

Marcello Pagano

**[JOTTER 1 A, WEEK ONE]**



Types of data

- Nominal Data

1 – male 2 – female	<b>Blood group:</b> • A – 1 • B – 2 • AB – 3 • O – 4
0 – alive 1 – dead	

**Nominally** numbers

- Name only
- No order
- Magnitude unimportant

In biostatistics and epidemiology we use numbers to tell our story. Numbers not only play a central role in our investigations, but they also allow us to use computers to do all the hard work, such as calculations, storage and graphics. Plus, the computers do all this without making any errors – we make the errors!

Often we do not need the full power of numbers for every application, although we still need limited properties, guided by our wish to accurately communicate with the computer. To make this point clear we classify our use of numbers into different class. For example, one kind of data is what we call a nominal data; when we label males as 0, females as 1, then that's nominal data. Another example of nominal data is if we use 0 to denote who's alive and 1 for denoting people who are dead. In both these examples, they are nominally numbers, just 0 or 1. The only property we're making use of the number system here is that 0 is different from 1.

We're not saying 1 is bigger than 0. We're not saying that 1 is one unit away from 0. Simply that 0 and 1 are different. This is the simplest example we have of nominal data. This is sometimes called binary data or dichotomous data, depending upon whether you prefer the Greek or the Latin root for two.

But it doesn't just have to have two values. For example, if we're looking at blood groups, here we would need four values: one each for blood groups A, B, AB and O.

Now, if you go to the Framingham Heart Study data set, you can take a look and see how many nominal data there are in that data set. So in summary, these are nominally numbers, only in name are they numbers. There's no order. The magnitude is unimportant. So that's nominal data.

### Types of data



- Nominal Data
- **Ordinal Data**

1. Mild
2. Moderate
3. Severe

Now, as we go up the ladder of complexity and properties of data, the next one up is ordinal data, where the order is important; for example, we might classify some disease as mild, moderate, or severe, where we might label mild as a 1, moderate as a 2, and severe a 3. We use the order of the data because 2 is a little bit more severe than 1, and 3 is a little bit more severe than 2. So the order is important. And this is called ordinal data.

### Recommended treatment strategy for schistosomiasis in preventive chemotherapy



Category	Prevalence among school-aged children	Action to be taken	
High-risk community	≥50% by parasitological methods (intestinal and urinary schistosomiasis) or ≥30% by questionnaire for visible haematuria (urinary schistosomiasis)	Treat all school-age children (enrolled and not enrolled) once a year	Also treat adults considered to be at risk (from special groups to entire communities living in endemic areas; see Annex 6 for details on special groups)
Moderate-risk community	≥10% but <50% by parasitological methods (intestinal and urinary schistosomiasis) or <30% by questionnaire for visible haematuria (urinary schistosomiasis)	Treat all school-age children (enrolled and not enrolled) once every 2 years	Also treat adults considered to be at risk (special risk groups only; see Annex 6 for details on special groups)
Low-risk community	<10% by parasitological methods (intestinal and urinary schistosomiasis)	Treat all school-age children (enrolled and not enrolled) twice during their primary schooling age (e.g. once on entry and once on exit)	Praziquantel should be available in dispensaries and clinics for treatment of suspected cases



Types of data

- Nominal Data
- **Ordinal** Data

ECOG performance status:

1. Mild 2. Moderate 3. Severe	0 Fully active 1 Ambulatory. Light work. 2 No work. Ambulatory > 50% 3 Ambulatory < 50% 4 Disabled
-------------------------------------	--

**Ordinal** data

- Order important
- Magnitude unimportant

Another example is provided by the Eastern Cooperative Oncology Group, which is a clinical trials group in cancer. In clinical trials they classify patients' performance status using a five-level classification system. It ranges from 0, where the patient is fully active, to 4, where the patient is disabled. And it progressively gets worse as one goes down the scale. So here the order is important. Thus ordinal data has more structure than nominal data.

So in summary, for ordinal data, the order is important, but the magnitude is unimportant. We're not saying here that the distance from 1 to 2 is the same as the distance from 3 to 4. The magnitude is not important.



**Rank data**

e.g. Ten leading causes of death in the USA —  
preliminary data for 2010  
(National Center for Health Statistics)

Rank	Cause	Number
1	Heart disease	599,413
2	Cancer	567,628
3	Chronic lower respiratory disease	137,353
4	Stroke	128,842
5	Accidents (unintentional injuries)	118,021
6	Alzheimer's disease	79,003
7	Diabetes	68,705
8	Influenza and Pneumonia	53,692
9	Nephritis group	48,935
10	Suicide	36,909

There is a special case of ordinal data that we use repeatedly. And that is called rank data. Rank data is sort of like when we just had the Olympics, the person who finishes first gets the gold medal. The person who finishes second gets the silver.

It doesn't matter how far behind the second is from the first. It's just that the second one finished second. So it could be a fraction of a second, to finish second, later than the first. Or it could be a few minutes. It doesn't matter. It's just the rank, the rank in which the data are ordered.

So here, for example, are the 10 leading causes of death in the US in 2010<sup>1</sup>. And we look and see that rank 1, or the highest or the most deaths, in this classification system were due to what was classified as heart disease. Number 2 was cancer. Number 3 was chronic lower respiratory disease, and so on.

We use the numbers only to order the data, thus the name, rank data. Sometimes, as we shall see later in the course when we come to what we call non-parametrics, we base our work on rank data and not the actual counts themselves that were used to obtain the ranks. It is amazing how much information just the ranks have and how useful they are..

#### Ten leading causes in 1993.



Rank	Cause	Number
1	Heart disease	739,860
2	Cancer	530,870
3	Stroke	149,740
4	Chronic lower respiratory disease	101,090
5	Accidents (unintentional injuries)	88,630
6	Influenza and Pneumonia	81,730
7	Diabetes	55,110
8	HIV infection	38,500
9	Suicide	31,230
10	Homicide and legal intervention	25,470

Sometimes the number actually confuses matters. So for example, if we look at the 10 leading

<sup>1</sup> [http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60\\_04.pdf](http://www.cdc.gov/nchs/data/nvsr/nvsr60/nvsr60_04.pdf)

causes of death in 1993, then the numbers are not at the same base level. Because there were more people in 2000 than there were in 1993. Maybe there were more deaths in 1993. We don't know because we only have the leading causes of death here, but they are: the 10 leading causes of death in 1993.



Ten leading causes in 1993.			
Rank	Cause	Number	Rank in 2010
1	Heart disease	739,860	1
2	Cancer	530,870	2
3	Stroke	149,740	4
4	Chronic lower respiratory disease	101,090	3
5	Accidents (unintentional injuries)	88,630	5
6	Influenza and Pneumonia	81,730	8
7	Diabetes	55,110	7
8	HIV infection	38,500	
9	Suicide	31,230	10
10	Homicide and legal intervention	25,470	

When we contrast 1993 and 2010, we see that the two causes ranked one and two have remained the same. Looking at rank 3 we see that stroke and chronic lower respiratory diseases have switched place. Rank 5 has remained the same, the unintentional injuries.

In 1993 we had HIV infection and homicide and legal intervention in the top 10. And neither of these have made the top 10 in 2010. So presumably, there has been some improvement in those two causes of death?

The point is we can't compare the number of deaths because the numbers refer to different bases, different groups of people. But we can refer to the ranks correctly, and we can make statements about how they vary over time. So rank data, is a special case of ordinal data.



## Types of data

- Nominal Data
- Ordinal Data
  - Rank Data
- Discrete Data
  - (Integer, Count data)

The next level up is discrete data, sometimes called integer data or count data. Discrete data is basically counting-- in mathematics we say you can put it into one-to-one correspondence with the integers.

Disease or injury	Deaths (millions)	Per cent of total deaths	Disease or injury	Deaths (millions)	Percent of total deaths
<b>World</b>					
1 Ischaemic heart disease	7.2	12.2	1 Lower respiratory infections	2.9	11.2
2 Cerebrovascular disease	5.7	9.7	2 Ischaemic heart disease	2.5	9.4
3 Lower respiratory infections	4.2	7.1	3 Diarrhoeal diseases	1.8	6.9
4 COPD	3.0	5.1	4 HIV/AIDS	1.5	5.7
5 Diarrhoeal diseases	2.2	3.7	5 Cerebrovascular disease	1.5	5.6
6 HIV/AIDS	2.0	3.5	6 COPD	0.9	3.6
7 Tuberculosis	1.5	2.5	7 Tuberculosis	0.9	3.5
8 Trachea, bronchus, lung cancers	1.3	2.3	8 Neonatal infections <sup>b</sup>	0.9	3.4
9 Road traffic accidents	1.3	2.2	9 Malaria	0.9	3.3
10 Prematurity and low birth weight	1.2	2.0	10 Prematurity and low birth weight	0.8	3.2
<b>Middle-income countries</b>					
1 Cerebrovascular disease	3.5	14.2	1 Ischaemic heart disease	1.3	16.3
2 Ischaemic heart disease	3.4	13.9	2 Cerebrovascular disease	0.8	9.3
3 COPD	1.8	7.4	3 Trachea, bronchus, lung cancers	0.5	5.9
4 Lower respiratory infections	0.9	3.8	4 Lower respiratory infections	0.3	3.8
5 Trachea, bronchus, lung cancers	0.7	2.9	5 COPD	0.3	3.5
6 Road traffic accidents	0.7	2.8	6 Alzheimer and other dementias	0.3	3.4
7 Hypertensive heart disease	0.6	2.5	7 Colon and rectum cancers	0.3	3.3
8 Stomach cancer	0.5	2.2	8 Diabetes mellitus	0.2	2.8
9 Tuberculosis	0.5	2.2	9 Breast cancer	0.2	2.0
10 Diabetes mellitus	0.5	2.1	10 Stomach cancer	0.1	1.8



For example, in the top left-hand corner are the number of deaths in millions from around the world as reported by the WHO<sup>2</sup>. We see that there were 7.2 million deaths from ischaemic heart disease and 5.7 million from cerebrovascular disease, and so on.

They also classify, each country into one of three groups, either the low-income group (the top, right-hand corner), the middle-income group (the bottom, left-hand corner), or the high-income group (the bottom, right-hand corner). The classification cutoffs were: low, income less than \$825; middle, income between \$825 and \$10,066; high, income more than \$10,066 per annum. You can see that as average income changes, the leading causes of death change.

If we're worried about the size of the groups, we can standardize by the population sizes by looking at the percentages in the last column. Then it makes sense to compare the different countries. We discuss standardization, below.

Another observation we can make is that the top left-hand corner chart is the whole world Now, we can look at these numbers. And this one here-- the top left-hand corner, the world-- is actually an average of the other three. And we'll talk about averages soon.

And what the average tells us, in some sense, should be the arithmetic average of these three panels. But it's not that. And it's a little bit more subtle than that. And it's actually a weighted average. But we'll get to that later this week.

But from this you can see that, for example the number 1 over here, ischaemic heart disease, it's not a 1 here, and it's not a 1 here-- it's number 2 here-- but it's a 1 here. So somehow what's happening in the high-income countries influences what happens as the overall, for the world average. So averages are fascinating things to deal with. And we'll attack that shortly.

---

<sup>2</sup> [http://www.who.int/healthinfo/global\\_burden\\_disease/2004\\_report\\_update/en/index.html](http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/index.html) Table 2 of Global Burden of Disease 2004 Update, WHO



## Types of data

- Nominal Data
- Ordinal Data
  - Rank Data
- Discrete Data
- Continuous Data

The last data category is continuous data. And this is what we usually think of as numbers. Between any two bounds, for example, any value is theoretically achievable.



### CONTINUOUS DATA

Between two bounds any value is (*theoretically*) achievable.

Examples: time  
length  
mass  
temperature . . . etc

Note:

All *measurements* we observe are discrete, *but* there is an advantage to *modeling* them as continuous.

*Digital* computers, as the name implies, only handle discrete numbers.

So for example, if you think of time, time is it is one that we typically think we can measure to any accuracy we want. And that makes it continuous. Or length, or mass, or temperature, et cetera, all of these are examples of continuous data. Now, we can get philosophical here and argue that all measurements-- and that's what we're really interested in; what can we measure? In reality, all measurements are discrete.

There is an advantage to modeling data as continuous, especially if we are going to use digital computers, which we do and will. Digital computers, as the name implies, only handle discrete numbers. They do not handle continuous numbers. For example, there are only a finite number of numbers on a computer. There is a largest number. We can talk about the smallest positive number, et cetera, something we cannot ordinarily talk about with continuous numbers.



## Types of data

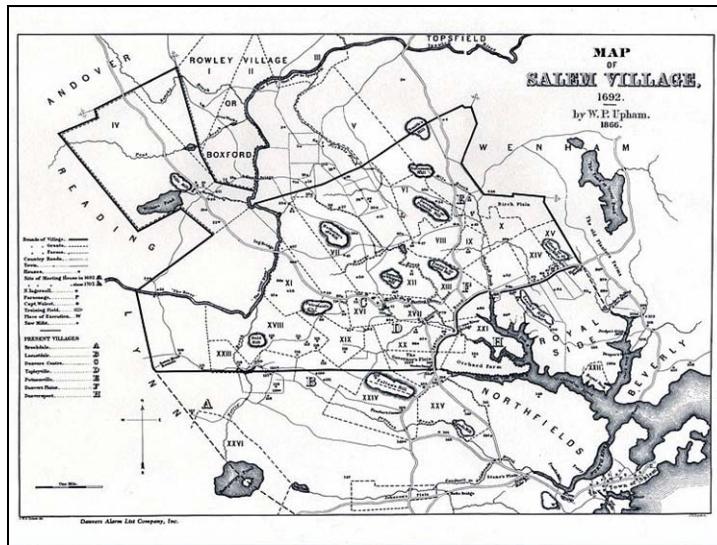
- Nominal Data
- Ordinal Data
  - Rank Data
- Discrete Data
- Continuous Data

So in summary, these are the kinds of data that we looked at. So this is our taxonomy. It's basically into four groups-- nominal data, ordinal data, discrete data, and continuous data. And the reason why we go into this much detail is because we are going to use different statistical methods when we have nominal data than we do when we have ordinal data than we do when we have discrete data and then we do when we have continuous data.



Sometimes, half a dozen figures will reveal, as with a lightning-flash, the importance of a subject which ten thousand labored words, with the same purpose in view, had left at last but dim and uncertain.

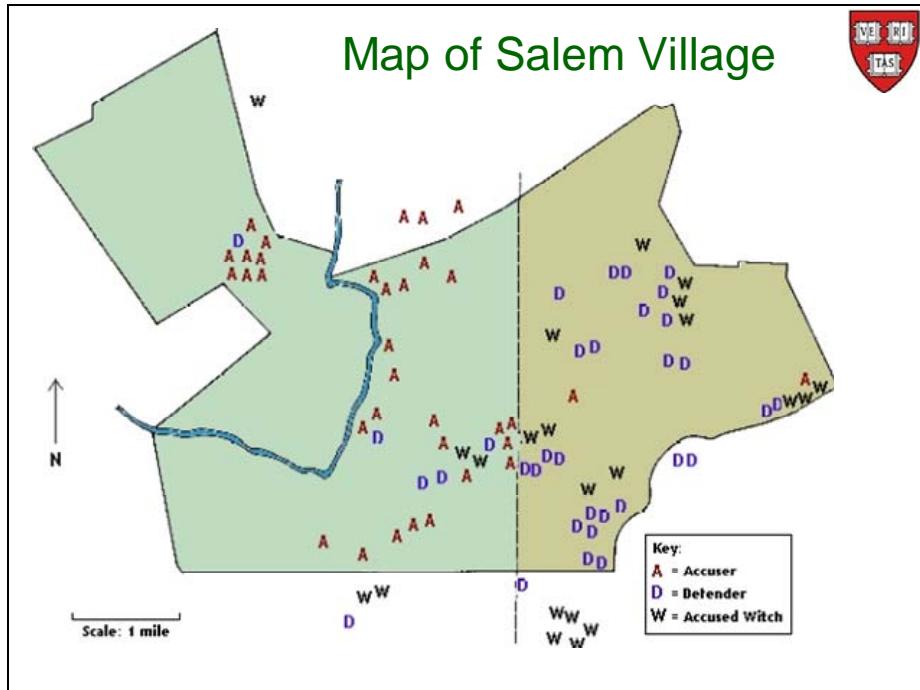
*Mark Twain — Life on the Mississippi, 1883*



One of the advantages of the computer is that it enables us to draw pictures. And a picture sometimes “is worth a thousand words”. At times a picture can reveal things that might not be obvious otherwise. Here's an example, a map of Salem Village.<sup>3</sup> Now, every year at Halloween, we all drive up to Salem Village and everybody celebrates the witches.<sup>4</sup> And it's a rather nasty custom considering that what we're celebrating is that some women got burned at the stake. Some got squeezed to death, et cetera.

<sup>3</sup> Benjamin C. Ray , The Geography of Witchcraft Accusations in 1692 Salem Village, *William and Mary Quarterly*, 3d Series, Volume LXV, Number 3, July 2008.

<sup>4</sup> See <http://www.smithsonianmag.com/history-archaeology/brief-salem.html> for an introduction to this topic.



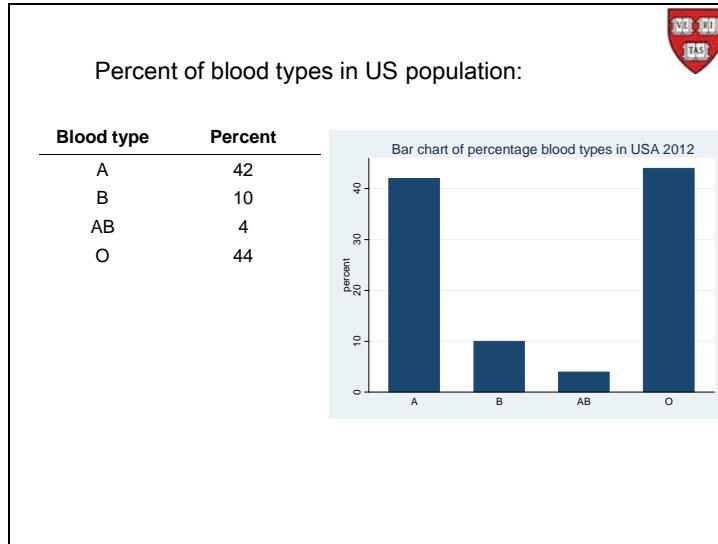
What we're told is that these women were possessed. This was in 1692. A historian decided to map out the addresses of the individuals accused of being possessed, the accusers and the defenders.<sup>5</sup> These are denoted by W, D and A, respectively in the map above.

And if you look, all the accusers tend to come from the left of the vertical line whereas the accused and their defenders are largely on the right of the vertical line. The classification is not perfect, but the point is that the graph reveals a strong geographic pattern that requires some explanation.

This makes us want to take advantage of the graphical options available on the computer.

---

<sup>5</sup> Benjamin C. Ray , The Geography of Witchcraft Accusations in 1692 Salem Village, *William and Mary Quarterly*, 3d Series, Volume LXV, Number 3, July 2008.



The first graphical device we look at is the bar chart. Let's go back to a categorical variable we spoke about: blood type, that takes on four values, A, B, AB, and O. If we are interested in the distribution amongst these four values we have the numbers from the Red Cross<sup>6</sup> who tell us that 42% of us in the US have blood type A, 10% of us have got blood type B, 4% type AB, and 44% type O. How can we display this graphically? What we can do is we can draw a bar chart. And here's a bar chart. The height of the bars are proportional to the percentage of the population with that blood type.

If we categorize these individuals into ethnic groups we get:

Harvard University Logo

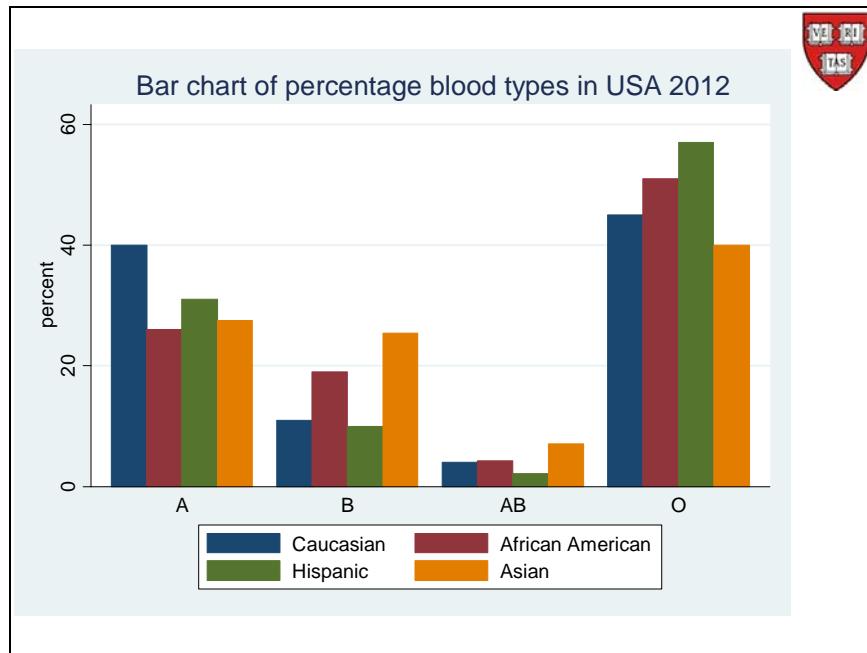
Ethnic group blood type breakdown according to the Red Cross

Type	Caucasian	African American	Hispanic	Asian
<b>A</b>	40	26	31	27.5
<b>B</b>	11	19	10	25.4
<b>AB</b>	4	4.3	2.2	7.1
<b>O</b>	45	51	57	40

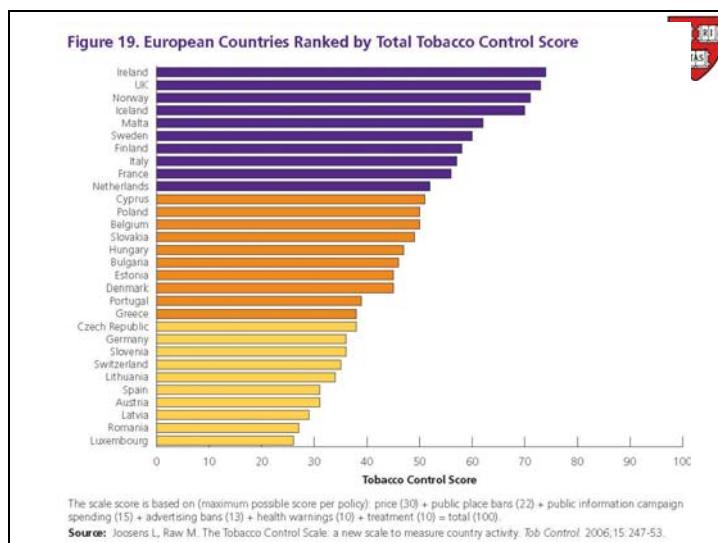
<http://www.redcrossblood.org/learn-about-blood/blood-types>

<sup>6</sup> <http://www.redcrossblood.org/learn-about-blood/blood-types>

Now, we look at this table and we could study this and get some understanding of how these distributions vary. Alternatively, if we draw the bar charts, put them side by side, we get an immediate picture of what's going on.



The first thing that hits us is, possibly, that there is an excess here of blood type B for Asian Americans, for example, and we notice disparate distributions in the different groups.



Here is a bar chart displaying how European countries were ranked by their total tobacco control score. The score is one that adds up to 100. And each country is judged by how much it is doing in its attempt to achieve tobacco control. The top country is Ireland. The next country down is UK, and then Norway, and so on. And we can see how this varies amongst the European countries.

### Barchart for Continuous Data



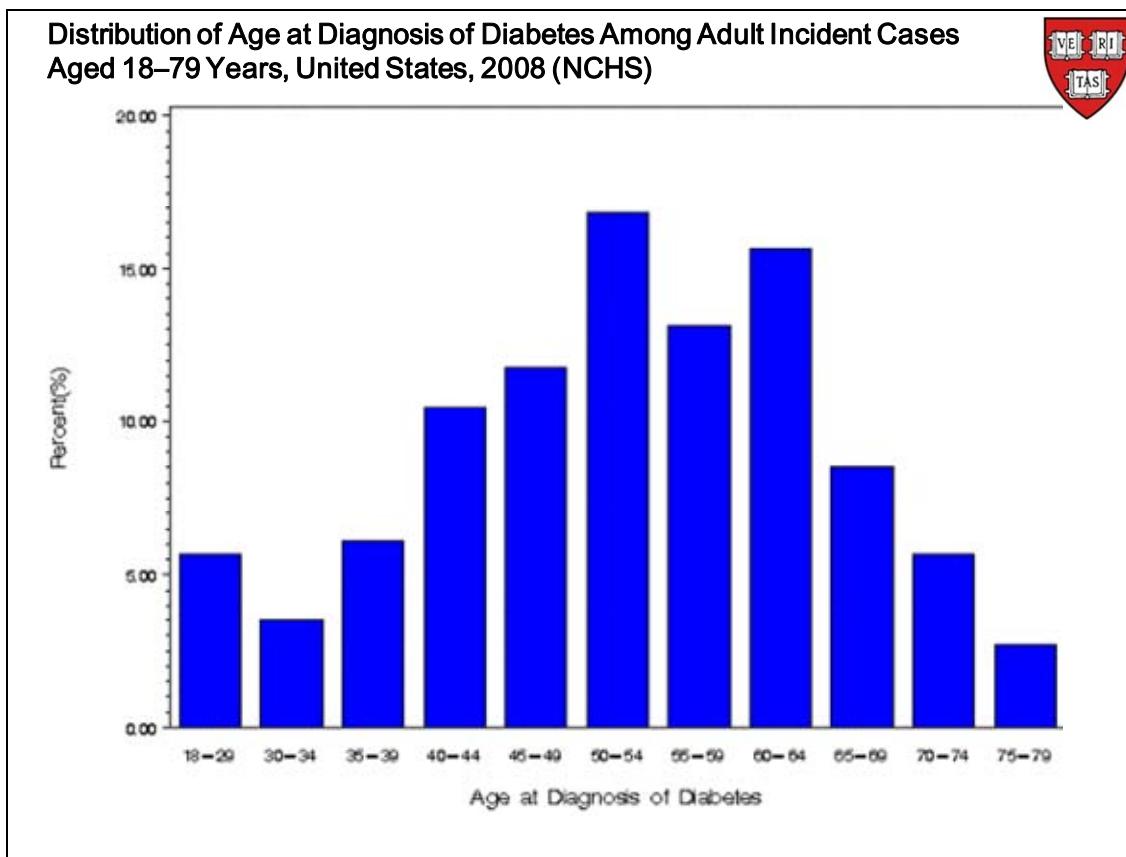
Age (Years)	Percent
18–29	5.7
30–34	3.5
35–39	6.1
40–44	10.5
45–49	11.8
50–54	16.8
55–59	13.1
60–64	15.6
65–69	8.5
70–74	5.6
75–79	2.7

We can also create a bar chart, from a continuous variable. So here is an example of the distribution of age of diagnosis of diabetes amongst incident cases<sup>7</sup> -- that means, who were diagnosed in the last year prior to tabulation, the United States in 2008. They created cells to report the data. For example, one cell is for 75 to 79 year olds, and there were 2.7% of the individuals in that cell. In the 70 to 74, it was 5.6% and so on. So the cells, now, play the role of

<sup>7</sup> Distribution of Age at Diagnosis of Diabetes Among Adult Incident Cases Aged 18–79 Years, United States, 2008  
<http://www.cdc.gov/diabetes/statistics/age/fig1.htm>

the label that we had for the categorical data-- like the blood types. But now these cells are contiguous.

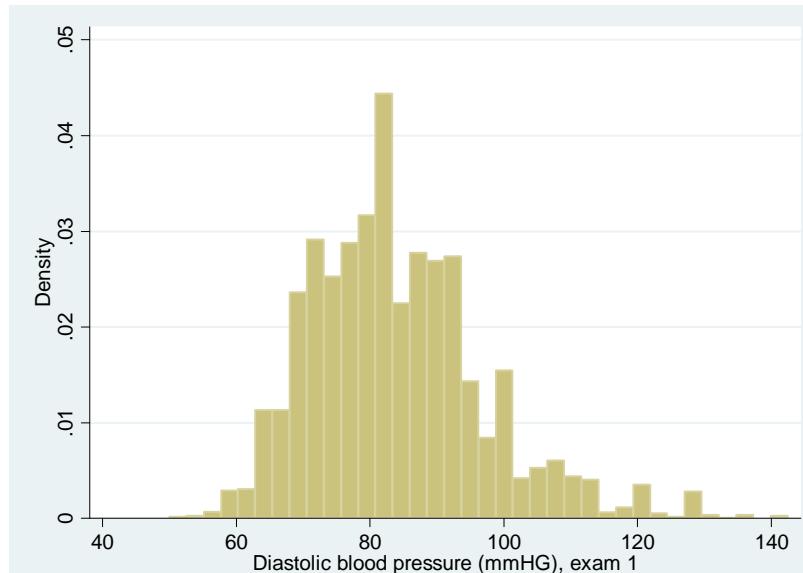
Note that the first cell is of width 12 years, whereas all the others are of width 5 years.



Looking at the distribution of the data, we have this almost bell-shaped distribution. But it is not quite right to come to this conclusion because the first cell is more than twice as wide as the others. So because of that design flaw, we lose the ability to fully evaluate the shape of the distribution.

Each of these cells is contiguous to its neighbors, so you should put them closer together, and there should be no space between the bars. When you do that, you get what is called a histogram.

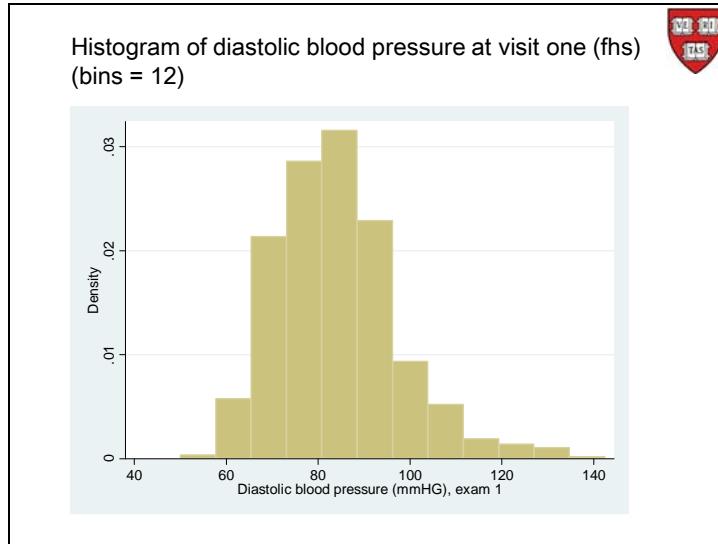
Histogram of diastolic blood pressure at visit one (fhs)



Here is a histogram of the diastolic blood pressure at visit one for our Framingham heart study. We can get an idea of how these data are distributed. There are a few down on the left. The mass of the data is in the middle, say between 60 and 120, but there are quite a number on the right.

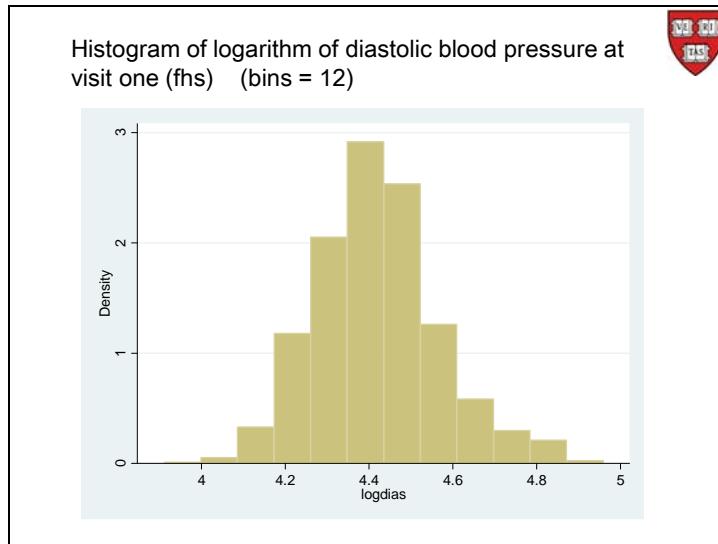
Now I just press the default buttons in Stata, and they chose the number of bins. To see the impact of the choice of the number of bins, try to see what happens on a famous data set,

<http://www.stat.sc.edu/~west/javahtml/Histogram.html>



We could have chosen fewer bins. For example, had we chosen 12 bins, we get a slightly different picture. And this is one of the problems with a histogram. When you are reading the literature and you see a histogram, how many cells were chosen, how wide are these cells and how much of an impact do these choices have on the kind of picture you get.

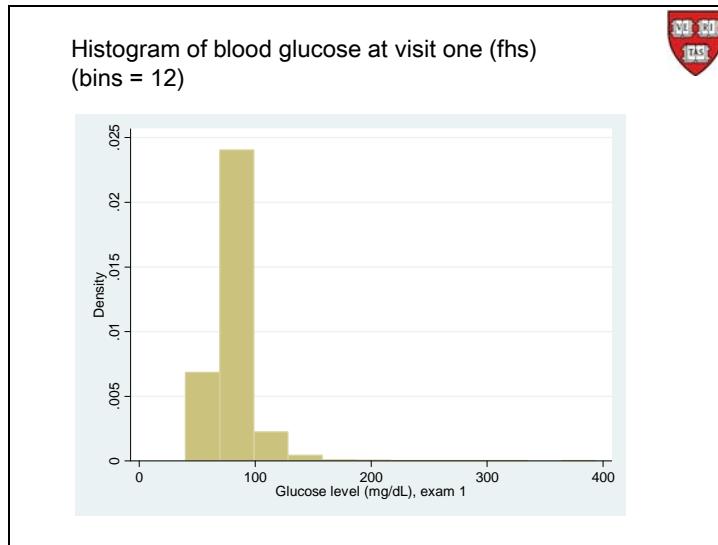
Putting that consideration aside for the moment, what we have here is a picture that almost looks symmetric about a central axis, except for an elongation on the right; what we call a long tail on the right. This kind of behavior is typical of data we shall see later in the course when we look at clinical trials data. When dealing with survival data, there is this very often a very long right tail.



We love symmetry in statistics, because it makes things a little bit easier to explain and also makes our techniques more powerful. And one thing we can do to our data to help us get closer to symmetry is to transform the data. So, instead of looking at the raw diastolic blood pressure--we can look at the logarithm of those numbers. So for each person define a new variable, which is related to the old one by just taking its logarithm, and what here is the histogram that is closer to being symmetric, since we pulled in the tail.

This is a common technique used in statistics: we do not just look at the raw data, but sometimes we transform the data, and here is one possible transform, namely, the logarithm.

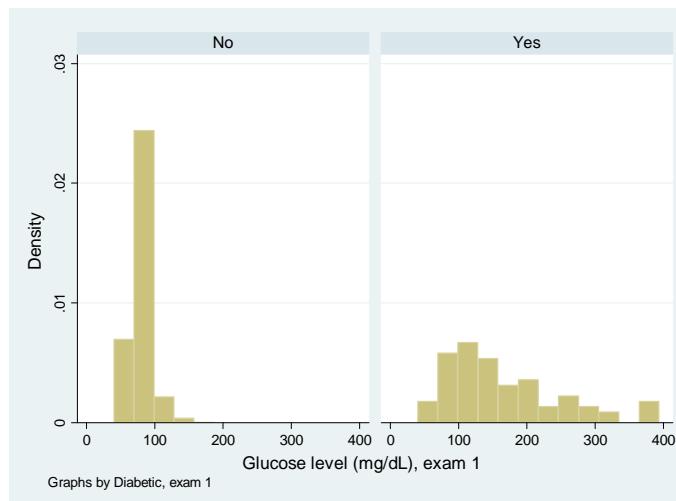
We need to be careful because by pulling the tail in we may lose an important aspect of the story. A case in point:



Here we have a histogram of the blood glucose at visit one in our Framingham heart study. And we see a right tail which is very much smaller, but it extends all the way to the right, almost to 400. Now the question is, what is happening at this right tail? Remember we are looking at the level of blood glucose. So what might be causing this right tail is the presence of diabetics.



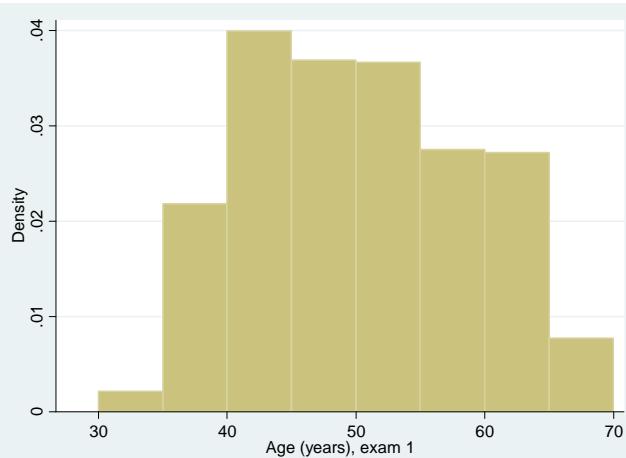
Histogram of blood glucose at visit one (fhs) with diabetics, at time 1, on right, and rest on left (bins = 12)



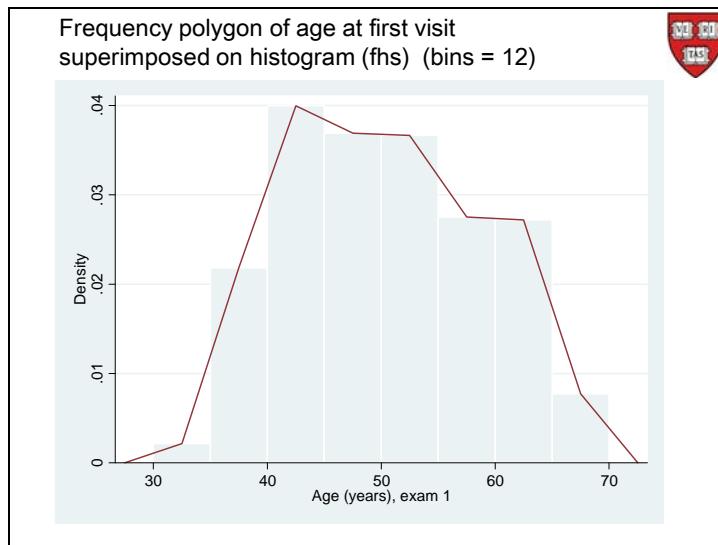
On the right we have the histogram for diabetics and on the left the histogram for the rest of the group. Now the highest value in the left group is 163, and there is a large amount of activity for the diabetic group on the right. This is now a very interesting part of the story, namely, what happens to diabetics. So be careful when you do the transformations. You might be hiding something that you don't necessarily want to hide.



Histogram of age at visit one (fhs)  
(bins = 12)

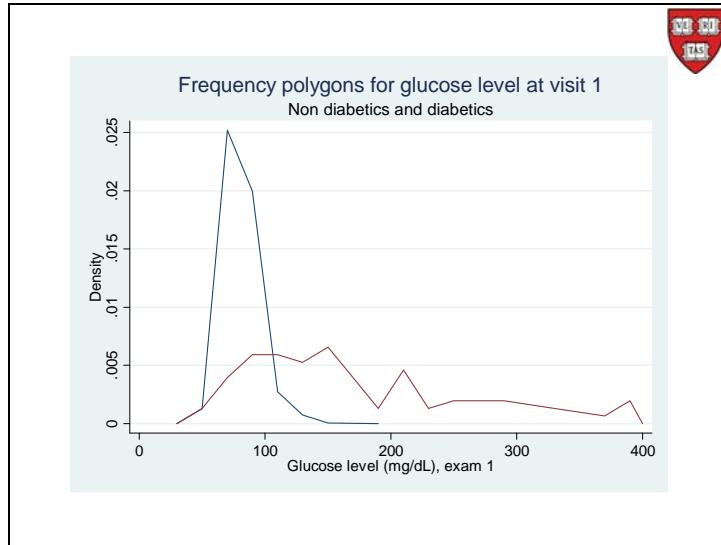
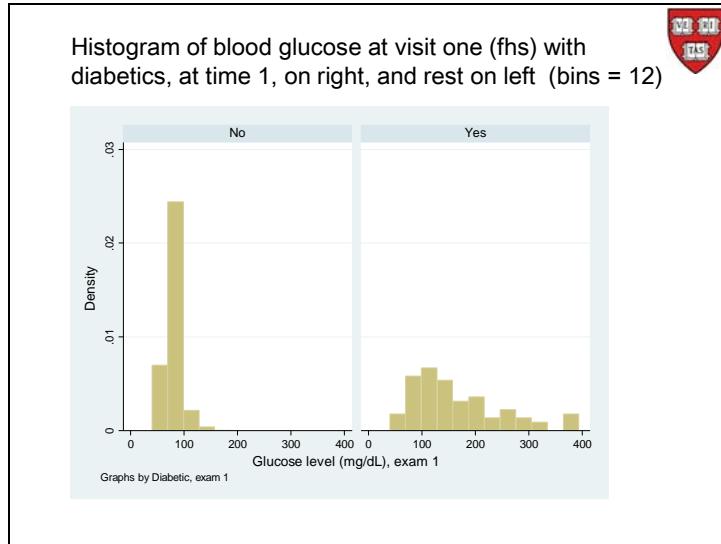


Here is another histogram, the histogram of the age of the participants at visit one.



To obtain a smoother view of the distribution of the data, one can draw what is called the frequency polygon. What the frequency polygon is, you take the midpoint of each one of these--take the midpoint of each one of these. And to maintain the area, you also anchor these at 0--an equal distance away, at either end--and then you join these points. And now we have exactly the same information as we had with the histogram, including maintaining the same area under both curves. Indeed, if we call the area equal to 1, then the area under the curve between any two horizontal points is the proportion of patients between those points.

This is called the frequency polygon. We see that the age is very concentrated between thirty five and seventy.



Returning to our histograms of glucose level at visit one for the diabetics and others, we can draw the frequency polygons for both and in fact superimpose them on each other. This is not as easy to do with histograms on top of each other. This makes quite clear the difference in the distribution of blood glucose levels between these two groups.

We return to frequency polygons especially when we look at models later in the course and we idealize the distributions to represent “infinite” populations.



**Cumulative Distribution of Age at Diagnosis of Diabetes  
Among Adult Incident Cases Aged 18–79 Years,  
United States, 2008**

Age (Years)	Percent	Cumulative Percent
18–29	5.7	5.7
30–34	3.5	9.2
35–39	6.1	15.3
40–44	10.5	25.8
45–49	11.8	37.6
50–54	16.8	54.4
55–59	13.1	67.5
60–64	15.6	83.1
65–69	8.5	91.6
70–74	5.6	97.2
75–79	2.7	99.9

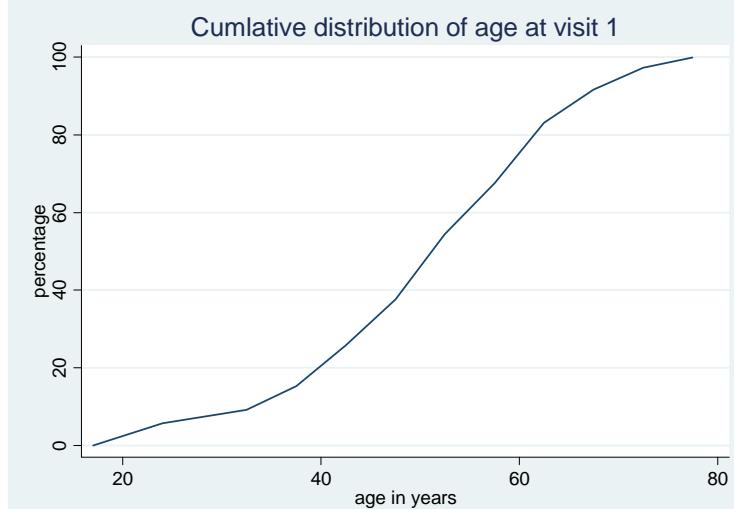
<sup>8</sup>

Before leaving this topic, let us look at one more way of displaying the distribution of the data. Return to the distribution of age at diagnosis. What we have is what percentages fall into each cell. We can also take the running sum of individuals. So, going down the table, we have 5.7% in the first cell and 3.5% in the second cell. So that means we have  $(5.7+3.5=)$  9.2% total in the first two cells. So 9.2% are 34 years or younger. Adding the third cell's 6.1% means that 15.3% are thus 39 years or younger. And so on, building up the third column in the table.

The third column is called the cumulative distribution--we accumulate the sum as we go down.

---

<sup>8</sup> <http://www.cdc.gov/diabetes/statistics/age/fig1.htm>



We can draw this third column, and we have the cumulative distribution function.

We can easily read summary statistics from this curve. For example, if we are interested in demarcating 50% of the people, we extend from the 50% point on the left axis to the curve and then down to the horizontal. Or if we're interested in 25%, or 75% of the population, we can read those values too from this curve.

Sometimes we look at 1 minus the cumulative distribution function and that is sometimes called the survival curve, and that is what we will be studying next week.



In summary, we can show the distribution of the variables with a bar chart, if the variable is nominal or ordinal (categorical), and a histogram otherwise.

A frequency polygon conveys the same information as a histogram. It also lends itself to the depiction of the cumulative distribution.

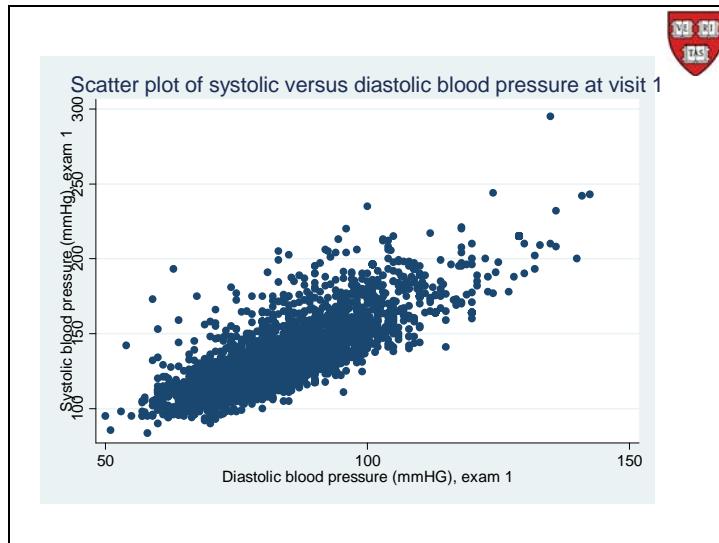
## Scatter plot



A scatter plot is a simple x-y plot where on the x-axis is one variable and on the y-axis the other and all couplets (x,y) of data are plotted.

So the next topic I'd like to talk about is the scatter plot and line plot. These are two more graphical devices and these are meant to show relationship between two variables.

A scatter plot is a simple xy plot where, on the x-axis is one variable and on the y-axis, the other variable. For example, if we measure on each patient, an x and a y, then plot the couplet (x,y) for each patient, and that is a scatter plot.



So for example, here's a scatter plot of systolic versus diastolic blood pressure at visit one. Not surprisingly at all, we see a pattern emerging that basically, these points fall into an ellipse that is pointed up and that the fit in the ellipse is rather tight. What this says to us is that typically a high diastolic blood pressure, goes hand in hand, with a high systolic blood pressure, and a low diastolic blood pressure is associated with a low systolic blood pressure..

This is called a scatter plot. And you're going to see scatter plots repeatedly because they are very useful at revealing patterns or associations.

Line plot

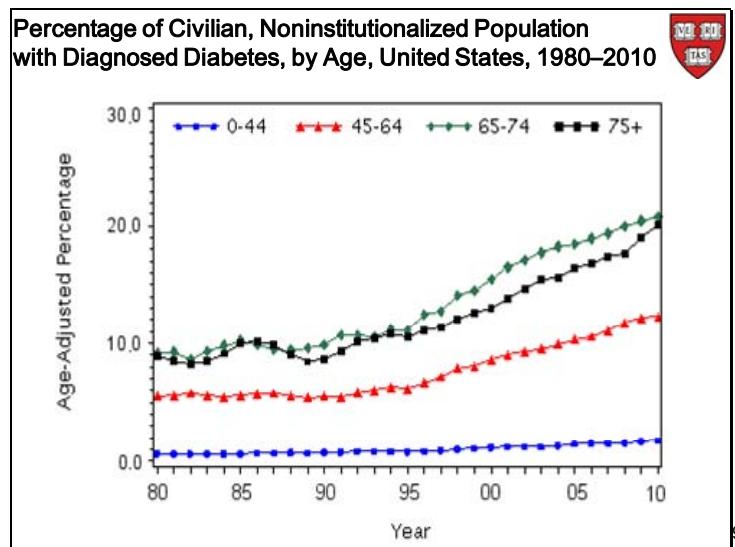
**Percentage of Civilian, Noninstitutionalized Population with Diagnosed Diabetes, by Age, United States, 1980–2010**



Year	Age				Year	Age			
	0–44	45–64	65–74	75+		0–44	45–64	65–74	75+
1980	0.6	5.5	9.1	8.9	1996	0.8	6.6	12.5	11.1
1981	0.6	5.6	9.2	8.4	1997	0.9	7.1	12.8	11.3
1982	0.6	5.8	8.6	8.3	1998	1.0	7.8	14.0	12.1
1983	0.6	5.6	9.3	8.5	1999	1.1	8.1	14.5	12.6
1984	0.6	5.4	9.8	9.1	2000	1.2	8.6	15.4	13.0
1985	0.6	5.6	10.2	10.0	2001	1.2	9.0	16.5	13.9
1986	0.7	5.7	9.9	10.1	2002	1.2	9.3	17.1	14.6
1987	0.7	5.8	9.5	9.8	2003	1.2	9.5	17.7	15.4
1988	0.7	5.6	9.4	9.0	2004	1.3	9.9	18.2	15.6
1989	0.7	5.4	9.6	8.4	2005	1.5	10.3	18.4	16.4
1990	0.7	5.5	9.9	8.6	2006	1.5	10.5	18.9	16.8
1991	0.8	5.4	10.7	9.3	2007	1.5	11.0	19.3	17.3
1992	0.8	5.8	10.6	10.1	2008	1.6	11.7	19.9	17.7
1993	0.8	6.0	10.5	10.4	2009	1.7	12.2	20.4	19.0
1994	0.8	6.3	11.1	10.8	2010	1.8	12.3	20.7	20.1
1995	0.8	6.2	11.1	10.6					

Here is a table displaying the distribution of diagnosis with diabetes from 1980 through 2010. We see that the percentage diagnosed increased by 200% (from 0.6% to 1.8%) for those aged 0–44 years, 124% (from 5.5% to 12.3%) for those aged 45–64 years, 127% (9.1% to 20.7%) for those aged 65–74 years, and 126% (8.9% to 20.1%) for those aged 75 years and older. In general, throughout the time period, the percentage of people with diagnosed diabetes increased among all age groups. In 2010, the percentage of diagnosed diabetes among people aged 65–74 (20.7%) was more than 11 times that of people younger than 45 years of age (1.8%).

So there's a lot of information here. There is a lot of information in this table. As a general trend we can see that these percentages are all going up with time, but it is difficult to see different speeds of increase in the different groups and any of the interrelationships that may exist.



Whereas, if we draw a picture using line plots, we can appreciate many more subtleties in the movements. For example, we can see that the two older groups, the 65 to 74 and the 75 up, overlap each other, they are roughly equal to each other. And they both go up, first, at a gentle pace, and then, from about 1995 on, for some reason, at a more accelerated pace.

The same is true for the middle age group, the 45 to 64 age group, the gentle increase until about 1995 and then a faster pace of increase.

The younger group, those who are zero to 44, have a much more gentle, almost flat, increase, but a little bit of an increase there, nonetheless. So it might be of interest to find out what happened in 1995, or thereabouts, that caused this increase in speed at which things happen.

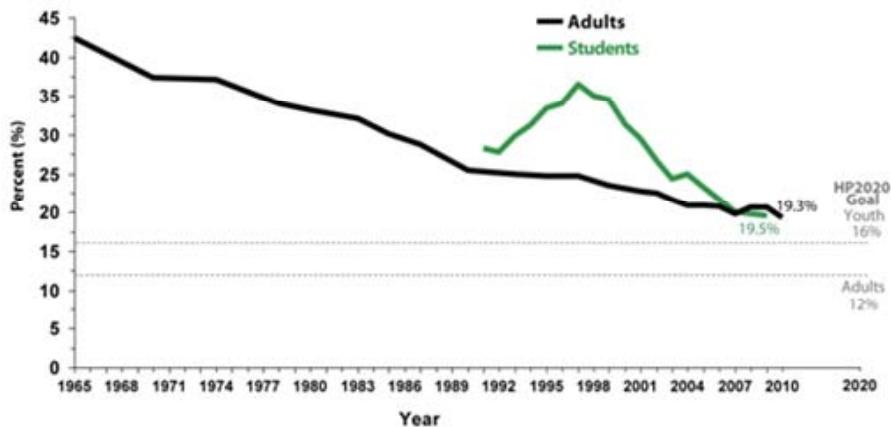
These percentages are really age adjusted percentages and not pure percentages. What does this mean? Rest assured it makes no difference in making the point that line plots are useful, but for the real meaning of what is going on here you will have to hang on to this course to find out what age adjustment means.

---

<sup>9</sup> <http://www.cdc.gov/diabetes/statistics/prev/national/figbyage.htm>



### Trends in Current Cigarette Smoking by High School Students\* and Adults\*\*—United States, 1965-2010



\*Percentage of high school students who smoked cigarettes on 1 or more of the 30 days preceding the survey (Youth Risk Behavior Survey, 1991-2009).

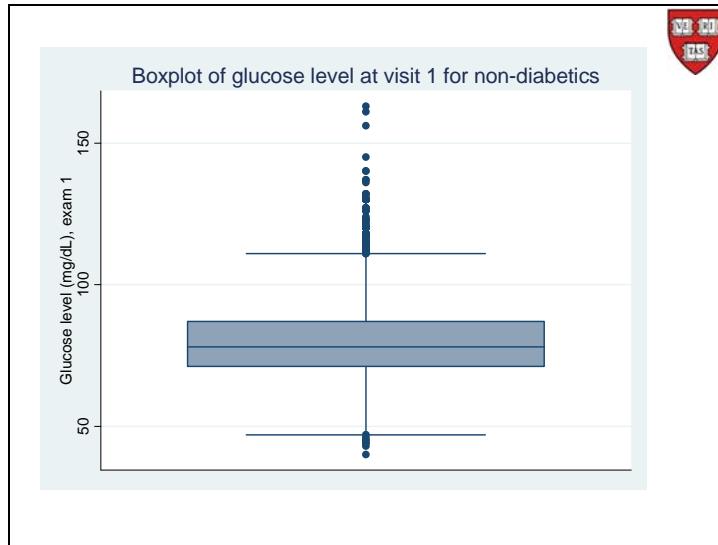
\*\*Percentage of adults who are current cigarette smokers (National Health Interview Survey, 1965-2010).

10

Here is another very interesting line graph, that shows what is happening with the trends in cigarette smoking for high school students and adults. Now the line for the high school students, does not go as far back as the line for adults, but that does not adversely affect our ability to show them both on the same graph..

<sup>10</sup> [http://www.cdc.gov/tobacco/data\\_statistics/tables/trends/cig\\_smoking/index.htm](http://www.cdc.gov/tobacco/data_statistics/tables/trends/cig_smoking/index.htm)

## Box plot



Here's an example of a boxplot. A box plot is a graphical way of summarizing the distribution of our data.

The bottom of the box is placed so that 25% of the data lies below the bottom of the box--the first quartile

The middle line in the box is the middle of the data, the median. So while 50% percent lies below, 50% is also going to lie above--the median.

And then the top of the box is where 75% of the data lies beneath the box--the third quartile.

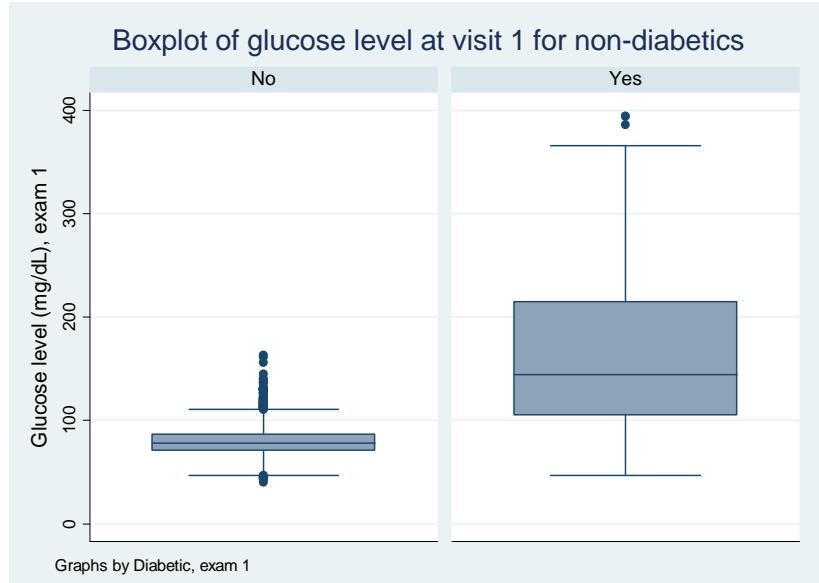
The distance between the top and the bottom of the box is called the interquartile range.

The whiskers are drawn out to try and get something like 98-some percent of the data.

And then everything outside of that, are called outliers. They lie outside these lines. Don't forget that we're basing this on something like 4,400 observations. So there should be quite a few out there.

We refer you to the Stata manual to see precisely how the box plot is drawn.

The boxplot helps us see whether the distribution is symmetric; for example, whether the median, is equidistant from each edge. Also see whether the lengths of the vertical lines are the same, and whether there are just as many outliers in the top as in the bottom.

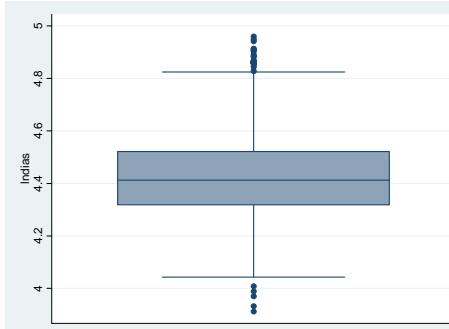
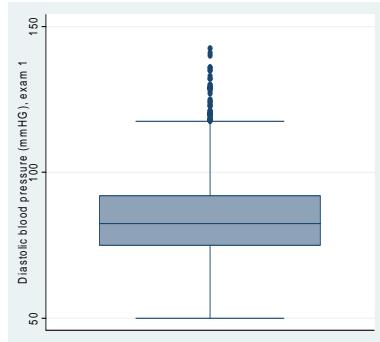


Here is another example, this time of side by side boxplots of two groups, the diabetics on the right and the rest on the left, again at visit 1, and looking at blood glucose level again.

We see that the distribution for the non-diabetics is very tight. The interquartile range is very, very small, with a few outliers at the top (high blood glucose levels) who might be pre-diabetics.

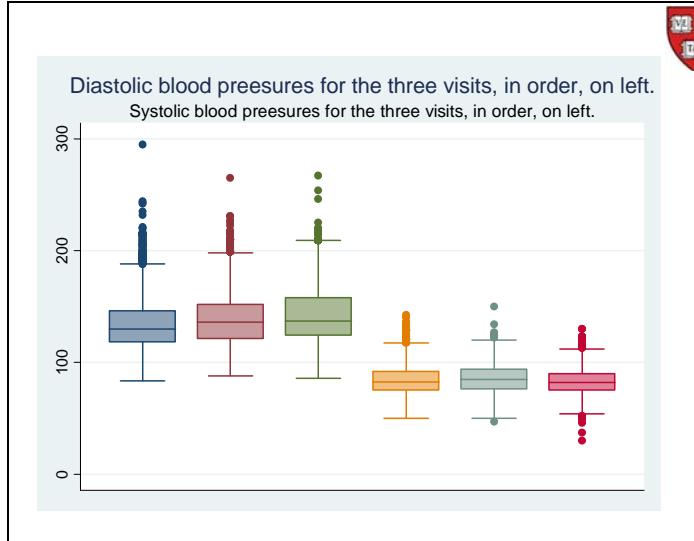
Those on the left have much better control of the blood glucose level than the diabetics on the right, not surprisingly. Plus we see a much longer tail for diabetics.

So we can get lots of information just from these five-number summaries.



Boxplots of diastolic blood pressure at visit1 on the left and for the logarithm of same on the right

Here is another side by side comparison. This time on the left is the diastolic blood pressure whereas on the right is the logarithm of the diastolic blood pressure. Recall that we looked at the logarithm transform to achieve symmetry. Do you think we were successful?



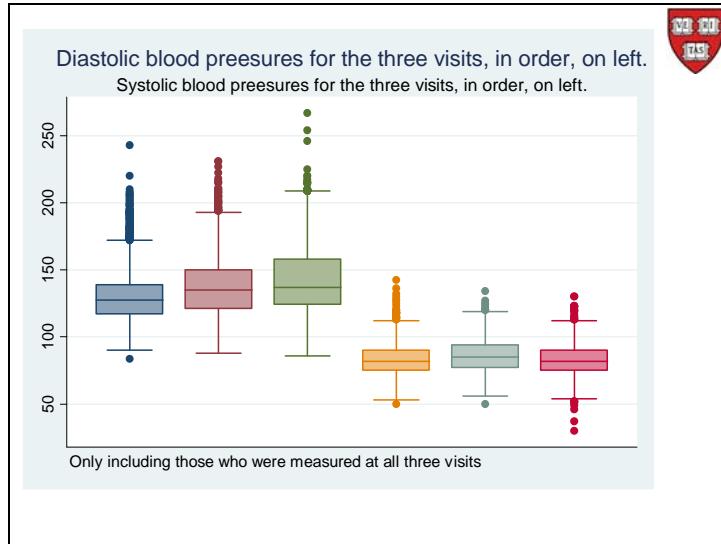
Box plots are also useful to place side by side over time. Here, the three on the left are the distributions of the systolic blood pressures for the three visits, and on the right the diastolic blood pressures for the same three visits.

And it looks like the median is increasing from visit one to visit two to visit three. Whereas the diastolic went up a little bit, but then at the third visit, it sort of goes down. So you can detect these kinds of movements by looking at these box plots side by side.

Now, be very careful when you do these kinds of comparisons over time. These are called longitudinal. So be careful when making longitudinal comparisons. The question you need to ask is, are we comparing the same people?

Is it the same people who showed up at the first visit as show up in the second visit? And the answer is no, they're not the same people. Is it the same people who then show up at visit three? No. Just by looking at the numbers who show up (not shown in the box plot) we know that different numbers show up at each visit, so it must be differing individuals.

Possibly the fact that these distributions go down between visit two and visit three might be related to the people who left the study. Whether they left because they were ill or they passed away or for whatever reason, we need to be careful before drawing conclusions about how the summaries behave..



For example, what we did here is redraw the above boxplots, but this time, only include those persons who were measured at all three visits. It looks like we are getting the same message when we include just those at all three visits, as we got before. This is comforting. This behavior is not peculiar to box plots, but I wanted to take this opportunity to warn you that when you make some comparisons over time make sure that you are comparing apples to apples. We return to this very important topic in the near future.

Marcello Pagano

### **[JOTTER 1B, WEEK ONE SUMMARIES]**

Week one deals with data types, graphics and summary statistics.

We'll be using the sigma, or summation notation quite often. Here is an example of what I mean:  
Suppose we have these 13 numbers



2.3, 2.15, 3.50, 2.60, 2.75, 2.82, 4.05,  
2.25, 2.68, 3.00, 4.02, 2.85.

Let me label them  $x_1, x_2, \dots, x_{13}$ . Now denote the operation of taking their sum, which is 38.35, by

$$\text{Sum} = \sum_{i=1}^n x_i$$

and set  $n=13$  in the formula.

[http://en.wikipedia.org/wiki/Summation#Capital-sigma\\_notation](http://en.wikipedia.org/wiki/Summation#Capital-sigma_notation)

We shall use the sigma or summation notation. Suppose we have these 13 numbers here: 2.3, 2.15, 3.50,..., 2.85. And let me label them  $x_1, x_2, \dots, x_{13}$ . We use the subscripts to denote the individual numbers. What we want to do is take their sum, which is 38.35.

We denote this operation by using the sigma notation. That is sigma, capital S in Greek, and we're going to have i, running from 1 to 13. If you are not familiar with the sigma notation, pause here and find some source where you can read up about this, for example, if you go to Wikipedia and look up the capital sigma notation, you will get a nice review of what this is.

We can now formally define the mean by dividing  
The sum of the numbers by however many of  
them there are:

$$\text{Mean} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

$$\text{Sum} = \sum_{i=1}^{13} x_i = 38.35$$

$$\text{Mean} = \bar{x} = \frac{38.35}{13} = 2.95$$

The mean is a measure of central tendency.

We can now formally define the mean by dividing the sum of the numbers by however many of them there are.

And the mean is a measure of central tendency, it tells us where the center of these number is. So going back to our example, the sum was 38.35. There were 13 numbers. And so if we divide by 13, we get that the mean is 2.95, and that's roughly the middle of where the numbers are.

Variable	Obs	Mean
age1	4434	49.8392
sex1	4434	1.548489
sysbp1	4434	134.3718
diabp1	4434	84.09303
cursmoke1	4434	.4772215
cigpd1	4415	8.174858
bmi1	4417	25.99249
diabetes1	4434	.0252594
heartrtel	4433	75.34446
glucose1	4047	81.61848

Now, let's take a look at some of the variables in our data set. Here are the first 10 of them. So at the first visit, let us look at the age. There were 4,434 people and their age had a mean of 49.83, or roughly 50.

The systolic blood pressure has mean of 134. The diastolic blood pressure averaged out at 84. Current smokers, now, here the current smokers was yes/no, was a 0/1 variable. So this tells us that roughly 48 percent of the people in the study had the value one. The other 52 percent had the value zero. So if one takes the mean of "smoker", that means that there were 48 percent smokers.

Same thing with sex, because sex is also a dichotomous variable, except you remember sex took on the values one and two. So what this says is that 55 percent of the sex variables at visit one had the value two. So two meant female, roughly 55 percent of this data set was female.

Cigarettes per day averaged at eight. Now be careful with this one. We are coming back to look at this one later, but there were a lot of people who weren't smokers. We have that 52 percent are nonsmokers. So what does this average actually tell us here?

BMI averaged at 26. Diabetes, very small, is once again, it's a 0/1 variable, so roughly 3 percent, 2.5 percent were at the value one, which means they had diabetes. And the heart rate averaged about 75. And the glucose averaged about at 81.6.

Now, be careful, because they were roughly 400 people or so who did not report their glucose level at their first visit.

In summary, by looking at the mean column, we start to get an idea about where these variables are centered.

## Robustness of the mean



Note what happens when one number, 4.02 say, becomes large, say 40.2 :

2.3, 2.15, 3.50, 2.60, 2.75, 2.82,  
4.05, 2.25, 2.68, 3.00, **40.2**, 2.85

Mean =  $\bar{x} = 5.73$

(versus 2.95, from before)

Mean is **sensitive** to every observation,  
it is not **robust**.

One of the characteristics of the mean is that it is affected by every single value. For example, take a look at what happens if we take one number, suppose we take the 4.02 and multiply it by 10 by mistake, say. Now it's 40.2. What happens to the mean?

When we calculate the mean it now is 5.73, in contrast to the old value of 2.95 when we had a 4.02 instead of 40.2.

So just moving this one number, has made a huge difference to the mean; indeed it has almost doubled it. What we have just demonstrated is that the mean is sensitive to every observation. It is not robust.

The median

Definition of the **Median**:



At least 50% of the observations are greater than or equal to the *median*, and at least 50% of the observations are less than or equal to the *median*.

$$2.15, \underline{2.25}, 2.30 - \text{ median} = 2.25$$

$$2.15, \underline{2.25}, \underline{2.30}, 2.60 -$$

$$\text{median} = \frac{1}{2} (2.25 + 2.30) = 2.275$$

Another statistic we can look at to inform us about the middle of the data is, is the median. The median is defined to be a number such that at least 50 percent of the observations are greater than or equal to the median, and at least 50 percent of the observations are less than or equal to the median.

So for example, if we have three numbers, 2.15, 2.25, and 2.30, then the median is the middle one, 2.25. In fact, this is true for any odd number of numbers, we always get the middle number. On the other hand, if we have an even number of numbers, like for example, four numbers such as 2.15, 2.25, 2.30 and 2.60, then the median, can be any number between the middle two, 2.25 and 2.30 here. By convention, we choose the mean of the middle two numbers. And so in this case, the median would be 2.275.

The median too, gives us an idea of where the middle of the data is.

Median Age at First Marriage, USA					
Year	Males	Females	Year	Males	Females
1890	26.1	22.0	1996	27.1	24.8
1900	25.9	21.9	1997	26.8	25.0
1910	25.1	21.6	1998	26.7	25.0
1920	24.6	21.2	1999	26.9	25.1
1930	24.3	21.3	2000	26.8	25.1
1940	24.3	21.5	2001	26.9	25.1
1950	22.8	20.3	2002	26.9	25.3
1960	22.8	20.3	2003	27.1	25.3
1970	23.2	20.8	2005	27.0	25.5
1980	24.7	22.0	2006	27.5	25.9
1990	26.1	23.9	2007	27.7	26.0
1993	26.5	24.5	2008	27.6	25.9
1994	26.7	24.5	2009	28.1	25.9
1995	26.9	24.5	2010	28.2	26.1

How useful is it? Consider this example. Here is the median age at first marriage in the USA between 1890 and 2010. We see that in 1890 the median age at first marriage for males was about 26. And now, it has gotten to be a little higher. In this decade, it's reached above 26. It went down in the middle of the century. It was down in the low 20s. But now it's picked up again.

How about females? We see a similar behavior: It was about 22. Then it went down. Then it picked up and then it sort of plateaued.

The difference in the medians between men and women was about four years back in 1890. It decreased to about two and a half. And now it's barely above two.

So this gives you some idea of what's happening with the middle of the population. Half the population is below, and half is above this. By all means check the data out and create your own favorite theory as to why this is happening, but the median does a good job in this case of summarizing the data.

## Mean and median

---

<sup>1</sup> <http://www.infoplease.com/ipa/A0005061.html>

U.S. Bureau of the Census; Web: [www.census.gov](http://www.census.gov).



. summ age1 sex1 sysbp1 diabp1 cursmoke1 cigpday1 bmi1 diabetes1 heartrte1 glucose1

Variable	Obs	Mean	Centile
age1	4434	49.8392	49
sex1	4434	1.548489	2
sysbp1	4434	134.3718	130
diabp1	4434	84.09303	82.5
cursmoke1	4434	.4772215	0
cigpday1	4415	8.174858	0
bmi1	4417	25.99249	25.64
diabetes1	4434	.0252594	0
heartrte1	4433	75.34446	75
glucose1	4047	81.61848	78

So now we've got two measures of central tendency. Do you use the mean or the median as a summary? If there is a choice, when do you use the one and when do you use the other?

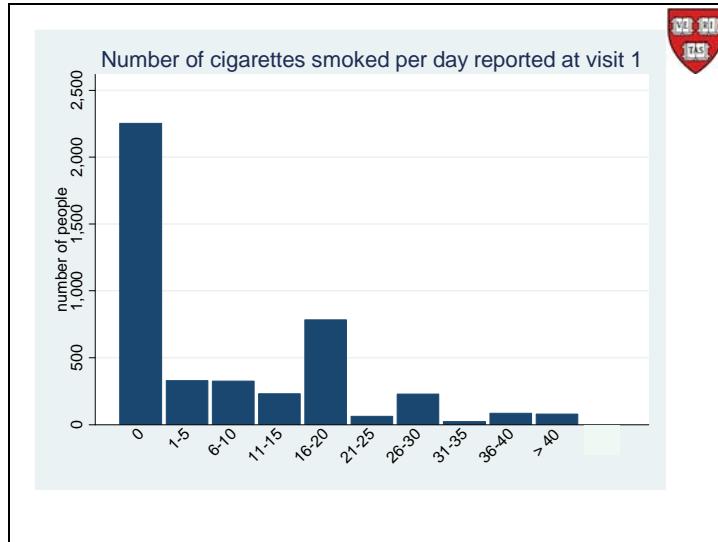
For example, if we return to the 10 variables for which we just calculated the mean: look at age and we see that the mean is about 50. The median is also roughly the same thing-- it's almost a year less. (By the way, this is the 50th centile, and that's how you get it from Stata. You get it from the detail option.)

Look at sex. Remember this was 0, 1 variable. Look at what happens to the median of a 0, 1 variable. It just tells you which of the, two values is more popular. And we said that the value 2 was 54 percent, so the median is not as informative, it's not as nuanced, as the mean is in the case of dichotomous variables.

Systolic blood pressure has the mean and median pretty much the same. The difference is about four. The diastolic blood pressure is pretty much the same, too. The "current smokers", the median, the same as sex, is not very informative-- it just tells you which one has got more than 50 percent.

The cigarettes per day is interesting. The median is 0. Why is that? A very similar argument as used with dichotomous variables: because more than half the people are nonsmokers, so more than half are at zero cigarettes per day. So once again, but for a slightly different reason, this is not very informative because it just tells you that more than 50 percent are at 0. And once again, we'll look at this carefully, soon. BMI—mean and median are very similar, 25.64; the same thing for the heart rate.

So we seem to come out of this, from these 10 variables thinking that for dichotomous variables or one where there's a big lump at the origin, or anywhere else the median is not very informative. So at the same time as the mean may be too sensitive, the median is not sensitive enough.



Now, let's take a look at our number of cigarettes smoked per day. Here is a bar chart, if you will, of the variable in question. Now how are you going to describe the center of this variable with a single number? It's not very easy to do, and I daresay it wouldn't be very informative even if you could do it.

But an interesting accompanying question is, why is this bar at 16-20 so big? The one at the origin we know because there are a number of people who do not smoke. But in order to answer this second bar we need to look at the data.



0	2,253
1	70
2	20
3	103
4	11
5	125
6	18
7	12
8	12
9	134
10	150
11	5
12	3
13	3
14	2
15	219
16	3
17	7
18	8
19	2
20	764
23	6
25	56
29	1
30	227
35	22
38	1
40	85
43	57
45	3
50	6
55	1
60	12
70	1
<hr/>	
Total	4,402

I've listed the values the variable takes together with a frequency count for each value (obtained by just tabulating this variable in Stata.) There are 4,402 individuals and more than 50 percent are at 0, so that explains the median.

But running up the frequencies from the bottom, there seems to be an interesting phenomenon in play here. Look at these large frequencies. For example at 40, 30, 20, 15 etc.. Why are those numbers very big?

Look at the largest, 764 people reported smoking 20 cigarettes per day. Well, one used to buy cigarettes in packs, and there were 20 cigarettes in a pack. And so what are these people telling us? That they smoked roughly one pack a day—actually it was surprising how people actually did smoke exactly one pack a day! Then, 40 represents two packs a day, and 43 maybe a little bit more than two packs a day. Then there are the ones with a pack that would last maybe two days, and they probably reported smoking 10 a day.

Why the nine is associated with a large frequency, I have no idea. The five, because it's a nice round number. And you will see this very often in data, and you should be careful of this phenomenon, namely people like to round off to the closest round number-- so, 5, 10, 15, 20 are often popular. This should not surprise you because one does not, by and large, smoke exactly the same number of cigarettes every day, my statement above notwithstanding. So the natural tendency is to quote a nice, round number.

## Mean vs Median



When to use mean or median:

Use both by all means.

Mean performs best when we have a symmetric distribution with thin tails.

If distribution is skewed, use the median.

Remember: the mean follows the tail.

Returning to the original question, mean or median? If the computer is going to be doing all the work, use them both, by all means.

When does the one perform better than the other? When we have continuous data, then the mean performs best when we have a symmetric distribution with thin tails. We don't want fat tails, because fat tails mean that there is a good chance of getting values far away from the center of the distribution, and the mean is not robust to those.

If the distribution is skewed, then my advice is to use the median. And this very often happens with survival data-- how long do people live?



Variable	Obs	Mean	Centile
age1	4434	49.8392	49
sex1	4434	1.548489	2
sysbp1	4434	134.3718	130
diabp1	4434	84.09303	82.5
cursmoke1	4434	.4772215	0
cigpday1	4415	8.174858	0
bmi1	4417	25.99249	25.64
diabetes1	4434	.0252594	0
heartrtel	4433	75.34446	75
glucosel	4047	81.61848	78

Remember, the mean follows the tail. Let's return to our ten variables. For example, with the systolic blood pressure, the mean is bigger than the median. With the diastolic blood pressure, the mean is bigger than the median. And that could be because we have a few large values on the right hand side.

The mean is much bigger than the median for cigarettes per day because we had a lot of people smoking a large number of cigarettes per day—a long right tail.

With the heart rate, the mean is bigger, as is the blood glucose value--remember, we have diabetics in the group. There may only be about 2.5 percent who are diabetic, but they would pull the right tail up, and that would explain why we have this difference between the mean and the median.

## Composition Formula

Grouped means										
	Total									
$X_i =$	$\frac{1}{4} \quad \frac{1}{2} \quad \frac{10}{3} \quad \frac{25}{1} = 69$									
$n_j =$	$4 \quad 2 \quad 3 \quad 1 = 10$									
$\frac{1}{10} \sum_{i=1}^{10} X_i = \frac{1}{10} \sum_{j=1}^4 n_j \bar{X}_j =$										
$\frac{1}{10} (4 \times 1 + 2 \times 5 + 3 \times 10 + 1 \times 25) = \frac{69}{10}$										
$= .4 \times 1 + .2 \times 5 + .3 \times 10 + .1 \times 25$										
$= \sum_{j=1}^4 p_j X_j \quad \text{where } p_j = \frac{n_j}{n}$										

So let's revisit the mean because the mean plays such a central role in statistics, and let's see what more we can learn about it. Consider this example. Suppose we want to find the mean of these numbers: 1, 1, 1, 1, 5, 5, 10, 10, 10, and 25. So what do we do?

First of all, we count out how many of them we have. And we've got 10 of them. Next we need their total: 4 plus 14, 24, 34, 44, 69. So the mean is 69 over 10. So that's 6.9.

Alternatively, I could count the other way, from right to left. I could say 25 plus 10 plus 10 plus 10, et cetera. And hopefully, I'll get back to 69. I want you to think of another way of summing this. Can you spot something? What have I done here? I've made a number of these things equal to each other. So can we take advantage of that?

Collecting like numbers before summing them, we have four 1s, two 5s, three 10s, and one 25. So we could say,  $4 \times 1 + 2 \times 5 + 3 \times 10 + 1 \times 25$ . And we would get the same total, 69, and divide by 10 to get the mean.

Now, I could divide the total by 10. Or I could go into each multiplicand and divide each one by 10. Let me rewrite it in decimals. So what I have is,  $0.4 \times 1 + 0.2 \times 5 + 0.3 \times 10 + 0.1 \times 25$ , and that is 6.9, as before because I haven't changed anything. So we see that the overall mean-- let's call that  $\bar{x}$ -- can be written as a summation of proportions times typical value.



## Properties of the weights (proportions)

1. Each of the ps (ns) is non-negative.
2. The ps sum to one(1).



Note that all the Xs within a group need not be equal.

$$\begin{aligned}\bar{X} &= \frac{1}{6}(1+2+3+4+6+8) \\ &= \frac{1}{6}(\{1+2+3\} + \{4+6\} + 8) \\ &= \frac{1}{6}\left(3\frac{\{1+2+3\}}{3} + 2\frac{\{4+6\}}{2} + 1\frac{8}{1}\right) \\ &= \frac{1}{6}(3 \times 2 + 2 \times 5 + 1 \times 8) \\ &= .5 \times 2 + .33 \times 5 + .17 \times 8 \\ &= \sum_{i=1}^3 p_i \bar{X}_i = 4\end{aligned}$$

Now, what if not all the x's within a group are equal? Does that make any difference? Can we still do the grouping? And the answer is yes. Let's just do it by example.

Suppose that the numbers I have are, 1, 2, 3, 4, 6, 8. And suppose I group the first three together, and then the next two, and lastly a group of size one for the last number. Now what is the sum of these six numbers?

Well, the sum of the first group is 6 divide that by 3 because there's 3 numbers. So that will give me the mean of the three. But if I want to keep the arithmetic the same, I have to multiply by 3. So here, I've got 2. For the second group the sum is 10. Their average would be 10 over 2. But then to retain the summation, I've got to multiply by 2. So 1 x 8 over 1.

So once again, if I want the mean-- I've got 6 numbers-- so I need to divide this sum by 6. So divide each group by 6. And now, what do I have? I've got the proportion for the first. And that's 3 over 6, with the first 50 percent. So it's 0.5.

And what is the average of them?  $\bar{x}_1$  would be 2 plus the proportion in the second group. So that would be 2 over 6, which is  $1/3$ , which is, let's say, 0.33 and times 5. And then plus  $1/6$  would be the  $P_3$ . And now, that's equal to 0.167, let's say. And then that's times 8. So once again, I can write this as summation  $P_j \bar{x}_j$ . So  $\bar{x}_1$  is 2,  $\bar{x}_2$  is 5,  $\bar{x}_3$  is 8.

So this is from  $j = 1, 2, 3$ . So this is exactly the same form as we had before. Namely, the overall mean  $\bar{x}$  is a weighted sum of these  $x_j$ 's. All these weights have to do is they have to sum up to 1. And secondly, each one of the  $P_j$ 's has to be greater than or equal to 0. We can call this the composition formula.



Thus a group mean can be represented as a weighted sum of the means within the groups. The weight of a particular group, or stratum, represents the proportion of the whole within that group.

$$\bar{X} = \sum_{j=1}^g p_j \bar{X}_j \quad \text{where } \sum_{j=1}^g p_j = 1$$
$$= \frac{1}{n} \sum_{j=1}^g n_j \bar{X}_j \quad \text{where } \sum_{j=1}^g n_j = n.$$

Overall mean made up of three groups



$$.5 \times 2 + .33 \times 5 + .17 \times 8 = 4$$

What happens if the mean of the first group goes up but the other two remain the same?

$$.5 \times 3 + .33 \times 5 + .17 \times 8 = 4.5$$

Indeed, the same effect, viz. the overall mean goes up, if one, some, or all of the individual group means go up.

Similarly, when the individual means go down, the overall mean goes down.

If some go down others go up, then we need to look at the Composite to see what happens.

<http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-the-honor-roll>

For example, suppose I've got a hospital. And so I've got a hospital. And what I want to do is, I want to find out what the mortality rate is for my hospital so that if I am looking for a hospital, I might look at the hospitals and see which of the hospital with the lowest mortality rate that I might think, OK, that's the best hospital. That's where I'd like to go if

I have to go to the hospital. Let's simplify the argument and say that there are three groups of people who go into the hospital.

Let's say, there's the young group. And let's say that there is the middle-aged or middle group. They are middle-aged. And then let's say, older folks. So those are the three. Now let's just, for argument's sake, say that the mortality rate associated with the young folks is something like 2 per 1,000, 10,000-- whatever it is. Let's say 2 per 1,000. Middle group, let's say is 5 per 1,000. And let's say, for the older group is 8 per 1,000.

Now, going back to our formula for our hospital. Suppose we've got a hospital, call it Hospital A, let's say. And let's say in Hospital A, 50 percent. So let's say 50 percent of the folks coming to a Hospital A are young folks. So what is their mortality? It would be  $50\% \times 2$ . And let's say that the middle group, there were 33 percent in the middle group. And so the mortality would be 5. And the last group would be-- it has to add up to 1, so this would be 17 percent. And that's at the 8.

So here's our formula from before. We had the proportion times the mean for that group. Proportion times the mean for that group. Proportion times the mean for that group. So that if we calculate the average-- maybe do the calculation here-- we get that the answer is 4. I actually chose the numbers so that the answer would be 4.

So the average then would be 4. And our units are per 1,000. So it would be 4 per 1,000. So there's our average for Hospital A that accepts this percentage at distribution.

Now, what I want you to do is go off and think about this for yourself a little bit. And see what happens when you have a hospital with a different percentages with a different mixture of patients. Always the same, keep these the same. So arguably, these hospitals that you're going to be playing with will have the same or comparable mortality rates associated with the three age groups.

But all I want you to do is play with these and see what happens to the overall average. And then come back to me and say, ah, maybe I shouldn't be looking at the overall average when I'm rating the hospitals. I should be looking at something a little bit more subtle, a little bit more standardized.





Return to original mean:

$$.5 \times 2 + .33 \times 5 + .17 \times 8 = 4$$

What happens to the mean if the third group gets to be a bit bigger (relatively)? E.g.

$$.5 \times 2 + .30 \times 5 + .20 \times 8 = 4.1$$

So you went off and you did your calculations. So let's see if our answers match, and if you come to the same conclusions I've come up with. So with hospital B, we said that hospital B, our distribution was 33 percent at the 2, plus 33 percent at the 5. Plus 34 percent, remember, all these percentages have to sum up to 100, at the 8.

And that gives a mean of 5.03 per thousand. Whereas hospital C had 75 percent of the young ones, might be like a children's hospital. And 25 percent in the middle group. Plus 0 at the top group, so this comes out to be 2.75. So here are three hospitals per thousand. So here are three hospitals, and what conclusion do you come to?

Well, if you're just going to base their ranking on the mortality, hospital C is the best hospital, because it's down at 2.75. Then hospital A is the number two hospital. So this would be the number one hospital, this would be the number two hospital because it's at four. And then hospital B is the worst hospital, rank number three.

But yet, at all three of them have exactly the same age specific mortality rates. Exactly the same. This one should have been an 8 also, but it gets multiplied by 0, so it doesn't matter. The only difference is the composition. That has nothing to do with the quality of the hospitals.

### Comparing composite or group means



When comparing two composite means make sure we are comparing likes. If the composition (weights or proportions) changes then the comparison of means is less meaningful.

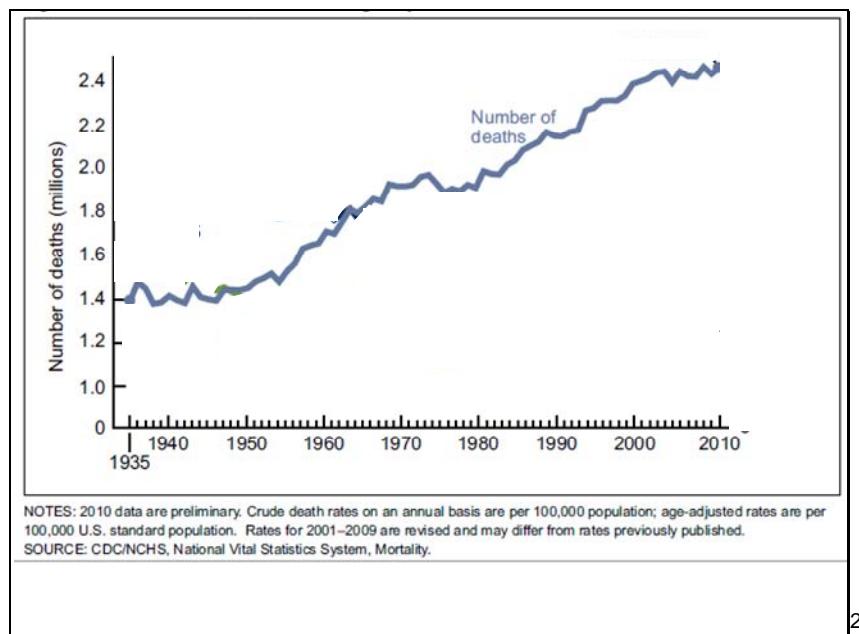
This gives rise to  
Index numbers, or  
Standardization methods

So the moral of the story is that when you are using the mean to make comparisons, make sure that your compositions are the same. Because the mean, remember, the mean can be written as summation over the groups. So it's a weighted sum of individual group means.

And you can vary the group means, or you can vary the proportions when you go to one hospital to the next. So be careful when you are comparing two means. Make sure that the compositions are the same.

## Standardization

## Rate



The challenge now is to find a statistical measure to evaluate the health of a group over time. For example, consider the USA, over the past century or so. There are a number of approaches we could take, but let us focus on mortality. We can argue that if health is better, the mortality should go down, and it is a hard endpoint that does not lend itself too easily to different interpretations.

If we judge mortality by looking at the number of deaths, then we have a graph that looks like this: on display is the number of deaths over the last 75 years. We see that the number of deaths have been going up.

<sup>2</sup> Donna L. Hoyert, Ph.D. 75 Years of Mortality in the United States, 1935–2010, NCHS Data Brief , Number 88, March 2012

This evaluation is troublesome because we believe that health has been improving, whereas this graph makes it look like it has gotten worse over these last 75 years. The problem is that the population has been increasing over the same period. So it is not surprising that the number of deaths has been increasing too, so the population size confounds the issue: the number of deaths can increase even if we are getting healthier—there are just more of us to die.

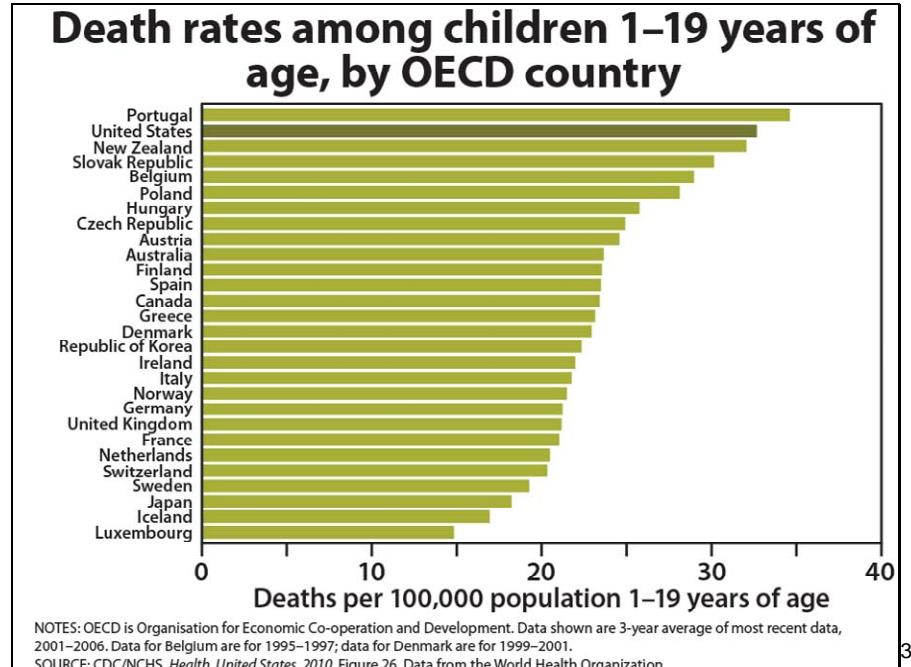
**Rate**


$$\text{Rate} = \frac{\text{numerator}}{\text{denominator}} \text{ per time unit}$$

- “Crude” rate, single number, summary
- allows for standardization
- makes comparisons more meaningful

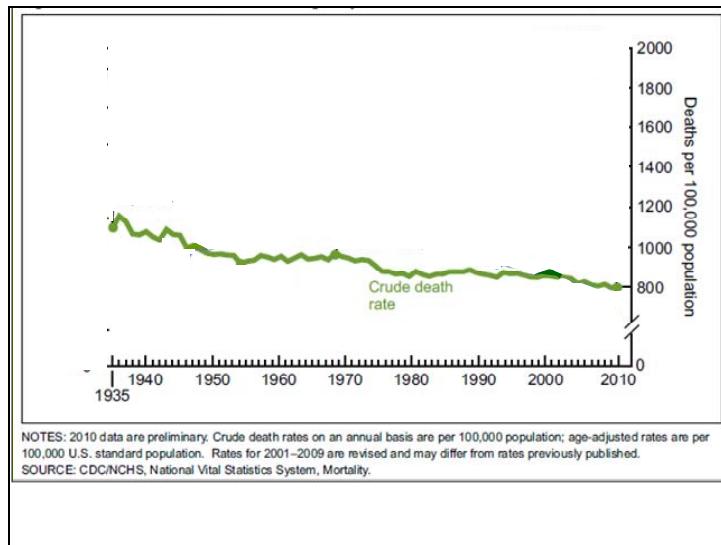
To overcome this problem, we can introduce a rate. A rate has a numerator, which could be the number of deaths, and a denominator, which could be the population size. Like that, we can correct for the population size as it gets bigger.

What makes this different, from a proportion, say, is that first this is per unit time. So for example, we could calculate a rate per year, and then our unit of time could be one year. If we do this, this will allow us to standardize—it makes comparisons more meaningful, because it takes into account the population size.



This is what we call a crude rate—it is a single-number summary. But it does allow us to make these comparisons. So for example, here is Portugal. And this is the mortality rate for 1 to 19 years of age. There's Portugal, and the population of Portugal is something like, oh, 10 million or so. So the US is, what, 300 million? So there's Portugal 10 million, versus 300 million. The comparison is still meaningful when we compare Portugal to United States, just as it is to compare it to Luxembourg, which only has, like, half a million. So the population size doesn't matter in this.

<sup>3</sup> <http://www.cdc.gov/nchs/hus/previous.htm> 2010 edition

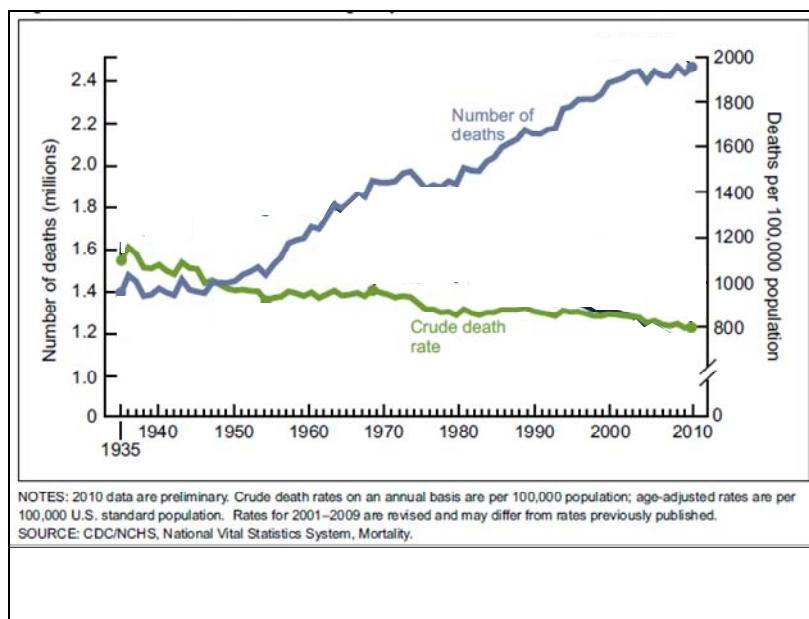


Now when we do that for the US, here's what we see. We see that the crude death rate, namely if we divide, each year, the number of deaths by the population size, we get something that looks like this. And voila, it's going down. And this makes us happy.

So the graph now also shows deaths per 100,000 population per year, and that rate is decreasing.

Secondly, note that a rate is not a proportion, since the latter would require that the numerator be part of the denominator. There is no such restriction with a rate.

For example, if we look at the number of colds in a season, if we look at the cold rate-- how many colds do you have per winter? You can have more than one. So the numerator would be the number of colds, and the denominator would be the population size. Thus the rate measures the number of recurring events in a season. So rates can be bigger than one.

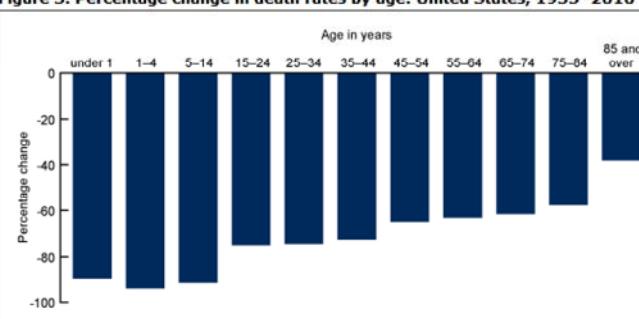


This graph shows both, the population size and the crude death rate. Let us think a little about the residual effect from year to year about what an improving death rate achieves.

Let us start by looking at the crude death rate over time. We see that in 1935 this rate was just a little bit less than 1,200. And now in 2010, we are about 800. So this difference, this delta, is less than 400. If it were 400, that would be a third. So this is about 30 percent or so, and we see a roughly a 30 percent improvement in the crude death rate.

The decrease in death rates over the 75 years varied by age.

**Figure 3. Percentage change in death rates by age: United States, 1935–2010**



NOTE: 2010 data are preliminary.  
SOURCE: CDC/NCHS, National Vital Statistics System, Mortality.

This crude death rate is measured on the whole population. Now consider the components that go into this calculation, namely the age specific (as measured within certain age groups) mortality rates. If we look at what happened in the individual age groups between 1935 and 2010, in those 75 years, we see that the under-one age group had an improvement of about 90 percent. The 1 to 4 had an even bigger improvement. The 5 to 14 age group shows roughly the same improvement. In the 15 to 44, is roughly in the 75 percent group. In fact, the smallest bar, the 85 and over, shows the smallest decrease, and that one is about 45 or so. It's in the 40s. It's more than 40 percent.



But the overall mortality rate is not showing the improvements we are seeing in the age specific mortality rates.

What is happening?

The rate is a mean. Let's go back to the mean and see what we can see about what the mean really is.

So how is it that the average of numbers-- each of which is over 40 percent --how is it that that average is less than 30 percent? Why aren't we seeing a similar improvement in the crude mortality rate that we see in its component parts?

So what is happening? Well, we're taking an average. So let's rethink and rediscover what an average really is.

Overall mean made up of three groups



$$.5 \times 2 + .33 \times 5 + .17 \times 8 = 4$$

What happens if the mean of the first group goes up  
but the other two remain the same?

$$.5 \times 3 + .33 \times 5 + .17 \times 8 = 4.5$$

Indeed, the same effect, viz. the overall mean goes up,  
if one, some, or all of the individual group means go up.

Similarly, when the individual means go down, the  
overall mean goes down.

If some go down others go up, then we need to look at the  
Composite to see what happens.



Return to original mean:

$$.5 \times 2 + .33 \times 5 + .17 \times 8 = 4$$

What happens to the mean if the third group gets to be a bit bigger (relatively)? E.g.

$$.5 \times 2 + .30 \times 5 + .20 \times 8 = 4.1$$

## Age adjustment



Thus a group mean can be represented as a weighted sum of the means within the groups. The weight of a particular group, or stratum, represents the proportion of the whole within that group.

$$\bar{X} = \sum_{j=1}^g p_j \bar{X}_j \quad \text{where } \sum_{j=1}^g p_j = 1$$

So what have we learned so far? We've learned that a mean, an overall mean, can be represented as a mean of means, or a weighted average of group means. So we've got  $g$  groups, each with their individual mean, and we can combine all of those together.

And we come up with the overall mean. Now each of these  $p_j$ s, as you've seen, are each greater than or equal to zero. And they sum up to one.

Rank	Hospital	Points	
1	Massachusetts General Hospital, Boston	30	
2	Johns Hopkins Hospital, Baltimore	30	
3	Mayo Clinic, Rochester, Minn.	28	
4	Cleveland Clinic	27	
5	Ronald Reagan UCLA Medical Center, Los Angeles	20	
6	Barnes-Jewish Hospital/Washington University, St. Louis	20	
7	New York-Presbyterian University Hospital of Columbia and Cornell, N.Y.	18	
8	Duke University Medical Center, Durham, N.C.	17	
9	Brigham and Women's Hospital, Boston	17	
10	UPMC-University of Pittsburgh Medical Center	16	

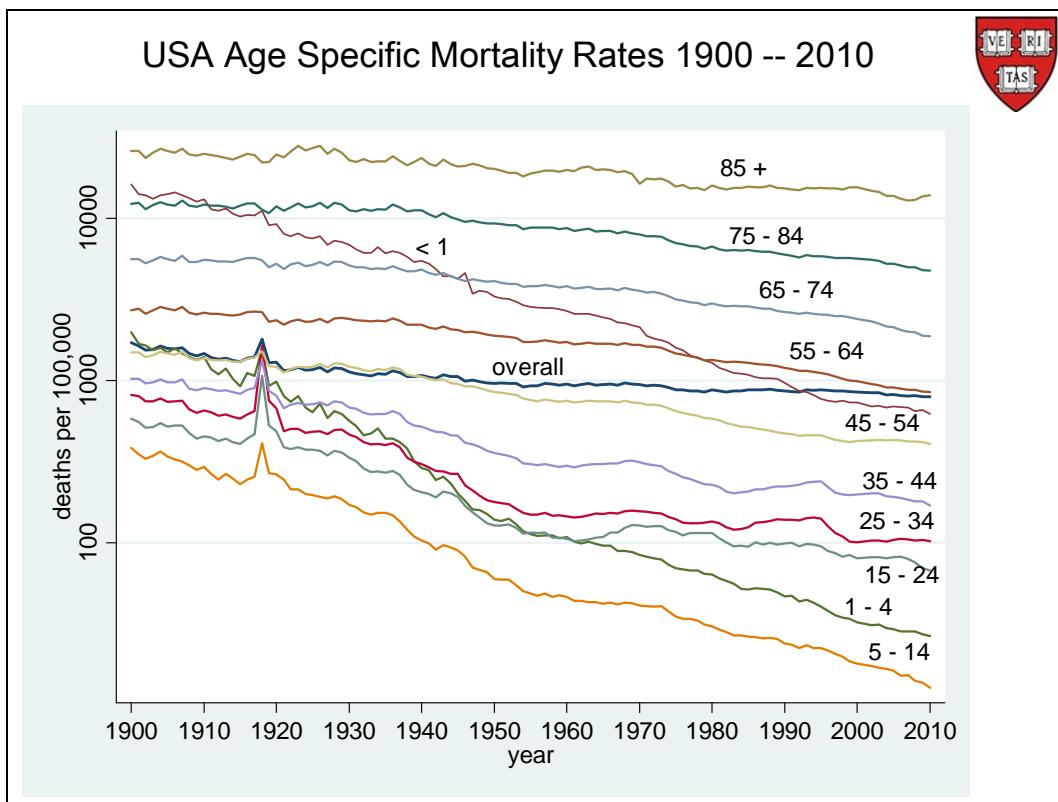
<http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-the-honor-roll>

4

Oh, and in fact, the example you are looking at with ranking of hospitals, the US News and World Reports every year do precisely this. They consider various subcategories, they get points from each subcategory, and then they sum these scores up, and come up with an overall score.

And they must be doing something right, because here they are for this year, and there are two Harvard hospitals in the top ten, including the top one, actually. And here's my alma mater, Johns Hopkins at number two. So they really must be doing something right.

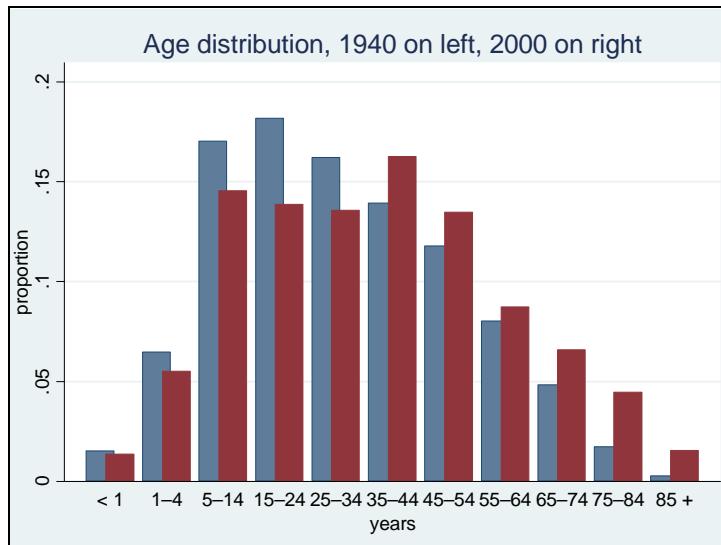
<sup>4</sup> <http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-the-honor-roll>



Returning to our task of studying the health of the USA over the last century, let us look at the age-specific mortality rates.

In the middle is also the overall, or crude, rate. Note that the crude rate is decreasing gently, whilst its component parts are decreasing at different speeds, but in the large at more aggressive speeds. The “less than 1” is the fastest. And even the 1 to 4 is also very fast. But there they are, by age group.

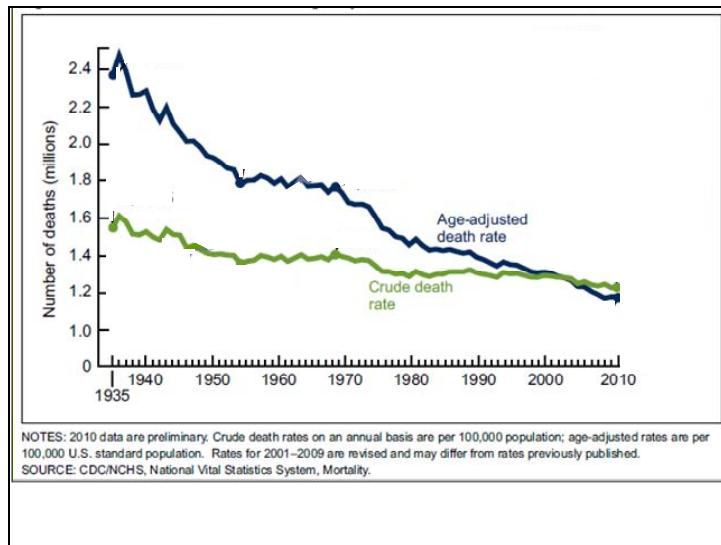
Before leaving this slide, I thought you'd be interested in the shapes of these rates. I took it back to 1900 to show you the effects of the 1917—1918 flu epidemic. What we see is that the spikes are not equally evident in all age groups, indeed, they are even lacking in some age groups. So the age groups in the population were differentially affected.



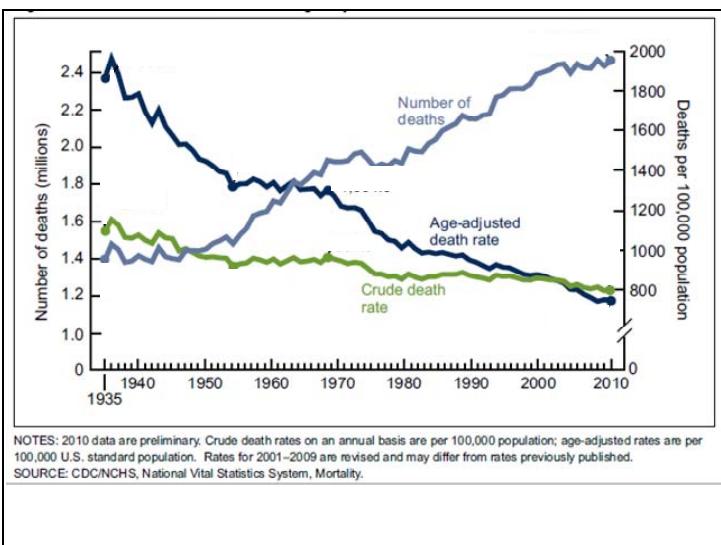
Returning to the problem at hand, if we think about the impact of time on our measurements, we realize that as our health improves, we, as a group, get older. As a result, we get penalized in the overall rate because, by and large, the older we are the higher the mortality rates we are subjected to. If we look at the composition of the US population in 1940—that is the blue bars--and compare that to the composition in the year 2000—that is the red bars--we see that, in the early age groups, the blues are above the reds. In the later age groups, the reds are above the blues. That means the population is getting older.

So, we learnt that comparing the overall mean, when ranking our hospitals whose patients came from three different age groups, was not fair if they had a different composition of patients, so too is it unfair to compare the crude mortality rate from year to year if the age composition of the population is changing year to year.

As a side observation: The two distributions above look similar in shape—a quick increase from birth followed by a slower decrease to the right—if we ignore the 35—44 and 45—54 age groups. These two groups are a little too large. They represent the baby boomers born after the Second World War.



So recognizing that the population is getting older with time, the crude death rate will provide a confounded picture of reality and thus will not satisfy our original requirement of a statistical measure of the population health. We could look at the age-specific mortality rates and those would tell the story quite accurately. But in our search for a single number, and its associated simplicity, we return to the mean and look at the composite character of the crude rate. We recognize that the weights are changing with time. So consider keeping the weights constant.



<sup>5</sup> Donna L. Hoyert, Ph.D. 75 Years of Mortality in the United States, 1935–2010, NCHS Data Brief , Number 88, March 2012

So then the question is, what weights should we use? Today, consider using a set of weights that roughly reflect today's age distribution. One such would be to use the 2000 age distribution.

So this is what is called the age-adjusted death rate. It is a construct that reflects having a population that looks like a particular year—currently the year 2000. So if we take our weighted average, with the weights given us by the population in the year 2000, that's exactly how that curve is calculated. So now, we can make comparisons from year to year without worrying about the confounding effect of the changing age composition.

In summary, what we've done is, we've gone from looking at the number of deaths, to the crude death rate, by just dividing by the population size, by making a construct. So this is a construct. So be careful. This is a construct which requires us deciding, making a judgment, what population to use to give us the weights. But once we do that, then we get a purer comparison.

#### Comparing composite or group means



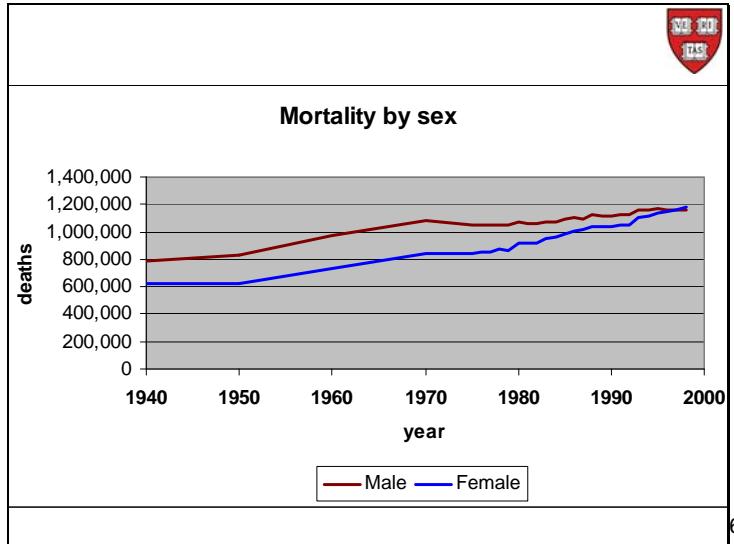
When comparing two composite means make sure we are comparing likes. If the composition (weights or proportions) changes then the comparison of means is less clear, and there is more confounding.

This gives rise to  
Index numbers, or  
Standardization methods

So when comparing two composite means, make sure we are comparing likes. If the composition changes, then the comparison of means is less clear and there's more confounding going on. For example, we're not just comparing rates, we are confounding the comparison with the age of the population.

This is exactly what goes on with index numbers. So if you're thinking of the consumer price index, or things like that. It's similar-- that's exactly how these things are calculated. These are the standardization methods.





6

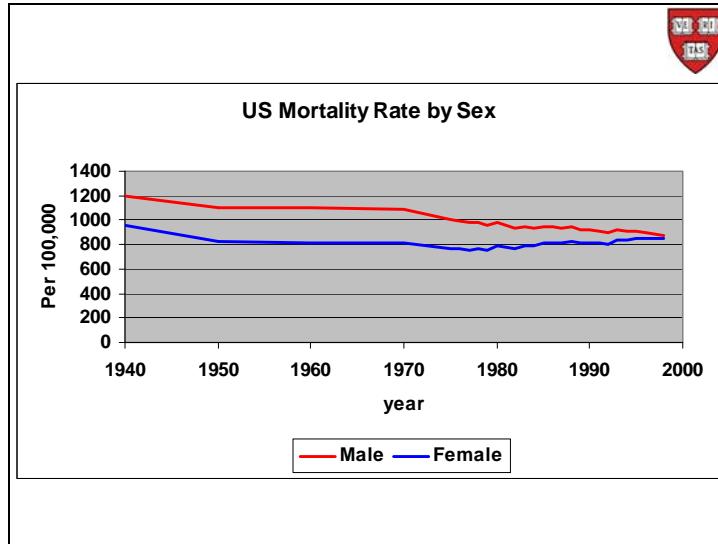
We must be cautious when we do age adjustment, if our intent is to compare two different groups. So for example, here is a sex breakdown of the crude mortality rate over the last 60 years in the USA. The top line is males and the bottom one is females.

We see that more men than women die each year, but the gap is closing up. Indeed, in the last year it looks like women are getting ready to overtake the men.

Once again, this is just counting the number of deaths, and we saw that that is not a good overall evaluation of our health because the population size could be changing, as indeed it has. But what is interesting here is that it has changed proportionately differently for males and females.

---

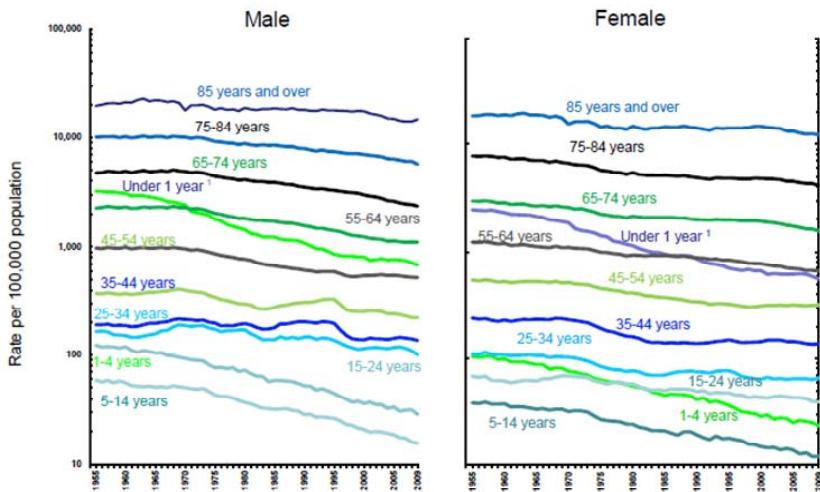
<sup>6</sup> [http://www.ssa.gov/oact/NOTES/as120/LifeTables\\_Body.html#wp1176553](http://www.ssa.gov/oact/NOTES/as120/LifeTables_Body.html#wp1176553)



So indeed, if we now correct by population size--be careful, look at the population of males and females separately.

So now the crossover at the right end has disappeared, although we still see a closing up of the gap. Should we stop here?

Figure 3. Death rates, by age and sex: United States, 1955-2009



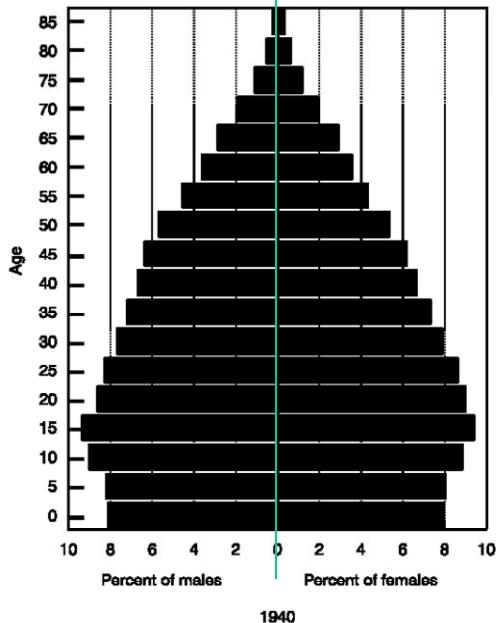
<sup>1</sup>Death rates for "Under 1 year" (based on population estimates) differ from infant mortality rates (based on live births); see Figure 7 for infant mortality rates and "Technical Notes" for further discussion of the difference. SOURCE: CDC/NCHS, National Vital Statistics System, Mortality.

Look at the age specific and, this time also sex specific, mortality rates over the same time period, we see a same similar picture. It takes a little time, but looking closely at this picture we see that the male rates are consistently higher than the respective female rates over the whole time period.

So it's going down, sure, but just as the population was getting older when the rates were going down for everybody, so too now we see that since they're going down differently and less for females, the females should be getting older but at a faster speed than the males.



## Age distributions USA in 1940



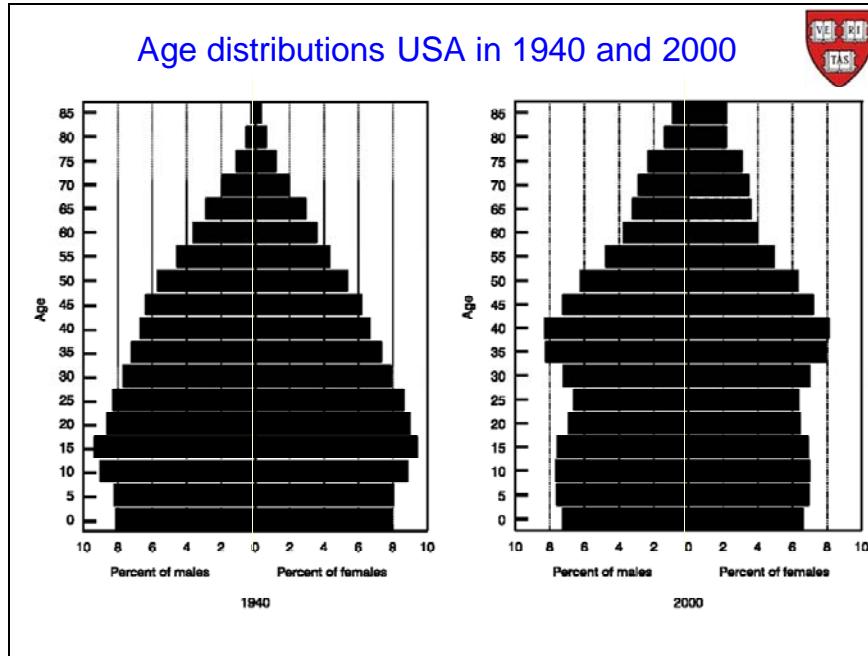
7

Look at the age pyramid. It is two vertical histograms—as opposed to the horizontal histograms you are accustomed to seeing—lying side-by-side. We see that in 1940, the vertical middle line at the zero point on the horizontal scale, demarcates males (on the left) from females (on the right). This line is approximately in the center, meaning that the age distributions for females and males are roughly the same.

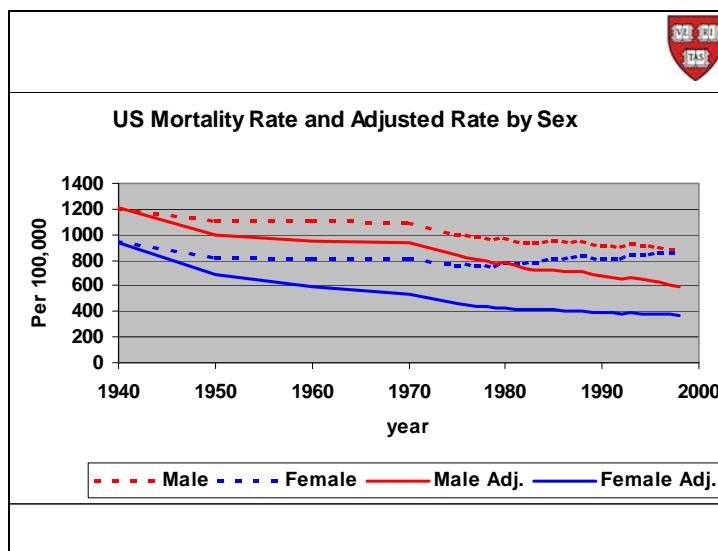
8

<sup>7</sup> Anderson RN, Rosenberg HM. Age Standardization of Death Rates: Implementation of the Year 2000 Standard. National vital statistics reports; vol 47 no. 3. Hyattsville, Maryland: National Center for Health Statistics. 1998.

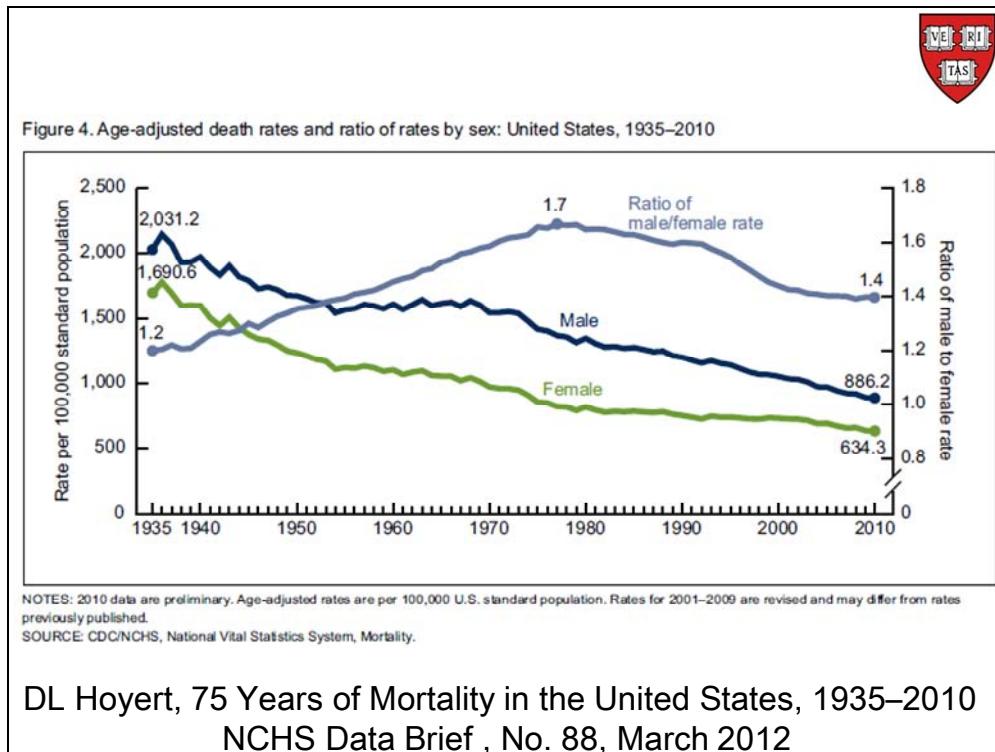
<sup>8</sup> Anderson RN, Rosenberg HM. Age Standardization of Death Rates: Implementation of the Year 2000 Standard. National vital statistics reports; vol 47 no. 3. Hyattsville, Maryland: National Center for Health Statistics. 1998.



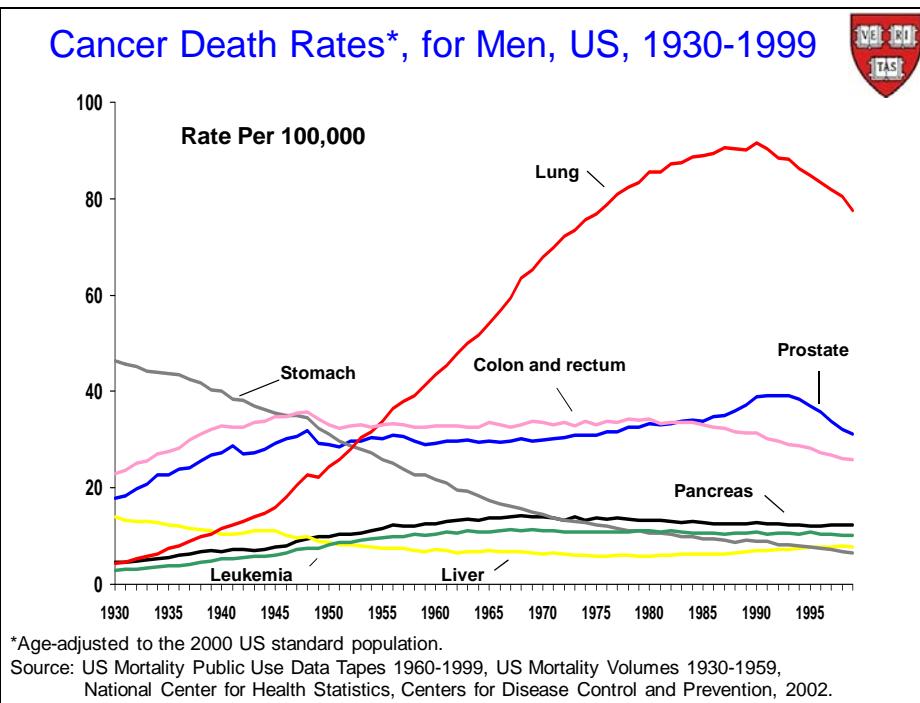
Whereas, after sixty years of differential mortality rates, by the time we get to the year 2000, look at what has happened to that center line. We see that the female age distribution is much more heavily weighted towards the older age groups than the males are. So age adjustment will have a differential effect on the two groups.



And indeed, that is what happens as evidenced by the graph above. You see this grouping together that we saw without the age adjustment has now disappeared. And the age adjustment shows that the gap is very much still there.



In fact, if we take the ratio of the male to female rate, it is coming down, since the late '70s, which is when it reached its peak of 1.7. It is now down to 1.4. But it is still very much there. So be careful and use age adjustments when making comparisons.

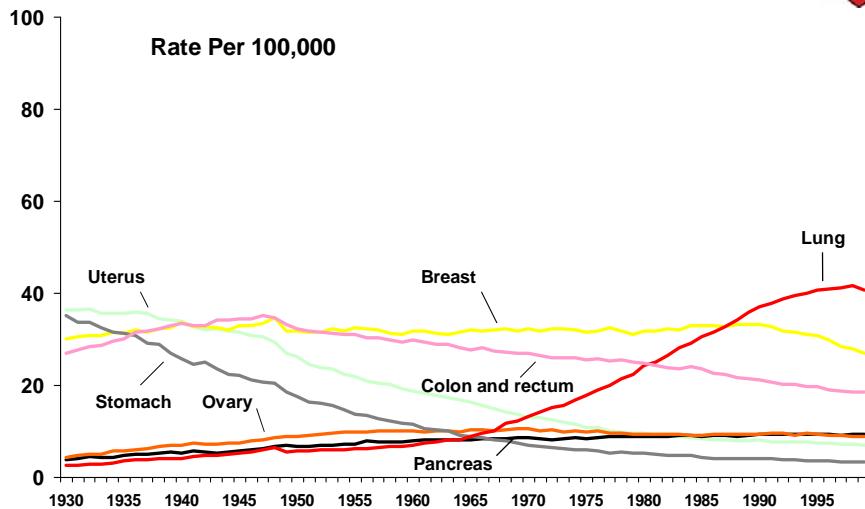


Once we do make age adjustment, then we can allow comparisons to be made over time in a meaningful way.

So here, for example, are cancer death rates for men between 1930 and 1999. These rates have been age adjusted to the 2000 US standard population. That means we have removed the impact of the size of the population and the aging of the population from the cancer mortality rates, making the comparisons over time more meaningful.

We can see that the stomach cancer rate is going down. Most rates are going down eventually. The prostate is coming down, too. Of course, the elephant in the room is what is happening to lung cancer. But even that is now coming down. It is huge relative to everything else, but it is coming down.

## Cancer Death Rates\*, for Women, US, 1930-1999

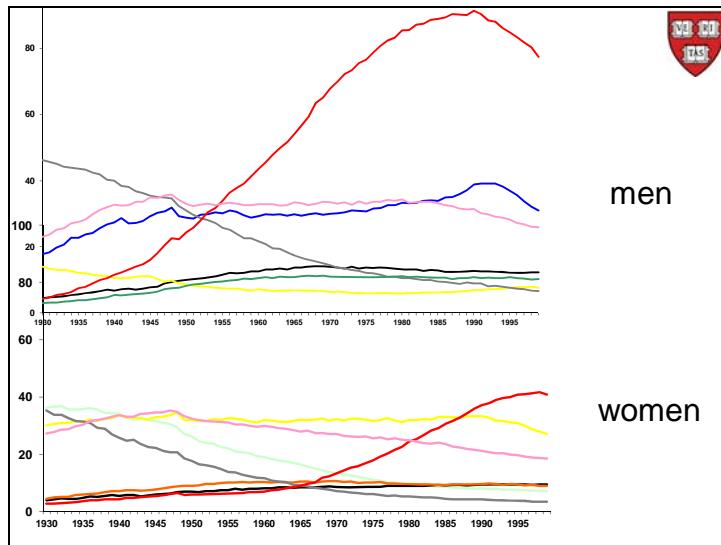


\*Age-adjusted to the 2000 US standard population.

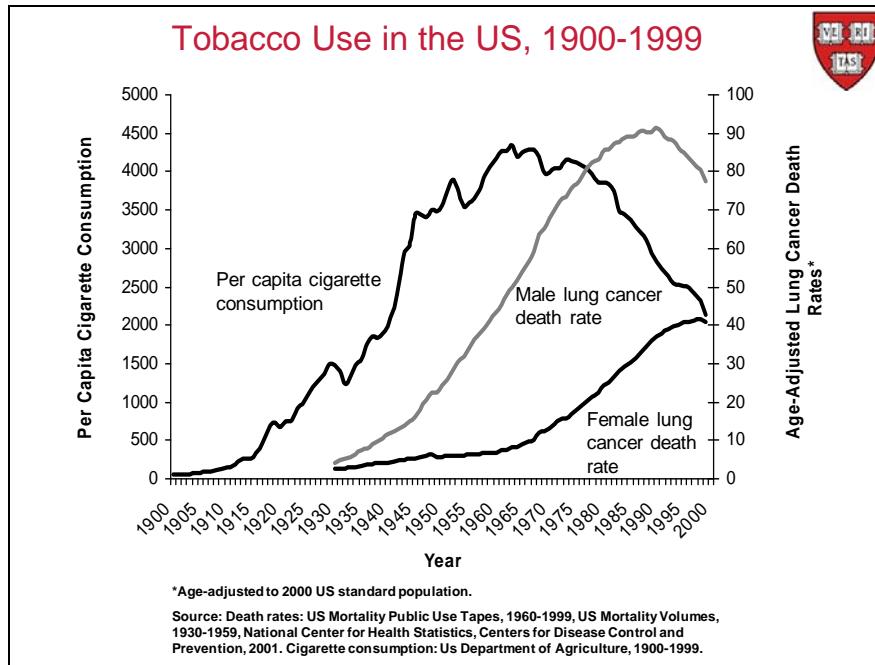
Source: US Mortality Public Use Data Tapes 1960-1999, US Mortality Volumes 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention, 2002.

With women, we are seeing something quite similar. Everything is sort of flat, except there's a little coming down for the colon and rectum cancers. But everything else is sort of pretty flat. There's stomach cancer has made a nice decline. And breast cancer's starting to make some sort of decline here. Let's hope that that goes on.

The lung cancer is pretty much on the increase, although at the very end there is a slight hint that something is going down.



If we put the two side-by-side, we see that the pictures are similar except for lung cancer.



Focusing on lung cancer, we can superimpose the per capita cigarette consumption, on the mortality rates and we see a parallelism between consumption and the male rate, with a lag here of what, about 30 years or so. And that is what makes it very difficult to

try and convince teenagers, or young people, not to smoke. They do not have much appreciation for guarding against something that may happen 30 years down the road.

With women, they are starting to improve, hopefully. The decrease in the per-capita consumption is mostly due to men, and thus the closer link between consumption and male mortality. But we see a hint of a downturn in the female mortality rate.

Remember that age adjustment is a construct. You have to decide what you use as your standard population, and that can impact your results. Just as you saw with the hospital ranking, if you pushed up the ones that you liked, that were doing well, that might impact the overall average.



In summary, standardization, for example, age adjustment, allows two groups of different compositions to be compared.

It achieves this by introducing a “standard” population, such as the US population in the year 2000. It is thus a construct.

So in summary, standardization, for example, age adjustment. allows two groups of different age compositions to be compared. And the way we do it is by taking a standard population to provide the age composition. So we standardize and make both groups look exactly the same. And then we can combine the rates accordingly.

But remember, it is a construct. We used to use 1940 as the standard. Now we use the USA 2000 population because it looks more similar to where we are right now. Possibly in the future, once the baby boomer worked his way through, for example, that might even change.

## Spread summaries

$\bar{x} = 2.95$	FEV <sub>1</sub>	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	
	2.30	-0.65	0.423	
	2.15	-0.80	0.640	
	3.50	0.55	0.303	
	2.60	-0.35	0.123	
	2.75	-0.20	0.040	
	2.82	-0.13	0.169	
	4.05	1.10	1.210	
	2.25	-0.70	0.490	
	2.68	-0.27	0.073	
	3.00	0.05	0.003	
	4.02	1.07	1.145	
	2.85	-0.10	0.010	
	3.38	0.43	0.185	
	Total	0.00	4.66	

Now, we are going to talk about the first component we said about statistics is variability. This module is now going to concentrate on how to actually measure this variability. So let's go back to our example where we had the 13 FEV<sup>9</sup> numbers.

So there they are-- 2.3, 2.15, et cetera. And we've got 13 of these. Now, we said that their center can be given by the mean, which in this case is 2.95. Now, we'd like to get an idea of how they vary around that center. So let's center every single observation by subtracting the 2.95 from every observation. And that is the blue, or second column.

So how do we summarize this? Why not take the mean of these? Just like we did before. We could, but if we did, we would run into problems because the mean of these so-called residuals is *always* 0. That's from the definition of what the mean is. It's the center of gravity. So the mean is always going to be 0—not very informative.

So the problem is that pluses knock out the minuses. So what can we do? Well, one thing we could do is get rid of the signs by taking absolute values. And that is called the mean absolute deviation if you take the average of that. We're not going to do that today.

What we're going to do is the other favorite way of getting rid of the signs, and that is by squaring. So if we square each residual, we get the third, or green column of positive numbers.

---

<sup>9</sup> FEV1 is forced expiratory volume of air that one breathes out in one second. This is often used to measure lung capacity.

## Variance



$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \geq 0$$

e.g.

$$= \frac{4.66}{12} = 0.39 \text{ liters}^2$$

So now what we can do is take the average of these. So take the sum, which is 4.66 and average that out. And that will give us the average of the squared deviations. We call that the variance. So the sum of the squared deviations, then averaged out. You will notice that I divided by  $n$  minus 1 and not  $n$ , to average.

But I'll tell you in a couple more visits why we did the  $n$  minus 1. But for the moment, just trust me. This average is what we call the variance.

We are not going to get into trouble unless we have only one observation—so  $n$  is one. In that case, we can't do anything anyway—we can't judge variations from a single observation.

Each residual is squared, so when we take their sum, that's going to be a non-negative quantity. And so the variance is always greater than or equal to 0. So the average of non-negative quantities is going to be non-negative. And so the variance is always going to be greater than or equal to 0.

The only time the variance can be 0 is if each one of the residuals is 0, which means each one of the  $x$ 's has to be equal to  $\bar{x}$ . In other words, there is no variability. So the variance equals 0 means you have no variability.

Applying this formula to our FEV1 numbers, we get that the sum, 4.66 divided by 12 because there were 13 numbers, and we get 0.39 liters squared. This is unfortunate because our original units were liters and we're not really interested in liters squared.

## Standard Deviation



Standard deviation =  $+\sqrt{\text{Variance}}$

e.g.

$$= \sqrt{0.39}$$

$$= 0.62 \text{ liters}$$

So what we can do is reverse this operation of taking the squares by taking the square root. And the resultant is called the standard deviation.

So the standard deviation is the square root of the variance. And we take the positive square root by convention. So in our example here, there is the positive square root of 0.39 is 0.62 liters.



```
. summ age1 sex1 sysbp1 diabp1 cursmoke1 cigpday1 bmi1 diabetes1 heartrte1 glucose1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age1	4434	49.8392	8.603956	32	70
sex1	4434	1.548489	.4976994	1	2
sysbp1	4434	134.3718	23.37108	83.5	295
diabp1	4434	84.09303	12.97376	50	142.5
cursmoke1	4434	.4772215	.4995372	0	1
cigpday1	4415	8.174858	11.23172	0	70
bmi1	4417	25.99249	4.027281	15.54	56.8
diabetes1	4434	.0252594	.1569295	0	1
heartrte1	4433	75.34446	12.14209	44	140
glucose1	4047	81.61848	21.25588	40	394

Let's apply this to the first 10 variables in our Framingham heart study data. And there they are. When I asked for summary in Stata, I also get the mean, the standard deviation, and the min and the max of each one of the observations.

We could also define what we call the range by taking the difference between the max and the min, and that, in some sense, is also a measure of variation. But we won't go into that very much more.

So let's look at the standard deviation. There is the standard deviation. So for example, the standard deviation for age, at age1, which is the age of the first visit, is 8.60, around the mean of 49. And around sex, it's 0.497 around a mean of 1.54, and so on.

Now, why are we looking for standard deviations? Well, it's giving us an idea of variability. OK, what does that tell us? Well, it tells us how tightly the observations clustered are around that mean. So remember, we said if we had a standard deviation of zero, we said the variance, but the square root of zero is zero, so if we have the standard deviation of zero then all of the observations are going to be equal to the mean, so there's no variability.

And as the standard deviation gets bigger, then the variability around that mean gets bigger. So the question then is, how big is big or how small is small? So for example, if

we look at the standard deviation around mean age, it's 8.6, whereas the standard deviation around the glucose level is 21. So is 8.6 small and is 21 large? We don't know.

### Empirical Rule:



If the distribution of a variable is *unimodal* and *symmetric* distribution, then

Mean  $\pm$  1 std. devs covers approx 67% obs.

Mean  $\pm$  2 std. devs covers approx 95% obs

Mean  $\pm$  3 std. devs covers approx all obs

Let's see if we can add a little bit more intelligence to that answer. This leads us to the empirical rule.

The empirical rule says, if the distribution of a variable is unimodal, what does that mean? Well, if we look at the distribution, we get something like a terrain with a single hill. It's unimodal. There's only one mode. The mode is the most popular value.

It's not bimodal. So bimodal might look like a camel's back. You've got a very popular value here and another locally very popular value there. That's called bimodal.

And this might be an indication that you've got a mixture of two populations in your distribution here. You might have one that's, say, predominantly short and another group that's predominantly tall. And that is probably better handled as two populations intermixed. Whereas, this population is more descriptive of a single group. So that's unimodal.

Then it says also that it has to be symmetric. So what is symmetric? Well, we looked at this earlier. And we said symmetry is judged by drawing a vertical line in the middle of the distribution, and spinning the distribution around that axis, you don't notice any difference. So basically it's what's going on on the left-hand side is the same as goes on on the right-hand side of that axis.

So if the distribution or variable is unimodal and symmetric—so this doesn't apply to everything. It doesn't apply to every distribution we have out there. It only applies to distributions that are unimodal and symmetric.

Then here's the magic. Create an interval by taking the mean and subtracting the standard deviation to create the left endpoint of the interval, and adding to the mean the standard deviation to create the right endpoint, then this interval contains approximately 67 percent, or  $2/3$ , of the observations. It covers approximately  $2/3$  of the observations.

So here's the magic. We look at the mean. We add its standard deviation, subtract the standard deviation. And now we've got an interval. And that's the interval where about  $2/3$  of the data fall.

So it's a rule. And that's almost like traffic laws in Italy. They are suggestive. The red light means it's sort of suggested that you should stop at that traffic light. But here, it's a rule.

And if we want to be a little bit more expansive, we could say mean plus or minus two standard deviations. And that will capture 95 percent of the observations. Mean plus or minus three standard deviations will capture all the observations.

This is the value of the standard deviation. It tells you how variable the data is. Whether it's small or large, all depends on this interval.

For example, how big is the mean plus or minus two standard deviations? That will depend on the context. If the context is such that this is too large an interval, then the standard deviation is too large.

If the context is such that the standard deviation gives you a small enough interval, then that's fine.

So whether a standard deviation is large or small depends very much on the context that you are dealing with.



Returning to the Framingham Heart Study dataset, we have that

```
. summ diabp1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
diabp1	4434	84.09303	12.97376	50	142.5

Mean  $\pm$  1 std. devs is 71.11927 , 97.06679

Mean  $\pm$  2 std. devs is 58.14551 , 110.04055

Mean  $\pm$  3 std. devs is 45.17175 , 123.01431

Let us investigate how well this rule does with the Framingham Heart Study data. For example, let's take a look at-- let's go back to the diastolic blood pressure at visit one. We get that the mean is 84.09, and the standard deviation is 12.97. So what the empirical rule says, if we can assume that the diastolic blood pressure is unimodal and approximately symmetric. Remember, we looked at this issue about the diastolic blood pressure to see whether it was symmetric or not. We'll come back to that in a minute, but if we can assume that, then, if we look at the mean plus or minus 1 standard deviation, we get this interval is from 71.11 to 97.07. If we calculate mean plus or minus two standard deviations, then the interval is from 54.15, let's say, to 110.04, and this interval should approximately encompass 95 percent of the data. For the mean plus or minus 3 standard deviations, we get roughly 45.17 to 123.01, and that should contain almost all the data.



The empirical rule; example

```
. summ diabp1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
diabp1	4434	84.09303	12.97376	50	142.5

should      actual

± 1 std. devs : (71.11927 , 97.06679)      67%      71%

± 2 std. devs : (58.14551 , 110.04055)      95%      95.99%

± 3 std. devs : (45.17175 , 123.01431)      all      98.89%

Compare the rule suggested results with what actually is happening, we get this last column in the slide above. There is good agreement between these last two columns, above.



The empirical rule; another example, log diastolic bp

```
. summ logdiasbp
```

Variable	Obs	Mean	Std. Dev.	Min	Max
logdiasbp	4434	4.420578	.1493413	3.912023	4.959342

should      actual      versus

± 1 std. devs : (4.27 , 4.56)      67%      70.3%      71%

± 2 std. devs : (4.12 , 4.72)      95%      95.15%      95.99%

± 3 std. devs : (3.97 , 4.87)      all      99.7%      98.89%

When we looked at the shape of the distribution of diastolic blood pressure we decided it was not quite symmetric; it had a bit of a long tail on the right. So we decided to transform the data and look at the logarithm of the diastolic blood pressure. When we

apply the Empirical Rule to the logarithm of the diastolic blood pressure we get the above results in the column headed “actual”. When we contrast that to the last column, which refers to the same calculations, but for the diastolic blood pressure, we see that it did make a little bit of a difference. It brought us down a little bit closer to the 67 percent and a little bit closer to the 95 percent and a little bit closer to the “all” by taking the logarithm, but not that much. So actually this empirical rule is robust to slight deviations from the assumptions of symmetry and unimodality. This is not a bad rule to abide by.

### Empirical Rule:



If the distribution of a variable is *unimodal* and *symmetric* distribution, then

$$t = \frac{\text{variable} - \text{mean}}{\text{standard deviation}}$$

Mean  $\pm$  1 std. devs covers approx 67% obs

Mean  $\pm$  2 std. devs covers approx 95% obs

Mean  $\pm$  3 std. devs covers approx all obs



### Empirical Rule:

If the distribution of a variable is *unimodal* and *symmetric* distribution, then

$$t = \frac{\text{variable} - \text{mean}}{\text{standard deviation}}$$

$|t| \leq 1$  approx 67% observations

$|t| \leq 2$  approx 95% observations

$|t| \leq 3$  approx all observations

One last thing before leaving the empirical rule, some people prefer to quote the empirical rule as follows. Create a new variable by looking at your old variable, subtract its mean and divide by its standard deviation. Call that t. And this is called a standardized (or Studentised) version of the variable.

Then the empirical rule says that we can compare this t quantity to an absolute number. Instead of saying mean plus 1 standard deviation, then these are the empirical rules. The mean plus or minus 1 standard deviation turns out to be just 1. In other words, this variable here, the standardized variable, has got standard deviation of 1. It's got a mean of zero and standard deviation of 1.

So this is maybe a simpler way to remember this, that mean plus or minus 2 standard deviations now becomes just t less than or equal to 2 in size, and mean plus or minus 3 standard deviations becomes t less than or equal to 3 in size. So this is an alternative way of stating the empirical rule.

The empirical rule; example					
					
<pre>. summ diabp1</pre>					
Variable	Obs	Mean	Std. Dev.	Min	Max
diabp1	4434	84.09303	12.97376	50	142.5

$t = \frac{\text{diabp1} - 84.09303}{12.97376}$	should	actual
$\pm 1$ std. devs : (-1 , 1)	67%	71%
$\pm 2$ std. devs : (-2 , 2)	95%	95.99%
$\pm 3$ std. devs : (-3 , 3)	all	98.89%

Just compare to t and what are the values that t takes. So t has to be less than or equal to 1, 2, or 3, and here we go. With our diastolic blood pressure, the mean was 84.09303, and the standard deviation was 12.97376, and so there's our t value. And then mean plus or minus standard deviations is between -1 and 1, mean plus 2 standard deviations is -2 and 2, mean plus 3 standard deviations is -3 and 3.

	
<p>In summary, the standard deviation, together with the mean, allows us to make summary statements about the distribution of our data.</p>	
<p><b>Do not forget the assumptions.</b></p>	
<p>There are other measures of dispersion, such as the range and interquartile range, and the mean absolute deviation, for example, that are sometimes used.</p>	

So, in summary, the standard deviation, together with the mean, allows us to make summary statements about the distribution of our data. Do not forget the assumptions. The assumptions are of symmetry and of unimodality.

There are other measures of dispersions, such as the range and interquartile range. Remember in the box plot, when we had the box plot like this? This is the interquartile range, the distance between the lower quartile and the upper quartile, and the MAD, the mean absolute deviation, and these are also sometimes used.

But, by far, the most ubiquitous is the standard deviation, and you see the value of this standard deviation.

Marcello Pagano

# [JOTTER 2 LIFE TABLES]

Material for week 2



## How long do we live?

Cannot answer for an *individual*,  
but can make a statement about  
a *group*.

Romans built on the observed  
constancies

[Halley](#) showed us how to calculate  
the [Lifetable](#)

The everlasting question of how long any one of us will live has no answer, of course, but the question is of importance not only to us, but to health policy professionals, to people planning retirement benefits, to insurance companies and to clinical investigators. Currently the best we can do is answer by studying the past, see what happened on average, and project that average onto the future. To enable us to do this in a principled manner, we turn to Halley's method for constructing the life table.

The real answer, of course, is we cannot answer the question, how long will I live? But what we can do is we can answer a question which is, how long do we live? So we can make a statement first about a group as opposed to an individual, and second that group's experience in the past. And that's the answer we provide to the question. How to make such statements so that they have some validity and value is what we now concentrate on.

The Romans observed some constancies that they were able to build on. For example, you need to be able to answer this question if you want to buy annuities, or if you want to buy insurance. How do you answer these questions? At first it was largely based on guesswork. Actually a lot of insurance companies went broke until Halley came along. And he showed us, about 300 years ago, how to calculate what is called the life table.<sup>1</sup> And we've been doing it that way for the last 300 years.

---

<sup>1</sup> E. Halley, An estimate of the degrees of the mortality of mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. *Philosophical Transactions of the Royal Society of London* 17 (1693), 596-610 and 654-656.



Lifetable tells us how long people live “on average”.

Useful for insurance companies, to determine annuities, plan for social security, etc....

It converts *cross-sectional* information into *longitudinal* cohort information.

The life table tells us how long people live on average, and more. It is used for making predictions of what our population will look like in the future.<sup>2</sup> These predictions guide, or, at least, should guide, our policies and plans for the future.

How the table is calculated is very cleverly done. It converts what we call cross-sectional information into longitudinal cohort information. Cross-sectional information is like taking a snapshot of today whereas longitudinal information like a film, or a video of the situation as it evolves over time.

So the idea is, we would love to track a population cohort for the next 100 years or so as it dies off, but we cannot afford to do that. We neither have the time nor the finances to do that. But we can take a snapshot today, or actually a short movie of two year's duration. Technically, we say we've got cross-sectional information and what we're really interested in answering this question is longitudinal or cohort information.

---

<sup>2</sup> [http://www.socialsecurity.gov/oact/NOTES/pdf\\_studies/study120.pdf](http://www.socialsecurity.gov/oact/NOTES/pdf_studies/study120.pdf)



## Average Life Span; or how long do we live?

Suppose we have a population of 10 people and we follow them till they die and here are their life spans:

1, 2, 10, 20, 35, 45, 50, 60, 70, 80

So the average life span is:

$$\frac{1+2+10+20+35+45+50+60+70+80}{10} = \boxed{37.3}$$

Let us start with an example with a population of ten individuals and calculate how long they live, on average. Here is how long each one of them lived, for some unit of time: 1,2,10,20,35,45,50,60,70 and 80. We can find their mean, which is 37.3.

What I want to do is I want to put some physical meaning into this and try a little demonstration for you in interpreting this calculation.

[VIDEO PLAYBACK]

Here we have some sticks. I've cut 10 of them, each stick represents one person, and the length of the stick is how long that person survived. So for example, this one is 80, because the longest was 80. And then there's the 70, the person who survived 70. All the way down to the two and the one. That little one there is a person, a little baby, only survived one unit of time.

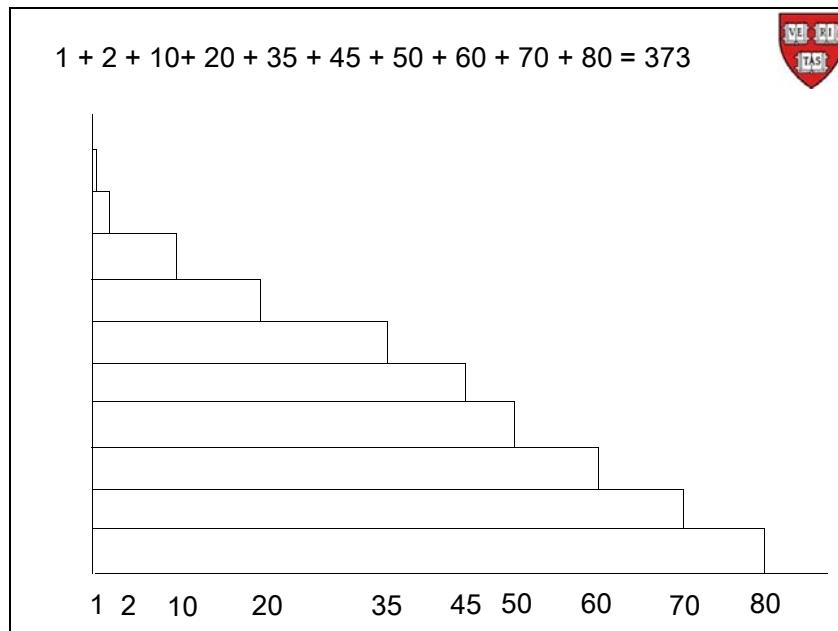
Now I want the average survival. So the mean survival should be the sum of all these lengths divided by 10. So one way to do this is I can lay these sticks out. So for example, there's the 80. So that's 80 units. And the next one, I put it in front. And so together they're 150 units. And then the next one, which was six, brings it up to 210 units. And so on, I string them all up like this, make sure I have all 10 of them. And then, what I can do is look at the final length here. And here's the little one. And then divide that by 10. And I've got 373 divided by 10. And so I've got that the average is 37.3.  
[Error in the film.]

Alternatively, what I could do is rearrange these sticks in a different direction. I can put the 80 at the bottom of the pile like this. And then I can, on top of that, put the 70. And

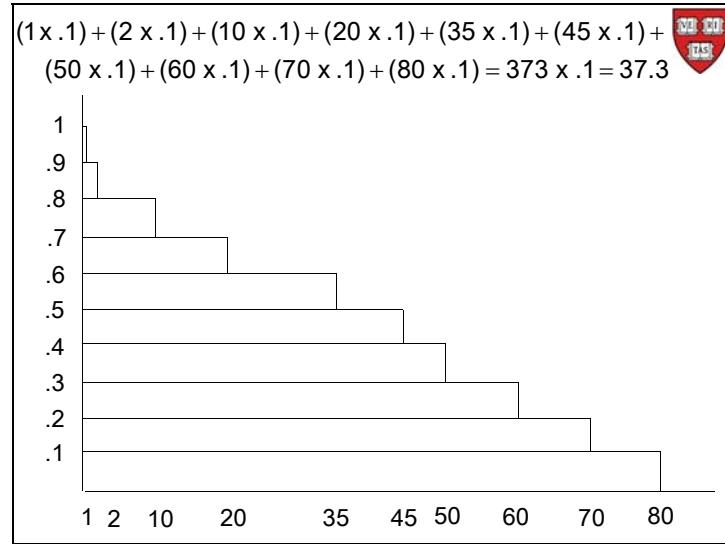
then on top of that, put the 60. And on top of that, the 50. And then the 45. And then the 35. And so on.

And now what I've got, I've got the same sticks. So the total length should be the same. But I've got them in a different configuration. So nothing's changed except the configuration. Let's leave the sticks and actually draw what happens before we explain anything further.

[END VIDEO PLAYBACK]

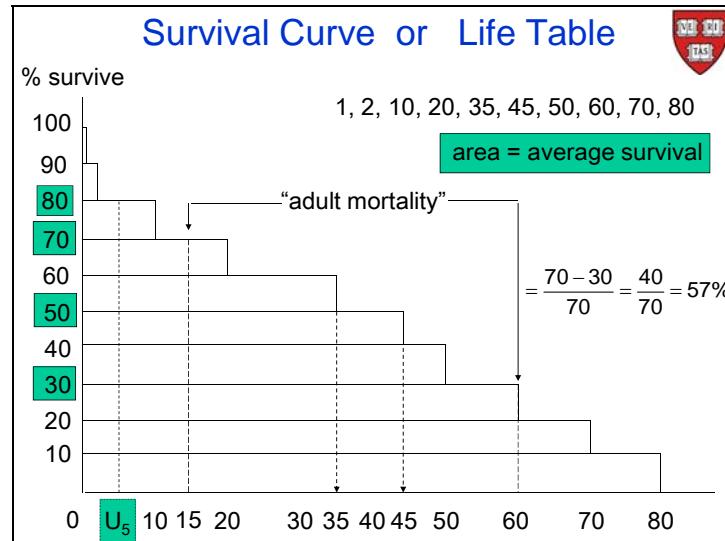


So here is an idealization of those sticks. Here, at the bottom, we have the stick that lasted 80 units. The next one up is 70 units, and so on. So here they are all 10 of them. And as we saw, if we put them lengthwise, it adds up to 373. But now what I want to do is I want to stack them like we did and here is the idealization. If I want to calculate the mean, we need to divide the 373 by 10.



Another way of thinking about this is to do the division before the summation. Now division by 10 is like multiplying by 0.1. Then you have to add two times 0.1, then add 10 times 0.1, and 20 times 0.1 and 35 times 0.1, et cetera. And that should give you 37.3. It's just simple arithmetic.

Consider what happens if we make the width of the stick 0.1. If we make the width of the stick 0.1, what is this quantity here, 80 times 0.1? 80 times 0.1 is just the area in the bottom rectangle. What about 70 times 0.1? 70 times 0.1 is the area under the second rectangle from the bottom. And so on. So that if we look at this curve, which starts off at one, because we had 10 sticks, each of width 0.1, and follow it all the way down, we get that the mean is just the area under that curve. This curve is called the survival curve or the life table.



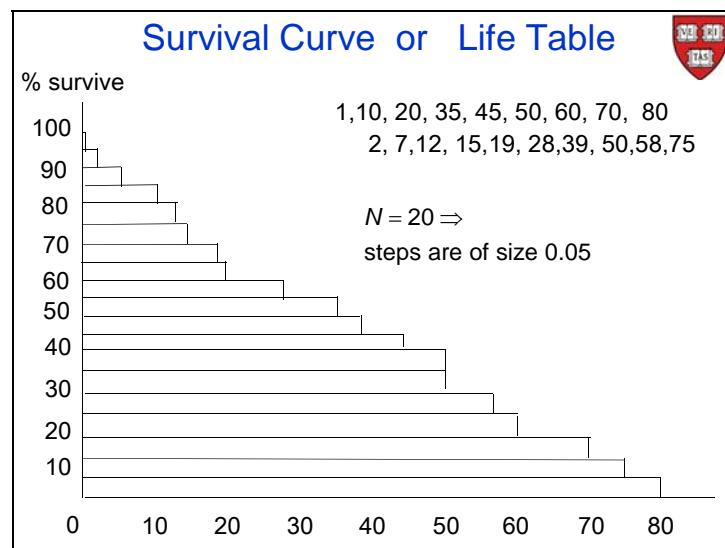
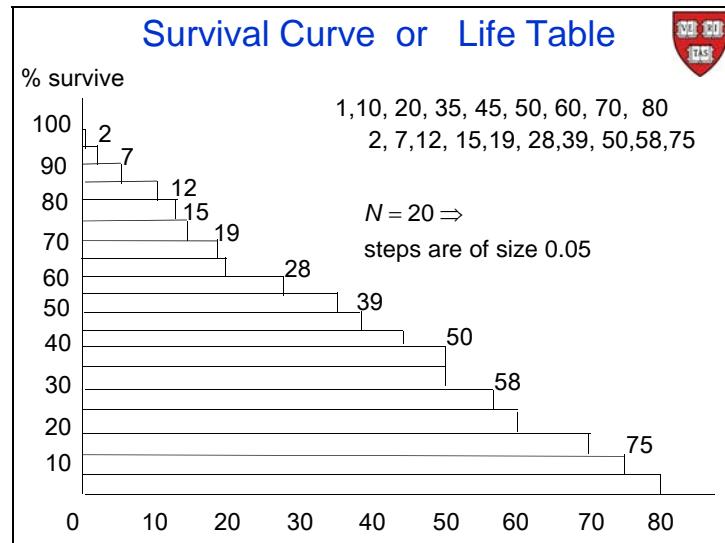
Change the X scale to have them equispaced so that they look more pretty and change the vertical scale to be percentages, so instead of one we will have 100%, 90%, 80%, et cetera. Then we can make the statement that the area is equal to the average survival. So the area under the survival curve is equal to the average survival. The units of the mean is the same as the units of the Xs.

We can also read off other quantities from this curve. For example, if we go up to 50%, on the vertical axis, then we can read off the median survival. Here, in this example, the median survival is going to be anywhere along the horizontal line at 50% which translates to 35 to 45. It's a flat spot, so any number between 35 and 45 will satisfy the definition of the median. Remember the definition of the median: at least 50% above, and at least 50% below. By convention, we can take the midpoint and the midpoint is 40. And that fits in well with the mean that is 37.3.

If this were a survival curve for a large group, such as a nation, then another summary that is often used is the under-five mortality. The way we get that is we look at age five on the horizontal axis and see what that corresponds to on the vertical axis. Here we see that 80% survive to age five. In other words, 20% do not. So the under five mortality in this example, is 20%. Very often you'll find this in the literature to summarize the health experience for a country.

Another summary statistic is the adult survival. The adult mortality, or adult survival, is asking the question, what happens between the ages of 15 and 60? So at 15, there are 70 people who make 15 and survive almost to 60, is 30%. So the adult mortality is defined as 70 minus 30—that is how many die between 15 and 60—and divide by the 70 that originally were there. And that gives us 57%. So with this set of data, the adult mortality is 57%.

So those are three ways of gleaned information from this survival curve, summary numbers associated with the survival curve.



Now what happens if instead of 10 people, we have 20 people. So suppose we got this extra data. And we got a new person at 2, 7, 12, 15 and so on. Then you can see what happens. Now that we have 20 the step sizes become 1/2 of what they were before. So the survival curve now looks something like this.

And you can imagine that the more and more and more people we get, the smoother and smoother this survival curve becomes. And we'll end up with something like this:



So as the group size gets bigger the steps get smaller and smaller, eventually leading to a “smooth” curve.

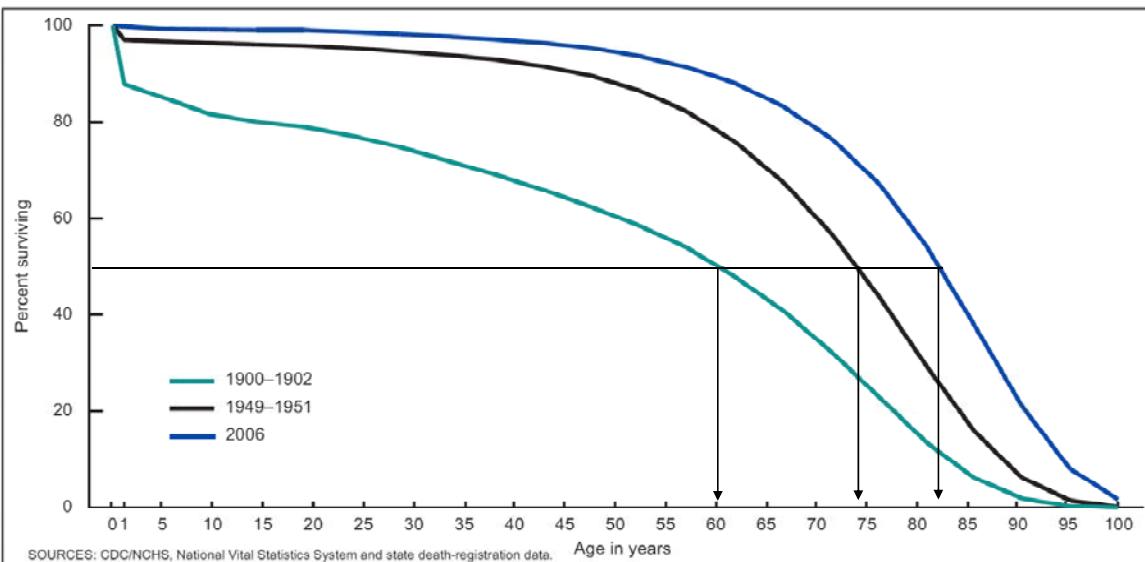


Figure 3. Percentage surviving, by age: Death-registration states, 1900–1902, and United States, 1949–1951 and 2006

So for example, here are three survival curves. And they all refer to the USA. The bottom one is the survival curve from 1900 to 1902. The middle one is the middle of the century, and it refers to the period 1949 to 1951. And the top one refers to 2006.

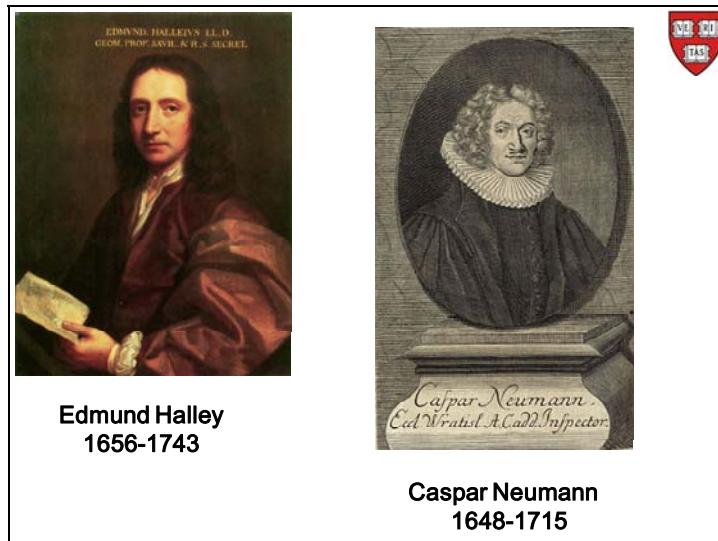
So here we go across the 20th century. And you can see what has happened to our survival experience. The first thing we notice is that the huge infant mortality, between the ages of zero and one, is slowly disappearing from the turn of the last century, to its middle to the beginning of this century. So here it was at the turn of the century. It's almost gone away. What that has done is it's sort of brought this curve up so that when we look at the average, remember the average is the area under the curve, in the first 50 years of the century, we can see how much we added to the average. In the second half of the century, we added even some more but not as much as we did in the first half of the century.

Another way of looking of looking at the average is to look at the median. As we cut across these three curves we see that at the beginning of the century it was 60, went up in the middle of the century to about 74 or so, and now it is at about 82, or 83. So we can see this progressive improvement as summarized by the median.

Now the question is how did we construct these curves? These particular ones were constructed by the National Center for Health Statistics. Did they wait 100 years, or so to construct them? Clearly not.

### Constructing a life table

To learn how these curves were constructed, we have to go back in time to Halley, Edmund Halley, he of comet fame.

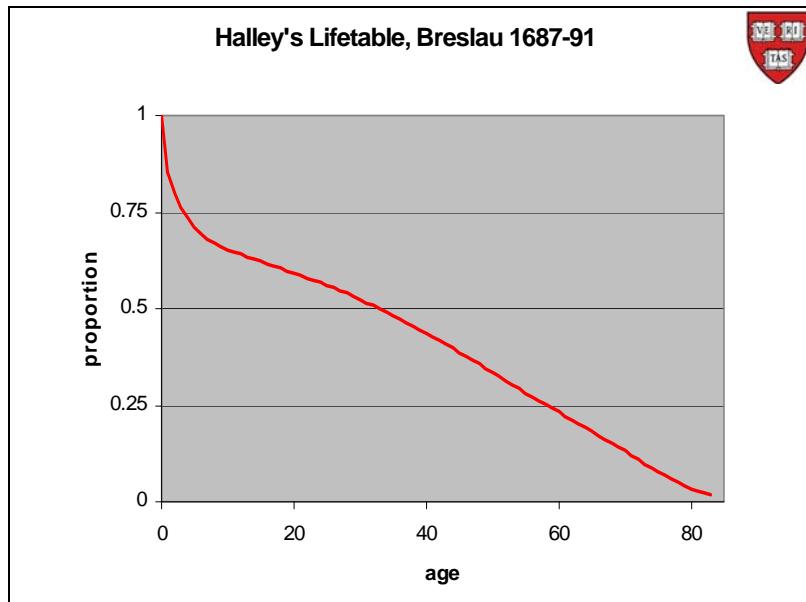


Edmund Halley wrote two papers in statistical methods, or probability, however you want to label them, at the end of the 17th century. And these papers presented his approach to tackling the problem of how to construct the life table. And it just so happened, that the reason why he was approached with the problem, is that someone, by the name of Caspar Neumann, had been collecting some fascinating data.

Neumann was way ahead of his time. What he had done, was collect five years of mortality data from the town of Breslau, which is where he lived, which I think at that time was part of Germany. Now it's part of Poland. He was way ahead of his time, first by collecting data on people, and second, in testing a statistical hypothesis.

He was a religious minister and he was very much interested in astrology. Apparently there is some theory in astrology that says the alignment of the planets when you are born is the same as when you die, or some such. I don't know precisely as I am ignorant of astrology. Anyway, he wanted to test this astrological theory, so he went about collecting data, to test the theory.

He tested it, showed that astrology was nonsense, and at the end of his investigation remained in possession of some wonderful data; five years' worth of it. So a member of the Royal Society sent these data to Halley, who constructed the very first life table using methods that we still use to this day.



And here is Halley's life table of Breslau at the end of the 17th century. And you can see that initial mortality, if one is to trust these data, that initial mortality is extremely high. Only about 66%, of those born reach age 10. So it is a big and precipitous drop. After that, the drop off is almost like a straight line. That's the life table.



Age. Curt.	Per- sons.										
1	1000	8	680	15	628	22	586	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt.	Per- sons.										
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Halley's Lifetable 1693

PHILOSOPHICAL  
TRANSACTIONS:  
Giving some  
**ACCOUNT**  
OF THE  
Present Undertakings, Studies and Labours  
OF THE  
**INGENIOUS,**  
In many  
Considerable Parts of the WORLD.  
FOL. XVII. For the Year 1693.  
L O M B O N .  
Printed for E. Smith and B. Wycliff, Printers to the Royal  
Society, at the Printers' Arms, in St. Paul's Church-yard, 1694.

Here it is in tabular form. In this form it is easier to explain what Halley did. He started off with 1,000 people. Now take the columns in pairs, and work your way down a pair of columns. So, for example, with the first two columns you start off with 1,000 people in their first year of life. How many will reach into their second year of life? He calculated that number to be 855. Of those, how many reach their third year of life? He calculated 798 would. And of those, 760 will reach the fourth year of life. And so on. Next we read the next pair of columns starting at age 8 and so on, working our way down the columns and then to the right. Eventually we see that of the original 1,000 people born, 20 reach their 84<sup>th</sup> year.

What he noticed, and what was really ingenious about this, is that he did it step by step. First we can ask, how did he go from 1,000 to 855? What he did was he observed what proportion of all the kids who in those five years died in their first year of life to get the mortality rate in the first year of life. That is the mortality rate he used to get the 855. Then he looked at the proportion of all kids who reached their first birthday but did not reach their second birthday. He then applied that proportion to the 855 to get the number 798. And then what proportion of those who reached their second birthday, reached their third birthday. And so on.



How are these curves constructed? Did we wait 100 years?

Concentrate on the 2006 curve between the ages of 65 and 70.

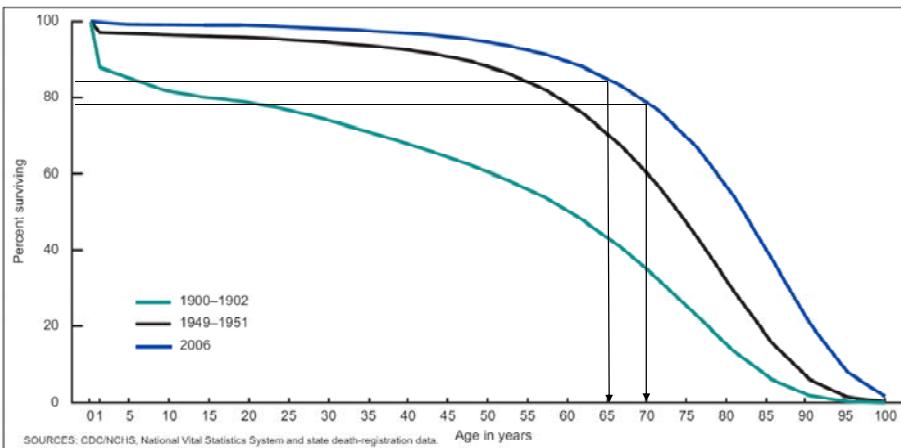


Figure 3. Percentage surviving, by age: Death-registration states, 1900–1902, and United States, 1949–1951 and 2006

We can apply the Halley logic to the modern tables. Focus on the 2006 curve, and ask what happens between the ages of 65 and 70, for example? We first concentrate on the people who have survived to age 65. Suppose we know that point on the curve. Then how do we get the point above age 70? Well if you just tell me what percentage of 65-year-olds will die before they reach 70, and that will give me the point on the curve above the age 70. Then I can repeat that logic to get to the next point, say above 71, and so on. Like that, we can build up the whole curve.

### Life Table



We don't know what will happen to the 65 year-olds in  $2007 + 65 = 2072$ ,  
BUT we do know what happened to the 65 year-olds in 2007, and the 66 year-olds in 2007 and the 67 year-olds in 2007 etc.....

The life table methodology constructs a life-experience for a cohort subjected to **current** mortality rates as it progresses through life, as if the current rates do not change.

Indeed, in 2007 we saw:

That is precisely what Halley did. He said to himself, we don't know what will happen to the 65-year-olds in 2072 because we're not there yet, and we cannot follow this cohort until 2072. But we do know what happened to 65 year olds in 2007. We'll just follow them up for a year or two and we'll see how many die in their subsequent year of life. Similarly I can also tell you what happened to 66 year olds in 2007. I can also tell you about 67 year olds in 2007, and so on.

The next step is to ask if we can project those age-specific mortality rates into the future. One way would be to assume they remain constant into the future. And that is exactly what Halley did.

So the life table methodology constructs a life experience for a cohort. It's a fictional cohort. But it's a cohort subjected to current mortality rates, as that fictional cohort progresses through life. As if the current rates don't change. It is a construct.

Age	Probability of dying between ages $x$ to $x + 1$	Number surviving to age $x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
0–1 . . . . .	0.006761	100,000	676
1–2 . . . . .	0.000460	99,324	46
2–3 . . . . .	0.000286	99,278	28
3–4 . . . . .	0.000218	99,250	22
4–5 . . . . .	0.000176	99,228	17
5–6 . . . . .	0.000164	99,211	16
6–7 . . . . .	0.000151	99,194	15
7–8 . . . . .	0.000140	99,179	14
8–9 . . . . .	0.000124	99,166	12
9–10 . . . . .	0.000105	99,153	10
10–11 . . . . .	0.000091	99,143	9
11–12 . . . . .	0.000094	99,134	9
12–13 . . . . .	0.000132	99,125	13
13–14 . . . . .	0.000209	99,112	21
14–15 . . . . .	0.000314	99,091	31
15–16 . . . . .	0.000426	99,060	42
16–17 . . . . .	0.000529	99,018	52
17–18 . . . . .	0.000627	98,965	62
18–19 . . . . .	0.000715	98,903	71
19–20 . . . . .	0.000796	98,832	79

So the second column in this table for 2007 tells us what age-specific mortality rates were actually observed from the one-two year followup starting in 2007 for the first 20 years of life. That's because I can only put 20 years of life on this slide.

If, in 2007, I look at all the babies born in 2007 and I follow them up for one year—so that means for babies born on January 1, 2007, I have to follow up until January 1, 2008, for babies born on January 2, I have to follow up to January 2, 2008, etc.. So every baby is followed for a year. That means it takes me two years to collect my data. Because for the babies born on December 31, I have to follow them up for two years. So for all babies born in 2007 the proportion who passed away within one year of being born is 0.006761.

Now, what about kids who were born in 2006, and thus turn two in 2007.in their second year of life in 2007? What proportion of them died within a year of their first birthday? And the answer is, 0.000460. So, once again I need only follow them up for two years. So for the ones who turned one on January 1, 2007, I have to follow up for a year to 2008. All the way down to December, I have to follow up for a year to do that. So within two years, I get both of these points. And the same is true for those between the ages of two and three. So those in their third year of life, I find by following them up for two years, that I get this number, which is 0.000286. So that's the proportion who die.

So in fact, this whole column here, this whole first column here, all these numbers and all the way down to 100, or whatever the lower number is, is based on observing this cohort over this very, very short amount of time. Okay, so these are the actual age-specific mortality rates for the 2007 year. Not for the 2007 cohort, but for the 2007 year.

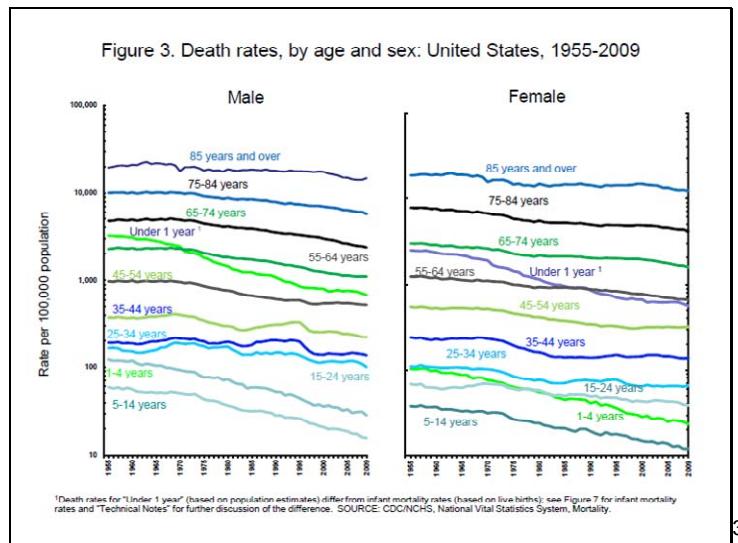
Now, to actually ask that other question, which is, all right, so after two years I know what happened to babies in their first year of life. What about the babies born in 2007 in their second year of life? Now I have to move into 2008, 2009 to get that. In the third year of life, I have to move into 2009, 2010. And so on, if I'm going to follow this cohort through life. And I can't. Very quickly, I run out of time because we're in 2012. All right? But, what I can do is I can apply these numbers I've got here to that cohort.

So let me construct a fictional cohort and subject it to the current mortality rates. So let me start with 100,000. Apply this proportion dying. So of those 100,000, we're saying, 0.006761 are going to die. So that gives me 676. So 676 are going to die in their first year of life. What does that leave with me with? Well 100,000 minus 676 which is 99,324. Let me look at this proportion of those that are going to die. So the 0.000460 times this will tell me how many of those kids will die in the following year. And that gives me 46. Subtract 46 from 99,324 gives me 99,278.

How many of those are going to die in the subsequent year? I multiply by 0.000286 and I get that 28. And subtract that from 278 and I get 99,250. And so on. And I keep on doing this.

So I'm creating a fictional cohort and applying today's age-specific mortalities to that. And what do I end up with here and carrying on down is my survival curve for this cohort of 100,000 people.

So remember, our survival curves start off with one. So to get that I just divide by 100,000 people. They introduced the 100,000 just to make the arithmetic look simpler. But just once you're done, just divide by 100,000 and then that's our survival curve.



3

The question then is, how good is this survival curve? Well we won't know for a hundred years or more, of course, but if we go back in time, we see that the age specific mortality rates are pretty constant. Most of them are going down. If there is a trend, that trend is going down. There are some exceptions such as the Vietnam War for males. If the trends continue to go down, then that will mean that our estimates of the survival curve are conservative and that people will actually live longer than predicted.

---

<sup>3</sup> [http://www.ssa.gov/oact/NOTES/as120/LifeTables\\_Body.html#wp1176553](http://www.ssa.gov/oact/NOTES/as120/LifeTables_Body.html#wp1176553)

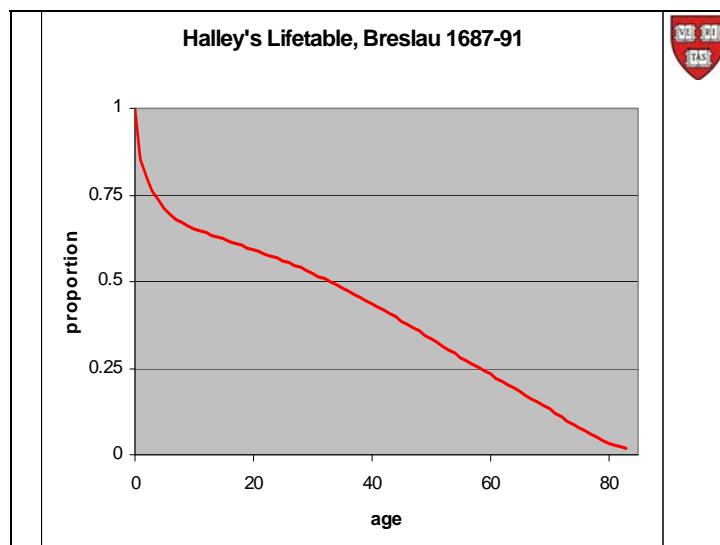
## Stable Population Assumption



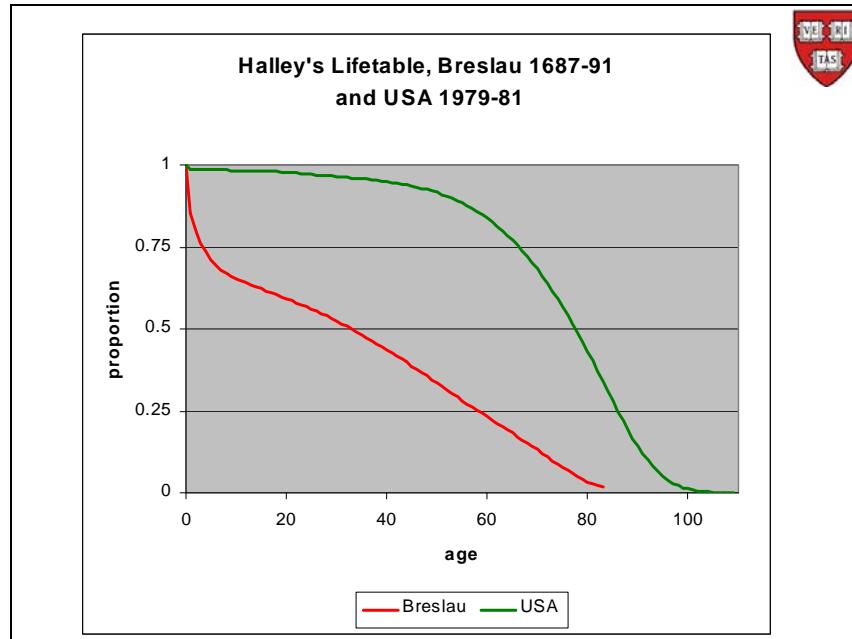
Halley did not have a population count for Breslau, so he made the extra assumption that the population was **stable** over the five years that Neumann counted deaths meaning that he assumed that the number of deaths equaled the number of births.

He actually then went on to infer the population size of Breslau during that period, from his lifetable!

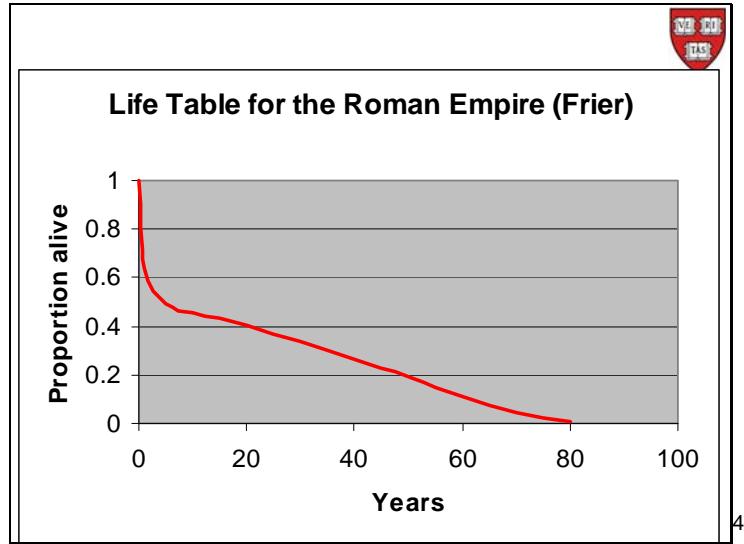
I should add that Halley made one extra assumption. For the 2007 life table we started from the proportions dying within age groups. To get those we had the number dying in the numerator and the number alive at the beginning in the denominator. Halley did not have a census so he did not know how many were alive, he had to estimate how many people were in Breslau. So he made to make an extra assumption, which was the stable population assumption, which basically says the number of kids born is equal to the number of kids dying within any one year. He recognized that he was a little bit off in his calculations because of that assumption. What we inherited from his work is a method for doing the calculations.



And I repeat, it has proved invaluable in that we still use it more than 300 years later.



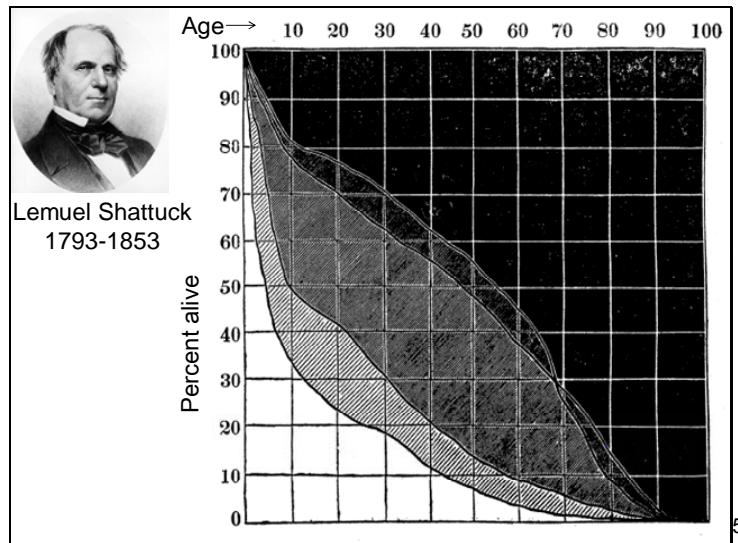
Here is Halley's life table of Breslau for the period 1687—91; it is the red line. We have superimposed the US life table for 1979—1981. We can see a few differences of what's happened. The biggest, of course, is what happens in the early years of life. What used to happen 300 years ago was much worse than what happens nowadays. Also note that now we sort of coast along until what, age 50 or so. And then we start dying at a higher rate. Whereas in Breslau's time, it was almost a constant rate from age 10 or so.



Here is what Frier thinks that the life table at the time of the Roman Empire looks like. And this is based on looking at tombs and records and seeing how old people were when they died and a very imaginative reconstruction of some data. If one is to believe these data, then we see a big precipice the first 5, 10 years of life and we are almost down to 40% by age ten.

---

<sup>4</sup> <http://www.richardcarrier.info/lifetbl.html> : Landlords and Tenants in Imperial Rome, Bruce W. Frier, Princeton University Press, 1980.



This one is my favorite. This is out of a report by Lemuel Shattuck. Lemuel Shattuck was the chair of a committee that did a survey of Boston in the middle of the 19th century, and he apparently wrote most of the report. He included only one graphic in all of it and it is this one.

He was making a case for improving public health in Boston. He attacked the quality of the air, the filth in the abodes, and the general poor level of hygiene in the city. Shattuck is known as the father of public health in this country. He outlined what a state department of public health should look like and, in fact, Massachusetts was the first state public health department and it based on his blueprint. Unfortunately, it happened after he died so he didn't get the credit that he deserved.

This graph shows four separate life tables. He chose a city in England that had roughly the same level of industrialization as Boston, Preston, and over roughly the same period of time. That is the top curve here. Be careful because here it crosses the next curve. The next curve down is Newton, Massachusetts, a rural town at the time. You see that Newton and Preston are very similar and as I said, by age 70, they crossed over.

Now the surprise is that just below that we get the life table for Boston. We see that, in Boston, about 50% of all kids died by age 10. And this he was attributing to the pollution and the living conditions and sanitation, et cetera, for the inhabitants of Boston.

---

<sup>5</sup> Shattuck picture: <http://www.mass.gov/eohhs/consumer/physical-health-treatment/health-care-facilities/public-health-hospitals/shattuck/mission-and-history.html>  
 © 2012 Commonwealth of Massachusetts.

He also makes the point that the poorer people had it even worse. The bottom curve is still Boston, but it represents the Catholics in Boston. That life table goes down to 30% by age ten. So by age 10, 70% of all kids died.

So a very powerful picture that actually did move the legislature into action.



Dean Briggs of Harvard says, "The peculiar evil in cigarettes I leave for scientific men to explain; I know merely that among college students the excessive cigarette smokers are recognized even by other smokers as representing the feeblest form of intellectual and moral life."

Healthy Living, Book Two  
C.A. Winslow 1920 p.185

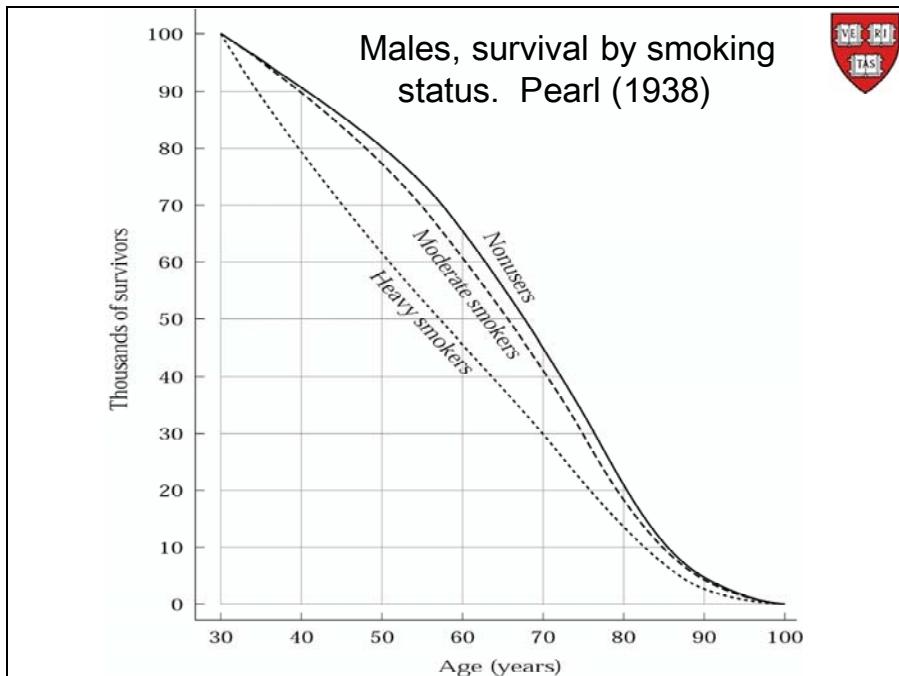
Another example, this time having to do with smoking. Here is what the dean at Harvard, Dean Briggs, had to say about smoking. And this was reported in 1920, and it is surprising that they recognized the evil in cigarettes as far back as then:

"I leave for scientific men to explain. I know merely that among college students the excessive cigarette smokers are recognized even by other smokers as representing the feeblest form of intellectual and moral life."

It was not until 1939 that Muller came out with his famous, but largely ignored, study that showed that smoking cause lung cancer.<sup>6</sup>

---

<sup>6</sup> Muller FH. Tabakmissbrauch und Lungencarcinom. Z Krebsforsch 1939;49:57-85.



Around that time, Ray Pearl published an article in 1938, of a study he carried out where he divided a cohort of men into three categories: “nonusers”, moderate smokers, and heavy smokers. Above are the survival curves derived from these three groups.

He started the table at age 30 primarily because he was interested in the effects of smoking, so he concentrated on adults. What we see is a much steeper drop for the heavy smokers when compared to the other two groups. The differences between the moderate smokers and the nonusers is there, but not as pronounced. If we look at the median survival, we see that the heavy smokers have a median age of about 57, or so, and the nonusers have a median age of about 67. So roughly speaking, just by looking at medians, one loses about 10 years of life on average because of membership in the heavy smoking group.

There is one other interesting observation to be made, and that is that death, like taxes, is inevitable. It happens to all of us. So what he drew here is that by age 100, everybody is dead. OK. But look at what's happening here. If we look at say, age 70, there's only 30% of heavy smokers left. So if we started with 100 heavy smokers, there are only 30 left by age 70, whereas for the nonusers, there are about what, 45% left at age 70. That means there would be 45 left of the original 100.

Now here's the interesting thing. By age 100, they're all going to die. So in the next 30 years, between 70 and 100, these 30 heavy smokers will die, whereas in the same amount of time 45 nonusers die. That tells us that if you have the same amount of time

to see 30 die as 45, then the 45 must die at a faster rate. That says to us that the mortality rate for the nonusers is going to be higher, than it is for the smokers.

This is an example of what has been called the healthy worker effect.<sup>7</sup> Summarised by Pearl as, "This is presumably an expression of the residual effect of the heavily selective character of the mortality in the earlier years of the groups damaged by the agent (in this case tobacco). On this view those individuals in the damaged groups who survive to 70 or thereabouts are such tough and resistant specimens that thereafter tobacco does them no further measurable harm as a group." [ Science, March 4 1938]

**TABLE 1**  
**THE DEATH RATE (1,000  $q_x$ ) AND SURVIVORSHIP ( $l_x$ ) FUNCTIONS, AT FIVE-YEAR INTERVALS, STARTING AT AGE 30, OF**  
**(a) NON-USERS OF TOBACCO; (b) MODERATE SMOKERS**  
**WHO DID NOT CHEW TOBACCO OR TAKE SNUFF;**  
**(c) HEAVY SMOKERS WHO DID NOT CHEW TOBACCO**  
**OR TAKE SNUFF. WHITE MALES**

Age	Non-users		Moderate smokers		Heavy smokers	
	1,000 $q_x$	$l_x$	1,000 $q_x$	$l_x$	1,000 $q_x$	$l_x$
30 . . . . .	8.18	100,000	7.86	100,000	16.89	100,000
35 . . . . .	8.78	95,883	9.63	95,804	21.27	90,943
40 . . . . .	10.01	91,546	11.89	90,883	23.91	81,191
45 . . . . .	12.04	86,730	14.80	85,129	25.69	71,665
50 . . . . .	15.16	81,160	18.61	78,436	27.49	62,699
55 . . . . .	19.82	74,538	23.67	70,712	30.09	54,277
60 . . . . .	26.73	66,564	30.49	61,911	34.29	46,226
65 . . . . .	36.88	57,018	39.83	52,082	41.20	38,328
70 . . . . .	51.69	45,919	52.84	41,431	52.72	30,393
75 . . . . .	73.02	33,767	71.28	30,455	72.33	22,838
80 . . . . .	103.22	21,737	97.95	19,945	100.44	14,494
85 . . . . .	142.78	11,597	136.50	10,987	139.48	7,865
90 . . . . .	197.49	4,753	190.23	4,686	193.68	3,292
95 . . . . .	273.2	1,320	265.1	1,366	268.9	938

Here are the actual numbers used by Pearl. So if you look, by age 70, of the non-smokers, there are 45,919. Of the moderate smokers, there's 41,431. And of the heavy smokers, there's 30,393. The mortality rates are roughly the same at age 70; for the three groups they are, 51.69, 52.84 and 52.72 per 1,000, respectively. At age 75, the mortality rates for the nonusers is 73, which is higher than for the heavy smokers, which is 72. For the nonusers at age 80, it is 103, whereas it is 100 for the heavy smokers. At

<sup>7</sup> Summary of Several Male Life Tables, William Ogle *Journal of the Royal Statistical Society*, Vol. 50, No. 4. (Dec., 1887), pp. 648-652.

age 85, it is 143 versus 139, and at age 90 it is 197 versus 193. Finally at age 95 it is 273 versus 269 per 1,000.

Do not, by any means read this to mean that you should not smoke until age 70 and then start smoking.

You will see in your epidemiology studies that if the outcome determines, or impacts, your membership in a group, then it makes subsequent conclusions suspect.

Now the other interesting thing is that his conclusion is that smoking is not very good for you. But what struck me when I first read this was that he published this in 1938. And he published it in a very reputable journal, Science. But somehow we ignored this message. It's all to the power of the tobacco industry that they were able to, because of advertising, and other means have such an impact on us.

### Healthy worker effect



Here, just as is usually the case in our experience in studies of this sort, the differences between the usage groups in specific mortality rates, as indicated by  $q_x$ , practically disappear from about age 70 on. This is presumably an expression of the residual effect of the heavily selective character of the mortality in the earlier years in the groups damaged by the agent (in this case tobacco). On this view those individuals in the damaged groups who survive to 70 or thereabouts are such tough and resistant specimens that thereafter tobacco does them no further measurable harm as a group.

### Smoking tobacco effect



However envisaged, the net conclusion is clear. In this sizable material the smoking of tobacco was statistically associated with an impairment of life duration, and the amount or degree of this impairment increased as the habitual amount of smoking increased.

#### Tobacco Smoking and Longevity

Raymond Pearl

*Science*, New Series, Volume 87, Issue 2253  
(Mar. 4, 1938), 216-217.

WHO

<http://www.theglobaleducationproject.org/earth/index.php>

I'd like to direct you to the World Health Organization. They have a lovely little website that does a lot of these calculations around the world, what life expectancy is around the world. So, go to the website and you'll see the address just above.

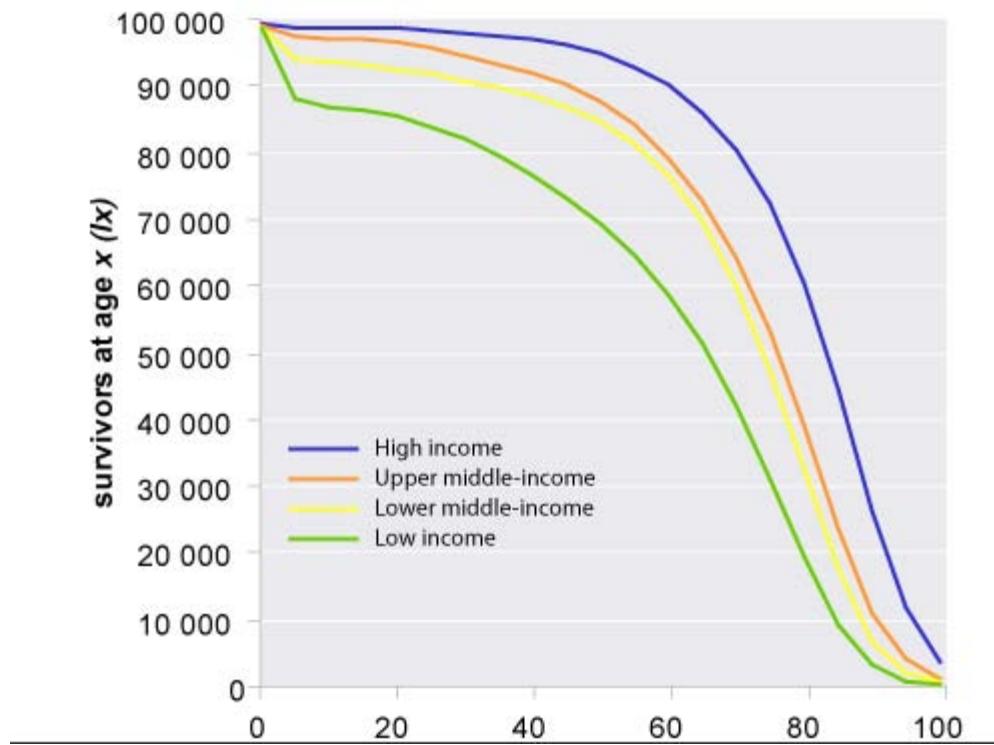
Once you get there, you should be able to view interactive graphics that are full of information.

You see a map of the world that allows you to zoom in on certain parts of the world. If you scroll over the country, you can see various information about the country. So, for example, Brazil-- life expectancy in Brazil is 73. If you go to India, life expectancy is 65, and so on.

And you can rank the countries alphabetically on the left. Or you can rank them by life expectancy. You can see that Malawi has the lowest. Afghanistan, Zambia, et cetera-- the lowest ones. If you want the highest ones, you just click again. We see that the highest is in Japan, and San Marino Republic in Italy, et cetera.

So you can have fun with this. And it's very instructive. And you can see how things vary around the world.

So it's worth a few minutes of your time to learn a lot about how life expectancy is distributed around the world.



[http://www.who.int/gho/mortality\\_burden\\_disease/life\\_tables/life\\_tables/en/index.html](http://www.who.int/gho/mortality_burden_disease/life_tables/life_tables/en/index.html)

One more site for you to visit to study some important differences in survival experiences around the world.

Marcello Pagano

### **[JOTTER 3 PROBABILITY]**

Definition of probability and its use in health models, including diagnostic tests and screening.

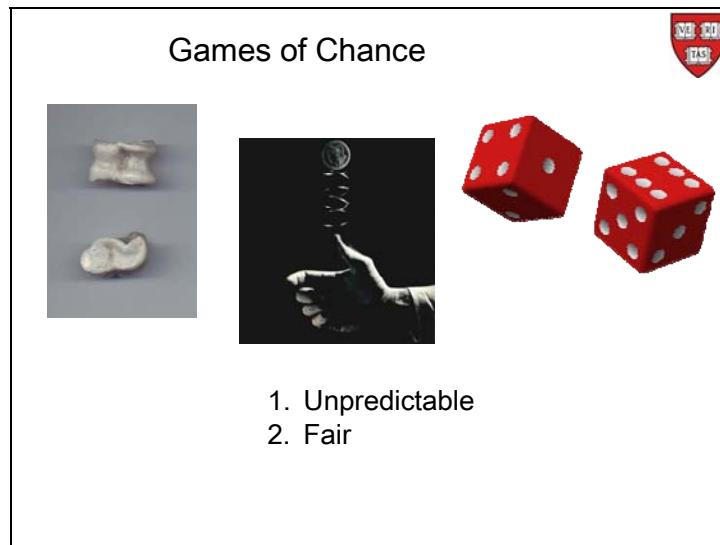
Doubt is not a pleasant condition, but certainty is absurd.

Voltaire 1894—1778

The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of times they are unable to account.

— Pierre-Simon Laplace      Introduction to *Théorie Analytique des Probabilités*

Probability is a way of quantifying uncertainty. We can use probability to measure risk so as to help us in decision making, and we also use it to assess our inference.



Historically, our main impetus for quantifying uncertainty was through games of chance<sup>1</sup>. One of the earliest instruments of chance is the astragalus, a bone which has four faces. You roll them out, and games can be based on which face is up and so on. Unfortunately, the astralagus was not ideal because some faces showed up more frequently than others thus making calculations difficult. So we idealized the situation by introducing dice.

A die has six faces with a one of the numbers, one through six, on each face—note the design for the five, it is called a quincunx. The idealization is that it is just as likely that

<sup>1</sup> It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

— Pierre-Simon Laplace *Théorie Analytique des Probabilités*

any one face will show when the die is rolled as any other face, and a number of games of chance depend on this ideal.

Lastly, to simplify things even more, we do not play dice too often, but we certainly quite often flip a coin to introduce an element of uncertainty into decision making.

There are two aspects of flipping a coin that are important. One is we flip it, for example, to choose who's going to serve in tennis, which side gets the sun, who gets the sun in their eyes, who gets to serve first. We flip because we cannot predict whether a head will come up or whether a tail will come up. (How useful would it be if a head always came up?)

So one, we've got this element of uncertainty. But equally important is this notion of fairness—flipping a coin is a fair way of deciding things.

What does that mean, to be fair? Certainly, if we flip the coin once, it is going to come up a head, or it is going to come up a tail. Excluding cheating, there is no way of telling which comes up.

The idea of fairness requires something more. In repeated trials we can find what we need: If we flip the coin repeatedly, then we expect that roughly half the time we shall get a head, and the other half we shall get a tail.

So this juxtaposition between unpredictability in the short run, but quite predictable in the long run is what makes probability fascinating and useful.



At any rate, I am convinced that He  
(God) does not play dice.

Albert Einstein (1879-1955)  
Letter to Max Born

God not only plays dice. He also  
sometimes throws the dice where  
they cannot be seen.

Stephen Williams Hawking (1942- )  
Nature 1975, 257.

It is also interesting how the acceptability of the notion of uncertainty has changed through the years. How it has now even become acceptable in the exact sciences, whereas it did not use to be. People much preferred certainty, determinism, instead of uncertainty, but we are learning.



In medicine, it has taken some time to accept the formalism of probability. What makes that ironic is that the first book on probability was actually written by a physician, by the name of Girolamo Cardano.

Cardano, because of his birth situation—he was born illegitimate, something that mattered in those days—was not accepted into the guild to practice medicine in Milano. So he turned to gambling on the side in order to earn a living. What he noticed—he was a brilliant man who wrote about 100 books—was these long run stabilities that he could take advantage of so as to outwit his opponents. The problem is it took more than 100 years for his book to get published, so it did not have the impact it deserved.



## Probability

- Element — **Event**
- Event — **set of descriptions**
  - proposition
  - **everyday sense**

An event can *occur* or *not occur*.

Use letters A, B, C, ... to denote events

We talk about the probability of an event happening, and an event is our elemental entity in the theory of probability.

An event is simply a set of descriptions; it is a proposition. We use it here in the everyday sense of the word. So just like when you were introduced to geometry, you started with the elemental point, and built on that to get a line, and then a plane, and so on, so too in probability. We start with an event and build up from there.

Events themselves we denote by capital letters, A, B, C, et cetera, and we focus on whether the event occurred or did not occur (or happened, or did not happen). Or in the future, will it or will it not occur. It makes it easier sometimes to think of the future, because you are then attempting to predict. But you can think retrospectively, too, just gets a little trickier.

In summary, we have that we are concerned about events that are unpredictable in the short run and much more predictable in the long run, and this juxtaposition is the essence of our use of probability.



## Operations on events

### 1º Intersect

The event "A intersect B," denoted  $A \cap B$ , is the event "both A and B."

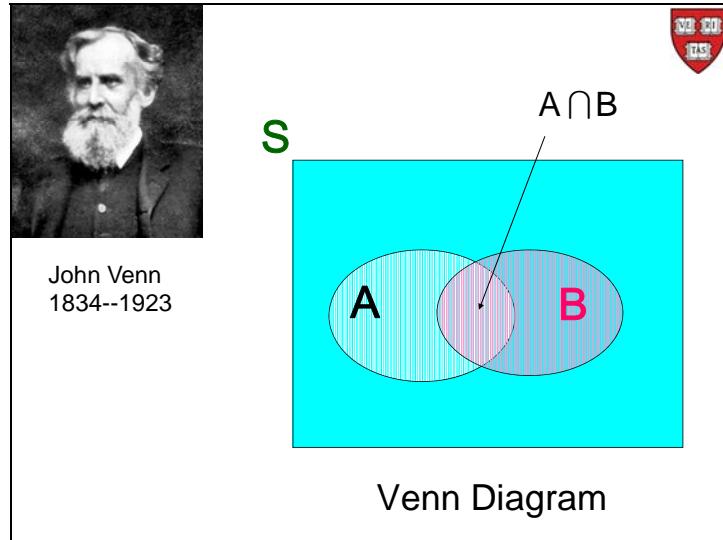
$A =$  "A 65 year old woman is alive at 70"  
 $B =$  "A 65 year old man is dead at 70"

$A \cap B =$  "A 65 year old woman is a widow at 70"

To repeat, probability is a measure defined on events. We have introduced events, so now let us start expanding on the grammar of events by combining events to create new events and thus create a whole library, a whole language with events.

The first operation we investigate is called the intersection: Given two events, say A and B, create a new event that occurs if both A and B occur. The notation we use to denote the new event is A upside down cup B;  $A \cap B$ .

For example, if the event A is that a 65-year-old woman is alive at 70, and the event B is that a 65-year-old man is dead at 70, and they are married, then the event A and B, denoted  $A \cap B$ , is that she will be a widow at age 70. Because we want both A and B to be true: she is alive at 70, and he is dead at 70.



To help us to think about events, the effect of combining them, and how to determine their probabilities, the philosopher John Venn created a wonderful construct that we now call the Venn diagram. It works as follows: You start off with the whole space,  $S$  that represents everything. Let us represent that by a rectangle. So any particular event you can imagine might be represented by a dot in this space  $S$ . Now consider the event  $A$ ; let us denote it by an ellipse. So if an event occurs that is inside of  $A$ , then we say the event  $A$  has occurred. For any event outside the ellipse  $A$ , we would say that the event  $A$  has not occurred. (Note how we use the same language whether we are talking about a single event or a collection of events.)

Now introduce the event  $B$ , another ellipse, say, and place it in  $S$ .

So we have four areas:

- (i) We start with the area outside of both ellipses. Events out there are neither  $A$  nor  $B$ .
- (ii) Now, think of the points within  $A$ . These can be classified as being in one of two groups: those not in  $B$ , and
- (iii) those in  $B$ .
- (iv) Finally, we have the events in  $B$  that are not in  $A$ .

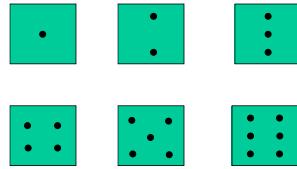
The intersection,  $A \cap B$ , is represented by the region where the two ellipses  $A$  and  $B$  (the events in (iii), above) overlap.

## 2° Union



The event "A union B," denoted  $A \cup B$ ,  
is "either A or B or both"

e.g.  
6 sided  
die

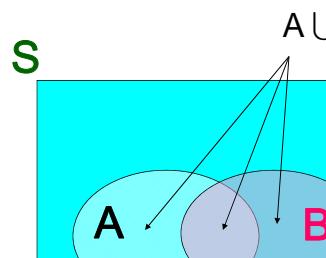


Roll twice:  
 $A = \text{"Roll a 7"} \quad B = \text{"Roll an 11"}$

$A \cup B = \text{"Roll a 7 or 11"}$

The other operation we need is the union. The union of A and B, denoted  $A \cup B$ , is the event that occurs if either A or B, or both, occur. It is not the exclusionary "or". It is either A, or B, or both.

So when we roll a six sided die, we either get a 1, 2, 3, 4, 5, or a 6. Suppose we roll it twice and sum the results of the two rolls. Define the event A to be we roll a 7, and the event B that we roll an 11. Then  $A \cup B$  is that we get a 7 or an 11 when we roll the die twice. So that is the union of two events.



Venn Diagram

Returning to our Venn diagram, we can see that  $A \cup B$  is the total space spanned by the joining of the two ellipses (to make almost a horizontal figure eight!).

Now we have the intersection that says that both have to happen at the same time, and the union that says either one or the other, or both happen at the same time.



### 3° Complement

“A complement,” denoted by  $A^c$ , is the event “not A.”

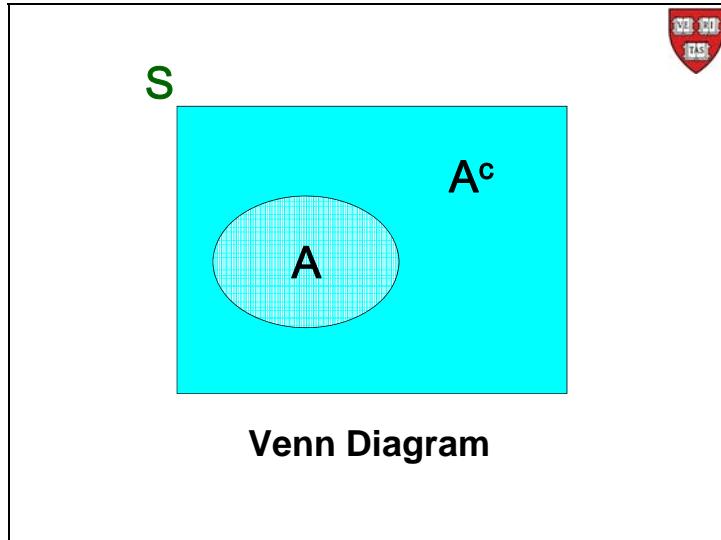
$A$  = “live to be 25”

$A^c$  = “do *not* live to be 25”

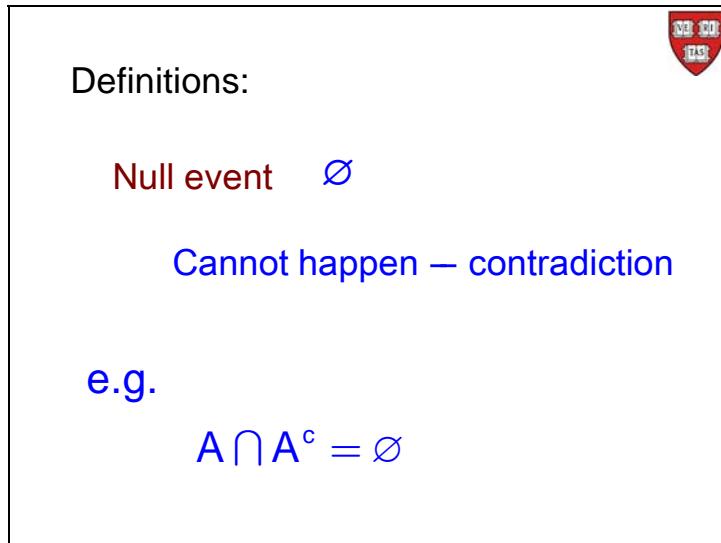
= “dead by 25”

,

Lastly, the third operation we need deals with the complement of an event. It is denoted by a superscript  $c$ ,  $A^c$ , and that denotes the event “not A”. So if the event  $A$  is, I live to be 25, then the event  $A^c$  is, I do not live to be 25; so dead by 25.



Returning to our Venn diagram, that part of the whole,  $S$ , that is not in the ellipse  $A$  is the event  $A^C$ .



It is amazing how much we can do just with those three operations. For example, we can define the null event, usually denoted by  $\emptyset$ . It is the event that cannot happen. It is a contradiction. So for example, the event  $A$  cannot happen at the same time as  $A^C$ . So  $A \cap A^c = \emptyset$ , the null event. It just cannot happen. You can't have it both ways.



Mutually exclusive events:

Cannot happen together:

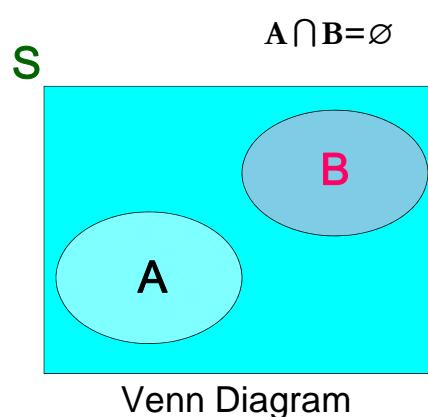
$$A \cap B = \emptyset$$

A = “live to be 25”

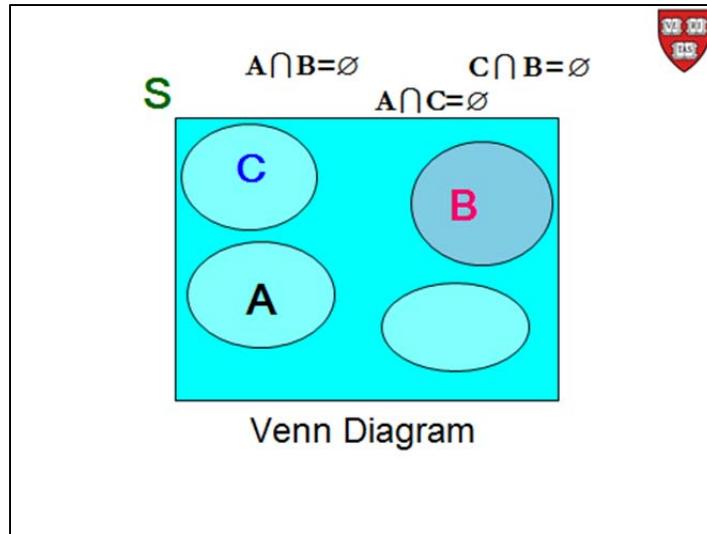
B = “die before 10<sup>th</sup> birthday”

This leads us to a very important collection of events, and these are mutually exclusive events. The definition of a pair of mutually exclusive events is that they cannot happen simultaneously. So, either one happens, the other, or neither, but you cannot have both of them happening together.

For example, if the event A is to live to be 25, and the event B is die before your 10<sup>th</sup> birthday. Then you cannot have both, and  $A \cap B = \emptyset$ . So this can be thought of as an extension of the idea of a complement.



In a Venn diagram, it means that the ellipses are separate.



This concept is extremely useful for attacking problems that are very complex, and involved, and where we have a huge amount of information. Then what we can do is break up the problem into little separate modules, knowing that only one of these modules can happen at a time. That then allows us to solve the problems one aspect at a time until we have the whole solved. This is the genesis of the idea of modularity. It is very important for computer programming, and other tasks, including how we present the material for this course to you! Here we call it mutually exclusive events.

## Probability

## Probability



If an experiment is repeated  $n$  times under essentially identical conditions and the event  $A$  occurs  $m$  times, then as  $n$  gets large the ratio  $\frac{m}{n}$  approaches the probability of  $A$ .

$$P(A) = \frac{m}{n}$$

$$\text{Odds of } A = \frac{m}{n-m} = \frac{P(A)}{1-P(A)}$$

Now that we have laid out the groundwork, we are ready to define probability, and we present what is known as the relative frequency definition of probability. There are others, but we do not pursue them.

The definition we are going to use is a practical one. Probability theory itself is a well-formed mathematical theory that is broad and fascinating and can keep us entranced for a lifetime, but we are not going to go down that path, here. We are going to focus on a practical use of the meaning of probability. Returning to our geometry analogy, when you first learned geometry, you learned about straight lines. We know that straight lines that abide according to the geometrical definition only exist in somebody's mind, whereas all around us all sorts of straight lines, that, for example, engineers construct in order to put up buildings, et cetera. Similarly, let us attempt a practical definition of probability.

We focus on the long range stability of events happening that Cardano, von Neumann and Halley took advantage of: Consider an experiment that we repeat  $n$  times under essentially identical conditions. For example, flip a coin over and over, under essentially identical condition. Now think of an event  $A$ , and suppose this event  $A$  occurs  $m$  times. Then, as  $n$  gets large, the ratio, or proportion,  $m$  over  $n$ , approaches the probability of  $A$ .

We have already noticed this in national mortality tables, except we did not call it probability we called them mortality rates, or proportions.

The definition is deliberately vague in certain spots, because we want a practical definition. So certain phrases such as identical conditions necessarily remain vague. What do we mean by that? Do we need to have the humidity exactly the same? Do we need to have the wind directions exactly the same? And then, what do we mean by  $n$

gets large? Is 3 a large  $n$ ? Is 3 million large? And what do we mean by approaches? Is this in a limit, as in mathematics?

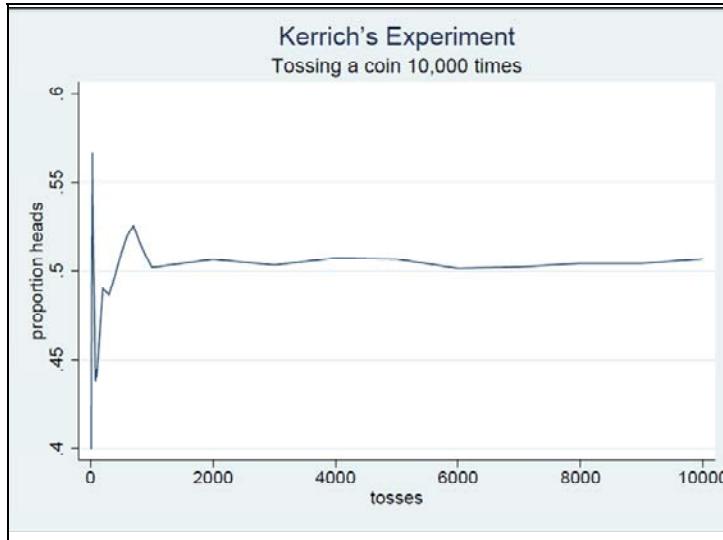
For example, we are going to be applying this definition when talking about people. We need to make statements such as; the probability that this patient will survive the operation is 0.4. Well, why is it 0.4? The answer possibly is, because 40% of previous patients have survived the operation. But were those prior patients identical to this patient? No, they were not, we are all individuals. But that is how we are going to apply this construct. This is the perpetual problem we face whenever we apply an ideal mathematical model to a practical and real situation. We must be prepared to understand when we can apply a model and when not to apply it.

Before leaving this definition, let us point out that there is a mathematically related quantity, and that is the odds of  $A$ . So rather than the probability of  $A$ , we can also talk about the odds of  $A$ .

The only difference is that the odds of  $A$ , which is also a ratio, is the ratio of the number of times  $A$  occurred to the number of times  $A$  did not occur. One should note that Cardano actually wrote his book entirely about odds! Nowadays one mostly hears odds mentioned only at the race track, sporting events, or, in this course, when we get to case control studies, et cetera. By and large, probably because of their training, people tend to favor probability, but the two are clearly related by a mathematical equation, and if you have the one you can uniquely retrieve the other, so their knowledge value is equivalent.

Another difference, which could be of minor importance, is that probability is symmetric around a half. At the edges we are certain—at zero we are certain it will not happen, at one we are certain it will happen. We have maximal uncertainty at the center, when  $p=1/2$ .

That point of symmetry ( $p=1/2$ ) with odds translates to one; below one we are arguing against the event happening, above one we favor the event happening. But the simple symmetry is no longer there, since, for example, at the left boundary, at zero, the odds are zero, whereas at the right boundary, at one, the odds are infinite. This is indicative of a more complex symmetry, namely the reciprocal symmetry—for example the odds of one third are one over the odds at 3, and thus the symmetry around the point one.



Just before the Second World War, a fellow by the name of John Edmund Kerrich was visiting Copenhagen in Denmark. The Germans invaded Copenhagen and Kerrich was South African, and so he was interred in Denmark for the duration of the war. Being imprisoned with nothing to do, and being a mathematician, he decided to start flipping a coin. He collected pieces of paper so he could keep tabs of his flips, and he had enough paper to note what happened when he flipped the coin 10,000 times.

Above you can see the results of his 10,000 tosses. Specifically, what is plotted is the proportion of heads as the number of flips increased. You can see that initially it is unpredictable (short range), and things go all over the place. But as time progresses (long range), the ratio stabilizes to almost a half (actually 0.5067 after 10,000 tosses of his *real* coin). He wrote a fascinating book about this experiment.<sup>2</sup>

---

<sup>2</sup> John Kerrich, *An experimental introduction to the theory of probability*, E. Munksgaard (Copenhagen) 1946. Also Reprinted in 1950 by BELGISK IMPORT COMPAGNI, LANDEMÆRKET 11, COPENHAGEN



Something must happen:

$$\frac{n}{n} = 1$$

$$\text{Pr(sure thing)} = 1$$

Impossible:

$$\frac{0}{n} = 0$$

$$\text{Pr(impossible)} = 0$$

So let us investigate some particular probabilities. First, something must happen. So if your event is that something happens (S) then if you repeat the experiment n times, something happens n times, so its probability is 1 and thus the probability of a sure thing is 1. On the other hand, at the other extreme, if the event is impossible then it never happens, so 0 out of n. So the probability of the impossible event is 0.



For any event A

$$m \leq n, \text{ so}$$

$$0 \leq \text{Pr}(A) \leq 1$$

Complement

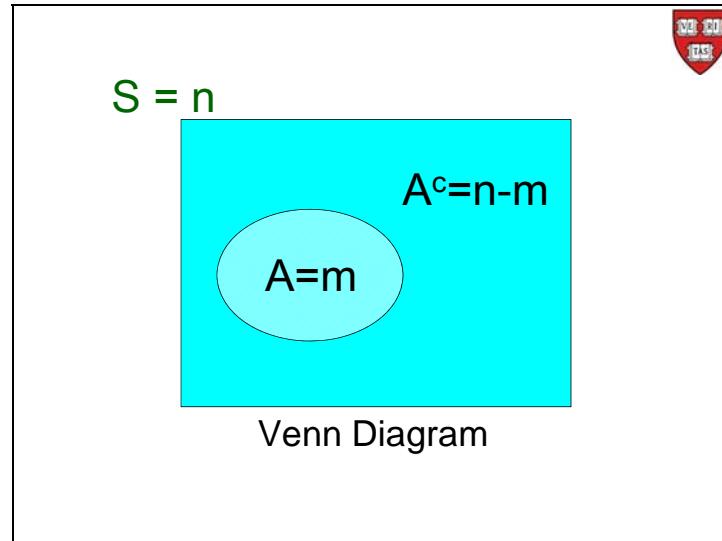
$$\text{P}(A) = \frac{m}{n}$$

$$\text{P}(A^c) = \frac{n-m}{n} = 1 - \text{P}(A)$$

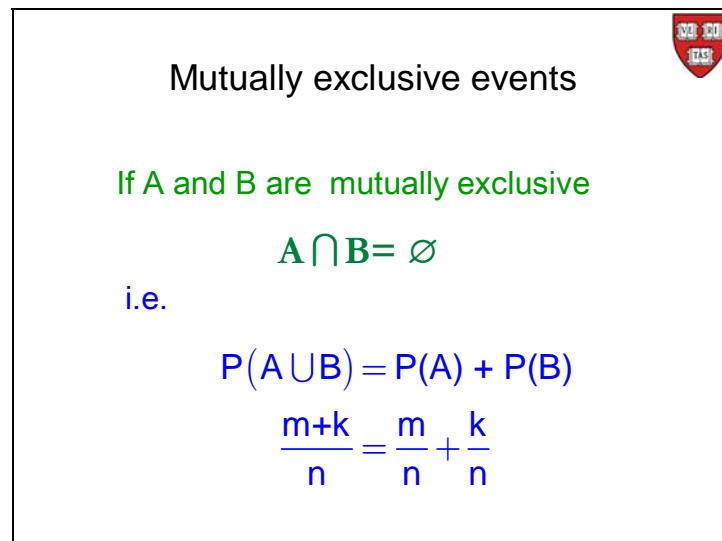
$$\text{P}(A) + \text{P}(A^c) = 1$$

Indeed, these two events define the limits, namely, the probability of any event is going to be between 0 and 1.

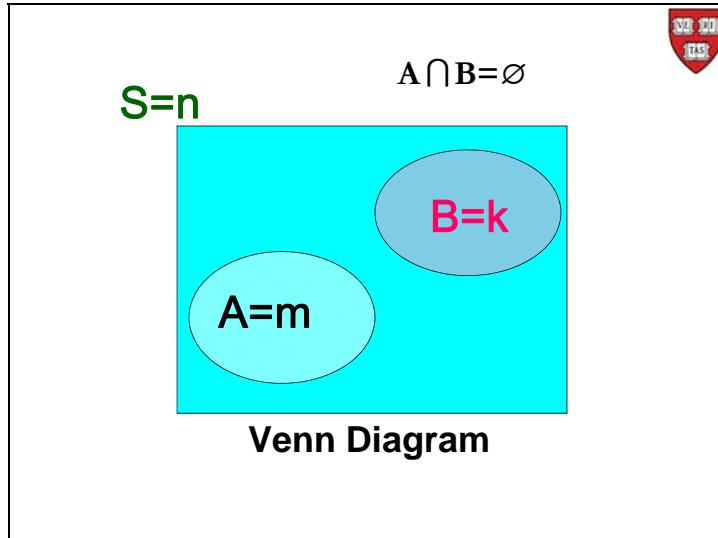
Remember the complement? Well, if A happened m times, that means the other n minus m times A did not happen, or  $A^c$  happened. So the complement happens n minus m times, and the probability of the complement is 1 minus the probability of the event. Or another way of saying that is the probability of A plus the probability of A complement is equal to 1.



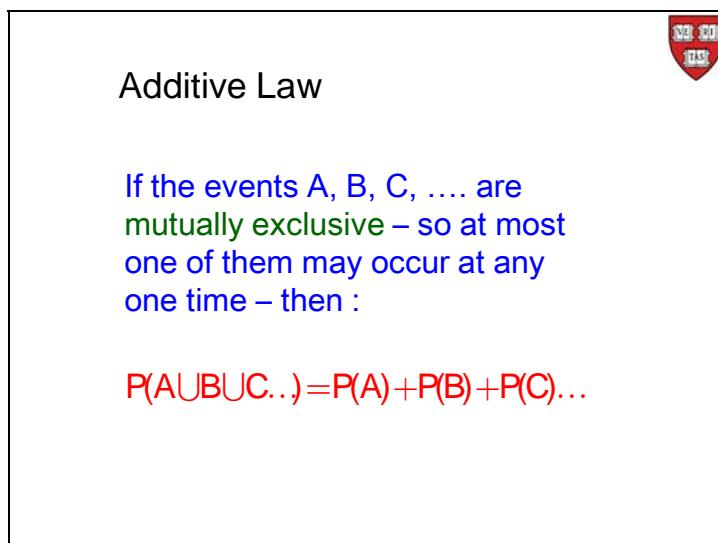
So here we have it in our Venn diagram.



Now, if A and B are mutually exclusive events, that means they cannot happen at the same time ( $A \cap B = \emptyset$ ). So the probability of A plus the probability of B is the probability of  $A \cup B$ , because the probability that A happened or B happened is just going to be the probability of A plus the probability of B.

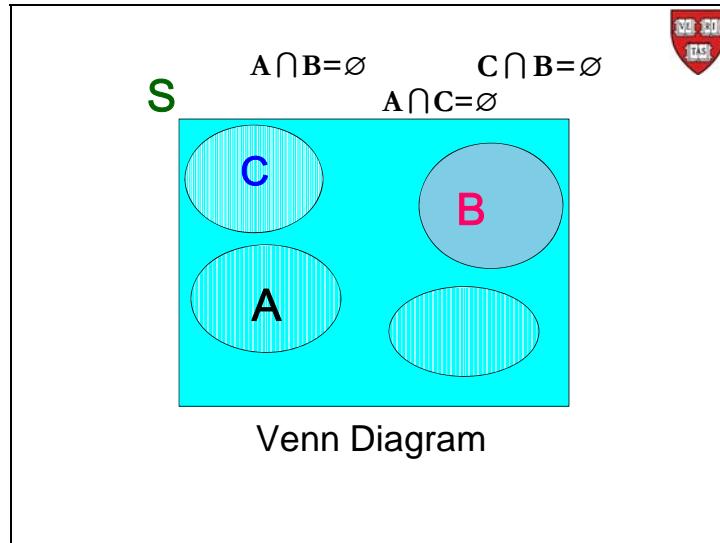


And we see this in our Venn diagram.

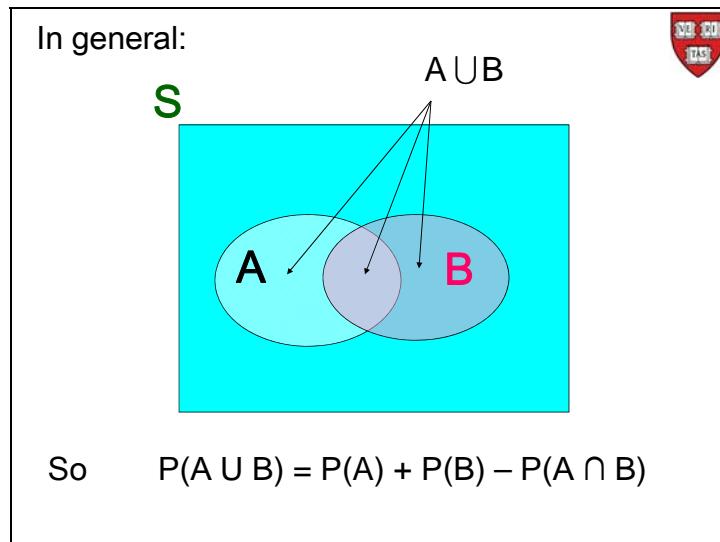


We can extend this to more than two mutually exclusive events and we give it a name: it is called the Additive Law. So the Additive Law of Probability tells us that if we have

mutually exclusive events then the probability of their union is the sum of their individual probabilities.



One way to think of probability when looking at Venn diagrams is to think of areas, and you will not go wrong. For example, with mutually exclusive events, as in the diagram above, the  $P(A \cup B \cup C \cup \dots) = P(A) + P(B) + P(C) + \dots$



In general, if events A and B are not mutually exclusive, then there is some overlap in the ellipses. If we now take the area of A and add it to the area of B, and think that that should be the probability of  $A \cup B$ , we'd be wrong. Why? Because we would be bringing the overlap of A and B, namely  $A \cap B$ , in twice into our calculation of the area of the union.

So the general formula is the probability of  $A \cup B$  is the probability of A plus the probability of B minus the probability of  $A \cap B$ , and that is the Additive Law.

## Conditional probability

### Conditional Probability



**Notation:**  $P(B | A)$

is the probability of B *given*, or knowing that the event A has happened.

**B**=“A person in US will live to be 70”

**A**=“A person is alive at age 65”

Then

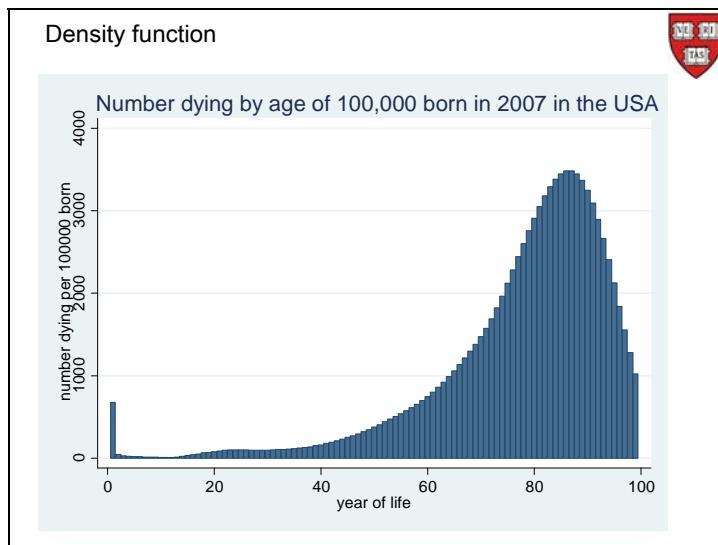
**$B | A$** =“A 65 year old person will be alive at 70”

Now that we have probability under our belts, we are almost there. We need to tackle one more concept, called conditional probability. As time evolves, we gather information, information that, if relevant, will possibly change our probabilities. Acknowledging that our probability depends on the information at hand, and how that probability changes, leads us to the concept of conditional probability.

To proceed, we need to expand our notation. Suppose that we retain interest in the probability of the event B, but suppose that now, we also want to bring into consideration the fact that the event A has happened. So what is the probability of B, given that the event A has happened? That is how we approach conditional probability. How do we modify  $P(B)$ , knowing that A has happened? Intuitively, if A is simpatico to B, then  $P$  of B, given A, denoted  $P(B|A)$ , should be bigger than  $P(B)$ . If, on the other hand, A is antithetical, or antipatico, to B, then  $P(B|A)$  should be smaller than  $P(B)$ .

The question is, how is the probability of B, given A, changed. For example, return to the USA life table. Suppose that the event B is that a person in the USA will live to be 70, and the event A, is that a person is alive at age 65. Then the event B, given A, is

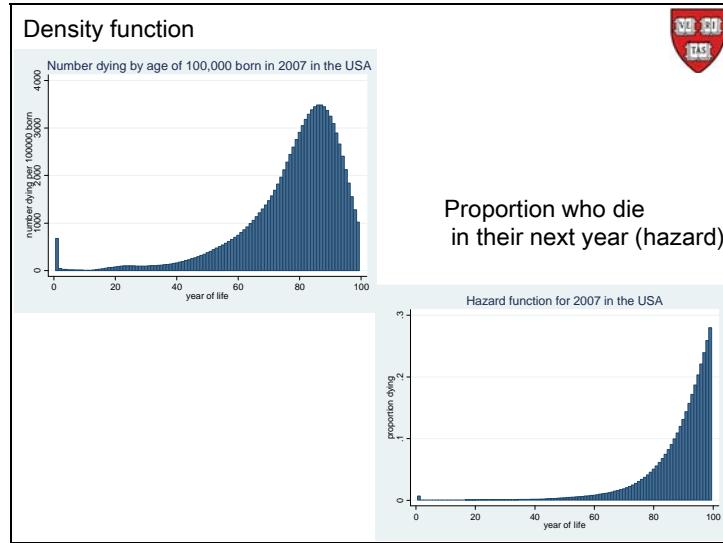
that a 65-year-old person will be alive at 70. That is different from the event B. So the question for the insurance salesman is, if you're selling insurance, would you charge the same thing for a baby to reach age 70 as you would a 65-year-old to reach age 70? Intuitively, the baby has to live another 70 years, including the stretch from 65 to 70, whereas the 65-year-old only has to live another five years—that stretch from 65 to 70. So these two events are not the same, thus their probabilities should be different. The question then is, how do we get from one probability to the other?



To make this point more empirically, here is what we saw when we looked at the life table for 2007 in the USA. This is how this fictional cohort of 100,000 people “born” in 2007, are going to die as the years progress. Remember, we said there was a big number around 80 to 95, or so, which is when most of us will die, and then it goes down. And George Burns said, "Very few are going to die beyond a 100," and that is right, because there are very few left. This is what we call the density function.



We also had, from the life table, the hazard function. The definition of the hazard function was, given that one is alive at a particular age, what is the probability that one dies within the next year? Note that now we do not have the curve thing coming back down beyond 95, because the probability of dying within the next year of life, when we are above 95, say, is very high.



So the density function and the hazard function address two different situations: The probability of dying at any particular age—the density function—versus the probability of dying in the next year, given that you are alive at the beginning of that year—the hazard function.

So there are the two different functions. On the left is the density function giving you the probability of dying at any time after birth, and on the right the hazard function. The

latter is a conditional probability—conditional on being alive at a particular age. Two related but different concepts.

Formula:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$



$B$  = “A person will be alive at 70”

$A$  = “A person will be alive at age 65”

$B | A$  = “A 65 year old person will be alive at 70”

$A \cap B$  = “A person will reach 65 & 70”

= “A person reaches 70”

Here is the formula to get you from one to the other. So the probability of B, given that A has happened, is the probability that both happen divided by the probability of A. I am, of course, assuming that we are not dividing by zero. So I'm assuming that the event A can happen. Returning to our example, if B was that a person will be alive at 70, A, that the person will be alive at 65, and B given A is that a 65-year-old person will be alive at 70, then the event  $A \cap B$  is that a person will reach 65, and a person will reach 70. In this example, if a person reaches 70, they will already have reached 65, so  $A \cap B = B$ , that is it reduces to the event that a person reaches 70.

Life table (segment) for the total population:  
United States, 2007



Age	Probability of dying between ages $x$ to $x + 1$	$l_x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
65–66 .....	0.013600	83,587	1,137
66–67 .....	0.014722	82,451	1,214
67–68 .....	0.015959	81,237	1,296
68–69 .....	0.017288	79,940	1,382
69–70 .....	0.018755	78,558	1,473
70–71 .....	0.020424	77,085	1,574

National Vital Statistics Reports, Vol. 59, No. 9,  
September 28, 2011

From the 2007 life table, we see that of the 100,000 who start off, 83,587 reach age 65, and 77,085 reach age 70.

Formula:  $P(B | A) = \frac{P(A \cap B)}{P(A)}$

e.g. 2007 lifetable: born: 100,000

$$65 : 83,587 \quad 70 : 77,085$$

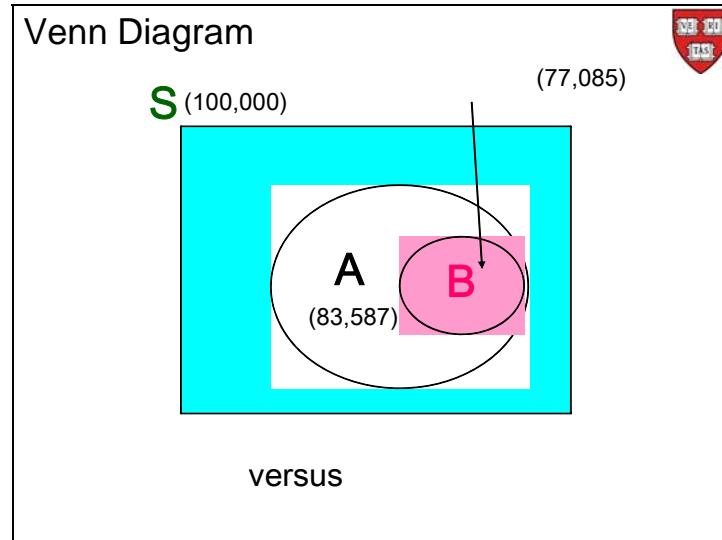
$$P(A \cap B) = \frac{77,085}{100,000} = 0.77$$

$$P(A) = \frac{83,587}{100,000}$$

$$P(B|A) = \frac{77,085/100,000}{83,587/100,000} = \frac{77,085}{83,587} = 0.92$$

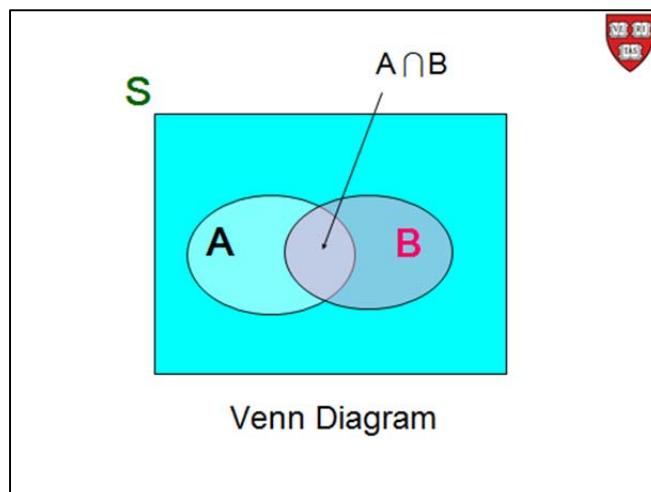
So if we bring all that information together,  $A \cap B$  is 77,085 divided by 100,000, so that is 0.77. So 77% of all babies born in this construct, this fictional cohort that we set up in 2007, 77% of those people will reach age 70. Also  $P(A)$  is 83,587 divided by 100,000. So dividing those two we get that the probability that a 65 year old reaches 70 is 92%.

So 92% of 65-year-olds will reach age 70. So it's much higher than at birth, as we intuited. That is how much bigger the probability becomes.



In the Venn diagram-- this is a special Venn diagram. Here A is the event you reach 65, and there were 83,587 such people. The event B was that you reach 70, and there were 77,085 such people. But B is also a sub-event of A (one must have reached 65 if one is to reach 70) so the 77,085 are also part of the 83,587—B is completely engulfed by A in the Venn diagram.

So if we calculated the conditional probability directly we would divide 77,085 by 83,587 and get 92%, which agrees with our earlier calculation that used the formula.



In general we can get an intuitive confirmation of the formula for conditional probability from the Venn diagram. If we associate  $P(B)$  with the area of  $B$  in the diagram, then that works if the area of  $S$  is one, otherwise we need to divide the area of  $B$  by the area of  $S$ . Now if we know that  $A$  has happened (so  $A$  is the new  $S$ ) then  $P(B|A)$  should be associated with the area of that part of  $B$  that contains  $A$ , namely  $B \cap A$ , so we want the area  $P(B \cap A)$ , but it needs to be normalized by the area of  $A$ —the new  $S$ —so divide it by  $P(A)$ , and we get  $P(B|A)$ .

## Multiplicative Law and Independence

From the formula for conditional probability,

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$



we get the multiplicative law

$$P(A \cap B) = P(A) P(B | A)$$

Note:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

From the formula for conditional probability, multiply both sides of the formula for the conditional probability by  $P(A)$  and you have the multiplicative law of probability. Note that even if  $P(A)=0$ , the law holds since then  $P(A \cap B)=0$ .



## Multiplicative law

$$P(A \cap B) = P(A) P(B|A)$$

Note:

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$\begin{aligned} \text{So } P(A \cap B) &= P(A) P(B | A) \\ &= P(B) P(A | B) \end{aligned}$$

A way of talking your way through the multiplicative law is to say, if both A and B are to happen, then if A happens first, then the probability I then want is the probability that B happens given that A has happened. This is just a minor help in remembering the law.

Note, by the way, that there's nothing special about A, nothing special about B. The whole development is symmetric in A and B, So we could either use probability of B given A, or the probability of A given B to obtain the multiplicative law. (Or, I could just as easily have A first or B first.)



## Independence

A and B are said to be independent if:

$$P(A \cap B) = P(A) P(B)$$

and since in general

$$P(A \cap B) = P(A) P(B | A)$$

So independence implies

$$P(B | A) = P(B)$$

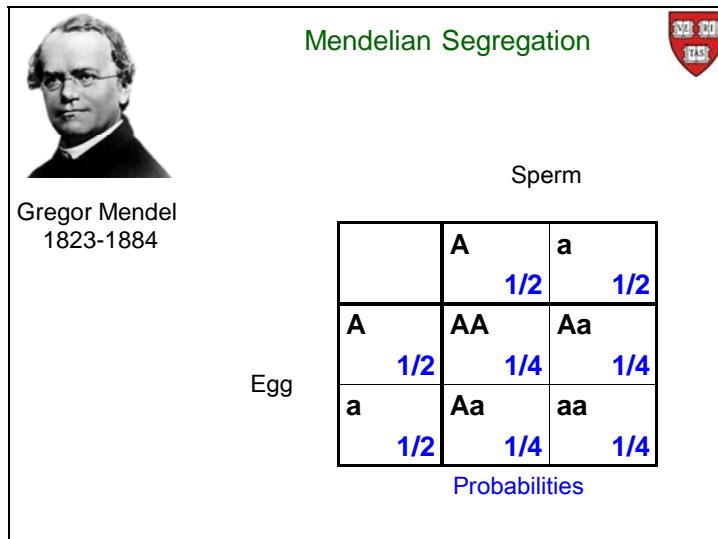
Similarly

$$P(A | B) = P(A)$$

The multiplicative law leads us to a very fundamental and important concept in probability, statistics, and epidemiology and that is the idea of independence. We say that two events, A and B, are independent if the probability of both A and B happening,  $P(A \cap B)$  is  $P(A) P(B)$ . That means that  $P(B|A) = P(B)$ . So knowing that A happens does not influence our probability of B happening. Similarly, by symmetry of A and B, we have that  $P(A)=P(A|B)$ . So knowing that B happens does not influence our probability that A happens. So the label, independent events, is well earned.

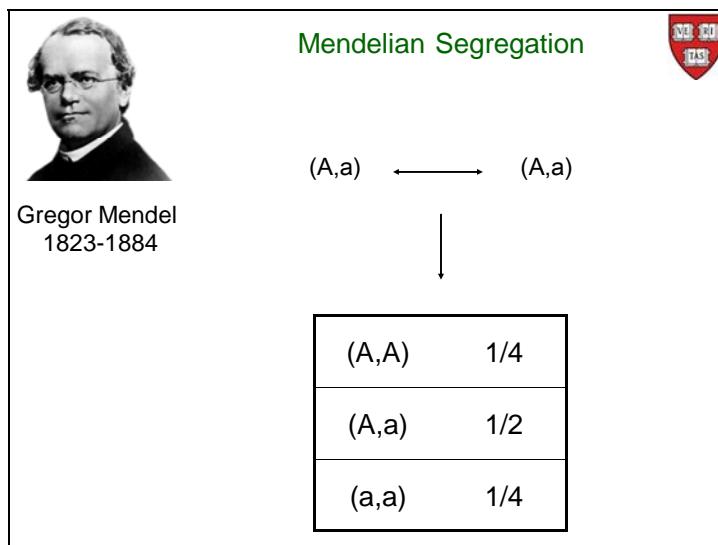
We use this condition repeatedly, for example, when we build up our knowledge by taking more and more and more patients. It will be very difficult if after observing the second patient we now have to go back to recalibrate what we learned from the first patient, and then after the third patient go back to the first two etc. So we assume that the patients are independent of each other, so what happened to the first patient is going to be independent of what happens to the second patient, independent of what happens to the third patient, and so on. Of course it must fit the situation or else the model is inappropriate, and we see an example of such a situation shortly. It is probably a little easier to accept with inanimate objects such as a fair roulette wheel. When you spin the ball around the wheel, surely what happened half an hour ago should not have any impact on where the ball lands up now. Surely the wheel does not have a memory! But judging by how people bet, this lack of memory is not believed by all.

The independence assumption needs to be justified, but we make it quite often, and it's extremely useful.

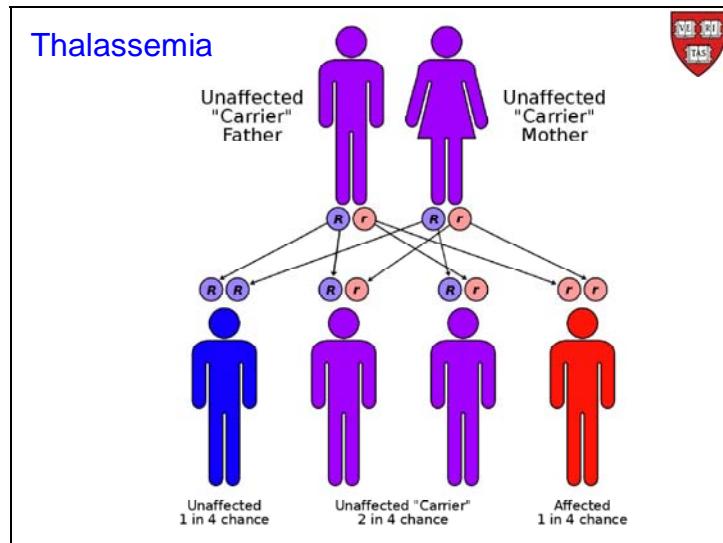


Here is one example of independence, and it has to do with Mendelian Segregation. If we have a sperm from a heterozygous male fertilizing an egg from a heterozygous female, then the first part of Mendelian Segregation tells us that it is equally likely that the sperm carries the dominant (so  $\frac{1}{2}$  the time) as the recessive allele ( $\frac{1}{2}$  the time), and it's equally likely to mate with the egg that carries the dominant ( $\frac{1}{2}$ ) or the recessive ( $\frac{1}{2}$ ) allele.

Step two says that you have independence. So what mates with what is independent of the allele in either the sperm or the egg. So the result is that you would get AA a quarter of the time (from the independence assumption  $\frac{1}{4} = \frac{1}{2} \times \frac{1}{2}$ ) so too with Aa, aA, and aa. Lastly, there is no difference between Aa and aA.



So finally, combining these four terms together, from a mating of heterozygous pairs, we should get dominant-dominant a quarter of the time, recessive-recessive a quarter of the time, and mixed half the time.



Indeed, this is what we see with thalassemia—or as it used to be called, Mediterranean anemia. This is what happens when you have an unaffected but carrier father, with an unaffected but carrier mother. It's a little deceptive because all the offspring are male in this diagram, but they don't have to be, of course. You can have any sex mixture of the offspring you want in there.

**Sally Clark and Roy Meadow**

Sally Clark was a British solicitor  
 Had a son in September of 1996.  
 He died in December of 1996.  
 Had a son in November of 1997.  
 He died in February of 1998.  
 She is accused and tried for murder.  
 Found guilty.

"Expert" witness, a pediatrician, Roy Meadow claims that the chance of two SIDS deaths in a family is "one in 73 million", and that carried the day.

Independent events are extremely important to us, and they play an integral role in a large number of models, but it is not always true that we have independent events. And if we falsely assume that there are independent events, then we can have an error that can have horrendous effects, such as what happened to Sally Clark.

Sally Clark was a British solicitor who had a son at the end of 1996. He died within a few months, in December of 1996.

She recovered from that event and had another son in November of 1997. This one died in February of 1998. She was accused of and tried for murder of both kids, and she was found guilty.

The defense argued that possibly these were cases of Sudden Infant Death Syndrome (SIDS) and so the prosecution called an expert witness—a pediatrician—one by the name of Roy Meadow, who went off and did some probability calculations, even though his expertise was not in probability but rather, in pediatrics. He guessed the chance that a particular family would have two SIDS deaths to be 1 in 73 million. He quite graphically explained that 1 in 73 million is like betting on 80 to 1 horses and having them win 4 races in a row<sup>3</sup>—in other words, 1 in 73 million is virtually impossible. This argument carried the day. She was found guilty, and she was imprisoned.

Meadow's Law:

one cot death is a tragedy, two cot deaths is suspicious and, until the contrary is proved, three cot deaths is murder.

The CESDI study looked at 472,823 live births.  
363 deaths were identified as SIDS.

$$P(\text{SIDS}) = 363/472,823 = 1/1300$$

Meadow testimony

$$(i) P(\text{SIDS}) = 1/8543$$

$$(ii) P(2 \text{ SIDS}) = (1/8543)^2 = 1/ 73 \text{ million.}$$

Fleming P, Bacon C, Blair P, Berry PJ (eds). *Sudden Unexpected Deaths in Infancy, The CESDI Studies 1993-1996*. London: The Stationery Office, 2000.

<sup>3</sup> <http://news.bbc.co.uk/2/hi/health/4432273.stm>

Meadow's law<sup>4</sup> was that one cot death is a tragedy. Two cot deaths-- that's what the British call SIDS—two cot deaths is suspicious, and until the contrary is proved, three cot deaths is murder. But the question still remained how did he get the 1 in 73 million?

It is of note that there is the CESDI report that studied 472,823 live births at roughly the same period in England, between 1993 and 1996, and they found 363 deaths that were identified as SIDS, or cot deaths<sup>5</sup>. So using their empirical evidence, the chance of a cot death should be 1 per 1,300 births.

Meadow argued that this family was a middle class family—both husband and wife were solicitors—so he used the divisor of 8,543, instead of 1,300. It is difficult to determine the basis for this creation.

But then he made the independence assumption. It then followed that, because of independence, the probability of two SIDS is  $1/8543$  squared, and that is how he got his 1 in 73 million.

In the CESDI report, they carried out a case-control study:



Among the 323 SIDS families studied, there were 5 previous SIDS,

$$P(\text{prev. SIDS in 323}) = 5/323 \approx 0.0155$$

Among the 1288 control families, there were 2 previous SIDS.

$$P(\text{prev. SIDS in 1288}) = 2/1288 \approx 0.00156$$

**It is dangerous to be right when the government is wrong.**  
Voltaire 1694 - 1778

Ray Hill, Multiple sudden infant deaths – coincidence or beyond coincidence?  
*Paediatric and Perinatal Epidemiology* 2004, 18, 320–326

If we return to the CESDI study, they have evidence of more than one SIDS death in a family. They used a case-control methodology to compare two groups, and what they found was that in the group with 323 SIDS deaths there were 5 previous SIDS, which

<sup>4</sup> Ray Hill, *Paediatric and Perinatal Epidemiology* 2004, 18, 320–326

<sup>5</sup> Fleming P, Bacon C, Blair P, Berry PJ (eds). *Sudden Unexpected Deaths in Infancy, The CESDI Studies 1993-1996*. London: The Stationery Office, 2000.

comes out to be 0.0155 and already raises some doubt about the 1 in 73 million figure. Then they compared this number to a control group of 1,288 families and found that the probability amongst that group of a previous SIDS was 0.00156.

So if there is one SIDS death already in the family, it is 10 times as likely that there will be another one. So this casts a doubt on Meadow's assumption of independence.

Voltaire had a saying for this one, too, "It is dangerous to be right when the government is wrong," and poor Sally Clark tragically suffered the consequences. She died shortly after they let her out of prison, which they did once the proper information was made available to the appeals court.



Clarification aid:

IF A and B are **mutually exclusive** then  
(**Additive Law**)

$$P(A \cup B) = P(A) + P(B)$$

IF A and B are **independent** then  
(**Multiplicative Law**)

$$P(A \cap B) = P(A) \times P(B)$$

Here is just a small clarification aid. We have introduced two situations where we have looked at properties of two events: one of them is when the two events are mutually exclusive, and then the additive law says that the probability of the union is the sum of the probabilities. The other dealt with independent events, in which case the multiplicative law says that the probability of the intersection is the product of the probabilities. Clearly, if two events are mutually exclusive they cannot be independent.

Students sometimes get a little bit confused about these two states. You can use as an aid to help you remember the distinction, union is like the addition of probability measures. Intersection is like the product of probability measures. So union is additive. You add things up and make them bigger when you unite them. Whereas, when you take an intersection it is like multiplying things and making them smaller. Just a little note to remind you, to make life a little bit easier for you.

## Bayes' Theorem

Return to the multiplication rule and now look at  $P(A | B)$ :



$$\begin{aligned} P(A | B) &= \frac{P(B \cap A)}{P(B)} \\ &= \frac{P(A) P(B | A)}{P(B)}, \end{aligned}$$

assuming  $P(B) > 0$ .

This is known as Bayes' Theorem

$$\begin{aligned} P(B) &= P(B \cap A) + P(B \cap A^c) \\ &= P(A) P(B | A) + P(A^c) P(B | A^c) \end{aligned}$$

Returning to the multiplication rule, let us look at the conditional probability,  $P(A|B)$ . It says that this probability is  $P(B \cap A)$  divided by  $P(B)$ . Once again, I am assuming that  $P(B) > 0$ . Returning to the multiplicative rule we have that  $P(B \cap A)$  can also be written as  $P(A) P(B|A)$ .

So after some straightforward algebra, we have on the left hand side a probability conditional on  $B$ , and on the right hand side the reverse: here we have that  $A$  is the conditioning event. This we are going to find very useful when we get to diagnostic testing.

This formula is called Bayes' Theorem. It is a beautiful theorem that is used repeatedly.

Sometimes it's not expressed this way, but rather the  $P(B)$  in the denominator is expanded, as shown. The first line in the expansion shows the use of the additive law, and the second equality shows the use of the multiplicative law.

## Diagnostic Tests



Diagnostic tests

$A = D = \text{"have disease"}$

$A^c = D^c = \text{"do not have disease"}$

$B = T^+ = \text{"positive screening result"}$

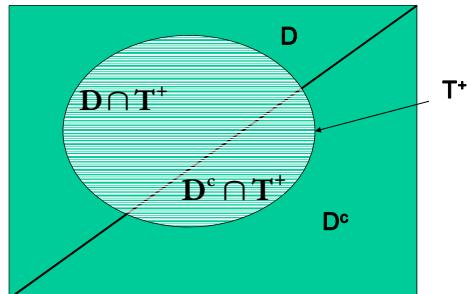
Find  $P(D | T^+)$

In diagnostic tests we associate the event A with the event D, the event having the disease in question. The event B is replaced with the letter T to denote testing, and  $T^+$  will denote testing positive for the disease.

Ideally if the test were perfect, anyone with the disease would test positive, and only people with the disease test positive. Unfortunately, there is little perfection in the world and that is certainly true of biological tests. We are thus interested how good a test is. Specifically, in the field we are interested in the probability that someone who tests positive for a condition or disease actually has that condition or disease.



Venn diagram of Bayes' Theorem



$$P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{P(D \cap T^+)}{P(D \cap T^+) + P(D^c \cap T^+)}$$

With a perfect test, of course, the probability we seek is one. But to see what happens with an imperfect test let us go back to our Venn diagram. For convenience, label the upper triangle, D, so it denotes the people with the disease. The lower triangle shows the people without the disease.

Now place an ellipse on the space to show the people who test positive. The diagonal line also intersects the ellipse, with the people in the ellipse above the line being those with the disease who test positive ( $D \cap T^+$ ), and those in the ellipse but below the diagonal the ones who do not have the disease but test positive nonetheless ( $D^c \cap T^+$ ).

If we had a perfect test, there would be nobody without the disease who tests positive, and there would be nobody with the disease who does not test positive.

The people who test positive who are not diseased ( $D^c \cap T^+$ ) are called false positives. They shouldn't be testing positive. They should be testing negative. Those who have the disease but are not testing positive are called the false negatives.

Bayes' Theorem is displayed at the bottom of the diagram.



Prior to testing				
	Has Disease D	Disease Free $D^c$		
Test Positive $T^+$	$P(T^+   D)$ sensitivity	$P(T^+   D^c)$		$P(T^+)$
Test Negative $T^-$	$P(T^-   D)$	$P(T^-   D^c)$ specificity		$P(T^-)$
	$P(D)$ prevalence	$P(D^c)$ $=1 - P(D)$		

Prior to using the test in the field we can quantify the properties displayed above. Create a so-called two-by-two table by cross-classifying individuals according to two binary variables: disease status and result of the test. Now fill the four cells of the table with the probability of being in that cell. Complete the table by filling in the probabilities of the margins. In the top left hand corner we have  $P(T^+|D)$ , and we call that the sensitivity of

the test—how sensitive it is at detecting the disease. In the bottom right hand corner we have the  $P(T^-|D^C)$ , and we call this the specificity—how specific is the test to the condition we are investigating, in other words will it just go positive for any condition, or just specifically for the one for which it was designed to test.

The column sums of the two cell probabilities are both one.

As often happens, the cells we name, are the ones where we want to achieve high values. So we would like tests with as high as possible sensitivities and specificities.

The people with the disease are sometimes called prevalent cases, and the probability of being in that state is called the prevalence.

The sensitivity and specificity of a test can be established before the test is used in the field, and guidance is available from the Food and Drug Administration<sup>6</sup>. Local conditions will determine the prevalence. All three of these quantities are of importance when testing an individual.



Post testing				
	Has Disease D	Disease Free $D^C$		
Test Positive $T^+$	$P(D   T^+)$ PPV	$P(D^C   T^+)$		$P(T^+)$
Test Negative $T^-$	$P(D   T^-)$	$P(D^C   T^-)$ NPV		$P(T^-) = 1 - P(T^+)$
	$P(D)$ prevalence	$P(D^C)$ $= 1 - P(D)$		

PPV=positive predictive value      NPV=negative predictive value

Post testing we can replace the probabilities in the two-by-two table. The margins remain the same. Prior to testing the column classification determines the conditioning event in the cell probabilities, whereas post-testing, the row classifier determines the conditioning. In the top left-hand corner we have the probability that someone who tests positive actually has the disease, and this is called the positive predictive value. In the

---

<sup>6</sup> <http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm>

bottom right-hand corner we have the probability that someone who tests negative does not have the disease. This is called the negative predictive value.

For this table the row sums are one.

From Bayes' theorem:

Positive predictive value:

$$P(D|T^+) = \frac{P(D) P(T^+ | D)}{P(D)P(T^+ | D) + P(D^c)P(T^+ | D^c)}$$

$$= \frac{\text{prevalence} \times \text{sensitivity}}{\text{prev} \times \text{sens} + (1-\text{prev}) \times (1-\text{specificity})}$$

As often happens, the diagonal entries are important. But the point to remember is that prior to testing, we have one set of measures; sensitivity, specificity. Post testing, we have another set of measures; the positive predictive value and the negative predictive value. What ties these quantities together is Bayes' theorem.

## Sensitivity and Specificity

Example: X-ray screening for tuberculosis



		Tuberculosis		
X-ray	Yes	No	Total	
Positive	22	51	73	
Negative	8	1739	1747	
Total	30	1790	1820	

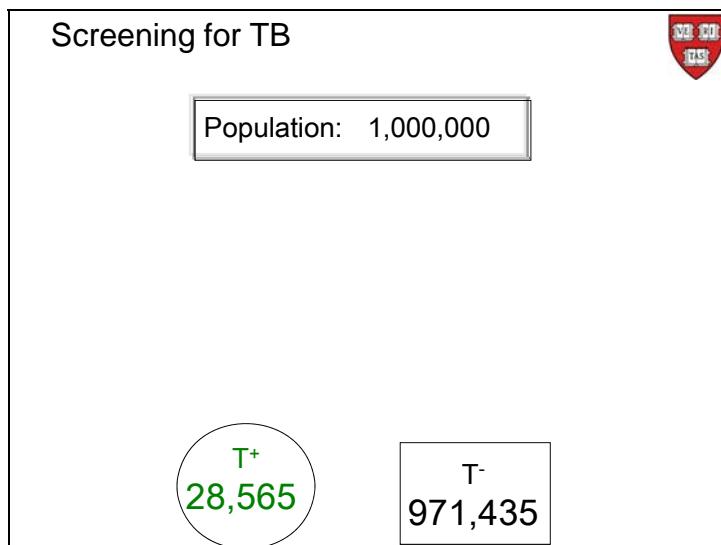
Sensitivity =  $\frac{22}{30} = .7333$   
 Specificity =  $\frac{1739}{1790} = .9715$

As an example of imperfect testing, suppose we want to use X-rays to screen for tuberculosis. In this study 30 people who had tuberculosis, had their X-rays read and 22 tested positive. Unfortunately 8 tested negative—the false negatives.

So the proportion of those with TB who tested correctly is 0.733 and that we can use as an estimate of the sensitivity.

For the 1790 without TB, 1739 correctly tested negative, but unfortunately 51 tested positive—the false positives. So we estimate the specificity to be .9715.

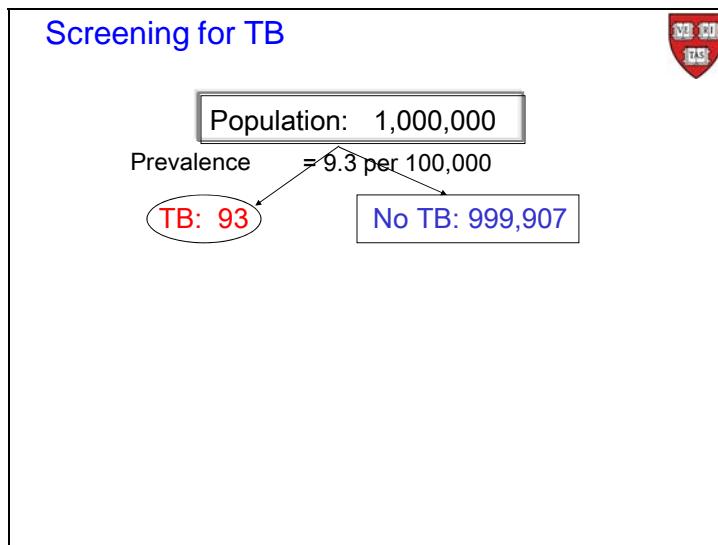
So this study establishes the properties of the testing procedure.



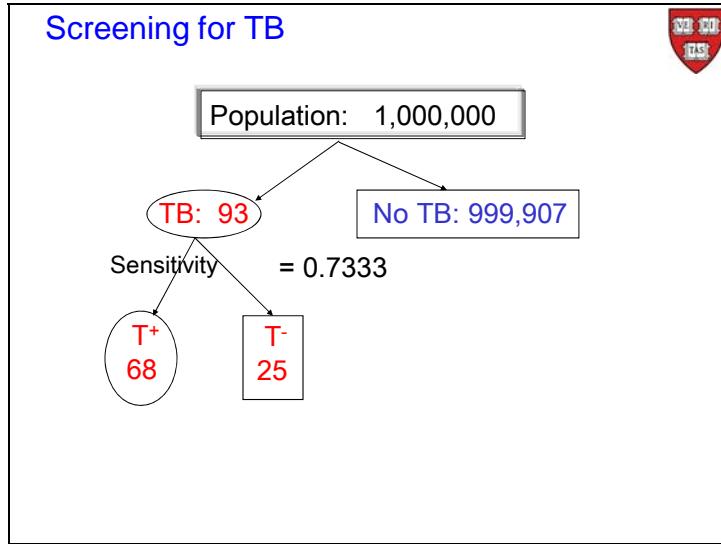
Having established the sensitivity and specificity, let us see what happens when the procedure is used in the field.

Suppose we have a population of a million people to be screened for TB and we use this test. When we implement a screening we find that 28,565 of this population actually tested positive, and the other 971,435 tested negative.

Before understanding the procedure, there is a third quantity we need, and that is the prevalence. How many of these one million people actually have TB? This thinking may seem circular, since determining how many people have TB is the point of the screening. But what we hope to do is look behind the scenes to see how these numbers are generated.

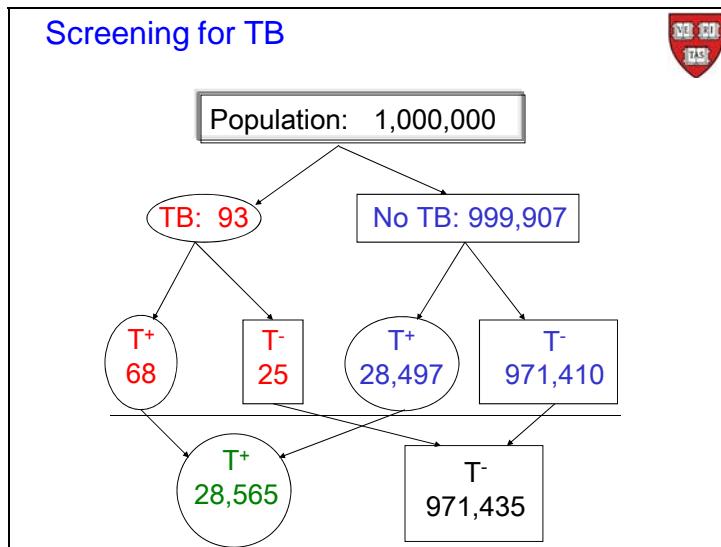


Let us suppose, for argument's sake, that the prevalence is 9.3 per 100,000. So that means 93 per million. We know that the test acts differently on those with TB and those without TB.



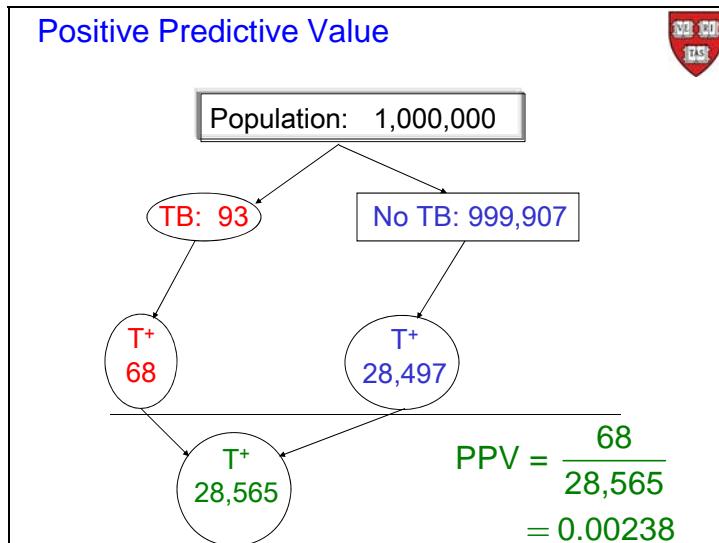
What would happen to these 93 when tested. The sensitivity is 0.7333 so 88 ( $=93 \times 0.7333$ ) would test positive and the other 25 would test negative.

---



For those without TB, we need to look to the specificity and see that of the 999,907 without TB, 971,410 ( $=999,907 \times 0.9715$ ) would test negative. That means the other 28,497 test positive. This is a large number of false positives.

We see the impact of this when we draw the line and see that beneath the line, the 28,565 who test positive are mostly (28,497) false positives. The total number of negatives is 971,435.



To determine the positive predictive value, we look at what proportion of those testing positive are true positives. In this case it is 68 of the 28,565. So the positive predictive value is 0.00238. That means that about 2 per thousand of those who test positive have TB.

Prior to the test we have:

$$P(D) = \frac{93}{1,000,000} = 0.000093$$

Post(er)ior to the test we have:

$$P(D | T^+) = \frac{68}{28.565} = 0.00238$$

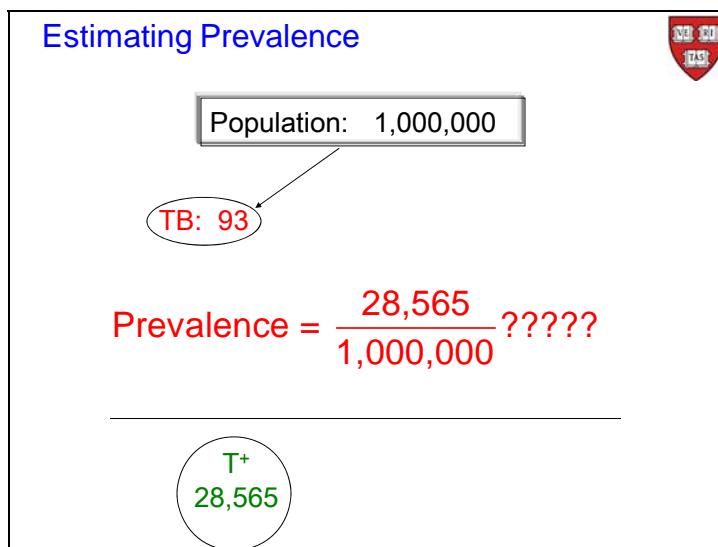
### Ratio:

$$\frac{0.00238}{0.000093} = 25.6$$

The problem here is that we are looking at a very rare disease. So if the specificity is not one we stand to observe a relatively large number of false positives, irrespective of how good the sensitivity is. A fairer way to appraise this number is to compare our ability to detect a case of TB before the screening (9.3 per 100,000, or about 1 in 10,000) with our ability after the test (approximately 2.4 per 1,000). So our detection capability has been increased by a factor of 25.6 (taking the ratio of pre to post probabilities of detection.)

If now we have a second test we can apply to those who screened positive on the X-ray test, then we can increase our detection probability even further since now we start with a prevalence of 2.38 per 1,000 as opposed to 9.3 per 100,000.

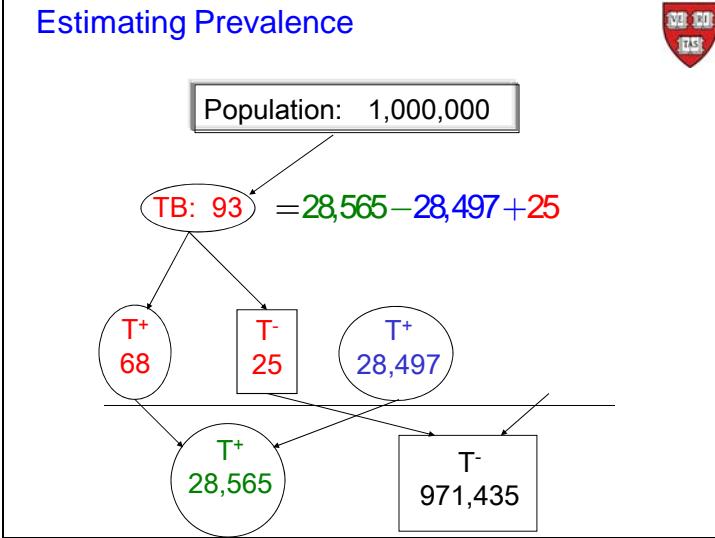
## Estimating prevalence



Often, as a result of screening we would like to estimate the prevalence of the disease. Indeed, what public health measures one takes might be dependent upon the prevalence. For example, if we are concerned with controlling schistosomiasis, then what the WHO prescribes as a public health measure very much depends on the prevalence of the disease.<sup>7</sup>

Given the results in our screening, how would we estimate the prevalence? If we divide all who tested positive by how many were tested, so  $28,565/1,000,000$  we would estimate the prevalence to be 28,565 per million and be off by a factor of 275 (=  $28,565/93$ ) because we do not have a perfect test.

<sup>7</sup> Page 42, Table A2.2 in [http://whqlibdoc.who.int/publications/2006/9241547103\\_eng.pdf](http://whqlibdoc.who.int/publications/2006/9241547103_eng.pdf)



To see how to obtain a good estimate of the prevalence, let us backtrack over the way the 28,565 were generated. From these 28,565 we would need to subtract the false positives (28,497) and add the false negatives (25) we lost. If we did that we would reduce the 28,565 down to 93.

$$\begin{aligned}
 93 &= 28,565 - 28,497 + 25 \\
 &= 28,565 - \\
 &\quad \{(1 - \text{prev}) \times (1 - \text{spec})\} 1,000,000 + 25
 \end{aligned}$$

We obtain the false positives from the  $1,000,000 \times (1 - \text{prevalence})$  who did not have the disease, and then multiply those by (1-specificity) to see how many tested positive.

Formula for estimating prevalence



$$93 = 28,565 - 28,497 + 25 \\ = 28,565 - \{(1 - \text{prev}) \times (1 - \text{spec}) - \\ \text{prev} \times (1 - \text{sens})\} 1,000,000$$

$$\text{prev} = \frac{\frac{28,565}{1,000,000} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$

$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$

The 25 false negatives, on the other hand we got from the 1,000,000xprevalence who tested negative, and then tested negative, so multiply by (1-sensitivity). Dividing through by 1,000,000 we get the formula above.

## Detection Limit

Conditions for formula to make sense



$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$

$$\text{"prop +ve"} - (1 - \text{spec}) \geq 0$$

$$\text{"prop +ve"} \geq (1 - \text{spec})$$


---

$$\text{sens} - (1 - \text{spec}) \geq 0$$

or

$$\text{sens} + \text{spec} \geq 1$$

We can think of our testing as utilizing an instrument, which like any other measurement instrument has its precision. Just like a cheap ruler to measure the length of an object

might only measure to the closest inch. On the other hand, if you had an electron microscope to measure the length of your object, you would have a much more precise measurement.

So what is the detection limit of our testing device? We can look at the prevalence formula and we know that prevalence has to be non-negative. That means that the numerator and denominator must both be of the same sign. Consider the case when they are both positive, leaving the other case for you to ponder.

Consider first the numerator. The proportion positive must thus be greater than one minus the specificity.

For the denominator to be positive we need that the sensitivity plus the specificity of the test must sum to more than one. Since we can get a sum of one by spinning a coin, this is not asking too much of our test. So let us assume that this is always the case. Indeed, with the X-ray we have a sensitivity of 0.733 and a specificity of 0.97, so their sum is indeed greater than one.

So the final constraint to ensure that we get a positive estimate of prevalence requires that the proportion positive be greater than one minus the specificity, and that is the lower bound to our detectability.

HIV newborn screening New York 11/87—3/90			
Region	Positive	Tested	Percent
NYS not NYC	601	346,522	0.17
NYC Suburban	329	120,422	0.27
Mid-Hudson	71	29,450	0.24
Upstate Urban	119	88,088	0.14
Upstate Rural	82	108,562	0.08
New York City	3650	294,062	1.24
Manhattan	799	50,364	1.59
Bronx	998	58,003	1.72
Brooklyn	1352	104,613	1.29
Queens	424	67,474	0.63
Staten Island	77	13,608	0.57

Here is an example of an HIV screening carried out amongst newborns in New York State, in the period November '87 through March '90. Now this is old data because they used to test all babies at birth, but then some politicians got involved and, as a result, this screening is no longer carried out and we do not have this monitoring information.

Over the period when we had the data, they reported separately for two regions: New York State (NYS), not New York City(NYC); and NYC. In NYS not NYC they tested 601 positive babies of the 346,522 tested. So the percent positive was 0.17%.

In NYC itself, there were 3,650 babies tested positive of the 294,062 births. So there, the rate was 1.24%.

Detection limit of instrument


$$\text{prevalence} = \frac{\text{"prop +ve"} - (1 - \text{spec})}{\text{sens} - (1 - \text{spec})}$$
$$\text{"prop +ve"} - (1 - \text{spec}) \geq 0$$
$$\text{"prop +ve"} \geq (1 - \text{spec})$$

So, for Upstate Urban NY where 119 tested positive out of 98,088 = 0.14% we need a specificity of better than

$$1 - 0.0014 = 0.9986 \text{ or } 99.86\%$$

Drilling down on these numbers, we can ask the question of whether some of these numbers truly represent infected babies. Our detection limits are such that the observed ratio must be greater than one minus the specificity. So for Upstate Urban NY where 119 tested positive of the 98,088 babies (=0.14%), that means that for those to be true positives would require that we have a specificity better than 99.86% (1-0.14%). This is an unrealistically large number for the specificity. So we don't expect that, indeed, these 119 were truly positive.

## ROC

Cotinine Level (ng/ml)	Smokers
0--13	78
14--49	133
50--99	142
100--149	206
150--199	197
200--249	220
250--299	151
300+	412
Total	1539

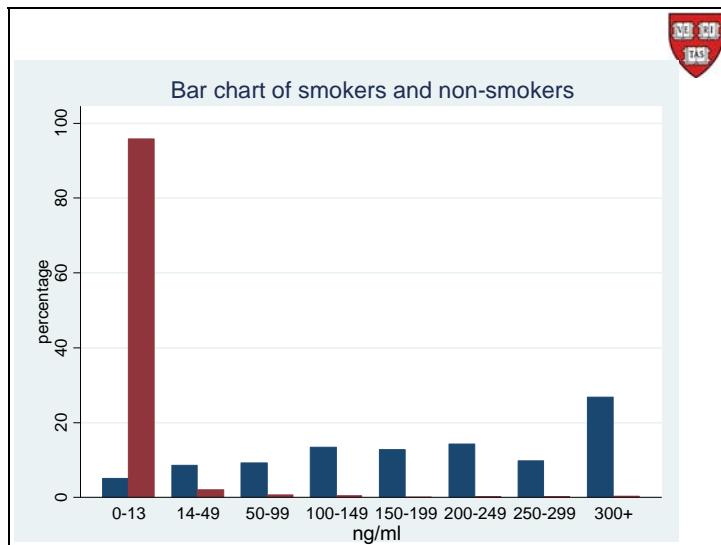
Continuing our study of imperfect tests, let us look at the rather common situation where a test is designed by measuring a certain biological quantity. If the test is applied to a sample that measures above a certain threshold then the test is declared positive, if below, then the result is negative. The determination of the threshold requires judgment. If set too high it would result in a number of false negatives, whereas if it is set too low it would result in a large number of false positives.

Here is a case in point. Cotinine is a metabolite for nicotine. Something like 40% of the nicotine metabolizes to cotinine. So measuring the cotinine level, in an individual's blood is a more reliable method than self-identification as a way to classify an individual as a smoker or non-smoker, so the theory goes.

Here are cotinine levels in this study of 1,539 identified as smokers. And 78 of them had very low cotinine levels, something less than 14 nanograms per milliliter (ng/ml), 133 of them had 14ng/ml to 49ng/ml. 142 of them have 50ng/ml to 99ng/ml, et cetera. So this is the distribution of cotinine level in smokers.

Cotinine Level (ng/ml)	Smokers	Non-smokers
0--13	78	3300
14--49	133	72
50--99	142	23
100--149	206	15
150--199	197	7
200--249	220	8
250--299	151	9
300+	412	11
Total	1539	3445

They also studied 3,445 non-smokers and found that 3,300 had less than 14 ng/ml in their system. They also found 72 with levels between 14ng/ml and 49 ng/ml, etc. as displayed above.



So if we draw the bar graph, the red bars are the non-smokers and the blue bars the smokers. To differentiate between the two on the basis of cotinine measure, it seems sensible to have a cutoff and if a person has cotinine below the cutoff, then classify that person as a non-smoker, and if above the cutoff, then classify that person as a smoker. Moving the cutoff right should mean more smokers will be falsely classified as non-smokers, and moving the cutoff to the left should mean that non-smokers will be falsely classified as smokers. A judgment has to be made about the judicious placement of the cutoff. Usually what enter into consideration for making this judgment are the consequences of potential errors.

The impact of the placement very much depends on the context. For example, in donated blood, then each unit of blood typically might impact some eight individuals because of the use of blood products. Thus when testing the donated blood, a false negative test for HIV or Hepatitis B, for example, might result in eight different people being infected as a result. The consequences of a false negative are thus dire.

On the other hand, a donation falsely labeled positive will result in the loss of one unit of blood. Of course, one must not ignore the false classification of the blood donor, but that can be rectified with further testing. The consequences of a unit of blood being falsely being labeled positive, do not seem to be as serious as a false negative.

In these examples, false negatives are highly consequential, whereas false positives might not be, but this is not always the direction of the imbalance. For example, with antenatal tests, one must be extremely careful about false positives<sup>8</sup>.

Cotinine Level (ng/ml)	Smokers	Non-smokers
0--13	78	3300
14--49	133	72
50--99	142	23
100--149	206	15
150--199	197	7
200--249	220	8
250--299	151	9
300+	412	11
Total	1539	3445

Returning to the cotinine numbers, consider the uncertainty in labeling by calculating the false positives and false negatives. Suppose we draw the cutoff line between 13 and 14, and call everybody below 14 a non-smoker and above 13 a smoker.

---

<sup>8</sup> <http://www.nlm.nih.gov/medlineplus/prenataltesting.html>

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	
50--99	142	23	
100--149	206	15	
150--199	197	7	
200--249	220	8	
250--299	151	9	
300+	412	11	
Total	1539	3445	

In that case, the sensitivity of this procedure would be, as measured by these data, that we would lose 78 of the 1,539 smokers. That leaves 1,461 which over 1,539 gives us the sensitivity associated with this cutoff.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	1328/1539
50--99	142	23	
100--149	206	15	
150--199	197	7	
200--249	220	8	
250--299	151	9	
300+	412	11	
Total	1539	3445	

What if we move the line down and use between 49 and 50 as our cutoff? Less than 50, would be called a non-smoker, and more than 49, would be called a smoker. Then we

will lose a further 133 smokers, and the sensitivity would now be 1,328 divided by 1,539.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	1461/1539
14--49	133	72	1328/1539
50--99	142	23	1186/1539
100--149	206	15	980/1539
150--199	197	7	783/1539
200--249	220	8	563/1539
250--299	151	9	412/1539
300+	412	11	
Total	1539	3445	

We can continue this logic line by line all the way down the table to obtain this table.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity
0--13	78	3300	0.95
14--49	133	72	0.86
50--99	142	23	0.77
100--149	206	15	0.64
150--199	197	7	0.51
200--249	220	8	0.37
250--299	151	9	0.27
300+	412	11	
Total	1539	3445	

Expressing these fractions as decimals, we see that the sensitivity goes from 0.95 to 0.86, to 0.77, to 0.64, and so on. So as we move the line down, our sensitivity decreases. But what you gain in the roundabout, you should be losing on the swings or vice versa, so let us look at what happens to the non-smokers—let us calculate the specificities.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	
50--99	142	23	0.77	
100--149	206	15	0.64	
150--199	197	7	0.51	
200--249	220	8	0.37	
250--299	151	9	0.27	
300+	412	11		
Total	1539	3445		

In anticipation of the next graph, let us not calculate the specificities but rather calculate one minus the specificity—namely, look at the proportions of the non-smokers who get incorrectly classified. So, starting at the top again, if the cutoff is between 13 and 14, then all but 3,300 of the 3,445 non-smokers would be correctly classified, or 145/3,445 would be the proportion incorrectly classified.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	73/3445
50--99	142	23	0.77	
100--149	206	15	0.64	
150--199	197	7	0.51	
200--249	220	8	0.37	
250--299	151	9	0.27	
300+	412	11		
Total	1539	3445		

Moving the cutoff down one line to between 49 and 50, then the proportion of non-smokers who would be misclassified would be 73/3,445.

Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	145/3445
14--49	133	72	0.86	73/3445
50--99	142	23	0.77	50/3445
100--149	206	15	0.64	35/3445
150--199	197	7	0.51	28/3445
200--249	220	8	0.37	20/3445
250--299	151	9	0.27	11/3445
300+	412	11		
Total	1539	3445		

Moving down the table one line at the time we finally get this table.

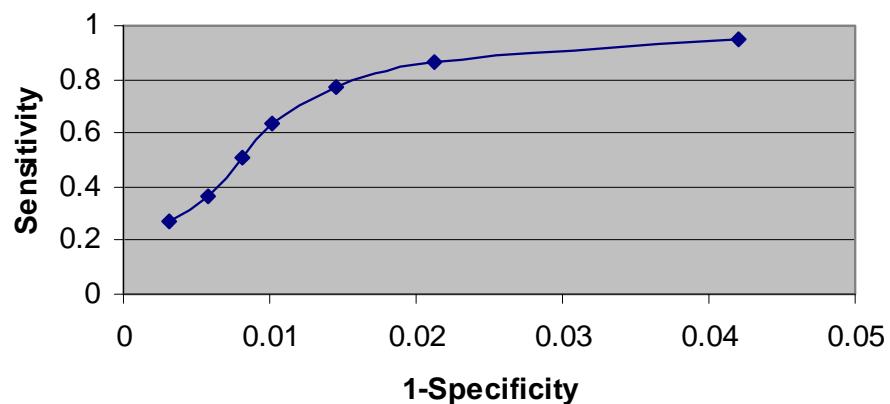
Cotinine Level (ng/ml)	Smokers	Non-smokers	Sensitivity	1—Specificity
0--13	78	3300	0.95	0.04
14--49	133	72	0.86	0.02
50--99	142	23	0.77	0.01
100--149	206	15	0.64	0.01
150--199	197	7	0.51	0.01
200--249	220	8	0.37	0.006
250--299	151	9	0.27	0.003
300+	412	11		
Total	1539	3445		

Finally, replacing all the fractions with their decimal equivalents we get this table. We see that as the cutoff is made higher (going down the table) the specificity goes up (one minus the specificity goes down) as, as we observed previously, the sensitivity goes down, as our intuition told us should happen.

We can plot the last two columns of the table against each other:



ROC Curve for Cotinine and Smoking

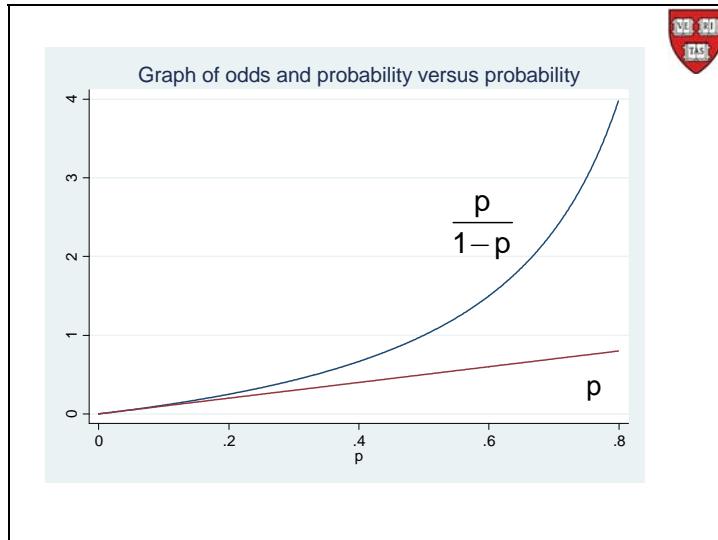


This plot is called the ROC, or Receiver Operator Characteristic, curve. This label comes from World War II signal detection.

This particular example is stopped at 0.05 on the horizontal axis, but it can continue all the way up to 1.

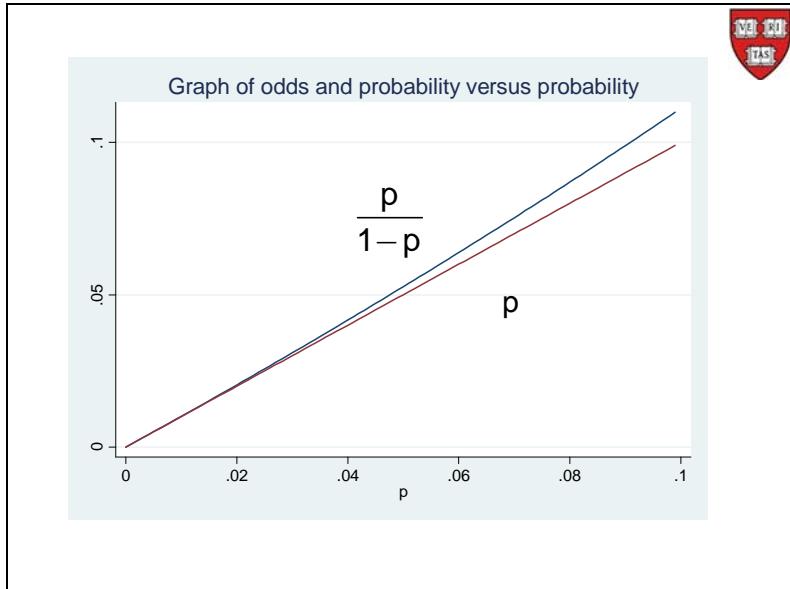
The ideal ROC would be zero at the origin, and one elsewhere. This one is not ideal, of course, few if any of any interest are, but it is quite good. To evaluate a test, or two compare two testing procedures, it is common practice to look at the area under the ROC curve (AUC), with the curve extending all the way to one on the horizontal. Presumably, when comparing two tests, the test with higher AUC is to be preferred.

## Probability and Odds



Let us briefly contrast probability with odds. If the probability of an event is  $p$ , then we said the odds of the event would be  $p$  over  $1 - p$ .

Mathematically, what happens to  $p$  over  $1 - p$ , as  $p$  gets close to 1 is that it goes to infinity, so above it is only drawn to  $p=0.8$ . In general, as  $p$  gets large the two, the probability and the odds, diverge. What is interesting is to look at them for small  $p$  when the two are close in value to each other.



Amplifying the left side of the last graph, we have this graph that only extends to when  $p$  equals 0.1. We see that they do start separating a little bit, but that for small  $p$  they are close to each other.

$p$	$p/(1-p)$ = odds	$\text{odds}-p$ = $\Delta$	$\Delta / p$ %	$\Delta / \text{odds}$ %
0.02	0.02	0.00	2.04	2.00
0.03	0.03	0.00	3.09	3.00
0.04	0.04	0.00	4.17	4.00
0.05	0.05	0.00	5.26	5.00
0.06	0.06	0.00	6.38	6.00
0.07	0.08	0.01	7.53	7.00
0.08	0.09	0.01	8.70	8.00
0.09	0.10	0.01	9.89	9.00
0.10	0.11	0.01	11.1	10.00

Here, when in tabular form, we see that when  $p$  is less than 0.07, the difference between the two is less than 0.005. For  $p$  less than or equal to 0.10, then even the relative difference is about 11% or less. The two are not the same, but when talking about rare events ( $p < 0.1$ ), there is not much difference between the two.

## Venn Diagram Tattoos



Marcello Pagano

# [JOTTER 4 – PROBABILITY MODELS]

Binomial Model, Central Limit Theorem, Poisson, and Normal Models

Life's most important questions are, for the most part, nothing but probability problems.

### Pierre-Simon Laplace

## Probability Models

Events associated with numbers;  
**Random Variables**

Dichotomous (Bernoulli):  $X = 0$  or  $1$

$$P(X=1) = p$$
$$P(X=0) = 1-p$$

e.g. Heads, Tails  
True, False  
Success, Failure  
Vaccinated, Not vaccinated



Now we are going to start applying what we learned about probability, and we start by applying probability to numbers and models. Mathematical models are an idealization, an idealization where there is right, wrong, exact, et cetera, and we now search how to use such models to approximate, or model, reality.

We start with random variables—random because we do not yet know exactly what value these variables take until we observe them. We know what values they potentially can take.

The simplest random variable is the dichotomous, or binary variable, which is also called the Bernoulli variable. It takes on one of two values; let us call them 0, and 1. We could call them 1, and 2, or any other two distinct values (nominal variable), but for the sake of definiteness, use 0 and 1.

To describe this variable probabilistically, we need to state the probability that it takes on the value 1. This also fixes the complement, namely the probability that it would take the value 0. The probability that it takes the value 1 is  $p$ . Therefore the probability that it takes the value zero is  $1-p$  and that then covers the whole spectrum (exhaustive).

And this is what we talk about with heads, tails; true, false; success, failure etc. This is where, for example, we spin a fair coin, and thus  $p=0.5$ . We also use this simple model in the very serious application to clinical trials, when we need to randomize a patient to one of two

treatments in such a way as to not show any favoritism for one treatment or the other. We return to clinical trials later.



e.g. Suppose that 80% of the villagers should be vaccinated. What is the probability that at random you choose a vaccinated villager?

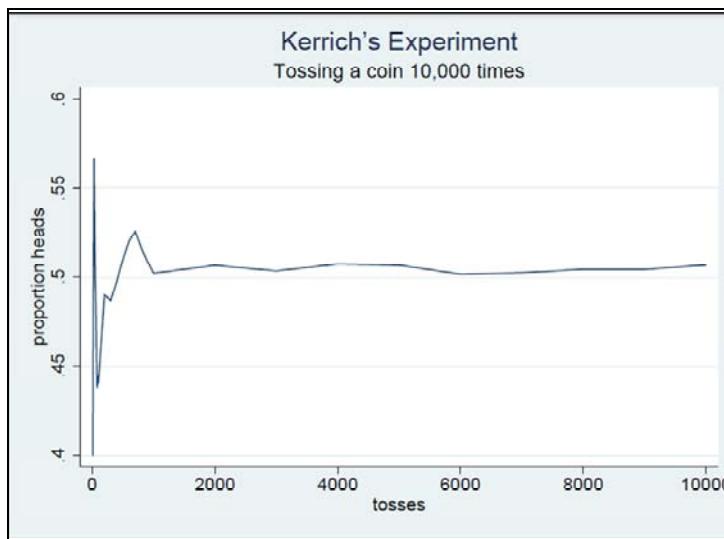
$1 \equiv$  success    (vaccinated person)  
 $0 \equiv$  failure    (unvaccinated person)

## 1 Trial

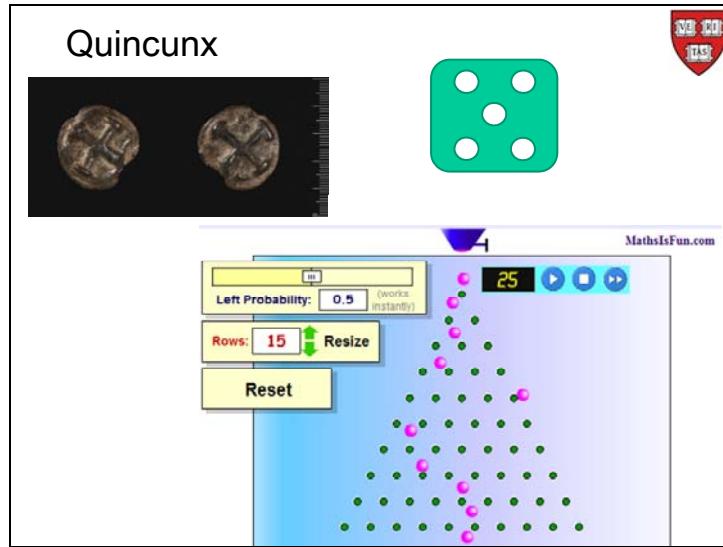
$$\begin{aligned}P(0) &= 1-p = 0.2 \\P(1) &= p = 0.8\end{aligned}$$

In our model we do not always have to have  $p = 0.5$ . For example, suppose we are concerned with vaccination coverage, then we might be aiming at a higher percentage, possibly such that 80% of the villagers should be vaccinated.

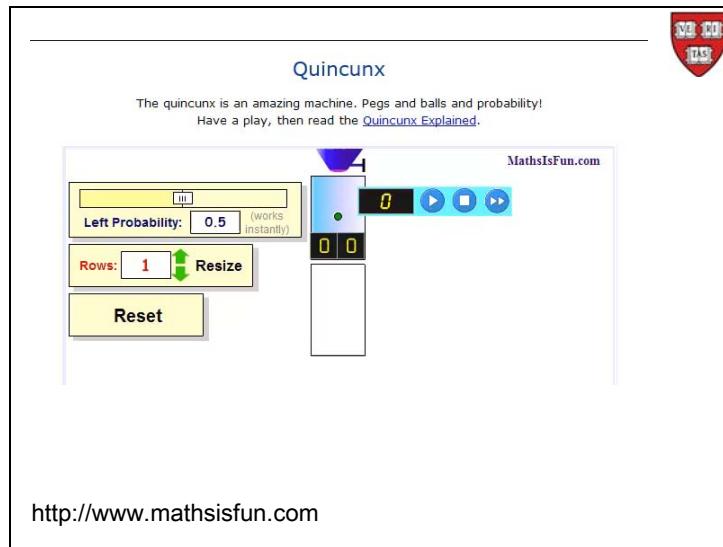
The way we model this is we say, what is the probability that at random you choose a vaccinated villager? You go into the village and choose a villager at random. What is the probability that that villager is vaccinated? We thus want to know what the probability is that we get a 1, if we label those vaccinated as 1's and those not vaccinated as 0's. So in this situation, if 80% of the villagers are vaccinated, then the probability of 1 would be 0.8.



Kerrich was faced with precisely such a situation. Remember, he spun a coin 10,000 times and he got 5,067 heads. Initially he saw a lot of variability in the ratio of heads to the number of tosses, but as time progressed it stabilized around approximately 0.5. We conceivably could repeat what Kerrich did, but remember, it took him years to perform this experiment.

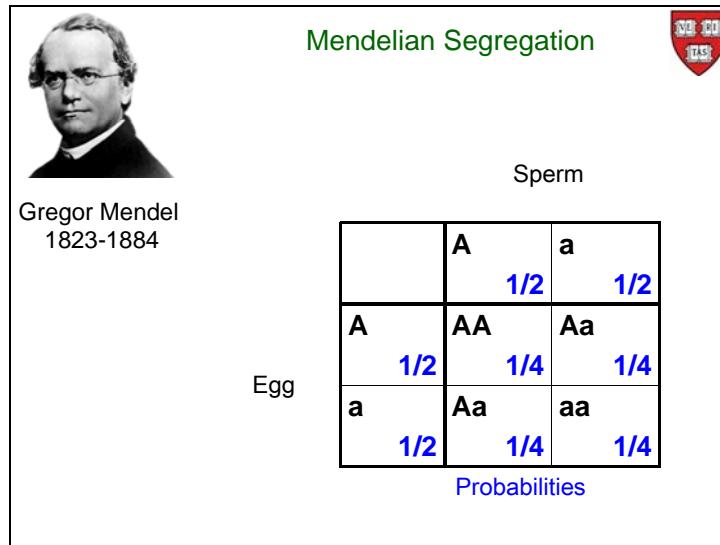


Alternatively, today we can do a similar experiment on a device that makes use of computer simulations and is called a quincunx. The name comes from an old Roman coin, shown above—quinc, of course, is from the Latin *quinque*, meaning five. The name is now attached to a design that puts five dots on a surface, just like the face that shows a five on your typical die. The same style triangular patterns occur all over the quincunx, so Francis Galton, who invented the quincunx, so named it. (We revisit Galton when we get to correlation.)

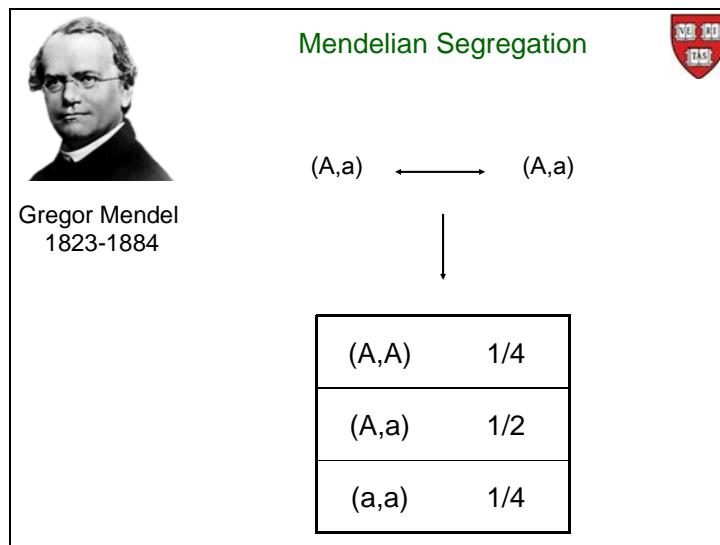


We can choose how many rows we want in our quincunx. Let us start with one row. Balls come out of the funnel at the top and bounce on the peg, either to the left or the right, and then collect in the wells at the bottom.

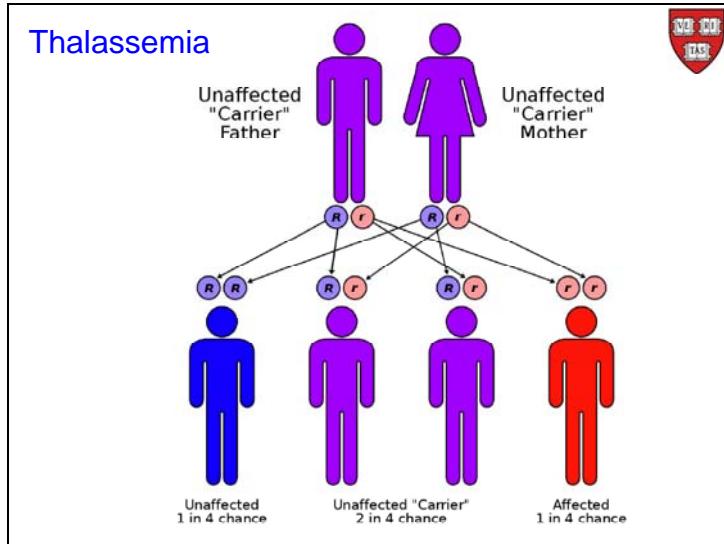
The way we've set it up is to choose "Left Probability" to be 0.5. That means that half the time the ball will bounce to the left, and half the time to the right, on average. What that means is that if you watch the wells at the bottom, then roughly half the balls gather in the left well, and half in the right well. The counter tells us exactly how many go in either well. This emulates tossing a fair coin.



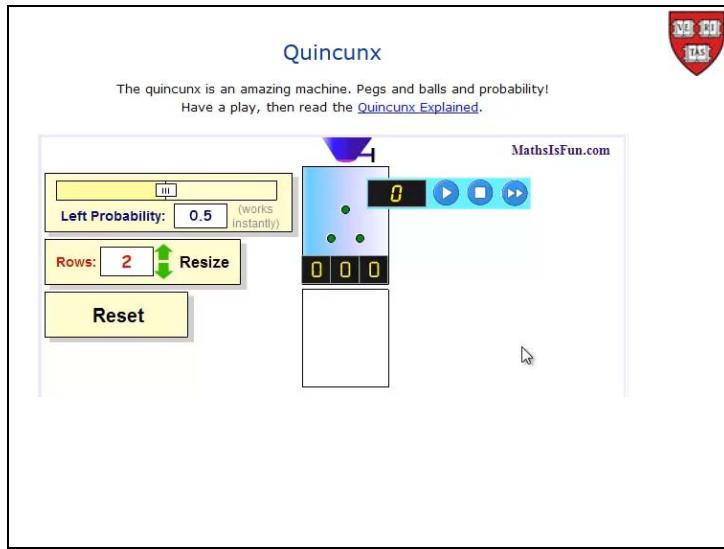
Returning to Mendelian segregation, we see that we would need to expand the Quincunx, if we wish to model that situation. Here we have two actions going on. First, what happens with the sperm and second, what happens with the egg. We can spin a fair coin to see whether the sperm carries the capital A or the little a, and similarly an independent fair coin to see whether the egg carries a capital A or a little a, before they join.



This is similar to spinning a coin twice (or equivalently spinning two coins) and associating AA with the outcome two heads, Aa with a head and a tail, and aa with two tails. We then end up with this probability distribution.

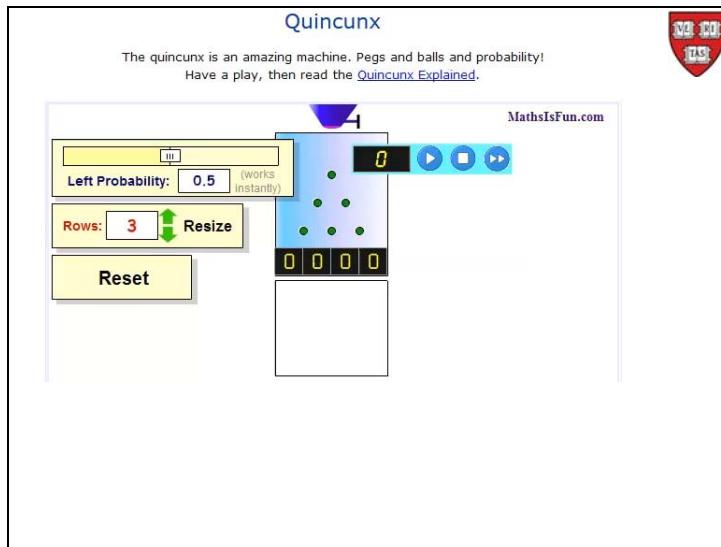


And we see that the ratios are 1 to 2 to 1, for the three categories (or the three wells in the Quincunx), which is exactly the ratios we saw with the Thalassemia application.



Returning to the Quincunx but this time set the number of rows to two. Now we see the ball first bounces on the top peg and then on a peg in the second row. To reach the leftmost well, the ball needs to bounce left and left. To reach the rightmost well, the ball has to bounce right and right. Whereas, in order to reach the center well the ball must bounce either left, and then right, or right, and then left. Since each path to the bottom is equally likely (because at each peg the ball goes left or right with equal probabilities, 0.5), that means that twice as many balls accumulate in the middle well (two ways of getting there) as in the outer two wells (only one way to get to either). So this looks like very much like the Mendelian example as well as the Thalassemia situation.

Thus theory tells us that we should get roughly a 1 to 2 to 1 ratio in the number of balls in the wells, and in the long run, if we let the Quincunx run for a while, this is what we expect to see.



We can continue to build up the complexity of the Quincunx by adding more and more rows. For example, for the three rowed Quincunx, then there is only one way to get to the outermost wells, as usual. This time there are three ways of getting to the two middle wells (left-left-right or left-right-left or right-left-left to reach well number two from the left, for example). So the ratio for the number of balls in the four wells is going to be 1:3:3:1. You can have a lot of fun yourself when you go to [mathsisfun.com](http://mathsisfun.com) (watch the s after math, this is an Australian site and they say, maths, not math) and play with the Quincunx and see what you get and how long “long-run” is.

This counting of paths to reach the bottom can quickly get tedious. But there is a device that has proven itself very useful over the years. We call it the Pascal triangle (after Blaise Pascal, 1623-1662), even though it was well known, even in Europe, before Pascal—we know that Tartaglia (1499-1557) knew about it.

Row 0					1			
Row 1				1		1		
Row 2			1		2		1	
Row 3		1		3		3		1
Row 4			1	4		6		4
Row 5	1		5		10		10	

We construct the triangle, which looks very much like a Quincunx, by having ones running down both edges either side. Then from top to bottom, all other entries are obtained by taking the sum of the two numbers in the row above on either side of the number you need to fill in. So  $2=1+1$ ,  $3=1+2$ ,  $3=2+1$ ,  $4=1+3$  etc.

Any row in the Pascal triangle will give you the ratios of balls in the wells of the Quincunx if you let it run for long enough with “Left Probability” set at 0.5. For example, with tossing a coin we look to Row 1; for the Thalassemia (Mendel) example we look to Row 2.



We call it Pascal's triangle, even though it was apparently known in Persia in the 11th century or so. Here it is in a Chinese text in the 14<sup>th</sup> century.

This is all of historical interest, but it does help us understand the Quincunx. In turn, the Quincunx helps us understand one of the most basic and beautiful models in statistics, the binomial model. Before we get there we need some notation.

## Factorial notation:

$$1 \times 2 = 2!$$

$$1 \times 2 \times 3 = 3!$$

$$1 \times 2 \times 3 \times 4 = 4!$$

⋮

$$1 \times 2 \times 3 \times \dots \times (n-1) \times n = n!$$

So,

$$3! = 6, \quad 4! = 24, \quad 5! = 120$$

By convention:  $0! = 1$

Consider the factorial notation: it is a positive integer followed by the exclamation mark.

To evaluate the expression for a particular integer, we multiply that integer by all positive integers smaller than it. So for example, 2 factorial, is 1 times 2, or 2. The product of 1, 2 and 3 is the same as 3 factorial; which, of course, equals 6. . By convention, we extend the factorial to include zero and define zero factorial to be one.



## Binomial Coefficients :

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad n = 1, 2, \dots$$
$$x = 0, 1, \dots, n$$

This leads us to the binomial coefficients, so called because they provide the coefficients in the binomial expansion of  $(a+b)^n$ . For  $n=1$ , we look to Row 1 in the Pascal triangle and get  $1a+1b$ . For  $n=2$ , we look to Row 2 and get  $(a+b)^2 = 1a^2 + 2ab + 1b^2$ . For  $n=3$ , we look to Row 3 and get  $(a+b)^3 = 1a^3 + 3a^2b + 3 a b^2 + 1 b^3$ , and so on.



Jakob Bernoulli  
1654-1705

## Binomial Distribution



A sequence of *independent Bernoulli trials (n)* with *constant probability of success at each trial (p)* and we are interested in the total number of successes (x).

e.g. In the 4<sup>th</sup> quarter of 1988 in Mass, of 21,835 births, 60 tested positive for HIV antibodies.

How many are infected?

Possible model: binomial with  $p \approx 0.25$  and  $n=60$ .

We now have all the background we need to introduce one of the most important fundamental models in all of statistics; the binomial model. All such models, including the model of independence that we have studied and Sally Clark got hurt with, have conditions which have to be satisfied in order for the model to be correctly applied. The condition of independence was not satisfied in the Sally Clark situation, and that led to a miscarriage of justice.

The conditions under which we can apply the binomial distribution, are: First, that we have a sequence of independent Bernoulli trials—named for the person Jakob Bernoulli, a brilliant

mathematician from a family of brilliant mathematicians, who did the early work on this model. Second, that there are a fixed number,  $n$ , of such trials. Third, that the probability of success at every single trial is a constant, call it  $p$ . Fourth, that we have a fixed number of trials.

For example, our “independent trials” might well be patients. We need to ask ourselves, does it make sense to think of these patients as independent? Will each patient have the same chance of success? If we can answer these in the affirmative, then we can use the binomial distribution. And the answers we get from the model will only be as good as how well the assumptions are satisfied.

If the assumptions are not satisfied, then we should not use the model. Note that we also need to have a fixed number of trials, and the binomial model is not appropriate when you continue the trials until you have a success, for example—think of having children till you have one of a particular sex.

If the conditions are satisfied, then we can apply this model. For example, in the fourth quarter of 1988, there were 21,835 births in the Commonwealth of Massachusetts which babies, at birth were tested for HIV. That test was an antibody test and 60 of these babies tested positive, which meant that the mothers of those 60 babies were infected with the HIV. The question then arises, in order to plan health services for these babies, how many of them can we expect to be infected with the HIV? At that time the mother to infant transmission rate was about 25%, so one possibility is to model the situation as a Binomial with  $n=60$  and  $p=0.25$ , if we assume the babies independent of each other.

This model is an idealization that hopefully yields some guidance. This is how we use models in biostatistics and epidemiology, in public health, in medicine et cetera.



Binomial Distribution

$X = \text{number of successes}$

$P(X) = \binom{n}{X} p^X (1-p)^{n-X} \quad X = 0, 1, 2, \dots, n$

$n = 1, 2, \dots$

Parameters:

$p = \text{probability of success}$

$n = \text{number of trials}$

The binomial model yields the binomial distribution that we would use to calculate the probability of obtaining  $X$  successes in  $n$  independent trials when the probability of success at each trial is  $p$ . So with the babies above, we would put  $n=60$  and  $p=0.25$ .

```

. gen p = binomialp(10,x,.5)
. list x p

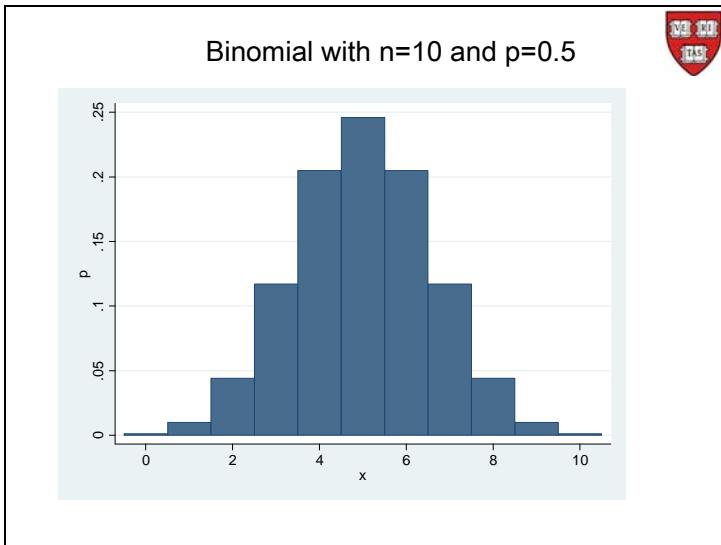
      x          p
 1. 0  .0009766
 2. 1  .0097656
 3. 2  .0439453
 4. 3  .1171875
 5. 4  .2050781
 6. 5  .2460938
 7. 6  .2050781
 8. 7  .1171875
 9. 8  .0439453
10. 9  .0097656
11. 10  .0009766

```

Of course you are not going to do these calculations yourself. You will ask Stata to do the hard work for you. In Stata it is a function called *binomialp* that has to be summoned. Here, for example are all 11 values it can take when  $n=10$  and  $p=0.5$ . So for  $x=0$  we get that  $p=0.0009766$ , which is also the probability of getting zero heads when tossing a fair coin 10 times.

The first thing to notice is the up and down symmetry around  $x=5$ . So the probability of zero heads is the same as the probability of 10 heads. So too the probability of 1 head is the same as the probability of 9 heads. And so on. This is due to  $p=0.5$  and the resultant symmetry and also in the arbitrariness in which side of the coin is called a head and which side is called a tail.

So that makes sense. Further, 5 is the most popular value (the mode), it is also actually the mean, which also fits in with intuition: if we spin a fair coin 10 times we expect 5 heads.



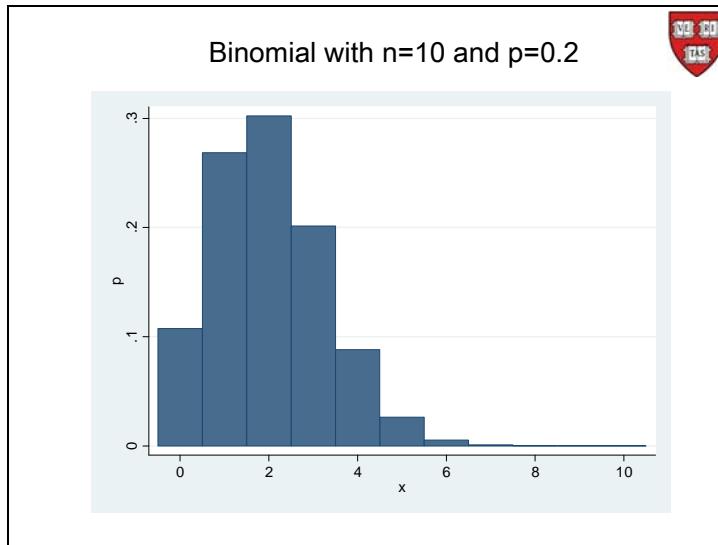
Let's plot it and this is what we get. So this is what your bin or the histogram underneath your bin in the Quincunx should look like if you run the Quincunx with 10 rows and  $p=0.5$ . Try it.

If you do, see how long it takes to get a shape that reasonably resembles this graph. How long is the "long run."

```
. gen p = binomialp(10,x,.2)
. list x p
```

	x	p
1.	0	.1073742
2.	1	.2684354
3.	2	.3019899
4.	3	.2013266
5.	4	.0880804
6.	5	.0264241
7.	6	.005505
8.	7	.0007864
9.	8	.0000737
10.	9	4.10e-06
11.	10	1.02e-07

If we change the  $p$  from 0.5 to 0.2, we would expect to see fewer successes and thus lose the symmetry. Here is the distribution. Now the mode and mean are at 2 (since  $10 \times 0.2 = 2$ ).

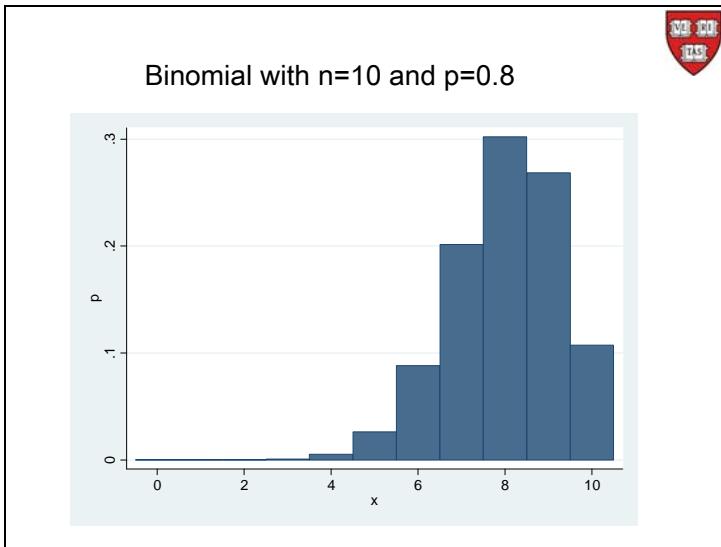


If we plot the distribution we get this graph. We see the symmetry gone and a tail appearing on the right.

```
. gen p = binomialp(10,x,.8)
. list x p
```

	x	p
1.	0	1.02e-07
2.	1	4.10e-06
3.	2	.0000737
4.	3	.0007864
5.	4	.005505
6.	5	.0264241
7.	6	.0880804
8.	7	.2013266
9.	8	.3019899
10.	9	.2684354
11.	10	.1073742

If we switch the probability of success from 0.2 to 0.8, then we should expect to see the mirror image of what we just saw—interchange your definitions of success and failure. So now our most popular value should be 8 ( $=10 \times 0.8$ ).



This is confirmed in the graph, as is the tail switching from the right to the left.

We can use this model directly if we ask about the 60 babies who tested positive for the antibodies to HIV where we said that the probability, of transmission, is 0.25 that any one of them actually has the virus. We have a fixed number of them, 60. Then if we assume the probability the same for each one of them and that the babies are independent of each other, we can use the binomial distribution to evaluate the probability for any number of them to be infected.

We can also reverse the logic. This graph gives us the probabilities for any particular number of successes. So let us ask, if we visit a village where the vaccination coverage is 80% (so  $p=0.8$ )

and I choose 10 villagers at random, what do I expect to see. In answer we can see where most of the mass is, mostly around 8, so that is what we expect to see. But what if I tell you that of the ten people, none were vaccinated; or one was; or two were; or three were. In each one of these cases you would say to me, "That is virtually impossible. Look, there is no probability mass down at the left end of the curve." So in such instances, what are we left with? Well we can argue that an extremely rare event has occurred, or we can seek an alternative explanation. One possibility is to question the validity of the assumption that  $p=0.8$ . In other words, is it possible that the vaccination crew missed this village in its last rounds? Or, was there a huge immigration of unvaccinated individuals into this village? This is how we use these models, namely to tell us what to expect and then to contrast that to what we observe, and then possibly question the assumptions so as to improve conditions.



For Binomial with  $n$  &  $p$

Then

$$\text{Mean} = np$$

$$\text{Stand. Dev.} = \sqrt{np(1-p)}$$

The binomial has two parameters,  $n$  and  $p$ . Remember, when we looked at the empirical rule, we said look that the mean and the standard deviation of a variable summarizes the distribution. So, if the conditions for the empirical rule are obeyed, then we could use it.

We saw that for  $p=0.5$  the binomial is perfectly symmetric. It also has finite tails, so they certainly go to zero very fast (they are thin). The mean for a binomial is  $np$ , and the variance is  $np(1-p)$ . So if we spin a fair coin 10 times, the mean is 5 and the standard deviation is  $\frac{1}{2} n^{1/2}$ , or 1.6.

This is exactly what you might have suspected; for example, if  $p=0.5$ , and you spin a coin 10 times, you expect half of those values, namely 5, to be heads.

The formula for the standard deviation is not intuitive at all. This takes some mathematics to work this out.



e.g.  $p=0.5$   $n=10$

$$\text{Mean} = np = 10 \times 0.5 = 5$$

$$\text{Stand. Dev.} = \sqrt{10 \times 0.5 \times 0.5} = 1.6$$

$$5 \pm 1.6 = (3.4, 6.6)$$

$$5 \pm 3.2 = (1.8, 8.2)$$

Applying the empirical rule to this, we get that two-thirds of the time we should get between 3.4 and 6.6 heads, and 95% of the time we should get between 1.8 and 8.2 successes.

These intervals are quite wide, and in fact are the widest since the worst variance ( $p(1-p)$ ) occurs when  $p=0.5$ .



e.g.  $p=0.25$   $n=60$

$$\text{Mean} = np = 60 \times 0.25 = 15$$

$$\text{Stand. Dev.} = \sqrt{60 \times 0.25 \times 0.75} = 3.4$$

$$15 \pm 3.4 = (11.6, 18.4)$$

$$15 \pm 6.8 = (8.2, 21.8)$$

If we return to our babies tested for HIV,  $p=0.25$  and  $n=60$ , and here are the two intervals.

We must be a little more cautious when applying the empirical rule here because  $p$  is no longer 0.5, so the symmetry no longer holds. We can, using Stata, check exactly what proportions fall into these intervals.

## Binomial Distribution



X = number of successes

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X} \quad X = 0, 1, 2, \dots, n$$
$$n = 1, 2, \dots$$

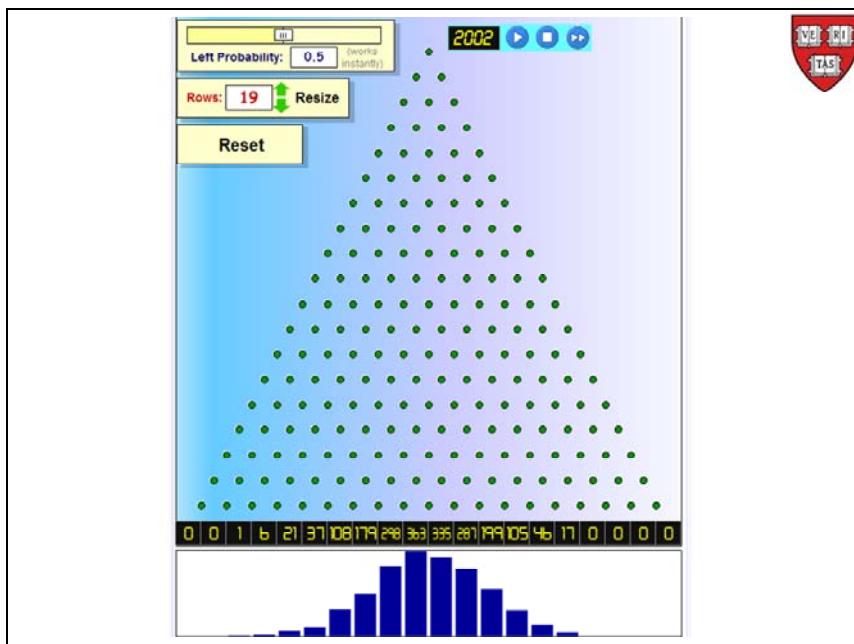
Parameters:

p = probability of success

n = number of trials

Something wonderful happens to the binomial when n becomes large. Today we have the computer to do all our calculations so we can be cavalier about hard calculations, relying on intelligent numerical analysts to do the right thing, but in the old days, all of these calculations had to be done by hand, and as a result, there were lots of errors. Also they were very difficult to do in this case when n was large, especially if p was small. The result was multiplying very big numbers by very small numbers and adding the resultant to small numbers and so on. The calculations took a tremendous amount of time and their accuracy was not to be trusted.

We are fortunate because of the computer we can see what happens when n is large. Make n as large as you can on the Quincunx; i.e. n=19. Now run it with p=0.5 for a long time. Go make yourself a sandwich in the kitchen and come back. What do you see? Here is what I saw,



It looks lovely. We see all these pink balls in a cascade, bouncing from peg to peg and building up a wonderful design at the bottom. Here is the result after 2002 balls came down—I tried to stop it at exactly 2000 but I was not successful. It's a lovely shape. The magic is that no matter how often you restart it and repeat this experiment, you get the same shape at the bottom! This is the magic of mathematics, the magic of statistics.

And this is what de Moivre discovered—he did not have the Quincunx to point the way, so you can imagine how brilliant he must have been! As a sidebar, De Moivre, apparently shared a characteristic with Cardano before him; namely, both are said to have predicted their own deaths. De Moivre based his prediction on mathematics: he discovered that he was sleeping an extra 15 minutes each night, so he posited that he would die the day he slept for 24 hours. He did.



**Binomial Distribution**

$X = \text{number of successes}$

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X} \quad X = 0, 1, 2, \dots, n$$

$$n = 1, 2, \dots$$
  

$$\frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}} \text{ approx. Normal}$$

The way to describe what De Moivre discovered is to say, if we look at our binomial variable,  $X$ , and standardize it by subtracting its mean,  $np$ , and dividing by its standard deviation,  $\sqrt{np(1-p)}$ , then this standardized variable, as  $n$  gets very large, can be treated like a standard normal variable.

## Continuous Random Variables

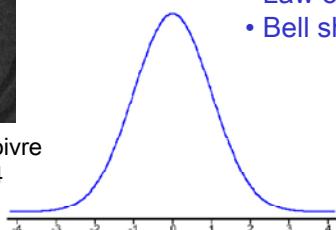


$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right\} \quad -\infty < x < \infty$$



Abraham de Moivre  
1667 – 1754

- Normal Distribution
- Gaussian (De Moivre)
- Law of errors
- Bell shaped



That means that the distribution, or density, will always have this shape, as defined by the mathematical expression above, as  $n$  becomes large. This is the distribution of a continuous random variable. A fantastic result.

Why this distribution is not called the Demoivrian and why it is called the Gaussian—after a person who came some one hundred years later—distribution is a puzzle the historians will have to unravel. This is the first example of what we revisit shortly, a central limit theorem, that is central to the practice of statistics. This is also sometimes called the normal distribution, because at one time we thought that every variable would have this distribution. It is also called the bell shaped curve, presumably because bells look like this.

The normal distribution, has two parameters,  $\mu$  and  $\sigma^2$ . So we are back to summarizing a distribution, just as we did in our empirical rule, by looking at the mean and the standard deviation. In the normal case, we get the entire distribution of the entire population by specifying just those two parameters. We return to it shortly.

## Poisson Distribution



**Poisson Distribution**



- 1. The probability an event occurs in the interval is proportional to the length of the interval.
- 2. An infinite number of occurrences are possible.
- 3. Events occur independently at a rate  $\lambda$ .

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

Another wonderful result occurs when we let  $n$  get large in the Binomial model, but this time we also let  $p$  simultaneously get very small. If this were to happen, we arrive at what is sometimes called the model for rare events, or more commonly, the Poisson model.

Mathematically, Poisson stipulated not only that  $n$  should get large, and that  $p$  get extremely small, but in such a way as to have their product,  $np$ , approach a constant, we call lambda,  $\lambda$ . We know that the mean of the Binomial is  $np$ , so  $\lambda$  is the mean of our new distribution.

When might such a model be appropriate? Think of a traffic intersection in your town. The probability of there being an accident when you observe the intersection for a minute, say, is probably negligible. But if you look at it long enough, say for a year, you observe two accidents, something of that order.

So there we have a situation where if you observe the process for a long time, so your  $n$  is really big, your  $p$ , the probability that you have an accident in any one of those minutes is extremely small—for example there are approximately  $60 \times 24 \times 365.25 = 525,960$  minutes in a year. So if we say that an accident is equally likely to happen at any one of those minutes—probably a gross oversimplification—and two happen a year, then  $p = 4 \times 10^{-6}$ ; a very small number. But the product, huge  $n$  times tiny  $p$ , to represent what happens in a year, can be a reasonable number—in our example it is 2.

So Poisson set up the conditions under which his model is correct—remember, we continue to work under the premise that models are idealizations that we wish to use to approximate reality—and they are:

1. The probability that an event occurs in an interval is proportional to the length of the interval (So if I watch it for twice as long, then my probability is going to be twice as big that I will observe an accident).

2. To make the mathematics possible, an infinite number of occurrences are possible (We know that can't happen, of course. But we'll see in a minute that that's not that big an assumption.)
3. The third assumption is a big one, and that is that events occur independently (So if we're watching the intersection and there was an accident yesterday or the day before, that's not going to make any difference to whether there's an accident today, let's say. So there is a certain amount of independence.)

These are Poisson's three conditions. If they hold, then the probability of having  $x$  events, or accidents, in a given year is given by the formula above.

Looking at the formula, we see a single parameter,  $\lambda$ . The conditions for the model are strict, but if you can apply the model, then you only need specify a single parameter.



For the Poisson one parameter:  $\lambda$

Mean =  $\lambda = np$

Variance =  $\lambda = np(1-p)$

$\approx np$

Let us look at this single parameter a little more closely. We defined it as equal to  $np$ . From the binomial we know that that is its mean, so lambda is also the mean of the Poisson. Now consider the variance of the binomial. As  $n$  gets large,  $p$  gets small, and  $np$  equals lambda, that means that the variance of the Poisson is also lambda. Thus lambda is both the mean and the variance.

This is why we often use the Poisson model when we study a phenomenon where the variance increases with the mean.



e.g. Probability of an accident in a year is 0.00024. So in a town of 10,000, the rate

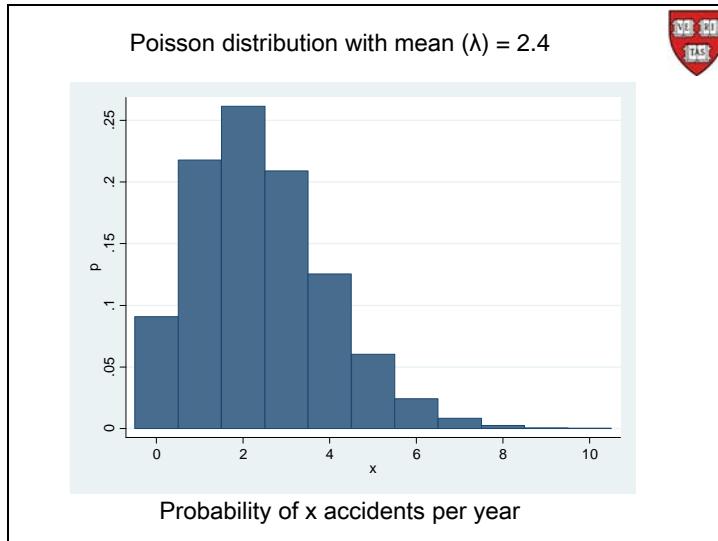
$$\lambda = np \\ = 10,000 \times 0.00024 = 2.4$$

$$P(X=0) = \frac{e^{-2.4}(2.4)^0}{0!} = 0.0907$$

$$P(X=1) = \frac{e^{-2.4}(2.4)^1}{1!} = 0.2177$$

Let us look at an example. Return to our accident example, except let us look at it from the perspective of each person. Suppose that the probability that any one person has an accident in a year is 0.00024; an extremely tiny probability. Suppose further that we are talking about a village that has 10,000 inhabitants. So  $np = \lambda = 2.4$ , and we expect a total of 2.4 accidents a year in this village.

So if we fit a Poisson to this situation, we put  $\lambda = 2.4$  into our Poisson formula to get that the probability of no accidents in a year is 0.0907, or approximately 0.1. So we expect that one in ten years we will not see any accidents in that village. The probability of one accident in a year is 0.2177. So we expect every five years to have a single accident in the village. And so on.



We can plot this distribution and here is what it looks like. So the most popular value (mode) is 2 and then things tail off after two. And you can see it's essentially 0 by the time we get to 10. So the fact that theoretically we can go to infinity is not a big deal, because we have most of the probability mass before we reach 10.

	<pre>. gen p = poissonp(2.4,x) . list x p</pre> <table border="1"> <thead> <tr> <th>x</th><th>p</th></tr> </thead> <tbody> <tr><td>1.</td><td>.090718</td></tr> <tr><td>2.</td><td>.2177231</td></tr> <tr><td>3.</td><td>.2612677</td></tr> <tr><td>4.</td><td>.2090142</td></tr> <tr><td>5.</td><td>.1254085</td></tr> <tr><td>6.</td><td>.0601961</td></tr> <tr><td>7.</td><td>.0240784</td></tr> <tr><td>8.</td><td>.0082555</td></tr> <tr><td>9.</td><td>.0024766</td></tr> <tr><td>10.</td><td>.0006604</td></tr> <tr><td>11.</td><td>.0001585</td></tr> </tbody> </table>	x	p	1.	.090718	2.	.2177231	3.	.2612677	4.	.2090142	5.	.1254085	6.	.0601961	7.	.0240784	8.	.0082555	9.	.0024766	10.	.0006604	11.	.0001585	
x	p																									
1.	.090718																									
2.	.2177231																									
3.	.2612677																									
4.	.2090142																									
5.	.1254085																									
6.	.0601961																									
7.	.0240784																									
8.	.0082555																									
9.	.0024766																									
10.	.0006604																									
11.	.0001585																									

Here is the Stata command we used to calculate the Poisson probabilities, and here are the first eleven values of the curve.

So the Poisson model is related to the Binomial by letting n go to infinity, and at the same time, let p go to zero in such a way that  $np = \lambda$ . That is a way of reaching the Poisson, but it also stands on its own as a useful model.

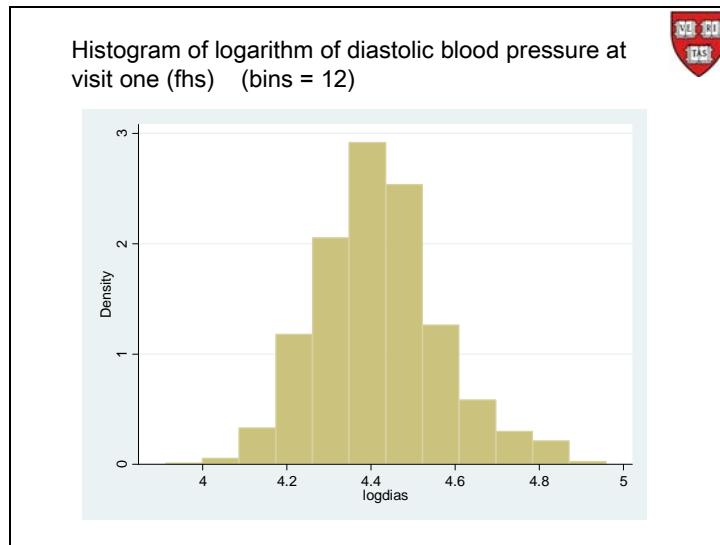
Returning to the De Moivre result, we also saw that simply letting n go to infinity in the binomial, we get the normal distribution. Let us return to that and study it more closely.

### The Normal Distribution

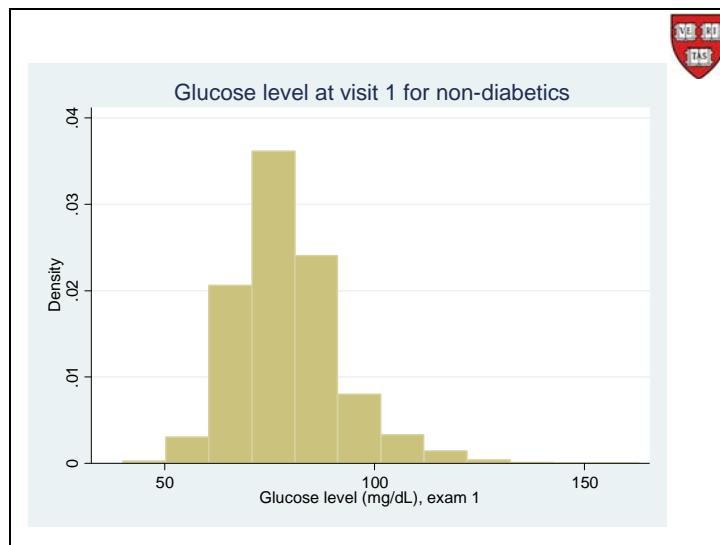


Here is the 10 Deutsche mark, the currency used in Germany before they started using the euro. We see a picture of Carl Friedrich Gauss (1777— 1855)—he of Gaussian distribution fame. In the middle of the mark, in the background, one can see the normal curve. We return to Gauss when we discuss regression theory and least squares.

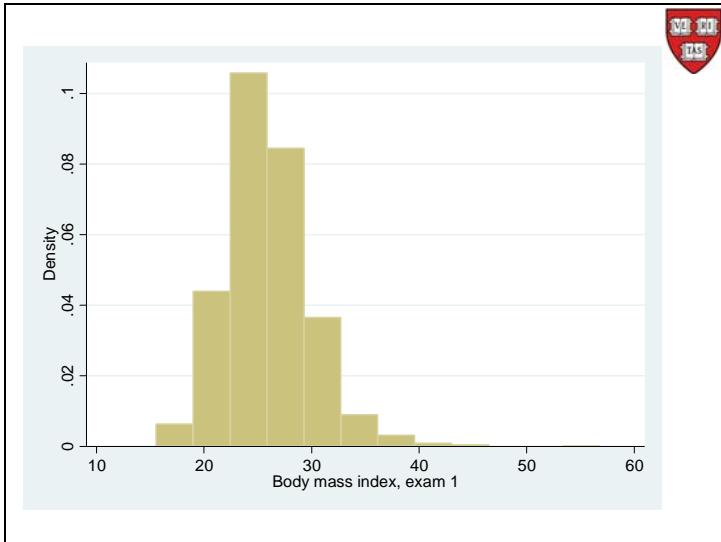
The normal distribution stands on its own, and not just as an approximation to the binomial.



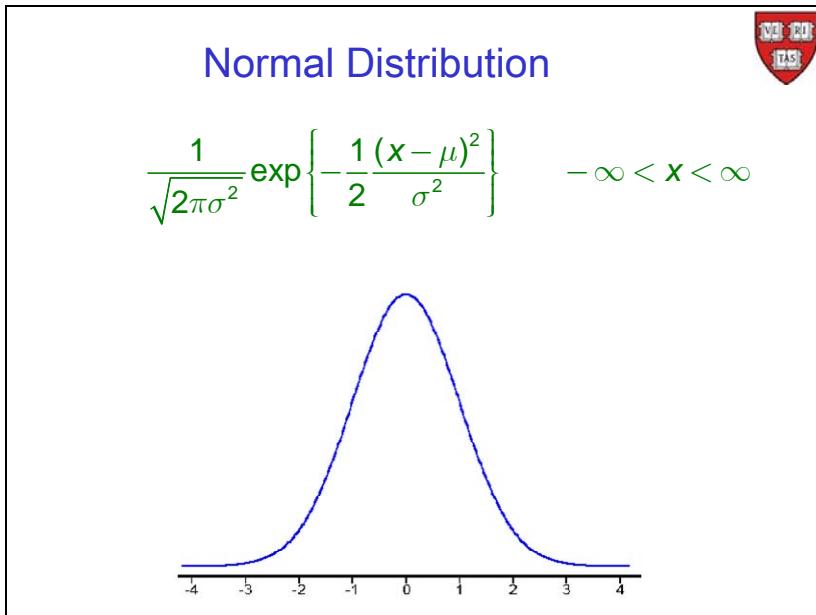
For example, here is the histogram for the logarithm of diastolic blood pressure at visit 1 in the Framingham Heart Study that we looked at earlier in this course. It seems approximately normally distributed.



Here is the glucose level at Visit 1 for the non-diabetics. Except for a small tail on the right, possibly pre-diabetics, they look approximately normally distributed.



Here is another example, the BMI, the Body Mass Index, at Exam 1, also from the Framingham Heart Study. They too look approximately normally distributed. So let us look at the normal distribution more closely.



Here is the equation for the normal density function, and a plot of one such, namely the one called the standard normal; when the mean,  $\mu = 0$  and the standard deviation,  $\sigma = 1$ . You can see the symmetry around zero. All normals are symmetric around their mean (zero in this case). The spread is determined by  $\sigma$ , so making  $\sigma$  larger makes the curve flatter and making  $\sigma$  smaller makes the curve more peaked.



### Properties:

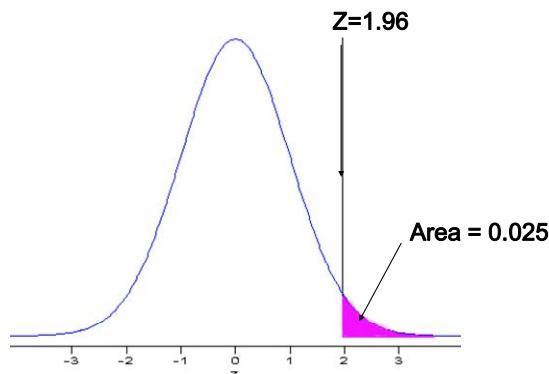
- symmetric about  $\mu$
- spread determined by  $\sigma$
- “Standard Normal”, with  $\mu = 0$  and  $\sigma = 1$ , has been tabulated.

For example, with  $z=1.96$  the area to the right, or probability, is 0.025.

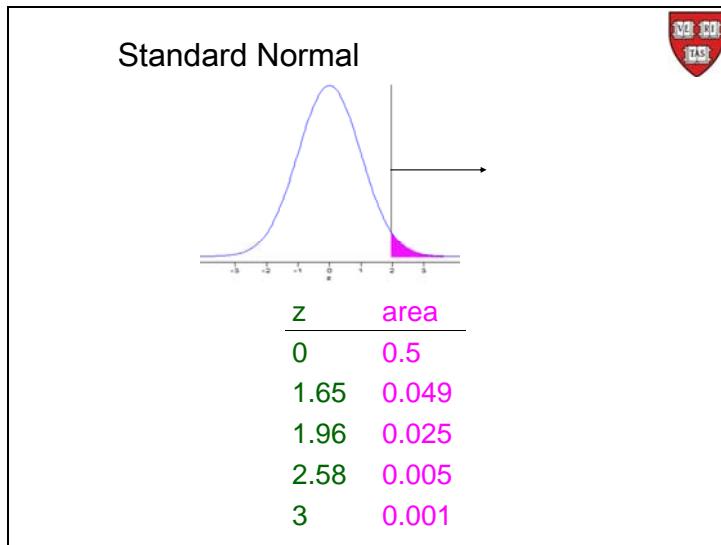
If you look for tabulated values of the normal distribution—unnecessary for us since we have Stata—then the standard normal is the one that gets tabulated.



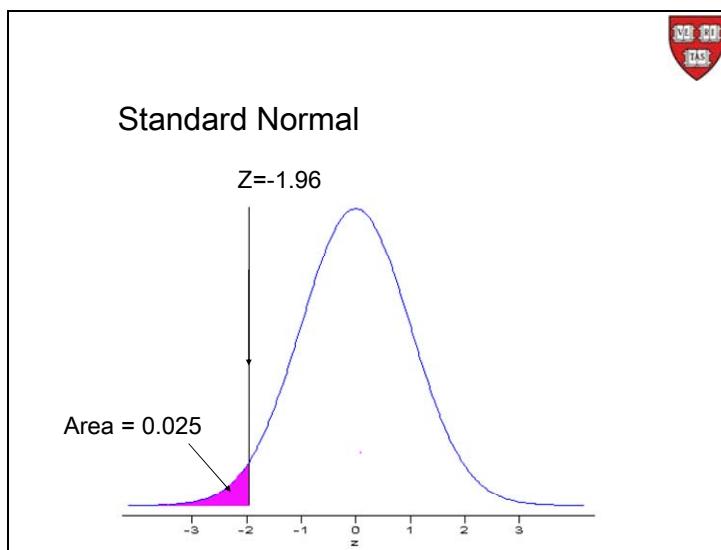
### Standard Normal

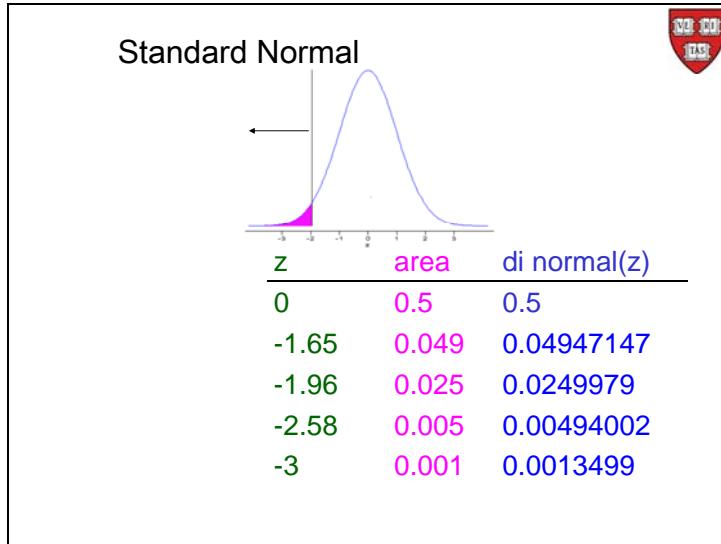


These tables give you the area under the curve, typically the area to the right of a particular  $z$  value. For example, in the standard normal, the area to the right of  $z=1.96$  is 0.025—remember that the total area, very much like the Venn diagram, and for the same reason, is equal to one.

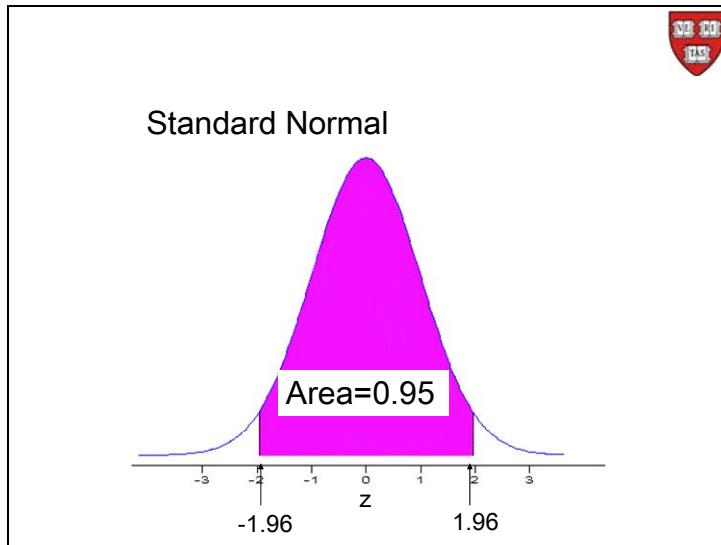


Other popular z values are:  $z = 0$ , then area to the right of 0 is 0.5, which makes sense since the curve is symmetric and the total area is one. The area to the right of  $z = 1.65$  is 0.049. When  $z = 3$ , the area to the right is 0.001.





The curve is symmetric, so the area to the left of  $-z$  is the same as the area to the right of  $z$ . Of course, the same can be said for the area to the right of  $-z$  being the same as the area to the left of  $z$ . The values above are all obtained from the Stata function *normal*.



Noting the symmetry, and that the total area is equal to one, we can also calculate the area between any two values. So, for example we have that the area between  $z = -1.96$  and  $z = 1.96$  is 0.95. (The area to the right of 1.96 is 0.025, therefore that is the area to the left of -1.96, and thus the area between -1.96 and 1.96 is  $1 - 0.025 - 0.025 = 0.95$ .)

That is exactly 95%. No more approximation, like we had in the empirical rule. This is the idealization of the empirical rule. The mean plus or minus 1.96 standard deviations gives us exactly 95% of the data. I leave it to you to calculate the exact area for plus or minus one, and three standard deviations.



### General Normal

Suppose  $X$  is a normal random variable with mean  $\mu$  and standard deviation  $\sigma$ , then

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal (mean zero, standard deviation one).

Let us look at a few ways of applying the normal distribution. When, in the past, we relied solely on tabulated values for the normal we reduced everything to the standard normal and proceeded from there—very similar to the idea of standardization we introduced a few weeks back. We can repeat that thinking here.

When dealing with a general normal variable that has a mean  $\mu$  and a standard deviation  $\sigma$ , then standardize it by subtracting  $\mu$  and dividing by  $\sigma$ :



### Predictive Interval

95% of the time:

$$-1.96 \leq Z \leq 1.96$$

$$-1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96$$

$$-1.96\sigma \leq X - \mu \leq 1.96\sigma$$

$$\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma$$

So let's look at an example of the use of this idea. We know that a standard normal will fall in the interval  $(-1.96, 1.96)$  95% of the time; that is what we call a *predictive interval*. We cannot tell you exactly what value you are going to observe, but 95% of the time, it will be in this interval.

We can translate this into a predictive interval around  $X$  by using the standardization formula, to get that 95% of the time,  $X$  will take a value that is in the interval  $(\mu - 1.96\sigma, \mu + 1.96\sigma)$ .

e.g. If  $X$  denotes systolic blood pressure, then approximately normal. For 18-74-year-old men in US the mean is 129 mm Hg and the stand. dev. is 19.8 mm Hg.

So

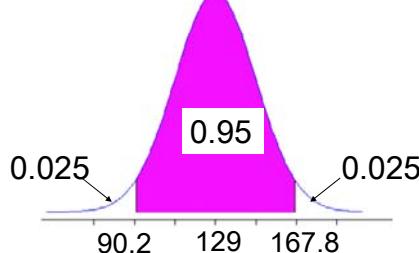
$$Z = \frac{X - 129}{19.8}$$

is standard normal. So if

$$1.96 = \frac{X - 129}{19.8}$$

then  $X = 167.8$

So for example, let us apply this to measuring systolic blood pressure on 18 to 74-year-old men in the US. We know that for this group, their systolic blood pressure is approximately normally distributed with mean  $\mu = 129$  mm Hg, and a standard deviation  $\sigma = 19.8$  mm Hg. So from the standardization formula we get that  $X = 167.8$ . So in this population we are going to get a value bigger than 167.8, 2.5% of the time. Similarly we can calculate 90.2 to be the lower 2.5% cutoff.



If we choose a person at random from this population, the probability is 0.975 that the person has systolic blood pressure less than 167.8.

Thus we can make statements like: If we choose a person at random from this population, the probability is 0.975 that the person has systolic blood pressure less than 167.8 mm Hg; or, that 95% of all men in this age group in the US have a systolic blood pressure between 90.2 and 167.8 mm Hg.

How many have blood pressure above  
150 mm Hg.?



$$Z = \frac{150 - 129}{19.8} = 1.06$$

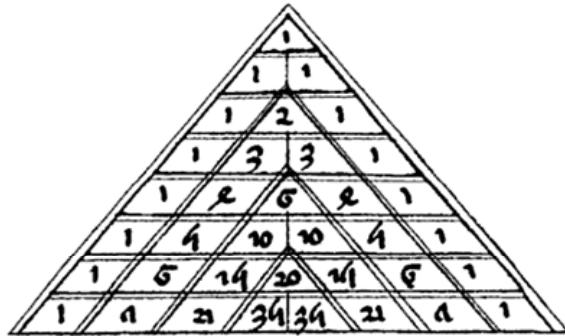
Stata:      > di normal(1.06)  
                > .8554277

So, approximately 14.5% of men in the US  
between the ages of 18 and 74 have systolic  
blood pressure above 150 mm Hg.

We can also turn things slightly around and ask statements directly on the x scale. For example, how many have blood pressure above 150 mm Hg? In order to answer that question we have to go from our x scale of 150, subtract the mean, divide by the standard deviation, to a statement about our standardized Z. And it turns out to be 1.06.

So we can ask Stata what is the area to the left of 1.06, and it comes back with 0.855. One minus this would be 14.5%. So, approximately 14.5% of men in the US between the ages of 18 and 74 have systolic blood pressure above 150 millimeters off mercury.

These are some of the questions we can use the models to answer.



De Arithetica by Jordanus de Nemore

Marcello Pagano

# [JOTTER-WEEK 5 SAMPLING DISTRIBUTIONS]

Central Limit Theorem, Confidence Intervals and Hypothesis Testing

## Inference

This is when the magic starts happening.

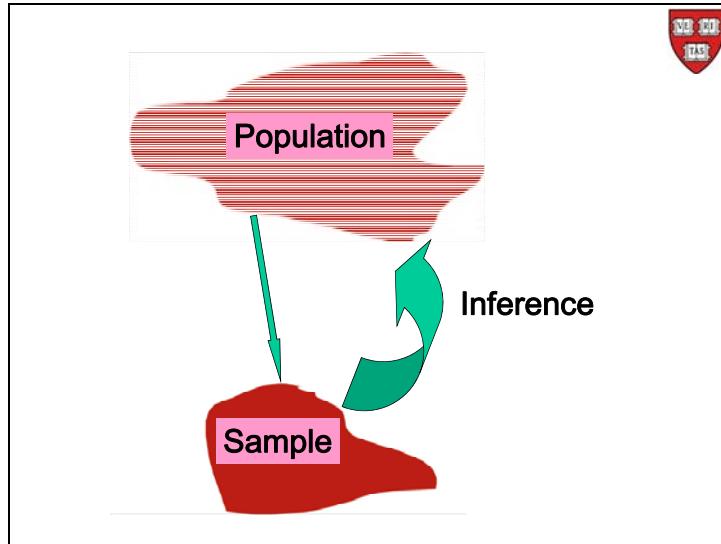


### Statistical Inference

Use of inductive methods to reason about the distribution of a population on the basis of knowledge obtained in a sample from that population.

First, let us define inference the way we use it in this course: Inference is the use of inductive methods to reason about the distribution of a population. What does that mean?

Well, we are interested in a population and the distribution of certain variables within this population. For example, how does cholesterol level vary within this population? How does it vary for women? How does it vary for men? Is there a difference between the two sexes in their cholesterol distributions? We want to be able to answer such questions about the population, without having to measure the whole population. We are going to infer what these answers might be on knowledge we obtain in a sample, or a subset of that population.



Schematically, what we have is a population that is the focus of our interest. We would love to measure everything on everybody in this population, because then our task would be completed. But we cannot do that. So we take a sample from this population. Then on the basis of this sample, we make inference about the population. This inference is our challenge in this module.

**Population**

- Can be real, can be conceptual.
- Can be past, current or future.
- The more homogeneous it is,  
the easier it is to describe.

e.g. Let us think of the Framingham Heart Study participants as our population.

Let us look at these three components one by one. First we have the population. It can be real. For example, it could be everyone you know, everybody in your town, everybody in your province, in your state, in your nation; just a group of people.

The population can be real or it can be conceptual. It could be everybody in the future who is destined to get cancer and we might be interested in how we are going to treat them. Or it could be even more conceptual: we could be asking the question, what would happen if we treated everybody with this particular treatment? If instead of that treatment we choose another. Then what would happen to the patients? These “what if” conceptual exercises can help us decide what treatments would be best for future patients.

The population can be in the past—what happened in the 1917-18 flu pandemic? It can be current. It can be in the future, as mentioned.

The more homogeneous the population the easier it is going to be for us to describe or measure it. So, for example, if we were all four feet tall, then that not only would be very easy to describe, but we would only need a sample of size one—measure a single person's height—to tell us how tall we all are. So the more homogeneous it is, the easier it is to describe, and the smaller the sample we will need.

Now, what we are going to do henceforth as an exercise is to take the Framingham Heart Study data on all 4,434 patients and treat them as our population. Ordinarily, we do not know this much about our population, but as a pedagogical license, let us consider them as our population. In this manner we know what the “truth” is that we ordinarily would be trying to estimate.



Our Population					
Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totchol1	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmi1	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

Here are the summaries of the first seven variables in our population. So, for example, approximately 35% of the 4,434 died. Roughly 16% experience angina. Looking at total cholesterol at the first visit, we see that the average was about 237. With this last variable we only have 4,382 measurements. For now we ignore the fact that some measurements are missing, but for the ones we have, they average out at about 237. Their standard deviation is about 44.6. And so on, for all the other variables in the data set. That is our population for now.



## Sample

- Must be representative  
(random).

e.g. Suppose we take a sample of size 49 from our population.

Second, take a sample from this population. If we want to make inference about the population on the basis of a sample from that population, then surely we would want our sample to be representative of the population. For example, if our population is 50% male, 50% percent female, we do not want a sample that's all male or all female especially if sex is an important consideration for the outcome that we are measuring. So we hope the sample is typical or representative of the population.

Now think about this logically a little bit, is it possible to have a truly representative sample. How do we know that the sample is representative? With respect to one variable, sex, it means that the sample should be 50% female, and thus 50% male.

Now what about age? Well we should have the same distribution of age in the sample as we have in the population; first amongst the females and then amongst the males. Then what about BMI? Well we need the same distribution of BMI amongst females within all the age groups, and amongst males within their age groups too. We must come to the conclusion that first, it is impossible to be truly representative if we literally mean that we have to have the same distribution on every single measure we can imagine. It is not possible unless the sample is as big as the population!

Number two, how would we know if we achieve representativeness in all dimensions? The only way we would know it is if we actually knew the population, in which case why are we measuring it? Why are we even taking a sample? So the answer to the question of whether our sample is truly representative, the answer is, we doubt it although we do not really know.

To overcome this problem we turn to a random device; we are going to take a random sample. What that means is that not everybody in the population will be in the sample, but everybody has an equal chance of being in the sample. And we do this because theory tells us that, *on average*, we will have a representative sample.



Variable	Obs	Mean	Std. Dev.	Min	Max
death	49	.3265306	.4738035	0	1
angina	49	.1632653	.3734378	0	1
totcholl1	48	233.6667	42.73836	157	410
sysbp1	49	140.3571	22.39745	96	196
diabp1	49	86.53061	12.65669	48	119
bmi1	48	27.21167	4.018565	19.2	36.65
glucose1	43	79.32558	12.2255	60	114

So let us try it with our new “population”. Suppose we take a random sample size 49 from our population. To do so we can call on Stata.

Stata has a “random” device built in. It’s not truly random. It’s called a pseudo-random number generator. And what you do with a pseudo-random number generator is give it a seed; some random number—the manual suggests taking out a dollar bill from your pocket and looking at the number on the bill. The number of my dollar bill is G72576466A. I removed the G and the A, and that is the seed that you see above.

Having set the seed, the computer goes about generating numbers seemingly at random. These numbers are unpredictable to someone who does not know the inner workings of the computer (us), but every time I give the same seed it returns the same sequence of “random” numbers, and thus the label “pseudo”. This property we need in science so as to enforce the much valued reproducibility of our experiments. On the other hand, we cannot predict what they are going to be, unless we have seen them before. So we have achieved an oxymoron, reproducible random numbers.

So you ask Stata to choose 49 people at random from the population. Returning to the seven variables we had chosen above (the first seven), we have 33% deaths in our sample, whereas we had 35% deaths in the population. The percent with angina is 16% in the sample, and the mean of the total cholesterol level is 233 and so on.

. summ death angina totcholl sysbp1 diabp1 bmil glucose1					
Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totcholl	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmil	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

. summ death angina totcholl sysbp1 diabp1 bmil glucose1					
Variable	Obs	Mean	Std. Dev.	Min	Max
death	49	.3265306	.4738035	0	1
angina	49	.1632653	.3734378	0	1
totcholl	48	233.6667	42.73836	157	410
sysbp1	49	140.3571	22.39745	96	196
diabp1	49	86.53061	12.65669	48	119
bmil	48	27.21167	4.018565	19.2	36.65
glucose1	43	79.32558	12.2255	60	114

So if we compare the summaries of these seven variables in our population of 4,434 to our sample of 49, we see that, by and large the sample reproduces the population summaries.

The range for total cholesterol level is from 107 to 696. The range in the sample, of course, has to be smaller. That is why we are forever breaking records in sports, et cetera.



Variable	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
death	4434	.3495715	.4768884	49	.3265306	.4738035
angina	4434	.1635092	.3698714	49	.1632653	.3734378
totcholl	4382	236.9843	44.6511	48	233.6667	42.73836
sysbp1	4434	132.9078	22.4216	49	140.3571	22.39745
diabp1	4434	83.08356	12.056	49	86.53061	12.65669
bmil	4415	25.84616	4.101821	48	27.21167	4.018565
glucose1	4037	82.18578	24.39958	43	79.32558	12.2255

Putting the variables side by side we see that the sample means and the population means are quite close to each other. Indeed, except for the glucose1, where the the sample standard deviation is one half the population standard deviation, we can say that the sample standard deviations are good estimators of the population standard deviations.

In general, you can see why it makes sense to use this sample to make inference about the population. The sample of size 49, turns out that that is a pretty big sample.

## Sample



- Must be representative  
(random).
- The bigger the sample,  
the better our inference.

In fact, the bigger the sample, the better our inference. We'll quantify this a little better later, but for now, let us explore this issue.

```
. set seed 72576466
. sample 10, count
(4424 observations deleted)
. summ death angina totcholi sysbp1 diabp1 bmil glucosel
```

Variable	Obs	Mean	Std. Dev.	Min	Max
death	10	.3	.4830459	0	1
angina	10	.1	.3162278	0	1
totcholi	10	238.9	72.50969	157	410
sysbp1	10	138.35	18.19043	106	173
diabp1	10	83.3	14.06374	48	98
bmil	10	25.642	3.375404	19.2	30.91
glucosel	10	81.2	12.05358	60	97

What I did is I set the seed the same as before. Now this time I asked for a sample of size 10. And because I did this—you should not do this, since it introduces too much predictability. But I did this on purpose so that these 10 are actually the first 10 that went into making up the 49 in the previous sample.

Variable	Obs	Mean	Std.D.	Obs	Mean	Std.D.	Obs	Mean	Std.D.
death	4434	.350	.48	49	.326	.47	10	.3	.48
angina	4434	.164	.37	49	.163	.37	10	.1	.32
totchol1	4382	237.0	44.7	48	233.7	42.7	10	238.9	72.5
sysbp1	4434	133.0	22.4	49	140.4	22.4	10	138.4	18.2
diabp1	4434	83.1	12.1	49	86.5	12.7	10	83.3	14.1
bmi1	4415	25.8	4.10	48	27.2	4.02	10	25.6	3.38
glucose1	4037	82.2	24.4	43	79.3	12.2	10	81.2	12.1

So here are the values. And once again, let me display them in such a way that our comparisons are visually easier to evaluate. And so here is the population. Deaths-- 35%. The first sample of 49 was 0.326. This one, of size 10, is 0.3. It's not as good as the earlier and bigger sample.

The prevalence of angina: 0.164, 0.163, 0.1. So that's not as good, either. The mean total cholesterol level, on the other hand, is better with the population value at 237, the sample of size 49 at 233.7, and the sample of size 10 at 238.9.

That is the problem with random samples, we cannot predict exactly how things are going to look!

We should also look at the sample standard deviations and see how they compare to the population values that they are estimating. In all instances except for the first and last variable (where they are approximately tied) the standard deviation for the sample of size 49 is closer to the "truth" than is the sample of size 10.

So things are random. But we shall see, in a moment, that as a general rule, a sample of size 49 is better than a sample of size 10. Now how big a sample do we need?

REAL CLEAR POLITICS General Election: McCain vs. Obama						
Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
<b>Final Results</b>	--	--	--	<b>52.9</b>	<b>45.6</b>	<b>Obama +7.3</b>
<b>RCP Average</b>	<b>10/29 - 11/3</b>	--	--	<b>52.1</b>	<b>44.5</b>	<b>Obama +7.6</b>
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jml	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

One counter intuitive result is that, unless we are talking about small populations where the sample is a sizable fraction of the population, how big a sample one needs does not depend on the size of the population. Case in point, here are the results from a Real Clear Politics of the prior US Presidential election that took place on the 4<sup>th</sup> of November 2008<sup>1</sup>. Also shown are the results of the last polls that were taken immediately before the election.

The final results had President Obama getting 52.9% of the vote. And this Marist poll predicted that he would get 52%. McCain actually got 45.6%. And they predicted 43%.

What Real Clear Politics did is also report the average of all those polls and came up with 52.1 for Obama and 44.5 for McCain. I do not think that you can get much closer than that.

What I find amazing is that close to 100 million people voted, and yet each of these polls was based on some number between one and three thousand people. You can predict how 100 million people are going to vote on the basis of what 1,000 people say?

That is the magic. This is the magic of random samples.

---

<sup>1</sup> <http://www.realclearpolitics.com/epolls/2008/president/national.html>



©Fir0002/  
Flagstaffotos



## Inference

From **part** infer about the **whole**.

- Uncertainty
- Probability

2

Having looked at the population and the sample, let us now look at the third component, inference. To repeat, we want to infer about the whole, the population, on the basis of a random sample from the population. And we want to do it in a principled way.

Our primary challenge is how to deal with the uncertainty inherent in what we wish to do. We are only privy to a small part of the whole population, and yet we have to generalize to the whole. We do not want to do like the Hindu parable of the blind men feeling the elephant, and depending upon what part of the elephant they were feeling, they projected or they inferred about the rest of the elephant, all leading to quite different pictures.

We now introduce probability into the argument to measure the uncertainty.

---

<sup>2</sup> [http://en.wikipedia.org/wiki/File:Asian\\_elephant\\_-\\_melbourne\\_zoo.jpg](http://en.wikipedia.org/wiki/File:Asian_elephant_-_melbourne_zoo.jpg)

## Sampling Distribution of the Mean



# Sampling Distribution

[Rice University](#)

[http://onlinestatbook.com/stat\\_sim/sampling\\_dist/index.html](http://onlinestatbook.com/stat_sim/sampling_dist/index.html)

We first introduce the concept of a sampling distribution. It will provide us the vehicle to quantify the uncertainty in our inference. And the way we do that is to go to the above website which is run by the statistics department at Rice University.



## Sampling Distribution

[Begin](#)

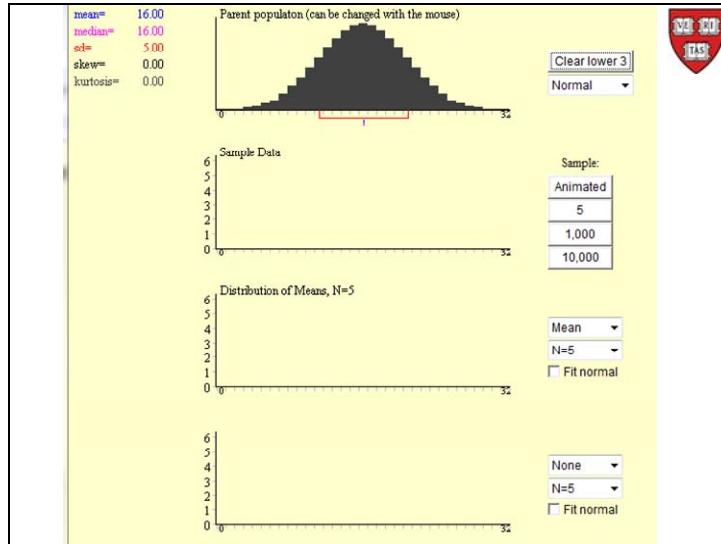
[Instructions](#)  
[Exercises](#)

### Instructions

Please wait until a button appears below the distribution. Then the Java applet will open in a new window.

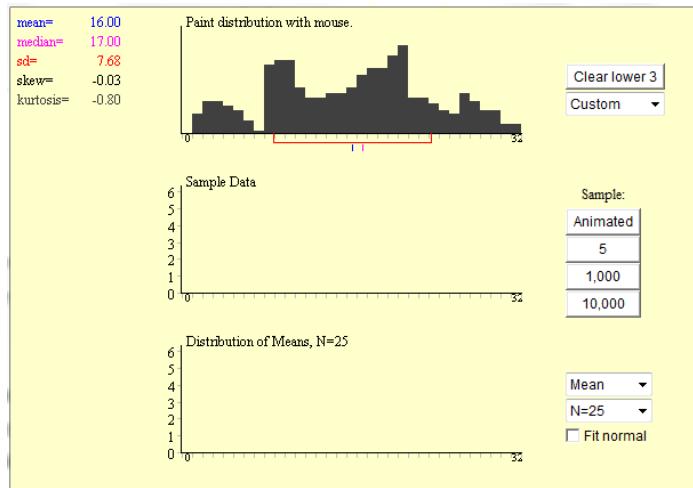
This Java applet lets you explore various sampling distributions. A normal distribution is displayed at the top. The distribution portrayed at the top of the applet is indicated by a small bell curve icon.

When you get there you click on “Begin”.



So we have before us a population. And that is the top panel here. We are going to take a sample from this population. And you will see the sample displayed in the second panel down. The third panel will keep a running summary of what we see in our samples because we're going to do the sampling repeatedly. In reality, of course, we only take a single sample most of the time. But for now, to get this concept of a sampling distribution under our belts, let's think of taking repeated samples, and pay attention to the summary.

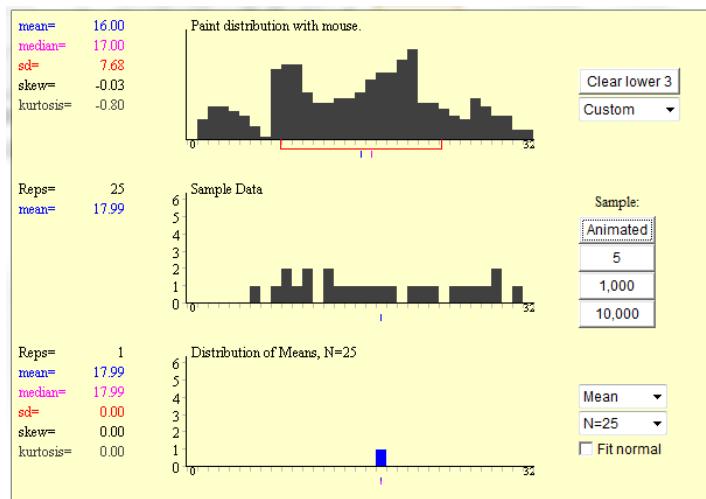
On the left at the top we see summaries about this population. For example, we see that the mean is 16 for this population, the median is 16, and the standard deviation is 5. We also see the skewness and kurtosis, but I am not going to be addressing them. I leave that up to you to look up, if you are interested. Also, for now, we ignore the bottom panel.



Let us start making choices. There are a number of statistics we can choose in the bottom-right corner. Let us stay with the mean. We have to decide on a sample size we want. Let us say N=25.

Now, let us look at the population that we have up here. We can choose the normal, or the uniform, or a skewed population, or we can make up our own custom population. Let us do that. Let us just make up a custom population.

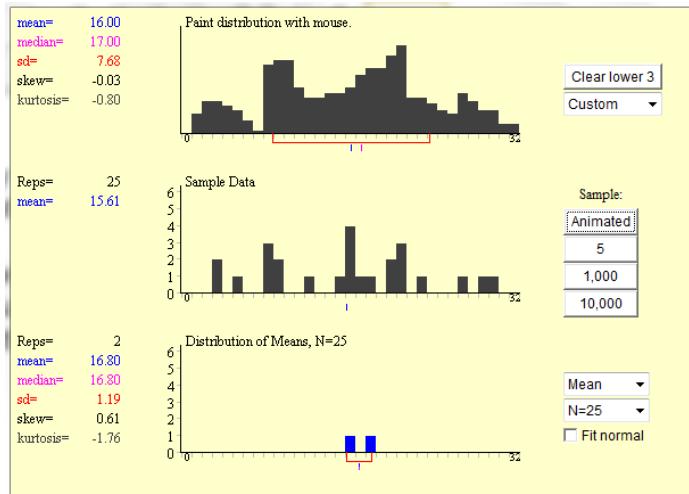
So here is the distribution of the population. I just made it up. I just love to do it with this and things work nicely. So this population has a mean of 16, a median of 17, and a standard deviation of 7.68.



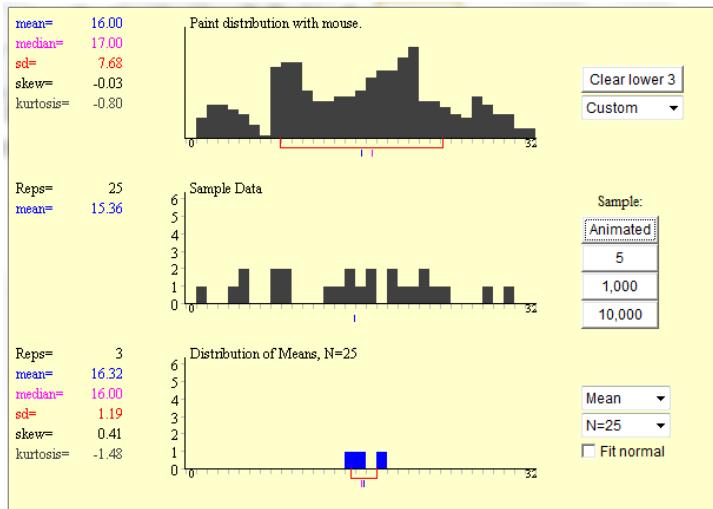
Now, we take a sample from this population. That's what happens when I click the "Animated" button on the right. So here, in the middle panel, are 25 observations from this population.

And here they range pretty much over the whole range, just like the population. We see on the left that the mean of these 25 values, the sample mean, is 17.99.

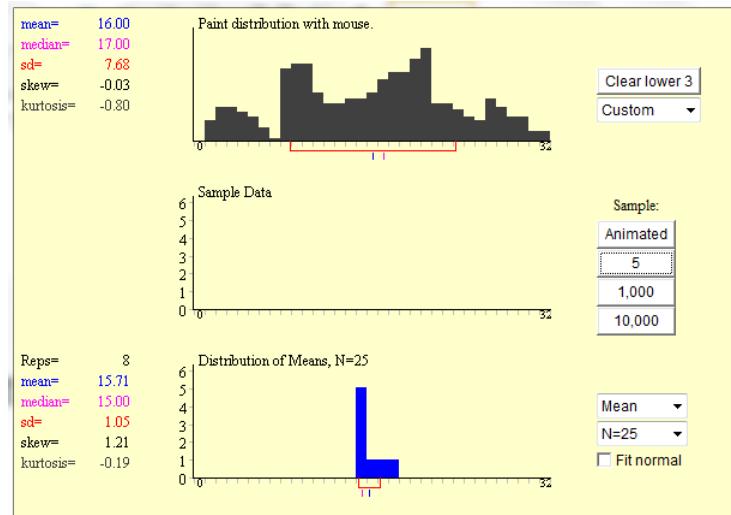
Remember we're trying to estimate the mean of the population, which is 16, on the basis of this sample of 25 observations. And the mean is 17.99. What the software does is also retain that mean, the mean of the sample, in the third panel down.



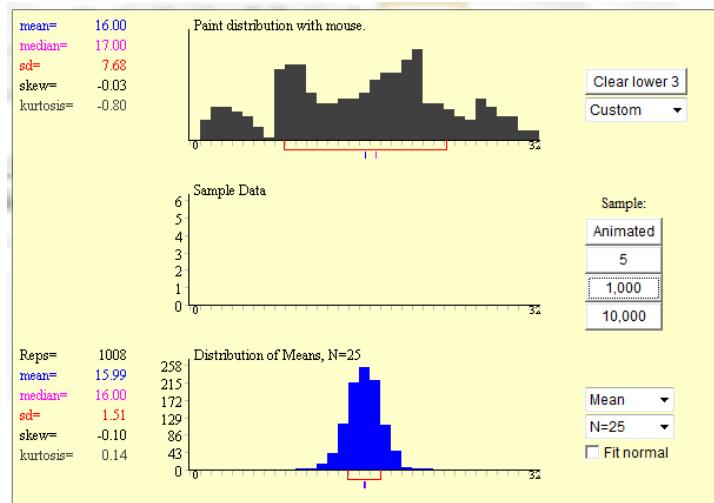
Let's take another sample of size 25. This time we'll get another mean. This mean is 15.61. We could use this to estimate the population mean of 16. But what we now also have (bottom left-hand corner) is that the mean of the two sample means—the earlier, 17.99, and the current, 15.61—is 16.80.



We can repeat this process. Henceforth we will have three means on the screen. The mean of the population (16), the mean of the current sample (15.36), and the mean of each of the sample means—three in this case—16.32. It looks like the last mean is getting closer to the mean of the population, 16—17.99, 16.88, and 16.32.

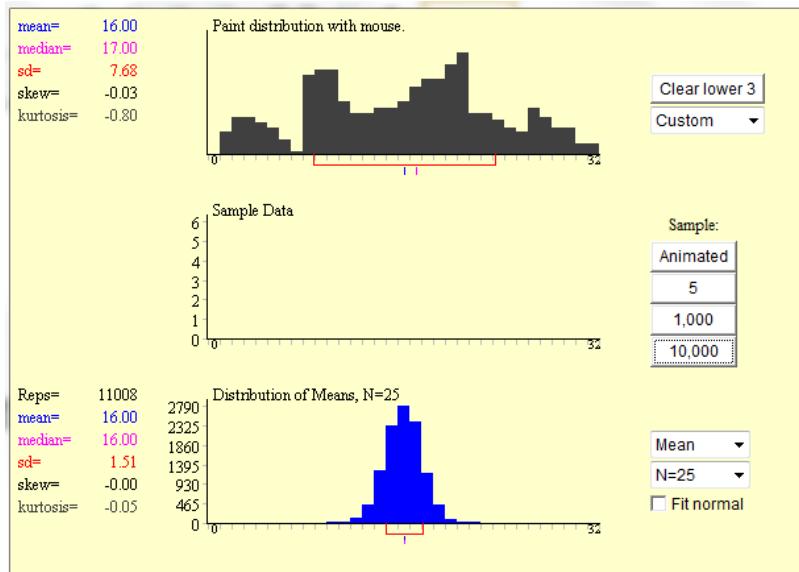


We can speed this up. We can take five samples by clicking the 5, just under “Animated”. So now we have taken eight samples, and we see that the mean of these eight is 15.71. We are getting closer to 16, the mean of the population.



Now we could keep on doing this. But let's just do it 1,000 times by clicking the button under the 5 on the right.

Now, the mean of the means is 15.99. It is getting ever closer to 16.

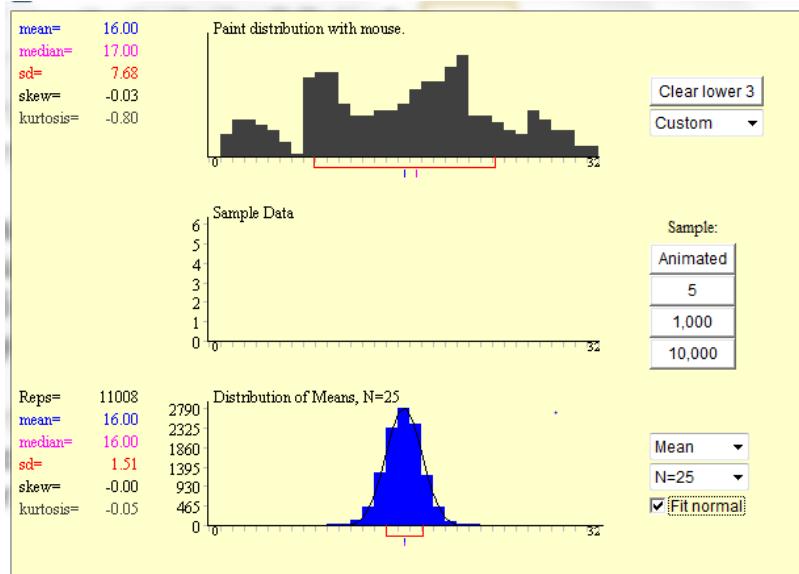


Let us do another 10,000 times by pressing the 10,000 button under the 1,000 on the right.

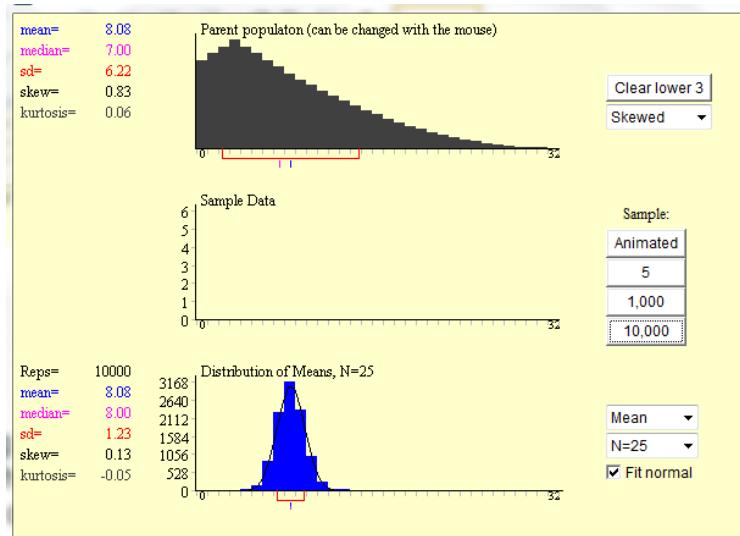
Now the mean of the means is 16. We have gotten there! Indeed, the statistical theorem states: as the sample size gets bigger (it is only 25 in this case) the mean of the means approaches the population mean. (To be correct, we should add that the variance of the population needs to exist.)

Let's look at the distribution of the mean of the means. They look tighter, they look more closely bunched together than the population. Of course, if we look at the spread of the distributions we see that the population standard deviation is 7.68, whereas the standard deviation of the distribution of the sample means is 1.51.

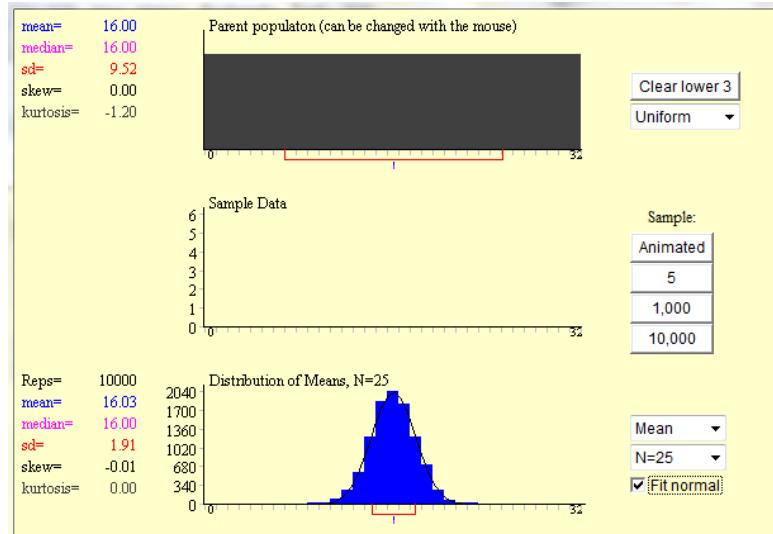
There is a relationship between the means of these two distributions—they are the same—so is there a relationship between these two standard deviations. The answer is in the affirmative. If we look at their ratio it is 5.1, or rounding off, approximately 5. This is the square root of the sample size we have been looking at all along, namely 25. And that is the relationship: divide the population standard deviation by the square root of the sample size to get the standard deviation of the distribution of the sample means. Further, this latter standard deviation has a name: it is called the *standard error*.



Lastly, what is the shape of the distribution of the sample means at the bottom of the picture, above? It is very much reminiscent of what we saw with the Quincunx, and for exactly the same reason. I clicked the box “Fit normal” at the bottom right-hand corner and the computer superimposed a normal curve, which, as you can see, does a good job of describing the distribution.

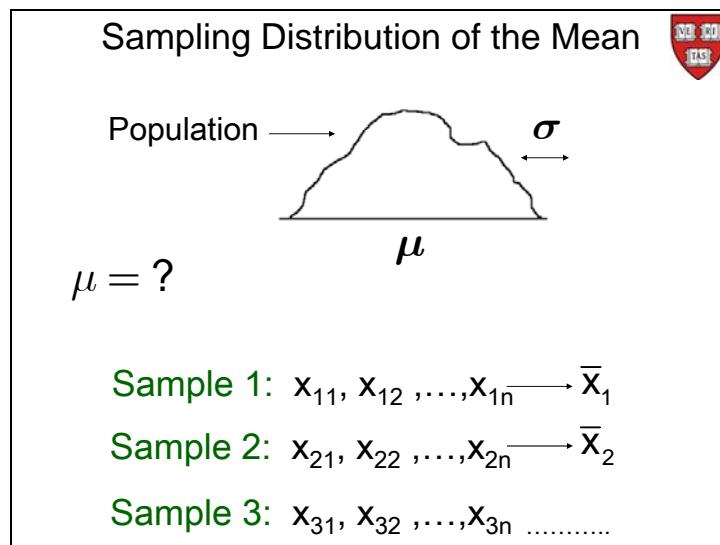


Let us return to the program and choose a skewed population distribution. Let us now take 10,000 samples of size 25, and we see, once again, a lovely bell shaped curve at the bottom.



If we choose a uniform population distribution, pretty much the same thing happens. The shape is close to normal.

You have just experienced magic. We always get this normal distribution when we look at the distribution of the sample means, from whatever population you can contrive in this program. If the sample size is big enough, here we took 25, we are always going to get this normal distribution at the bottom.



So let us summarize. We start with an arbitrary population. It has a mean,  $\mu$ , and standard deviation,  $\sigma$ <sup>3</sup>. We don't know what these values are typically, and we would like to make inference about the value of this population mean.

Then this concept of a sampling distribution is, first of all, take a sample. Then calculate the mean of the sample. And discard the sample and just retain the sample mean.

Do this again tomorrow. Take another sample, keep the sample mean, get rid of the original observations. And keep on doing this repeatedly. OK

### The Central Limit Theorem



Population:  $\mu, \sigma$

Samples of size:  $n$

Sample means:  $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots$

**Distribution of  $\bar{X}$**

- 1. Has mean  $\mu$
- 2. Has standard deviation  $\frac{\sigma}{\sqrt{n}}$
- 3. Is Normal as  $n \rightarrow \infty$

We typically reserve Greek letters for the quantities we do not know, or the parameters that refer to the population. So the population has mean,  $\mu$ , and standard deviation,  $\sigma$ . We are going to take repeated samples of size  $n$ . Once again, typically we just take one sample. But for now we are talking about sampling distributions, so let us talk about repeated samples. Then for each one of these samples we get a sample mean. Let us focus on the distribution of these sample means.

We saw this distribution displayed in the third panel down. That distribution has mean,  $\mu$ , the same as the population distribution, as dictated by theory. That is result number one.

This distribution has standard deviation, or standard error,  $\sigma/\sqrt{n}$ , which, if  $n > 1$ , is smaller than the population standard deviation, and gets smaller as  $n$ , the sample size, gets larger.

So, when we think of how the sample mean varies from sample to sample, first it does so around a mean that is the same as the population mean. That makes sense.

Number two, it varies much less than the population values—and the factor by how much less, is the square root of  $n$ . Third, it is normally distributed or has a bell-shaped curve. And this is

---

<sup>3</sup> For the purists, we are assuming that  $\sigma < \infty$ .

technically true as  $n$  goes to infinity. That is the theory. This is a most amazing theorem. De Moivre showed this to be true when we took samples from a binomial population, but this now is stated for sampling from quite general population distributions.

All these results require that the sample size  $n$  goes to infinity. Of course, we do not have infinite samples, nor are we likely to get any in the near future, so how do we use this result? We argue that this theorem provides us large sample approximations. What do we mean by large? It turns out that the closer the population distribution is to itself being bell shaped, the smaller the sample we need to have good approximations.

This theorem is called the *central limit theorem*. It actually is central, if I am allowed the pun, to just about all the inference we are going to be making in this course. It is extremely important to make sure you understand the meaning of what we have been saying up to now. Go to the Rice site and experiment for yourself.



**Example:** In our population, the total cholesterol level at visit one had

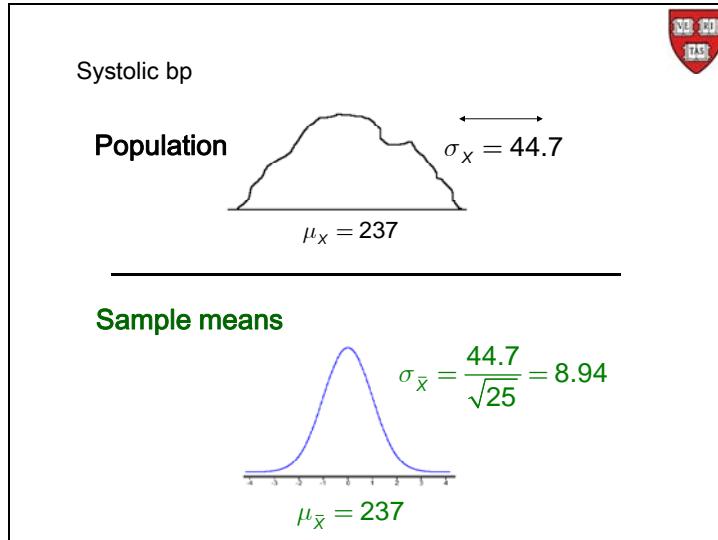
$$\mu_x = 237 \text{ mg / 100ml}$$

$$\sigma_x = 44.7 \text{ mg / 100ml}$$

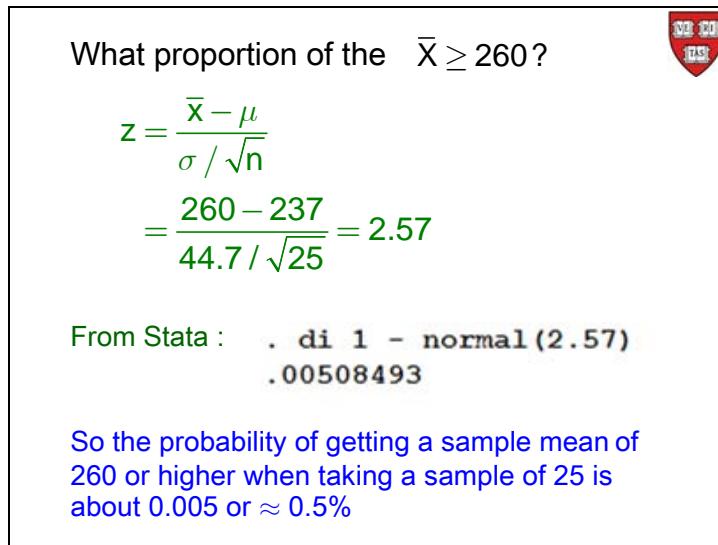
Take *repeated* samples of size 25 from this population. What proportion of these samples will have means  $\geq 260$  mg/100ml ?

How do we use the central limit theorem? Let us take a look at an example. We know that in our population, the total cholesterol level at visit one has a mean of 237 and a standard deviation of 44.7, rounding things off. Now, if we take repeated samples of size 25 from this population, what proportion of these samples will have a mean that is greater than 260?

The mean  $\mu = 237$ , and  $\sigma = 44.7$ , because we know the population from which we are sampling. I repeat, that is not how things typically work in practice, but we are carrying out an intellectual exercise. Let us ask the question, how often will the sample mean—not the population mean—how often will the sample mean, when we take samples of size 25, have a value bigger than or equal to 260?



We can use the central limit theorem because we've got a population with mean 237, standard deviation of 44.7. And we're going to take the sample mean. And we know from the central limit theorem that if we do this repeatedly, those sample means will be distributed around a mean of 237 and a standard deviation of 44.7 divided by the square root of the sample size, which in this case gives us 8.94.



So let us return to our standardized variable by subtracting from 260 its mean, 237, and dividing by its standard deviation, and the answer we get is 2.57. So the question asked about  $X$  bar gets translated into a statement about  $Z$  asking, how often is  $Z$  greater than or equal to 2.57? The answer is 0.005.

So we can answer the question, the probability of getting a sample mean of 260 or higher when taking a sample of size 25 from this population is about 0.5%. So it's something that will happen

probably roughly once in 200 times. So this might be our definition of a rare event. So very rarely, will we see a sample mean of 260 or bigger.

Throughout these calculations we are working under the assumption that n=25 is a large enough sample so that the central limit theorem provides us a good approximation. But that is how we can use the central limit theorem.

## Sample Size



Our Population					
Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
totchol1	4382	236.9843	44.6511	107	696
sysbp1	4434	132.9078	22.4216	83.5	295
diabp1	4434	83.08356	12.056	48	142.5
bmi1	4415	25.84616	4.101821	15.54	56.8
glucose1	4037	82.18578	24.39958	40	394

Now, how can we use the central limit theorem to answer sensible questions? We saw one question we just answered. Another question that's very often asked is, how big a sample do I need? And the answer of course, is for what? But let's look at one situation where we can answer this question by being more precise in our question.



How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within  $\pm 25$  mg/100ml of the population mean?

$$\Pr\{-25 \leq \bar{X} - \mu \leq 25\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \leq \frac{\bar{X} - \mu}{44.7/\sqrt{n}} \leq \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \leq Z \leq \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Rightarrow \frac{25}{44.7/\sqrt{n}} = 1.96 \quad \Rightarrow n = 12.3 \Rightarrow n = 13$$

$$\Rightarrow \frac{12.5}{44.7/\sqrt{n}} = 1.96 \quad \Rightarrow n = 49.1 \Rightarrow n = 50$$

Now suppose we do not know, as is usually the case, what the population is. And we wish to take a sample. So the question might be phrased this way. "How big a sample do we need to be 95% sure that the sample mean for total cholesterol level is within plus or minus 25 milligrams per 100 milliliters of the population mean?" So we state how sure we need to be and the degree of accuracy we seek.

Here, we use the ubiquitous 95%. So how big a sample do we need to be 95% sure? So I am not going to be certain, when I get my answer, but I shall be 95% sure that what we did will get me to within 25 units of the population mean. I am not certain that what I am going to do will be within the limit that you have set, but I am 95% sure. That's the best I can do.

So let us standardize. Divide by the standard deviation of the sample mean, or the standard error. Which, because we know the population standard deviation—I am being a little bit unrealistic here, I am acting as if I know the population standard deviation, I realize that. We remove that assumption shortly, but let us just answer this question first.

Divide through by the standard error to get our standardized Z. So now we are asking what is the probability that Z will be between minus this quantity and plus the quantity? We want that to be 0.95. So we know, that a standard Z, is between  $-1.96$  and  $1.96$  95% of the time. And in order to do that, because n is the only unknown in here, we can solve for n. And it turns out that n of 13 will allow us to be 95% sure that we are within plus or minus 25 units of the population mean by using the sample mean.

Suppose we want to lower that 25. Let's halve it to 12.5. Then we would need a sample of size 50.



## Sample Size

So, in general if we want to be 95% sure that the sample mean will be within  $\pm \Delta$  of the population mean, then we need a sample of size

$$\left( \frac{1.96\sigma}{\Delta} \right)^2$$

where  $\sigma$  is the population standard deviation.

In summary, if we want to be 95% sure that the sample mean is within plus or minus delta of the population mean, then we need a sample of size n that equals this quantity, where sigma is the population standard deviation.

This formula is appealing. The bigger is our sigma, the more variable is our population. And we have been saying all along, the less homogeneous the population, the bigger the sample we are going to need. And the smaller we make delta, the more precise you want to be, then the larger your sample size will have to be. So these are the two controlling factors here, how close you want to be and how variable is your population.

## Confidence Interval



Confidence Interval on  $\mu$  ( $\sigma$  known)

$$\Pr \left\{ -1.96 \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq 1.96 \right\} = 0.95$$

$$\Pr \left\{ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right\} = 0.95$$

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

is a 95% confidence interval for  $\mu$ .

The next topic we explore is the ever popular confidence interval. I am going to keep on acting as if we know sigma. As I said, we will remove that shortly. But for the time being, just to keep the conversation straightforward without extra complexity, let's just act as if we know sigma.

So here is our standardized variable. From the central limit theorem, we know that  $X\bar{}$  has got mean  $\mu$  and standard deviation (standard error)  $\sigma/\sqrt{n}$ . Let us assume that  $n$  is large enough so that we can make the statement that our standardized variant will take on values between minus 1.96 and 1.96 95% of the time.

Sigma is positive. Square root of  $n$  is positive. So let us multiply through by the standard error, and that will not disturb the inequalities. Now subtract  $X\bar{}$  from both sides, and then multiply everything by minus 1 and reverse the inequalities. And there you have it, after just a little bit of algebra. These two probability statements are exactly the same. But look at what we have done: From sample to sample, in the first probability statement, the standardized variable is what varies. By doing that little bit of algebra, from sample to sample, what now varies are the limits of our interval— $\mu$  remains fixed.

So we have created what is called a random interval. We now have an interval that is from  $X\bar{}$  minus, to  $X\bar{}$  plus,  $1.96 \sigma/\sqrt{n}$ . This interval, is going to vary from sample to sample, because  $X\bar{}$  does. Furthermore, this interval will cover  $\mu$  95% of the time. And, of course, roughly 5% of the time it will not cover  $\mu$ . Unfortunately, on any one occasion we do not know whether we are in the 95% or the 5% !

This interval is called a confidence interval. We are 95% confident before we do any calculations, that the resultant interval will cover  $\mu$ . Once we have done our calculations, we do not know. Maybe the interval includes  $\mu$ . Maybe it does not. What we have here is a recipe, one

that has a 95% chance of success. That is what a confidence interval is. It is a rule that has a 95% chance of success; success being measured by whether the interval contains  $\mu$ .

### Confidence Interval on $\mu$ ( $\sigma$ known)



**Before** taking the sample:

$$\Pr \left\{ \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right\} = 0.95$$

the interval:

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

has a 95% chance of covering  $\mu$ .

So before we take the sample, we can make the statement that there is a 95% chance that it will do the right thing. So we are 95% confident that we will capture  $\mu$  with this interval.

### 95% confidence interval for total cholesterol mean



e.g. If we take a sample of size  $n = 49$  from our Framingham population where  $\sigma = 47.7$  mg/100ml for the total cholesterol level, then the interval

$$\left( \bar{X} - 1.96 \frac{47.7}{\sqrt{49}}, \bar{X} + 1.96 \frac{47.7}{\sqrt{49}} \right)$$

$$(\bar{X} - 13.4, \bar{X} + 13.4)$$

has a 95% chance of covering  $\mu$   
(which we know is 237.0 mg/100ml).

Consider the total cholesterol level at visit one from our Framingham population. If we take a sample of size 49 and assume that the standard deviation is 47.7, then our 95% confidence interval is  $X\text{-bar}$  minus 13.4, to  $X\text{-bar}$  plus 13.4, and that interval has a 95% chance of covering the true mean.

### 95% confidence interval for total cholesterol mean



We observed  $\bar{x} = 233.7$  so the 95% confidence interval is (220.3 , 247.1).

Here we know the answer because we know the population, but in general, this interval may or may *not* contain the mean of the population, we do not know. **But** we followed the rules that give us a 95% chance of being correct- confident.

In our example, we know what the true mean is, so we can check to see if, indeed, it worked or not, because we're just doing this as an exercise. So let us see what happens here. We actually observe a sample mean of 233.7. So our 95% confidence interval is 220.3 to 247.1. And in this case, we know it covered it. Typically, we don't know if it covered it or not.

### Predictive versus Confidence Interval

#### Predictive Interval



If we have a Normal distribution with mean  $\mu$  and standard deviation  $\sigma$ , then for a single observation  $X$ ,

$$Z = \frac{X - \mu}{\sigma}$$

$$\Pr \left\{ -1.96 \leq \frac{X - \mu}{\sigma} \leq 1.96 \right\} = 0.95$$

$$\Pr \{ \mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma \} = 0.95$$

There is a closely related concept to confidence intervals and that is the predictive interval. We have seen this before. What we have said is, if we are to sample from a normal population, where do we expect an observation to fall? We do not know, but we can give an interval with an

There is a closely related concept to confidence intervals and that is the predictive interval. We have seen such an interval before. The way we introduced it is: if we are to sample from a normal population, where do we expect an observation to fall? We do not know, but we can give an interval with an associated probability as an answer to that question. That is what a predictive interval is.

We start with a variable, standardize it, and then the standardized variable will take a value between  $\mu - \sigma$  and  $\mu + \sigma$  95% of the time. Before we measure the variable  $X$ , whatever value it takes, we can predict that 95% of the time that value is within two standard deviations of the mean. That is the predictive element.

So,  $(\mu - 1.96\sigma, \mu + 1.96\sigma)$

is a **predictive** interval for  $X$ , just as

$$\left( \mu - 1.96 \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

is a **predictive** interval for  $\bar{X}$ , and

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

is a **confidence** interval for  $\mu$ .



We have the predictive interval for  $X$ , what about  $\bar{X}$ ? Because of the central limit theorem, we have a parallel result for  $\bar{X}$  if we replace the standard deviation with the standard error.

This is in contrast to what we have just seen, namely the 95% confidence interval. The latter we calculate *after* we have taken our sample. So the confidence interval is *ex post*, whereas the predictive interval is *pre*.



e.g. So for total cholesterol at visit 1,  $\mu = 237$  and  $\sigma = 44.7$ , so

$$(\mu - 1.96\sigma, \mu + 1.96\sigma) = (149.4, 324.6)$$

is a 95% **predictive** interval for  $X$ , just as

$$\left( \mu - 1.96 \frac{\sigma}{\sqrt{n}}, \mu + 1.96 \frac{\sigma}{\sqrt{n}} \right) = (224.5, 249.5)$$

is a 95% **predictive** interval for  $\bar{X}$ , and

$$\begin{aligned} \left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) &= (\bar{X} - 12.5, \bar{X} + 12.5) \\ &= (221.2, 246.2) \end{aligned}$$

is a 95% **confidence** interval for  $\mu$ .

Applying these ideas to our total cholesterol level at visit one, the mean of the population is 237, with a standard deviation of 44.7. So we would predict that 95% of the time that we choose an individual from this population, that individual, we will have a value that is between 149.4 and 324.6.

If we choose a sample of size 49 from this population then we predict that 95% of the time the mean of such a sample will be between 224.5 and 249.5. This is a much tighter interval because the standard error is one seventh the standard deviation.

Once we have that the sample mean is 233.7, we can construct the confidence interval. We see that we are successful with this confidence interval because it covers the true mean of 237.

### Width of the Confidence Interval

#### Width of Confidence Interval



		<u>width</u>
95%	$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$	$3.92 \frac{\sigma}{\sqrt{n}}$
99%	$\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$	$5.16 \frac{\sigma}{\sqrt{n}}$

- As confidence increases (95% to 99%) the width of the interval increases.
- As the sample size increases, the width decreases.

Typically we would like to make the confidence interval as tight as possible. One cheap way to achieve that is to place less confidence in the interval. For example, we see that if we want to be 99% confident then that interval has width  $2 \times 2.58 \times$  the standard error. On the other hand, to be 95% confident the interval only has width  $2 \times 1.96 \times$  the standard error; about 76% the size.

The other variable that determines the width of the interval is the standard error, which we know we can decrease by increasing the sample size.



n	95% CI for $\mu$	Interval width
10	$\bar{X} \pm 0.620\sigma$	$1.240\sigma$
100	$\bar{X} \pm 0.196\sigma$	$0.392\sigma$
1000	$\bar{X} \pm 0.062\sigma$	$0.124\sigma$

Smaller is  $\sigma$ , the tighter are the bounds  
– more homogeneous.

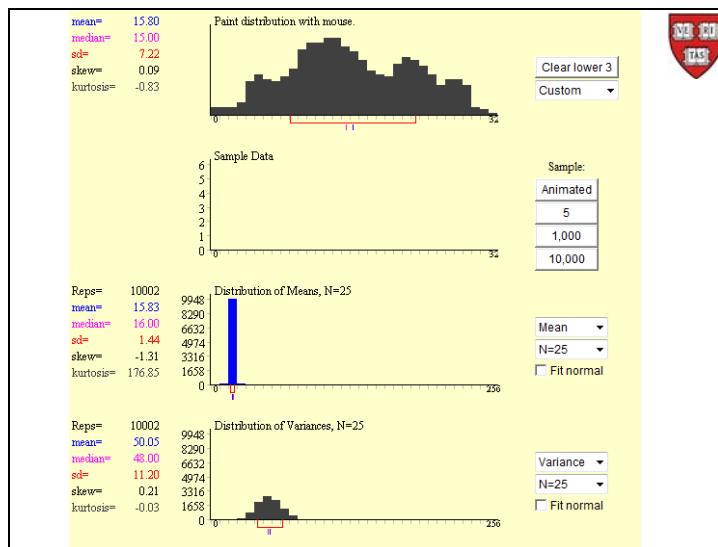
Here are three n's to show this decrease in width, chosen so that we see that it is the square root of n that determines the width and not n. So to see a decrease of one tenth, we need to go from n=10 to n=1,000.

We have no control over  $\sigma$ , it is a population parameter, but we see that the bounds are a function of sigma, reflecting what we have been saying all along: The more homogeneous our population, the more confident we are in whatever inference we make.

It is not quite right to say that we have no control over  $\sigma$ , as we see later in the course when we can stratify the population, we can look at different, and more homogeneous, sections of the population in turn and make our inferences in turn for each stratum. For example, instead of targeting the population as a whole, if there is a difference between men and women, and men are homogeneous, and the women homogeneous, then we might first make inference for the women and then for the men. That way we take advantage of the smaller, sex defined, standard deviations. More about this when we get to sampling later in the course.

## Unknown $\sigma$ – Student's t-distribution

To study the situation when  $\sigma$  is unknown let us return to the website at Rice, but this time let us also look at the bottom tableau.



Let's stick with our  $N$  of 25. And in this one, let's stick with the mean as we did before, and have fun drawing a population distribution. This one has a mean of 15.8 and a standard deviation of 7.22. In the bottom panel let us ask for the variance.

After two animated samples we can jump ahead and run 10,000 samples. Now, we have that the mean of the means is 15.83, very close to the population mean. Plus the standard error is 1.44 which is one fifth of 7.20 which is very close to the population standard deviation.

Now let us look at the bottom panel. The mean of the variances is 50.05. How does this relate to the population variance. The software gives the standard deviation as 7.22, thus the population variance is  $(7.22)^2 = 52.13$ . So it seems like the mean of the sample means is the same as the population variance, and indeed that is what theory shows. Indeed, this is the reason why we use  $(n-1)$  as the divisor in the definition of the sample variance; to obtain this property. This is what we call *unbiasedness*. We say that the sample variance is an unbiased estimator of the population variance—just like the sample mean is an unbiased estimator of the population mean.

A few weeks ago I told you that I would let you know why we divide by  $n$  minus 1, and not  $n$ . Had we divided by  $n$ , we would not have gotten an unbiased estimator of the population variance. Promise met.

What if  $\sigma$  is unknown?

Student's t

$$Z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$
 has  $n-1$  degrees of freedom

Sample: size  $n$   
sample mean  $\bar{X}$   
sample standard deviation  $s$

Population:  $X$  is approx. normal  
mean  $\mu$   
standard deviation  $\sigma$

  
William Sealy Gosset  
1876 – 1937

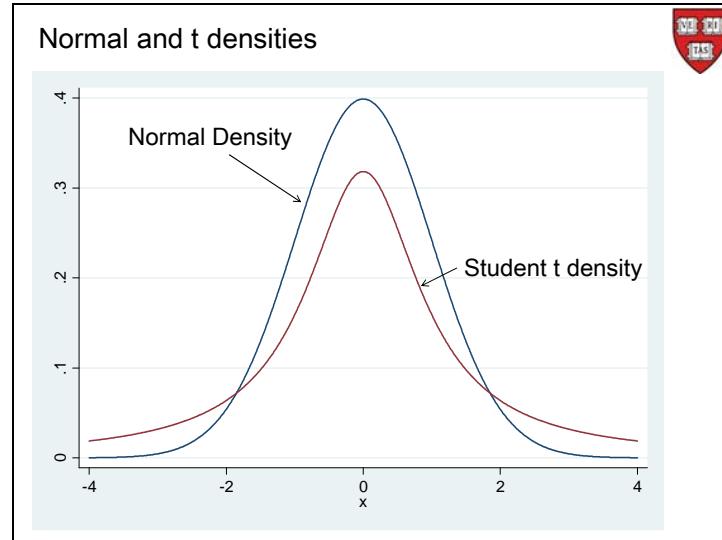
We make use of all this information we have gathered by looking at our standardized variate, Z, and if we do not know  $\sigma$ , we replace it with the sample standard deviation, s. To distinguish it from Z, we call this modified standardized variate t. This is sometimes called a Studentized variable after William Gosset who wrote under the nom de plume, Student. He did that to guard his work identity—he worked at the Guinness Brewery.

Student was the person who discovered the sampling distribution of this t. It is not as simple as Z in that we now require that the population distribution of X is at least approximately normal. Then the distribution of t varies with the size of the sample n. We call  $n-1$  the *degrees of freedom* of the t distribution.

n fixed Std.dev.

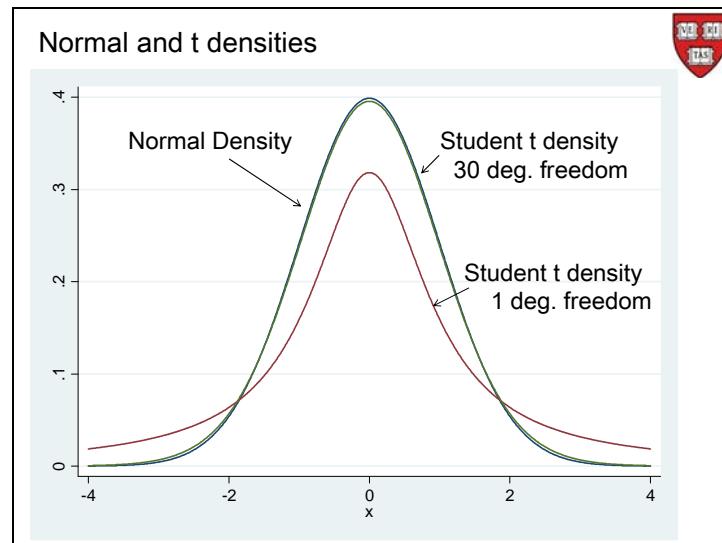
$\bar{x}_1$	$s_1$	$\frac{\bar{x}_1 - \mu}{s_1 / \sqrt{n}}$
$\bar{x}_2$	$s_2$	$\frac{\bar{x}_2 - \mu}{s_2 / \sqrt{n}}$
$\bar{x}_3$	$s_3$	$\frac{\bar{x}_3 - \mu}{s_3 / \sqrt{n}}$
$\vdots$	$\vdots$	$\vdots$
$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$		$\frac{\bar{X} - \mu}{S / \sqrt{n}}$

When we compare the distribution of the t to the Z we expect that there will be more variability in the former as the sample standard deviation varies from sample to sample, whereas in contrast  $\sigma$  in the definition of Z, remains constant.



Our intuition is borne out by the theory. Here is the normal density superimposed on a Student's t. You can see that the tails on the t are much wider than the normal, and thus the variance is larger. Also, in the same vein, the normal is much more peaked and tighter around the origin.

Now, what I've drawn here is about as bad as the Student t can get. This is when the Student t has one degree of freedom, which is also sometimes called the Cauchy distribution, and the tails are so fat that the mean does not even exist for this distribution, let alone the variance.



By the time you get to about 30 degrees of freedom, the normal and the t are almost indistinguishable. In this graph I have actually plotted three curves, the two from the previous graph plus a t-density with 30 degrees of freedom. As you see that there is very little difference between the student with 30 degrees of difference and the normal density.

In the old days, when we looked up values in tables, we would say, oh, if it has more than 30 degrees of freedom, just use the normal distribution. Stata does not do that, because it can calculate quantities exactly for you, but if stuck on a desert island without Stata and only a book of tables...

## Hypothesis Testing



### Hypotheticodeductive Method

Karl Popper's essence of the scientific method:

1. Set up falsifiable hypotheses
2. Test them

*Conjectures and refutations: the growth of scientific knowledge.*  
NY Routledge & Kegan Paul, 1963

So up to now we have made inference by either just estimating the population value in the case of the mean or the standard deviation, or by constructing confidence intervals for the mean. We could also construct confidence intervals for the standard deviation, but we will not in this course. Instead we are going to look at another popular method of making inference. It is called hypothesis testing.

Hypothesis testing forms part of the hypothetical deductive method that we have been using for almost 200 years now, to increase our knowledge in science. Karl Popper called it the essence of the scientific method:

First set up a falsifiable hypothesis. That differentiates science from other forms of knowledge: The hypotheses you set up must be falsifiable.

Second, design your experiment, get your data, and test your hypothesis.

And, as we saw, there is going to be uncertainty in our inference.



We know that total cholesterol levels in \*our\* Framingham population are distributed with mean  $\mu=237$  mg/100 ml and standard deviation  $\sigma = 44.7$  mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 230 \text{ mg/100 ml}?$$

Is it reasonable to assume that this is a sample from our population?

What if there were another possible explanation:  
The group was on a cholesterol lowering regimen?

Let us take a look at an example. We know that the total cholesterol levels in our population are distributed with mean  $\mu = 237$ , and  $\sigma = 44.7$ . We know that, but let us act as if we did not know it. Now, suppose somebody approaches us and says, I have a sample of 49 cholesterol levels, and their sample mean is 230 milligrams per 100 milliliters.

Is it likely that these 49 actually came from your Framingham population, or from a population with the same characteristics as your Framingham population? Is it possible that we could take a sample of 49 from our population and come up with a sample mean of 230?

Maybe what you would do is calculate that the standard deviation is 44.7, so the standard error is approximately 6 or  $6\frac{1}{2}$  ( $=44.7/7$ ). So 230 is approximately one standard error away from 237, and so the central limit theorem tells us that quite often we will be within plus or minus 1 standard error of the population mean. So it would be reasonable in this case to assume that this is a sample from our population.

What if instead of a sample mean of 230, your friend had a sample mean of 223? Is it likely that this sample came from our population? Well, now we are talking about a deviation which is approximately two standard errors from the mean. We might now hesitate.

What if instead your friend had come with a sample mean of 215? Now we are talking about three standard errors away from the population mean. Maybe now we start looking for other explanations of where this sample came from. Were these folks on some cholesterol lowering medication?

This, in general, is how we approach the testing of hypotheses.



Use of 95% confidence interval to infer value of  $\mu$  ( $\mu = 237?$ )

$$\left( \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \right) \rightarrow \left( \bar{X} \pm 1.96 \frac{47.7}{\sqrt{49}} \right) \rightarrow (\bar{X} \pm 13.4)$$

has a 95% chance of including  $\mu$ .

If $\bar{X}$	95% confidence interval
230	( 216.6, 243.4 )
223	( 209.6, 236.4 )
215	( 201.6, 228.4 )

We could possibly use the 95% confidence interval to infer values of  $\mu$  from the sample mean. The results of the calculations in the three instances are displayed above. The first confidence interval includes 237, the other two do not. In all instances we can say, I have followed a recipe that has a 95% chance of success. In the first case, I can say, the data do not refute the hypothesis that  $\mu = 237$ . Whereas in each of the next two I can say, either  $\mu = 237$  and something that has a 95% chance of success did not happen, or  $\mu \neq 237$ .



Alternatively,

IF

$\mu = 237$  and  $\sigma = 47.7$  and we take a sample of size  $n=49$  from this population, then the Central Limit Theorem tells us that the sample mean is approximately normally distributed with mean  $\mu = 237$  and std. dev.  $47.7/\sqrt{49}$ ; i.e.

$$\Pr \left\{ -1.96 \leq \frac{\bar{X} - 237}{47.7/\sqrt{49}} \leq 1.96 \right\} = 0.95$$

$$\Pr \left\{ 223.6 \leq \bar{X} \leq 250.4 \right\} = 0.95$$

The two approaches are consistent.

Alternatively, instead of confidence intervals we can construct a predictive interval. So in this case, we can say, when sampling from our population, a 95% predictive interval for the sample mean of samples of size 49, is (223.6, 250.4).

So in the first instance the sample mean is within these bounds, so our prediction worked and we can say we observed an event that had a 95% of happening, and thus is consonant with our hypothesized population. On the other hand in the second or third situation the 95% event did not happen, so either a rare event, one that has a 5% chance of happening, happened, or

maybe we should reject the hypothesis that this sample came from this hypothesized population.

The conclusions you come to in those two situations, the confidence interval or the predictive interval these two approaches are completely consistent. There are some people who would tell you otherwise, but do not believe them. They are not correct. These two approaches are completely consistent.

## Formalism of Hypothesis Testing

Let us look a little deeper into this second approach. It actually has a lot in common with our legal system as it pertains to criminal cases.



Individual on trial. Did he commit the crime?		
Evidence	Trial	
Jury	Person	
	Innocent	Guilty
Not Guilty	✓	✗
Guilty	✗	✓

We start with an individual on trial to answer the question of whether he committed the crime.

The evidence is presented at trial. The person truly is innocent and did not commit the crime or the person is guilty. The jury needs to decide whether the person is or is not guilty. Person is assumed innocent. So the jury just decides whether not guilty or guilty.

If the jury decides the person is not guilty when in fact the person is innocent, then the jury is doing the right thing. If on the other hand, the jury decides the person is guilty when the person in fact, did commit the crime, then once again, the jury is doing the right thing. The jury is only going to decide guilty or not guilty. So potentially it could make one of those two decisions. But in reality, after the trial, only one of these decisions is made, and thus only one row is relevant.

If on the other hand we're in one of the off diagonal situations, then the jury has made a mistake. So finding a guilty person not guilty or finding an innocent person guilty-- those are the mistakes that juries can make. Of course, at most a single jury is only going to possibly make one of those mistakes. But potential is there for either of those mistakes to be made, in general.



Test of Hypothesis that  $\mu = \mu_0$ ?

Sample      Analysis

Us	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject	✓	Type II
Reject	Type I	✓

In hypothesis testing, rather than decide whether a person committed a crime, we decide whether the mean of the population is equal to  $\mu_0$ . For example, is  $\mu = 237$ ?

To help us decide we take a sample (evidence) and we replace the trial with our analysis of the sample, and we replace the jury, and we have to make a decision about the population. We are faced with a hypothesis that the mean of the population is a given amount. And we are going to decide to reject our hypothesis or not on the basis of a sample from that population.

Now let us look at these possible errors. We label the first one Type 1 and the second one Type 2. Returning to the analogy with the criminal case, the Romans used to say better that 10 guilty men go free than that one innocent person be found guilty. So they had this 10 to 1 ratio of these probabilities of these types of errors. Today we have diluted that a bit, possibly because we value life a little less, but we now say that better that eight guilty men go free than that one innocent person be found guilty.

In hypothesis testing in statistics, we label these probabilities as alpha or beta. So the probability of a Type 1 error is alpha. And the probability of a Type 2 area is beta.



Probability of Type I error is  $\alpha$   
i.e. the probability of rejecting the null hypothesis when it is true.

Probability of Type II error is  $\beta$   
i.e the probability of **not** rejecting the null hypothesis when it is false.

$1-\beta$  is the *power* of the test.

So the probability of rejecting the null hypothesis when it is true is called alpha. And the probability of a Type 2 error, which is not rejecting the null hypothesis when we should be rejecting it, is beta. So in both instances, we'd like to make these as small as possible.

This is very similar, of course, to what we did with diagnostic testing. Except with diagnostic testing, we took a more positive view on life. And we spoke about sensitivity and specificity, the probabilities of doing the right thing, whereas here we label the probabilities of doing the wrong things. Ideally, we wanted both the sensitivity and the specificity to be one.

Here in hypothesis testing, we of course want to minimize our chances of making mistakes so ideally we would want both alpha and beta to be zero.

In terms of beta, sometimes we look at 1 minus beta. And that is called the *power* of the test. We delve more into the power issue, later.



Recap: Hypothesis testing about  $\mu$  :

1<sup>o</sup> Hypothesize a value ( $\mu_0$ )

2<sup>o</sup> Take a random sample (n).

Is it **likely** that the sample  
3<sup>o</sup> came from a population with  
mean  $\mu_0$  ( $\alpha = 0.05$ ) ?

So to recap, we looked at hypothesis testing about  $\mu$ . The first thing we did was we hypothesized a particular value for  $\mu$  in the population. For example, this may represent the status quo. We call that  $\mu_0$ , naught for the null hypothesis.

Then we take a random sample of size  $n$  from that population, and ask the question, is it likely that the sample came from a population with this hypothesized mean,  $\mu_0$ .

We do not want to make it unlikely that we reject this hypothesis, when in fact, it is not true. We can set it up so we do not reject it 95% of the time, say, when it is true. Or, another way of saying the same thing, we only want to run a risk of  $\alpha = 0.05$ , that is a 5% chance of making this mistake, when we should not.



Decide on statistic:  $\bar{X}$

Determine which values of  $\bar{X}$  are consonant with the hypothesis that  $\mu = \mu_0$  and which ones are not.

Look at  $\frac{\bar{X} - \mu_0}{\sigma}$  and decide.

One sided or two?

So we first decide on the statistic to use. Since we are talking about the population mean  $\mu$ , the statistic that we might decide on is the sample mean,  $X$ -bar. We saw, from the central limit theorem, that the mean of the sampling distribution of the sample mean is the population mean.

This allows us to determine which values of  $X$ -bar are consonant with the null-hypothesis and which are less likely. A predictive interval would be one way of doing that. Another way of determining it is to look at the value of the standardized variable. If it is too large, then reject the null hypothesis.



Need to set up 2 hypotheses to cover *all* possibilities for  $\mu$ .

Choose one of three possibilities:

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

Now we need to be careful deciding whether the standardized variable is too “large”. To determine how to operationalize this we need to look at three possibilities when looking at the pair—namely, the null and alternative hypotheses.

If we fall into the category we call “two sided” as shown in the slide above, then a value outside our usual (-1.96, 1.96) interval would only happen 5% of the time if our null hypothesis is true. There is no directionality implied in our null hypothesis.

On the other hand, if we have the first set of “one sided” hypotheses, as labeled above, then a standardized variable that is positive would be consonant with the null hypothesis. So it could not be “too large” in the positive direction, and we would seek too large in the negative direction to reject the null hypothesis. The fact that  $\alpha=0.05$ , allows us to reject the null hypothesis even if we have a small negative value for the standardized statistic.

In contrast to the previous situation, if we have the third set of hypotheses, above, then we would seek too large a positive value of the standardized statistic to reject the null hypothesis.

This is the issue of whether we have one-sided or two-sided hypotheses. The decision of where we are in this trichotomy should not be influenced by the data. Ideally, we should make our decision before even looking at the data at all. The decision should be made on scientific considerations. The theory we have presented here does not support decisions of sidedness based on having seen the data.



Look at

$$H_0: \mu = \mu_0$$

$$H_A: \mu \neq \mu_0$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

and reject  $H_0$  if  $Z$  is too large, + or -, e.g.

Reject if  $Z$  is  $>1.96$  or  $<-1.96$ , then

$$\Pr(\text{reject } H_0 \text{ when true}) = 0.05$$

Let us take a look at a couple of applications of hypothesis testing. Suppose we want to test in the two-sided framework, that  $\mu = \mu_0$ . Then we set up our  $Z$ , or  $t$  if we do not know  $\sigma$ . We reject the null hypothesis if  $Z$  is too large. That means if  $|Z| > 1.96$ . That, of course presupposes that our  $\alpha = 0.05$ .

$$H_0: \mu = 237$$

$$H_A: \mu \neq 237$$

$$\sigma = 47.7 \text{ mg/100ml}$$

Sample of 49 non-hypertensives have:

$$\bar{x} = 221.9 \text{ mg/100ml}$$

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

So reject the null hypothesis.

So for example, if we go back to our population, and we stated that the total cholesterol level in our population in Framingham at visit one was 237 with a standard deviation of 47.7.

Now, I took a sample of 49 non-hypertensives from this population. If I'm not mistaken, I used the random seed from the bill like before, and I got a sample mean of 221.9. Now the question is, would you believe that this was a sample from the general population; namely, a population with a mean of 237, and a standard deviation of 47.7?

We calculate the standardized variable: the observed sample mean was 221.9, the population mean was 237, the standard deviation was 47.7, and the sample size was 49. Thus  $Z$  is -2.37 which is less than minus 1.96, and so we reject the null hypothesis.

```

. set seed 725764662

. drop if hyperten==1
(3252 observations deleted)

. sample 49 , count
(1133 observations deleted)

. mean totchol1

Mean estimation                               Number of obs = 49

+-----+
|      Mean   Std. Err. [95% Conf. Interval]
+-----+
| totchol1 | 221.8776    4.614348    212.5998    231.1553
+-----+

. di (221.8776-237) / (44.7/7)
-2.3681611

```

So on the basis of these 49 non-hypertensives, we reject the hypothesis that they come from a population with this high a cholesterol level.

What you do with the information is up to you, but the formalism of the hypothesis testing makes us reject the null hypothesis.

### P-value

Some prefer to quote the [p-value](#). The p-value answers the question, “What is the probability of getting as large, or larger, a discrepancy?” ( $\mu - \bar{X}$ )

$$\begin{aligned} \Pr(Z > 2.37 \text{ or } Z < -2.37) &= 2\Pr(Z > 2.37) \\ &= 2 \times 0.0222 \\ &= 0.044 \end{aligned}$$

---

Stata:      . di normal(-2.0106348)  
              .02218202

Rather than, looking to see whether the Z is bigger than 1.96 or smaller than -1.96 to reject the null hypothesis, some people prefer quoting what is called the p-value. The p-value answers the question, what is the probability of observing as large or larger a discrepancy than the one I observed? In this case, after we standardize things, our standardized value was 2.37. So the p-value will answer the question of what is the probability of getting a deviation as large or larger than 2.37 after we standardize? Because this is a two-sided calculation, that means we want to know the probability that our standard, Z, is bigger than 2.37 or less than minus 2.37. And the answer is p = 0.044.

Some then argue that since the p-value is less than 0.05, we should reject the null hypothesis.

It is up to you how you set these critical points and to defend your choice. The choice of 0.05 is quite ubiquitous, and we see it everywhere in science.



Blood glucose level of healthy persons has  $\mu = 9.7 \text{ mmol/L}$  and  $\sigma = 2.0 \text{ mmol/L}$

$H_0 : \mu \leq 9.7$
$H_A : \mu > 9.7$

Sample of 64 diabetics yields  
 $\bar{x} = 13.1 \text{ mmol/L}$   
 $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = 13.60$   
p-value << 0.001

Here is another example. This time we are looking at blood glucose levels of healthy persons. If we look at a sample of 64 diabetics, then we are in a one-sided hypothesis. If the mean is less than or equal to 9.7, we will be perfectly happy, but we are concerned about whether it's bigger than 9.7.

In this case, of course, it doesn't make any difference, because the Z value is 13.6 and this would be rejected as either a one- or two-sided test statistic. So whether we did a one-sided or two-sided, test really makes no difference.


$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$\alpha = 0.05$

$H_0 : \mu = \mu_0$	Reject if $ z  > 1.96$
$H_A : \mu \neq \mu_0$	
$H_0 : \mu \geq \mu_0$	Reject if $z < -1.645$
$H_A : \mu < \mu_0$	
$H_0 : \mu \leq \mu_0$	Reject if $z > 1.645$
$H_A : \mu > \mu_0$	

The issue about one-sided or two-sided tests is quite controversial in the field. We see that  $Z$  has to be larger to reject a two sided hypothesis than a one-sided one. Since some journals would prefer to publish “significant” results than non-significant—man bites dog, rather than dog bites man, sort of thing—and since publications are desirable for promotions and advancement in general, if an author tries to publish a one-sided hypothesis test, he or she runs the risk of being accused of all sorts of dastardly deeds, and editors will have none of it. This is not just an academic controversy, as some in the pharmaceutical industry have argued that the FDA should consider one-sided drug testing.

We should also point out that a similar development with one and two sided confidence intervals can be carried out, so that does not overcome the objections.

Marcello Pagano

## [JOTTER 6 TWO-MEANS]

The testing of one mean is extended to two or more samples (ANOVA) with a fuller description of power.



## One sample hypothesis testing about the mean

### 1. Set up the null hypothesis

$$(H_0: \mu = \mu_0)$$

### 2. Set up the alternative

$$(H_A: \mu \neq \mu_0)$$

### 3. Choose $\alpha$ -level

$$(\alpha = 0.05)$$

### 4. Take a sample and calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Let us quickly review the steps we take to test an hypothesis with a single sample: Step one, we set up the null hypothesis. For example, we can make a statement about the mean of the population; that it is equal to some  $\mu_0$ . Step two, set up the complement of the null hypothesis. That is the alternative hypothesis; here we have the two-sided case. Step three, choose an alpha level; choose the ubiquitous 0.05.

Step four is to take a sample of size  $n$  and then calculate either  $Z$  or  $t$ , depending on whether we know  $\sigma$ , or not. In either case, what we are comparing the sample mean to the hypothesized population mean to see how far apart they are. If they are very close to each other, then we shall say that the data is consonant with the hypothesis. If they are far apart, then we argue that the data do not seem to be supporting the null hypothesis and so we reject the null hypothesis. The remaining point is how to judge what is small and what is big?

To this end, we divide by the appropriate standard deviation—we standardize;  $Z$  if we know  $\sigma$ ,  $t$  otherwise. The  $Z$ , or the  $t$ , can now be compared to what we expect to see if the null hypothesis is true.



5. Calculate appropriate p-value.

6. Compare p-value to  $\alpha$ .

7. Either reject  $H_0$ , or not.

Alternatively,

5. Find cutoff ( $\pm 1.96$ )

6. Compare z or t to cutoff

7. Either reject  $H_0$ , or not.

We then compare these values to their appropriate cut-offs. So, for example, we can calculate the appropriate p-value for our statistic, compare that p-value to  $\alpha$ , and either reject  $H_0$  or not. If the p-value is less than  $\alpha$ , we shall reject; if not, we shall not reject  $H_0$ .

Alternatively, we can look directly for the cut-off of the statistic. For example, if we were looking at Z our cut-off would be plus or minus 1.96. So reject  $H_0$  if the magnitude of Z is bigger than 1.96; do not reject if it is less than 1.96 in magnitude. In the case of t, find the appropriate cut-off for the given degrees of freedom, and then proceed as with Z. In either case, we are going to come up with the decision to either reject the null hypothesis, or not.



### Comparative Situations – Two samples

- Before and After
- Treatment and Control
- Two groups

$$H_0 : \mu_1 - \mu_2 = \Delta$$

Now let us move on and generalize the situation to the case when we have two samples, and not just one. For example, if we have a before and after situation—weighing in before going on a diet and after going on a diet to judge whether the diet is any good.

Alternatively, we might have two groups of individuals: one group gets an experimental treatment and the other group, the control group, gets the standard treatment. Once again, it could be the same persons in the groups—for one week you give the experimental treatment and then you let that wear off, and then the next week you give that person a control treatment, for example.

Or you might have two groups of patients, 50 patients, get the treatment. Another, separate group of 50 patients, get the control. And that is a treatment and control situation.

Our null hypothesis here is going to deal with the difference in the two means, let us call it delta, and we are going to hypothesize something about delta. Now the most common hypothesis is that delta is equal to 0; that is that there is no difference, or no effect.

This is the generic way to set up the two sample problem.


$$H_0 : \mu_1 - \mu_2 = \Delta \quad ?$$
$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\text{standard deviation}}$$

Question: Are two samples independent?

1. No (dependent, before/after)
2. Yes (different people)

As in the one-sample situation we need to determine our statistic. The thinking is very similar to the one-sample situation in that we look at the difference between the two sample means as our basic statistic. Also, as before, we compare this difference to the hypothesized difference,  $\Delta$ , and then all that remains is to evaluate the size of this difference by dividing by an appropriate standard deviation.

The whole trick is going to be what we put in there for the standard deviation. As far as current theory has it, we need to classify our situation into one of two, depending on whether the two samples are independent, or not.

If they are not, for example, if we have a before and after situation on the same individual, or we are looking at the right eye of a rat and comparing it to the left eye of the same rat, then they are not independent. Then we perform one set of calculations. On the other hand, if the samples are independent—so we are talking about different people in the two groups—for example, we may be doing a male versus female, or possibly a young versus old comparison, or maybe we break up the population into two groups, one group gets one treatment and the other group gets the control treatment—then we perform a different set of calculations.

```
. gen totchol = totchol2 - totchol1  
(675 missing values generated)
```

```
. summ totchol
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totchol	3759	13.16574	33.30605	-159	321

Let us first look at the dependent situation, and then we deal with the independent situation.

Suppose we are interested in what happens to the total cholesterol level between the first and second visit. We like the idea of each person serving as their own control, so we plan to measure each person twice, once at each visit.

To analyze these data, generate a new variable, call it *totchol*, defined to be the difference between a person's cholesterol level at visit 2 and at visit 1. Once we do this, Stata returns the message that there are 675 missing values generated. This should be of concern to us, if this was a scientific enquiry, because what this is telling us is that there are 675 individuals who do not have both cholesterol readings at time 2 and at time 1. That might very well impact any sort of inference we want to make about the change in cholesterol level because these 675 who are missing might have a very different story to tell from the ones who are not missing.

Let us move on and not worry about this right now. Let us summarize *totchol* to find its mean is 13, there are 3,759 observations, and the standard deviation is 33, and not everybody shows an increase, so there are some negative numbers, but on average there is an increase in total cholesterol over the two visits.

```

. set seed 725764662

. sample 49 , count
(4385 observations deleted)

. mean totchol

Mean estimation                               Number of obs = 36


```

	Mean	Std. Err.	[95% Conf. Interval]
totchol	13.41667	4.545602	4.188603 22.64473

Now let us take a sample of size 49 from our population. Now when we ask for the mean of this variable we find that the sample consists of only 36 of the 49 observations with *totchol* defined.<sup>1</sup>

```

. ttest totchol == 0

One-sample t test


```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
totchol	36	13.41667	4.545602	27.27361	4.188603 22.64473

```

mean = mean(totchol)                                t = 2.9516
Ho: mean = 0                                         degrees of freedom = 35
Ha: mean < 0           Ha: mean != 0           Ha: mean > 0
Pr(T < t) = 0.9972       Pr(|T| > |t|) = 0.0056      Pr(T > t) = 0.0028

```

Now we can test the null hypothesis that the difference in the means is zero. We see that the 95% confidence interval for the difference in cholesterol levels means between visit 2 and visit 1, is (4.2, 22.6). This does not include 0. So we know that the confidence interval approach to testing the null hypothesis, that the difference is 0, would reject that null hypothesis at the 5% level.

Now let's look at the hypothesis testing approach. The t statistic is 2.95 and there are 35 degrees of freedom, and if the alternative is that the mean is not 0, or the two-sided alternative, has a p-value attached to it of 0.0056. So we would reject the null hypothesis at the 5% level.

---

<sup>1</sup> This is different from the video because I did not set the seed in the video.

1 Dependent



$$\begin{aligned} d_1 &= x_{12} - x_{11} \\ d_2 &= x_{22} - x_{21} \quad \rangle \Delta = \mu_2 - \mu_1 \\ &\vdots \\ d_n &= x_{n2} - x_{n1} \end{aligned}$$

So a test about the difference in the means is a test about the mean of the differences.

$$H_0 : \Delta = ?$$

So in summary, in the dependent case look at each individual and take the difference between the value at time 2 and the value at time 1. Then ignore the individual x's and concentrate on the individual differences.

We treat these differences just like we treated a single sample before, and call the mean of these differences,  $\Delta$ ; the difference between the two means.

So now we are back in the single sample situation with  $n$  observations, based on the  $n$  differences, and we can set up hypotheses about  $\Delta$ —the most common hypothesis being  $H_0: \Delta=0$ .

1 Dependent



So treat the  $d$ 's as the data and perform a one-sample t-test:

$$\begin{aligned} \bar{d} &= \frac{1}{n} \sum_{j=1}^n d_j \\ s^2 &= \frac{1}{n-1} \sum_{j=1}^n (d_j - \bar{d})^2 \\ t &= \frac{\bar{d} - \Delta}{s / \sqrt{n}} \quad (n-1) \text{ d.f.} \end{aligned}$$

So we proceed exactly as before: Calculate the sample mean of the  $d$ 's; calculate the sample variance of the  $d$ 's; and then take the mean of the  $d$ 's, and divide by the standard error.

Insert the hypothesized value of  $\Delta$  and under the null hypothesis, this statistic is distributed as a t with  $(n-1)$  degrees of freedom. And this is exactly what Stata reports.

2 Independent



$$H_0 : \mu_1 - \mu_2 = \Delta \quad ?$$

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \Delta}{\text{standard deviation}}$$

OK. So now we've got the dependent situation under our belts. Let's look at the independent situation. So we have two independent samples, and the hypothesis that we want to test is exactly the same as in the dependent case, namely a statement about the value of the difference between the two population-means; call it  $\Delta$ .

And once again, the most common hypothesis is  $H_0: \Delta=0$ . So there is no difference in these two groups. There is no difference between males and females. So this is our t just like before, the difference though is to be how we calculate the standard deviation.

2 Independent



Population (Normal)

Pop. 1

Pop. 2

$\mu_1$

$\mu_2$

$\sigma_1$

$\sigma_2$

Sample

$n_1$

$n_2$

$\bar{x}_1$

$\bar{x}_2$

$s_1$

$s_2$

$$H_0 : \mu_1 - \mu_2 = \Delta$$

Let us establish some notation. We have two populations, and let us distinguish them by use of the subscripts 1 and 2. We take a sample of size  $n_1$  from the first population and of size  $n_2$  from the second population. Whereas, in the dependent case, these two sizes had to be the same, in this, independent case, they do not.

2 Independent



$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In order to decide the  $s^2$ s and  
the degrees of freedom we need  
to know whether, or not,  $\sigma_1 = \sigma_2$

One possible  $t$  that we could think of is to take the variance of the first sample divided by  $n_1$  then the second variance and divide it by  $n_2$ , and take their sum and use that as our variance.

Unfortunately we are not done. Things get a little more complicated. We need to ask a second question. That second question is going to depend very much on what we know about the relative sizes of the two standard deviations—I am assuming we do not know the actual values of the two sigmas. What we need to know is if they are equal or not.

If the two population standard deviations are equal, then we call that the *homoscedastic* case; if not, we call that the *heteroscedastic* case. We can actually use our sample standard deviations to carry out a preliminary test to decide which of the two cases we are in, but that is not recommended.

## 2 Independent



### (a) Homoscedastic Case

If  $\sigma_1 = \sigma_2$  (which can be tested)  
we can use a common value:

$$s^2 = s_1^2 = s_2^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\text{d.f.} = n_1 + n_2 - 2$$

In the homoscedastic case, we create a single estimator of the common variance, displayed above. Plug that into the t-statistic and it is distributed with  $n_1 + n_2 - 2$  degrees of freedom.

Now here we see what price we paid when we had the dependent case. In the dependent case we had  $n_1 - 1$  degrees of freedom, whereas here we have twice as many (assuming equal sample sizes). The degrees of freedom are related to the sample size(s), so the more degrees of freedom we have, the better it is.

## 2 Independent



### (b) Heteroscedastic Case

If  $\sigma_1 \neq \sigma_2$  (recommended)

Use individual sample standard deviations  
and degrees of freedom,  $\nu$ :

$$a = \frac{s_1^2}{n_1} \quad \text{and} \quad b = \frac{s_2^2}{n_2}$$

$$\nu = \frac{(a+b)^2}{\frac{a^2}{(n_1 - 1)} + \frac{b^2}{(n_2 - 1)}}$$

Now let us look at the heteroscedastic case. This is the way I would recommend you proceed in general; namely, act as if you are in the heteroscedastic case. You lose a few degrees of freedom, but for those you buy a little protection from an assumption (equality of the standard deviations) you do not have to make, and even if true, should not cause problems.

The complications with the heteroscedastic case are: (i) the degrees of freedom are a little bit more complex (for Stata) to calculate; and, (ii) they are not exact, but provide an approximation. The one you see above is one approximation; Stata provides us with a choice.

```

. tab hyperten , summ(totchol1)

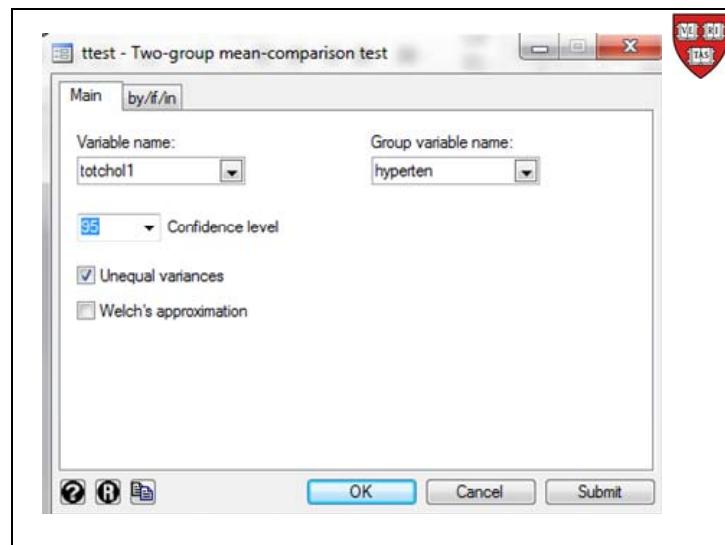
      Incident   Summary of Total cholesterol
Hypertension | (mg/dL), exam 1
              n          Mean     Std. Dev.    Freq.
              No        227.98031  42.644789  1168
              Yes       240.25638  44.919635  3214
              Total      236.98425  44.651098  4382

. set seed 72576466

. sample 49, count by(hyperten)
(4336 observations deleted)

```

So let us go to Stata and see how easy it is to actually do these calculations. First, let us set the problem up: If we split the population into two groups, those who are not hypertensive at the beginning (*hyperten* is no,  $N_2 = 1168$ ) and those who are ( $N_1 = 3214$ ), and for whom we have total cholesterol level readings at visit 1. For those who are not hypertensive coming in, their mean total cholesterol is about 228, and those who did have hypertension coming in, their cholesterol level is about 240. So indeed there is a small difference in total cholesterol between these two groups. Let us proceed to take samples, of size 49, from each of these groups and see if we can detect this difference based on inference from our samples.



Now we go to Statistics > Summary tables > Classical tests of hypotheses, and what we want is the > Two-group mean comparison test. So there it is.

The variable that we want to test is going to be our total cholesterol level at time 1, so it's *totchol1*. And the Group variable name is *hyperten*. As mentioned above, I recommend you choose the "Unequal variances" option.

We could check the Welch approximation box and get a different approximation, but let us first leave it at the default and get what is called the Satterthwaite approximation. I leave it up to you to check the Manual to see investigate the details about the differences between these two, if you are interested. By all means rerun your analysis using the one you did not previously use and see if you get different results. They usually give very similar answers.



. ttest totchol1, by(hyperten) unequal						
Two-sample t test with unequal variances						
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
No	48	220.7083	5.953868	41.24961	208.7307	232.686
Yes	48	236.0833	6.501557	45.04411	223.0039	249.1628
combined	96	228.3958	4.455026	43.65016	219.5515	237.2402
diff		-15.375	8.815826		-32.88079	2.130787
		diff = mean(No) - mean(Yes)			t = -1.7440	
		Ho: diff = 0			Satterthwaite's degrees of freedom = 93.2813	
		Ha: diff < 0			Ha: diff != 0	Ha: diff > 0
		Pr(T < t) = 0.0422			Pr( T  >  t ) = 0.0844	Pr(T > t) = 0.9578

The first thing we notice on the output is that we lost two observations, one from each sample, because of missing data. So of the 48 who do not have hypertension, their mean is 220. And for the 48 with hypertension, their mean was 236. These means are lower than in the population, but the standard deviations are pretty close to what they are in the population.

The differences are reported in the "diff" row. The difference in the sample means is  $220.7083 - 236.0833 = -15.375$ . Those differences have a reported standard error of 8.81. And if we want to use the confidence interval approach to test the hypothesis of any differences, we would find that the 95% confidence interval is (-32.88, 2.13) which includes the value 0. So by using the confidence interval approach, we would not reject the null hypothesis that the two means are the same.

If we perform a hypothesis test we see at the bottom that the p-value associated with a two sided test is 0.0844, and so we would not reject the null hypothesis of equality of the two group means.

Explore what happens with the Welch approximation and also explore what happens had you made the homoscedastic assumption.

## Type 2 Error – Power



One sample hypothesis testing about the mean

1. Set up null hypothesis

$$(H_0: \mu = \mu_0)$$

2. Set up the alternative

$$(H_A: \mu \neq \mu_0)$$

3. Choose  $\alpha$ -level

$$(\alpha = 0.05)$$

4. Take a sample and calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Here is an outline of what we have done up to now to test a hypothesis about the mean of a population. Here are the first four steps we take.



5. Calculate appropriate p-value.

6. Compare p-value to  $\alpha$ .

7. Either reject  $H_0$ , or not.

Alternatively,

5. Find cutoff ( $\pm 1.96$ )

6. Compare z or t to cutoff

7. Either reject  $H_0$ , or not.

Then we take the fifth step, sixth, and seventh steps whichever way we go, whether we calculate the appropriate p-value or whether we do it by using our cutoffs for the appropriate alpha.



$\Pr(\text{rejecting } H_0 \mid H_0 \text{ is correct}) = \alpha$

$\Pr(\text{not rejecting } H_0 \mid H_A \text{ is correct}) = \beta$

$\Pr(\text{rejecting } H_0 \mid H_A \text{ is correct}) = \text{Power}$

Remember: Power =  $1 - \beta$

What we have not paid attention to up to now is what happens to our power, or the type II error. We have set the alpha at 0.05, and not mentioned beta. So what can be said about the probability of not rejecting  $H_0$  when we should be rejecting it? Or, if we like to be positive in life, we can talk about 1 minus beta, namely the power.



- Omnidox, broad spectrum antibiotic, recalled < 6 mos, severe reactions; 3 deaths
- Versed, a sedative, < 18 mos, 86 adverse reactions including 46 deaths.
- Fenoterol, many years, relieve asthma attacks, increased the risk of death.
- Oraflex, antiinflammatory (arthritis) < 3 mos, 72 deaths (USA & UK.)

Is this something we should be concerned about in real life? The very dramatic example of when this happens is in clinical trials, which usually form the knowledge base for the FDA to decide whether certain drugs should be allowed out onto the market.

An important aspect of these clinical trials is to study the drugs' side effects. Here are some drugs that were deemed safe enough to be allowed out onto the market, but were subsequently

recalled because of their safety profiles. Possibly the null hypothesis of not having serious drug effects was not rejected.

The first one, Omnitrox, was recalled less than six months after it had been released. There were all sorts of severe reactions associated with it—it was a broad spectrum antibiotic—including three deaths. Versed, was out for about 18 months, and had 46 deaths associated with it. And so on. So the question is, can we avoid this? It seems like we cannot, even though the FDA has a terrific track record.

It seems like risks are inevitable, as we cannot cover all possibilities. For example, suppose a negative effect takes six years to manifest itself. If the clinical trial lasts five years, you are not going to see it. This is an argument for post-marketing monitoring to keep an eye on what is going on.

*Associated Press* news researcher Rhonda Shafner:



- 2010: Mylotarg -- Risks: Liver disease
- 2009: Raptiva -- Risks: A rare brain infection
- 2007: Zelnorm -- Risks: Increased risk of heart problems
- 2007: Permax -- Risks: Heart valve damage
- 2005: Cylert -- Risks: Liver problems, including death
- 2005: Bextra -- Risks: May increase the risk of heart attacks and strokes; also may cause rare but serious skin conditions
- 2005: Tysabri -- Risks: Rare, but life-threatening side effect (Drug returned to market in 2006 under a restricted distribution.)
- 2004: Vioxx -- Risks: Heart attacks, strokes
- 2001: Baycol -- Risks: Severe damage to muscle, sometimes fatal

2

---

<sup>2</sup> <http://www.drug-injury.com/druginjurycom/2010/07/unsafe-drug-recall-decision-determination-factors-medicines-withdrawn-us-fda-history.html>

2000: Lotronex -- Risks: Intestinal damage from reduced blood flow

2000: Propulsid -- Risks: Fatal heart rhythm abnormalities

2000: Rezulin -- Risks: Severe liver toxicity

1999: Hismanal -- Risks: With other drugs or high dose can cause fatal heart rhythm

1999: Raxar -- Risks: Fatal heart rhythm abnormalities

1998: Posicor -- Risks: Dangerous interaction with other drugs

1998: Duract -- Risks: Severe liver damage

1998: Seldane -- Risks: Fatal heart rhythm abnormalities

1997: Pondimin -- Risks: Heart valve abnormalities

1997: Redux -- Risks: Heart valve abnormalities

These recalled drugs are historical recalls, but the problem persists. Here is a list of more recent recalls compiled by Rhonda Shafner of the Associated Press, that covers the last fourteen years or so.

Lest we go away with the wrong impression, the opposite can happen too, namely a good drug taken off the market for the wrong reasons.

### Bendectin Story



- Hyperemesis gravidarum—morning sickness
- 1956 Bendectin introduced (FDA approved) (known as Debendox in the UK and Diclectin in Canada) is a mixture of pyridoxine (Vitamin B-6), and doxylamine.
- 1979 *National Enquirer* attributes “hideous birth defects” to bendectin. ‘Experts’ compare it to thalidomide.
- 1983 After millions of dollars in litigation costs for alleged birth defects (including *Daubert v. Merrell Dow Pharmaceuticals* (1993)) Bendectin removed from market.

Let me tell you the Bendectin story. Bendectin was a drug that was introduced in 1956 to combat morning sickness (NVP—nausea and vomiting of pregnancy). So for pregnant women suffering from morning sickness, they had this drug, Bendectin. In the UK and Canada it has a different name. It is available all across the world except in the US. And when we ask, why, one

discovers that in 1979 the National Enquirer published an article about Bendectin. In the article they had phrases like "hideous birth defects" are associated with Bendectin. They had statements by "experts" that compared Bendectin to thalidomide. And back in those days, the word thalidomide was—it still is—a very scary thought for a pregnant woman. But back in those days, it raised all sorts of horrible mental pictures.

For those of you who do not know what the National Enquirer is, I do not know how to explain it. It is the sort of publication you see when you are waiting in line in the supermarket to check out your purchases and you see those—I do not want to use the word rag but—kinds of newspapers that you really should not be spending any money buying but you read while you are waiting in line for a good laugh.

## Aftermath

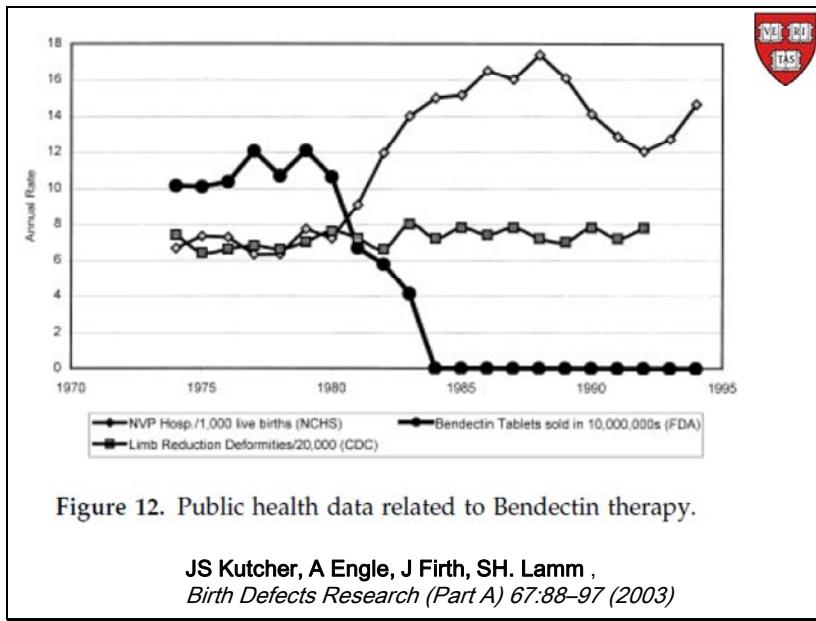


- 30 epidemiological studies, WHO, FDA & March of dimes, concur that Bendectin is safe
- Not one court case lost
- Since 1983, CDC
  - no significant decrease in incidence of birth defect
  - hospitalizations for hyperemesis gravidarum has doubled

But what happened subsequent to that story, by 1983 after millions of dollars in litigation costs for alleged birth defects, Bendectin was removed from the market by the company that manufactured it. Now it is sad because at that time that was the only drug that had been approved by the FDA to handle morning sickness. In its defense, there were 30 large epidemiological studies done all over the world by very reputable organizations such as the WHO, the World Health Organization, the FDA, the March of Dimes.

They all came to the same conclusion, namely that Bendectin is safe.

So you had on the one side the National Enquirer claiming something, and then you had the preponderance of scientific evidence and the scientific community saying that what was published was nonsense. Indeed, not one court case was lost, but the manufacturer did not wish to take any further risks. So sometimes it is not science that dictates on scientific issues.



Some twenty years after the withdrawal of Bendectin from the market, still with no replacement drug, a study was done to shed some light on the issue. It is an ecological study, namely done at the country level and not at the individual level, but it is interesting nonetheless. The argument follows along the line: if you have a period when Bendectin is on the market and it's supposed to be doing some harm, then when you take it off the market the frequency of that harm should decrease. Look at the above graph of three lines.

The solid black line that starts at the top on the left and ends at the bottom on the right represents Bendectin sales in the US. It drops precipitously in the early 80s to reflect the fact that it was taken off the market. Prior to being taken off the market, the sales had been substantial.

Now remember that the drug was being blamed for causing birth defects. The middle line is the number of birth defects in the country—about as level a line as you can imagine. Taking the drug off the market does not seem to have had any impact on this line.

What about the poor pregnant women suffering from morning sickness? Focus on the third line (NVP) in the graph, the one that almost perfectly mirrors the Bendectin sales line. The line represents hospitalizations for the side effects of morning sickness, standardized by the number of births.

Clearly, this is an ecological study so all sorts of other factors may be influencing what is going on in this graph, but we are not just observing it is not a single relationship, that could be easily explained away, we are seeing two patterns: (i) it seems that taking Bendectin off the market was not followed by any decrease in birth defects, and (ii) the number of hospitalizations, for precisely the effects Bendectin was supposed to alleviate, increased after Bendectin was taken off the market.



Melvin Belli  
1907-1996

Bendectin and Birth Defects  
The Challenges of Mass Toxic Substances Litigation  
(1997) University of Pennsylvania Press p. 106, 124

Michael D. Green

The other question is why did the National Enquirer do this in the first place? Why did they publish the article? Well it turns out that they were fed the story by Melvin Belli, a lawyer, who had an interest in these mass tort cases and was suing the manufacturers of Bendectin and stood to make a lot of money had they been found culpable.



Good place to start is

<http://www.fda.gov/Safety/Recalls/default.htm>

And don't miss

<http://www.fda.gov/Food/FoodSafety/FSMA/ucm249087.htm>

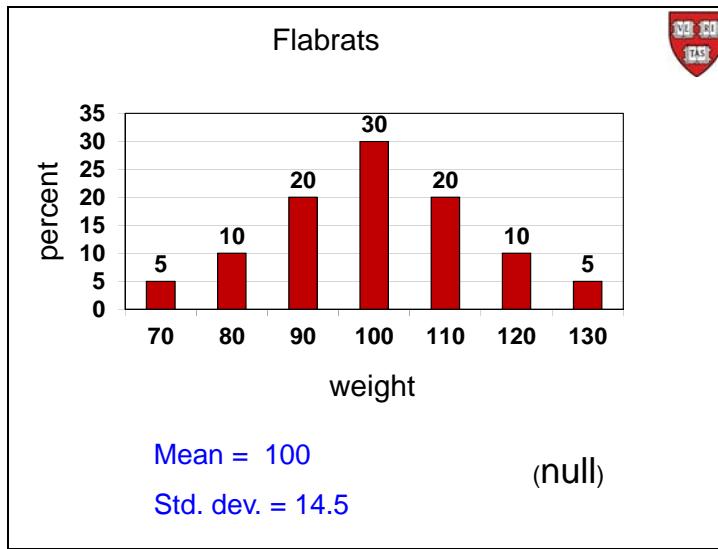
If you are interested in this topic, a good place to start is the FDA (Food and Drug Administration). They track recalls, and they do it for medications, for medical devices, et cetera. A definite do-not-miss is the bottom URL. It is a little difficult sometimes to find what you are really looking for in these large government agencies, but it is all there if you look carefully.

## Power



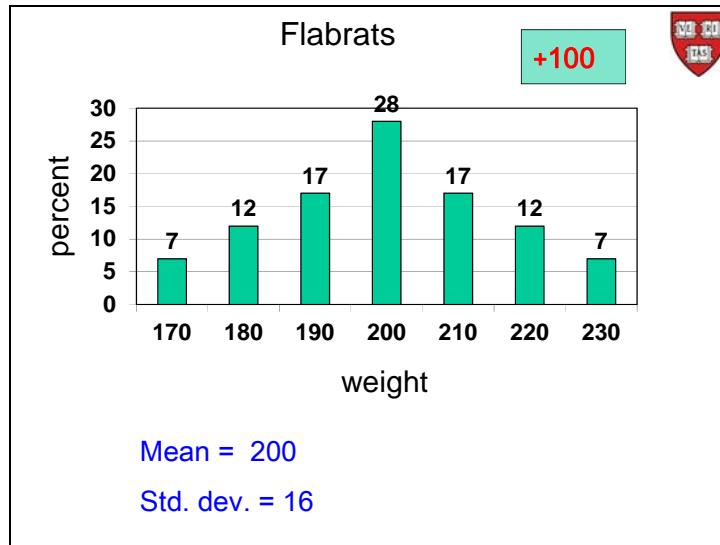
Flabrat

The next topic we want to investigate is power. And to help us explain power, we turn to our friend the flabrat. Some years ago, when I first talked about flabrats, a student gave me this picture and said, here is a flabrat. This flabrat is eating a little bit of leaf here. And it's not really a flabrat, but anyway returning to statistics, flabrats are fabulous because they are sensitive to their diets and their weight. So we use them in the lab and just measure their weights.



Here is a bar chart describing the weights of flabrats out in the wild. Not that they are very wild, but we see that 5% weigh 70 units, 10% weigh 80 units, 20% weigh 90 units, and so on. If you need a unit of weight, you can think of your favorite unit, such as a gram if you want to, but it

does not have to be, and if we put this distribution into our statistical calculator, we find out that its mean is 100 and its standard deviation is 14.5. Let us call this our null distribution.

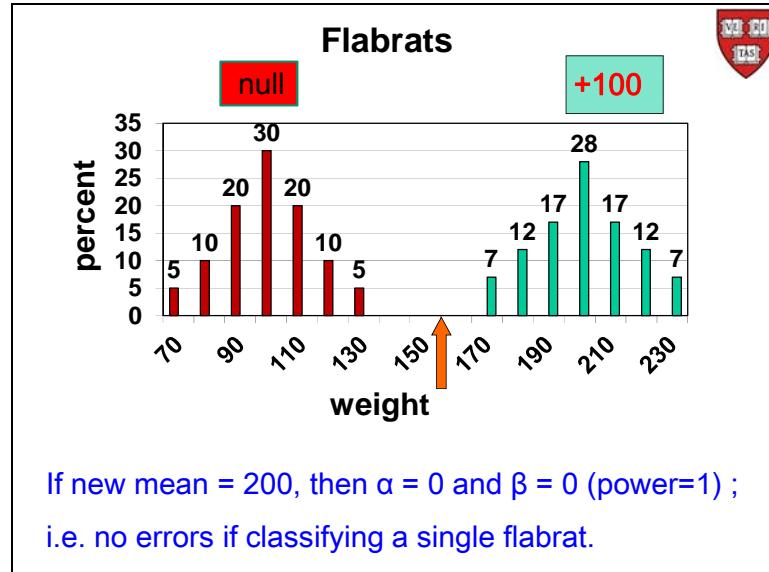


If we feed our flabrats a diet called the “Plus 100” diet, we shift the whole distribution 100 units to the right and the mean becomes 200, but the standard deviation also changed a bit. Instead of having 5 on the extremities, we have 7% on the extremities, amongst other changes. Now the standard deviation is 16. (This was done so as not to confuse you later.)

Here is the problem: we have two labs set up on either side of a corridor, and the lab on the left houses the flabrats with their natural diet (null), and on the lab on the right houses the flabrats who are on the “Plus 100” diet.

One morning you come to work and there is a flabrat in the corridor, and you say to yourself, oh my gosh, where did this flabrat come from? Does this flabrat belong in the left lab or the right lab?

The challenge is to properly classify the wandering flabrat.

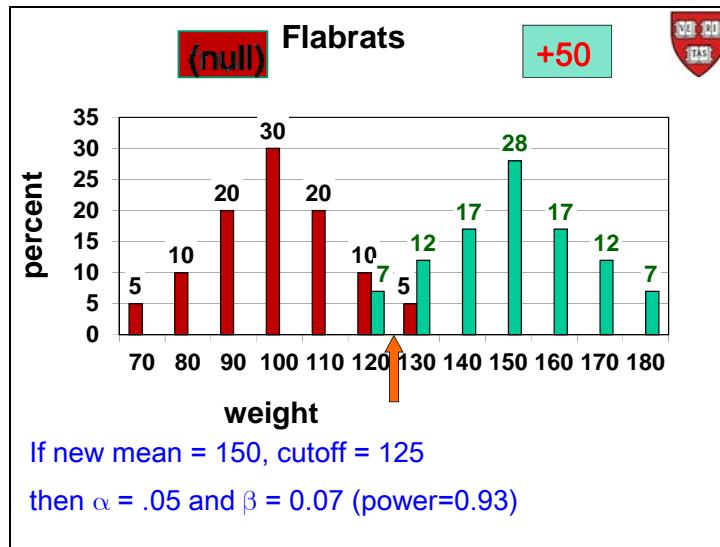


If we put the two distributions side by side, how would you decide to classify the wandering flabrat? Surely, the decision should be made on how much the wandering flabrat weighs. So you can pick it up, weigh it, and then decide.

Then a sensible criterion might be, choose a cutoff and a flabrat weighing less than the cutoff will be sent to the left lab, the one housing the null distribution, and a flabrat weighing more than the cutoff will be sent to the right lab, the one housing the “Plus 100”.

Placing the cutoff to the right of 130 and to the left of 170 will give us perfect discrimination because the two curves do not overlap.

If we place this in a hypothesis testing framework, we see that both our alpha and beta are going to be zero, and thus our power is one.



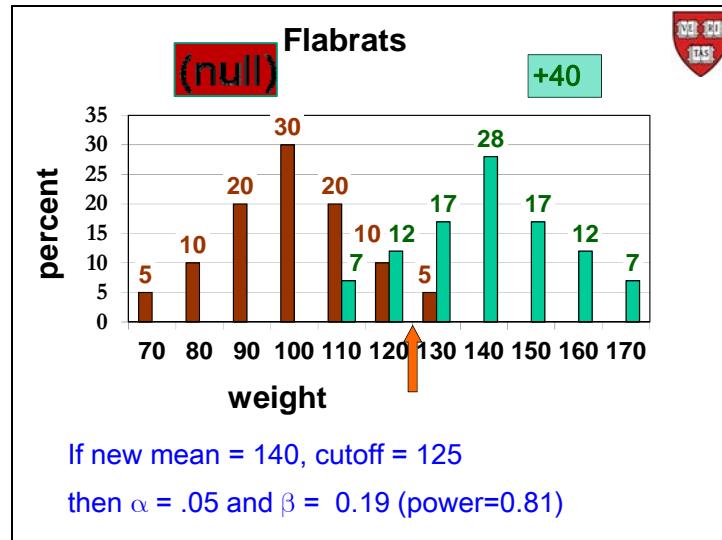
Your friend, one floor up, is also studying flabrats, except she is using a “Plus 50” diet as her alternative diet.

If we place the “Plus 50” distribution next to the null, we see overlap: there are flabrats in both groups who weigh 120 units (10% of the null and 7% of the “Plus 50’s) and 130 units (5% of the null and 12% of the “Plus 50’s). So now our discrimination is not as clean, and we may have some confusion—for example, if our wandering flabrat weighs 120 or 130 units.

But it still makes sense, since the “Plus 50’s tend to weigh more than the nulls, to use a weight cutoff as our discriminator, even though some of the nulls weigh more than some of the “Plus 50’s. And that is the source of our potential errors.

Suppose we want to keep our alpha error at most 0.05, then that means the cutoff must be to the right of 120. If we make 125 our cutoff, then alpha is 0.05 and beta is automatically 0.07 (and so the power is 93%). Anything much bigger, let us say, more than 130, will make our alpha smaller but it will increase our beta, and thus decrease our power.

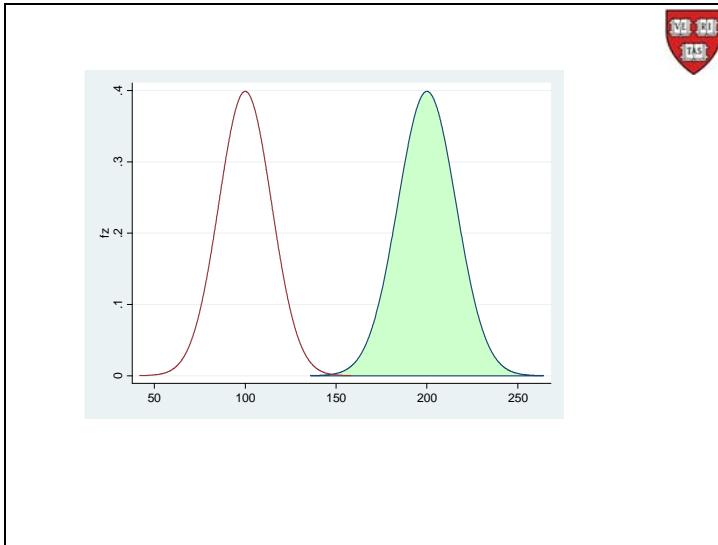
So, in general, shifting the cutoff to the left increases alpha and decreases beta (increases power), with the opposite effect if we shift the cutoff to the right (decrease alpha, increase beta, decrease power). This is exactly what happens with hypothesis testing—we can associate the null with the flabrat coming from the null distribution, and the alternative with the flabrat coming from the “Plus 50” distribution. In this case we say we have 93% power to distinguish between the null and the “Plus 50” distribution on the basis of a single observation.



Your friend, two floors up, is also studying flabrats but with the subtler “Plus 40” diet. Now we see a bigger overlap.

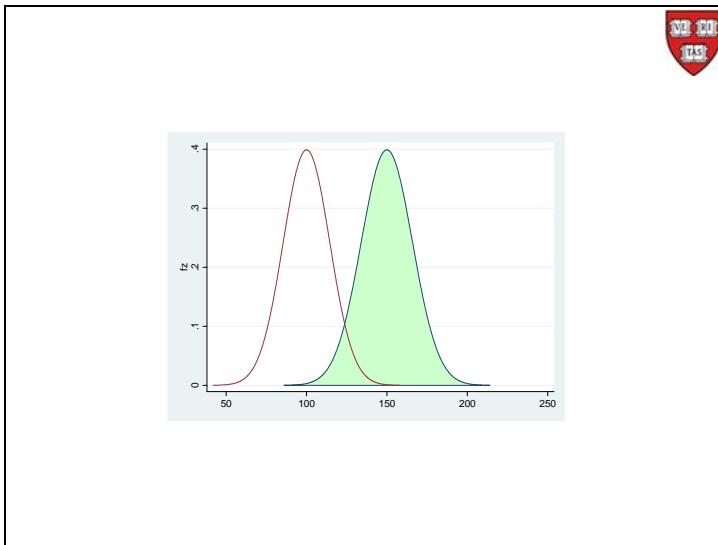
Suppose that to retain the alpha at 5% you keep your cut-off at 125, then what does this do to the power (and beta)? Because of the bigger overlap we see that not only do we have the 7% falling at 110, but we also have 12% falling at 120, and thus both groups weigh less than the cutoff of 125, and thus a wandering flabrat from either of these groups would be classified as

being in the null. So our beta now, because of the larger overlap, has increased to 0.19, dragging the power down to 81%.

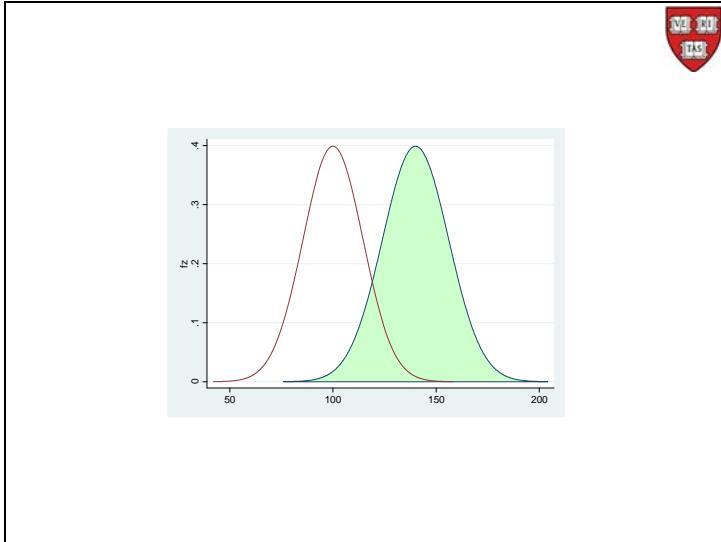


To smooth things out, let us bid farewell to the flabrats and look at smooth densities instead of our barcharts. It is much easier to now slide the alternative up and down the horizontal scale.

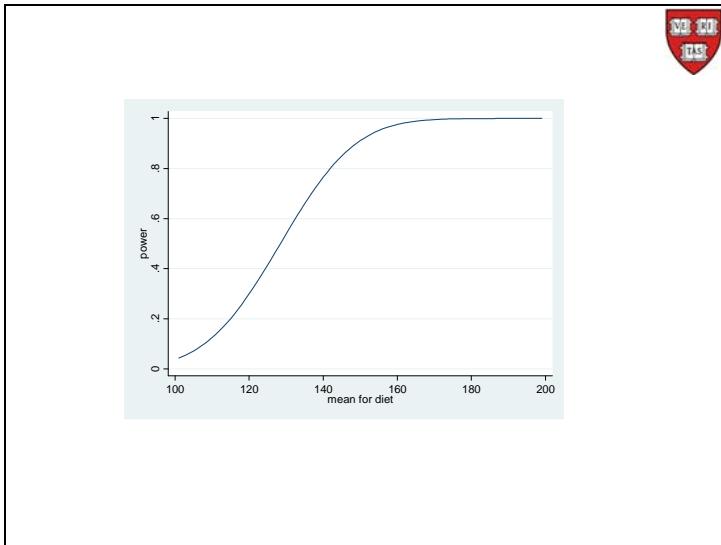
Above you see the analog of the “Plus 100” diet. Since these are two normal curves they both theoretically go off to infinity at both ends, so even at “Plus 100” there will be a tiny bit of an overlap of the two curves, and neither alpha nor beta ever is perfectly zero.



For the “Plus 50”, above, the overlap between the two distributions is much more noticeable than it was for the “Plus 100” situation, above. Now the mean is at 150, closer to the null mean of 100, than when the mean was at 200 for the “plus 100”.



This is the chart for the “Plus 40” diet; much greater overlap because now the mean for the alternative is at 140.



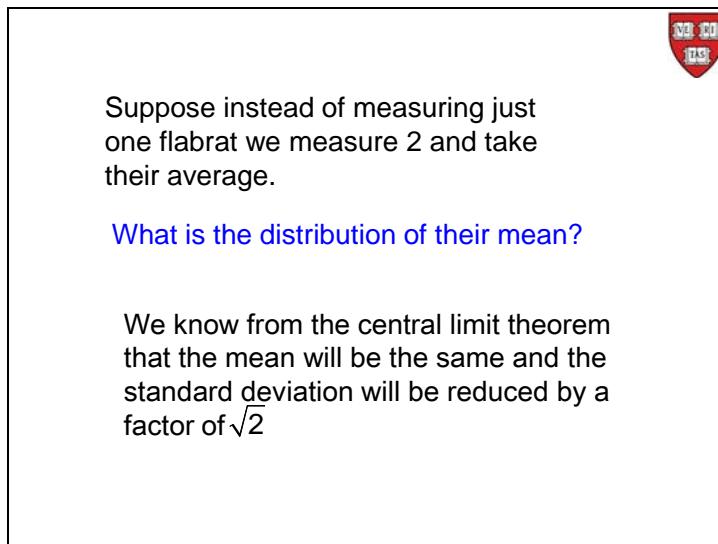
We can encapsulate this changing alternative with this curve, called the *power curve*, which displays the probability of rejecting the null hypothesis as a function of how far away the mean of the alternative is from the null mean. So you can see if there is a small difference between

the null and the alternative diet, we are not going to have very much power to detect the difference. So if the wandering flabrat comes from a population whose mean is 140, reading up to the curve from 140 on the horizontal, we have approximately 90% power of properly classifying such a flabrat.

In summary, the power increases as the two population means become further apart. Intuitively, if the diet that the wandering flabrat is on is going to make him or her that much heavier, then it is going to be that much easier to distinguish him or her from the flabrat who is on the null diet. So the power increases as the delta increases, if all the while we hold alpha constant.

So what is causing the loss in power? It is the overlap—the region where both populations have representation. Will we always have overlap? Are we stuck with the amount of overlap we have? We saw that to decrease the overlap we can pull the curves apart (increase delta, the distance between the means). Another way, if we maintain the same distance, is to make the curves tighter around their means. This can be achieved by decreasing the standard deviations. Since we are stuck with the standard deviations because they are fixed in the population, how else can we decrease them?

This is reminiscent of the Central Limit Theorem. There we saw a decrease in the standard deviations—more precisely the standard errors—when we considered the distribution of the sample means and we allowed the sample size to increase. So extending this idea here, we can look at two distributions with means delta apart and decrease the overlap if instead of looking at the distribution of an individual; we looked at the distribution of means of individuals from both distributions.



Suppose instead of measuring just one flabrat we measure 2 and take their average.

**What is the distribution of their mean?**

We know from the central limit theorem that the mean will be the same and the standard deviation will be reduced by a factor of  $\sqrt{2}$

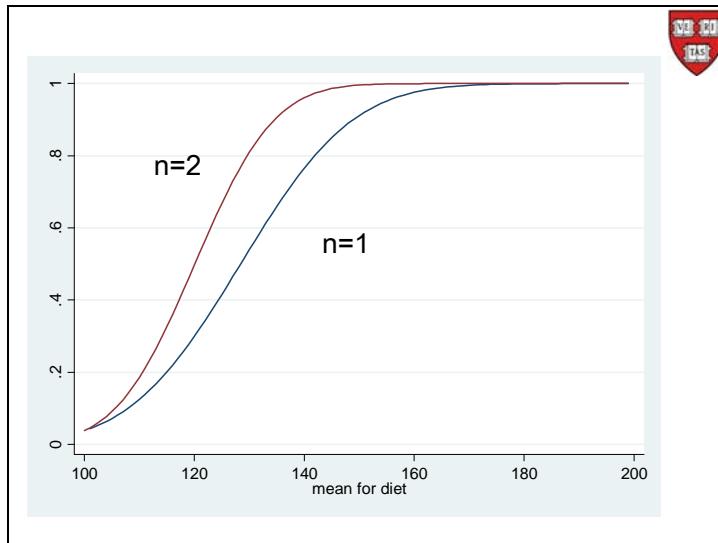


So in terms of our flabrats, if instead of finding one wandering flabrat out in the corridor we found 2 flabrats out wandering, then instead of weighing and classifying them individually, just like we did before for the single wanderer, assume that they both came from the same lab<sup>3</sup>. Then we can take their average, and classify them according to their average weight. (Remember from the Rice simulation that we saw much more variability in the top panel, the population, than we say in the third panel down, where we saw the distribution of the sample

<sup>3</sup> This is my story and I can make up the assumptions!

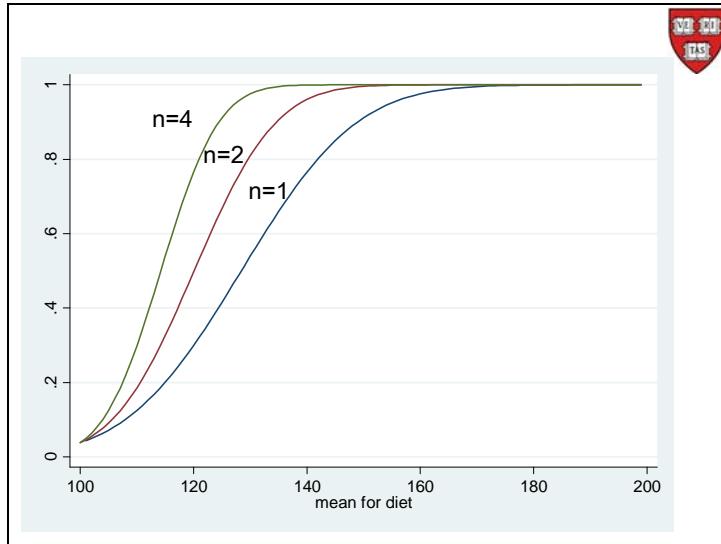
means—in the top panel the appropriate quantity to measure variability was the standard error, whereas in the third panel the appropriate quantity was the smaller standard error.)

## Power and Sample Size



So now that you have convinced yourself that you can increase the power and lower the alpha, or keep the alpha constant, by increasing the sample size, you can see that the three are interrelated—namely, alpha, power (beta) and the sample size. Indeed, fixing two, fixes the third. (For those who like to think in these terms, that means we have three variables but only two degrees of freedom.)

For the moment, let us keep alpha constant, and look at the power curves and compare the power curve when the sample size is 1, to the power curve when the sample size is 2. We have precisely these two power curves plotted above. We can see, for example that when the mean of the alternative is 120 (which would be for the “Plus 20” diet) then we do not have a very high chance of properly classifying a wandering flabrat on this diet; indeed, the power is about 30%, or so. On the other hand, if 2 flabrats had escaped, then the power zooms up and almost doubles to about 60%.



If we had 4 wandering flabrats, so now we have a total prison outbreak, then let us superimpose the power curve for  $n=4$ . This curve dominates the other two curves to reflect an increase in power at every point of the domain, except, of course at the null, namely 100, because we have kept that at 5%.

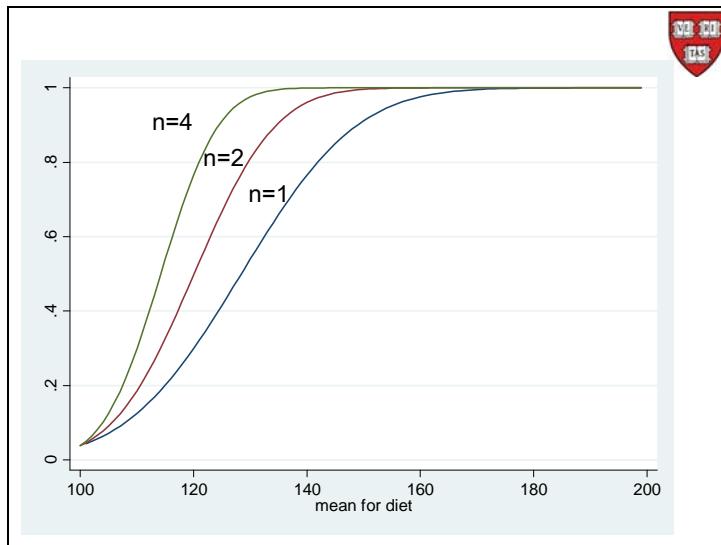
So now even at 120 (the “Plus 20” diet), with 1 wandering flabrat we said the power it was about 30, but we go up to 4 wandering flabrats, then we have almost 90% power—let us say 87%.

### In summary:

Thus the power increases as:

1. Real  $\mu$  gets further away from the hypothesized  $\mu_0$  ( $\Delta$  gets larger).
2. Sample size increases.

So, in summary, from before we have that the power increases the more the two means of the distributions separate, and now we can add that the power also increases if the sample size increases.



Before leaving this graph let us make one more observation. We just read the graph by finding a point on the horizontal axis (such as 120) and shooting up and reading the values on the three curves. Now let us look at what happens if we look at first identifying a point on the vertical axis and looking horizontally until we hit the three curves.

For example, let's say we zero in on 0.8. We can ask, if I design my study to have 80% power, how big a difference between the means will I be able to detect? Of course, when we say "be able to detect" that has to be interpreted as detect with a certain power of detection. (No certainty in this course!) had power of 0.8, how small a difference would I be able to

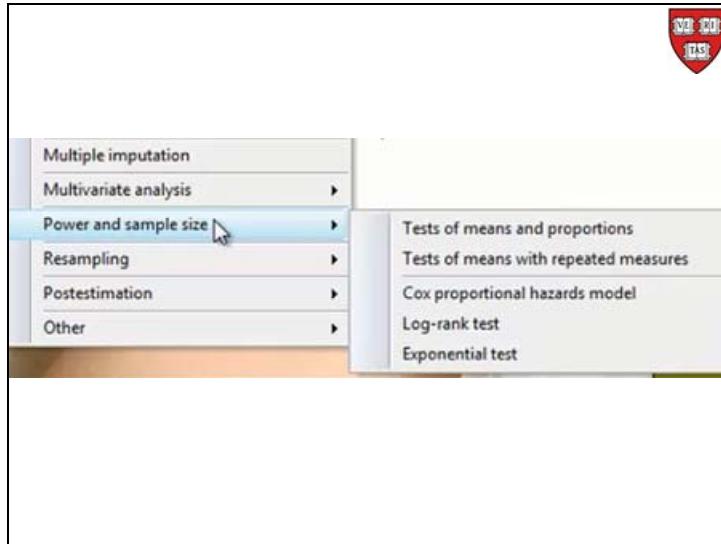
The answer is in the graph if we draw a horizontal line at 0.8. That line crosses the "n=4" curve at let us say 117; the "n=2" curve at 128; and, the "n=1" curve at 141. What that means is that we have 80% power to detect a "Plus 17" diet with a sample of size 4—actually we should say 17, or higher, because this power curve goes above 80% to the right of 117.

On the other hand, if we only have a sample of size 2, then we need to have a bigger separation of the means, namely to 128, or a "Plus 28" diet, to have an 80% chance of detecting a difference. That is, means of 128, or higher, to have at least an 80% chance of detecting a difference.

Finally, with a single wandering flabrat, to have at least an 80% chance of detecting the difference we need to have the alternative mean be 141, or higher.

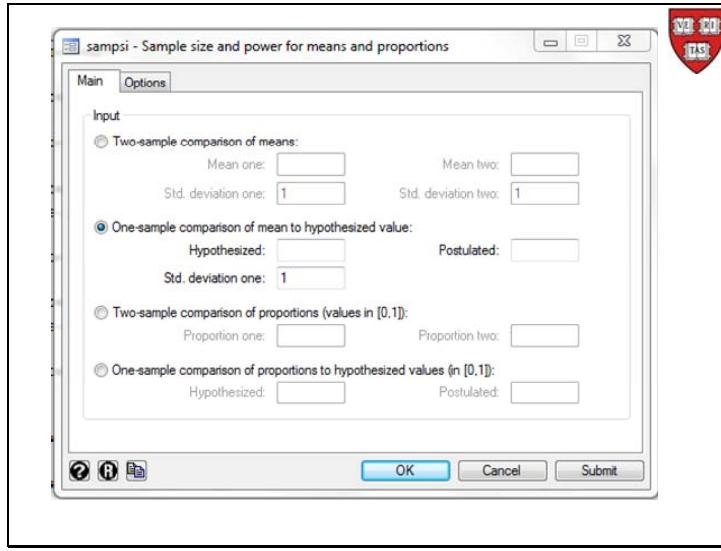
This is very often the way we design studies: namely, decide how big a difference you wish to detect—for example, what difference would be clinically meaningful—decide what power would be acceptable, and determine how big a sample size you will need to satisfy those constraints.

These considerations should remind us of the thinking we did when we looked at diagnostic testing. We had two kinds of errors and it was the accuracy of our measuring instrument that determined how precise our measurements could be. Here we can think of our measuring instrument as being the sample and the statistical analysis we do on that sample, and one way of buying a more precise instrument is to get a larger sample.



Let us turn to Stata to help us do the necessary calculations that guide us when designing a study. We first look at the one-sample situation and then at the two-sample case:

We start by clicking on “Statistics” and choosing “Power and sample size” and then “Tests of means and proportions”, to get:



Let's look at the second one, “One-sample comparison of mean to hypothesized value:”. Now what it wants from us are three quantities: the hypothesized value of the mean, the postulated value of the mean (where you want to calculate the power), and the standard deviation.

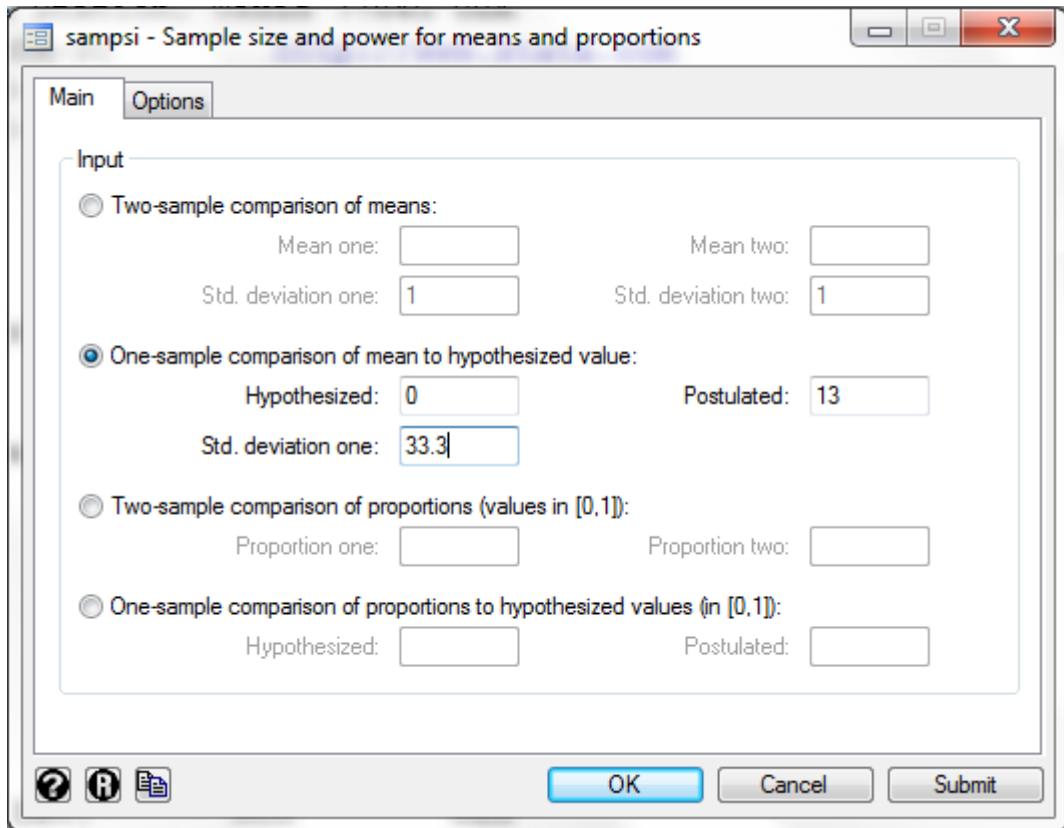
Before answering these questions let us set this aside as we set up an example.

```
. gen totchol = totchol2 - totchol1  
(675 missing values generated)
```

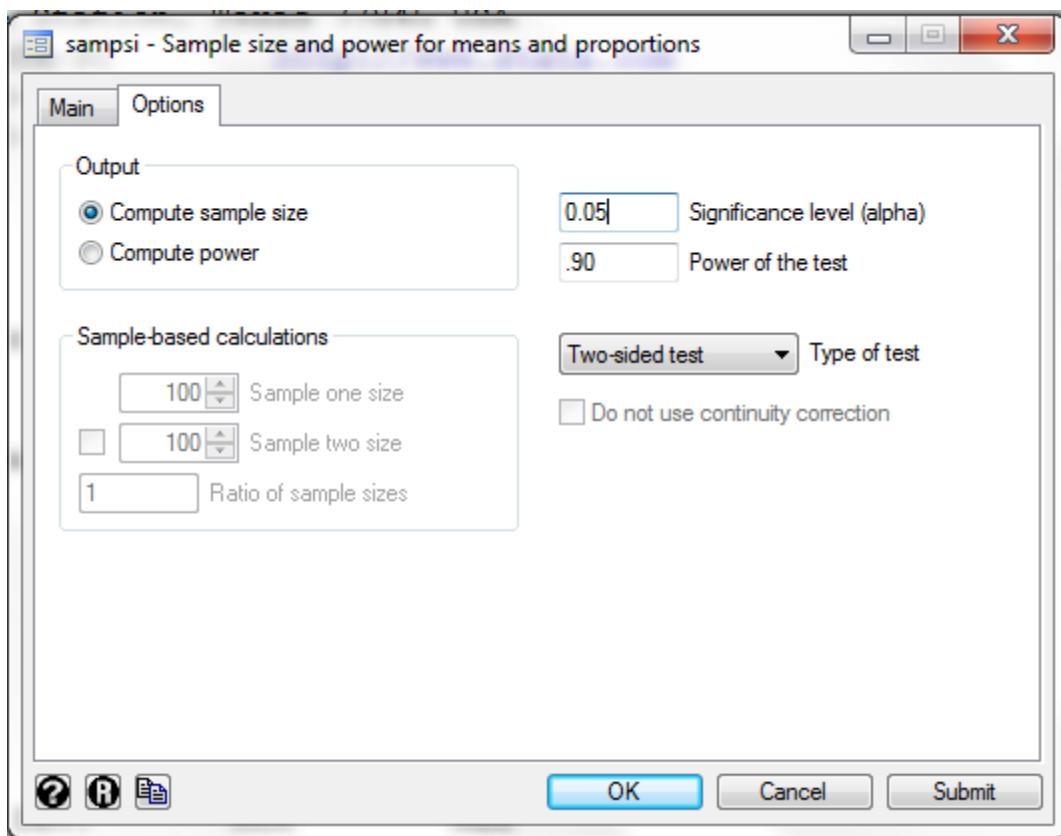
```
. summ totchol
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totchol	3759	13.16574	33.30605	-159	321

We had previously defined *totchol* to be the difference between the total cholesterol at visit 2 minus the total cholesterol at visit 1 in the Framingham Heart Study. When we summarize this variable we find that its mean is 13 and its standard deviation is 33.



Returning to our program, *sampsii*, let us fill in the hypothesized value as zero—so we are hypothesizing that there is no change in total cholesterol between the two visits—that the standard deviation is 33.3—ordinarily, here we need something that approximates the truth and one relies on past experiences with such observations, short of that, one cannot proceed—and then the postulated value. We have used 13, but this is the point where the power will be calculated. It should make subject-matter sense.

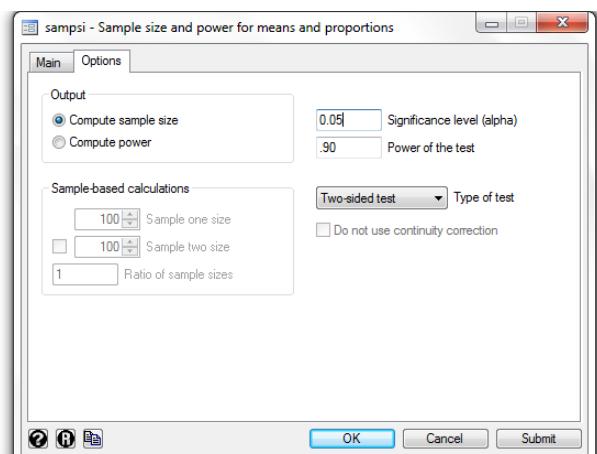


Now click on the “Options” tab to get this menu.

What alpha do we want? Let us put our friend 0.05 in for alpha, and it suggests a power of 90%, so let us leave that. And let us look at a two-sided test. And so what we're going to ask Stata to do for us is the compute the sample size.

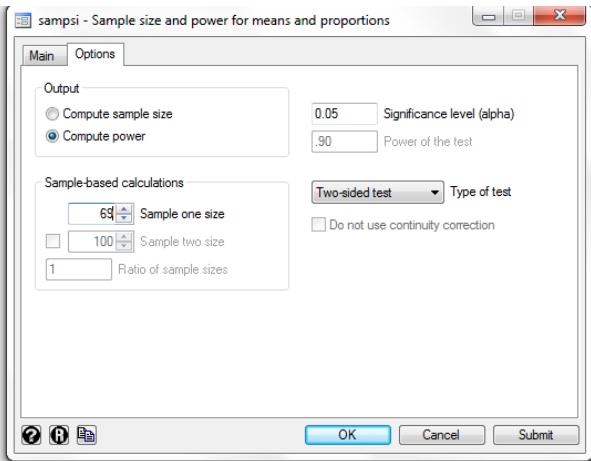
Now when we click “Submit” we are asking the question, how big a sample do I need if I am going to test the hypothesis that the population mean is zero, at the alpha of 0.05 and I want a a 90% chance of detecting a difference of 13.

```
. sampsiz 0 13, sd1(33.3) alpha(0.05) onesample
Estimated sample size for one-sample comparison of mean
to hypothesized value
Test Ho: m =      0, where m is the mean in the population
Assumptions:
    alpha =    0.0500  (two-sided)
    power =    0.9000
  alternative m =      13
        sd =     33.3
Estimated required sample size:
    n =      69
```



Stata returns the value 69.

```
. sampsi 0 13, sd1(33.3) alpha(0.05) n1(69) onesample
Estimated power for one-sample comparison of mean
to hypothesized value
Test Ho: m =      0, where m is the mean in the population
Assumptions:
    alpha =    0.0500  (two-sided)
alternative m =      13
    sd =     33.3
sample size n =      69
Estimated power:
    power =   0.9002
```



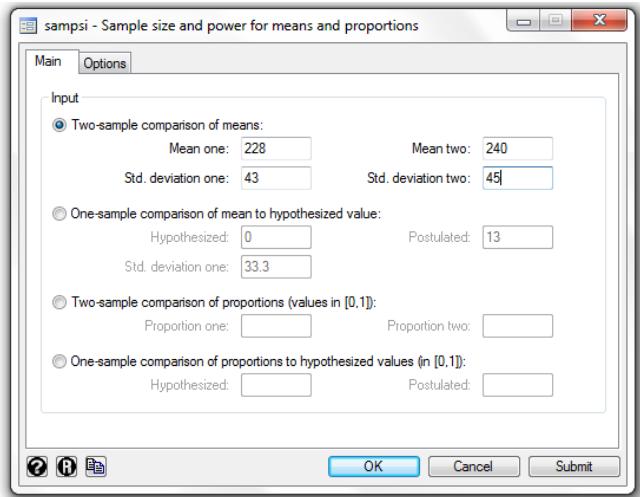
We could have done it the other way around asking Stata to compute the power for us (all the while at 13) when we have a sample of size 69 by clicking on “Compute power” and filling in the appropriate window in “Sample-based calculations”.

The answer we get—as expected—is 90.02%.

You can explore any number of “what if” scenarios by changing any of the numbers the program asks for, and get an idea of how uncertainty and precision vary together with the amount of knowledge needed etcetera.

```
. tab hyperten , summ(totchol1)

  Incident Hypertension | Summary of Total cholesterol
  Incident Hypertension | (mg/dL), exam 1
  Incident Hypertension | Mean Std. Dev. Freq.
  No             227.98031 42.644789 1168
  Yes            240.25638 44.919635 3214
  Total          236.98425 44.651098 4382
```



Let us now look at the two sample comparison of means. We first looked at this when we looked at the difference in total cholesterol level at visit 1 between those who had had a hypertensive event and those who had not. Here is the summary of those two groups.

The means are 228 and 240, so let us use those means in Stata. The standard deviations for the two groups are at 43 and 45, so let us use those too.

```
. sampsi 228 240, sd1(43) sd2(45) alpha(0.05)

Estimated sample size for two-sample comparison of means

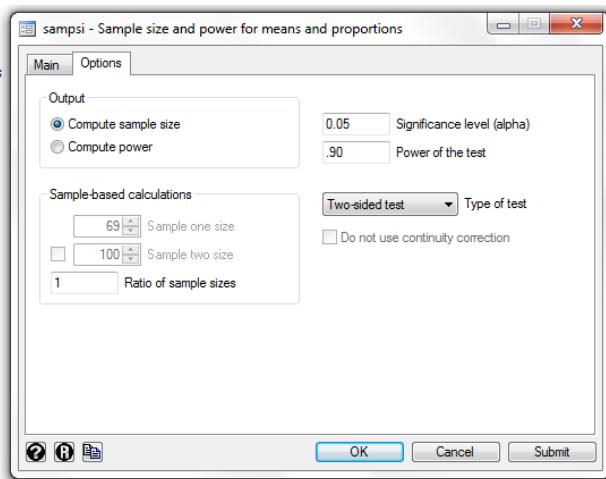
Test Ho: m1 = m2, where m1 is the mean in population 1
          and m2 is the mean in population 2

Assumptions:

alpha = 0.0500 (two-sided)
power = 0.9000
m1 = 228
m2 = 240
sd1 = 43
sd2 = 45
n2/n1 = 1.00

Estimated required sample sizes:

n1 = 283
n2 = 283
```



Now, let us look at our options. Let us leave the significance level at 0.05 and the power at 0.90 and leave the “Type of test” at Two-sided test, and submit that.

What we get back is that we are going to need a sample of size 283 from each of the two groups. (We can also play with the ratio of the two sample sizes, but I leave that to you to discover why you may want to do that.)

What if we want to spot a bigger difference, say between 228 and 250, then we go back to the previous menu and change the 240 to 250 and resubmit to get that we would need a much smaller sample of 85 from each of the groups.

These sample size calculations are very important when one is designing a study because they often dictate the difference between what is, and what is not a feasible study. Once again, I would recommend you explore different scenarios with this program to see how all these parameters are interrelated.

## ANOVA



One population:

$$H_0 : \mu = \mu_0$$

Two populations:

$$H_0 : \mu_1 - \mu_2 = \Delta$$

Multiple populations:

$$H_0 : \mu_1 = \dots = \mu_k \quad (k \geq 2)$$

So far we have studied testing the mean of single population and then the means of two populations. What happens when we have more than two populations? This question was very useful in agricultural experimentations in the past when one was restricted to a single growing season per year, so it was beneficial to perform more than one experiment a year. In medicine we are faced with a similar problem when investigating bleak situations such as lung cancer where curative treatments are not forthcoming and one has to investigate a large number of potential treatments and there is some savings in time and effort in doing them simultaneously.

There are a number of ways to approach this problem, but I first want to concentrate on the simple one where we want to test the single hypothesis that the means of each of the populations are equal to each other.



Data, k independent, random samples:

Population:	1	2	...	k
	$x_{1,1}$	$x_{2,1}$	...	$x_{k,1}$
	$x_{1,2}$	$x_{2,2}$	...	$x_{k,2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample size	$n_1$	$n_2$	...	$n_k$
Sample mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Sample sd's	$s_1$	$s_2$	...	$s_k$

So we start with  $k$  independent random samples. For each we have a sample size, a sample mean and a sample standard deviation.

**Populations:**



means:  $\mu_1, \dots, \mu_k$   
s.d.'s :  $\sigma_1, \dots, \sigma_k$

**Null hypothesis:**

$$H_0 : \mu_1 = \dots = \mu_k$$

**Assumptions:**

1. Populations are normal.
2. Homoscedasticity:  $\sigma_1 = \dots = \sigma_k$
3. Independent samples ( $k$ )

The  $k$  populations have means and standard deviations, and we wish to test the hypothesis that these  $k$  population means are all equal to each other.

We only look at the situation when all  $k$  populations are normally distributed, the populations are homoscedastic, and the samples are independent. One could remove these assumptions, but that is beyond the scope of this course.

We could test the null hypothesis by testing for each pair  $i,j$  the hypotheses:

$$H_0 : \mu_i = \mu_j \quad i, j = 1, 2, \dots, k$$



But if we did proceed in this fashion consider the type I error rate: e.g.  $k = 5$  so there are 10 hypotheses we need to test.

Then ask ourselves, what is the probability that we get it right 10 times, even if the null hypothesis is true?

Hint :  $(0.95)^{10} = 0.6$       Alternatively,.....

One way to test the hypothesis that all  $k$  population means are equal is to use what we have developed so far in the course and compare all the populations pairwise; since we know how to test the equality of the means of two populations.

Think a little about what would happen to your error rate if you proceed in this fashion. For example if  $k$  is equal to 5, then you would have 5 combination 2, or 10 tests to do. What is the probability that at least one of these is wrong. That is one minus the probability that you get all 10 right. If all ten tests were independent and each was tested at alpha is 0.05, then the probability that you get at least one wrong, if the null hypothesis is true, is  $(1-0.95^{10}) = 0.40$ . Thus your 5% has ballooned to 40%. Granted not all the tests are going to be independent, but in some sense then things might be even worse.

So let us hope that this is not the best approach we can take. There is another approach called the *analysis of variance*, and as the name implies, we study the behavior of the sample variances to direct us to an answer.

#### Homoscedasticity assumption and within variance



From the assumptions we have that  $s_1, s_2, \dots, s_k$  all estimate  $\sigma$  the common value of the standard deviation in each of the groups.

So, combine to get a better estimate:

$$s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \dots + (n_k - 1)s_k^2}{n_1 + n_2 + \dots + n_k - k}$$

This is the “within” variance estimator.

Consider the assumption we have made of homoscedasticity. If that assumption is correct, then we have  $k$  estimates of a common quantity; that being the common population variance. Call it  $\sigma^2$ . We can estimate it by taking a weighted average (since each of the  $n_i$  may be different) of each of the sample variances. Call this estimate,  $s_w^2$ . This is called the “within” variance estimator (generated from within the  $k$  samples).

To justify this estimator we call on the homoscedasticity assumption. When we test a hypothesis we usually start with, if the null hypothesis is true we expect to see... and proceed from there.

Homoscedasticity assumption and between variance



IF the null hypothesis is true (all means are equal).

Looking at the k groups, it's as if we were sampling k times from the same population.

So what do we expect to see if the null hypothesis is true? So if the null hypothesis is true, we expect all the means to be equal.

Data, k independent, random samples:



Population:	1	2	...	k
	$x_{1,1}$	$x_{2,1}$	...	$x_{k,1}$
	$x_{1,2}$	$x_{2,2}$	...	$x_{k,2}$
	:	:	:	:
Sample size	$n_1$	$n_2$	...	$n_k$
Sample mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Sample sd's	$s_1$	$s_2$	...	$s_k$

So let us look at these k groups. If the means are equal and we have homoscedasticity, then it's as if we were sampling k times from the same (normal) population.

The sample sizes may be different.

Data, k independent, random samples:



Sample size	$n_1$	$n_2$	...	$n_k$
Sample mean	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_k$
Sample sd's	$s_1$	$s_2$	...	$s_k$

Let us focus on the summary statistics. Possibly different sample sizes, but each of these sample means is estimating the same overall mean (under the null) and each of these sample standard deviations is estimating the same standard deviation (homoscedasticity assumption).

So from the central limit theorem,  $\bar{X}_1$  is a sample from a population that has mean  $\mu$  and standard deviation  $\sigma / \sqrt{n_1}$ ,  $\bar{X}_2$  a sample from a population that has mean  $\mu$  and standard deviation  $\sigma / \sqrt{n_2}$ , and so on. Both of the first two will give me information about  $\sigma$ . Indeed, all k of them will. Combining these k we can construct a combined estimate of  $\sigma^2$

So we can get a better estimate of  $\mu$  by combining all k estimators:

$$\begin{aligned}\bar{x} &= \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + \dots + n_k \bar{x}_k}{n_1 + n_2 + \dots + n_k} \\ &= \frac{\sum x}{n} \quad \text{where } n = n_1 + n_2 + \dots + n_k\end{aligned}$$

And another estimator of  $\sigma^2$ , the “between” estimator

$$s_B^2 = \frac{n_1(\bar{x}_1 - \bar{x})^2 + \dots + n_k(\bar{x}_k - \bar{x})^2}{k - 1}$$

First we combine them all to get an estimate of the common  $\mu$ . We want to construct a weighted mean because we want to weight the larger sample sizes more than the small ones. In fact what we have done is tantamount to ignoring what sample the observations comes from and just summing them all up and dividing by how many observations we have; our usual way of calculating the mean.

We can now see how the individual  $\bar{X}$  vary around the sample mean, and that should be related to the standard error. (Remember the definition of the standard error, it tells us how much the sample means vary around the overall mean.) We call this variance estimator the between estimator,  $s^2_B$ .



Ronald Fisher  
1890–1962



If the null hypothesis is true, these two estimates of  $\sigma$ , namely  $s_B$  and  $s_W$ , should be about the same. So as a measure of the null hypothesis, compare them:

$$F = \frac{s_B^2}{s_W^2}$$

is Snedecor's F with  $(k-1)$  and  $(n-k)$  degrees of freedom.

The genius of it all is to now just compare these estimators of the same quantity. We can take their ratio. It is called the F statistic<sup>4</sup>. This should be approximately 1 if the null-hypothesis is true. So just like we have had statistics and their distributions, for example the Z and the t, we now have the F and we can use it to determine p-values we can attach to a hypothesis.

Let us take a look at an example of the analysis of variance (ANOVA).

I have created a new variable, and called it *sexh*. And what this variable does is it breaks up our population into four groups. The first two refer to men, and the bottom two refer to women. And what the h stands for is hypertension. So I'm going to look at incident hypertension.

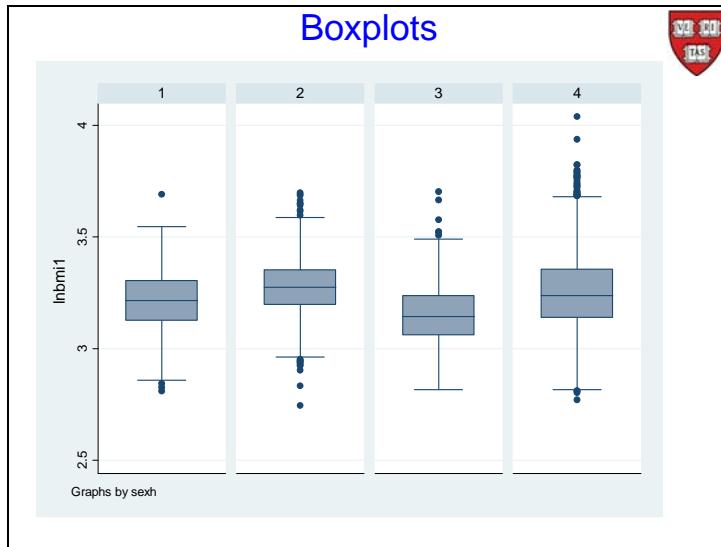
And so the first group (*sexh*=1) is going to be men without hypertension; the second (*sexh*=2) is men with hypertension; the third group (*sexh*=3) is women without hypertension; and the fourth

---

<sup>4</sup> The statistic is related to one Fisher had proposed in one of the most important papers about ANOVA he wrote in the 1920s, but this one was proposed and tabulated by George Snedecor who named it F in honor of Fisher. Apparently Fisher was none too pleased to have seen his statistic modified.

group ( $\text{sexh}=4$ ) is women with hypertension. What we would like to find out is whether there is a difference between these four groups, with respect to an outcome.

Let us define a new outcome, the logarithm of BMI. Remember, with the analysis of variance, we need to assume normality, so that is why we look at the logarithm of BMI; it is closer to being normally distributed than the raw BMI.



To get a feel for the data, let us look the box plots of these groups. From the boxplot, it looks like the BMI is slightly higher for men and women with hypertension, and it looks like the difference is bigger for women than it is for men.

Let's see what happens if we take a sample from this population, and submit it to the an ANOVA. So let us take a sample of size 25 from each of these four groups.



Summary of lnbmil			
sexh	Mean	Std. Dev.	Freq.
1	3.265275	.14501408	25
2	3.3025174	.15209203	25
3	3.1640145	.10438534	25
4	3.289139	.17237256	24
Total	3.254894	.15335159	99

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	.294016251	3	.098005417	4.63	0.0046
Within groups	2.01062129	95	.021164435		
Total	2.30463754	98	.02351671		

Bartlett's test for equal variances:  $\text{chi2}(3) = 5.7978$  Prob>chi2 = 0.122

The Stata command for the analysis of variance is `oneway`—there exist more complex analyses, this is the simplest one, and that is why it is called `oneway`.

We see that we lost one person in group 4; a missing value. Let us ignore that. The means and standard deviations for the four groups are reported. We see, in agreement with what we saw with the boxplots for the whole population, that those without hypertension (groups 1 and 3) have lower BMIs (log BMIs to be precise), and that women do better than men (have lower BMIs).

Here is the analysis of variance table. The column labeled MS gives us the two quantities we called  $s^2_B$  and  $s^2_W$ , above. Their ratio is in the column labeled F, and the p-value associated with the null hypothesis of equality of the four population means is given in the last column, 0.0046.

So at the 0.05 level we reject the null hypothesis of the equality of the population means.

The last line also gives us the results of Bartlett's test for homoscedasticity. This test is somewhat sensitive to departures from normality, but there it is and what it tells us is that the data seems to be consonant with a hypothesis that we have homoscedasticity. In other words our assumption seems to be acceptable.

WESLEYAN UNIVERSITY

```
. oneway lnbmil sexh, tab bon
```

sexh	Summary of lnbmil		
	Mean	Std. Dev.	Freq.
1	3.265275	.14501408	25
2	3.3025174	.15209203	25
3	3.1640145	.10438534	25
4	3.289139	.17237256	24
Total	3.254894	.15335159	99

Source	Analysis of Variance			F	Prob > F
	SS	df	MS		
Between groups	.294016251	3	.098005417	4.63	0.0046
Within groups	2.01062129	95	.021164435		
Total	2.30463754	98	.02351671		

Bartlett's test for equal variances:  $\chi^2(3) = 5.7978$  Prob> $\chi^2 = 0.122$

So where we left it was at this point where we said we will reject the null hypothesis that all the means are equal. Now here are all the means we observed, and you can see that group one and four look very close together, and both are close to group two. So what has caused us to reject the overall hypothesis?

Now remember the overall hypothesis was that all the means are equal. So any departure from this overall equality could be the cause of us rejecting the whole. It would be interesting to find out the cause(s).

There is something we can do. As an option in the *oneway* command I added comma *bon*. What did that give us?



Carlo Bonferroni  
1892--1960

### Bonferroni Correction:

If we wish to perform all possible pairs of comparisons, then there are  $\binom{k}{2}$  such comparisons. So to have an overall level of  $\alpha$ , one needs to perform each individual test at level

$$\alpha^* = \frac{\alpha}{\binom{k}{2}} \quad \text{or} \quad \alpha^* \frac{k!}{2!(k-2)!} = \alpha$$

The word *bon* is an abbreviation for the Bonferroni Correction. There are other ones we could use, but this one is quite conservative, and thus might be the best to use.

If we wish to perform all possible pairs of comparisons—so for example, 1 versus 2, and 1 versus 3, and 1 versus 4, and 2 versus 3, and 2 versus 4, and 3 versus 4, thus all the pairwise comparisons—then way to do that is to take your alpha, let's say 0.05, and divide it by k combination 2; the total number of possible pairwise comparisons.

So in this case k was 4, so k combination 2 is 6. So divide alpha by 6, and then test each one of these pairwise comparisons at this ( $\alpha/6$ ) level, to still maintain an overall level of alpha.

Or if you multiply by k combination 2, that is your overall alpha, and this is what Stata reports for us.



Comparison of lnkmi1 by sexh (Bonferroni)			
Row Mean- Col Mean	1	2	3
2	.037242 1.000		
3	-.10126 0.094	-.138503 0.007	
4	.023864 1.000	-.013378 1.000	.125125 0.020

So what the call to *bon* did when we ran this command with Stata is to produce this table. In the table we look at the couplet in each row/column combination. The upper number is the mean of the group identified by the row label minus the mean of the group identified by the column label. The lower number is the modified p-value (to accommodate the Bonferroni correction) associated with the test of the hypothesis that the difference between the groups identified by the row/column identifiers, is zero.

So suppose we wish to maintain our overall alpha of 0.05, then the only two pairs for whom we would reject the null hypothesis of no difference in the means are the pairs, 2 and 3, and 3 and 4—because those two p-values are the only ones less than 0.05.

So the one group that is sticking out is group number 3, the women who are not hypertensive. They are not different from men who are not hypertensive, but they are significantly different from both the men and the women who are hypertensive. No pair of mean differences amongst the other three groups is found to be significantly different from zero. So the non-hypertensive women seem to be the cause of the rejection of the overall null hypothesis in the original analysis of variance.

Marcello Pagano

# [JOTTER 7 CONTINGENCY TABLES]

Inference about the binomial parameter  $p$ . Sample size calculations. Odds ratios. Berkson's fallacy. Yule Effect.

We now have under our belts how to test and estimate the mean of a single population, of two populations, and more than two populations, when we have normal data. In other words, when dealing with continuous measures. Now let us switch attention to the situation when we have count data. Let us start with the binomial situation. Now what we have already seen about the binomial model is that it is useful for modeling dichotomous data: yes, no; male, female; alive, dead, et cetera.

**Binomial Distribution**



**X = number of successes**

$$P(X) = \binom{n}{X} p^X (1-p)^{n-X} \quad X = 0, 1, 2, \dots, n$$

n = 1, 2, ...

**Parameters:**

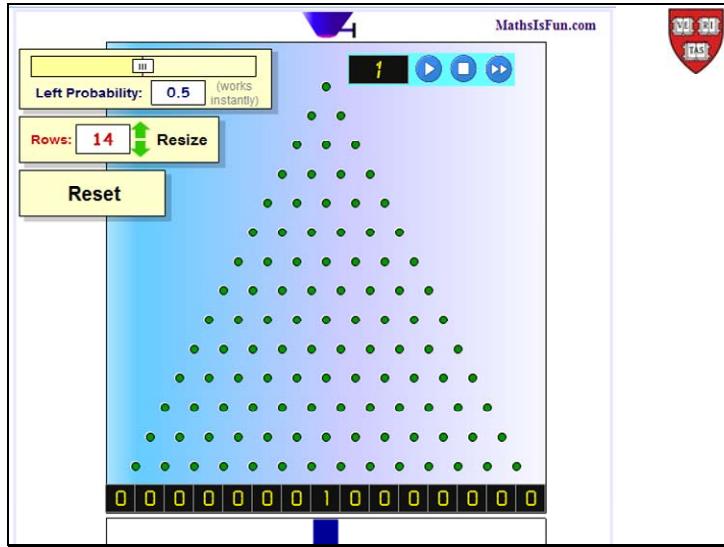
p = probability of success  
 n = number of trials

Mean = np      and  
 standard deviation =  $\sqrt{np(1-p)}$

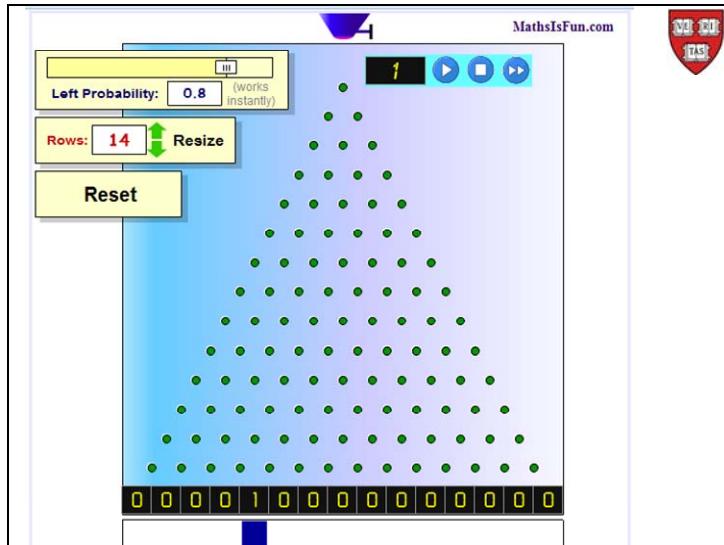
Now our challenge is how to estimate p in the binomial model. We usually assume that n is known, and now we are going to estimate p, the probability of a success at every trial, when we have n independent trials.

For example, we might think of a village where p is the prevalence of vaccinated children, and we take a random sample of 14 children. We can model the situation of sampling from such a village by using a binomial model. So, the mean number of children we find vaccinated in such villages, as we visit village to village, is np, with a standard deviation of  $\sqrt{np(1-p)}$ , and we put n=14 in these formulae.

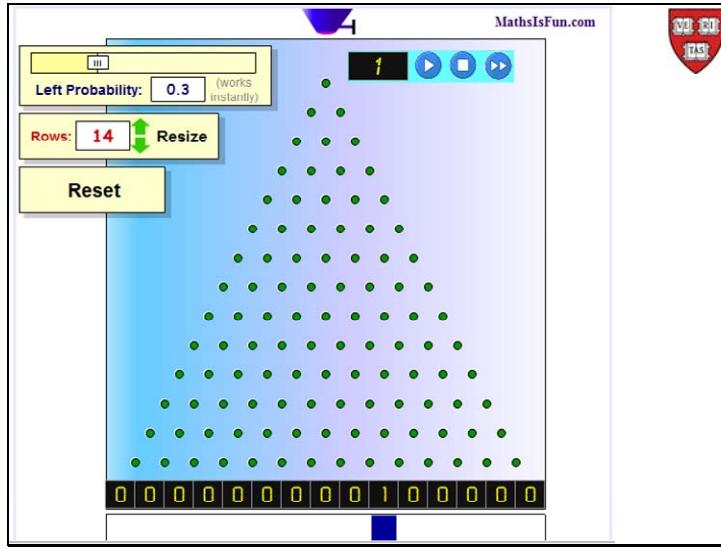
In the usual spirit of hypothesis testing, we approach inference about p by saying, if this is the truth, what do I expect to see? We can go to the Quincunx to see how the number of vaccinated children we actually do find in a village, varies from sample to sample.



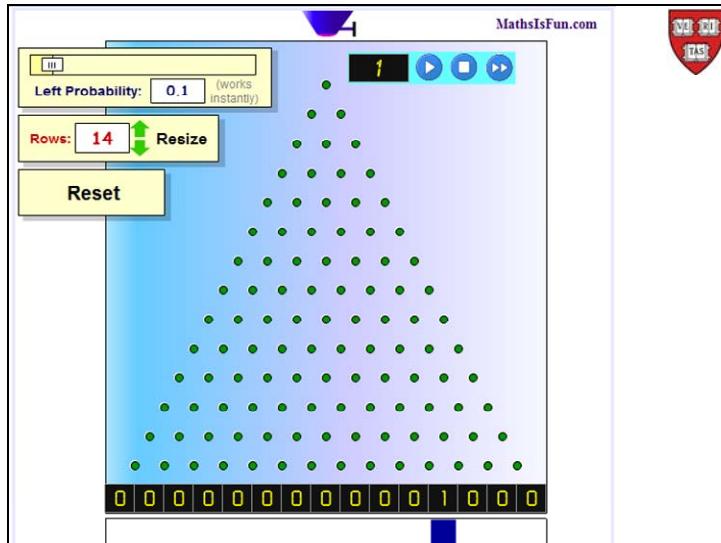
So here is the quincunx. I let one ball fall down with 14 rows—to simulate our asking 14 children in the village. I set the left probability at 0.5—to simulate a village with 50% vaccination coverage. We expect the observation to come down the center since at each peg it is equally likely to go left as it is to go right. It turns out, that it fell into the center bin! If the ball falls into the leftmost bin, that is finding zero children vaccinated. In the next bin to the right, is like finding one child vaccinated, and so on, all the way to the rightmost bin which represents our finding all 14 kids vaccinated.



If I now change the “Left Probability” to 0.8 and rerun the simulation, the ball should fall left of center in our view of the Quincunx; as, indeed, it does.

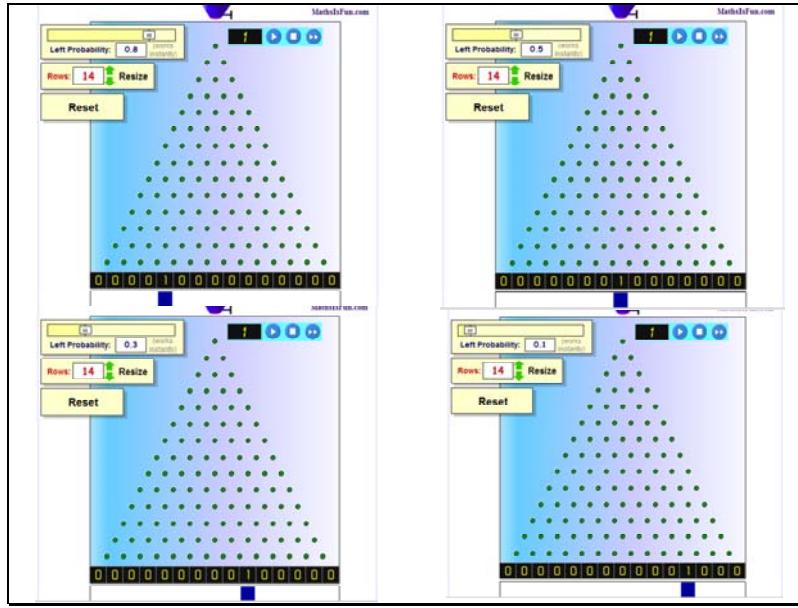


Next, I changed the left probability to 0.3. This should result in the ball ending up right of center, as indeed it does. (The left probability is the probability of being not vaccinated. Sorry about that, but I did not design the Quincunx!)

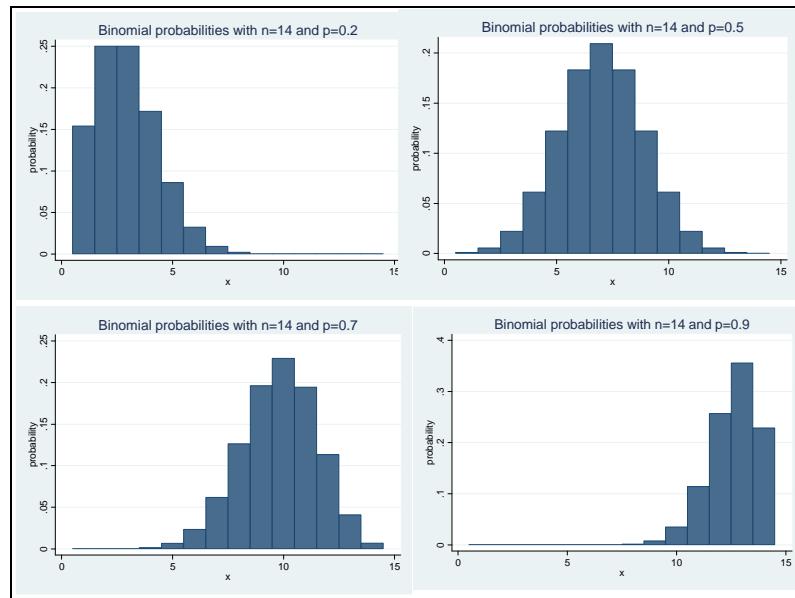


Lastly, when we set the left probability to 0.1, then the ball should end up even more towards the right end, as it does.

So here is the challenge: Suppose we know where the ball ended up—how many kids did we find vaccinated in the village—determine from that information what the value of the “Left Probability” is—or what the village (not the sample, we know that) vaccination coverage, or prevalence, is.



So, for example, here are the four snapshots of the Quincunx that we just looked at. It is not going to land in exactly the same spot every time, as we have seen often enough.

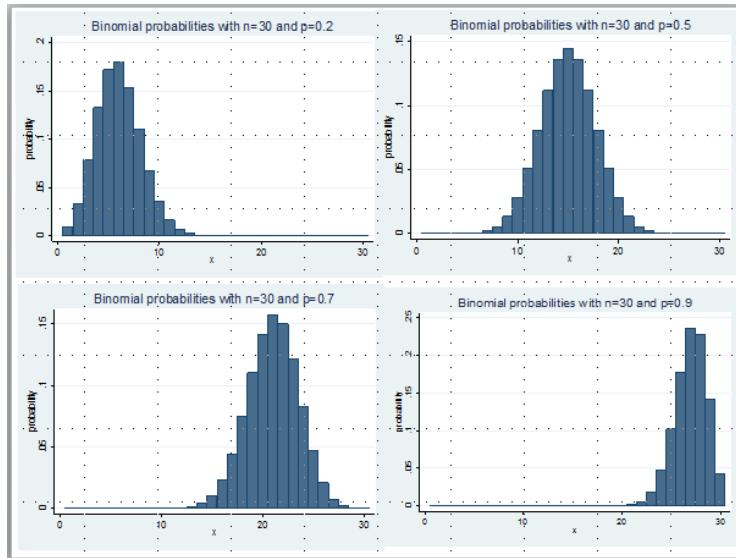


If we repeatedly run the Quincunx, we get pictures like these to reflect the distribution of the balls in the bins. These are actual theoretical binomial distributions generated by Stata, but we know that if we run the Quincunx ad infinitum, this is what we would see.

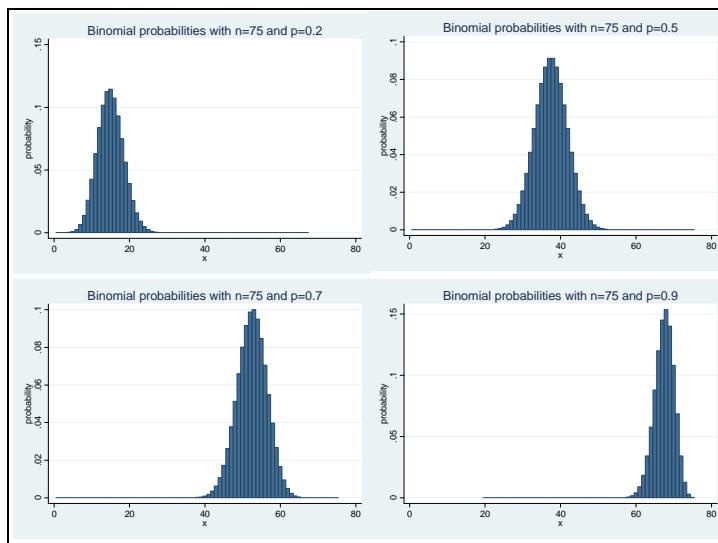
These distributions peak at points that travel from left to right and their location monotonically depends on the size of  $p$ , the probability of a success at a single trial; the peak travels from left

(when  $p$  is small) to right (when  $p$  is large). Note that the  $p$ 's in these pictures correspond to (1-“Left Probability”) in the Quincunx pictures.

These peaks are important as they direct us to where we expect most of the balls to land. They mostly land around the peak because that is where most of the probability mass is. This is a fancy way of saying that when we run the Quincunx, the height of the bin tells us how popular that bin is, and the most popular bin is the one that corresponds to the peak,  $np$ .

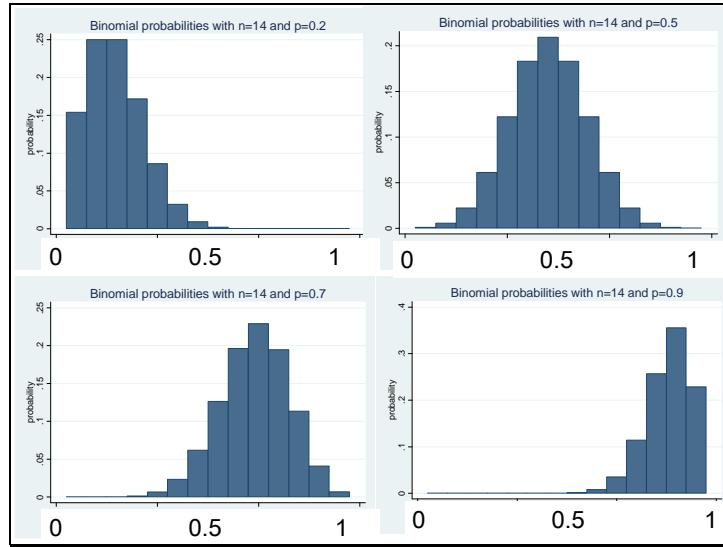


If we increase the number  $n$ —the number of children we check in the village—from 14 to 30, this is what we see: a tighter agglomeration of the probability mass around the peak, or mean. And so, just like with the flab rats, it is much easier to distinguish between any two of these four distributions than it was with  $n=14$ .

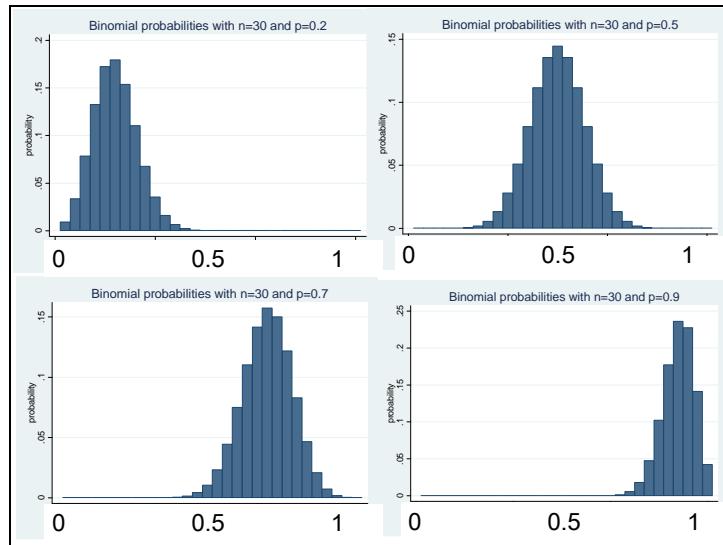


When we make  $n$  even larger, to 75, say, then it is even easier to distinguish these four cases since they are even more separated—little overlap.

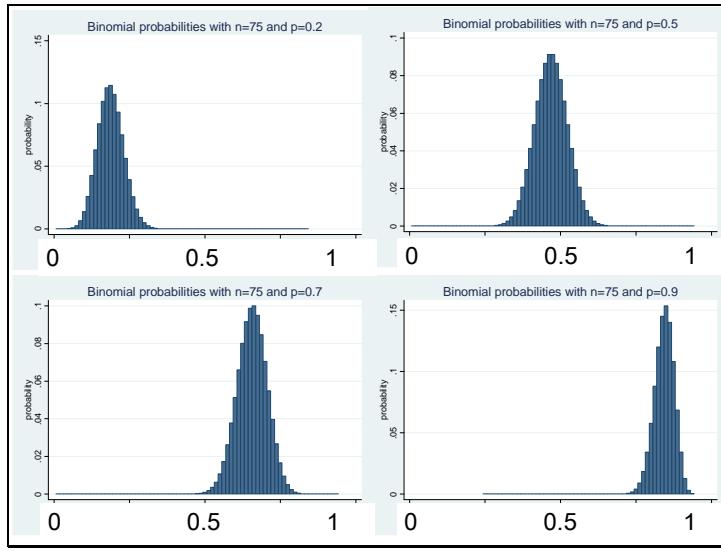
So it becomes easier to distinguish between these  $p$ 's as my sample size is bigger. You notice though that the scales on the plots change as  $n$  changes, so these comparisons are not quite fair. So let us rescale the plots to reflect not our counting the number of successes, but rather looking at the proportion of successes—just divide the total number of successes by  $n$ , the sample size. In that way our horizontal axis should now go from zero to one.



Here are the plots of the distribution of the proportion of successes when  $n=14$ .



Here are the same plots when  $n=30$ .



And here are the same plots when  $n=75$ .

### Estimator of $p$



$n$  trials,  $x$  successes  $= \sum_{i=1}^n d_i$

where  $d_i = 1$  if  $i^{\text{th}}$  trial is a success,  
 $= 0$  if  $i^{\text{th}}$  trial is a failure.

So  $\hat{p} = \frac{x}{n} = \frac{1}{n} \sum_{i=1}^n d_i$

So what we have drawn is the sampling distribution of the proportion of successes,  $\hat{p}$ , as opposed to our earlier plots of the total number of successes. Since this is the number of successes (which peaked at  $np$ ) divided by  $n$ , the number of trials, what that means is that the sampling distribution of  $\hat{p}$  will peak at  $p$ , the quantity we are trying to estimate.

We should have expected this because if we define each trial to be a  $d$  that is equal to zero or one depending on whether we had a failure or a success, then  $\hat{p}$  is simply the sample mean of

these d's, and the central limit theorem tells us that the mean of the sampling distribution of the  $\hat{p}$  is p. So on average, across the villages, we get the population value, p. So we have an *unbiased* estimator of p.

### Estimator of p



$$n \text{ trials, } x \text{ successes} = \sum_{i=1}^n d_i$$

where  $d_i = 1$  if  $i^{\text{th}}$  trial is a success,  
 $= 0$  if  $i^{\text{th}}$  trial is a failure.

So  $\hat{p} = \frac{x}{n} = \frac{1}{n} \sum_{i=1}^n d_i$

is approximately normal with mean p  
and standard deviation  $\sqrt{p(1-p)/n}$

Viewing our estimator as the sample mean, albeit of zero-one variables, allows us to appeal to the central limit theorem to say that for large sample sizes, n, this sample mean is approximately normally distributed with mean p and standard deviation (standard error)  $\sqrt{p(1-p)/n}$ .

### Standardization



So

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

is approximately standard normal.

In order to create confidence intervals or perform hypothesis testing with respect to p, we can collect all this information and create our standardized Z, as above.

What makes this standardization a little more complex than what we have become accustomed to, in our prior inference, is that the parameter of interest,  $p$ , appears both in the numerator of  $Z$  and in its denominator, whereas before we had  $\mu$ , the parameter of interest, in the numerator and a separate  $\sigma$  in the denominator.

One solution to too many appearances of, and this is the so-called Wald solution, is to argue that we are interested in the  $p$  inasmuch as it is the mean, so estimate the standard deviation (as we did before when we replaced the unknown  $\sigma$  by the sample standard deviation,  $s$ ) by replacing  $p$  where it appears in the standard deviation (denominator) by  $\hat{p}$ . This is equivalent to using the sample standard deviation to estimate the population standard deviation, just as we did before in going from the  $Z$  to the  $t$  with continuous data. This solution is not to be recommended as it can lead to problems when  $p$  is close to either boundary (zero or one).

### Confidence intervals (Wilson)



$$\Pr \left\{ -1.96 \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq 1.96 \right\} = 0.95$$

$$\Pr \left\{ \frac{(\hat{p} - p)^2}{p(1-p)/n} \leq (1.96)^2 \right\} = 0.95$$

So approximate CI, solve for  $p$ s that satisfy:

$$(\hat{p} - p)^2 \leq (1.96)^2 p(1-p) / n$$

$$(\hat{p} - p)^2 - (1.96)^2 p(1-p) / n \leq 0$$

Another approach that also utilizes the DeMoivre result is the Wilson method. This starts by simply squaring the standardized  $Z$ . This has squaring has the advantage of eliminating both the pesky square-root in the denominator and the plus, minus interval that  $Z$  needs to fall into for the confidence interval to be satisfied. We recognize that the resulting single inequality defines a parabola in  $p$  that can subsequently be solved.

Let us apply this thinking to our Framingham Heart Study data.

. summ death angina hospmi stroke cvd hyperten diabetes1					
Variable	Obs	Mean	Std. Dev.	Min	Max
death	4434	.3495715	.4768884	0	1
angina	4434	.1635092	.3698714	0	1
hospmi	4434	.1023906	.3031955	0	1
stroke	4434	.0935949	.2912972	0	1
cvd	4434	.2609382	.4391958	0	1
hyperten	4434	.7334235	.4422189	0	1
diabetes1	4434	.0272891	.162943	0	1

Here are seven dichotomous variables with their respective proportions, p's, in the population of 4,434 people in our data set. These are the parameters we would like to infer when we take a sample from this population. So, for example, we have that deaths are approximately 35% of the population. Angina is about 16%, and so on down to hypertension at 73%, and our old friend diabetes at the first visit is roughly 2.7%.

Now take a sample of size 20 from this population.

. sample 20, c (4414 observations deleted)					
Variable	Obs	Mean	Std. Err.	Wilson [95% Conf. Interval]	
death	20	.25	.0968246	.1118617	.4687009
angina	20	.1	.067082	.0278665	.3010336
hospmi	20	.15	.0798436	.0523687	.3604189
stroke	20	.05	.048734	.0088814	.2361312
cvd	20	.25	.0968246	.1118617	.4687009
hyperten	20	.9	.067082	.6989664	.9721335
diabetes1	20	0	0	0	.1611252*

(\*) The Wilson interval was clipped at the lower endpoint

Now that we have the sample, the *ci* command in Stata will give us the summaries, above, and the confidence intervals, the Wilson ones in our case because we use the Wilson option. (Do not forget the *bi* option to tell Stata we have binomial data.)

Let us look at this output from Stata. We can look at each line, one by one, starting with the first one, death. First we see that  $\hat{p}$  is 0.25 with a standard error of approximately 0.1. The

associated (Wilson) 95% confidence interval for p is (0.11, 0.47). Remember the true value of p (which we know in our exercise because we are acting as if we know the population) is 0.35. So the confidence interval did indeed cover the population value in this case.

All the others lend themselves to similar interpretations until we get to diabetes1. We have to be careful when the mean is very small, close to zero, or very large, close to one. In those cases we sometimes run into problems as we do here. Here we have that the proportion of diabetics is small, 0.027, and as a result none appeared in our sample of 20. So  $\hat{p}$  is zero and the Wilson estimator has problems. Thus the clipping message, and in this case the coverage probability may not be 95%.

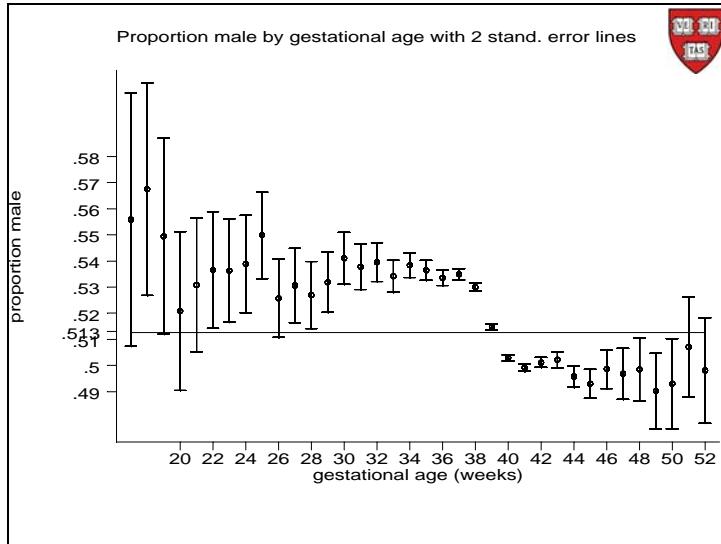


. ci death angina hospmi stroke cvd hyperten diabetes1				
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
death	20	.25	.0993399	.0420791 .4579209
angina	20	.1	.0688247	-.0440518 .2440518
hospmi	20	.15	.0819178	-.0214559 .3214559
stroke	20	.05	.05	-.0546512 .1546512
cvd	20	.25	.0993399	.0420791 .4579209
hyperten	20	.9	.0688247	.7559482 1.044052
diabetes1	20	0	0	0 0

In the *ci* command, had we not asked for the Wilson confidence interval what we would have gotten by default is the misleadingly called “exact” answer. First, we need look in the help file to see what this confidence interval really is since it is not properly labeled—we are just told that it is a/the(?) “95% Conf. Interval”.

Presumably it is called exact because it uses the exact, or model binomial sampling distribution of  $\hat{p}$ , as opposed to the DeMoivre normal approximation. In reality all these confidence intervals are approximate. Where this “exact” method does its approximating is in the confidence level—it does not give you a 95% confidence interval. What it does produce is an interval that has at least 95% confidence attached to it, but the true value may be much larger. It gives the exact answer to the wrong question.

Ordinarily, there is not much difference between these confidence intervals, especially for large sample sizes. If you need to trust one, go with the Wilson unless your n is tiny, in which case the confidence interval is not very informative, anyway.



Here is an example with huge  $n$  where I used the Wald approximation. Here we report on roughly 4 million singleton births in a year in the US.

What the circles refer to are the proportion males as a function of gestational age. The solid line at 0.513 is the overall average, ignoring gestational age. The plus/minus bars are capped off at plus or minus two standard errors. This is typically the graph one gets from year to year. It does not vary much.

The first thing we notice about this pattern is that when the interval is wide (small), then we have a small (large)  $n$ , since each  $\hat{p}$  is roughly the same size. So we see that most of us are born roughly in the 36- to 42-week interval.

Secondly, these numbers are based on birth certificates, which are largely considered administrative documents, so I do not know how much trust to have at either end of the gestation scale. So let us concentrate mostly in the 22-week to 44-week window.

What we are seeing in this window, and even beyond, is a pattern that favors males in the early gestational periods, but that switches, rather smoothly, to favoring females at the later gestational ages. Why this pattern, I do not know.

## Hypothesis Testing



$$H_0 : p = 0.082$$

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ &= \frac{0.115 - 0.082}{\sqrt{0.082(1-0.082)/52}} \\ &= 0.87 \end{aligned}$$

$$p\text{-value} = 0.384$$

Let us now investigate tests of hypotheses about the parameter  $p$  in the binomial. Once again, there are two approaches one can take. Let us first look at the central limit theorem approach, namely the normal approximation. Here the fact that  $p$  appears both in the numerator and in the denominator of the standardized  $Z$  is not an issue.

Consider a sample of 52 individuals, where 6 survived to five years post diagnosis of cancer, and we want to test the hypothesis that the survival proportion in the population is 0.082. We can set up our standardized  $Z$  and calculate it to be 0.87. Check with Stata to find out that the two-sided  $p$ -value is 0.384. So we do not reject the null-hypothesis.

## Stata Output



```
. prtesti 52 6 0.082 , count
One-sample test of proportion                                x: Number of obs =      52
                                                              
Variable          Mean    Std. Err.          [95% Conf. Interval]
x                .1153846   .0443047          .0285491   .2022202
                                                              
p = proportion(x)                                     z =      0.8774
Ho: p = 0.082
                                                              
Ha: p < 0.082          Ha: p != 0.082          Ha: p > 0.082
Pr(Z < z) = 0.8099          Pr(|Z| > |z|) = 0.3802          Pr(Z > z) = 0.1901
```

Rather than do any of these calculations ourselves, we can get Stata to do them for us if we use the command *prtesti*, as above.

The output from Stata looks very much like the output from the t-test. So that is using the normal approximation, which for large samples such as this is perfectly fine.

. bintesti 52 6 0.082									
N	Observed k	Expected k	Assumed p	Observed p					
52	6	4.264	0.08200	0.11538					
<hr/>									
Pr(k >= 6)	= 0.251946	(one-sided test)							
Pr(k <= 6)	= 0.868945	(one-sided test)							
Pr(k <= 1 or k >= 6)	= 0.317935	(two-sided test)							
<hr/>									
Ho: proportion = .082									
-- Binomial Exact --									
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]					
				<hr/>					
	52	.1153846	.0443047	.0435439	.2344114				

The other approach is to argue that we have a binomial model, so why not use the calculations appropriate for a binomial and not rely on the large sample approximations afforded us by DeMoivre's central limit theorem. You may invoke this approach by using the Stata command, *bintest*. The results of a call to the "immediate" version of it, *bintesti*, is displayed above. We see that the results here look very much what the large sample approximation showed us.

This approach is fine if you are doing a one-sided test. It is slightly controversial if you are doing a two-sided test since the sampling distribution of the test statistics is not symmetric and is discrete, so it is unclear how what mass gets lumped into the rejection region.

So there are the options for you to do hypothesis testing.

## Sample Size Estimation



Suppose we wish to test the hypothesis  $H_0 : p \leq 0.082$  at the  $\alpha = 0.01$  level, and we want power of 0.95 at  $p=0.2$ . How big a sample do we need?

For  $\alpha = 0.01$  the  $z = 2.32$ . So since

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}},$$

a  $z$  of 2.32 corresponds to a  $\hat{p}$  of :

$$\hat{p} = 0.082 + 2.32\sqrt{0.082 \times 0.918/n}$$

Just as with the continuous variables when we calculate the sample size required to achieve a particular power, so too with the discrete variables, we can perform similar calculations. So, for example, if we are testing a hypothesis that  $p$  is less than or equal to 0.082—once again, let us just investigate the one-sided test—and we want to test this at the 0.01 level. And suppose we want a power of 95% or 0.95, when  $p$  equals 0.2. So if  $p=0.2$  we want to be 95% sure we reject the null hypothesis that  $p=0.082$ .

Just like the continuous variable, we need these four quantities: what the null hypothesis is; at what level we wish to test that hypothesis; at what point of the alternative to calculate the power; and, what power do we wish to have. So since we chose the 0.01 level, that means we would look at a  $Z$  of 2.32. So let us look at our  $Z$ , and we want to have it equal to 2.32. This thus provides us a relationship between  $\hat{p}$  and  $n$ .

Now this sort of calculation is a little bit complex, and before you get too concerned, rest assured that we appeal to Stata to do the hard work for us. This is here for those who, like me, enjoy these details!

If we want to reject with probability 0.95 (when  $p=0.2$ ), then  $z = -1.645$ .



So a  $z$  of  $-1.645$  corresponds to:

$$-1.645 = \frac{\hat{p} - 0.20}{\sqrt{0.20(1-0.2)/n}}$$

$$\hat{p} = 0.20 - 1.645\sqrt{0.2 \times 0.8/n}$$

Remember

$$\hat{p} = 0.082 + 2.32\sqrt{0.082 \times 0.918/n}$$

So  $n = 120.4$ . Thus round off to  $n=121$ .

Now turn attention to the power. When  $p$  is equal to 0.2, we want to have a power of 0.95. That means that  $Z$  must equal -1.645 when  $p=0.2$ . That provides us a second relationship between  $\hat{p}$  and  $n$ .

Solving these two equations relating  $\hat{p}$  and  $n$  we get that  $n=120.4$ . Rounding off gives us that an  $n$  of 121 will satisfy the alpha and power requirements we require.

**Stata:**

```
.sampsiz 0.082 0.2 , alpha(0.01) power(.95) onesamp oneside
```

Estimated sample size for one-sample comparison of proportion  
to hypothesized value

Test Ho:  $p = 0.0820$ , where  $p$  is the proportion in the population

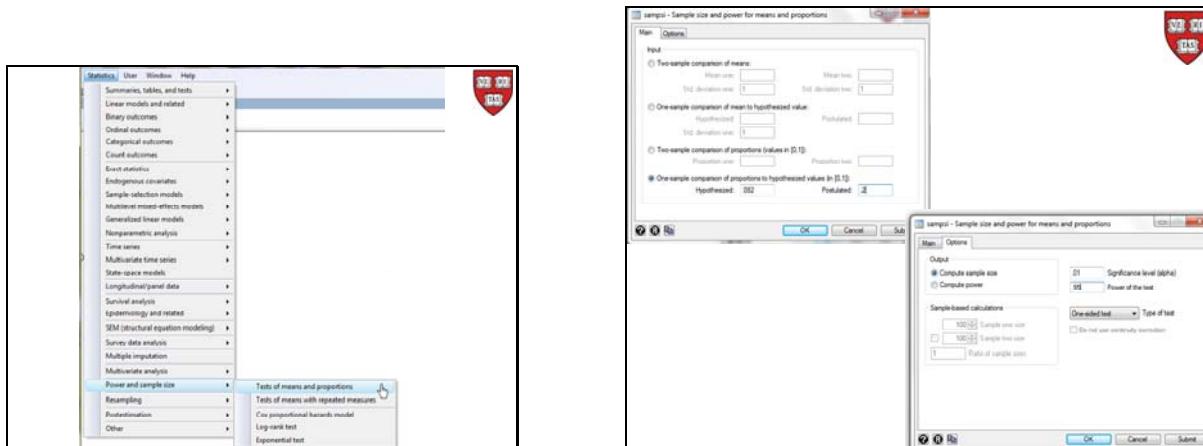
Assumptions:

alpha = 0.0100 (one-sided)  
power = 0.9500  
alternative p = 0.2000

Estimated required sample size:

n = 121

The Stata command that will do all these calculations for us is *sampsiz*. You can type it in, or you can use the pull down menus:



First click on Statistics, then on Power and sample size, and then on Tests of means and proportions. Go down to the One-sample comparison of proportions, and enter the hypothesized 0.082. The “postulated” slot is where you place the 0.2, namely the value of  $p$  where you wish to calculate the power.

Then you click on the tab labeled Options. This next menu is where you fill in your alpha level, the power you wish, and whether you want to use a one-sided or two-sided test. Then clicking Submit will get Stata to calculate the sample size for you. Alternatively you can click on

“Calculate the power” in the top left-hand corner and give Stata a sample size and it will calculate the power for you.

## Two-sample situation

Now let us look at the two-sample p. What do I mean by that? What this means is we have two populations, with population I having prevalence  $p_1$ , and population II having prevalence  $p_2$ .

We take a sample from each of these two populations and on the basis of these samples, make inference about the relative sizes of  $p_1$  and  $p_2$ . The primary hypothesis of interest is that these two are equal, but others can be entertained.

There are two ways to proceed: One is to look at their differences—so look at  $p_1 - p_2$ , and ask is this difference equal to zero, which is exactly what we did with the two-mean situation in the continuous case. I am going to leave this approach for you to explore. The Stata programs are all there.

Another approach this is to look at their ratio; namely  $p_1/p_2$ , or the *relative risk*. In this approach you need be careful when  $p_2$  is equal to 0. The “null” value would be that this ratio is one.

Of course, once we think of ratios we think of our friend the odds. Because of the relationship between probabilities and odds, we know that saying the relative risk is one is equivalent to saying that the relative odds, or the odds ratio, is one.

The two approaches are equivalent as far as testing the null hypothesis that the ratio is one, but if one approach has better statistical properties, then by all means follow that approach. And that is why we look at the odds ratio in this situation.

## Review of Odds

Odds

pp 144 et seq.



If the probability of an event A is p, then the odds of the event are  $p/(1-p)$ , or  $p : (1-p)$

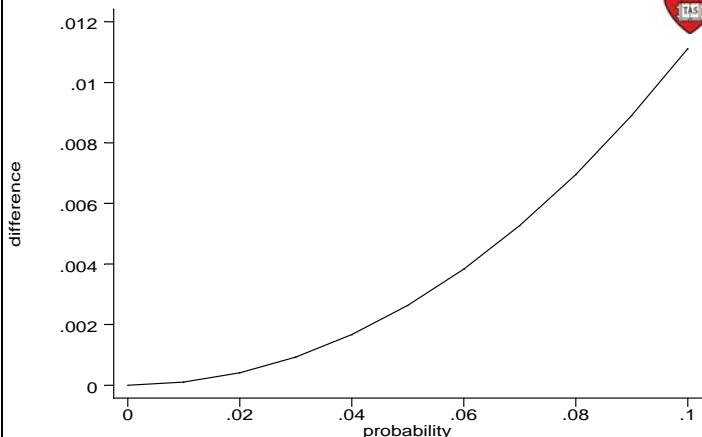
If  $p$  is small then odds  $\approx p$  :

$$\frac{p}{1-p} \approx p(1+p) \text{ for } p \approx 0$$

Let us take a quick refresher on relative odds. Remember, if the probability of an event A is p, then the odds of the event are p over  $1 - p$ .

Sometimes you see this stated as “p to  $1 - p$ ”, or the ratio of p to  $1 - p$ . You will also in the literature that if p is tiny, then the odds are approximately the same as the probability, p. And that is because if you are going to expand one over  $1 - p$  for small p, then that is approximately the same as  $1 + p$ . And so the odds are approximately equal to p, because you leave out the  $p^2$  term.

Graph of the odds of an event minus the probability of the event



Plotted above is the difference between the odds of an event and its probability, when the probability is less than 0.1. You can see that the difference is sizable on the right, but very small on the left.



## Odds versus probability

Probability	Odds	Odds
0	0	0
1/4	1/3	1 : 3
1/3	1/2	1 : 2
1/2	1	1 : 1
2/3	2	2 : 1
3/4	3	3 : 1
1	$\infty$	$\infty$

And just to remind you, if the probability is 0 the odds are 0. If the probability is 1, the odds are infinite. With probability smaller than  $\frac{1}{2}$ , the corresponding odds are less than 1. With probability  $\frac{1}{2}$  the odds are one, and with probability greater than  $\frac{1}{2}$ , the odds are bigger than 1. So, when the event is less likely to happen than not happen, the odds are less than one. If the event is as likely to happen as not, the odds are one. If the event is more likely to happen as not, the odds are greater than one.

Sometimes you see odds as stated in this way, other times they are stated as a ratio. For example, odds of 1/3 are sometimes stated as 1 to 3. Odds of 1/2 are 1 to 2. Evens, 1 to 1. When you get above 1, then it's 2 to 1, 3 to 1, instead of 2 it's 2 to 1, 3 to 1, cetera. So this just depends on the culture you are in.



## Relative Odds or Odds Ratio

Suppose we have a disease  
(e.g. lung cancer)

And two groups  
(e.g. smokers, non-smokers)

Relative odds (OR)

$$= \frac{P(D|S)}{1-P(D|S)} / \frac{P(D|S^c)}{1-P(D|S^c)}$$

D ≡ disease S ≡ smoker

$S^c$  ≡ non-smoker

Now here is the real strength of the odds, and that is when you look at the odds ratio or relative odds. They are closely related to our use of Bayes' theorem: Suppose we have a disease such as lung cancer. And we have got two groups of people, smokers and nonsmokers. Now what are the relative odds of the disease, for smokers versus nonsmokers? So we first find the odds of the disease for smokers. Then do the same for non-smokers. Then take the ratio of these two odds.

So that's the relative odds of the disease for smokers relative to nonsmokers.

### Symmetry of odds ratio



From Bayes theorem:

$$P(D|S) = \frac{P(D)P(S|D)}{P(D)P(S|D) + P(D^c)P(S|D^c)}$$

So odds of disease for smokers:

$$\frac{P(D|S)}{1 - P(D|S)} = \frac{P(D)P(S|D)}{P(D^c)P(S|D^c)}$$

So odds ratio of disease, smokers to non-smokers

$$\begin{aligned} OR &= \frac{P(D|S)}{1 - P(D|S)} / \frac{P(D|S^c)}{1 - P(D|S^c)} \\ &= \frac{P(D)P(S|D)}{P(D^c)P(S|D^c)} / \frac{P(D)P(S^c|D)}{P(D^c)P(S^c|D^c)} \end{aligned}$$

Now recall Bayes' theorem, first for smokers and then for non-smokers. We see that we can write the odds of the disease for smokers as a ratio where the numerator is the product of the probability of having the disease and the probability of being a smoker, given that one has the disease. Similarly, the denominator is the same product but for those without the disease.

After developing a similar expression for the odds of disease for the non-smokers, we can take their ratio, and simplify by canceling some common terms.



## Symmetry of odds ratio

$$\begin{aligned} \text{OR} &= \frac{P(S|D)}{P(S|D^c)} / \frac{P(S^c|D)}{P(S^c|D^c)} \\ &= \frac{P(S|D)}{P(S^c|D)} / \frac{P(S|D^c)}{P(S^c|D^c)} \\ &= \frac{P(S|D)}{1-P(S|D)} / \frac{P(S|D^c)}{1-P(S|D^c)} \end{aligned}$$

But this is the odds ratio of being a smoker, for diseased versus non-diseased.

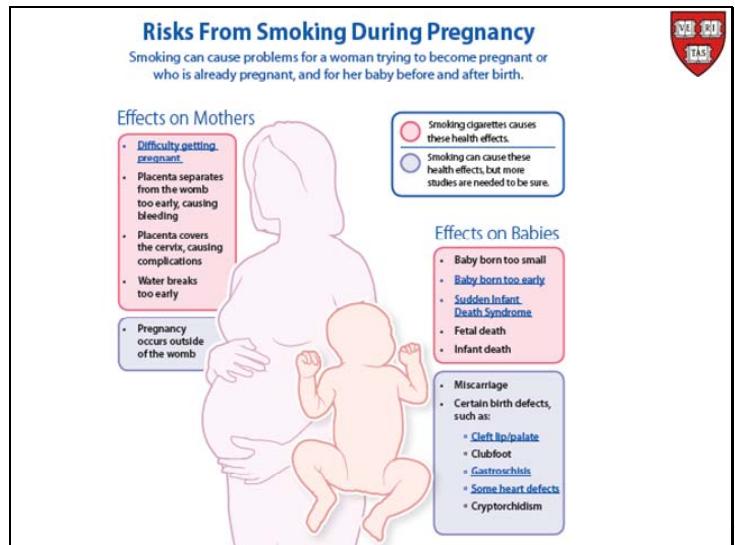
After rearranging the terms we recognize that the odds ratio can also be written as the ratio of the odds of being a smoker, given one has the disease, and the odds that one is a smoker amongst those non-diseased!

This rather surprising reversal says that the relative odds of disease amongst smokers and non-smokers, is exactly the same as the odds of smoking amongst those diseased and non-diseased. So if you are interested in the relative odds of getting the disease by smoking, then that is exactly the same as the relative odds of smoking and having the disease. So sampling amongst smokers and non-smokers and then determining their disease status in time, is the same as sampling amongst diseased and non-diseased individuals and determining their smoking status.

That means you can follow smokers over a lifetime to see what the odds are of getting lung cancer. Do the same for non-smokers. Take the ratio of those two odds to see the relative odds. Alternatively, find a group inflicted with lung cancer and see what the odds are that they are smokers. Do the same with a comparable group of individuals who do not have lung cancer. Calculate the relative odds for these two groups, and the Bayes' theorem tells us that these two relative odds are the same.

That is the theory behind case control studies.

## Case Control Example



1

Let us look at an example of a case control study that deals with the risks of smoking during pregnancy. This poster is from the Centers for Disease Control, and is part of their campaign to alert pregnant women about the ill effects of smoking during pregnancy. It lists the effects on the mother, on the left, and the effects on babies, on the right, and none of these look very good for you.

One outcome I want you to pay attention to is fetal death. All these side effects are serious, of course, but fetal death is particularly pertinent to the study at hand:

---

<sup>1</sup> <http://www.cdc.gov/reproductivehealth/TobaccoUsePregnancy/PDF/SmokingPregRisk.pdf>

## Preeclampsia



Alternative names: pregnancy-induced hypertension; toxemia

### Definition:

The development of swelling, elevated blood pressure, and protein in the urine during pregnancy.

### Causes, incidence, and risk factors:

The exact cause of preeclampsia has not been identified.

Numerous theories of potential causes exist, including genetic, dietary, vascular (blood vessel) and neurological factors. None of the theories have yet been proven. Preeclampsia occurs in approximately 5% of all pregnancies. Increased risk is associated with first pregnancies, teenage mothers, or mothers more than 40 years old, African-American women, multiple pregnancies, and women with a past history of diabetes, hypertension, or kidney disease.

The study deals with preeclampsia, also known as toxemia, so let me define it briefly for those of you who do not know what it is. First, the condition does not occur too often. It has a prevalence of only about 5% of all pregnancies. It typically happens late in the pregnancy, and is very dangerous. Indeed, the only way of saving the mother's life is to deliver the baby immediately; typically with a cesarian-section.

### "Urinary cotinine concentration confirms the reduced risk of preeclampsia with tobacco exposure"



K.Lain, R.Powers, M.Krohn, R.Ness, et al  
Am.J.Obs & Gyn 1999; 181 (5)(Nov):1192-1196.

50 women with preeclampsia (>35 weeks gestage)

matched with 50 controls (gestage, date, & BMI)

Assayed urine for cotinine.

35 patients had detectable cotinine levels:

11 (22%) of women with preeclampsia &

24 (48%) of control women.

Let us look at the study called, "Urinary cotinine concentration confirms the reduced risk of preeclampsia with tobacco exposure." The reason it attracted my attention is the rather surprising claim in the title that there might be a beneficial outcome to smoking during pregnancy.

Remember the cotinine concentration. Cotinine is a metabolite for nicotine, and so mothers with high cotinine in their urine presumably can be identified as being smokers, or ingesting nicotine.

The study is a case control study, and they took 50 women with preeclampsia—here it was determined at more than 35 weeks of gestational age—and they matched those 50 women with 50 controls. And they were matched on gestational age, on date, and BMI, Body Mass Index.

They found that there were 35 patients of the 100, with cotinine in the urine, and how they split up is very interesting: 11 of the women with preeclampsia, 50, had high cotinine, whereas 24 of the control had high cotinine.

So just looking at this you say, wait a minute, there are many more smokers amongst the control group, in fact more than twice as many. So it looks like, indeed, the title is correct.

### Example cont.



Odds of smoking, toxemic women is:

$$\frac{11}{50} / \frac{39}{50} = \frac{11}{39}$$

Odds of smoking, control women is:

$$\frac{24}{50} / \frac{26}{50} = \frac{24}{26}$$

So

$$OR = \frac{24 \times 39}{26 \times 11} = 3.27$$

Let us calculate the odds ratio, and it turns out to be 3.27. So we have more than a tripling of the odds for the toxemic women. What is going on?



Neyman Fallacy

Prevalence-incidence bias

We came in at 35 weeks gestation to compare smokers to non-smokers. What happened in the first 34 weeks of gestation?  
Would that be relevant?

We need to be careful when considering case control studies that evolve over time where the condition for entry into the study might be related to a measure associated with the study. One of the most important problems you can have is what is called the Neyman Fallacy. It is also called the Prevalence-incidence bias. It's very much like the healthy worker effect.

Entry into the study required that a woman be pregnant for 35 weeks, whence they compared smokers to non-smokers. But what about what happened in the first 34 weeks of gestation? Could that be relevant? Remember, the CDC warned about fetal deaths being attributed to smoking. So if you took 50 women who smoke at the beginning of their pregnancy, and took 50 comparable women who did not smoke, at the beginning of their pregnancy, then after 35 weeks would you still have two groups of 50? Chances are, you would not. So looking for the effects of smoking at the tail end of the pregnancy, when in fact it has had an effect throughout the pregnancy, is misleading. Just as it would have been in the Pearl smoking and longevity study to only look at the three groups he studied, after they turned 70. This is a nonsense study.

**Discrete Outcomes**



Consider whether electronic fetal monitoring (EFM) has an impact on caesarean decision.

Sample 5,824 deliveries:  
of these 2,850 were EFM exposed and 2,974 were not.

358 of the 2,850 had c-sections as did 229 of the 2,974.

Binomial with n huge.

Returning to our primary aim, which is to look at discrete outcomes, let us start with whether two binary outcomes are related. Let me introduce this topic with an example.

Consider whether Electronic Fetal Monitoring, let's call it EFM, has an impact on the decision to have a Cesarean section, not. In this study they looked at 5,824 deliveries. Of these, in roughly half, 2,850, the women were subjected to EFM, and in other roughly half, 2,974, they were not.

Now consider how many of the women had c-sections: 358 of the 2,850 who had EFM had a c-section, and of the remaining 2,974 who were not exposed to EFM, 229 had a c-section.

So we could treat this as two binomials and use the methods already developed, but the  $n$  is exceptionally huge, so let us explore another avenue, one that is impervious to large  $n$ , and one that is more easily generalizable to the situation when we have more than two groups and also when we deal with nominal variables that take on more than two possible values.



Do the two exposures differ?

### Chi square test

Proceed as usual:

1. If there is no difference (null hypothesis) what do we expect to see?
2. How does this compare to what we have observed? (statistic & its distribution)

The next test we look at is the very popular chi square test. It proceeds very much like all the tests we have encountered so far: You start with an if statement established by your null-hypothesis. Then you ask, if the null hypothesis is true—for example, that there is no difference between those who were EFM exposed and those who were not, if that is our null our null hypothesis—then what do we expect to see? What did we see? Compare the two.

We are going to have to come up with a statistic that allows us to compare the two, and establish how this statistic varies from sample to sample—in other words, determine the sampling distribution of the statistic. In other words, the same approach we have repeatedly in this course.



### Data-Contingency table

Caesarean Delivery	EFM Exposure		Total
	Yes	No	
Yes	358	229	587
No	2,492	2,745	5,237
Total	2,850	2,974	5,824

If the c-section rate is unaffected by EFM exposure, then ignore column classification and go with totals.

Let us go back to our numbers, displayed above in a two by two table. If the null-hypothesis that the c-section rate is the same for those exposed to EFM as it is for those not exposed to EFM, then the 587 with c-sections should be distributed amongst the two groups in the same ratio as the membership in the two groups; namely in the ratio 2850 : 2974. Since 587/5824 is approximately 10%, we expect 10% of the 2850 and 10% of the 2974 to have had c-sections. This is ideally what to expect, but we also expect some sampling variability around this ideal. How to account for that?



```
. tab csec efm
```

csec	efm		Total
	no	yes	
no	2,745	2,492	5,237
yes	229	358	587
Total	2,974	2,850	5,824

So for example when we go and we get Stata to do these calculations for us,



```
. tab csec efm , row col
```

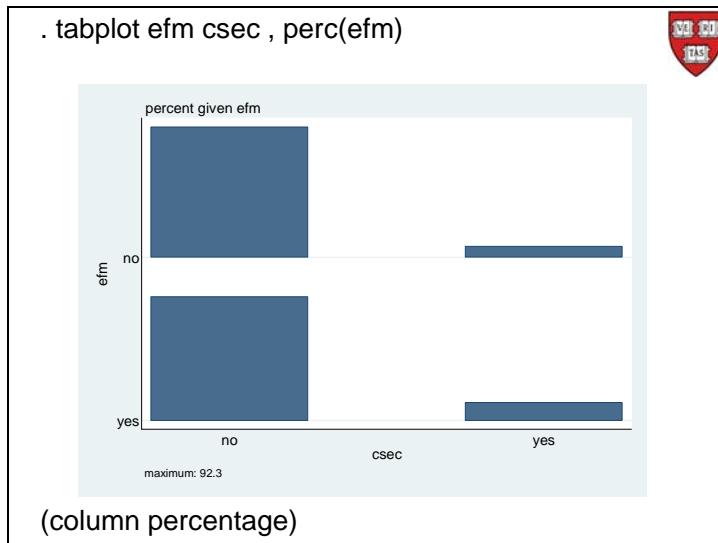
Key			
frequency			
row percentage			
column percentage			
Key			
frequency			
row percentage			
column percentage			
csec	efm		
	no	yes	Total
no	2,745	2,492	5,237
	52.42	47.58	100.00
	92.30	87.44	89.92
yes	229	358	587
	39.01	60.99	100.00
	7.70	12.56	10.08
Total	2,974	2,850	5,824
	51.06	48.94	100.00
	100.00	100.00	100.00

we can ask Stata to show us the actual frequency, that's just like the last slide, but also show me the row percentages. So for example, 52.42% of the no c-section did not have EFM, and the other 47% did have EFM.

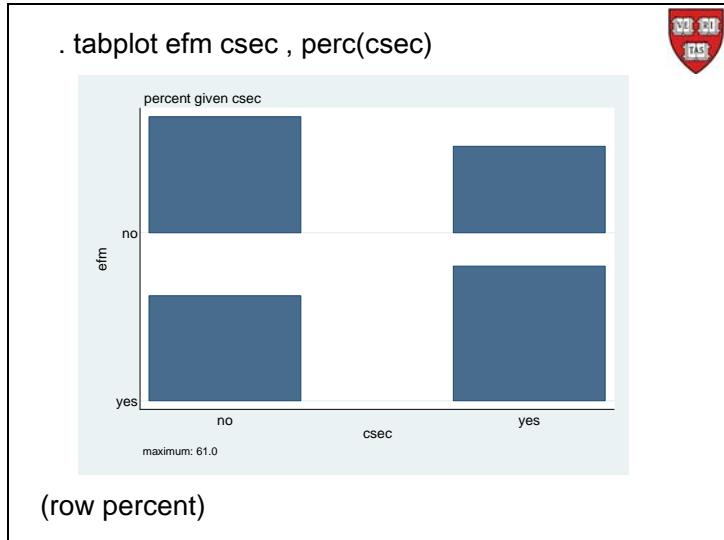
Inspecting this table by row, we see that of those who did have a c-section, 39% did not have EFM exposure, and 61% did. This distribution is suspiciously different from the overall tally where 51% had EFM exposure, and 48.94% did not—by design, this was supposed to be 50-50.

Alternatively we can look at this table by column. We see from the margin that about 10% had c-sections. In those not exposed to EFM we see that only 7.7% had c-sections, in contrast to those in the exposed to EFM column where 12.56% had c-sections.

This table does not look balanced. It seems like there is a relationship between the row classification and the column classification. Is this difference attributable to chance alone?



Another way of looking at this table is to plot it. You can use the *tabplot* command, to draw a bar graph for the cell counts. Here I have used the percentage EFM option; that means that we condition on EFM. So it is like asking Stata for the row percentage. So horizontally these should sum up to 100%. It looks like, the bottom right cell is a little bit thicker than the one above, in other words these two patterns do not seem to be parallel to each other.



We could have taken the column percentages, instead. We now see the lack of column parallelism much more evident than before.

### Probability of c-section

From the totals we can estimate:

$$\Pr\{\text{c-section}\} = \frac{587}{5,824} = 0.101$$

$$\Pr\{\text{no c-section}\} = \frac{5,237}{5,824} = 0.899$$

To judge whether we can attribute the evident differences between what we expected to see (parallelism) and what we saw, let us return to the numbers to quantify our expectations under the null: From the totals we can estimate, as we did before, the probability of a c-section—and that is 0.101, or roughly 10%. That means that the probability of no c-section is 0.899. So let's apply this probability of c-section to each of the two groups, those who had the EFM and those who did not.

## Expectations



What do we expect to see if EFM has no effect?

EFM exposed (2,850 mothers):

Expect:  $0.101 \times 2850 = 287$  c-secs  
and:  $0.899 \times 2850 = 2563$  vaginal

No EFM (2,974 mothers)

Expect:  $0.101 \times 2974 = 300$  c-secs  
and:  $0.989 \times 2974 = 2674$  vaginal

So what do we expect to see if EFM has no effect? Well then we'd expect that of the EFM exposed mothers, all 2,850 of them, roughly 10% should have a c-section. So we expect 0.101 times the 2850. So that's 287. So we expect that 287 of these mothers would have had a c-section. And that means that roughly 90% percent of them, 2850, which gives us 2563, would have had a vaginal delivery.

Let us do the same thing now for the mothers who were not EFM exposed. So of these 2974, roughly 10% percent of them, which comes out to be 300, we expect to have had c-sections, and the other 2,674 to have had vaginal deliveries.

## Contingency table



Expected, if independence of row and column classification is true, in boxes:

C-sect	EFM Exposure?		Total
	Yes	No	
Yes	358	287	587
No	2492	2563	5237
Total	2850	2974	5824

Let us gather these numbers and place them in boxes within our two by two table so we can contrast what we expect with what we actually saw.

Within each cell we see that actually observed. We see that in the yes-yes cell (top left hand corner) we expected to see 287, but we actually saw many more, namely 358. Since across the table, for each row and column, the sum of the observed and the sum of the expecteds is the same, if one expected value is too large, that means that the entry in the next row, or column, must be too small.

### Symmetry



Note that we could have worked on the rows instead of the columns and gotten the same results:

$$\Pr\{\text{EFM}\} = \frac{2850}{5824} = .489$$

$$\Pr\{\text{no EFM}\} = \frac{2974}{5824} = .511$$

Note we could have worked along the rows instead of the columns and gotten the same results. We could have argued, for example, that 48.9% of all the women were subjected to EFM.

### Expectations



Of the c-sections (587 mothers):

Expect:  $.489 \times 587 = 287$  had EFM

and:  $.511 \times 587 = 300$  no EFM

Of vaginal deliveries (5,237 mothers)

Expect:  $.489 \times 5237 = 2563$  EFM

and:  $.511 \times 5237 = 2674$  no EFM

So applying the probability of EFM to the mothers who had c-sections, we expect that 48.9% of these 587, or 287 had an EFM, and the remaining 300 not to have had EFM. Similarly, of the 5,237 who had vaginal deliveries, 48.9%, or 2,563, to have been exposed to EFM, and the other 2674, to have not been exposed to EFM.

This set of expected numbers is exactly what we calculated above, and if there is no relationship between the row and column classifications, then this is what we expect to see.

Contingency table					
		Expected, if independence of row and column classification is true, in boxes:			
C-sect	EFM Exposure?			Total	
	Yes	No	Total		
Yes	358	287	229	300	587
No	2492	2563	2745	2674	5237
Total	2850		2974		5824

To compare what we have seen to what we expected under the null, we can attempt to do what we did before, but we have four contrasts to make here, whereas in the past (with the mean, for example) we only had one comparison to make. We could take the four differences and sum them up, but when we do that we always get zero; just like when we tried to average out all the distances from the mean. In that case we were led to the standard deviation by taking the squares of the deviations and we can do the same thing here.

A problem with this approach is that the cells are not the same size to begin with. A discrepancy of 10 observations in the first row is of much more consequence, when comparing expected to observed, in a row where the average size is 300, than a discrepancy of ten in the second row where the average cell size is about 2,600. So we can standardize before summing by dividing by the expected value in the cell.



## Chi Square Goodness of fit

(Table page A-26)

$$x^2 = \sum_{\text{cells}} \left\{ \frac{(obs - exp)^2}{exp} \right\}$$
$$d.f. = (\# \text{ rows} - 1)(\# \text{ columns} - 1)$$

The  $X^2$  statistic, introduced by Pearson, aggregates, over all the cells, the observed minus the expected squared, divided by the expected. Its sampling distribution is approximated by a Chi-squared distribution. The surprising thing is that this is almost independent of the number of observations we have. It is also quite robust to the sampling plan—for example, it works here even though we chose 50% of the women to have EFM, and it would have worked the same whether we had just spun a coin for each woman to determine whether she gets EFM or not.

And just as with the t, we need degrees of freedom. And degrees of freedom are calculated as the number of rows minus 1 times the number of columns minus 1. In this case it was a 2x2 table, so it's 2 minus 1 times 2 minus 1, which is equal to 1.

Intuitively, the reason why this is so is because once you fix one of the cell numbers, the other three are available by subtraction. So you have only one degree.

## Continuity correction factor



In 2x2 tables (only) we apply  
a continuity correction factor:

$$x^2 = \sum_{\text{cells}} \left\{ \frac{(|obs - exp| - 0.5)^2}{exp} \right\}$$
$$d.f. = (2 - 1)(2 - 1) = 1$$

There is an exception to the above rule. In the single instance when we have a 2x2 table, as opposed to the forthcoming bigger tables, we get a better approximation to the sampling distribution of the  $\chi^2$  statistic if we make a so called *continuity correction*. Before squaring each cell, we decrease the absolute difference within the cell by 0.5.

### Example

For the EFM and c-section example, above:

$$\begin{aligned}\chi^2 &= \frac{(|358 - 287| - .5)^2}{287} + \frac{(|229 - 300| - .5)^2}{300} + \\ &\quad \frac{(|2492 - 2563| - .5)^2}{2563} + \frac{(|2745 - 2674| - .5)^2}{2674} \\ &= 37.95\end{aligned}$$

$$\chi^2_{1,0.001} = 10.83 \Rightarrow p\text{-value} < 0.001$$

So here is a sample calculation one would do for the current study. The p-value is less than 0.05, so we reject the null hypothesis that there is no relation between the row and column classifications. That means that we conclude that it is unlikely that the discrepancy between the observed and expected happened just by chance and that we believe that the EFM caused more c-sections to be performed than would have been done without the EFM.

### Stata output:

	Exposed	Unexposed	Total	Proportion Exposed
Cases	358	229	587	0.609
Controls	2492	2745	5237	0.4758
Total	2850	2974	5824	0.4894
	Point estimate			[95% Conf. Interval]
Odds ratio	1.722035		1.446551 2.049976 (Cornfield)	
Attr. frac. ex.	.4192916		.3087003 .5121894 (Cornfield)	
Attr. frac. pop	.2557178			
chi2(1) = 37.95 Pr>chi2 = 0.0000				

If we go straight to Stata, we would use the `cc` command (case control). We see that this agrees with our hand calculation.

## Rxc Tables



. tab diabetes1 sex1 , chi col			
Key			
frequency			
column percentage			
Diabetic, exam 1	Sex, exam 1		
	Male	Female	Total
No	1,885 96.97	2,428 97.51	4,313 97.27
Yes	59 3.03	62 2.49	121 2.73
Total	1,944 100.00	2,490 100.00	4,434 100.00

Pearson chi2(1) = 1.2217 Pr = 0.269

Returning to our Framingham Heart Study we can look at the relationship between diabetes at visit one and the subjects' sex. The above analysis says that there does not appear to be any relationship and we do not reject the null hypothesis of independence of the row and column classifications.

The nice thing about the chi squared test is that we can extend it to more than two by two tables. In fact, we can extend it to, as we say in the jargon, r by c tables; that means any number of rows (r) and any number of columns (c).

. tab diabetes3 sex1 , chi col



Key			
frequency column percentage			
Diabetic, exam 3	Sex, exam 1		
	Male	Female	Total
No	1,267 91.35	1,742 92.86	3,009 92.22
Yes	120 8.65	134 7.14	254 7.78
Total	1,387 100.00	1,876 100.00	3,263 100.00
Pearson chi2(1) = 2.5293 Pr = 0.112			

Consider extending this very same example by looking at the diabetes status at the third exam. We get that the p-value is 0.112, so at first blush we do not reject the null hypothesis that by the time they did the exam 3 that these two classifiers, diabetes status and sex, are still independent of each other.

But recall that for a longitudinal study (one that progresses over time) we have to take care especially if there are subjects dropping out of the study. Are we still comparing the same group at visit 3 as we were at visit 1? Well at visit 3 we have 3,263 in contrast to the 4,434 we had at visit 1. So we have lost 1,171 people. We have lost a considerable number of people.

If these 1,171 were lost at random, then we might not be concerned. The problem arises if there is any relationship between the classifiers under study; diabetic status and sex. If there were a relationship, we might now get a distorted impression by visit 3.

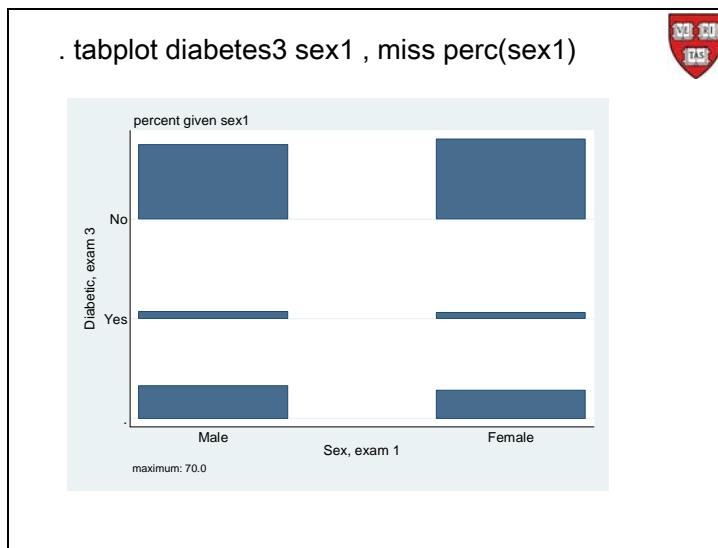
. tab diabetes3 sex1 , chi col miss



Key			
frequency column percentage			
Diabetic, exam 3	Sex, exam 1		
	Male	Female	Total
No	1,267 65.17	1,742 69.96	3,009 67.86
Yes	120 6.17	134 5.38	254 5.73
.	557 28.65	614 24.66	1,171 26.41
Total	1,944 100.00	2,490 100.00	4,434 100.00
Pearson chi2(2) = 11.4694 Pr = 0.003			

If we ask Stata to include the missings by appending the option *miss* in our tab command we get the above output.

Now that the missing 1,171 are included in our analysis, we see that there is a 4% differential between the male missing and the female missing. In fact, if we look at the distribution down the two columns we see a difference. The chiu-squared p-value is 0.003, so we would reject the null hypothesis that the row classification and the column classification are independent of each other.



If we *tabplot* this table, we see the above. We see a larger proportion males missing than females. These two columns are not parallel. Why? What is happening? I leave it for you to find out. I have no answer. But the statistic here is telling us that something is going on that might be of interest.

**r x c Tables**

e.g. Accuracy of Death Certificates

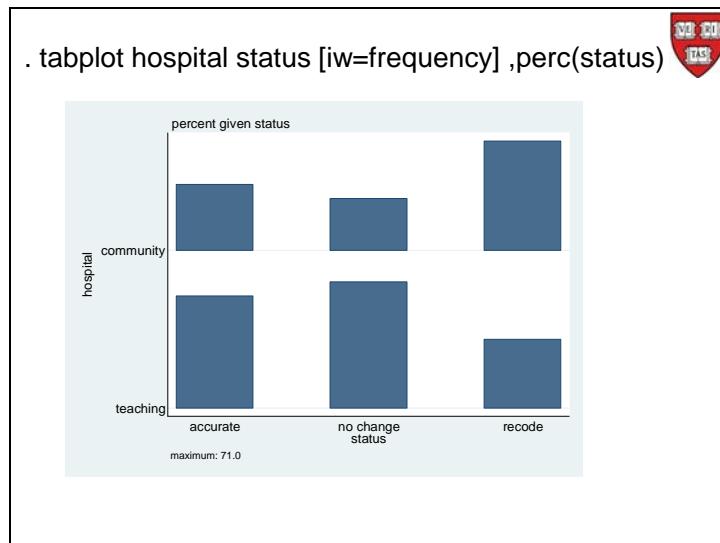
---

		Certificate Status			Total
Hospit.	Conf.	Inacc.	Incorr.		
		Accur.	No Ch.	Recode	
Comm.	157	18	54		229
Teach.	268	44	34		346
Total	425	62	88		575

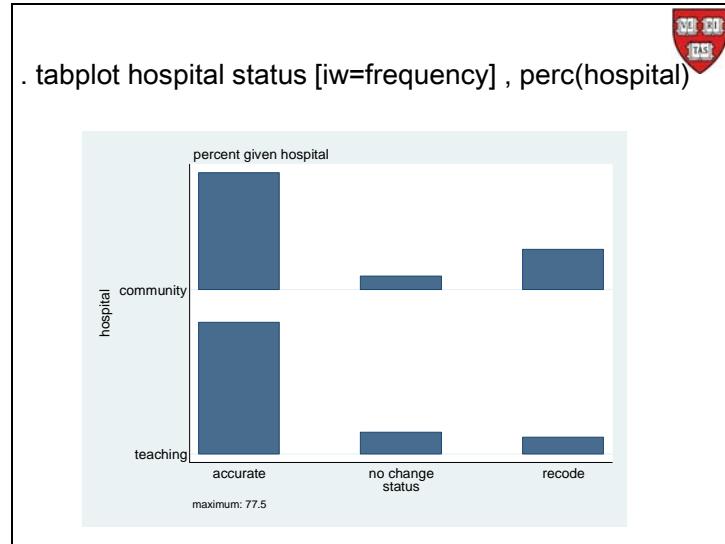
Here is another example. It approaches another, general interesting question. How accurate are death certificates? This study looked at 575 death certificates. And the row classification is whether the hospital that completed the death certificate was a community hospital or whether it was a teaching hospital. I believe this was in the state of Connecticut.

Now of these 575, 229 came from the community hospital and 346 came from the teaching hospitals. So roughly a third came from the community hospitals and 2/3 from the teaching hospitals. They compared these 575 death certificates to the medical records to check for any inaccuracies and found that of 425 of the 575 were accurately filled in. Of the remaining ones that had inaccuracies they classified them into two groups, those where the inaccuracies were not serious enough to require recoding of the death certificates (62 certificates), and those that did require a recoding of the death certificates (88 certificates).

Now the null hypothesis is that the row classification is independent of the column classification. That means that the accuracy with which these certificates were recorded was independent of the type of hospital. So since roughly a third of the certificates came from the community hospitals, that one-third ratio should maintain in each of the three columns, under the null hypothesis.



Now if we look at the tabplot for this table with the `perc(status)` option, so that the column sums are 100%, we see some variability around this roughly one-third ratio; the largest deviation is in the “recode” column. These two rows of bar graphs do not seem to be parallel.



We can look at it in the other direction, namely with the `perc(hospital)` option so that the row sums are 100%, again we see that the two rows of bar graphs are not parallel.

Certificate Status				
Hospital	Confirmed Accurate	Inaccurate No Change	Incorrect Recoded	Total
Comm.	157 169.3	18 24.7	54 35.0	229
Teach.	268 255.7	44 37.3	34 53.0	346
Total	425	62	88	575

$\chi^2 = 21.62$   
 $d.f. = (2-1)(3-1) \Rightarrow p\text{-value} < 0.001$

tabi 157 18 54 \ 268 44 34

Our suspicions are confirmed when we submit the table to Stata to find that the p-value associated with the  $\chi^2$  is less than 0.001. So we reject the null hypothesis that the row and column classifiers are independent of each other.

We thus see that the  $\chi^2$  statistic extends simply to the larger tables.

## McNemar Test

McNemar's Test				
Paired Dichotomies				
e.g. Pairs matched on age & sex:				
Diabetes	M.I.	Total		
	Yes	No		
Yes	46	25	71	
No	98	119	217	
Total	144	144	288	

An important two by two table is when the data actually represent matched pairs. For example, this study matched 144 pairs of individuals on their diabetes status and on whether they had suffered a myocardial infarction or not. They were matched on age and sex. So we have 288 observations, true, but they should more properly be analyzed as 144 couplets.

The test designed for such an analysis is called the McNemar Test. This is the discrete analog of what we did with the two sample, dependent t.

Table			
M.I.	No M.I.	Total	
Diabetes	Diabetes	No Diabetes	
Diabetes	9	37	46
No Diabetes	15	82	98
Total	25	119	144

We reclassify the 144 couplets by registering the status of each pair according to the diabetes status of the MI individual versus the diabetes status of the member of the pair without MI. The

argument then goes, if we are interested whether there is a relationship between MI and diabetes, then consider the diagonal elements in the above table. If both members of the couple have diabetes, or neither do, then that is not going to provide any information about the relationship between MI and diabetes. So the diagonal elements are non-informative and we can discard them.

Now consider the off diagonal cells. If those with diabetes but no MI are the same in number as those with MI but no diabetes, then there is no relationship between MI and diabetes. SO we can test for a relationship by looking at these off diagonal cells and test whether it is plausible to think that their difference can be attributed purely to chance. That is the McNemar test.

Some find it troubling that we ignore the diagonal cells, feeling somehow that there is some information there; for example you could have a zillion on this diagonal and 37 and 16 in the off diagonal cells, as we have here, and the McNemar test would give the same answer we are about to calculate.

## Chi-squared



Discordant entries: 37 & 16

$$X^2 = \frac{[|37 - 16| - 1]^2}{37 + 16}$$

$$= 7.55$$

$$X^2_{1,010} = 6.63$$

$$X^2_{1,001} = 10.83$$

$$.001 < p < .010$$

Stata ignores the correction factor, 1

Returning to the McNemar test, we see the formula above. The 1 is a continuity correction that Stata ignores. We then compare this  $X^2$  to a Chi-square with one degree of freedom to see that the p-value is less than 0.05, so we reject the null hypothesis, and conclude that there is a relationship between diabetes and MI.

**Stata:**

		Controls		Total
Cases		Exposed	Unexposed	
Exposed		9	37	46
Unexposed		16	82	98
Total		25	119	144

McNemar's chi2(1) = 8.32 Pr>chi2 = 0.0039  
Exact McNemar significance probability = 0.0055  
Proportion with factor  
Cases .3194444  
Controls .1736111 [95% conf. interval]  
difference .1458333 .0427057 .2489609  
ratio 1.84 1.208045 2.802546  
rel. diff. .1764706 .0676581 .285283  
odds ratio 2.3125 1.25512 4.45228 (exact)



This is the Stata output. They get 8.32 for the statistic, instead of the 7.55 we got, but qualitatively the conclusions are the same.

## Odds Ratio Review

### Relative Odds or Odds Ratio



Suppose we have a disease  
(e.g. lung cancer)

And two groups  
(e.g. smokers, non-smokers)

Relative odds (OR)

$$= \frac{P(D|S)}{1-P(D|S)} / \frac{P(D|S^c)}{1-P(D|S^c)}$$

D ≡ disease S ≡ smoker

S<sup>c</sup> ≡ non-smoker

Let us return to the odds ratio for quantifying the relationship between two dichotomous classifications—the row and the column classification in our 2x2 table.

We saw the advantage of the odds ratio for case control studies, but it also quantifies these relationships in general.

Recall that 1 is the null value. It is the value for which there is no relationship between the row and column classifiers. So this pivotal value of 1 can be used as a point from which we can measure dependence; the further away the larger the dependence. The only hitch is that the differences are not symmetric around one. For example, m and 1/m are equally distant from 1 in some sense; for example odds of 3:1 are the same as 1:1/3. So the odds of 1/3:1 is just a reversal of what we consider success and what we consider failure. So if the distance from 1 is important, some people restrict themselves to restructuring the problem so as to just deal always with odds bigger than 1. We can always do this just by rephrasing the statement.

Another view is to always deal with the logarithm of odds and feeling that that domain is closer to linearity when dealing with ratios; the log of 1/3 is minus the log of 3, and that provides symmetry around 1 whose log is zero. We deal with this more fully when we look at logistic regression in the penultimate week.

Theory for odds ratio



	Exposed	Unexposed	Total
Disease	a	b	a+b
No Disease	c	d	c+d
Total	a+c	b+d	n

$$\widehat{OR} = \frac{\hat{P}(D|E)/(1-\hat{P}(D|E))}{\hat{P}(D|E^c)/(1-\hat{P}(D|E^c))}$$

$$= \frac{(a/a+c)/(c/a+c)}{(b/b+d)/(d/b+d)}$$

Treating the odds ratio as a population parameter—unfortunately we do not have a Greek letter reserved for this, we just use OR, so we are being inconsistent—we are interested in estimating it on the basis of a sample from that population. The natural estimator is the one above, obtained from using the obvious estimators of the probabilities shown.

After we cancel the a+c and the b+d we are left with ac divided by bd—the product along one diagonal divided by the product along the other diagonal; a simple formula to remember.

## Theory for odds ratio



	Exposed	Unexposed	Total
Disease	a	b	a+b
No Disease	c	d	c+d
Total	a+c	b+d	n

$$\widehat{OR} = ad/bc$$

$$\widehat{se}[\ln(\widehat{OR})] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

It turns out that the sampling distribution of the log of the estimator of the odds ratio, ( $\ln(ad/bc)$ ) is approximately normal and we can estimate the standard error simply by taking the sum of the reciprocals of the cell entries.<sup>2</sup> This is a very cute formula.

That is the theory for odds ratios.

## Data



Caesarean Delivery	EFM Exposure		Total
	Yes	No	
Yes	358	229	587
No	2,492	2,745	5,237
Total	2,850	2,974	5,824

$$\widehat{OR} = 1.72 \quad \ln(\widehat{OR}) = 0.542$$

$$\widehat{se}[\ln(\widehat{OR})] = 0.089$$

$$(0.542 - 1.96 \times 0.089, 0.542 + 1.96 \times 0.089)$$

$= (0.368, 0.716)$  is a 95% CI for  $\ln(OR)$

So  $(1.44, 2.05)$  is a 95% CI for  $OR$

Returning to our electronic fetal monitoring exposure and c-section study where we found that the chi-squared analysis had a small p-value. Approaching the same data using what we have learnt about odds ratios, we have the above analysis. The confidence interval for OR does not include one, so our inference here is in agreement with the chi-squared analysis. This analysis, though, is more informative because we quantify the dependence via the OR.

<sup>2</sup> If a cell entry is zero, just replace it with 0.5 for the sake of this formula.

### Stata output:

				Proportion	
	Exposed	Unexposed	Total	Exposed	
Cases	358	229	587	0.609	
Controls	2492	2745	5237	0.4758	
Total	2850	2974	5824	0.4894	
	Point estimate		[95% Conf. Interval]		
Odds ratio	1.722035		1.446551	2.049976	(Cornfield)
Attr. frac. ex.	.4192916		.3087003	.5121894	(Cornfield)
Attr. frac. pop	.2557178				
+-----					
	chi2(1) =		37.95	Pr>chi2	= 0.0000

The Stata output requires some studying, and I leave that to you to do.

## Berkson's Fallacy



**Example**

2,784 people interviewed of whom 257 hospitalized

Respiratory Disease		Total	P(respiratory disease)
Yes	No		
Circ dis.	7	29	36
			7/36 = 0.19
No circ dis	13	208	221
			13/221 = 0.06
Total	20	237	257
			20/257 = 0.08

$\chi^2 = 4.9 \Rightarrow p < .05$     $\widehat{OR} = 3.86$

We must include an example of what can go wrong with the study of discrete data; although this is certainly not peculiar to dichotomous data. This is an example of what is called Berkson's fallacy. Berkson is the one who quantified the fallacy, not commit the error. It is an example of sampling error, and it is the same error committed by Ray Pearl, about 100 years ago. He noticed from autopsy records that very rarely did people die of both cancer and tuberculosis. So he decided that the tuberculosis must be a protector against cancer. So as a treatment for the cancer patients, he proposed to inject patients with the tuberculin.

This was before the days of informed consent, and he actually did treat one patient thusly. The patient passed away, and he was about to treat a second patient when the authorities made him put a stop to his study. Apart from the obvious ethical issues involved judging by our current sensitivities, the question we need to ask is who gets autopsied? Does this pool of autopsied patients represent a random sample of dead patients? If not, is it then correct to infer what would happen to general patients from autopsied patients?

Consider the study above where 2,784 people were interviewed. Of those, 257 were hospitalized. So let us look at those 257: each was classified as either having a circulatory disease or not, and whether they had a respiratory disease or not. We see the resultant 2x2 table above.

When we subject this table to a chi-squared analysis we see that the  $\chi^2$  value is 4.9, the p-value is less than 0.05, and we reject the null hypothesis that there is no relationship between the row and column classifiers. In other words, this sure looks like there is a relationship between respiratory and circulatory ailments. Indeed, the odds ratio is a sizable 3.86.

We can understand what is happening if we look at the people with circulatory disease, seven of them had respiratory disease; about 19%. On the other hand, who had no circulatory disease, the probability of respiratory disease is 6%. That explains the estimated odds ratio.

But let us return to the original 2,784 who were interviewed. Who were these 287 that we chose of those 2,784? In a sense these were the sickest patients. These were the ones who were hospitalized. They certainly were not a random subset.

### Example continued



Whereas if we consider the whole sample:

Respiratory Disease		Total	P(resp disease)
Yes	No		
Circ dis.	22	171	193
No circ dis	202	2,389	2,591
Total	224	2,560	2,784
			224/2784 = 0.08

$$\chi^2 = 2.7 \Rightarrow p > 0.1 \quad \widehat{OR} = 1.52$$

Let us return to the whole sample only to find that this time the chi-squared analysis does not lead to the rejection of the null hypothesis and the odds ratio estimate is now reduced to 1,52. So on the basis of the 2,784 patients we conclude that there is no relationship no relationship between respiratory and circulatory diseases.

So what happened? What happened is we looked at the sub sample. Those who were more sick than the rest.



### Hospitalization Rates:

Circulatory Disease	Respiratory Disease	
	Yes	No
Yes	7/22=31.8%	29/171=17.0%
No	13/202=6.4%	208/2389=8.7%

Indeed, if we look at the hospitalization rates and how they vary with respect to the classifications we were interested in studying, we see a large discrepancy with an over representation of the group that provided the impetus for the results we saw.

Thus we got a distorted view of the relationship between respiratory diseases and circulatory diseases. Thus Berkson's fallacy results from not getting a representative sample.

### Yule Effect—Simpson's Paradox

Women who could be classified as smokers/non-smokers in a 20 year follow-up of a one-in-six survey of the electoral roll in 1972-1974 in Whickham, UK.

	Smokers	Non-smokers	Total
Dead	139	230	369
Alive	443	502	945
Mortality	0.239	0.314	0.281

DR Appleton, JM French, and MPJ Vanderpump, Ignoring a Covariate: An Example of Simpson's Paradox , *The American Statistician*, Nov 1996, Vol. 50, No. 4

One last topic: I cannot leave contingency tables without showing you this topic. This is something that is called the Yule effect, or, Simpson's Paradox, and that is because Simpson wrote about this about 50 years after Yule, so it makes sense that we call it the Simpson Paradox.

Let me introduce it by example. The researchers chose one in six persons in the electoral roll between 1972 and 1974, in Wickham, in the UK. Let us focus on the women in the sample, because that is the group the study reported.

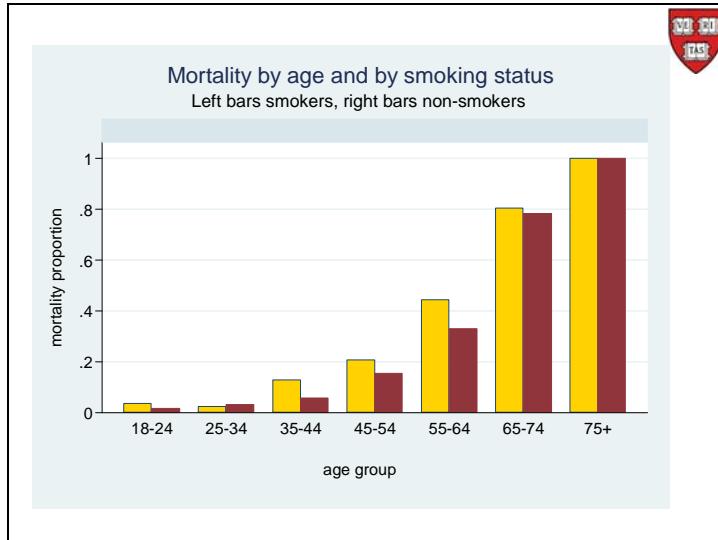
Each woman was classified as a smoker or a non-smoker, and then they returned 20 years later to do a follow up, and saw who had survived and who had passed away. What they discovered is shown above, where we see that the mortality rate amongst smokers is 0.239, whereas amongst the non-smokers it is a higher 0.314.

What is going on here? This is very odd. Before you run off and buy yourself some cigarettes, let us look a little more closely and see what is going on here.

Age	Smokers			Non-smokers		
	Alive	Dead	Mortality	Alive	Dead	Mortality
18-24	53	2	.036	61	1	.016
25-34	121	3	.024	152	5	.032
35-44	95	14	.128	114	7	.058
45-54	103	27	.208	66	12	.154
55-64	64	51	.443	81	40	.331
65-74	7	29	.806	28	101	.783
75+	0	13	1	0	64	1
Total	443	139	.239	502	230	.314

Let us break down the two groups into age categories because we learnt that the mean can hide a lot of sins and the composition formula warned us of what can happen when comparing two different groups.

If we look at each of these age groups, we see that, except for the 25-34 age group, the mortality rates for the smokers are higher than the non-smokers. In fact, in the 25-34 age group if there had been one more death amongst the smokers, then the mortality in that group would also fall in line.



The point is well made in this graphic. We see that the yellow bars are bigger than the red bars, except in the second age group. But that has such a small mortality that it is unlikely that could be driving the overall average being higher for the red group than for the yellows.

Having learnt from the composition formula we look for the makeup of the two groups of women.

Age	Smokers				Non-smokers			
	Alive	Dead	Proportion	Mortality	Alive	Dead	Proportion	Mortality
18-24	53	2	.0945	.036	61	1	.0847	.016
25-34	121	3	.2131	.024	152	5	.2145	.032
35-44	95	14	.1873	.128	114	7	.1653	.058
45-54	103	27	.2234	.208	66	12	.1066	.154
55-64	64	51	.1976	.443	81	40	.1653	.331
65-74	7	29	.0619	.806	28	101	.1762	.783
75+	0	13	.0223	1	0	64	.0874	1
<b>Total</b>	<b>443</b>	<b>139</b>	<b>1</b>	<b>.239</b>	<b>502</b>	<b>230</b>	<b>1</b>	<b>.314</b>

When we look at the composition of the two groups we see that the smokers are younger than the non-smokers: the proportions in the under 64 age groups are all higher for the smokers, except for the 25-34 age group which is larger in the third decimal place for the non-smokers. The older age groups, the 65 and over, are where the non-smokers have higher representation than the smokers.

Marcello Pagano

# [JOTTER 8 SURVEY SAMPLING]

Stratified sampling, cluster sampling, biases, randomized response.

1996 Presidential Election				
Poll	N	Clinton	Dole	Perot
Harris	1339	51	39	10
ABC-Wash Pst	703	51	39	10
CBS-NYT	1519	53	35	12
NBC-WStJ	1020	49	37	14
Gall-CNN-USA	1200	52	41	7
Reuters/Zogby	1200	44	37	19
Pew Research	1211	49	36	15
Hotline/Battlg	1000	45	36	19
Actual	96m	49.3	40.7	10

<sup>1</sup>

Possibly the most popular use of surveys, especially during presidential campaigns, is to make predictions about the outcome of the election. In the political realm, such surveys are called polls.

If we look at the 1996 presidential elections, when President Clinton was running against Senator Dole with Perot as a third candidate, we have here eight polls, each represents the last poll taken by that company before the actual election.

In the actual election, 96 million people voted, with Clinton getting 49.3% of the vote, Dole getting 40.7% of the vote, and Perot getting 10% of the vote.

When we compare the poll results to the actual numbers who voted for each candidate, given in the bottom line, we see a good agreement.

But the fascinating thing is evident when we look at the Harris poll, for example. The Harris poll was based on asking 1,339 people to give us their opinion of how they would vote. Now this is 1,339 versus the 96 million who actually voted. How can 1,339 people predict what 96 million people are going to do? Indeed, the mean sample size for these eight polls is 1,149. So based on what just over a thousand people say, these pollsters predict how 96 million people will vote, at some future time.

This is the magic. How well did the pollsters do in subsequent elections?

---

<sup>1</sup> <http://www.ncpp.org/files/1936-2000.pdf>

2000 Presidential Election				
Poll	N	Gore	Bush	Nader
Harris (phone)	1348	47	47	5
ABC-Wash Pst	826	45	48	3
CBS	1091	45	44	4
NBC-WStJ	1026	44	47	3
Gall-CNN-USA	2350	46	48	4
Reuters/Zogby	1200	48	46	5
Pew Research	1301	47	49	4
Battleground	1000	45	50	4
<b>Actual</b>	<b>96m</b>	<b>48</b>	<b>48</b>	<b>3</b>

<sup>2</sup>

When we look at the results for the 2000 elections, once again 96 million had their votes counted. The reported results had them at a virtual tie; 48% each.

And once again, on the basis of an average of 1,268 responders per poll, the predictions were rather good.

2004 Presidential Election		
Projector	Bush	Kerry
Harris	49	48
ABC-Wash Pst	49	48
CBS	49	47
NBC-WStJ	48	47
USA/Today/Gall	49	47
Zogby	49.4	49.1
Pew Research	51	48
Battleground	51.2	47.8
<b>Actual</b>	<b>50.75</b>	<b>48.3</b>

<sup>3</sup>

<sup>2</sup> <http://www.pollingreport.com/wh2gen1.htm>

<sup>3</sup> <http://www.pollingreport.com/2004.htm>

Once again in 2004, where I have not included the number polled, but they remain at roughly what they were in the previous years.

CANDIDATE ESTIMATE ERROR -Preliminary Report	Start Date	End Date	Voter Sample	MoE +/ -	Obama	McCain	Unallocated
<b>UNOFFICIAL RESULT - 11/21/08</b>							
<b>52.7% 46.0%</b>							
<b>Projections-Undecideds Allocated</b>							
1 GWU/Battleground-Tarrance(R)	2-Nov	3-Nov	800	3.5%	50%	48%	
2 GWU/Battleground-Lake(D)	2-Nov	3-Nov	800	3.5%	52%	47%	
3 Rasmussen	1-Nov	3-Nov	3,000	1.8%	52%	46%	
4 Investors Business Daily/TIPP	1-Nov	3-Nov	981	3.1%	52%	44%	
5 Harris Interactive (Internet)	30-Oct	3-Nov	3,946	1.6%	52%	44%	
6 Gallup /USA Today	31-Oct	2-Nov	2,472	2.0%	55%	44%	
7 McClatchy/Ipsos	30-Oct	2-Nov	760	3.6%	53%	46%	
8 Democracy Corps, GQR (Dem)	30-Oct	2-Nov	1,000	3.1%	53%	44%	
9 Pew Research Center	29-Oct	1-Nov	2,587	1.9%	52%	46%	
<b>Final Trial Heats-No allocation</b>							
10 Marist College	3-Nov	3-Nov	804	3.5%	52%	43%	
11 ABC News/Washington Post	31-Oct	3-Nov	2,904	1.8%	53%	44%	
12 Daily Kos (D)/Research 2000	1-Nov	3-Nov	1,100	3.0%	51%	46%	
13 American Research Group	1-Nov	3-Nov	1,200	2.8%	53%	45%	
14 NBC News/Wall St Journal	1-Nov	2-Nov	1,011	3.1%	51%	43%	
15 Zogby - Reuters	31-Oct	3-Nov	1,226	2.8%	54%	43%	
16 FOX News/Opinion Dynamics	1-Nov	2-Nov	971	3.1%	50%	43%	
17 CBS News	31-Oct	2-Nov	932	3.2%	51%	42%	
18 Hotline-Diagco/FD	31-Oct	2-Nov	887	3.3%	50%	45%	
19 CNN/Opinion Research Corp	30-Oct	1-Nov	714	3.7%	53%	46%	
<b>ESTIMATE AVERAGE/POLL ERROR</b>							
<b>52% 44%</b>							

4

The 2008 election had results very similar to the previous elections. Roughly the same poll sizes and roughly the same accuracy of the predictions when compared to what actually took place.

How is it that 1,000, or so, people can predict what 100 million will vote?

---

<sup>4</sup> [http://www.ncpp.org/files/08FNLncppNatlPolls\\_010809.pdf](http://www.ncpp.org/files/08FNLncppNatlPolls_010809.pdf)

 **Mark Blumenthal** W Become a fan  
mark@huffingtonpost.com



## 2012 Poll Accuracy: After Obama, Models And Survey Science Won The Day

Posted: 11/07/2012 8:04 am EST

**Pollster Model Correctly Predicts Outcome in All States**  
Results as of Nov. 7, 2012, 5:01 a.m. ET

State	Electoral Vote		Pollster Estimate			Unofficial Election Result			
	State	Cumul.	Obama	Romney	Margin	% in	Obama	Romney	Margin
Pennsylvania	20	<b>237</b>	50.1	44.2	<b>+5.8 D</b>	99%	51.9	46.8	<b>+5.1 D</b>
Wisconsin	10	<b>247</b>	50.4	45.8	<b>+4.7 D</b>	99%	52.8	46.1	<b>+6.7 D</b>
Nevada	6	<b>253</b>	50.0	46.5	<b>+3.6 D</b>	98%	52.3	45.7	<b>+6.6 D</b>
Ohio	18	<b>271</b>	49.2	45.8	<b>+3.4 D</b>	99%	50.1	48.2	<b>+1.9 D</b>
Iowa	6	<b>277</b>	48.6	46.0	<b>+2.6 D</b>	99%	52.1	46.5	<b>+5.6 D</b>
New Hampshire	4	<b>281</b>	49.2	46.8	<b>+2.4 D</b>	90%	52.0	46.7	<b>+5.3 D</b>
Virginia	13	<b>294</b>	48.7	46.8	<b>+1.9 D</b>	99%	50.8	47.8	<b>+3.0 D</b>
Colorado	9	<b>303</b>	48.6	46.8	<b>+1.7 D</b>	71%	50.5	47.3	<b>+3.2 D</b>
Florida	29	<b>332</b>	48.4	47.9	<b>+0.5 D</b>	100%	49.8	49.3	<b>+0.5 D</b>
North Carolina	15	<b>206</b>	47.3	48.8	<b>+1.6 R</b>	100%	48.4	50.6	<b>+2.2 R</b>

5

This last election (2012) the pollsters also predicted down to the State level and got all fifty results correct. Indeed, this is the second presidential election when one pollster, Nate Silver<sup>6</sup>, achieved this feat. Here are the ten States most difficult to predict. These are the results given for the “modelers” who averaged individual pollster’s results. We know from the Central Limit Theorem (standard deviation versus standard error) that they should do better, but it is still amazing that they actually got all fifty correct.

<sup>5</sup> [http://www.huffingtonpost.com/2012/11/07/2012-poll-accuracy-obama-models-survey\\_n\\_2087117.html](http://www.huffingtonpost.com/2012/11/07/2012-poll-accuracy-obama-models-survey_n_2087117.html)

<sup>6</sup> <http://fivethirtyeight.blogs.nytimes.com/>

NCHS 

The National Center for Health Statistics

The Vital Statistics Program,

The National Health Survey Program

<http://www.cdc.gov/nchs/>

The National Center for Health Statistics<sup>7</sup>, who maintain the Vital Statistics program for the US, have also taken, in the last half-century a number of surveys to measure and monitor the health of the population

NCHS 

The first three of these national health examination surveys were conducted in the 1960s:

1. 1960-62—National Health Examination Survey I (NHES I);
2. 1963-65—National Health Examination Survey II (NHES II); and
3. 1966-70—National Health Examination Survey III (NHES III).

All 3 surveys had an approximate sample size of 7,500 individuals.

1. 1971-75—National Health and Nutrition Examination Survey I (NHANES I);
2. 1976-80—National Health and Nutrition Examination Survey II (NHANES II);
3. 1982-84—Hispanic Health and Nutrition Examination Survey (HHANES); and
4. 1988-94—National Health and Nutrition Examination Survey (NHANES III).

+ +

They started out early 1960s with the National Health Examination Survey. They repeated that survey twice more in that decade. Then they expanded to the National Health and Nutrition Examination surveys, starting in 1971 and extending into the mid-90s. These very important health surveys provided invaluable information, and the center continues to provide a wealth of statistical information. It is a wonderful resource, and well worth a visit.

---

<sup>7</sup> <http://www.cdc.gov/nchs/>

DHS



Demographic and Health Surveys (USAID)

<http://www.measuredhs.com/>

Another important source of health related survey data—although this is much more global—is one funded by USAID, and it is the collection of Demographic and Health Surveys (DHS)<sup>8</sup>. Go visit them. Lauren has a presentation about this topic this week.

Accuracy and Precision



- Accuracy – how close to reality the measure represents
  - Depends on how much bias
- Precision – how confident are we in the reproducibility of the results.
  - Depends on variability

As with all measuring instruments, and certainly surveys fall into that category, we need to know how trustworthy the results are. With inferential tools we like to consider two aspects, and we differentiate between them here: accuracy and precision.

---

<sup>8</sup> <http://www.measuredhs.com/>

Accuracy attempts to quantify how close to the truth the measure is. Another way of explaining this is to look at the opposite view and see whether we have bias in our measurement. For example, if we only survey men, then it is not really telling us about the population. At best the survey will only contain information about half the population. So it is biased in that respect.

Does the bias matter? It may or may not depending on what it is we are measuring; is there a sex difference in what it is we are measuring? The answer depends on the context, but typically we do not wish to take the chance that it does matter.

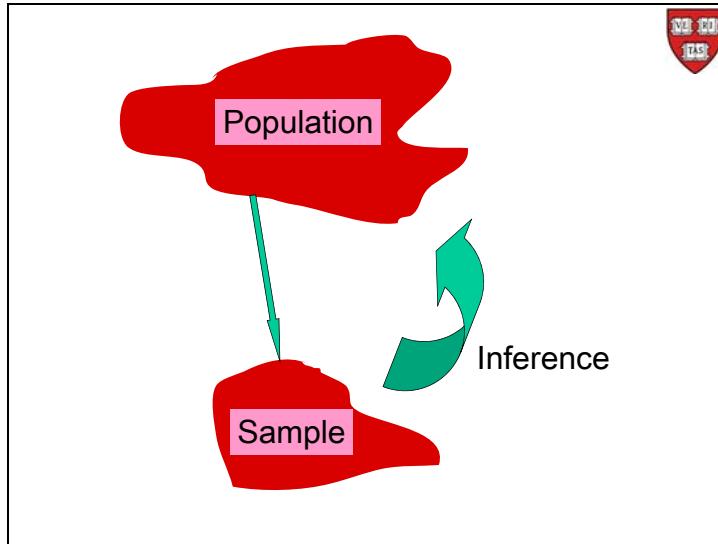
The precision associated with the survey tells us how confident we are in the reproducibility of the results. This very much depends on how variable the population is that we are attempting to measure. If everyone in the population is equal to each other, then we need a sample of size one, and it is very precise. So the precision depends upon the population variability and the size of the survey.

## Precision

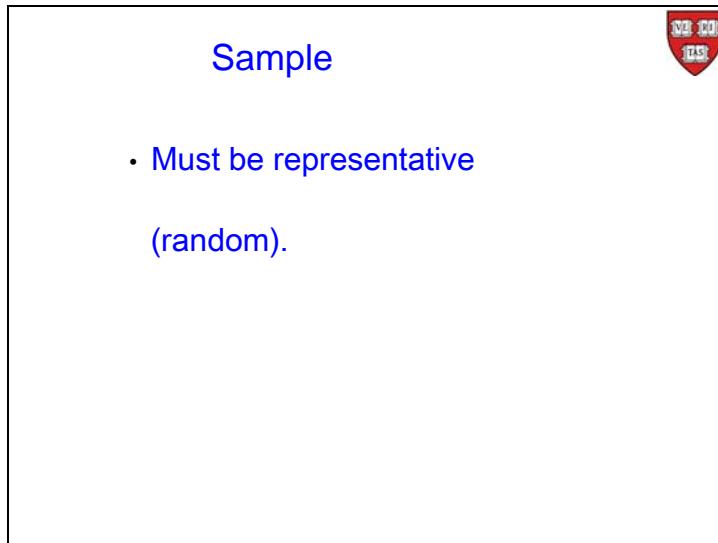


- Function of...
  - Sample size
  - Way that sample was selected
- Some designs yield more precise estimates than others.

Precision is a function of the sample size and, as it turns out, the way that sample was selected. Some designs yield more precision than others, and yet retain representativeness by incorporating external information we may have about the population under study. We elaborate on this point as we go along.



Returning to our introduction to inference, we said that we have a population from which we take a sample, and on the basis of this sample make inference about the population.



Up to now, we have taken a simple random sample from the population—everyone in the population has an equal chance of being chosen to be in the sample. The reason for this choice is that we decided that a representative sample was unobtainable in general, so we resort to random samples to obtain representativeness in aggregate.



## Sampling Theory

Up to now we have assumed:

1. Population infinite
2. Simple random sample

Suppose we take a sample of size  $n$   
from a population of size  $N$  with  
mean  $\mu$  and standard deviation  $\sigma$

So far, we have not made any mention of the population size. We have been going along with what is called the infinite population assumption; namely that the population is huge. Let us investigate that assumption a little.

To this end, introduce some notation: suppose we are to take a sample of size  $n$  from a population of size  $N$  that has a mean  $\mu$  and a standard deviation  $\sigma$ .

Does the population size make any difference to our inference? And the answer is, yes, the population size is small, or the sample is big relative to the population size.

## Sampling fraction



Sampling without replacement  
(no duplication)

$$\text{Sampling fraction: } f = \frac{n}{N}$$

Simple random sample:

Each individual has an equal chance,  $f$ ,  
of being included in the sample.

(More about this equiprobability later.)

What is important, when quantifying our inference, is the sampling fraction,  $f = n/N$ .

We now have the notation to tell us what we mean by simple random sample. It means that every person in the population has the same probability of being in the sample, and that probability is  $f$ . So let us say that we are going to take a sample of 100 people from our population of 1,000. Then  $f$  is equal to  $1/10$ , and that is the probability each person has of being in the sample. Technically, we are talking about *sampling without replacement*, which means that once a person has been chosen to be in the sample then that person is removed from the prospective pool of people to be subsequently chosen for the sample.

This chance of  $f$  is for simple random sample. With some sampling designs we now consider, this may not hold, even if the sampling is random at some level, different individuals will have different probabilities of being chosen. These probabilities are sometimes called weights.

The number  $f$  is called the *sampling fraction*.

### Finite population correction factor



Central limit still holds, except:

$$\begin{aligned}s.e. &= s.d.(\bar{X}) = \frac{\sigma}{\sqrt{n}} \sqrt{1-f} \\ &\approx \frac{\sigma}{\sqrt{n}} \text{ if } f \approx 0\end{aligned}$$

So, if population is huge,  $f \approx 0$   
and population size is not important.

That means that  $n$  and not  $f$  determines the precision.

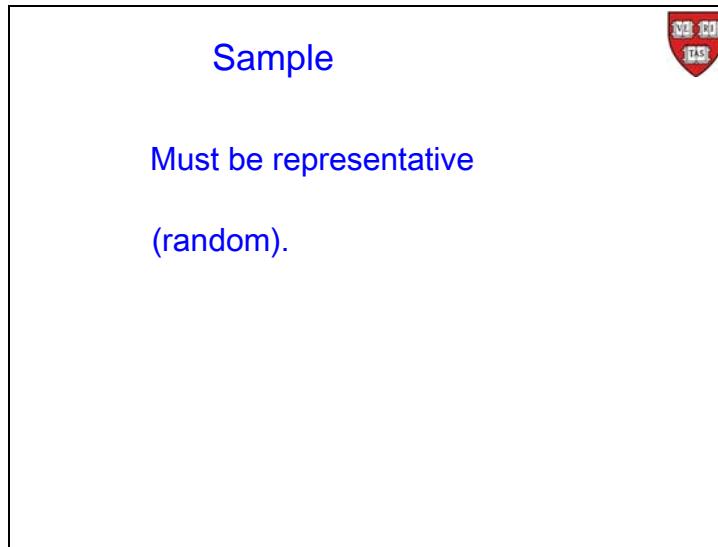
The central limit theorem still holds even when sampling from a finite population. The difference is that we must make the population size explicit. This manifests itself in the formula for the standard deviation of  $\bar{X}$ —namely the standard error—is actually as shown above.

We recognize the  $\sigma/\sqrt{n}$ ; that is what we had before. But now we have what is called the *finite population correction factor*:  $\sqrt{1-f}$ . If  $f \approx 0$ , then we are back to the formula we have become accustomed to, and there is nothing new to be learnt. So if the population is very large and the sample does not represent a sizable fraction of the population, then we can assume that  $f$  is approximately zero and ignore it.

In summary, if we have a huge population—120 million voters in the US, say—and we are going to choose a sample of 1,000 people and ask them about how they are going to vote, then that gives us an  $f$  that is 1 over 120,000. So  $\sqrt{1-f} \approx 1$ . Thus it makes no practical difference to the value of the standard error, if we ignore this factor, or not.

The important point to notice is that the population size  $N$  only enters into the argument via  $f$ , so if we ignore  $f$  as being too small, then it is the sample size  $n$  that is important in the standard error formula and not the population size  $N$ . This is something that must seem counterintuitive, judging by the number of people who do not believe this at first—namely, if I am going to take a sample of size 1,000, it does not matter if I take this 1,000 from the City of New York whose population is approximately 10 million, or if I take that 1,000 from the US, whose population is 300 million or so, or if I take that 1,000 from the world, which has a population of 7 billion. It does not matter as far as the standard error of my estimate is concerned.

This is true as long as we are selecting a truly simple random sample. Remember that that means that everyone in the population has an equal chance of being selected, so that may be the cause if the reticence felt by some to the statement that it is the sample size that matters, and not the population size. Possibly, one is more willing to believe this assumption is satisfied with a smaller rather than with a larger population, such as everyone in the world, but I am speculating now.



Let us return to thinking about the requirements we have of the sample; namely, that it be representative. Intuitively, this makes sense since we are going to base our inference about the population on what we find out in the sample. What does representative really mean, since it feels like it should fit the bill perfectly?

The screenshot shows a web page from Investopedia. At the top left is the Investopedia logo with the word "INVESTOPEDIA". At the top right is a small red shield-shaped logo with white text. Below the header, the title "Representative Sample" is displayed. Underneath the title is a sub-header "Filed Under » Statistics". To the left of the main content area is a small icon of an open book. The main content area starts with the definition of "Representative Sample": "Definition of 'Representative Sample'" followed by a detailed explanation. To the left of this text is another logo consisting of a red circle with a white letter "I" inside. At the bottom of the page is a URL: "http://www.investopedia.com/terms/r/representative-sample.asp#axzz2APWIHYEF".

I went onto the web and I found a location called, Investopedia<sup>9</sup>. They define what they mean by a representative sample, and I quote, "it is a subset of a statistical population." Now I don't know what is meant by a statistical population, as opposed to just a population, but to continue, "that accurately reflects the members of the entire population."

The example they give is, in a classroom of 30 students in which half the students are male and half are female, a representative sample might include six students, three males and three females. I do not know what they mean by "might include." I suspect they do not mean that it might not include them either. I suspect they mean it should include three males and three females because that would then be representative of the population of 30 students; because half are male and half are female. It seems like to them a representative sample is a miniature version of the population.

They go on to say, "when a sample is not representative, the result is known as a sampling error." Do not believe everything you read.

They state, "Using the classroom example again, a sample that includes six students, all of whom are male, would not be a representative sample. Whatever conclusions were drawn from studying the six male students would not be likely to translate to the entire group since no female students were studied."

There are some things wrong with that statement. First, what they seem to be saying is that a representative sample must be 50% female and 50% male. This works fine with a single factor, such as sex—assuming we do not take an odd numbered sample!

---

<sup>9</sup> <http://www.investopedia.com/terms/r/representative-sample.asp#axzz2DoRioP00>



	Female	Male
Number	15	15
Proportion	50%	50%

	Female	Male
Young	9 (60%)	6 (40%)
Older	6 (40%)	9 (60%)

Other factors, e.g. height, weight, .....

But why stop at a single factor. Suppose that in this classroom, we judge students as being young for that class or older for the class. And suppose the breakdown is as shown above.

So not only do we need to have a 50-50 sex split, but we also must have that amongst the females, we have that 60% are young and 40% are older, and amongst the males, we want that 40% are young and 60% are older, if the sample is to be representative.

So you can imagine what is going to happen once we start looking at additional factors: height, for example. Some are tall, some are short. Do we have to have the correct representation of height amongst the young females, the older females, the young males and the older males? How about weight? Should we have correct representation of weight distribution amongst the tall, young, females; amongst the short, young, females; etc..

Very quickly, we are going to be unable to satisfy these requirements.

There are two things wrong with this approach: One is the fact that once we start looking at more and more and more factors, it becomes more and more complex as we define more and more precise cells to divide up the population, and the sample, and we are not going to be able to find people to fit into each cell. A trivial example, if we stop at just two factors we need to fill four cells (young-female, older-female, young-male and older-male), but if we only need a sample of size three, we cannot proceed. In general if there are  $k$  factors and we assume, in order to simplify the argument, that each factor is only at two levels. Then we would need a sample of size exactly  $2^k$  to claim we have a representative sample.

Problem number two is how is it that we know this much about your population? To be truly representative we need to know how all these factors, assuming we have all the factors accounted for, interact with each other. For example, as above, how the age distribution varies with sex. Knowing this much about the population then raises the issue about the need for sampling?

The profession had this argument some 100 years ago, and consensus was reached that rather than rely on the “experts” to decree what is and what is not representative, we turn to randomization to provide a sample. This then has the advantage that we can quantify the precision and accuracy with which we make our inference. That is not to say that we cannot incorporate important information into our sampling methods, as we show below.

Further, we must be careful in how we actually communicate with individuals. For example, if we carry out a phone survey, then, obviously, only people who own phones can be sampled. Second, we know that the ownership of cell phones and/or land lines are very much dependent on the age, and other demographic factors, of the owner<sup>10</sup>. So if we ignore this fact—for example, if we ignore people reachable only through cell phones—then the resultant survey might well be biased towards one group or another.



If you have knowledge to bring to the table, by all means use it—for example, if you know you have 50% female and 50% male, and sex will impact the outcome of interest, then by all means ensure that your sample is 50% female and 50% male.

Once you have satisfied your marginal constraints, we turn to,

#### Simple Random Samples

The point of all this is if you have knowledge to bring to the table, by all means, bring it. For example, if you have 50% female and 50% male, and sex will impact your outcome of interest, then make sure that your sample is 50% female and 50% male. Do not leave it up to chance because chance is not going to give you a 50-50 split every time.

But be careful, as argued above, do not impose too many of these constraints, because you quickly run out of degrees of freedom. Further, once you impose a constraint you lose the ability to estimate the prevalences of those factors. For example, if you impose a 50-50 sex ratio in your, you can no longer estimate the sex ratio in the population.

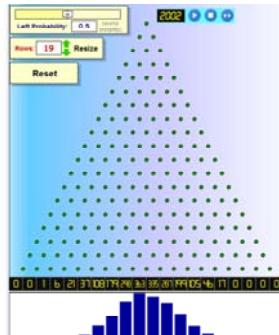
---

<sup>10</sup> <http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201112.htm#differences>

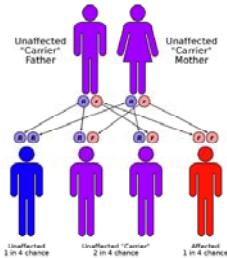
Chance is but the expression of man's ignorance.



Pierre-Simon Laplace  
Essai philosophique sur les probabilités, 1814



### Thalassemia



Laplace said, "chance is but the expression of man's ignorance." And here we are aiming for a simple random sample and rely on chance as a foundation for our inference!

We know that probability is not very helpful, except to provide uncertainty, in the short run, and the other reason why it is attractive is that it provides us with predictable, steady behavior in the long run. We saw this repeatedly with the Quincunx, and in real life we looked at the long run behavior of families with Thalassemia.

We bank our inference on this long run stability.

### Sampling Frame



In order to ensure that everyone has an equal chance of being sampled, we need what is called a sampling frame; or an itemization of every person in the population.

If the two are not the same (population and sampling frame) then we are actually taking a random sample from the sampling frame.

To obtain a random sample we need to ensure that everyone have an equal chance of being chosen. This is achievable if we have a list of everyone in the population. That list is called the

*sampling frame*. In a sense the two are inextricably related; the population ideally defines the sampling frame, but operationally the sampling frame defines the population from which we can sample.

For example, if you wish to make inference about everyone who lives in a city, your list may consist of all the home addresses in the city. That means you do not have homeless people represented in your sample, and as a result you do not have a sample of everyone in the city.

Another way of stating that is that you may have a random sample of people who have homes in the city.

If your sampling frame consists of everybody who voted at the last election, you are not getting everybody in your town. If you are a jury clerk who chooses potential jurors, if you use such a list then not everyone in the city has an equal chance of being on a jury, which might be a requirement by law.

We also need a random device that can operate on the sampling frame and yield a random sample. Let us assume we have such. The government has rules that should be obeyed by random samples. Let us assume we obey those rules.<sup>11</sup>



### WEIRD Samples

Psychologists rely on Western, Educated, Industrial, Rich, and Democratic subjects.

"The findings suggest that members of WEIRD societies, including young children, are among the least representative populations one could find for generalizing about humans."

J Henrich, SJ Heine, & A Norenzayan, The weirdest people in the world? *Behavioral and Brain Sciences* (2010) 33, 61–135

Sometimes the samples are not quite as random as we would like to believe. Psychologists speak about WEIRD samples<sup>12</sup>. What they mean is—WEIRD is an acronym for Western, Educated, Industrial, Rich, and Democratic subjects—that their subjects may not truly be as representative as they thought and as a result the theories that they have promulgated on the basis of their studies may be less generalizable than they thought.

---

<sup>11</sup> [http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards\\_stat\\_surveys.pdf](http://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf)

<sup>12</sup> [Henrich J, Heine SJ, Norenzayan A](#). The weirdest people in the world? *Behav Brain Sci.* 2010 Jun;33(2-3):61-83; discussion 83-135. Epub 2010 Jun 15.

Now we do something similar in our health studies. If in need of a hospital set of patients, our students typically go to teaching hospitals to find them. Most medical studies are also done at teaching hospitals. For example, if we need a sample of children, there is a children's hospital just down the street. It is a wonderful hospital. The research it performs is terrific. But are those children at that hospital representative of all children everywhere in the world?

Possibly health studies are not going to be as badly influenced as the social scientists who are mainly concerned with cultural events. There are a number of health studies, of course, that are very heavily dependent on cultural mores, but it will be interesting to find out as time evolves how dependent our medical studies suffer from this phenomenon.

### Stratified Sampling

Suppose we have information about our population that we wish to incorporate into our survey design and analysis. For example, the country we are surveying has provinces and we want to make inference both at the provincial level and at the national level. Or there may be structural information: for example, we may know that the population is half female and half male. Making use of this external information should prove beneficial in our inference about the population in improving both precision and accuracy. Further, when calculating information about the whole, it is more informative to along the way also be able to obtain information about subgroups that make up the whole.

For example, obtaining information for each province and then combining those provincial estimates to obtain an estimate about the country yields information about how the country aggregate is distributed amongst the provinces. This is an example of *stratified sampling*.

For example, if we have 6 provinces, and suppose we want a sample of size 60 for our country, then taking 10 from each province will give me information about each province. Whereas, had I taken a simple random sample of 60 from the country, it is possible (more than 1 in a thousand) that I get zero, or one person from one of the provinces. So stratified sampling seems more informative than simple random sampling.

Further, there is no dictate that says we need to take the same number of individuals from each province. Intuitively, we may wish to spend more effort (larger sample) in a province that is more variable than one that is more homogeneous.

These are some of the optimizations we can exercise with stratified sampling.



Note that all the Xs within a group need not be equal.

$$\begin{aligned}
 \bar{X} &= \frac{1}{6}(1+2+3+4+6+8) \\
 &= \frac{1}{6}(\{1+2+3\} + \{4+6\} + 8) \\
 &= \frac{1}{6}\left(3\frac{\{1+2+3\}}{3} + 2\frac{\{4+6\}}{2} + 1\frac{8}{1}\right) \\
 &= \frac{1}{6}(3 \times 2 + 2 \times 5 + 1 \times 8) \\
 &= .5 \times 2 + .33 \times 5 + .17 \times 8 \\
 &= \sum_{i=1}^3 p_i \bar{X}_i = 4
 \end{aligned}$$

A central role in stratification is the construct that breaks the whole into parts and then recombines the results from each to get the result for the whole. Our initial focus is with the overall mean, or the total prevalence, so recall the composition formula that explicitly shows how to combine group means to obtain the overall mean. This formula plays a central role when dealing with stratified sampling.



### HIV Antenatal Clinic Surveillance (Sentinel or Convenient Sample)

ANC HIV surveillance has been carried out among women attending antenatal clinics in more than 115 countries worldwide.  
(2006 – 600 sites in sub-Saharan Africa)

Annually or bi-annually and they provide ready and easy access to a cross-section of sexually active pregnant women from the general population.

Used to “assess trends” in the epidemic over time.

In generalised epidemics, HIV prevalence among pregnant women has been considered a good approximation of prevalence among sexually active men and women aged 15–49 years.

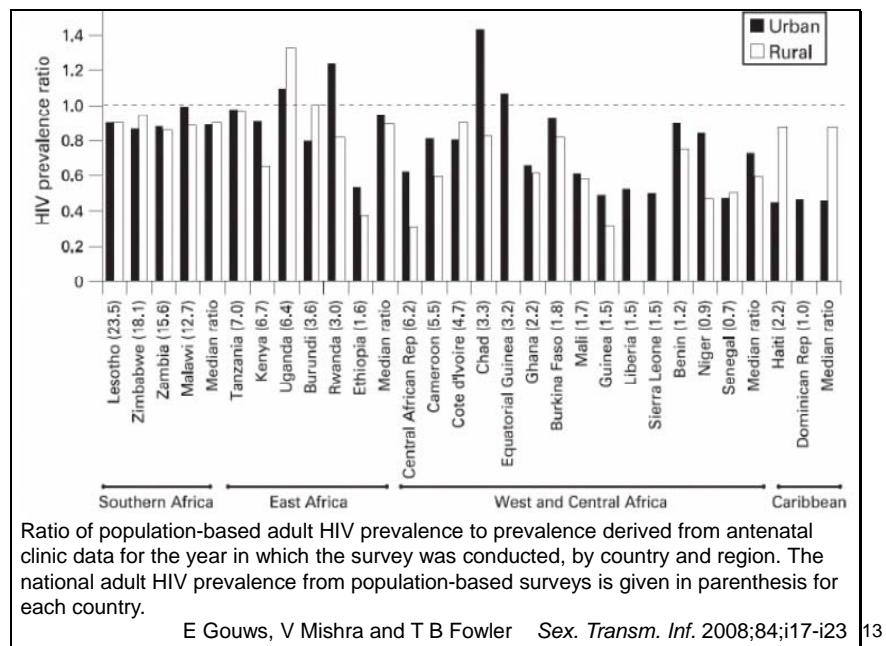
Considering the whole as made up of groups, or parts, is often quite informative because when we look at the composition formula we can see the effect of ignoring any groups. We need everybody to be represented. Consider a case in point, the estimation of HIV prevalence, and incidence, at a global level. A common practice in disease surveillance is to make use of

*sentinel systems.* A sentinel system concentrates on some pre-chosen hospitals, let us say, to use them as sentinels, or proxies, for what is going on in the population.

One that is very widely used, especially in sub-Saharan Africa, to monitor, the HIV epidemic, makes use of information available from Antenatal Clinics (ANC). Women go to ANCs to get prenatal care, and part of prenatal care is to collect and store blood samples from the patients. These blood samples can be tested for the presence of the HIV and thus one can obtain an excellent measure of the prevalence, and incidence, of HIV infection amongst this group of women.

The difficulty arises when one attempts to use these measures to infer values for the population as a whole. A simple argument that this group of women represents the whole population is problematic in a number of ways: these are young, sexually active women who can access the ANC. So neither are men, nor are women in general represented—for example, women not represented include those not in their child bearing years, those who are not sexually active (that is, of course, a critical consideration when dealing with a sexually transmitted disease), and women who do not live near an ANC, or afford themselves of any ANC services.

Are these just theoretical considerations, or does it make a difference to our estimators?



<sup>13</sup> E Gouws, V Mishra and T B Fowler, Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data *Sex. Transm. Inf.* 2008;84:i17-i23

In this study the authors compare the results of the ANC Surveillance, to results from a DHS done at approximately the same time. The graphic shows the ratio of the population based prevalence (DHS) to that obtained from the ANC. The dotted line represents when the ratio is unity, and thus is the national average—reported parenthetically next to the country's name.

The solid blocks are for urban areas, and the open blocks represent rural areas. A few of these bars cross the unit line, but the great majority of them fall beneath that line. That means that the prevalences at the Antenatal Clinic are consistently, except for the four bars in the center, higher than the national averages, as we expect.

Some argue that, these ANC averages may not provide good estimates of the prevalence, but they do track the epidemic, and provide a good idea of how things are moving; in other words, a good estimate of the incidence.<sup>14</sup> For this to be true we would have to argue that the same mechanism that caused the prevalence to go up or down in sexually active women of child bearing age who have access to ANCs will have the same effect on the rest of the population. In order for this to be true, it would have to be a perfect storm that needs to remain in place throughout the life of the sentinel system.

### Bias



UNAIDS/WHO estimates of national adult HIV prevalence have been based on prevalence data collected over time from pregnant women attending antenatal clinics.

But without an element of randomness, we cannot claim unbiasedness, and we cannot measure the bias.

Modest proposal to retain the knowledge in the convenient sample but introduce unbiasedness.

Hedt & Pagano Statistics in Medicine 2010

What we have is a biased system since we are only measuring a subset of the population. Can this bias be removed? And the answer is yes, but it requires some sampling, preferable random, of the rest of the population. We can take advantage of the information from the ANC, especially since the economics of the situation make large samples readily available from the ANC. The methods are covered in this report<sup>15</sup>.

---

<sup>14</sup> This hope that two (prevalence) curves that are not equal will have the same derivatives (incidences) associated with them, can be labeled wishful thinking.

<sup>15</sup> [Hedt BL, Pagano M](#). Health indicators: eliminating bias from convenience sampling estimators. *Stat Med*. 2011 Feb 28; 30(5): 560-8.

## Stratified Sampling



Aim: Increase precision for same cost.

Suppose the population is made up of  $g$  groups,  
with group

Sizes:	$N_1$	$N_2$	...	$N_g$
Means:	$\mu_1$	$\mu_2$	...	$\mu_g$
Std. Devs.	$\sigma_1$	$\sigma_2$	...	$\sigma_g$
samples	$n_1$	$n_2$	...	$n_g$
Sample means	$\bar{X}_1$	$\bar{X}_2$	...	$\bar{X}_g$

Consider stratified sampling in general. Suppose our population is composed of  $g$  groups, or strata. And suppose that, for each stratum, say the  $i^{\text{th}}$ , we know the size of the stratum,  $N_i$ , but we do not know its mean,  $\mu_i$ , or its standard deviation  $\sigma_i$ . We are interested in estimating these parameters by taking random samples of size  $n_i$  from each of the  $g$  strata. Let us also assume that the population size is  $N$  and the total sample size is  $n$ .

## Stratified Sampling



$$\mu = \sum_{i=1}^g \frac{N_i}{N} \mu_i$$

So, estimate by

$$\hat{x} = \sum_{i=1}^g \frac{N_i}{N} \bar{X}_i$$

whose variance is

$$\sum_{i=1}^g \left( \frac{N_i}{N} \right)^2 \frac{\sigma_i^2}{n_i} \left( 1 - \frac{n_i}{N_i} \right)$$

which is minimized by choosing.

We are interested in the overall population mean, which from our composition formula we know we can express as above. So it makes sense that we estimate this overall mean by the weighted average,  $\hat{x}$ , we show above. It is a weighted average of the individual strata means.

We see that the weights are proportional to the size of the population within each stratum. Thus if we are to use this estimator, we need to know the stratum sizes.

We can also calculate the variance of this estimator, whose square root is the standard error. We have control of the sample sizes from each stratum, so if we want to choose a design to minimize this standard error, we can modify the individual sample sizes—a straightforward mathematical problem.

### Stratified Sampling



$$n_i = n \frac{N_i \sigma_i}{\sum_{i=1}^g N_i \sigma_i}$$

Thus, to maximize precision, the sampling fraction in each stratum should be proportional to the standard deviation in that stratum, i.e.

$$\frac{n_i}{N_i} = \sigma_i \times \text{constant}$$

If the cost per observation is  $c_i$ , then to maximize the precision for a fixed cost:

$$\frac{n_i}{N_i} = \frac{\sigma_i}{\sqrt{c_i}} \times \text{constant}$$

Here is the solution: to minimize the standard error of the estimator of the population mean, choose the sampling fraction in each stratum to be proportional to the standard deviation of the population in that stratum. If the sampling cost varies between strata, then minimizing the standard error within a fixed cost leads to the solution shown above.

Intuitively, cost aside, this answer makes sense. What it says is spend your effort where the variability is greatest. At an extreme, for example, if everyone within a stratum were equal to each other, we need only take a single observation from that stratum. This is the statistical version of the squeaky wheel getting the attention.

## Sampling Weights

### Sample Weights



If the probability that a person is sampled is not  $f$ , then this must be reflected in the analysis.

e.g. Suppose we have 2 strata, the first is of size  $2n$  and the second is of size  $n$ . If we take a single sample from each stratum ( $x_1$  and  $x_2$ ), then each can be said to "represent" their stratum ( $2n$  and  $n$ ). Any subsequent estimator using both observations must reflect this. For example, if we want to estimate the overall mean, then we would give  $x_1$  weight  $2n$  and  $x_2$  weight  $n$ . So the estimator would be,

$$\bar{x} = \frac{2n x_1 + n x_2}{2n + n}$$

In summary then, let us talk about sample weights. We stated that for simple random samples,  $f$ , the sampling fraction, is the (same) probability that anyone in the population is in the sample. This may vary for other sampling schemes. We deviate from this equality when we use other sampling designs. For example, with a stratified scheme, the probability of each individual being in the sample is not constant, but varies across individuals.

This variability in the probability of being in the sample matters when it comes to the calculation of the estimators of the population parameters. The probability with which a person is chosen to be in the sample should be reflected in the formula for the estimator, just as we saw it was in the calculation of  $\hat{x}$ , above in the case of stratified sampling. This allowed us to ensure that the mean of the sampling distribution of  $\hat{x}$  is  $\mu$ , the population mean—we label this property unbiasedness.

Here is a simple example. If we have two strata, and suppose that the size of the first stratum is  $2n$ , and the size of the second stratum is  $n$ ; so one stratum is twice as big as the other. Now, take a single sample from each of the strata.

Now the observation from the first stratum can be thought of as representing  $2n$  people, whereas the one from the second stratum only represents  $n$  people. So intuitively the one from the first stratum should carry twice as much weight as the one from the second stratum in the estimation of the population mean.

This is the rationale behind sampling weights. So, when analyzing sample surveys, sometimes, such as in DHS surveys, each observation comes with its own sampling weight. In a sense that weight represents how much information this one observation has relative to the others.

How to determine strata?



To **maximize precision** of estimation,  
construct strata so that:

1. Their averages are as different  
as possible.
2. The standard deviations within a  
stratum are as small as possible.

If we have any latitude in determining the strata, then how should we choose them? If we want to maximize the precision of the estimation, then choose the strata to be as homogeneous as possible within a stratum, and as different, or heterogeneous as possible across strata.

## Cluster Sampling

Cluster Sampling



*"Wellbee" says*  
**BE WELL!**  
*take*  
**ORAL**  
**POLIO**  
**VACCINE**



- *tastes good*
- *works fast*
- *prevents polio*

<http://phil.cdc.gov/phil/details.asp?pid=7224>

Another common sampling strategy is to use cluster sampling. The idea of cluster sampling is similar, and at the same time dissimilar, to stratified sampling. The idea is roughly to create

clusters that resemble the population, and then sample a few of them to measure, and then measure each one thoroughly. This typically turns out to be much cheaper than stratified sampling, and not as accurate.

One of the biggest cluster sampling designs in the 20<sup>th</sup> century was used to overcome one of the biggest scourges we had in the 20<sup>th</sup> century, but which declined precipitously in the second half of the 20<sup>th</sup> century, and that is polio. The challenge was how to test the polio vaccine, once it had been introduced? They decided to use a cluster design to test 1.8 million children (440,000 received the vaccine, 210,000 received a placebo, and 1.2 million served as controls) the biggest clinical trial in history<sup>16</sup>.

### Cluster Sampling



Substantial loss in precision but cheaper.

1. Divide the population into clusters (just like strata).
2. Choose which clusters to measure (total or sample).

For maximum precision form clusters so that individuals within a cluster vary as much as possible.  
(Two-stage, ... )

There is a substantial loss in precision in a cluster sample, but it is much cheaper than either simple random sampling or stratified sampling. One starts by dividing the population into clusters, just as we did for stratified sampling. Now, instead of choosing every single one of those clusters, you choose which clusters to measure. So you might choose one, two, three clusters. And then you measure those clusters.

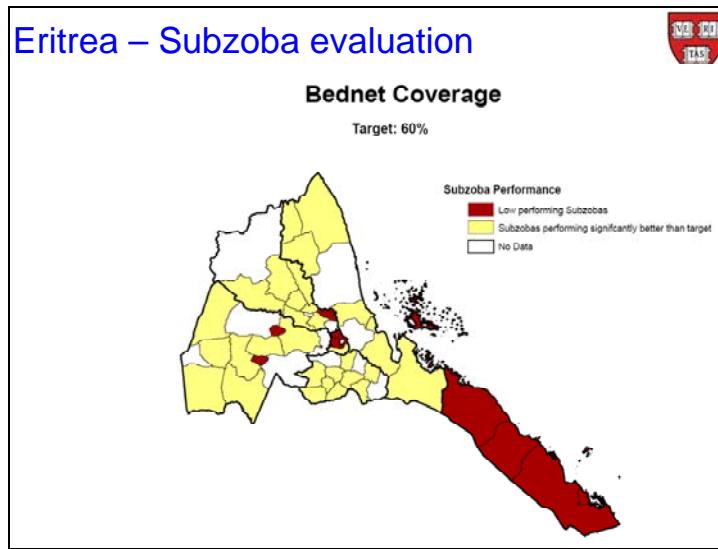
After some thought, if you are only going to measure a few clusters, then if you want a good representation of the population, you would want each of the chosen clusters to be as representative of the population as possible. So you want to maximize variability within a cluster; in contrast to stratified sampling where we sought homogeneity within a stratum.

One can also do what is called two-stage cluster sampling. There you choose some clusters, and then within those clusters you might take a simple random sample and not sample everyone within a cluster.

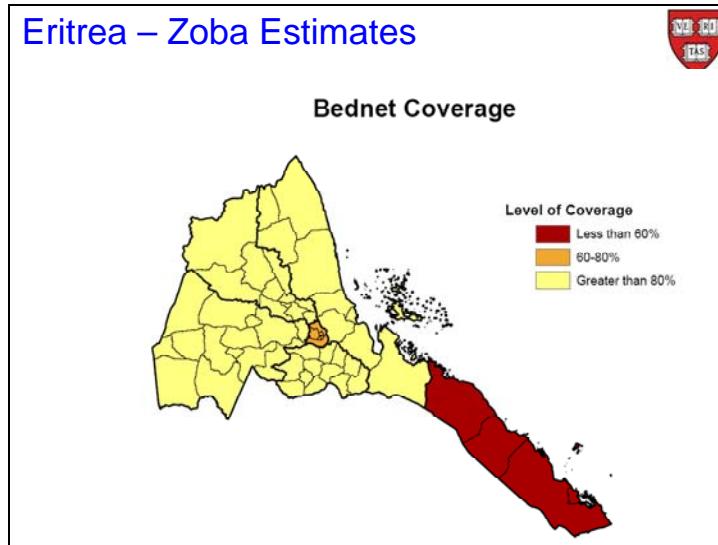
Indeed, you can mix and match as many layers of clustering and stratification and simple random sampling as you wish.

---

<sup>16</sup> American Journal of Public Health, Volume 45, Issue 5\_Pt\_2 (May 1955)



Here is an example of cluster sampling. In the struggle against malaria in Eritrea, they needed to evaluate the use of bed nets countrywide. The target they set was 60% coverage, and they wanted to classify each locale as having reached the target or not. This map is at the subzoba level. The subzobas are classified into one of three colours: red, low performance ( $<60\%$ ) ; yellow, high performance ( $\geq60\%$ ); and white, no data. These classifications were based on a simple random sample in each of the subzobas chosen.



The subzobas can then be combined to form zobas, and the sampled subzobas provide estimates for the zobas.



People in the same cluster are more likely to be similar to one another than to people in a different cluster.

So choosing another person from the same cluster will not be as informative as choosing someone from another cluster.  
 e.g. In a malnutrition study, choosing 3 children in the same household is not as informative as 3 children at random in a village, which in turn is not as informative as 3 children at random in a province, etc.

The **design effect** (DEFF) is the ratio of the variance under the design used, to the variance with simple random sampling,  
 e.g. DEFF = 2 means you need twice as big a sample to get the same variance as you would get with a simple random sample.

One of the problems with cluster sampling is that people in the same cluster are more likely to be more similar to one another than they are to people in different clusters.

So for example, one (cluster) design might consider each household a cluster and we sample every child between the ages of two and five in that household. In contrast, we might consider the household as the sampling unit and only possibly (assuming such a child exists in the household) sample one child between the ages of two and five within each household. It is not unreasonable to believe that kids within a household are more likely to share common characteristics than kids across different households. For example, when doing a vaccination study, if the parents are responsible enough to have one child vaccinated, chances are they will get all the kids vaccinated. Or kids within the same village, when doing a malnutrition study, are more likely to look similar than kids in other villages. So there is a correlation, if you will, between people within the same cluster.

So here is the problem, suppose you have sampled one kid. If you now choose the second kid in the same household or the same village, it is going to be much cheaper, because you do not have to go across town, or to another village to choose another kid. But because of the shared genes or environment, it is not going to be as informative as choosing that second kid across town or in another village.

So how much do you lose? Well, it all depends on how close the kids are to each other; how related they are to each other. If you have a cluster of physicians in a clinic who were all trained at the same school by the same teachers, chances are they are all going to act the same way when faced with the same patient.

To quantify this effect statisticians have come up with what is called the DEFF; called the *design effect*. The design effect is the ratio of the variance under the design used to the variance with simple random sampling. For example, cluster sampling is going to give you a bigger variance because you don't have as good representation, so the DEFF is going to be greater than one.

For example, if the design effect is two, so the ratio of the variances is two, that means you will need twice as big a sample with your cluster design than you would with a simple random sample to get the same precision. So this is the basis for differentiating between designs when considering costs and accuracy.

Design effects of one-and-a-half to two, or three are usually what one experiences in the field when you have a well-designed study. But big DEFFs have been documented too, all the way up to 30, and you can understand why, because of the relationship between people within the same cluster.

### Sources of error



So far we have been talking about biases & imprecision caused by *sampling* variability.

Other sources of error:

1. Selection
2. Non-response
3. Recall
4. Lying

The generic survey involves asking selected people questions and recording their answers. As a result, different types of errors can crop into surveys, beyond what we have learnt to expect due to sampling variability.

Let us consider just four types of possible error: selection bias, non-response bias, recall bias, and the simplest one that occurs when the person surveyed lies.

**Selection**



Literary Digest and Presidential Election of 1936

Landon	1,293,669
Roosevelt	972,897

Final Returns in The Digest's Poll of Ten Million Voters

Good reputation predicting 1920 - 1932 elections.

(Majority of respondents had voted for Hoover!)

See autopsy results.

Selection bias is as the name implies due to the fact that we were biased in who we chose to be in the sample. We have already mentioned the WEIRD sample phenomenon as well as the ANC HIV surveillance. Another, classic example of selection bias is something that happened in 1936. The Literary Digest, a very reputable publication, predicted the outcome of the presidential election of 1936, and they had Landon, who got 1.3 million votes in their survey, to beat Roosevelt, who got only 972,000 votes in their survey. Of course, we know that they were wrong.

As a result of their wrong call, the Literary Digest actually went out of business. They had a terrific track record in predicting correctly every four years from 1916 to 1932, but with this one bad prediction they lost their credibility.

The problem is that this was quite predictable. You see, this was the Depression. They had sent out 10 million cards and about 2 million were returned. First, how did they choose the 10 million to whom they sent the cards and secondly which 2 million responded? Remember this was the depression, people tended to spend their money wisely. Further, they also asked how voters had voted in 1932, so they could see how well the responders represented the voting public.

And this is exactly the problem we had with Ray Pearl when he looked at autopsy results. Who gets autopsy? Do not say dead people. Not everybody who dies has an equal chance of being autopsied. That is what you need to ask yourself with any survey, are we getting a random sample of the population we think we are surveying.

**Non-response**



Survey of sexual abuse of patients by US psychiatrists.

Surveyed	5,574
Responders	1,442 (hostile 19)
Response rate	26%

Admit having sex with patients:

<b>Male</b>	<b>Female</b>
1,057	257
7.1%	3.1%

``Psychiatrist-patient sexual contact: results of a national survey, I: Prevalence''  
N. Gatrell et al. A.J. Psy. 143 (1986) 1126-31

The non-response bias is another important bias. Just as we saw with the Literary Digest, 10 million cards were sent out and only 2 million responded. Are those 2 million who did respond representative of the 10 million? Of the 8 million who did not respond? This is sometimes called the high school reunion effect. Who shows up at a high school reunion? Typically, it is the people who have been successful in life, and want to show off to their high school friends. They typically are not representative of those who do not show up.

Here is another example of this bias. A survey was carried out by Nancy Gatrell to determine the pervasiveness of the problem of psychiatrists sexually abusing their patients. So she sent out 5,574 questionnaires, and 1,442 responded, and that included 19 who just scrawled cuss words on the responses. So the generous response rate was just 26%.

Amongst the responders, 7.1% of the males admitted to having had sexual relations with their patients, and 3.1% of the females admitted to having had sexual relations with their patients.

Is it reasonable to think that these 7.1% and 3.1% rates can be used as estimators of the practice in general? What about the non-responders? Is it possible that one of the reasons why the 74% did not respond might be related to possibly self-admission of something that is expressly prohibited in the Hippocratic oath? In other words, might there be a relation between not responding and whether they had sexually abused their patients, or not?

If the 26% who responded acted in pretty much the same way as the 74% who responded, then there is no bias and the result of the low response rate is that the standard error will be based on 1,442 responders and not 5,574. If, on the other hand, there is some relationship between not responding and the outcome we are measuring, then the estimators based on the 26% are biased estimators of what we are seeking to measure.

### Benjamin Franklin 1759 :



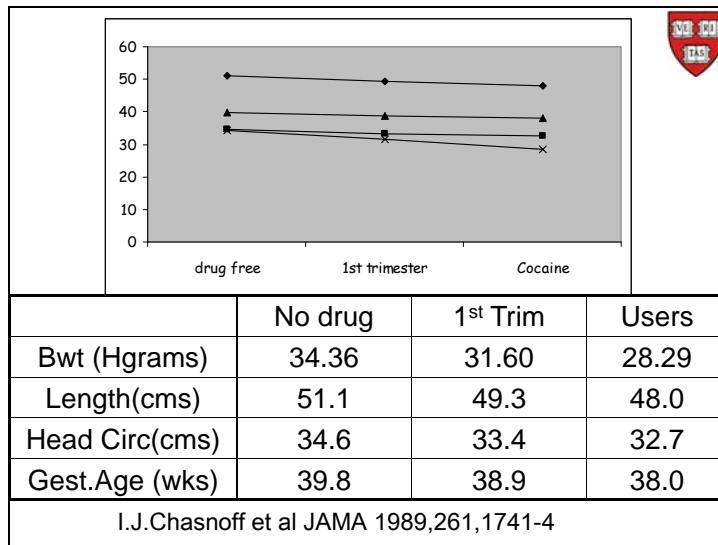
As the practice of Inoculation always divided people into parties, some contending warmly for it, and others as strongly against it; the latter asserting that the advantages pretended were imaginary, and that the Surgeons, from view of interest conceal'd or diminish'd the true numbers of deaths occasion'd by Inoculation, and magnify'd the number of those who died of the Small-pox in the common way: It was resolved by the Magistrates of the town, to cause a strict and impartial enquiry to be made by the Constables of each ward, who were to give in their returns upon oath; and that the enquiry might be made more strictly and impartially, some of the partisans for and against the practice were join'd as assistants to the officers, and accompany'd them in their progress through the wards from house to house. Their several returns being receiv'd and summ'd up together, the numbers turn'd out as follows,

The trustworthiness of surveys is apparently an old problem. Benjamin Franklin<sup>17</sup> reports in 1759 on a survey he did to gauge the effectiveness of smallpox inoculation. The idea of whether or not to inoculate was a rather heated topic at the time, as it is again today! So they decided to do a study. The solution was, "So it was resolved by the magistrates of the town to cause a strict and impartial inquiry to be made by the constables of each ward. So they sent the constables out to do the survey."

Had the Small-pox in the common way		Of these died		Received the distemper by Inoculation		Of these died	
Whites	Blacks	Whites	Black	Whites	Black	Whites	Black
5059	485	452	62	1974	139	23	7
[9.3%]				[1.4%]			
<b>Survey following Smallpox outbreak in Boston in 1753-4, as reported by Benjamin Franklin; percentages added.</b>							

<sup>17</sup> Benjamin Franklin, Some account of the success of inoculation for the smallpox in England and America, London W.Strahan, 1759

Here are the results of the study, and they seem convincingly in favor of inoculation. And we know that the constables took care of this data. So we can trust it.



Here is another study where lying plays a role, although the rationale for lying is elusive.

The study was done on pregnant women to try and determine the impact on birth of the use of illicit drugs (such as marijuana and cocaine) during the pregnancy, and here are the results for cocaine.

The authors obtained all the right permissions to make the study ethical including to have the mothers tested for cocaine every time they came in for an antenatal visit.

Simultaneously, as they were doing the testing, they also asked the mothers about their drug use. As a result they had two answers to the drug usage question: one verbally provided by the mother (who had also given permission to be tested), and the other from the biological test.

The analysis of pregnancy outcomes was performed as a comparison of three groups: (i) those mothers who partook of no illicit drugs, (ii) those who quit illicit drug usage before the first trimester of their pregnancy, and (iii) those who used illicit drugs beyond the first trimester of their pregnancy.

They looked at these four pregnancy outcomes: (i) birth weight of the baby, (ii) length of the baby, (iii) head circumference of the baby, and (iv) the baby's gestational age. Above we see the four results, and they all show a significant ( $\alpha = 0.05$ ) trend, when the classifications into the three groups is done on the basis of the biological tests.

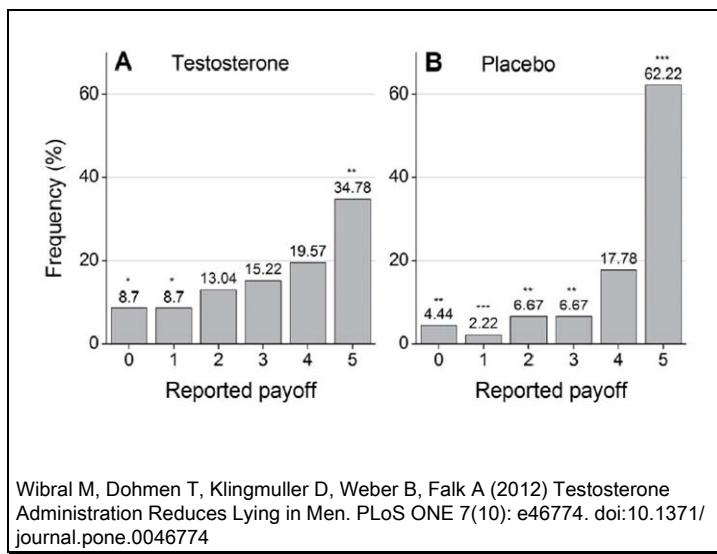
Yet when they redid this analysis, but using this time basing the classification into the three groups on the basis of the mothers' verbal responses, the results lost their significance. This is rather surprising that we reach qualitatively different conclusions because of lying, even when there really was not much incentive to lie.

### Be a man: tell the truth

Higher levels of [testosterone](#) have been implicated in some negative qualities associated with men, such as impaired empathy. Recent research, though, is showing some upsides of the same hormone; now, a new study finds that [testosterone inhibits lying](#). Men were given either testosterone or a placebo and were later asked to roll a six-sided die in private, and whatever number the men reported rolling would determine their payoff for the experiment. Men with higher testosterone reported they'd scored payoffs far closer to the rate of probability. *While men who were administered the placebo reported rolling the highest payoff 62% of the time, those who were administered testosterone reported rolling the highest payoff just 35% of the time.*

*Wibral, M. et al., "Testosterone Administration Reduces Lying in Men," PLoS ONE (October 2012).*

AN article that just appeared led to the above headline, "Be a man," it says. "Tell the truth." The study<sup>18</sup> was of 91 men where roughly half the men got a shot of testosterone, and the other half were administered a placebo. They then had the men roll a die to get a payoff whose size depended on the role of the die. The catch was that no one watched the role of the die and each man reported the value of the die when claiming the payoff.



Here is the distribution of the payoffs. They were given fair die, so the expected value was that all these bars should be approximately the same size. The contention of the paper is that there

<sup>18</sup> Wibral M, Dohmen T, Klingmu N Iler D, Weber B, Falk A (2012) Testosterone Administration Reduces Lying in Men. PLoS ONE 7(10): e46774. doi:10.1371/journal.pone.0046774

is a significant difference between these two bar graphs, with the placebo group showing a greater tendency to lie, as measured by too large a bar on the right hand side of the graphic.

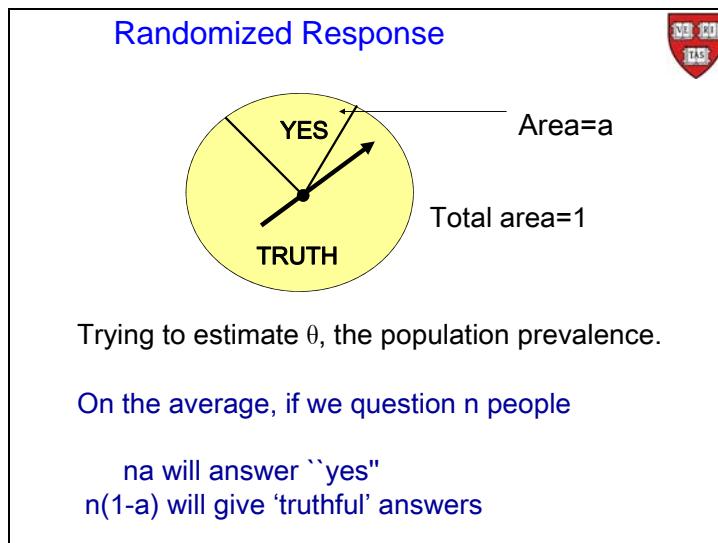
Column = die face reported							
							. tabi 4 4 6 7 9 16 \2 1 3 3 8 28 , row chi
row	1	2	3	4	5	6	Total
1	4 8.70	4 8.70	6 13.04	7 15.22	9 19.57	16 34.78	46 100.00
2	2 4.44	1 2.22	3 6.67	3 6.67	8 17.78	28 62.22	45 100.00
Total	6 6.59	5 5.49	9 9.89	10 10.99	17 18.68	44 48.35	91 100.00

Pearson chi2(5) = 8.3882 Pz = 0.136

1 = Testosterone group  
2 = Placebo group

Their analysis is somewhat suspect, given that they perform all sorts of subtable analyses possibly in order to get significance? Anyway, when I fed the above table to Stata you can see that the Chi-squared analysis shows the differences between the two groups to be insignificant.

## Randomized Response



So is there anything we can do to overcome lying? Possibly not.

Ex post adjustments, for example take 80% of the ANC prevalence estimates and use those as population estimates, are unsatisfactory ad hoc solutions, especially if the truth can go in either direction. For example, in some nutritional studies, people who think they are too thin will tell you that they eat more than they actually have, and at the other end of the spectrum, people who think they are overweight will under-report how much they have eaten.

There is one design, introduced in the 1950s, that utilizes probability to mask the individual response, and thus elicit more truthful behaviour. The idea is simple, before asking a question a randomization device, hidden from the interviewer, is introduced and this device tells the person being interviewed how to respond. Since the device is hidden from the interviewer, the individual response cannot be linked to the individual being interviewed with any certainty. This thus protects the privacy of the individual, yet at the same time we can gauge aggregate behavior, which is our original intent in taking a survey, anyway!

For example, suppose you are carrying out a survey and you ask the person being interviewed to spin a pointer, which is hidden from your view, before responding to a question. The spinner can come to rest in one of two areas: if in the one, then the person being interviewed is instructed to answer, yes, whatever the question; and if the pointer falls in the other area, then the person should answer the question truthfully.

That way, if the area of the “yes” part is  $a$ , if the area of the whole is 1, and if we can consider this to be an unbiased spinner, then the probability that the person being interviewed is instructed to say “yes” by the pointer, is  $a$ . We can think of this  $a$  as the obfuscation factor; the larger we make  $a$ , the more uncertainty we introduce into the study. This would argue for a small  $a$ , but too small an  $a$  will defeat the purpose of the obfuscation, which was to convince the person being interviewed that we cannot link them to the truthful situation, unless it is a “no”. Presumably the “no” is the non-controversial label.

This should remind you of an imperfect diagnostic test. We have introduced a specificity of  $a$  by making asking that proportion of those who should answer “no”, to answer “yes”. Yet at the same time we hope that this will increase the sensitivity of this device; ideally to one. We are introducing some imprecision, hoping to gain some veracity.

## Randomized Response



So, how many will answer "Yes"? (m)

On average:

na because of the dial  
plus

$n(1-a) \theta$  because answering truth

In a particular sample:

$$m = n a + n(1-a) \hat{\theta}$$

$$\hat{\theta} = \frac{\{m/n\} - a}{1 - a}$$

The modification we need to make to our answers are the same as those we made to accommodate an imperfect diagnostic test.

## Appraisal



Advantage :

Lessens "lying" fraction

Disadvantages:

1. Increases variability of estimator (versus idealized "everyone telling truth")
2. Costly --- difficult to use in mail survey

The advantage, we hope, is that this device lessens the lying fraction. This disadvantage is that it increases the variability but with respect to what? With respect to the idealized 1, which does not exist, which is when everybody tells you the truth.

It is a costly solution. It is difficult to use in a mail survey, for example. It requires explaining, and that may introduce errors. It has been used, but I am surprised it has not been used more often.

## Survey



Phone survey to determine illicit marijuana use --- 1986.

Estimates of prevalence:

Direct questioning estimate 40%

Randomized response estimate 64%

This design was used in a phone survey in 1986, where the person being interviewed was told to get three pennies and spin them. If the coins landed as all heads, then the person interviewed was instructed to say, yes, whatever the question. If the coins landed all tails, the answer was to be, no. The person interviewed was instructed to answer the question truthfully for any other configuration of heads and tails. This design is a little bit different than the one above; 1/8 of the people will say "yes," 1/8 of the people will say "no," and the remainder will (hopefully) tell the truth. So we not only have a specificity of 7/8, but now we also have a sensitivity of 7/8, by design, assuming everyone tells the truth.

They used this design to survey the usage of marijuana (which was illegal then). They first carried out the survey without the coins and 40% of those interviewed said they used marijuana. They then redid the survey, but this time they incorporated the flipping of the coins. With this randomized response design, the estimate of marijuana use increased from 40% to 64%, a 60% increase.

Marcello Pagano

# [JOTTER 9 CORRELATIONS AND NON-PARAMETRIC TESTS]

Pearson's Correlation Coefficient, Spearman's correlation coefficient, sign test, Wilcoxon signed rank test, Wilcoxon rank sum test.

**Relationships between variates**

The odds ratio is a means of quantifying a relationship between two dichotomous variates:

e.g. exposure (yes,no) and disease (yes,no)

We now continue our exploration of the relationship between two variables. Today we look at the correlation coefficient to attempt to quantify the relationship between two continuous variables, but before doing that let us briefly review what we did to measure the relationship between two dichotomous variables. To quantify that relationship we looked at the odds ratio.

Key

frequency	row percentage	column percentage
-----------	----------------	-------------------

OR=2.11

What is your sex?	When you wash your hair in the shower, do you...		Total
	Face away	Face the	
Female	2,378 60.91 55.87	1,526 39.09 37.46	3,904 100.00 46.87
Male	1,878 42.43 44.13	2,548 57.57 62.54	4,426 100.00 53.13
Total	4,256 51.09 100.00	4,074 48.91 100.00	8,330 100.00 100.00

Pearson chi2(1) = 283.5208 Pr = 0.000

Let us digress a little and look at an example of a study of the relationship between dichotomous variables. These data come from the survey you were asked to fill in if you were one of the early enrollers in this course. You guys were asked about rinsing your hair in the shower in the morning and whether you faced the showerhead or do you turn around and rinse your hair from the back of the head?

Here is how 8,330 of you responded: Of the female responders, 60% answered that they faced away, and 40% said that they faced the shower head. On the other hand, the males, responded

almost the opposite: 42% faced away, and 58% faced the shower head. I do not know why, but every time I ask this question in a class, it comes out this way. It is a 60/40 split, one way or the other.

A number of people have suggested that this behavior is perfectly predictable because it has only to do with hair length; with longer hair it is easier or preferable to rinse facing away from the shower head, and women tend to have longer hair than men. Puzzle solved.

To empirically check this theory, we also asked you to tell us whether you considered yourself to have long hair or not.

Key		OR=0.67			
		When you wash your hair in the shower, do you...			
		Face away	Face the	Total	
No		2,649 47.79 62.02	2,894 52.21 70.78	5,543 100.00 66.30	
Yes		1,622 57.58 37.98	1,195 42.42 29.22	2,817 100.00 33.70	
Total		4,271 51.09 100.00	4,089 48.91 100.00	8,360 100.00 100.00	
Pearson chi2(1) = 71.6252 Pr = 0.000					

And you responded as shown in the above table. Of those who did not consider themselves to have long hair, 52% faced the shower and 48% did not. Of those who considered themselves to have long hair, 49% faced the shower, 51% did not. So not quite as sharp a distinction as with sex—indeed, the odds ratio is 1.48 this time, instead of the 2.11 odds ratio above, when the dichotomy was based on sex—but still significantly different from one, with a reported p-value of 0.000.

So now we are confused, is it sex-related or hair-length-related?



-> Doyouconsideryourhairtobe = Yes			
Key			
	frequency	row percentage	column percentage
What is your sex?	When you wash your hair in the shower, do you...		
Female	Face away	Face the	Total
Female	1,418 61.41 87.86	891 38.59 75.13	2,309 100.00 82.46
Male	196 39.92 12.14	295 60.08 24.87	491 100.00 17.54
Total	1,614 57.64 100.00	1,186 42.36 100.00	2,800 100.00 100.00
Pearson chi2(1) = 76.6095 Pr = 0.000			
-> Doyouconsideryourhairtobe = No			
Key			
	frequency	row percentage	column percentage
What is your sex?	When you wash your hair in the shower, do you...		
Female	Face away	Face the	Total
Female	958 60.14 36.32	635 39.86 22.03	1,593 100.00 28.85
Male	1,680 42.77 63.68	2,248 57.23 77.97	3,928 100.00 71.15
Total	2,638 47.78 100.00	2,883 52.22 100.00	5,521 100.00 100.00
Pearson chi2(1) = 137.0243 Pr = 0.000			

We are, of course, concerned with the Yule effect (or Simpson's paradox), so let us investigate this a little deeper. Let us first look at the people who considered their hair to be long.

There were 2,800 of you who did, and when we looked at the sex-related odds ratio we find it is 2.4. (61% of females faced away and 40% males faced away.)

Amongst the 5,521 of you who considered that you did not have long hair, the sex-related odds ratio was 2.0. (60% of females faced away and 43% males faced away.)

Neither of these is too far from the 2.11 odds ratio when length of hair was ignored, above, and both odds ratios associated with a reported p-value of 0.000 for the null hypothesis that the odds ratio is one. So the sex difference maintains in both groups.

And so the same sort of sex-related relationship maintains amongst those who consider their hair be long, as in the group who did not consider their hair to be long. So hair length does not seem to matter for this relationship.



-> What is your sex = Female				-> What is your sex = Male					
Key		frequency		Key		frequency			
		row percentage				row percentage			
Do you consider your hair to be long?		When you wash your hair in the shower, do you...							
No		Face away	Face the	Total		Total			
No		958 60.14 40.32	635 39.86 41.61	1,593 100.00 40.83					
Yes		1,418 61.41 59.68	891 38.59 58.39	2,309 100.00 59.17					
Total		2,376 60.89 100.00	1,526 39.11 100.00	3,902 100.00 100.00					
Pearson chi2(1) = 0.6422 Pr = 0.423									
OR=0.95				OR=1.1					
Do you consider your hair to be long?		When you wash your hair in the shower, do you...							
No		Face away	Face the	Total		Total			
No		1,680 42.77 89.55	2,248 57.23 88.40	3,928 100.00 88.89					
Yes		196 39.92 10.45	295 60.08 11.60	491 100.00 11.11					
Total		1,876 42.45 100.00	2,543 57.55 100.00	4,419 100.00 100.00					
Pearson chi2(1) = 1.4524 Pr = 0.228									

One last analysis in order to convince ourselves: let us look within each sex. First with females we see that the hair-length related odds ratio is 0.95, and with males the hair-length related odds ratio is 1.1. In neither case would we reject the null hypothesis that the odds ratio is one. This should convince us that it is not hair-length related, once we know the sex of the responder. So we are still confounded by the reason for this sex-related phenomenon of whether we face the shower head or not when rinsing our hair.

One last aside from this last slide: you may hear skeptics complain that if your sample size is large enough you can always find significance. Here we see two very large samples, 3,902 in the one sample and 4,419 in the other and in neither case did we reject the null hypothesis that the odds ratio is one.

## Relationships between variates



The **odds ratio** is a means of quantifying a relationship between two dichotomous variates:

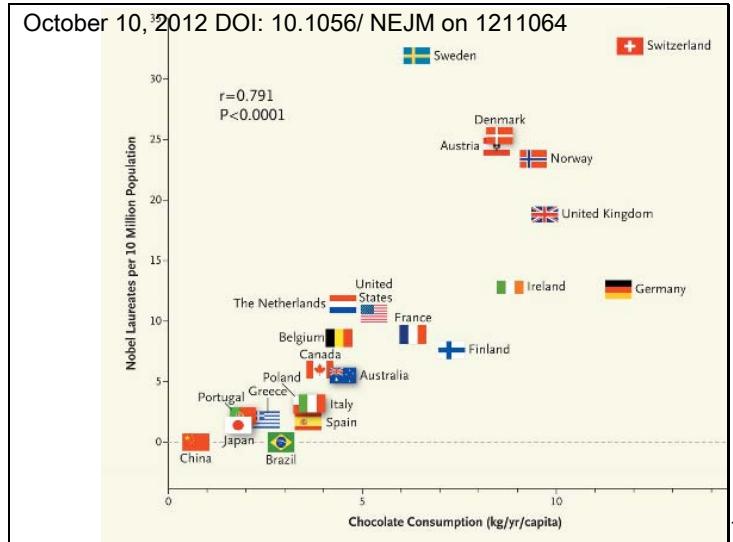
e.g. exposure (yes,no)  
and disease (yes,no)

What if the two variates are not dichotomous? – Generalize.

Quantify

Let us now turn our attention to quantifying the relationship between two variables when they are not both dichotomous. Let us first look at when both variables are continuous.

Pearson's Correlation Coefficient



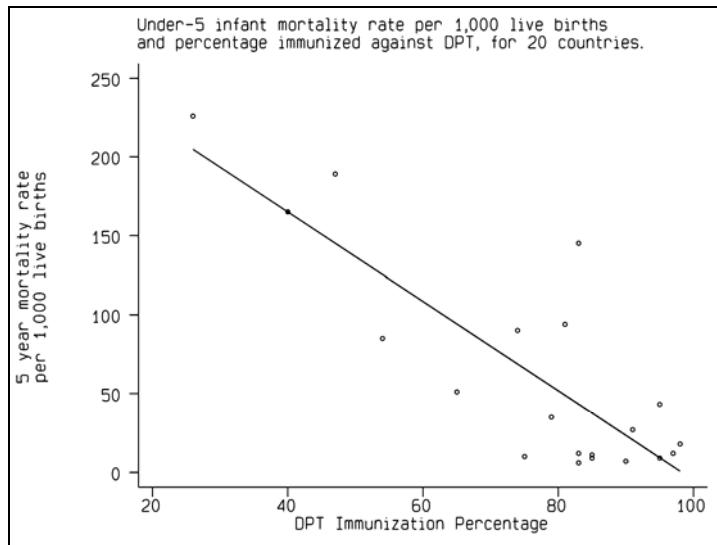
This graph just appeared in the New England Journal. On the bottom axis we have the per capita chocolate consumption—kilograms per year. On the vertical scale we have the number of Nobel laureates per 10 million inhabitants in a country. And what we see is a quasi-linear relationship between chocolate consumption and the number of Nobel laureates per country.<sup>1</sup>

This is your classical correlation analysis. The author found a very high correlation, and deduced, tongue in check, no doubt, that chocolate has some impact on your brain cells, to explain the correlation.

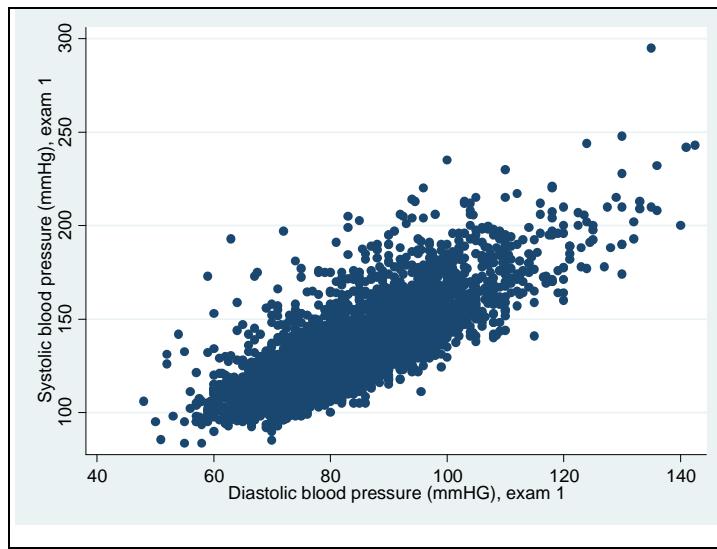
Now, there is a problem with this study. Not to be a wet blanket, but if you want to be serious about this, this study suffers from what is called the ecological fallacy. We return to this issue before this week is out.

---

<sup>1</sup> Occasional Notes: **Chocolate Consumption, Cognitive Function, and Nobel Laureates**, Franz H. Messerli, M.D. October 10, 2012 DOI: 10.1056/NEJMOn1211064

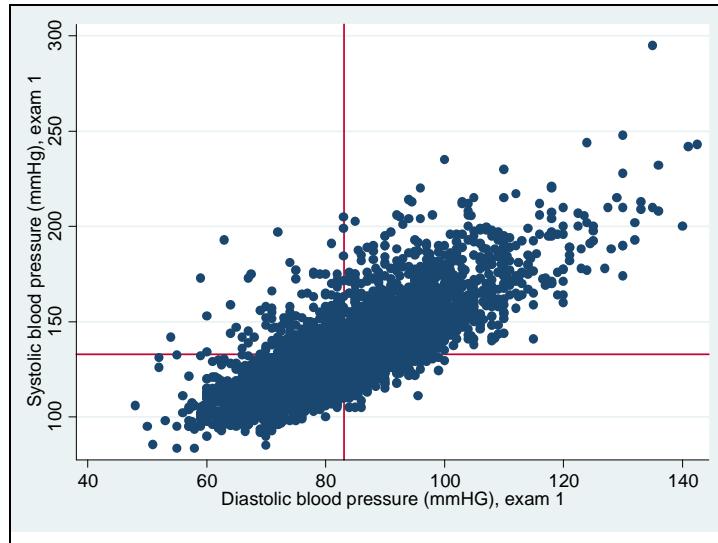


Here is another example. Once again we are dealing with countries. And here we have the five year mortality rate per 1,000 live births on the vertical axis. That is an indicator often used to monitor the health level of a country. On the horizontal we have the DPT (diphtheria, pertussis and tetanus) immunization percentage in a country. This too is a country level graph, and whatever conclusions we can draw from this should be at the country level.

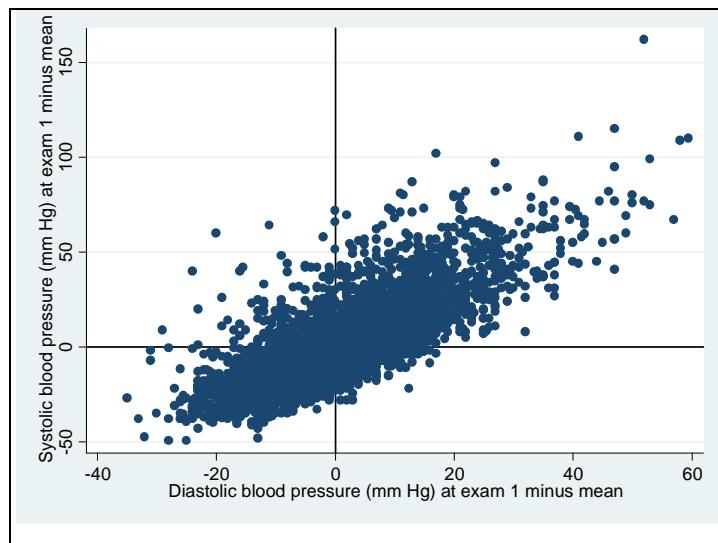


before. We see that as the diastolic blood pressure goes up, the systolic blood pressure goes up, and vice versa. They vary together. They co-vary. They both vary. But they co-vary. They both go off in the same direction, hand in hand.

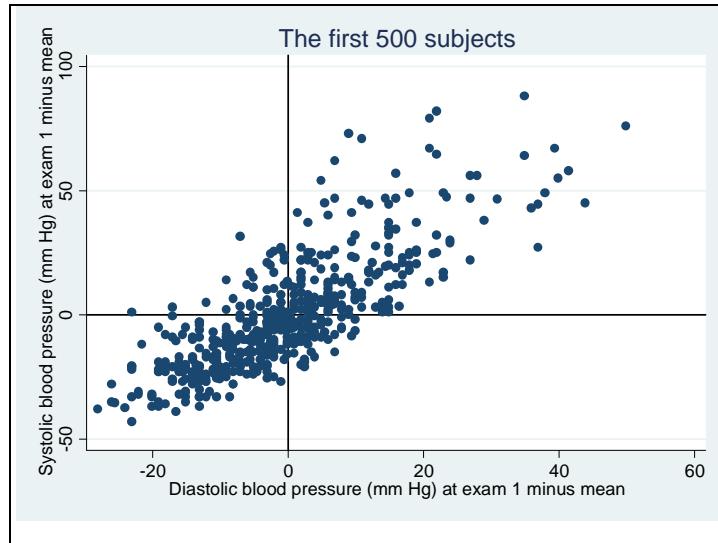
Can we quantify this behavior?



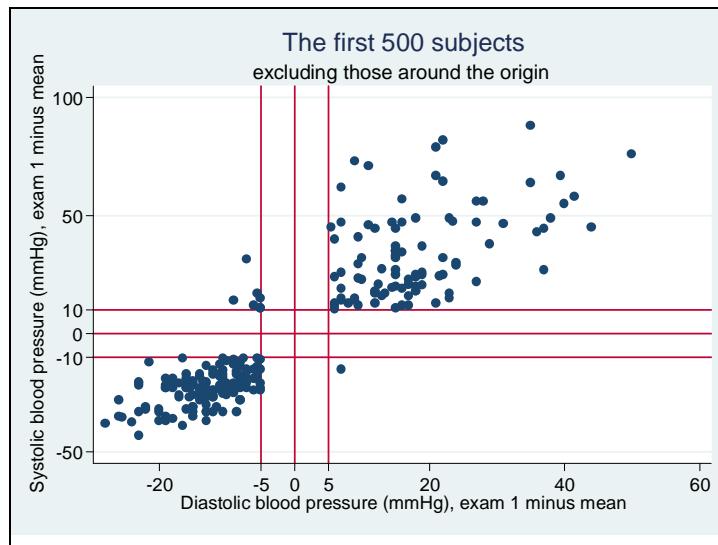
To this end, consider centering the graph by introducing axes at the means of the two. So let me subtract the mean of the systolic pressures from the systolic readings and the mean of the diastolic pressures from the diastolic readings.



After relabeling the axes we get this picture. Since there are too many dots on this graph let us just consider the first 500 subjects.



This allows us to get a better picture of what is going on.



Let us further attempt to understand what is happening by excluding all the points with small, in absolute value, components to get the picture above. It shows four quadrants only, and those are points both of whose components are sizable. Concentrating on just these values we see that most of them fall into two quadrants where the signs of both the components are the same: large positive diastolic readings go with large positive systolic readings, and large negative diastolic readings go with large negative systolic readings. You should not erase points, of

course, but this is just temporarily and for expository purposes. The challenge remains of capturing this behavior in a single number.

	diabp1	mdiabp1	sysbp1	msysbp1
1.	70	-13.08356	106	-26.9078
2.	81	-2.08356	121	-11.9078
3.	80	-3.08356	127.5	-5.4078
4.	95	11.91644	150	17.0922
5.	84	.91644	130	-2.9078
6.	110	26.91644	180	47.0922
7.	71	-12.08356	138	5.0922
8.	71	-12.08356	100	-32.9078
9.	89	5.91644	141.5	8.5922
10.	107	23.91644	162	29.0922
11.	76	-7.08356	133	.0922
12.	88	4.91644	131	-1.9078
13.	94	10.91644	142	9.0922
14.	88	4.91644	124	-8.9078

Mean  
 diabp1=83  
 sysbp1=133

Let us return to looking at these numbers in a table rather than on the graph. Here are the first 14 values. The columns labeled diabp1 and sysbp1 contain the original data. These are transformed by subtracting the respective means to get the columns mdiabp1 and msysbp1. Concentrating on these two columns, we see that large negative values are paired as are large positive values, by and large.

To capture this behavior we can think back to the variance.

$\bar{x} = 2.95$	FEV <sub>1</sub>	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	2.30	-0.65	0.423
	2.15	-0.80	0.640
	3.50	0.55	0.303
	2.60	-0.35	0.123
	2.75	-0.20	0.040
	2.82	-0.13	0.169
	4.05	1.10	1.210
	2.25	-0.70	0.490
	2.68	-0.27	0.073
	3.00	0.05	0.003
	4.02	1.07	1.145
	2.85	-0.10	0.010
	3.38	0.43	0.185
	Total	0.00	4.66

## Variance



$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \geq 0$$

e.g.

$$= \frac{4.66}{12} = 0.39 \text{ liters}^2$$

With the variance we multiplied each deviation by itself and then found the average squared-deviation.

## Covariance



$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Covariance} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Now we have two variables, so instead of squaring the one deviation and getting the variance, we can multiply the two deviations, average them out and get what is called the *covariance*. This tells us how much the x and the y, co-vary, or vary together.

Pearson's Correlation Coefficient.  
Product moment correlation.



n pairs:  $(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)$

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

$$= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

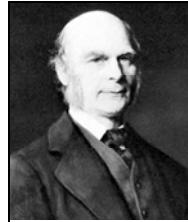
Note that  $-1 \leq r \leq 1$

0.7842 diastolic and systolic blood pressure for FHS at first visit

If we also standardize our variables by dividing the deviations by their respective standard deviations, then we get  $r$ , or what is called the *sample correlation coefficient*, or the *product moment correlation coefficient*, or *Pearson's Correlation Coefficient*.

One immediate result of this standardization is that, from the Cauchy inequality,  $-1 \leq r \leq 1$ .

In the example with diastolic and systolic blood pressure at visit one, we get that the correlation coefficient is 0.7842—very high and close to 1, but not quite 1.




**Correlation**

Francis Galton  
1822-1911

Karl Pearson  
1857-1936

Measure of **linear** relationship between two continuous random variables.

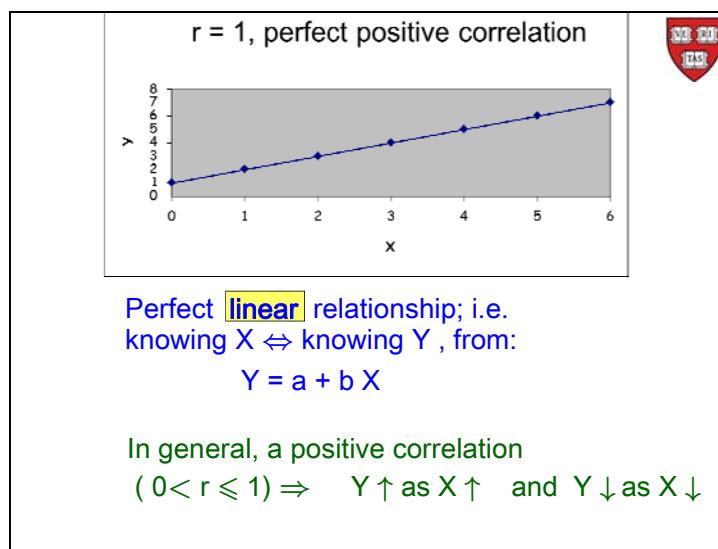
Correlation Coefficient

$$\rho = \text{average} \left[ \left( \frac{X - \mu_X}{\sigma_X} \right) \left( \frac{Y - \mu_Y}{\sigma_Y} \right) \right]$$

Pearson did the early mathematical work on correlation, but its introduction is due Francis Galton. Galton was Charles Darwin's cousin, and the story goes that he was a little bit jealous of his cousin's fame. Galton also invented the word eugenics, and left us with that perversion of his cousin's research.

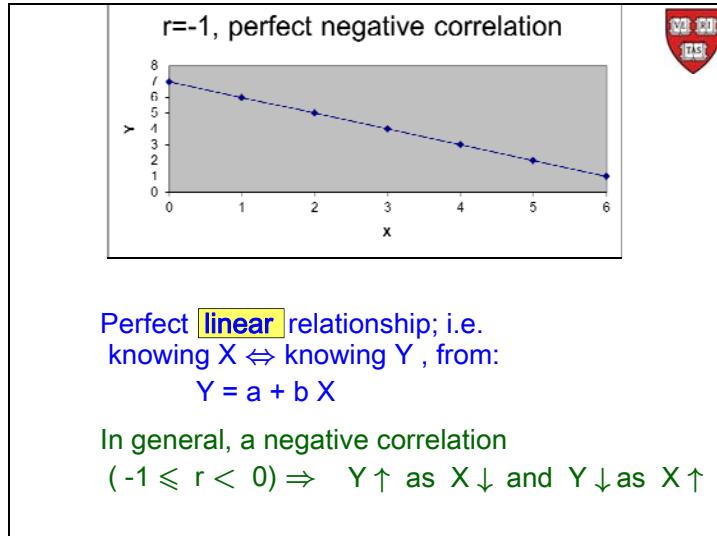
Take care that this coefficient is a measure of *linear* relationship. This word linear people sometimes ignore, but you do that at your own peril. As we discuss this coefficient, you should appreciate what that qualifier means.

Within the population, we follow our tradition of using Greek letters to label parameters. Here we use the letter  $\rho$  to denote the population parameter that we are attempting to estimate by using the  $r$  from a sample.



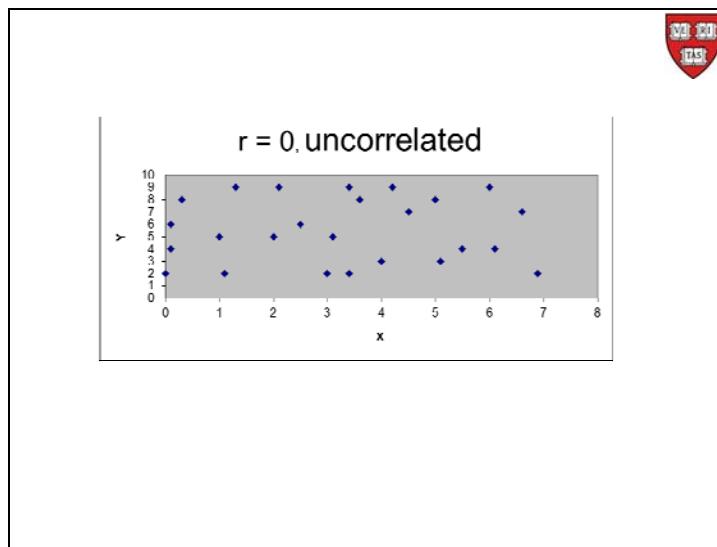
There are some special, extreme values of  $r$ , and  $\rho$ , that are worth noting. One is at the right end of the scale, at the value 1. Then we have perfect positive correlation, and that is the largest correlation one can have. What it actually means is that there is a straight line relating the two variables. They are thus basically a single random variable. The slope  $b$  is positive, so the variables increase, and decrease, together.

In general, a positive correlation means that the two variables tend to increase, and decrease, together. The relationship is only perfect at the extreme of one.

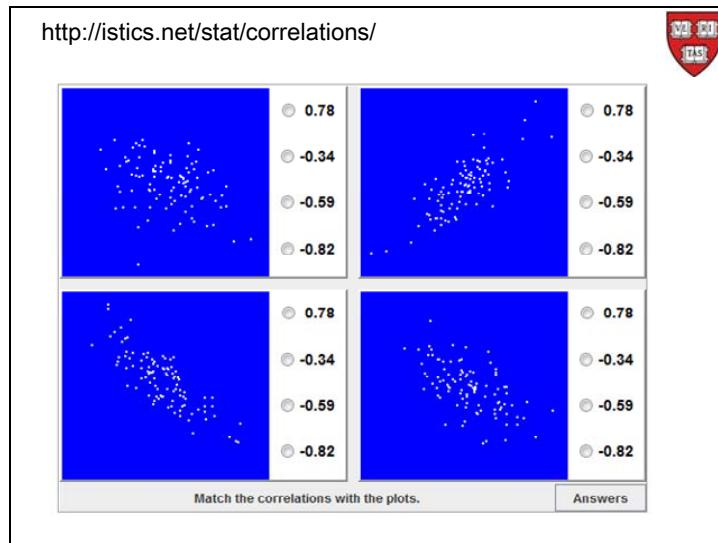


The other extreme is when  $r$ , and  $p$ , equal minus one. This happens when we have a perfect relationship, as with the situation above when the correlation was one, except that this time the variables go in opposite directions; so the slope  $b$  is negative, and thus one variable goes up as the other goes down

In general, when you do not have this perfect negative correlation, but the correlation coefficient is still negative, then on average one variable goes up as the other one comes down.



The other special value the correlation coefficient can take is right in the center, and that is when it is equal to zero. When that happens we say that the variables are *uncorrelated*. This might remind you of independence, but do not be confused. What this means is that, and here comes that important word again, there is no *linear* relationship between the variables.



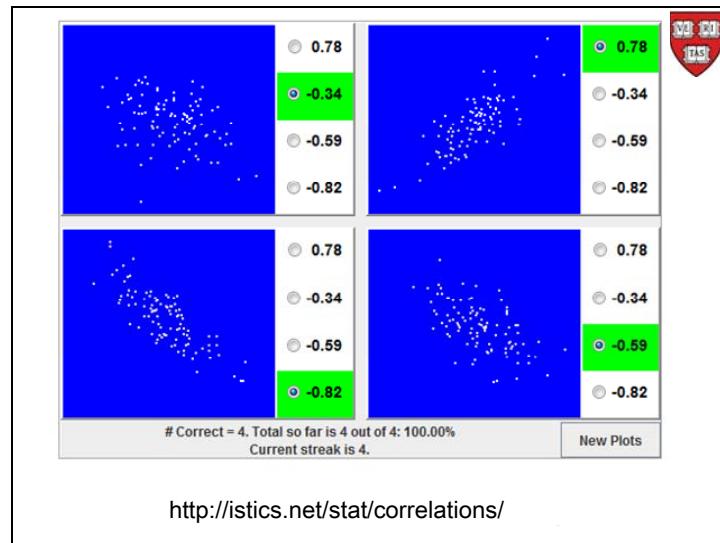
What I would like you to do is go to this website<sup>2</sup>. There you will find a game I find mesmerizing. They throw up four scatter plots of clouds of points. On the right hand columns of each plot you find four correlation coefficients, the same four at each plot. The game is to link a graph with a correlation coefficient, and you get scored on how many you get correct. That way you can get a feel for what the correlation coefficient is measuring.

For example, with this panel, we see that the top right-hand corner is sloping positively, so that one should get the 0.78 choice. The others get gradually more negative from -0.34 to -0.59 to -0.82. The trick in making the identification is to think back to the three graphs above. For the extremes at plus or minus one we had no variability around the line, whereas at a correlation of zero we had maximal variability. So grade these three according to the amount of variability in the scatters: the top left-hand corner probably has the maximum variability, so it should be identified with the -0.34. Of the bottom two, the one on the left looks tighter than the one on the right.

You make your choices, and then Answers.

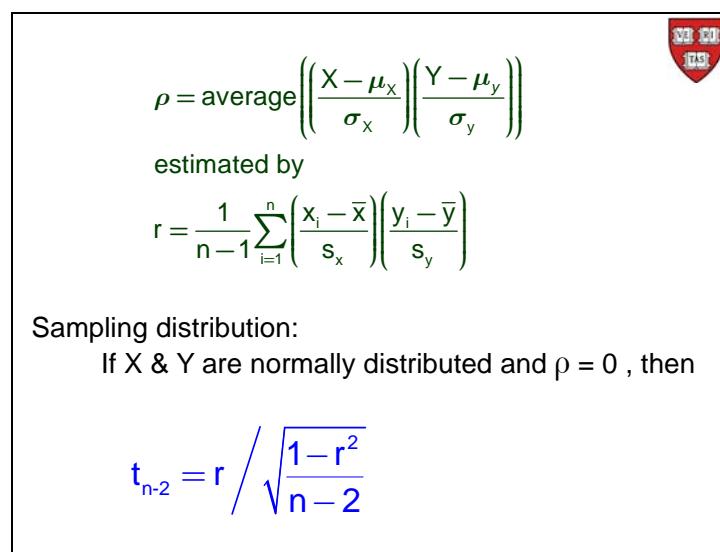
---

<sup>2</sup> <http://istics.net/stat/correlations/> or <http://www.istics.net/Correlations/> if you prefer Java.



In this case I was correct on all four. So after all these years I was able to be right! You go head and have some fun and get some feel for how sample correlation coefficients vary.

### Inference on $\rho$



We have a population parameter  $\rho$ , and we would like to make inference about this parameter on the basis of a sample from this population. Suppose we wish to test a hypothesis about  $\rho$ . The only hypothesis we cover in this course is the one that says that  $\rho = 0$ . So we test the hypothesis that two variables, X and Y, are uncorrelated.

So what we need is the sampling distribution of  $r$  when  $\rho$  is equal to 0. The slide gives us the sampling distribution if the two variables, X and Y, are normally distributed.



e.g.  $r = -0.829$  for DPT example

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

$$= -0.829 \sqrt{\frac{20-2}{1-(-0.829)^2}} = -6.29$$

versus t with 18 degrees of freedom, so  $p < 0.001$ .

So reject  $H_0 : \rho = 0$ .

In the DPT example, above, we have that  $r$  was -0.829, so our  $t=-6.29$  and the p-value is less than 0.001, and so we would reject, at the 5% level, the null hypothesis.

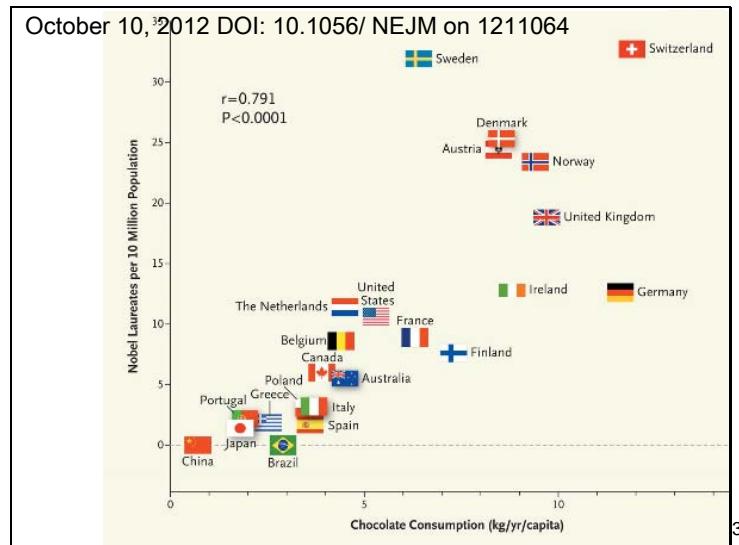
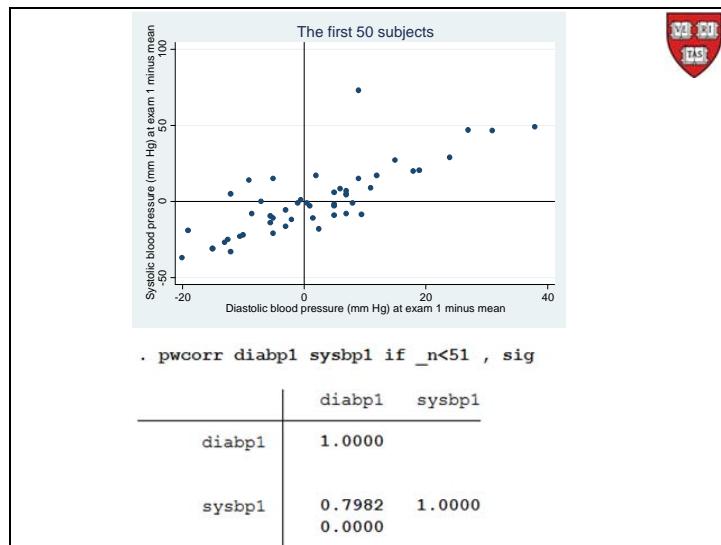


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

In the top left-hand side of the graph we see that  $r=0.791$  and the p-value, associated with the null that  $p=0$ , is less than 0.0001. So we would reject that hypothesis.



<sup>3</sup> Franz H. Messerli, M.D. Occasional Notes, **Chocolate Consumption, Cognitive Function, and Nobel Laureates**, October 10, 2012 DOI: 10.1056/NEJMOn1211064

Choosing the first 50 subjects in our Framingham heart study, just as an exercise, we test the hypothesis that  $p = 0$ , using the command `is pwcorr`, and we get that the correlation is 0.7982 and the p-value is underneath it. It is 0.0000. So the p-value is less than 0.00005. So on the basis of these 50 observations (and this is not a proper study since this is not a random sample, just me exercising pedagogical license) we would reject the null hypothesis that diastolic and systolic blood pressure at visit one were uncorrelated.



### Misconceptions:

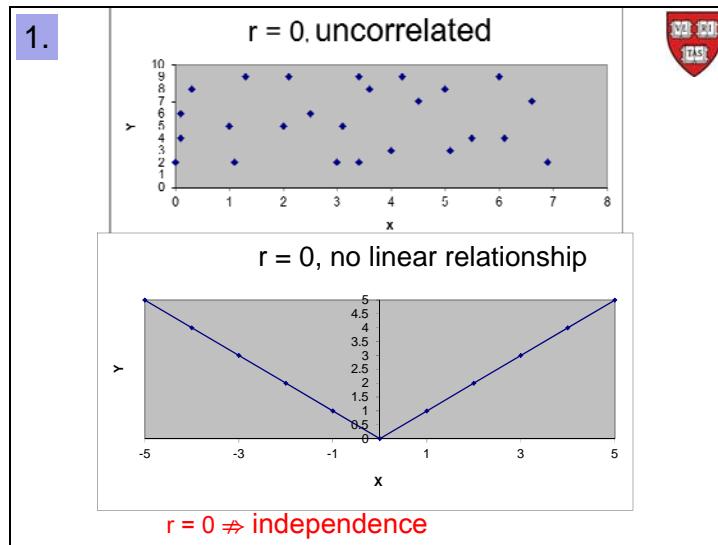
1. Correlation = 0 implies independence
2. Correlation implies causality
3. Ecological Fallacy

Now that you have seen the correlation coefficient, we look at three important misconceptions some people have about correlations.

The first misconception is that when the correlation equals zero, this implies independence. Correlation equals zero does not imply independence, the two variables are merely uncorrelated.

The second misconception is a very touchy one, and that is that correlation implies causality. It does not. It simply implies correlation. The two are not synonymous.

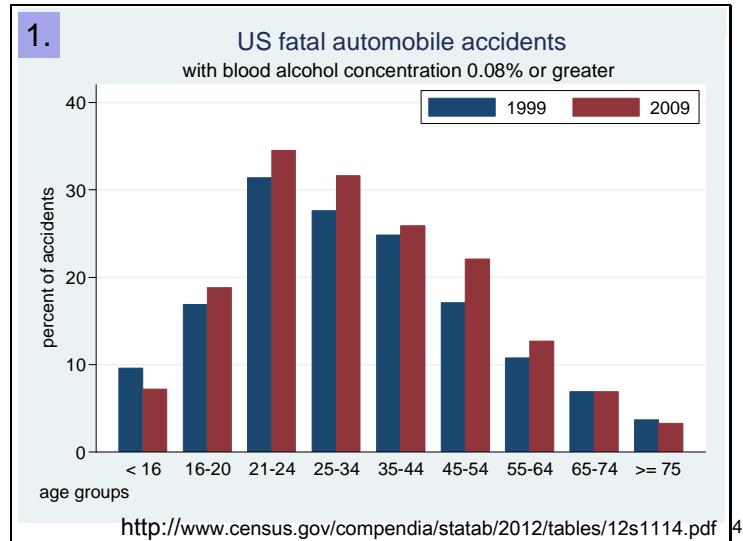
And the third misconception is that correlation at the ecological level implies correlation at the personal level—the ecological fallacy we have just seen. Let us take these misconceptions one by one.



You get an  $r=0$  from the top graph when the points in the scatter have no discernible pattern, but you also get zero correlation in the bottom graph. Although this one is reminiscent of the extreme correlations of plus and minus one, indeed the pattern it is a combination of those two patterns, you get an  $r$  that equals zero.

The perfect linear relationships on the positive side and on the negative side just sort of cancel each other out. You get this relationship when  $Y = |X|$ . It does not mean that  $X$  and  $Y$  are independent, of course, just uncorrelated.

This non-monotonic relationship, first down then up, can happen in nature, it is not just a mathematical artifact.



<sup>4</sup> <http://www.census.gov/compendia/statab/2012/tables/12s1114.pdf>

Here is an example of a U-shaped relationship (except that it is an upside-down U, but that does not affect the argument). The height of the bars represent the percent of all fatal automobile accidents in the US that involve blood level of alcohol of 0.08%, or above for the drivers. The blue bars refer to 1999 and the red bars to ten years later, namely 2009, and each pair of bars refer to a different age group.

In those 10 years we see a small improvement in the less than 16s, even though most of those do not have their driver's license so that is problematic. There has been a decrease in those over 75 and a flat spot in the 65 to 74. But with everybody else between the ages of 16 and 64 there was an increase in the percentage of fatal accidents that involved alcohol. So something is going wrong here.

That is not why I am showing you this. Of course, it is good if I give you the don't drink and drive, message, but the reason I am showing you this is that this looks very much like that U-shaped relationship we just saw, and here too the correlation coefficient is not very high. But the fact that there is a relationship between the age of the drivers and this outcome we are measuring, seems indisputable.

So what we are seeing is an example of the fact that a non-linear relationship is not measured well by the correlation coefficient. There are a number of situations like this where we just have to be very careful. So correlation zero does not imply independence.

2. Short of  $\rho = \pm 1$ , a high correlation does not imply a cause & effect relationship.



The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning.

Stephen Jay Gould  
*The Mismeasure of Man*,  
 W.W.Norton & Co. 1981, p. 242

So moving on to the second misconception, here is what Stephen Jay Gould had to say about that.

2.



All causation as we have defined it is correlation, but the converse is not necessarily true, i.e. where we find correlation we cannot *always*<sup>\*</sup> predict causation. In a mixed African population of [black Africans] and Europeans, the former may be more subject to smallpox, yet it would be useless to assert darkness of skin (and not absence of vaccination) as a cause.

\* [stress added m.p.]

The Grammar of Science  
Karl Pearson  
London, Adam and Charles Black, 1900

Indeed, Pearson recognized this problem back in 1900 shortly after he introduced the correlation coefficient. So this direction has been well established.

2.

Oct 31 12:11 PM Harvard Crimson



### Correlation Still Doesn't Equal Causation in Soda Studies

The report links aspartame to increased risks of leukemia, lymphoma, and non-Hodgkin's lymphoma,.....Boston public high school students has shown that students who identified as heavy soda drinkers were more likely to engage in violent behavior.....

Neither study decisively proves the harmful effects of soda, so until more intensive studies are preformed, **it looks like you're safe to enjoy a glass of your favorite soda** without worrying too much about the possibility of either cancer risks or increased violence.

<http://www.thecrimson.com/article/2012/10/31/soda-studies-harvard/>

5

But just because correlation does not imply causality it does not mean that because there is correlation then there cannot be a causal. This might sound silly, but the tobacco companies

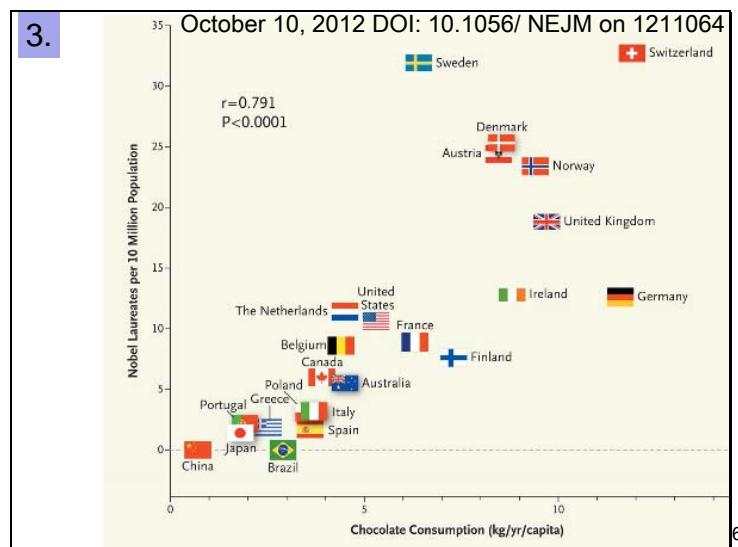
---

<sup>5</sup> <http://www.thecrimson.com/article/2012/10/31/soda-studies-harvard/>

kept singing this tune for some 75 years: you have only shown a correlation and correlation does not imply causality.

Here is an article that appeared in the Harvard Crimson. It was commenting on two studies that had come out of the Harvard School of Public Health dealing with the consumption of soft drinks. One of the studies showed a correlation between aspartame—an artificial sweetener used in the soft drinks—and increased risks of leukemia, lymphoma, and non-Hodgkin's lymphoma. The other study correlated high school students who were identified as heavy soda drinkers were also more likely to engage in violent behavior.

Now this undergraduate reporter had been rightly taught that correlation still does not equal causation, but she incorrectly went one step too far when she said, "It looks like you're safe to enjoy a glass of your favorite soda."



Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

The last misconception is the ecological fallacy. It is a mathematical property of the correlation coefficient that it increases when considering two groups as opposed to the correlation between the individuals in the group.

Case in point, above, when considering the correlation coefficient between countries' chocolate consumption and their Nobel prizes received. At that level of aggregation, the correlation turns

<sup>6</sup> Franz H. Messerli, M.D. Occasional Notes, **Chocolate Consumption, Cognitive Function, and Nobel Laureates** October 10, 2012 DOI: 10.1056/NEJMOn1211064

out to be 0.791. Before you run out and drive up the price of chocolate, this calculation tells us nothing about the correlation coefficient when calculated at the individual level. That is the ecological fallacy; namely believing that it does.

3.



Ecological fallacy: Assuming that correlations measured at an aggregated level imply the same at an individual level.

**Are people who drink “hard” water containing higher levels of calcium and/or magnesium less likely to suffer cardiovascular disease?**

[http://www.who.int/water\\_sanitation\\_health/gdwqr\\_evision/cardiofullreport.pdf](http://www.who.int/water_sanitation_health/gdwqr_evision/cardiofullreport.pdf)

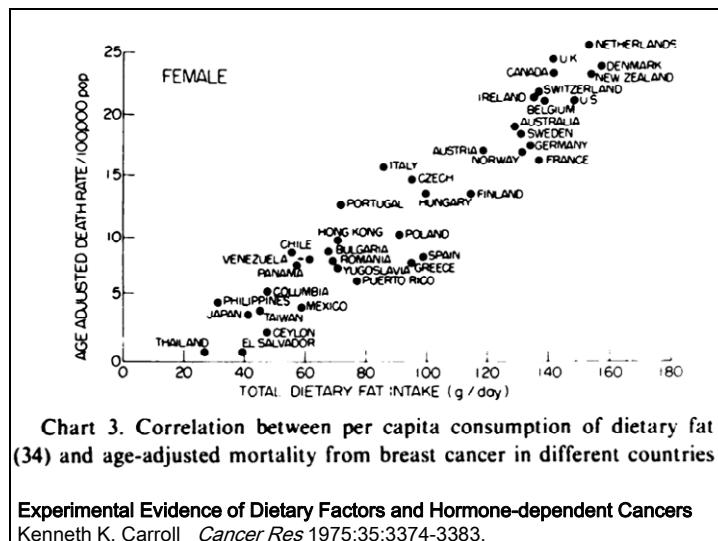
There is an ongoing debate that has lasted a number of years that is trying to establish whether people who drink hard water containing higher levels of calcium and or magnesium are less likely to suffer cardiovascular diseases.

A large number of studies have investigated the potential health effects of drinking-water *hardness*. Most of these have been *ecologic* and have found an inverse relationship between water hardness and cardiovascular mortality. Inherent weaknesses in the ecologic study design limit the conclusions that can be drawn from these studies.



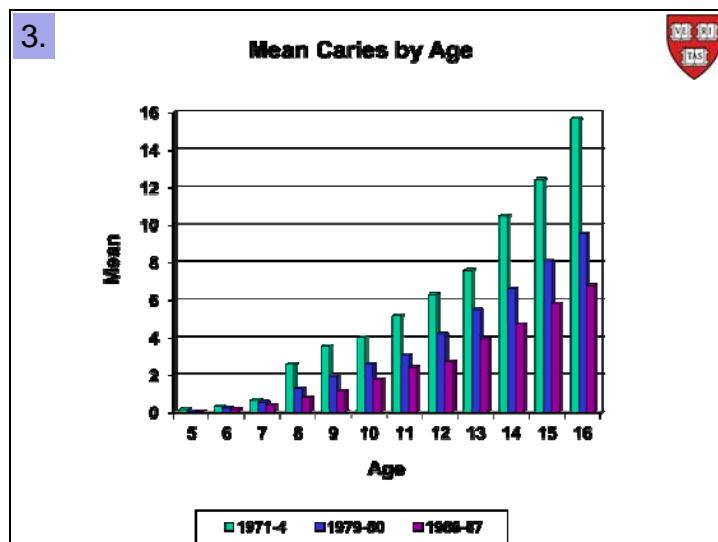
Based on identified *case-control* and *cohort studies*, there is no evidence of an association between water hardness or calcium and acute myocardial infarction or deaths from cardiovascular disease (acute myocardial infarction, stroke and hypertension). There does not appear to be an association between drinking-water magnesium and acute myocardial infarction. However, the studies do show a negative association (i.e. protective effect) between cardiovascular mortality and drinking-water magnesium. Although this association does not necessarily demonstrate causality, it is consistent with the well known effects of magnesium on cardiovascular function.

Above is shown the address of the WHO website, and you may wish to go there to follow the debate. They recognize the shortcomings of an ecological study and what they are looking for is something better, possibly case control studies and maybe cohort studies to settle the debate.



Here is another example of an ecological study that shows the per capita consumption of dietary fat and age adjusted mortality from breast cancer in different countries. We see a wonderful, wonderful lozenge shaped relationship, but it is not telling us anything about the individuals involved. It is telling us something about the aggregate level, and the aggregate correlation is going to be higher than the individual correlation. We need further study to go into this.

But let me repeat, correlation at the aggregate level does not mean that there might not be causality at the individual level too.



Consider this case in point. We looked at how the mean carries varied by age across three surveys; the first in 1971-74, the second about eight years later, and the third about seven years after that. What we saw was that in each age group the number of carries goes down over those three studies. The explanation was that the more and more locales fluoridated their water over

the period of the studies. But this linkage between fluoride and cavities was first suspected because different regions in the country had naturally different fluoride content in the water and that was correlated with stronger teeth. As a result fewer dentists need fill in cavities.

There are any number of relationships, such as asbestos and mesothelioma, that start off as observation of correlations and that subsequently are proven to be causal. I repeat, just because it is correlation does not necessarily mean that it is not causal also. Indeed, it might be the first tip off, as it was with fluoride and asbestos, to lead to the right direction to discover the causal path.

### Spearman's Correlation Coefficient

**Robustness**



Charles E Spearman  
1863—1945



Note also that  $r$  is sensitive to outliers & it measures linear relationship.

Alternative: Spearman's rank correlation – same as Pearson's but replace observations by their ranks.

Pearson's correlation coefficient is not robust and is sensitive to all the observations. (You can wait till next week to understand this more fully, or research the issue for yourself.) Further, this coefficient is quantifying a linear relationship. So what Spearman—a person whose name is related to intelligence testing and all that entails—came up with a very clever idea. He argued that since we may or may not have normality, and its associated linearity, replace the observations with their ranks. Then calculate the correlation coefficient between the ranks. That is what we today call the Spearman correlation coefficient.

Tied Ranks:					
X :	1.7	2.3	2.3	3.4	
Ranks:	1	2	2	4	
		0.5 (2+3)			
	1	2.5	2.5	4	
<hr/>					
X :	1.7	2.3	2.3	2.3	4
	1	3	3	3	4
	$\frac{2+3+4}{3} = 3$				

First, a quick reminder about ranks: if we have these four observations, 1.7, 2.3, 2.3, and 3.4, we might give them ranks 1, 2, 3, 4, except that the two middle ones are equal—we call those tied ranks.

There are a number of ways to handle tied ranks—for example, argue that each of the 2.3 are second smallest so each should get rank 2. The most common compromise is to average out the ranks that would have been given if we had slightly jiggled the data to break the ties, but not enough to reorder them. For example change the 2.3s to 2.31 and 2.32. Then the original 2.3s would get ranks 2 and 3. Average those out and associate the rank of 2.5 with each of the original 2.3. This is what sometimes happens in sporting events when the prize money is averaged out between contestants who tie.

It does lead to fractional ranks, although the worst it can get is, as here, introduction of a 0.5 in the ranks, and that happens if even numbers of people reach a tie. If odd numbers of people reach a tie, then the ranks remain whole. I leave that for you to prove for yourself.

e.g. fake numbers:



Raw Data			Ranks	
i	x	y	x <sub>r</sub>	y <sub>r</sub>
1	1.3	14.3	2	2
2	1.7	14.7	4	3
3	0.8	18.0	1	4
4	1.4	12.1	3	1

$$r_s = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_{ri} - \bar{x}_r}{s_{x_r}} \right) \left( \frac{y_{ri} - \bar{y}_r}{s_{y_r}} \right)$$

So returning to Spearman, operationally this is what you do: first rank the x amongst themselves, then rank the y amongst themselves, all the while retaining the order of the data to maintain the proper linkage. Then ignore the original data and act as if the ranks are the original data. Now calculate the Pearson correlation coefficient with these ranks. That is the Spearman correlation coefficient, r<sub>s</sub>.

e.g. fake numbers:

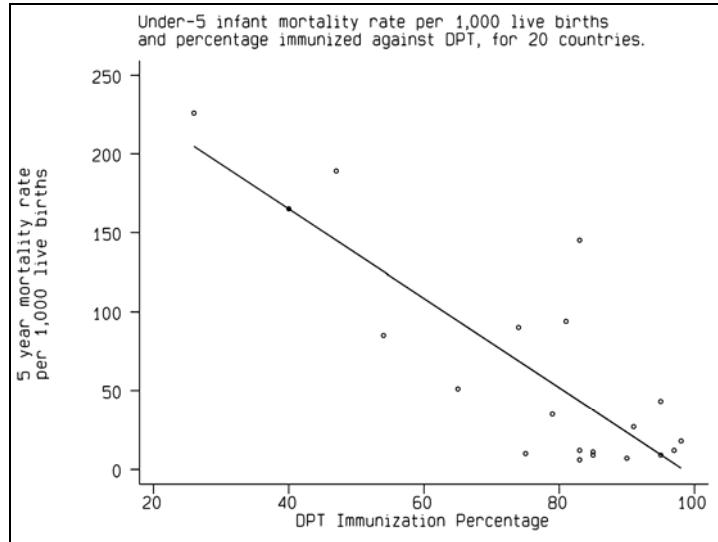


Raw Data			Ranks		d
i	x	y	x <sub>r</sub>	y <sub>r</sub>	d
1	1.3	14.3	2	2	0
2	1.7	14.7	4	3	1
3	0.8	18.0	1	4	-3
4	1.4	12.1	3	1	2

$$r_s = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_{ri} - \bar{x}_r}{s_{x_r}} \right) \left( \frac{y_{ri} - \bar{y}_r}{s_{y_r}} \right)$$

$$= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

One nice mathematical observation is this formula; useful if you do not have a computer to do your calculations for you!



Returning to our DPT example, chances are that neither the immunization percentages nor the mortality rates are normally distributed, so we might be tempted to calculate the Spearman's correlation coefficient in this case.

Nation	%Immun.	Rank	Death /1000	Rank	d	d <sup>2</sup>
Ethiopia	26	1	226	20	-19	361
Bolivia	40	2	165	18	-16	256
Senegal	47	3	189	19	-16	256
Brazil	54	4	85	14	-10	100
Mexico	65	5	51	13	-8	64
Turkey	74	6	90	15	-9	81
U.K.	75	7	10	5	2	4
USSR	79	8	35	11	-3	9
Egypt	81	9	94	16	-7	49
Japan	83	10	6	1	9	100
Greece	83	11	12	7.5	3.5	12
India	83	12	145	17	-6	36
Italy	85	13	11	6	7	56
Canada	85	14	9	3.5	10	100
Finland	90	15	7	2	13	169
Yugoslavia	91	16	27	10	6	36
France	95	17	9	3.5	14	196
China	95	18	43	12	5.5	30
USA	97	19	12	7.5	11.5	132
Poland	98	20	18	9	11	121
Total						2169

Here is the tabular data and all the steps necessary to calculate Spearman's correlation coefficient.



In the DPT example:

$$\begin{aligned} r_s &= 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \\ &= 1 - \frac{6(2,169)}{20(399)} \\ &= -0.631 \end{aligned}$$

Versus Pearson's:

$$r = -0.829$$

And the Spearman's correlation coefficient is -0.631. Remember that how we rank, from smallest to largest, or vice versa, is arbitrary, so we need to look how we ranked here in order to understand what the coefficient is saying. In this case we ranked the immunization with the lowest coverage getting a rank of 1, and then going up, and the lowest mortality getting a rank of 1, on then up. So a negative correlation means that as the coverage goes up the mortality goes down, as we might expect.

It turns out that with these data, the Pearson's was not that far different from the Spearman's, but it does buy us a little safety to know both, anyway.



To test **correlation** of two characteristics  
(only has power against  $\rho \neq 0$ )

$$\begin{aligned} t_s &= r_s \sqrt{\frac{n-2}{1-r_s^2}} \\ &= -0.631 \sqrt{\frac{18}{1-(0.631)^2}} \\ &= -3.45 \end{aligned}$$

versus t with 18 degrees of freedom,  
so  $0.001 < p < 0.01$

7

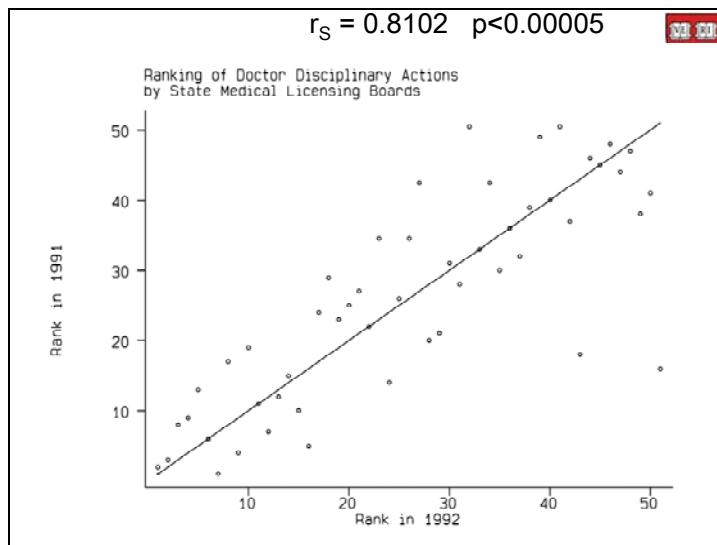
---

<sup>7</sup> I have edited this slide. The word correlation appears where the word independent used to be. It was a mistake.

Note that in the Spearman's correlation coefficient the word linear does not appear. That is because it is not necessarily measuring just a linear relationship between the two variables. Mathematically, we say that it is measuring the strength of a monotonic relationship. (Monotonic means that the variables travel together. Either they both go up (down) together, or one goes up as the other goes down. A special case of this is a straight line. With a positive slope they both go up (down) together, and with a negative slope one goes up as the other goes down. But the straight line is not the only relationship that is monotonic. Think of non-straight line (or non-linear) relationships such as weight gain with age; height with age; etcetera.)

So Spearman is a generalization of Pearson, but it does not solve the U-shaped relationship problem. (Test it for yourself with the same example we used above for Pearson, namely  $Y=|X|$ . That relationship yields both a Pearson and a Spearman of zero.)

Testing hypotheses with Spearman's correlation is almost identical to the Pearson case, even sharing the shortcoming that it only has power against the null hypothesis that rho is zero.



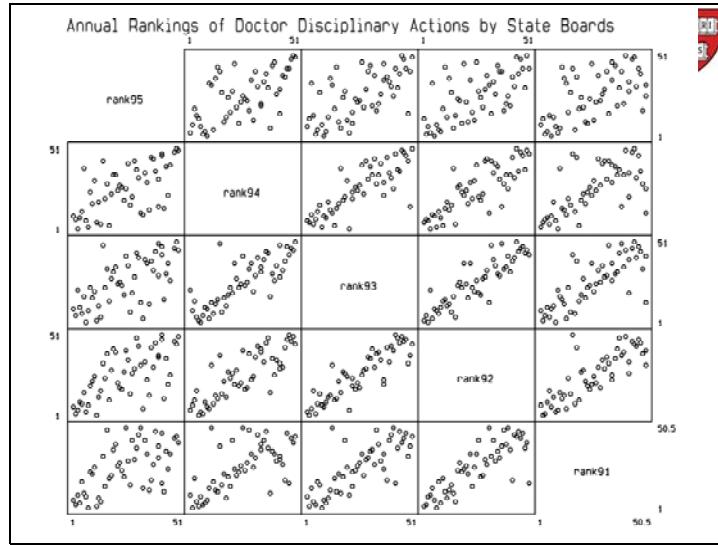
Here is another situation where you would certainly want to calculate the Spearman correlation coefficient. This shows the rankings of States (plus the District of Columbia) by a watchdog group of the medical profession in the US<sup>8</sup>. They keep an eye on the disciplinary action taken by the state medical boards.

This is a plot of the ranks in 1991 against the ranks in 1992. We see that apart from a few outliers, there seems to be a very strong correlation between these two years. In support of this observation we have that Spearman's rank correlation is 0.8102. So, here, we do not have to

<sup>8</sup> <http://www.citizen.org/Page.aspx?pid=183>

assume anything about the distribution of what it is that these people are measuring. We are just looking at their ranks.

I found it rather surprising, when I first saw this, that from one year to the next the medical boards' actions are so highly correlated. Why this should be I do not know. Do all bad doctors go to the same State?



Here is a matrix plot to capture this behavior over the five years, 1991 to 1995. To read a matrix plot we look at each diagonal entry. That variable defines all the horizontal axes for the plots in that column, and all the vertical axes for all the plots in that row. Looking at the plot as a whole, we see that the set of the lower triangle of plots is almost the same as the upper triangle set, so I could have suppressed one set of them—Stata gives me that option—but the role of column and row axes are interchanged in these two sets, and I wanted to retain that.

When looking at this particular set of variables it is interesting to look along diagonal rows. We have spoken of the main diagonal, it holds the names of the variables. If we move one up, so to speak, we get the plots when the two years (vertical and horizontal axes) are one-year apart. We get the same comparison by going down one diagonal from the main diagonal. If we go up (down) two diagonals, then we get the comparisons of years that are two-years apart. And so on.

The pattern that we are seeing is that they seem pretty tight when they are one-year apart, but that that tightness decreases as we go further away from the main diagonal; i.e. when the comparison is being made of two years further apart in time. (Remember the correlation game you played!)

```

. spearman rank95 rank94
Number of obs =      51
Spearman's rho =     0.6035
Test of Ho: rank95 and rank94 independent
Pr > |t| =     0.0000

. pwcorr rank95-rank91
| rank95  rank94  rank93  rank92  rank91
-----+
rank95 |  1.0000
rank94 |  0.6057  1.0000
rank93 |  0.6321  0.8071  1.0000
rank92 |  0.6441  0.8168  0.8808  1.0000
rank91 |  0.5833  0.6292  0.7643  0.8102  1.0000

```

Here are the Spearman coefficients to accompany these plots. We see the same striation pattern as we see in the plots.

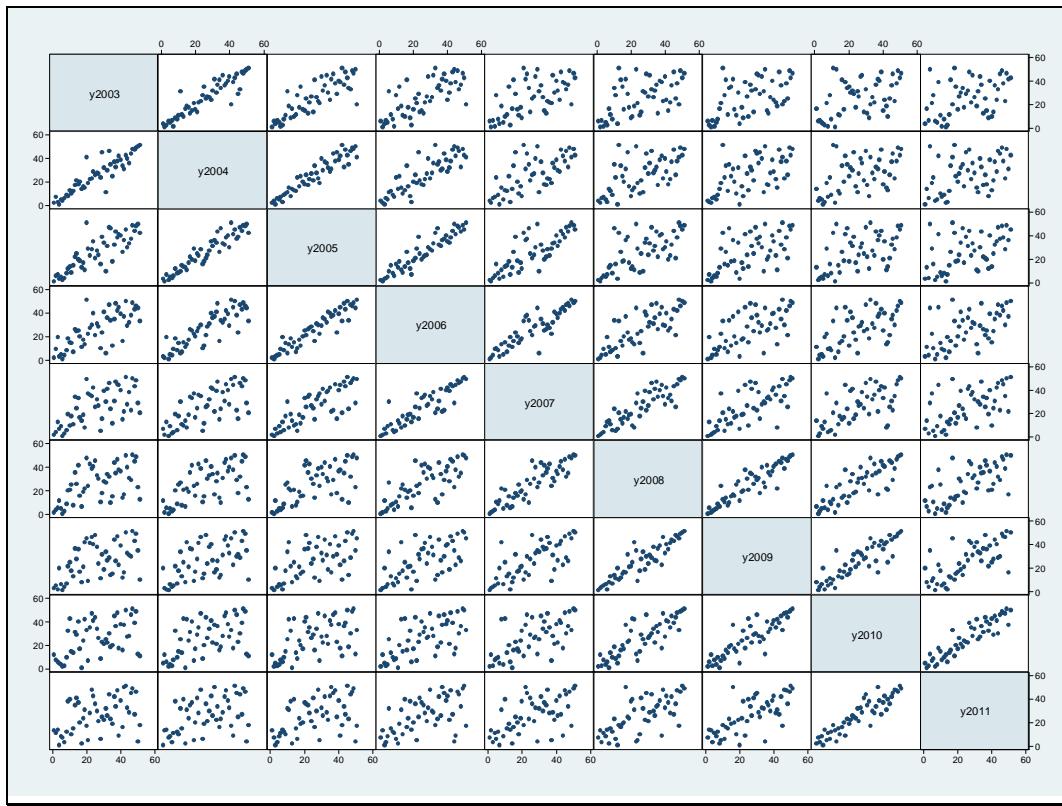
```

. pwcorr rank95-rank91 , bonf sig
| rank95  rank94  rank93  rank92  rank91
-----+
rank95 |  1.0000
|
rank94 |  0.6057  1.0000
|  0.0000
rank93 |  0.6321  0.8071  1.0000
|  0.0000  0.0000
rank92 |  0.6441  0.8168  0.8808  1.0000
|  0.0000  0.0000  0.0000
rank91 |  0.5833  0.6292  0.7643  0.8102  1.0000
|  0.0001  0.0000  0.0000  0.0000

```

You can also ask for the significance value with the command *pwcorr*. Here it is for the Bonferroni, and we see that they are all significant at the 5% level.

I was confused about why this correlation existed, but it was 15 years, or so ago. So recently I went searching for more current data and here it is for the years 2003 to 2011.



And here is that same matrix. This time the years range from 2003 to 2011. If anything, the striation pattern looks even more pronounced now than it did fifteen years ago.

	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010
y2003	<b>1.0000</b>							
y2004	0.9142	<b>1.0000</b>						
y2005	0.8287	0.9472	<b>1.0000</b>					
y2006	0.7353	0.8584	0.9395	<b>1.0000</b>				
y2007	0.6213	0.6976	0.8093	0.8927	<b>1.0000</b>			
y2008	0.5459	0.6207	0.6580	0.7658	0.8824	<b>1.0000</b>		
y2009	0.4996	0.5721	0.5879	0.6494	0.7319	0.9138	<b>1.0000</b>	
y2010	0.4703	0.5075	0.5360	0.6319	0.6437	0.8260	0.8757	<b>1.0000</b>
y2011	0.4280	0.4401	0.4771	0.5670	0.5733	0.6791	0.7206	0.9182

Public Citizen's Health Research Group Ranking of the Rate of State Medical Boards' Serious Disciplinary Actions, 2009-2011  
SM Wolfe, C Williams, and A Zaslow , May 17, 2012  
<http://www.citizen.org/documents/2034.pdf>

Here is the matrix of the Spearman correlation coefficient. The striation pattern is quite evident.

	y2003	y2004	y2005	y2006	y2007	y2008	y2009	y2010
y2003	<b>1.0000</b>							
y2004	0.9142	<b>1.0000</b>						
y2005	0.8287	0.9472	<b>1.0000</b>					
y2006	0.7353	0.8584	0.9395	<b>1.0000</b>				
y2007	0.6213	0.6976	0.8093	0.8927	<b>1.0000</b>			
y2008	0.5459	0.6207	0.6580	0.7658	0.8824	<b>1.0000</b>		
y2009	0.4996	0.5721	0.5879	0.6494	0.7319	0.9138	<b>1.0000</b>	
y2010	0.4703	0.5075	0.5360	0.6319	0.6437	0.8260	0.8757	<b>1.0000</b>
y2011	0.4280	0.4401	0.4771	0.5670	0.5733	0.6791	0.7206	0.9182

And the Bonferroni results are here, and they are all significant at the 5% level, except for the bottom left corner. (Stata has changed how it reports these results from 15 years ago when it showed the p-value. Now it shows which coefficients are significant, at the chosen level of significance.)

What does this show? Is this just an example of something we know very well, which is that these medical boards tend to have a number of physicians sitting on them, so this is just a

manifestation of the observation that groups of people cannot police themselves? I leave to you to interpret these graphs.

## Non-parametrics

[http://en.wikipedia.org/wiki/Sex\\_ratio](http://en.wikipedia.org/wiki/Sex_ratio) 11/10/2012



The CIA estimates that the current world wide sex ratio *at birth* is 107 boys to 100 girls.<sup>[3]</sup> In 2010, the global sex ratio was 986 females per 1,000 males and trended to reduce to 984 in 2011.<sup>[4]</sup>

$$107 \text{ m : } 100 \text{ f} \quad \Pr(\text{male}) = \frac{107}{207} = 0.5169$$

$$(935 \text{ f : } 1000 \text{ m})$$

$$984 \text{ f : } 1000 \text{ m} \Rightarrow \Pr(\text{male}) = \frac{1000}{1984} = 0.5040$$

$$986 \text{ f : } 1000 \text{ m} \quad \Pr(\text{male}) = \frac{1000}{1986} = 0.5035$$

Spearman's rank correlation coefficient is your first example of something that is called a non-parametric statistic because it does not make any assumptions, which can be characterized up to some parameters, about the distribution of the population.

The first example I want to show you has to do with the sex ratio. So I went to Wikipedia and this is what I found: the CIA estimates that the current worldwide sex ratio at birth is 107 boys to 100 girls<sup>9</sup>. Also, The Times of India gave the ratio for 2010 and for 2011, and there they were<sup>10</sup>.

I think there is something wrong with these numbers, because if you put them all on the same scale—1,000 males, say—then I do not think that in three years one sees this much disparity.

But accepting that, there are two points I want to make with this. The first is that we are still talking about the sex ratio at birth. We seem to be stuck on this theme for many years now. It has been used as an indicator for any number of things. Currently there is concern about the worrisome practice of using technology to influence this ratio by producing more males.

<sup>9</sup> [https://www.cia.gov/library/publications/the-world-factbook/fields/print\\_2018.html](https://www.cia.gov/library/publications/the-world-factbook/fields/print_2018.html)

<sup>10</sup> [http://articles.timesofindia.indiatimes.com/2011-08-17/india/29895810\\_1\\_ratio-abortion-and-craze-craze-for-male-child](http://articles.timesofindia.indiatimes.com/2011-08-17/india/29895810_1_ratio-abortion-and-craze-craze-for-male-child)

The other point struck me when reading the Times of India. After listening to the complaints of some students in the past who claim not to be able to understand odds, I see before me a newspaper that presumably appeals to a large, general readership talking about odds! Surely they want their readership to understand what they are writing, and yet they feel free to use odds.

For those of you who do not appreciate odds, we can calculate the probabilities, and there they are.



**John Arbuthnot**  
1667-1735



**Claim:**  
Divine providence, not chance,  
governs the sex ratio at birth.

John Arbuthnot (1710) [``An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes."](#)

*Philosophical Transactions of the Royal Society*

Why I brought up the sex ratio at birth is John Arbuthnot. He not only was a physician, to the queen yet, the inventor of John Bull, but he also was a man way ahead of his time both in measuring characteristics of people, and in the formulation of a statistical hypothesis test. This he did in the context of the sex ratio, which he used in order to utilize statistics to prove there is a god. This is one of the first times hypothesis testing was used in the literature.

This claim he made in his article, “An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes,” in the *Philosophical Transactions of the Royal Society*, in 1710.



Data:

Christenings, London 1629--1710 where for 82 years he observed more boys christened than girls.

If  $\text{Pr}(\text{boy}) = 0.5$ ,

$$\text{Prob}(82 \text{ years}, \# \text{ boys} > \# \text{ girls}) = (0.5)^{82}$$

$$= \frac{1}{4\ 8360\ 0000\ 0000\ 0000\ 0000\ 0000}$$

``From whence it follows, that it is Art, not Chance, that governs."

The way he went about his study was to look at the christening records in London for the period 1629 to 1710. For those 82 years he counted that there were more boys than girls christened. Chance to him meant that the sex ratio had to be 0.5. So to see 82 years, out of 82 years observed, where there were more boys than girls had that minute p-value above. Getting this p-value was much too small to believe Chance was governing this, so Art must have been at work!

Now, why chance had to be exactly 0.5 and not any other number he did not defend. This was one of the first examples, the other being Neumann and the data from Breslau used to disprove astrology, of the use of hypothesis testing.



Resting energy expenditure (kcal/day): cystic fibrosis & healthy, matched on age, sex, height and weight.

Pair	CF	Healthy	Diff	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1634	1162	472	+
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+

What Arbuthnot did is what we now call the Sign Test. For each year he counted the number of boys christened and subtracted the number of girls born, and he came up with 82 plus signs. If  $p$ , the probability a boy is christened is 0.5, then he would have expected about 41 years where there would be more boys and 41 years where more girls were christened.

Here is a more modern example. We have 13 pairs of individuals, 13 with cystic fibrosis and another 13 controls without, who were matched with the cases according to age, sex, height, and weight. The outcome measured in each patients was the resting energy expenditure.

Taking the differences in the outcomes in each pair we see 11 positive and 2 negative differences.



### The Sign Test

Let D be the number of positive differences:

$$Z_+ = \frac{D - (n/2)}{\sqrt{n/4}}$$

$$= \frac{11 - 6.5}{\sqrt{13/4}} = 2.50$$

$p=0.006 + 0.006 = 0.012 < 0.05$

For small n use the binomial.

We can use a normal approximation, as above, or call on Stata.



. signtest CF = Healthy		
Sign test		
sign	observed	expected
positive	11	6.5
negative	2	6.5
zero	0	0
all	13	13

One-sided tests:

Ho: median of CF - Healthy = 0 vs.  
Ha: median of CF - Healthy > 0  
Pr(#positive >= 11) =  
Binomial(n = 13, x >= 11, p = 0.5) = 0.0112

Ho: median of CF - Healthy = 0 vs.  
Ha: median of CF - Healthy < 0  
Pr(#negative >= 2) =  
Binomial(n = 13, x >= 2, p = 0.5) = 0.9983

Two-sided test:

Ho: median of CF - Healthy = 0 vs.  
Ha: median of CF - Healthy != 0  
Pr(#positive >= 11 or #negative >= 2) =  
min(1, 2\*Binomial(n = 13, x >= 11, p = 0.5)) = 0.0225

We see from the Stata output that the null hypothesis that the sign is just as likely to be positive as negative has a p-value of 0.0225, and thus we reject that hypothesis at the 5% level.

The meaning of a probability of 0.5 of a positive or negative sign of the difference is that the median of the distribution of that difference is zero. In other words, that the distributions of the controls and the cases both have the same median.

## Wilcoxon Signed Rank Test

Resting energy expenditure (kcal/day): cystic fibrosis & healthy, matched on age, sex, height and weight.				
Pair	CF	Healthy	Diff	Sign
1	1153	996	157	+
2	1132	1080	52	+
3	1165	1182	-17	-
4	1460	1452	8	+
5	1634	1162	472	+
6	1493	1619	-126	-
7	1358	1140	218	+
8	1453	1123	330	+
9	1185	1113	72	+
10	1824	1463	361	+
11	1793	1632	161	+
12	1930	1614	316	+
13	2075	1836	239	+

The Sign test only looks at the signs of the differences and not much else. So small differences and large differences are all counted the same. Frank Wilcoxon argued that there is some information to be gained by looking at the magnitude of these differences.

Ranks usually 1 top 2 next .....				
2010 U.S. Suicides, by methods & sex				
Method	Male	Rank	Female	Rank
Firearms	16,962	1	2,430	2
Poisoning	3,573	3	3,026	1
Suffocation	7,592	2	1,901	3
Other	2,150	4	730	4
Total	30,277		8,087	

Consider this report of the methods of suicide in the US in 2010<sup>11</sup>. If we rank the data for males and females we find that for males the most popular method involves the use of firearms. The next most popular is what they used to call strangulation and suffocation, presumably one hangs oneself. For females, poison is the most popular way of committing suicide. Firearms is the second most popular and suffocation is third.

From looking at the rankings of the methods used, we can see that there is a qualitative difference between the sexes. If we want to act as if these were samples and we want to do some sort of t-test, then we would have to claim some normality, but if these numbers are not normally distributed, can we do something with the ranks?



**One sample, or paired, Wilcoxon**  
 e.g. Reduction in forced vital capacity (FVC) for a sample of patients with cystic fibrosis:  
**Frank Wilcoxon**  
 1892 - 1965

Wilcoxon came up with tests analogous to the t-tests but instead of using the raw data, he first ranks the data and then uses the ranks. Indeed, Wilcoxon is to Student as Spearman is to Pearson. First we look at the one-sample version of the Wilcoxon by applying it to a study of a drug to ameliorate the effects of cystic fibrosis. Forced vital capacity (FVC) is the volume of air that one can expel from ones lungs in 6 seconds. In this study they compared the reduction in FVC for patients with cystic fibrosis over two similar periods of time, once when taking the drug and once when on placebo.

---

<sup>11</sup> [http://webappa.cdc.gov/sasweb/ncipc/mortrate10\\_us.html](http://webappa.cdc.gov/sasweb/ncipc/mortrate10_us.html)

Pat	Placebo	Drug	Diff	Rank	"Signed Rank"	
1	224	213	11	1	1	
2	80	95	-15	2	2	
3	75	33	42	3	3	
4	541	440	101	4	4	
5	74	-32	106	5	5	
6	85	-28	113	6	6	
7	293	445	-152	7	7	
8	-23	-178	155	8	8	
9	525	367	158	9	9	
10	-38	140	-178	10	10	
11	508	323	185	11	11	
12	255	10	245	12	12	
13	525	65	460	13	13	
14	1023	343	680	14	14	
Totals (T)				86	19	

Here are fourteen pairs of readings on the fourteen patients. We obtain the differences, as in the Sign test (or the dependent t-test), and then rank the absolute values of the differences.

We then look at the total ranks of those associated with positive differences and the total ranks of those with negative differences. If the distribution of the differences is symmetric around zero, then one would intuitively expect that these two sums should be approximately the same. This is the Wilcoxon signed rank test.

$H_0: \text{Median}_X = 0$
Under $H_0$
$\text{mean}_T = \mu_T = \frac{1}{4}n(n+1)$
$\sigma_T = \sqrt{n(n+1)(2n+1)/24}$
For large n:
$Z = \frac{T - \mu_T}{\sigma_T}$ approx. stand. normal

In contrast to the Sign test, we are not only taking the sign of the difference into account, but also the size of these differences by looking at the ranked differences. The Z-statistic to use, which is displayed above, has an approximately normal sampling distribution.



For the above example,  $T = 19$   
and  $n = 14$

$$\begin{aligned} Z &= \frac{19 - 14(15) / 4}{\sqrt{14 \times 15 \times 29 / 24}} \\ &= -2.10 \end{aligned}$$

$$p = 0.036$$

In this instance, p value is 0.036, and so we would reject the null hypothesis of equality. If we look at the data we see that the sum of the positive ranks is too high. That means that the placebo seems to cause a bigger loss than the drug, so the drug looks effective, on the basis of these data.



```
. signrank placebo = drug
Wilcoxon signed-rank test

sign |   obs  sum ranks  expected
-----+
positive |    11      86    52.5
negative |     3      19    52.5
zero |     0       0     0
-----+
all |    14     105    105

unadjusted variance   253.75
adjustment for ties   0.00
adjustment for zeros   0.00
-----
adjusted variance     253.75

Ho: placebo = drug
z =  2.103
Prob > |z| =  0.0355
```

## Wicoxon Rank Sum Test<sup>12</sup>



Low Exposure				High Exposure			
nMA	Rank	nMA	Rank	nMA	Rank	nMA	Rank
34.5	2	51.0	23	28.0	1	51.0	23
37.5	6	52.0	25.5	35.0	3	52.0	25.5
39.5	7	53.0	28	37.0	4.5	53.0	28
40.0	8	54.0	31.5	37.0	4.5	53.0	28
45.5	11.5	54.0	31.5	43.5	9	54.0	31.5
47.0	14.5	55.0	34.5	44.0	10	54.0	31.5
47.0	14.5	56.5	36	45.5	11.5	55.0	34.5
47.5	16	57.0	37	46.0	13		
48.7	19.5	58.5	38.5	48.0	17		
49.0	21	58.5	38.5	48.3	18		
51.0	23			48.7	19.5		
Total ranks		467		Total ranks		313	

The Wilcoxon two sample test, called the Wilcoxon Rank Sum Test, is what we would use when the two samples are independent. Here are data to study children suffering from phenylketonuria (PKU), a disorder associated with the inability to metabolize the protein phenylalanine. The children have been divided into two groups, the first with average serum levels of phenylalanine less than 10.0 mg/dl and the second group with average phenylalanine levels above 10.0 mg.dl. The study wishes to compare the normalized mental scores for these two groups, without assuming normality of the distributions of these scores.

The test first pools the two samples in order to rank all of them together. Then take the sum of the ranks in each of the groups. Intuitively if the two population distributions are equal to each other and the two sample sizes are equal, then these two sums of ranks should be approximately equal. If the two sample sizes are different we can average things out to make the two comparable.

---

<sup>12</sup> Sometimes called the Mann-Whitney, or Wilcoxon-Mann-Whitney, or any permutation of these names.

Wilcoxon-Mann-Whitney Test  
for two independent samples.



I	Ranks	II	Ranks
$x_1$	$R_1$	$y_1$	$R_{n_1+1}$
$x_2$	$R_2$	$y_2$	$R_{n_2+2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x_{n_1}$	$R_{n_1}$	$y_{n_2}$	$R_{n_1+n_2}$

$$W = \sum_{j=1}^{n_1} R_j$$

$H_0$  : Sample I and II same population

Under  $H_0$



$$\text{Average}(W) = \mu_W = n_s(n_s + n_L + 1)/2$$

$$\sigma_W = \sqrt{\frac{n_s n_L (n_s + n_L + 1)}{12}}$$

For large  $n_s, n_L$  ( $>10$ )

$$Z = \frac{W - \mu_W}{\sigma_W}$$

is approximately standard normal

So you'd expect the two ranks-- the two sums of the ranks be the same, if the two sample sizes are the same. If not, you'll do the appropriate weighting.



Example:  $W = 313$

$$\begin{aligned}\mu_w &= n_s(n_s + n_L + 1) / 2 \\ &= 18(18 + 21 + 1) / 2 = 360 \\ \sigma_w &= \sqrt{n_s n_L (n_s + n_L + 1) / 12} \\ &= \sqrt{18(21)(18 + 21 + 1) / 12} = 35.5 \\ z &= \frac{W - \mu_w}{\sigma_w} \\ &= \frac{313 - 360}{35.5} = -1.32\end{aligned}$$

$p = 0.186$

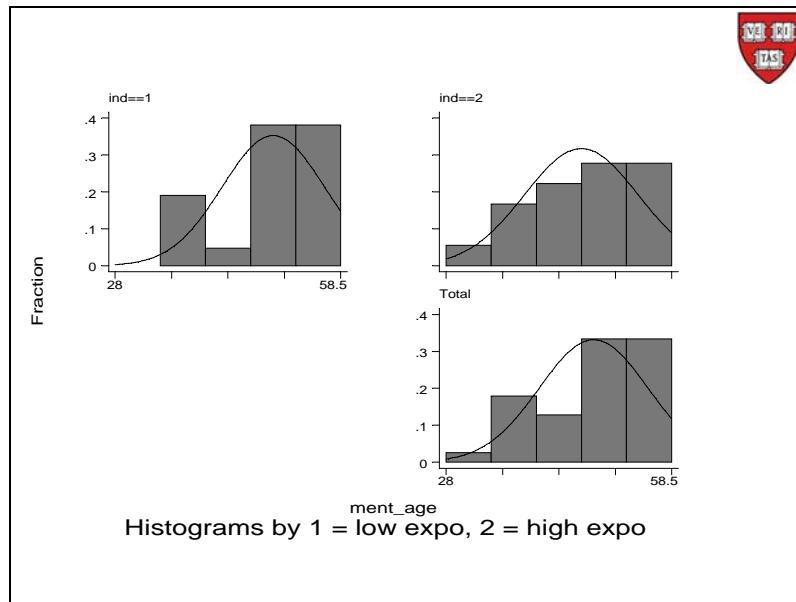
In this example, the standard Z is -1.32, so the p-value is 0.186, and we do not reject the null hypothesis.



```
. tab ind, summ( ment_age )

1 = low    |
expo, 2 = |      Summary of ment_age
high expo |      Mean        Std. Dev.   Freq.
-----+-----+
1 | 49.366667  6.9038636   21
2 | 46.277778  7.6722997   18
-----+-----+
Total | 47.941026  7.3384968   39
```

Here are the summary statistics for the raw data before we did the ranking. Pre-today, or pre you learning Wilcoxon, you might have done a t-test. These two means look approximately the same.



If we want to use the t-test we could look at the histograms. Unfortunately, none of the three (the two individual groups and the combined groups) histograms look particularly normal. So maybe we shouldn't be doing the t-test, but it does not matter, we are halfway there so let us do look at both tests.

```
. ranksum ment_age , by(ind)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
      ind |   obs   rank sum  expected
-----+
      1 |    21     467     420
      2 |    18     313     360
-----+
combined |    39     780     780
unadjusted variance   1260.00
adjustment for ties   -3.19
-----
adjusted variance   1256.81
Ho: ment_age(ind==1) = ment_age(ind==2)
z =  1.326
Prob > |z| =  0.1849
```

And now here's the rank-sum. If we did the Wilcoxon rank sum test we see that the p value is 0.1849,

ttest ment_age , by(ind) unequal						
Two-sample t test with unequal variances						
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
1	21	49.36667	1.506547	6.903864	46.22407	52.50927
2	18	46.27778	1.808378	7.6723	42.46243	50.09312
combined	39	47.94103	1.1751	7.338497	45.56216	50.31989
diff		3.088889	2.353702		-1.691287	7.869065
Satterthwaite's degrees of freedom: 34.6139						
Ho: mean(1) - mean(2) = diff = 0						
Ha: diff < 0		Ha: diff ~ 0		Ha: diff > 0		
t = 1.3124		t = 1.3124		t = 1.3124		
P < t = 0.9010		P >  t  = 0.1980		P > t = 0.0990		

And if we do the t-test, the p value is 0.198. So in this set of data where the normality assumption is not too strictly adhered to, the t-test and the Wilcoxon rank sum test provide very similar results. It seems like the t-test is a little more robust than expected.

Once you have your data on the computer and have access to a decent set of computer routines and there are three different ways to do something, then explore. You will learn something about your data whether you get the same answer three times, or when you get different answers that will require some explaining.

Wilcoxon versus Student's t :

**Advantage**

We do not need to assume that the population is Normally distributed for Wilcoxon to be applicable – robust.

**Disadvantage**

When in fact the parent population is Normal, the Wilcoxon is less efficient (approximately 95% efficient).

In summary, when should we use the Wilcoxon, and when should we use Student's t? The advantage of the Wilcoxon is that it does not require us to assume anything about the parent or population distribution of the variable in question—in particular we have no need to assume normality, whereas that is an assumption we need to make for Student's t. The Wilcoxon is more robust in that sense, and also in that it deals with ranks that are more robust to outliers.

But you are not going to get something for nothing, so how do we pay for this? Well, here's the disadvantage. If in fact you were justified in making your normality assumption—so if in fact it would be legitimate for you to use the t, how much do you lose by using the Wilcoxon? And the answer is not that much. When in fact you have normal data, the Wilcoxon is about 95% efficient, versus Student's t. So if you have 20 observations in your sample, the Wilcoxon is about as efficient as if you had 19 observations and use the t appropriately. I like the Wilcoxon, and as I said, it's not that expensive. And it's a good back up to see how different, if at all, it is to using the t.

Marcello Pagano

# [JOTTER 10 LINEAR REGRESSION]

Simple linear regression, least squares, indicator variables, multiple regression, subset regression

Review:

Straight line:  $y=a+bx$

Today we start on our study of regression; in particular, *linear regression*. It is very close to what we have just done with correlation. It is a continuation of our study of how variables vary together. Now we want to allow more structure and not stop at just two variables.

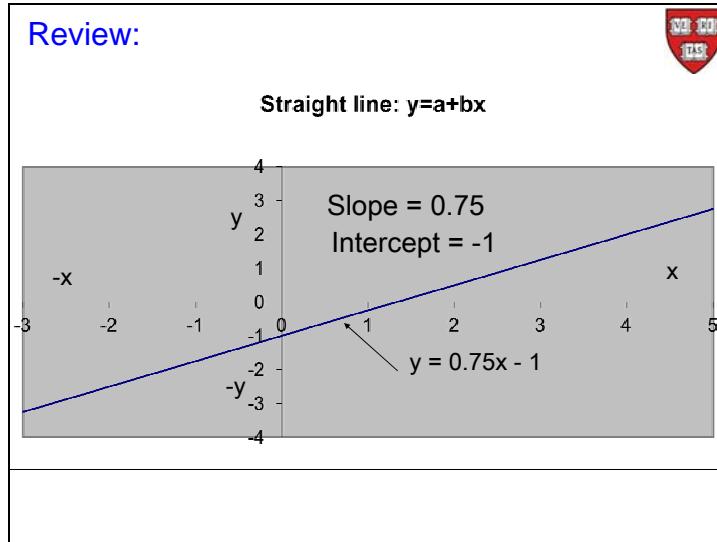
First a very quick review of your high school mathematics, and when you looked at a straight line. Here is a straight line drawn on a Cartesian coordinate system together with its definition. A straight line describes the relationship between two variables,  $x$  and  $y$ . So given an  $x$  I can find a single  $y$ , and vice versa, given a  $y$  I can find an  $x$ . I have a little trouble with this generalization when we have a flat or a vertical line. More about those two extreme cases later, but they are the only two that give us problems.

The reason we are looking at this is that up to now we have been estimating a single population parameter, like the mean, the standard deviation, or the correlation coefficient. Now we are going to define a line in the population and try and estimate it. This is not as complex as it sounds because the line is defined by two parameters,  $a$  and  $b$ —maybe I should have used the Greek letters  $\alpha$  and  $\beta$  to be consistent with our previous work, but I did not want to be fancy. So now instead of estimating a single population parameter, we are being asked to estimate two of them.

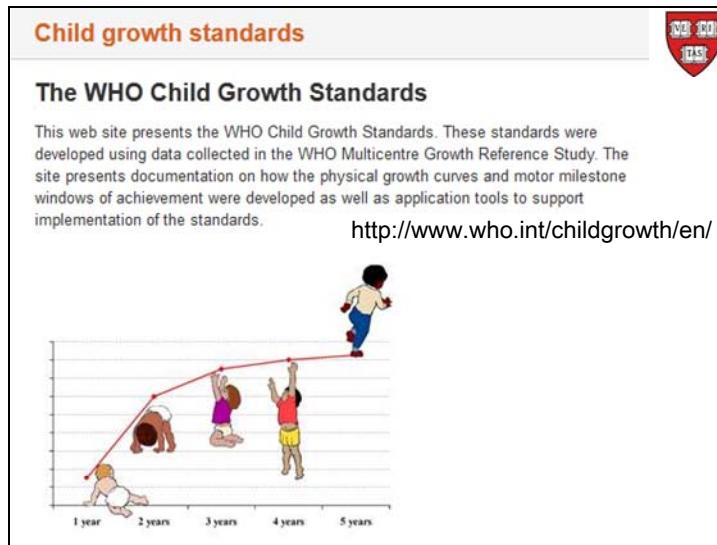
The parameter  $a$  is the intercept. So, for example, if you put  $x$  equals 0 into the equation of the line, you get that  $y$  is equal to  $a$ . So  $y=a$  is where the line crosses the  $y$ -axis.

The parameter  $b$  is the slope of the line: If we start at a point on the line, and we increase  $x$  by one,  $y$  gets increased by the amount  $b$ .

So now I hope all these memories of straight lines have come back to you.

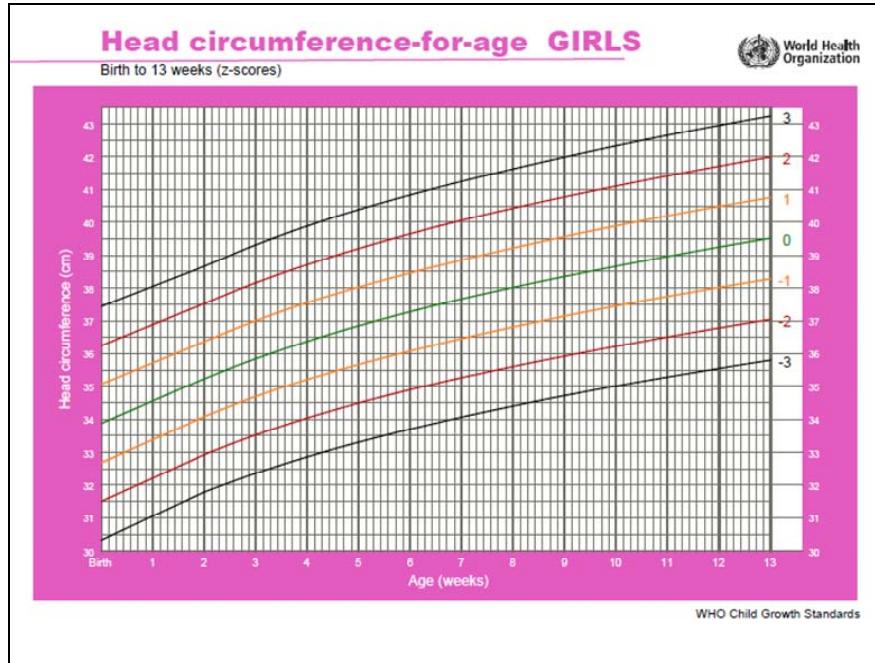


Let us look at a particular straight line, above, that has slope and intercept as stated.



The WHO is much concerned about making statements about how we should be growing. Here is a typical picture showing a monotonic relationship between age on the horizontal axis and some generic measure of growth that should be progressing as we get older. We could be talking of weight, head circumference, or any number of properties associated with growth.

These graphs typically show you how things should progress, and since we do not all progress at the same rate, the graphs also show bounds within which "normal" growth is somehow defined. Of course, falling outside these bounds is sometimes used as a warning that something could be amiss.



Here is an example. This is a chart from the WHO website<sup>1</sup> that shows how head circumference increases with age for girls.

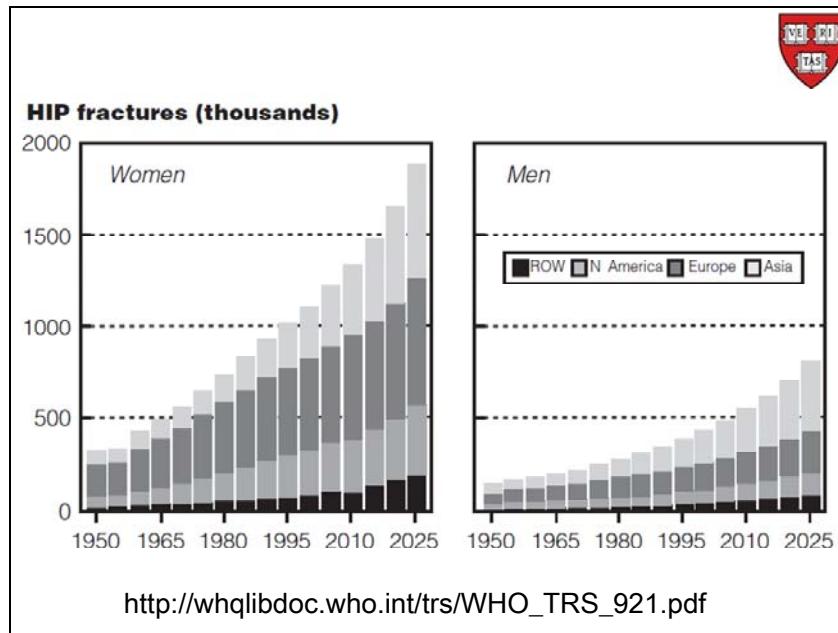
The green line in the middle is the mean. To explain this line: for every fixed age there is a population of girls that age. If we measured the head circumferences for each of these girls and calculated the mean of those numbers, we would get the reading on the green line. So, at birth, for example, that mean is 34 cms, the mean head circumference at birth for baby girls. At eight weeks, the mean is 38 cms, and that is the mean head circumference for baby girls who are eight weeks old. And so on. We get those readings by looking at the scale on that lovely, purply color.

If at each age group we standardize the reading on each girl, get the Z-score—remember, we standardize by subtracting the mean for that age group and then dividing by the standard deviation for that age group—then the green, or mean line is the line when the Z-score is zero. Looking at the right of the chart we see the green line so identified.

We also see the other colored lines identified as Z-score lines for various values of Z ( $\pm 1$  (orange),  $\pm 2$  (red), and  $\pm 3$  (blue)). So these lines give us, for each fixed age, predictive intervals, if you believe in the empirical rule, for percentages of the population. In fact, these head circumferences, for each fixed age, have a distribution that is approximately normal. So the orange lines demarcate approximately the middle two-thirds of the population; the red lines demarcate approximately the middle 95% of the population; and the blue lines demarcate approximately the middle 99.8% of the population.

<sup>1</sup> [http://www.who.int/childgrowth/standards/second\\_set/chf\\_hcfa\\_girls\\_z\\_0\\_13.pdf](http://www.who.int/childgrowth/standards/second_set/chf_hcfa_girls_z_0_13.pdf)

So this chart guides us by showing us how the mean head circumference grows with age, and it also shows us the distributions around these means. It answers the question of what is normal, however we wish to define it, and what is not normal, if we define the latter by measuring the head circumference—for example, we can use it to define macro- and microcephaly.



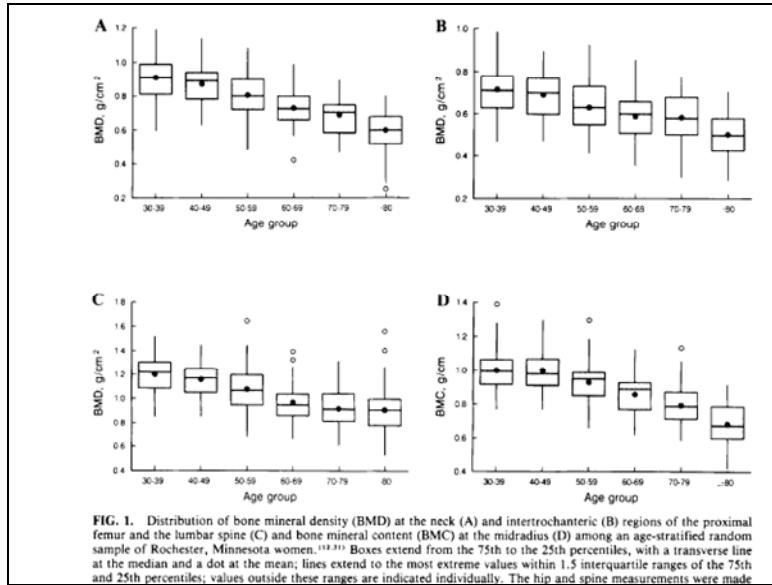
This is a common approach of describing maturation. Sometimes a more complex approach is taken. For example, let us focus on osteoporosis.

One troubling manifestation of osteoporosis is bone fractures; for example hip fractures. These graphs come from the WHO to show us that the numbers of hip fractures are increasingly with time around the world.<sup>2</sup>

Be careful, part of the chart is based on real numbers and part of the chart is based on predictions; it goes out to 2025 and here we sit in 2012. Also, it is a problem for both men and women, but it looks like a much bigger problem for women than it is for men, so let us concentrate on women for this discussion.

---

<sup>2</sup> Note the use of stacked bar charts. Here it works very well, but this graphic technique can sometimes lead to confusing results.



Bone density typically decreases with age. Here are four different parts of the body, where this decrease is evident.

$$T_{\text{score}} = \frac{\text{bone density} - \text{mean of 30 year-old}}{\sigma}$$

World Health Organization Definitions Based on Bone Density Levels	
Level	Definition
Normal	Bone density is within 1 SD (+1 or -1) of the young adult mean.
Low bone mass	Bone density is between 1 and 2.5 SD below the young adult mean (-1 to -2.5 SD).
Osteoporosis	Bone density is 2.5 SD or more below the young adult mean (-2.5 SD or lower).
Severe (established) osteoporosis	Bone density is more than 2.5 SD below the young adult mean, and there have been one or more osteoporotic fractures.

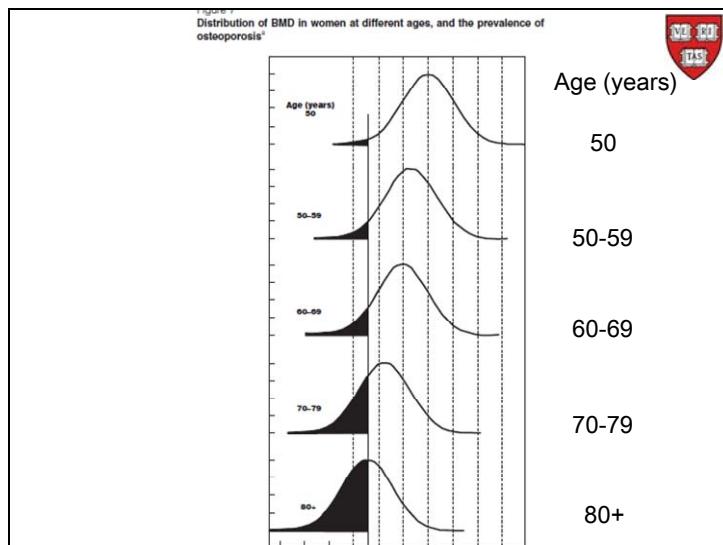
[http://www.niams.nih.gov/Health\\_Info/Bone/Bone\\_Health/bone\\_mass\\_measure.asp#d](http://www.niams.nih.gov/Health_Info/Bone/Bone_Health/bone_mass_measure.asp#d)

What the WHO have done here with the  $T_{\text{score}}$  is define a way of calculating a measure for each person. This resembles standardization but it is different because they subtract a constant mean, namely a number they get by measuring the mean of a 30 year old. These measures are group specific, so if we are dealing with women, we subtract the mean of 30-year-old women.

Had we subtracted the mean of people within the appropriate age group, then we would have the Z-score, just as we did above with head circumference.

The reason for this T-score is to now interpret it as we would a Z-score, but with a difference in the group quantification associated with our favorite numbers; say  $\pm 1$ ,  $\pm 2$ , and  $\pm 3$ . Because the mean we use in defining the T-score is not the correct mean to get the 67%, 95%, 99.8% interpretations we are accustomed to with the Z-score, we need other interpretations for the T-score. What we know is that the correct mean goes down with age, so now the T-score is no longer age interpretable without further modification, if we are interested in population percentages.

In practice, the T-score is used for diagnostic purposes: If the T is above -1, then the person is called "Normal"<sup>3</sup>; if the T-score is between -1 and -2.5, then the person is labeled to have "Low bone mass"<sup>4</sup>; and, if the T-score is less than -2.5 then the person is diagnosed to have "Osteoporosis". The last label is explained above. How many people fall into these categories depends on the age of the person—to repeat, because bone density goes down with age.



Bone density is approximately normally distributed with a drift to the left (smaller mean) as age increases, as we see above. The T-score measures distance from the fixed, 30-year-olds mean. So the age-specific proportion of individuals below a fixed point will increase as age increases, as displayed above<sup>5</sup>.

With this construct of the T-score and having it define osteoporosis, we see that they are linking bone density with osteoporosis. And they are doing it in such a way that as age increases, you are getting more and more people with osteoporosis, and the speed with which this increase is occurring obeys the law given by the normal distribution and the increase in the black area

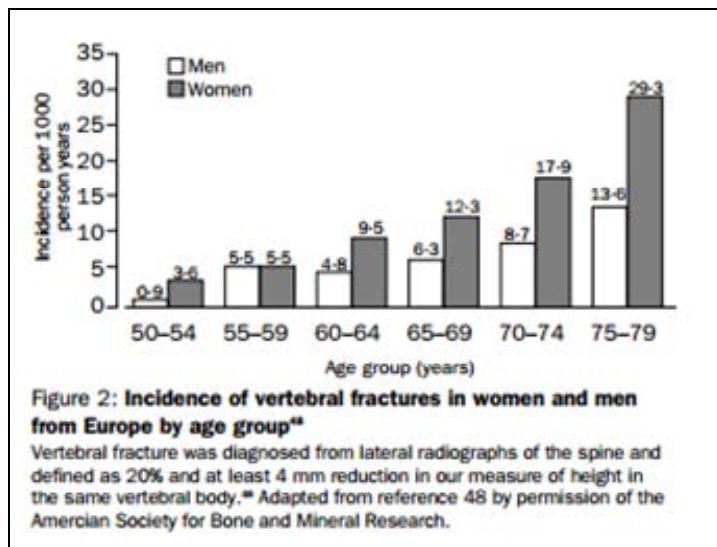
<sup>3</sup> The +1 in the slide must be a mistake.

<sup>4</sup> Sometimes labeled Osteopenia.

<sup>5</sup> [http://whqlibdoc.who.int/trs/WHO\\_TRS\\_921.pdf](http://whqlibdoc.who.int/trs/WHO_TRS_921.pdf)

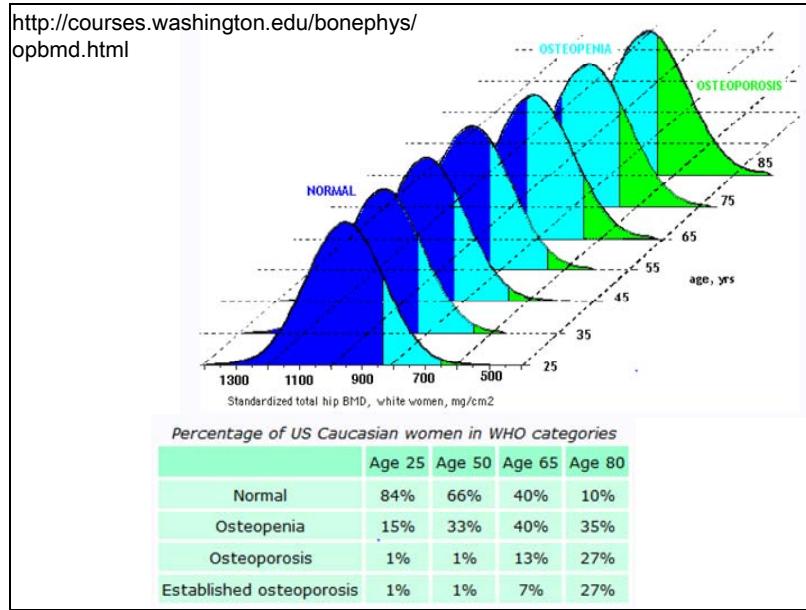
above. This is a rather stringent requirement. It dictates how many individuals are to be labeled with normal, osteopenia, osteoporosis, and severe osteoporosis at every age. It would be interesting to see the empirical verification of this classification.

So it is not that they are just trying to scare you with these T-scores, there does seem to be some thought on how these numbers increase. That this speed is correct, I have not been able to find a justification.



Above are some results in that direction<sup>6</sup>.

<sup>6</sup> SR Cummings and LJ Melton, Epidemiology and outcomes of osteoporotic fractures, *The Lancet*, 359• May 18, 2002 • [www.thelancet.com](http://www.thelancet.com)



Above is a wonderful graphic attempting to show everything we have just been discussing about the T-score.

Note how the distributions are shifting with age, so if the cutoffs remain the same, the proportions in the various classes change as is evidenced by the colors. The distributions change with age. The standard deviations remain the same—homoscedasticity—but the means remain change.

This idea that at every age group we have a distribution is an important one, for the next topic, regression. We have seen this before when we looked at the sex ratio as a function of gestational age; and when we looked at head circumference changing with age. How these changes are linked to each other is our next topic: regression.



## Regression

Galton – “regression to the mean”

**Distribution of one variable (Y)**

- response variable
- (dependent variable)
- osteoporosis, head circumference ...

**as another is varied (X)**

- explanatory variable
- (independent variable)
- bone density, age ...

As opposed to correlation, it is not symmetric in the variables; try to quantify relationship; predict.

The idea behind regression goes something like this: we look at the distribution of one variable, call it Y—for example, osteoporosis, or head circumference—as another variable, call it X—which might be bone density in the case of osteoporosis, or age in the case of head circumference. We look at the distribution of one of them, Y, to see how it is affected by the other variable X.

So the Y is sometimes called the response, or dependent variable, and X the explanatory, or independent variable. Of interest is how the distributions of the response variable vary as we change our explanatory variable. So, for example, as age increases, how does the distribution of head circumference vary? We start trying to answer this question by first zeroing in on a particular characteristic of the response variable, and that is its mean.

So the question we ask is, how do the means of the Ys vary as the value of X varies? This is the technical meaning of the word regression, and it is due to Francis Galton. He actually originally called it regression to mediocrity, but in those days mediocrity meant mean, so you sometimes see it referred to as regression to the mean.

What he had discovered was that when he measured certain characteristics of children and their reported parents in England, there were interesting relationships. For one, if you looked at sons and reported fathers, the tall man would be associated with tall sons, although they were not as tall as the reported fathers. So too with short sons: the short man would be associated with a short reported son, but the son was not as short as the reported father. So, in both of these cases, when going from one generation to the next, the heights of the younger generation had them becoming “closer to the mean”; a regression to the mean.

Regression is different from correlation, although you will see shortly that they are closely related. Correlation between X and Y is symmetric—it is the same as the correlation between Y and X. Regression is statistically different in that we describe the relation between the mean of the Y as X varies, and that is not the same as describing the mean of the X as Y varies. Also, we use regression differently in that the roles of Y and X are different; X is more the

independent variable, the explanatory variable. It is the variable we may be able to set or control some, whereas Y is what results from having set the X. Y is the outcome or dependent variable.

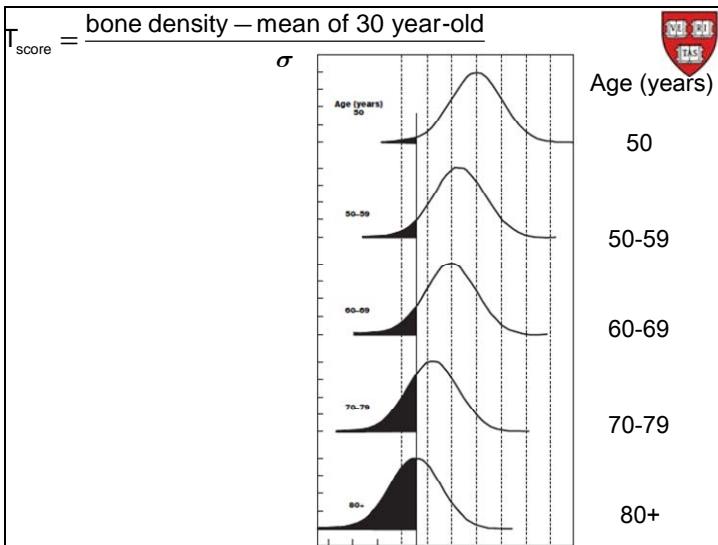


Given a population where we take two measurements on each person, say X and Y.

Fix a value of X and consider all the Ys for that given X.

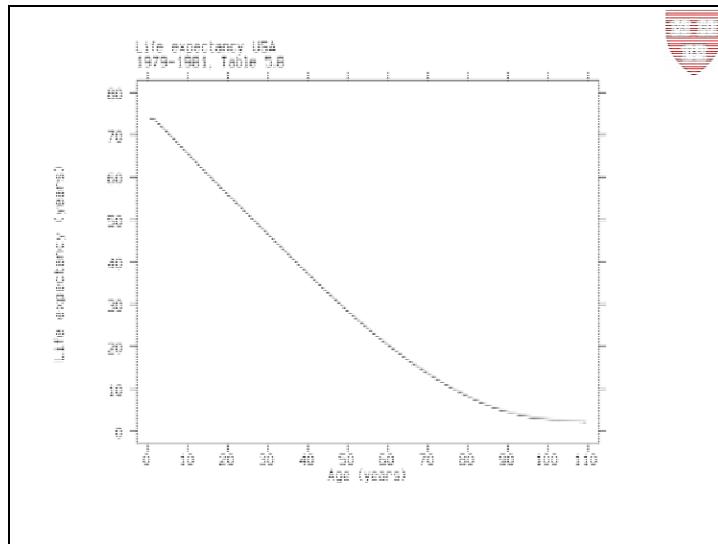
The regression line of Y on X are the means of the Ys for given Xs — it is a function of X.

Here is a formal definition of the regression line. We usually abbreviate it to regression, but we really are referring to a regression line.



<sup>7</sup> [http://whqlibdoc.who.int/trs/WHO\\_TRS\\_921.pdf](http://whqlibdoc.who.int/trs/WHO_TRS_921.pdf)

We can get an idea of what the regression line is from the above. This data set is less than ideal to depict the thought, because the age is not shown exactly, but rather in groups. Take the pedagogical license of acting as if the age was at approximately the center of the intervals: 50, 55, 65, 75 and 85. Now visualize a graph with those ages plotted on the horizontal axis, and on the vertical axis the means of these normal density functions (their centers). Join up those points. That is your regression line of bone density on age.



Here is another example. Look at this curve where for each age on the horizontal, the vertical value of the curve gives the mean life remaining for persons that age. That is the regression line of residual survival on age—how much time, on average, a person still has to live.

Note that the line is quite straight from about age two to about age sixty. If that were so then we would say we have linear regression over that region. We are going to be focusing on linear regression at first.



Here is another example of a regression line (obtainable by joining the dots.) This is the same group for which we earlier saw the sex ratio at birth—the group of about 4.2 million singleton births in one year, 1991, in the US. This time we look at the mean of the mothers' ages for each gestational age.

We also see the plus and minus two standard errors around the mean ages to see the 95% confidence intervals for the means. The widths of these intervals are quite tiny because the sample sizes are so large.

Given the prior statement about what is the meaning of independent and dependent variable, one could argue that we have the classification in reverse.



When we do interchange the roles, this is the regression line of gestational age of the baby on the mothers' ages. The curve is quite smooth until we reach mothers' ages in the mid-forties, but these data reflect conditions in the early nineties, so the points on the right hand side are means of very few mothers.

The whole regression is not linear, but one can see two or three linear segments in this graph. We focus on linear regression for the rest of this chapter.



Low birthweight (<1500grams)

Head circumference (Y)

$$\mu_Y = 27\text{cms}$$

$$\sigma_Y = 2.5\text{cms}$$

and approx. normal.

So, e.g. 95% of kids

$$\mu_Y \pm 1.96\sigma_Y$$

i.e. (22.1, 31.9) is a 95% predictive interval

First we look at what we call *simple linear regression*. It is so called because we have a single outcome variable, a single explanatory variable, and the regression line is straight. We extend this shortly to the situation when we have more than one explanatory variable, and then we consider a certain class of non-linear regressions.

Return now to the group of low birth weight infants. Here we assume that they weighed less than 1,500 grams at birth. Let us look at the distributions of their head circumferences, and how these distributions vary with gestational age. So it is a little similar to the data from the WHO, we looked at above, except that this is a breakdown of a certain subset of that population, at birth.

What we find is that this population is approximately normally distributed with mean 27 centimeters and a standard deviation of 2.5 centimeters. So we can set up predictive intervals etcetera. For example, the 95% predictive interval would be roughly from 22.1 to 31.9 centimeters.

The question we can now ask is can we be more precise in our predictive interval? Is there some other information we can bring to bear that will allow us to construct a tighter predictive interval, let us say? In other words, can we legitimately use a smaller standard deviation? This is what regression allows us to accomplish, and the answer is yes, if we can find a good explanatory variable.



$X$	$\mu_{y x}$	$\sigma_{y x}$
26 wks	24 cms	1.6 cms
29 wks	26.5 cms	1.6 cms
32 wks	29 cms	1.6 cms
:	:	:
All	27 cms	2.5 cms

If we concentrate on babies born at 26 weeks gestational age, we find that their head circumferences are approximately normally distributed with mean 24 centimeters and a standard deviation around that of 1.6 centimeters. So immediately we see that these kiddies have smaller heads (mean 24 versus 27), and more homogeneous (standard deviation 1.6 versus 2.5) than when we ignore their gestational age, and just consider all these kids as a single group. So it looks like gestational age is going to earn the privilege of being called an explanatory variable, in this setting.

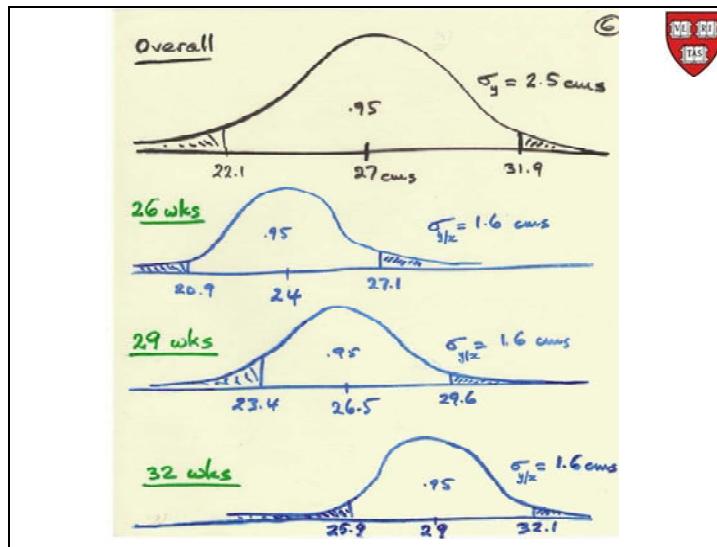
Before we continue, note the new notation we introduce. The mean and standard deviations— $\mu$  and  $\sigma$ , Greek letters because we are talking about the population of these kids for the moment—have subscripts to show that we are talking of the head circumferences, our  $y$ s, and their dependencies on the gestational age,  $x$ .

Now concentrate on the 29-weekers. Their mean head circumference is slightly larger than the 26-weekers—26.5 centimeters versus 24 centimeters—although the standard deviation remains constant at 1.6 centimeters.

Similarly with the 32-weekers, the mean head circumference edges up to 29 centimeters and the standard deviation remains at 1.6 centimeters.

So overall it seems that the mean head circumference, the regression line, goes up with gestational age, and we have homoscedasticity (constant standard deviations) around this regression line. This homoscedasticity is a special property that does not always hold.

Suppose this is an accurate representation of the whole; namely, this regression line and homoscedasticity holds for the other gestational ages. Then by taking the gestational age of the baby into account we have a more homogeneous group and get a tighter predictive interval for the head circumferences.



Schematically, here is what is happening. This is my attempt at free hand. Overall, when we consider the head circumference of all low birth weight infants, the appropriate distribution is the black one on top; a normal with mean 27 centimeters and standard deviation 2.5 centimeters. Breaking this population down into groups defined by the explanatory variable, the gestational age of the infant at birth, we get the blue distributions displayed above; each has standard deviation 1.6 centimeters (homoscedasticity) that is smaller than the overall standard deviation of 2.5 centimeters, and thus the groups are more homogeneous, but their means vary. These distributions drift from the left of the picture to the right as the gestational age increases to reflect the fact that the head circumferences increase with gestational age.

This drift means that the predictive intervals will also drift from left to right. Of course there are kids at 27 weeks, 28 weeks, etcetera but my drawing skills only extend so far and had I tried to include them all in the picture we would have ended up with a mess<sup>8</sup>.

<sup>8</sup> or a Jackson Pollock masterpiece?

If  $X$  and  $Y$  are normally distributed

$$\sigma_{y|x}^2 = (1 - \rho^2) \sigma_y^2$$

where  $\rho$  is the correlation coefficient between  $X$  and  $Y$ .

Now focus on the means of these groups and plot them against gestational age. That is the regression line of head circumferences on gestational age. The means fall on a straight line. We return to that below, but first concentrate on the standard deviations of these various groups. What is the relation between these conditional standard deviations and the overall standard deviation?

If we have linear regression and the distributions are normal, as we do here, and we have homoscedasticity, as we do here, then the variance of the gestational-age-specific distributions are related to the overall variance (ignoring gestational age, just considering all these babies together) by the formula above. Thus the correlation coefficient squared reflects the proportional reduction in the variance afforded by considering the explanatory variable, gestational age, in the argument.

A special case of this is when the correlation is plus or minus one, in which case this, so called, *residual variance*, is zero. (This forms the basis of the tip we used to guess the values of the correlation coefficients when faced with those four scatter plots in the correlation game!)

The other special value of the correlation coefficient is when it is zero. Then there is no reduction in the variance. That means that the independent, or explanatory, variable in question is uncorrelated with the outcome variable and linear regression would not be helpful in defining more homogeneous sub-groups. The explanatory variable does not explain.

So here is connection number one between the correlation coefficient and regression: It is the relative reduction in the variance—or standard deviation—of your measurement, when you introduce the explanatory variable into the argument. When we consider the estimation of the regression line, below, you will be introduced to a statistic, R-squared, that estimates the squared correlation coefficient above. Its importance is predicated by the above equation.

**Variance reduction**



$\sigma_Y = 2.5 \text{ cms}$  &  $\sigma_{Y|X} = 1.6 \text{ cms}$

If X and Y are normally distributed

$$\sigma_{y|x}^2 = (1 - \rho^2) \sigma_y^2$$

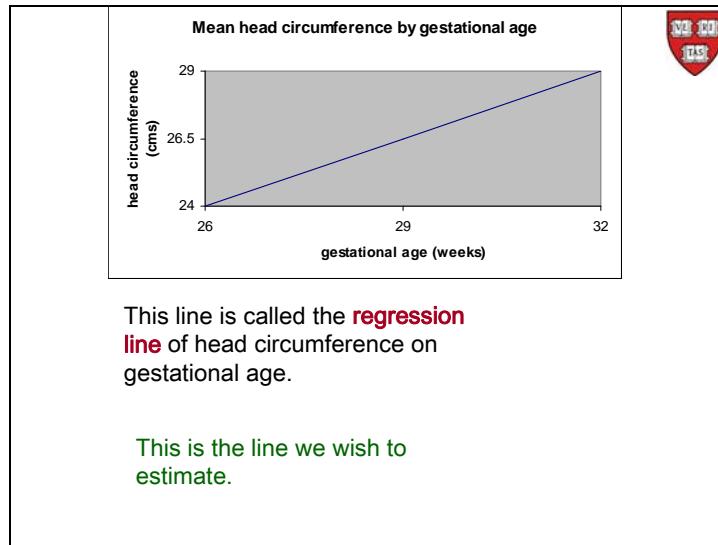
where  $\rho$  is the correlation between X and Y.

$$(1.6)^2 = (1 - \rho^2)(2.5)^2$$

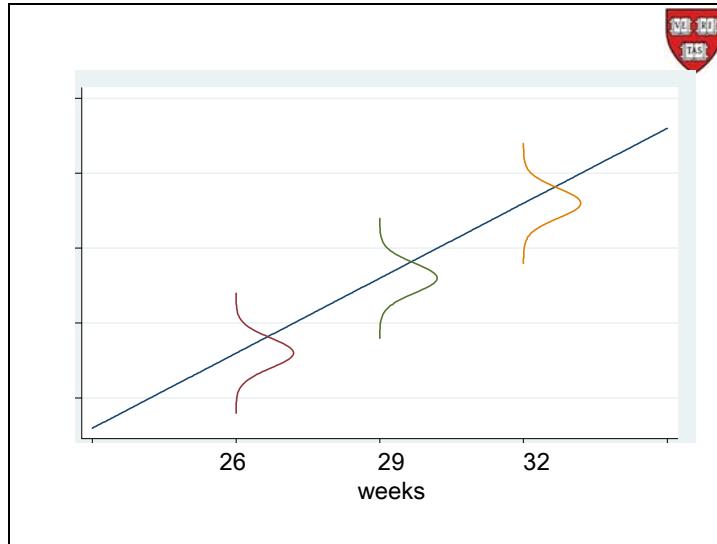
$$\rho = \pm 0.768$$

Note that if  $\rho = 0 \Leftrightarrow \sigma_y = \sigma_{y|x}$   
and x is no help in explaining y.

We can use the equation above to calculate the value of the correlation squared. To then determine which square root to use, we determine the sign by arguing that the head circumference increases with gestational age and thus the correlation is positive.



Returning to the regression line of head circumference on gestational age, this is the line that we eventually want to estimate. Up to now we have inferred values of single parameters such as a mean, or prevalence, or variance, or an odds-ratio, a correlation, etcetera. Now we wish to estimate a relationship in the population that is represented by a line. So inference becomes a little more complex.



The line we want is the line where the means of the distributions lie. To depict that we have coming out of the paper, in a three dimensional plot that not great, but there it is. Actually, I have to thank Nick Cox for drawing this for me.

**Framework:**

1. Linearity  
The regression line is straight.
2. Homoscedasticity  
The stand. dev.  $\sigma_{yx}$  is constant

Needed for inference:

---

3. Gaussian  
For a given x (age) the y (head circumference) are normally dist.
4. Independence  
We have n independent pairs of obs.

The framework within which we operate is: (i) linearity, so we assume that the regression line we want is a straight line; (ii) homoscedasticity, so we assume that each of the distributions coming out of the paper, above, has the same standard deviation. These two assumptions are sufficient to get us going and estimate the regression line in a reasonable fashion. Subsequently to make further inference, we assume that (iii) we have normal data, i.e. the distributions coming out of the paper, above, are each normal; and (iv) our usual assumption that we have independent data.

All of these assumptions can be relaxed, but we delay that to your next course on regression.

### Correlation and slope



If  $X$  &  $Y$  are jointly normal  
(for each fixed  $X$  ( $Y$ ) then  $Y$  ( $X$ )  
is normal) then

$$\mu_{y|x} = \alpha + \beta x$$

and

$$\beta = \frac{\sigma_y}{\sigma_x} \rho$$

Which shows the relation between  
correlation and slope of regression.

Now we come to connection number two between the correlation coefficient and regression: If we look at our regression line and call the slope  $\beta$ , then there we have the relation between the slope of the regression line and the correlation coefficient. So if  $Y$  and  $X$  have the same standard deviation—for example if we have standardized both variables so they each now have standard deviation equal to one—then, in that particular case, the correlation coefficient is the slope of the regression line.

In general, the interesting relationship is when the correlation coefficient is zero (uncorrelated) then we see that the slope of the regression line is zero. So the population means of the  $Y$ s does not depend on the explanatory variable  $X$ . [Theory alert: skip to next paragraph if not interested in theory.] With the conditions we have imposed here, namely normality and homoscedasticity, then we see, since the whole distribution in the normal case is determined by the mean and the variance, that flat regression line, or uncorrelated is synonymous with independence between  $X$  and  $Y$ . [End of theory alert.]

## Equivalence

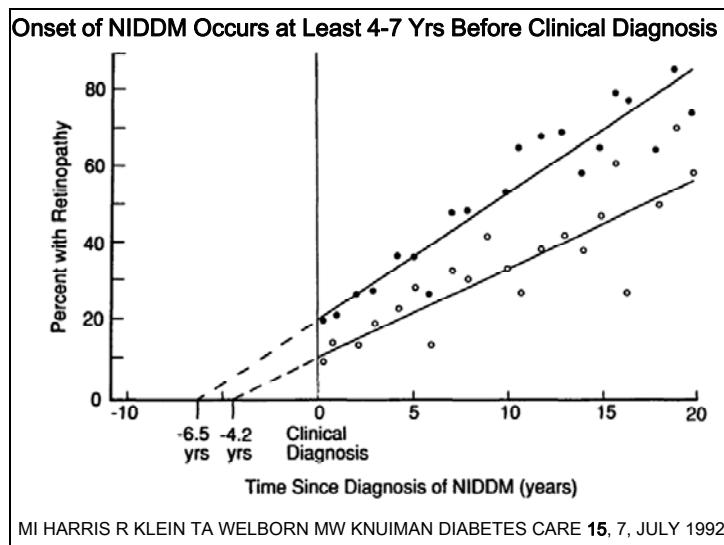


So with normal data the following 3 hypotheses are equivalent:

$$\begin{aligned}
 H_0: \rho &= 0 \\
 \Leftrightarrow \\
 H_0: \beta &= 0 \\
 \Leftrightarrow \\
 H_0: \sigma_y &= \sigma_{y|x}
 \end{aligned}$$

Collecting those two relations between the correlation coefficient and the regression line together, we come up with the equivalence of these three null hypotheses.

## Least Squares



All too often in science we are faced with the generic, here are some points through which I would like to draw a straight line, problem and that sets us off in search of the line that best represents, or summarizes, those points. What we mean by that word 'best' sounds, of course, subjective.

In regression, we seek to determine the regression line associated between an outcome variable and an explanatory variable. Frequently, we cannot afford to fix the value of the explanatory  $x$  and find a huge number of  $y$ s to determine their mean. Then fix another, different  $x$  and repeat the process, and so on and then lay all the means out only to discover that because of sampling variability, experimental error, what have you, the points still do not lie on a perfect straight line. So how do we obtain this population regression line?

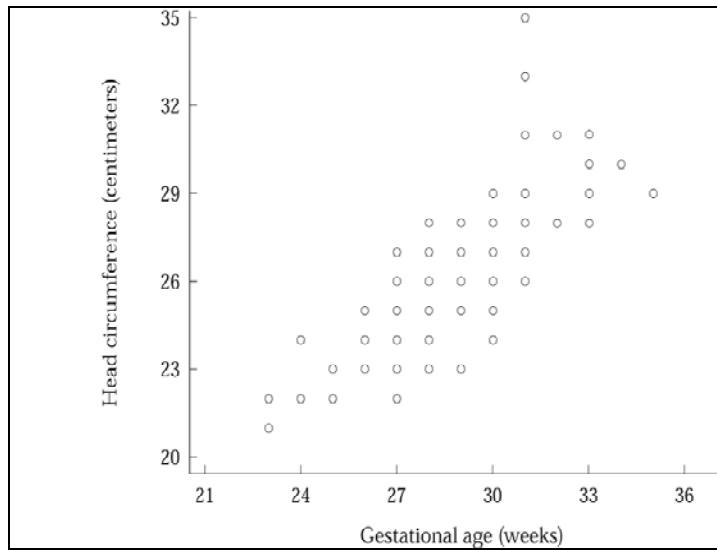
We are going to proceed with how we have learnt to do inference. We are going to take a sample of points and infer from them what the population parameters are. In this case the population parameter is a straight line defined by two parameters, the intercept and the slope of the line.

So let us return to the generic curve fitting problem where we have points through which we would like to fit a straight line. Above we see an example where they looked at the prevalence of people with diabetes related retinopathy plotted as a function of time since diagnosis with Type II diabetes mellitus (NIDDM)<sup>9</sup>.

This study was done in two countries—in Australia and in the US—and we see two sets of points, one set for each country. Their argument is that if we fit a straight line through those points, we would be capturing the relationship that seems to exist when considering the increase in prevalence with time. They then proceed to extend the line to the left of the zero, the time of clinical diagnosis of NIDDM, to argue that Type II diabetes exists before current clinical diagnoses are done—about 4.2 years in the one and 6.5 years in the other country—as seen by this incubation period for retinopathy. We leave that argument to the experts, and concentrate on the line fitting issue. So we concentrate on the generic problem of trying to pass a straight line through a collection of points.

---

<sup>9</sup> MI Harris, R Klein, TA Welborn, MW Knuiman, Onset of NIDDM Occurs at Least 4-7 Yr Before Clinical Diagnosis, *DIABETES CARE*, 15, # 7, JULY 1992



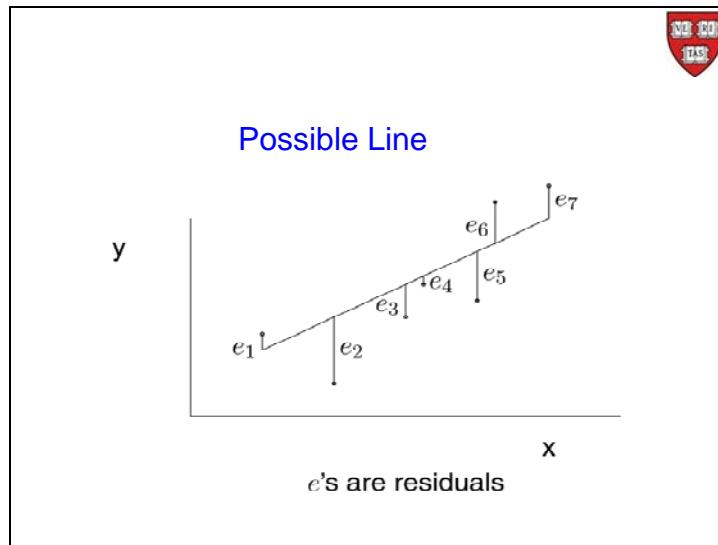
So here are head circumferences for 100 low birth weight, premature babies we introduced above. Care should be taken in reading this graph because some points represent more than one baby. See Page 27, below, for a jittered version of this graph.

For each gestational age we could find the mean head circumferences and fit a straight line through those points, but chances are they would not lie on a straight line. So let us forego that approach and just treat these as general points and see where that leads us.

There are an infinite number of lines that we could draw and claim for each one that it “represent” these points. Heuristically, on the basis of an eyeball test, sometimes facetiously called the intraocular test if you want to impress someone, we can discard a large number of these lines. In an attempt to devise a more objective method for choosing between candidate lines, for each line we draw we can see how well it does by each point by calculating the vertical distance between the line and the point. Ideally we would love all those distances to be zero, but that is not going to happen, unless the points all lie on a straight line.

The length of these vertical distances between each point and the straight line are called *residuals*. Since we cannot make them all zero, we can seek to make them as small as possible, and since we are talking about a number of them, we can consider their sum. If we do that we run into the very same problem we ran into when we looked at the variance. Then we had that when we subtracted the mean from every observation then the sum became zero, so too here, any line that goes through the center of gravity—the point  $(\bar{x}, \bar{y})$ —results in the sum of the residuals being zero, because the positive residuals cancel out the negative residuals. (What you consider a positive and what you consider a negative residual depends on whether you are calculating distance from the line to the point, or the distance from the point to the line, and whether up is plus and down minus, or *vice versa*.)

So get rid of the signs of the residuals!

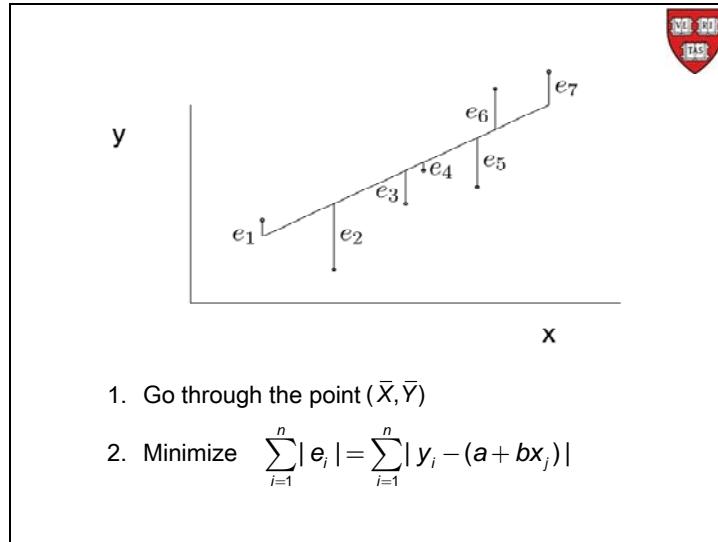


So here's, for example,  $e_1$  through  $e_7$ . Look at the absolute value of those residuals and make their sum as small as you can.



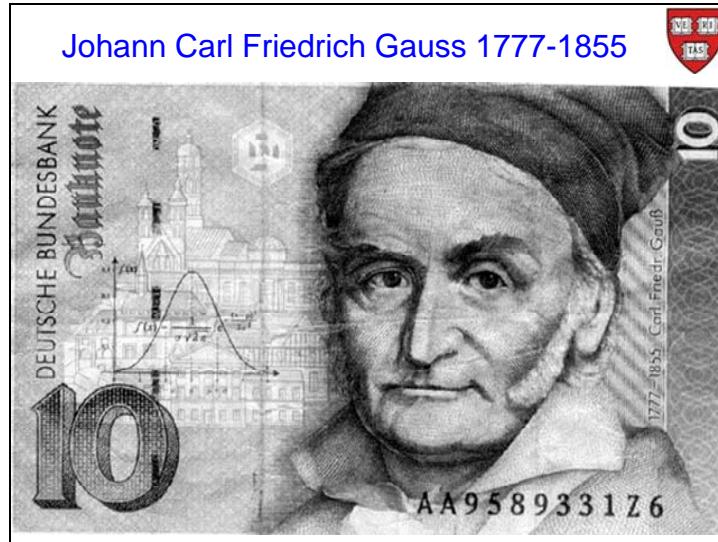
And that was the brilliant idea of Ruggero Giuseppe Boscovich<sup>10</sup>. He was attempting to estimate the distance to the moon, and he had a number of conflicting measures, when he came up with this idea.

<sup>10</sup> The article by Churchill Eisenhart, "Boscovich and the combination of observations" in *Roger Joseph Boscovich, S.J., F.R.S., 1711-1787 : Studies of His Life and Work on the 250th Anniversary of His Birth*, Whyte, Lancelot Law, Ed, Fordham University Press., New York, 1961.



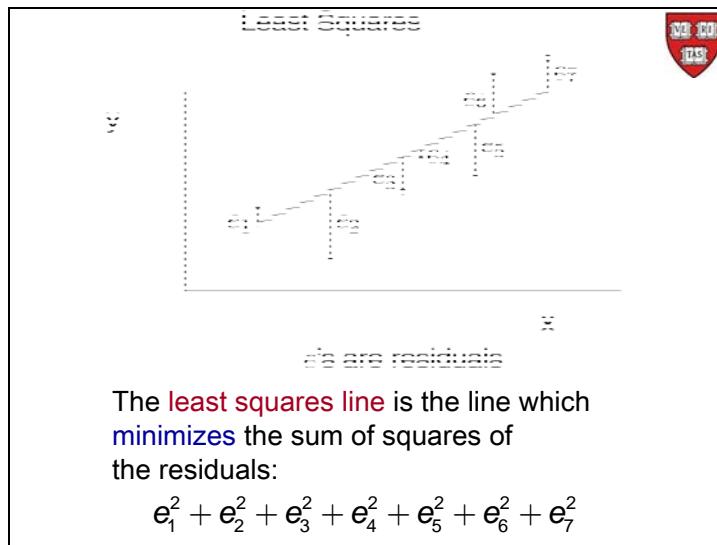
Not only did Boscovich have the idea, but he also gave the algorithm for actually calculating the best straight line; *viz.* the line that goes through the center of gravity and minimizes the sum of the absolute values of the residuals.

There was nothing wrong with his solution, it is truly a lovely algorithm and the basis is very strong, but it did not extend to more than one explanatory variable. As a result it was superseded by the idea of Gauss—although, as usual, there is a dispute about who actually invented it!<sup>11</sup>



<sup>11</sup> Stephen M. Stigler, Gauss and the Invention of Least Squares, *Annals Statistics*, 9, 1981, 465-474.

Remember Gauss? We looked at this Deutsche Mark because of the normal curve.

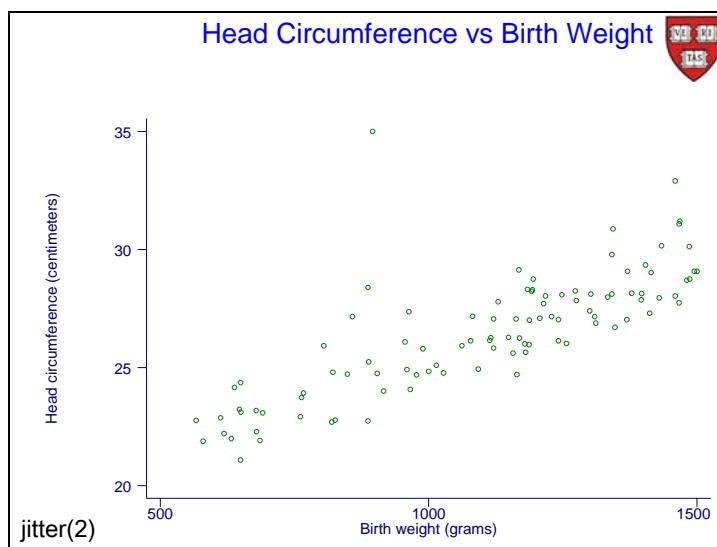


Gauss not only suggested that we minimize the sum of the squares of the residuals—just like when we looked at the variance, we squared the distances from the mean before averaging them out to get rid of the problem of the positive residuals cancelling out the negative ones—but he also provided a wonderfully easy way to actually calculate the slope and intercept of such a line—which, of course, is what Stata does for us. His line, just as Boscovitch's, also goes through the center of gravity.

So amongst all lines that go through the center of gravity, Boscovitch's line minimizes the sum of the absolute residuals, and Gauss's least squares line minimizes the sum of the squares of the residuals. This is a little bit of a shortcoming of least squares in that large residuals get their values squared, thus making them even larger, and this allows them to exert too much influence sometimes. More about that shortly when you explore least squares, below.

	Headcirc	length	weight	tox	momage	sbp	sex	grmhem	gestage	apgar5
27	41	1360	0	37	43	1	0	29	7	10
29	40	1490	0	34	51	1	0	31	8	10
30	38	1490	0	32	42	2	0	33	0	10
28	38	1180	0	37	39	2	0	31	8	10
29	38	1200	1	29	48	2	0	30	7	10
23	32	680	0	19	31	1	1	25	0	10
22	33	620	1	20	31	1	0	27	7	10
26	38	1060	0	25	40	2	0	29	9	10
27	30	1320	0	27	57	2	0	28	6	10
25	34	830	1	32	64	2	0	29	9	10
23	32	880	0	26	46	2	0	26	7	10
26	39	1130	0	29	47	2	1	30	6	10
27	38	1140	0	24	63	2	0	29	8	10
27	39	1350	0	26	56	2	0	29	1	10
26	37	950	0	25	49	1	0	29	8	10

Here are the first fifteen observations in the dataset, lowbwt (low birth weight) a sample of 100 babies whose birth weight is less than 1500 grams. The first variable (headcirc) is head circumference at birth. The next (length) is the length (height, except babies do not stand up!) of the baby in cms, followed by the babies' birth weight (weight) in grams. Whether the mother was toxemic (1) or not (0) is noted in the next variable (tox). Her age is recorded in the next variable (momage) as is her systolic blood pressure (sbp) in mm/Hg. The babies sex is next recorded (sex) where 0=female. Next is recorded whether the baby had a brain hemorrhage, in the germinal matrix (grmhem) with 1=yes. The baby's gestational age at birth is reported in weeks (gestage), and finally the Apgar score at 5 minutes is in the last variable (apgar5).



So here are the plotted data, except that the data has been jittered. I refer you to the Stata manual to see what that means exactly, but roughly what that means is that each data point is

moved a smidgen so that no two points fall exactly on each other, since if that were the case it would look like we had a much smaller data set, as happened on Page 23.

The relationship between head circumference and birth weight is arguably linear in this scatter plot.

. regress headcirc gestage					
Source	SS	df	MS	Number of obs = 100	
Model	386.867366	1	386.867366	F( 1, 98) = 152.95	
Residual	247.882634	98	2.52941463	Prob > F = 0.0000	
Total	634.75	99	6.41161616	R-squared = 0.6095	
				Adj R-squared = 0.6055	
				Root MSE = 1.5904	
headcirc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gestage	.7800532	.0630744	12.37	0.000	.6548841 .9052223
_cons	3.914264	1.829147	2.14	0.035	.2843817 7.544147

Fitted (least squares) regression line:  

$$\text{headcirc} = 3.914 + 0.78 \text{ gestage} + e$$
where standard dev of e = 1.59

The Stata command for fitting the least squares line is simply *regress* followed first by the response variable (*headcirc* in this case) and then by the explanatory variable (*gestage* in this case). What Stata then does is it calculates the least squares line for us, and provides us with the statistics to allow us to go ahead with the inference we wish to make.

From this we first see that there are 100 data points. Then the bottom two lines tell us about the two variables defining the regression line estimate. The bottom line is the constant (*\_cons*) and the line above that is for the gestational age (*gestage*) coefficient.

So we see that the least squares line has constant 3.914 and slope 0.78. That means that for every week extra gestational age, the head circumference at birth will increase by 0.78 centimeters.

When we think of these two numbers as estimates of population parameters, we can see their standard errors, their t-statistic to test the null that each is zero, in turn, and the associated p-values (0.000 for the coefficient associated with gestational age, and 0.035 for the constant). We are usually not concerned with the constant term in situations like this, since it is telling us about something, crossing the axis, at gestational age zero which is meaningless and far outside the experimental region, anyway. The 95% confidence interval for the gestational age coefficient in the population regression line is (0.65, 0.91). So we would reject the null hypothesis that the head circumference is not impacted by gestational age, on the basis of these 100 observations.

The last statistic to look at is the R-squared, as advertised above, which is the estimate of the correlation coefficient squared, or reduction in variance. I would recommend looking at the adjusted R-squared, especially when we have more than one explanatory variable.

---

I leave the transcript of what happened with the visit to the National Council of Teachers of Mathematics website, but strongly urge you to watch that video and visit the site.

---

So let's go to a website run by the National Council of Teachers of Mathematics. And they've got this lovely little applet here that we can use to get some idea of how least squares works. In particular, what I'd like to show you is that least squares can be very sensitive to 1 or 2 points. And since we saw that the least squares line is also intimately related with the correlation coefficient, this will reinforce the statement that I made last week that the correlation coefficient is also sensitive to outliers.

So here's what we do. We go and we plonk down some points. So here are a few points. Or so here there's, what, two, four, five. There's another point. And then you can say, show the line. So this line is the least squares line through those five points. And we see that it's almost a perfect fit.  $R$  is equal to 1 with an intercept of 4.1 and a slope of 1.36.

Now the question to ask is, what happens if a point or two gets shifted around. Well if we go here and we shift this point around, move to there, look at that. It didn't have much impact. The slope is still 1.34. The  $r$  is 0.99, so it didn't do much. Move this one a little bit. Move the center one's a little bit and nothing much happens. With this one line started moving a little bit, but we're still at 0.98 for  $r$ .

Now what happens on the other hand, if we choose one of the extreme points. And this is when we're going to start worrying, and we start seeing things happen. Especially if we move it. Look at this. It's dragging the line with it. So if we move it radically—so for example, if we move it all the way down here then look at that. The correlation coefficient now is minus 0.27. So it's gone from a very high positive to negative.

And things get worse if I also move this one, so all of a sudden, we get a completely different picture of what's going on. So we're at  $r=-0.64$ . And you might say, oh well look you moved two of the five, so you'd expect this to happen. Well OK, what happens if we add some more points along the original line. And look at this, it's not having a tremendous effect.

Yeah, it's having some effect, but look at that. So here we've got 17 points. So now two of the 17 are way off, but the other 15 of the 17 don't have enough influence to carry the day. So we can carry on like this if you want, and look at what it's going to take to shift the line around. It's going to take quite a bit to bring the line around. It still only a quarter of the way there. We're still at only 0.24.

Still just these two points, these two extreme points, carry a lot of influence. So that's the moral of this story. Now what I want you to do, is I would like for you to do this by yourself. The

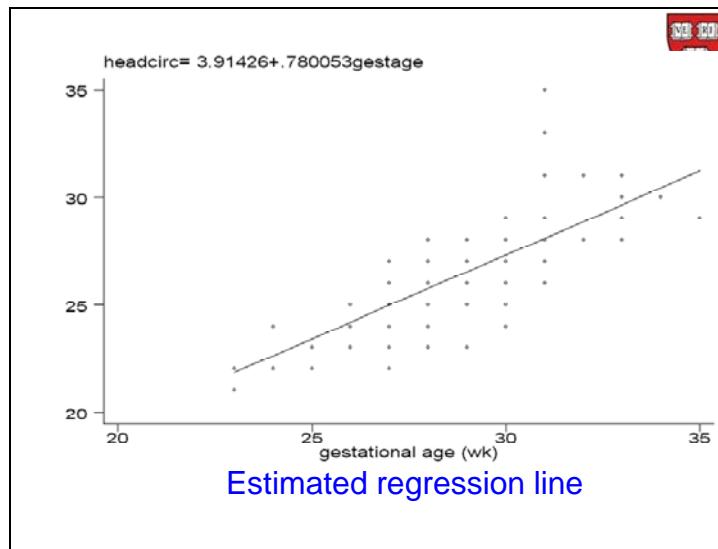
instructions are down here on the screen, and you can see how to do this. It is very easy. And there's the website down there too. All right.

Now, while you're here, we can also revisit that statement that I'd made to you, remember, if you had the v shaped, if you have  $y$  is equal to the absolute value of  $x$ . And I said, oh look, that's probably going to give you a correlation of 0. Well let's try that and see what happens when we fit a line here. Look at that, 0.05. So that bears out that statement that I made last week with a correlation coefficient.

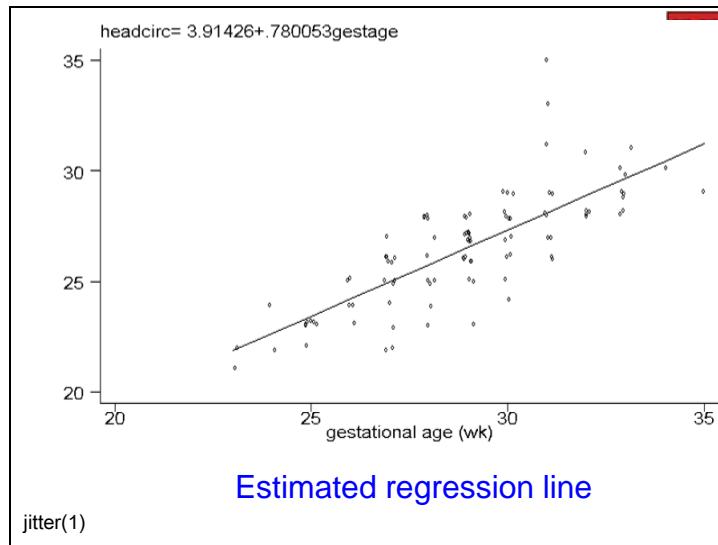
In fact, remember we looked at the fatal automobile accidents and what percentage had an alcohol content above 0.08. So we had age along the horizontal, and then we had the percentages on the vertical. And once again, there it is. The correlation coefficient is equal to 0.00. This is a little bit more extreme than what we had before, but you get the idea.

So go there and try this for yourself, and you'll get a little better feel for how the least squares line works. We could have stuck with what Boskovic did, but that's for the next course. The way you will learn about mean absolute deviation, or median regression, or  $L^1$  regression it's called. There are more robust ways of doing regression, but go ahead and enjoy yourselves.

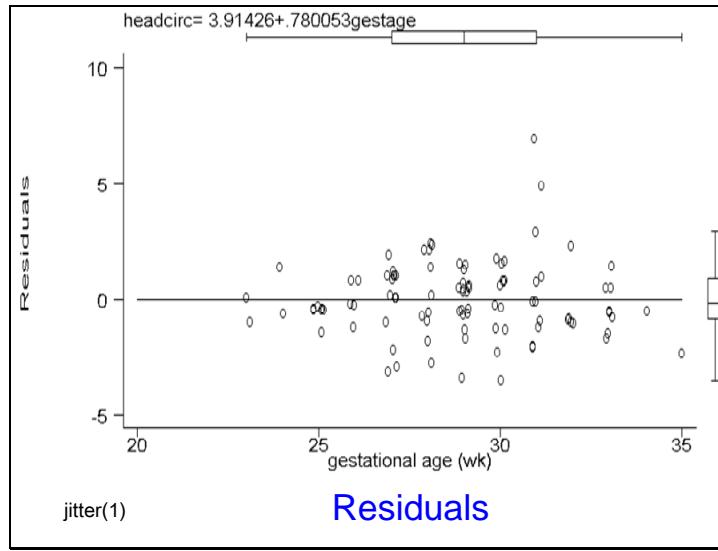
---



Returning to the original data—non-jittered—showing the estimated least squares regression line. In this case it looks like the straight line is not a bad fit. What if it were not? Before answering that question let us see what might lead us to such a conclusion.

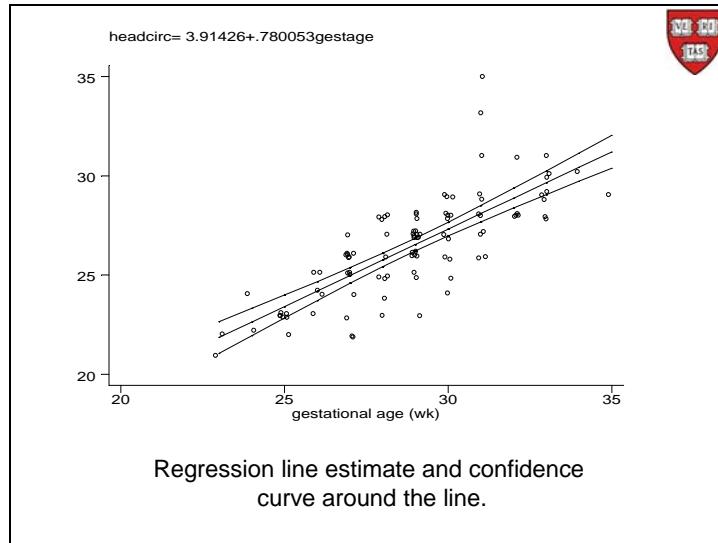


Let us look at a jittered version of the data, and that should not change our opinion that the relationship is linear.



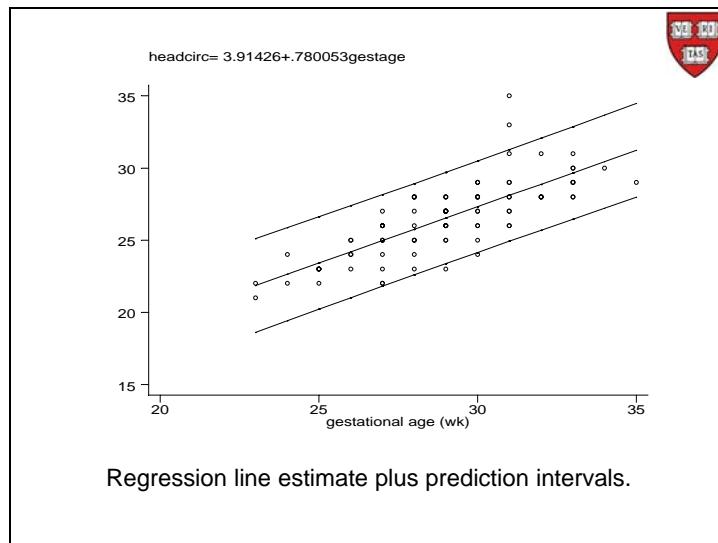
We can also subtract regression line from all the points thus creating the residuals. Now remember we said that we were assuming homoscedasticity for these. On looking at the graph we do not see any residual shape in these—they are not all negative at the left and positive at the right, for example—they look pretty much evenly distributed about zero, and their size is

evenly distributed and thus our assumption of linearity and homoscedasticity seems well supported by the data.



Remember the line is a statistic, it can vary from sample to sample. We can place a 95% confidence interval around this line that has the following interpretation: For any fixed gestational age, this interval serves as a 95% confidence interval for the mean (the regression line) of all kids with that gestational age at birth.

We see that the interval is tighter near the center than it is on the extremities, and that intuitively makes sense because most of our data are near the center, and so that is where we have most of our information.



We might also be interested in where 95% of the babies will be, and that can be answered with this estimated prediction interval. For example, our least squares estimate of where 95% of the babies' head circumferences are for babies who are born at 25 weeks gestational age, then that is given by this graph when evaluated at 25 weeks.

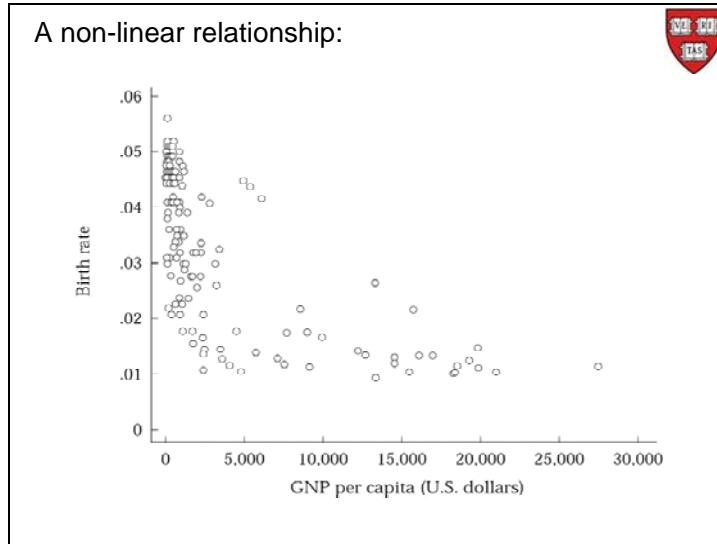


Strategy for regression:

1. Draw a scatter plot of the data.
2. Check residuals.
3. Plot residuals versus fitted ys.  
There should be no discernible pattern

So in summary, what is our general strategy for simple linear regression? We start by drawing a scatter plot of the data. That will give us an idea of what the relationship is and whether we in fact have linearity or not. Then once we fit the model we should plot our residuals to see if there are any discernible patterns. The patterns to look for are first to see if there are any trends going up or down or U-shaped etcetera that would reveal non-linearity. Second we should also look for patterns of non-homoscedasticity—for example are the residuals flute-shaped to denote more variability at one extreme or the other.

There should not be any discernible patterns. Also be on the lookout for outliers since they may overly affect your results, as you noticed when exercising with the applet above.



Here is an example of a scatter plot we would rather not see if we are about to fit a linear regression line. This plot shows the birth rate for some countries against their per capita gross national product (GNP) in US dollars. The relationship between the two is certainly not linear.

Some shapes we can rectify and transform the data to obtain a linear relationship. What do we mean by transform and what kind of transform to investigate. By transform we are talking about instead of looking at the measure directly, a transform of it might be more appropriate. For example, we all know about the inverse square law in physics, Newton's law of gravity is one example of this. It says that the intensity of some quantity is inversely proportional to the square of the distance from the source of that quantity. Now, if instead of distance, we invented a new quantity called  $tsidtsid$ , say, which is calculated by taking the reciprocal of the square of the distance, then the intensity would now be linearly related to  $tsidtsid$ . So we would have the  $tsidtsid$  law and we could draw a linear regression of intensity on  $tsidtsid$ .

Conversely, if instead of the intensity of the quantity we measure the reciprocal of the square root of the quantity, then it would have a linear relationship with the distance.

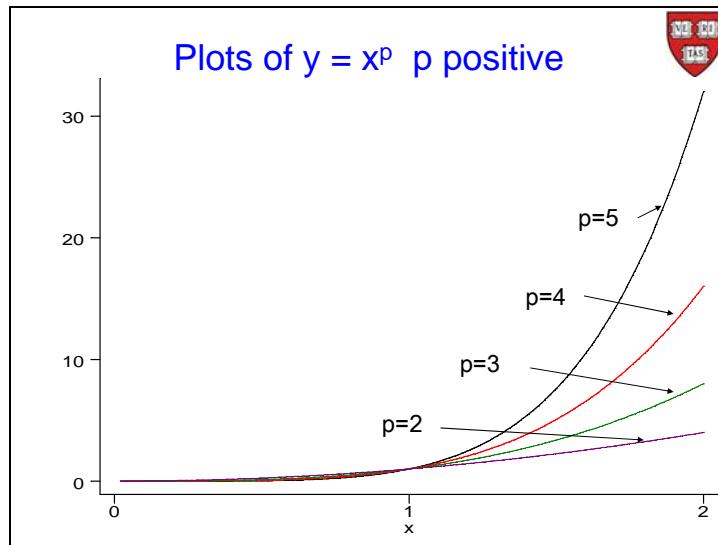
What all this is saying is if

$$I = \frac{1}{d^2} \quad \text{then if} \quad tsidtsid = \frac{1}{d^2} \quad \text{then} \quad I = tsidtsid.$$

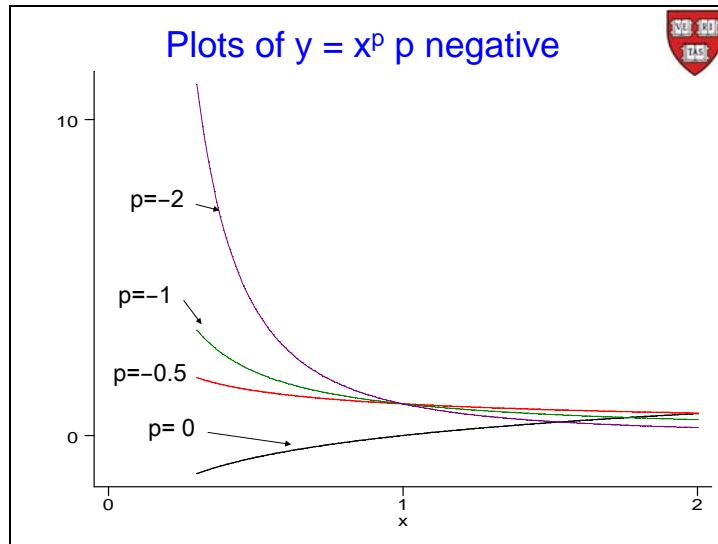
Or, alternatively, if above, then

$$\frac{1}{\sqrt{I}} = d.$$

So sometimes by exploring scatter plots we can find a way to transform our data to achieve linearity.



Consider this family of curves. They show the relationship of  $y=x^p$  for various positive values of  $p$ . As  $p$  gets larger, the rate at which  $y$  increases with  $x$  gets steeper and steeper. So if we take the inverse transformation,  $x^{1/p}$ , then that would slow the rate of increase of  $y$ , and in fact achieve linearity.



With the power  $p$  negative we get different shapes. Like this we can build up an armamentarium of transformations to bring to the problem of straightening out curves.



Options for running the regression in such cases:

1. Do a non-linear regression.

2. Transform the ys or the xs:

i.e. look at  $y^p$  or  $x^p$  e.g.

$$\begin{aligned} y^{-2}, y^{-1}, y^{-0.5}, \ln(y), y^{0.5}, y, y^2 \\ x^{-2}, x^{-1}, x^{-0.5}, \ln(x), x^{0.5}, x, x^2 \end{aligned}$$

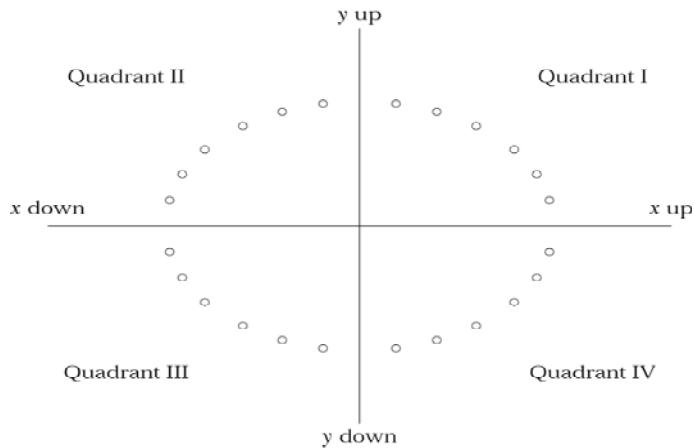
going up the ladder →

going down the ladder ←

We can create a ladder of transformations starting on the left with a large negative value for p and moving right by increasing the value of p. To fill out the ladder we can replace p=0 with the logarithm. That way we can search up and down the ladder to see if anything straightens out our relationship to allow us to fit a linear regression.



To straighten the relationship:

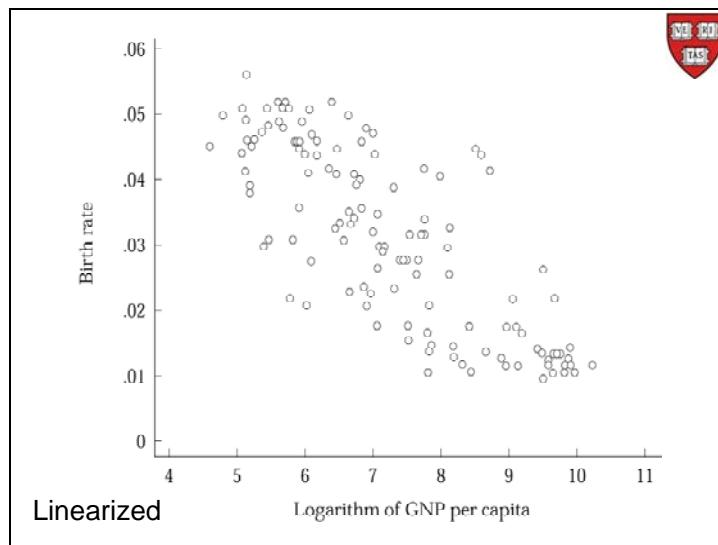


To guide us in our search of the appropriate transformation we have this quadrant guide due to John Tukey<sup>12</sup>. The way we use it is we start with the power of y is zero and the power of x is zero. Then identify the shape of the scatter plot by seeing which quadrant it falls into. If the

<sup>12</sup> Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. Also look at <http://onlinestatbook.com/2/transformations/tukey.html>

shape that best describes the scatter plot is that in Quadrant I, then either up the power on x, and or up the power on y to achieve linearity. Quadrant II would suggest upping the power on y and or lowering the power on x. And so on.

Why upping the power on y might be different than lowering the power on x might be because of the error structure that best describes the situation at hand.



Returning to the birth rate – GNP relationship above, we might identify that with the shape in Quadrant III, and that suggests looking at a decrease in the power of either x or y. In this case, replacing y with its logarithm, results in this scatter plot which reveals a relationship which is much closer to linearity.

Sometimes these transformations work and sometimes they do not, but they are always worth a try.

## Multiple Linear Regression



Multiple Regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q + \varepsilon$$

**Assume:**

1. For fixed  $x_1, \dots, x_q$ ,  
 $y$  is **normally** distributed with  
 mean  $\mu_{y|x_1, \dots, x_q}$   
 and standard deviation  $\sigma_{y|x_1, \dots, x_q}$

All too often a single explanatory variable is not sufficient to fully explain the changes in the distribution of a response variable in which case we might introduce more explanatory variables. To handle these extra variables, let us consider our assumptions just like we did before.

We still assume normality of our response variable, except that now we want that this normality to be true for a fixed collection of xs. So instead of just gestational age we might also want to include the babies' birth weights, the mothers' ages and so on, bring in a number of factors.

Now fix all those, and we shall have a number of ys, just like before, and we are going to assume that they are normally distributed, with a particular mean, and those means are going to lie on a plane. Around those means we are going to have standard deviations.

continued



2.  $\mu_{y|x_1, \dots, x_q}$  is linear in  $x_1, \dots, x_q$   
i.e.

$$\mu_{y|x_1, \dots, x_q} = \alpha + \beta_1 x_1 + \dots + \beta_q x_q$$

3. Homoscedasticity  
i.e.

$$\sigma_{y|x_1, \dots, x_q} \text{ is constant}$$

4. The Ys are independent.

$$\text{Minimize } \sum_{i=1}^n (y_i - a - b_1 x_1 - \dots - b_q x_q)^2$$

We assume that the mean is the linear in all xs, and whereas before we fit a straight line, now we are going to fit a plane. And we are going to retain our assumption of homoscedasticity. Once again, the ys are independent. So it's the same assumptions as before, except we're talking about a plane instead of a line. Not a big change.

Once again, we, or rather, Stata, are going to minimize the sum of squares of residuals.

. regress headcirc gestage weight							
Source	SS	df	MS				Number of obs = 100
Model	477.326905	2	238.663453	F( 2, 97) =	147.06		
Residual	157.423095	97	1.6229185	Prob > F =	0.0000		
Total	634.75	99	6.41161616	R-squared =	0.7520		
				Adj R-squared =	0.7469		
				Root MSE =	1.2739		
headcirc	Coeff.	Std. Err.	t	P> t	[95% Conf. Interval]		
gestage	.4487328	.067246	6.67	0.000	.3152682	.5821975	
weight	.0047123	.0006312	7.47	0.000	.0034596	.005965	
_cons	8.300015	1.570943	5.26	0.000	5.174251	11.44170	

$$\text{headcirc} = 8.3 + 0.45 \text{ gestage} + 0.0047 \text{ weight} + e$$

where standard dev of  $e = 1.27$

Versus:

$$\text{headcirc} = 3.914 + 0.78 \text{ gestage} + e$$

where standard dev of  $e = 1.59$

Here is an example of multiple regression. The Stata command is as before except we now add another explanatory variable to the list, namely *weight*, the birth weight of the babies. Let us see if adding this extra explanatory variable helps us or not in understanding head circumference.

We see the two regression lines, the first without birth weight and the second with birth weight added. The coefficient for gestational age has decreased some in the presence of birth weight. The interpretation of this is that in the first equation gestational age, which is correlated with birth weight, was carrying the explanation afforded by both, whereas with birth weight in the equation it can do its own work. Once again, by looking at both, the p-value and confidence interval associated with birth weight we see that it is a significant explanatory variable. We also see that the standard deviation of the residuals has gone down from 1.59 to 1.27 centimeters, a 20% decrease. So it seems that when looking at head circumference of a baby, both its gestational age and its weight at birth are important.

Note that the coefficients are not pure numbers, they both depend on the units of measurement being utilized. For example, the 0.45 coefficient of gestational age has units of cms/week, and the 0.0047 coefficient for weight is for cms/gram. Had the babies weights been measured in Kilograms, then the coefficient would have been 4.7 cms/Kg.

The other statistic to keep an eye on is the R-squared. It went from 0.6 to 0.75, a good improvement.

The main problem with multiple regression is that the way we have done the modeling here is we have acted as if these two explanatory variables, gestational age and weight, are additive. Is it possible that there is some interaction between gestational age and weight? Once we have opened the Pandora's box the two variables, and later with even more, may also interact with each other in a complicated way. We return to that but first introduce another class of variables.

## Indicator Variables

**Indicator Variables**



e.g. toxemia {  
 1 = yes  
 0 = no

Estimated regression equation:

$$\hat{y} = 1.50 + .874 \text{ gestage} - 1.41 \text{ tox}$$

For toxemics:

$$\hat{y} = .083 + .874 \text{ gestage}$$

For non-toxemics:

$$\hat{y} = 1.50 + .874 \text{ gestage}$$

An important set of variables we can introduce as explanatory variables are the dichotomous variables. They are the ones we have been using to build tables. In the regression context they are called indicator variables. The ideas we are about to develop extend in a straightforward manner to categorical variables also, but let us just look at the 2-valued, simple ones.

Sometimes in the literature you will see these called dummy variables, but why inflict that on them.

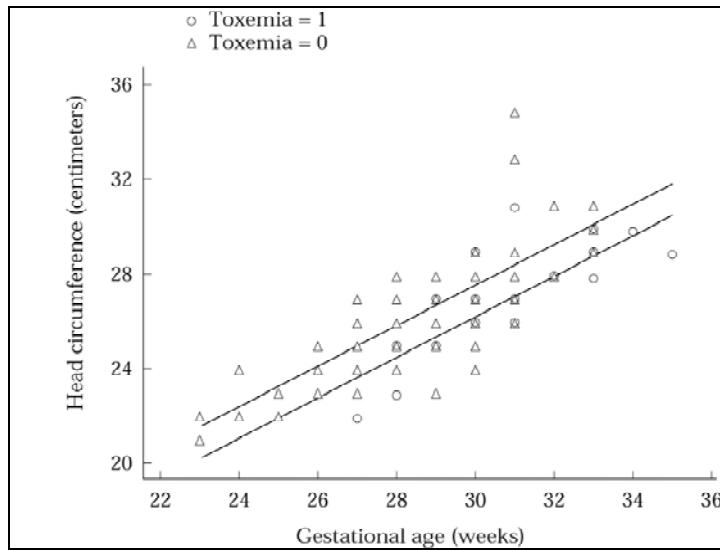
For example, let us look at toxemia. If the mother was toxemic at the time of delivery, then we let the *tox* variable take the value 1. If the mother was not, then this variable will take on the value 0. Above we see the result of fitting a regression line with the variables gestational age and toxemia.

We can view this as two regression lines, one when the mother was toxemic and one when she was not. Above we see the definitions of these two lines, and we note that they are parallel with different intercepts.



regress headcirc gestage tox					
Source	SS	df	MS	Number of obs = 100 F( 2, 97) = 91.18 Prob > F = 0.0000 R-squared = 0.6528 Adj R-squared = 0.6456 Root MSE = 1.5074	
Model	414.342993	2	207.171497		
Residual	220.407007	97	2.27223718		
Total	634.75	99	6.41161616		
headcirc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gestage	.87404	.065608	13.32	0.000	.7438262 1.004254
tox	-1.412335	.4061539	-3.48	0.001	-2.218438 -.6062316
_cons	1.495575	1.867993	0.80	0.425	-2.211874 5.203024

We obtained all this information from this Stata output. The indicator variable denoting toxemia is significant, although the R-squared is not as good as it was when we introduced birthweight into the regression.



Here is a plot of the two parallel lines. What these say is that an extra week of gestational age as far as the head circumference goes has the same average effect whether the mother suffered from toxemia or not. They start from different starting points, but the effect of an extra week is the same.

We can investigate whether this makes physiological sense or not.

Two-sample t test with unequal variances					
Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
0	79	26.27848	.2831046	2.516289	25.71486 26.8421
1	21	27.09524	.5559409	2.547641	25.93557 28.25491
combined	100	26.45	.2532117	2.532117	25.94757 26.95243
diff		-.8167571	.6238738		-2.088858 .4553441
diff = mean(0) - mean(1) t = -1.3092					
Ho: diff = 0 Satterthwaite's degrees of freedom = 31.1801					
Ha: diff < 0 Pr(T < t) = 0.1000		Ha: diff != 0 Pr( T  >  t ) = 0.2000		Ha: diff > 0 Pr(T > t) = 0.9000	

We can perform a t-test between the two groups of infants and find we cannot reject the null hypothesis of no difference in head circumference even though we did find a difference when we included birth weight in the model.

Toxemia happens later in the pregnancy. We saw that later in the pregnancy, as measured by gestational age, does have an impact on head circumference. Also, birth weight increases with

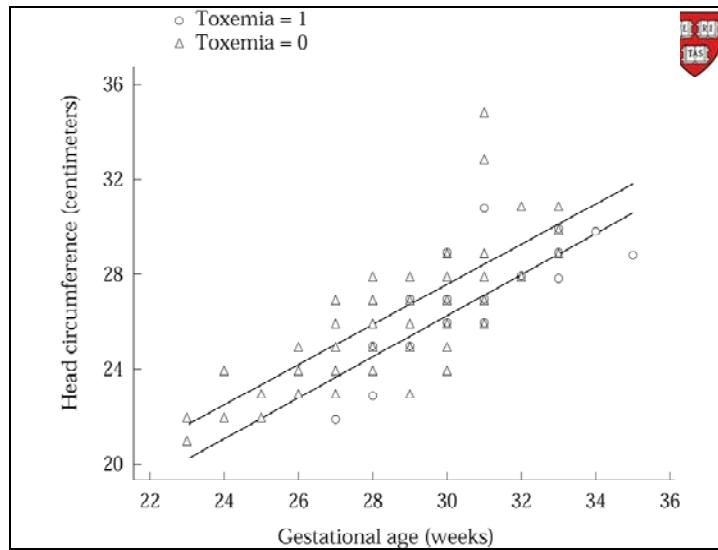
gestational age. So how all of these explanatory variables work together is quite complex. Is there an interaction between these explanatory variables?

																																									
<pre>. gen gestox = gestage*t tox</pre>																																									
<pre>. regress headcirc gestage tox gestox</pre>																																									
<table border="1"> <thead> <tr> <th>Source</th><th>SS</th><th>df</th><th>MS</th><th colspan="3"></th></tr> </thead> <tbody> <tr> <td>Model</td><td>414.52584</td><td>3</td><td>138.17528</td><td colspan="3">Number of obs = 100</td></tr> <tr> <td>Residual</td><td>220.22416</td><td>96</td><td>2.29400167</td><td colspan="3">F( 3, 96) = 60.23</td></tr> <tr> <td>Total</td><td>634.75</td><td>99</td><td>6.41161616</td><td colspan="3">Prob &gt; F = 0.0000</td></tr> </tbody> </table>						Source	SS	df	MS				Model	414.52584	3	138.17528	Number of obs = 100			Residual	220.22416	96	2.29400167	F( 3, 96) = 60.23			Total	634.75	99	6.41161616	Prob > F = 0.0000										
Source	SS	df	MS																																						
Model	414.52584	3	138.17528	Number of obs = 100																																					
Residual	220.22416	96	2.29400167	F( 3, 96) = 60.23																																					
Total	634.75	99	6.41161616	Prob > F = 0.0000																																					
<table border="1"> <thead> <tr> <th>headcirc</th><th>Coef.</th><th>Std. Err.</th><th>t</th><th>P&gt; t </th><th>[95% Conf. Interval]</th><th></th></tr> </thead> <tbody> <tr> <td>gestage</td><td>.8646116</td><td>.073898</td><td>11.70</td><td>0.000</td><td>.7179251</td><td>1.011298</td></tr> <tr> <td>tox</td><td>-2.815032</td><td>4.985147</td><td>-0.56</td><td>0.574</td><td>-12.71047</td><td>7.080407</td></tr> <tr> <td>gestox</td><td>.0461658</td><td>.1635213</td><td>0.28</td><td>0.778</td><td>-.2784214</td><td>.370753</td></tr> <tr> <td>_cons</td><td>1.762912</td><td>2.102255</td><td>0.84</td><td>0.404</td><td>-2.410031</td><td>5.935855</td></tr> </tbody> </table>						headcirc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		gestage	.8646116	.073898	11.70	0.000	.7179251	1.011298	tox	-2.815032	4.985147	-0.56	0.574	-12.71047	7.080407	gestox	.0461658	.1635213	0.28	0.778	-.2784214	.370753	_cons	1.762912	2.102255	0.84	0.404	-2.410031	5.935855	
headcirc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]																																				
gestage	.8646116	.073898	11.70	0.000	.7179251	1.011298																																			
tox	-2.815032	4.985147	-0.56	0.574	-12.71047	7.080407																																			
gestox	.0461658	.1635213	0.28	0.778	-.2784214	.370753																																			
_cons	1.762912	2.102255	0.84	0.404	-2.410031	5.935855																																			

We can generate a new variable called gestox which is equal to gestational age times toxemia. This is called an interaction term. It should tell us whether the babies born of mothers who suffered from toxemia act differently from those whose mothers did not have toxemia. In the latter case gestox=0, whereas it is not zero in the former. That means that the two regression lines will no longer be parallel.

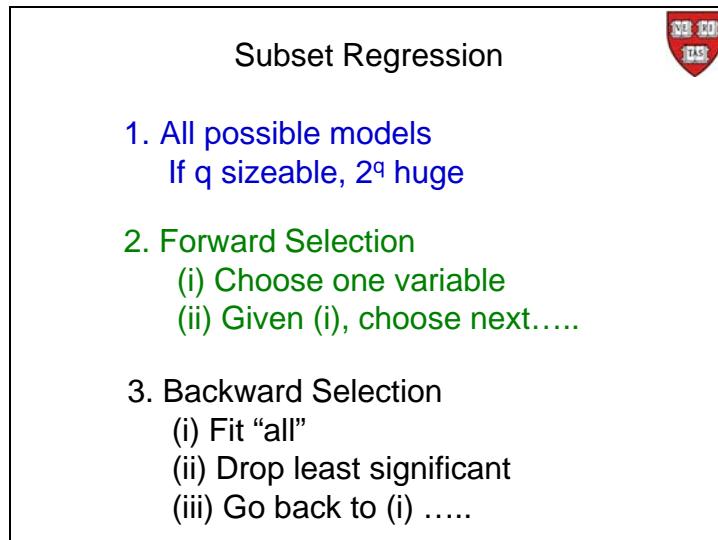
Once we let this variable enter into the model we see that now toxemia is no longer important. It is not important either by itself or as part of the interaction term. The only explanatory variable left of any import is gestational age.

The R-squared has gone down to 0.64. This should not be happening, we should be improving the explanatory capability of a model by adding explanatory variables, not decreasing the capability.



If we plot the regression lines now, they are almost, but not quite, parallel. We have run into a problem of collinearity we explore below. This is what can happen when we explore which variables to include in our regression model when we have the choice of many variables, as we do in this small data set, even. Not only do the variables represent different explanatory aspects of the outcome variable in question, they may also interact with each other in a way that makes the model fitting much more complex.

### Subset Regression



Which explanatory variables and in what form to include into our regression model is called the subset regression problem. All too often in a study the problem being studied is serious and

patients, and their families, are being asked to give of themselves to be studied mostly to help future patients so we do not wish to miss something important. Add on top of that the scientists' favorite theories and as a result there are usually a large number of candidates to be entertained as possible explanatory variables. On the other hand, if this study is to be of use to future patients it has to be generalized and we do not want to have results that are peculiar only to this particular set of patients. That usually argues for a parsimonious model with a few, choice, and important explanatory variables.

There are three main strategies for attacking the subset regression problem. The general problem has not yet been solved, so we proceed by relying very much on our own personal experience with model fitting.

One approach is to investigate all possible models. If there are  $q$  possible explanatory variables that means there are  $2^q$  possible models, ignoring interaction terms, that can be fit. If the number of observations is large, then each one of these models may take some time to fit, and if on top of that,  $q$  is sizable then you could spend the rest of your life, like Tycho Brahe, fitting models to data. This approach also creates a huge Bonferroni problem, if you will, of related tests.

We could rank the models by looking at the respective R-squareds, and that is one way to proceed. Not advisable, but a way.

Another approach is what is called the forward selection approach. Here what you do is find the best single variable somehow. The one with the highest R-squared, or the highest t-value, and then force that into the model. Then with that first one in the model look for the next explanatory variable to include, and then the third, and so on. The problem with this approach is that once you have the first two, there is no guarantee that that is the best two-explanatory variable model.

Finally there is the backward selection approach. In this one you fit all the explanatory variables into the model and then work your way down by eliminating the least important ones, one by one. Once again, there is no guarantee that when you end up with a model with  $m$  explanatory variables, it is the best  $m$  explanatory variable model.

So here are three approaches that you can take, and there are some computer programs that have been automated to do all this. I would strongly recommend that you do not use any of these. They are problematic. For one, it is quite probable that you do not get the same answer from all three approaches, and it is often very difficult to understand why.



## Collinearity

i.e. two or more of the **explanatory** variables are correlated to the extent that they convey essentially the same information.

One of the problems is that you run into is what is called collinearity. The most extreme case of collinearity occurs if you had somebody's weight measured in kilograms in one variable, and in another variable that person's weight measured in pounds. If you plot those two variables against each other the relationship would be a straight line—thus collinear. When two variables are collinear then you really only have one variable, not two, the second variable does not add any more information that is not contained in the first variable. The two variables are perfectly correlated with each other.

It can get even more complicated than that. You might have three variables who are collinear in the sense that two of them can predict the third one exactly. And so on, with multiple variables.

And it gets even more complex when the variables are not perfectly correlated with each other, but almost perfectly correlated, or highly correlated with each other. You might still then run into interpretation and fitting problems. And that is the collinear problem.

Results:



	No interaction term	Interaction term
Coeff	-1.412	-2.815
Std. Err.	0.406	4.985
T-stat	-3.477	-.565
P-value	0.001	0.574
R <sup>2</sup>	0.653	0.653
Adj. R <sup>2</sup>	0.646	0.642

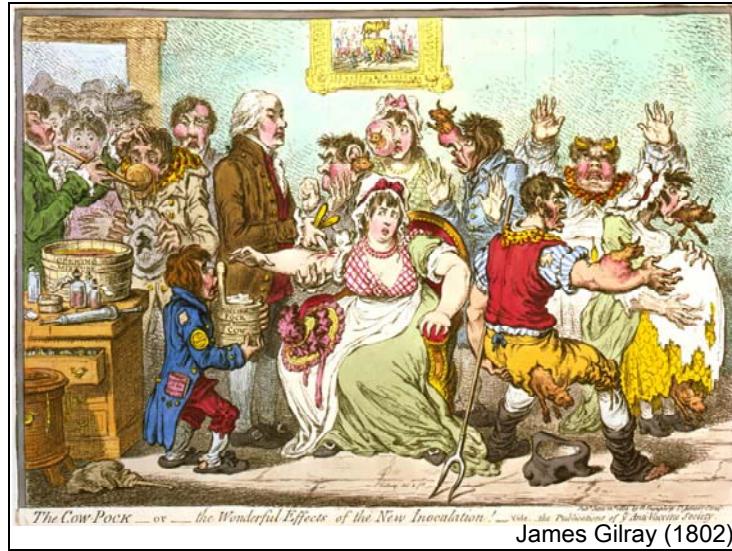
That is the problem we have run into when looking at toxemia. If you look at the two models, one with no interaction term and then the one with the interaction term, and you looked at the toxemia coefficient. In the first model it is -1.442. In the second model it almost doubled in value to -2.815. Plus, when you look at the standard error associated with this term, it increased tenfold. In the meantime the p-value went from 0.001 to 0.574 and there was no change in the R<sup>2</sup>. Such a radical change in these statistics, brought about by the introduction of a single explanatory variable (accompanied by no change in the R<sup>2</sup>) is an indication of instability; exactly the sort of instability brought about by collinearity.

In this case, these unstable results make sense, because we know that toxemia tends to happen later in the pregnancy, and thus toxemia is related with higher gestational age. So the introduction of the interaction term between toxemia and gestational age is attempting to introduce three very interrelated terms in to the regression. Take care when this happens.

Marcello Pagano

# [JOTTER 11 LOGISTIC REGRESSION]

Odds ratio, logistic function, dichotomous response, weighted least squares



This week we stay on our theme of regression and we are going to talk about logistic regression—we extend the idea of regression to the situation where our outcome variable is dichotomous.

To introduce the topic we look at a famous cartoon by James Gillray in 1802<sup>1</sup>. It deals with the reaction to what at the time was the rather revolutionary idea of vaccination. This was the first time a disease from another species was injected into man for beneficial purposes. To make the point of the anti-vaccine crowd, you see various ways of caricaturing the practice, by having cows protrude from all sorts of places.

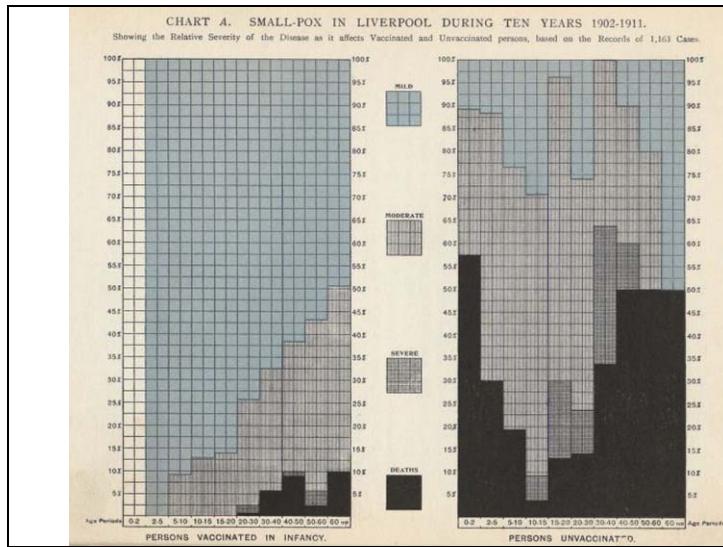
TABLE SHOWING NUMBER OF SMALL-POX CASES (948 VACCINATED AND 220 UNVACCINATED)  
AT EACH AGE-PERIOD, WITH RELATIVE DEGREES OF SEVERITY.

	Under 1 year	1-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55 and up	Total Vacc.	Total Un-vacc.							
A. Modified discrete and discrete	No cases	3	7	3	31	6	54	9	89	1	249	11	163	No cases	63	1	20	2	10	1	686	37
B. Profuse discrete and semi-confluent	No cases	9	No cases	15	3	15	8	19	14	20	77	21	72	5	29	3	13	3	8	No cases	224	110
C. Confluent and death	No cases	17	No cases	8	No cases	5	No cases	3	No cases	9	7	10	13	9	9	6	2	5	2	1	33	73
Total	No cases	29	7	26	34	26	62	31	101	30	333	42	248	14	101	10	35	10	20	2	943	320
Deaths alone	0	17	0	8	0	5	0	1	0	4	3	6	13	8	9	5	1	5	2	1	28	60

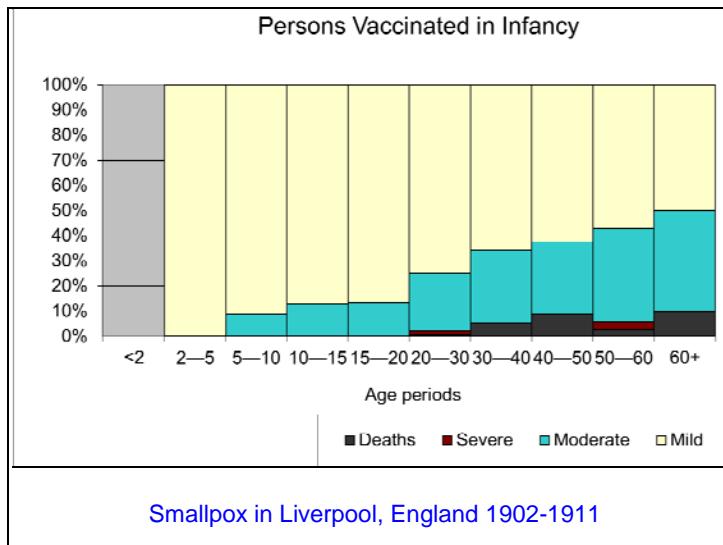
William Hanna, Studies in Smallpox and Vaccination,  
William Wood and Company, New York 1913.

<sup>1</sup> *The Cow-Pock or the Wonderful Effects of the New Inoculation* by James Gillray (1757-1815) was published in England on June 12, 1802 by the Anti-Vaccine Society.

Smallpox was a quite lethal problem at that time, 200 years ago, as it was even 100 years ago, when this book was written<sup>2</sup>. It gives a thorough statistical description of a smallpox outbreak in the city of Liverpool that was part of a general outbreak they had in England in 1902-1903.



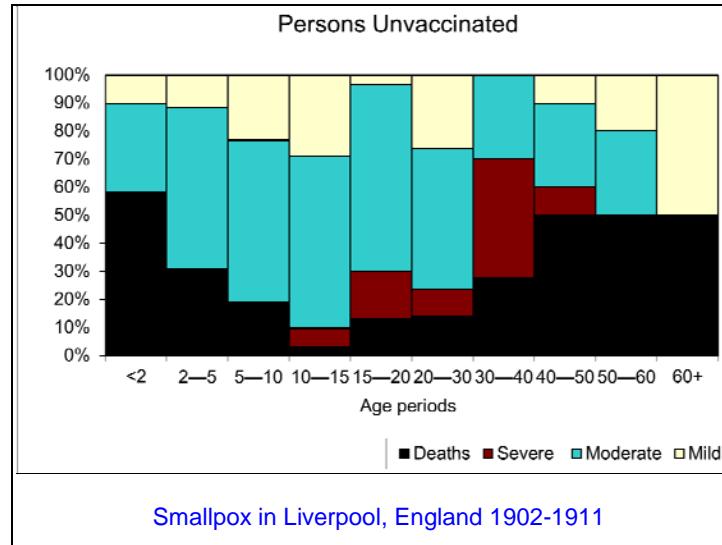
Above is a graphic from the book. It shows, as a function of age, the health impact of the epidemic. The panel on the left is for people who had been vaccinated at birth, and the panel on the right is for those not so vaccinated. One hundred years after the cartoon, vaccination had taken hold, although not everyone was vaccinated. The black bars show the number of people who died, by age. So you can see that the number of deaths is very much less for those who were vaccinated.



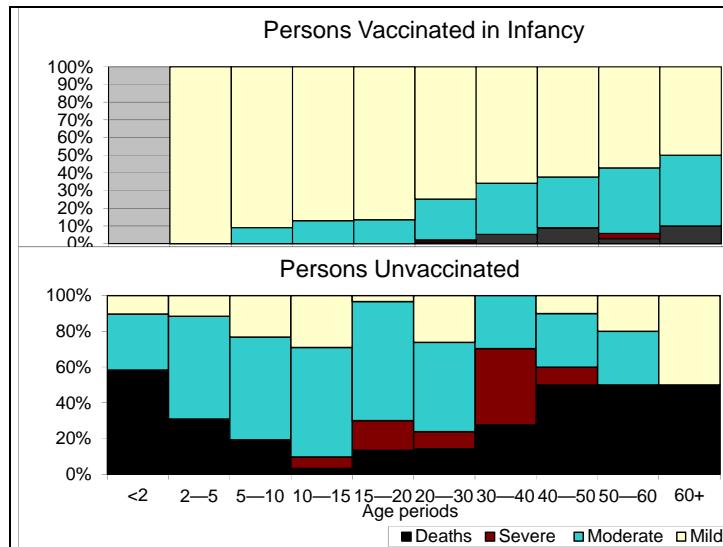
What I did was I reproduced those two graphs using modern techniques. This graph is for those who were vaccinated—the left panel above. Just focusing on the deaths we see a slight trend as people age.

<sup>2</sup> William Hanna, Studies in Smallpox and Vaccination, William Wood and Company, New York 1913.

Also, if we look at the frequencies of “Moderate” effects, those too increase with age. All in all, what we see is consonant with the beliefs that the effects of vaccinations wear off with time—most of these individuals were vaccinated shortly after birth.



For those unvaccinated, we see a large number of deaths—much more black on the graph than for the previous one.

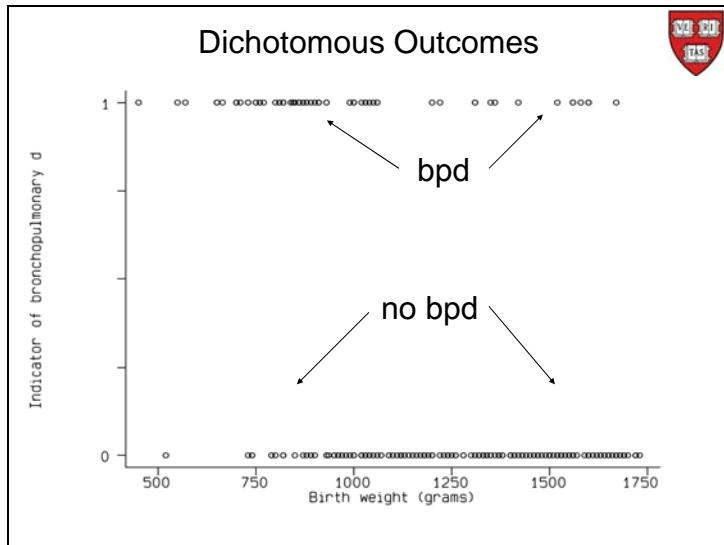


Indeed, if we place these two together, we can see that there is a big difference between the two graphs, arguing that vaccination has a great beneficial effect.

Focusing on the Vaccinated, if we look at the proportion who die in each age group we see an age effect. Certainly with the young ones, nobody below the age of 20 dies. Also, looking at the percentage who suffer Moderate, or worse, effects those too build up with age.

Age	Vaccinated			Unvaccinated		
	Numb.	Died	Mortal	Numb.	Died	Mortal
<2	0	0	0	29	17	58.0%
2–5	7	0	0	26	8	30.6%
5–10	34	0	0	26	5	19.0%
10–15	62	0	0	31	1	3.2%
15–20	103	0	0	30	4	13.3%
20–30	333	3	0.9%	42	0	14.2%
30–40	248	13	5.2%	14	8	33.3%
40–50	101	9	8.9%	10	5	50.0%
50–60	35	1	2.8%	10	5	50.0%
≥60	20	2	10.0%	2	1	50.0%
Total	943	28	2.9%	220	60	27.2%

What we are seeing is that the probability, or the proportion, who are dying is impacted by age. This reminds us of regression—we can regress the proportion who die on age. And here are the numbers to support this suggestion. If we look at the vaccinated we can see that the mortality goes up with age, almost monotonically. It is not quite smooth because of the 50 to 60 age group, but it does seem like age might affect the mortality rate, in general. Of course, we now know that the effect of vaccination does wear off with time—vaccination occurred in infancy.



Here is another example. We are looking at babies who suffer from bpd (bronchopulmonary dysplasia). This condition gets measured at about age 29 days, roughly a month after birth, amongst premature infants, and it indicates progressive lung inflammation. This study<sup>3</sup> was carried out on infants born weighing less than 1750 grams and each child was categorized as 0 (no bpd), or 1 (bpd) and the results are plotted above, as a function of birth age.

We can see that at the left of the graph (low birth weight) we see a higher intensity of infants with bpd, whereas on the right (relatively higher birth weight) we see a higher intensity of infants without bpd. This picture would be indicative of what we would expect to see if the probability of bpd goes down as birth weight increases. There are kids with and without bpd at either end of the scale, so we are not observing a phenomenon that deterministically decides the classification, but rather something more nuanced.

For example, if we create a window of width 250 grams, say, and slide this window from left to right, then the *proportion* of babies in the window who have bpd will go down as we move across.

---

<sup>3</sup> Van Marter, L. J., Leviton, A., Kuban, K. C. K., Pagano, M., and Allred, E. N., Maternal Glucocorticoid Therapy and Reduced Risk of Bronchopulmonary Dysplasia," *Pediatrics*, **86**, September 1990, 331-336.

Bpd by birth weight					
Birthweight (gms)	Bpd	N	Prop.	Odds	
0-950	49	68	0.721	2.58	
951-1350	18	80	0.225	0.29	
1351-1750	9	75	0.120	0.14	
Total	76	223	0.341	0.52	

A simple manifestation of this idea would be to take a snapshot of the window in three locations. Here I have arbitrarily chosen these three intervals, for no other reason than to make this point: the proportion with bpd goes down as we go from top to bottom. That means, as the birth weight increases. This behavior, of course, is evident in the odds of bpd, too—they go down as birth weight increases. The challenge is to be somewhat more formal in quantifying this relationship between the probability, or odds, of bpd and birth weight.

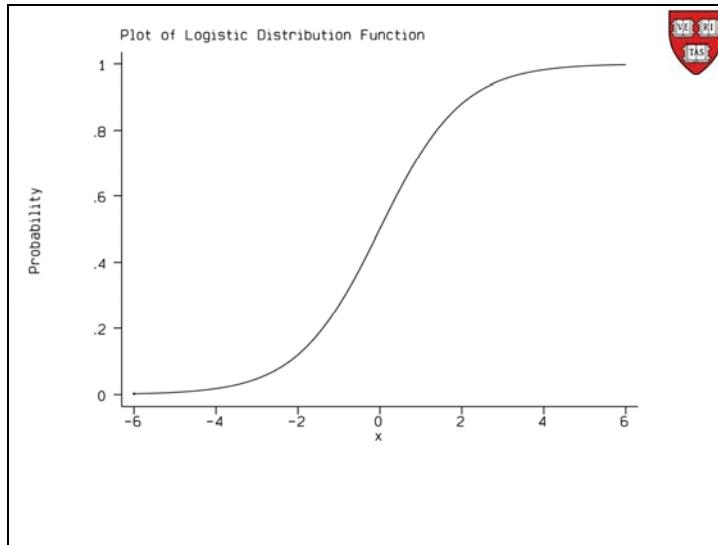
Logistic regression		
In linear regression, continuous Y:		
$\mu_{Y x} = \alpha + \beta x$		
In case of dichotomous Y:		
$\mu_{Y x} = \Pr(Y = 1   x) = p_{Y x}$		
1. $p_{Y x} = \alpha + \beta x$		
2. $p_{Y x} = e^{\alpha + \beta x}$		
3. $p_{Y x} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$		

Linear regression allowed us to quantify the relationship between a response variable and explanatory variables. For example, with simple linear regression, the regression line was the straight line relationship between the mean of the Ys and the explanatory, x.

We can mimic what we did there by first calculating the mean of the Ys. Now Y is a Bernoulli variable that takes on the values 0 and 1, so we know that its mean is the probability that it takes on the value 1. So suggestion 1, would have us fitting a straight line to that probability. That is clearly unsatisfactory because for one a straight line can be negative. So suggestion 2 would have us taking the exponential of the

straight line—that would solve the problem about becoming negative. Unfortunately, that still leaves us with the problem that this function can be greater than one and thus would not be a good model for a probability.

Suggestion 3 satisfies all the constraints on a probability. Not only that, it is monotonic in  $x$ —that means that the probability goes up with  $x$  (as  $x$  gets larger, the probability gets larger) or the probability goes down with  $x$  (as  $x$  gets larger the probability gets smaller). One can prove this mathematically, or you can convince yourself by drawing this function for yourself. You will then also see that what determines whether it goes up with  $x$  getting larger, or down with  $x$  getting larger, is the sign of the  $\beta$ —positive  $\beta$  means probability goes up with  $x$  getting larger. Note, of course, that if  $\beta=0$ , then the probability is not impacted by the value of  $x$ . In that case we would conclude that  $Y$  is independent of  $X$ .



If we plot this as a function of  $x$ , it looks something like this— $\alpha=0$  and  $\beta=1$  in this case. This is what the logistic function looks like for a positive  $\beta$ . If  $\beta<0$ , then we get the mirror image of this function that starts at one in the top-left corner and monotonically decreases to zero in the bottom right-hand corner—in other words, one minus this function.

So if we were to use this function to model the probability of bpd as a function of birth weight, we would expect a negative  $\beta$ . Of course, we could have switched the roles of zero and one in the definition of bpd, and then we would expect a positive  $\beta$ , as drawn above, because then the curve would represent the probability of *not* having bpd.

**Log odds and logistic**

If  $p_{Y|x} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

then  $\ln\left(\frac{p_{Y|x}}{1-p_{Y|x}}\right) = \alpha + \beta x$

e.g. Log odds =

$$\ln\left(\frac{p_{bpd|bwt}}{1-p_{bpd|bwt}}\right) = 4.0343 - 0.0042 \times bwt$$

where  $p_{bpd|bwt}$  is the probability of bpd for a given birthweight (bwt).

If we solve the logistic function for  $\alpha + \beta x$  we get the above equation for the log odds. So, simple logistic regression fits a straight line to the log of the odds.

So for example, in our bpd example, the straight line that we would fit to the log odds has intercept 4.03 and slope -0.0042. These numbers were obtained by using least squares to fit the line to the data. To be precise, it is not the least squares we are accustomed to using but, since we do not have homoscedasticity, we need to use a technique called weighted least squares. Another approach, that yields the same answer, is what is called, maximum likelihood. The technicalities associated with both are beyond the scope of this course, but this is what Stata calculates for us.

**Stata:**

In model form....beta

```
. logit bpd birthwt
Iteration 0:  log likelihood = -143.06998
Iteration 1:  log likelihood = -113.06062
Iteration 2:  log likelihood = -111.86443
Iteration 3:  log likelihood = -111.86031
Iteration 4:  log likelihood = -111.86031

Logistic regression                                         Number of obs     =      223
                                                               LR chi2(1)      =     62.42
                                                               Prob > chi2    =     0.0000
                                                               Pseudo R2       =     0.2181

Log likelihood = -111.86031


```

bpd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
birthwt	-.0042291	.0006409	-6.60	0.000	-.0054852 -.0029731
_cons	4.034291	.6957851	5.80	0.000	2.670578 5.398005

There are two Stata commands we can use. One is *logit* and it works very much like *regress*, and the output we get looks quite similar to what we get with *regress*, and it is interpreted much the same way.

So, in this case, we get that birth weight is significant. We see that the 95% confidence interval does not include zero, and since the coefficient is negative we conclude that, as birth weight goes up, the odds, or the probability, of getting BPD goes down. The log of the odds of getting bpd goes down by 0.0042 for every gram increase in birth weight.

**Stata:**

In odds form....odds ratio

```
. logistic bpd birthwt
```

Logistic regression

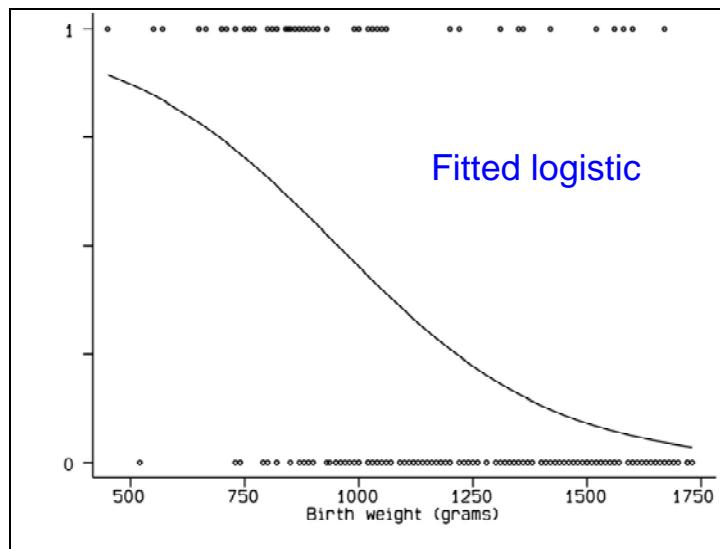
Number of obs	=	223
LR chi2(1)	=	62.42
Prob > chi2	=	0.0000
Pseudo R2	=	0.2181

Log likelihood = -111.86031

bpd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
birthwt	.9957798	.0006382	-6.60	0.000	.9945298 .9970313
_cons	56.50287	39.31386	5.80	0.000	14.44831 220.9652

We can look at the odds directly by using the other command in Stata to fit this model, and that is the command *logistic*. It reports the odds ratio, rather than reporting the betas. Remember, we are fitting the straight line to the log of the odds, so if we report the odds that is the same as reporting the exponentials of the last screen. That means, for example, that for every gram increase in birth weight, the odds of getting bpd gets multiplied (because we are taking the exponentials) by 0.9957798. (The  $\log(0.9957798) = -0.0042291$ .)

The interpretation, of course, does not change, it is just a different way to report the same results.



Here is the fitted logistic. So if we had an idealized version of the moving window we described earlier, then as it moves from left to right, this curve shows the proportion in the window with bpd.

Stata:		In model form....beta																		
<pre>. logit bpd birthwt</pre>																				
<pre>Iteration 0:  log likelihood = -143.06998 Iteration 1:  log likelihood = -113.06062 Iteration 2:  log likelihood = -111.86443 Iteration 3:  log likelihood = -111.86031 Iteration 4:  log likelihood = -111.86031</pre>																				
<pre>Logistic regression   Number of obs =      223  LR chi2(1) =     62.42  Prob &gt; chi2 =    0.0000  Pseudo R2 =     0.2181</pre>																				
<table border="1"> <thead> <tr> <th>bpd</th><th>Coeff.</th><th>Std. Err.</th><th>z</th><th>P&gt; z </th><th>[95% Conf. Interval]</th></tr> </thead> <tbody> <tr> <td>birthwt</td><td>-.0042291</td><td>.0006409</td><td>-6.60</td><td>0.000</td><td>-.0054852 -.0029731</td></tr> <tr> <td>_cons</td><td>4.034291</td><td>.6957851</td><td>5.80</td><td>0.000</td><td>2.670578 5.398005</td></tr> </tbody> </table>			bpd	Coeff.	Std. Err.	z	P> z	[95% Conf. Interval]	birthwt	-.0042291	.0006409	-6.60	0.000	-.0054852 -.0029731	_cons	4.034291	.6957851	5.80	0.000	2.670578 5.398005
bpd	Coeff.	Std. Err.	z	P> z	[95% Conf. Interval]															
birthwt	-.0042291	.0006409	-6.60	0.000	-.0054852 -.0029731															
_cons	4.034291	.6957851	5.80	0.000	2.670578 5.398005															

Returning to the beta coefficients, we see that the coefficient of birth weight is small. The units of this are 1/grams, so if we were to measure the babies' birth weights in Kilograms, this coefficient would become -4.2291. One should be cognizant of the units of the measurement before opining about its size.

Observed versus fitted			
Birthweight (gms)	Obsrvd bpd	Fitted bpd	N
0-950	0.721	0.646	68
951-1350	0.225	0.324	80
1351-1750	0.120	0.082	75
Total	0.341	0.341	223

One way to judge how well the model has done is to return to trichotomy we created before. In green we see the observed proportions. For each child we also now have a fitted probability of getting bpd—for a child weighing x grams, solve for p in the equation: the  $\log(p/(1-p)) = 4.034291 - 0.0042291 \times$ . The red numbers above are the averages of these p for the babies in the respective groups.

So the “fitted” values compare somewhat with the actual, observed values. There is some room for improvement.

## Back to Liverpool 1902-11, vaccinated



```
. logit dead total age, or
Logistic regression for grouped data
Number of obs      =      943
LR chi2(1)        =     19.58
Prob > chi2       =    0.0000
Pseudo R2         =    0.0777

_outcome | Odds Ratio   Std. Err.      z    P>|z|   [95% Conf. Interval]
          |
age      | 1.064589   .0147107    4.53   0.000   1.036144   1.093816
_cons   | .00361   .0020815   -9.75   0.000   .0011661   .0111764
```

Returning to the Liverpool outbreak, we can fit a logistic model to those data. Unfortunately, we did not have the raw data, only what we presented above, so in the reconstruction we had to cheat a little bit since we only know the ages within a category. What we did is take the midpoint of each category and used that to represent the age of all the individuals in that category. Now that we have so called blocked data—we have data only in blocks, not individual data—we can use the Stata command called *logit*, it is the command to use for data presented in this fashion.

So here are the results for such a logistic analysis on the people who had been vaccinated prior to the outbreak in Liverpool. This says that the odds are multiplied by 1.06 every time we go up one year in age. We see that this is significant, so that age is an important explanatory variable.

## Logistic growth model



Aside, population models from human ecology:

$r$  is net population growth rate per individual

$N_t$  is the population size at time  $t$

$$N_{t+1} - N_t = r N_t$$

(Malthus, doubles every 25 years) — exponential growth

If  $K$  is number sustainable by environ

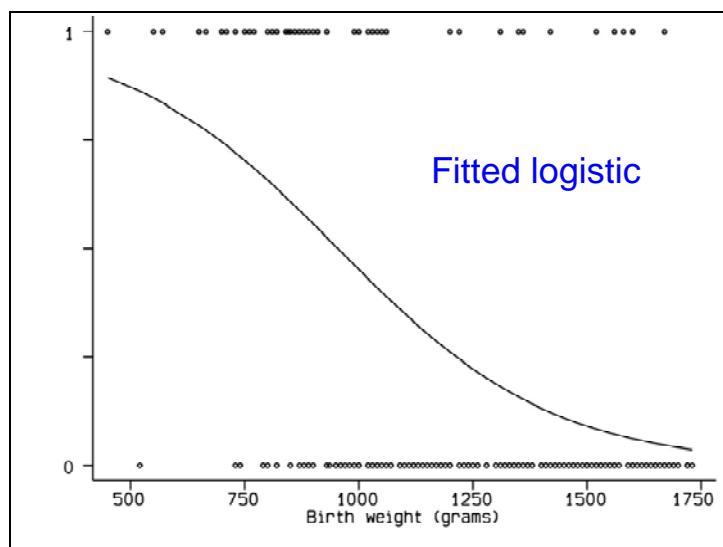
$$N_{t+1} - N_t = r N_t (1 - N_t / K) \quad \text{— logistic growth}$$

As an aside, I just wanted to point out that the logistic growth model also does appear in ecology. If you look at population size at time  $t$ , call that  $N_t$ . Suppose each individual in the population gives birth to  $r$  individuals in the next generation. Then if you look at the population growth, we see that expressed in the equation above.

This is what Malthus was concerned about almost two centuries ago, because the  $r$  he envisioned led to a doubling of the population approximately every 25 years. He was quite prescient. And that is worrisome. The value of  $r$  is all important: if  $r$  is less than 1, then we will eventually die out;  $r$  is equal to 1, then we just remain the same size; and, if  $r$  is bigger than 1, then that leads to exponential growth.

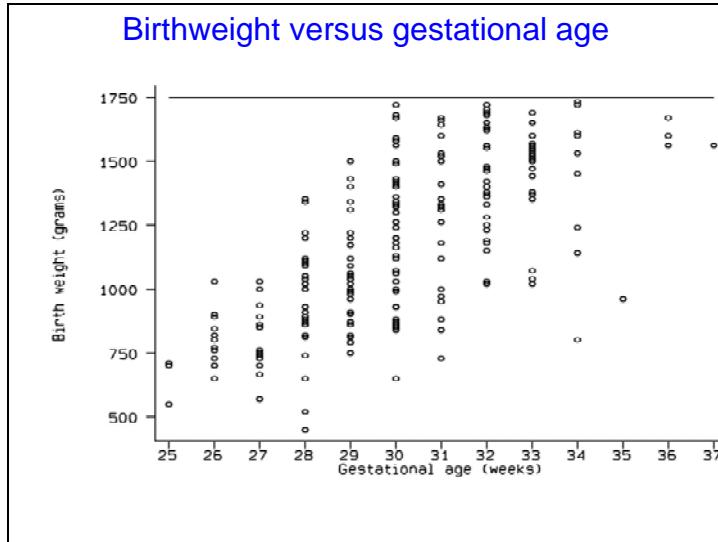
Now what happens to this model if the surroundings cannot maintain the exponential growth? For example, suppose you can only sustain  $K$  individuals in your environment. Incorporating that constraint into the growth model leads to logistic growth.

### Multiple Logistic



So we fitted the logistic to a single variable—birth weight—and found that the fit was good, but it left some room for improvement. Logistic regression works just like linear regression. We can do simple logistic regression with a single variable, and we can increase the number of explanatory variables to get multiple logistic regression.

We run into exactly the same problems we had with multiple linear regression: which explanatory variable will be best? Is there room for other explanatory variables? What about interaction? What about multicollinearity? And we handle them in a manner very similar to the way we did in the case of linear regression.



Consider two candidates for inclusion in the regression as explanatory variables; birth weight and gestational age. Here is a scatter plot of these two. We definitely see a relationship between the two even with the study cutoff that the babies had to weigh less than 1750 grams at birth. This relationship is, of course, of no surprise.

So instead of birth weight, we could consider incorporating gestational age as an explanatory variable.

**Bpd by gestational age:**

Gestational age	bpd	n	Proport.
<29 wks	40	58	0.690
29—30 wks	26	73	0.356
>30 wks	10	92	0.109



Does this yield information over and above birth weight?

Once again, for exploratory purposes, let us break down the gestational age into three categories. As expected, we see a “dose response”, or monotonic relationship with proportion with bpd. A high proportion (69%) in the smallest gestational age group, have bpd. A smaller proportion (35.6%) in the middle age group, have bpd. The smallest proportion (10.9%) with bpd is in the group with the biggest gestational ages. We thus see evidence of a decrease in the chance of having bpd with gestational age.

Birthweight by gestational age			
	Percent with bpd:		
Birth wt (gms)	Gestational age (weeks)		
	<29	29-30	>30
0-950	0.805 (n=41)	0.714 (n=21)	0.167 (n=6)
951-1350	0.412 (n=17)	0.194 (n=36)	0.148 (n=27)
1351-1750	- (n=0)	0.250 (n=16)	0.085 (n=59)

Before entering the gestational age as an explanatory variable, let us check for collinearity between gestational age and birth weight. Consider this 3-by-3 table. Once again, this is just for expository purposes, but if we had collinearity we would expect to only have single entries in a row—for example, just entries in the diagonal cells—certainly not entries in all cells but one. Indeed, in each row we see a monotonic decrease in the proportion with bpd, thus arguing that even accounting for birth weight, we see an impact of gestational age.

Having this exploratory investigation under our belts, let us fit the model with both explanatory variables.

. logit bpd birthwt gestage						
Iteration 0:	log likelihood = -143.06998					
Iteration 1:	log likelihood = -106.45551					
Iteration 2:	log likelihood = -104.6221					
Iteration 3:	log likelihood = -104.6146					
Iteration 4:	log likelihood = -104.6146					
Logistic regression		Number of obs = 223				
		LR chi2(2) = 76.91				
		Prob > chi2 = 0.0000				
Log likelihood = -104.6146		Pseudo R2 = 0.2688				
bpd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
birthwt	-.0024097	.0007925	-3.04	0.002	-.003963	-.0008564
gestage	-.3982616	.1129995	-3.52	0.000	-.6197366	-.1767865
_cons	13.02725	2.932117	4.72	0.000	8.000408	19.57409

Here is the output from Stata. As expected, we see that both coefficients are negative, so the chances of bpd go down as either birth weight or gestational age goes up. Both p values are less than 0.05, so both are considered significant. Of course, we could have arrived at the same conclusion by looking at the two confidence intervals.

							
<pre>. logistic bpd birthwt gestage</pre>							
Logistic regression							
Number of obs = 223							
LR chi2(2) = 76.91							
Prob > chi2 = 0.0000							
Log likelihood = -104.6146 Pseudo R2 = 0.2688							
<hr/>							
bpd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]		
birthwt	.9975932	.0007906	-3.04	0.002	.9960448	.9991439	
gestage	.6714864	.0758777	-3.52	0.000	.5380862	.8379587	
_cons	1011810	2966744	4.72	0.000	3230.552	3.17e+08	

We could have come to the same conclusion with the *logistic* command.

Red is observed bpd proportion, green is fitted				
Birth wt (gms)	Gestational age (weeks)			
	<29	29-30	>30	
0-950	0.805	0.714	0.167	
	(n=41)	(n=21)	(n=6)	
951-1350	0.766	0.514	0.320	
	(n=17)	(n=36)	(n=27)	
	0.412	0.194	0.148	
1351-1750	0.527	0.354	0.152	
	(n=0)	(n=16)	(n=59)	
	-	0.250	0.085	

To investigate the fit, we can repeat the exercise we carried out above with fitted values. Averaging the fitted probabilities of each baby in each cell we get the green numbers above. This shows an improvement over the model with only birth weight as an explanatory variable.

## Indicator Variables



toxemia = 1 if mother had toxemia  
= 0 otherwise

```
. logit bpd toxemia

Iteration 0:  log likelihood = -143.06998
Iteration 1:  log likelihood = -141.64339
Iteration 2:  log likelihood = -141.63956
Iteration 3:  log likelihood = -141.63956

Logistic regression                                         Number of obs =      223
                                                               LR chi2(1) =       2.86
                                                               Prob > chi2 =    0.0908
                                                               Pseudo R2 =     0.0100
Log likelihood = -141.63956
```

bpd	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
toxemia	-.7719484	.4821775	-1.60	0.109	-.1716999 .1731021
_cons	-.5717863	.1494999	-3.82	0.000	-.8648008 -.2787718

$$\ln\left(\frac{\hat{p}_{Y|\text{toxemia}}}{1-\hat{p}_{Y|\text{toxemia}}}\right) = -0.572 - .772 \text{ toxemia}$$

Let us look at indicator variables with logistic regression because it should remind us of something we have already seen. Recall that we looked at toxemia as an indicator variable that indicated whether the mother suffered from toxemia during the pregnancy. Here is the result of fitting a logistic regression with just toxemia as the explanatory variable.

We see that the coefficient is -0.77 and is judged not to be significant. But here is the best fitting model if we include toxemia.



$$\ln\left(\frac{\hat{p}_{Y|\text{toxemia}}}{1-\hat{p}_{Y|\text{toxemia}}}\right) = -0.572 - .772 \text{ toxemia}$$

$$\ln\left(\frac{\hat{p}_{Y|\text{has toxemia}}}{1-\hat{p}_{Y|\text{has toxemia}}}\right) = -0.572 - .772$$

$$\ln\left(\frac{\hat{p}_{Y|\text{no toxemia}}}{1-\hat{p}_{Y|\text{no toxemia}}}\right) = -0.572$$

$$\ln(\text{OR}) = -.772$$

$$\Rightarrow \text{OR} = e^{-.772} = 0.46$$

So the log odds of bpd are as above. The difference between the log odds when the mother has toxemia and when the mother does not have toxemia, is the log of the odds ratio, and that is -0.772. So the odds ratio is 0.46.



. cc toxemia bpd			
	Exposed	Unexposed	Proportion Exposed
Cases	6	23	0.2069
Controls	70	124	0.3608
Total	76	147	0.3408
	Point estimate	[95% Conf. Interval]	
Odds ratio	.4621118	.1472474	1.243955 (exact)
Prev. frac. ex.	.5378882	-.2439547	.8527526 (exact)
Prev. frac. pop	.1940834		
chi2(1) = 2.66 Pr>chi2 = 0.1028			

Alternatively, we can consider the 2x2 table of bpd versus toxemia. When we analyze that table, as above, we see that the odds ratio is the same 0.462 as we got from the logistic regression analysis.

So an indicator variable and logistic regression is just another way of looking at the tables we had analyzed before. What the logistic regression approach allows us to do now is to analyze tables considering a collection of dimensions—any number of explanatory variables—together with interaction terms, etcetera.



### Final model:

	coeff	Std.err.	p-value
bwt	-1.341	0.274	.0000
gestage	-0.932	0.310	.0026
toxemia	-1.359	0.624	.0293
sex	0.757	0.384	.0489
Mat. steroids	-0.921	0.425	.0302
constant	1.827	0.420	.0000

Maternal and neo-natal factors affecting bpd  
based on 223 premature, small babies

Indeed, the final model for these data is above, and we see a collection of dichotomous (toxemia, sex, maternal steroids (yes/no)) and continuous explanatory variables (birth weight and gestational age). This is a much richer class of models to fit to a dichotomous response variable than is afforded by analysis as

2x2 tables. Note that the final model includes toxemia—possibly because toxemia happens later in the pregnancy and is related to gestational age.

Marcello Pagano

# [JOTTER 12 SURVIVAL ANALYSIS]

Survival curves, product limit method, censoring, log rank test

Survival Analysis 

Regression where the end-point, or dependent variable, is positive.

e.g. **survival**

- time to an event
- e.g. **response**
- failure**
- pregnancy**
- infection**
- strike end**

latency period

We continue our study of regression. This week we look at what is generically called survival analysis—essentially the study of positive outcome variables—and how to incorporate explanatory variables into the argument. This is an important topic that can be applied in a number of situations, not the least is the study of clinical trials, but we only have enough time to introduce you to the topic.

Quick review: In our study of regression, we first looked at simple linear regression. That meant that the means of the outcome variable, for fixed values of the explanatory variable, all lay on a straight line.

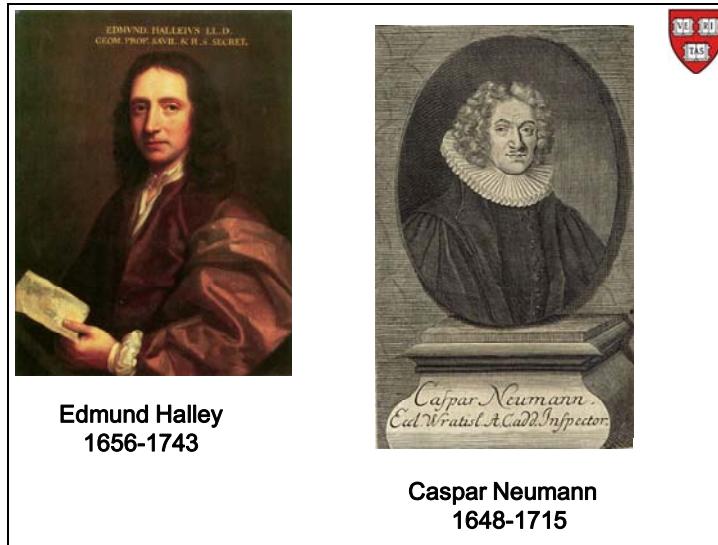
Sometimes, if the plotted line of the means is not straight, we can *transform*, or change the units of the measurements to obtain a straight line. We then extended these methods to multiple-regression by incorporating more than one explanatory variable.

When faced with a dichotomous outcome variable, its mean is a probability, and thus is constrained to take values between zero and one. As a result, a straight line cannot provide a good model for dichotomous regression, and so we turned to logistic regression (other transformations are available, of course).

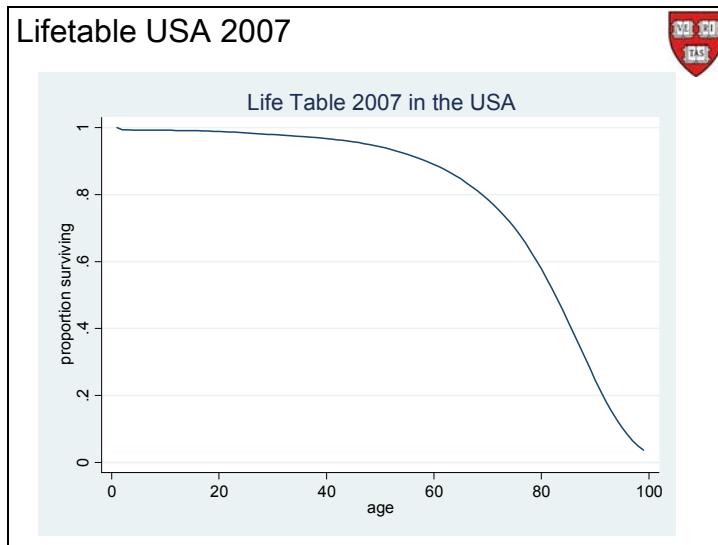
Now we look at the situation when the outcome variable is non-negative. For example, we might be interested in the time to an event, such as the time it takes to respond to cancer treatment or how long it takes for the tumor to shrink; or the time to failure, such as death; or in a fertility study, how long does it take for a woman to get pregnant; or when studying HIV transmission, how long before a partner gets infected with the virus.

All of these are non-negative variables, so once again we probably would not want to use linear regression in general, since a non-horizontal straight line is both positive and negative. Plus, more importantly, with survival data, for example, linear methods are not optimal.

When these outcome variables are positive, the first idea might be to simply transform them by taking their logarithms, and thus removing the positivity constraint. This approach is fine, although it needs to be modified some to include the possibility of an outcome of zero, but there are ways of handling that circumstance. But such an approach does not handle incomplete or censored data, too well. We explain the meaning of this below, after we introduce the new method we recommend. First, though we revisit how we handled survival data at the population level, a few weeks ago, because it is quite closely related to the proposed, new method.



Let us return to Edmund Halley and recall how his method for creating the life table out of Neumann's data.



Here is the graph we get from applying Halley's method to the data for 2007 in the USA. Along the horizontal we have age and on the vertical we have proportion surviving to a particular age. This way we see the survival experience for this constructed cohort of individuals.

Age	Probability of dying between ages $x$ to $x + 1$	$q_x$
0–1	0.006761	
1–2	0.000460	
2–3	0.000286	
3–4	0.000218	
4–5	0.000176	
5–6	0.000164	
6–7	0.000151	
7–8	0.000140	
8–9	0.000124	
9–10	0.000105	
10–11	0.000091	
11–12	0.000094	
12–13	0.000132	
13–14	0.000209	
14–15	0.000314	
15–16	0.000426	
16–17	0.000529	
17–18	0.000627	
18–19	0.000715	
19–20	0.000796	

To remind ourselves, in order to construct this survival curve, or life table, we start with the life span broken into age intervals, and then obtain the conditional probabilities that someone entering an interval will die within that interval—the hazard function. With this set of probabilities we can trace the mortality, or life experience of a cohort subjected to these hazards, as it progresses through time. And that is the life table—it is a way to convert the hazard function into an associated survival function.

Age	Probability of dying between ages $x$ to $x + 1$	$l_x$	$d_x$
	$q_x$		
0–1	0.006761	100,000	676
1–2	0.000460		
2–3	0.000286		
3–4	0.000218		
4–5	0.000176		
5–6	0.000164		
6–7	0.000151		
7–8	0.000140		
8–9	0.000124		
9–10	0.000105		
10–11	0.000091		
11–12	0.000094		
12–13	0.000132		
13–14	0.000209		
14–15	0.000314		
15–16	0.000426		
16–17	0.000529		
17–18	0.000627		
18–19	0.000715		
19–20	0.000796		

In particular, to start the curve off, suppose 100,000 babies are born (enter the first interval). Of these, a proportion of 0.006761 die. That means 676 die.

Age	Probability of dying between ages $x$ to $x + 1$	Number surviving to age $x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
0–1	0.006761	100,000	676
1–2	0.000460	99,324	46
2–3	0.000286		
3–4	0.000218		
4–5	0.000176		
5–6	0.000164		
6–7	0.000151		
7–8	0.000140		
8–9	0.000124		
9–10	0.000105		
10–11	0.000091		
11–12	0.000094		
12–13	0.000132		
13–14	0.000209		
14–15	0.000314		
15–16	0.000426		
16–17	0.000529		
17–18	0.000627		
18–19	0.000715		
19–20	0.000796		

That leaves 99,324 to reach their first birthday. Now, of those a proportion of 0.00460 will die. That means 46 ( $= 99,324 \times 0.00460$ ) of these will die before reaching their second birthday.

Age	Probability of dying between ages $x$ to $x + 1$	Number surviving to age $x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
0–1	0.006761	100,000	676
1–2	0.000460	99,324	46
2–3	0.000286	99,278	28
3–4	0.000218		
4–5	0.000176		
5–6	0.000164		
6–7	0.000151		
7–8	0.000140		
8–9	0.000124		
9–10	0.000105		
10–11	0.000091		
11–12	0.000094		
12–13	0.000132		
13–14	0.000209		
14–15	0.000314		
15–16	0.000426		
16–17	0.000529		
17–18	0.000627		
18–19	0.000715		
19–20	0.000796		

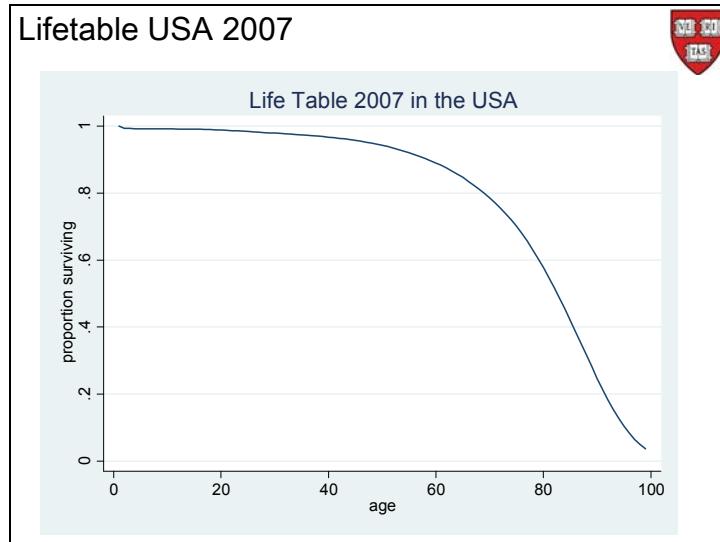
That means 99,278 ( $= 99,324 - 46$ ) reach their second birthday. Of those, the proportion who die before they reach their third birthday is 0.000286. So of these, 28 ( $= 99,278 \times 0.000286$ ) die before they reach their third birthday. So  $99,278 - 28$  reach their third birthday.

And continuing this logic, we can tumble down the table to generate the survival experience of all 100,000 in this constructed cohort.

Age	Probability of dying between ages $x$ to $x + 1$	$l_x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
0-1 .....	0.006761	100,000	676
1-2 .....	0.000460	99,324	46
2-3 .....	0.000286	99,278	28
3-4 .....	0.000218	99,250	22
4-5 .....	0.000176	99,228	17
5-6 .....	0.000164	99,211	16
6-7 .....	0.000151	99,194	15
7-8 .....	0.000140	99,179	14
8-9 .....	0.000124	99,166	12
9-10 .....	0.000105	99,153	10
10-11 .....	0.000091	99,143	9
11-12 .....	0.000094	99,134	9
12-13 .....	0.000132	99,125	13
13-14 .....	0.000209	99,112	21
14-15 .....	0.000314	99,091	31
15-16 .....	0.000426	99,060	42
16-17 .....	0.000529	99,018	52
17-18 .....	0.000627	98,965	62
18-19 .....	0.000715	98,903	71
19-20 .....	0.000796	98,832	79

Now divide the third column ( $l_x$ ) by 100,000 to turn the numbers into proportions and get the survival curve shown below.

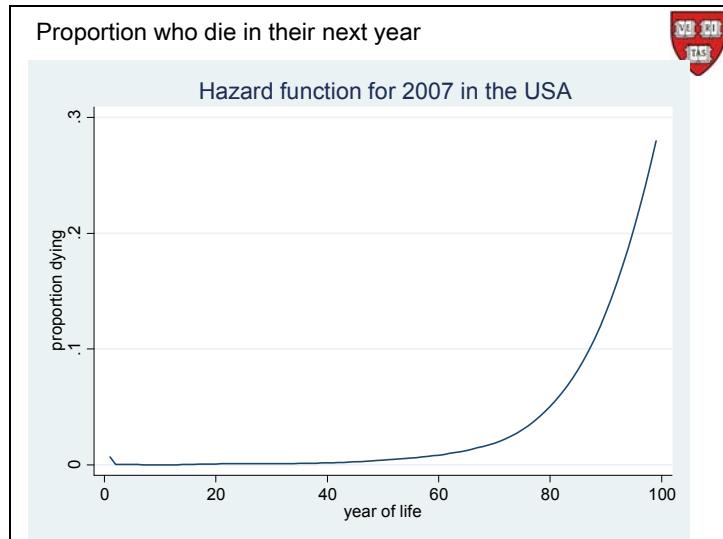
In passing, it should be noted that we could have taken another (more positive?) approach by looking at  $(1-q_x)$  to find the probability of surviving the interval, and thence a particular  $l_x$  entry is simply the product of the  $(1-q_x)$  and the  $l_x$  from the previous row. For example,  $(1 - 0.006761)100,000 = 99,324$  ;  $(1 - 0.000460)99,324 = 99,278$  ; etcetera. Why? (Answer, below.)



This is just a reminder of how we calculated the survival curve.

Age	Probability of dying between ages $x$ to $x + 1$	$l_x$	$d_x$
	$q_x$		
0–1	0.006761	100,000	676
1–2	0.000460	99,324	46
2–3	0.000286	99,278	28
3–4	0.000218	99,250	22
4–5	0.000176	99,228	17
5–6	0.000164	99,211	16
6–7	0.000151	99,194	15
7–8	0.000140	99,179	14
8–9	0.000124	99,166	12
9–10	0.000105	99,153	10
10–11	0.000091	99,143	9
11–12	0.000094	99,134	9
12–13	0.000132	99,125	13
13–14	0.000209	99,112	21
14–15	0.000314	99,091	31
15–16	0.000426	99,060	42
16–17	0.000529	99,018	52
17–18	0.000627	98,965	62
18–19	0.000715	98,903	71
19–20	0.000796	98,832	79

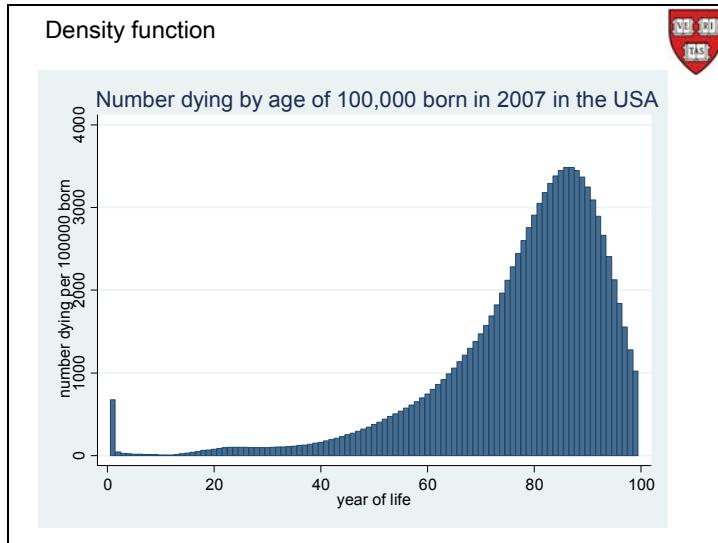
It is also interesting to look at the column labeled  $q_x$ . This gives us the values of the *hazard function* as  $x$  varies. That is the conditional probability of dying within an interval, given that one has survived to the beginning of the interval. So, for example, if we consider two-year-olds we see that the probability is 0.000286 that one of them does not reach their third birthday.



Plotting the hazard function shows that there is a blip in the first year of life, but subsequently, until age 40 or 50, the chance of dying within the next year of life is negligible. After that it increases exponentially. (Check this linearity out for yourself by plotting this function using a log scale on the vertical axis.)

Age	Probability of dying between ages $x$ to $x + 1$	$l_x$	Number dying between ages $x$ to $x + 1$
	$q_x$		
0-1	0.006761	100,000	676
1-2	0.000460	99,324	46
2-3	0.000286	99,278	28
3-4	0.000218	99,250	22
4-5	0.000176	99,228	17
5-6	0.000164	99,211	16
6-7	0.000151	99,194	15
7-8	0.000140	99,179	14
8-9	0.000124	99,166	12
9-10	0.000105	99,153	10
10-11	0.000091	99,143	9
11-12	0.000094	99,134	9
12-13	0.000132	99,125	13
13-14	0.000209	99,112	21
14-15	0.000314	99,091	31
15-16	0.000426	99,060	42
16-17	0.000529	99,018	52
17-18	0.000627	98,965	62
18-19	0.000715	98,903	71
19-20	0.000796	98,832	79

The last column, the one labeled  $d_x$ , can also be plotted to show when each of the 100,000 in the cohort, die. It tells us how densely the individuals are packed into each age category and thus is called the density function (when normalized; i.e. divide by 100,000.)



Technically the density function should have total area equal to one—easier to see if we divide the vertical scale by 100,000. From this we see that a relatively big number of babies die in their first year of life, but then very few die until the late teens. Then mortality picks up, with most of us dying in our late eighties, early nineties, and then decreases since there are not too many of us who live to our late nineties.

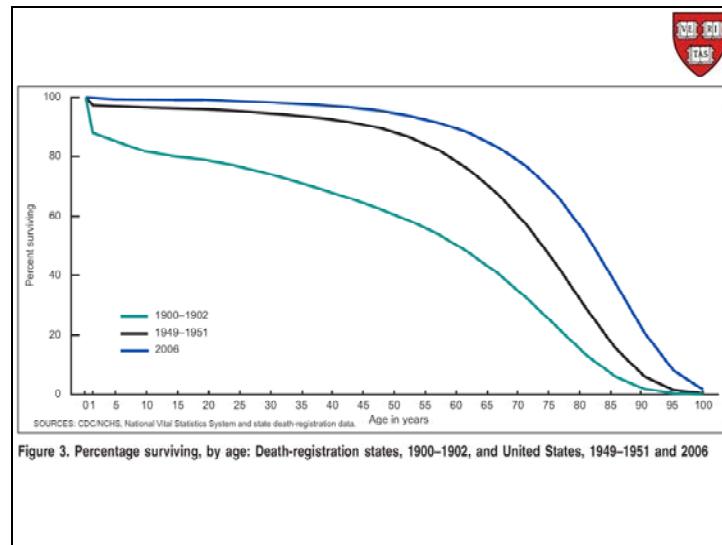
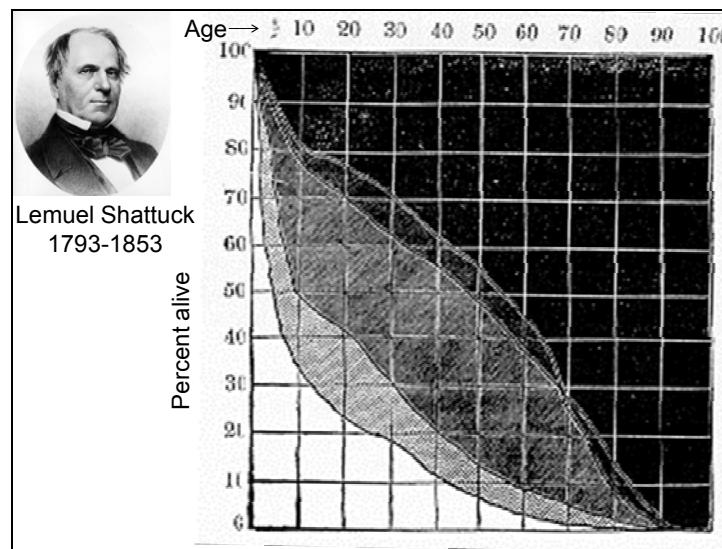


Figure 3. Percentage surviving, by age: Death-registration states, 1900–1902, and United States, 1949–1951 and 2006

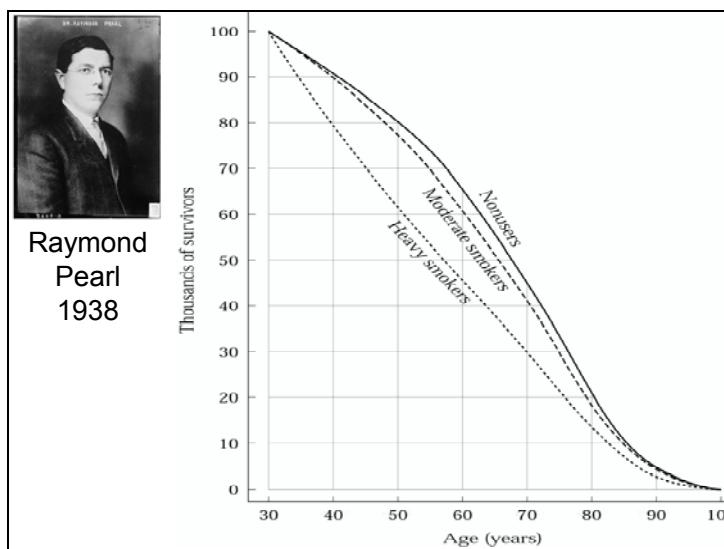
It is interesting to compare these life tables over the last century. At the beginning of the last century we see a very high mortality in the first year of life, which we do not see anymore. This improvement in mortality in the first years of life has had a huge impact on our life expectancy—remember that life expectancy is the area under the curve—because lifting the curve at the left resulted in lifting the curve throughout the life span.

The length of life has not changed much in the last century. What has changed is the proportions living longer—the curves do not extend much beyond where they have extended in the past, just the level of the curve on the right has increased.



These are the life tables Lemuel Shattuck published<sup>1</sup> in the document that was instrumental in the forming of the first State Health Department in the US: Massachusetts.

<sup>1</sup> Report to the Committee of the City Council appointed to obtain the census of Boston for the year 1845: embracing collateral facts and statistical researches, illustrating the history and condition of the population, and their means of progress and prosperity, Lemuel Shattuck, Boston (Mass.) John H. Eastburn, City Printer (1846)



We also looked at the depiction of the study of 7,000 individuals by Raymond Pearl comparing the three groups—nonsmokers, moderate smokers, and heavy smokers<sup>2</sup>. These curves can be read to evaluate the effects of smoking.

Table 3. Cumulative percent of never-married males and females 15–19 years of age who have ever had sexual intercourse before reaching selected ages, by age, race, and Hispanic origin: United States, 1988, 1995, and 2002

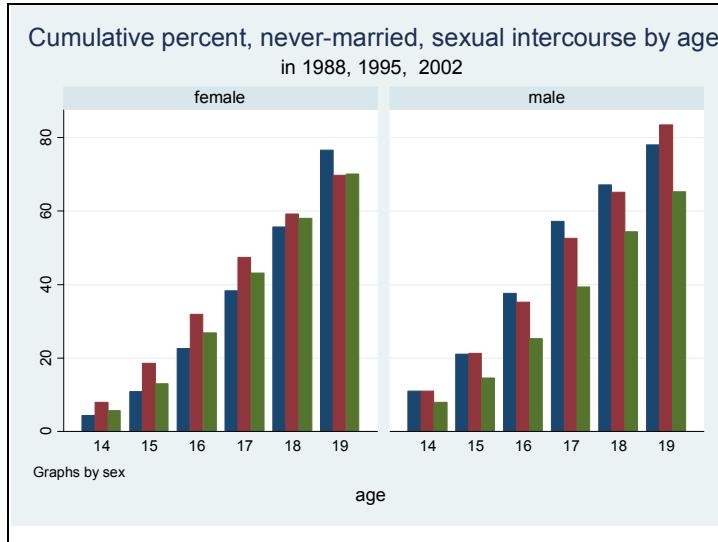
Table 3-Supplement

Characteristic	Female			Male		
	1988	1995	2002	1988	1995	2002
All never-married <sup>1</sup> . . . . .	51.1	49.3	45.5	60.4	55.2	45.7
Age						
14 years . . . . .	4.4	8.0	5.7	11.0	11.0	7.9
15 years . . . . .	10.9	18.6	13.0	21.1	21.3	14.6
16 years . . . . .	22.6	31.9	26.8	37.6	35.2	25.3
17 years . . . . .	38.3	47.4	43.1	57.2	52.6	39.4
18 years . . . . .	55.7	59.2	58.0	67.1	65.1	54.3
19 years . . . . .	76.5	69.7	70.1	78.0	83.4	65.2

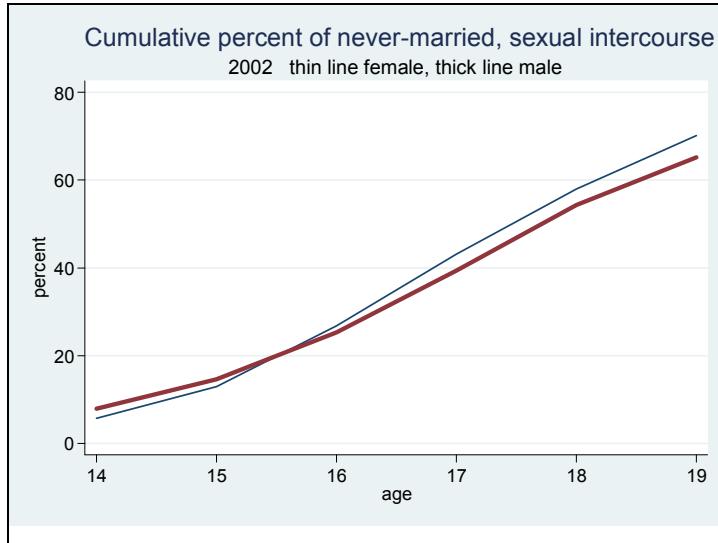
Survival methods can also be used to examine the time to an event, such as the first time to sexual intercourse. These are the results of three such surveys from the Centers for Disease Control and Prevention, for studying teenage behavior<sup>3</sup>.

<sup>2</sup> Pearl, R., 1938 Tobacco smoking and longevity. Science 87: 216–217.

<sup>3</sup> Abma JC, Martinez, GM, Mosher, WD., Dawson, BS. Teenagers in the United States: Sexual activity, contraceptive use, and childbearing, 2002. National Center for Health Statistics. Vital Health Stat 23(24). 2004.



We can plot these data as bar graphs. We can see that, as it should, the bars increase monotonically from left to right for each sex and for each survey.



Choosing a particular year, 2002 here, one can plot the cumulative curves. These cumulative curves are complementary to the survival curves in that the later go from 100% monotonically down to zero, whereas these curves monotonically increase from zero (not shown because we only have information in the 14-19 year age groups) to 100%.

With survival curves, if instead of plotting the percent who survived we plot the percent who die, we would then get a cumulative curve that monotonically increases with time; just like the one above. One minus the survival curve gives us a cumulative curve, namely a curve representing the percent who die.

## Product Limit Method

**Longitudinal**

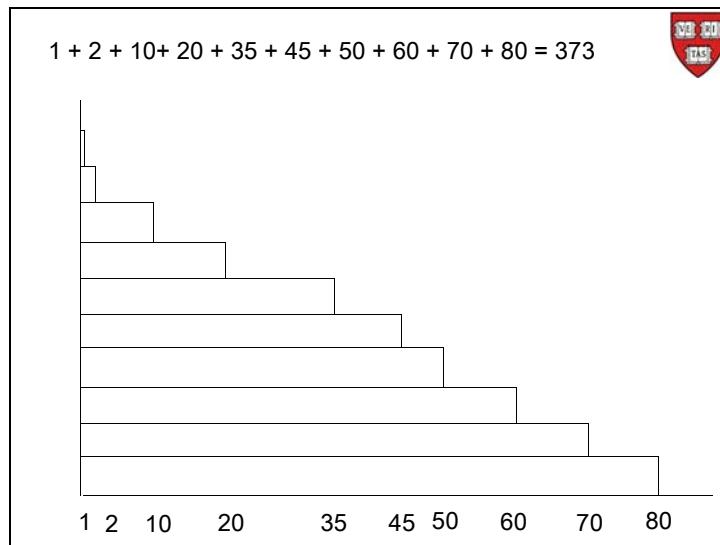


Note that all of these are cross-sectional studies attempting to describe a longitudinal process.

Another approach is to follow the cohort longitudinally.

In all of these examples that we have seen up to now, except for one, we tried to quantify a longitudinal experience—people living through their life spans—by using cross-sectional data—what happens to people over a year of life, by and large. It is a cohort we have constructed, to try to emulate what we might observe were we to follow them for a hundred years, or more.

What if we did have a longitudinal cohort, such as we had with the sticks—this is the exception we noted, above? How would we then estimate the survival curve? This is the longitudinal approach we now follow.



When we looked at the 10 sticks, we actually followed them until they all died, and this is the resulting survival curve. It is the properties of such a curve that we now study.

## Longitudinal



Note that all of these are cross-sectional studies attempting to describe a longitudinal process.

Another approach is to follow the cohort longitudinally.

Consider a sampling approach, i.e. not the whole population.

Exactly what we look at is the situation where we want to make inference about a population survival curve, when having information on only a random sample from that population—think of the ten sticks as a random sample of ten individuals from the population. So we look at making inference about another curve in the population, namely the survival curve.

## AIDS Data



Interval from AIDS to death Hemoph. < 41

Patient	Survival (months)
1	2
2	3
3	6
4	6
5	7
6	10
7	15
8	15
9	16
10	27
11	30
12	32

A case in point: Here is a sample of 12 individuals from an old study of haemophiliacs under the age of 41, and we see their survival beyond their diagnosis with AIDS. And the question that was being asked in this aspect of this study was, whether there is an age effect on survival subsequent to being diagnosed with AIDS.

### Life table approach:

[t, t+1)	# dying at t
0-1	0
1-2	0
2-3	1
3-4	1
4-5	0
5-6	0
6-7	2
7-8	1
8-9	0
9-10	0
10-11	1

We can take a life table approach and split the time interval into months. Here are the first eleven such intervals. And we can note how many died in each interval—the left endpoint of the interval is considered in the interval, whereas the right endpoint is considered to be in the subsequent interval.

### Life table approach:

[t, t+1)	# survive to t+	# dying at t
0-1	12	0
1-2	12	0
2-3	11	1
3-4	10	1
4-5	10	0
5-6	10	0
6-7	8	2
7-8	7	1
8-9	7	0
9-10	7	0
10-11	6	1

Now create another column that keeps a tally of how many survive to the beginning of the interval. To continue with our prescription for creating a life table, we need to know what the probabilities are of failing within an interval.

### Life table approach:

[t, t+1)	Prob dying in t to t+1	# survive to t+	# dying at t
0-1	0/12	12	0
1-2	0/12	12	0
2-3	1/12	11	1
3-4	1/11	10	1
4-5	0/10	10	0
5-6	0/10	10	0
6-7	2/10	8	2
7-8	1/8	7	1
8-9	0/7	7	0
9-10	0/7	7	0
10-11	1/7	6	1

Create a new column that calculates the probability of dying within the interval; the number who die in the interval divided by the number entering the interval. This gives us all we need to create a life table.

### Life table approach:

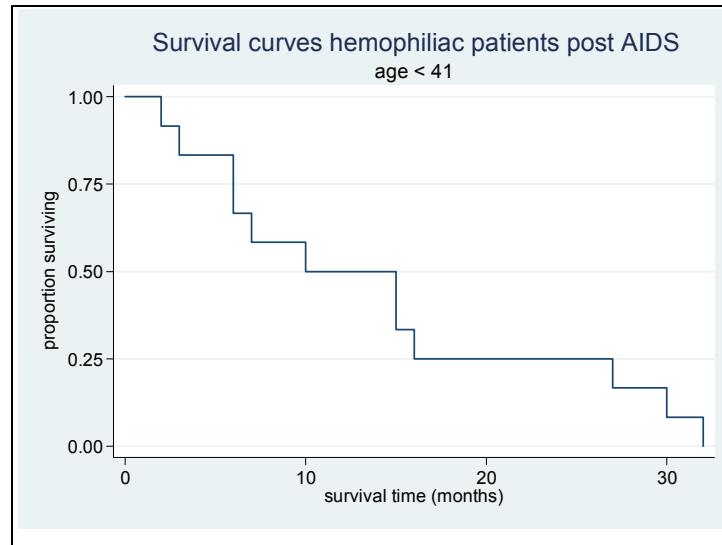
[t, t+1)	Prob dying in t to t+1	# survive to t+	# dying at t	S(t)
0-1	0/12	12	0	1
1-2	0/12	12	0	1
2-3	1/12	11	1	11/12=0.92
3-4	1/11	10	1	10/12=0.83
4-5	0/10	10	0	10/12=0.83
5-6	0/10	10	0	10/12=0.83
6-7	2/10	8	2	8/12=0.67
7-8	1/8	7	1	7/12=0.58
8-9	0/7	7	0	7/12=0.58
9-10	0/7	7	0	7/12=0.58
10-11	1/7	6	1	6/12=0.50

And voila, we've got the survival curve. So the survival curve is just this, the number surviving of the 12 who came in. To standardize, we divide by 12, to give us the last column.

That would be the life table approach to this summary. We could plot this survival curve, but before doing that, let us do some cleaning up. We see that nothing happens to the curve except when an event—a death in this example—happens. For example, the curve starts at 1 and continues there until at two months it goes down to 0.92. Then at three months it goes down to 0.83 and stays there until it reaches six months. Thus we could replace the 0-1 and 1-2 intervals with a 0-2 interval, and the 3-4, 4-5, and 5-6 intervals with a 3-6 interval. Or even, replace all the intervals by just the time points at which an event happens.

Product Limit Method: only at deaths				
[t, t+)	Prob dying in t to t+	# survive to t+	# dying at t	S(t)
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3	1/11	10	1	10/12=0.83
6	2/10	8	2	8/12=0.67
7	1/8	7	1	7/12=0.58
10	1/7	6	1	6/12=0.50
15	2/6	4	2	4/12=0.33
16	1/4	3	1	3/12=0.25
27	1/3	2	1	2/12=0.17
30	1/2	1	1	1/12=0.08
32	1	0	1	0

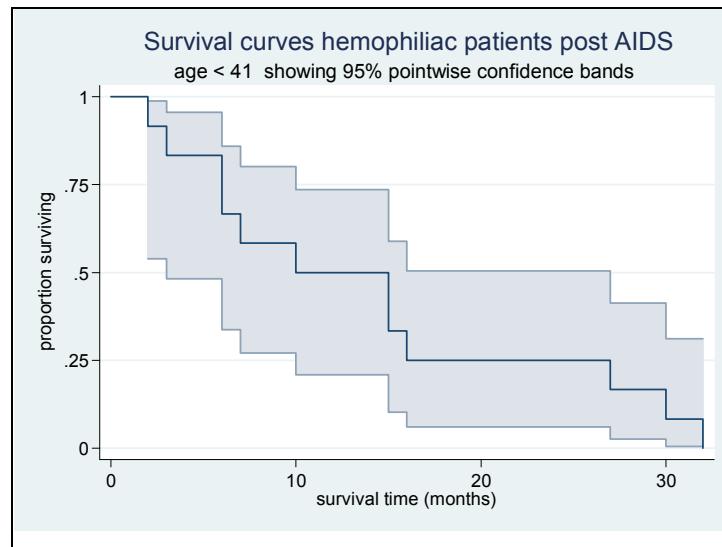
This is the product limit method of estimating the survival curve. It represents a minor change over the life table method.



Here is what the survival curve looks like, and it reminds us of the stick example we considered earlier.

Product Limit Method: only at deaths				
[t, t+)	Prob dying in t to t+	# survive to t+	# dying at t	S(t)
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3	1/11	10	1	10/12=0.83
6	2/10	8	2	8/12=0.67
7	1/8	7	1	7/12=0.58
10	1/7	6	1	6/12=0.50
15	2/6	4	2	4/12=0.33
16	1/4	3	1	3/12=0.25
27	1/3	2	1	2/12=0.17
30	1/2	1	1	1/12=0.08
32	1	0	1	0

Looking at the rightmost column, the survival curve, we see that it is calculated as a ratio; the number alive of the number who started. This reminds us that these 12 patients were a sample of patients. In particular we have the estimate of the survival at that point and we can construct a confidence interval at any point in time<sup>4</sup>.



We can collect all these confidence intervals at every time point and display them as the shaded region above.

<sup>4</sup> The theory for these confidence intervals are beyond the scope of this course. For the interested reader we refer you to, Greenwood, M. 1926. The natural duration of cancer. *Reports on Public Health and Medical Subjects* 33: 1–26.

So when you have a sample, you can treat this curve just like any statistic we have seen before, *and in particular* there is uncertainty associated with it, which we can quantify.



. sts list if age==1						
failure _d: death analysis time _t: surv						
Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf. Int.]
2	12	1	0	0.9167	0.0798	0.5390 0.9878
3	11	1	0	0.8333	0.1076	0.4817 0.9555
6	10	2	0	0.6667	0.1361	0.3370 0.8597
7	8	1	0	0.5833	0.1423	0.2701 0.8009
10	7	1	0	0.5000	0.1443	0.2085 0.7361
15	6	2	0	0.3333	0.1361	0.1027 0.5884
16	4	1	0	0.2500	0.1250	0.0601 0.5048
27	3	1	0	0.1667	0.1076	0.0265 0.4130
30	2	1	0	0.0833	0.0798	0.0051 0.3111
32	1	1	0	0.0000	.	.

Here is the *Stata* command *which will display* the numbers used *in* the above graph.

### Censored Observations

### Incomplete Longitudinal — Censoring

Suppose some of the patients are still not “dead” at the time the analysis is performed —

censored observations

One of the many complications associated with longitudinal studies, is *something* that we *call* censoring:

Consider a typical clinical trial. *It has* a start date when the trial *opens* to accepting patients; let us say January 1, 2006. Not all the patients arrive that day, but they accumulate over time, *and we typically assume that they arrive at random*. Suppose *the endpoint of interest* is survival. What that typically means is that *survival of the patients is measured beyond a start point, such as the point in time when the patient joins the trial*. In the prior graph it was the survival time after being diagnosed with AIDS. There is an implicit assumption of uniformity here; for example, that two months survival after being diagnosed in

January is to be considered the same as two months survival after being diagnosed in June. We make this assumption, and that allows us to draw a single curve incorporating all patients as starting from the same point.

Now suppose the study has been going on for a year and you would like to monitor the trial. For example, you may wish to conclude that the treatments are not too toxic, or, if two treatments are being compared, that one treatment is not far better than the other, since it would not be ethical to continue to put patients on the lesser treatment. So you decide to do an interim analysis of the trial. We should note that there are a number of famous trials that have been stopped earlier than planned because of interim results. Care must be taken when stopping trials early, and the statistical estimation methods subsequently used must reflect this early termination, otherwise they are statistically unsound.<sup>5</sup>

Returning to the analysis of an ongoing trial, the measurements may reflect that some of the patients have reached the endpoint, death, say, while others have not reached the endpoint and are still alive. How to calculate the survival curve? These “partial” observations—patients you have observed for some time but have not yet reached the endpoint—do provide partial information about the time to the endpoint, but not as much information as those who have died. For example, for the patient who is still alive three months after entering the study we only know that survival will be more than three months without knowing the exact survival time.

We could discard these partial data points, but by discarding the information we have on these patients we run the risk of introducing bias into our study. For example, if all the patients on the one treatment in a randomized trial are dead at the time we do our analysis, whereas all the patients on the other treatment are alive, then surely that could be very informative. So discarding such data is not a general solution.

These data are called censored observations. Why the censoring occurred may be informative—for example, if we see that a patient is responding poorly to a medication we would take that patient off the trial, and even though that would be a censored observation, it is also informative since we may classify it as a treatment failure—and that leads to a whole area of research that we do not pursue further, here. So we assume henceforth that the censoring occurs at random, or is non-informative—we cannot read anything into the censoring. For example, the family moved out of town and that is why we lost the patient—often labeled, lost to follow up.

Let us look at an example with uninformative censoring.

---

<sup>5</sup> Bassler D, Briel M, Montori VM, Lane M, et al, Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA. 2010 Mar 24;303(12):1180-7.

Product Limit Method: only at deaths				
[t, t+)	Prob dying in t to t+	# survive to t+	# dying at t	S(t)
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3+	1/11	10	1	10/12=0.83
6	2/10	8	2	8/12=0.67
7	1/8	7	1	7/12=0.58
10+	1/7	6	1	6/12=0.50
15	2/6	4	2	4/12=0.33
16	1/4	3	1	3/12=0.25
27	1/3	2	1	2/12=0.17
30	1/2	1	1	1/12=0.08
32	1	0	1	0

Returning to our hæmophiliac example, let us change these data and say, just for argument's sake, that the two patients, the one who died at 3 months and the other who died at 10 months, were both censored (indicated by a + sign); i.e. they were still alive at those time points, but we know nothing about them beyond those time points.

What this means is that Columns two through five should now be modified.

Product Limit Method: only at deaths				
[t, t+)	Prob dying in t to t+	# survive to t+	# dying at t	S(t)
0	0/12	12	0	1
2	1/12	11	1	11/12=0.92
3+	0/11	11	0	11/12=0.92
6	2/10	8	2	
7	1/8	7	1	
10+	0/7	7	0	
15	2/6	4	2	
16	1/4	3	1	
27	1/3	2	1	
30	1/2	1	1	
32	1	0	1	0

The first change to make is that zero persons died at both 3 months and 10 months. So the correct entries in Column two are 0/11 and 0/7, respectively. Also the entries in Column three should be 11 and 7, because 11 survived 3 months—now patient at time 3 did not die, so 11 reached time 3 and 11 were still alive immediately subsequent to time 3—and similarly for time 10 months—7 reached and survived 10 months. That means there should not be any change to the survival curve at either 3 months or 10 months. The changes to column four are clear. How to change column 5, requires some thought.

Product Limit Method: only at deaths					
[t,t+)	Prob dying in t to t+	# survive to t+	Prob surv In t to t+	# dying at t	S(t)
0	0/12	12	12/12	0	1
2	1/12	11	11/12	1	11/12=0.92
3+	1/11	10	10/11	1	10/12=0.83
6	2/10	8	8/10	2	8/12=0.67
7	1/8	7	7/8	1	7/12=0.58
10+	1/7	6	6/7	1	6/12=0.50
15	2/6	4	4/6	2	4/12=0.33
16	1/4	3	3/4	1	3/12=0.25
27	1/3	2	2/3	1	2/12=0.17
30	1/2	1	1/2	1	1/12=0.08
32	1	0	0	1	0

Let us return to the complete data tableau and introduce a fourth column, the more optimistic view of column two; namely, the probability of surviving the interval  $t$  to  $t+$  (whereas column two is the probability of dying in the interval  $t$  to  $t+$ ). So at every row, columns two and four need to sum to one.

The advantage of this new column (column 4) is that it can be used to calculate the probability of survival to any point. Pick a row. Let us say we choose the “6 month” row, and ask, what is the probability of surviving beyond 6 months?

One way to answer this is to observe that we must first survive to just before 6 months and then survive through the sixth month. From the above table, the probability of surviving to just before 6 months is  $10/12=0.83$ . Then the probability of surviving from 6 to just beyond 6 is  $8/10$ . So the probability of surviving to just beyond 6 months is  $10/12 \times 8/10 = 8/12 = 0.67$ —the multiplicative rule of probability:  $P(A \cap B) = P(A)P(B|A)$ .

We can carry out this logic for every row and thus generate the whole table.

Product Limit Method: only at deaths			
[t, t+)	Prob dying in t to t+	Prob surv In t to t+	S(t)
0	0/12	12/12	1
2	1/12	11/12	$11/12=0.92$
3+	1/11	10/11	$10/11=0.83$
6	2/10	8/10	$8/10=0.67$
7	1/8	7/8	$7/8=0.58$
10+	1/7	6/7	$6/7=0.50$
15	2/6	4/6	$4/6=0.33$
16	1/4	3/4	$3/4=0.25$
27	1/3	2/3	$2/3=0.17$
30	1/2	1/2	$1/2=0.08$
32	1	0	0

This slide schematically shows how to generate the whole survival curve. This is the recipe we use below to incorporate the censoring.

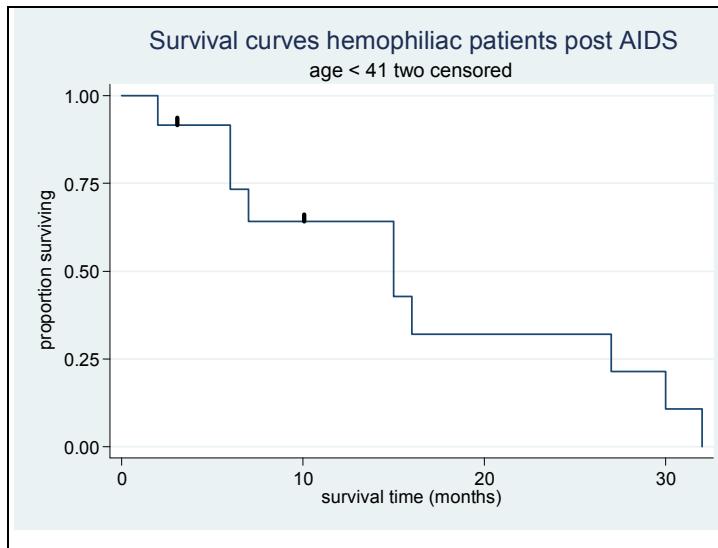
Product Limit Method: only at deaths				
[t, t+)	Prob dying in t to t+	# survive to t+	Prob surv In t to t+	S(t)
0	0/12	12	12/12	1
2	1/12	11	11/12	0.92
3+	0/11	11	11/11	$11/11 \times 0.92 = 0.92$
6	2/10	8	8/10	$8/10 \times 0.92 = 0.73$
7	1/8	7	7/8	$7/8 \times 0.73 = 0.64$
10+	0/7	7	7/7	$7/7 \times 0.64 = 0.64$
15	2/6	4	4/6	$4/6 \times 0.64 = 0.43$
16	1/4	3	3/4	$3/4 \times 0.43 = 0.32$
27	1/3	2	2/3	$2/3 \times 0.32 = 0.21$
30	1/2	1	1/2	$1/2 \times 0.21 = 0.11$
32	1	0	0	0.00

Let us return to the censored observations and apply this logic to the new situation.

First let us look at the observation at month three. Now nobody dies, so column 2 is 0/11 and column 3 is 11. Thus column 4 becomes 11/11. Similarly at month ten, column 2 is 0/7 and column 3 is 7 and column 4 is 7/7.

Now with the new column 4 we can generate a new column 5, just like we did before in the complete data (no censoring) case.

This is called the product limit method.

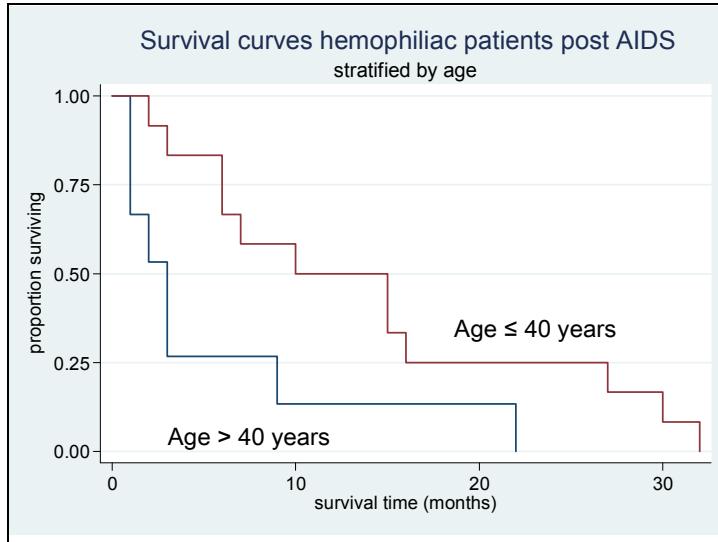


Here is the resultant curve—sometimes called the Kaplan-Meier curve. Two marks, one at 3 months and one at 10 months, show where the censored observations happened. So at the censored observations the curve remains flat and there are no steps.

Intuitively, one can see from the curve what the method does. Returning to the sticks or the curve when everyone died, then each step size was the same,  $1/10$  for the sticks and  $1/12$  for these 12 haemophiliacs. That is exactly what happens in this curve, which includes censored observations, too in the beginning—at 2 months 1 person dies and the curve goes down  $1/12$ . Now when we reach the person who is censored at 3 months, the curve does not go down, but we must still utilize the  $1/12$  of the whole that this person represents. What this method does is distributes that  $1/12$  equally to the 10 people to the right of 3 months. Now each person to the right is going to be “worth”  $1/12$  plus  $1/12 \times 1/10 = 11/10 \times 1/12 = 0.092$ . So at 6 months when 2 die, the curve should go down by  $0.184$ , which indeed it does ( $0.92 - 0.73 = 0.19$ ), up to roundoff. Similarly at 7 months, when another patient dies, the curve goes down by  $0.09$ .

This time at 10 months when another patient is censored, that patient is carries weight of  $0.09$  that needs to be distributed to the remaining 5 patients to the right of 10 months, so each of these five will carry weight of  $0.09 + 1/5 \times 0.09 = 0.108$ . That is the size of the steps the curve will take for every death that occurs.

This redistribution of the weights to the right of censored observations is like saying: as long as this person is alive the person enters into the denominator. Once the person leaves the study, I only know that the person is still alive and will die subsequent to the point in time when they left the study. How survival behaves beyond this point we can appeal to those still in the study, so let us say that this person who has left the study is equally likely to behave like any one of the remaining patients, and thus this method of equidistribution of the weight.



We now have seen how to estimate the survival curve and treat it like a statistic and quantify the uncertainty by calculating the confidence interval at points along the curve. We can also start to think about making comparative statements about two, or more, survival curves. For example, we can estimate the survival curves for the two samples of haemophiliac patients once they have been stratified by age—say age 40—and ask whether these two curves are statistically different.

In order to answer that question, there are a number of statistical tests one can perform. A popular one is the logrank test. It calculates a “distance” between these two curves and, as usual, then determines how likely it is that the distance as large, or larger, can be attributed to chance.

**Logrank Test**

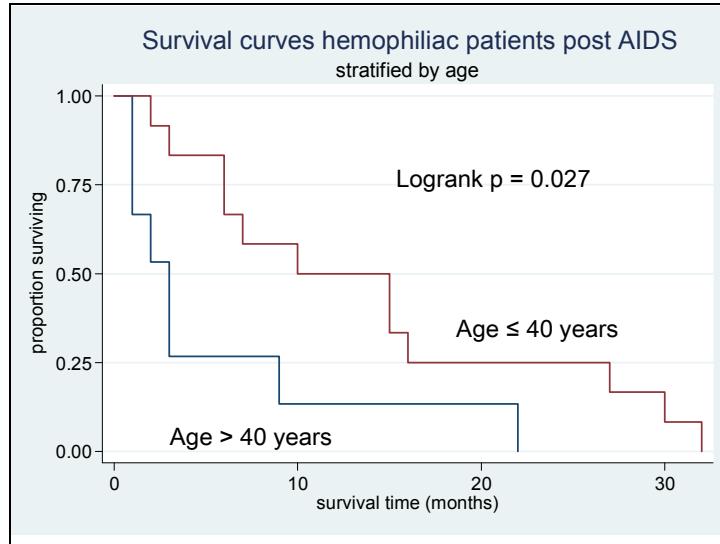


At each *death* point construct a 2x2 table:

	Dead	Alive	Total
Treat 1			
Treat 2			
Total			

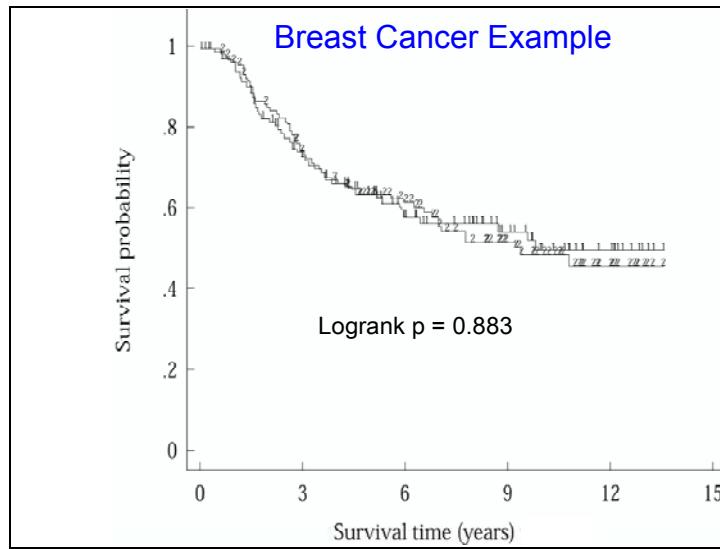
Then treat as g independent 2x2 tables in Mantel-Haenszel

The logrank test looks at every time point when either, or both, of the curves makes a change (where a death occurred), and creates a 2x2 table, as above. If there are a total of  $g$  distinct death times between the two groups, then we have  $g$  such tables. The logrank test then treats these as  $g$  independent 2x2 tables and appeals to the Mantel-Haenszel method for handling 2x2 independent tables.



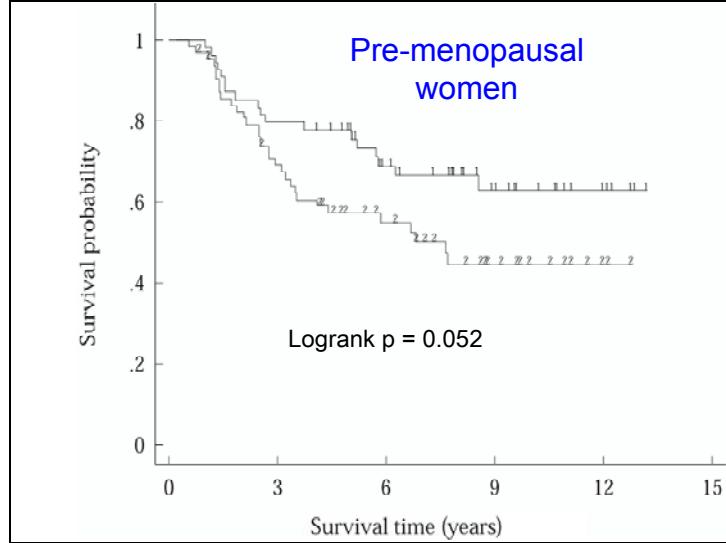
The result of the logrank test in this example yields a p-value of 0.027. So, on the basis of these two samples, we reject the null hypothesis that there is no survival difference between these two age groups at the 5% level.

We can also think about bringing in other covariates to explain the differences, and indeed, we can entertain the thought of extending regression consideration to the survival setting. That might lead us to the Cox model, but that is beyond the scope of this course. We look at one last example to show that even in a survival situation, we must still remain alert to the same problems we had before.

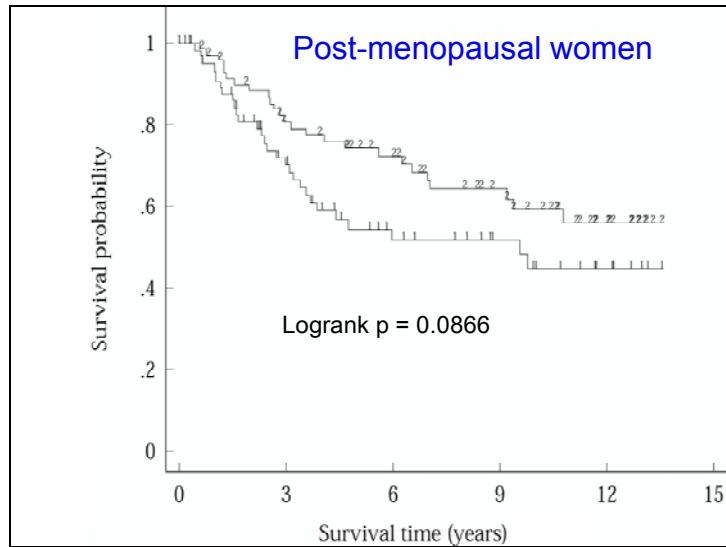


Here is a breast cancer study to compare survival experiences for two treatments. There is a lot of censoring evident, and to distinguish the two treatments, a "1" is used to indicate the censoring times for the first treatment, and a "2" indicates the censored values for patients on the second treatment.

From this graph it looks like there is no evidence of treatment differences as far as survival is concerned. The log rank p-value is 0.883 agreeing with our visual assessment that the difference between the two curves could easily be explained by sampling variability.



Instead of looking at the whole sample, let us just look at premenopausal women. Now we see a separation of the two curves, with the "1's on top. The log rank p-value is much less than before although, at 0.052, is not significant at the 5% level.



When we look at post-menopausal women we once again see a separation of the survival curves. The separation is not sufficiently large to be significant at the 5% level by the log rank test, since the p-value is 0.0866.

It is interesting to note that both curves show separation when we look at the subgroups determined by menopausal status, neither significant, but much larger separations than when we look at the women as a single group; ignoring classification by menopausal status. Also, note that the positioning of the curves has flipped; whereas treatment “1” seems to be superior for the pre-menopausal women, treatment “2” seems to be superior for post-menopausal women. The reason for showing you this example was for you to see that no matter how complicated the outcome, or statistic, we are measuring, the Yule effect (Simpson’s Paradox) can rear its ugly head if you ignore a covariate.