



Marcello Pagano

[JOTTER 1 A, WEEK ONE]

Types of data

- Nominal Data

1 – male

2 – female

0 – alive

1 – dead

Blood group:

- A – 1
- B – 2
- AB – 3
- O – 4

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

- Mild
- Moderate
- 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 as 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	$\geq 50\%$ by parasitological methods (intestinal and urinary schistosomiasis) or $\geq 30\%$ 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	$\geq 10\%$ 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	0 Fully active
2. Moderate	1 Ambulatory. Light work.
3. Severe	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 2010¹. 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

¹ 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.

World			Low-income countries ^a (< \$825)		
Disease or injury	Deaths (millions)	Per cent of total deaths	Disease or injury	Deaths (millions)	Per cent of total deaths
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 ^b	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
Middle-income countries			High-income countries (>\$10,066)		
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 WHO². 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.

² 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*

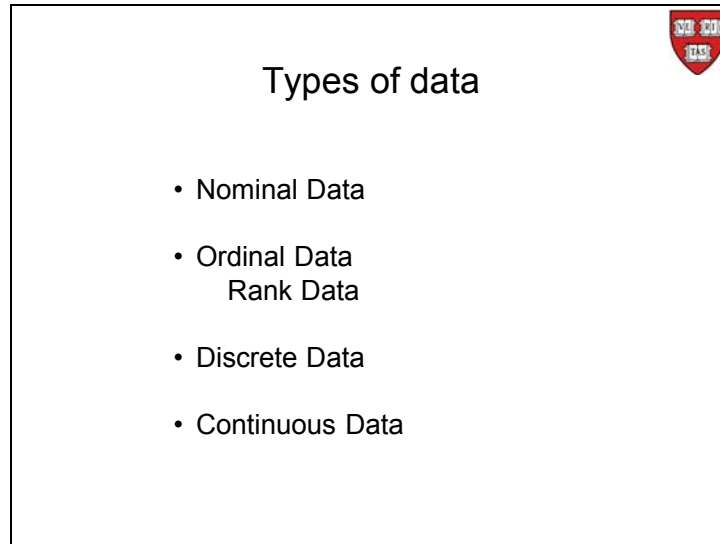
Note:

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.

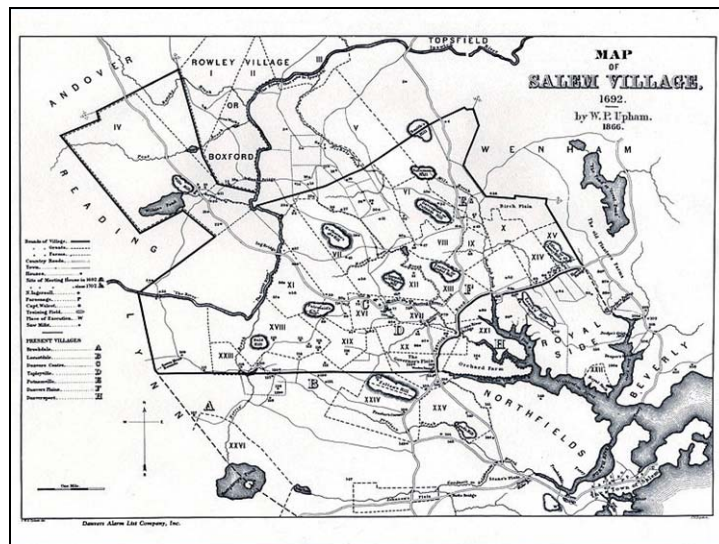


So in summary, these are the kinds of data that we looked at. So this is our taxonomy. It's basically into four groups-- nominal data, ordinal data, discrete data, and continuous data. And the reason why we go into this much detail is because we are going to use different statistical methods when we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

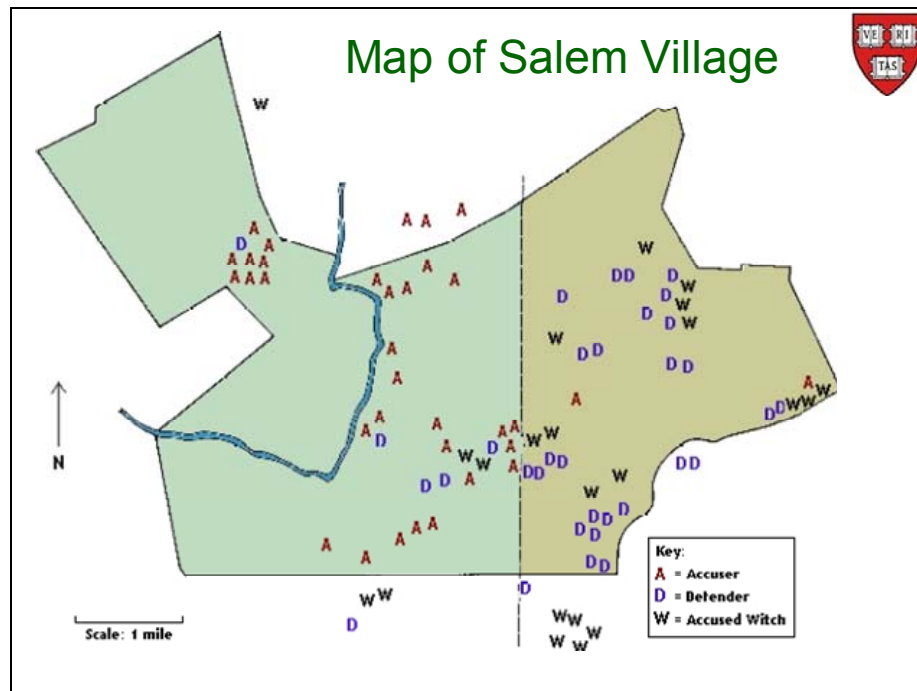
Mark Twain — Life on the Mississippi, 1883



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.³ Now, every year at Halloween, we all drive up to Salem Village and everybody celebrates the witches.⁴ 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.

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

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

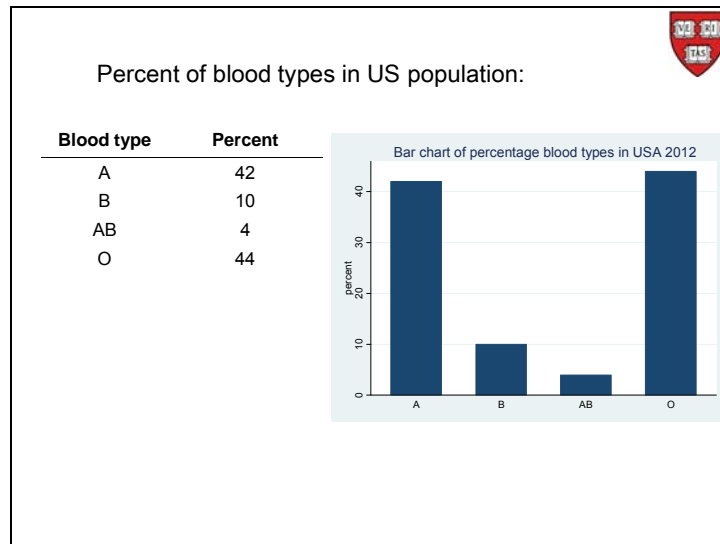


What we're told is that these women were possessed. This was in 1692. A historian decided to map out the addresses of the individuals accused of being possessed, the accusers and the defenders.⁵ These are denoted by W, D and A, respectively in the map above.

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.

⁵ 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 Cross⁶ 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:



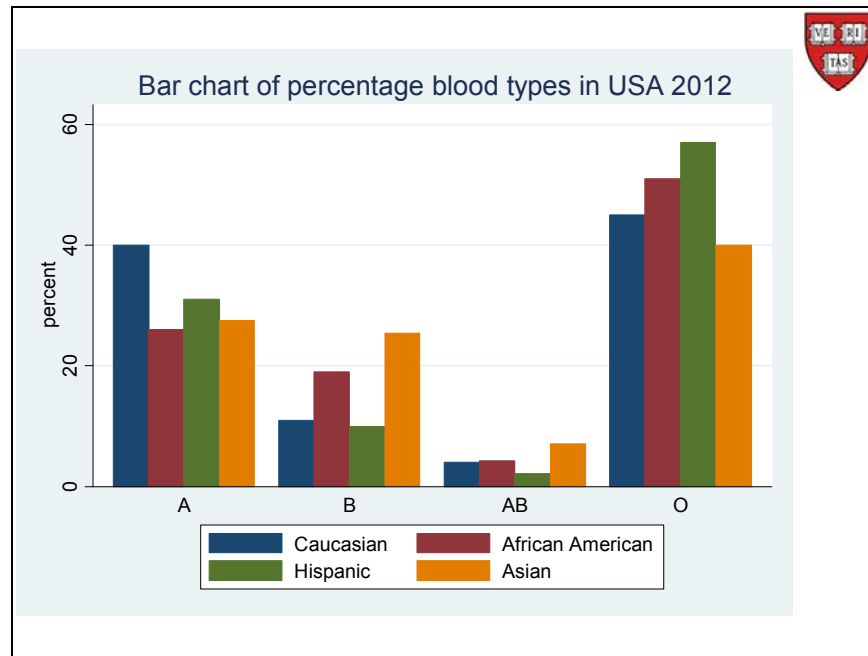
Ethnic group blood type breakdown according to the Red Cross

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

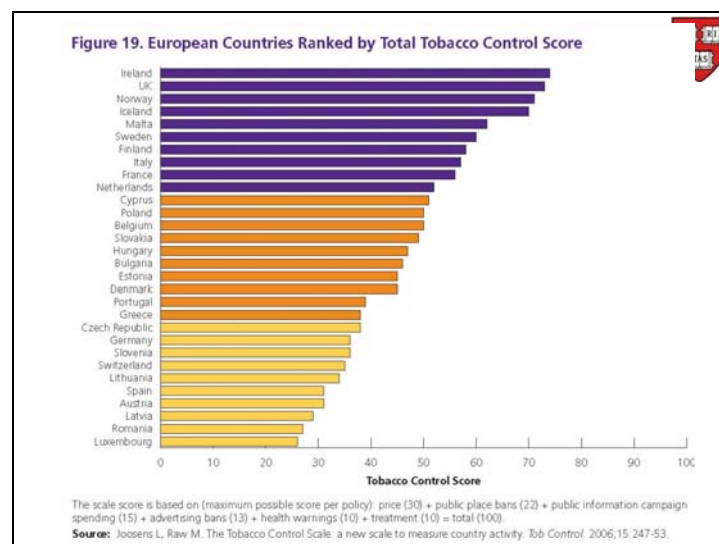
<http://www.redcrossblood.org/learn-about-blood/blood-types>

⁶ <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

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



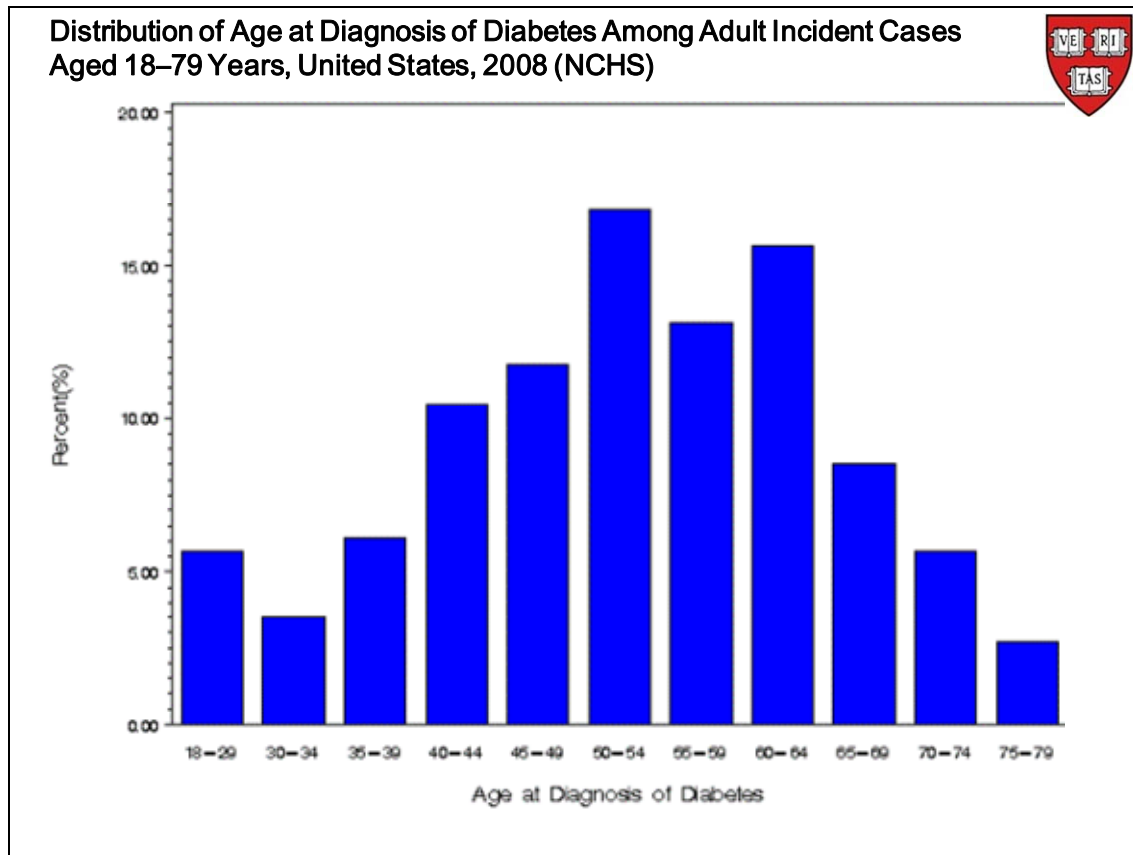
Age (Years)	Percent
18–29	5.7
30–34	3.5
35–39	6.1
40–44	10.5
45–49	11.8
50–54	16.8
55–59	13.1
60–64	15.6
65–69	8.5
70–74	5.6
75–79	2.7

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 cases⁷ -- 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

⁷ 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>

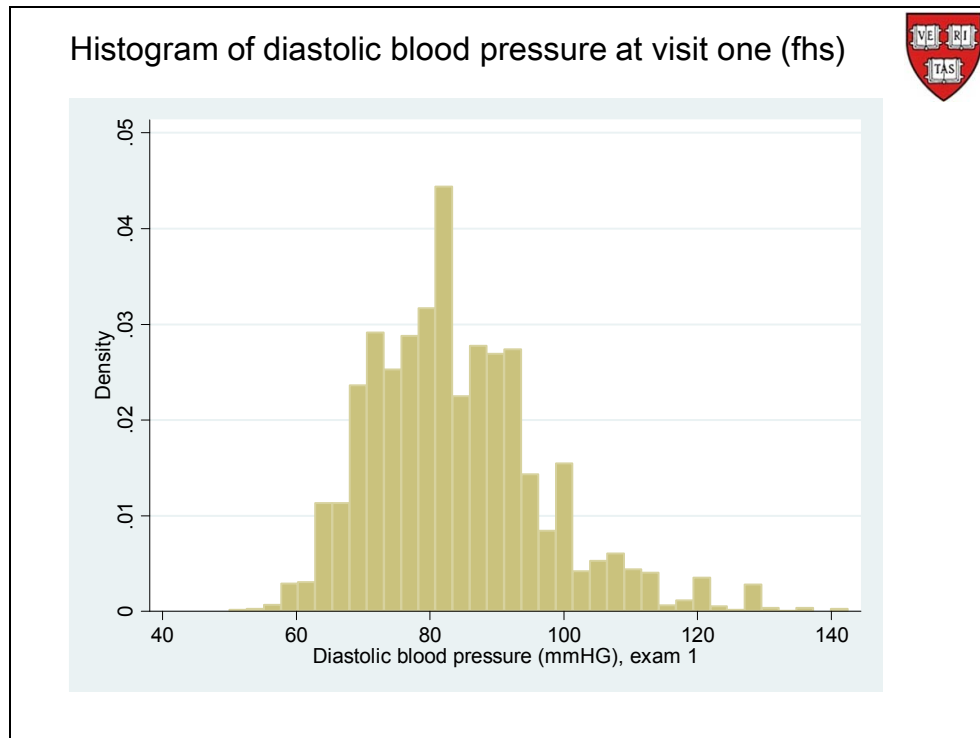
the label that we had for the categorical data-- like the blood types. But now these cells are contiguous.

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.

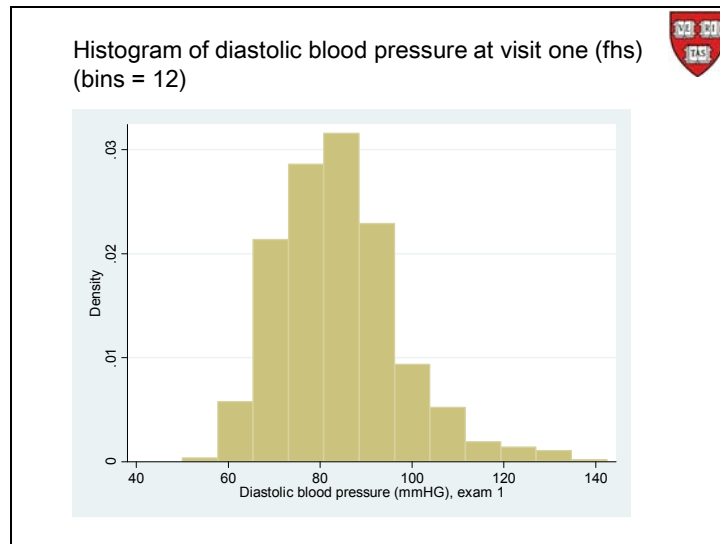
Each of these cells is contiguous to its neighbors, so you should put them closer together, and there should be no space between the bars. When you do that, you get what is called a histogram.



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.

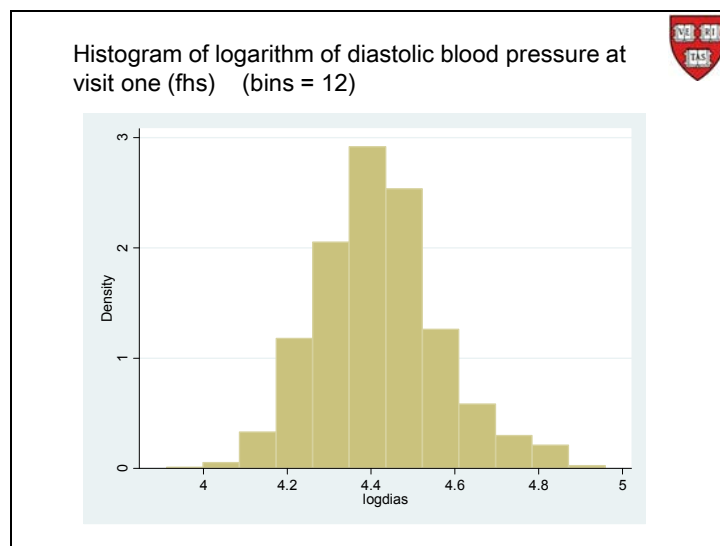
Now I just press the default buttons in Stata, and they chose the number of bins. To see the impact of the choice of the number of bins, try to see what happens on a famous data set,

<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.

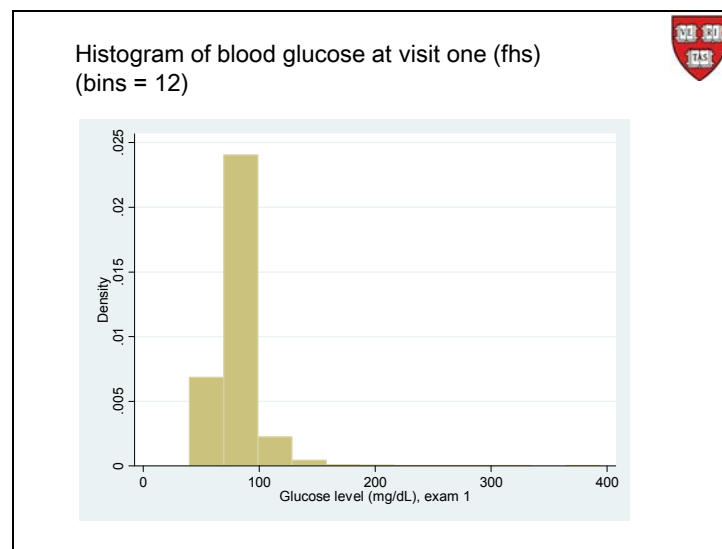
Putting that consideration aside for the moment, what we have here is a picture that almost looks symmetric about a central axis, except for an elongation on the right; what we call a long tail on the right. This kind of behavior is typical of data we shall see later in the course when we look at clinical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

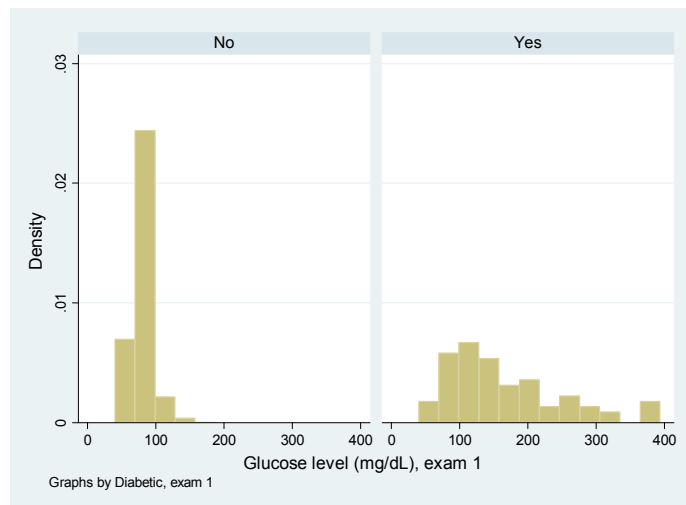
This is a common technique used in statistics: we do not just look at the raw data, but sometimes we transform the data, and here is one possible transform, namely, the logarithm.

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



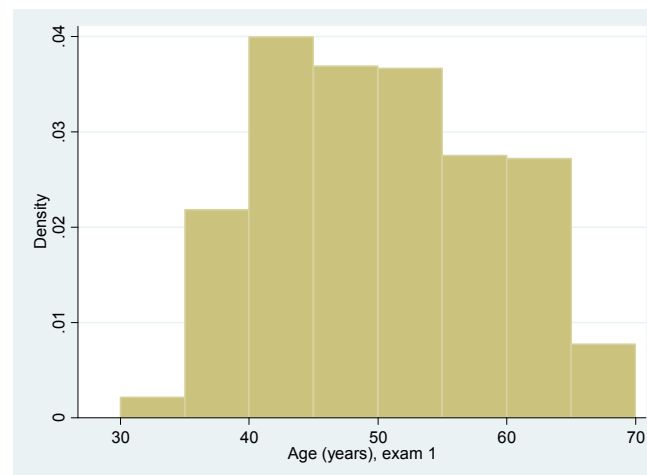
Here we have a histogram of the blood glucose at visit one in our Framingham heart study. And we see a right tail which is very much smaller, but it extends all the way to the right, almost to 400. Now the question is, what is happening at this right tail? Remember we are looking at the level of blood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

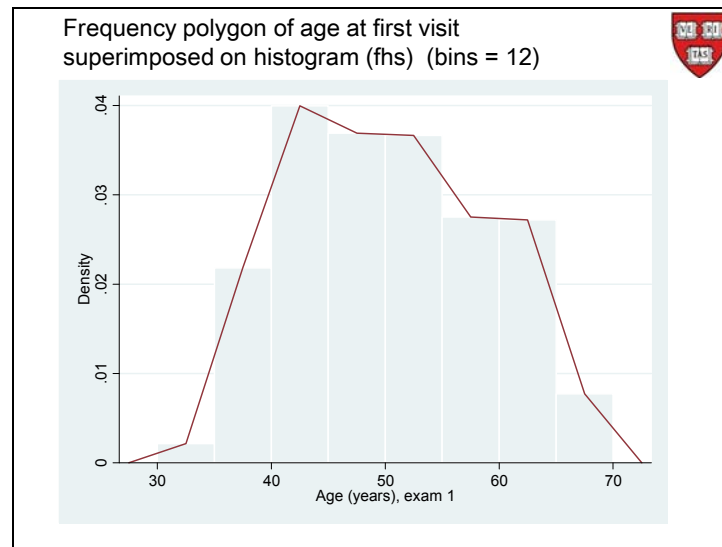


On the right we have the histogram for diabetics and on the left the histogram for the rest of the group. Now the highest value in the left group is 163, and there is a large amount of activity for the diabetic group on the right. This is now a very interesting part of the story, namely, what happens 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)

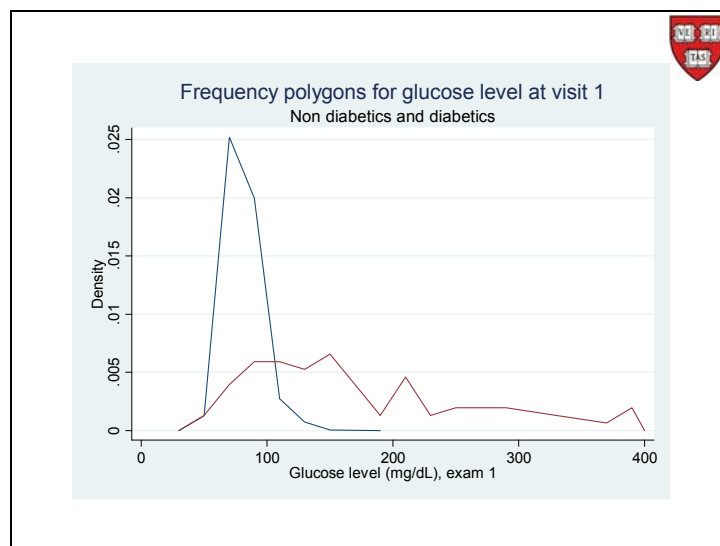
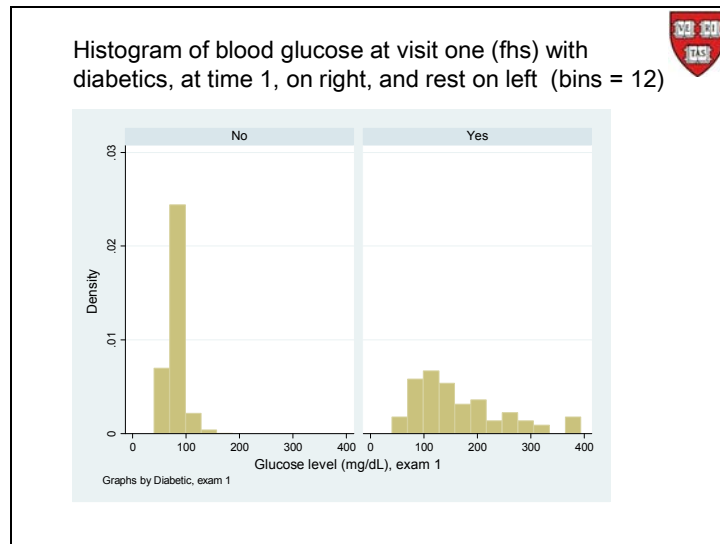


Here is another histogram, the histogram of the age of the participants at visit one.



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**



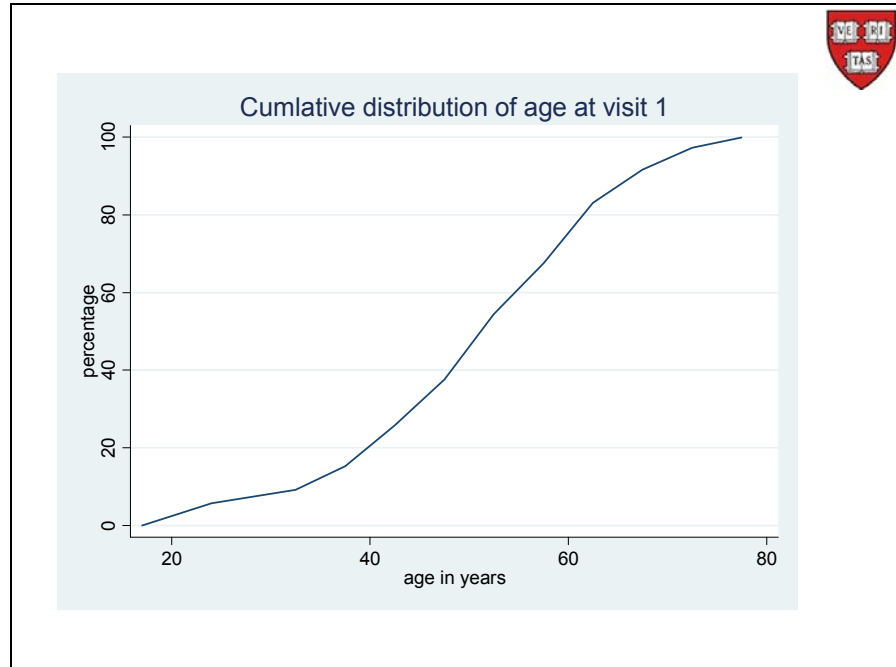
Age (Years)	Percent	Cumulative Percent
18–29	5.7	5.7
30–34	3.5	9.2
35–39	6.1	15.3
40–44	10.5	25.8
45–49	11.8	37.6
50–54	16.8	54.4
55–59	13.1	67.5
60–64	15.6	83.1
65–69	8.5	91.6
70–74	5.6	97.2
75–79	2.7	99.9

8

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.


⁸ <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.

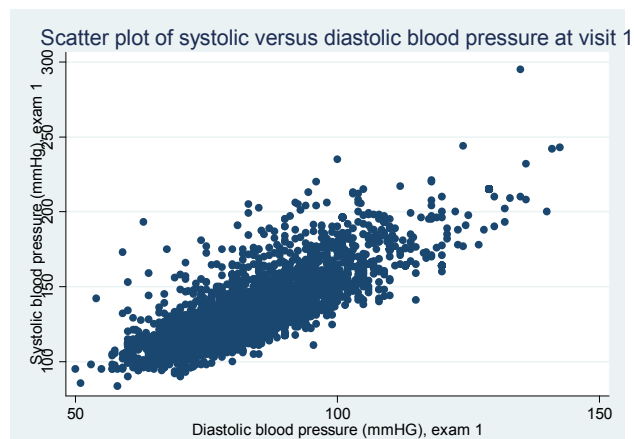
A frequency polygon conveys the same information as a histogram. It also lends itself to the depiction of the cumulative distribution.

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..

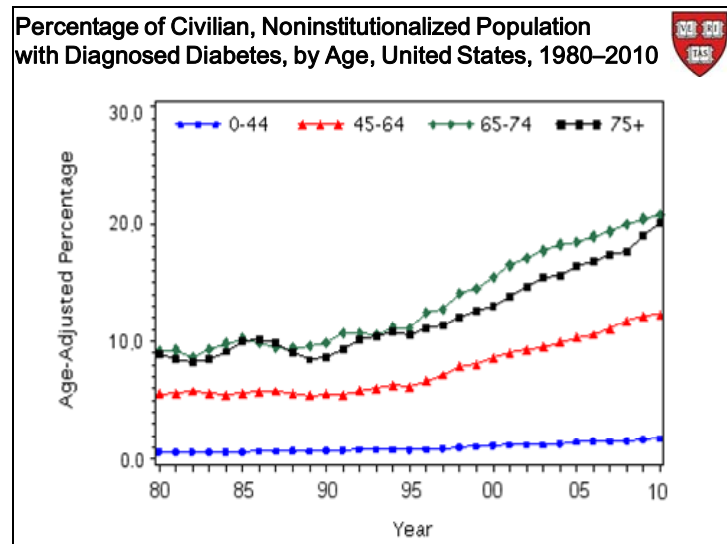
This is called a scatter plot. And you're going to see scatter plots repeatedly because they are very useful at revealing patterns or associations.

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

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.



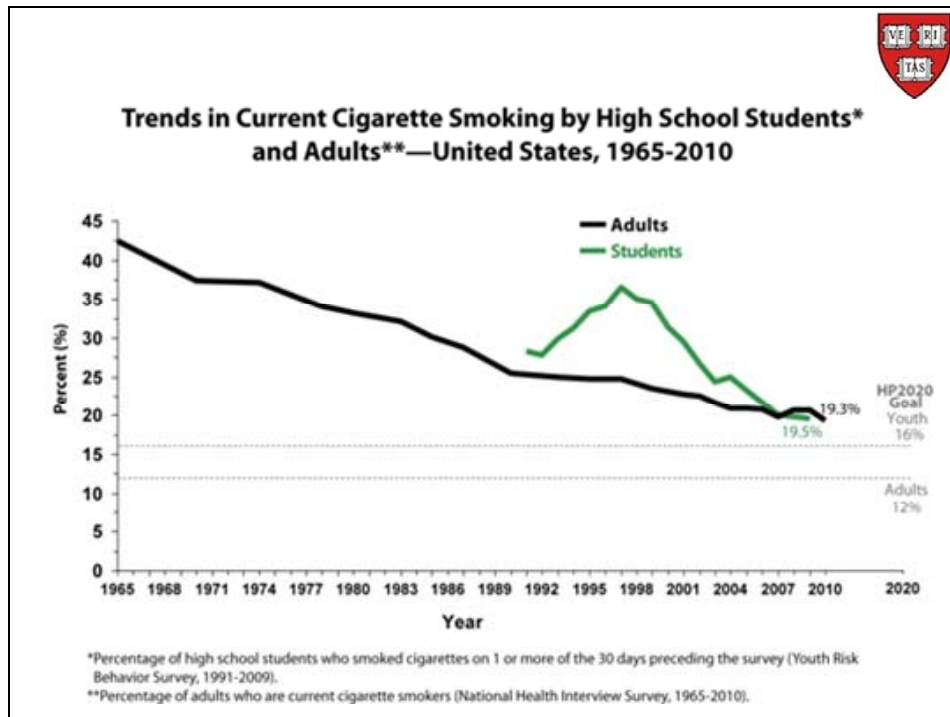
Whereas, if we draw a picture using line plots, we can appreciate many more subtleties in the movements. For example, we can see that the two older groups, the 65 to 74 and the 75 up, overlap each other, they are roughly equal to each other. And they both go up, first, at a gentle pace, and then, from 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 95 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.

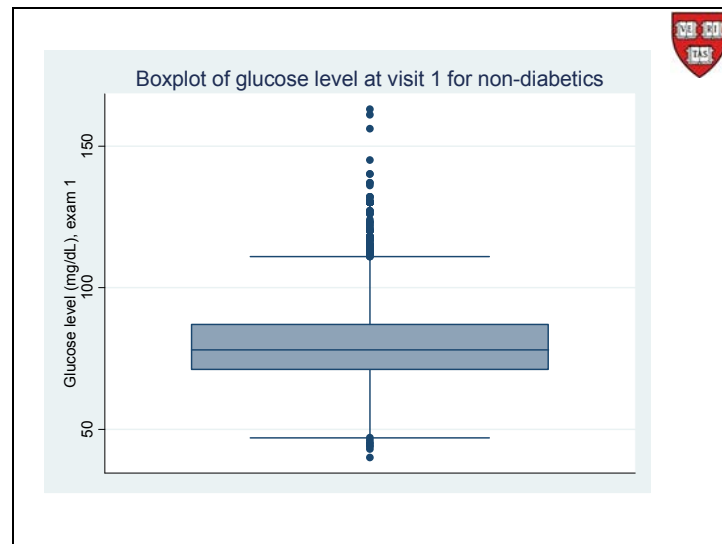
These percentages are really age adjusted percentages and not pure percentages. What does this mean? Rest assured it makes no difference in making the point that line plots are useful, but for the real meaning of what is going on here you will have to hang on to this course to find out what age adjustment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

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



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..

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.

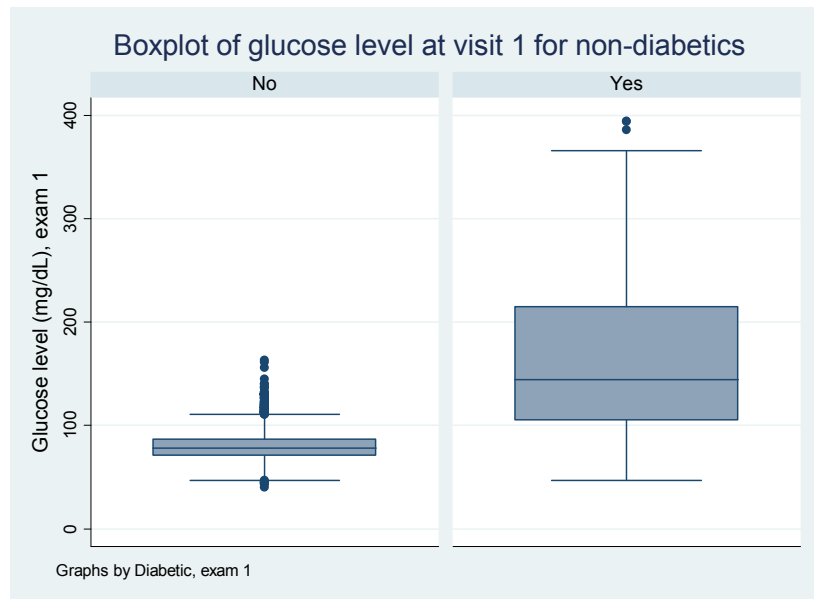
The distance between the top and the bottom of the box is called the interquartile range.

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.

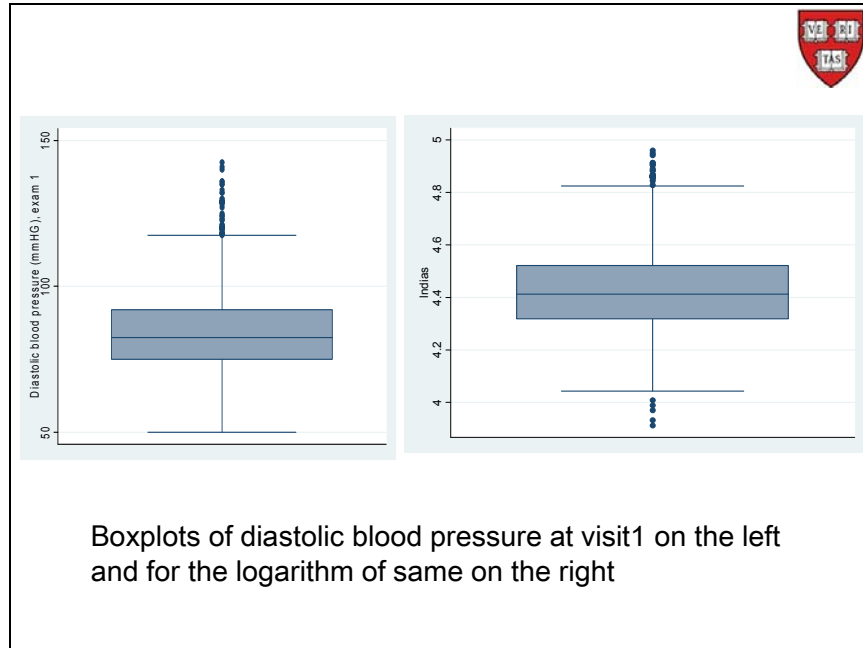


Here is another example, this time of side by side boxplots of two groups, the diabetics on the right and the rest on the left, again at visit 1, and looking at blood glucose level again.

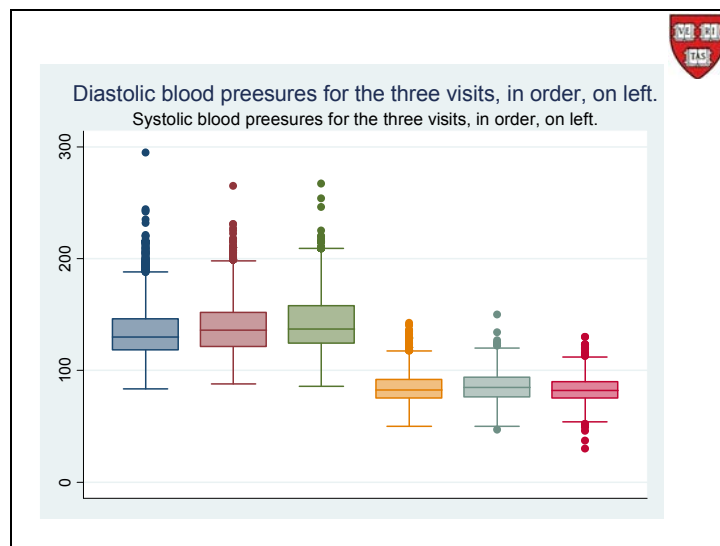
We see that the distribution for the non-diabetics is very tight. The interquartile range is very, very small, with a few outliers at the top (high blood glucose levels) who might be pre-diabetics.

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.



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?



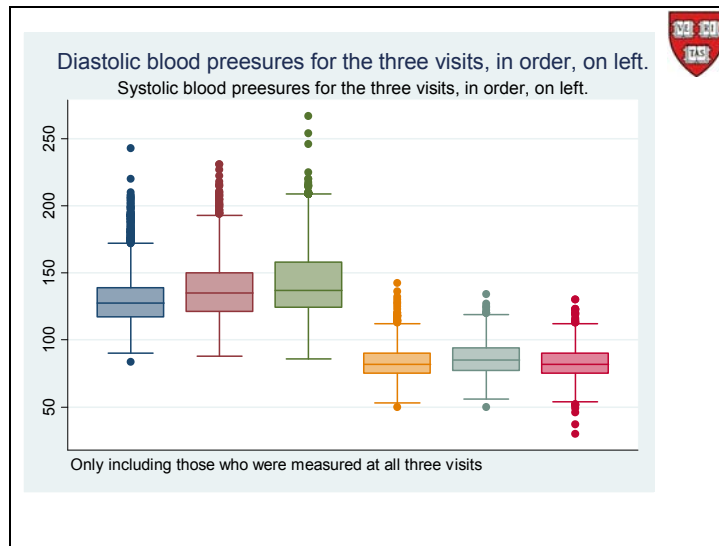
Box plots are also useful to place side by side over time. Here, the three on the left are the distributions of the systolic blood pressures for the three visits, and on the right the diastolic blood pressures for the same three visits.

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 amount 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.