

A decorative graphic in the top-left corner consisting of two red squares of different sizes and a thin vertical black line extending downwards from the bottom-left corner of the larger square.

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.



. sum age1 sex1 sysbp1 diabp1 cursmoke1 cigpday1 bmi1 diabetes1 hearttrtel glucose1

Variable	Obs	Mean
age1	4434	49.8392
sex1	4434	1.548489
sysbp1	4434	134.3718
diabp1	4434	84.09303
cursmoke1	4434	.4772215
cigpday1	4415	8.174858
bmi1	4417	25.99249
diabetes1	4434	.0252594
hearttrtel	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, **2.25**, 2.30 – median = 2.25

2.15, **2.25**, **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>



```
. summ age1 sex1 sysbp1 diabp1 cursmoke1 cigpday1 bmi1 diabetes1 hearttrtel 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
hearttrtel	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.)

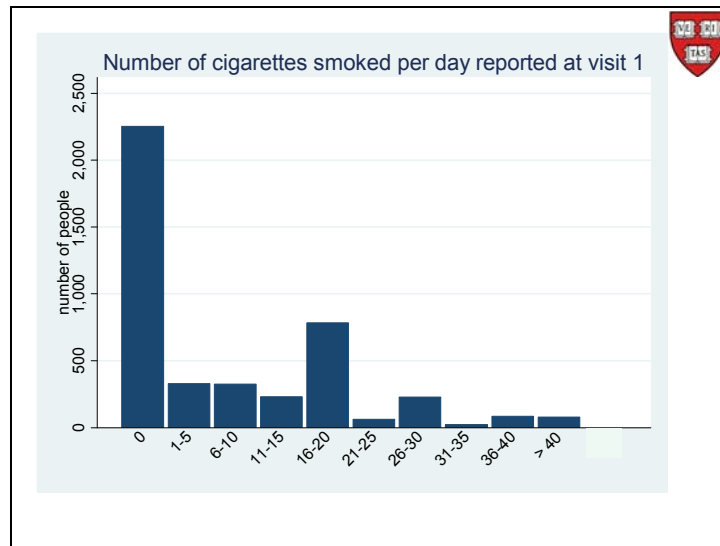
Look at sex. Remember this was 0, 1 variable. Look at what happens to the median of a 0, 1 variable. It just tells you which of the, two values is more popular. And we said that the value 2 was 54 percent, so the median is not as informative, it's not as nuanced, as the mean is in the case of dichotomous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



```

. sum age1 sex1 sysbp1 diabp1 cursmoke1 cigpday1 bmi1 diabetes1 hearttrte1 glucosel

```

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
hearttrte1	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



$X_i =$	1	1	1	1	5	5	10	10	10	25	Total
											= 69
$n_j =$											= 10

$$\frac{1}{10} \sum_{i=1}^{10} X_i = \frac{1}{10} \sum_{j=1}^4 n_j \bar{X}_j =$$

$$\frac{1}{10} (4 \times 1 + 2 \times 5 + 3 \times 10 + 1 \times 25) = \frac{69}{10}$$

$$= .4 \times 1 + .2 \times 5 + .3 \times 10 + .1 \times 25$$

$$= \sum_{j=1}^4 p_j \bar{X}_j \quad \text{where } p_j = \frac{n_j}{n}$$

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  $p_i$  (ns) is non-negative.
2. The  $p_i$  sum to one(1).



Note that all the  $X_i$  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  $1 \times 8$  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?  $\bar{x}$  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  $\bar{x}_j$ '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 percent  $\times$  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.

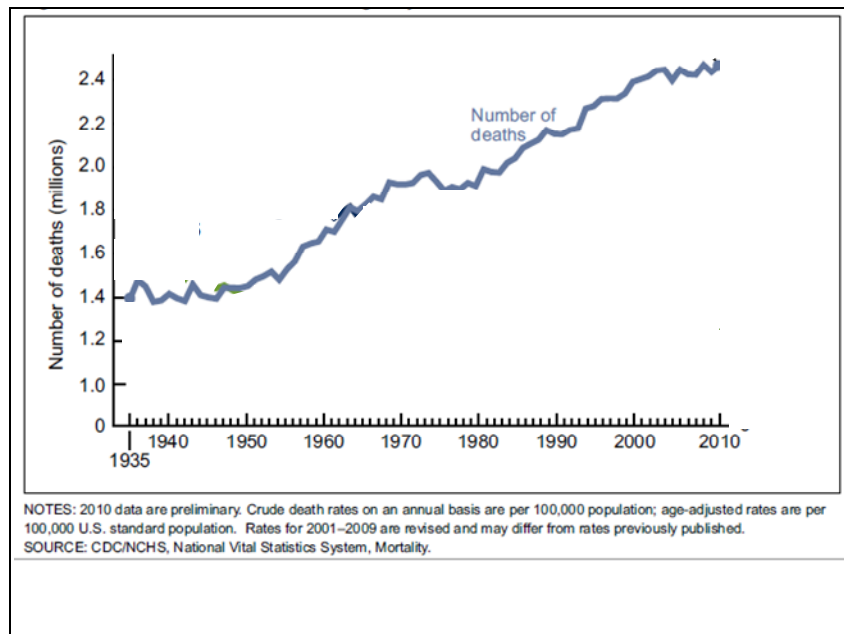
This gives rise to  
Index numbers, or  
Standardization methods

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.

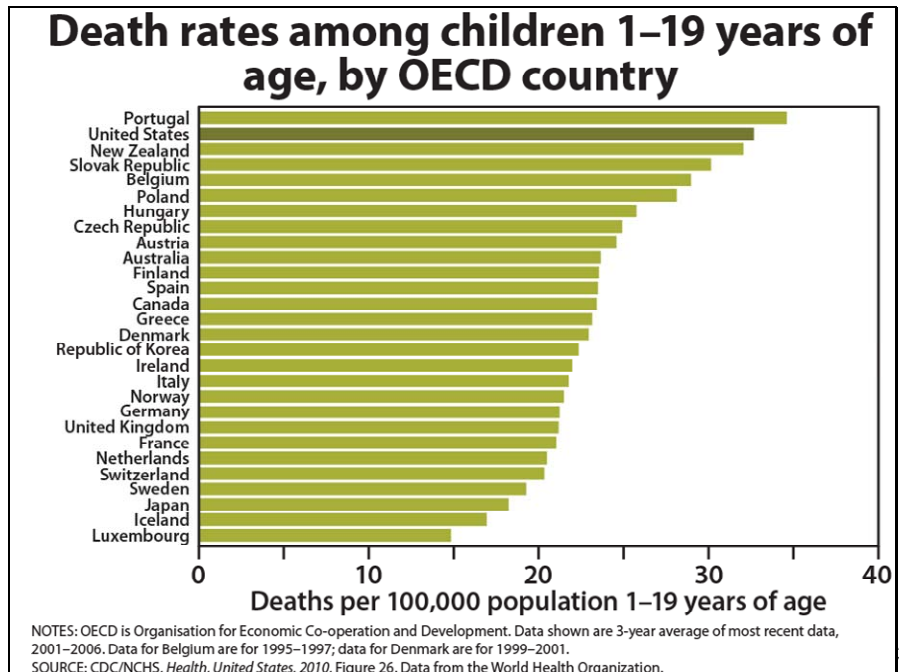
### Rate


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

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

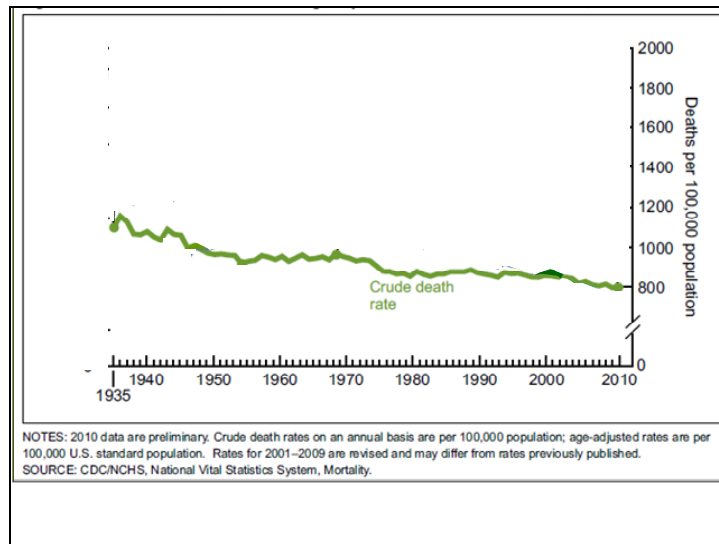
To overcome this problem, we can introduce a rate. A rate has a numerator, which could be the number of deaths, and a denominator, which could be the population size. Like that, we can correct for the population size as it gets bigger.

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



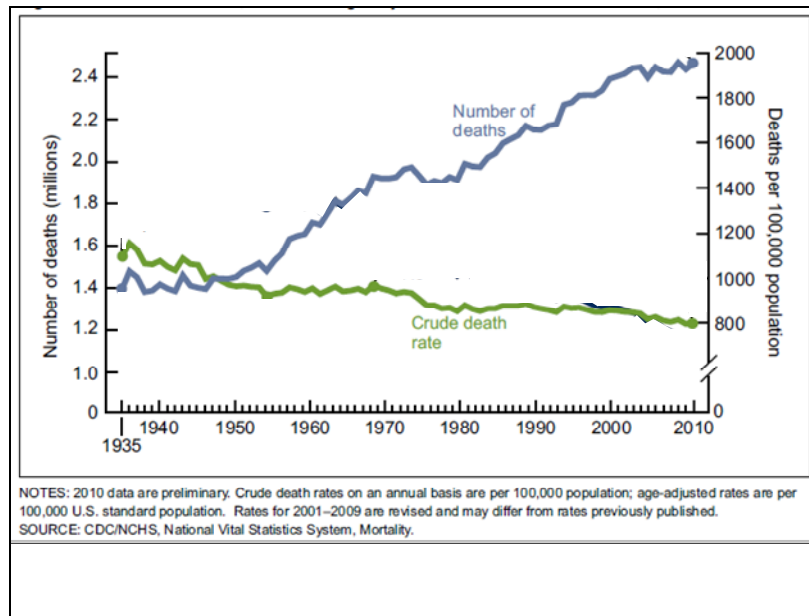
Now when we do that for the US, here's what we see. We see that the crude death rate, namely if we divide, each year, the number of deaths by the population size, we get something that looks like this. And voila, it's going down. And this makes us happy.

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.

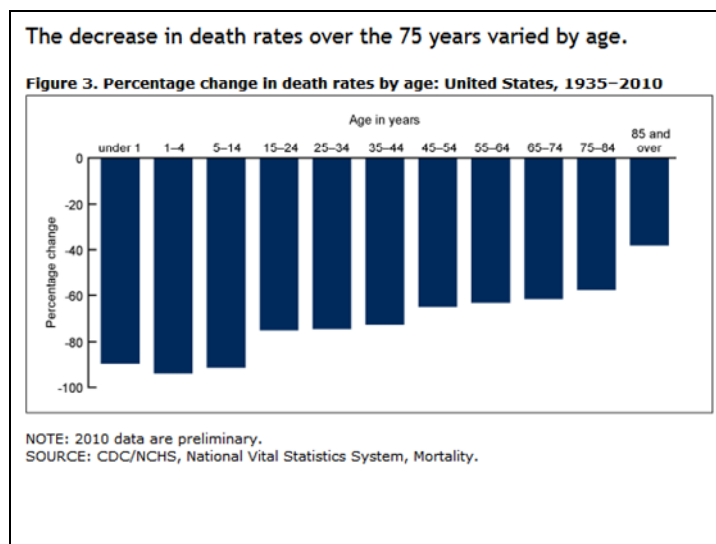
For example, if we look at the number of colds in a season, if we look at the cold rate--how many colds do you have per winter? You can have more than one. So the numerator would be the number of colds, and the denominator would be the population size. Thus the rate measures the number of recurring events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



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.

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

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

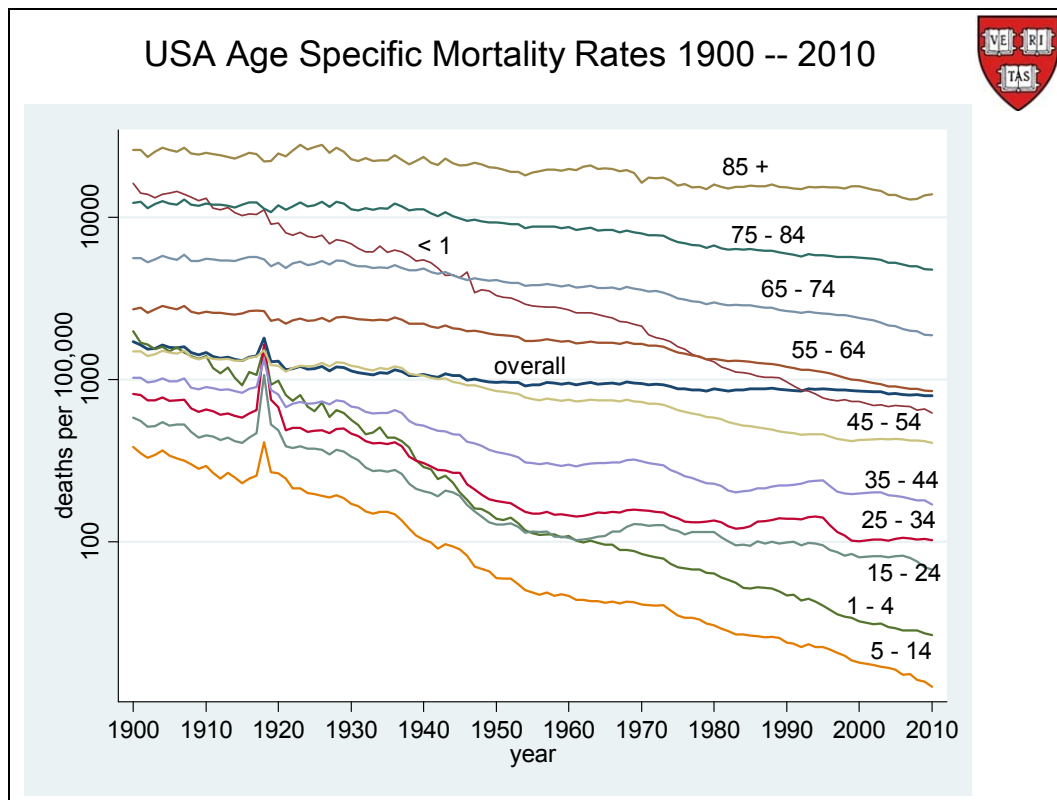
<http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-the-honor-roll>



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.

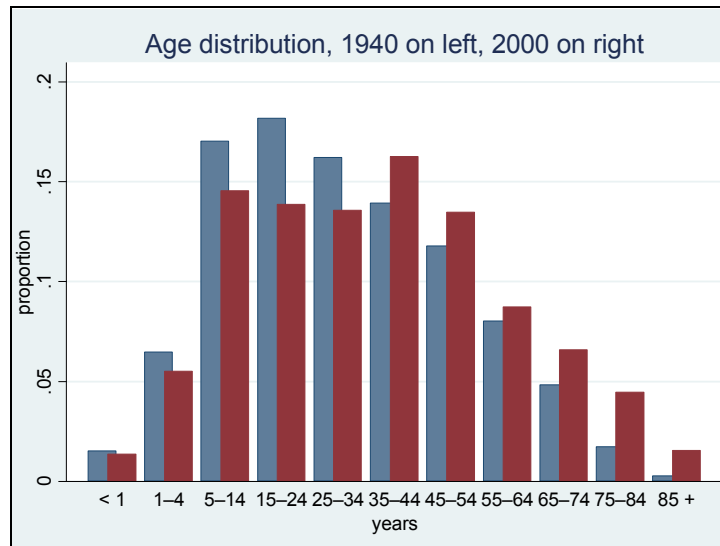
<sup>4</sup> <http://health.usnews.com/health-news/best-hospitals/articles/2012/07/16/best-hospitals-2012-13-the-honor-roll>



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.

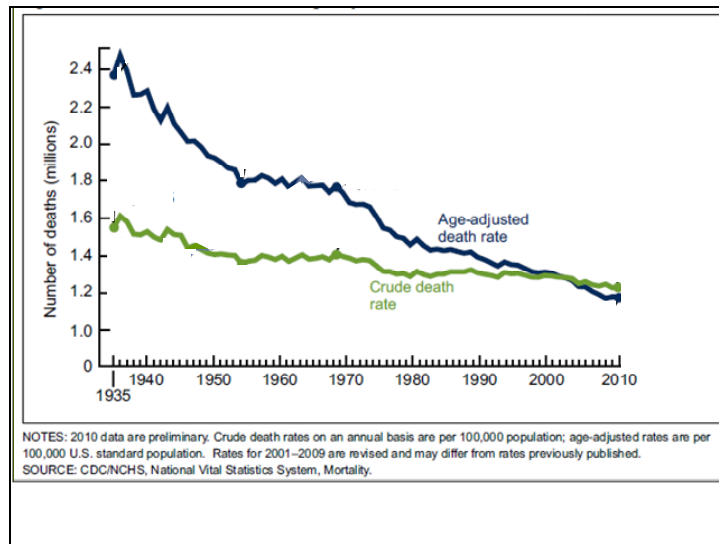
Before leaving this slide, I thought you'd be interested in the shapes of these rates. I took it back to 1900 to show you the effects of the 1917—1918 flu epidemic. What we see is that the spikes are not equally evident in all age groups, indeed, they are even lacking in some age groups. So the age groups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



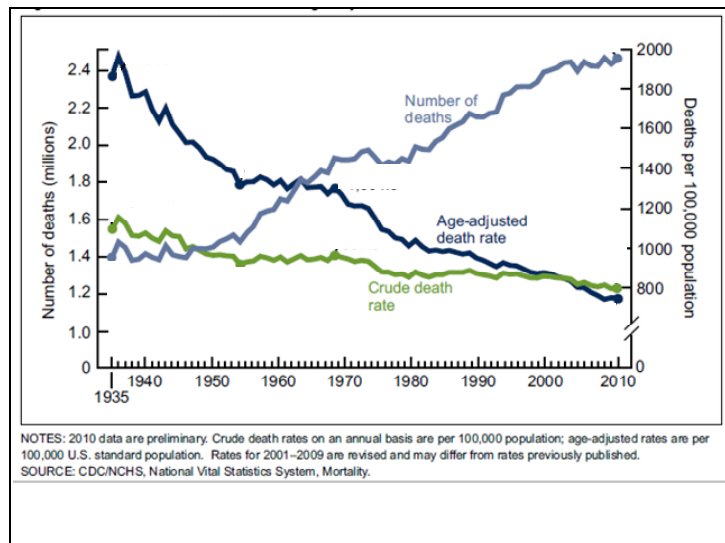
Returning to the problem at hand, if we think about the impact of time on our measurements, we realize that as our health improves, we, as a group, get older. As a result, we get penalized in the overall rate because, by and large, the older we are the higher the mortality rates we are subjected to. 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



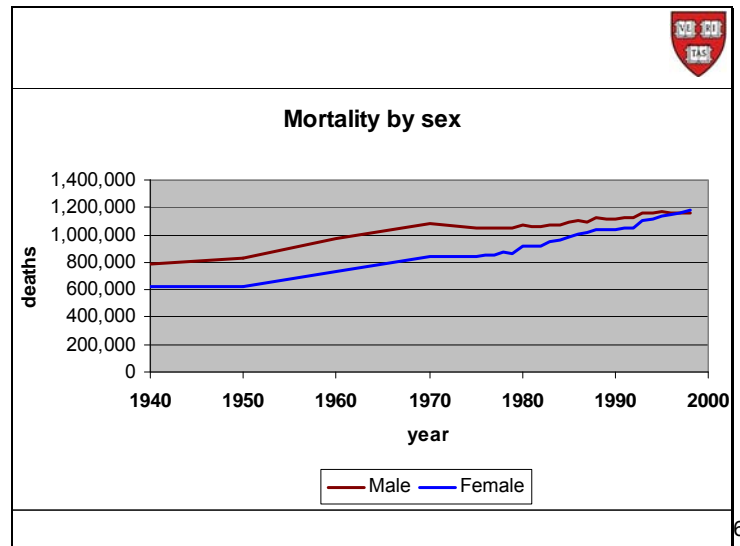
When comparing two composite means make sure we are comparing likes. If the composition (weights or proportions) changes then the comparison of means is less clear, and there is more confounding.

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.



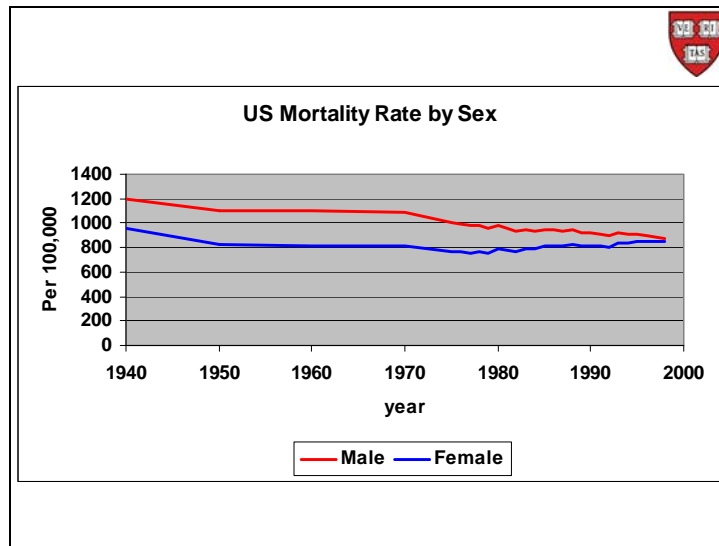


We must be cautious when we do age adjustment, if our intent is to compare two different groups. So for example, here is a sex breakdown of the crude mortality rate over the last 60 years in the USA. The top line is males and the bottom one is females.

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.

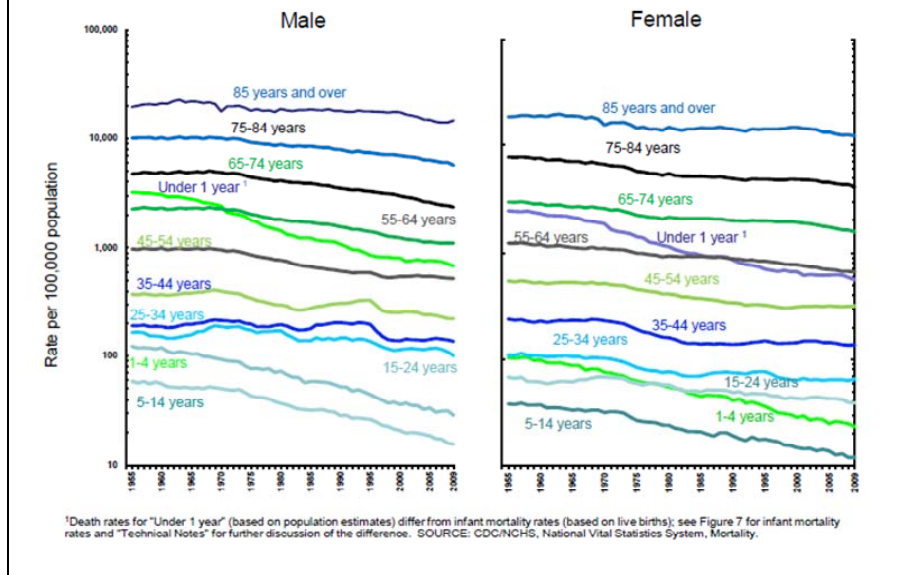
<sup>6</sup> [http://www.ssa.gov/oact/NOTES/as120/LifeTables\\_Body.html#wp1176553](http://www.ssa.gov/oact/NOTES/as120/LifeTables_Body.html#wp1176553)



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

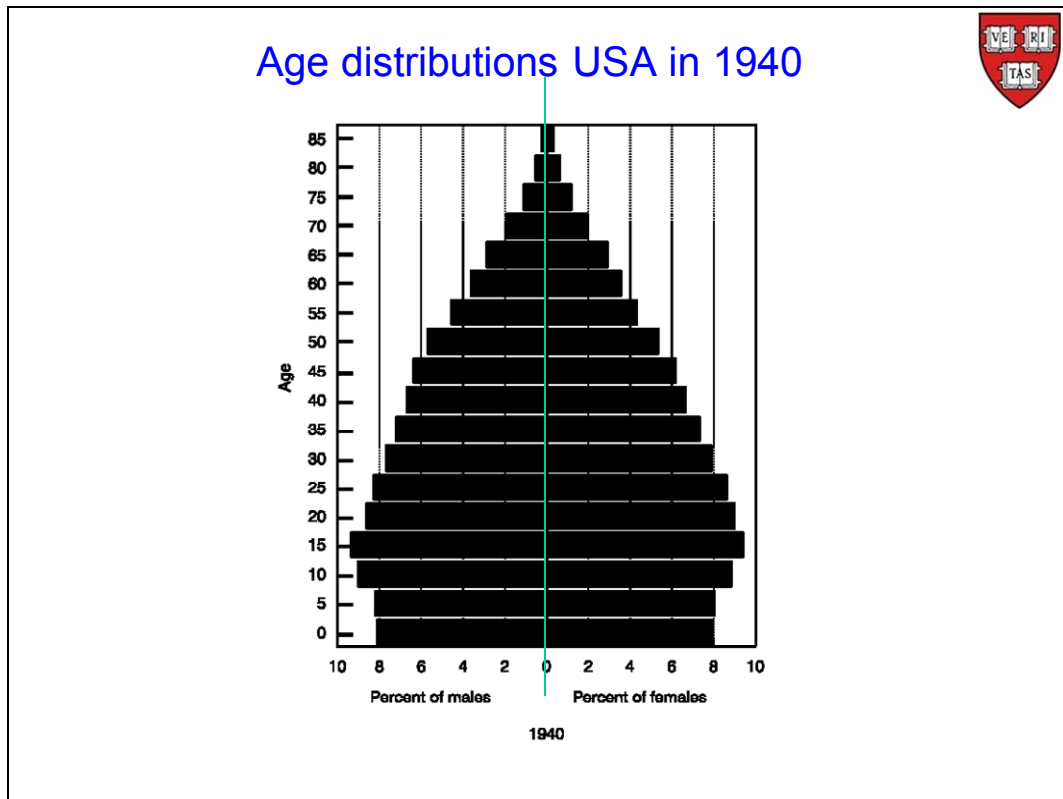
So now the crossover at the right end has disappeared, although we still see a closing up of the gap. Should we stop here?

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



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.

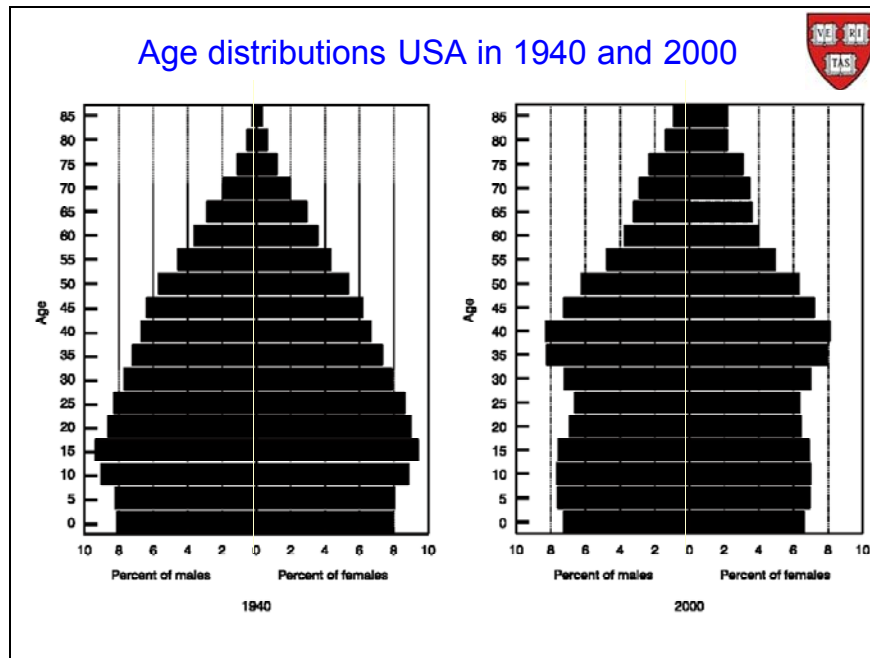
So it's going down, sure, but just as the population was getting older when the rates were going down for everybody, so too now we see that since they're going down differently and less for females, the females should be getting older but at a faster speed than the males.



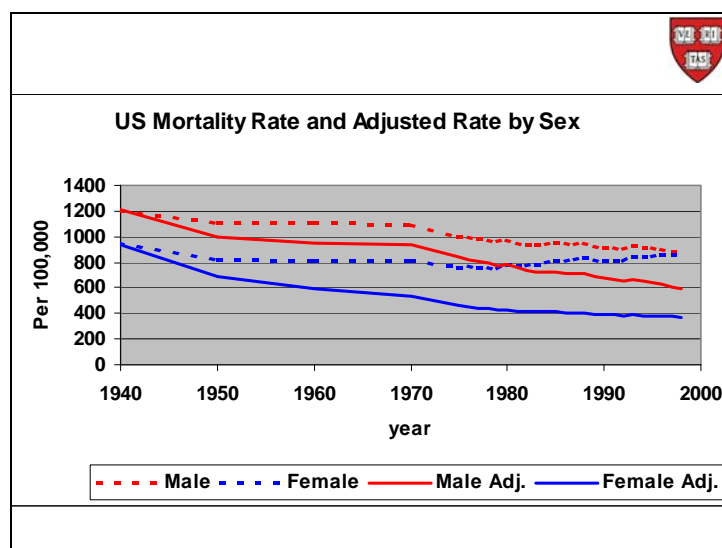
Look at the age pyramid. It is two vertical histograms—as opposed to the horizontal histograms you are accustomed to seeing—lying side-by-side. We see that in 1940, the vertical middle line at the zero point on the horizontal scale, demarcates males (on the left) from females (on the right). This line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

<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.

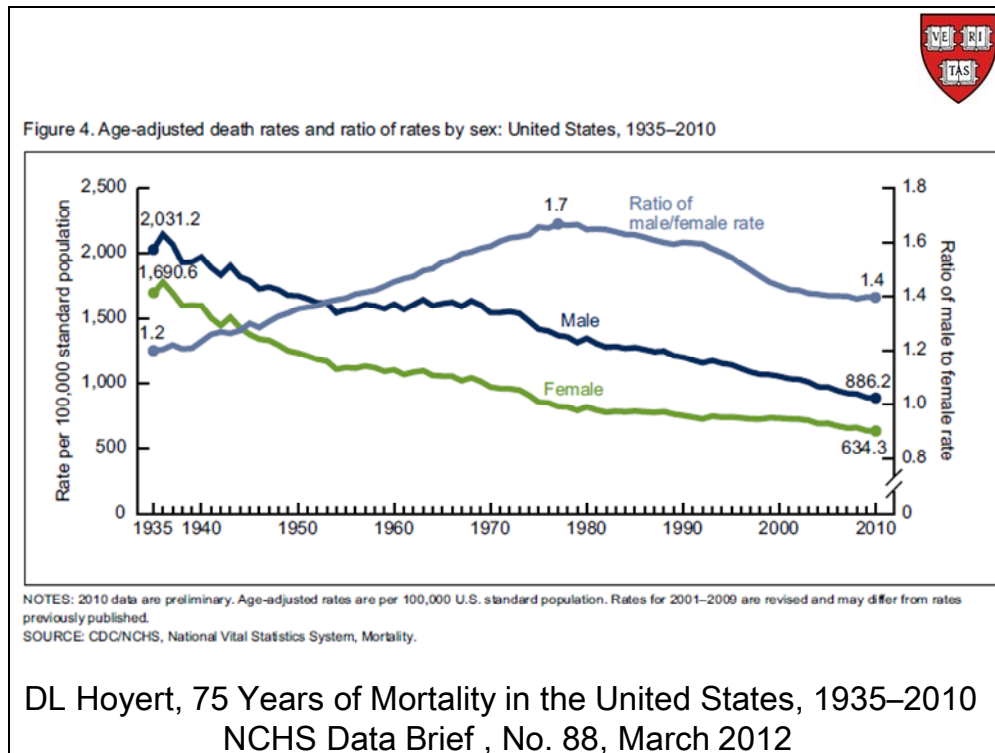
<sup>8</sup> Anderson RN, Rosenberg HM. Age Standardization of Death Rates: Implementation of the Year 2000 Standard. National vital statistics reports; vol 47 no. 3. Hyattsville, Maryland: National Center for Health Statistics. 1998.



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.

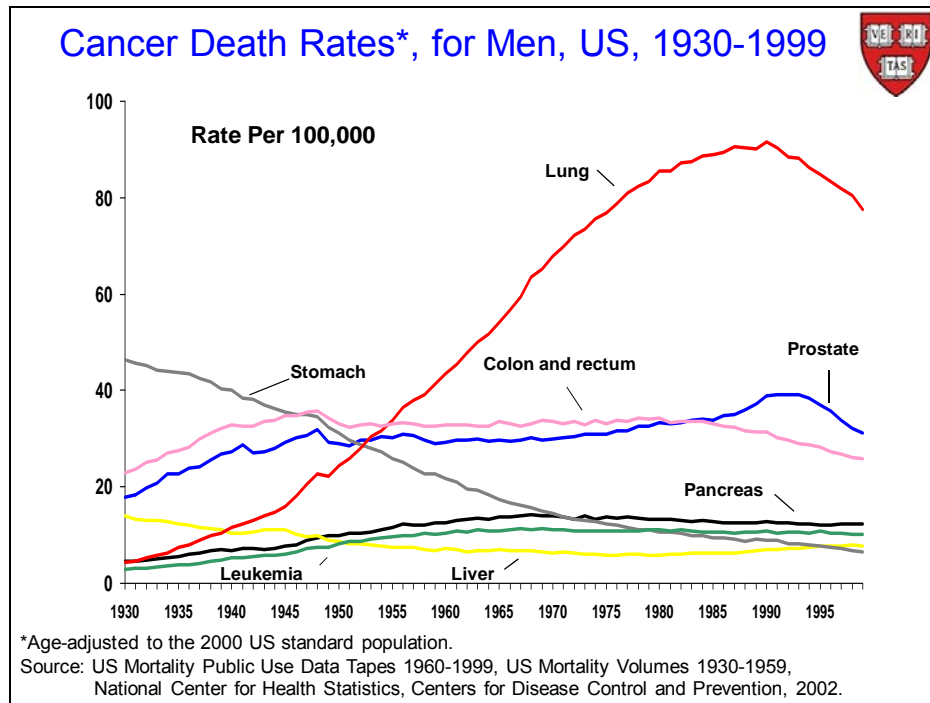


And indeed, that is what happens as evidenced by the graph above. You see this grouping together that we saw without the age adjustment has now disappeared. And the age adjustment shows that the gap is very much still there.



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.



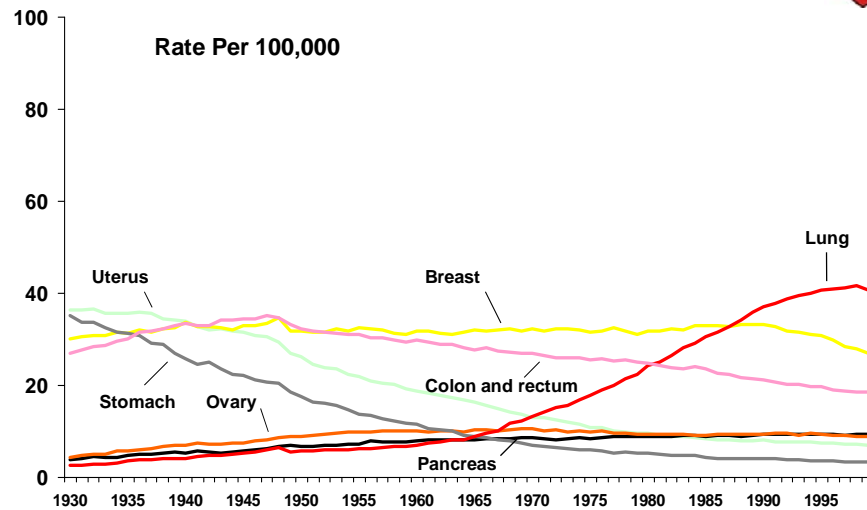


Once we do make age adjustment, then we can allow comparisons to be made over time in a meaningful way.

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

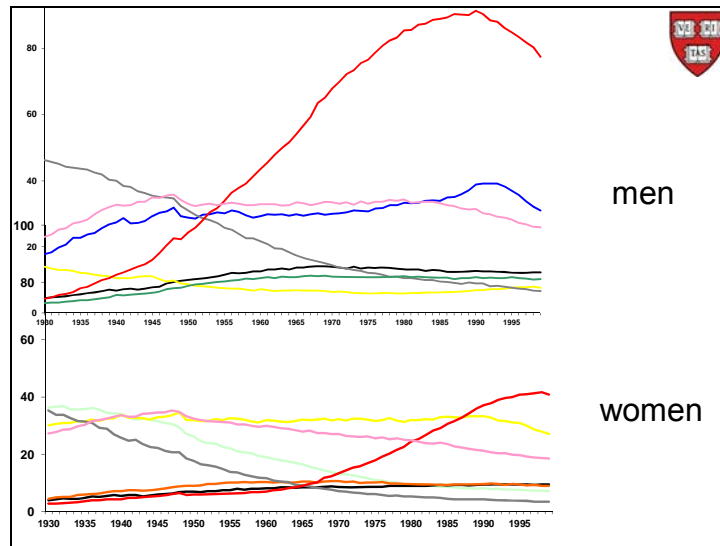


\*Age-adjusted to the 2000 US standard population.

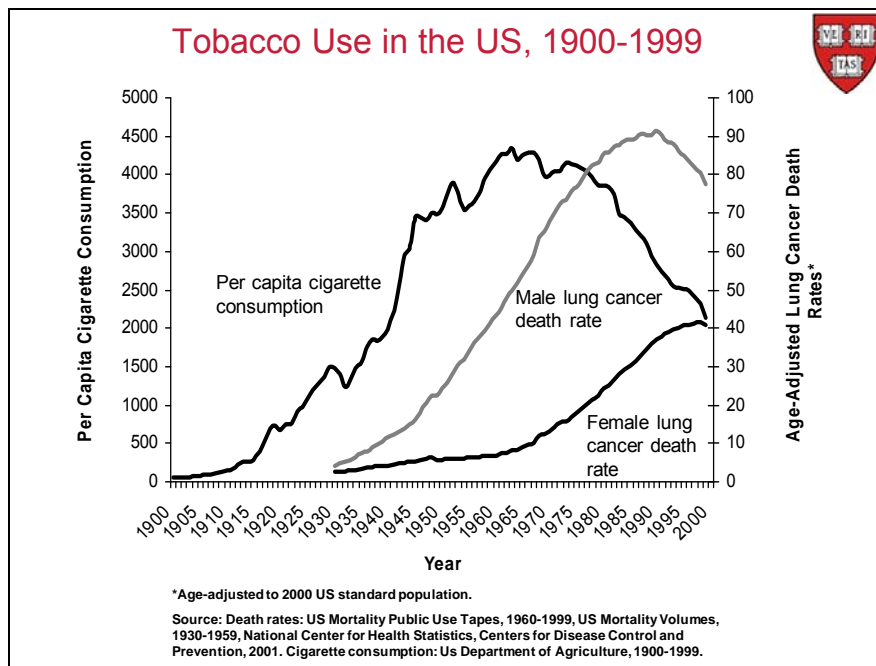
Source: US Mortality Public Use Data Tapes 1960-1999, US Mortality Volumes 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention, 2002.

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>	$(x_j - \bar{x})$	$(x_j - \bar{x})^2$
	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<sub>1</sub><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 *a/ways* 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> FEV<sub>1</sub> 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 hearttrtel glucosel
```

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
hearttrtel	4433	75.34446	12.14209	44	140
glucosel	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 camels 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 depictive 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

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

$$|t| \leq 1 \quad \text{approx 67\% observations}$$

$$|t| \leq 2 \quad \text{approx 95\% observations}$$

$$|t| \leq 3 \quad \text{approx all observations}$$



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					
. summ diabp1					
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	
± 1 std. devs : (-1 , 1)			67%	71%	
± 2 std. devs : (-2 , 2)			95%	95.99%	
± 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>Do not forget the assumptions.</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.