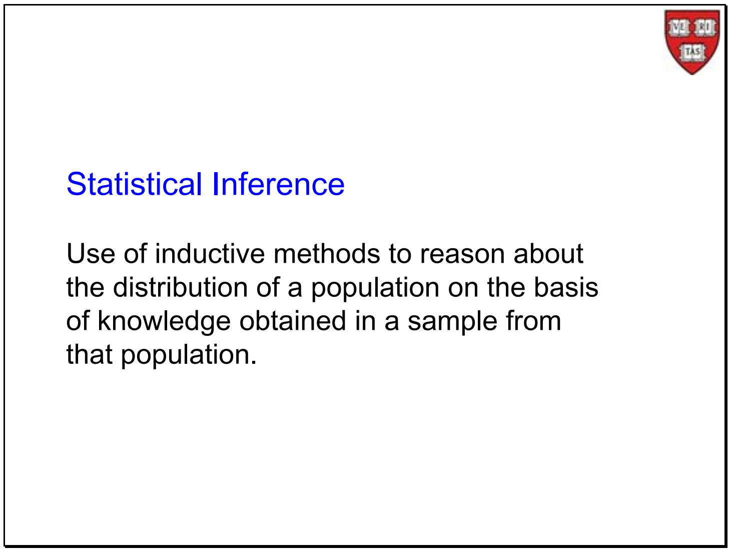
Marcello Pagano

[JOTTER-WEEK 5 SAMPLING DISTRIBUTIONS]

Central Limit Theorem, Confidence Intervals and Hypothesis Testing

Inference

This is when the magic starts happening.

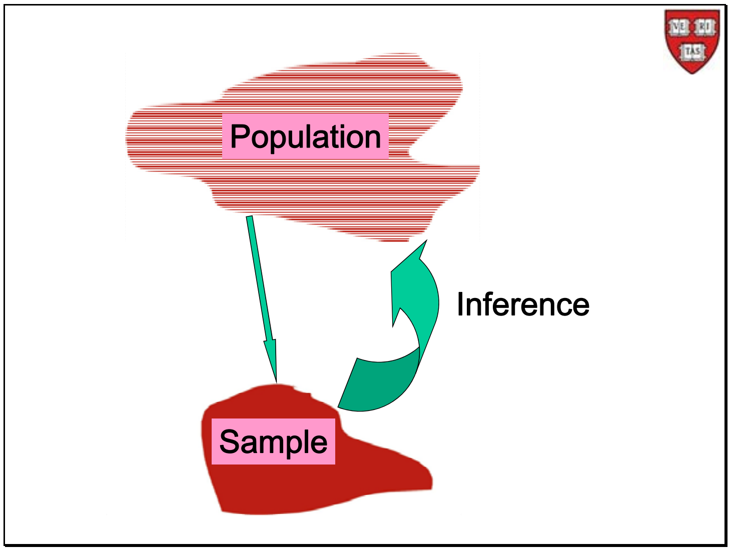


Statistical Inference

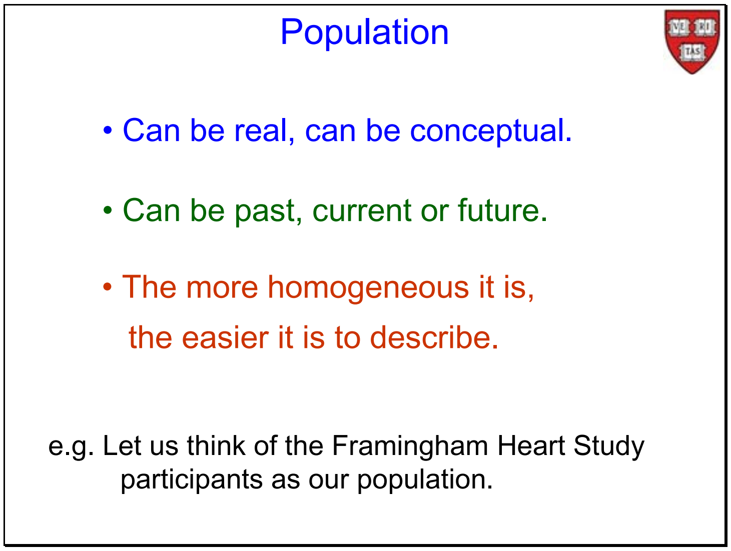
Use of inductive methods to reason about the distribution of a population on the basis of knowledge obtained in a sample from that population.

First, let us define inference the way we use it in this course: Inference is the use of inductive methods to reason about the distribution of a population. What does that mean?

Well, we are interested in a population and the distribution of certain variables within this population. For example, how does cholesterol level vary within this population? How does it vary for women? How does it vary for men? Is there a difference between the two sexes in their cholesterol distributions? We want to be able to answer such questions about the population, without having to measure the whole population. We are going to infer what these answers might be on knowledge we obtain in a sample, or a subset of that population.



Schematically, what we have is a population that is the focus of our interest. We would love to measure everything on everybody in this population, because then our task would be completed. But we cannot do that. So we take a sample from this population. Then on the basis of this sample, we make inference about the population. This inference is our challenge in this module.


Population

- Can be real, can be conceptual.
- Can be past, current or future.
- The more homogeneous it is, the easier it is to describe.

e.g. Let us think of the Framingham Heart Study participants as our population.

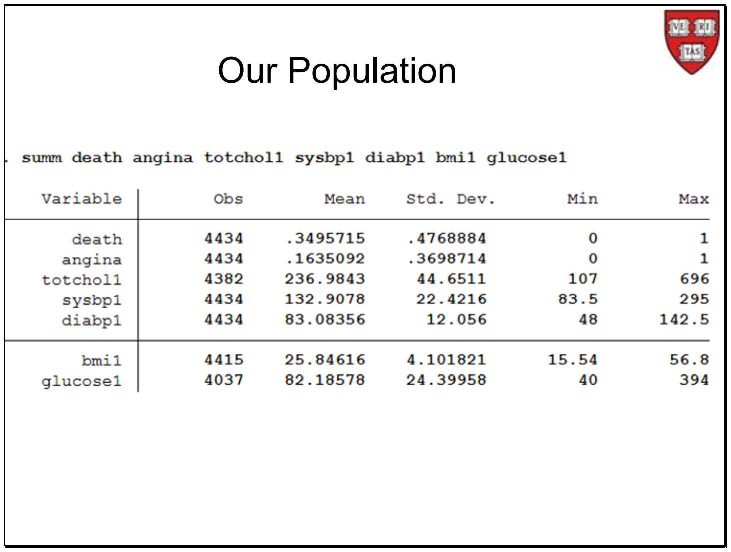
Let us look at these three components one by one. First we have the population. It can be real. For example, it could be everyone you know, everybody in your town, everybody in your province, in your state, in your nation; just a group of people.

The population can be real or it can be conceptual. It could be everybody in the future who is destined to get cancer and we might be interested in how we are going to treat them. Or it could be even more conceptual: we could be asking the question, what would happen if we treated everybody with this particular treatment? If instead of that treatment we choose another. Then what would happen to the patients? These “what if” conceptual exercises can help us decide what treatments would be best for future patients.

The population can be in the past—what happened in the 1917-18 flu pandemic? It can be current. It can be in the future, as mentioned.

The more homogeneous the population the easier it is going to be for us to describe or measure it. So, for example, if we were all four feet tall, then that not only would be very easy to describe, but we would only need a sample of size one—measure a single person's height—to tell us how tall we all are. So the more homogeneous it is, the easier it is to describe, and the smaller the sample we will need.

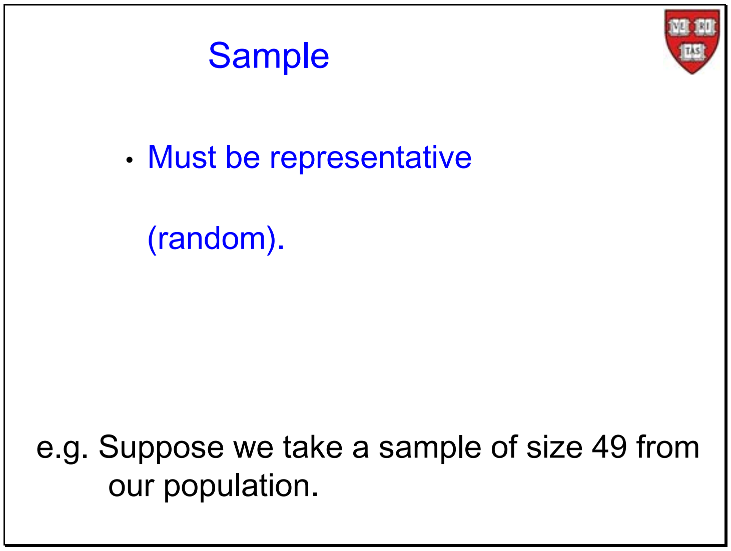
Now, what we are going to do henceforth as an *exercise* is to take the Framingham Heart Study data on all 4,434 patients and treat them as our population. Ordinarily, we do not know this much about our population, but as a pedagogical license, let us consider them as our population. In this manner we know what the "truth" is that we ordinarily would be trying to estimate.



Our Population

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4434    .3495715    .4768884          0          1
      angina |      4434    .1635092    .3698714          0          1
    totchol1 |      4382    236.9843     44.6511        107        696
      sysbp1 |      4434    132.9078     22.4216       83.5        295
      diabp1 |      4434    83.08356      12.056         48      142.5
-------------+--------------------------------------------------------
        bmi1 |      4415    25.84616    4.101821      15.54       56.8
    glucose1 |      4037    82.18578    24.39958         40        394
```

Here are the summaries of the first seven variables in our population. So, for example, approximately 35% of the 4,434 died. Roughly 16% experience angina. Looking at total cholesterol at the first visit, we see that the average was about 237. With this last variable we only have 4,382 measurements. For now we ignore the fact that some measurements are missing, but for the ones we have, they average out at about 237. Their standard deviation is about 44.6. And so on, for all the other variables in the data set. That is our population for now.



Sample

- Must be representative

  (random).

e.g. Suppose we take a sample of size 49 from our population.

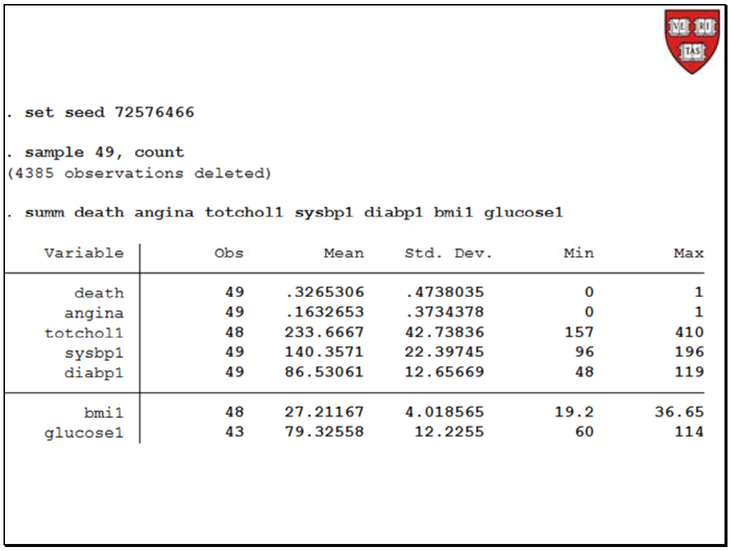
Second, take a sample from this population. If we want to make inference about the population on the basis of a sample from that population, then surely we would want our sample to be representative of the population. For example, if our population is 50% male, 50% percent female, we do not want a sample that's all male or all female especially if sex is an important consideration for the outcome that we are measuring. So we hope the sample is typical or representative of the population.

Now think about this logically a little bit, is it possible to have a truly representative sample. How do we know that the sample is representative? With respect to one variable, sex, it means that the sample should be 50% female, and thus 50% male.

Now what about age? Well we should have the same distribution of age in the sample as we have in the population; first amongst the females and then amongst the males. Then what about BMI? Well we need the same distribution of BMI amongst females within all the age groups, and amongst males within their age groups too. We must come to the conclusion that first, it is impossible to be truly representative if we literally mean that we have to have the same distribution on every single measure we can imagine. It is not possible unless the sample is as big as the population!

Number two, how would we know if we achieve representativeness in all dimensions? The only way we would know it is if we actually knew the population, in which case why are we measuring it? Why are we even taking a sample? So the answer to the question of whether our sample is truly representative, the answer is, we doubt it although we do not really know.

To overcome this problem we turn to a random device; we are going to take a random sample. What that means is that not everybody in the population will be in the sample, but everybody has an equal chance of being in the sample. And we do this because theory tells us that, *on average*, we will have a representative sample.



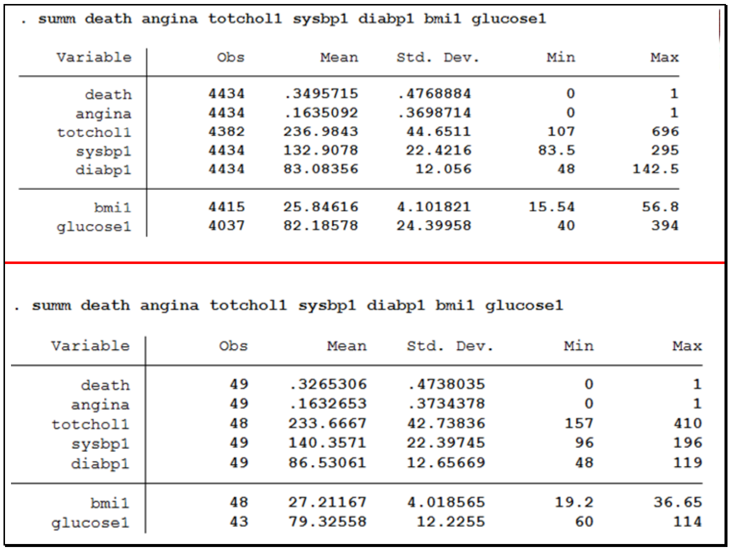
```
. set seed 72576466

. sample 49, count
(4385 observations deleted)

. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |        49    .3265306    .4738035          0          1
      angina |        49    .1632653    .3734378          0          1
    totchol1 |        48    233.6667    42.73836        157        410
      sysbp1 |        49    140.3571    22.39745         96        196
      diabp1 |        49    86.53061    12.65669         48        119
-------------+--------------------------------------------------------
        bmi1 |        48    27.21167    4.018565       19.2      36.65
    glucose1 |        43    79.32558     12.2255         60        114
```

So let us try it with our new “population”. Suppose we take a random sample size 49 from our population. To do so we can call on Stata.

Stata has a “random” device built in. It's not truly random. It's called a pseudo-random number generator. And what you do with a pseudo-random number generator is give it a seed; some random number—the manual suggests taking out a dollar bill from your pocket and looking at the number on the bill. The number of my dollar bill is G72576466A. I removed the G and the A, and that is the seed that you see above.

Having set the seed, the computer goes about generating numbers seemingly at random. These numbers are unpredictable to someone who does not know the inner workings of the computer (us), but every time I give the same seed it returns the same sequence of “random” numbers, and thus the label “pseudo”. This property we need in science so as to enforce the much valued reproducibility of our experiments. On the other hand, we cannot predict what they are going to be, unless we have seen them before. So we have achieved an oxymoron, reproducible random numbers.

So you ask Stata to choose 49 people at random from the population. Returning to the seven variables we had chosen above (the first seven), we have 33% deaths in our sample, whereas we had 35% deaths in the population. The percent with angina is 16% in the sample, and the mean of the total cholesterol level is 233 and so on.

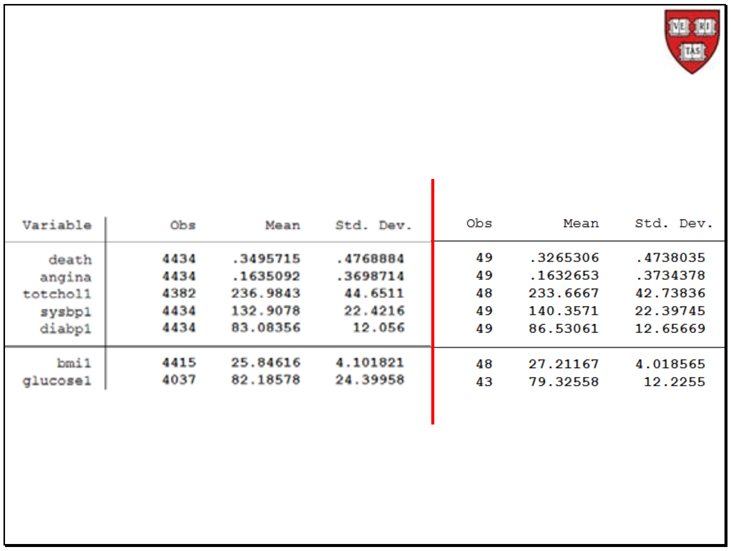


```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4434    .3495715    .4768884          0          1
      angina |      4434    .1635092    .3698714          0          1
    totchol1 |      4382    236.9843     44.6511        107        696
      sysbp1 |      4434    132.9078     22.4216       83.5        295
      diabp1 |      4434    83.08356      12.056         48      142.5
-------------+--------------------------------------------------------
        bmi1 |      4415    25.84616    4.101821      15.54       56.8
    glucose1 |      4037    82.18578    24.39958         40        394
```

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |        49    .3265306    .4738035          0          1
      angina |        49    .1632653    .3734378          0          1
    totchol1 |        48    233.6667    42.73836        157        410
      sysbp1 |        49    140.3571    22.39745         96        196
      diabp1 |        49    86.53061    12.65669         48        119
-------------+--------------------------------------------------------
        bmi1 |        48    27.21167    4.018565       19.2      36.65
    glucose1 |        43    79.32558     12.2255         60        114
```

So if we compare the summaries of these seven variables in our population of 4,434 to our sample of 49, we see that, by and large the sample reproduces the population summaries.

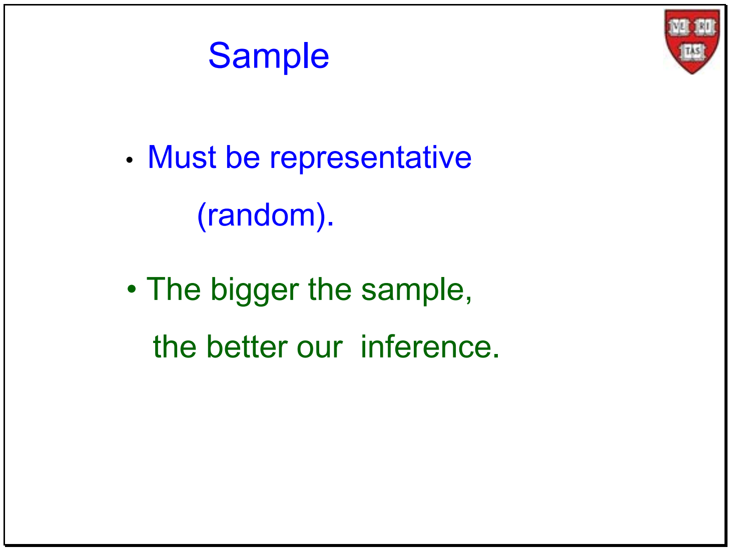
The range for total cholesterol level is from 107 to 696. The range in the sample, of course, has to be smaller. That is why we are forever breaking records in sports, et cetera.



```
    Variable |       Obs        Mean    Std. Dev. |  Obs        Mean    Std. Dev.
-------------+--------------------------------------+----------------------------
       death |      4434    .3495715    .4768884  |   49    .3265306    .4738035
      angina |      4434    .1635092    .3698714  |   49    .1632653    .3734378
    totchol1 |      4382    236.9843     44.6511  |   48    233.6667    42.73836
      sysbp1 |      4434    132.9078     22.4216  |   49    140.3571    22.39745
      diabp1 |      4434    83.08356      12.056  |   49    86.53061    12.65669
-------------+--------------------------------------+----------------------------
        bmi1 |      4415    25.84616    4.101821  |   48    27.21167    4.018565
    glucose1 |      4037    82.18578    24.39958  |   43    79.32558     12.2255
```

Putting the variables side by side we see that the sample means and the population means are quite close to each other. Indeed, except for the glucose1, where the the sample standard deviation is one half the population standard deviation, we can say that the sample standard deviations are good estimators of the population standard deviations.

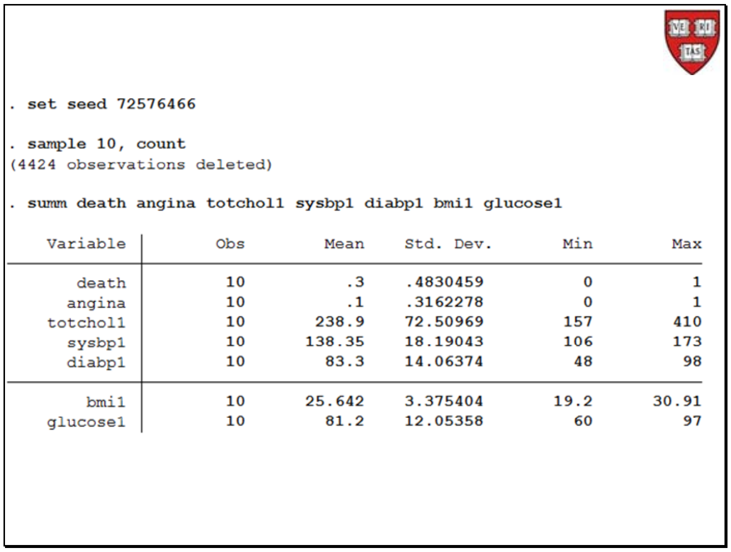
In general, you can see why it makes sense to use this sample to make inference about the population. The sample of size 49, turns out that that is a pretty big sample.



Sample

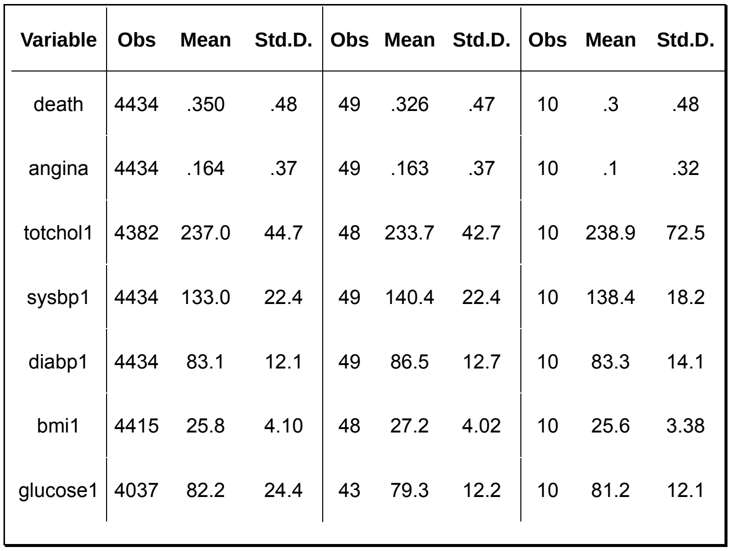
- Must be representative (random).
- The bigger the sample, the better our inference.

In fact, the bigger the sample, the better our inference. We'll quantify this a little better later, but for now, let us explore this issue.



```
. set seed 72576466

. sample 10, count
(4424 observations deleted)

. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |        10          .3    .4830459          0          1
      angina |        10          .1    .3162278          0          1
    totchol1 |        10       238.9    72.50969        157        410
      sysbp1 |        10      138.35    18.19043        106        173
      diabp1 |        10        83.3    14.06374         48         98
-------------+--------------------------------------------------------
        bmi1 |        10      25.642    3.375404       19.2      30.91
    glucose1 |        10        81.2    12.05358         60         97
```

What I did is I set the seed the same as before. Now this time I asked for a sample of size 10. And because I did this—you should not do this, since it introduces too much predictability. But I did this on purpose so that these 10 are actually the first 10 that went into making up the 49 in the previous sample.



Variable	Obs	Mean	Std.D.	Obs	Mean	Std.D.	Obs	Mean	Std.D.
death	4434	.350	.48	49	.326	.47	10	.3	.48
angina	4434	.164	.37	49	.163	.37	10	.1	.32
totchol1	4382	237.0	44.7	48	233.7	42.7	10	238.9	72.5
sysbp1	4434	133.0	22.4	49	140.4	22.4	10	138.4	18.2
diabp1	4434	83.1	12.1	49	86.5	12.7	10	83.3	14.1
bmi1	4415	25.8	4.10	48	27.2	4.02	10	25.6	3.38
glucose1	4037	82.2	24.4	43	79.3	12.2	10	81.2	12.1

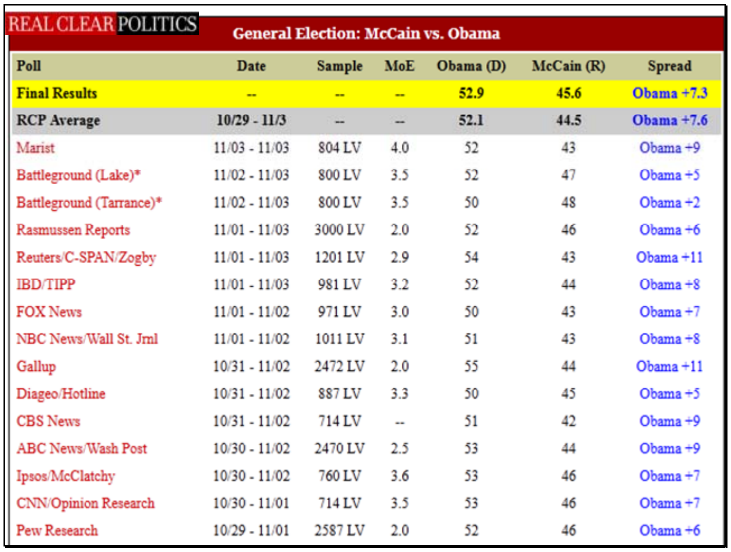
So here are the values. And once again, let me display them in such a way that our comparisons are visually easier to evaluate. And so here is the population. Deaths-- 35%. The first sample of 49 was 0.326. This one, of size 10, is 0.3. It's not as good as the earlier and bigger sample.

The prevalence of angina: 0.164, 0.163, 0.1. So that's not as good, either. The mean total cholesterol level, on the other hand, is better with the population value at 237, the sample of size 49 at 233.7, and the sample of size 10 at 238.9.

That is the problem with random samples, we cannot predict exactly how things are going to look!

We should also look at the sample standard deviations and see how they compare to the population values that they are estimating. In all instances except for the first and last variable (where they are approximately tied) the standard deviation for the sample of size 49 is closer to the “truth” than is the sample of size 10.

So things are random. But we shall see, in a moment, that as a general rule, a sample of size 49 is better than a sample of size 10. Now how big a sample do we need?



REAL CLEAR POLITICS General Election: McCain vs. Obama

Poll	Date	Sample	MoE	Obama (D)	McCain (R)	Spread
Final Results	--	--	--	**52.9**	**45.6**	**Obama +7.3**
RCP Average	**10/29 - 11/3**	--	--	**52.1**	**44.5**	**Obama +7.6**
Marist	11/03 - 11/03	804 LV	4.0	52	43	Obama +9
Battleground (Lake)*	11/02 - 11/03	800 LV	3.5	52	47	Obama +5
Battleground (Tarrance)*	11/02 - 11/03	800 LV	3.5	50	48	Obama +2
Rasmussen Reports	11/01 - 11/03	3000 LV	2.0	52	46	Obama +6
Reuters/C-SPAN/Zogby	11/01 - 11/03	1201 LV	2.9	54	43	Obama +11
IBD/TIPP	11/01 - 11/03	981 LV	3.2	52	44	Obama +8
FOX News	11/01 - 11/02	971 LV	3.0	50	43	Obama +7
NBC News/Wall St. Jrnl	11/01 - 11/02	1011 LV	3.1	51	43	Obama +8
Gallup	10/31 - 11/02	2472 LV	2.0	55	44	Obama +11
Diageo/Hotline	10/31 - 11/02	887 LV	3.3	50	45	Obama +5
CBS News	10/31 - 11/02	714 LV	--	51	42	Obama +9
ABC News/Wash Post	10/30 - 11/02	2470 LV	2.5	53	44	Obama +9
Ipsos/McClatchy	10/30 - 11/02	760 LV	3.6	53	46	Obama +7
CNN/Opinion Research	10/30 - 11/01	714 LV	3.5	53	46	Obama +7
Pew Research	10/29 - 11/01	2587 LV	2.0	52	46	Obama +6

One counter intuitive result is that, unless we are talking about small populations where the sample is a sizable fraction of the population, how big a sample one needs does not depend on the size of the population. Case in point, here are the results from a Real Clear Politics of the prior US Presidential election that took place on the 4th of November 2008[1]. Also shown are the results of the last polls that were taken immediately before the election.

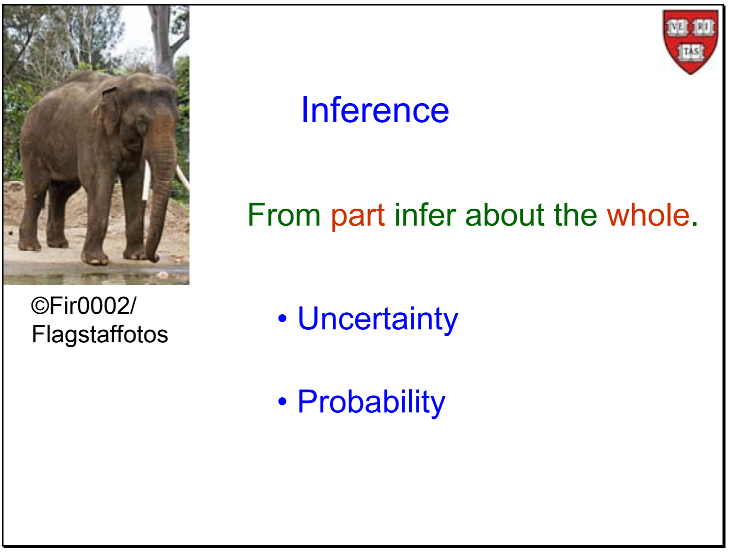
The final results had President Obama getting 52.9% of the vote. And this Marist poll predicted that he would get 52%. McCain actually got 45.6%. And they predicted 43%.

What Real Clear Politics did is also report the average of all those polls and came up with 52.1 for Obama and 44.5 for McCain. I do not think that you can get much closer than that.

What I find amazing is that close to 100 million people voted, and yet each of these polls was based on some number between one and three thousand people. You can predict how 100 million people are going to vote on the basis of what 1,000 people say?

That is the magic. This is the magic of random samples.

[1] http://www.realclearpolitics.com/epolls/2008/president/national.html



©Fir0002/
Flagstaffotos

Inference

From part infer about the whole.

- Uncertainty
- Probability

2

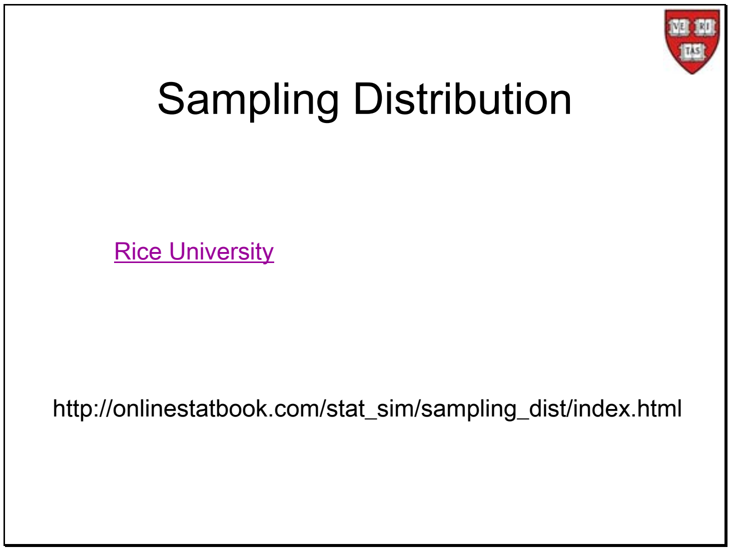
Having looked at the population and the sample, let us now look at the third component, inference. To repeat, we want to infer about the whole, the population, on the basis of a random sample from the population. And we want to do it in a principled way.

Our primary challenge is how to deal with the uncertainty inherent in what we wish to do. We are only privy to a small part of the whole population, and yet we have to generalize to the whole. We do not want to do like the Hindu parable of the blind men feeling the elephant, and depending upon what part of the elephant they were feeling, they projected or they inferred about the rest of the elephant, all leading to quite different pictures.

We now introduce probability into the argument to measure the uncertainty.

[2] http://en.wikipedia.org/wiki/File:Asian_elephant_-_melbourne_zoo.jpg

Sampling Distribution of the Mean

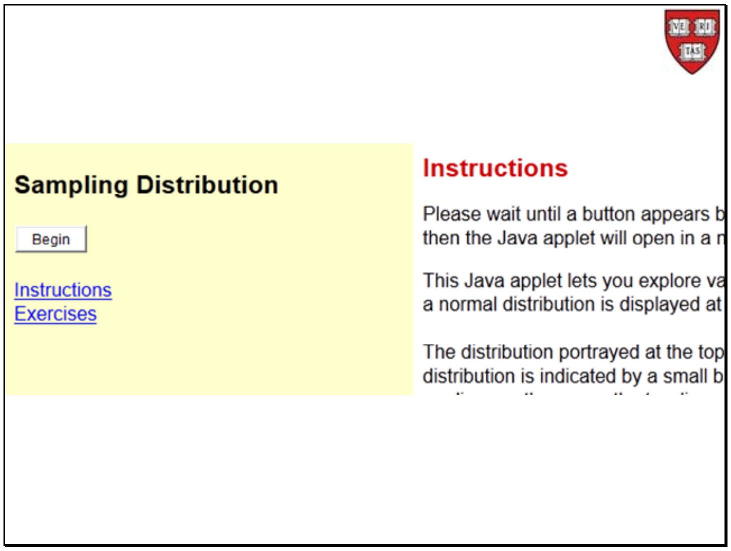


Sampling Distribution

Rice University

http://onlinestatbook.com/stat_sim/sampling_dist/index.html

We first introduce the concept of a sampling distribution. It will provide us the vehicle to quantify the uncertainty in our inference. And the way we do that is to go to the above website which is run by the statistics department at Rice University.



Sampling Distribution

Instructions
Exercises

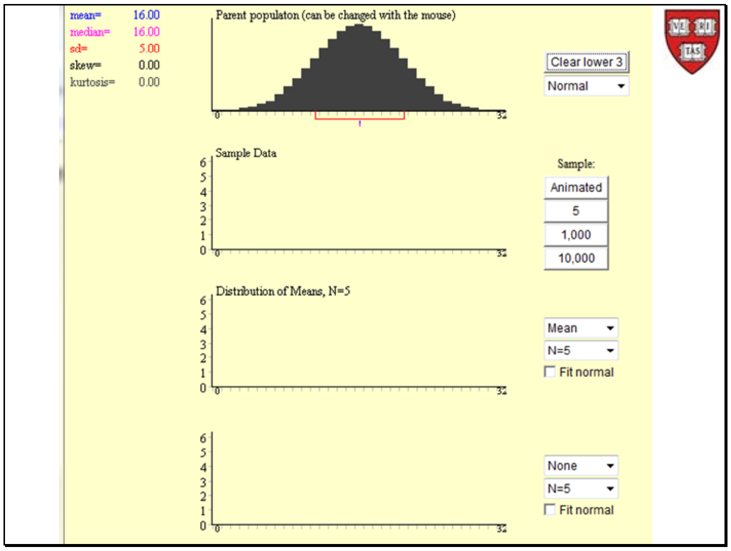
Instructions

Please wait until a button appears b
then the Java applet will open in a n

This Java applet lets you explore va
a normal distribution is displayed at

The distribution portrayed at the top
distribution is indicated by a small b

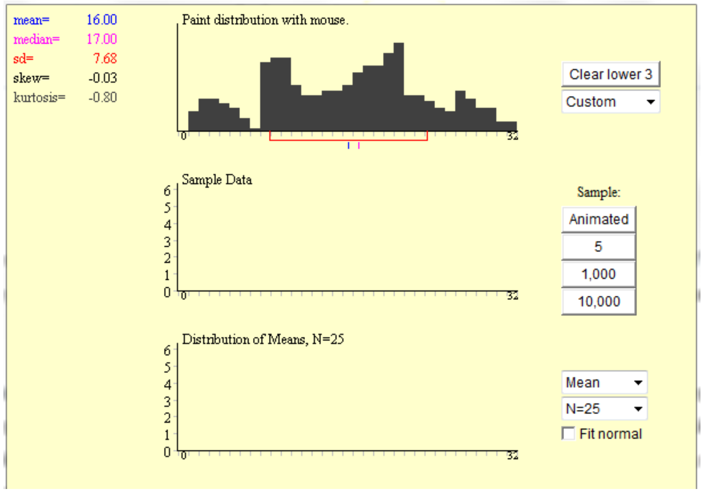
When you get there you click on “Begin”.



mean= 16.00
median= 16.00
sd= 5.00
skew= 0.00
kurtosis= 0.00

So we have before us a population. And that is the top panel here. We are going to take a sample from this population. And you will see the sample displayed in the second panel down. The third panel will keep a running summary of what we see in our samples because we're going to do the sampling repeatedly. In reality, of course, we only take a single sample most of the time. But for now, to get this concept of a sampling distribution under our belts, let's think of taking repeated samples, and pay attention to the summary.

On the left at the top we see summaries about this population. For example, we see that the mean is 16 for this population, the median is 16, and the standard deviation is 5. We also see the skewness and kurtosis, but I am not going to be addressing them. I leave that up to you to look up, if you are interested. Also, for now, we ignore the bottom panel.

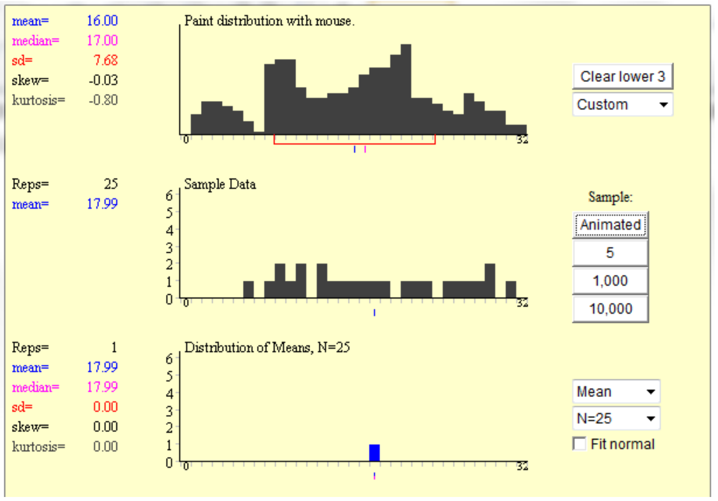


mean= 16.00
median= 17.00
sd= 7.68
skew= -0.03
kurtosis= -0.80

Let us start making choices. There are a number of statistics we can choose in the bottom-right corner. Let us stay with the mean. We have to decide on a sample size we want. Let us say N=25.

Now, let us look at the population that we have up here. We can choose the normal, or the uniform, or a skewed population, or we can make up our own custom population. Let us do that. Let us just make up a custom population.

So here is the distribution of the population. I just made it up. I just love to do it with this and things work nicely. So this population has a mean of 16, a median of 17, and a standard deviation of 7.68.



mean= 16.00
median= 17.00
sd= 7.68
skew= -0.03
kurtosis= -0.80

Reps= 25
mean= 17.99

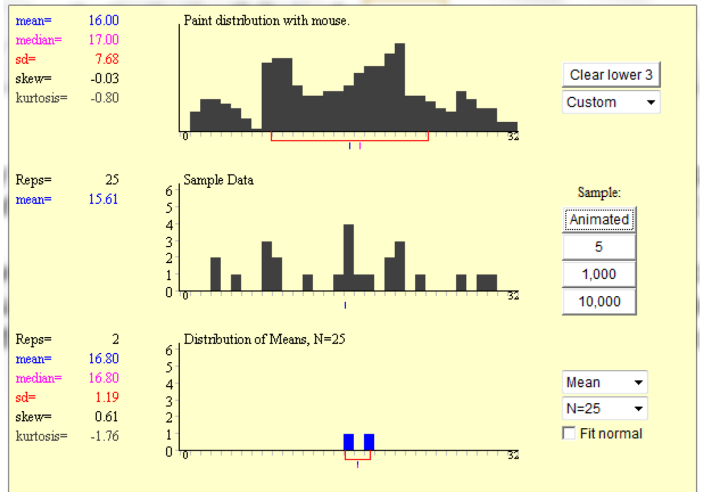
Reps= 1
mean= 17.99
median= 17.99
sd= 0.00
skew= 0.00
kurtosis= 0.00

Sample: Animated
Mean
N=25
□ Fit normal

Now, we take a sample from this population. That's what happens when I click the “Animated” button on the right. So here, in the middle panel, are 25 observations from this population.

And here they range pretty much over the whole range, just like the population. We see on the left that the mean of these 25 values, the sample mean, is 17.99.

 Remember we're trying to estimate the mean of the population, which is 16, on the basis of this sample of 25 observations. And the mean is 17.99. What the software does is also retain that mean, the mean of the sample, in the third panel down.



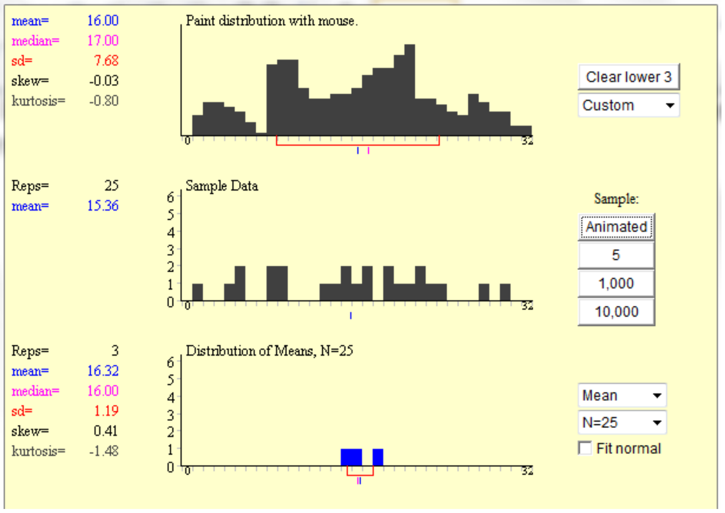
mean= 16.00
median= 17.00
sd= 7.68
skew= -0.03
kurtosis= -0.80

Reps= 25
mean= 15.61

Reps= 2
mean= 16.80
median= 16.80
sd= 1.19
skew= 0.61
kurtosis= -1.76

Custom
Mean
N=25
☐ Fit normal

Let's take another sample of size 25. This time we'll get another mean. This mean is 15.61. We could use this to estimate the population mean of 16. But what we now also have (bottom left-hand corner) is that the mean of the two sample means—the earlier, 17.99, and the current, 15.61—is 16.80.



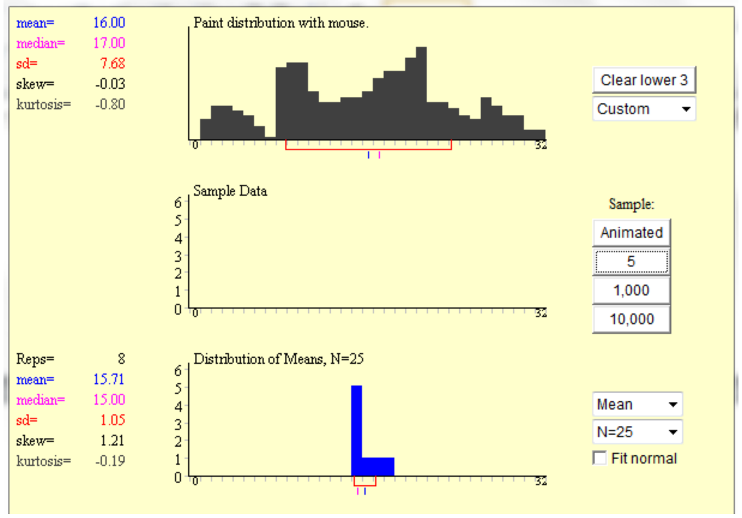
mean= 16.00
median= 17.00
sd= 7.68
skew= -0.03
kurtosis= -0.80

Reps= 25
mean= 15.36

Reps= 3
mean= 16.32
median= 16.00
sd= 1.19
skew= 0.41
kurtosis= -1.48

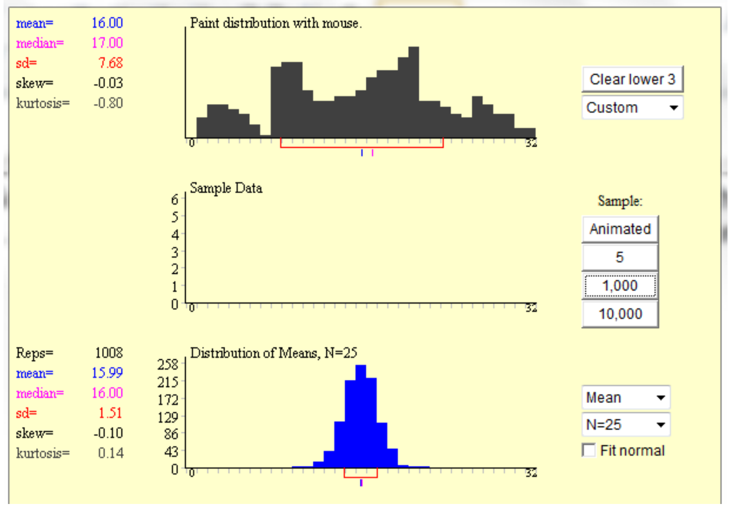
Custom
Mean
N=25
☐ Fit normal

We can repeat this process. Henceforth we will have three means on the screen. The mean of the population (16), the mean of the current sample (15.36), and the mean of each of the sample means—three in this case—16.32. It looks like the last mean is getting closer to the mean of the population, 16—17.99, 16.88, and16.32.



```
mean=       16.00
median=     17.00
sd=          7.68
skew=       -0.03
kurtosis=   -0.80

Reps=           8
mean=       15.71
median=     15.00
sd=          1.05
skew=        1.21
kurtosis=   -0.19
```

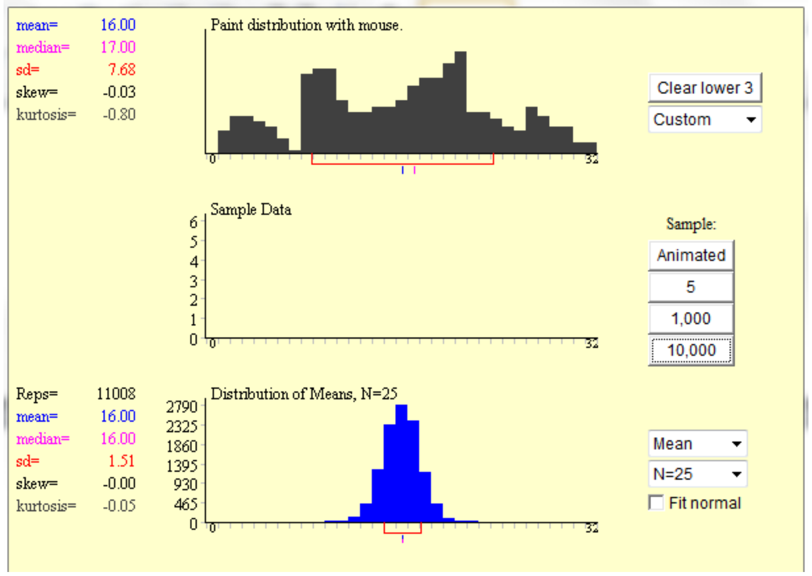
We can speed this up. We can take five samples by clicking the 5, just under "Animated". So now we have taken eight samples, and we see that the mean of these eight is 15.71. We are getting closer to 16, the mean of the population.



```
mean=       16.00
median=     17.00
sd=          7.68
skew=       -0.03
kurtosis=   -0.80

Reps=        1008
mean=       15.99
median=     16.00
sd=          1.51
skew=       -0.10
kurtosis=    0.14
```

Now we could keep on doing this. But let's just do it 1,000 times by clicking the button under the 5 on the right.

Now, the mean of the means is 15.99. It is getting ever closer to 16.



```
mean=       16.00
median=     17.00
sd=          7.68
skew=       -0.03
kurtosis=   -0.80
```

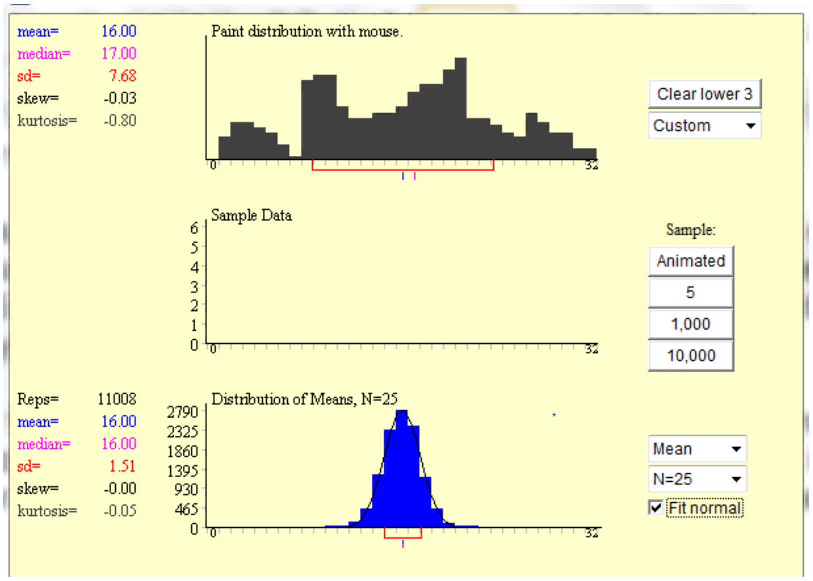
```
Reps=       11008
mean=       16.00
median=     16.00
sd=          1.51
skew=       -0.00
kurtosis=   -0.05
```

Let us do another 10,000 times by pressing the 10,000 button under the 1,000 on the right.

Now the mean of the means is 16. We have gotten there! Indeed, the statistical theorem states: as the sample size gets bigger (it is only 25 in this case) the mean of the means approaches the population mean. (To be correct, we should add that the variance of the population needs to exist.)

Let's look at the distribution of the mean of the means. They look tighter, they look more closely bunched together then the population. Of course, if we look at the spread of the distributions we see that the population standard deviation is 7.68, whereas the standard deviation of the distribution of the sample means is 1.51.

There is a relationship between the means of these two distributions—they are the same—so is there a relationship between these two standard deviations. The answer is in the affirmative. If we look at their ratio it is 5.1, or rounding off, approximately 5. This is the square root of the sample size we have been looking at all along, namely 25. And that is the relationship: divide the population standard deviation by the square root of the sample size to get the standard deviation of the distribution of the sample means. Further, this latter standard deviation has a name: it is called the *standard error*.



```
mean=      16.00
median=    17.00
sd=         7.68
skew=      -0.03
kurtosis=  -0.80

Reps=      11008
mean=      16.00
median=    16.00
sd=         1.51
skew=      -0.00
kurtosis=  -0.05
```

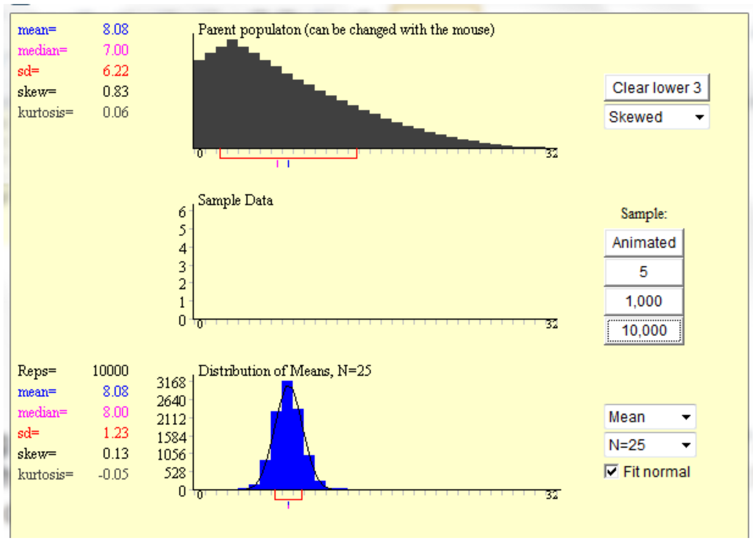
Custom

Mean

N=25

☑ Fit normal

Lastly, what is the shape of the distribution of the sample means at the bottom of the picture, above? It is very much reminiscent of what we saw with the Quicunx, and for exactly the same reason. I clicked the box “Fit normal” at the bottom right-hand corner and the computer superimposed a normal curve, which, as you can see, does a good job of describing the distribution.



```
mean=      8.08
median=    7.00
sd=        6.22
skew=      0.83
kurtosis=  0.06

Reps=     10000
mean=      8.08
median=    8.00
sd=        1.23
skew=      0.13
kurtosis= -0.05
```

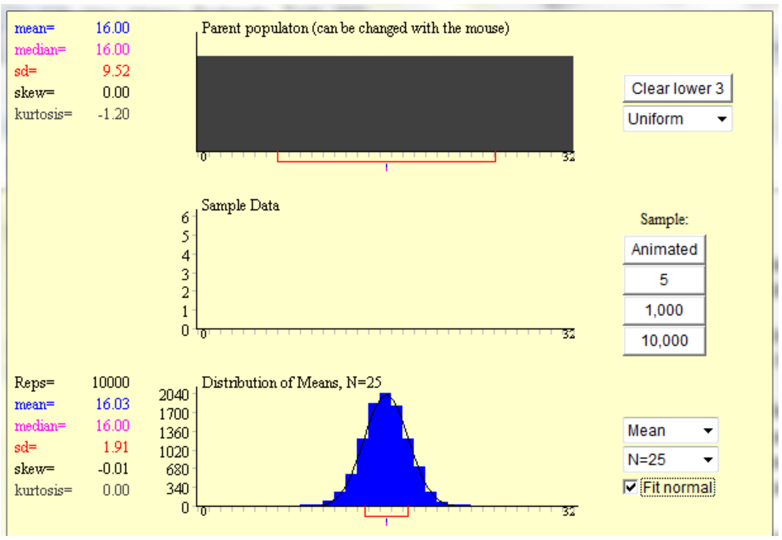
Skewed

Mean

N=25

☑ Fit normal

Let us return to the program and choose a skewed population distribution. Let us now take 10,000 samples of size 25, and we see, once again, a lovely bell shaped curve at the bottom.



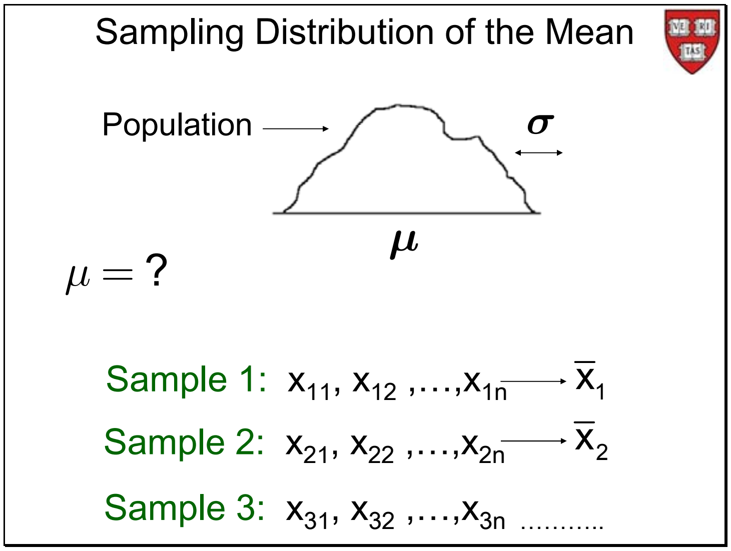
mean= 16.00
median= 16.00
sd= 9.52
skew= 0.00
kurtosis= -1.20

Reps= 10000
mean= 16.03
median= 16.00
sd= 1.91
skew= -0.01
kurtosis= 0.00

Uniform
Mean
N=25
☑ Fit normal

If we choose a uniform population distribution, pretty much the same thing happens. The shape is close to normal.

You have just experienced magic. We always get this normal distribution when we look at the distribution of the sample means, from whatever population you can contrive in this program. If the sample size is big enough, here we took 25, we are always going to get this normal distribution at the bottom.



Sampling Distribution of the Mean

Population ⟶ σ μ

$\mu = ?$

Sample 1: $x_{11}, x_{12}, \ldots, x_{1n} \longrightarrow \bar{x}_1$

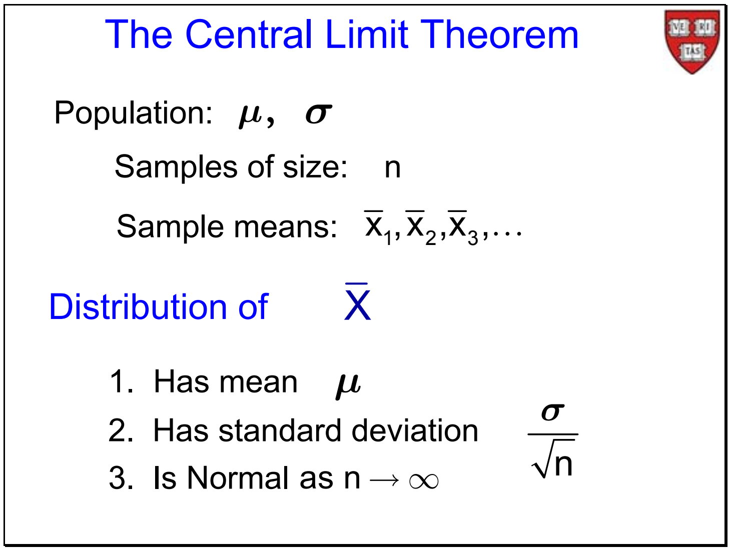
Sample 2: $x_{21}, x_{22}, \ldots, x_{2n} \longrightarrow \bar{x}_2$

Sample 3: $x_{31}, x_{32}, \ldots, x_{3n}$ ....

So let us summarize. We start with an arbitrary population. It has a mean, μ, and standard deviation, σ[3]. We don't know what these values are typically, and we would like to make inference about the value of this population mean.

Then this concept of a sampling distribution is, first of all, take a sample. Then calculate the mean of the sample. And discard the sample and just retain the sample mean.

Do this again tomorrow. Take another sample, keep the sample mean, get rid of the original observations. And keep on doing this repeatedly. OK



The Central Limit Theorem

Population: $\mu, \quad \sigma$

Samples of size: n

Sample means: $\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots$

Distribution of $\bar{X}$

1. Has mean μ
2. Has standard deviation $\dfrac{\sigma}{\sqrt{n}}$
3. Is Normal as $n \to \infty$

We typically reserve Greek letters for the quantities we do not know, or the parameters that refer to the population. So the population has mean, μ, and standard deviation, σ. We are going to take repeated samples of size n. Once again, typically we just take one sample. But for now we are talking about sampling distributions, so let us talk about repeated samples. Then for each one of these samples we get a sample mean. Let us focus on the distribution of these sample means.

We saw this distribution displayed in the third panel down. That distribution has mean, μ, the same as the population distribution, as dictated by theory. That is result number one.

This distribution has standard deviation, or standard error, σ/√n, which, if n>1, is smaller than the population standard deviation, and gets smaller as n, the sample size, gets larger.

So, when we think of how the sample mean varies from sample to sample, first it does so around a mean that is the same as the population mean. That makes sense.

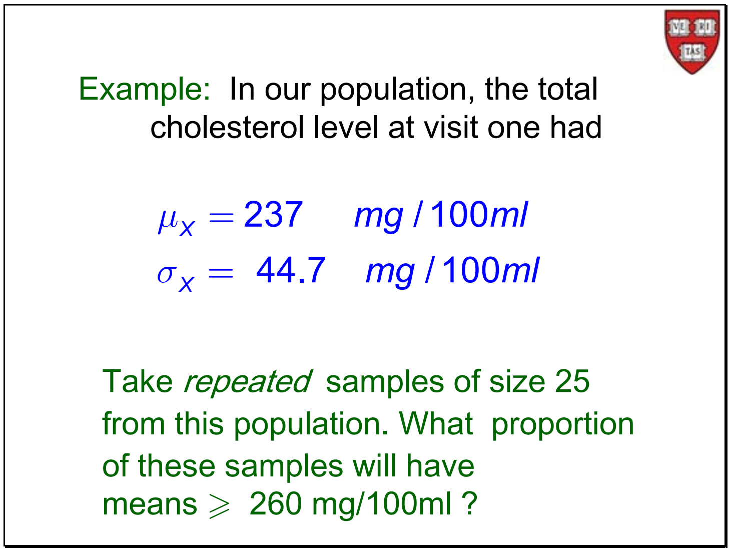
Number two, it varies much less than the population values—and the factor by how much less, is the square root of n. Third, it is normally distributed or has a bell-shaped curve. And this is

[3] For the purists, we are assuming that $\sigma < \infty$.

technically true as n goes to infinity. That is the theory. This is a most amazing theorem. De Moivre showed this to be true when we took samples from a binomial population, but this now is stated for sampling from quite general population distributions.

All these results require that the sample size n goes to infinity. Of course, we do not have infinite samples, nor are we likely to get any in the near future, so how do we use this result? We argue that this theorem provides us large sample approximations. What do we mean by large? It turns out that the closer the population distribution is to itself being bell shaped, the smaller the sample we need to have good approximations.

This theorem is called the *central limit theorem*. It actually is central, if I am allowed the pun, to just about all the inference we are going to be making in this course. It is extremely important to make sure you understand the meaning of what we have been saying up to now. Go to the Rice site and experiment for yourself.



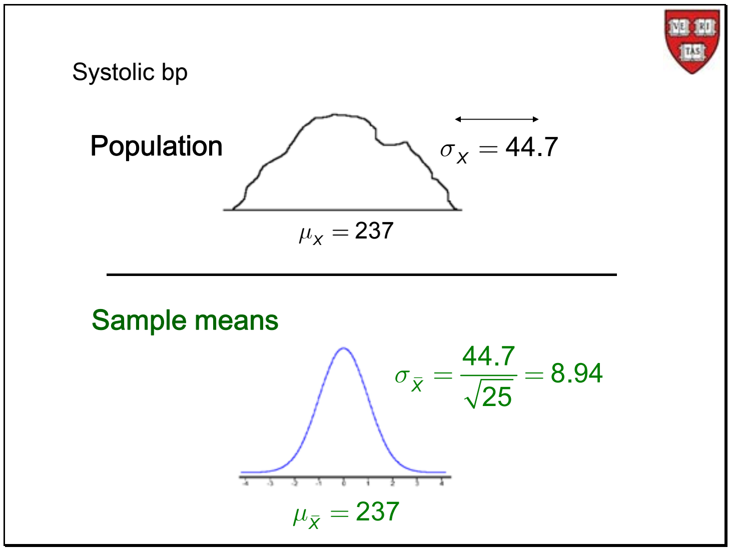
Example: In our population, the total cholesterol level at visit one had

$$\mu_X = 237 \quad mg/100ml$$
$$\sigma_X = 44.7 \quad mg/100ml$$

Take *repeated* samples of size 25 from this population. What proportion of these samples will have means $\geqslant$ 260 mg/100ml ?

How do we use the central limit theorem? Let us take a look at an example. We know that in our population, the total cholesterol level at visit one has a mean of 237 and a standard deviation of 44.7, rounding things off. Now, if we take repeated samples of size 25 from this population, what proportion of these samples will have a mean that is greater than 260?

The mean μ = 237, and σ =44.7, because we know the population from which we are sampling. I repeat, that is not how things typically work in practice, but we are carrying out an intellectual exercise. Let us ask the question, how often will the sample mean— not the population mean— how often will the sample mean, when we take samples of size 25, have a value bigger than or equal to 260?



Systolic bp

Population

$\sigma_X = 44.7$

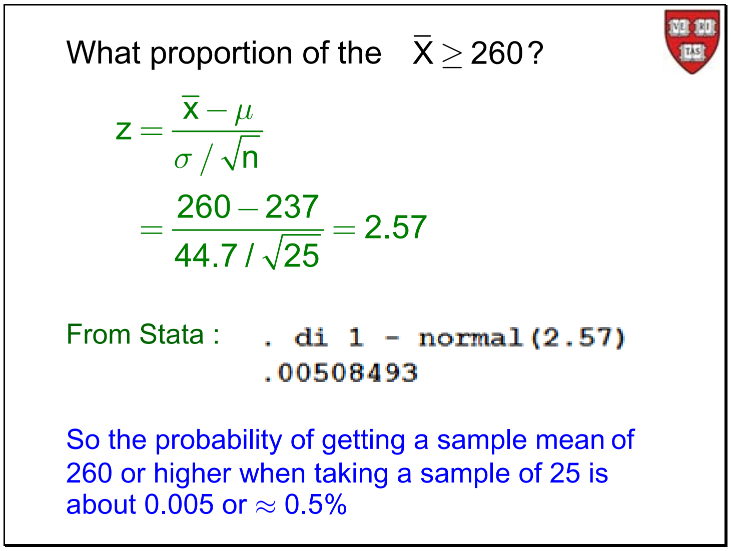
$\mu_X = 237$

Sample means

$\sigma_{\bar{X}} = \frac{44.7}{\sqrt{25}} = 8.94$

$\mu_{\bar{X}} = 237$

We can use the central limit theorem because we've got a population with mean 237, standard deviation of 44.7. And we're going to take the sample mean. And we know from the central limit theorem that if we do this repeatedly, those sample means will be distributed around a mean of 237 and a standard deviation of 44.7 divided by the square root of the sample size, which in this case gives us 8.94.



What proportion of the $\bar{X} \geq 260$?

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$= \frac{260 - 237}{44.7 / \sqrt{25}} = 2.57$$

From Stata :

```
. di 1 - normal(2.57)
.00508493
```

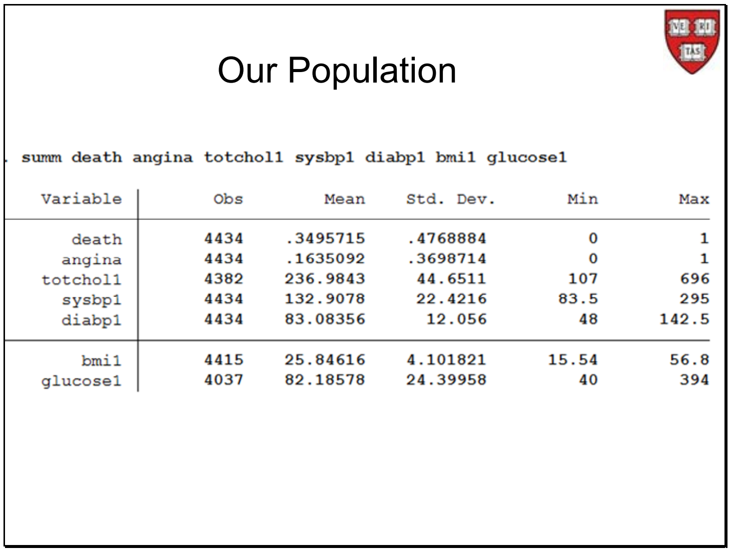
So the probability of getting a sample mean of 260 or higher when taking a sample of 25 is about 0.005 or ≈ 0.5%

So let us return to our standardized variable by subtracting from 260 its mean, 237, and dividing by its standard deviation, and the answer we get is 2.57. So the question asked about X bar gets translated into a statement about Z asking, how often is Z greater than or equal to 2.57? The answer is 0.005.

So we can answer the question, the probability of getting a sample mean of 260 or higher when taking a sample of size 25 from this population is about 0.5%. So it's something that will happen probably roughly once in 200 times. So this might be our definition of a rare event. So very rarely, will we see a sample mean of 260 or bigger.

Throughout these calculations we are working under the assumption that n=25 is a large enough sample so that the central limit theorem provides us a good approximation. But that is how we can use the central limit theorem.

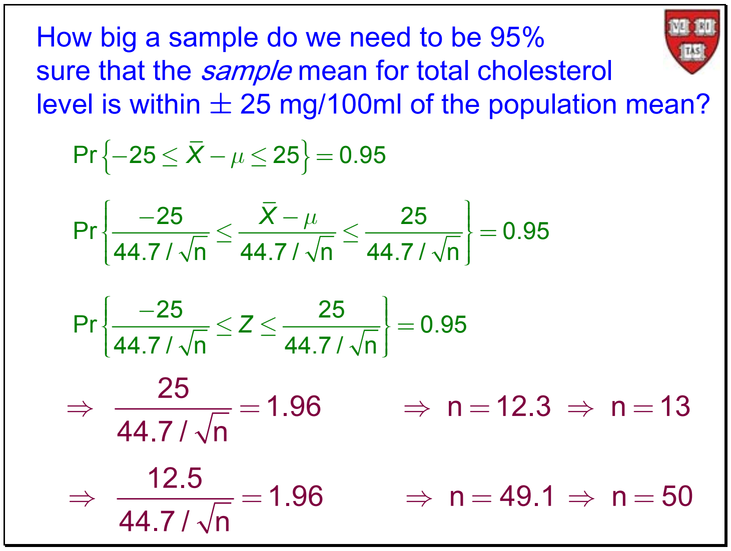
Sample Size



Our Population

```
. summ death angina totchol1 sysbp1 diabp1 bmi1 glucose1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       death |      4434    .3495715    .4768884          0          1
      angina |      4434    .1635092    .3698714          0          1
    totchol1 |      4382    236.9843     44.6511        107        696
      sysbp1 |      4434    132.9078     22.4216       83.5        295
      diabp1 |      4434    83.08356      12.056         48      142.5
-------------+--------------------------------------------------------
        bmi1 |      4415    25.84616    4.101821      15.54       56.8
    glucose1 |      4037    82.18578    24.39958         40        394
```

Now, how can we use the central limit theorem to answer sensible questions? We saw one question we just answered. Another question that's very often asked is, how big a sample do I need? And the answer of course, is for what? But let's look at one situation where we can answer this question by being more precise in our question.



How big a sample do we need to be 95% sure that the *sample* mean for total cholesterol level is within ± 25 mg/100ml of the population mean?

$$\Pr\left\{-25 \le \bar{X} - \mu \le 25\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le \frac{\bar{X} - \mu}{44.7/\sqrt{n}} \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Pr\left\{\frac{-25}{44.7/\sqrt{n}} \le Z \le \frac{25}{44.7/\sqrt{n}}\right\} = 0.95$$

$$\Rightarrow \frac{25}{44.7/\sqrt{n}} = 1.96 \qquad \Rightarrow n = 12.3 \Rightarrow n = 13$$

$$\Rightarrow \frac{12.5}{44.7/\sqrt{n}} = 1.96 \qquad \Rightarrow n = 49.1 \Rightarrow n = 50$$

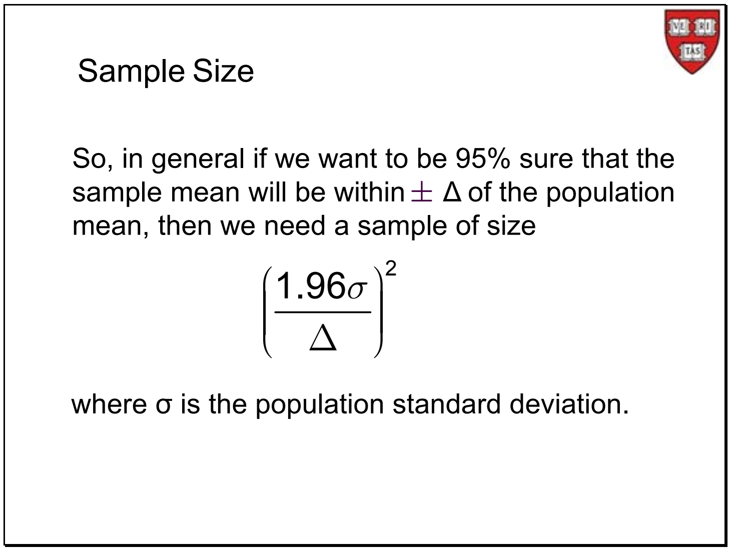
Now suppose we do not know, as is usually the case, what the population is. And we wish to take a sample. So the question might be phrased this way. "How big a sample do we need to be 95% sure that the sample mean for total cholesterol level is within plus or minus 25 milligrams per 100 milliliters of the population mean?" So we state how sure we need to be and the degree of accuracy we seek.

Here, we use the ubiquitous 95%. So how big a sample do we need to be 95% sure? So I am not going to be certain, when I get my answer, but I shall be 95% sure that what we did will get me to within 25 units of the population mean. I am not certain that what I am going to do will be within the limit that you have set, but I am 95% sure. That's the best I can do.

So let us standardize. Divide by the standard deviation of the sample mean, or the standard error. Which, because we know the population standard deviation—I am being a little bit unrealistic here, I am acting as if I know the population standard deviation, I realize that. We remove that assumption shortly, but let us just answer this question first.

Divide through by the standard error to get our standardized Z. So now we are asking what is the probability that Z will be between minus this quantity and plus the quantity? We want that to be 0.95. So we know, that a standard Z, is between –1.96 and 1.96 95% of the time. And in order to do that, because n is the only unknown in here, we can solve for n. And it turns out that n of 13 will allow us to be 95% sure that we are within plus or minus 25 units of the population mean by using the sample mean.

Suppose we want to lower that 25. Let's halve it to 12.5. Then we would need a sample of size 50.



Sample Size

So, in general if we want to be 95% sure that the sample mean will be within ± Δ of the population mean, then we need a sample of size

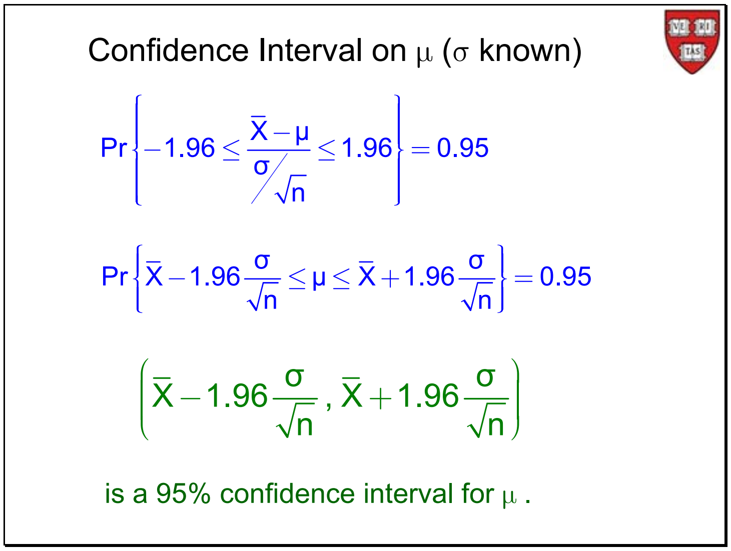
$$\left(\frac{1.96\sigma}{\Delta}\right)^2$$

where σ is the population standard deviation.

In summary, if we want to be 95% sure that the sample mean is within plus or minus delta of the population mean, then we need a sample of size n that equals this quantity, where is sigma is the population standard deviation.

This formula is appealing. The bigger is our sigma, the more variable is our population. And we have been a saying all along, the less homogeneous the population, the bigger the sample we are going to need. And the smaller we make delta, the more precise you want to be, then the larger your sample size will have to be. So these are the two controlling factors here, how close you want to be and how variable is your population.

Confidence Interval



Confidence Interval on μ (σ known)

$$\Pr\left\{-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right\} = 0.95$$

$$\Pr\left\{\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

$$\left(\bar{X}-1.96\frac{\sigma}{\sqrt{n}}\,,\ \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for μ .

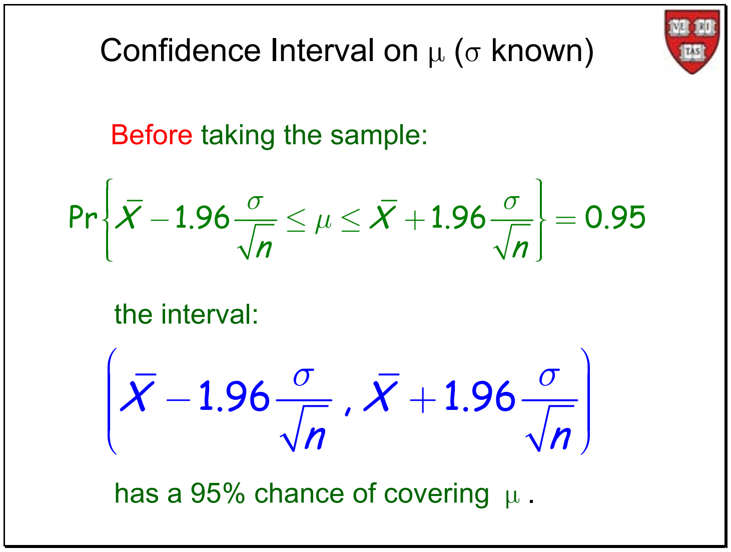
The next topic we explore is the ever popular confidence interval. I am going to keep on acting as if we know sigma. As I said, we will remove that shortly. But for the time being, just to keep the conversation straightforward without extra complexity, let's just act as if we know sigma.

So here is our standardized variable. From the central limit theorem, we know that X-bar has got mean μ and standard deviation (standard error) σ/√n. Let us assume that n is large enough so that we can make the statement that our standardized variant will take on values between minus 1.96 and 1.96 95% of the time.

Sigma is positive. Square root of n is positive. So let us multiply through by the standard error, and that will not disturb the inequalities. Now subtract X-bar from both sides, and then multiply everything by minus 1 and reverse the inequalities. And there you have it, after just a little bit of algebra. These two probability statements are exactly the same. But look at what we have done: From sample to sample, in the first probability statement, the standardized variable is what varies. By doing that little bit of algebra, from sample to sample, what now varies are the limits of our interval—μ remains fixed.

So we have created what is called a random interval. We now have an interval that is from X-bar minus, to X-bar plus, 1.96 σ/√n. This interval, is going to vary from sample to sample, because X-bar does. Furthermore, this interval will cover μ 95% of the time. And, of course, roughly 5% of the time it will not cover μ. Unfortunately, on any one occasion we do not know whether we are in the 95% or the 5% !

This interval is called a confidence interval. We are 95% confident before we do any calculations, that the resultant interval will cover μ. Once we have done our calculations, we do not know. Maybe the interval includes μ. Maybe it does not. What we have here is a recipe, one that has a 95% chance of success. That is what a confidence interval is. It is a rule that has a 95% chance of success; success being measured by whether the interval contains μ.



Confidence Interval on μ (σ known)

Before taking the sample:

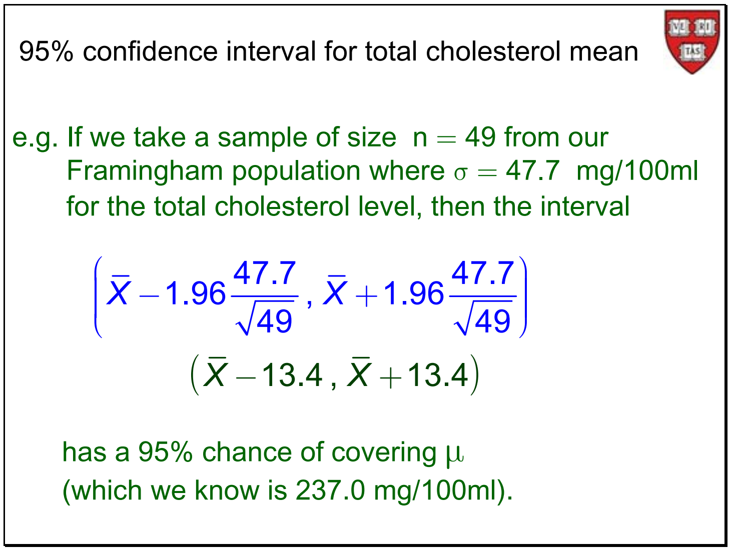
$$\Pr\left\{\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$$

the interval:

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,,\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

has a 95% chance of covering μ .

So before we take the sample, we can make the statement that there is a 95% chance that it will do the right thing. So we are 95% confident that we will capture μ with this interval.



95% confidence interval for total cholesterol mean

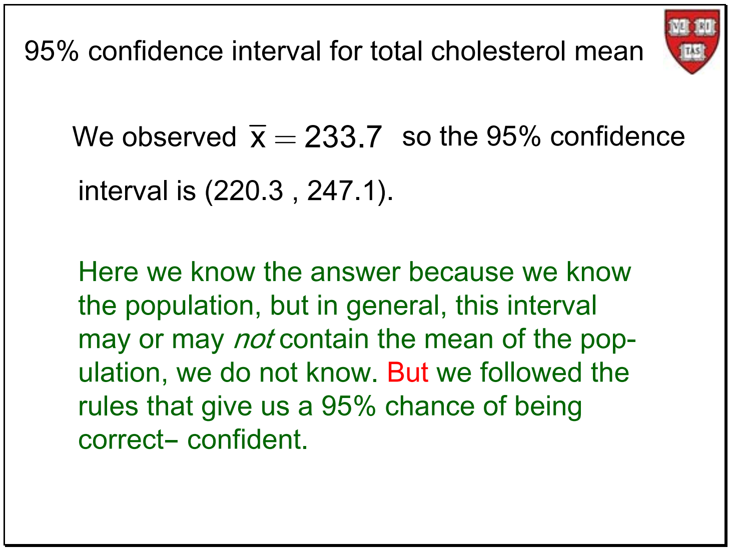
e.g. If we take a sample of size $n = 49$ from our Framingham population where $\sigma = 47.7$ mg/100ml for the total cholesterol level, then the interval

$$\left(\bar{X} - 1.96\frac{47.7}{\sqrt{49}}\,,\ \bar{X} + 1.96\frac{47.7}{\sqrt{49}}\right)$$

$$\left(\bar{X} - 13.4\,,\ \bar{X} + 13.4\right)$$

has a 95% chance of covering μ (which we know is 237.0 mg/100ml).

Consider the total cholesterol level at visit one from our Framingham population. If we take a sample of size 49 and assume that the standard deviation is 47.7, then our 95% confidence interval is X-bar minus 13.4, to X-bar plus 13.4, and that interval has a 95% chance of covering the true mean.



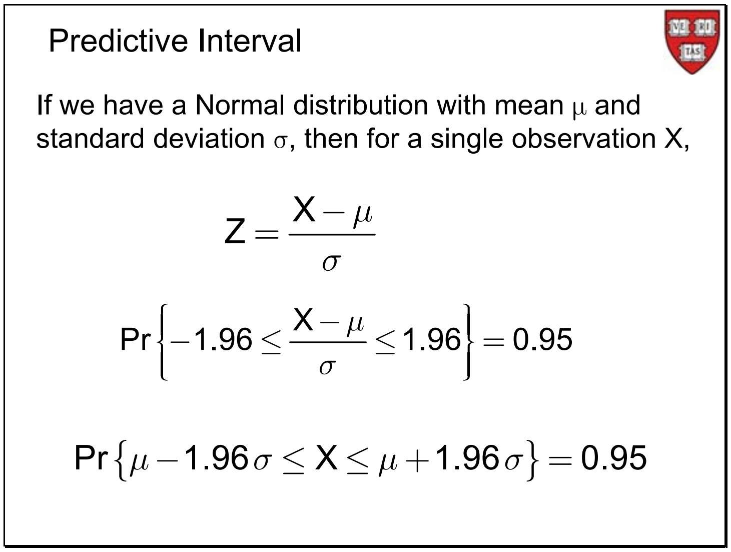
95% confidence interval for total cholesterol mean

We observed $\overline{x} = 233.7$ so the 95% confidence interval is (220.3 , 247.1).

Here we know the answer because we know the population, but in general, this interval may or may *not* contain the mean of the population, we do not know. But we followed the rules that give us a 95% chance of being correct– confident.

In our example, we know what the true mean is, so we can check to see if, indeed, it worked or not, because we're just doing this as an exercise. So let us see what happens here. We actually observe a sample mean of 233.7. So our 95% confidence interval is 220.3 to 247.1. And in this case, we know it covered it. Typically, we don't know if it covered it or not.

Predictive versus Confidence Interval



Predictive Interval

If we have a Normal distribution with mean μ and standard deviation σ, then for a single observation X,

$$Z = \frac{X - \mu}{\sigma}$$

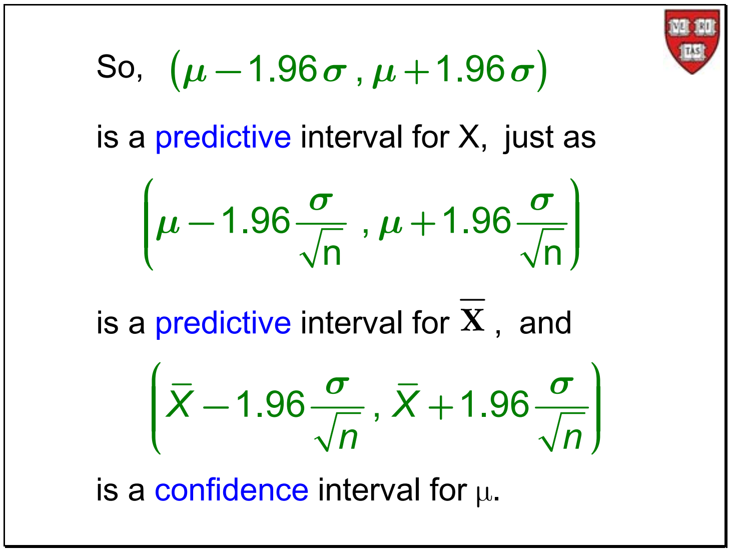
$$\Pr\left\{-1.96 \le \frac{X - \mu}{\sigma} \le 1.96\right\} = 0.95$$

$$\Pr\left\{\mu - 1.96\sigma \le X \le \mu + 1.96\sigma\right\} = 0.95$$

There is a closely related concept to confidence intervals and that is the predictive interval. We have seen this before. What we have said is, if we are to sample from a normal population, where do we expect an observation to fall? We do not know, but we can give an interval with an

There is a closely related concept to confidence intervals and that is the predictive interval. We have seen such an interval before. The way we introduced it is: if we are to sample from a normal population, where do we expect an observation to fall? We do not know, but we can give an interval with an associated probability as an answer to that question. That is what a predictive interval is.

We start with a variable, standardize it, and then the standardized variable will take a value between μ – σ and μ + σ 95% of the time. Before we measure the variable X, whatever value it takes, we can predict that 95% of the time that value is within two standard deviations of the mean. That is the predictive element.



So, $(\mu - 1.96\sigma\,,\, \mu + 1.96\sigma)$

is a predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \mu + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

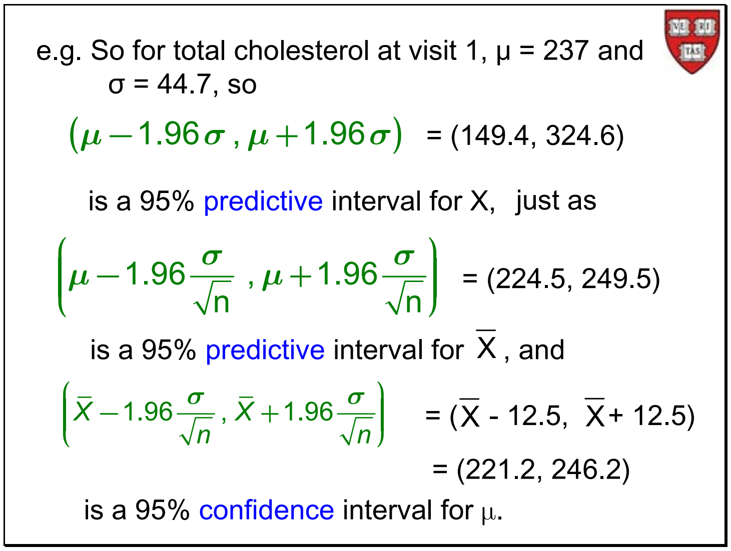
is a predictive interval for $\bar{\mathbf{X}}$, and

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\,,\, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

is a confidence interval for μ.

We have the predictive interval for X, what about X-bar? Because of the central limit theorem, we have a parallel result for X-bar if we replace the standard deviation with the standard error.

This is in contrast to what we have just seen, namely the 95% confidence interval. The latter we calculate *after* we have taken our sample. So the confidence interval is ex post, whereas the predictive interval is pre.



e.g. So for total cholesterol at visit 1, μ = 237 and σ = 44.7, so

$$(\mu - 1.96\sigma , \mu + 1.96\sigma) = (149.4, 324.6)$$

is a 95% predictive interval for X, just as

$$\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} , \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = (224.5, 249.5)$$

is a 95% predictive interval for $\bar{X}$, and

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} , \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = (\bar{X} - 12.5, \bar{X} + 12.5)$$

$$= (221.2, 246.2)$$

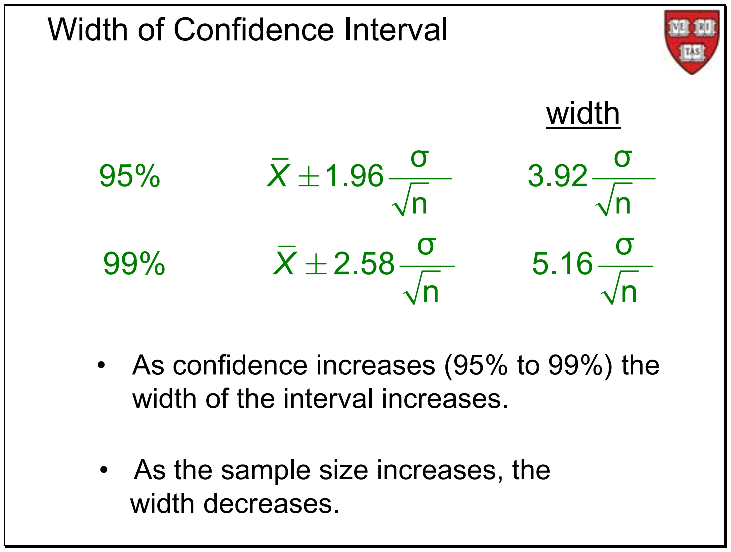
is a 95% confidence interval for μ.

Applying these ideas to our total cholesterol level at visit one, the mean of the population is 237, with a standard deviation of 44.7. So we would predict that 95% of the time that we choose an individual from this population, that individual, we will have a value that is between 149.4 and 324.6.

If we choose a sample of size 49 from this population then we predict that 95% of the time the mean of such a sample will be between 224.5 and 249.5. This is a much tighter interval because the standard error is one seventh the standard deviation.

Once we have that the sample mean is 233.7, we can construct the confidence interval. We see that we are successful with this confidence interval because it covers the true mean of 237.

Width of the Confidence Interval



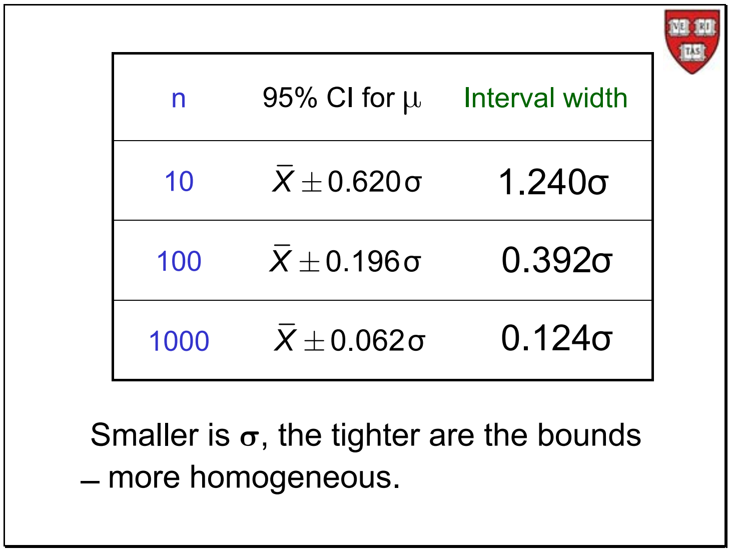
Width of Confidence Interval

		width
95%	$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$	$3.92\frac{\sigma}{\sqrt{n}}$
99%	$\bar{X} \pm 2.58\frac{\sigma}{\sqrt{n}}$	$5.16\frac{\sigma}{\sqrt{n}}$

- As confidence increases (95% to 99%) the width of the interval increases.
- As the sample size increases, the width decreases.

Typically we would like to make the confidence interval as tight as possible. One cheap way to achieve that is to place less confidence in the interval. For example, we see that if we want to be 99% confident then that interval has width 2 x 2.58 x the standard error. On the other hand, to be 95% confident the interval only has width 2 x 1.96 x the standard error; about 76% the size.

The other variable that determines the width of the interval is the standard error, which we know we can decrease by increasing the sample size.



n	95% CI for μ	Interval width
10	$\bar{X} \pm 0.620\sigma$	1.240σ
100	$\bar{X} \pm 0.196\sigma$	0.392σ
1000	$\bar{X} \pm 0.062\sigma$	0.124σ

Smaller is σ, the tighter are the bounds – more homogeneous.

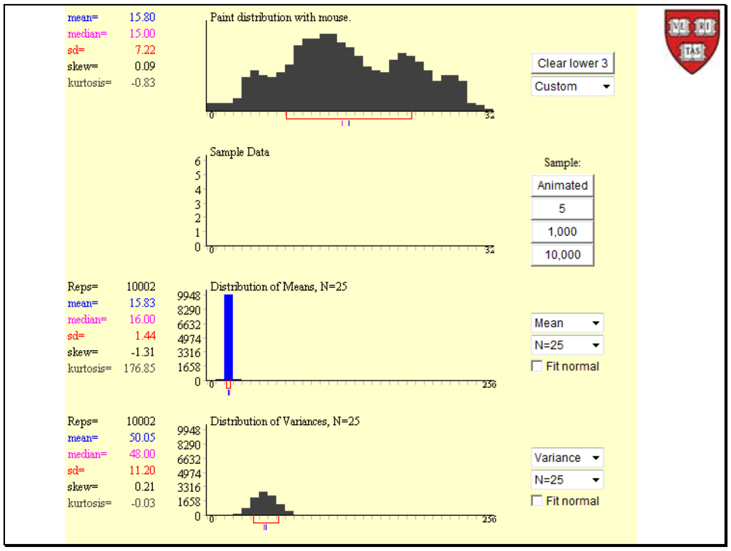
Here are three n's to show this decrease in width, chosen so that we see that it is the square root of n that determines the width and not n. So to see a decrease of one tenth, we need to go from n=10 to n=1,000.

We have no control over σ, it is a population parameter, but we see that the bounds are a function of sigma, reflecting what we have been saying all along: The more homogeneous our population, the more confident we are in whatever inference we make.

It is not quite right to say that we have no control over σ , as we see later in the course when we can stratify the population, we can look at different, and more homogeneous, sections of the population in turn and make our inferences in turn for each stratum. For example, instead of targeting the population as a whole, if there is a difference between men and women, and men are homogeneous, and the women homogeneous, then we might first make inference for the women and then for the men. That way we take advantage of the smaller, sex defined, standard deviations. More about this when we get to sampling later in the course.

Unknown σ – Student's t-distribution

To study the situation when σ is unknown let us return to the website at Rice, but this time let us also look at the bottom tableau.



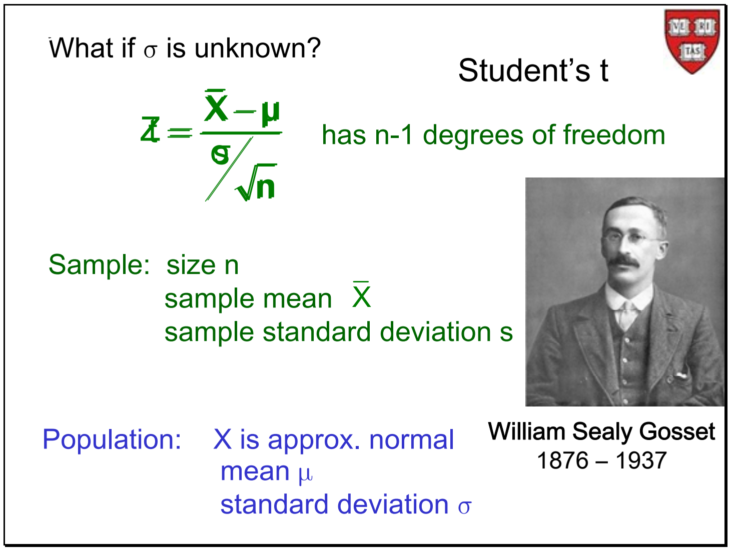
```
mean=       15.80
median=     15.00
sd=          7.22
skew=        0.09
kurtosis=   -0.83

Reps=       10002
mean=       15.83
median=     16.00
sd=          1.44
skew=       -1.31
kurtosis=  176.85

Reps=       10002
mean=       50.05
median=     48.00
sd=         11.20
skew=        0.21
kurtosis=   -0.03
```

Let's stick with our N of 25. And in this one, let's stick with the mean as we did before, and have fun drawing a population distribution. This one has a mean of 15.8 and a standard deviation of 7.22. In the bottom panel let us ask for the variance.

After two animated samples we can jump ahead and run 10,000 samples. Now, we have that the mean of the means is 15.83, very close to the population mean. Plus the standard error is 1.44 which is one fifth of 7.20 which is very close to the population standard deviation.

Now let us look at the bottom panel. The mean of the variances is 50.05. How does this relate to the population variance. The software gives the standard deviation as 7.22, thus the population variance is $(7.22)^2$ = 52.13. So it seems like the mean of the sample means is the same as the population variance, and indeed that is what theory shows. Indeed, this is the reason why we use (n-1) as the divisor in the definition of the sample variance; to obtain this property. This is what we call *unbiasedness*. We say that the sample variance is an unbiased estimator of the population variance—just like the sample mean is an unbiased estimator of the population mean.

A few weeks ago I told you that I would let you know why we divide by n minus 1, and not n. Had we divided by n, we would not have gotten an unbiased estimator of the population variance. Promise met.



What if σ is unknown?

Student’s t

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

has n-1 degrees of freedom

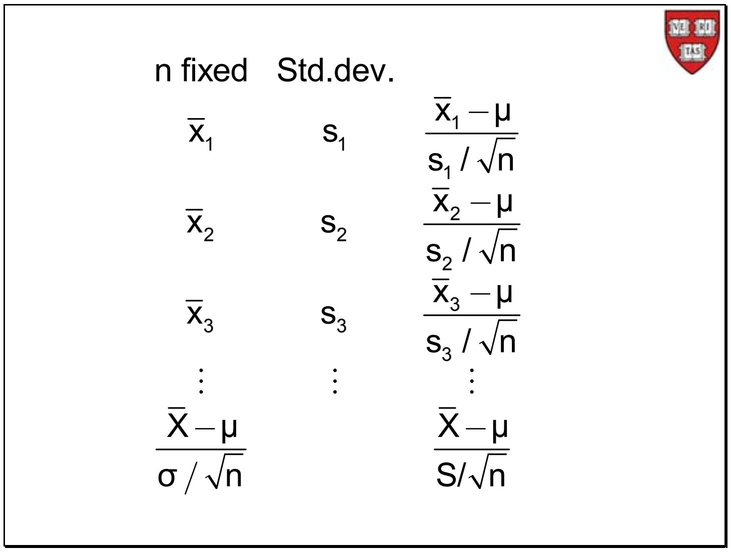
Sample: size n
sample mean $\bar{X}$
sample standard deviation s

Population: X is approx. normal
mean μ
standard deviation σ

William Sealy Gosset
1876 – 1937

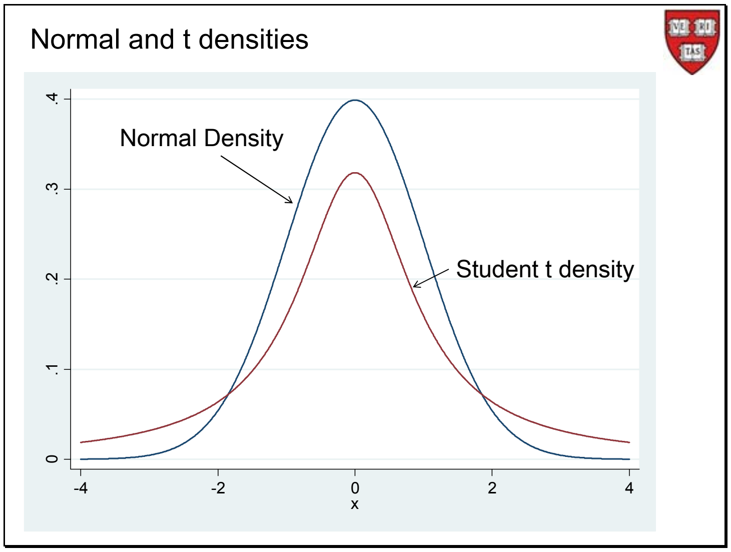
We make use of all this information we have gathered by looking at our standardized variate, Z, and if we do not know σ, we replace it with the sample standard deviation, s. To distinguish it from Z, we call this modified standardized variate t. This is sometimes called a Studentized variable after William Gosset who wrote under the nom de plume, Student. He did that to guard his work identity—he worked at the Guinness Brewery.

Student was the person who discovered the sampling distribution of this t. It is not as simple as Z in that we now require that the population distribution of X is at least approximately normal. Then the distribution of t varies with the size of the sample n. We call n-1 the *degrees of freedom* of the t distribution.



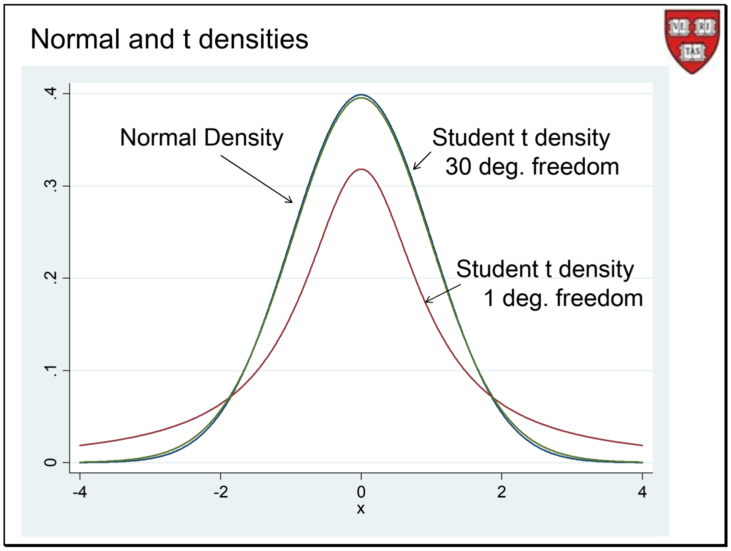
n fixed	Std.dev.	
$\bar{x}_1$	s_1	$\frac{\bar{x}_1-\mu}{s_1/\sqrt{n}}$
$\bar{x}_2$	s_2	$\frac{\bar{x}_2-\mu}{s_2/\sqrt{n}}$
$\bar{x}_3$	s_3	$\frac{\bar{x}_3-\mu}{s_3/\sqrt{n}}$
⋮	⋮	⋮
$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$		$\frac{\bar{X}-\mu}{S/\sqrt{n}}$

When we compare the distribution of the t to the Z we expect that there will be more variability in the former as the sample standard deviation varies from sample to sample, whereas in contrast σ in the definition of Z, remains constant.



Our intuition is borne out by the theory. Here is the normal density superimposed on a Student's t. You can see that the tails on the t are much wider than the normal, and thus the variance is larger. Also, in the same vein, the normal is much more peaked and tighter around the origin.

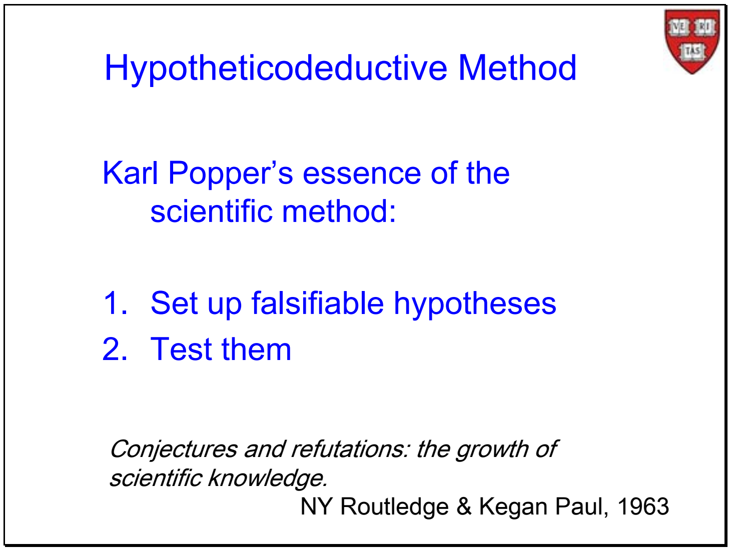
Now, what I've drawn here is about as bad as the Student t can get. This is when the Student t has one degree of freedom, which is also sometimes called the Cauchy distribution, and the tails are so fat that the mean does not even exist for this distribution, let alone the variance.



By the time you get to about 30 degrees of freedom, the normal and the t are almost indistinguishable. In this graph I have actually plotted three curves, the two from the previous graph plus a t-density with 30 degrees of freedom. As you see that there is very little difference between the student with 30 degrees of difference and the normal density.

In the old days, when we looked up values in tables, we would say, oh, if it has more than 30 degrees of freedom, just use the normal distribution. Stata does not do that, because it can calculate quantities exactly for you, but if stuck on a desert island without Stata and only a book of tables…

Hypothesis Testing



Hypotheticodeductive Method

Karl Popper's essence of the scientific method:

1. Set up falsifiable hypotheses
2. Test them

Conjectures and refutations: the growth of scientific knowledge.
NY Routledge & Kegan Paul, 1963

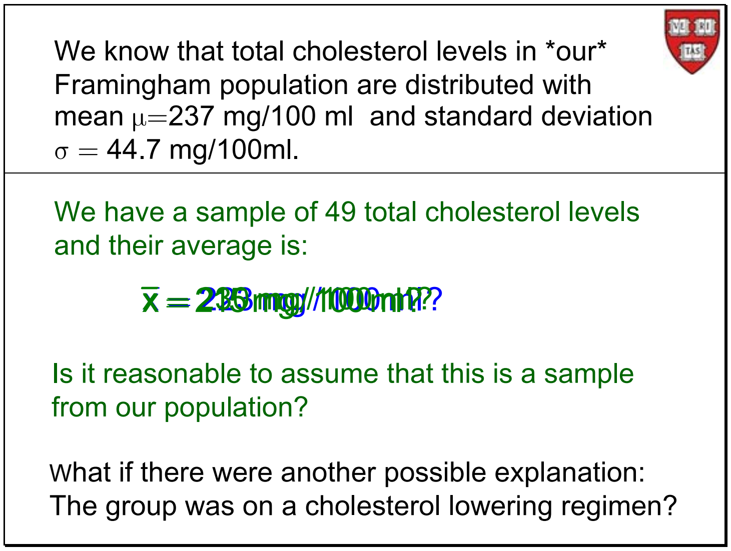
So up to now we have made inference by either just estimating the population value in the case of the mean or the standard deviation, or by constructing confidence intervals for the mean. We could also construct confidence intervals for the standard deviation, but we will not in this course. Instead we are going to look at another popular method of making inference. It is called hypothesis testing.

Hypothesis testing forms part of the hypothetical deductive method that we have been using for almost 200 years now, to increase our knowledge in science. Karl Popper called it the essence of the scientific method:

First set up a falsifiable hypothesis. That differentiates science from other forms of knowledge: The hypotheses you set up must be falsifiable.

Second, design your experiment, get your data, and test your hypothesis.

And, as we saw, there is going to be uncertainty in our inference.



We know that total cholesterol levels in *our* Framingham population are distributed with mean μ=237 mg/100 ml and standard deviation σ = 44.7 mg/100ml.

We have a sample of 49 total cholesterol levels and their average is:

$$\bar{x} = 230 \text{ mg}/100 \text{ ml}$$
$$\bar{x} = 223 \text{ mg}/100 \text{ ml??}$$
$$\bar{x} = 215 \text{ mg}/100 \text{ ml??}$$

Is it reasonable to assume that this is a sample from our population?

What if there were another possible explanation: The group was on a cholesterol lowering regimen?

Let us take a look at an example. We know that the total cholesterol levels in our population are distributed with mean μ = 237, and σ = 44.7. We know that, but let us act as if we did not know it. Now, suppose somebody approaches us and says, I have a sample of 49 cholesterol levels, and their sample mean is 230 milligrams per 100 milliliters.

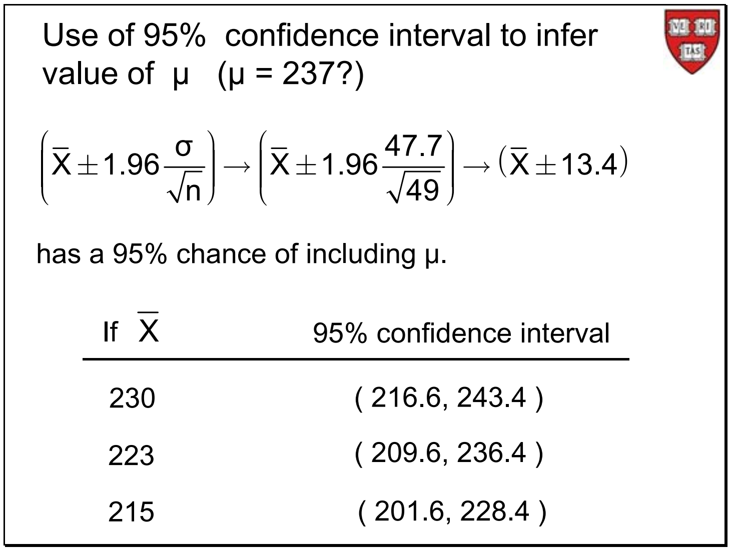
Is it likely that these 49 actually came from your Framingham population, or from a population with the same characteristics as your Framingham population? Is it possible that we could take a sample of 49 from our population and come up with a sample mean of 230?

Maybe what you would do is calculate that the standard deviation is 44.7, so the standard error is approximately 6 or 6½ (=44.7/7). So 230 is approximately one standard error away from 237, and so the central limit theorem tells us that quite often we will be within plus or minus 1 standard error of the population mean. So it would be reasonable in this case to assume that this is a sample from our population.

What if instead of a sample mean of 230, your friend had a sample mean of 223? Is it likely that this sample came from our population? Well, now we are talking about a deviation which is approximately two standard errors from the mean. We might now hesitate.

What if instead your friend had come with a sample mean of 215? Now we are talking about three standard errors away from the population mean. Maybe now we start looking for other explanations of where this sample came from. Were these folks on some cholesterol lowering medication?

This, in general, is how we approach the testing of hypotheses.



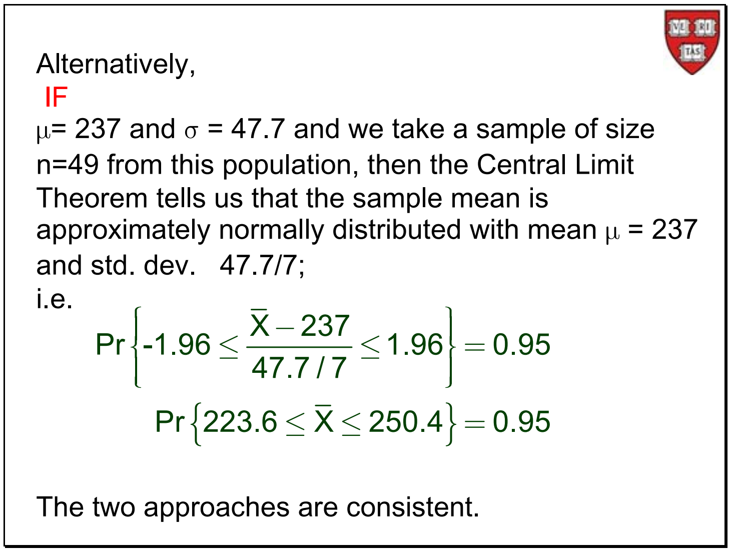
Use of 95% confidence interval to infer value of μ (μ = 237?)

$$\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right) \rightarrow \left(\bar{X} \pm 1.96\frac{47.7}{\sqrt{49}}\right) \rightarrow \left(\bar{X} \pm 13.4\right)$$

has a 95% chance of including μ.

If $\bar{X}$	95% confidence interval
230	(216.6, 243.4)
223	(209.6, 236.4)
215	(201.6, 228.4)

We could possibly use the 95% confidence interval to infer values of μ from the sample mean. The results of the calculations in the three instances are displayed above. The first confidence interval includes 237, the other two do not. In all instances we can say, I have followed a recipe that has a 95% chance of success. In the first case, I can say, the data do not refute the hypothesis that μ = 237. Whereas in each of the next two I can say, either μ = 237 and something that has a 95% chance of success did not happen, or μ ≠ 237.



Alternatively,
IF
μ= 237 and σ = 47.7 and we take a sample of size n=49 from this population, then the Central Limit Theorem tells us that the sample mean is approximately normally distributed with mean μ = 237 and std. dev. 47.7/7;
i.e.

$$\Pr\left\{-1.96 \le \frac{\bar{X} - 237}{47.7/7} \le 1.96\right\} = 0.95$$

$$\Pr\left\{223.6 \le \bar{X} \le 250.4\right\} = 0.95$$

The two approaches are consistent.

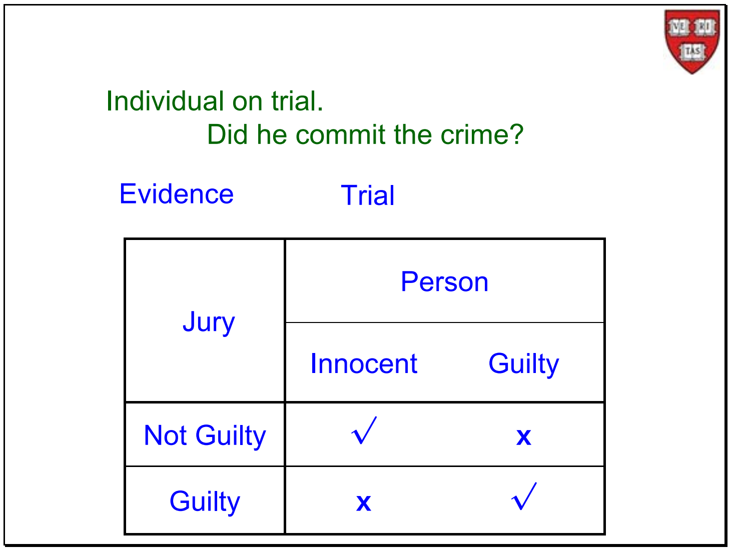
Alternatively, instead of confidence intervals we can construct a predictive interval. So in this case, we can say, when sampling from our population, a 95% predictive interval for the sample mean of samples of size 49, is (223.6, 250.4).

So in the first instance the sample mean is within these bounds, so our prediction worked and we can say we observed an event that had a 95% of happening, and thus is consonant with our hypothesized population. On the other hand in the second or third situation the 95% event did not happen, so either a rare event, one that has a 5% chance of happening, happened, or maybe we should reject the hypothesis that this sample came from this hypothesized population.

The conclusions you come to in those two situations, the confidence interval or the predictive interval these two approaches are completely consistent. There are some people who would tell you otherwise, but do not believe them. They are not correct. These two approaches are completely consistent.

Formalism of Hypothesis Testing

Let us look a little deeper into this second approach. It actually has a lot in common with our legal system as it pertains to criminal cases.



Individual on trial.
Did he commit the crime?

Evidence   Trial

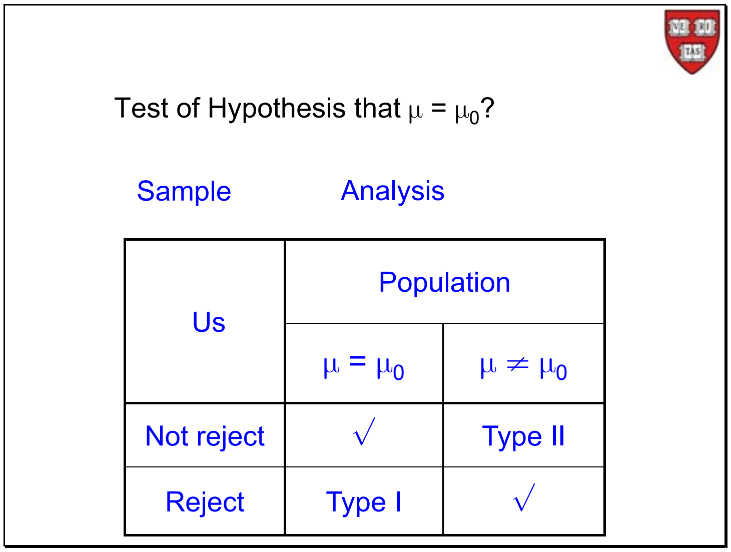
Jury	Person	
	Innocent	Guilty
Not Guilty	√	x
Guilty	x	√

We start with an individual on trial to answer the question of whether he committed the crime.

The evidence is presented at trial. The person truly is innocent and did not commit the crime or the person is guilty. The jury needs to decide whether the person is or is not guilty. Person is assumed innocent. So the jury just decides whether not guilty or guilty.

If the jury decides the person is not guilty when in fact the person is innocent, then the jury is doing the right thing. If on the other hand, the jury decides the person is guilty when the person in fact, did commit the crime, then once again, the jury is doing the right thing. The jury is only going to decide guilty or not guilty. So potentially it could make one of those two decisions. But in reality, after the trial, only one of these decisions is made, and thus only one row is relevant.

If on the other hand we're in one of the off diagonal situations, then the jury has made a mistake. So finding a guilty person not guilty or finding an innocent person guilty-- those are the mistakes that juries can make. Of course, at most a single jury is only going to possibly make one of those mistakes. But potential is there for either of those mistakes to be made, in general.



Test of Hypothesis that $\mu = \mu_0$?

Sample Analysis

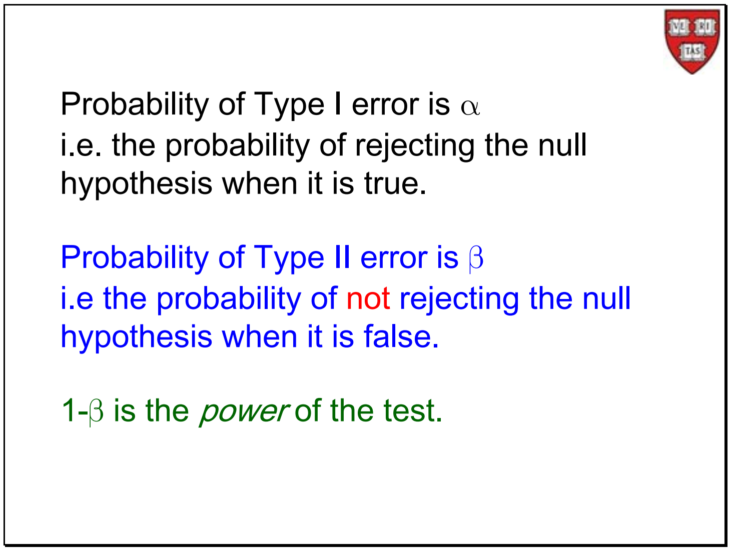
Us	Population	
	$\mu = \mu_0$	$\mu \neq \mu_0$
Not reject	√	Type II
Reject	Type I	√

In hypothesis testing, rather than decide whether a person committed a crime, we decide whether the mean of the population is equal to μ_0. For example, is $\mu = 237$?

To help us decide we take a sample (evidence) and we replace the trial with our analysis of the sample, and we replace the jury, and we have to make a decision about the population. We are faced with a hypothesis that the mean of the population is a given amount. And we are going to decide to reject our hypothesis or not on the basis of a sample from that population.

Now let us look at these possible errors. We label the first one Type 1 and the second one Type 2. Returning to the analogy with the criminal case, the Romans used to say better that 10 guilty men go free than that one innocent person be found guilty. So they had this 10 to 1 ratio of these probabilities of these types of errors. Today we have diluted that a bit, possibly because we value life a little less, but we now say that better that eight guilty men go free than that one innocent person be found guilty.

In hypothesis testing in statistics, we label these probabilities as alpha or beta. So the probability of a Type 1 error is alpha. And the probability of a Type 2 area is beta.



Probability of Type I error is α
i.e. the probability of rejecting the null hypothesis when it is true.

Probability of Type II error is β
i.e the probability of not rejecting the null hypothesis when it is false.

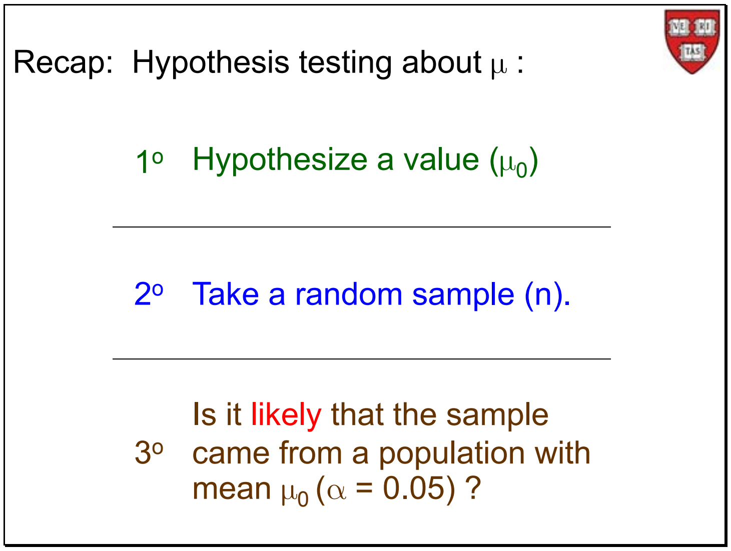
1-β is the *power* of the test.

So the probability of rejecting the null hypothesis when it is true is called alpha. And the probability of a Type 2 error, which is not rejecting the null hypothesis when we should be rejecting it, is beta. So in both instances, we'd like to make these as small as possible.

This is very similar, of course, to what we did with diagnostic testing. Except with diagnostic testing, we took a more positive view on life. And we spoke about sensitivity and specificity, the probabilities of doing the right thing, whereas here we label the probabilities of doing the wrong things. Ideally, we wanted both the sensitivity and the specificity to be one.

Here in hypothesis testing, we of course want to minimize our chances of making mistakes so ideally we would want both alpha and beta to be zero.

In terms of beta, sometimes we look at 1 minus beta. And that is called the *power* of the test. We delve more into the power issue, later.



Recap: Hypothesis testing about μ :

1° Hypothesize a value (μ_0)

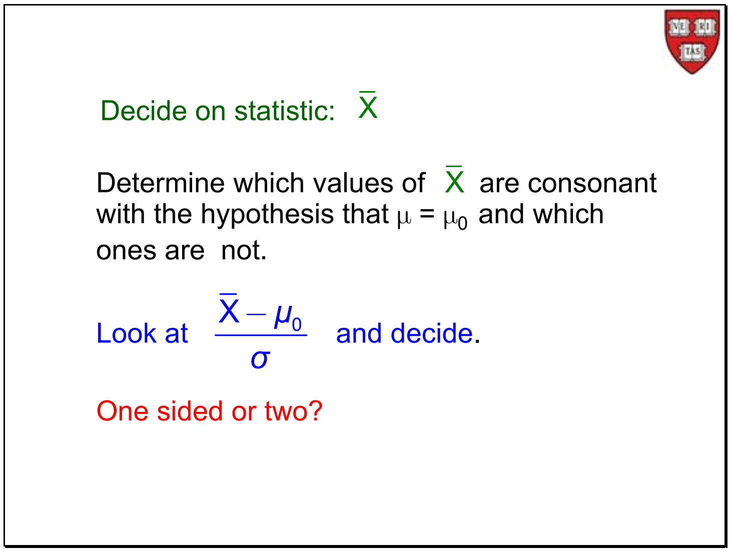
2° Take a random sample (n).

3° Is it likely that the sample came from a population with mean μ_0 ($\alpha = 0.05$) ?

So to recap, we looked at hypothesis testing about mu. The first thing we did was we hypothesized a particular value for μ in the population. For example, this may represent the status quo. We call that $\mu_{0,}$ naught for the null hypothesis.

Then we take a random sample of size n from that population, and ask the question, is it likely that the sample came from a population with this hypothesized mean, μ_0.

We do not want to make it unlikely that we reject this hypothesis, when in fact, it is not true. We can set it up so we do not reject it 95% of the time, say, when it is true. Or, another way of saying the same thing, we only want to run a risk of α = 0.05, that is a 5% chance of making this mistake, when we should not.



Decide on statistic: $\bar{X}$

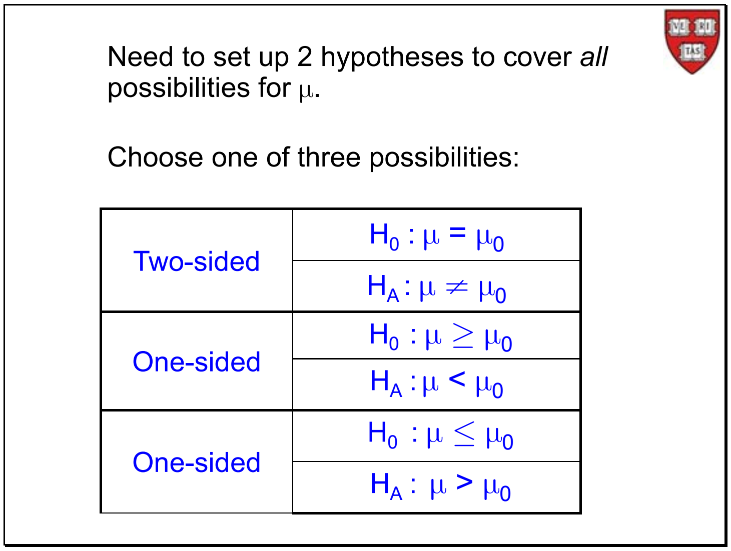
Determine which values of $\bar{X}$ are consonant with the hypothesis that $\mu = \mu_0$ and which ones are not.

Look at $\dfrac{\bar{X} - \mu_0}{\sigma}$ and decide.

One sided or two?

So we first decide on the statistic to use. Since we are talking about the population mean μ, the statistic that we might decide on is the sample mean, X-bar. We saw, from the central limit theorem, that the mean of the sampling distribution of the sample mean is the population mean.

This allows us to determine which values of X-bar are consonant with the null-hypothesis and which are less likely. A predictive interval would be one way of doing that. Another way of determining it is to look at the value of the standardized variable. If it is too large, then reject the null hypothesis.



Need to set up 2 hypotheses to cover *all* possibilities for μ.

Choose one of three possibilities:

Two-sided	$H_0 : \mu = \mu_0$
	$H_A : \mu \neq \mu_0$
One-sided	$H_0 : \mu \geq \mu_0$
	$H_A : \mu < \mu_0$
One-sided	$H_0 : \mu \leq \mu_0$
	$H_A : \mu > \mu_0$

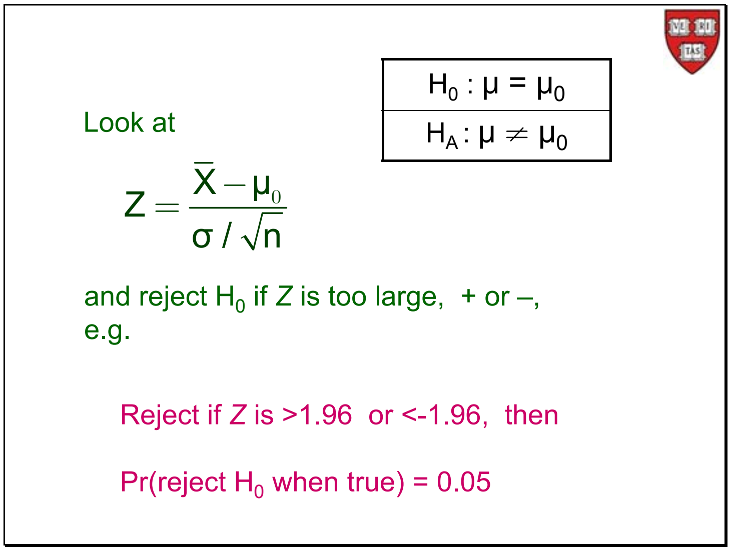
Now we need to be careful deciding whether the standardized variable is too “large”. To determine how to operationalize this we need to look at three possibilities when looking at the pair—namely, the null and alternative hypotheses.

If we fall into the category we call “two sided” as shown in the slide above, then a value outside our usual (-1.96, 1.96) interval would only happen 5% of the time if our null hypothesis is true. There is no directionality implied in our null hypothesis.

On the other hand, if we have the first set of “one sided” hypotheses, as labeled above, then a standardized variable that is positive would be consonant with the null hypothesis. So it could not be “too large” in the positive direction, and we would seek too large in the negative direction to reject the null hypothesis. The fact that α=0.05, allows us to reject the null hypothesis even if we have a small negative value for the standardized statistic.

In contrast to the previous situation, if we have the third set of hypotheses, above, then we would seek too large a positive value of the standardized statistic to reject the null hypothesis.

This is the issue of whether we have one-sided or two-sided hypotheses. The decision of where we are in this trichotomy should not be influenced by the data. Ideally, we should make our decision before even looking at the data at all. The decision should be made on scientific considerations. The theory we have presented here does not support decisions of sidedness based on having seen the data.



$H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

Look at

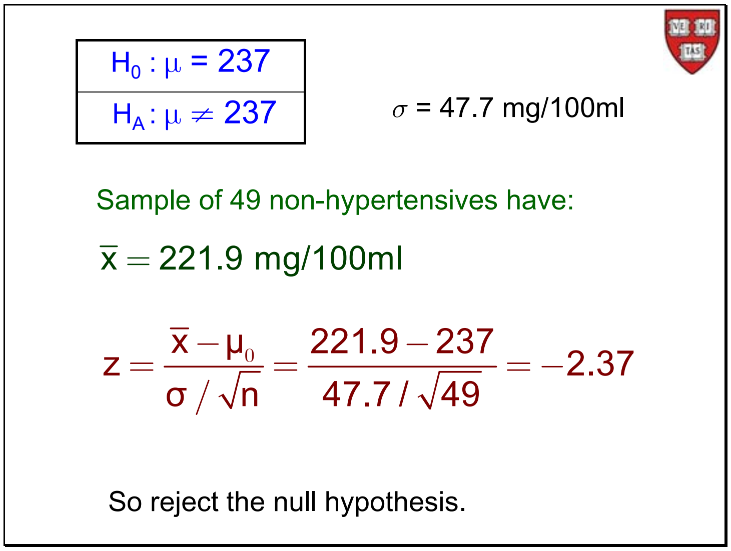
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

and reject H_0 if *Z* is too large, + or –, e.g.

Reject if *Z* is >1.96 or <-1.96, then

Pr(reject H_0 when true) = 0.05

Let us take a look at a couple of applications of hypothesis testing. Suppose we want to test in the two-sided framework, that $\mu = \mu_0$. Then we set up our Z, or t if we do not know σ. We reject the null hypothesis if Z is too large. That means if |Z| > 1.96. That, of course presupposes that our $\alpha = 0.05$.



$H_0 : \mu = 237$

$H_A : \mu \neq 237$

σ = 47.7 mg/100ml

Sample of 49 non-hypertensives have:

$$\bar{x} = 221.9 \text{ mg/100ml}$$

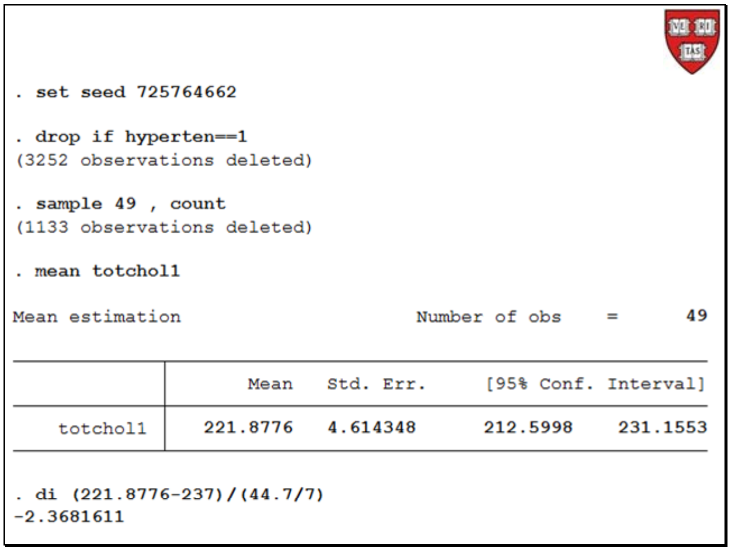
$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{221.9 - 237}{47.7 / \sqrt{49}} = -2.37$$

So reject the null hypothesis.

So for example, if we go back to our population, and we stated that the total cholesterol level in our population in Framingham at visit one was 237 with a standard deviation of 47.7.

Now, I took a sample of 49 non-hypertensives from this population. If I'm not mistaken, I used the random seed from the bill like before, and I got a sample mean of 221.9. Now the question is, would you believe that this was a sample from the general population; namely, a population with a mean of 237, and a standard deviation of 47.7?

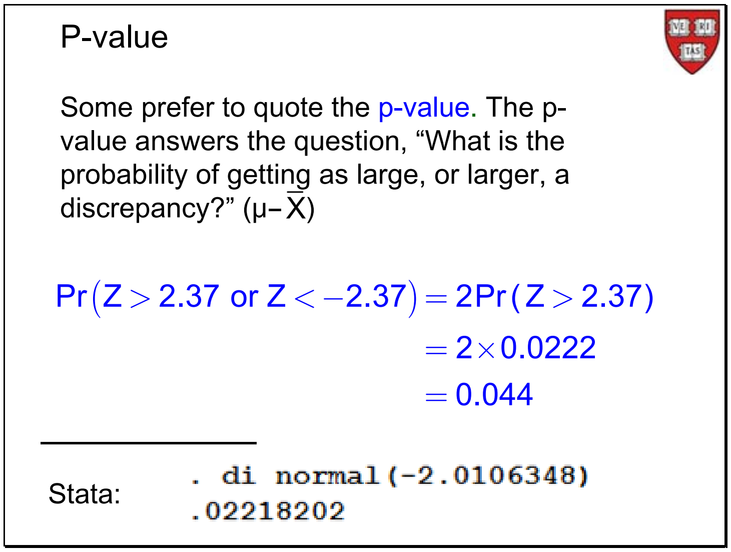
We calculate the standardized variable: the observed sample mean was 221.9, the population mean was 237, the standard deviation was 47.7, and the sample size was 49. Thus Z is -2.37 which is less than minus 1.96, and so we reject the null hypothesis.



```
. set seed 725764662

. drop if hyperten==1
(3252 observations deleted)

. sample 49 , count
(1133 observations deleted)

. mean totchol1

Mean estimation                     Number of obs    =       49

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    totchol1 |   221.8776   4.614348      212.5998    231.1553
--------------------------------------------------------------

. di (221.8776-237)/(44.7/7)
-2.3681611
```

So on the basis of these 49 non-hypertensives, we reject the hypothesis that they come from a population with this high a cholesterol level.

What you do with the information is up to you, but the formalism of the hypothesis testing makes us reject the null hypothesis.



P-value

Some prefer to quote the p-value. The p-value answers the question, “What is the probability of getting as large, or larger, a discrepancy?” $(\mu - \overline{X})$

$$\begin{aligned}\Pr(Z > 2.37 \text{ or } Z < -2.37) &= 2\Pr(Z > 2.37) \\ &= 2 \times 0.0222 \\ &= 0.044\end{aligned}$$

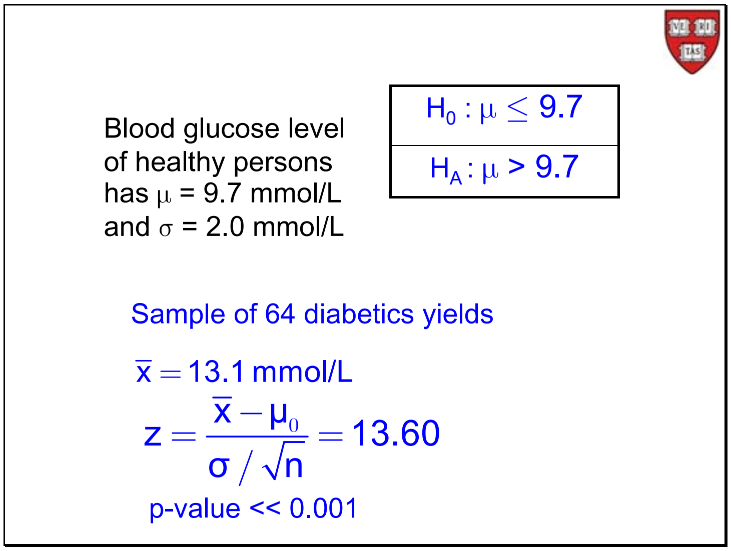
Stata:

```
. di normal(-2.0106348)
.02218202
```

Rather than, looking to see whether the Z is bigger than 1.96 or smaller than -1.96 to reject the null hypothesis, some people prefer quoting what is called the p-value. The p-value answers the question, what is the probability of observing as large or larger a discrepancy than the one I observed? In this case, after we standardize things, our standardized value was 2.37. So the p-value will answer the question of what is the probability of getting a deviation as large or larger than 2.37 after we standardize? Because this is a two-sided calculation, that means we want to know the probability that our standard, Z, is bigger than 2.37 or less than minus 2.37. And the answer is p = 0.044.

Some then argue that since the p-value is less than 0.05, we should reject the null hypothesis.

It is up to you how you set these critical points and to defend your choice. The choice of 0.05 is quite ubiquitous, and we see it everywhere in science.



Blood glucose level of healthy persons has μ = 9.7 mmol/L and σ = 2.0 mmol/L

$H_0 : \mu \leq 9.7$
$H_A : \mu > 9.7$

Sample of 64 diabetics yields

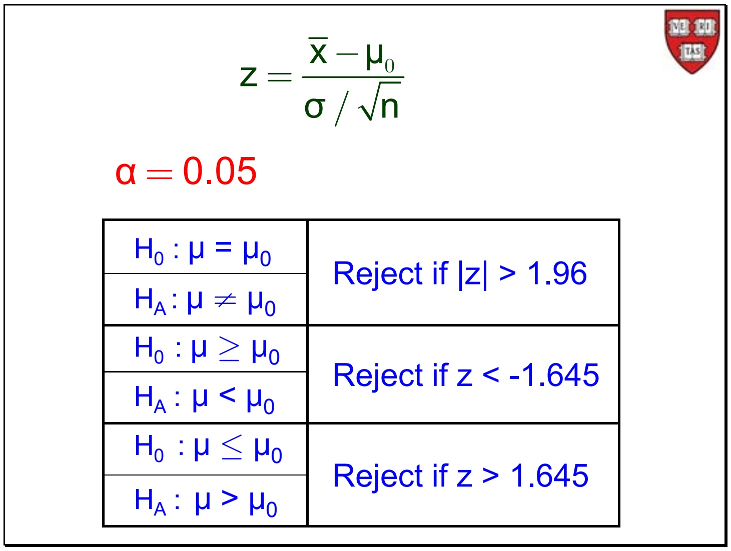
$\bar{x} = 13.1$ mmol/L

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = 13.60$$

p-value << 0.001

Here is another example. This time we are looking at blood glucose levels of healthy persons. If we look at a sample of 64 diabetics, then we are in a one-sided hypothesis. If the mean is less than or equal to 9.7, we will be perfectly happy, but we are concerned about whether it's bigger than 9.7.

In this case, of course, it doesn't make any difference, because the Z value is 13.6 and this would be rejected as either a one- or two-sided test statistic. So whether we did a one-sided or two-sided, test really makes no difference.



$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

$\alpha = 0.05$

$H_0 : \mu = \mu_0$ $H_A : \mu \neq \mu_0$	Reject if $\lvert z \rvert > 1.96$
$H_0 : \mu \geq \mu_0$ $H_A : \mu < \mu_0$	Reject if $z < -1.645$
$H_0 : \mu \leq \mu_0$ $H_A : \mu > \mu_0$	Reject if $z > 1.645$

The issue about one-sided or two-sided tests is quite controversial in the field. We see that Z has to be larger to reject a two sided hypothesis than a one-sided one. Since some journals would prefer to publish "significant" results than non-significant—man bites dog, rather than dog bites man, sort of thing—and since publications are desirable for promotions and advancement in general, if an author tries to publish a one-sided hypothesis test, he or she runs the risk of being accused of all sorts of dastardly deeds, and editors will have none of it. This is not just an academic controversy, as some in the pharmaceutical industry have argued that the FDA should consider one-sided drug testing.

We should also point out that a similar development with one and two sided confidence intervals can be carried out, so that does not overcome the objections.