PSY 201: Statistics in Psychology

Lecture 17

Sampling distribution of the mean

Marvel at my predictive powers!

Greg Francis

Purdue University

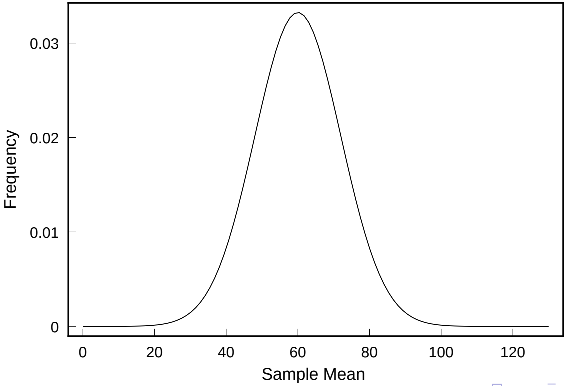
Fall 2019

SAMPLING

- suppose we have a **population** with a mean μ and a standard deviation σ
- suppose we take a **sample** from the population and calculate a sample mean $\overline{X}_1$
- suppose we take a **different** sample from the population and calculate a sample mean $\overline{X}_2$
- suppose we take a **different** sample from the population and calculate a sample mean $\overline{X}_3$

DISTRIBUTION

- the different $\overline{X}_i$ sample means that are calculated will be related to each other because they all come from the same population, which has a population mean of μ
- we can consider a **distribution** of the sample means (same idea as distribution of sum of dice roles)



DISTRIBUTION

- this distribution involves frequencies of **means** rather than frequencies of **scores**
- for most of inferential statistics we do **not** deal with the frequency distribution of scores
- A sampling distribution is the underlying distribution of values of the statistic under consideration, from all possible samples of a given size.
- currently, the statistic is the sample mean $\overline{X}$

SAMPLING DISTRIBUTION

- how do we get the sampling distribution?
- e.g., suppose you have a population of 5 people with math scores
  - and you take sample sizes of 3
- you must consider every possible group of 3 people from the population
  - turns out there are 10 such groups
- NOTE: the number of samples is greater than the size of the population!

CENTRAL LIMIT THEOREM

- fortunately, there are theorems that tell us what the distribution will look like
- as the sample size (n) increases, the sampling distribution of the mean for simple random samples of n cases, taken from a population with a mean equal to μ and a finite variance equal to σ^2, approximates a normal distribution
- another theorem based on unbiased estimation tells us that the mean of the **sampling distribution** is μ

STANDARD ERROR

- theorems on unbiased estimates also give us the sampling distribution variance and standard deviation
- denote the sampling distribution variance as

$$\sigma^2_{\overline{X}}$$

- it turns out that

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

- where
  - σ^2 = variance in the population
  - n = size of sample

STANDARD ERROR

- of course the standard deviation of the sampling distribution is the square root of the variance

$$\sigma_{\overline{X}} = \sqrt{\sigma^2_{\overline{X}}}$$

- or

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

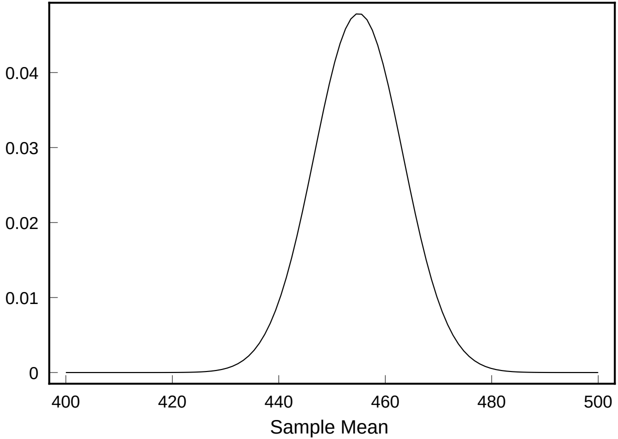
- also called the **standard error of the mean**

WHY BOTHER?

- suppose you know that for a population, $\mu = 455$ and $\sigma = 100$ (an example involving SAT scores)
- then we know the following about a **sampling distribution** involving samples sizes of 144 students
    - The distribution is normal.
    - The mean of the distribution is 455.
    - The standard error of the mean is $100/\sqrt{144} = 8.33$.

WHY BOTHER?

- this is something we can work with!



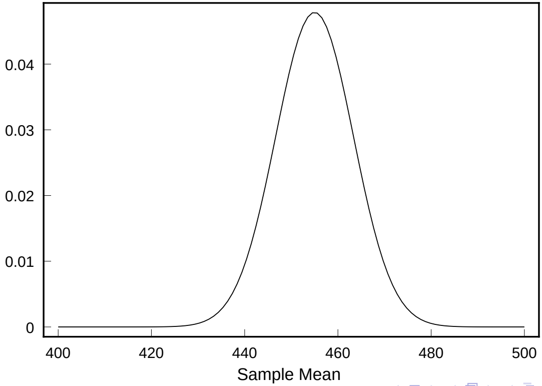
- calculate percentages, proportions, percentile ranks

PROBABILITY

- we can answer questions like
- what is the probability of randomly selecting a sample with a mean $\overline{X}$ such that

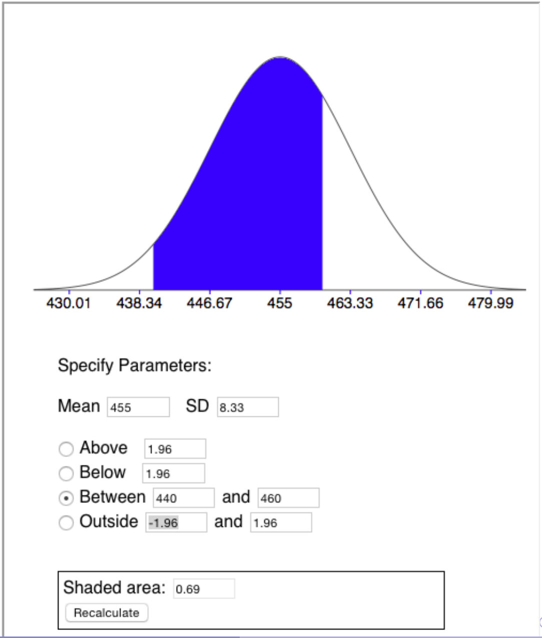
$$440 < \overline{X} < 460 \text{ ?}$$

- area under the curve



PROBABILITY

- everything is just like before
- area under the curve
- We use the normal distribution calculator with Mean=455 and SD=8.33



SAMPLING DISTRIBUTION

- the sampling distribution has two critical properties
  - As sample size (n) increases, the sampling distribution of the mean becomes more like the normal distribution in shape, even when the population distribution is not normal.
  - As the sample size (n) increases, the variability of the sampling distribution of the mean decreases (the standard error decreases).

SHAPE

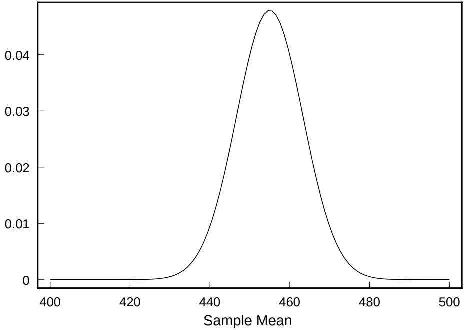
- with large sample sizes, all sampling distributions of the mean look like normal distributions
- means the conclusions we draw from sampling distributions are not dependent on the shape of the population distribution!
- a remarkable result that is due to the central limit theorem

VARIABILITY

- from our calculation of standard error:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- we see that increasing n makes for smaller values of $\sigma_{\overline{X}}$
- e.g. for $n = 144$ in our previous example $\sigma_{\overline{X}} = 8.33$

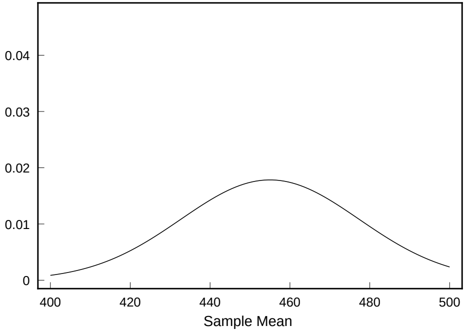


VARIABILITY

- but if $n = 20$,

$$\sigma_{\overline{X}} = \frac{100}{\sqrt{20}} = 22.36$$

- compare to the 8.33 with $n = 144$

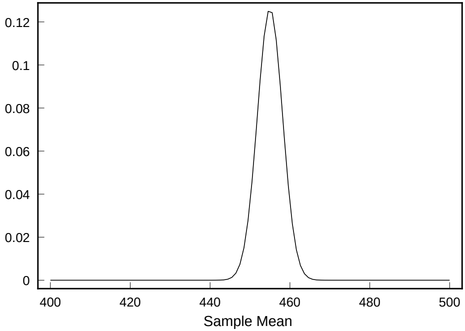


VARIABILITY

- OR if $n = 1000$,

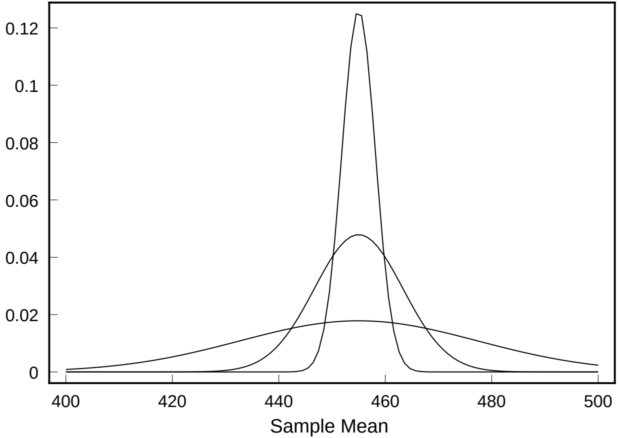
$$\sigma_{\overline{X}} = \frac{100}{\sqrt{1000}} = 3.16$$

- compare to the 8.33 with $n = 144$



VARIABILITY

- increasing the sample size decreases the variability of sample means
- makes sense if you think about it



SAMPLING

- to use the sampling distribution like we want to, we must have **random** samples
- without random sampling, our calculations about probability of sample means are not valid (this will get more important later)
- lots of methods of sampling that emphasize different aspects of the data

WHY STATISTICS WORKS

- we have two ways of finding the sampling distribution of the mean
  - gather lots of samples, calculate means and standard deviations (virtually impossible)
  - calculate mean and standard deviation of the population, use central limit theorem (relatively easy)
- the central limit theorem allows us to do inferential statistics, without it, much of this course would not exist (actually there is one other way to do statistics...)

EXAMPLE

- let's create a sampling distribution
- two things
  - Write down the height of your father (in inches) on the papers going around the room.
  - Sample the height measure of 10 people close to you.

EXAMPLE

- I'll sit down and calculate the **population** mean (μ) and standard deviation (σ)
- you calculate the **sample** mean ($\overline{X}$) for the 10 scores you have

$$\overline{X} = \frac{\Sigma X_i}{10}$$

EXAMPLE

- OK, I get

$$\mu = \frac{\Sigma X_i}{N} =$$

$$\sigma = \sqrt{\frac{\Sigma(X_i)^2 - \left[\left(\Sigma X_i\right)^2 / N\right]}{N}} =$$

EXAMPLE

- with this information, I can **predict** the frequency of sample means each of you calculated
- I predict that most of you calculated sample means close to

$$\overline{X} = \mu =$$

- moreover, I predict that the distribution of sample means is **normal**
- lets plot the sample means you calculated

EXAMPLE

- Let's calculate the standard deviation of the sampling distribution of the mean heights as

$$\sigma_{\overline{X}} = \sqrt{\frac{\Sigma(\overline{X})^2 - \left[\left(\Sigma\overline{X}\right)^2/N\right]}{N}} =$$

- I predict that it will be very close to

$$\frac{\sigma}{\sqrt{10}} =$$

CONCLUSIONS

- sampling distribution of the mean looks like a normal distribution
- methods of calculating mean and standard deviation **if** μ and σ are known
- samples must be randomly selected

NEXT TIME

- hypothesis testing
- using the sampling distribution (in what looks to be reverse!)
- null hypothesis

Why I don't use herbal medicines.