Sampling distributions & the sampling distribution of the mean

Lectures 11 & 12

Readings: GW 7 & SS 7.4

Parameters and statistics

- Population - Entire group of items/individuals we want information about.
- Sample - The part of the population we actually examine in order to gather information.
- A parameter is a number that describes the population. It is fixed, but we don't know its value.
- A statistic is a number that describes a sample. Its value is known, but it varies from sample to sample.
- We often use statistics to estimate the unknown parameter

Statistical Inference

- Statistical inference draws conclusions about a population on the basis of data from a sample.
- It also provides us with a statement of how much confidence we can place in our conclusions.
- We are in many cases interested in the mean value a variable takes in the population.

Efficiency of statistics

• Individual scores are random draws from a population

• The sample mean is a guess about the true population mean

• But how accurate (or efficient) is the sample mean?

• Or, I could say, what is the standard deviation of the sample mean

• I want to estimate the SD of the mean of *n* observations, i.e., how much the mean is expected to vary from sample to sample

•But I only get to observe *one* sample

The sampling distribution

Imagine that you could draw a sample and calculate a mean or median or SD or whatever statistic again and again from a population.

What would that distribution of this statistic look like?

You're conceptualizing a sampling distribution.

What is its expected value and standard deviation?

If you know this, you can answer how likely it is that a sample with a given mean (or median or SD) was drawn from a population with known mean (or median or SD)

The sampling distribution

- …is a distribution of sample statistics (means, medians, etc.)
- …is a theoretical distribution that describes all possible means, medians, etc., and the probability of obtaining each value.
- …can be visualized using simulations, but must be imagined when collecting real data.

Three “Amazing” Facts about Sampling Distributions of the Mean

1. They are approximately normal
   - **When data in population are normally distributed and even if they are not, assuming large n**
2. They are centered at μ of the population they are drawn from
   - **Mean is unbiased**
3. Their standard deviation equals the standard deviation of the individual scores divided by the square root of the sample size (<u>standard error of the mean</u>)

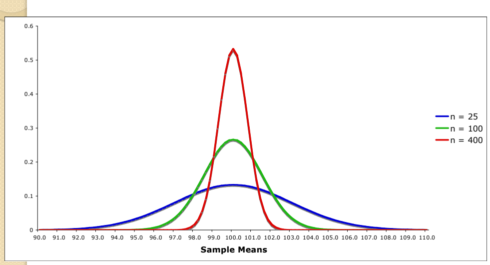
$$SEM = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distributions as a function of *n*

- Assume IQ: $\mu = 100;\ \sigma = 15$
- Sampling Distribution of Sample Means if n = 25
  - Normal $E(\bar{X}) = 100; \sigma_{\bar{X}} = \frac{15}{\sqrt{25}} = 3$
- Sampling Distribution of Sample Means if n = 100
  - Normal $E(\bar{X}) = 100; \sigma_{\bar{X}} = \frac{15}{\sqrt{100}} = 1.5$
- Sampling Distribution of Sample Means if n = 400
  - Normal $E(\bar{X}) = 100; \sigma_{\bar{X}} = \frac{15}{\sqrt{400}} = 0.75$

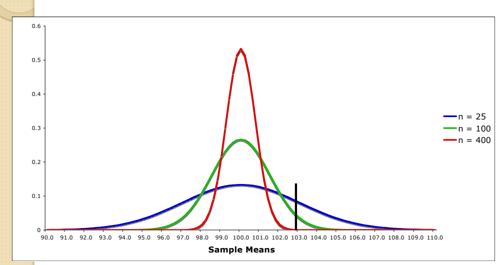
Sampling Distributions of the Mean

$\mu = 100;\ \sigma = 15;\ n = 25,\ 100,\ 400$



Sampling Distributions of the Mean

How likely is a Sample Mean IQ of 103 or greater?



Sampling Distributions of the Mean

How likely is a Sample Mean IQ of 103 or greater?

- n = 25 $z_{103} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{103 - 100}{3} = 1.00$ p = .1587
- n = 100 $z_{103} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{103 - 100}{1.5} = 2.00$ p = .0228
- n = 400 $z_{103} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{103 - 100}{.75} = 4.00$ p < .0001

Sampling Distributions of the Mean

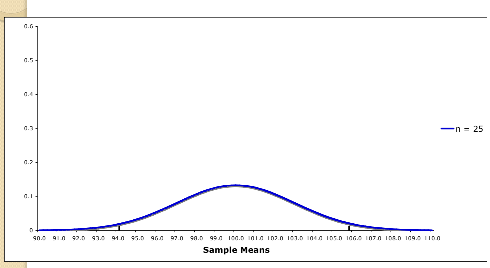
What Values of the Mean would occur 95% of the time?

- What Z score in a normal distribution separates the most extreme 5% of the scores from the middle-most 95% of the scores?
- ±1.96
- n = 25; standard error of the mean = 3.00

$$\pm 1.96 = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - 100}{3}$$
$$\bar{X} < 100 - 1.96(3.00) = 94.12$$
$$\bar{X} < 100 + 1.96(3.00) = 105.88$$

Sampling Distributions of the Mean

What Values of the Mean would occur 95% of the time?



Sampling Distributions of the Mean

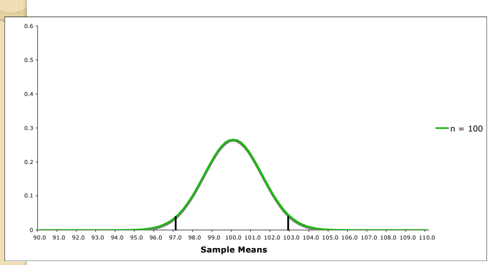
What Values of the Mean would occur 95% of the time?

- What Z score in a normal distribution separates the most extreme 5% of the scores from the middle-most 95% of the scores?
- ±1.96
- n = 100, standard error of the Mean = 1.50

$$\pm 1.96 = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - 100}{1.50}$$
$$\bar{X} < 100 - 1.96(1.50) = 97.06$$
$$\bar{X} < 100 + 1.96(1.50) = 102.94$$

Sampling Distributions of the Mean

What Values of the Mean would occur 95% of the time?



Sampling Distributions of the Mean

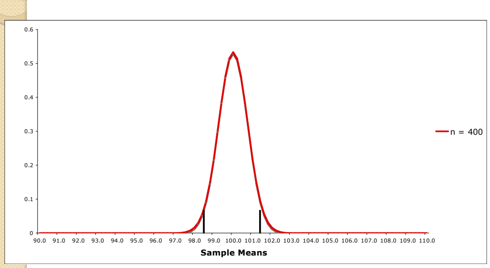
What Values of the Mean would occur 95% of the time?

- What Z score in a normal distribution separates the most extreme 5% of the scores from the middle-most 95% of the scores?
- ±1.96
- n = 400 , standard error of the Mean = 0.75

$$\pm 1.96 = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - 100}{.75}$$
$$\bar{X} < 100 - 1.96(.75) = 98.53$$
$$\bar{X} < 100 + 1.96(.75) = 101.47$$

Sampling Distributions of the Mean

What Values of the Mean would occur 95% of the time?



Deriving the standard error of the mean

This section is for your own edification regarding *why* SEM = SD/sqrt(n).

You will not be tested on it.

Background on random variables

If X is a random variable, var(X) is its variance
Sum of two variables X1 and X2 = X1 + X2

Variance sum law:
Var(X1 + X2) = var(X1) + var(X2)
Var(X1 - X2) = var(X1) + var(X2)

Constant multiplication rule:
If I multiply a random variable X by 2, I get 2X
Var(2X) = 2^2 * var(X)
Var(aX) = a^2var(X)

The sampling distribution for a mean

- Imagine I measure two subjects x1 and x2
- They are drawn from random variables X1 and X2, respectively
- I assume they come from identical distributions
- Their mean, or the sample mean, is (X1 + X2) / 2
- What is the variance of that sample mean?
- This tells me how accurate the sample mean is.
- Why? Sqrt(var) = st. deviation = how far off the true mean I typically am

$$find : \mathrm{var}(\frac{X1+X2}{2})$$

SEM derivation

Variance of sampling distribution (for mean)

Assume independent X1 and X2!

$$\mathrm{var}(\frac{X1+X2}{2}) = (1/2)^2(\mathrm{var}(X1)+\mathrm{var}(X2))$$

$$\mathrm{var}(X1) = \mathrm{var}(X2) = \mathrm{var}(X)$$

Assume X1 & X2 have identical distribution, with same variance!

$$\mathrm{var}(\frac{X1+X2}{2}) = (1/2)^2 * 2 * \mathrm{var}(X) = \frac{2\,\mathrm{var}(X)}{4}$$

$$define : \mathrm{var}(X1) = \mathrm{var}(X2) = \sigma_X^2$$

$$\mathrm{var}(\frac{X1+X2}{2}) = \frac{\sigma^2{}_X}{2}$$

Variance of sampling distribution for mean of 2 subjects

SEM derivation

Standard deviation of sampling distribution (for mean)

$$\mathrm{var}(\frac{X1+X2}{2}) = \frac{\sigma^2{}_X}{2}$$

$$SD(\frac{X1+X2}{2}) = \sqrt{\frac{\sigma^2{}_X}{2}} = \frac{\sigma_X}{\sqrt{2}}$$

Std. of mean of 2 variables

$$\mathrm{var}(\frac{X1+X2...+X_N}{N}) = \frac{1}{N^2}\mathrm{var}(X1+X2+...X_N) = \frac{1}{N^2}\left[\mathrm{var}(X1)+\mathrm{var}(X2)+...\mathrm{var}(X_N)\right] = \frac{1}{N^2}\left[N\sigma^2\right] = \frac{\sigma^2}{N}$$

Std. of mean of n variables

$$SD(\frac{X1+X2...Xn}{n}) = \frac{\sigma_X}{\sqrt{n}}$$

Each subject is a random variable
--> n subjects

The sampling distribution

This means

If $\hat{\sigma}$, or **s**, is our estimate of the sample standard deviation (average deviation of an individual from the sample mean)

$\frac{\hat{\sigma}}{\sqrt{n}}$ Is our estimate of how far off the sample mean is, on average, from the true population mean

This is the **standard error of the mean**
…our estimate of the standard deviation of the sampling distribution of means