NNSE 784
Advanced Analytics Methods

Instructor: F Doyle (CESTM L210)

MW 4:30 – 5:50, NFN 203

Slide Set #8
Sampling Distributions
&
Central Limit Theorem
Continued

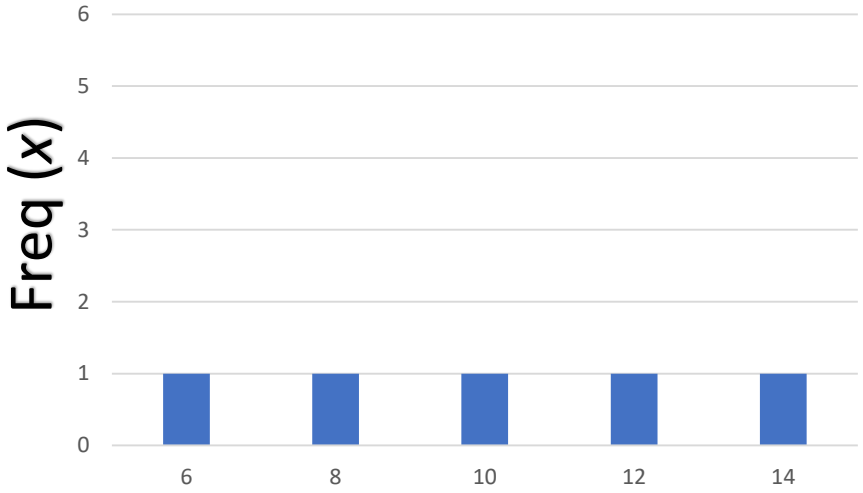
Lecture Outline

- Review Quiz #1
- Normal Distribution as a Probability Distribution - recap
- Central Limit Theorem – recap
- Distribution of the Sample Mean – example problem
- Distribution of the Difference Between Sample Means – introduction via example problem

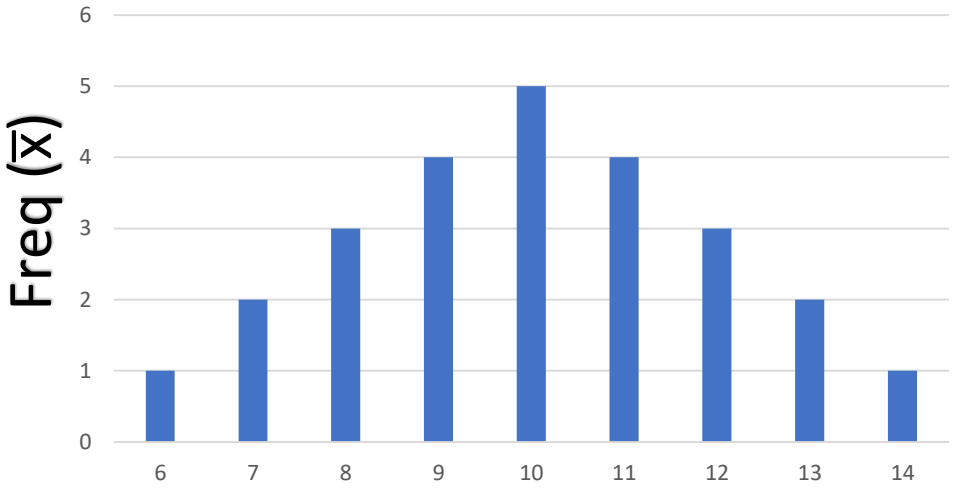
Characteristics of the Normal Distribution

- Symmetrical about its mean, μ (no skew)
- The mean, median and mode are all equal.
- **The area under the curve about the x-axis is one unit (it is a probability distribution)**
- 68-95-99.7 rule:
  - 68% of the area under the curve is within +/- 1 standard deviation
  - 95% of the area under the curve is within +/- 2 standard deviation
  - 99.7% of the area under the curve is within +/- 3 standard deviation
- Completely determined by parameters μ and σ
  - Different values of μ shift the distribution left or right on the x axis
  - Different values of σ determine the spread of the distribution

Distribution of Sample Mean - recap



Distribution of population

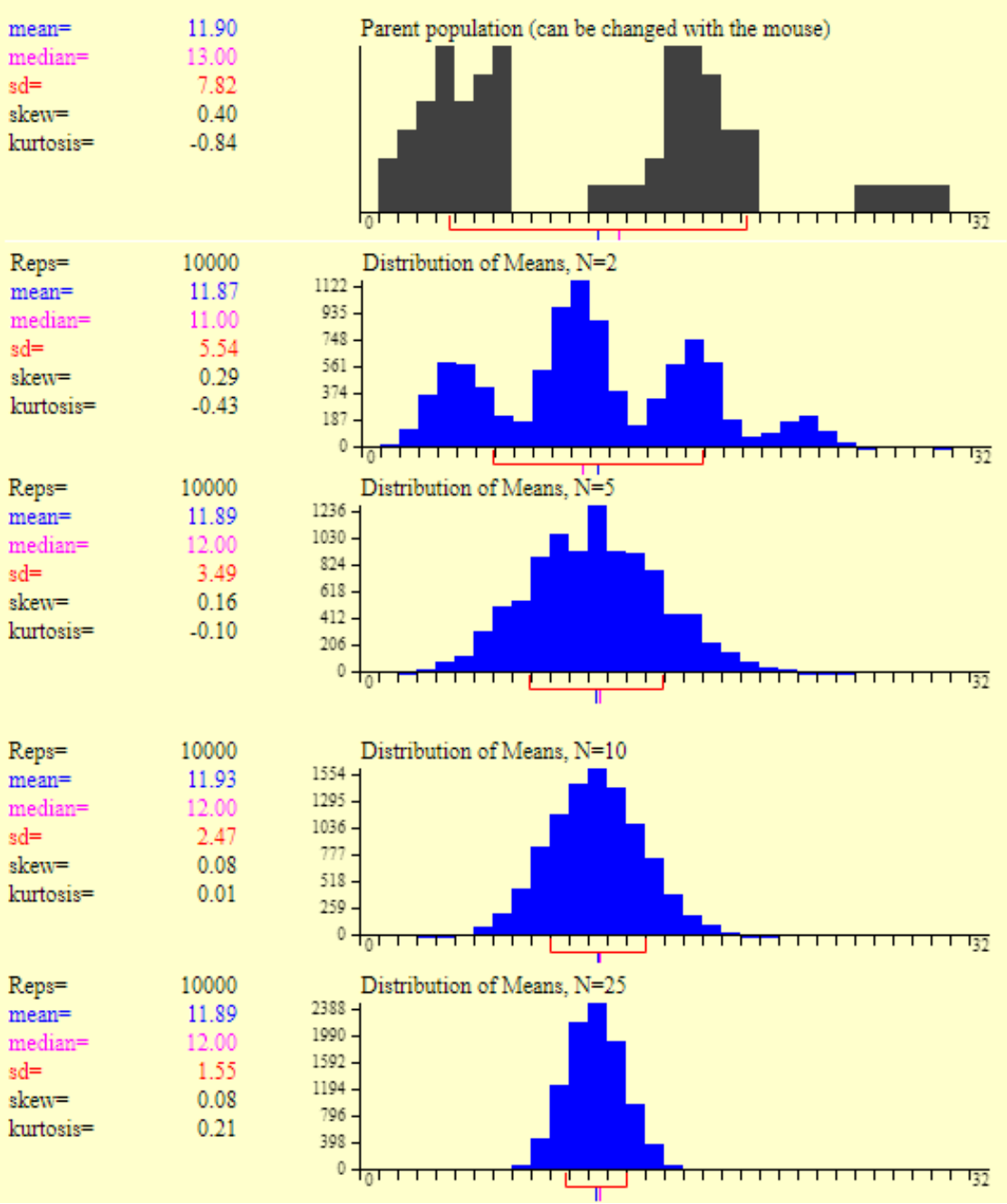


Sampling distribution of $\bar{x}$

($n = 2$)

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N^n} = \frac{6 + 7 + 7 + 8 + \cdots + 14}{25} = \frac{250}{25} = 10$$

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$



Central Limit Theorem - Recap

Not only is the parent population not normal, it is composed of discontinuous blocks!

However, note the shape of the sampling distributions as *n* increases.

If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$. This is true no matter what the sample size n is.

If the population distribution is not Normal, the central limit theorem tells us that the sampling distribution of $\bar{x}$ will be approximately Normal in most cases if $n \geq 30$.

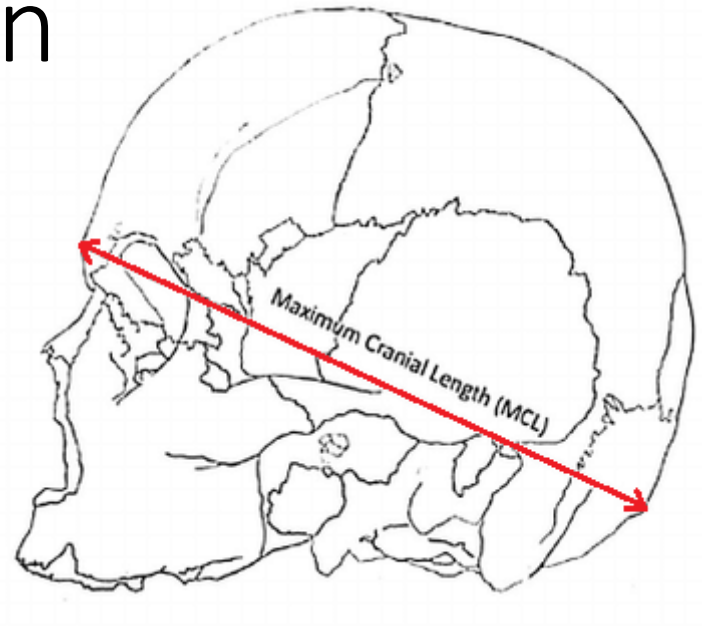
Graphs generated at site: https://onlinestatbook.com/stat_sim/sampling_dist/index.html

Distribution of the Difference Between Sample Means

What Is the Point?

- We've looked at:
  - Descriptive statistics
  - Basic probability
  - One sampling distribution – the sampling distribution of the mean
- We are inching toward Inferential Statistics which is all about determining ("inferring") aspects of populations (parameters of and relationships between) with some associated probability, by analyzing samples of those populations
- We are not quite there yet, but we can start putting pieces together
- The simplest application of what we've learned so far, with respect to the distribution of the sample mean, is determining the probability of obtaining a sample with a mean of some specified magnitude from a known population.

Sampling Distribution of the Mean Example Problem



Suppose it is known that in a certain large human population cranial length is approximately normally distributed with a mean of 185.6 mm and a standard deviation of 12.7 mm.

What is the probability that a random sample of size 10 from this population will have a mean greater than 190?

We know that we can use a z-score to get a probability for a value in normal population.

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

Example Problem

Population mean = 185.6 mm

Standard deviation = 12.7 mm.

What is the probability that a random sample of size n=10 from this population will have a mean greater than 190?

We need a z-score for a value in the sampling distribution.

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

We know distribution of sampling mean has variance: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ **therefore** $\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

Example Problem

Population mean = 185.6 mm

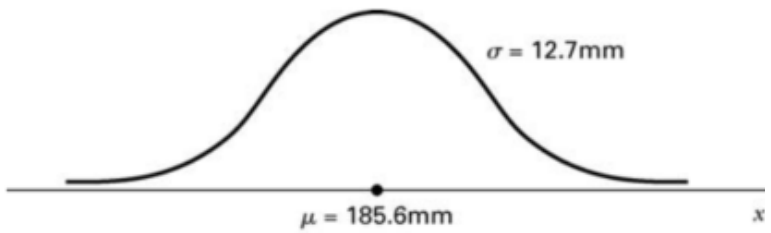
Standard deviation = 12.7 mm.
What is the probability that a random sample of size n=10 from this population will have a mean greater than 190?

We need a z-score for a value in the sampling distribution.

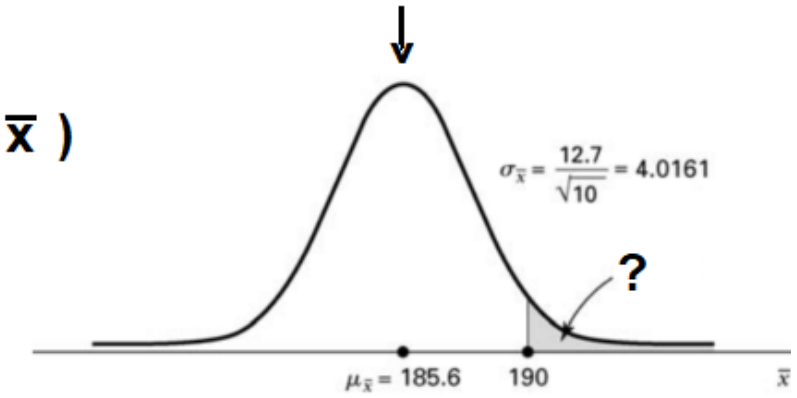
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}} = \frac{190 - 185.6}{4.0161} = \frac{4.4}{4.0161} = 1.10$$

Example Problem

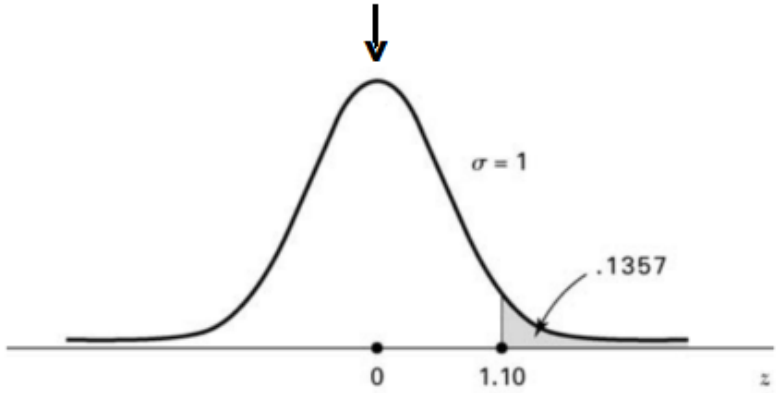
Population Distribution



Sampling Distribution (for $\bar{x}$)



Standard Normal Distribution




```
from scipy.stats import norm
# Calculate the cumulative probability
# of a z-score less than 1.10 and
#subtract it from 1 as we are looking
#for the probability of getting a
#result that high or higher
prob = 1-norm.cdf(1.10)

# Print the p-value
print(prob)
```

```
0.13566606094638267
```

Distribution of the Difference Between Two Sample Means

- Often, we are interested in two populations.
- Specifically, we may want to know something about the difference between two population means (for example… think about the experimental populations of treatment and control)
- Knowledge regarding sampling distributions of the difference between two means is useful for this type of analysis.

Example Problem

- Suppose we have two populations of individuals—one population (population 1) has experienced some condition thought to be associated with mental retardation, and the other population (population 2) has not experienced the condition. The distribution of intelligence scores in each of the two populations is believed to be approximately normally distributed with a standard deviation of 20.

- Suppose, further, that we take a sample of 15 individuals from each population and compute for each sample the mean intelligence score with the following results: $\bar{x}_1 = 92$ and $\bar{x}_2 = 105$. If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger ($\bar{x}_1 - \bar{x}_2$) between sample means?

Problem (and much of the other statistics materials) taken from: Biostatistics – A Foundation for Analysis in the Health Sciences, 10th Ed. Year 2013, Authors Daniel and Cross, Publisher Wiley, Hoboken NJ

Difference Between Sample Means

- To answer this, we need to know the sampling distribution of the relevant statistic, the *difference between two sample means,* $(\bar{x}_1-\bar{x}_2)$
- We saw in the last lecture that a population of size N=5 sampled at size n=2, yields 25 possible sample combinations ($5^2=25$)
- If we were to look at a second population (same N and n) and take all combinations of samples between the two, we would have 625 (25*25 =625). Here we are looking at populations of unknown, but assumedly larger sizes and sample sizes of 15. We will therefore not build the full distribution, but rather attempt to conceptualize the process by which it would be built.

Example Table for Constructing the Sampling Distribution of $\bar{x}_1 - \bar{x}_2$

Working Table for Constructing the Distribution of the Difference Between Two Sample Means

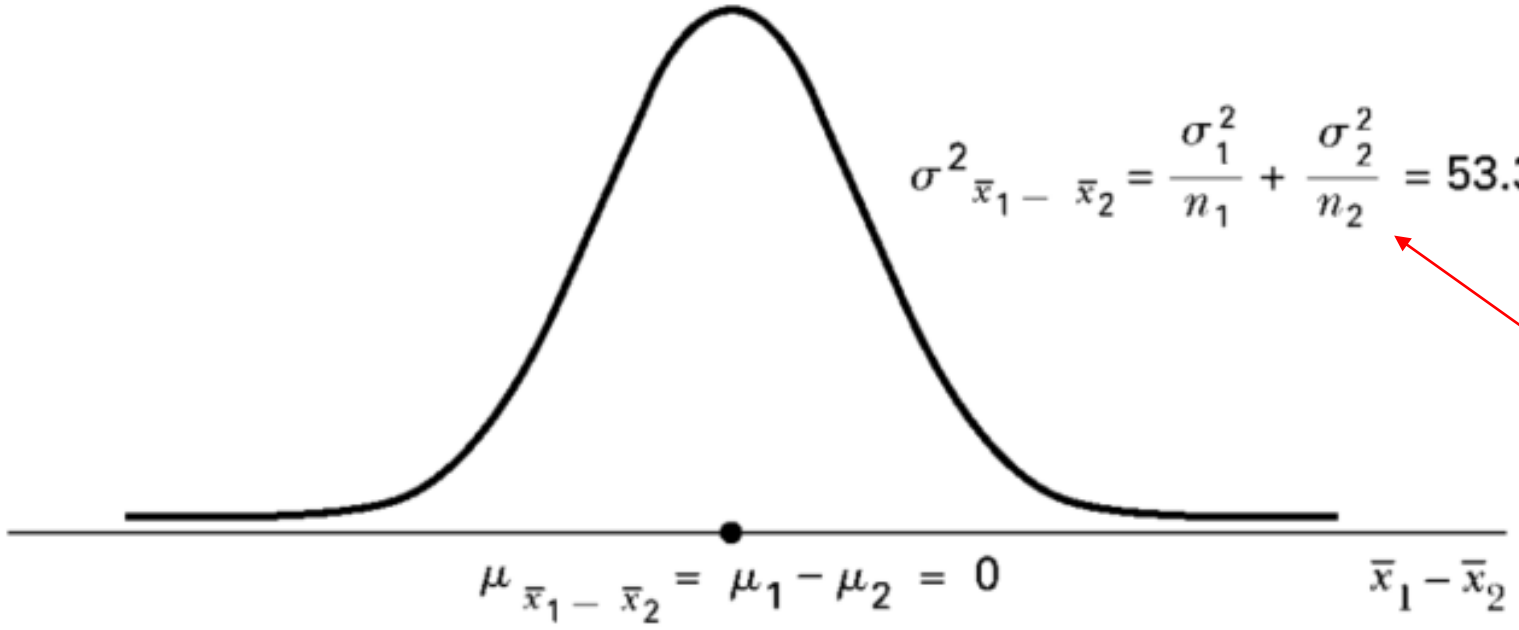
Samples from Population 1	Samples from Population 2	Sample Means Population 1	Sample Means Population 2	All Possible Differences Between Means
n_{11}	n_{12}	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{11} - \bar{x}_{12}$
n_{21}	n_{22}	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{11} - \bar{x}_{22}$
n_{31}	n_{32}	$\bar{x}_{31}$	$\bar{x}_{32}$	$\bar{x}_{11} - \bar{x}_{32}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$n_{N_1} C_{n_1} 1$	$n_{N_2} C_{n_2} 2$	$\bar{x}_{N_1} C_{n_1} 1$	$\bar{x}_{N_2} C_{n_2} 2$	$\bar{x}_{N_1} C_{n_1} 1 - \bar{x}_{N_2} C_{n_2} 2$

Just sample numbers, don't get hung up on the order

Population specifier, not exponent!

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$: Characteristics

In our problem:
$\bar{x}_1 = 92$
$\bar{x}_2 = 105$
Difference = -13



Remember, we said:
$\sigma = 20$ and $n = 15$

- The mean equals $\mu 1 - \mu 2$ (the difference between the two population means)
  - In this case the plot reflects no difference between these
- The variance equals $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$
  - The overall variance of the sampling distribution is affected by both contributing distributions and this is accounted for by summing them

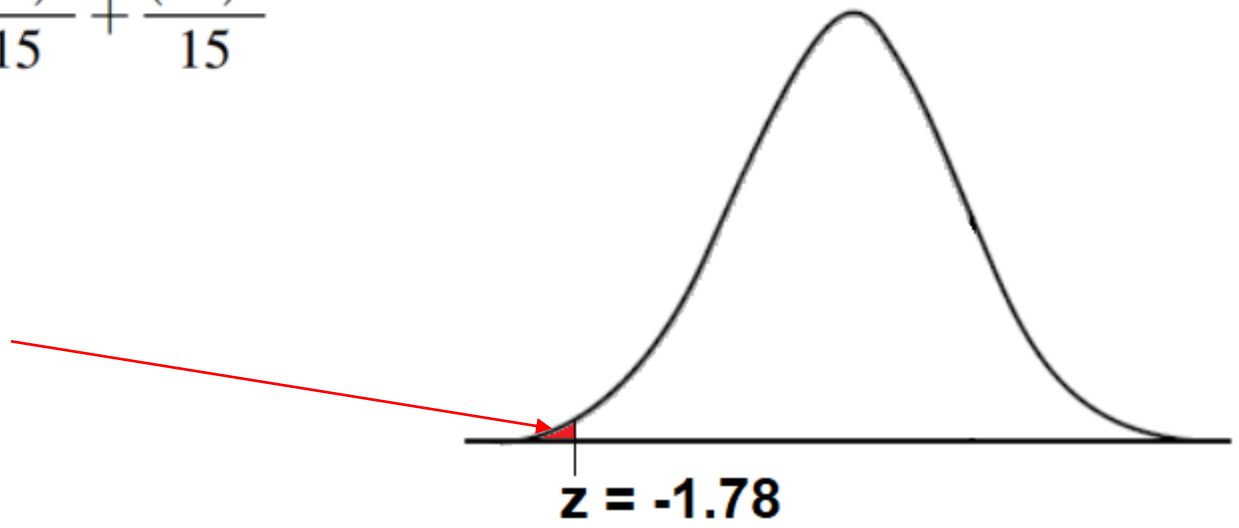
Convert to a z-score

- We showed previously that a normal sampling distribution can be converted to a standard normal distribution by modifying the formula for z. This is the version we use for the current problem:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{-13 - 0}{\sqrt{\frac{(20)^2}{15} + \frac{(20)^2}{15}}} = \frac{-13}{\sqrt{53.3}} = \frac{-13}{7.3} = -1.78$$

In our problem:
$\bar{x}_1 = 92$
$\bar{x}_2 = 105$
Difference = -13

Book states the probability equates to this area under the curve which is equal to .0375. **Is this correct?**



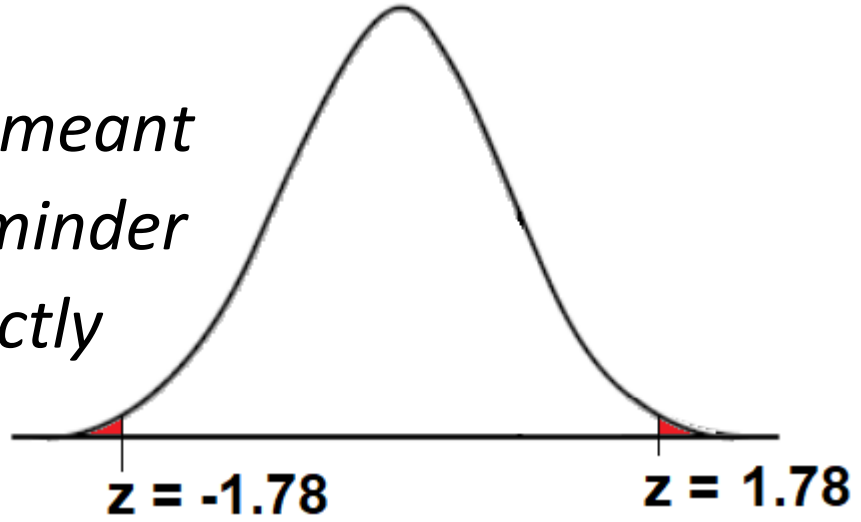
Is Proposed Solution Correct?

Question was:

"If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger ($\bar{x}_1 - \bar{x}_2$) between sample means?"

That seems to indicate we are looking for a difference of that magnitude or greater, regardless of which population is higher, which would involve two separate regions of the distribution for a cumulative total probability of .075.

Given the background of the problem, they likely meant for a unidirectional difference, but it is a good reminder that it is important to be clear with regard to exactly what problem you are trying to solve.



Sampling From Normal Populations

The procedure we have just followed is valid even when the sample sizes, n1 and n2, are different and when the population variances, σ_1^2 and σ_2^2 have different values. The theoretical results on which this procedure is based may be summarized as follows:

Given two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the sampling distribution of the difference, $\bar{x}_1 - \bar{x}_2$, between the means of independent samples of size n_1 and n_2 drawn from these populations is normally distributed with mean $\mu_1 - \mu_2$ and variance $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$.

Sampling from Nonnormal Populations

- Many times you may be faced by one of the following problems:
  1. You are sampling from a non-normal population
  2. You are sampling from a population whose form is unknown
- A solution to these problems is to take large sample sizes, since we know that when the sample sizes are large, the central limit theorem applies and the distribution between the two sample means will be at least approximately normally distributed with a mean equal to $\mu_1 - \mu_2$ and a variance of $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. Our process to find the probabilities associated with specific values of the statistics would then be the same as that given for sampling from normally distributed populations.