Some Important Sampling Distribution (III)

Huining Kang

HuKang@salud.unm.edu

August 19, 2020

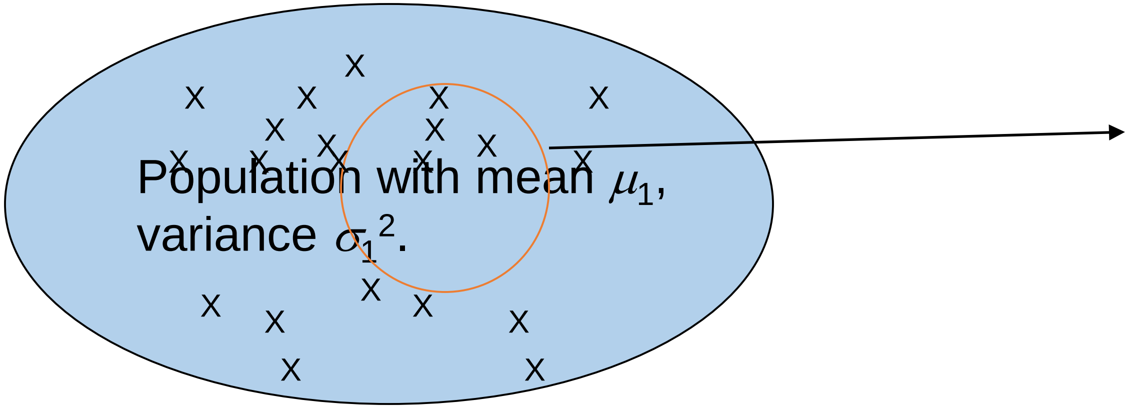
Outline

- Contents
  - Review
    - Some useful properties of the random variables
    - Mean, variance and distribution of a sample mean
  - Distribution of the difference between two sample means
    - Introduction
    - Characteristics of the difference between two sample means
    - Sampling from normal population
    - Sampling from non-normal population
- Learning outcome
  - Understand the ideas
  - Know the characteristics (mean and variance) and distribution of the difference between two sample means and know how to calculate the probabilities related to the difference statistic

Review

- Some useful properties of the random variables
  - $\mu_{\alpha \cdot X} = E(\alpha \cdot X) = \alpha \cdot E(X) = \alpha \cdot \mu_X$
  - $\sigma^2_{\alpha \cdot X} = Var(\alpha \cdot X) = \alpha^2 Var(X) = \alpha^2 \cdot \sigma^2_X$
  - $\mu_{X+Y} = \mu_X + \mu_Y$,
  - $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$, if X and Y are independent
  - $\alpha X + \beta Y \sim normal$, if $X \sim normal$ and $Y \sim normal$
  - The Central Limit Theorem
- Mean and variance of sample mean $\bar{X}$
  - $\mu_{\bar{X}} = \mu, \quad \sigma^2_{\bar{X}} = \sigma^2/n$
- Distribution of sample mean $\bar{X}$
  - $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ holds (or approximately holds) under the three conditions:
    - Population is normally distributed
    - Population is non-normally distributed and the sample size is large
    - Population has a unknown functional distribution form and the sample size is large

Distribution of the difference between two sample means

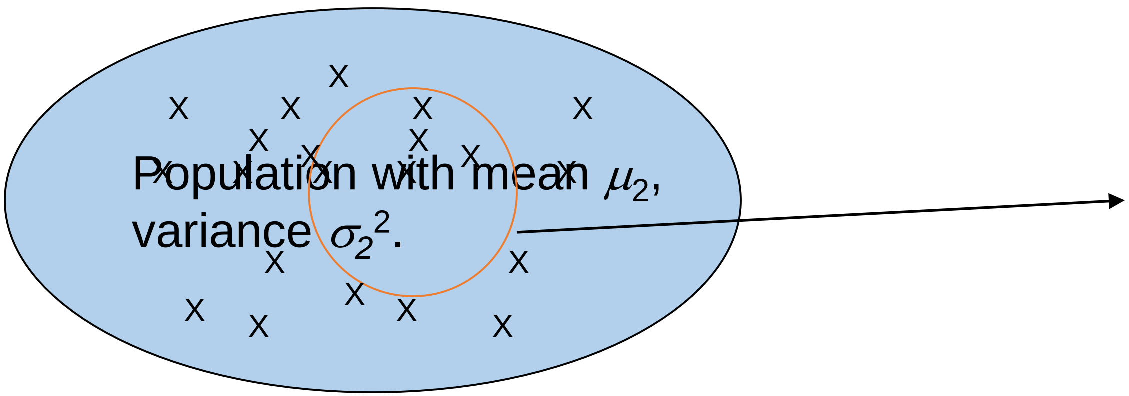


Simple random sample (SRS) of size n_1 is

$$X_{11}, X_{21}, \cdots, X_{n_1 1}$$

Sample mean is

$$\overline{X}_1 = \frac{X_{11} + X_{21} + \cdots + X_{n_1 1}}{n_1}$$



Simple random sample (SRS) of size n_2 is

$$X_{12}, X_{22}, \cdots, X_{n_2 2}$$

Sample mean is

$$\overline{X}_2 = \frac{X_{12} + X_{22} + \cdots + X_{n_1 2}}{n_2}$$

- Parameter of interest
  - Difference of population means (usually unknown): $\mu_1 - \mu_2$
- Related statistic
  - Difference of sample means: $\bar{X}_1 - \bar{X}_2$
- Questions of interest
  - Sampling distribution of $\bar{X}_1 - \bar{X}_2 = ?$ $\mu_{\bar{X}_1 - \bar{X}_2} = ?$ $\sigma^2_{\bar{X}_1 - \bar{X}_2} = ?$

Distribution of the difference between two sample means

- Population 1
  - Population mean, variance and standard deviation: $\mu_1, \sigma_1^2, \sigma_1$
  - Simple random sample: $X_{11}, X_{21}, \cdots, X_{n_1 1}$
  - Sample mean: $\bar{X}_1 = \frac{X_{11}+X_{21}+\cdots+X_{n_1 1}}{n_1}$
- Population 2
  - Population mean, variance and standard deviation: $\mu_2, \sigma_2^2, \sigma_2$
  - Simple random sample: $X_{12}, X_{22}, \cdots, X_{n_2 2}$
  - Sample mean: $\bar{X}_2 = \frac{X_{12}+X_{22}+\cdots+X_{n_2 2}}{n_2}$
- Parameter of interest
  - $\mu_1 - \mu_2$
- Difference between two sample means
  - $\bar{X}_1 - \bar{X}_2$
- Questions of interests
  - Distribution of $\bar{X}_1 - \bar{X}_2 =?$ $\mu_{\bar{X}_1 - \bar{X}_2} =?$ $\sigma^2_{\bar{X}_1 - \bar{X}_2} =?$

Distribution of the difference between two sample means

- Characteristics
  - $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$
    - $\mu_{\bar{X}_1-\bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma^2_{\bar{X}_1-\bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$
    - $\sigma^2_{\bar{X}_1-\bar{X}_2} = \sigma^2_{\bar{X}_1+(-\bar{X}_2)} = \sigma^2_{\bar{X}_1} + \sigma^2_{(-\bar{X}_2)} = \sigma^2_{\bar{X}_1} + (-1)^2\sigma^2_{\bar{X}_2}$
    $= \sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$
  - $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
- Converting to Z (Z-score transformation)

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1-\bar{X}_2}}{\sigma_{\bar{X}_1-\bar{X}_2}}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Distribution of the difference between two sample means

- Characteristics
  - $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Converting to Z

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1-\bar{X}_2}}{\sigma_{\bar{X}_1-\bar{X}_2}}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Sampling from normal populations
  - Population distributions: $N(\mu_1, \sigma_1^2)$, $N\left(\mu_2, \sigma_2^2\right)$
  - Sampling distribution of $\bar{X}_1 - \bar{X}_2$ is normal distribution

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

Distribution of the difference between two sample means

- Characteristics
  - $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
  - $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Converting to Z

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_{\bar{X}_1-\bar{X}_2}}{\sigma_{\bar{X}_1-\bar{X}_2}}$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Sampling from non-normal populations
  - Sampling distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal distribution, when n_1 and n_2 are large

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{a}{\sim} N(0, 1)$$

Distribution of the difference between two sample means

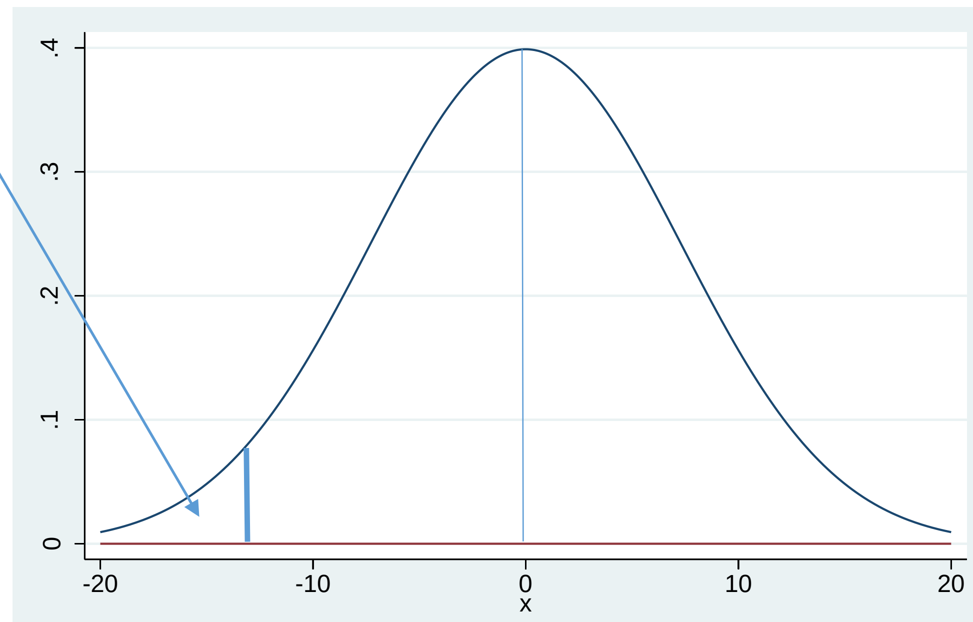
- Example 5.4.1 (Page 145, 10^{th} Ed)
  - Two populations of individuals
    - Population 1 has experienced some condition thought to be associated with mental retardation
    - Population 2 has not experienced the condition
  - Random variable of interest
    - Intelligence scores (X) in each of the distribution follows the normal distributions with a standard deviation $\sigma = 20$.
  - Take a sample of size 15 in each population and calculate the mean intelligence scores:
    - $\bar{x}_1 = 92, \;\; \bar{x}_2 = 105.$
  - Assumption
    - There is no difference in intelligence score between the two populations
  - Question
    - What is the probability of observing a difference this large or larger $(\bar{x}_1 - \bar{x}_2)$ between sample means?

Distribution of the difference between two sample means

- Example 5.4.1 (Cont.)
  - Population characteristics
    - Population means – mean intelligence scores: μ_1, μ_2 (unknown)
    - Population standard deviations: $\sigma_1 = \sigma_2 = \sigma = 20$
    - Population distributions: Normal
  - Samples
    - Sample sizes: $n_1 = n_2 = n = 15$
    - Sample means – mean intelligence scores of 15 individuals from each population: $\bar{X}_1, \bar{X}_2$
    - Statistics of interest: $\bar{X}_1 - \bar{X}_2$
    - Sample observations: $\bar{x}_1 = 92, \bar{x}_2 = 105, \bar{x}_1 - \bar{x}_2 = -13.$
  - Assumption
    - $\mu_1 = \mu_2$
  - Question
    - $P(\bar{X}_1 - \bar{X}_2 \leq \bar{x}_1 - \bar{x}_2) =?$, i.e., $P(\bar{X}_1 - \bar{X}_2 \leq -13) =?$

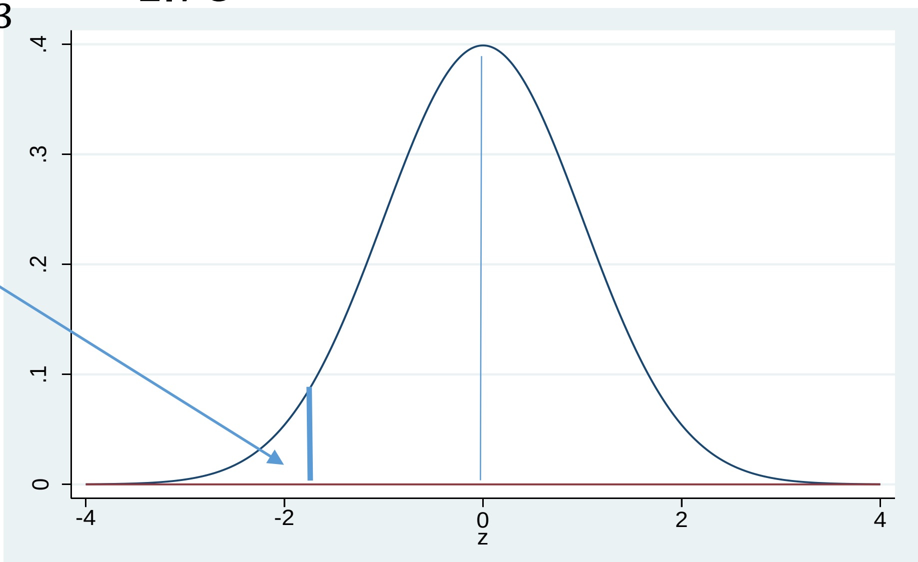
Distribution of the difference between two sample means

- Example 5.4.1 (Cont.)
  - Solution 1
    - $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
    - Mean: $\mu_1 - \mu_2 = 0$
    - Standard deviation: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{20^2}{15} + \frac{20^2}{15}} = 7.3$
    - $\bar{X}_1 - \bar{X}_2 \sim N(0, 7.3^2)$
    - $P(\bar{X}_1 - \bar{X}_2 \leq \bar{x}_1 - \bar{x}_2) = P(\bar{X}_1 - \bar{X}_2 \leq -13) = \Phi\left(\frac{-13-0}{7.3}\right) = .0375$
    - Stata command
      - disp normal(-13/7.3)
      - .03747077



Distribution of the difference between two sample means

- Example 5.4.1 (Cont.)
  - Solution 2
    - $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
    - Mean: $\mu_1 - \mu_2 = 0$
    - Standard deviation: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{20^2}{15} + \frac{20^2}{15}} = 7.3$
    - $Z = \frac{\bar{X}_1 - \bar{X}_2}{7.3} \sim N(0,1)$
    - Observed value of Z: $z = \frac{-13}{7.3} = -1.78$
    - $P(\bar{X}_1 - \bar{X}_2 \leq \bar{x}_1 - \bar{x}_2)$
    - $= P(Z \leq -1.78)$
    - $= \Phi(-1.78)$
    - $= .0375$
  - Stata Command
    - Disp normal(-1.78)
    - 03753798



Distribution of the difference between two sample means

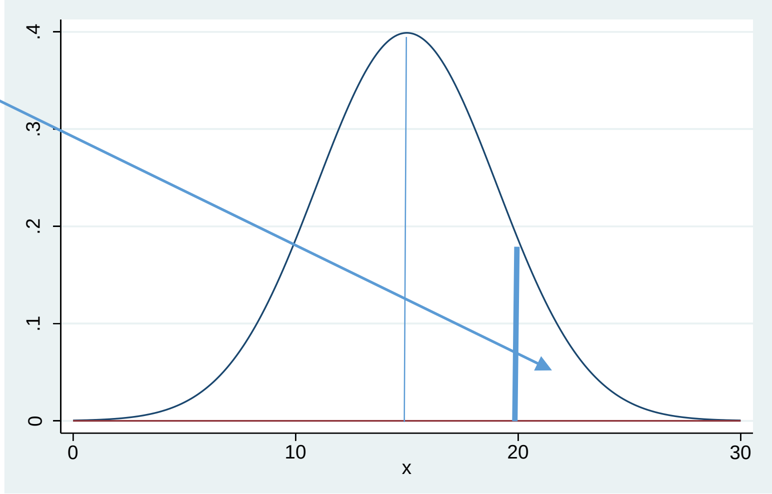
- Example 5.4.2 (Page 148, 10th Ed)
  - Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between client type one and client type two by 20 or more minutes?

Distribution of the difference between two sample means

- Example 5.4.2 (Cont.)
  - Population characteristics
    - Population means – average lengths of home visit: $\mu_1 = 45, \mu_2 = 30$
    - Population standard deviations: $\sigma_1 = 15,\ \sigma_2 = 20$
    - Population distributions: unknown
  - Samples
    - Sample sizes: $n_1 = 35, n_2 = 40$ (Large sample)
    - Sample means:
      - $\bar{X}_1 =$ average of 35 visits to the first type of clients
      - $\bar{X}_2 =$ average of 40 visits to the first type of clients
    - Statistics of interest: $\bar{X}_1 - \bar{X}_2$
  - Question
    - $P(\bar{X}_1 - \bar{X}_2 \geq 20) =?$

Distribution of the difference between two sample means

- Example 5.4.2 (Cont.)
  - Solution 1
    - $\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$
    - Mean: $\mu_1 - \mu_2 = 45 - 30 = 15$
    - Standard deviation: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{15^2}{35} + \frac{20^2}{40}} = 4.0532$
    - $\bar{X}_1 - \bar{X}_2 \sim N(15, 4.0532^2)$
    - $P(\bar{X}_1 - \bar{X}_2 \geq 20) = 1 - P(\bar{X}_1 - \bar{X}_2 \leq 20) = 1 - \Phi\left(\frac{20-15}{4.0532}\right) = .1087$
    - Stata command
      - disp 1 - normal((20-15)/4.0532)
      - .10867726



Distribution of the difference between two sample means

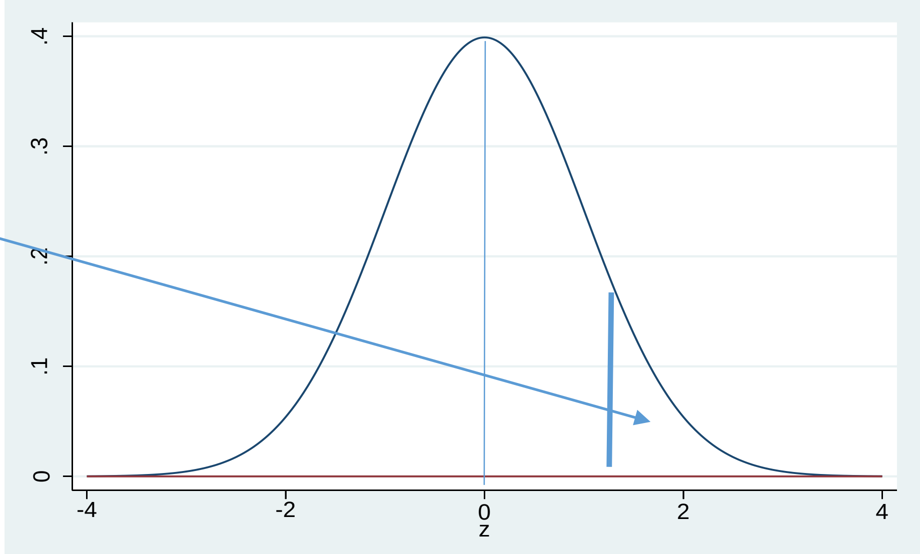
- Example 5.4.1 (Cont.)
  - Solution 2
    - $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$
    - Mean: $\mu_1 - \mu_2 = 45 - 30 = 15$
    - Standard deviation: $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{15^2}{35} + \frac{20^2}{40}} = 4.0532$
    - $Z = \frac{\bar{X}_1 - \bar{X}_2 - 15}{4.0532} \sim N(0,1)$
    - $P(\bar{X}_1 - \bar{X}_2 \geq 20)$

      $= P\left(Z \geq \frac{20-15}{4.0532}\right)$

      $= 1 - P\left(Z \leq \frac{20-15}{4.0532}\right)$

      $= 1 - \Phi\left(\frac{20-15}{4.0532}\right)$

      $= .1087$



Summary

- Most important formulas to remember
  - $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$
  - $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$

Homework 10

- Due date: August 31, 2020 (same as the previous homework)
- Exercises 5.4.3 and 5.4.5 (Page 150, 10^{th} Ed)