Outlier

Section Summary | Introduction to Statistics
The Central Limit Theorem

Sample Statistics and The Law of Large Numbers

1) Sample vs. Population
    a) It is often impossible to get data for an entire population
    b) Instead, we can collect data for a small sample of the population
    c) We can then use the statistics of that sample to estimate the population parameters
    d) Parameter: a summary measure or characteristic of an entire population
        i) Population Mean: μ
        ii) Population Variance: σ^2
    e) Statistic: a summary measure or characteristic of a sample
        i) Sample Mean: $\overline{x}$
        ii) Sample Variance: s^2
        iii) Sample Standard Deviation: s
2) Law of Large Numbers
    a) We can estimate population parameters by comparing the statistics of many samples
    b) Larger samples tend to produce more accurate estimations of the population
    c) $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} \overline{x_i} \to \mu \; as \; n \to N$

    where n is the sample size and N is the population size.
3) Sample Distributions
    a) Statistics will change from sample to sample
    b) We should ask several questions about our samples
        i) What is the mean of our sample statistic across all samples?
        ii) What is the variance of our sample statistic across all samples?
        iii) If we make a histogram of that statistic for all samples, what is the shape of that histogram?
        iv) How do the mean, variance, and histogram shape of this statistic behave as we alter the size of the sample?
    c) As the sample size increases, the mean of the statistic should approach the population parameter

d) As the sample size increases, the variance between different samples should decrease
e) As the sample size increases, the histogram of the samples should look more like a normal distribution with smaller and smaller standard deviation

Central Limit Theorem

1. CLT for Means
    a. Sample means will be normally distributed
    b. $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
    c. $\bar{x} = sample\ mean$
    d. $\mu = population\ mean$
    e. $\sigma = population\ standard\ deviation$
    f. $n = sample\ size$
    g. By taking many samples we can fit the sample means to a normal distribution and estimate the population mean and standard deviation
    h. As the sample size increases, the distribution of sample means becomes more normal and the width of the distribution decreases
    i. The rule of thumb for very large samples is to keep $n \geq 30$
    j. z-transformation of CLT: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
    k. With this z-transformation we can calculate certain probabilities
    l. If we have the population mean and standard deviation, we can calculate the probability of selecting a sample with a given range of means
    m. Given the sample mean and standard deviation, we can calculate the probability of being a certain distance from the actual population parameter (this requires approximating the population standard deviation with the sample standard deviation)
2. Standard Error
    a. $Standard\ Error = \frac{\sigma}{\sqrt{n}}$
    b. This shouldn't be confused with the standard deviation
    c. Standard deviation of the population, σ, gives the variability in the population's data

d. Standard deviation of the sample, s, gives the variability in a sample's data

e. Standard error, $\frac{\sigma}{\sqrt{n}}$, gives the variability of all the sample means

3. CLT for Sums

a. The sum of random samples is also normally distributed

$$\sum_{i=1}^{n} x_i \sim N(n\mu, \sigma\sqrt{n})$$

b. z-transformation:

$$z = \frac{\sum_{i=1}^{n} x_i - n\mu}{\sigma\sqrt{n}}$$

c. Given the population mean, μ, and population standard deviation, σ, you can calculate the accumulation of a certain result after n data points.