Discussion 5 Review

1. A collection of RVs X_1, X_2, ..., X_n are said to be **independent and identically distributed**, or **iid**, if the following things are true:

   - They are all independent from one another. That is, the realization of any one of them does not change the probability distribution of any other one.

   - They all have exactly the same probability distribution.

2. Estimation

   - Sample mean: $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

   - Sample variance of X: $\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

   - Sample standard deviation of X: $\hat{\sigma} = S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$

   - The **bias** in an estimator $\hat{\theta}$ is defined as:

     $$bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

     If the the bias is equal to zero, the estimator $\hat{\theta}$ is called **unbiased** for θ. All other things being equal, smaller bias is better.

   - The variance of an estimator $\hat{\theta}$ is defined as $VAR(\hat{\theta})$. All other things being equal, smaller variance is better. The square root of the variance is usually called the standard deviation or SD. However, when we are talking about estimating a parameter, we instead use the term **standard error** or **SE**, to remind us that this is the amount of error in estimation. Thus the square root of the variance of an estimator will be denoted $SE(\hat{\theta})$.

   - $E(\bar{X}) = E(\frac{X_1+X_2+...+X_n}{n}) = \frac{\mu+\mu+...+\mu}{n} = \mu.$

   - $VAR(\bar{X}) = VAR(\frac{X_1+X_2+...+X_n}{n}) = \frac{\sigma^2+\sigma^2+...+\sigma^2}{n^2} = \frac{\sigma^2}{n}.$

   - $SE(\bar{X}) = \sqrt{VAR(\bar{X})} = \frac{\sigma}{\sqrt{n}}.$

   - Estimated standard error of $\bar{X}$: $\widehat{SE(\bar{X})} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{S}{\sqrt{n}}.$

3. A **normal quantile-quantile plot** or **normal QQ plot** can be used to evaluate normality. If the data appears to be drawn from a normally distributed population, the points in the plot will usually fall on a roughly straight line.

4. The Central Limit Theorem can be stated as follows. Let X_1, X_2, ..., X_n be a collection of iid RVs with $E(X_i) = \mu$ and $VAR(X_i) = \sigma^2$. For large enough n, the distribution of $\bar{X}$ will be approximately normal with $E(\bar{X}) = \mu$ and $VAR(\bar{X}) = \frac{\sigma^2}{n}$. That is,

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

The required size for n depends on the nature of the true distribution of X_i. The closer the distribution of X_i is to normal, the smaller n is required for the approximation to be good. Usually about $n = 30$ is sufficient.

5. Confidence Intervals

- The interpretatoin for a confidence interval constructed for a population parameter θ, is that if you had theoretically taken many samples from the population, and created a different interval for each sample, $100(1-\alpha)\%$ of them would cover the true value of θ. This is usually shortened to saying we have $100(1-\alpha)\%$ **confidence** that the interval covers θ.

- When using $\bar{X}$ to estimate μ, if the X_i are normal and σ is known, or n is large enough for the CLT to work, then a $100(1-\alpha)\%$ CI for μ is given by:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

- When using $\bar{X}$ to estimate μ, if the X_i are normal, σ is unknown, and the sample size is small, then a $100(1-\alpha)\%$ CI for μ is given by:

$$\bar{X} \pm t_{(n-1,\alpha/2)}\frac{S}{\sqrt{n}}.$$

- The general form for a CI often looks like:

$$\text{estimate} \pm \text{multiplier} * \text{estimated SE(estimator)}$$

- When intending to create a $100(1-\alpha)\%$ CI for μ, assuming normality and a large sample size, the n required to achive a half-width of no larger than H is given by:

$$n = \frac{(z^2_{\alpha/2})(\sigma^2)}{H^2}.$$