Review - Week 10

Read: Chapters 20-23

Review:

There are two types of errors one can make when performing significance tests:

- Type I error – The null hypothesis is true, but we mistakenly reject it (False positive).
- Type II error – The null hypothesis is false, but we fail to reject it (False negative).

The probability that a hypothesis test will correctly reject a false null hypothesis is the power of the test.

To find power, we must specify a particular alternative parameter value as the true value. The difference between the null hypothesis value and the true value of a model parameter is the effect size.

Typically reducing the Type I error leads to an increase in the Type II error. The only way to reduce both types of error is by increasing the sample size.

Exercise 1: The owner of a store is concerned since only 30% of people who enter his store make a purchase. His assistant thinks his prices are too high and suggests that he gives a 5% rebate on all purchases for a two-week trial period. The owner counts the number of shoppers that enter the store and how many make a purchase, to determine if there has been a significant increase in the proportion of people who make purchases, in which case he will make the rebate permanent.

(a) Help the owner formulate the appropriate null and alternative hypotheses.
(b) In this context, what is a Type I error and what impact would such an error have on the store?
(c) In this context, what is a Type II error and what impact would such an error have on the store?
(d) In this context, what is meant by the power of the test?
(e) The owner decides to extend the trial period to last for one month so that he can gather more data. Will the power of the test increase, decrease or stay the same?
Review:

We often want to compare the proportion of people, in two populations or groups, which have some characteristic. Suppose we take samples of size $n_1$ and $n_2$ from each of the two populations and obtain the sample proportions $\hat{p}_1$ and $\hat{p}_2$, respectively.

To compare the two population proportions we use the difference between the sample proportions:

$$D = \hat{p}_1 - \hat{p}_2.$$ 

When $n_1$ and $n_2$ are large, $D$ is approximately $N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$.

A level $C$ confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

where $z^*$ is the value for the standard normal density curve with area $C$ between $-z^*$ and $z^*$.

A significance test for $H_0: p_1 = p_2$ uses the two-proportion $z$-statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

with P-values from the N(0,1) distribution. Here $\hat{p}$ is the pooled estimate of the common value of $p_1$ and $p_2$, which is given by

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}.$$ 

Exercise 1: Random samples of size $n_1 = 55$ and $n_2 = 63$ were drawn from Population 1 and 2, respectively. The samples yielded $X_1 = 15$ and $X_2 = 15$.

(a) Construct a 90% confidence interval for $p_1 - p_2$.
(b) Construct a 99% confidence interval for $p_1 - p_2$.
(c) Test $H_0: p_1 - p_2 = 0$ against $H_a: p_1 - p_2 > 0$. Can you reject $H_0$ at the 5% level of significance?
**Exercise 2:** A SRS of adult Americans were asked if they supported a certain proposal up for debate in Congress. Of the 800 people that were asked their opinion, 360 supported the proposal while the rest opposed it. In a follow up survey performed a month later, 900 people were asked their opinion and 460 people replied that they supported the proposal.

(a) Is the difference between the two polls significant at the 1% level? Construct an appropriate test of hypothesis.
(b) Construct a 95% confidence interval for the difference between the two polls.
(c) Construct a 90% confidence interval for the difference between the two polls.
(d) Judging by the confidence intervals in parts (b) and (c), is the difference between the polls significant at the 5% and 10% level?
Review:

The sample mean, $\bar{y}$, follows a $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ model. We can estimate $\sigma$ using the sample standard deviation, 

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}.$$ 

The standard error of the sample mean is $SE(\bar{y}) = \frac{s}{\sqrt{n}}$.

When $\sigma$ is unknown, we need to use the standard error in place of the standard deviation and the resulting standardized sample mean 

$$t = \frac{\bar{y} - \mu}{s/\sqrt{n}}$$

follows a $t$-model with $n-1$ degrees of freedom.

The $t$-models are symmetric, single-peaked, and bell shaped. They are similar in shape to the standard normal model, though their spread is slightly larger. As the degrees of freedom increase, the $t$-model becomes increasingly similar to a normal model.

A level $C$ confidence interval for $\mu$ is 

$$\bar{y} \pm t^*_n \frac{s}{\sqrt{n}}$$

where $t^*_n$ is the value under the $t_{n-1}$ model with area $C$ between $-t^*_n$ and $t^*_n$.

A significance test for the statement $H_0: \mu = \mu_0$ is based on the $t$ statistic: 

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

where $n$ is the size of the simple random sample.

If $n<15$ use $t$-procedures only if the data are close to normal. If $15 \leq n < 40$, use $t$-procedures as long as there are no outliers or strong skewness. If $n \geq 40$ the $t$-procedures can be used even for clearly skewed distributions.
**Exercise 1:** A random sample of $n$ observations is drawn from a normal population with mean $\mu$. Specify the critical value needed to construct a confidence interval with confidence level $C$ for each of the following choices of $C$ and $n$.

(a) $C = 99\%$, $n=15$
(b) $C = 99\%$, $n=30$
(c) $C = 90\%$, $n=9$
(d) $C = 95\%$, $n=21$

**Exercise 2:** Suppose a SRS of size 10 is made from a normally distributed population with mean $\mu$ and standard deviation $\sigma$, both unknown. Using the obtained data we were able to calculate $\bar{y} = 98.7$ and $s = 20$.

(a) Find a 90% Confidence Interval for $\mu$.
(b) Find a 95% Confidence Interval for $\mu$.
(c) Find a 99% Confidence Interval for $\mu$.

**Exercise 3:** Suppose we test the null hypothesis $H_0: \mu = 20$ against the alternative $H_a: \mu > 20$, for a normal population with mean $\mu$ and standard deviation $\sigma$, both unknown. A SRS of 25 observations is drawn from the population. Using the obtained data we were able to calculate $\bar{y} = 22.6$ and $s = 5.8$. Would you reject $H_0$ at the 5% level of significance?