Notes for the Final Exam; Statistics 301/371; Summer 2014

There will be no questions on using computational websites. There will be no questions on hypothesis testing: no test statistics, no formulas for P-values, no power; nothing. You will be given $t^*$ for 95% intervals and $z^*$ for levels 80%, 90%, 95%, 98%, 99% and the nearly certain interval, 99.87%. These are the only table values that you will need for the exam.

Also, I recommend that you review the material in Sections 18.4.2 and 20.4.2.

Chapter 12: Inference for a Binomial $p$. The approximate 95% confidence interval (CI) estimate of $p$ is:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$  

You need to know the following general facts about CIs. A CI can be written as $[l, u]$, where $l[/u]$ is the lower [upper] bound of the interval, and $l \leq u$. A CI is too small [large], if, and only if, $u [l]$ is smaller [larger] than the parameter being estimated. Every CI is either too small, too large or correct; only Nature knows which it is.

An approximate CI is centered at $\hat{p}$.

Chapter 13: The Poisson Distribution. If $X$ has a Poisson distribution with parameter $\theta$, then both its mean and variance equal $\theta$. If $Y$ has a binomial distribution with parameters $n$ and $p$ with

$$n \text{ large, } p \text{ close to zero and } np < 25;$$

then probabilities for $Y$ can be approximated well by using the Poisson distribution with parameter $\theta = np$. In particular, an exact confidence interval for $\theta$ is an approximate confidence interval for $np$, which, of course, will yield an approximate confidence interval for $p$.

Suppose that you have a Poisson Process (PP) with rate $\lambda$ per hour. Let $X$ denote the number of successes obtained by observing this process for $t$ hours. Then $X$ has a Poisson distribution with parameter $\theta = t\lambda$. Be careful: I sometimes like to mix units; for example, if the rate is per hour I might tell you that the PP is observed for $m$ minutes.

Suppose that $X_1$ has a Poisson distribution with parameter $\theta_1$ and $X_2$ has a Poisson distribution with parameter $\theta_2$. If $X_1$ and $X_2$ are statistically independent, then $Y = X_1 + X_2$, has a Poisson distribution with parameter $\theta_1 + \theta_2$.

Suppose that $X$ has a Poisson distribution with parameter $\theta$; then the approximate CI for $\theta$ is

$$x \pm z^*\sqrt{x}.$$  

Chapter 14: Prediction. You need to know both prediction interval (PI) formulas for the binomial and the PI for a PP. You do not need to know any of the chapter’s rules for means and variances.

We plan to observe $m$ future Bernoulli trials and want to predict the total number of successes, $Y$, that will be obtained.

• If $p$ is known, then the approximate 95% prediction interval for $Y$ is:

$$mp \pm 1.96\sqrt{mpq}.$$  

• If $p$ is unknown, we need previous data from the process which consists of $x$ successes in $n$ trials, yielding $\hat{p} = x/n$ and $\hat{q} = 1 - \hat{p}$. Define $r = m/n$, the ratio of the future to the past. The approximate 95% prediction interval for $Y$ is:

$$rx \pm 1.96\sqrt{r(1 + r)x\hat{q}}.$$  

We have past data from a PP consisting of $x$ successes in time $t_1$. We plan to observe the
same PP in the future for time \( t_2 \). Note that \( t_1 \) and \( t_2 \) must be in the same units; e.g., both hours or both minutes. The PI for the number of successes in the future observation of the PP is:

\[
r'x \pm z^* \sqrt{r'x(1 + r')}, \text{ where } r' = t_2/t_1.
\]

**Chapter 15: Comparing Two Binomial Populations.** The 95% confidence interval estimate of \((\hat{p}_1 - \hat{p}_2)\) is:

\[
(\hat{p}_1 - \hat{p}_2) \pm 1.96 \sqrt{(\hat{p}_1\hat{q}_1)/n_1 + (\hat{p}_2\hat{q}_2)/n_2}.
\]

You need to know about Simpson’s Paradox: We have a collapsed table of data which we describe with \( \hat{p}_1 \) and \( \hat{p}_2 \). We have two component tables, each of which has its own values of \( \hat{p}_1 \) and \( \hat{p}_2 \). The component tables must be consistent with the collapsed table.

We have Simpson’s Paradox occurring if either of the following occurs:

- We have \( \hat{p}_1 > \hat{p}_2 \) in the collapsed table and \( \hat{p}_1 < \hat{p}_2 \) in both component tables.
- We have \( \hat{p}_1 < \hat{p}_2 \) in the collapsed table and \( \hat{p}_1 > \hat{p}_2 \) in both component tables.

**Chapter 16: Two Dichotomous Responses.** You need to know the formula for conditional probability:

\[
P(A|B) = P(AB)/P(B).
\]

In formal mathematics, \( P(A|B) \) is read as the probability that \( A \) will occur given that \( B \) occurs. Less formally, for finite populations (i.e., not trials) it can be read as of those with feature \( B \), the proportion that also have feature \( A \).

All Chapter 16 data are presented as follows:

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( a )</td>
<td>( b )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>( A^c )</td>
<td>( c )</td>
<td>( d )</td>
<td>( n_2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( m_1 )</td>
<td>( m_2 )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

The CI for \( P(A) - P(B) \) is:

\[
\left( \frac{b - c}{n} \right) \pm \left( \frac{z^*}{n} \right) \sqrt{n(b + c) - (b - c)^2} n - 1
\]

We also consider the structure of populations in Chapter 16. Population counts are presented as follows.

<table>
<thead>
<tr>
<th></th>
<th>( B )</th>
<th>( B^c )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
<td>( N_{AB} )</td>
<td>( N_{AB^c} )</td>
<td>( N_A )</td>
</tr>
<tr>
<td>( A^c )</td>
<td>( N_{A^cB} )</td>
<td>( N_{A^cB^c} )</td>
<td>( N_{A^c} )</td>
</tr>
<tr>
<td>Total</td>
<td>( N_B )</td>
<td>( N_{B^c} )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

There are two classes of problems for which population counts are important.

First, let \( A [A^c] \) denote that a disease is present [absent]. Let \( B [B^c] \) denote that the screening test result is positive [negative]. The following combinations have names:

- \( A \) with \( B \) is a correct positive screening test result.
- \( A \) with \( B^c \) is a correct negative screening test result.
- \( A \) with \( B^c \) is an incorrect negative (false negative) screening test result.
- \( A^c \) with \( B \) is an incorrect positive (false positive) screening test result.

Second, let \( B [B^c] \) denote a bad [not bad] outcome. Let \( A [A^c] \) denote the presence [absence] of a risk factor. The population relative risk is defined to be

\[
(\frac{N_{AB}/N_A}{N_{A^cB}/N_{A^c}})\div.
\]

The population odds ratio is defined to be

\[
\frac{N_{AB}N_{A^cB^c}}{N_{A^cB}N_{AB^c}}.
\]

With data from a case-control study, the point estimate of the odds ratio is \((ad)/(bc)\) and the relative risk cannot be estimated.
Chapters 17 and 18: Inference for One Numerical Population. For a measurement response, the population is a pdf. A pdf is a function with the property that its total area equals 1 and probabilities are obtained by calculating areas.

Gosset’s confidence interval estimate of the population mean, $\mu$, is:

$$\bar{x} \pm t^* \left( \frac{s}{\sqrt{n}} \right).$$

Remember that the degrees of freedom for finding $t^*$ is $(n - 1)$.

The data we use above for a confidence interval estimate of $\mu$ can be used to predict one future observation; the prediction interval is:

$$\bar{x} \pm t^* s \sqrt{1 + \left( \frac{1}{n} \right)}.$$

As above, $df = (n - 1)$ for $t^*$.

Both of the formulas above are exact if the population is a Normal curve; otherwise, both are approximations.

If the population is a pdf and $n > 20$, it is possible to obtain an approximate CI for the population median, $\nu$. Proceed as follows:

- Sort the data.
- Calculate

$$k' = \frac{n + 1}{2} - \frac{z^* \sqrt{n}}{2}.$$

If $k'$ is an integer (never happens) set $k = k'$. Otherwise, round $k'$ down to the nearest integer $k$.

- The confidence interval is: $[x_{(k)}, x_{(n+1-k)}]$; in words, the $k$th smallest observation to the $k$th largest observation.

Also, for any value of $k$, with $k < n/2$, the interval $[x_{(k)}, x_{(n+1-k)}]$ can be viewed as a prediction interval for one future (independent) response from the same population. Its exact probability level is

$$1 - \left[ \frac{2k}{n + 1} \right].$$

Chapters 19 and 20: Comparing Two Numerical Response Populations. For independent random samples from the first and second populations we obtain:

$$\bar{x}, s_1, s^2_1 \text{ and } \bar{y}, s_2, s^2_2 \text{ and } n_1.$$

You are not responsible for Case 1. For Case 2, the pooled variance is:

$$s^2_p = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{n_1 + n_2 - 2}.$$

The 95% CI for $\mu_1 - \mu_2$ is:

$$(\bar{x} - \bar{y}) \pm t^* s_p \sqrt{\left( \frac{1}{n_1} \right) + \left( \frac{1}{n_2} \right)},$$

where $df = (n_1 + n_2 - 2)$.

For paired data, within pair $i$ we calculate $d_i = x_i - y_i$.

Our data become:

$$\bar{d}, s_d, s^2_d \text{ and the number of pairs } m.$$

Gosset’s 95% confidence interval estimate of the mean of the population of differences is:

$$\bar{d} \pm t^* \left( \frac{s_d}{\sqrt{m}} \right),$$

where $df = (m - 1)$ for $t^*$.

For paired data, we have the following relationships between the $x$’s the $y$’s and their differences, the $d$’s:

$$\bar{d} = \bar{x} - \bar{y} \text{ and } s^2_d = s^2_1 + s^2_2 - 2rs_1s_2,$$

where $r$ is the correlation coefficient for the $m$ pairs of $x$’s and $y$’s. Also, the mean of the population of differences, $\mu_d$, is equal to $\mu_1 - \mu_2$.

Chapters 21 and 22: Simple Linear Regression. We have data from $n$ units, called cases in regression. Each case yields two numbers, denoted by $X$—the predictor—and $Y$—the response. We represent the data by the
We compute the following summaries of these data, using the same notation as Chapters 19 and 20: \( \bar{x}, \bar{y}, s_1 \) and \( s_2 \). We restrict attention to data sets for which both \( s_1 \) and \( s_2 \) are positive numbers; i.e., we are not interested in data sets for which all the \( x_i \)’s or all the \( y_i \)’s are identical.

The correlation coefficient, denoted by \( r \), summarizes the strength and direction of the linear relationship between \( Y \) and \( X \). You don’t need to know the six features of \( r \) that are given in the Course Notes.

A researcher wants to find the line that best describes or fits the data. We study the line that is best according to principle of least squares. We calculate

\[
b_1 = r\left(\frac{s_2}{s_1}\right) \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}.
\]

The line \( \hat{y} = b_0 + b_1 x \) is the best line based on the principle of least squares. It goes by a variety of names, including: the regression line, the best line, the least squares line and the least squares regression line.

Recall that case \( i \) has values \( x_i \) and \( y_i \), its predictor and response. It also has

\[
\hat{y}_i = b_0 + b_1 x_i \quad \text{and} \quad e_i = y_i - \hat{y}_i,
\]

called its predicted response and residual, respectively. Thus, case \( i \) has four numbers associated with it:

\( x_i, y_i, \hat{y}_i \) and \( e_i \).

In addition to the relationships between these shown above, for every data set \( \sum e_i = 0 \). Thus, the mean of the residuals is always 0; the variance of the residuals is denoted by \( s^2 \).

On occasion, a researcher will want to calculate \( \hat{y} \) for a value of \( x \) that is not in the data set. This is considered to be ok, unless the value of \( x \) falls outside the range of the data’s \( x \)-values.

You don’t need to know about the various sums of squares (SSTO, SSE, SSR), but you do need to know that \( R^2 = r^2 \).

You need to know how to use regression output from Minitab. In particular, you need to know the following. Note that in the items below, the df for \( t^* \) is \((n - 2)\).

1. The 95% confidence interval estimate of the population slope, \( \beta_1 \) is:

\[
b_1 \pm t^* \text{SE}(b_1).
\]

2. For a given value of \( X \), which I will denote by \( x_0 \), the population mean value of the response is \( \mu_0 = \beta_0 + \beta_1 x_0 \). The 95% confidence interval estimate of \( \mu_0 \) is:

\[
\text{Fit} \pm t^* \text{SE(Fit)},
\]

where ‘Fit’ equals \( b_0 + b_1 x_0 \). Note that the value \( x_0 \) must be included as one of the \( x \)’s in the computer output. Thus, you can obtain ‘Fit’ from the computer output and don’t need to do any arithmetic to obtain it. Also, SE(Fit) will be in the computer output.

3. We know the value of \( X \) for a future case, call it \( x_{n+1} \). I want to find the 95% prediction interval for its response. First, calculate the variance of the prediction:

\[
\text{Var(pred)} = [\text{SE(Fit)}]^2 + s^2.
\]

The 95% prediction interval is:

\[
\text{Fit} \pm t^* \sqrt{\text{Var(pred)}}.
\]