Notes for the Final Exam; Statistics 301/371; Fall 2015; Professor Wardrop

**General.** You will be asked about the t-distribution website linked to our course webpage. Make sure you understand the three boxes below the default t-score. In particular, for a given problem, know what to put in the first two boxes and how to interpret the site’s response in the third box.

You will be given \( t^* \) and \( z^* \) for 90% and 95% intervals. These are the only table values that you will need for the exam. If you believe you need one that is not provided, then you need to rethink the problem.

A CI (confidence interval) is correct if it includes or contains the parameter being estimated. A CI is incorrect if it does not. There are two ways for a CI to be incorrect: it can be too small or too large. If we denote a CI as \([L, U]\) then

- A CI is **too large** if, and only if, \( L \) is greater than the parameter being estimated; and
- A CI is **too small** if, and only if, \( U \) is less than the parameter being estimated.

**Chapter 15: Comparing Two Binomial Populations.** The 95% confidence interval estimate of \((p_1 - p_2)\) 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 \( A \).

If we rewrite the formula for conditional probability we get the following multiplication rule:

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

Given one probability and two conditional probabilities you need to be able to construct an entire \( 2 \times 2 \) table of probabilities.

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><strong>Total</strong></td>
<td>( m_1 )</td>
<td>( m_2 )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

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

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

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><strong>Total</strong></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.
- $A^c$ with $B$ is a correct negative screening test result.
- $A$ with $B^c$ is an incorrect negative (false negative) screening test result.
- $A^c$ with $B^c$ is an incorrect positive (false positive) screening test result.


$$ \left( \frac{N_{AB}}{N_A} \right) \div \left( \frac{N_{A^cB}}{N_{A^c}} \right). $$

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.

You need to know about the three types of random samples from a finite population with two dichotomous features:

- **Type 1**: Overall random sample from the population.
- **Type 2**: Consider the following two populations: the population of members with feature $A$ and the population of members with feature $A^c$. We select independent random samples from these two populations. Because $A$ and $A^c$ define the rows of the table of population counts, we refer to this as being independent random samples from the rows.
- **Type 3**: Consider the following two populations: the population of members with feature $B$ and the population of members with feature $B^c$. We select independent random samples from these two populations. Because $B$ and $B^c$ define the columns of the table of population counts, we refer to this as being independent random samples from the columns.

When we assume we have i.i.d. dichotomous trials, then we have Type 1 sampling. This means that we may use the above CI formula for $[P(A) - P(B)]$ as well as the Chapter 15 CI formula for $(p_1 - p_2)$.

**Chapters 17 and 18: Inference for One Numerical Population.** You will not be asked any questions about testing or P-values from these two chapters.

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:

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

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.
The observed value of the test statistic for the null hypothesis \( \mu = \mu_0 \) is
\[
t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}.
\]
For the alternative \( > \) the P-value is the area under the t-curve to the right [left] of \( t \). The P-value for \( \neq \) is twice the smaller of the previous two.

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

**Chapters 19 and 20: Comparing Two Numerical Response Populations.** For independent random samples from the first and second populations we obtain:
\[
\bar{x}, s_1, s_1^2 \text{ and } n_1; \text{ and } \bar{y}, s_2, s_2^2 \text{ and } n_2.
\]
You are not responsible for Case 1. For Case 2, the pooled variance is:
\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.
\]

The 95% CI for \( \mu_1 - \mu_2 \) is:
\[
(\bar{x} - \bar{y}) \pm t^* s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},
\]
where \( df = (n_1 + n_2 - 2) \).

For paired data, within pair \( i \) we calculate
\[
d_i = x_i - y_i.
\]

Our data become:
\[
\bar{d}, s_d, s_d^2 \text{ and the number of pairs } n.
\]

Gosset’s 95% confidence interval estimate of the mean of the population of differences is:
\[
\bar{d} \pm t^* (s_d / \sqrt{n}),
\]
where \( df = (n - 1) \) for \( t^* \).

For paired data, we have the following relationships between the \( x \)'s the \( y \)'s and their differences, the \( d \)'s:
\[
\bar{d} = \bar{x} - \bar{y} \text{ and } s_d^2 = s_1^2 + s_2^2 - 2rs_1s_2,
\]
where \( r \) is the correlation coefficient for the \( n \) 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 \( n \) pairs: \((x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\).

We compute the following summaries of these data, using the same notation as Chapters 19 and 20: \( \bar{x}, \bar{y}, s_1 \text{ and } s_2 \). We restrict attention to data sets for which both \( s_1 \text{ and } s_2 \) 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

$$b_1 = r(s_2/s_1) \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

The line $\hat{y} = b_0 + b_1x$ is the best line based on the principle of least squares.

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

$$\hat{y}_i = b_0 + b_1x_i \text{ and } e_i = y_i - \hat{y}_i,$$

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

$$x_i, y_i, \hat{y}_i \text{ and } e_i.$$

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 $s^2$.

Note that if you have two points on a regression line, you can determine the line by twice plugging into the equation

$$\hat{y} = b_0 + b_1x.$$

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:

$$b_1 \pm t^*SE(b_1).$$

2. For a given value of $X$, which I will denote by $x_0$, the population mean value of the response is $\mu_0 = \beta_0 + \beta_1x_0$. The 95% confidence interval estimate of $\mu_0$ is:

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

where ‘Fit’ equals $b_0 + b_1x_0$. 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 an independent 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:

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

The 95% prediction interval is:

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