Notes for the Second Midterm Exam; Statistics Statistics 371, Lecture 3; Fall 2014

Miscellaneous. You will not be asked any questions about output from vassarstats.

Chapters 8 and 9: Dichotomous Responses; Critical Regions and Statistical Power. In Chapter 8 we consider the situation in which the response is a dichotomy. In this situation, we present our data in a table of counts:

| Treatment | Response | Proportions | | | |
|-----------|----------|-------------|----------|----------|
|           | S        | F           | Total    |           |
| 1         | a        | b           | n₁       | ̂p₁ = a/n₁ ̂q₁ = b/n₁ |
| 2         | c        | d           | n₂       | ̂p₂ = c/n₂ ̂q₂ = d/n₂ |
| Total     | m₁       | m₂          | n        |           |

The observed value of the test statistic is \( x = ̂p₁ - ̂p₂ \). As shown in the Course Notes, this test statistic, \( X \), is equivalent to our earlier test statistic \( U \). Thus, if \( n₁ = n₂ \), then the sampling distribution of \( X \) is symmetric around 0.

Exact P-values for \( X \) can be obtained from the internet. You need to remember the following properties, stated below for \( X \), but also true for \( U \):

- For any number \( x \), \( P(X \geq x) + P(X \leq x) = 1 + P(X = x) \).
- If the sampling distribution is symmetric around 0, then for any number \( x \),
  \[ P(X \geq x) = P(X \leq -x). \]

Given the exact sampling distribution for \( U \) and a target value of \( \alpha \), you need to be able to find the critical region for the alternative \( > \) and for the alternative \( < \). You will not be required to find the critical region for the alternative \( \neq \). The method is presented below for the test statistic \( U \). This method can be applied, in the obvious way, for the test statistic \( R_1 \).

- If the alternative is \( > \) and \( \alpha = 0.05 \), then you need to be able to find the number \( c \) that satisfies:
  \[ P(U \geq c) = 0.05. \]
- If the alternative is \( < \) and \( \alpha = 0.05 \), then you need to be able to find the number \( c \) that satisfies:
  \[ P(U \leq c) = 0.05. \]
- In each of the cases above, you will need to use trial-and-error to obtain the value of \( c \). Also, you need to be able to do this for any value of \( \alpha \), not just 0.05.

More often, I do not give you the exact sampling distribution of \( U \) or \( R_1 \). I will give my findings based on a computer simulation experiment. For example, I might give you the following table obtained under the assumption that the Skeptic is correct.
<table>
<thead>
<tr>
<th>( U )</th>
<th>( R_1 )</th>
<th>( R_1 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject ( H_0 )</td>
<td>( a = 9,374 )</td>
<td>( b = 126 )</td>
<td>( a + b = 9,500 )</td>
</tr>
<tr>
<td>Reject ( H_0 )</td>
<td>( c = 128 )</td>
<td>( d = 372 )</td>
<td>( c + d = 500 )</td>
</tr>
<tr>
<td>Total</td>
<td>( a + c = 9,502 )</td>
<td>( b + d = 498 )</td>
<td>( m = 10,000 )</td>
</tr>
</tbody>
</table>

Make sure you know what all of these numbers/cells represent. For example, for 9,374 assignments both tests correctly fail to reject; for 498 assignments test \( R_1 \) incorrectly rejects; and for 128 assignments \( R_1 \) correctly fails to reject while \( U \) incorrectly rejects.

**If instead of the Skeptic being correct, a particular version of the alternative is correct,** we get a table with the same labels, but now the topic is power.

<table>
<thead>
<tr>
<th>( U )</th>
<th>( R_1 )</th>
<th>( R_1 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject ( H_0 )</td>
<td>( a = 7,500 )</td>
<td>( b = 100 )</td>
<td>( a + b = 7,600 )</td>
</tr>
<tr>
<td>Reject ( H_0 )</td>
<td>( c = 300 )</td>
<td>( d = 2,100 )</td>
<td>( c + d = 2,400 )</td>
</tr>
<tr>
<td>Total</td>
<td>( a + c = 7,800 )</td>
<td>( b + d = 2,200 )</td>
<td>( m = 10,000 )</td>
</tr>
</tbody>
</table>

As with the earlier table, you need to know what each number/cell represents. In this table, the approximate powers are 0.2400 for \( U \) and 0.2200 for \( R_1 \). If we denote the exact powers by \( r_U \) and \( r_R \), respectively, then the nearly certain interval for \( r_U - r_R \) is:

\[
\frac{(c - b)}{m} \pm \frac{3}{m} \sqrt{b + c}.
\]

The nearly certain interval for \( r_R - r_U \) is

\[
\frac{(b - c)}{m} \pm \frac{3}{m} \sqrt{b + c}.
\]

Note that both of these intervals are approximations to and simplifications of the intervals given in the Course Notes. In my opinion the approximation is good provided \( m \geq 1,000 \).

**Chapter 10: Populations: Getting Started.** There are two types of problems in Chapter 10.

First, I will describe a Chance Mechanism and you determine probabilities using some combination of multiplying and adding. Below are two examples.

**Chance Mechanism:** I have a box with four cards, numbered 1, 2, 3 and 4. I plan to select two cards at random with replacement. Let \( T \) be the total of the two cards and define \( X_1 \) and \( X_2 \) as in the Course Notes. I want to determine \( P(T = 7) \). This is equal to

\[
P(X_1 = 3 \text{ and } X_2 = 4) + P(X_1 = 4 \text{ and } X_2 = 3) = (1/4)(1/4) + (1/4)(1/4) = 2/16 = 1/8.
\]

**Chance Mechanism:** Same as above, but I will select two card at random without replacement. Again, I want \( P(T = 7) \) which can be written the same as above. When we calculate, however, we now get:

\[
(1/4)(1/3) + (1/4)(1/3) = 2/12 = 1/6.
\]

Second, you are given a table of joint probabilities for two random variables \( X_1 \) and \( X_2 \), you need to be able to determine the probability of a variety of events. Identify the cells in the table that correspond to the event of interest and sum their probabilities; using the multiplication rule typically does not work.
Chapter 11: Bernoulli Trials (BT). If we have BT, then, in addition to the multiplication rule being true, we know that if $X$ is the total number of successes in $n$ BT, then the sampling (or probability) distribution of $X$ is binomial, written $\text{Bin}(n, p)$, where $p$ is the probability of success on a trial. For $(n \leq 6)$ make sure you can evaluate the following by hand.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x},$$

for $x = 0, 1, 2, \ldots, n$.

If $X$ has a binomial distribution, its mean is $np$ and its variance is $npq$.

You need to know about the random variables $R$ (the number of runs), $V$ (the length of the longest run of successes) and $W$ (the length of the longest run of failures). For example, suppose that 15 dichotomous trials yield:

11 0 1 0 1 1 1 0 0 0 0 1 0 1.

Then the observed values of these random variables are $r = 9$, $v = 3$ and $w = 4$.

Inference for $R$ is performed conditional on the value of $x$. In particular, define $c = 2x(n-x)$. With this definition, the mean and standard deviation of the sampling distribution of $R$ is:

$$\mu = 1 + (c/n) \quad \text{and} \quad \sigma = \sqrt{\frac{c(c - n)}{n^2(n - 1)}}.$$ 

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

$$\hat{p} \pm 1.96\sqrt{(\hat{p}\hat{q})/n}.$$ 

Note that this approximate CI is centered at $\hat{p}$.

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.

If $\hat{p} = 0.50$, then the exact CI is centered at $\hat{p}$; for other values of $\hat{p}$ it isn’t.

With the help of computer output I can provide for you, the exact value of $\alpha$ and the exact power can be obtained for the tests of Chapter 12; make sure you understand how to do this.

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 $n$ large, $p$ 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 $p$. 

3
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. The following formula gives the approximate CI for $\theta$

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

Once you have this CI for $\theta$, it can be converted to a CI for $\lambda$

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 $\theta_1 + \theta_2$.

**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:

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