Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 25 (MWF) Tests based on proportions (one sample)

Suhasini Subba Rao
Categorical data: The Binomial revisited

• In several situations we have to deal with categorical data.

• In this section we consider categorical data where only one of two outcomes is possible, and we are interested in the number of these outcomes.

• Example:

  – A company wants to know whether people are aware of a new product. They would do a survey, by asking people on the street. For each person surveyed the answer could be yes or no. We can formally write this as, let $X_i$ be the answer of the $i^{th}$ person interviewed, where

  $X_i = \begin{cases} 
  1 & \text{if yes} \\
  0 & \text{if no}
  \end{cases}$

  Suppose the probability $P(X_i = 1) = P(\text{Person says yes}) = p$. This means the probability that a randomly selected person has heard of the product is $p$. 
A student does a multiple choice exam, where the answers in each of the questions are completely independent of each other. Let $X_i$ be the grade of the $i$th question, where

$$X_i = \begin{cases} 
1 & \text{if correct} \\
0 & \text{if not correct}
\end{cases}$$

$$P(X_i = 1) = P(\text{question is correct}) = p.$$ 

Each time we do a statistical test it is done at the 5% level. Let $X_i$ be the outcome of the $i$th test (reject or not), then

$$X_i = \begin{cases} 
1 & \text{reject} \\
0 & \text{do not reject}
\end{cases}$$

If the null is true, $P(X_i = 1) = 0.05$.

Suppose that out of 5 randomly selected people on the street 3 have heard of the product, or the student gets 3 questions out of 5 correct.

We write this as $S_5 = X_1 + X_2 + X_3 + X_4 + X_5 = 3$ (this is the number
of people who said yes or the number of correct questions).

• To calculate the probability that $S_5$ takes a certain number we use the binomial distribution:

$$P(S_5 = 3) = \frac{5!}{3! \cdot 2!} p^3(1 - p)^2,$$

(you don’t have to remember this formula!) or use Software (recommended!)

• We now we use the Binomial distribution in statistical inference.
Example 1: Scores in an exam

- Suppose a midterm is a multiple choice, with five different options for each question.

If I guess each answer, there is a 0.2 chance of getting each question correct. By guessing, I can score anything from none of the questions correct to all correct. If I guess, the grade is random and follows a distribution. Let us look at the example that there are 15 questions in the exam. The distribution for the number I get correct, is a Binomial(15,0.2):
• We see from the plot that there is a 25% chance of my scoring 3 out of 15 (when I randomly guess) - just look at the height corresponding to 3. This makes sense, since if I was randomly guessing, on average I would get 20% of 15 correct, which is 3.

• Suppose Mike scores 5 out of 15, does this suggest that they were not randomly guessing? We can articulate this as a hypothesis test and use the plot on the previous slide to answer the question. Let p denote the probability of getting an answer correct. If \( p = 0.2 \) then this means randomly guessing. If \( p > 0.2 \), than some knowledge has been put into answering the question. The test can be written as \( H_0 : p \leq 0.2 \) against \( H_A : p > 0.2 \). In the exam, Mike scores 5 out of 15 (which corresponds getting 33% of the questions correct), how likely is it for Mike to get 5 or more correct by just randomly guessing. From the previous page we see this is 16.4% (this is the p-value), as this is quite large (larger than the 5% significance level), we cannot reject the null. This means, he
could have easily have scored 5 out of 15 if he were randomly guessing, though we will never know whether he was guessing or not (may be he knew the answer to those 5 questions and not the others). This offers one explanation as to why we cannot accept the null

- We can do the test formally in Statcrunch, and we get the output:

  ![Statcrunch output](image)

  - However, the p-values do not match exactly. This is because the above
test uses the normal approximation (more of this later).
Scores in an Exam: Example 2

- Let us suppose that the exam was a multiple choice exam, again with each question having a choice of 5.

- Rick scores 33 out of 100 (in this example like the last the person in question gets one third of the questions correct, however the conclusions will be different since the sample size has gone from 15 to 100). Is there evidence to suggest that he was not randomly guessing? The interesting point is that the proportion of questions he got correct is identical to the previous example (33%).

- Again we are testing as $H_0 : p \leq 0.2$ against $H_A : p > 0.2$ and we need to calculate the probability of scoring 33 or more when he is randomly guessing.
• From the previous plot we see that this corresponds to 0.15%, this is the p-value (the chance of scoring 33 or more by randomly guessing) as this probability is ‘small (less than the usual 5% significance level), this suggests Rick may not have been guessing.

• The difference between this case as the previous example is the standard error. To see what the standard error is we do the test in Statcrunch:
The standard error, under the null hypothesis being true, is \( \sqrt{0.2 \times 0.8/100} = 0.04 \), in the previous example the standard error was \( \sqrt{0.2 \times 0.8/15} = 0.103 \). By increasing the number of questions in the exam the estimate of \( p \) (the true proportion of questions one should get correct) becomes more reliable.

- However, the eagle eyed amongst you may have observed that the p-values of 0.13\% (using the Binomial distribution) and the 0.06\% given
above are a little different (this was because it was approximated by using the normal distribution). This is because the p-value of 0.13% is the correct p-value (obtained using the exact distribution) and 0.06% is simply the normal approximation.

• For one-sided tests of proportions using the Binomial distribution gives the accurate p-values. However, the binomial distribution cannot be used to construct confidence intervals or to do tests based on two proportions. In these instances, using the normal approximation can be very useful.
Polls of Gay marriage

A recent Gallup poll found that 64.2% (90 people) of a sample of 140 individuals were ‘pro-gay marriage.

- These are some news headlines. Are the following headlines accurate?
  - The New York Times: Survey suggests that the majority of the general public support Gay marriage:
    - The Wall Street Journal: Poll suggest that over 60% of the population support gay marriage.

Are these reports accurate?

To answer this question we write them as a hypothesis tests.
The New York Times reporting

- Majority means a proportion is over 50%. This means we are testing $H_0 : p \leq 0.5$ against $H_A : p > 0.5$.

- The Binomial is below:

![Binomial chart]

From the plot we see that the probability is 0.045%. This probability tells us that the chance of 90 or more people out of 140 saying they are
pro-gay marriage when public is equally divided is so 0.045% as this is very small (and a lot of less than the 5% significance level) we can reject the null. It seems likely that public opinion is in favour of gay marriage, ie. It is over 50%.
We want to see if there is evidence in the data that over 60% of the population support Gay marriage, this means testing $H_0 : p \leq 0.6$ against $H_A : p > 0.6$.

The Binomial is below:

From the plot we see that the probability is 17.1%. This probability tells us that the chance of 90 or more people out of 140 saying they are pro-gay...
marriage when public is opinion is divided 60:40 is 17.1%. As this is quite large (over the 5%), data could easily come from a population where the opinion is 60% (and not over 60%) pro-gay marriage. Therefore there is no evidence to back the alternative.
**Why bother with the normal approximation?**

- As we illustrated above, obtaining the exact p-values by using the true, Binomial distribution, is very useful.

- However, if we wanted to construct a confidence interval for the proportion, the Binomial can no longer be used.

- Consider the case of a student who so far has scored 22 out of 30 in his multiple choice midterms - he wants to know his final grade.

- It is impossible to know his grade for sure but we can give an interval (a confidence interval), where with a high degree of confidence we believe his final grade will lie.
The normal approximation to the binomial distribution

• Condition for this to work: If the number of success and the number of failures in the sample are greater than 5 the the normal distribution gives a relatively good approximation to the binomial. The number 5 for both successes and failure ensures that the binomial distribution is close symmetric about the mean even when $p$ is close to zero or one.

• Remember the mean of $Y_n$ is $np$ and the variance is $np(1-p)$.

• Since $Y_n = \sum_{i=1}^{n} X_i$, then the central limit theorem says that the average $Y_n/n = \frac{1}{n} \sum_{i=1}^{n} X_i$ has approximately the distribution

$$\hat{p} = \frac{Y_n}{n} \sim N(p, \frac{p(1-p)}{n}).$$
You saw the ‘normality’ happen in the above slides and also in Lecture 8. In other words \( \frac{Y_n}{n} \) is approximately normal with mean \( p \) and standard deviation \( \sqrt{p(1-p)/n} \).

- As always, the all important value is the standard error \( \sqrt{\frac{p(1-p)}{n}} \). It will tell us how ‘close’ the proportion estimator \( \hat{p} \) is to the true proportion \( p \). Furthermore, as the sample size \( n \) grows, the standard error get smaller. The estimator improves.

- Since \( \hat{p} = \frac{Y_n}{n} \) has the mean \( p \) it is an estimator of \( p \). Secondly we can use the normality result to construct confidence intervals and hypothesis test.
### Confidence intervals

- We use the same ideas as before. Since \( \frac{Y_n}{n} \sim \mathcal{N}(p, \frac{p(1-p)}{n}) \) the 100\((1 - \alpha)\)% confidence interval is

\[
\left[ \frac{Y_n}{n} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, \frac{Y_n}{n} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right],
\]

where \( z_{\alpha/2} \) is the normal distribution evaluated at \( \alpha/2 \) (recall \( z_{0.025} = 1.96 \)).

- Of course \( p \) is unknown which is why we are constructing a CI for it, so the standard error \( \sqrt{\frac{p(1-p)}{n}} \) is unknown. Instead we replace \( p \) with its estimator \( \hat{p} = \frac{Y_n}{n} \).
We still get approximately $\hat{p} = \frac{Y_n}{n} \sim \mathcal{N}(p, \frac{\hat{p}(1-\hat{p})}{n})$ and pretty much the same interval as before

$$\left[ \frac{Y_n}{n} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \frac{Y_n}{n} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right].$$

We do not use the $t$ distribution.
Testing

- We can also do a hypothesis test in the same way. We test \( H_0 : p = p_0 \) against the alternative \( H_A : p \neq p_0 \) (or \( H_0 : p \leq p_0 \) against \( H_A : p > p_0 \) etc.).

  - Remember, we want to see how likely we are to get the estimate \( \hat{p} = \frac{Y}{n} \) when the true global proportion is \( p_0 \). If the true global is \( p_0 \) the the standard error is \( \sqrt{p_0(1-p_0)/n} \) (not the \( \sqrt{\hat{p}(1-\hat{p})/n} \) used to construct a confidence interval).

  - We make a z-transform

\[
    z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}
\]

  and look up the z-tables according to the correct hypothesis.
Example 1 (testing)

Elections are due to take place next week. Ernie and Bert are the two candidates.

There has been some speculation that Ernie may win the elections. To test this hypothesis a poll of 1000 randomly selected people was taken and the data is summarised below.

<table>
<thead>
<tr>
<th></th>
<th>Ernie</th>
<th>Bert</th>
</tr>
</thead>
<tbody>
<tr>
<td>number</td>
<td>600</td>
<td>400</td>
</tr>
</tbody>
</table>

(a) What test would you use to test the hypothesis that Ernie will win?

(b) Based on this sample, is there evidence to suggest that Ernie will win? Remember, to state clearly the null and alternative (do the test at the 5% level).
(c) Construct a 95% CI for the proportion of the population who will vote for Ernie.
Solution

(a) We are interested proportion of people who will vote for Ernie. This means we should a one-sample test for proportions.

(b) Suppose $p$ are the proportion of people who will vote for Ernie. Ernie will win if $p > 0.5$. Hence we want to test the hypothesis $H_0 : p \leq 0.5$ against $H_A : p > 0.5$.

Our estimate of $p$ is $\hat{p} = 600/1000 = 0.6$. Now we want to use $\hat{p} = 0.6$ is significantly larger than 0.5 (to see whether we will reject the null and accept the alternative).

The standard error is $\sqrt{0.5 \times 0.5/1000} = 0.0158$ and the z-transform corresponding to the test is

$$z = \frac{0.6 - 0.5}{0.0158} = 6.32.$$
The test is pointing to the right of 6.32 (this is more than 6 standard deviations from the mean - it is huge). Therefore the p-value is the area of the normal to the right of 6.32, which is almost zero ($1.29 \times 10^{-10}$). This is so small, there is a large amount of evidence to suggest that alternative is true and that Ernie will win the election.

![One sample Proportion with summary](image)

- We now construct a 95% CI for the true proportion $p$. To construct a 95% confidence the standard error is $\sqrt{0.6 \times 0.4/1000}$. Note this is different to the standard error used for the test, this is because we are
looking at the likelihood of the data under any given distribution, so we now we need to put into our standard error the best estimator of the proportion that we have. The 95% confidence interval is

\[
[0.6 - 1.96 \times \sqrt{\frac{0.6 \times 0.4}{1000}}, 0.6 + 1.96 \sqrt{\frac{0.6 \times 0.4}{1000}}] = [0.57, 0.63].
\]

In other words, this proportion is likely to be 60% with a margin of error 3%.

This interval tells us that we believe that Ernie will get between 57%-63% of the vote. This is the corresponding output:
Lecture 25 (MWF) Tests based on proportions

95% confidence interval results:

* p : proportion of successes for population

Method: Standard-Wald

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Count</th>
<th>Total</th>
<th>Sample Prop.</th>
<th>Std. Err.</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>600</td>
<td>1000</td>
<td>0.6</td>
<td>0.015431934</td>
<td>0.56963634</td>
<td>0.63036364</td>
</tr>
</tbody>
</table>
Confidence intervals and different standard errors for different $p$

A student scores 22 out of 30 in a multiple choice exam, construct a 95% confidence interval for their final grade.

- The estimate of the final grade based on the information so far is $\hat{p} = \frac{22}{30} = 0.73$.

- The standard error is $\sqrt{0.73 \times (1 - 0.73) / 30} = 0.0807$.

- The 95% confidence interval for his final grade is

$$[0.73 - 1.96 \times 0.0807, 0.73 + 1.96 \times 0.0807] = [57\%, 88\%].$$

Therefore if the student does not change his academic behaviour we believe with 95% confidence his mean final grade lies in the interval
57%-88%. This interval is so wide it is hard to predict what his grade will be.

- It is very interesting to note if his current grade were 29 out of 30. Then his standard error will be $\sqrt{0.96 \times (1 - 0.96)/30} = 0.03$ and the 95% confidence interval for the final grade is

$$\left[\frac{29}{30} \pm 1.96 \times 0.03\right] = [90.2\%, 103\%].$$

This means that with 95% confidence we believe that if he does not change his behaviour he would get an A. (Observe, by using the normal distribution we do not ‘respect’ the fact that percentages cannot be over 100%).

- We now try a similar example. Another student’s score is 15 out of 30. This gives a standard error $\sqrt{0.5 \times (1 - 0.5)/30} = 0.091$ and the 95%
confidence interval for the final grade is

\[
\left[ \frac{15}{30} \pm 1.96 \times 0.091 \right] = [32.1\%, 67.9\%].
\]

• This is an extremely wide interval - due to the standard error being so large.

• The main observation is that even though the sample size is the same in the three examples above, the standard error is smallest when \( p = 0.96 \). This is because \( p = 29/30 = 0.96 \) is close to one. The closer \( p \) is to either one or zero, the smaller the standard error will be. The closer \( p \) is to 0.5 the larger the standard error.

These differences in spread can be seen even in the Binomial distribution for different \( p \) but same \( n \).
The above is a plot of how the standard error of an estimator changes with the true proportion.

For a person who always gets a 100% in an exam \((p = 1)\), the standard error for the grade in any given exam is a 0 - this is because that person will get 100% in that exam. There is no variation. The same is true person who gets everything wrong. However, if a person is a 50% person, the variability from exam to exam will be the most.

This plot is useful when designing experiments and choosing a sample size.
Choosing the sample size

• Let us return to the example of Ernie and Bert. The 95% confidence interval for those who will vote for Ernie is

\[
[0.6 - 1.96 \times \sqrt{\frac{0.6 \times 0.4}{1000}}, 0.6 + 1.96 \sqrt{\frac{0.6 \times 0.4}{1000}}] = [0.57, 0.63] = [57, 63]\
\]

• The Margin of Error is \(0.03 = 1.96 \sqrt{\frac{0.6 \times 0.4}{1000}}(= 3\%).\)

• In general the Margin of Error of the 95% confidence interval is

\[
MoE = 1.96 \sqrt{\frac{p(1-p)}{n}}.
\]
Two factors determine how large the margin of error is (i) the sample size \(n\) and (ii) the proportion \(p\). If \(p = 1\), then the s.e. = 0 and our estimator is actually to the true value (which is one).

- In general, the margin of error for a \(100(1 - \alpha)\%\) CI is

\[
MoE = z_{\alpha/2} \sqrt{\frac{p(1 - p)}{n}}.
\]

- To obtain the sample size for a given MoE we use

\[
n = \frac{z_{\alpha/2}^2 p(1 - p)}{MoE^2}.
\]

For example, if we want to construct a 95\% CI, then we use \(z_{\alpha/2} = 1.96\).
• Of course $p$ is unknown. But we have shown that when the sample size is fixed that the confidence interval is largest when $p = 1/2$. Therefore to be very cautious we can set $p = 0.5$. This will ensure that the true margin of error cannot be larger than the pre-specified margin of error.
Example 3

- What sample size should we need in order to achieve a margin of error of no more than 2% for a 90% confidence interval for the proportion of arthritis patients taking ibuprofen who suffer severe side effects?

- Solution: Since the standard error is maximised at $p = 0.5$

We replace $p$ with 0.5 and solve the equation

$$n = \left( \frac{1.64^2 \times 0.5 \times 0.5}{0.02^2} \right) = 1681$$
Example 4

• Suppose we **know** that the proportion of the population who get a side effect cannot be greater than 15%. Now what sample size should we choose?

• **Solution** If the upper bound is reliable, then we should include this in the sample size calculation as this will reduce the sample size required (recall that standard error and thus margin of error gets larger the closer to proportion to 0.5):
Using the plot we see that margin of error (for a given sample size) is maximised at 0.15 (if the proportion cannot be larger than 0.15). Using this information the sample size should be:

$$n = \left( \frac{1.64^2 \times 0.15 \times 0.85}{0.02^2} \right) = 858.$$

- Compared with the previous sample size of 1681, we observe that we have managed to substantially reduce the sample size when we use prior knowledge.