Basic Statistical Tests

Brian Vegetabile

2017 Statistics Bootcamp
Department of Statistics
University of California, Irvine

September 22th, 2016
Test of Proportions I

- The first test is the basic test of proportions.
- Consider a sample $X_1, X_2, \ldots, X_n$ of Bernoulli trials and testing whether or not the parameter of the distribution is $p_0$.
- That is

\[
H_0 : p = p_0 \\
H_1 : p \neq p_0
\]

- Our first concern is finding a test statistic that will allow us to answer this question.
Test of Proportions II

- We know that $\bar{X}$ is a consistent estimator for $p$ and $Var(X) = p(1 - p)$ so by the CLT and under the null hypothesis

$$Z = \sqrt{n} \frac{\bar{X} - p_0}{\sqrt{p_0(1 - p_0)}} \xrightarrow{L} N(0, 1)$$

- Now we will want to reject $H_0$ for values of $Z$ that are too large, i.e. if $|Z| \geq c$

- In order to choose $c$ we choose a significance level, $\alpha$, to control the Type I error rate
  - How often we are willing to make a type I error (rejecting $H_0$ when it is true)

- For a normal distribution this becomes $|Z| \geq \Phi^{-1}(1 - \alpha/2)$ for an equal tail test.
Test of Proportions III

Our rule then is

\[ \text{Reject } H_0 \text{ if } |Z| \geq \Phi^{-1}(1 - \alpha/2) \]

If we are interested in finding the \( p \)-value under the null hypothesis for this test we can compute \( \Phi(Z) \).

Finally, we calculate an estimate of \( z \) and perform the test procedure now that we have constructed our test statistic, the rejection region, and set up a rule for deciding to reject the null.
Testing the Mean with Known Variance I

- Consider a population where we believe the data is distributed normally.
- One test we could perform is to test if the mean value of this population is equal to some value $\mu_0$.
- Recall from the Central Limit Theorem

$$\sqrt{n} \frac{\bar{X} - \mu}{\sigma} \xrightarrow{L} N(0, 1)$$
Testing the Mean with Known Variance II

- Notice that the normal distribution is a parameterized by both it’s mean and variance, thus to use this result we must assume that know the population variance.

- Assuming that the variance is known, this proceeds very similarly to our proportion example.
Testing the Mean with Known Variance III

i) Construct a Null and Alternative Hypothesis

\[ H_0 : \mu = \mu_0 \]
\[ H_1 : \mu \neq \mu_0 \]

ii) Under the null hypothesis find a test statistic

\[ Z = \sqrt{n}\frac{\bar{X} - \mu_0}{\sigma} \xrightarrow{L} N(0, 1) \]

iii) Decide on a critical region consistent with an appropriate significance level

\[ |Z| > c = \phi^{-1}(1 - \alpha/2) \]

iv) Obtain a sample and calculate the observed test statistic and compare with the critical region

v) Report your conclusions in the context of the problem
Testing the Mean with Known Variance IV

- If we know the variance this is a very simple procedure which can allow us to decide if the observed data agrees with our hypothesis of the population mean.
- Additionally this procedure could allow us to provide $p$-values and other summaries to our collaborators based on the normal distribution.
- Unfortunately, we do not always know the variance of the population beforehand and must estimate it.
- This necessitates understanding the distribution of the sample variance.
Distribution of Sums of Normal Random Variables I

- Provided are some results about sums of normally distributed random variables which are useful for hypothesis testing.

- The proofs for these results should be investigated at your own pace and will come up throughout the year as you investigate transformations of random variables.
Consider a random sample $X_1, X_2, \ldots, X_n$ where each $X_i \sim N(\mu, \sigma^2)$ and let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Then

a) $\bar{X}$ and $S^2$ are independent random variables

b) $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

c) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ that is a chi-squared random variable with $n - 1$ degrees of freedom

Additionally notice that the above holds for **normally distributed random variables**.

Additionally notice that we do not need to rely on asymptotics for the previous hypothesis test!
The $t$-Distribution I

- Again we provide some unproven relationships which define special random variables.
- You will most likely show this, but it will be important that you at least know the relationship between the normal distribution and the chi-square distribution and how they are used to create $t$ random variable.
The \( t \)-Distribution II

- Consider two independent random variables \( Y \) and \( Z \) such that \( Y \sim \chi^2_n \) and \( Z \sim N(0, 1) \).
- We define a transformation of these random variables \( T \) such that
  \[
  T = \frac{Z}{\sqrt{Y/n}}
  \]
- The distribution of \( T \) is called the \( t \) distribution with \( n \) degrees of freedom.
Test of the Mean with Unknown Variance I

- We’ll use those facts to create a test for the population mean when the variance is unknown.
- This test is often referred to as a *t*-test and you should recognize it from your basic statistics courses.
- We outline a bit more of the technicalities then you may have seen in your intro stats course.
Test of the Mean with Unknown Variance II

▶ Consider a random sample \(X_1, X_2, \ldots, X_n\) where each \(X_i \sim N(\mu, \sigma^2)\).

▶ We would like to test the hypothesis that the mean \(\mu\) is some value \(\mu_0\), without knowing the population variance.

▶ As with our last example, the null and alternative hypotheses remain the same

\[
H_0 : \mu = \mu_0 \\
H_1 : \mu \neq \mu_0
\]

▶ We now must find an appropriate test statistic.
Test of the Mean with Unknown Variance III

Based on previous results, we know that

\[ Z = \sqrt{n} \frac{\bar{X} - \mu}{\sqrt{\sigma^2}} \sim N(0, 1) \quad \text{and} \quad W = \frac{n - 1}{\sigma^2} S^2 \sim \chi^2_{n-1} \]

and we want to create a statistic that does not contain \( \sigma^2 \).

Therefore if we divide \( \sqrt{\frac{S^2}{\sigma^2}} \), we have that

\[ T = \frac{\sqrt{n} \frac{\bar{X} - \mu}{\sqrt{\sigma^2}}}{\sqrt{\frac{S^2}{\sigma^2}}} = \sqrt{n} \frac{\bar{X} - \mu}{\sqrt{S^2}} \]
Test of the Mean with Unknown Variance IV

- Considering $T$, we notice that it is actually the ratio of two random variables

$$T = \frac{X}{\sqrt{\frac{W}{n-1}}}$$

- Therefore by our previous results we know that

$$T \sim t_{n-1}$$
Test of the Mean with Unknown Variance V

- This implies that we can use the test statistic $T$ for testing the population mean when the variance is unknown.
- Now we must define a critical region for the test statistic for an appropriate significance level.
- Let’s investigate the $t$-Distribution a little further.
We see that $t$-distribution is very similar to a normal distribution in shape and spread.

Also, it appears that for large degrees of freedom the $t$ distribution approaches a normal distribution.
Therefore we can choose symmetric points from the $t$ distribution such that

$$|T| > c$$

For a specific $\alpha$ level we will define these points as $T_{\alpha/2}$ and $T_{1-\alpha/2}$.

Now that we have defined the critical region for a specific significance level, we can create our testing procedure.
Test of the Mean with Unknown Variance VIII

i) Construct a Null and Alternative Hypothesis

\[ H_0 : \mu = \mu_0 \]
\[ H_1 : \mu \neq \mu_0 \]

ii) Under the null hypothesis find a test statistic

\[ T = \sqrt{n} \frac{\bar{X} - \mu_0}{\sqrt{S^2}} \sim t_{n-1} \]

iii) Decide on a critical region consistent with an appropriate significance level

\[ |T| > T_{\alpha/2} = T_{1-\alpha/2} \]

iv) Obtain a sample and calculate the observed test statistic and compare with the critical region

v) Report your conclusions in the context of the problem
Comparing Multiple Populations

- All three of these tests can be modified to consider testing the difference in population means between two groups under study.
- I won’t review all of them in the interest of time, but it is worth picking up a basics statistics textbook and reviewing:
  - Two-Proportion Test
  - Two-Sample Test of the Mean with Variance Known
  - Two-Sample Test of the Mean with Unknown Variance (Two-Sample t-Test)
  - Paired Tests