STAT 311: Hypothesis Testing

Y. Samuel Wang

Summer 2016

Logistics

- Homework 6 posted
- Lab 6 posted as well

Hypothesis Testing Procedure

A hypothesis test consists of the following steps

1. Determine the Null and Alternative hypotheses
2. Determine the Null distribution
3. Gather data / calculate a test statistic
4. Calculate a p-value
5. Draw conclusions

P-values

P-values (probability values) explicitly quantify **the probability of a test statistic as (or more) extreme as the one I actually did observe if the null hypothesis is true**.

- A smaller p-value denotes stronger evidence against the null hypothesis
- Typically a cut-off of .05 is used for statistical significance
- The cut off is called the **level** or **size** of the hypothesis test

P-values

P-values (probability values) explicitly quantify **the probability of a test statistic as (or more) extreme as the one I actually did observe if the null hypothesis is true**.

- A smaller p-value denotes stronger evidence against the null hypothesis
- Typically a cut-off of .05 is used for statistical significance
- The cut off is called the **level** or **size** of the hypothesis test

P-values do **not** mean- "How likely is the null hypothesis?"

- Correct: $P(\text{Data}|\text{Null Hypothesis})$
- Incorrect: $P(\text{Null Hypothesis}|\text{Data})$

If the p-value is large we do not "accept" the null hypothesis, we simply **"fail to reject"**

Details about Hypothesis Testing

The specifics of a hypothesis test depend on the details

Parameter	H_0	H_A	Test Statistic	Null Distribution
p	$p = p_0$	$p \overset{>}{\underset{<}{\neq}} p_0$	$\hat{p}$	$\mathcal{N}\left(p_0, \frac{p_0(1-p_0)}{n}\right)$
$p_1 - p_2$	$p_1 - p_2 = 0$	$p_1 - p_2 \overset{>}{\underset{<}{\neq}} 0$	$\hat{p}_1 - \hat{p}_2$	$\mathcal{N}\left(0, \hat{p}_0(1 - \hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ $\hat{p}_0 = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$
μ	$\mu = \mu_0$	$\mu \overset{>}{\underset{<}{\neq}} \mu_0$	$\sqrt{n}\frac{\bar{x} - \mu_0}{s_x}$	$\mathcal{T}_{n-1}$
$\mu_1 - \mu_2$	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \overset{>}{\underset{<}{\neq}} 0$	$\frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$\mathcal{T}_{\min(n_1 - 1, n_2 - 1)}$

For the T distributions, if the degrees of freedom are greater than 30, we can use a normal distribution, since the difference is negligible

Example: Single proportion

Each month, the employees of Alaska Airlines receive a \$50 bonus, if at least 78% of all flights are on-time. On Feb 2, 2016, Alaska Airlines had X arrivals at Sea-Tac Airport. Of the X arrivals, Y arrived on time. Based on our sample, do we have reason to believe that the true proportion of all on-time arrivals at Sea-Tac is less than 78%?

$$H_0 : p = .78$$

$$H_A : p < .78$$

Example: Single proportion

Each month, the employees of Alaska Airlines receive a \$50 bonus, if at least 78% of all flights are on-time. On Feb 2, 2016, Alaska Airlines had 146 arrivals at Sea-Tac Airport. Of the 146 arrivals, 99 arrived on time[1]. Based on our sample, do we have reason to believe that the true proportion of all Alaska arrivals at Sea-Tac which are on-time is less than 78%?

$$H_0 : p = .78$$

$$H_A : p < .78$$

Based on our sample, we have $\hat{p} = .68$

[1] http://transborder.bts.gov/xml/ontimesummarystatistics/src/index.xml

Example: Single proportion

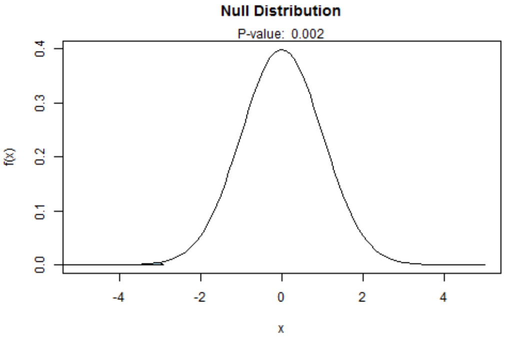
Under the null hypothesis, $p = .78$,

$$\hat{p} \sim \mathcal{N}(p_0, p_0(1-p_0)/n) = \mathcal{N}(.78, 0.0011) =$$

Notice that we use p_0 for the null distribution and not $\hat{p}$

Example: Single proportion

The z-score is $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{.68-.78}{\sqrt{.78(1-.78)/146}} = -2.91$, so under the null distribution, the probability of seeing a sample as or more extreme is .002.



Example: Single proportion

Since the p-value is less than .05, we would reject the null hypothesis that the true proportion of on-time arrivals is .78.

Example: Single proportion

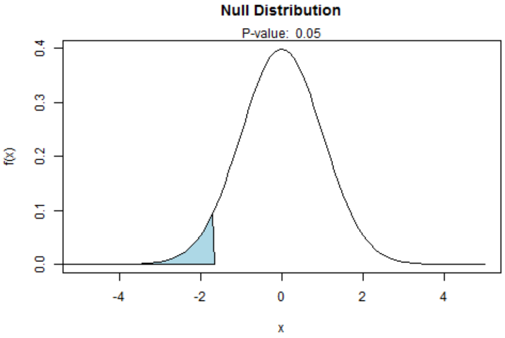
Since the p-value is less than .05, we would reject the null hypothesis that the true proportion of on-time arrivals is .78.

Can we generalize this conclusion?

- Is this date representative of other days in February?
- Is Sea-Tac representative of all Alaska arrivals?

Rejection Region

Instead of looking at a p-value of an observed value, we could've just first decided a cut-off, then translate that back into a length. For instance, if we decide on a cut-off of .05, that means we would reject the null if the test statistic is less than -1.645.



We could then translate this z-score back to an observed $\hat{p}$.

Rejection Region

We could then translate this z-score back to an observed $\hat{p}$.

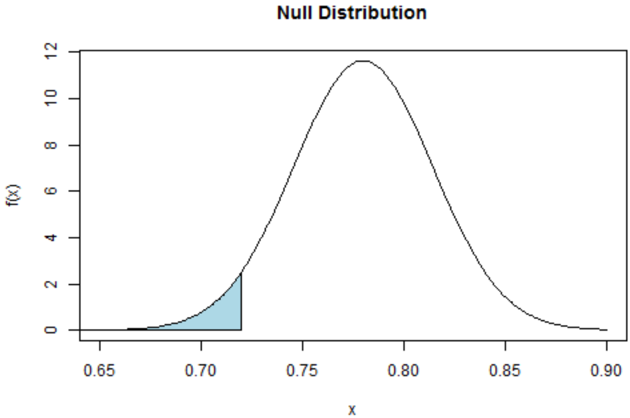
$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = -1.645$$

$$\Rightarrow \hat{p} = p_0 - 1.645\sqrt{p_0(1 - p_0)/n} = .72$$

Thus, an equivalent procedure to rejecting if the p-value is less than .05, is rejecting if $\hat{p}$ is less than .72. The region of the null distribution less than .72 is called the **Rejection Region**.

Rejection Region

Thus, an equivalent procedure to rejecting if the p-value is less than .05, is rejecting if $\hat{p}$ is less than .72. The region of the null distribution less than .72 is called the **Rejection Region**.



Example: Difference in proportions

A 2015 Pew Research Poll investigated the internet and social media habits of Americans[2]. One particular question asked respondents whether or not they used auto-deleting messaging apps (ie Snapchat). One might be interested in whether the proportion of men and women which use the these apps differs.

$$H_0 : p_{men} - p_{women} = 0$$

$$H_A : p_{men} - p_{women} \neq 0$$

[2] http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/

Example: Difference in proportions

A 2015 Pew Research Poll investigated the internet and social media habits of Americans[2]. One particular question asked respondents whether or not they used auto-deleting messaging apps (ie Snapchat). One might be interested in whether the proportion of men and women which use the these apps differs.

$$H_0 : p_{men} - p_{women} = 0$$

$$H_A : p_{men} - p_{women} \neq 0$$

Note that this is a two sided test, so the p-value will be $P(|\hat{p}_{men} - \hat{p}_{women}| > \text{Observed})$ assuming the Null Hypothesis is true

[2] http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/

Example: Difference in proportions

A 2015 Pew Research Poll investigated the internet and social media habits of Americans. One particular question asked respondents whether or not they used auto-deleting messaging apps (ie Snapchat). One might be interested in whether the proportion of men and women which use the these apps differs. The released data gives $\hat{p}_{men} = .17$ and $\hat{p}_{women} = .18$. It does not give the sample size for each sub-population, but the total sample size is 1907, so let's assume that's evenly divided with $n_{men} = 953$ and $n_{women} = 954$.

Example: Difference in proportions

Under the null hypothesis, where $p_{men} = p_{women}$,

$$\hat{p}_{men} - \hat{p}_{women} \sim \mathcal{N}\left(0, \hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

where $\hat{p}_0 = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$. Note that this null distribution actually doesn't specify what the common value is, just that they are equal. In this case, $\hat{p}_0 = .175$ so the null distribution is

$$\mathcal{N}\left(0, 0.0174^2\right)$$

which gives a z-score of

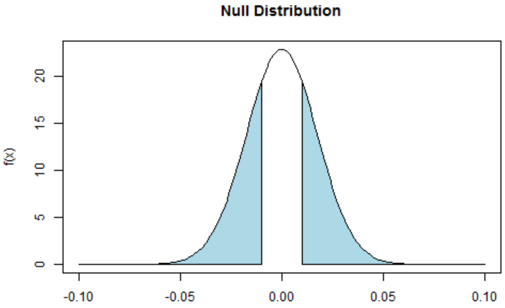
$$\frac{.01 - 0}{.0174} = .575$$

so the p-value is $P(|Z| > .575) = 0.282 \times 2$

Example: Difference in proportions

Because it is a two sided test, we actually "double" the p-value we see in the table. This is because we are only interested in the absolute value $|\hat{p}_{men} - \hat{p}_{women}|$.

$$P(|\hat{p}_{men} - \hat{p}_{women}| > .01|\text{Null Hypothesis})$$
$$= P(|Z| > .575) = P(Z > .575) \times 2 = 0.282 \times 2$$



Example: Difference in proportions

Because the p-value is greater than .05, we fail to reject the Null hypothesis that there is no difference between the proportion of men and women who use auto-delete messaging services

Example: Difference in proportions

Because the p-value is greater than .05, we fail to reject the Null hypothesis that there is no difference between the proportion of men and women who use auto-delete messaging services

Note that we **do not** accept the Null hypothesis! We simply do not have evidence to reject it.

Example: Single Mean

In 2013, a class action suit was brought against Subway, claiming that the footlong subs were not actually 12 inches long. Subway eventually agreed to settle the lawsuit for $525,000. Let μ denote the average sub length.

$$H_0 : \mu = 12$$

$$H_A : \mu < 12$$

Example: Single Mean

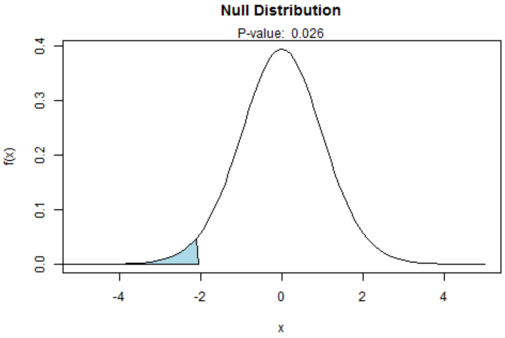
In 2013, a class action suit was brought against Subway, claiming that the footlong subs were not actually 12 inches long. Subway eventually agreed to settle the lawsuit for $525,000. Let μ denote the average sub length.

$$H_0 : \mu = 12$$
$$H_A : \mu < 12$$

Suppose we measure 25 sandwiches and get an average length of 11.91 inches and $s_x = .22$ which yields $\sqrt{n}\frac{\bar{x}-\mu_0}{s_x} = -2.05$

We know that $\sqrt{n}\frac{\bar{x}-\mu_0}{s_x} \sim T_{24}$, so we can find the probability of a value as or more extreme as -2.05 from a T_{24}.

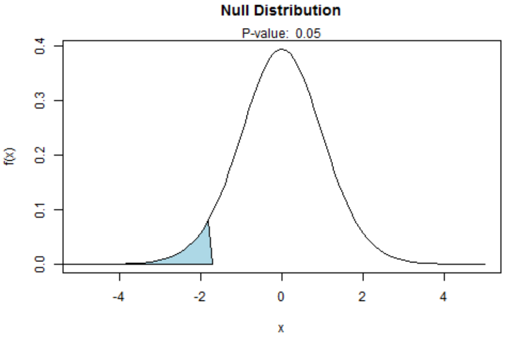
Example: Single Mean



Under the null hypothesis, we would only expect to see a test statistic as or more extreme than -2.05 about .026 of the time. Thus, in a hypothesis test with level .05, we would reject the null hypothesis that the average length of the footlong subs is actually 12 inches long.

Rejection Region

Instead of looking at a p-value of an observed value, we could've just first decided a cut-off, then translate that back into a length. For instance, if we decide on a cut-off of .05, that means we would reject the null if the test statistic is less than -1.71.



The area to the left of -1.71 is called the **Rejection Region**.