# STAT 311: Hypothesis Testing

Y. Samuel Wang

Summer 2016

# Logistics

- Homework 5 due now
- Homework 6 posted later today

# Roadmap

So far we have learned about...

- Describing (univariate) Data
- Describing relationships between data
- How to gather data
- How data happens
- How to make decisions with data

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

Figure: Balling out at the gym last week

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

- You come and observe 2 of my games, where I score an average of 18 points
- Do you still believe my claim?

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

- You come and observe 2 of my games, where I score an average of 10 points
- Do you still believe my claim?

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

- You come and observe 15 of my games, where I score an average of 10 points
- Do you still believe my claim?

# Hypothesis Testing Intuition

I'm a really good basketball player and on average I score 20 points a game

- You come and observe 15 of my games, where I score an average of 10 points
- Do you still believe my claim?

How do you decide whether to reject my claim or not? How much evidence is enough?

# Null Hypotheis

The **Null Hypothesis** is a statement about the status quo. It typically states that some parameter is equal to 0, or that there is nothing happening. Typically denoted using $H_0$.

- Under the null hypothesis, we would expect our observations to have a certain distribution. This is called the **Null Distirbution**
- In most cases, the researcher hopes to disprove the null hypothesis
- Examples-
  - I score 20 points on average
  - Taking this medicine has no effect on a patient's probability of recovery
  - There is no relationship between education level and income

# Alternative Hypotheis

The **Alternative Hypothesis** is the "opposite" of the Null Hypothesis. Typically denoted using $H_1$ or $H_A$.

- In most cases, the researcher hopes to gather evidence in support of the alternative hypothesis
- Examples-
  - I score less than 20 points on average
  - Taking this medicine has an effect on a patient's probability of recovery
  - There is a positive relationship between education level and income

# Alternative Hypotheis

If the null hypothesis specifies that a parameter is equal to a specific value $\mu_0$ (typically, but not always 0), the alternative hypothesis can have a few different forms

$$H_0 : \mu = \mu_0$$

- $H_A : \mu \neq \mu_0$ (Two sided Alternative)
- $H_A : \mu < \mu_0$ (One sided Alternative)
- $H_A : \mu > \mu_0$ (One sided Alternative)

# Specificying Hypotheis

For the previous examples, we can formalize the hypothesis

- Do I score 20 points on average?
  - $H_0 : \mu_{points} = 20$ I score on avg 20 points a game
  - $H_A : \mu_{points} < 20$ I score less than 20 points on average
- Does this medicine change the probability of recovery?
  - $H_0 : p_{medicine} - p_{control} = 0$ Taking this medicine has no effect on a patient's probability of recovery
  - $H_A : p_{medicine} - p_{control} \neq 0$ The probability of recovery is different for patients who take the medicine
- Does education have an effect on income?
  - $H_0 : \beta_{edu} = 0$ There is no relationship between education level and income
  - $H_A : \beta_{edu} > 0$ There is a positive relationship between education level and income

# Test Statistic

Once we have established our hypothesis, we gather data (through an experiment or observational study). We then summarize our data into a **test statistic**. This is typically the natural estimate of the parameter of interest (what we made a statement about in our hypothesis)

# Test Statistic

Once we have established our hypothesis, we gather data (through an experiment or observational study). We then summarize our data into a **test statistic**. This is typically the natural estimate of the parameter of interest (what we made a statement about in our hypothesis)

If we assume the null hypothesis is true, we would expect our test statistic to follow the null distribution (these are the sampling distributions we talked about last week). When we actually observe data and calculate the test statistic, we can make a statement about how likely the data is under the null hypothesis

# 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"**

# Basketball Example

If I do actually score 20 pts on average, for me

- To score an avg of 18 points in the 2 games you observe is probably pretty likely (large p-value)
- To score an avg of 10 points in the 2 games you observe is probably less likely (smaller p-value)
- To score an avg of 10 points in the 15 games you observe is probably very unlikely (very small p-value)

# 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

# How do we choose an appropriate cut-off

A hypothesis test is just like a medical test. We can have false positives, false negatives, etc. The cut-off I use (the level of my test) is also the probability of a Type 1 error. Setting an appropriate cut-off depends on the context.

| | Null Hypothesis is True | Null Hypothesis is False |
|---|---|---|
| Reject Null | Type 1 error | True Positive |
| Fail to Reject Null | True Negative | Type II error |

# How do we choose an appropriate cut-off

A hypothesis test is just like a medical test. We can have false positives, false negatives, etc. The cut-off I use (the level of my test) is also the probability of a Type 1 error. Setting an appropriate cut-off depends on the context.

| | Null Hypothesis is True | Null Hypothesis is False |
|---|---|---|
| Reject Null | Type 1 error | True Positive |
| Fail to Reject Null | True Negative | Type II error |

The sensitivity of my hypothesis test $P(\text{Reject}|H_0 \text{ is false})$ is known as the **power** of my test

# Cautions against P-values

A low p-value simply means that the data we saw is improbable under the null hypothesis, not that it is impossible.

- If I run 100 experiments, I would expect a 1/20 event (event with p-value of .05) to happen about 5 times
- Even if the null hypothesis is correct every time, I will accidentally reject the null hypothesis 1/20 of the time
- If you plan on testing many things at once, you need to account for that in the cut-off for statistical significance
- Forming your hypothesis after you have collected and looked through your data can sometimes discover unexpected true results, but more often than not results in erroneous results. This is sometimes called **Data snooping** or **data fishing**.

`https://xkcd.com/882/`

# Cautions against P-values

The "Reproducibility Project: Psychology" set out to reproduce 100 findings published in top psychology journals in 2008.

- Project found that they could only replicate the findings of 39% of the articles
- Most people are not intentionally trying to publish wrong findings, but if you do enough experiments, you will stumble onto results that are statistically significant even though they are not true

`http://science.sciencemag.org/content/349/6251/aac4716/`

# Replicability

The "Reproducibility Project: Psychology" set out to reproduce 100 findings published in top psychology journals in 2008.

- Project found that they could only replicate the findings of 39% of the articles
- Most people are not intentionally trying to publish wrong findings, but if you do enough experiments, you will stumble onto results that are statistically significant even though they are not true

`http://science.sciencemag.org/content/349/6251/aac4716/`

# Details about Hypothesis Testing

The specifics of a hypothesis test depend on the details

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

# Examples