Handout 1.2: Introduction to Hypothesis Testing

Dr. Bean - Stat 5100

1 Why Hypothesis Testing?

In statistics, hypothesis tests are a way to determine if an **observed** difference is **significant** or simply due to chance. The key ingredients of a hypothesis test are:

- A null and alternative hypothesis.
- An observed statistic taken from a sample of the population.
  - Example: $t = \frac{\bar{X}-\mu_0}{SE\{X\}}$
- An assumed probability distribution for the test statistic IF the null hypothesis is true.

The climax of a hypothesis test is the determination of the **p-value**. If the p-value is small (< 0.05), we reject the null hypothesis, and if it is not small, we fail to reject the null hypothesis.
Note that we *never* accept the alternative hypothesis, we simply *fail to reject* the null hypothesis.

(Individual) What is a p-value?

The probability of obtaining our sample statistic, or one more extreme, if the null hypothesis was true.

(Groups) Determine whether or not the following statements are true or false:

- **The p-value is the probability that the null hypothesis is true. (FALSE)**
- **We reject a null hypothesis when the p-value is small. (TRUE)**
- **If the p-value is very small, it is not possible that the null hypothesis is true. (FALSE)**
- **The difference between the sample mean and the population mean is all that matters in the test statistic. (FALSE)**

(Individual) Why is the assumed distribution for the test-statistic such a big deal?

We use the assumed probability distribution is how we determine the p-value, and the p-value is how we determine the "significance" of our results. If the probability distribution is not appropriate then the p-value will be worthless.

How can we know the distribution of the test statistic?

- Through visualizations: Histograms, qqplots, boxplots.
- More often, the **Central Limit Theorem** assures us that the test statistic will follow a normal probability distribution.

2 Example:

Researchers have studied how the amount of sunlight bamboo is exposed to affects the speed of growth. One study compared the growth of 50 bamboo shoots grown under standard conditions to the growth of 49 bamboo shoots that had been exposed to 10% more sunlight. The growth was measured 40 days after planting. The observed mean and sd for bamboo under the standard growing conditions were 32.04 inches and 5.82 inches while the observed mean and sd for the more sunlight bamboo were 28.61 inches and 6.32 inches.
Hypothesis:

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

We will never know the values of μ_1 and μ_2. They are **population parameters** that we could only know if we sampled every person in the population.

Rather, we **estimate** the values of these population parameters with our samples, giving us values of $\bar{X}_1 = 32.04$ and $\bar{X}_2 = 28.61$ and an observed difference of $\bar{X}_1 - \bar{X}_2 = -3.43$.

If the null hypothesis was true then the observed difference would follow a t-distribution centered at 0 with a standard deviation (assuming pooled variances) of 1.221. This also means that our observed value of $t = \frac{-3.43}{1.221} = -2.81$. The p-value associated with our observation is 0.006.

The p-value in this setting is:

> The probability of having an observed difference between the two bamboo groups as, or more, extreme than -3.43 IF the null hypothesis (no difference) was in fact true.

The p-value says that our observed difference would have been **very unlikely** (less than a 1/100 chance) if the null hypothesis was actually true. This gives us evidence to **reject the null hypothesis** and conclude that the growth rate of bamboo is different when sunlight conditions change.

To summarize:

1. We make a claim about the value of the population parameters.
2. We test the claim by obtaining statistics from a sample of the population.
3. We determine the probability of obtaining our sample statistic (or something more extreme) IF the null hypothesis was true.
4. If the probability of our observation is LOW, we reject the null hypothesis, if it is NOT LOW, then we fail to reject the null hypothesis.

(Individual) Suppose the p-value for the above example had been 0.02 instead of 0.006. Would you conclusions change? What about if the p-value had been 0.13? How about 0.98?

Same conclusions for 0.02, but fail to reject the null hypothesis for p-values of 0.13 and 0.98.

(Individual) What is the different between a practical difference and a "significant" difference?

As seen before, a practical difference is not always significant, but a significant difference is not always practical.
When sample sizes are LARGE, nearly every difference is flagged as significant, even if the actual difference between groups is small.

3 Inference vs Prediction

In linear modeling, we assume that the population follows the model:

$$Y = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

In these models, we can conduct **inference** to determine if the linear relationship between an explanatory variable X_k and the response variable Y is significant.

HOWEVER, we can also use these same models to try and make accurate **predictions** of Y.

Models that are accurate tend to have significant coefficients, but models with significant coefficients are not always accurate.

Our approach to linear modeling in this class changes slightly when our primary interest is establishing significance, vs being accurate.

(Groups) Can you think of an example when our primary motive for creating a model is to create accurate predictions? How about an example where the primary motive is determining the significance of the coefficients?

Accuracy: Predicting the market value of a house given square footage, lot size, etc.

Significance: Determining if there is a statistically significant gender bias in pay, after accounting for other demographic factors.