Lecture 16: Testing Hypotheses

- Introduction to Hypothesis Testing
- Hypothesis Testing Setup
- Examples

Introduction

– So far in this class, we have been talking about making inference on a population using observed data.

– In line with that, we have talked about building confidence intervals for population parameters by relying on central limit theorem and bootstrap.

– Today we will extend that discussion to what is called **hypotheses testing**. Loosely speaking, hypotheses testing is asks the question: how strongly does the data support my preconceived hypothesis about a population parameter?

– This is the same as asking: if my preconceived hypothesis about a population parameter is true, how extreme is the data I just observed?

Is There a Gender Discrimination?

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

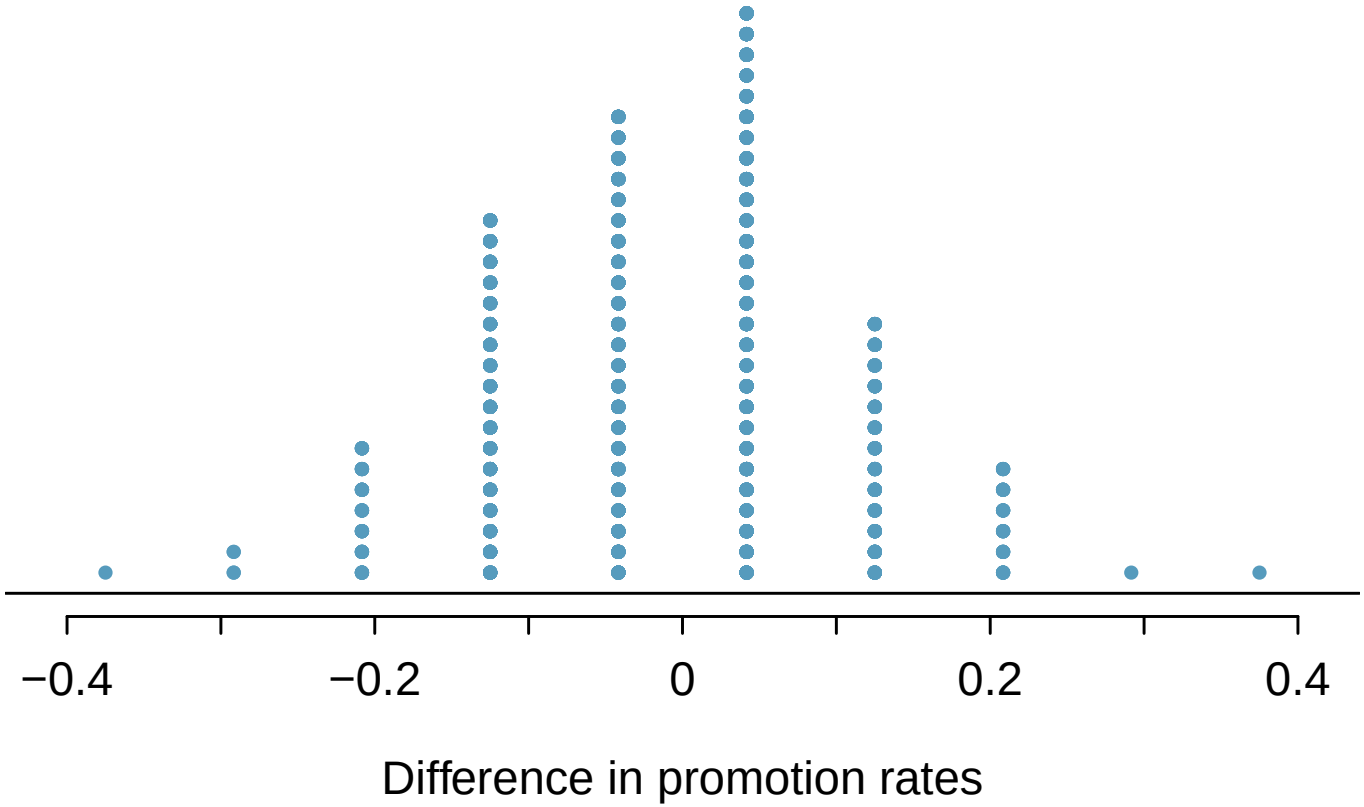
$$\hat{p}_{\text{males}} = 21/24 \approx 0.88$$

$$\hat{p}_{\text{females}} = 14/24 \approx 0.58$$

Possible explanations:

- Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance. $\rightarrow$ *null* - (nothing is going on)
- Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance. $\rightarrow$ *alternative* - (something is going on)

Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we decided to reject the null hypothesis in favor of the alternative.

Hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via *simulation* or traditional methods based on the *central limit theorem* (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the **hypothesis testing** framework using an example on testing a claim about a population mean.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the average number of exclusive relationships college students have been in to be (2.7, 3.7). Based on this confidence interval, do these data support the hypothesis that college students on average have been in more than 3 exclusive relationships.

- The associated hypotheses are:

  H_0: $\mu = 3$: College students have been in 3 exclusive relationships, on average

  H_A: $\mu > 3$: College students have been in more than 3 exclusive relationships, on average
- Since the null value is included in the interval, we do not reject the null hypothesis in favor of the alternative.
- This is a **quick-and-dirty** approach for hypothesis testing. However it doesn't tell us the likelihood of certain outcomes under the null hypothesis, i.e. the **p-value**, based on which we can make a decision on the hypotheses.

Number of college applications

A similar survey asked how many colleges Duke students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability.
- We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended)

$$H_0 : \mu = 8$$

- We test the claim that the average number of colleges Duke students apply to is greater than 8

$$H_A : \mu > 8$$

Number of college applications - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Students in the sample should be independent of each other with respect to how many colleges they applied to.

(b) Sampling should have been done randomly.

(c) The sample size should be less than 10% of the population of all Duke students.

(d) There should be at least 10 successes and 10 failures in the sample.

(e) The distribution of the number of colleges students apply to should not be extremely skewed.

Number of college applications - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Students in the sample should be independent of each other with respect to how many colleges they applied to.

(b) Sampling should have been done randomly.

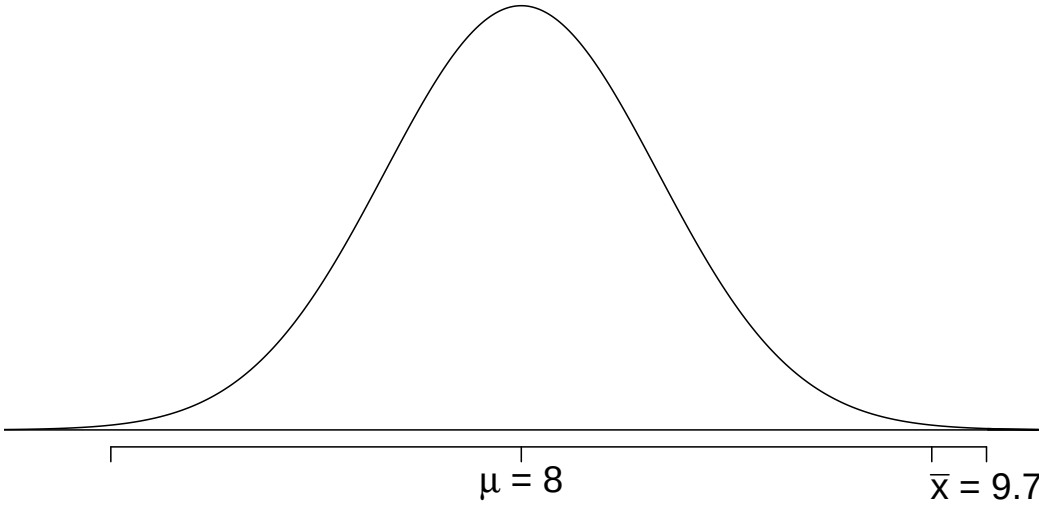
(c) The sample size should be less than 10% of the population of all Duke students.

(d) *There should be at least 10 successes and 10 failures in the sample.*

(e) The distribution of the number of colleges students apply to should not be extremely skewed.

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



$$\overline{X} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

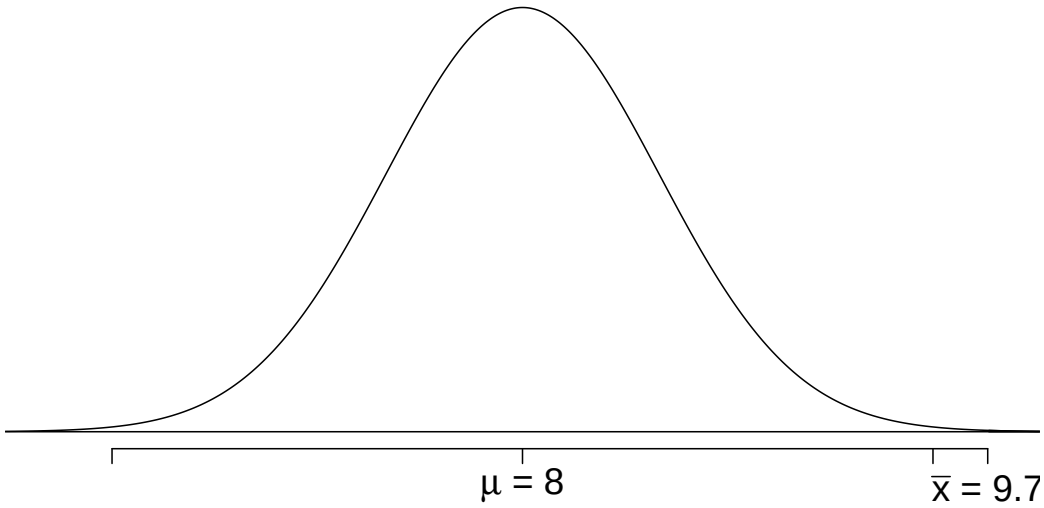
Yes, and we can quantify how unusual it is using a p-value.

p-values

- We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* H_0.

Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\overline{X} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject* H_0.
- The data provide convincing evidence that Duke students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *unlikely to be due to chance or sampling variability*.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

(a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

(d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

(e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

(a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(b) *Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.*

(c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

(d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

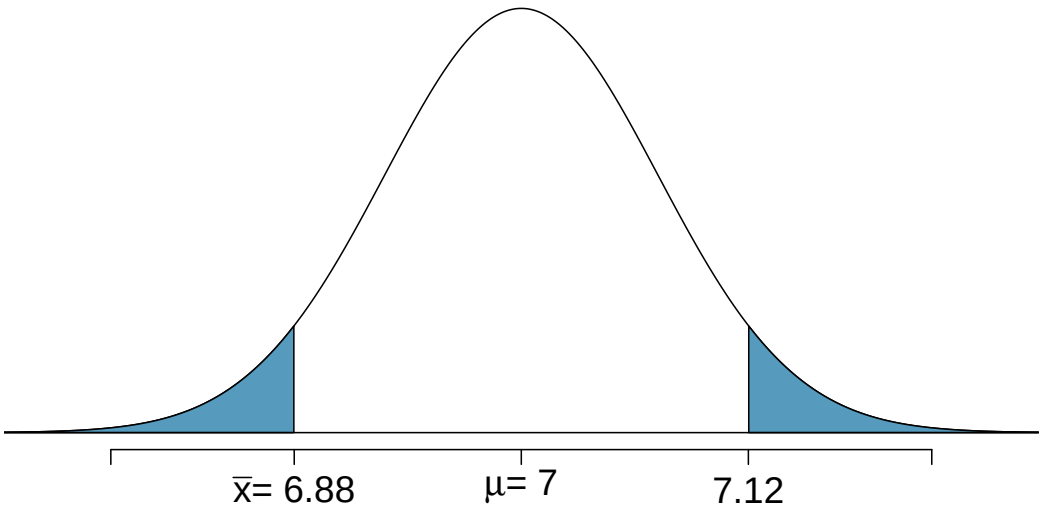
(e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

- If the research question was "Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?", the alternative hypothesis would be different.

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

- Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$

Steps for Testing Hypotheses

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

Hypothesis testing for a population mean

1. Set the hypotheses
   - $H_0 : \mu = null\ value$
   - $H_A : \mu <$ or $>$ or $\neq\ null\ value$
2. Calculate the point estimate
3. Check assumptions and conditions
   - Independence: random sample/assignment, 10% condition when sampling without replacement
   - Normality: nearly normal population or $n \geq 30$, no extreme skew – or use the t distribution
4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\overline{X} - \mu}{\text{SE}},\ where\ \text{SE} = \frac{s}{\sqrt{n}}$$

5. Make a decision, and interpret it in context
   - If p-value $< \alpha$, reject H_0, data provide evidence for H_A
   - If p-value $> \alpha$, do not reject H_0, data do not provide evidence for H_A

Hypotheses Tests

A hypothesis test (significance test) is a way to decide whether the data strongly support one point of view or another.

There are many kinds of significance tests, but all involve:

- a null and alternative hypothesis
- a test statistic
- a significance probability (P-value).

The following gives an overview of most of the different kinds of significance tests.

Hypotheses Tests

Step 1: Pick the null and alternative hypotheses.

The null and alternative are two contradictory statements about a parameter and their union is the set of all possible parameter values. For example:

$$H_0 : \theta \leq 90 \qquad \textbf{vs.} \qquad H_1 : \theta > 90$$

where θ is a generic parameter. **Usually, the null hypothesis H_0 is a current belief while the alternative hypothesis H_1 or H_A is the one that leads to new action, or the outcome you would like to prove wrong.**

Example 1: H_0: The mean cable strength $\geq$ 3 tons.
H_1 : The mean cable strength $<$ 3 tons.

Example 2: H_0 : The sd in income $\leq$ \$5,000
H_1 : The sd in income $>$ \$5,000.

In example 1, accepting that mean cable strength $<$ 3 tons probably **leads to a new action** and for example 2, sd in income $>$ \$5,000 **leads to a new action**.

Hypotheses Tests

Step 2: Calculate the test statistic.

The test statisic is a one-number summary of all the information in the sample **regarding the correctness of the alternative hypothesis**. Different kinds of hypothesis tests (e.g., about means, proportions, differences of means, differences of proportions, etc.) require different test statistics.
Soon we shall list many standard cases.

Step 3: Find the *P*-value (or significance probability).

Use a table to find the *P*-value. This is **"the probability of obtaining data that is as or more supportive of the alternative hypothesis than the data that were observed, when the null hypothesis is correct."**
This interpretation of the *P*-value is a bit subtle.

Hypotheses Tests

I: The Three Possible Pairs of Null and Alternative Hypotheses

1. $H_0: \theta = \theta_0$ versus $H_A: \theta \neq \theta_0$
2. $H_0: \theta \leq \theta_0$ versus $H_A: \theta > \theta_0$
3. $H_0: \theta \geq \theta_0$ versus $H_A: \theta < \theta_0$

Here θ represents a generic parameter. It could be a population mean, a population proportion, the difference of two population means, or many other things.

The θ_0 is the **null value**. Often it is a value specified in a contract, regulation, or clinical trial.

Hypotheses Tests

II Possible Test Statistics In this class, we mostly consider simple test statistics of the form

$$ts = \frac{pe - \theta_0}{se} = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})}$$

Where

- pe or $\hat{\theta}$ is the point estimate for θ.
- θ_0 is the null value.
- se is the standard error of $\hat{\theta}$.

Clearly, when $\hat{\theta}$ is an average (or sum) and θ_0 is true, the law of large numbers and the central limit theorem kicks in, so that ts is like a "z-transformation" and it has a standard normal or student-t distribution and we can find the probability that it exceeds some critical value.

Hypotheses Tests

a. For a test on the population mean we take θ to be the population mean μ. If you know the population σ, or for $n > 31$ with

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum(X_i - \overline{X})^2}$$

as an estimate of the σ, then you get the significance probability from a z-table and the test statistic is:

$$ts = \frac{\overline{X} - \mu_o}{\hat{\sigma}/\sqrt{n}}.$$

b. For the previous case, if you have a sample of size $n \leq 31$ and you use $\hat{\sigma}$ to estimate the population σ, then the significance probability comes from a t_{n-1} table and again the test statistic is:

$$ts = \frac{\overline{X} - \mu_o}{\hat{\sigma}/\sqrt{n}}.$$

Hypotheses Tests

c. For a test about a proportion, $\theta = p$. The significance probability comes from a z-table and the test statistic is:

$$ts = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

d. For a test of the difference of two means, $\theta = \mu_1 - \mu_2$. Assuming that the sample sizes from each population satisfy $n_1 > 30$ and $n_2 > 30$, then the significance probability comes from a z-table and the test statistic is:

$$ts = \frac{(\overline{X}_1 - \overline{X}_2) - \theta_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

If $n_1 > 30$ and $n_2 > 30$ isn't satisfied, use a t-table with $\min(n_1 - 1, n_2 - 1)$ degrees of freedom (use $n_1 + n_2 - 2$ only if equal variances of the two groups are assumed).

Hypotheses Tests

e. For a test of the difference of two proportions, take $\theta = p_1 - p_2$. Use a z-table for the significance probability and the test statistic:

$$ts = \frac{(\hat{p}_1 - \hat{p}_2) - \theta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

f. For $n > 30$, with $\theta = \mu_1 - \mu_2$, and n paired differences $X_i - Y_i$, use the z-table for the significance probability. The test statistic is:

$$ts = \frac{(\overline{X} - \bar{Y}) - \theta_0}{\hat{\sigma}_d / \sqrt{n}}$$

Here $\hat{\sigma}_d$ is the estimated standard deviation of the n differences. This is called a **paired difference test** and can have greater power than the two-sample tests in II.d and II.e. **We will talk about what power means later**. If $n \leq 30$, use a t-table with $n - 1$ degrees of freedom.

Hypotheses Tests

III Significance Probability or P-value

The significance probability of the test statistic depends upon the hypothesis chosen in Part I. For that choice, let W be a random variable with a z or t_{n-1} distribution, as indicated in Part II. Then for each possible pair of null and alternative hypotheses in Part I,

1. The significance probability is $\mathbb{P}[W \leq -|ts|] + \mathbb{P}[W \geq |ts|]$.
2. The significance probability is $\mathbb{P}[W \geq ts]$.
3. The significance probability is $\mathbb{P}[W \leq ts]$.

The significance probability is **THE CHANCE OF OBSERVING DATA THAT SUPPORTS THE ALTERNATIVE HYPOTHESIS AS OR MORE STRONGLY THAN THE DATA YOU HAVE SEEN, WHEN THE NULL HYPOTHESIS IS CORRECT.**

Hypotheses Tests

With all the pieces, a decision can be reached on whether to **reject the null hypothesis if there is too much evidence against it**, or to **fail to reject the null hypothesis if there isn't enough evidence to suggest it is false**.

Note: We don't actually "accept the alternative hypothesis" since intuitively, our decision is based on a sample from the population.

Thus, formally, for each possible pair of null and alternative hypotheses in Part I, **reject H_0 if the significance probability or p-value is less than a chosen error rate α (usually 0.05 or 0.01).**

Examples

Example: Suppose you have a new oil additive that may extend the life of an engine. You give it to 25 random motors and find that the average lifespan is 78 months, and the standard deviation in their lifespan is 12. You know that with unmodified oil, the mean lifespan is 72 months, and hope to show that your additive improves that.

The first step is to choose the null and alternative hypotheses. You put what you want to prove in the alternative, so this is case I.2 of the previous taxonomy:

H_0: The mean lifetime with the new additive $\leq$ 72 months.

H_A : The mean lifetime with the new additive $>$ 72 months.

The second step is to find the test statistic. We are in case II.b of the taxonomy, so:

$$ts = \frac{\overline{X} - \mu_0}{sd/\sqrt{n}} = \frac{78 - 72}{12/\sqrt{25}} = 2.5$$

Examples

The third step finds the significance probability. If you use hypotheses I.**2**, then you use rule III.**2**.

The significance probability, or P-value, is $\mathbb{P}[t_{24} > 2.5]$ and from the t-table in the book, 2.5 isn't actually on the table but we can provide bounds for it (note that we need to find 2.5 on the table and find the corresponding α value). Thus,

$$0.01 = \mathbb{P}[t_{24} > 2.492] > \mathbb{P}[t_{24} > 2.5] > \mathbb{P}[t_{24} > 2.797] = 0.005.$$

So if the null hypothesis is true and the additive does not help, then you have between a 1% chance and a 0.5% chance of observing the result in your experiment, which is so rare. This is pretty persuasive that the additive helps.

Also, if we let $\alpha = 0.05$, then the p-value is less than α and we can reject the null hypothesis that the new additive does not help and conclude that it does help.

Recap

Today we learned about hypotheses testing and how to setup testing problems.

In the next lecture, we will learn about what α really means and why we use it as a cut-off in hypotheses testing (and confidence intervals). The duality between confidence intervals and hypotheses testing

Suggested reading:

- D.S. Sec. 9.1, 9.2, 9.4, 9.6
- OpenIntro3: Sec. 4.3