

## Comparing Two Groups

- Chapter 7 describes two ways to compare two populations on the basis of independent samples: a **confidence interval for the difference in population means** and a **hypothesis test**.
- The basic structure of the confidence interval is the same as in the previous chapter — an estimate plus or minus a multiple of a standard error.
- Hypothesis testing will introduce several new concepts.

## Standard Error of $\bar{y}_1 - \bar{y}_2$

The **standard error** of the difference in two sample means is an empirical measure of how far the difference in sample means will typically be from the difference in the respective population means.

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

An alternative formula is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{(SE(\bar{y}_1))^2 + (SE(\bar{y}_2))^2}$$

This formula reminds us of how to find the length of the hypotenuse of a triangle.

(Variances add, but standard deviations don't.)

## Two Independent Samples

**Bret Larget**

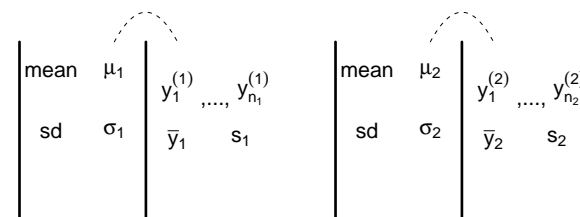
Department of Statistics

University of Wisconsin - Madison

October 17, 2003

## Setting

- Model two populations as buckets of numbered balls.
- The population means are  $\mu_1$  and  $\mu_2$ , respectively.
- The population standard deviations are  $\sigma_1$  and  $\sigma_2$ , respectively.
- We are interested in **estimating  $\mu_1 - \mu_2$**  and in testing the hypothesis that  $\mu_1 = \mu_2$ .



## Sampling Distributions

The sampling distribution of the difference in sample means has these characteristics.

- **Mean:**  $\mu_1 - \mu_2$
- **SD:**  $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- **Shape:** Exactly normal if both populations are normal, approximately normal if populations are not normal but both sample sizes are sufficiently large.

## Theory for Confidence Interval

standard deviations, then

$$\Pr \left\{ -1.96 \leq \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq 1.96 \right\} = 0.95$$

where we can choose  $z$  other than 1.96 for different confidence levels. This statement is true because the expression in the middle has a standard normal distribution.

But in practice, we don't know the population standard deviations. If we substitute in sample estimates instead, we get this.

$$\Pr \left\{ -t \leq \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t \right\} = 0.95$$

We need to choose different end points to account for the additional randomness in the denominator.

## Pooled Standard Error

If we wish to assume that the two population standard deviations are equal,  $\sigma_1 = \sigma_2$ , then it makes sense to use data from both samples to estimate the common population standard deviation.

We estimate the common population variance with a weighted average of the sample variances, weighted by the degrees of freedom.

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The **pooled standard error** is then as below.

$$SE_{\text{pooled}} = s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

## Theory for Confidence Interval

The recipe for constructing a confidence interval for a single population mean is based on facts about the sampling distribution of the statistic

$$T = \frac{\bar{Y} - \mu}{SE(\bar{Y})}$$

Similarly, the theory for confidence intervals for  $\mu_1 - \mu_2$  is based on the sampling distribution of the statistic

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE(\bar{Y}_1 - \bar{Y}_2)}$$

where we **standardize** by subtracting the mean and dividing by the standard deviation of the sampling distribution.

If **both populations are normal** and if **we know the population**

## Confidence Interval for $\mu_1 - \mu_2$

The confidence interval for differences in population means has the [same structure](#) as that for a single population mean.

$$(\text{Estimate}) \pm (t \text{ Multiplier}) \times SE$$

The only difference is that for this more complicated setting, we have [more complicated formulas](#) for the standard error and the degrees of freedom.

Here is the df formula.

$$df = \frac{(SE_1^2 + SE_2^2)^2}{SE_1^4/(n_1 - 1) + SE_2^4/(n_2 - 1)}$$

where  $SE_i = s_i/\sqrt{n_i}$  for  $i = 1, 2$ .

As a check, the value is often close to  $n_1 + n_2 - 2$ . (This will be exact if  $s_1 = s_2$  and if  $n_1 = n_2$ .) The value from the messy formula will always be between the smaller of  $n_1 - 1$  and  $n_2 - 1$  and  $n_1 + n_2 - 2$ .

## Example

A calculator or R can compute the margin of error.

```
> se = sqrt(1.34^2 + 1.3^2)
> tmult = qt(0.975, 190)
> me = round(tmult * se, 1)
> se
[1] 1.866976
> tmult
[1] 1.972528
> me
[1] 3.7
```

We are 95% confident that the mean reduction in systolic blood pressure due to the biofeedback treatment in a population of similar individuals to those in this study would be between 6.1 and 13.5 mm more than the mean reduction in the same population undergoing the control treatment.

## Theory for Confidence Interval

It turns out that the sampling distribution of the statistic above is [approximately](#) a  $t$  distribution where the degrees of freedom should be estimated from the data as well.

Algebraic manipulation leads to the following expression.

$$\Pr \left\{ (\bar{Y}_1 - \bar{Y}_2) - t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{Y}_1 - \bar{Y}_2) + t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right\} = 0.95$$

We use a  $t$  multiplier so that the area between  $-t$  and  $t$  under a  $t$  distribution with the estimated degrees of freedom will be 0.95.

## Example

### Exercise 7.12

In this example, subjects with high blood pressure are randomly allocated to two treatments. The [biofeedback group](#) receives relaxation training aided by biofeedback and meditation over eight weeks. The control group does not. Reduction in systolic blood pressure is tabulated here.

	Biofeedback	Control
$n$	99	93
$\bar{y}$	13.8	4.0
SE	1.34	1.30

For 190 degrees of freedom (which come from both the simple and messy formulas) the table says to use 1.977 (140 is rounded down) whereas with R you find 1.973.

## Example Assuming Equal Variances

For the same data, were we to assume that the population variances were equal, the degrees of freedom, the standard error, and the confidence interval are all slightly different.

```
> t.test(height ~ color, var.equal = T)
      Two Sample t-test

data: height by color
t = 1.1064, df = 40, p-value = 0.2752
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4804523  1.6428053
sample estimates:
mean in group green  mean in group red
      8.940000         8.358824
```

## Logic of Hypothesis Tests

All of the hypothesis tests we will see this semester fall into this general framework.

1. State a **null hypothesis** and an **alternative hypothesis**.
2. Gather data and compute a **test statistic**.
3. Consider the **sampling distribution of the test statistic assuming that the null hypothesis is true**.
4. Compute a **p-value**, a measure of how consistent the data is with the null hypothesis in consideration of a specific alternative hypothesis.

## Example Using R

### Exercise 7.21

This exercise examines the growth of bean plants under red and green light. A 95% confidence interval is part of the output below.

```
> ex7.21 = read.table("lights.txt", header = T)
> str(ex7.21)
'data.frame':      42 obs. of  2 variables:
 $ height: num  8.4 8.4 10 8.8 7.1 9.4 8.8 4.3 9 8.4 ...
 $ color : Factor w/ 2 levels "green","red": 2 2 2 2 2 2 2 2 2 ...
> attach(ex7.21)
> t.test(height ~ color)
      Welch Two Sample t-test

data: height by color
t = 1.1432, df = 38.019, p-value = 0.2601
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4479687  1.6103216
sample estimates:
mean in group green  mean in group red
      8.940000         8.358824
```

## Hypothesis Tests

- **Hypothesis tests** are an alternative approach to statistical inference.
- Unlike confidence intervals where the goal is **estimation with assessment of likely precision of the estimate**, the goal of hypothesis testing is to ascertain whether or not data is **consistent** with what we might expect to see assuming that a hypothesis is true.
- The logic of hypothesis testing is a probabilistic form of **proof by contradiction**.
- In logic, if we can say that a proposition  $H$  leads to a contradiction, then we have proved  $H$  false and have proved  $\{\text{not}H\}$  to be true.
- In hypothesis testing, if observed data is highly unlikely under an assumed hypothesis  $H$ , then there is strong (but not definitive) evidence that the hypothesis is false.

## Wisconsin Fast Plants Example

- In an experiment, seven Wisconsin Fast Plants (*Brassica campestris*) were grown with a treatment of Ancyamidol (ancy) and eight control plants were given ordinary water.
- The **null hypothesis** is that the treatment has no effect on plant growth (as measured by the height of the plant after 14 days of growth).
- The **alternative hypothesis** is that the treatment has an effect which would result in different mean growth amounts
- A summary of the sample data is as follows. The eight control plants had a mean growth of 15.9 cm and standard deviation 4.8 cm. The seven ancy plants had a mean growth of 11.0 cm and standard deviation 4.7 cm.
- The question is, is it reasonable to think that the observed difference in sample means of 4.9 cm is due to **chance variation alone**, or is there evidence that some of the difference is due to the ancy treatment?

## Example: Calculate a Test Statistic

In the setting of a difference between two independent sample means, our test statistic is

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Your book adds a subscript,  $t_s$ , to remind you that this is computed from the **sample**.)

For the data, we find this.

```
> se = sqrt(4.8^2/8 + 4.7^2/7)
> se
[1] 2.456769
> ts = (15.9 - 11)/se
> ts
[1] 1.994489
```

The standard error tells us that we would expect that the observed difference in sample means would typically differ from the population difference in sample means by about 2.5 cm.

## Logic of Hypothesis Tests

5. Assess the strength of the evidence against the null hypothesis in the context of the problem.

We will introduce all of these concepts in the setting of testing the equality of two population means, but the general ideas will reappear in many settings throughout the remainder of the semester.

## Example: State Hypotheses

Let  $\mu_1$  be the population mean growth with the control conditions and let  $\mu_2$  be the population mean with ancy.

The null and alternative hypotheses are expressed as

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

We state statistical hypotheses as statements about population parameters.

## Example: Find the Sampling Distribution

The sampling distribution of the test statistic is a  $t$  distribution with degrees of freedom calculated by the messy formula. This useful R code computes it. If you type this in and save your work space at the end of a session, you can use it again in the future.

```
> getDF = function(s1, n1, s2, n2) {  
+   se1 = s1/sqrt(n1)  
+   se2 = s2/sqrt(n2)  
+   return((se1^2 + se2^2)^2/(se1^4/(n1 - 1) + se2^4/(n2 - 1)))  
+ }  
> getDF(4.8, 8, 4.7, 7)  
[1] 12.80635
```

## Example: Interpreting a P-Value

The smaller the p-value, the more inconsistent the data is with the null hypothesis, the stronger the evidence is against the null hypothesis in favor of the alternative.

Traditionally, people have measured **statistical significance** by comparing a p-value with arbitrary **significance levels** such as  $\alpha = 0.05$ . The phrase “statistically significant at the 5% level” means that the p-value is smaller than 0.05.

In reporting results, it is best to report an actual p-value and not simply a statement about whether or not it is “statistically significant”.

## Example: Calculate a Test Statistic

If the population means are equal, their difference is zero. This test statistic tells us that the actual observed difference in sample means is 1.99 standard errors away from zero.

## Example: Compute a P-Value

To describe how likely it is to see such a test statistic, we can ask what is the probability that chance alone would result in a test statistic at least this far from zero? The answer is the area below  $-1.99$  and above  $1.99$  under a  $t$  density curve with 12.8 degrees of freedom.

With the  $t$ -table, we can only calculate this p-value within a range. If we round down to 12 df, the  $t$  statistic is bracketed between 1.912 and 2.076 in the table. Thus, the area to the right of 1.99 is between 0.03 and 0.04. The p-value in this problem is twice as large because we need to include as well the area to the left of  $-1.99$ . So,  $0.06 < p < 0.08$ .

With, R we can be more precise.

```
> p = 2 * pt(-ts, getDF(4.8, 8, 4.7, 7))  
> p  
[1] 0.06783269
```

## Rejection Regions

Suppose that we were asked to make a decision about a hypothesis based on data. We may decide, for example to **reject the null hypothesis** if the  $p$ -value were smaller than 0.05 and to **not reject the null hypothesis** if the  $p$ -value were larger than 0.05.

This procedure has a **significance level** of 0.05, which means that if we follow the rule, there is a probability of 0.05 of rejecting a true null hypothesis. (We would need further assumptions to calculate the probability of not rejecting a false null hypothesis.)

Rejecting the null hypothesis occurs precisely when the test statistic falls into a **rejection region**, in this case either the upper or lower 2.5% tail of the sampling distribution.

## Comparing $\alpha$ and $P$ -values

- In this setting, the significance level  $\alpha$  and  $p$ -values are both areas under  $t$  curves, but they are not the same thing.
- The significance level is a prespecified, arbitrary value, that does not depend on the data.
- The  $p$ -value depends on the data.
- If a decision rule is to reject the null hypothesis when the test statistic is in a rejection region, this is equivalent to **rejecting the null hypothesis when the  $p$ -value is less than the significance level  $\alpha$** .

## Example: Summarizing the Results

For this example, I might summarize the results as follows.

There is slight evidence ( $p = 0.068$ , two-sided independent sample  $t$ -test) that there is a difference in the mean height at 14 days between Wisconsin Fast Plants grown with ordinary water and those grown with Ancyimidol.

Generally speaking, a confidence interval is more informative than a  $p$ -value because it estimates a difference in the units of the problem, which allows the reader with background knowledge in the subject area to assess both the **statistical significance** and the **practical importance** of the observed difference. In contrast, **a hypothesis test examines statistical significance alone**.

## Relationship between $t$ tests and confidence intervals

The rejection region corresponds exactly to the test statistics for which a 95% confidence interval contains 0.

We would reject the null hypothesis  $H_0 : \mu_1 - \mu_2 = 0$  versus the two-sided alternative at the  $\alpha = 0.05$  level of significance **if and only if** a 95% confidence interval for  $\mu_1 - \mu_2$  does not contain 0.

We could make similar statements for general  $\alpha$  and a  $(1 - \alpha) \times 100\%$  confidence interval.

## More on $P$ -Values

Another way to think about  $P$ -values is to recognize that they depend on the values of the data, and so are **random variables**. Let  $P$  be the  $p$ -value from a test.

- If the null hypothesis is true, then  $P$  is a **random variable distributed uniformly between 0 and 1**.
- In other words, the probability density of  $P$  is a flat rectangle.
- Notice that this implies that  $\Pr\{P < c\} = c$  for any number  $c$  between 0 and 1. If the null is true, there is a 5% probability that  $P$  is less than 0.05, a 1% probability  $P$  is less than 0.01, and so on.
- On the other hand, if the alternative hypothesis is true, then the distribution of  $P$  will be not be uniform and instead will be shifted toward zero.

## More $P$ -value Interpretations

A verbal definition of a  $p$ -value is as follows.

The  $p$ -value of the data is the probability calculated assuming that the null hypothesis is true of obtaining a test statistic that deviates from what is expected under the null (in the direction of the alternative hypothesis) at least as much as the actual data does.

**The  $p$ -value is not the probability that the null hypothesis is true.** Interpreting the  $p$ -value in this way will mislead you!

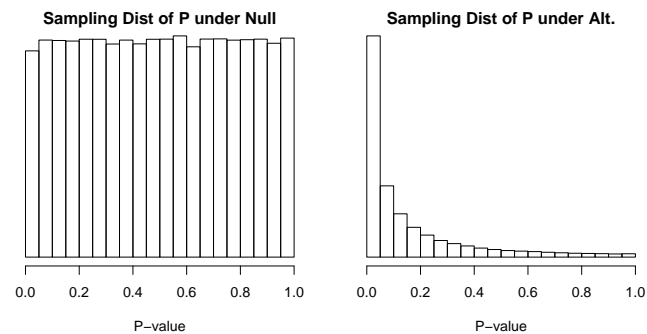
## Type I and Type II Errors

There are two possible decision errors.

- Rejecting a true null hypothesis is a **Type I error**.
- You can interpret  $\alpha = \Pr\{\text{rejecting } H_0 | H_0 \text{ is true}\}$ , so  $\alpha$  is the probability of a Type I error. (You cannot make a Type I error when the null hypothesis is false.)
- Not rejecting a false null hypothesis is a **Type II error**.
- It is convention to use  $\beta$  as the probability of a Type II error, or  $\beta = \Pr\{\text{not rejecting } H_0 | H_0 \text{ is false}\}$ . If the null hypothesis is false, one of the many possible alternative hypotheses is true. It is typical to calculate  $\beta$  separately for each possible alternative. (In this setting, for each value of  $\mu_1 - \mu_2$ .)
- **Power** is the probability of rejecting a false null hypothesis.  $\text{Power} = 1 - \beta$ .

## Simulation

We can explore these statements with a simulation based on the Wisconsin Fast Plants example. The first histogram shows  $p$ -values from 100,000 samples where  $\mu_1 - \mu_2 = 0$  while the second assumes that  $\mu_1 - \mu_2 = 5$ . Both simulations use  $\sigma_1 = \sigma_2 = 4.8$  but the calculation of the  $p$ -value does not.



## Example (cont.)

Here is a table of the results in a hypothetical population of 100,000 people.

		True Situation		Total
		Healthy ( $H_0$ is true)	Ill ( $H_0$ is false)	
Test	Negative (do not reject $H_0$ )	94,050	200	94,250
Result	Positive (reject $H_0$ )	4,950	800	5,750
Total		99,000	1,000	100,000

Notice that of the 5750 times  $H_0$  was rejected (so that the test indicated illness), the person was actually healthy  $4950/5750 = 86\%$  the time!

A rule that rejects  $H_0$  when the  $p$ -value is less than 5% only rejects 5% of the true null hypotheses, but this can be a large proportion of the total number of rejected hypotheses when the false null hypotheses occur rarely.

## Exercise 7.54

The following data comes from an experiment to test the efficacy of a drug to reduce pain in women after child birth. Possible pain relief scores vary from 0 (no relief) to 56 (complete relief).

Treatment	$n$	Pain Relief Score	
		mean	sd
Drug	25	31.96	12.05
Placebo	25	25.32	13.78

**State hypotheses.**

Let  $\mu_1$  be the population mean score for the drug. and  $\mu_2$  be the population mean score for the placebo.

$$H_0: \mu_1 = \mu_2 \quad H_A: \mu_1 > \mu_2$$

## Example for P-value Interpretation

In a medical testing setting, we may want a procedure that indicates when a subject has a disease. We can think of the decision **healthy** as corresponding to a null hypothesis and the decision **ill** as corresponding to the alternative hypothesis.

Consider now a situation where 1% of a population has a disease. Suppose that a test has an 80% chance of detecting the disease when a person has the disease (so the power of the test is 80%) and that the test has a 95% of correctly saying the person does not have the disease when the person does not (so there is a 5% chance of a false positive, or false rejecting the null).

## One-tailed Tests

- Often, we are interested not only in demonstrating that two population means are different, but in demonstrating that the difference is in a particular direction.
- Instead of the two-sided alternative  $\mu_1 \neq \mu_2$ , we would choose one of two possible one-sided alternatives,  $\mu_1 < \mu_2$  or  $\mu_1 > \mu_2$ .
- For the alternative hypothesis  $H_A: \mu_1 < \mu_2$ , the  $p$ -value is the area to the left of the test statistic.
- For the alternative hypothesis  $H_A: \mu_1 > \mu_2$ , the  $p$ -value is the area to the right of the test statistic.
- If the test statistic is in the direction of the alternative hypothesis, the  $p$ -value from a one-sided test will be half the  $p$ -value of a two-sided test.

## Exercise 7.54

---

### Summarize the results.

There is fairly strong evidence that the drug would provide more pain relief than the placebo on average for a population of women similar to those in this study ( $p = 0.038$ , one-sided independent sample  $t$ -test).

Notice that this result is “statistically significant at the 5% level” because the  $p$ -value is less than 0.05.

For a two-sided test, the  $p$ -value would be twice as large, and not significant at the 5% level.

## Permutation Tests

---

- The idea of a [permutation test](#) in this setting is quite straightforward.
- We begin by computing the difference in sample means for the two samples of sizes  $n_1$  and  $n_2$ .
- Now, imagine taking the group labels and mixing them up ([permuting them](#)) and then assigning them at random to the observations. We could then again calculate a difference in sample means.
- Next, imagine doing this process over and over and collecting the [permutation sampling distribution](#) of the difference in sample means.
- If the difference in sample means for the actual grouping of the data is atypical as compared to the differences from random groupings, this indicates evidence that the actual grouping [is associated with the measured variable](#).
- The  $p$ -value would be the proportion of random relabellings with sample mean differences at least as extreme as that from the original groups.

## Exercise 7.54

---

### Calculate a test statistic.

```
> ts = (31.96 - 25.32)/sqrt(12.05^2/25 + 13.78^2/25)
> ts
[1] 1.813664
```

### Find the null sampling distribution.

The book reports a  $t$  distribution with 47.2 degrees of freedom. We can check this.

```
> degf = getDF(12.05, 25, 13.78, 25)
> degf
[1] 47.16131
```

### Compute a (one-sided) $p$ -value.

```
> p = 1 - pt(ts, degf)
> p
[1] 0.03804753
```

## Validity of $t$ Methods

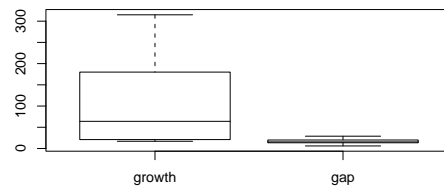
---

- All of the methods seen so far are formally based on the assumption that populations are normal.
- In practice, they are valid as long as the sampling distribution of the difference in sample means is approximately normal, which occurs when the sample sizes are large enough (justified by the Central Limit Theorem).
- Specifically, we need the sampling distribution of the test statistic to have an approximate  $t$  distribution.
- [But what if the sample sizes are small and the samples indicate non-normality in the populations?](#)
- One approach is to [transform](#) the data, often by [taking logarithms](#), so that the transformed distribution is approximately normal.
- The textbook suggests a nonparametric method called the [Wilcoxon-Mann-Whitney](#) test that is based on converting the data to ranks.
- I will show an alternative called a [permutation test](#).

## Example

Soil cores were taken from two areas, an area under an opening in a forest canopy (the gap) and a nearby area under an area of heavy tree growth (the growth). The amount of carbon dioxide given off by each soil core (in mol CO<sub>2</sub>g soil/hr).

```
> growth = c(17, 20, 170, 315, 22, 190, 64)
> gap = c(22, 29, 13, 16, 15, 18, 14, 6)
> boxplot(list(growth = growth, gap = gap))
```



## Permutation Tests

- With very small samples, it is possible to enumerate all possible ways to divide the  $n_1 + n_2$  total observations into groups of size  $n_1$  and  $n_2$ .
- An R function can carry out a permutation test.

## Example Permutation Test in R

```
> library(exactRankTests)
> perm.test(growth, gap)
      2-sample Permutation Test
```

```
data: growth and gap
T = 798, p-value = 0.006371
alternative hypothesis: true mu is not equal to 0
```

There is very strong evidence ( $p = 0.0064$ , two sample permutation test) that the soil respiration rates are different in the gap and growth areas.