Chapter 6

Section 6.4-D: Distribution of a Difference in Means

Recall that in chapter 3.1 we already learned about the distribution of the sample proportion.

A Sampling Ditribution is the distribution of sample statistics computed for different samples of the same size from the same population. A sampling distribution shows us how the sample statistic varies from sample to sample.

Properties of Sampling distribution:

- Center: If samples are randomly selected, the sampling distribution will be centered around the population parameter. (for difference of two population mean: μ_1-μ_2)
- Shape: For most of the statistics we consider, if the sample size is large enough the sampling distribution will be symmetric and bell-shaped.

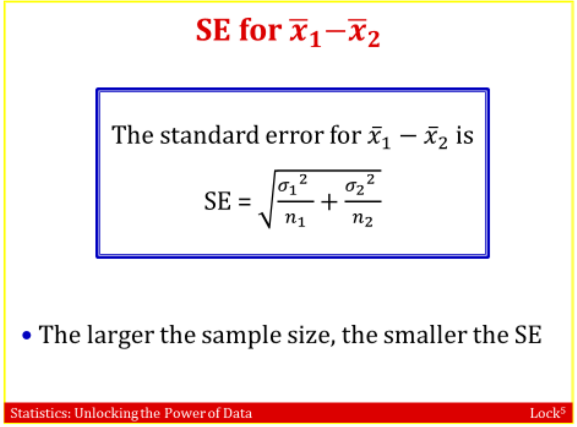
In Chapter 5 we saw the generic formula for the sampling distribution:

Statistic ~ N(parameter, SE)

We can change the generic formula to describe the sampling distribution of differences in sample means:

$$\bar{x}_1 - \bar{x}_2 \sim N(\mu_1 - \mu_2,\ SE)$$

Notice that we need the standard error. The formula for the standard error when describing the sampling distribution of a difference in means is …



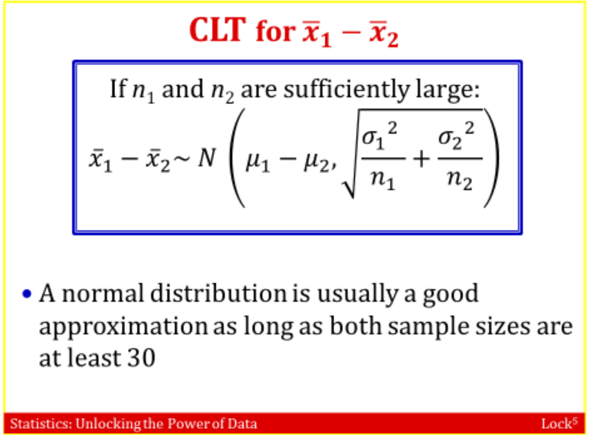
SE for $\bar{x}_1 - \bar{x}_2$

The standard error for $\bar{x}_1 - \bar{x}_2$ is

$$SE = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

- The larger the sample size, the smaller the SE

Statistics: Unlocking the Power of Data Lock[5]



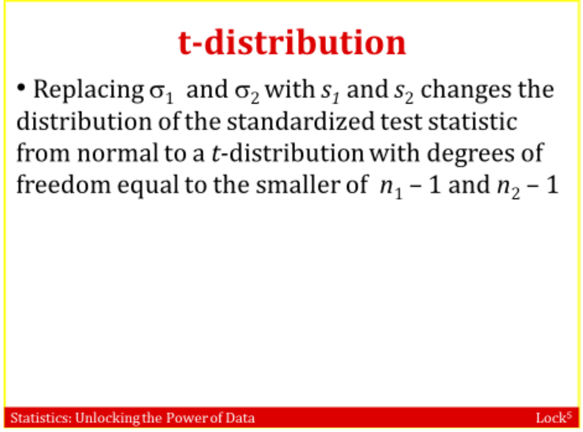
CLT for $\bar{x}_1 - \bar{x}_2$

If n_1 and n_2 are sufficiently large:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}\right)$$

- A normal distribution is usually a good approximation as long as both sample sizes are at least 30

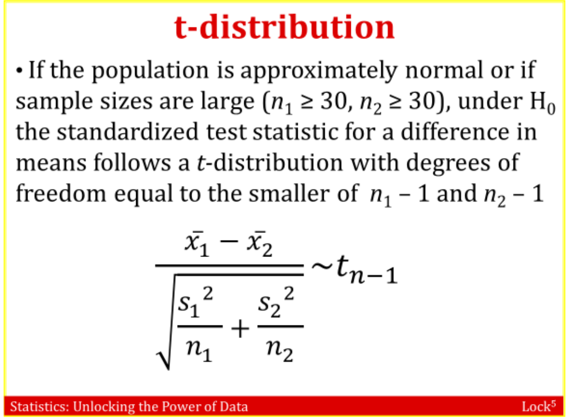
Statistics: Unlocking the Power of Data Lock[5]



t-distribution

- Replacing σ_1 and σ_2 with s_1 and s_2 changes the distribution of the standardized test statistic from normal to a *t*-distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$

Statistics: Unlocking the Power of Data Lock[5]



t-distribution

- If the population is approximately normal or if sample sizes are large ($n_1 \geq 30$, $n_2 \geq 30$), under H_0 the standardized test statistic for a difference in means follows a *t*-distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \sim t_{n-1}$$

Statistics: Unlocking the Power of Data Lock[5]

Example 1: Salaries of Baseball Players

Of the 855 major league baseball players in the 2012 season, there were 423 pitchers and 432 batters playing other positions. The average salary for the pitchers was 3.189 million dollars with a standard deviation of 4.288, while the average salary for the batters was 3.683 million dollars with a standard deviation of 5.060. Suppose that we take random samples of 30 pitchers and 50 batters and calculate the difference in mean salary between the two groups (Pitchers – Batters).

(a) Describe the shape, mean, and standard error of the distribution of differences in means.

The distribution will be bell-shaped with a mean of $\mu_P - \mu_B = 3.189 - 3.683 = -0.494$ *million dollars and a standard error of*

$$SE = \sqrt{\frac{4.288^2}{30} + \frac{5.060^2}{50}} = 1.061.$$

(b) How many degrees of freedom would you use in this situation for a t-distribution when doing inferences for the difference in sample means?

Since the sample sizes are $n_P = 30$ *and* $n_B = 30$ *the degrees of freedom for a t-distribution for the difference in means would be* $30 - 1 = 29$ *(the smaller of the df for the two samples).*

Example 2: Use a t-distribution to answer this question. Assume the samples are random samples from distributions that are reasonably normally distributed, and that a t-statistic will be used for inference about the difference in sample means. State the degrees of freedom used.

Find the endpoints of the t-distribution with 2.5% beyond them in each tail if the samples have sizes n1=13 and n2=9.

Enter the exact number for the degrees of freedom and round your answer for the endpoints to two decimal places.

Degrees of freedom = 8

endpoints = ± 2.306

Example 3: Use a t-distribution to answer this question. Assume the samples are random samples from distributions that are reasonably normally distributed, and that a t-statistic will be used for inference about the difference in sample means. State the degrees of freedom used.

Find the proportion in a t-distribution below -2.7 if the samples have sizes n1=12 and n2=32.

Enter the exact answer for the degrees of freedom and round your answer for the area to four decimal places.

Degrees of freedom = 11

proportion = 0.0103

Section 6.4-CI: Confidence Interval for a Difference in Means

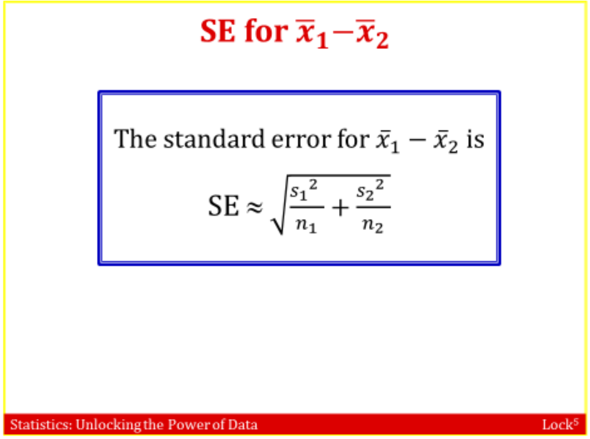
In Chapter 5 we saw the generic formula for a confidence interval:

Statistic ± Critical Value · SE

We can change the generic formula to describe a confidence interval for a difference in means:

$$\bar{x}_1 - \bar{x}_2 \pm critical\ value\ \times SE$$

Notice that we need the standard error. The formula for the standard error when describing a confidence interval for a difference in means is …

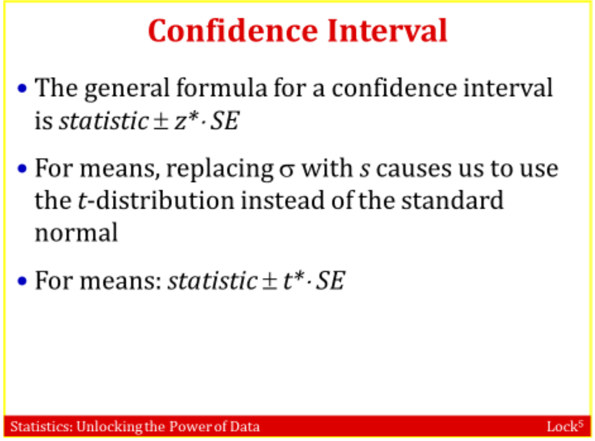


SE for $\bar{x}_1 - \bar{x}_2$

The standard error for $\bar{x}_1 - \bar{x}_2$ is

$$\text{SE} \approx \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

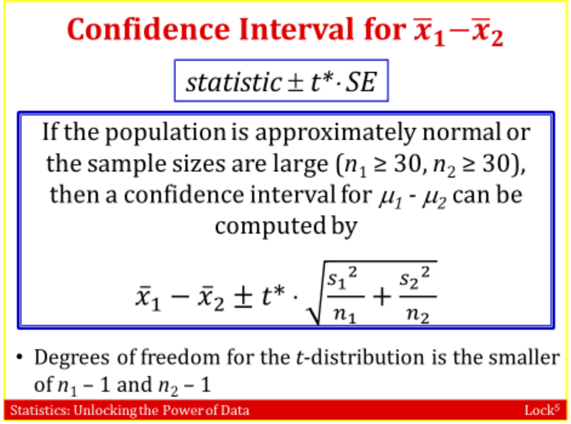
Statistics: Unlocking the Power of Data Lock5



Confidence Interval

- The general formula for a confidence interval is *statistic* ± *z** · *SE*
- For means, replacing σ with *s* causes us to use the *t*-distribution instead of the standard normal
- For means: *statistic* ± *t** · *SE*

Statistics: Unlocking the Power of Data Lock5



Confidence Interval for $\bar{x}_1 - \bar{x}_2$

$\boxed{statistic \pm t^* \cdot SE}$

If the population is approximately normal or the sample sizes are large ($n_1 \geq 30, n_2 \geq 30$), then a confidence interval for μ_1 - μ_2 can be computed by

$$\bar{x}_1 - \bar{x}_2 \pm t^* \cdot \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}$$

- Degrees of freedom for the *t*-distribution is the smaller of n_1 – 1 and n_2 – 1

Statistics: Unlocking the Power of Data Lock5

Example 1: Does diet cola wash calcium out of our systems? A study to investigate this question randomly assigned 16 healthy women to drink 24 ounces of either diet cola or water. Their urine was collected for three hours after ingestion and calcium excretion was measured. For the 8 diet cola drinkers, mean amount of calcium excreted was 56.0 mg with a standard deviation of 4.93. For the 8 water drinkers, the mean was 49.1 mg with a standard deviation of 3.64. Neither distribution had any significant outliers or skewness. (Why is this important?) Find and interpret a 95% confidence interval for the difference in mean amount of calcium excreted between diet cola drinkers and water drinkers. Does diet cola appear to have an effect?

The sample sizes are quite small but we are told that there are no significant outliers or skewness so the conditions are met to use the t-distribution. We use 7 degrees of freedom for the t-distribution. The confidence interval for $\mu_c - \mu_w$, where μ_c is the mean amount of calcium excreted by women who drink diet cola and μ_w is the mean amount of calcium excreted by women who drink water, is given by:

$$\text{Statistic} \pm t^* \cdot SE$$

$$(\bar{x}_c - \bar{x}_w) \pm t^* \cdot \sqrt{\frac{s_c^2}{n_c} + \frac{s_w^2}{n_w}}$$

$$(56.0 - 49.1) \pm 2.364 \cdot \sqrt{\frac{4.93^2}{8} + \frac{3.64^2}{8}}$$

$$6.9 \pm 5.122$$

$$1.778 \quad \text{to } 12.022$$

We are 95% sure that the mean amount of calcium excreted by diet cola drinkers is between 1.778 mg and 12.022 mg higher than the mean amount of calcium excreted by water drinkers. All values in this interval are positive, so we are 95% sure that the mean for diet cola drinkers is higher than the mean for water drinkers. Yes, diet cola does appear to have an effect on calcium excretion.

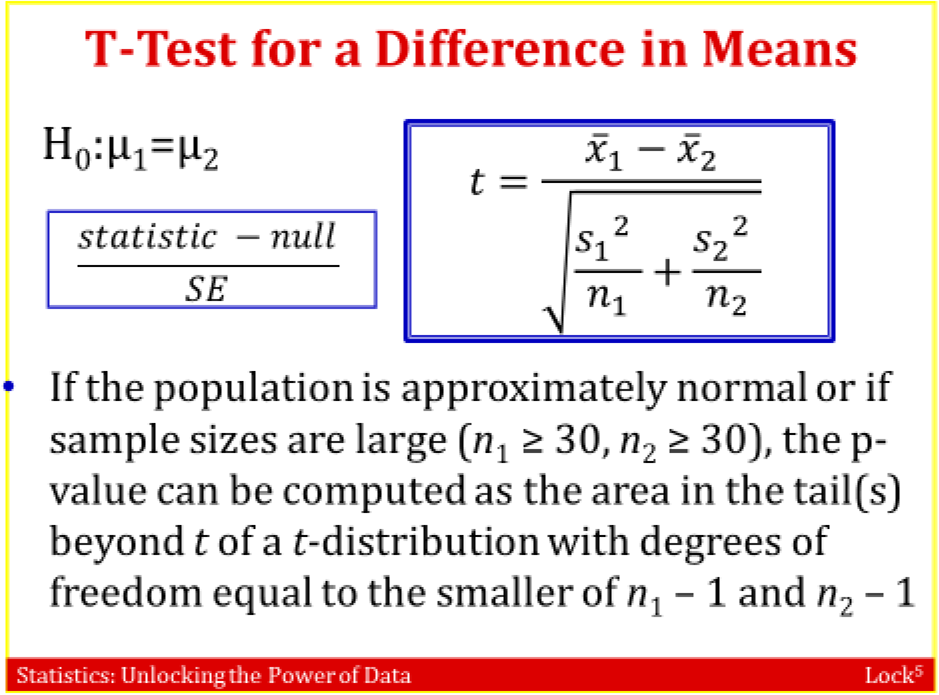
Section 6.4-HT: Hypothesis Test for a Difference in Means

In Chapter 5 we saw the generic formula for a test statistic:

$$\text{Test Statistic} = \frac{\text{Statistic} - \text{Value under H}_0}{\text{SE}}$$

We can change the generic formula to describe a test statistic for a difference of means:

$$\text{Test Statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\text{SE}}$$



T-Test for a Difference in Means

$H_0:\mu_1=\mu_2$

$$\boxed{\frac{statistic\ - null}{SE}}$$

$$\boxed{t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}}$$

- If the population is approximately normal or if sample sizes are large ($n_1 \geq 30$, $n_2 \geq 30$), the p-value can be computed as the area in the tail(s) beyond t of a t-distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$

Statistics: Unlocking the Power of Data Lock^5

Example 1: Cognition Score and Alcohol

A recent study asked college students to indicate the level of alcohol use, and in this example, we compare the group of students who said they were light drinkers to the group of students who said they were heavy drinkers. Each student was also given several cognitive skills tests and assigned a cognition z-score based on the performance on these tests. The 83 students who said they were light drinkers had a mean cognition z-score of 0.1302 with a standard deviation of 0.75, while the 16 students who said they were heavy drinkers had a mean cognition z-score of –0.2338 with a standard deviation of 0.65. Test, using a 5% significance level, to see if there is evidence that heavy drinkers have a lower mean cognitive level than light drinkers. If the results are significant, can we conclude from this study that heavy drinking affects cognitive ability?

We are testing H_0: $\mu_L = \mu_H$ vs H_a: $\mu_L > \mu_H$, where μ_L represents the mean cognition level for college students who say they are light drinkers and μ_H represents the mean cognition level for college students who say they are heavy drinkers.

The test statistic is $t = \frac{Statistic-Null}{SE} = \frac{(\bar{x}_L-\bar{x}_H)-0}{\sqrt{\frac{s_L^2}{n_L}+\frac{s_H^2}{n_H}}} = \frac{0.1302-(-0.2338)}{\sqrt{\frac{0.75^2}{83}+\frac{0.65^2}{16}}} = 1.998.$

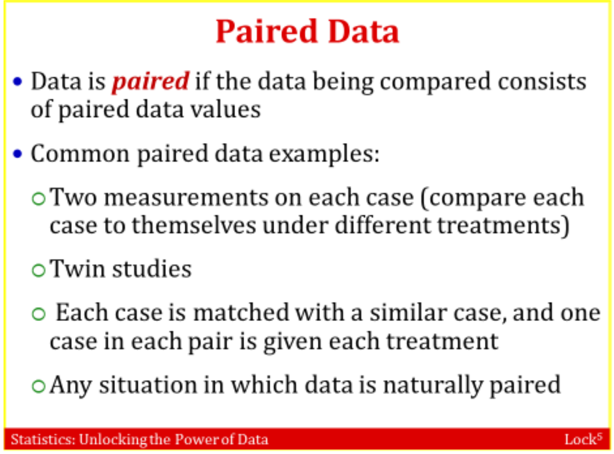
This is a right-tail test, and we use a t-distribution with 15 degrees of freedom. We see that the p-value is 0.032.

At a 5% significance level, we reject H_0 and conclude that students who say they are heavy drinkers have lower cognition scores than students who say they are light drinkers. The results are significant, but we cannot conclude that heavy drinking affects cognitive ability since the data come from an observational study not an experiment. There are many possible confounding variables. (See if the students can name some!)

Quick Self-Quiz: Restaurant Tips and Credit Cards

We analyze the percent tip left on 157 bills from the First Crush bistro in Northern New York State. The mean percent tip left on the 106 bills paid in cash was 16.39 with a standard deviation of 5.05. The mean percent tip left on the 51 bills paid with a credit card was 17.10 with a standard deviation of 2.47. Test to see if there is a difference in the mean percent of the bill left as a tip between bills paid with a credit card and bills paid with cash.

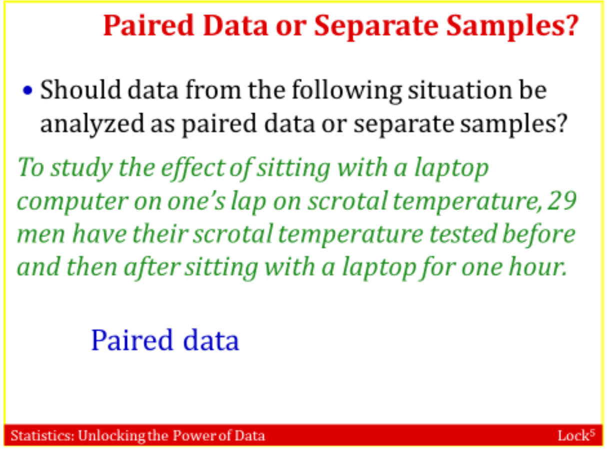
Section 6.5: Paired Difference in Means



Paired Data

- Data is ***paired*** if the data being compared consists of paired data values
- Common paired data examples:
  - Two measurements on each case (compare each case to themselves under different treatments)
  - Twin studies
  - Each case is matched with a similar case, and one case in each pair is given each treatment
  - Any situation in which data is naturally paired

Statistics: Unlocking the Power of Data Lock^5



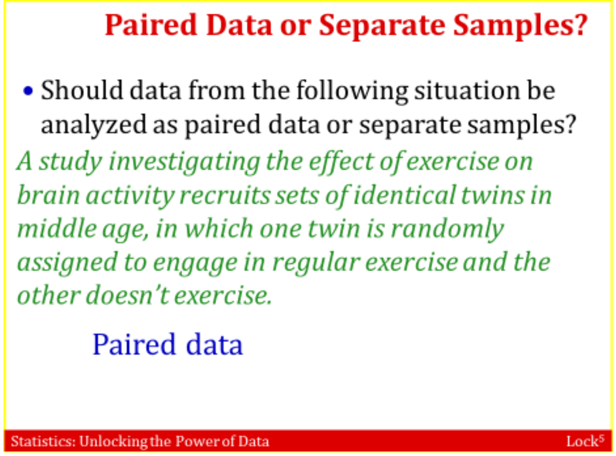
Paired Data or Separate Samples?

- Should data from the following situation be analyzed as paired data or separate samples?

To study the effect of sitting with a laptop computer on one's lap on scrotal temperature, 29 men have their scrotal temperature tested before and then after sitting with a laptop for one hour.

Paired data

Statistics: Unlocking the Power of Data Lock^5



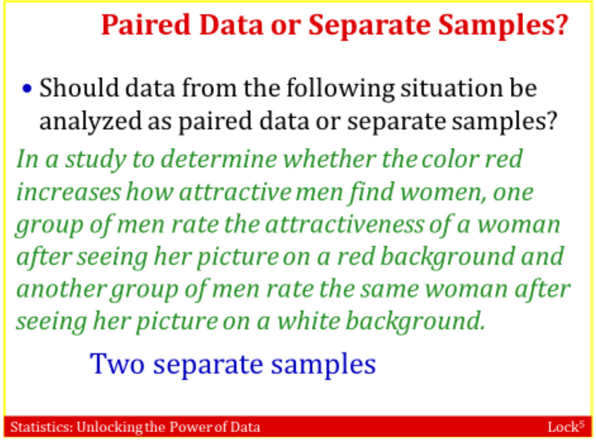
Paired Data or Separate Samples?

- Should data from the following situation be analyzed as paired data or separate samples?

A study investigating the effect of exercise on brain activity recruits sets of identical twins in middle age, in which one twin is randomly assigned to engage in regular exercise and the other doesn't exercise.

Paired data

Statistics: Unlocking the Power of Data Lock^5



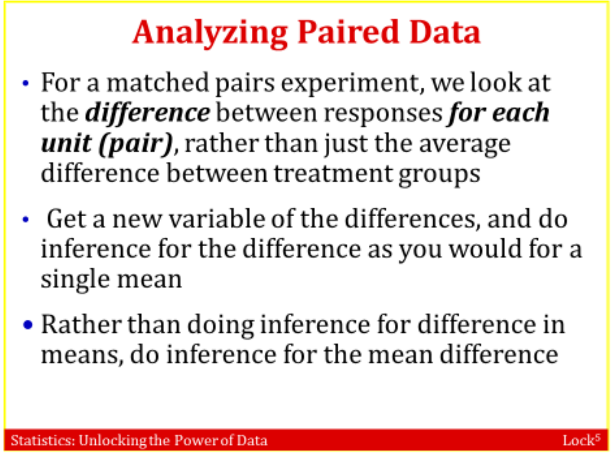
Paired Data or Separate Samples?

- Should data from the following situation be analyzed as paired data or separate samples?

In a study to determine whether the color red increases how attractive men find women, one group of men rate the attractiveness of a woman after seeing her picture on a red background and another group of men rate the same woman after seeing her picture on a white background.

Two separate samples

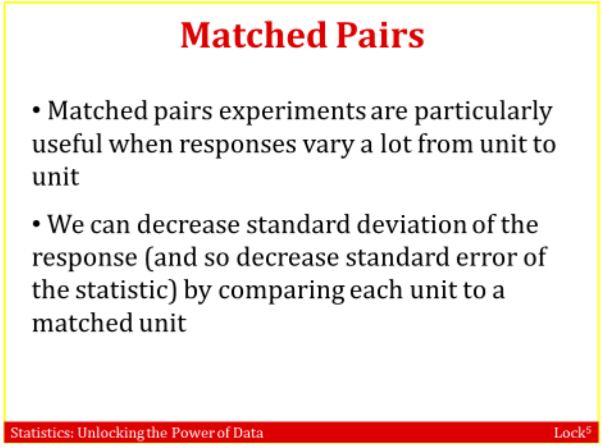
Statistics: Unlocking the Power of Data Lock^5



Analyzing Paired Data

- For a matched pairs experiment, we look at the ***difference*** between responses ***for each unit (pair)***, rather than just the average difference between treatment groups
- Get a new variable of the differences, and do inference for the difference as you would for a single mean
- Rather than doing inference for difference in means, do inference for the mean difference

Statistics: Unlocking the Power of Data Lock[5]



Matched Pairs

- Matched pairs experiments are particularly useful when responses vary a lot from unit to unit
- We can decrease standard deviation of the response (and so decrease standard error of the statistic) by comparing each unit to a matched unit

Statistics: Unlocking the Power of Data Lock[5]

Inference for a Difference in Means with Paired Data

To estimate the difference in means based on paired data, we first subtract to compute the difference for each data pair and compute the mean $\bar{x}_d$, the standard deviation $\boldsymbol{s_d}$, and the sample size $\boldsymbol{n_d}$ for the sample differences.

Provided the differences are reasonably normally distributed (or the sample size is large), a confidence interval for the difference in means is given by

$$Statistic \pm t^{*} \cdot SE = \bar{X}_d \pm t^{*}\frac{s_d}{\sqrt{n_d}} \qquad SE \approx \frac{s_d}{\sqrt{n_d}}$$

where $\boldsymbol{t^*}$ is a percentile from a t-distribution with $\boldsymbol{n_d} - 1$ degrees of freedom.

To test $\boldsymbol{H}_0 : \boldsymbol{\mu_d} = 0$ vs $\boldsymbol{H_a} : \boldsymbol{\mu_d} \neq 0$ (or a one-tail alternative) we use the t-test statistic

$$t = \frac{statistic - Null\ value}{SE} = \frac{\bar{X}_d - 0}{(s_d / \sqrt{n_d})}$$

If the differences are reasonably normally distributed (or the sample size is large), we use a t-distribution with $\boldsymbol{n_d} - 1$ degrees of freedom to compute a p-value for this test statistic.

Example: CAOS Comparisons

The CAOS (Comprehensive Assessment of Outcomes in Statistics) exam is an online multiple-choice test on concepts covered in a typical introductory statistics course. Students take one version before the start of the course and another version after the course ends. Before and After scores for a possible random sample of 10 students are shown in the table. (An actual random sample of scores are given in Exercise C.68 on page 455 of the text.)

Student	A	B	C	D	E	F	G	H	I	J
Before	43	40	48	65	60	48	43	38	43	55
After	60	45	55	80	85	71	52	35	54	55
Difference	***17***	***5***	***7***	***15***	***25***	***23***	***9***	***-3***	***11***	***0***

a). We are interested in determining whether taking the course increases students' understanding, as measured by this test. Why should we do a paired design rather than two separate groups?

There seems to be a great deal of variation in the scores between the students both before and after the course. We are interested in the increase in score for individuals, and pairing the data reduces the random variation.

b). Find the differences (*After – Before*) for all 10 students. State the mean, standard deviation, and sample size for the differences here:

See the differences above as the new fourth row of the table. Using these 10 values, we calculate that the mean is ___10.9__, the standard deviation is _9.219__, and the sample size is _10__.

c). Test to see if scores at the end of the course are higher, on average, than scores at the beginning of the course. Show all details of the test.

We are testing H_0: $\mu_d = 0$ vs H_a: $\mu_d > 0$ where μ_d represents the mean increase on a student's score after taking an introductory statistics course.

The test statistic is $t = \frac{Statistic - Null}{SE} = \frac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{n_d}}} = \frac{10.9 - 0}{\frac{9.219}{\sqrt{10}}} = 3.74$

This is a right-tail test, and we use a t-distribution with df = 9 to find the p-value. We see that the p-value is 0.0023.

We reject H_0 and conclude that there is strong evidence that, on average, scores increased after taking an introductory statistics course.

d). what is the average increase on the exam after taking the course? Compute and interpret a 95% confidence interval for the improvement in mean CAOS scores between the *before* and *After* scores.

$$\text{Statistic} \pm t^* \cdot \text{SE}$$
$$\bar{x}_d \pm t^* \cdot \frac{s_d}{\sqrt{n_d}}$$
$$10.9 \pm 2.262 \cdot \frac{9.219}{\sqrt{10}}$$
$$10.9 \pm 6.59$$

4.31 to 17.49

We are 95% sure that the mean increase for students on the CAOS exam after taking an introductory statistics class is between 4.31 points and 17.49 points.