Parametric inference on means

Overview

Inference on means

Review and continuation of a single mean

- Distribution, confidence intervals, and hypothesis tests

The difference between two means

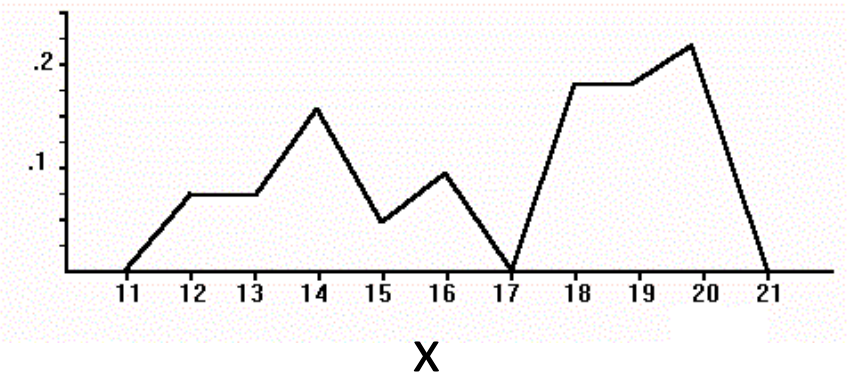
- Distribution, confidence intervals, and hypothesis tests

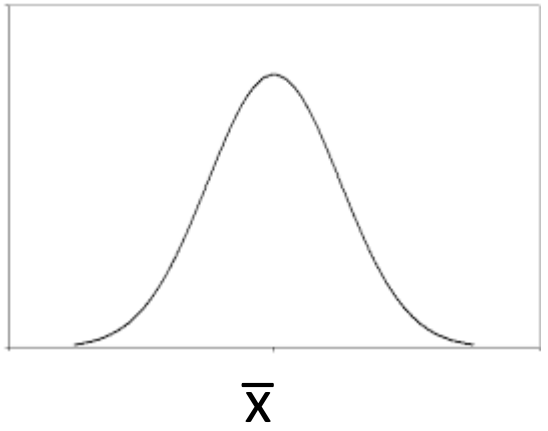
The difference between two means when the data is paired

Review: parametric inference on a single mean

Central Limit Theorem for Sample means

The sampling distribution of sample means ($\bar{x}$) *from* ***any population distribution*** will be normal, provided that the sample size is large enough



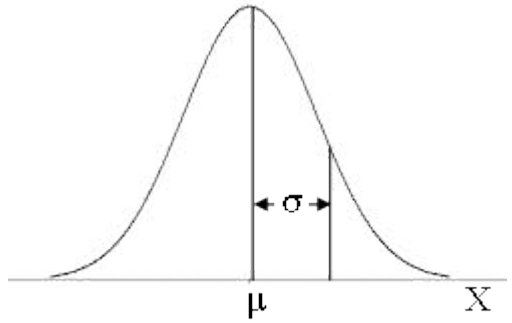


The more skewed the distribution, the larger sample size we will need for the normal approximate to be good

Sample sizes of 30 are usually sufficient. If the original population is normal we can get away with smaller sample sizes

Central Limit Theorem for Sample means

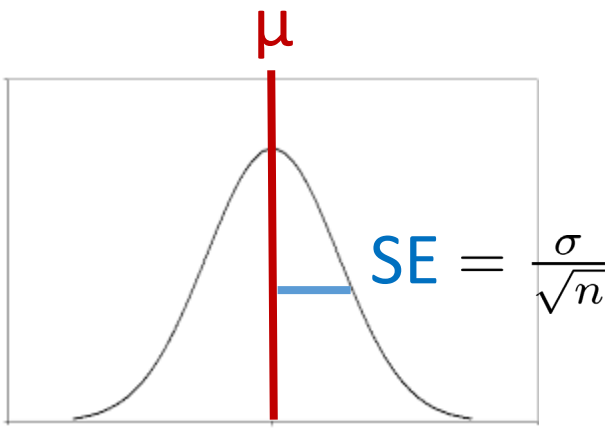
All normal distributions density models have two parameters N(**μ, σ**)



For modeling the ***sampling distribution*** of the sample means (**$\bar{x}$**):

- The center of the N(**μ, σ**) density model (μ) is the population mean **μ**
- The spread of the N(**μ, σ**) density model (σ) is the SE which is given by the formula: $SE = \frac{\sigma}{\sqrt{n}}$

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Ok, everything is cool so far, but…

Why is it usually impossible to use the following formula to compute the standard error?

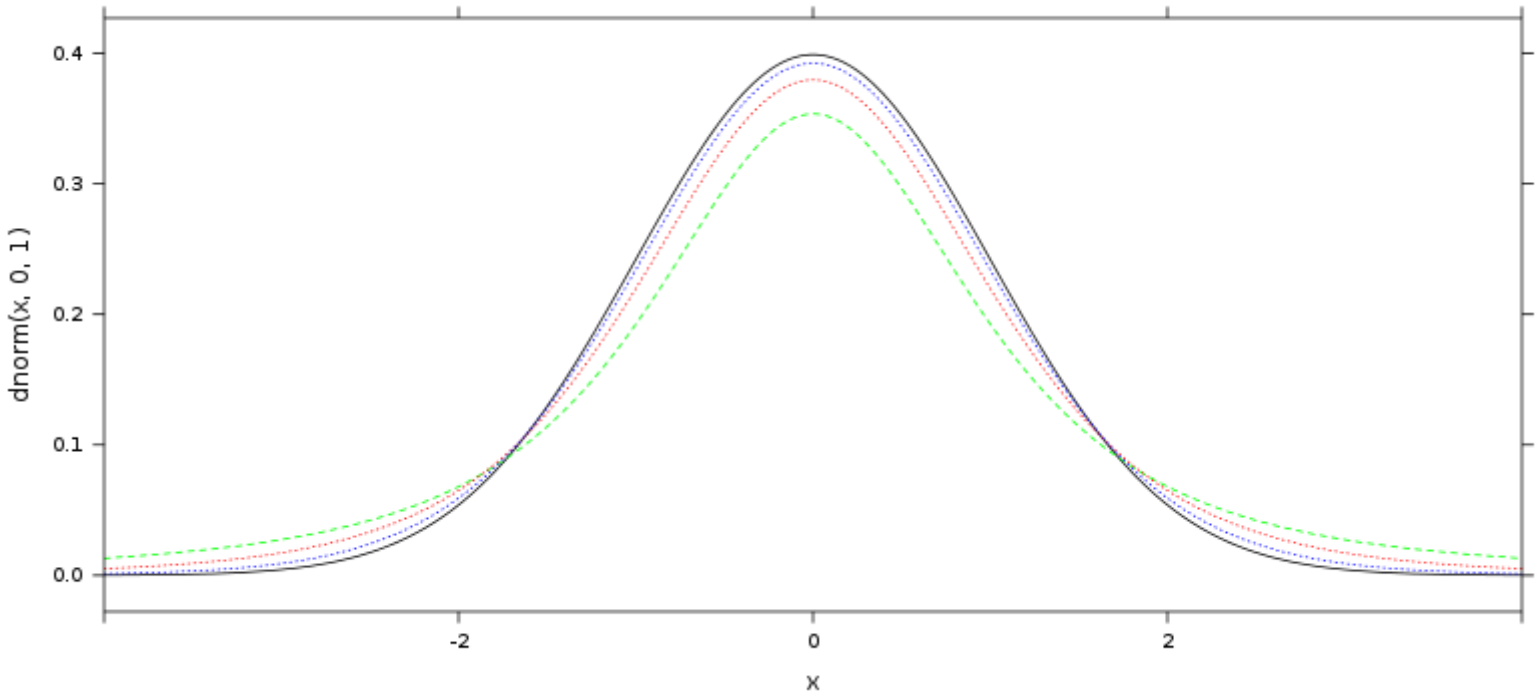
$$SE = \frac{\sigma}{\sqrt{n}}$$

Only Plato knows σ

If we substitute ***s*** for ***σ*** the sampling distribution is not exactly normal

- i.e., substituting $SE = \frac{s}{\sqrt{n}}$ for $SE = \frac{\sigma}{\sqrt{n}}$ leads to a t-distribution!

t-distributions



N(0, 1), df = 2, df = 5, df = 15

Summary: The Distribution of Sample Means ($\bar{x}$) using the Sample Standard Deviation

When choosing random samples of size n from a population with mean μ, the distribution of the sample means has the following characteristics

Center: The mean is equal to the population mean μ

Spread: The standard error is estimated using $SE = \frac{s}{\sqrt{n}}$

Shape: The standardized sample means approximately follows a t-distribution with n-1 degrees for freedom (df)

For small sample sizes ($n \leq 30$), the t-distribution is only a good approximation if the underlying population has a distribution that is approximately normal

The Distribution of Sample Means Using the Sample Standard Deviation

$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

The fine print - this works if:

The underlying population has a distribution that is approximately normal or n > 30)

Confidence Interval for a single mean

A confidence interval for a population mean μ can be computed based on a random sample of size n using:

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t* is an endpoint chosen from a t-distribution with n-1 df to give the desired confidence level

- i.e., use the qt(prob, df) to get t*)

The t-distribution is appropriate if the distribution of the population is approximately normal or the sample size is large (n >= 30)

Are we mostly bacteria?

A study by Qin etl al (2010) found that the average number of unique genes in gut bacteria, from a sample of 99 healthy European individuals was 564 million, with a standard deviation of 122 million

Use the t-distribution to find a 95% confidence interval for the mean number of unique genes in gut bacteria for European individuals

nature > articles > article

nature

Article | Open Access | Published: 04 March 2010

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin, Ruiqiang Li, [...] Jun Wang

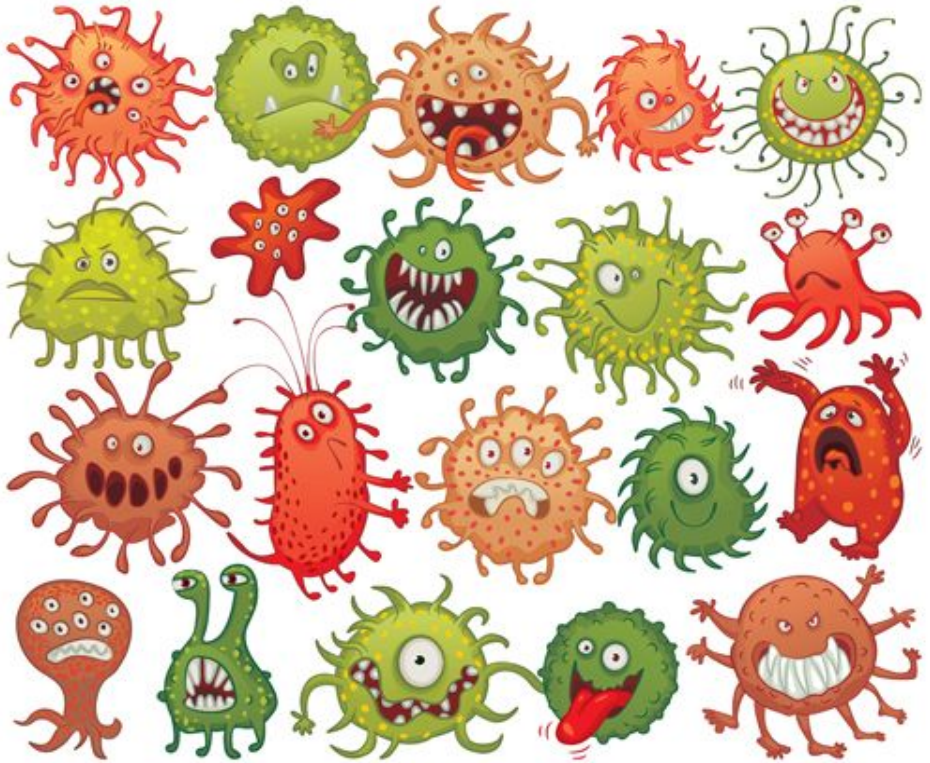
Nature **464**, 59–65(2010) | Cite this article

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$

R: qt(area, df)

Are we mostly bacteria?

$$\overline{x} \pm t^* \frac{s}{\sqrt{n}}$$



$\bar{x}$ = 564,000,000

s = 122,000,000

n = 99

t* = qt(.975, df = 98) = 1.98

564,000,000 ± 1.98 · 122,000,000/9.95 = [539,667,529 588,332,471]

Parametric hypothesis test for a single mean μ

When the distribution of a statistic under H_0 is **normal**, we compute a standardized test statistic using:

$$z = \frac{Sample\ \ Statistic\ -\ Null\ \ Parameter}{SE}$$

When testing hypotheses for a single mean we have:

- H_0: $\mu = \mu_0$ (where μ_0 is specific value of the mean)

Thus the null parameter is μ_0 and the sample statistics is $\bar{x}$ so we have:

$$z = \frac{\overline{x} - \mu_0}{SE}$$

Parametric test for a single mean μ

We can estimate the standard error by $SE = \frac{s}{\sqrt{n}}$

however this makes the statistic follow a t-distribution with n-1 degrees of freedom rather than a normal distribution

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

This works if n is large or the data is reasonably normally distributed.

Because we are use a t-distribution to find the p-value, this is called a t-test

t-Test for Single Mean

To test:

H_0: $\mu = \mu_0$ vs.

H_A: $\mu \neq \mu_0$ (or a one-tailed alternative)

We use the t-statistic: $t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$

A p-value can be computed using a t-distribution with n-1 degrees of freedom

- Provided that the population is reasonable normal (or the sample size is large)

Home prices in New Jersey

The average US house sells for about $265,000

A sample of n = 30 houses in New Jersey showed had an average price of $\bar{x}$ = $388,500, with a standard deviation of s = $224,700

Is the average price of a house in New Jersey significantly greater than the US average?

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

pt(t, df = deg_of_free)

Home prices in New Jersey

H_0: $\mu = 265{,}000$ vs H_A: $\mu > 265{,}000$

$\bar{x} = 388{,}500$

$s = 224{,}700$

$n = 30$

$df = 29$

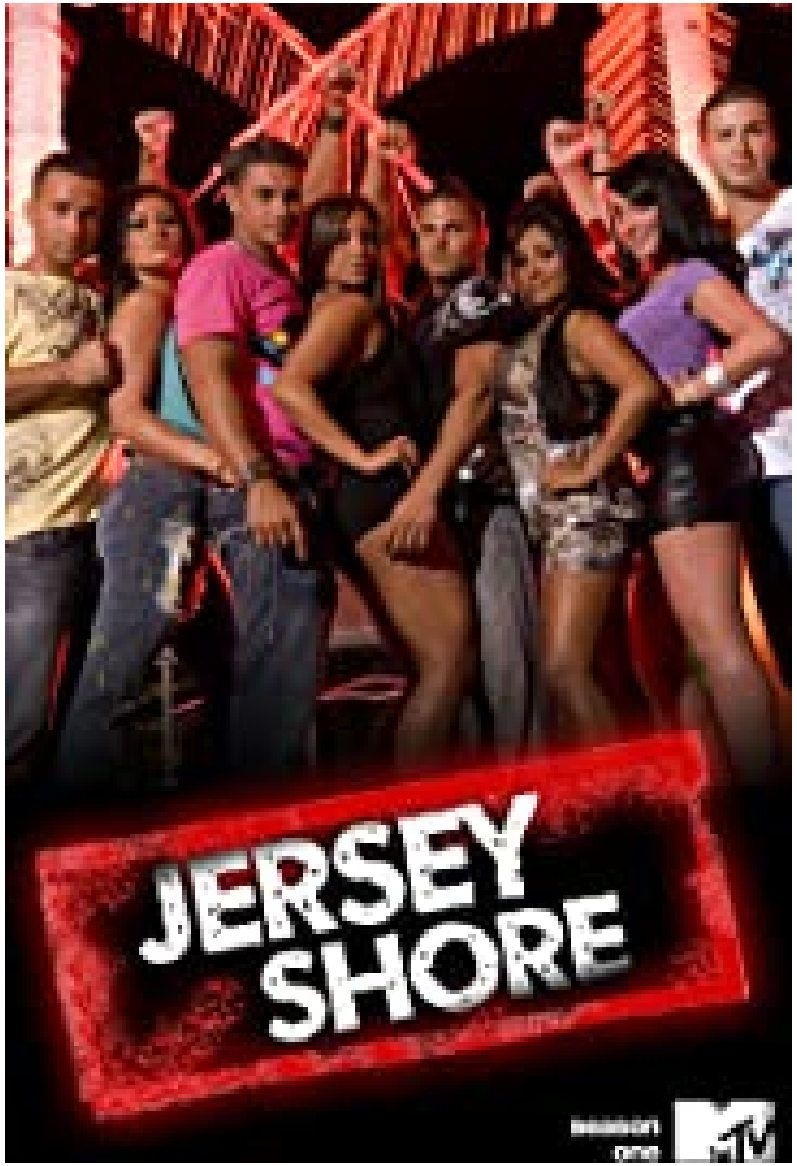
$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

SE = 224700 /sqrt(30) = 41,024

t = (388500 – 265000)/41024 = 3.01

P-value: pt(3.01, df = 29, lower.tail = FALSE)

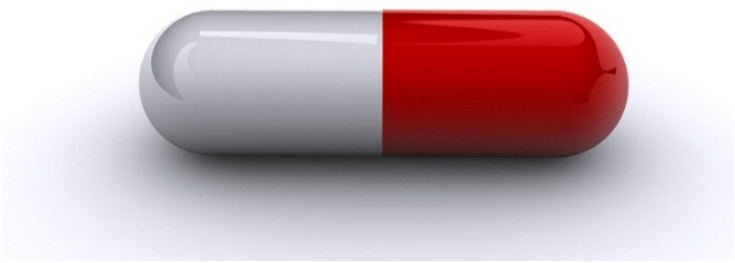
= 0.0027



Parametric inference for the difference between two means

Distribution of differences in means

What is an example of a *hypothesis test* for comparing the difference between two means?



The distribution of differences of means (and consequently inferences about differences in means) is similar to what we have seen for proportions and a single mean

Central Limit Theorem for Differences in Two Sample Means

Suppose we have two populations where

- Population 1 has: mean μ_1 and standard deviation σ_1
- Population 2 has: mean μ_2 and standard deviation σ_2

Suppose we also have samples from these populations of size n_1 and n_2

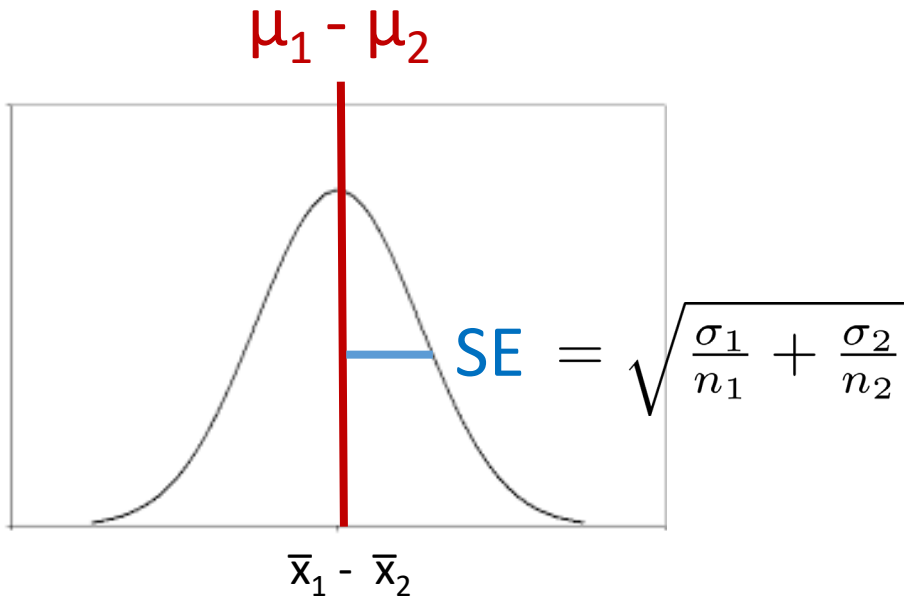
The distribution of the differences in two samples means $\bar{x}_1 - \bar{x}_2$ is:

- Approximately normal if both sample sizes are large (≥ 30)
- Has a center at $\mu_1 - \mu_2$
- Has standard deviation given by:

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

Distribution of differences in means

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



The standard error of differences of means

Similar to the standard error for means from a single sample we do not know σ.

We can substitute s for σ

Our sample statistic (difference of means) comes from a t-distribution
(provided n is large or the data is not too skewed)

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We will use the minimum of $n_1 - 1$, or $n_2 - 1$ as a conservative estimate of the df

Summary: the distribution of differences in sample means

When choosing random samples of size n_1 and n_2 from populations with means μ_1 and μ_2, the distribution of the differences in two samples means, $\bar{x}_1$ and $\bar{x}_2$ has the following characteristics:

Center: The mean is equal to the difference in populations means $\mu_1 - \mu_2$

Spread: The standard error is: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Shape: The standardized differences in sample means follow a t-distribution with degrees of freedom approximately equal to the smaller of $n_1 - 1$ and $n_2 - 1$

For small sample sizes ($n_1 < 30$, or $n_2 < 30$), the t-distribution is only a good approximation if the underlying population has a distribution that is approximately normal

Parametric confidence intervals for the difference between two means

Confidence interval for a difference in two means

If we have large samples (or samples that are reasonably normally distributed) of sizes n_1 and n_2 from two different groups, we can construct a confidence interval for $\mu_1 - \mu_2$, the difference in means between those two groups, using:

$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where, $\overline{x}_1$ and $\overline{x}_2$ are the means and s_1 and s_2 are the standard deviations

The t* value is a quantile from a t-distribution to give the desired confidence level

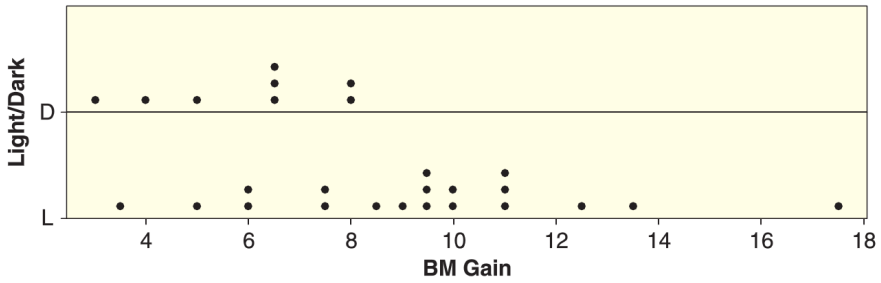
Use the smaller of $n_1 - 1$ and $n_2 - 1$ to give the degrees of freedom

More on mice eating late at night gaining weight

Another study examining how much weight was gained by mice eating late at night (as determined by keeping a light on at night) had the following characteristics:

27 mice were randomly divided into 2 groups:

- The 8 mice in darkness gained an average of 5.9g with a standard deviation of 1.9g
- The 19 mice with light at night gained an average of 9.4 grams with a standard deviation of 3.2g



$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Find and interpret the 80% confidence interval for the difference in weight gained

More on mice eating late at night gaining weight

$\bar{x}_L = 9.4$ $\bar{x}_D = 5.9$ $\bar{x}_L - \bar{x}_D = 3.5$

$s_L = 3.2$ $s_D = 1.9$

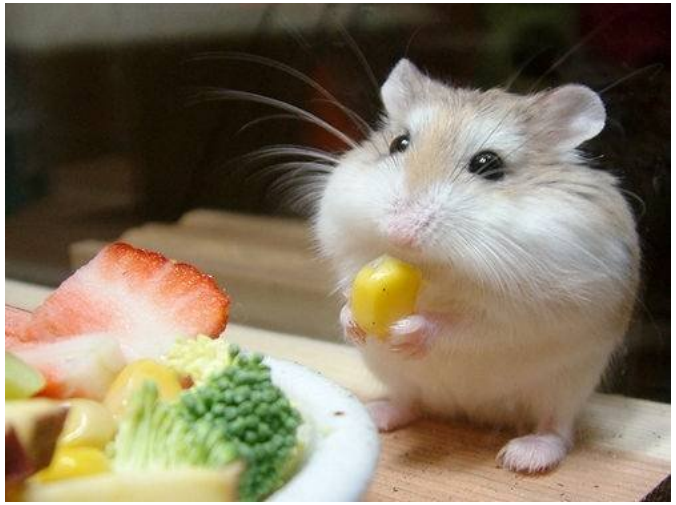
$n_L = 19$ $n_D = 8$

SE = sqrt(((3.2)^2)/19 + ((1.9)^2)/8) = .995

df = 7

t* = qt(.90, df=7) = 1.415

3.5 ± 1.415 * .995 = (2.09 4.91) more gained in light condition



$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Parametric hypothesis tests for the difference between two means

Test for difference in means

As we've seen several times now, we can create a z-score for hypothesis tests using:

$$z = \frac{Sample\ \ Statistic\ -\ Null\ \ Parameter}{SE}$$

The sample statistic here is: $\bar{x}_1 - \bar{x}_2$

For the difference of means, the SE is: $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Using this SE means that we have to use a t-distribution rather than a standard normal distribution

Two-sample t-Test for a Difference in Means

To test H_0: $\mu_1 = \mu_2$ vs. H_A: $\mu_1 \neq \mu_2$ (or a one-tailed alternative) based on sample sizes of n_1 and n_2 from the two groups, we use the two-sample t-statistics

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

Where $\overline{x}_1$ and $\overline{x}_2$ are the sample means and s_1 and s_2 are the standard deviations for the respective samples

We can use the t-distribution if the sample is large (>30) or if the population is reasonably normal. We can use the df as the smaller of $n_1 - 1$ or $n_2 - 1$, or technology to get a better approximation

Do right or left handed men make more money?

A study randomly sampled 2295 American men

- 2027 men were right-handed, 268 men were left-handed
- right-handers earned $13.10/hr, left-handers earned $13.40/hr
- The standard deviation for both groups was $7.90

Test the hypothesis that there is a difference in earnings between right and left handed men

- 1. State the null and alternative hypothesis
- 2-4. Find the t-statistic and p-value
- 5. Interpret the conclusions

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

Do right or left handed men make more money?

H_0: $\mu_R = \mu_L$ H_A: $\mu_R \neq \mu_L$

mean_right_handed <- 13.10
mean_left_handed <- 13.40

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

n_right_handed <- 2027
n_left_handed <- 268

var_both <- (7.90)^2
SE <- sqrt((var_both/n_right_handed) + (var_both/n_left_handed)) = 0.51

t_value <- (mean_right_handed – mean_left_handed)/SE = -.584

P_value <- 2 * pt(t_value, df = n_left_handed -1) = .56
(we don't use 1 – pt(t_value, df = n_left_handed -1) because the t-value is negative here)

Parametric paired sample hypothesis tests for the difference between two means

Are grades significantly higher on a second quiz?

A sample of grades on the first two quizzes in an introductory statistics class are given in the table below for n = 10 students

Student	1	2	3	4	5	6	7	8	9	10
First Quiz	72	95	56	87	80	98	74	85	77	62
Second Quiz	78	96	72	89	80	95	86	87	82	75

Did students scored higher on average on the second quiz?

Run a hypothesis test to see if there is a statistically significant different

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

Are grades significantly higher on a second quiz?

H_0: $\mu_{quiz1} = \mu_{quiz2}$ vs. H_A: $\mu_{quiz2} > \mu_{quiz1}$

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

```
quiz1 <- c(72, 95, 56, 87, 80, 98, 74, 85, 77, 62)
quiz2 <- c(78, 96, 72, 89, 80, 95, 86, 87, 82, 72)
```

```
SE <- sqrt( var(quiz1)/10 + var(quiz2)/10)  = 5.01
```

```
t_stat <- (mean(quiz2) - mean(quiz1))/SE   = 1.02
```

```
p_val <- pt(t_stat, 9, lower.tail = FALSE)  =  .168
```

Are we convinced that there was not a statistically significant difference in the average quiz scores?

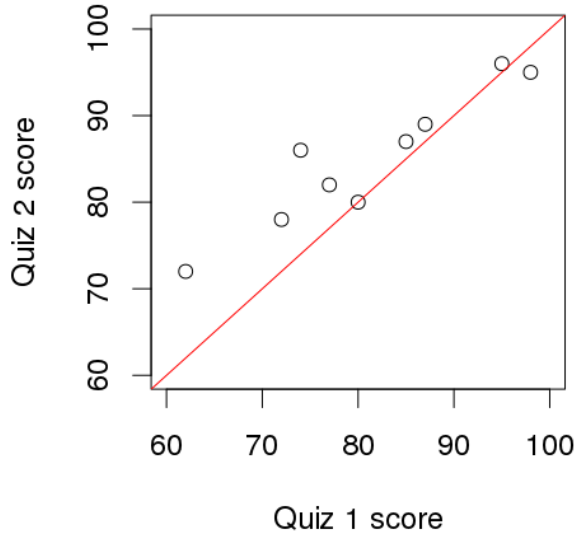
Are grades significantly higher on a second quiz?

A sample of grades on the first two quizzes in an introductory statistics class are given in the table below for n = 10 students

Student	1	2	3	4	5	6	7	8	9	10
First Quiz	72	95	56	87	80	98	74	85	77	62
Second Quiz	78	96	72	89	80	95	86	87	82	75

Notice that the scores between quiz 1 and quiz 2 are not independent since they come from the same students

Some students are just score higher overall



Are grades significantly higher on a second quiz?

A sample of grades on the first two quizzes in an introductory statistics class are given in the table below for n = 10 students

Student	1	2	3	4	5	6	7	8	9	10
First Quiz	72	95	56	87	80	98	74	85	77	62
Second Quiz	78	96	72	89	80	95	86	87	82	75

Notice that the scores between quiz 1 and quiz 2 are not independent since they come from the same students

Some students are just score higher overall

If we can take into account the fact that some students score better than others this, could reduce some of variability in the data and could lead to a more powerful test

- i.e. a test that is better able to reject the null hypothesis H_0 when it is false

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

We can then compute the mean $\bar{x}_d$ the standard deviation $\bar{s}_d$, and the sample size n_d for the sample difference to test…

H_0: $\mu_d = 0$

H_A: $\mu_d \neq 0$

we use the t-statistic: $$t = \frac{\overline{x}_d}{s_d / \sqrt{n_d}}$$

Are grades significantly higher on a second quiz?

A sample of grades on the first two quizzes in an introductory statistics class are given in the table below for n = 10 students

Student	1	2	3	4	5	6	7	8	9	10
First Quiz	72	95	56	87	80	98	74	85	77	62
Second Quiz	78	96	72	89	80	95	86	87	82	75

Let's convert this to the differences in scores between the quiz 2 and quiz 1

We can now run a one sample t-test for H_0: $\mu_d = 0$ vs. H_A: $\mu_d > 0$ which should be be better able to reject H_0

$$t = \frac{\overline{x}_d}{s_d / \sqrt{n_d}}$$

Are grades significantly higher on a second quiz?

quiz_diff <- quiz2 – quiz1

$$t = \frac{\overline{x}_d}{s_d / \sqrt{n_d}}$$

SE_diff <- sd(quiz_diff)/sqrt(10) = 1.88

t_stat <- mean(quiz_diff)/SE_diff = 2.71

p_val <- pt(t_stat, df = 9, lower.tail = FALSE) = .012


12
3
6
9