Parametric inference on means

Overview

Review and continuation of inference on a single mean

- Distribution, confidence intervals, and hypothesis tests

Inference on the difference between two means

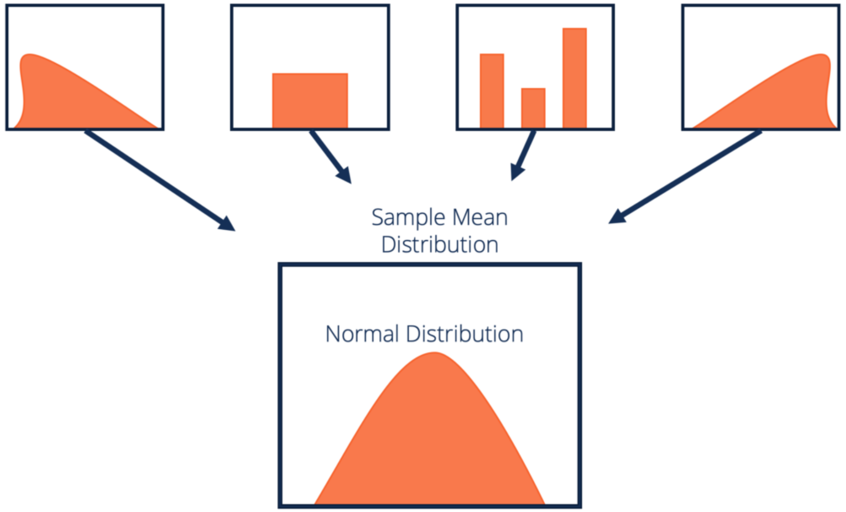
- Distribution, confidence intervals, and hypothesis tests

Inference on the difference of means for paired data

Review and continuation of parametric inference on a single mean

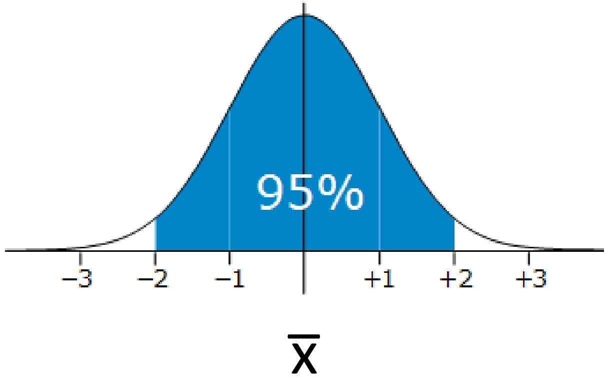
Review: Central limit theorem

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



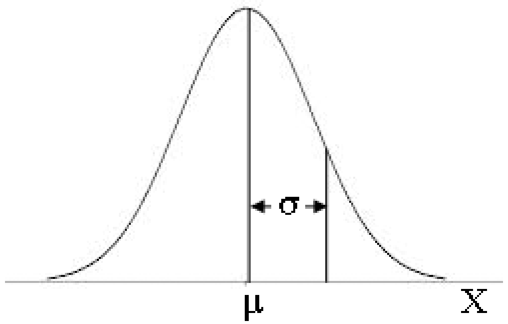
The more skewed the distribution, the larger sample size we will need for the normal approximate to be good

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

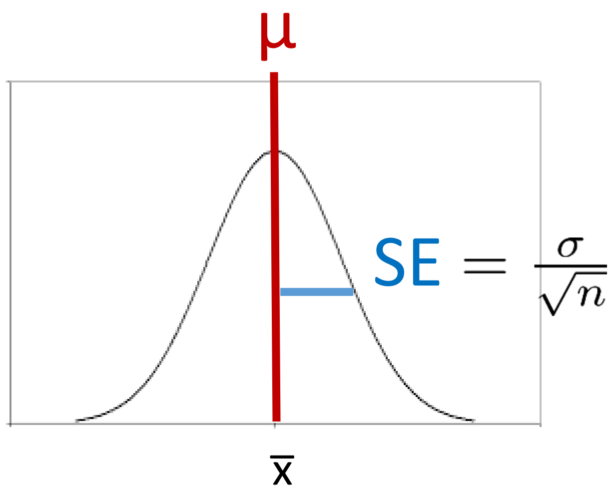
All normal distributions density models have two parameters N(**μ, σ**)



For modeling the ***sampling distribution*** of the sample means (**x̄**):

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

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



A formula for estimating the standard error

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

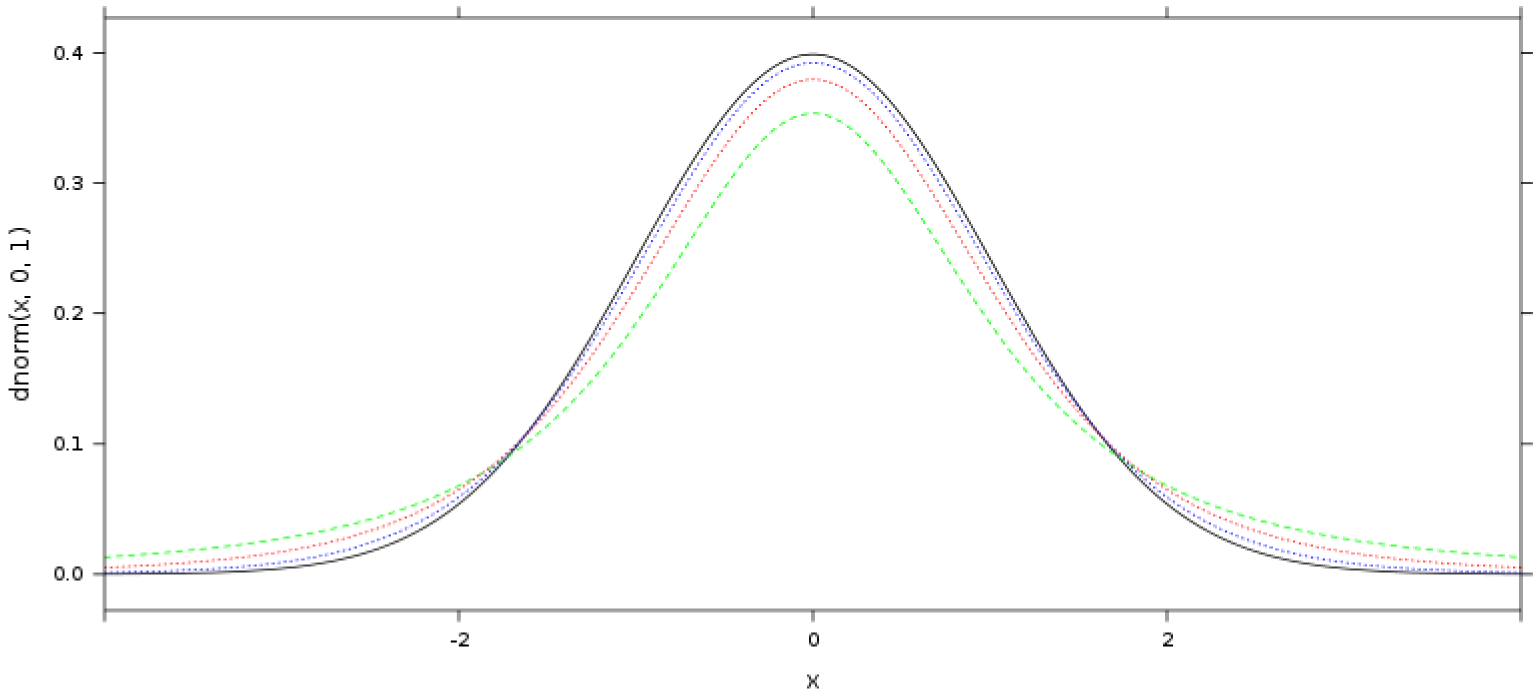
Only Plato knows σ

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

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

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)

Review: 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:

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

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) or mosaic::ct() 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)

How many grams of fiber do people get in a day?

A study by Nierenberg et al (1989) investigated the relationship between dietary factors, and plasma concentrations of carotenoids

Let’s use the data they collected to create a 98% confidence interval for the number of grams fiber US adults get in a day

```
nutrition_df <- read.csv("NutritionStudy.csv")
fiber <- nutrition_df$Fiber
```

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

Try it in R at home!

Parametric hypothesis test for a single mean μ

When the **null distribution** 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:

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

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 using a t-distribution to find the p-value, this is called a **t-test**

t-test for a 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: $$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

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)

The Chips Ahoy! Challenge

In the mid-1990s a Nabisco marking campaign claimed that there were at least 1000 chips in every bag of Chips Ahoy! cookies

A group of Air Force cadets tested this claim by dissolving the cookies from 42 bags in water and counting the number of chips

They found the average number of chips per bag was 1261.6, with a standard deviation of 117.6 chips

Test whether the average (mean) number of chips per bag is greater than 1000. Do the results confirm Nabisco's claim?

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

pt(t, df = deg_of_free)

Let's try it in R!

The Chips Ahoy! Challenge

H_0: $\mu = 1000$ vs H_A: $\mu > 1000$

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

$\bar{x} = 1261.6$

s = 117.6

n = 42

df = 41

SE = 117.6/sqrt(42)

t = (1261.6 – 1000)/18.141 = 14.42

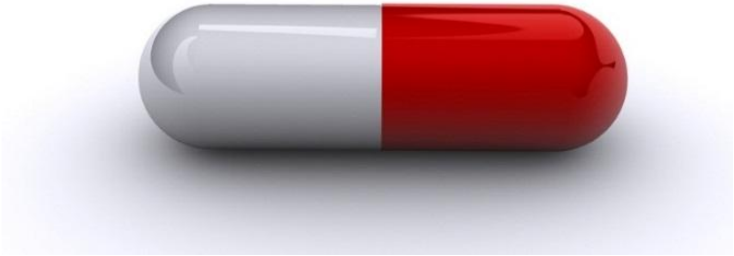
P-value: pt(14.32, df = 41) < 10^-16

Does this verify chips ahoy!’s claim?

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

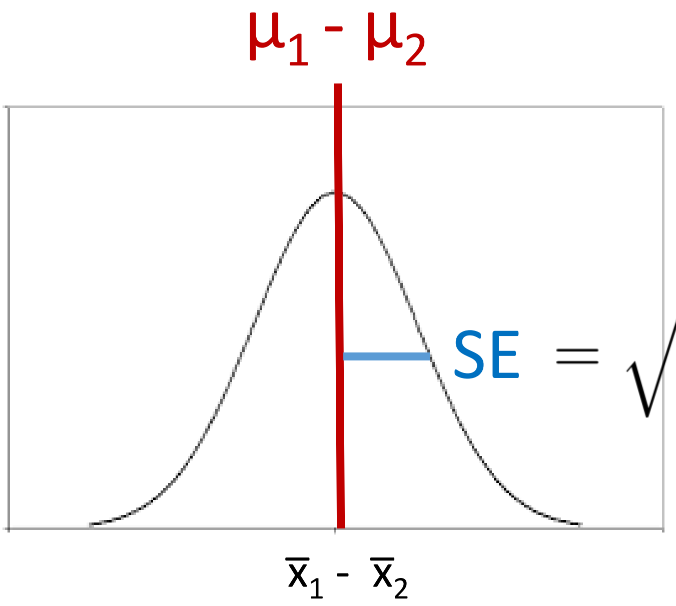
The distribution of the differences in two samples means $\bar{x}_1 - \bar{x}_2$ is:

- 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

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



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

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

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

Our sample statistic (difference of means) comes from a t-distribution

(provided n is large or the data is not too skewed)

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

Parametric confidence intervals for the difference between two means

Confidence interval for a difference in two means

Suppose we have large (or reasonably normally distributed) samples of sizes n_1 and n_2 from two different groups

We can construct a confidence interval for μ_1 - μ_2, the difference in means between those two groups, using:

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

We use the smaller of n_1 -1 and n_2 - 1 to give the degrees of freedom

Who eats more fiber, males or females?

Let's use the Nierenberg et al (1989) data to find a 95% confidence interval for the differences in the number of grams of fiber eaten in a day between males and females

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

Let's try it in R!

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 formula for the SE means we need to use a t-distribution for the null distribution

Two-sample t-test for a difference in means

Suppose we would like to test:

- H_0: $\mu_1 = \mu_2$
- H_A: $\mu_1 \neq \mu_2$ (or a one-tailed alternative)

Suppose we also have sample sizes of n_1 and n_2 from the two groups

We use this “two independent sample t-test” using our test statistic:

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

We can use the df as the smaller of $n_1 - 1$ or $n_2 - 1$, or technology to get a better approximation

(this works provided n is large or the data is not too skewed)

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: $\mu_R \neq \mu_L$

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

```
mean_right_handed <- 13.10
mean_left_handed <- 13.40

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
```

Parametric paired sample hypothesis tests for the difference between two means

Are grades significantly higher on a second exam?

A sample of grades on two exams 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 exam	72	95	56	87	80	98	74	85	77	62
Second exam	78	96	72	89	80	95	86	87	82	75

Did students score higher on average on the second exam?

We could run a hypothesis test to see if there is a statistically significant different

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

Are grades significantly higher on a second exam?

H_0: $\mu_{exam1} = \mu_{exam2}$ vs. H_A: $\mu_{exam2} > \mu_{exam1}$

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

```
exam1 <- c(72, 95, 56, 87, 80, 98, 74, 85, 77, 62)
exam2 <- c(78, 96, 72, 89, 80, 95, 86, 87, 82, 72)
```

```
SE <- sqrt( var(exam1)/10 + var(exam2)/10)
```
= 5.01

```
t_stat <- (mean(exam2) - mean(exam1))/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 exam?

A sample of grades on the first two exams 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 Exam	72	95	56	87	80	98	74	85	77	62
Second Exam	78	96	72	89	80	95	86	87	82	75

Notice that the scores between exam 1 and exam 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, 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 compute the difference for each data pair

We can then compute the mean $\overline{x}_d$ the standard deviation $\overline{s}_d$, 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: $$t = \frac{\overline{x}_d}{s_d / \sqrt{n_d}}$$

Let’s try it in R!

Are grades significantly higher on a second quiz?

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

exam_diff <- exam2 - exam1

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

t_stat <- mean(exam_diff)/SE_diff = 2.71

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