

Parametric inference on means

Overview

Final project presentations by Natalie and Rhys

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

Announcements: self-evaluations and class evaluations

Please fill out self-evaluations and class evaluations on the hub

Self-evaluations: please honestly assess

- What went well
- What you struggled with
- What are your future plans for using what you learned

Class evaluations

- Overall how was the class
- What worked well and what would be good to know to make the class better in the future
 - E.g., which worksheets/classes were helpful, which could have been better

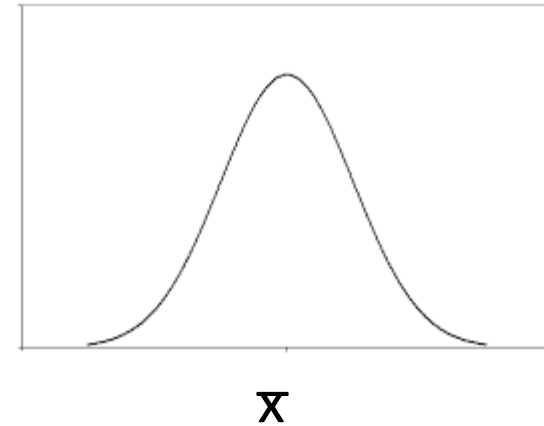
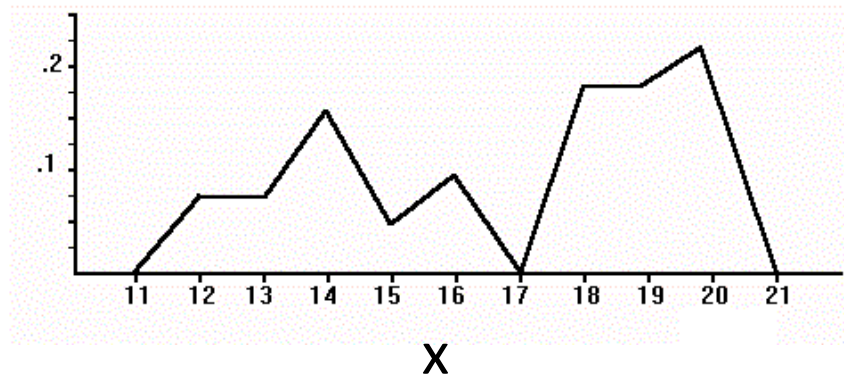
Final project presentations

Natalie and Rhys

All other presentations are due this Sunday December 9th

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

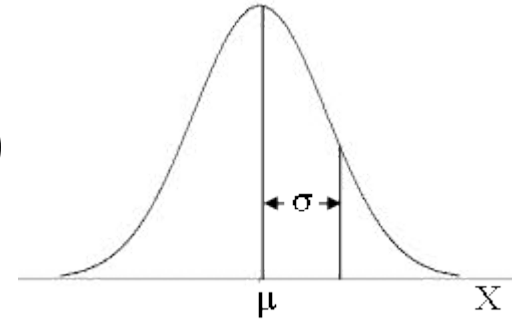


The more skewed the distribution, the larger sample size we will need for the normal approximate to be good.

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

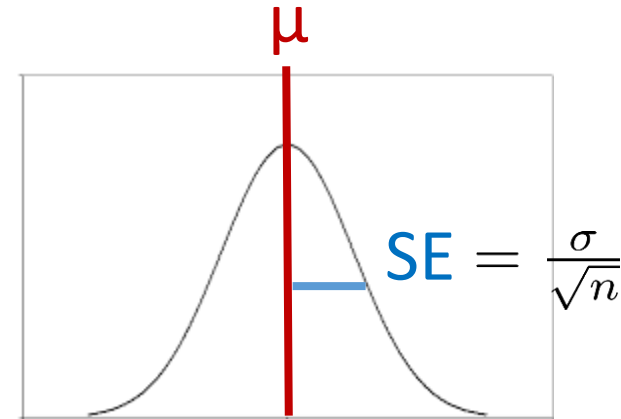
All normal distributions density models have two parameters $N(\mu, \sigma)$



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

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

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



Ok, everything is cool so far, but...

Why (in reality) is it not possible to use the following formula to compute the standard error?

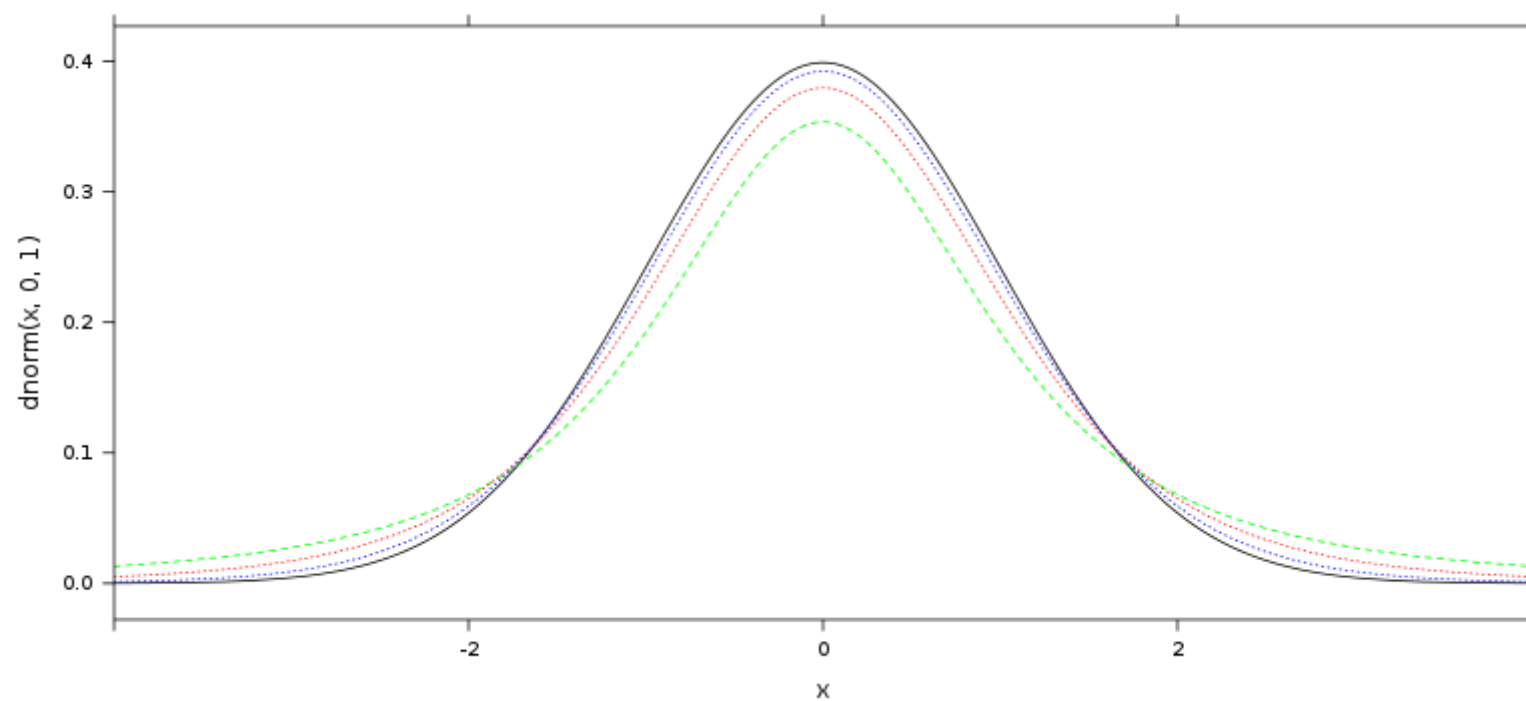
$$SE = \frac{\sigma}{\sqrt{n}}$$

← 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



$N(0, 1),$

$df = 2,$

$df = 5,$

$df = 15$

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:

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

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

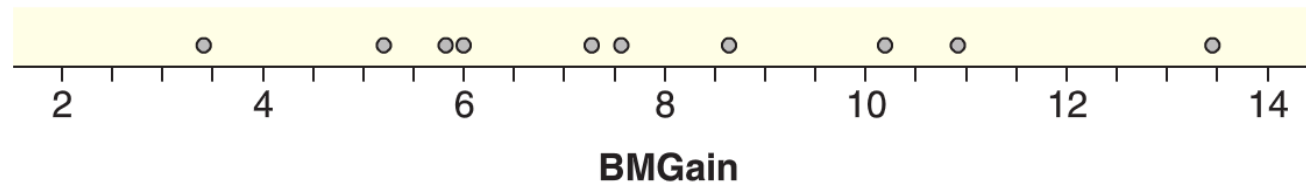
- i.e., use the `qt(prob, df)` to get t^*)

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

Review: Light at night makes mice fat

A study kept a light on at night which allowed mice to eat at night when they typically are resting. These mice gained a significant amount of weight compared to control mice kept in darkness which ate the same amount of calories.

The 10 mice with light gained an average of 7.9g with a standard deviation of 3.0g.



Find the 90% CI for the amount of weight gained

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

R: qt(area, df)

Light at night makes mice fat

What is the parameter we are estimating?

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

$$\bar{x} = 7.9,$$

$$s = 3,$$

$$n = 10$$

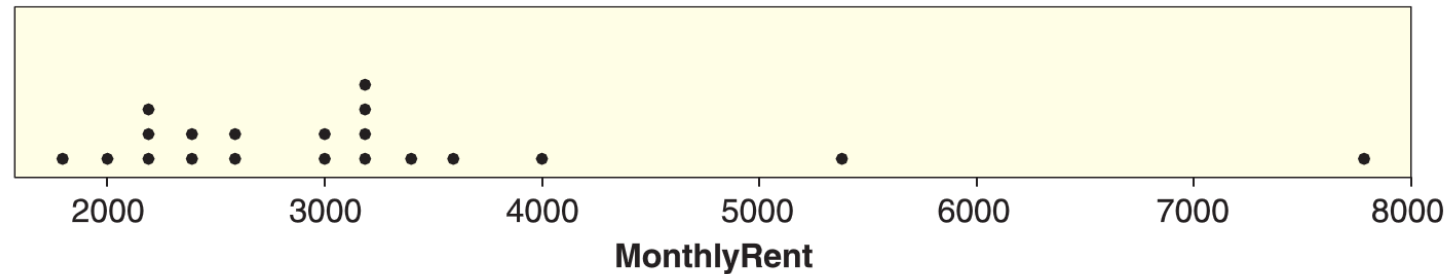
$$t^* = qt(.95, df = 9) = 1.833$$

$$7.9 \pm 1.833 \cdot 3/\sqrt{10} = (6.16 \text{ } 9.64)$$



NYC 1 bedroom apartment prices

The rental price for 20 one bedroom apartments in NYC were collected from Craig's list and plotted below



How can we get a 95% confidence interval for the mean price of a one bedroom apartment?

We can't use the t-distribution here because of the small sample size and outliers!!!

We need to use the bootstrap to estimate the SE and then use the stat + 2 * SE

- Ok to use 2 here because are dealing with a normal distribution when using the bootstrap

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:

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

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

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

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

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 marketing 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 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)



The Chips Ahoy! Challenge

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

$\bar{x} = 1261.6$

$s = 117.6$

$n = 42$

$df = 41$

$SE = 117.6/\sqrt{42}$

$t = (1261.6 - 1000)/18.141 = 14.42$

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

Does this verify chips ahoy!'s claim?

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



John Tukey quote



“Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question”

- *The future of data analysis*. Annals of Mathematical Statistics 33 (1), (1962), page 13.

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

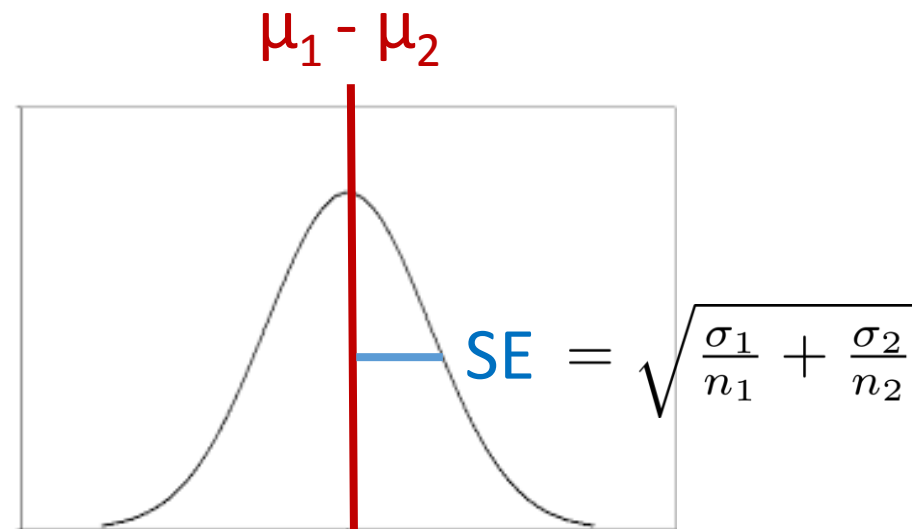
The distribution of the differences in two samples means $\bar{x}_1 - \bar{x}_2$ is:

- 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

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:

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

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

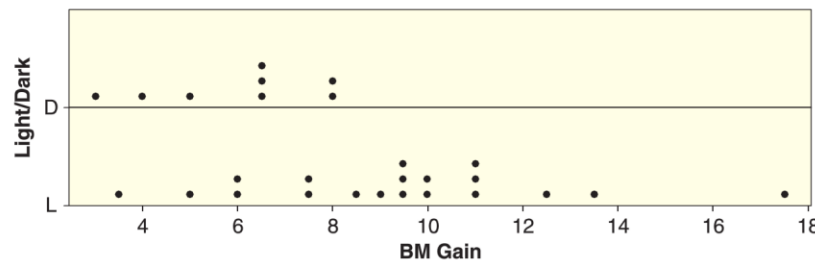
The t^* value is a percentile from a t-distribution to give the desired confidence level. Use the smaller of $n_1 - 1$ and $n_2 - 1$ (or R) to give the degrees of freedom

More on mice eating late at night and getting fat

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.



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

Find and interpret the 99% confidence interval for the difference in weight gained.

More on mice eating late at night and getting fat

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

$$s_L = 3.2 \quad s_D = 1.9$$

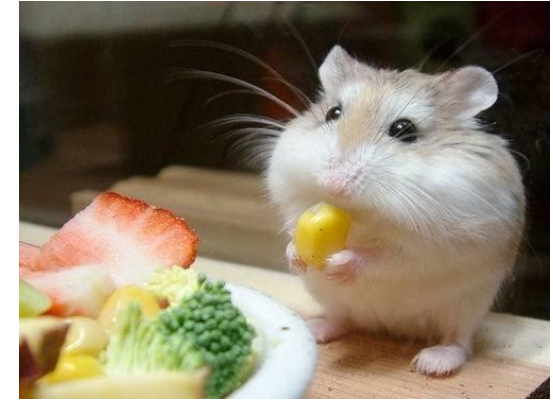
$$n_L = 19 \quad n_D = 8$$

$$SE = \text{sqrt}(((3.2)^2)/19 + ((1.9)^2)/8) = .995$$

$$df = 7$$

$$t^* = \text{qt}(.995, df=7) = 3.5$$

$$3.5 \pm 3.5 * .995 = (0.0175 \quad 6.9825) \text{ more gained in light}$$



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

Test for difference in means

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

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

What is the null hypothesis (and null parameter) here?

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 $H_0: \mu_1 = \mu_2 = 0$, 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

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

Where $\bar{x}_1$ and $\bar{x}_2$ 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.

Final projects

Presentations due on midnight on Sunday Dec 9th

Please use the final project presentation template...

Worksheet 12

Optional: Due on Wednesday Dec 12th

```
> source('/home/shared/intro_stats_2016/cs206_functions.R')  
> get_worksheet(12)
```