Parametric inference on a single mean

Overview

Example of a final project presentation

Quick review of parametric inference for proportions

Inference on means overview

Parametric confidence intervals for a single mean

Parametric hypothesis tests for a single mean

Final project

Final project: analyze your own data set

Final project report: a 5-8 page R Markdown document that contains:

1. Background information:
    - What question you will answer and why it is interesting
    - Where you got the data, and any prior analyses
2. Descriptive plots
3. A hypothesis tests using resampling and parametric methods
4. A confidence interval using the bootstrap and parametric methods
5. A conclusion and reflection
6. Optional: an appendix with extra code (appendix can go over the 8 page limit)

A list of a few data sets you can use are on Canvas

There is an R Studio Cloud workspace that has a template for the final project

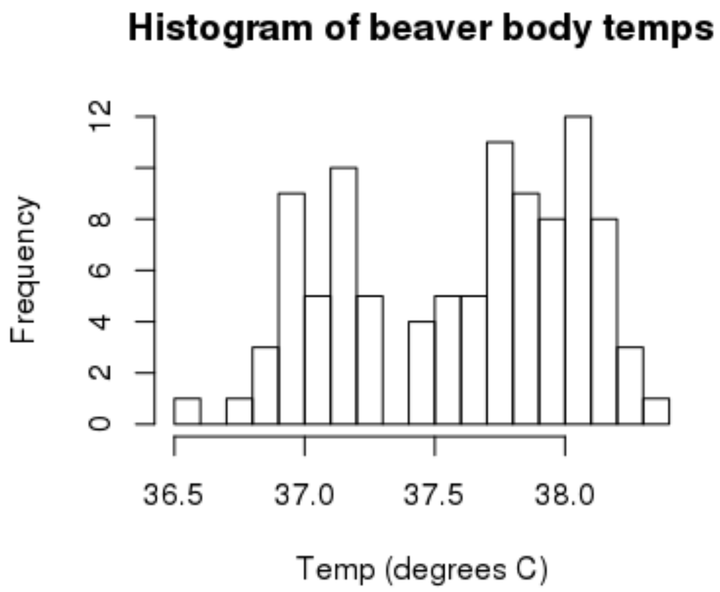
Question: do beavers have the same body temperature as humans?

Motivation and data

Motivation: There is a labor shortage in the construction industry

- Beavers are a hard working species of animals
- If beavers have the same body temperature as humans (37℃), perhaps they can be employed in the construction industry

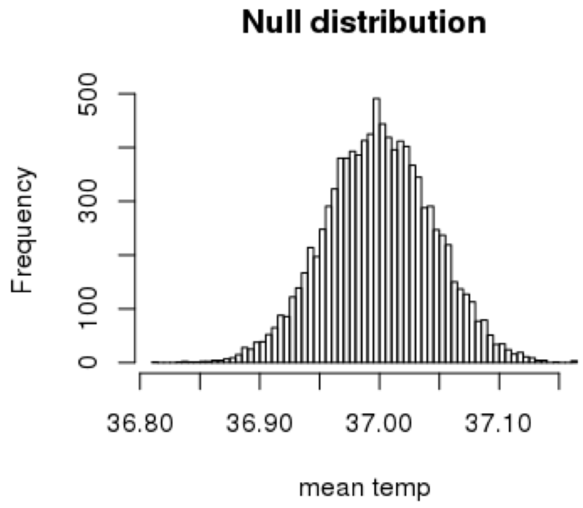
The data:



- Body temperatures collected from 400 beavers*
- Data from:
  - Lange et al (1994). In time-series analyses of beaver body temperatures. https://vincentarelbundock.github.io/Rdatasets/doc/boot/beaver.html

*not the real data

Results

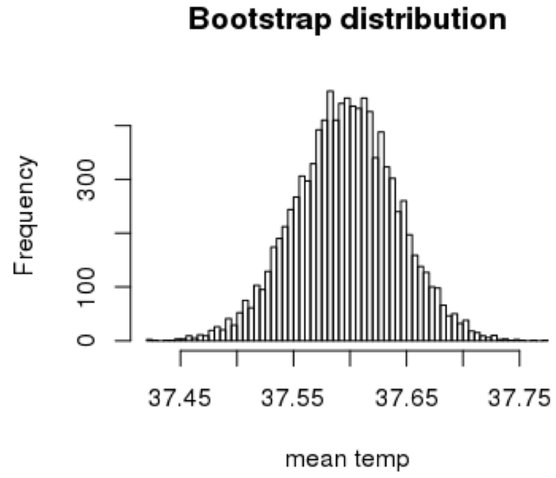


The average human body temperatures is $\mu = 37°C$

Hypothesis test

- H_0: $\mu = 37$ H_A: $\mu \neq 37$
- p-value based on a permutation test: $\bar{x} = 37.6$, p-value = 0
- p-value based on a t-test: $t = 13.35$, $df = 99$, p-value = 0

95% confidence interval for the mean beaver body temp



- Bootstrap: [37.51 37.68]
- t-distribution: [37.51 37.68]

Conclusions

Conclusion: Beavers do not seem to have the same body temperatures as humans

37℃ humans vs. 37.6℃ beavers

Implications: Due to their higher body temperatures, if beavers join the construction industry they might be too good at their jobs leading to job loss of human workers

Caveats: human body temperatures might not be exactly 37℃

Loading data for your project into R…

Review of inference on proportions

Central Limit Theorem for Sample Proportions

When taking samples of size n from a population with a proportion parameter π, the distribution of the sample proportions p̂ has the following characteristics:

Shape: If the sample size is sufficiently large, the distribution is reasonably normal

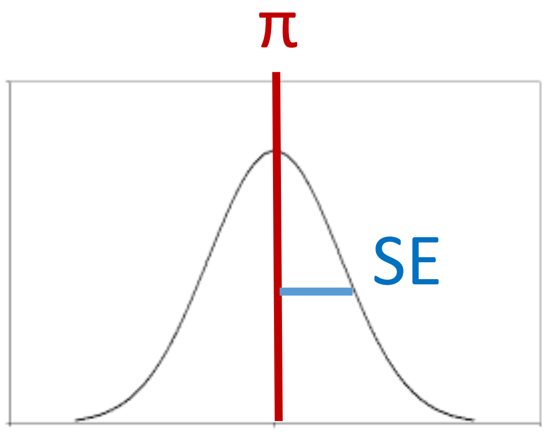
Center: The mean is equal to the population proportion π

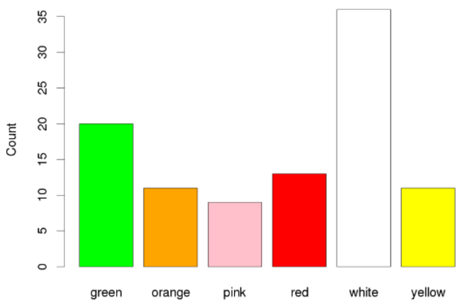
Spread: The standard error is: $SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

$$\hat{p} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$

A normal distribution is a good approximation as long as:

$n\pi \geq 10$ and $n(1-\pi) \geq 10$





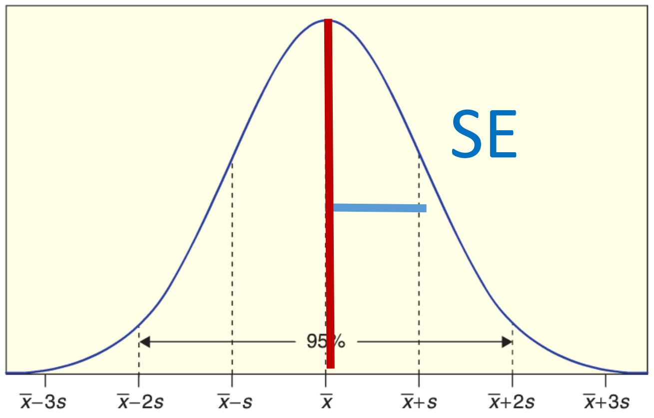
π_{red}

$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

π



$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$

$$\hat{p} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}})$$

Sampling distribution!

Confidence intervals for a single proportion

Confidence interval for a single proportion

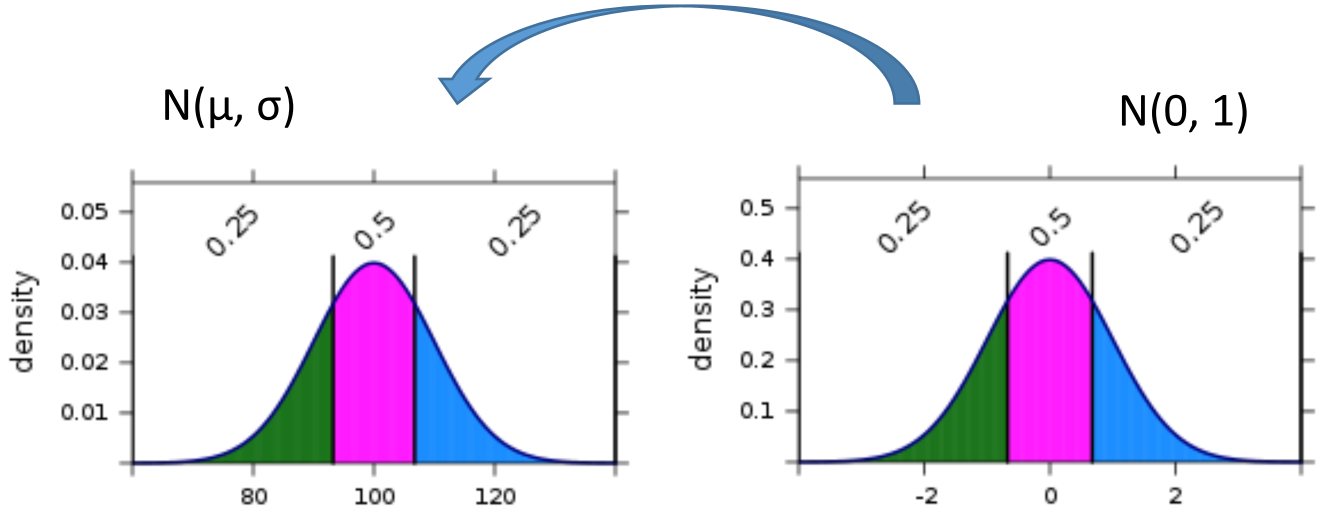
$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note we are substituting p̂ for π

Confidence intervals for using normal distributions

To convert from Z ~ N(0, 1) to any X ~ N(μ, σ), we use the transformation:

$$X = \mu + Z \cdot \sigma$$



For confidence intervals using normal distributions, we start with z* as the margin of error of error for N(0, 1)

We convert it to the scale our actual data is on using: $CI = stat \pm z^* \cdot SE$

One true love?

A survey asked 2625 people whether they agreed with the statement "There is only one true love for each person"

1812 of the respondents disagreed

Compute a 90% confidence interval for the proportion who disagreed

$$\hat{p} \ \pm \ z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

One true love?

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

n <- 2625

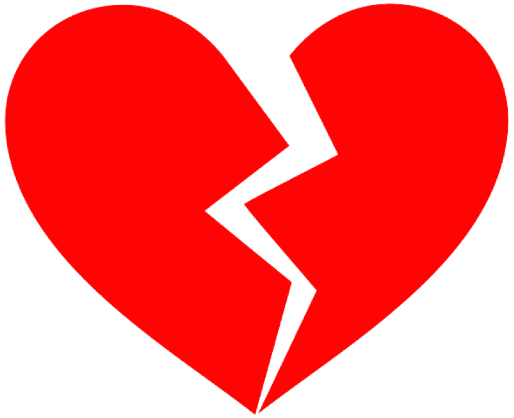
p_hat <- 1812/2625 = .69

SE <- sqrt((p_hat * (1 – p_hat))/n)

z_star <- qnorm(.95, 0, 1) = 1.64

ME <- z_star * SE   = .032

CI <- c(p_hat - ME, p_hat + ME) = [.658 .723]

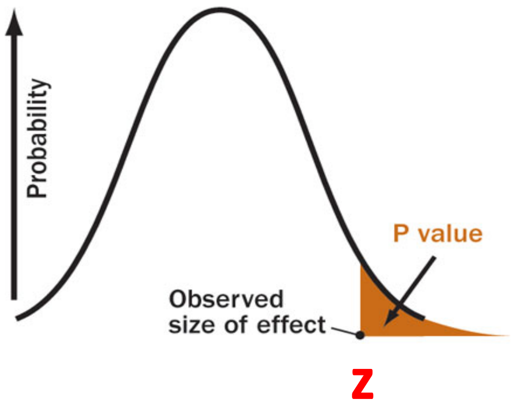


Hypothesis tests for a single proportion

Test statistic for a single proportion: H_0: $\pi = \pi_0$

$$z = \frac{stat_{obs} - param_0}{SE} \qquad\qquad z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Our z statistic comes from a standard normal distribution z ~ N(0, 1)



H_A: $\pi > \pi_0$

pnorm(z, 0, 1, lower.tail = FALSE)

Adult persistence of head-turning asymmetry

Background

- Most people are right handed, right eye dominant, etc.
- Biologists have suggested that human embryos tend to turn their heads to the right as well.

German bio-psychologist Onur Güntürkün conjectured that this tendency manifests itself in other ways, so he studies which ways people turn their heads when they kiss.

Adult persistence of head-turning asymmetry

He and his researchers observed kissing couples in public places and noted whether the couple leaned their heads to the right or left

They observed 124 couples, ages 13-70 years

Adult persistence of head-turning asymmetry

A neonatal right-side preference makes a surprising romantic reappearance later in life.

A preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first six months after birth[1,2] constitutes one of the earliest examples of behavioural asymmetry and is thought to influence the subsequent development of perceptual and motor preferences by increasing visual orientation to the right side[3,4]. Here I show that twice as many adults turn their heads to the right as to the left when kissing, indicating that this head-motor bias persists into adulthood. My finding may be linked to other forms of sidedness (for example, favouring the right foot, ear or

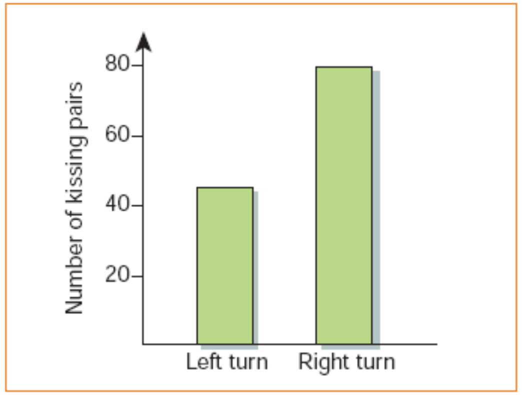


Figure 1 The number of couples who turn their heads to the right rather than to the left when kissing predominates by almost 2:1 (64.5%: 35.5%; $n = 124$ couples).

Of the 124 couples observed, 80 leaned their heads to the right while kissing

- Let's run a parametric hypothesis test for proportions on this data

Complete the following steps to analyze the data

1. State Null and Alternative in symbols and words

2. Calculate the observed z-statistic of interest
$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

3-4. Use the standard normal distribution to get a p-value

5. Make a decision about whether the results are statistically significant

Adult persistence of head-turning asymmetry

1. $H_0: \pi = 0.5$ $H_A: \pi > 0.5$ $.5 * 124 = 62 > 10$

2. $\hat{p} = 80/124 = .64$

   SE = sqrt((.5 * (1 - .5))/124) = 0.045

   z = (.64 - .50)/.045 = 3.12

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

3-4. `p_val <- pnorm(3.12, 0, 1, lower.tail = FALSE)` `# p-value = 0.0009`

5. Decision?

Parametric inference on means

Inference on means

1. From the central limit theorem, the distribution of sample means $\bar{x}$, has what shape?
   - A: Normal!

2. And what value (symbol) is the sampling distribution of $\bar{x}$ center at?
   - A: μ

3. What other piece of information would be need to plot the sampling distribution of $\bar{x}$?
   - A: SE

4. And how can we get the SE?
   - A: We could use the bootstrap, or…

Standard Error of Sample Means

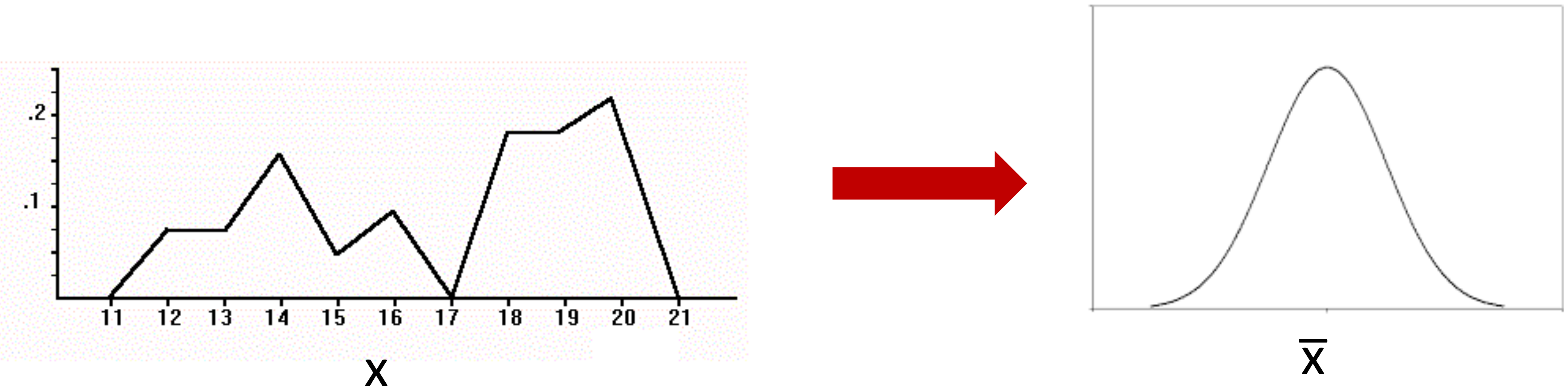
When choosing random samples of size n from a population with mean μ and standard deviation σ, the standard error of the sample means is:

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

The larger the sample size (n), the smaller the standard error

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

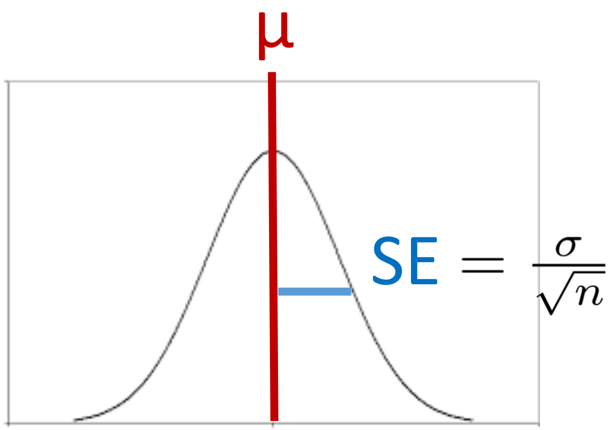
Sample sizes of 30 are usually sufficient. If the original population is normal we can get away with smaller sample sizes

Central Limit Theorem for Sample means

For random samples of size **n** from a population with mean **μ** and standard deviation **σ**…

the distribution of the sample means (**x̄**) is reasonably normal if the sample size is sufficiently large (n ≥ 30), with the mean **μ** and standard error $SE = \frac{\sigma}{\sqrt{n}}$

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



Ok, everything is cool so far, but…

For proportions, ***we used our sample estimate of p̂ for the population parameter π*** and to compute the standard error, and the sampling distribution was still normal so everything worked

$$\hat{p} \sim N(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}) \qquad \hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

When computing the sampling distribution for the sample mean , x̄, if we substitute s for σ it turns out that this sampling distribution is not exactly normal ☹

- i.e., if we substitute $SE = \frac{s}{\sqrt{n}}$ for $SE = \frac{\sigma}{\sqrt{n}}$ the distribution not normal ☹

The good news

Fortunately about 100 years ago William Sealy Gosset figured out that this sampling distribution where s is substituted for σ has another parametric form called a t-distribution

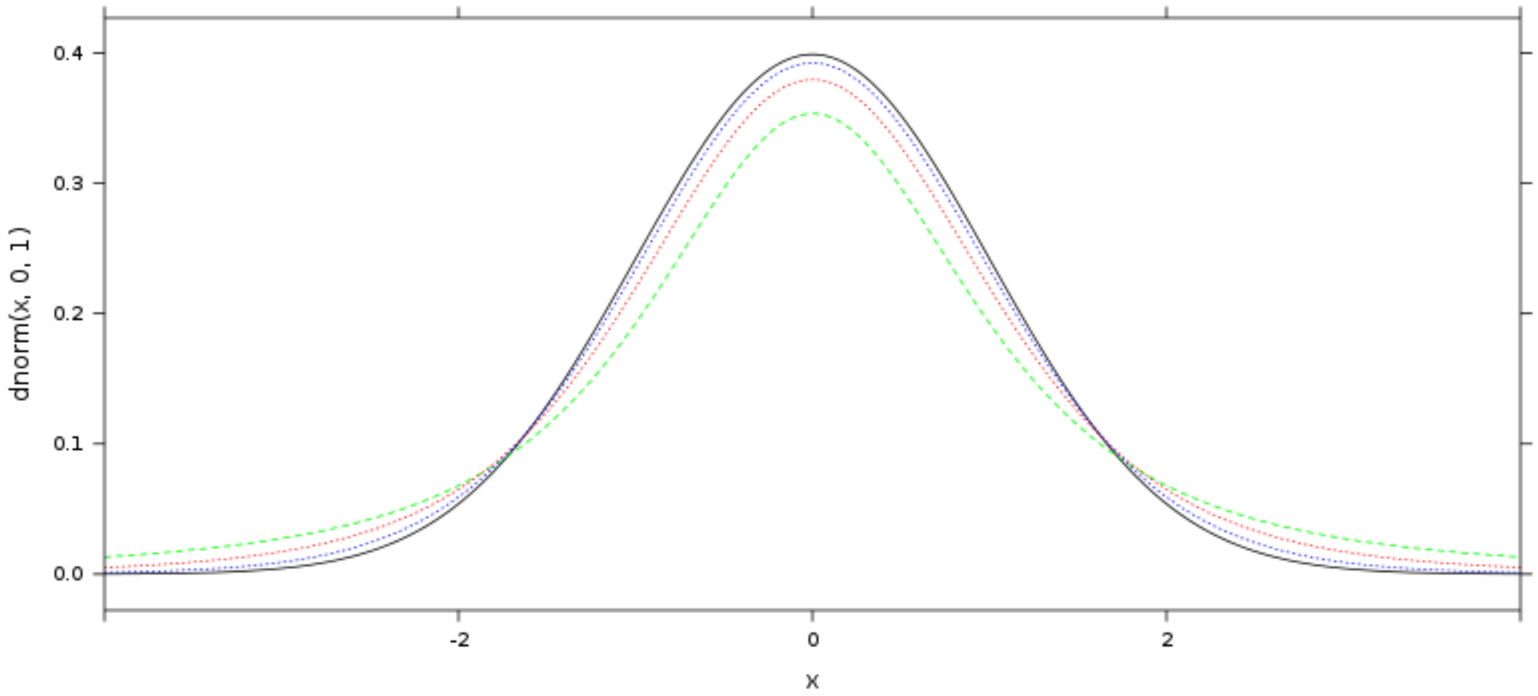
The t-distribution becomes more normal as the sample size *n* grows larger

There is an additional parameter called *the degrees of freedom*, that tells us which t-distribution to use

When working with $\bar{x}$ for a sample size of n, and we use a t-distribution with n-1 degrees of freedom

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

t-distributions



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

The Distribution of Sample Means ($\overline{x}$) 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

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

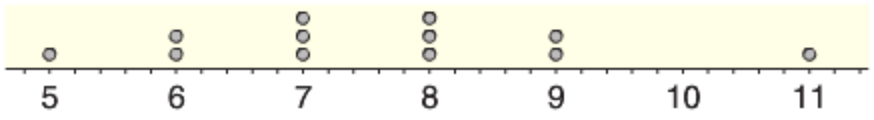
The fine print - this works if:

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

Is the t-distribution appropriate?

A

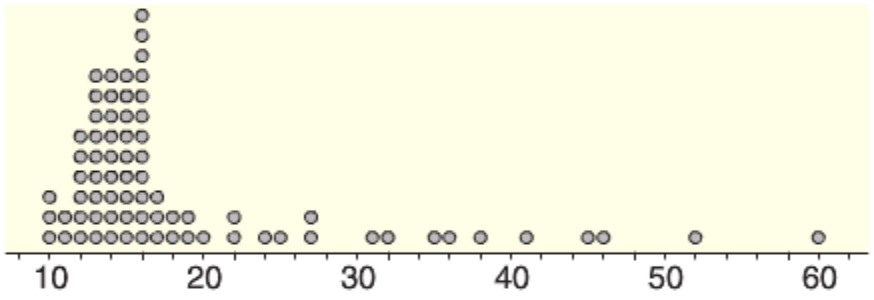
A sample with $n = 12$, $\overline{x} = 7.6$, and $s = 1.6$



Distribution seems normal
so OK to use t-distribution

B

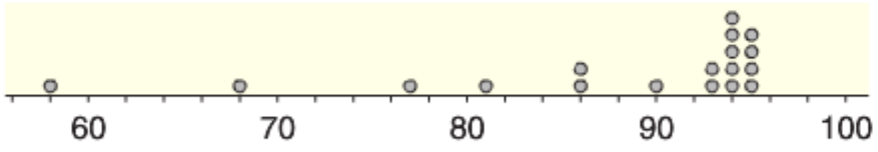
A sample with $n = 75$, $\overline{x} = 18.92$, and $s = 10.1$



Sample size is larger than n = 30
so OK to use the t-distribution

C

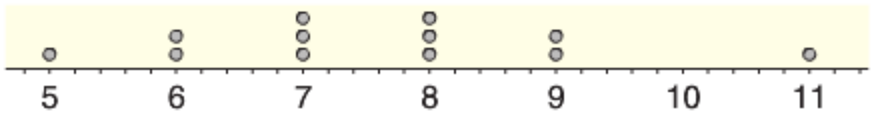
A sample with $n = 18$, $\overline{x} = 87.9$, and $s = 10.6$



Population distribution does not look normal and n < 30 so NOT ok to use the t-distribution

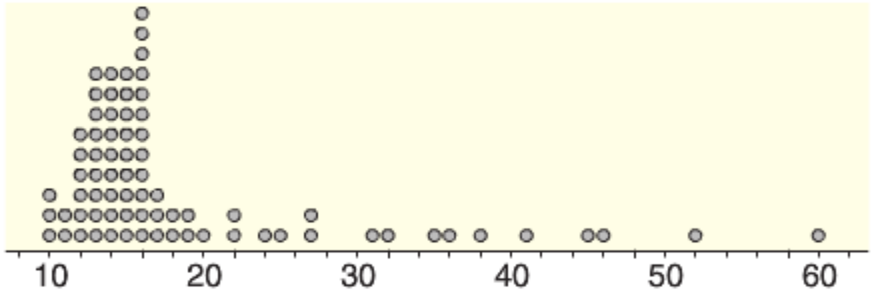
Is the t-distribution appropriate?

A A sample with $n = 12, \bar{x} = 7.6$, and $s = 1.6$



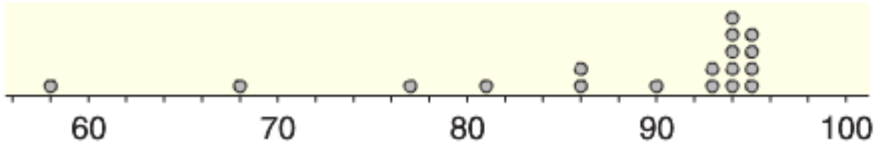
SE = 1.6/sqrt(12) = 0.462
df = 11

B A sample with $n = 75, \bar{x} = 18.92$, and $s = 10.1$



SE = 10.1/sqrt(75) = 1.166
df = 74

C A sample with $n = 18, \bar{x} = 87.9$, and $s = 10.6$



For A and B calculate the SE and the degrees of freedom

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

Calculating probabilities and quantiles from a t-distribution

$\Pr(T \leq t;\ deg\_of\_free)$ = `pt(t, df = deg_of_free)`

quantiles: `qt(area, df = deg_of_free)`

If a sample size is n = 16, write R code to calculate:

1. The 2.5th and 97.5th percentiles
2. Find the probability that a t-statistic is more than 1.5
3. Calculate these same values for the standard normal

Calculating probabilities and quantiles from a t-distribution

If a sample size is n = 16, calculate:

1. Calculate the 2.5th and the 97.5th percentiles
2. Find the probability that a t-statistic is more than 1.5
3. Calculate these same values for the standard normal

1. qt(c(.025, .975), df=15) = [-2.13 2.13]

2. 1 - pt(1.5, df=15) = 0.077

3. A) qnorm(c(.025, .975), 0, 1) = [-1.96 -1.96]

   B) 1 - pnorm(1.5, 0, 1) = 0.067

Parametric confidence intervals for a single mean

Confidence Interval for a single mean

For a normally distributed variable (e.g., a proportion), we saw that we could create a confidence interval with the formula:

Sample statistics $\pm z^* \times$ SE

We can use an similar formula for the sample mean which comes from a t-distribution with mean μ and $SE = \frac{s}{\sqrt{n}}$

A confidence interval for **μ** is:

Sample statistics $\pm$ $t^* \times$ SE

$= \bar{x} \pm t^* \times$ SE

Summary: 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:

- $\overline{x}$ is the sample mean and s is the sample standard deviation
- t* is an endpoint chosen from a t-distribution with n-1 df to give the desired confidence level

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

Light at night makes mice gain weight

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

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

R: qt(area, df)

Light at night makes mice gain weight

What is the parameter we are estimating?

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

$\overline{x}$ = 7.9,

s = 3,

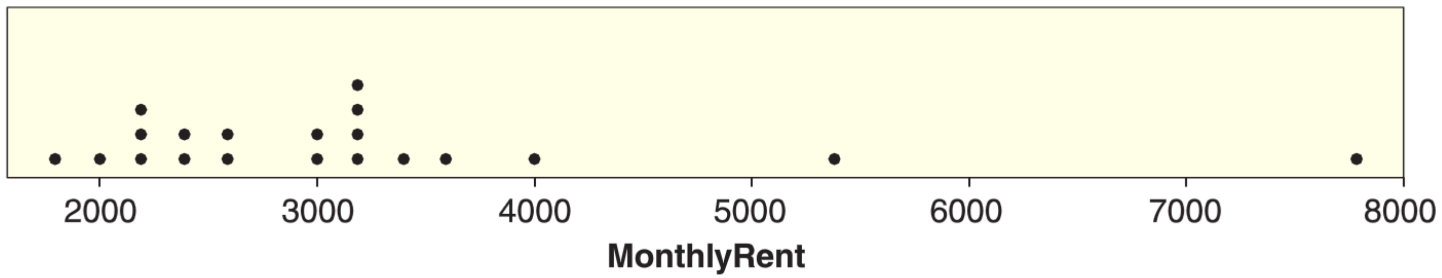
n = 10

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

7.9 ± 1.833 · 3/3.16 = (6.16 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

Hypothesis tests for a single mean

Parametric 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:

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

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 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 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)

The Chips Ahoy! Challenge

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

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

$\bar{x} = 1261.6$

$s = 117.6$

$n = 42$

$df = 41$

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?

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.