Parametric inference on means

Overview

Example of a final project presentation

Quick review of worksheet 11 and inference on proportions

Inference on means

A single mean

- Distribution, confidence intervals, and hypothesis tests

The difference between two means

- Distribution, confidence intervals, and hypothesis tests

Do beavers have the same body temperature as humans?

Ethan Meyers

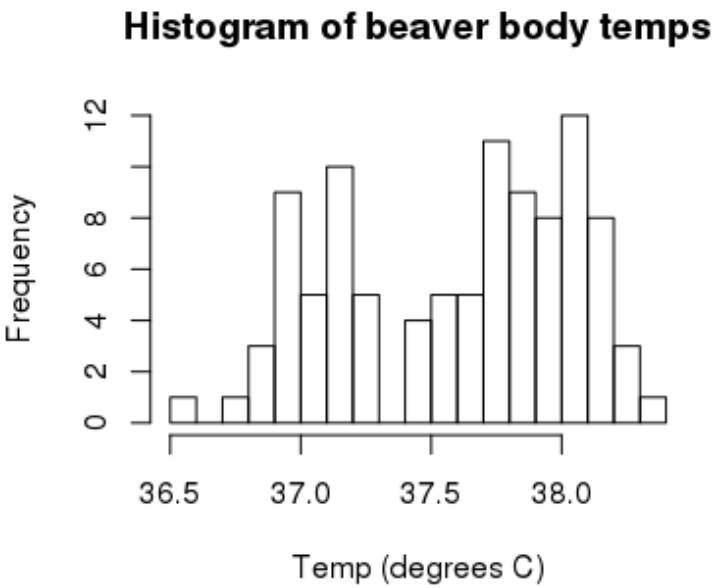
Motivation and data

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 (37C), perhaps they can be employed in the construction industry

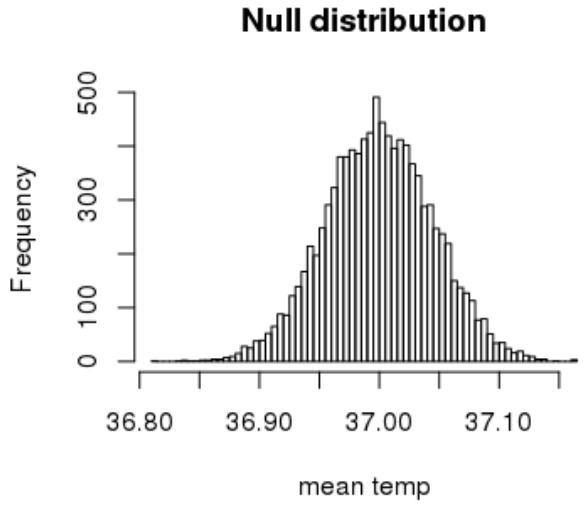
The data:

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



*not the real data

Results



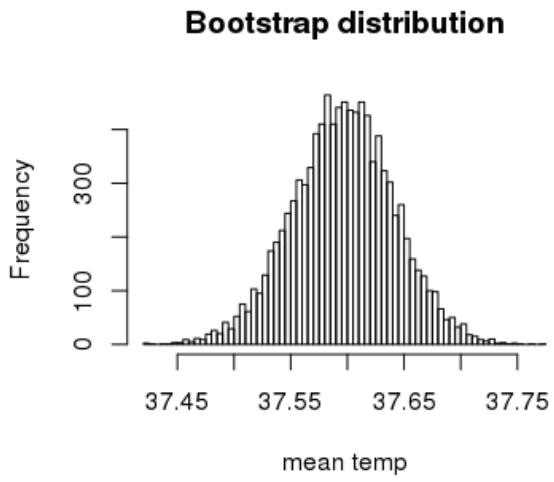
The average human body temperatures is $\mu = 37^{\circ}C$

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℃

Final project presentation guidelines

Must be no longer than 5 minutes!

A final project presentation template can be downloaded from:
bit.ly/CS206_final_presentation

Final project presentation slides are due next Sunday, Dec. 9th at 11:59pm

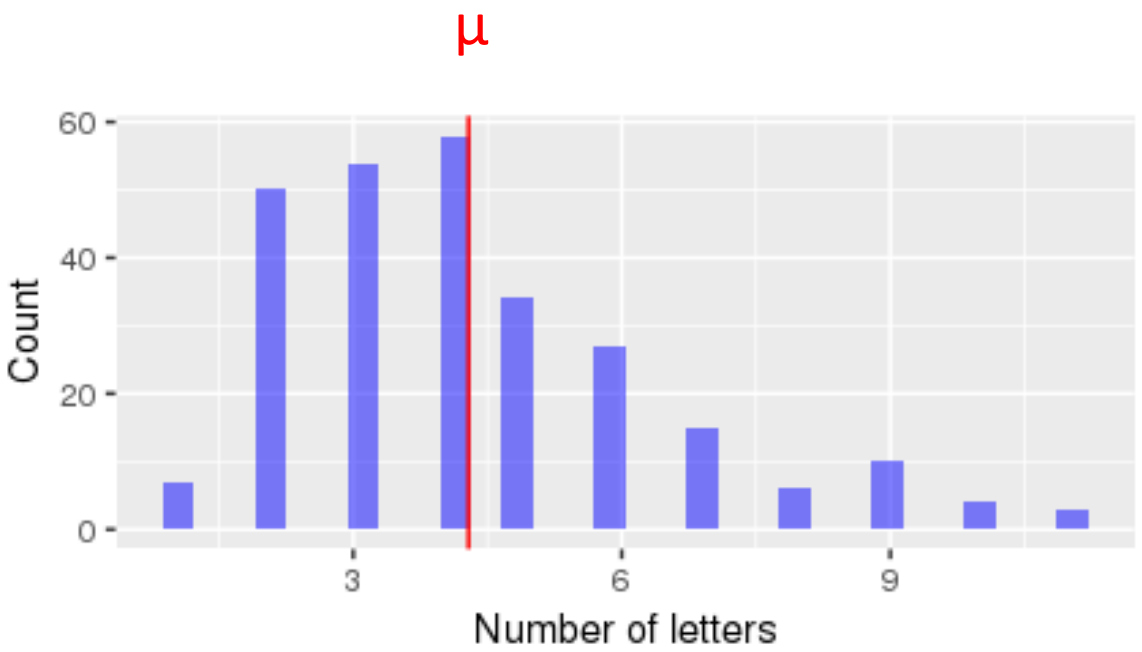
Questions?

Review: art time!

Draw:

- Population for a categorical variable (e.g., sprinkles)
- 1 sample that has 100 points
- 9 more samples that have 100 points
- Plato
- A population parameter with appropriate symbol
- Sample statistics with appropriate symbol
- A sampling distribution!

Gettysburg address word length sampling distribution



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

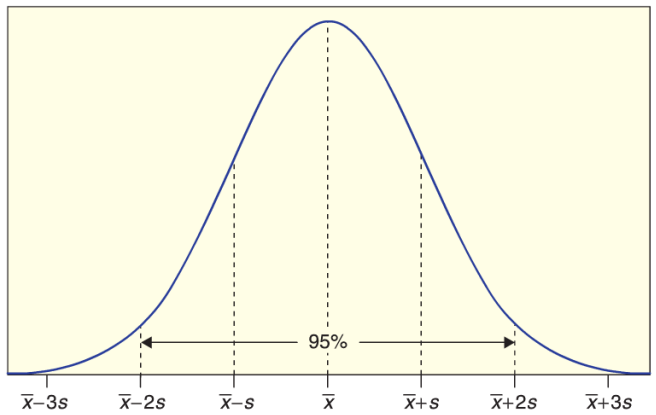
$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

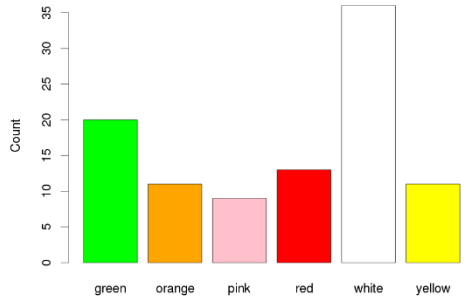
3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

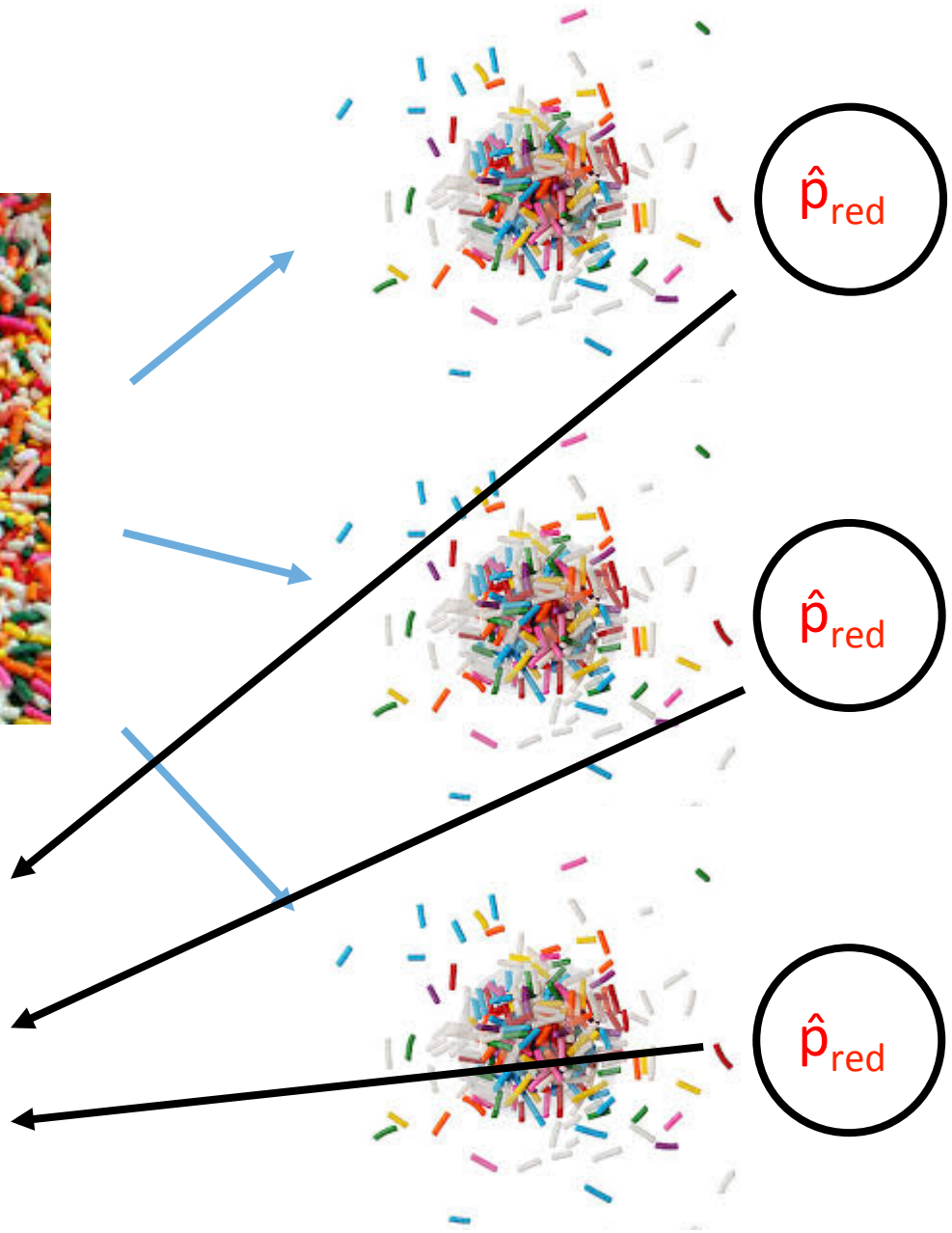


Sampling distribution!

Gettysburg sampling distribution app

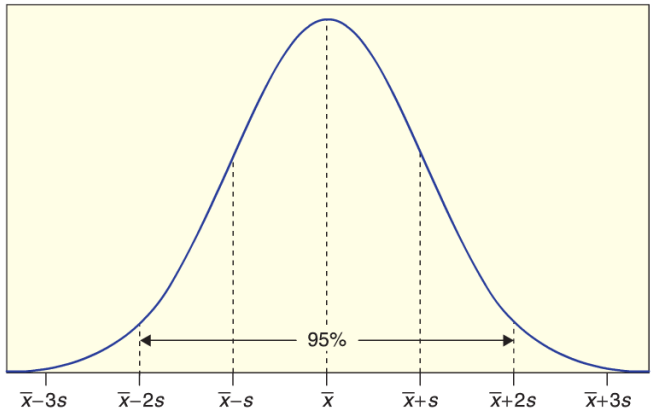


π_{red}



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

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



Sampling distribution!

Worksheet 11 – question 6.4

Q: Suppose we have a population proportion π = 0.27. Let's create the sampling distribution for n = 50 observations:

- a) First, find the mean and SE for the sampling distribution
- b) Then plot the sampling distribution

A: The central limit theorem says:

- The sampling distribution for **proportions (p̂)** is **normally distributed** and is **centered at the value of the population parameter (π)**

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

μ = .27

σ or SE = .063

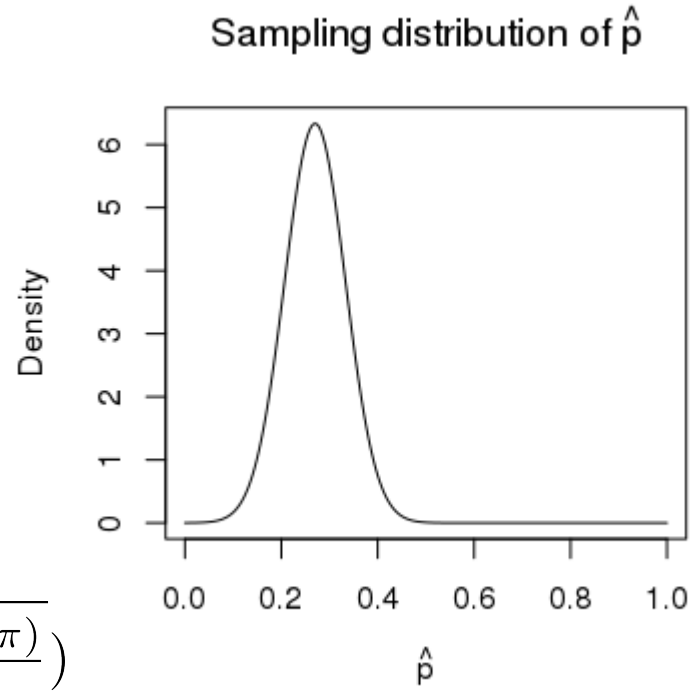
Worksheet 11 – question 6.4

Q: Suppose we have a population proportion π = 0.27. let's create the sampling distribution for n = 50 observations:

- a) First, find the mean and SE for the sampling distribution
- b) Then plot the sampling distribution

b)

```
x_vals <- seq(0, 1, by = 0.001)
density_curve <- dnorm(x_vals, 0.27, .063)
plot(x_vals, density_curve, type = "l")
```



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

Worksheet 11 – other questions?

Review of inference on proportions

Central Limit Theorem for Sample Proportions

For random samples of size n from a population with a proportion π, the distribution of the sample proportions has the following characteristics:

Center: The mean is equal to the population proportion, π

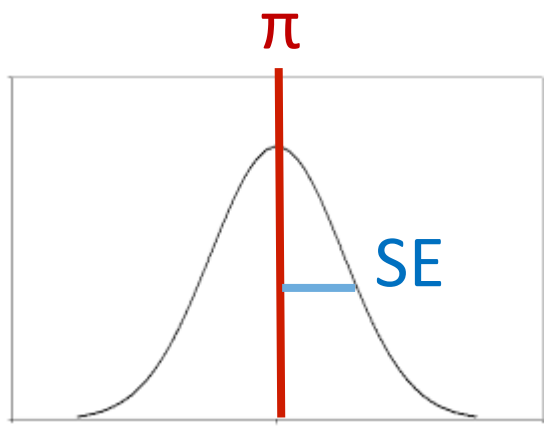
Spread: The standard error is: $SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

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

A normal distribution is a good approximation as long as:

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

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



Confidence intervals and tests 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 $\hat{p}$ for π

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

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

Our z statistic comes from a standard normal distrinution $z \sim N(0, 1)$

Sinister lawyers

10% of American population is left-handed.

A study found that out of a random sample of 105 lawyers, 16 were left-handed.

Test whether the proportion of left-handed lawyers is greater than the proportion found in the American population.

1. State the null and alternative hypothesis

2-4. Calculate the statistic of interest and calculate the p-value
- Hint: the pnorm() function will be useful

5. What do you conclude?

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

Sinister lawyers

1. State the null and alternative hypothesis
   - H_0: $\pi = .10$
   - H_A: $\pi > .10$

2-4. Calculate the statistic of interest and the p-value
   - $\hat{p} = 16/105 = .152$
   - SE = sqrt((.10 * (1 - .10))/105) = .029
   - z = (.152 - .10)/.029 = 1.79
   - 1 - pnorm(z, 0, 1) = .037

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

5. What do you conclude?

Inference on means

1. From the central limit theorem we know that 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

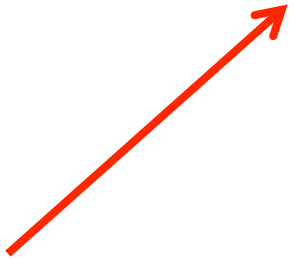
3. What other piece of information would be need to plot the sampling distribution of $\bar{x}$?
    - A: SE

4. And how can we get the SE?
    - A: 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:

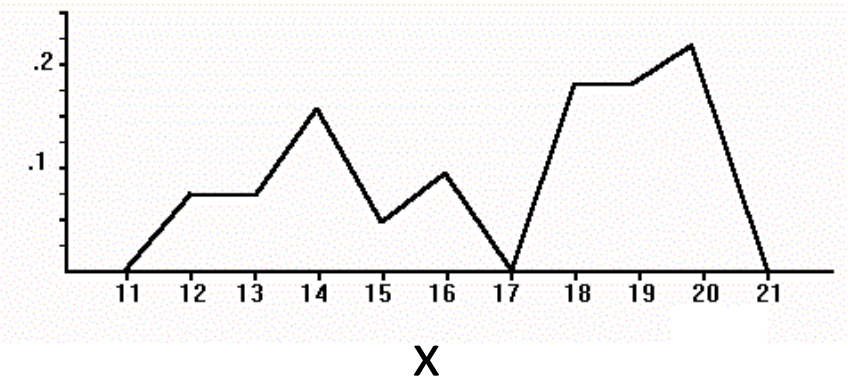
$$SE = \frac{\sigma}{\sqrt{n}}$$

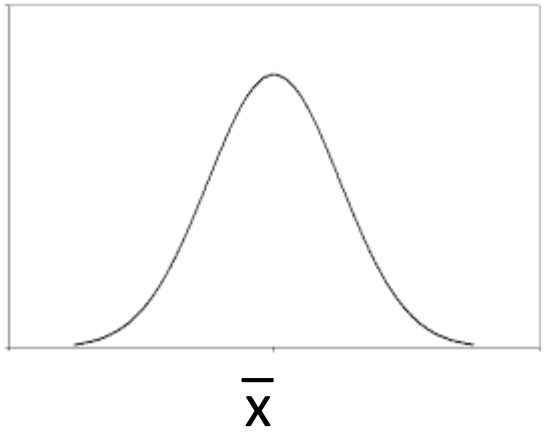


The larger the sample size (n), the smaller the standard error

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.

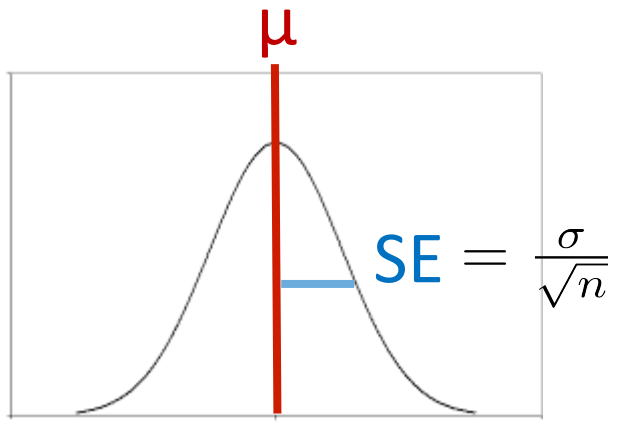
Sample sizes of 30 are usually sufficient. If the original population is normal we can get away with smaller sample sizes

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 ($\bar{\mathbf{x}}$) 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 , $\bar{x}$, 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

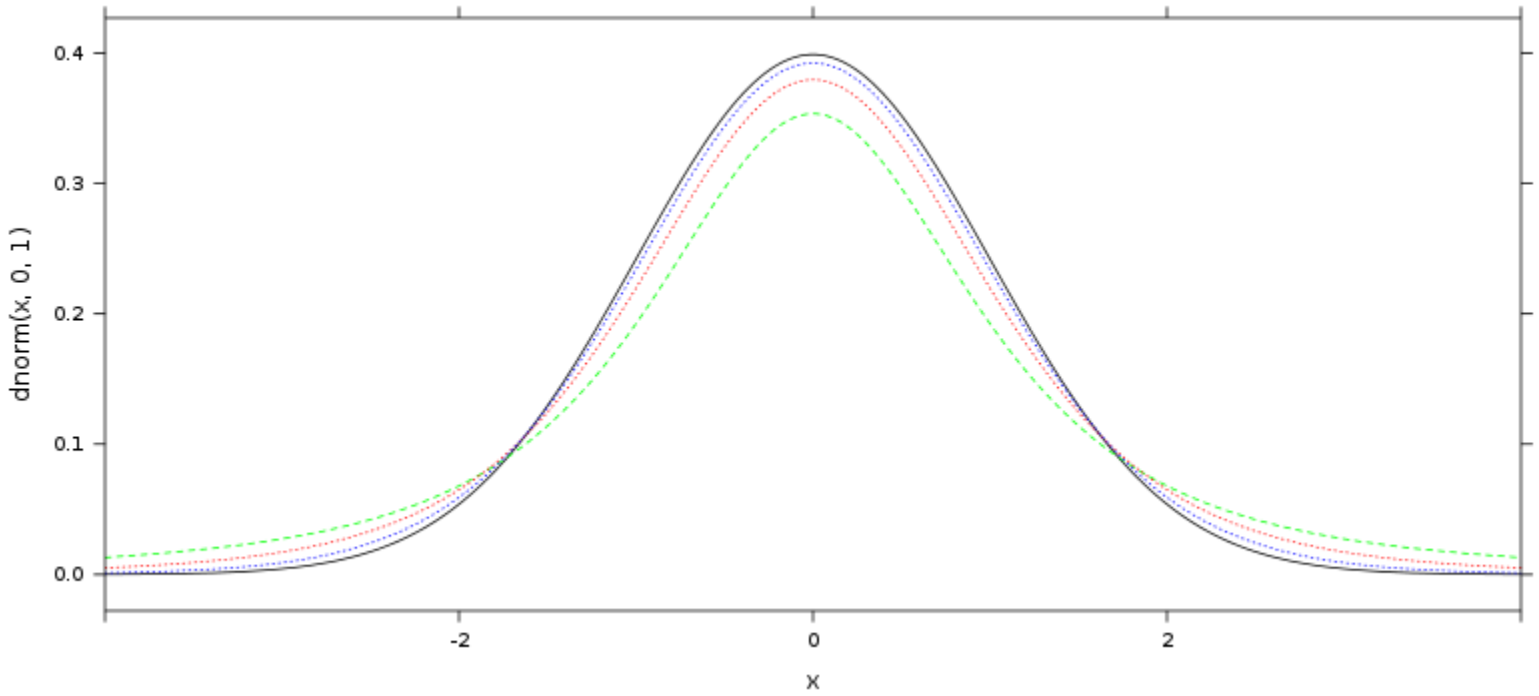
Fortunately about 100 years ago William Sealy Gosset figured out that this sampling distribution where s is substituted for σ has another parametric form called a t-distribution

The t-distribution becomes more normal as the sample size grows, so 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

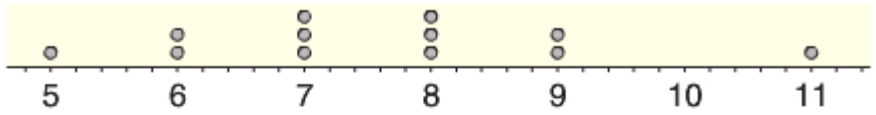
$$\frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

The fine print - this works if:

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

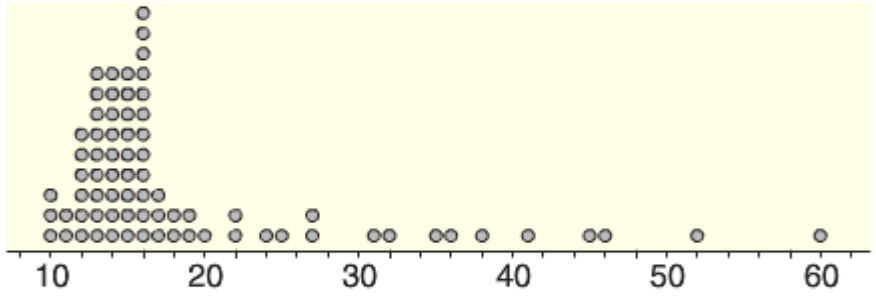
Is the t-distribution appropriate?

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



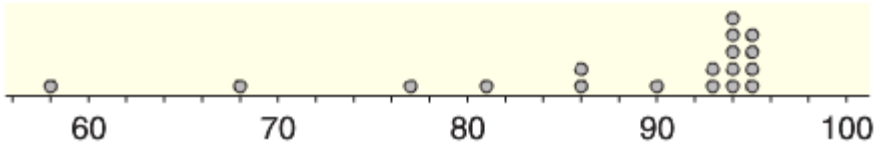
Distribution seems normal
so OK to use t-distribution

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



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

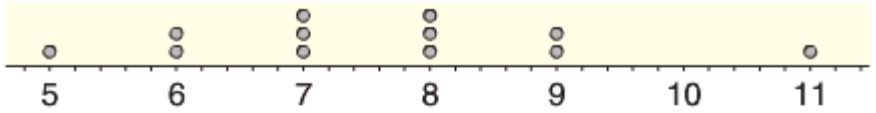
C A sample with $n = 18, \bar{x} = 87.9$, and $s = 10.6$



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

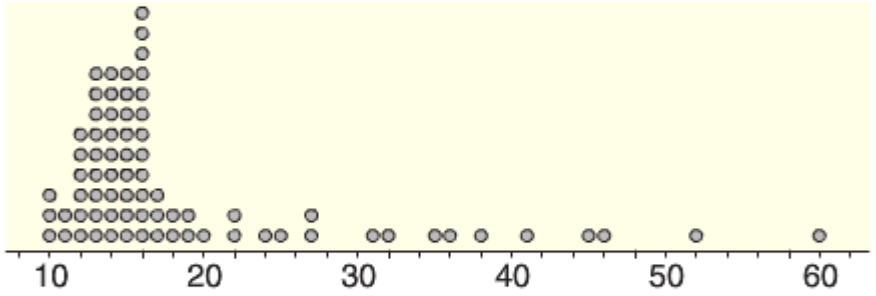
Is the t-distribution appropriate?

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



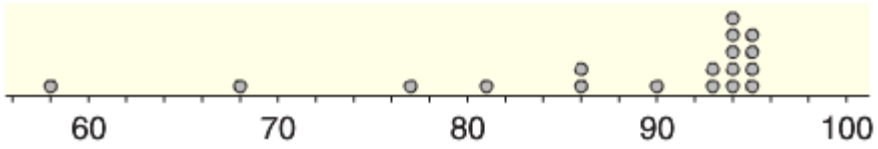
SE = 1.6/sqrt(12) = 0.462
df = 11

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 2.5^{th} and 97.5^{th} 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 2.5^{th} and the 97.5^{th} 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(.025, 0, 1) = [-1.96 -1.96]

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

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:

$$\text{Sample statistics} \pm z^* \times \text{SE}$$

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:

$$\text{Sample statistics} \pm t^* \times \text{SE}$$

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

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, and 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 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

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

R: qt(area, df)

Light at night makes mice fat

What is the parameter we are estimating?

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

$\bar{x}$ = 7.9,

s = 3,

n = 10

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

7.9 ± 1.833 · 3/3.16 = (6.16 9.64)