Parametric inference on proportions

Overview

Example of a final project

Quick review of Normal distributions and hypothesis tests/CI using normal distributions

Parametric inference on proportions

- Distribution of a sample proportion
- Confidence interval for a single proportion
- Tests for a single proportion

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 also an R Markdown template for the final project on Canvas

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

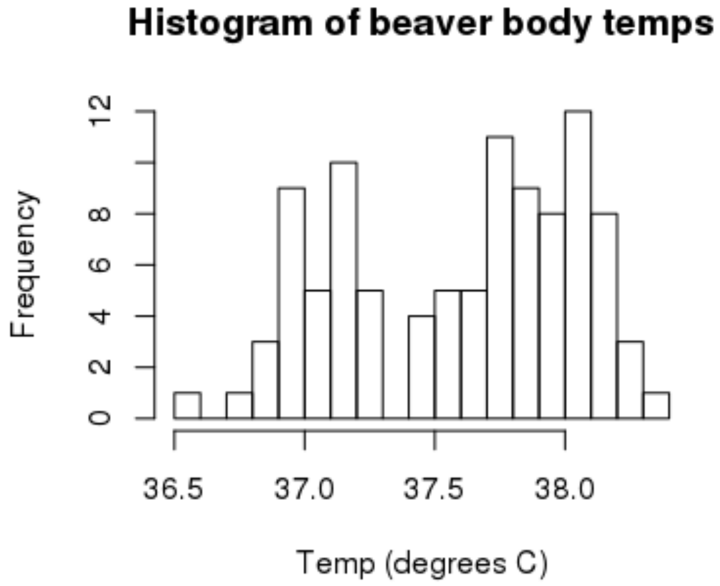
Motivation and data

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

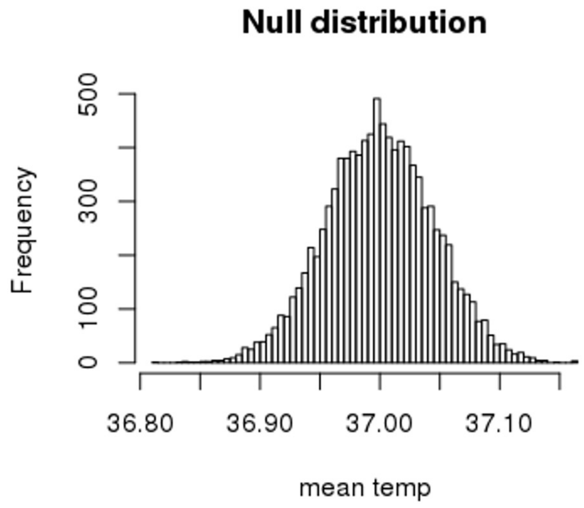
The data:

- 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

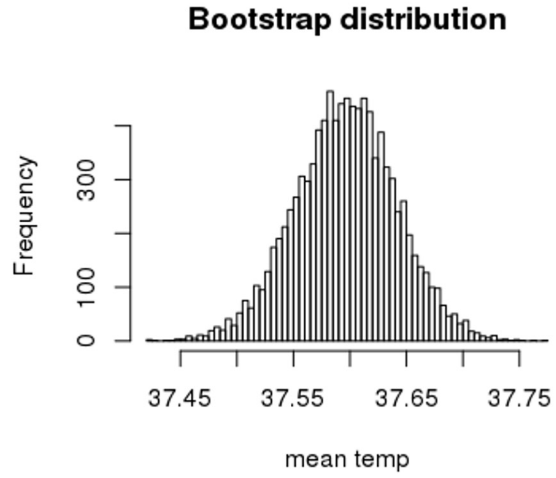


The average human body temperatures is $\mu = 37°C$

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

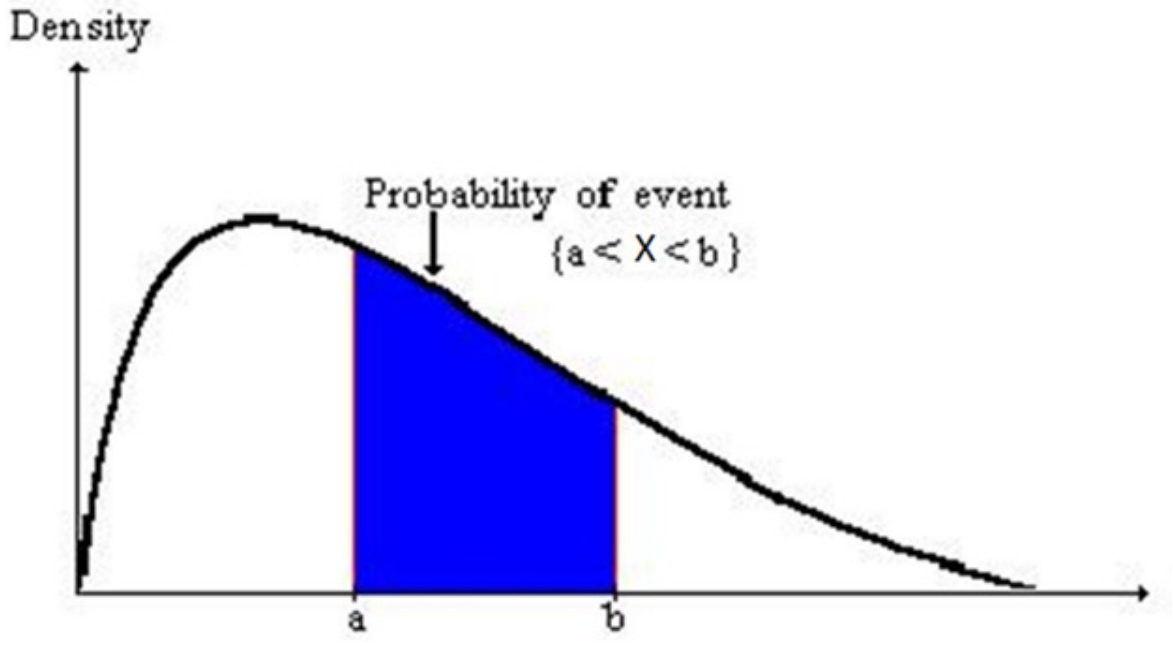
Caveats: human body temperatures might not be exactly 37℃

Quick review of Normal distributions

Density Curves

The probability that a random number X will be in the interval [a, b] can be modeled using the <u>area under a density curve</u>

Pr(a < X < b) is the area under the curve from a to b



D**ensity curve** are functions f(x) that have two key properties:

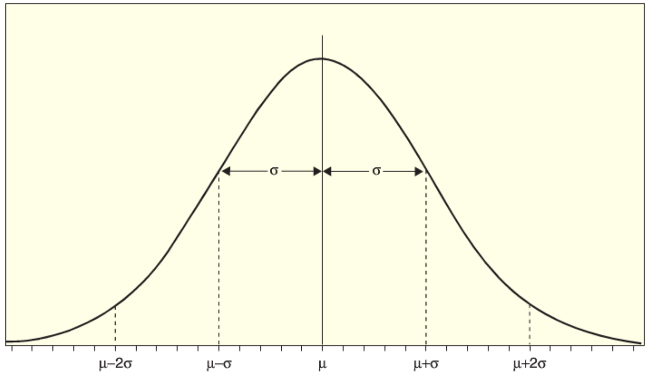
1. The total area under the curve f(x) is equal to 1

2. The curve is always ≥ 0

The Normal Density Curve

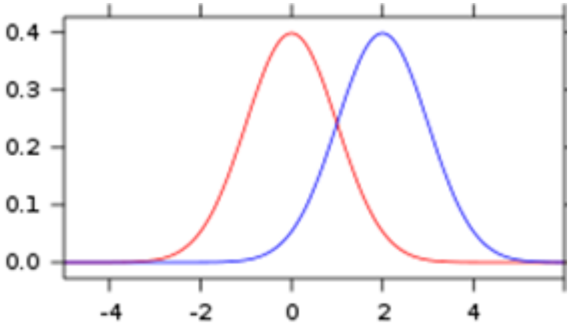
Normal distributions are a family of bell-shaped curves with two parameters

- The mean: μ
- The standard deviation: σ

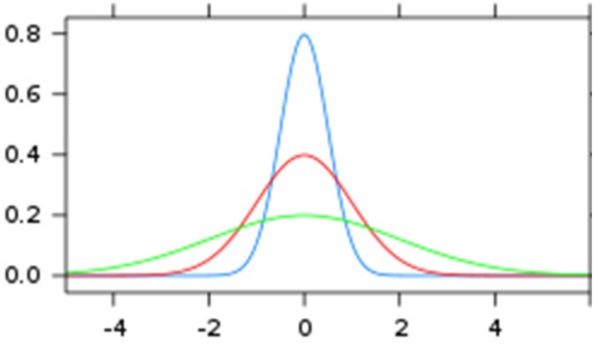


Notation: X ~ N(μ, σ)

Changing μ



Changing σ



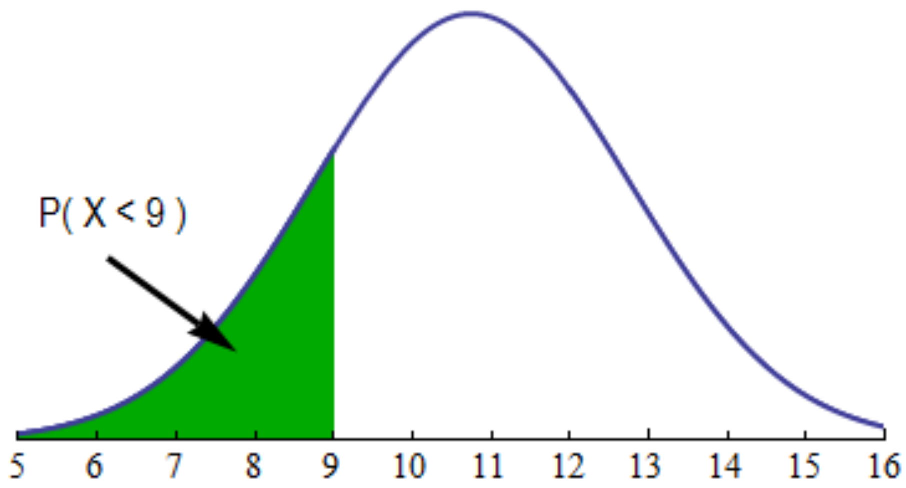
Densities, probabilities and quantiles from normal distributions

We can plot the density curve using:

dnorm(x_vec, mu, sigma)

We can get the probability that we would get a random value less than x using:

pnorm(x_vec, mu, sigma)



We can get the quantile values using:

qnorm(area, mu, sigma)

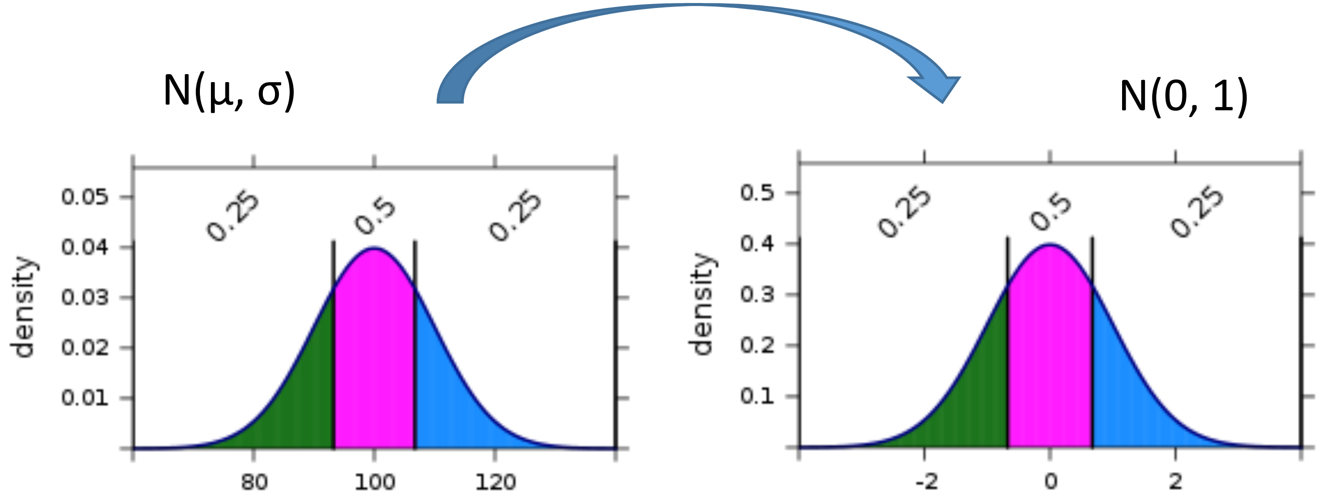
Standard Normal N(0, 1)

It is convenient to work with the **standard normal** distribution:

$$Z \sim N(0, 1) \qquad \text{i.e.,} \quad \mu = 0, \;\; \sigma = 1$$

We transform any normally distributed random variable $X \sim N(\mu, \sigma)$ to the standard normal distribution $Z \sim N(0, 1)$ using:

$$Z = (X - \mu)/\ \sigma$$



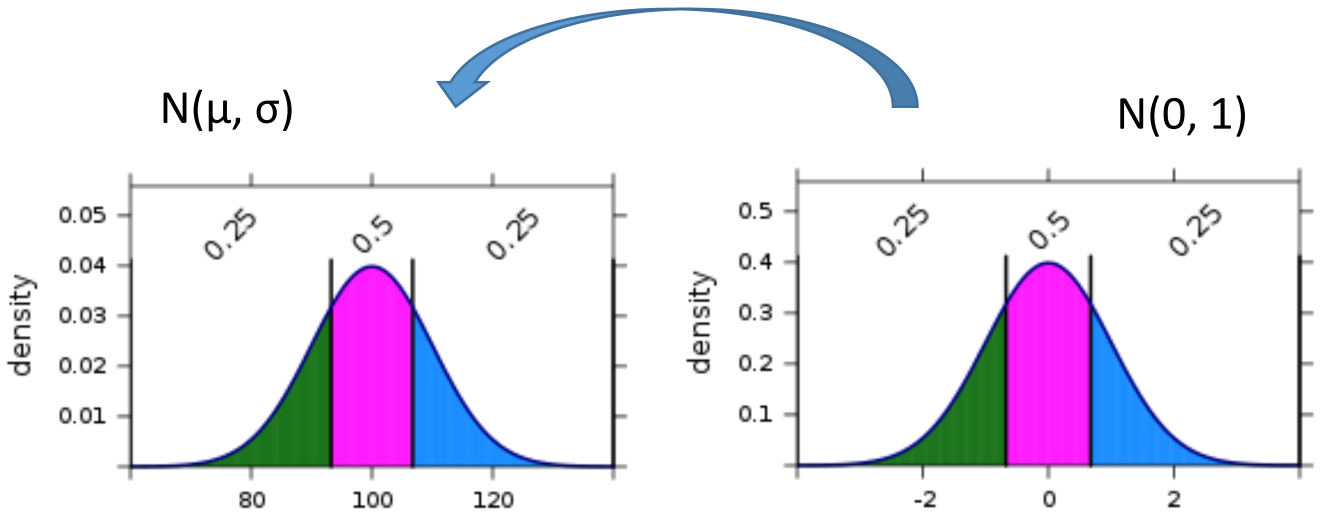
Standard Normal N(0, 1)

It is convenient to work with the **standard normal** distribution:

$$Z \sim N(0, 1) \qquad \text{i.e.,} \quad \mu = 0, \ \sigma = 1$$

To convert from $Z \sim N(0, 1)$ to any $X \sim N(\mu, \sigma)$, we reverse the standardization with:

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

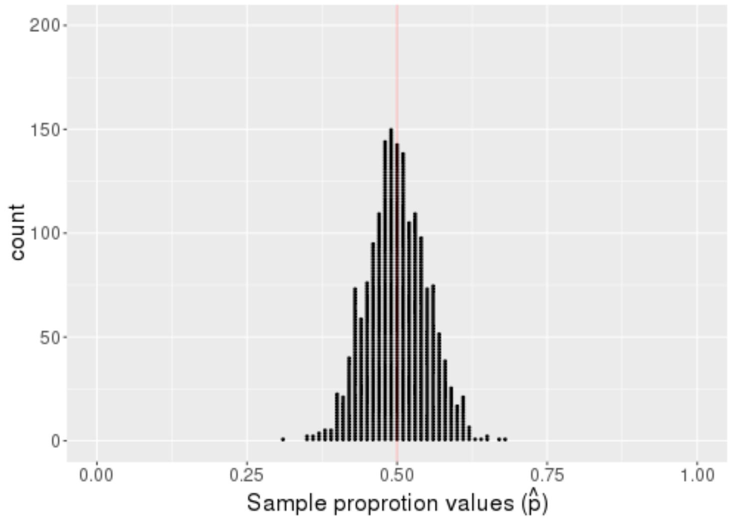


Central limit theorem

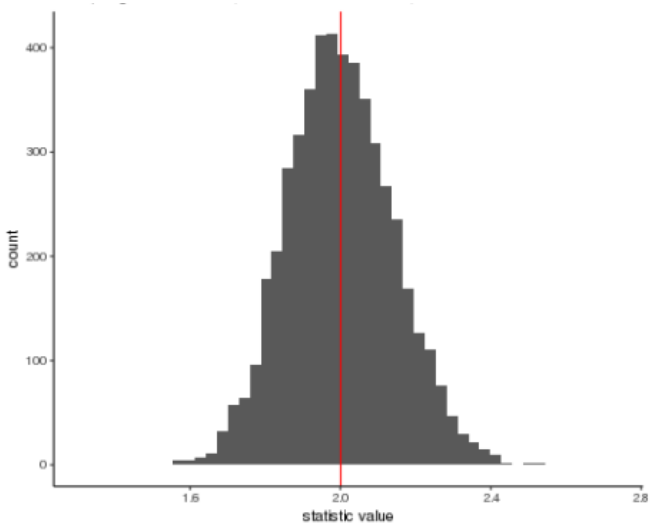
For random samples with a sufficiently large sample size (n), the distribution of sample statistics for a mean ($\bar{x}$) or a proportion ($\hat{p}$) is:

- normally distributed
- centered at the value of the population parameter

proportion ($\hat{p}$)



mean ($\bar{x}$)

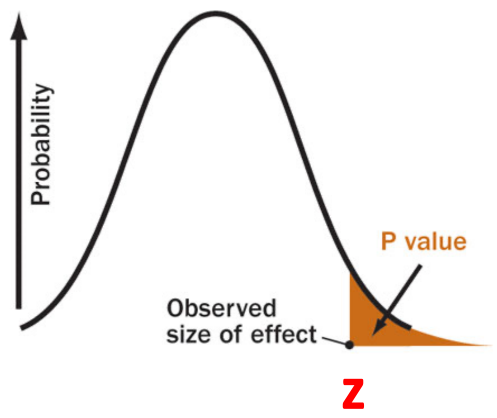


Hypothesis tests based on a Normal Distribution

When the null distribution is normal, it is often convenient to use a standard normal test statistic using:

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

The p-value for the test is the probability a standard normal value is beyond this standardized test statistic



$$\Pr(Z \geq z_{obs}\ ;\ \ \mu = 0,\ \ \sigma = 1)$$

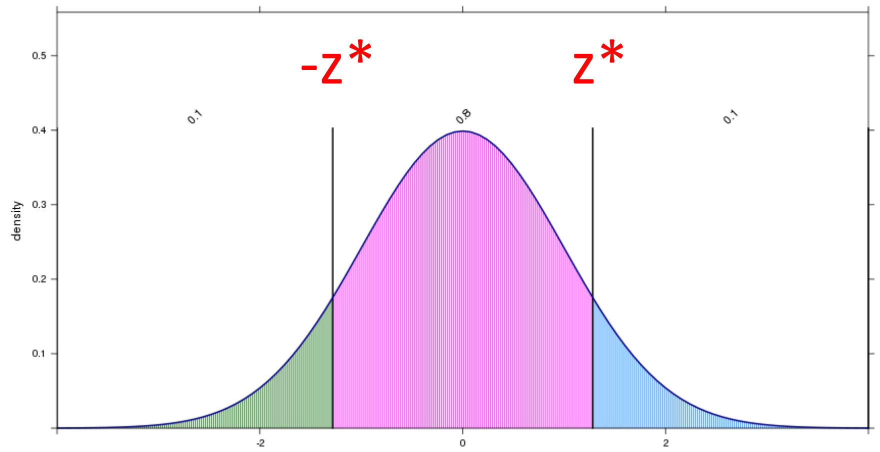
pnorm(z, 0, 1, lower.tail = FALSE)

Confidence intervals based on a Normal Distribution

If the distribution for a statistic is normal with a standard error SE, we can find a confidence interval for the parameter using:

sample statistic $\pm$ $z^* \times$ SE

where z^* is chosen so that the area between $-z^*$ and $+z^*$ in the standard normal distribution is the desired confidence level



Confidence level	80%	90%	95%	98%	99%
Z*	1.282	1.645	1.960	2.326	2.576

```
z_stars <- qnorm(c(.90, .95, .975, .99, .995), 0, 1)
```

Parametric inference on proportions

Review: questions about proportions

Q1: What symbols have we been using for the parameter and statistic for proportions?

- Parameter: π
- Statistic: $\hat{p}$

Q2: What are examples of confidence intervals and hypotheses tests we've run for proportions?

- Hypothesis tests: Doris and Buzz, Paul, Joy, etc.
- Confidence intervals: proportion of red sprinkles, etc.

Review: questions about proportions

Q3: What does the shape of a sampling distribution for a proportions $\hat{p}$ look like?

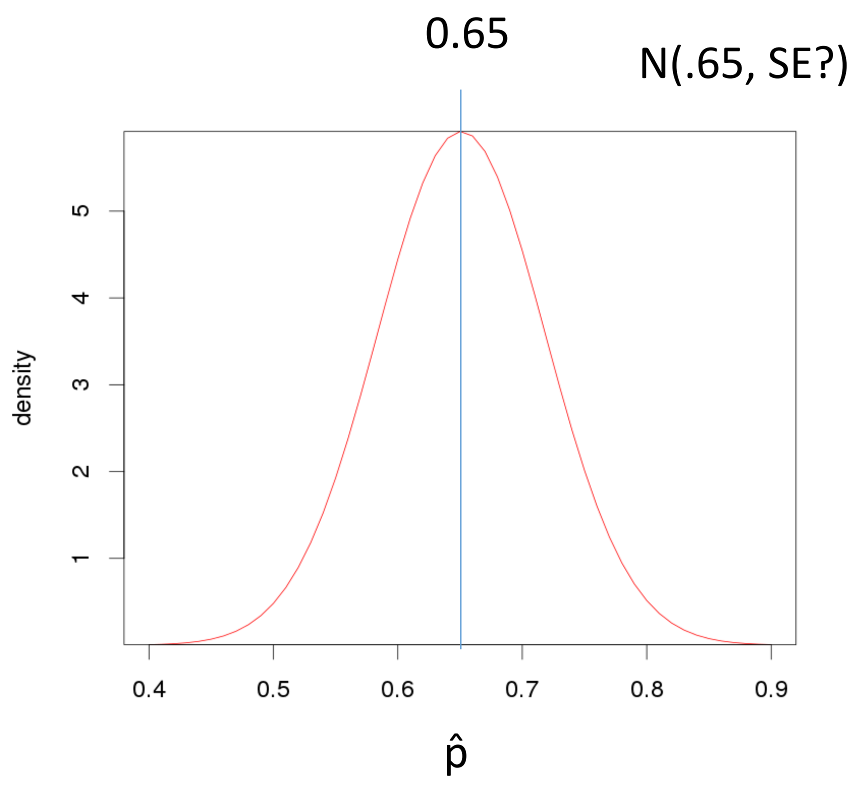
- A: normal!
  - (If the sample size n is larger enough)

Q4: Suppose $\pi = .65$, and $n = 50$, could you draw the sampling distribution for $\hat{p}$?

- A: It is centered at 0.65, but what is the spread (SE)?

We could use the bootstrap to estimate the SE with SE*

Alternatively, we can use a math/theory



Standard Error for Sampling Proportions

When choosing random samples of size n from a population with proportion π, the standard error (SE) of the sample proportions is given by:

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

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

SE for percentage of houses owned

65.1% of all houses are owned   ($\pi = .651$)

If we randomly selected 50 houses…

a) What would the standard error (SE) of sampling distribution for the proportion of owned houses ($\hat{p}$) be?

b) What would this sampling distribution look like?

What if we randomly selected 200 houses?

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

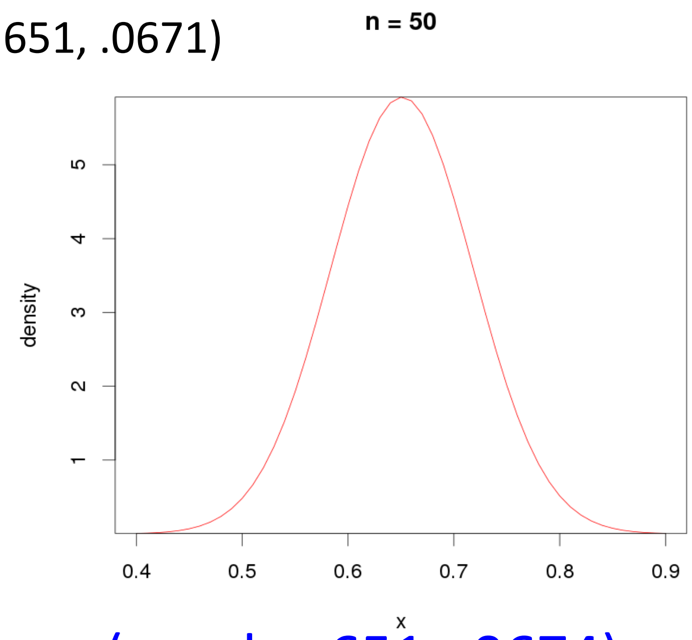
SE for percentage of houses owned

65.1% of all houses are owned

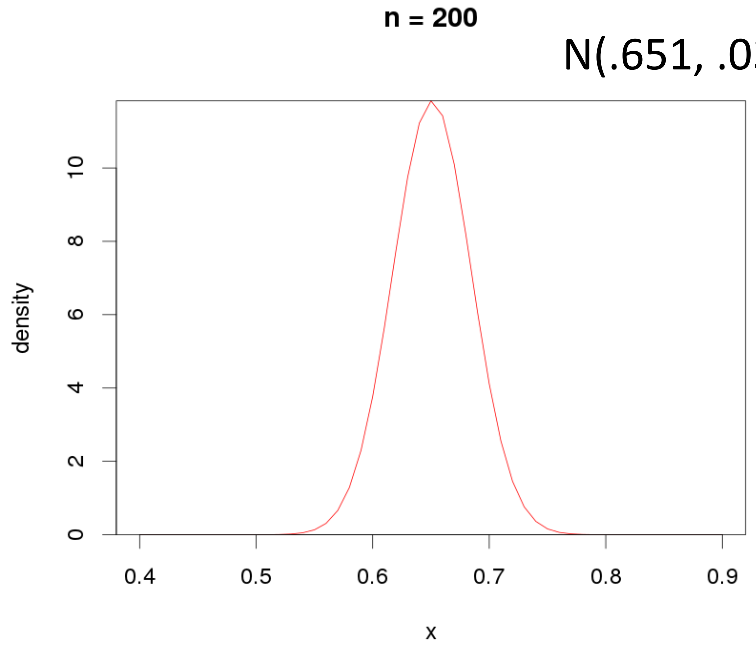
- $\pi = .651$
- When n = 50: SE = .0674
- When n = 200: SE = .0337

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

N(.651, .0671)



N(.651, .0337)



y_vals <- dnorm(x_vals, .651, .0674)

How large of a sample size n is needed for the sampling distribution of p̂ to be normal?

n = 50

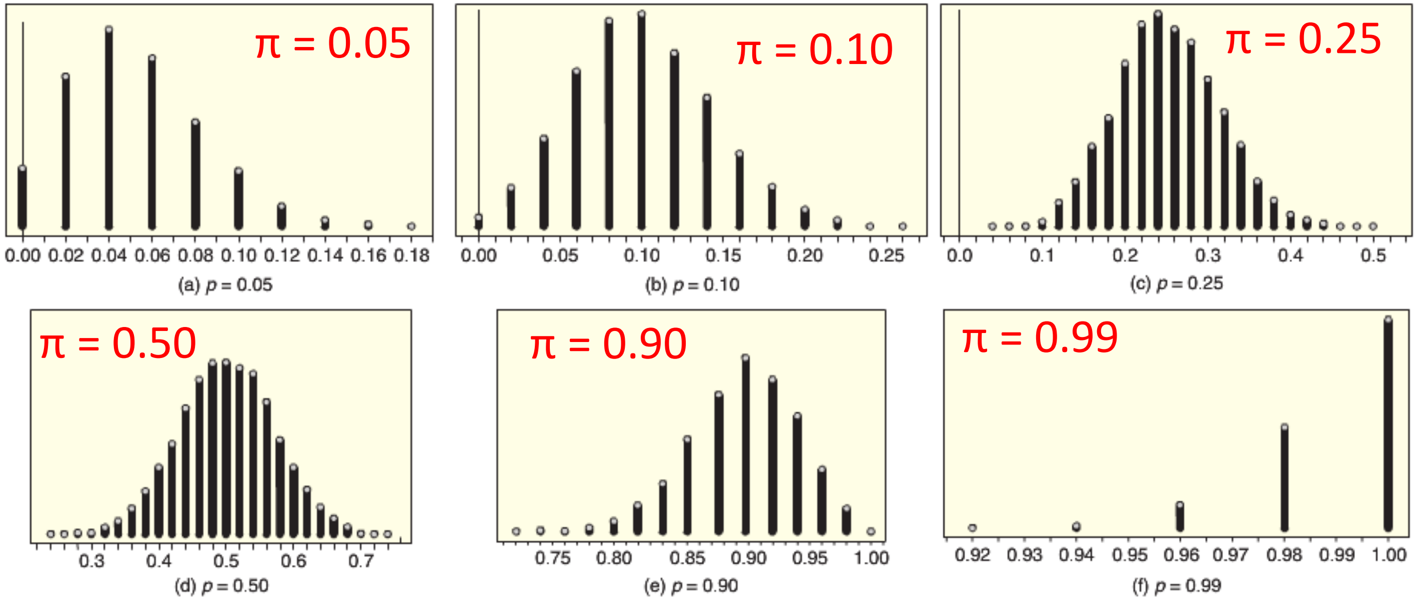


Figure 6.2 *Distributions of sample proportions when $n = 50$*

π = 0.10

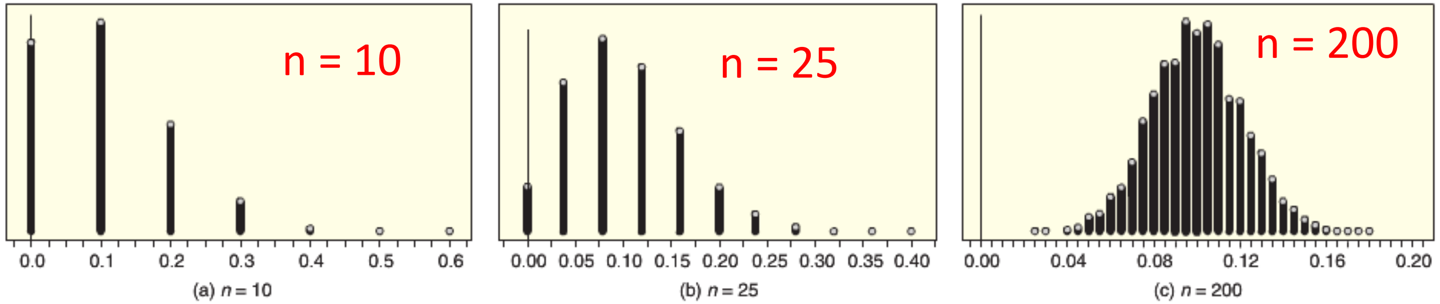


Figure 6.3 *Distributions of sample proportions when $p = 0.10$*

How large of a sample is needed for the normal approximation?

The normal approximation is reasonable good when we see 10 "positive" outcomes and 10 "negative" outcomes

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

Summary: Central Limit Theorem for Sample Proportions

For samples of size n from a population with a proportion π,
the distribution of the sample proportions 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

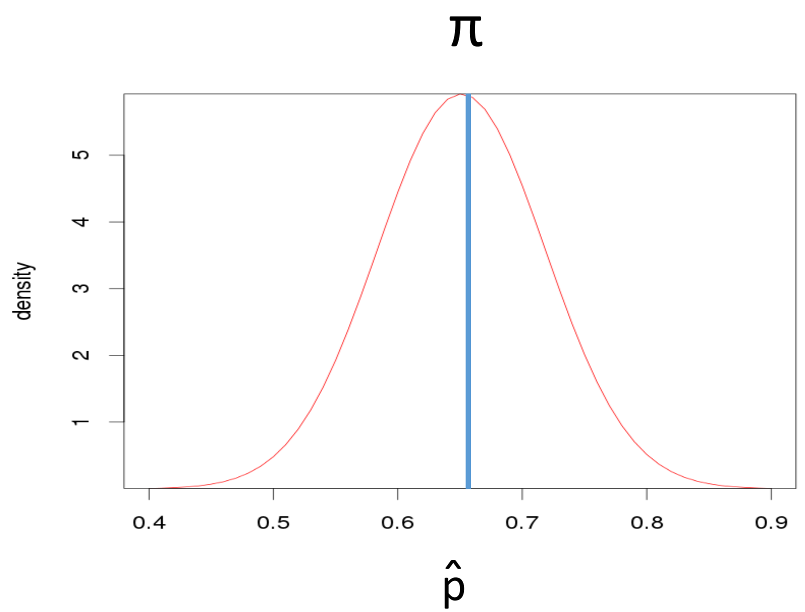
Spread: The standard error is: $SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

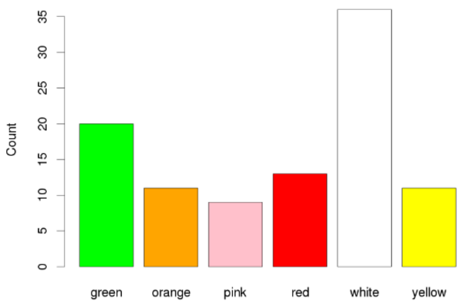
The larger the sample size, the more like a normal distribution it becomes.
A normal distribution is a good approximation as long as:

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

Summary: Central Limit Theorem for Sample Proportions

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





π_{red}

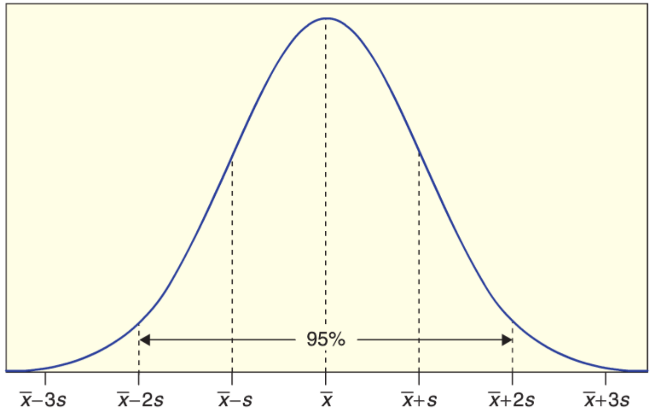
$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

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

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

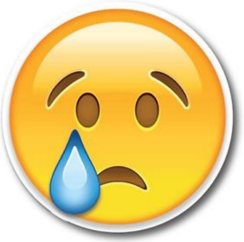


Sampling distribution!

Standard Error for Sampling Proportions

Note: we don't usually know π, so we can't compute the standard error exactly using the formula:

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



However, we can substitute p̂ for π and then we can get an estimate of the standard error:

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

Comparing formula SE to the bootstrap SE

In previous classes we have used the bootstrap to get an estimate of the standard error SE*

How could we do this for the green sprinkles?

```
bootstrap_dist  <-  do_it(100000) *  {
    boot_sample <- sample(my_sprinkles, replace = TRUE)
    sum(boot_sample  == 'green')/100
}

bootstrap_SE <- sd(bootstrap_dist)
```

Color
White
Red
Red
White
Green
White
.
.
.
White
Green

n = 100 sprinkles

Comparing formula SE to the bootstrap SE

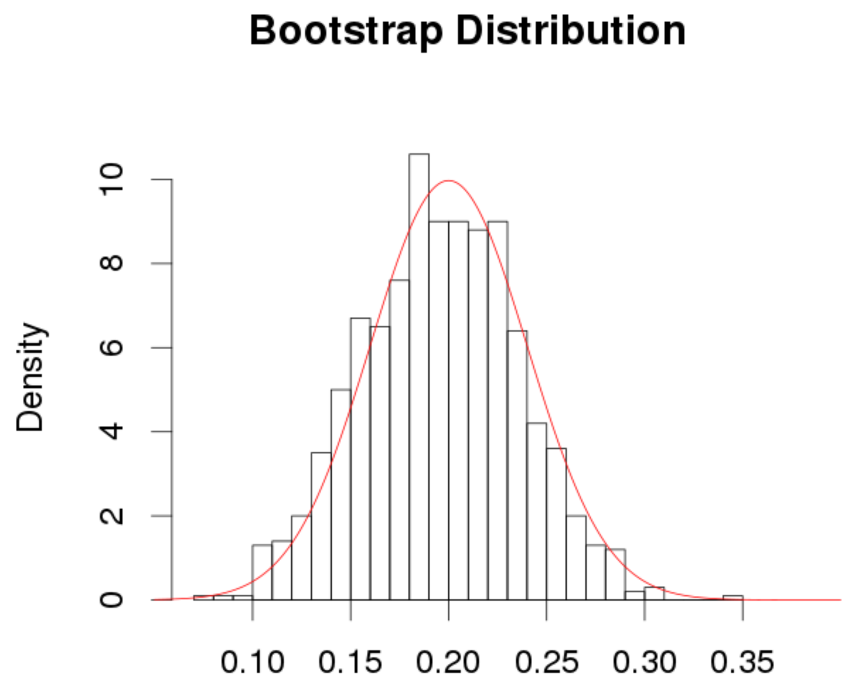
For my green sprinkles I got:

- Bootstrap SE = 0.039959
- Formula SE = 0.04

$\hat{p} = 0.20$

$n = 100$

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




```
SE <- sqrt(  (.2 * (1 - .2)  ) /100   )
```

Parametric confidence intervals for proportions

Confidence intervals for a single proportion

Suppose we have a sample of size n of categorical data

Suppose that n is large enough so that $n\pi \geq 10$ and $n(1 - \pi) \geq 10$

A confidence interval for a population proportion π can be computed from our random sample of size n using:

Equation for SE

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

Where $\hat{p}$ is the sample proportion and z^* is a standard normal endpoint to give the desired confidence level

Sprinkle example

To create a confidence interval for proportion of green sprinkles π_{green} we take a sample of size n = 100

1	orange
2	red
3	green
4	white
5	white
6	white
7	white
8	white
9	red

My green sprinkles

20 of the 100 sprinkles were green

What is a 95% confidence interval for the population proportion π of green sprinkles?

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

My green sprinkles

p̂ = 20/100 = .20

n = 100

SE = .04

z* = 1.96 (for 95% CI)

CI = 0.1216 to 0.2784

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

$$.20 \;\pm\; 1.96 \cdot \sqrt{\frac{.2\cdot(1-.2)}{100}}$$

Parametric hypothesis tests for proportions

Test for single proportions

To compute p-values when the null distribution is normal we use:

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

In the context of proportions our null hypothesis is of the form H_0: $\pi = \pi_0$

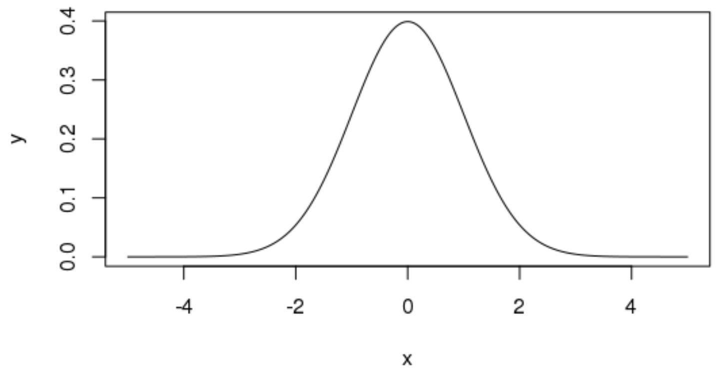
Our formula for z then becomes:

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

Test for single proportions

To test for H_0: $\pi = \pi_0$ vs H_A: $\pi \neq \pi_0$ (or the one-tail alternative), we use the standardized test statistic:

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



Where p̂ is the proportion in a random sample of size n

Provided the sample size is reasonable large (usual conditions), the p-value of the test is computed using the standard normal distribution

Do more that 25% of US adults believe in ghosts?

A telephone survey of 1000 randomly selected US adults found that 31% of them say they believe in ghosts. Does this provided evidence that more than 1 in 4 US adults believe in ghosts?

1. State the null and alternative hypothesis

2. Calculate the statistic of interest

3-4. Calculate the p-value

Hint: the pnorm() function will be useful

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

5. What do you conclude?

Do more that 25% of US adults believe in ghosts?

Step 1:

H_0: $\pi = .25$

H_A: $\pi > .25$

Step 2:

$\hat{p} = .31$

$n = 1000$

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

```
SE <-  sqrt( (.25 * (1 - .25))/1000)
z_val  <-  (.31 - .25)/SE
```

z_val is 4.38

Do more that 25% of US adults believe in ghosts?

Step 1:

H_0: $\pi = .25$

H_A: $\pi > .25$

Step 2:

z_val <- 4.38

Step 3-4:

p-value = pnorm(z_val, 0, 1, lower.tail = FALSE)

Step 5:

Indeed, very strong evidence!

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
   - pnorm(z, 0, 1, lower.tail = FALSE) = .037

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

5. What do you conclude?