Parametric inference on proportions

Overview

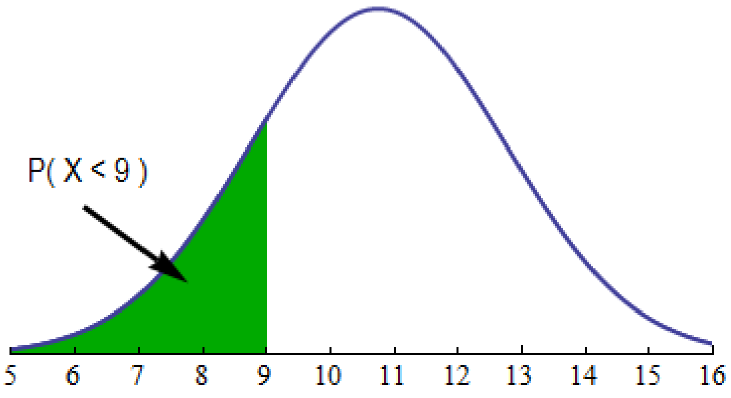
Review and continuation of using normal distributions for inference

Parametric inference on proportions

- Normal sampling distributions for proportions and formulas for the standard error
- Parametric confidence intervals for proportions
- Parametric hypothesis tests for proportions

Review and continuation of using normal distributions for inference

Review: Normal probability functions



Generate random data

- rnorm(m, mean, sd)

Plot the density curve

- dnorm(x_vec, mean, sd)

Get the probability that we would get a random value less than x: P(X < x)

- pnorm(x_vec, mean, sd)

Get the quantile value for a given proportion of the distribution

- qnorm(area, mean, sd)

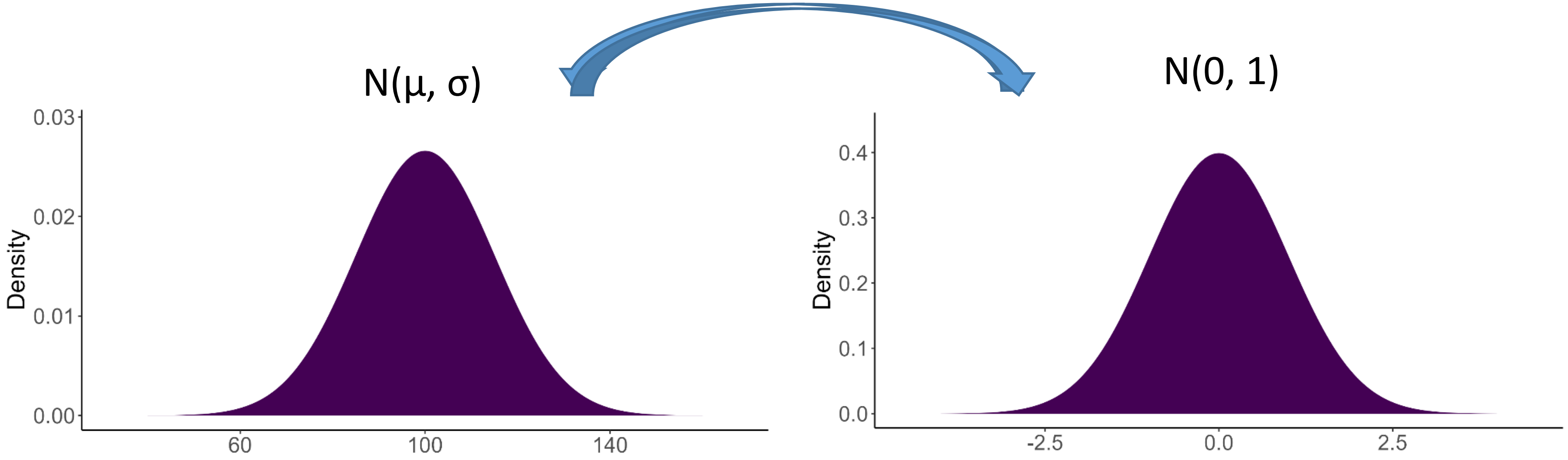
Review: Converting to the standard normal distribution

We can apply a z-score transformation to any normally distributed random variable $X \sim N(\mu, \sigma)$ to convert it to the standard normal distribution $Z \sim N(0, 1)$:

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

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

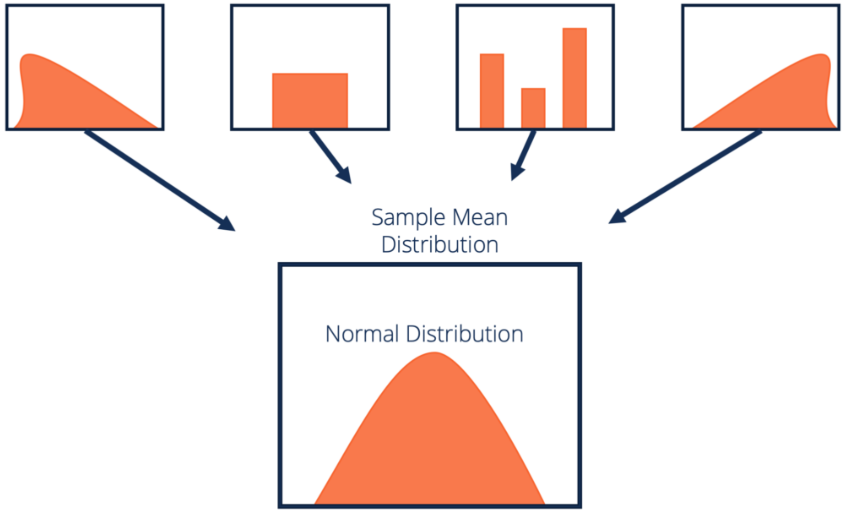
$$X = \mu + Z \cdot \sigma$$



Review: Central limit theorem

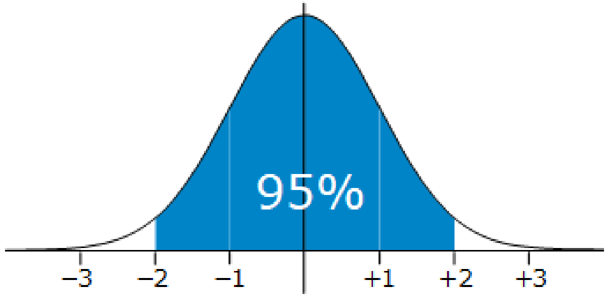
For random samples with a sufficiently large sample size (n), the distribution of sample statistics for a mean ($\bar{x}$) or a proportion ($\hat{p}$) is:

- Normally distributed
- Centered at the value of the population parameter



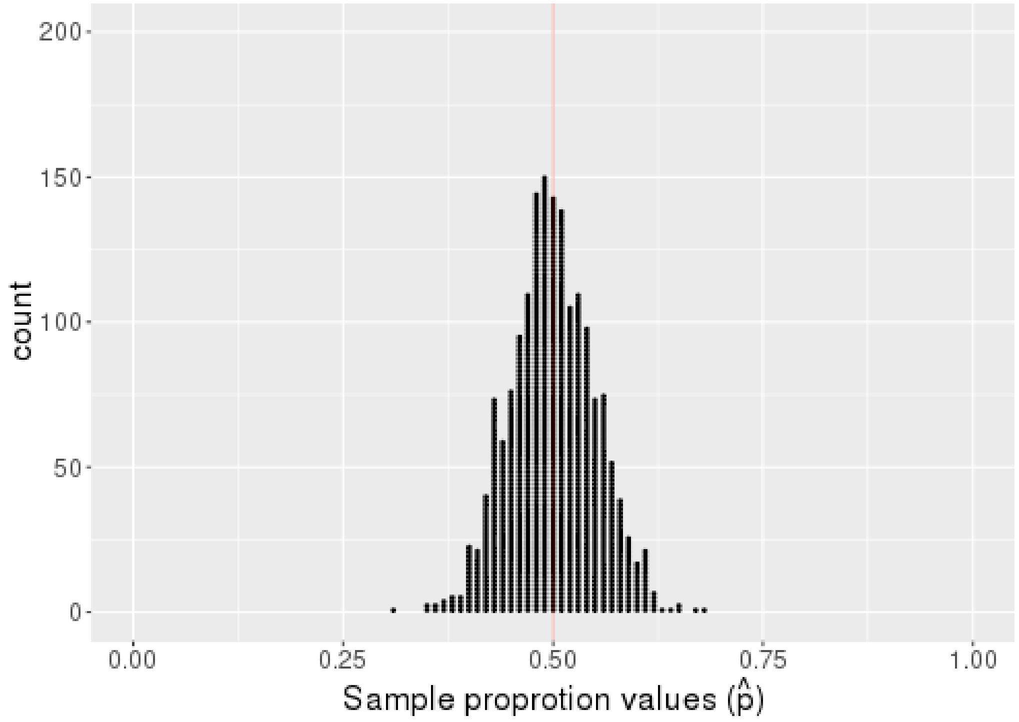
Upshot: We can create confidence intervals and run hypothesis tests using the normal distribution

- Rather than using computational methods like the bootstrap or a randomization methods to create a null distribution!

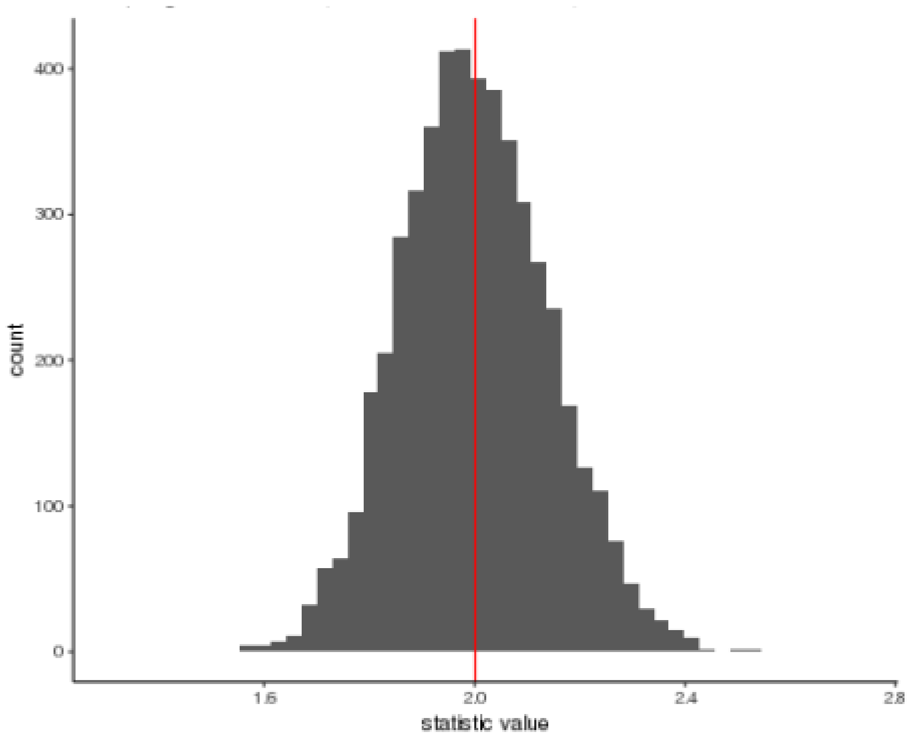


Review: Central limit theorem

proportion ($\hat{p}$)



mean ($\bar{x}$)



Proportion sampling distribution app

Sampling/Bootstrap distribution app

The plan

For large n, the sampling distributions of $\bar{x}$ and $\hat{p}$ have normal distributions

We can convert any normal distribution $N(\mu, \sigma)$, into a standard normal distribution $N(0, 1)$

We can then use the standard normal distribution for inference

- I.e., To create confidence intervals and run hypothesis tests

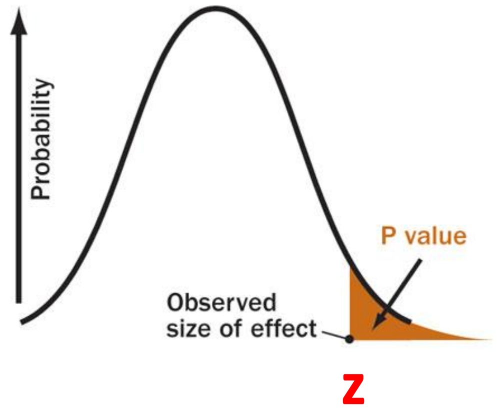
Hypothesis tests and confidence intervals using a normal distribution

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



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

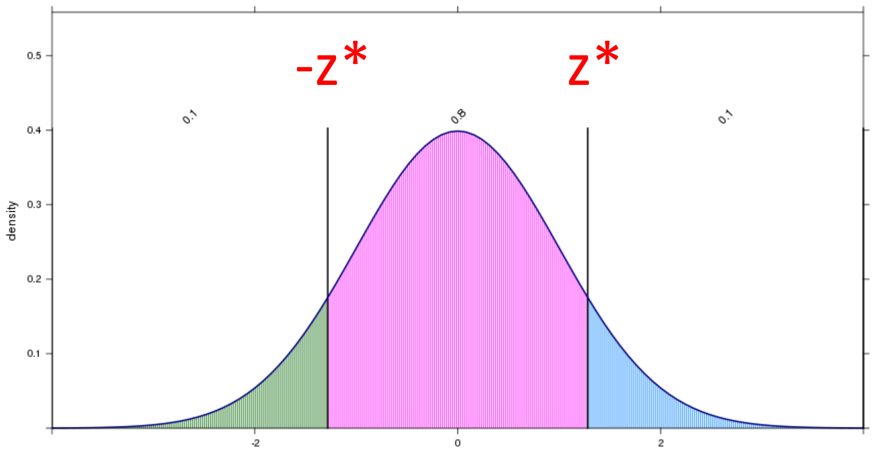
pnorm(z, 0, 1, lower.tail = FALSE)

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)
```

```
mosaic::cnorm()
```

Do goalies guess the direction of a penalty shot less than 50% of the time?

1. Start by stating H_0 and H_A

H_0: $\pi = .5$

H_A: $\pi < .5$

2. Calculate the observed statistic

- Goal keepers correctly guessed the direction 41% of the time out of 128 kicks
- With SE* = 0.043
  - (could you calculate this SE*?)

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

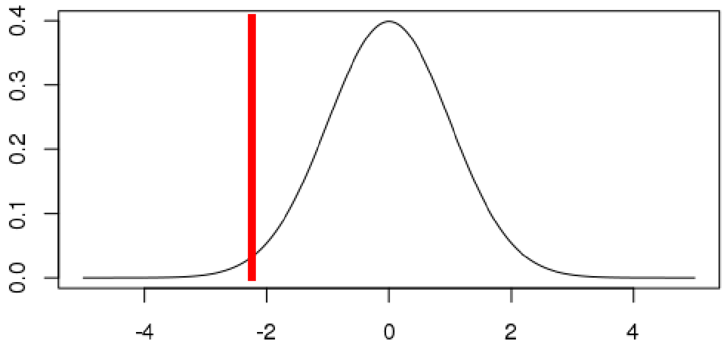
Can you compute a z-statistic?

Let's try the rest in R!

Do goalies guess the direction of a penalty shot less than 50% of the time?

Steps: 3-4. What is the probability one would get a z-statistic as small or smaller than -2.093 from a standard normal distribution?

Standard normal null distribution



pnorm(-2.093, 0, 1)

[Normal area app P(X ≤ x)]

p-value = 0.018

Step 5?

Parametric inference on proportions

Review: questions about proportions

Q_1: What symbols have we been using for the parameter and statistic for proportions?

- Parameter: π
- Statistic: $\hat{p}$

Q_2: What are examples of confidence intervals and hypotheses tests we've run for proportions?

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

Review: questions about proportions

Q_3: What does the shape of a sampling distribution for a proportions $\hat{p}$ look like?

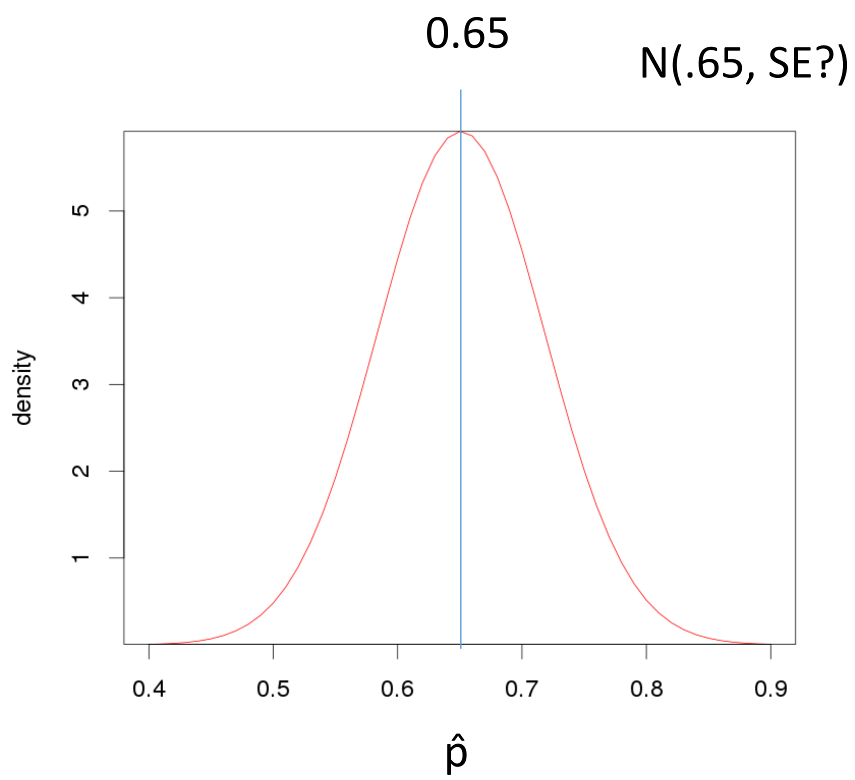
- A: normal!
  - (If the sample size n is larger enough)

Q_4: 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

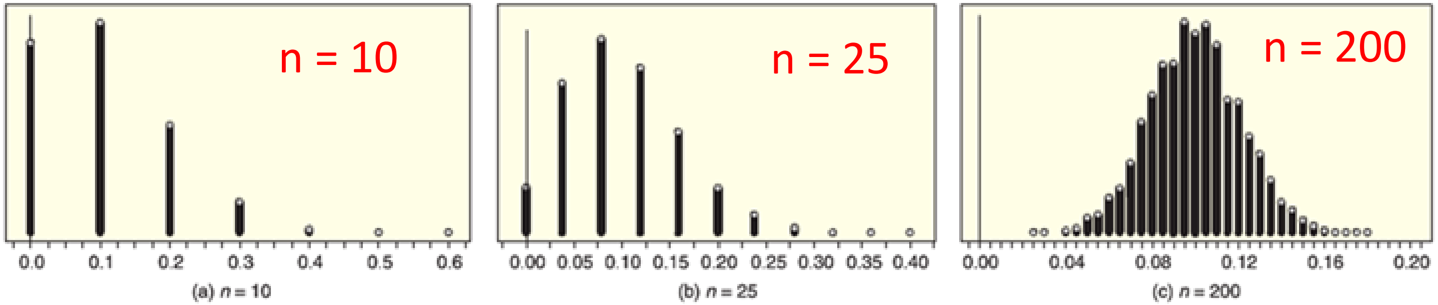
When choosing random samples of size n from a population with proportion π, the standard error (SE) of the sample proportions is given by:

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

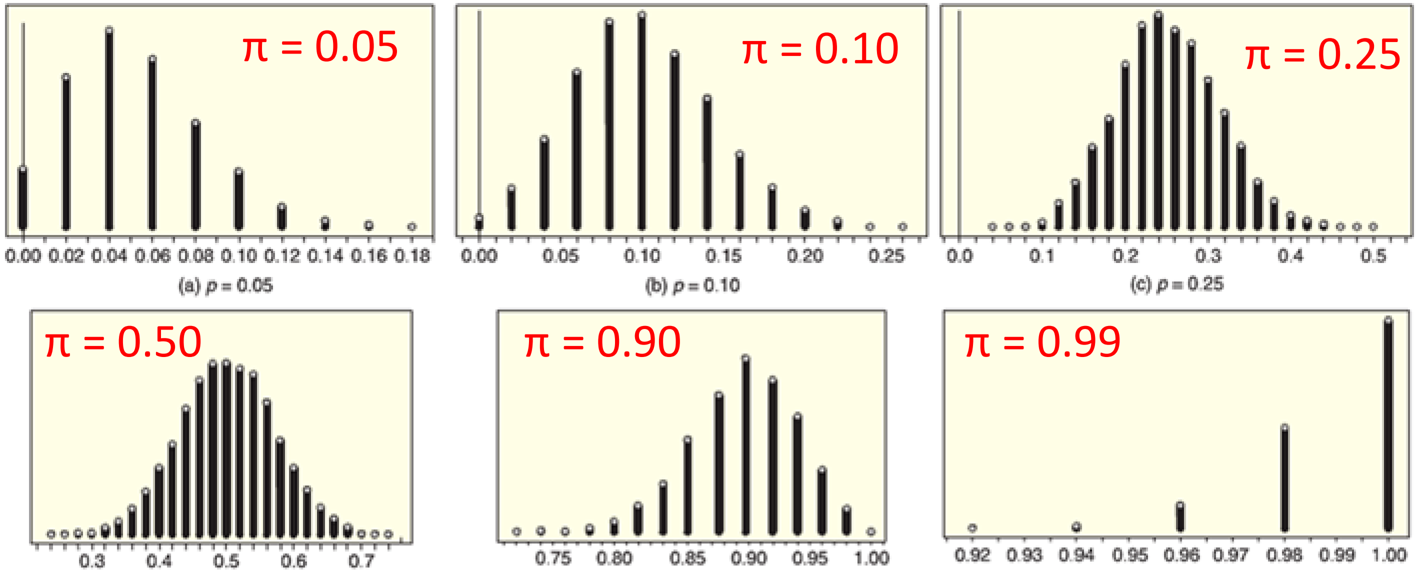
The larger the sample size (n) the smaller the standard error (SE)

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

Fixed $\pi = 0.10$
Changing n



Fixed $n = 50$
Changing π



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

The normal approximation is reasonably 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: $SE = \sqrt{\frac{\pi(1-\pi)}{n}}$

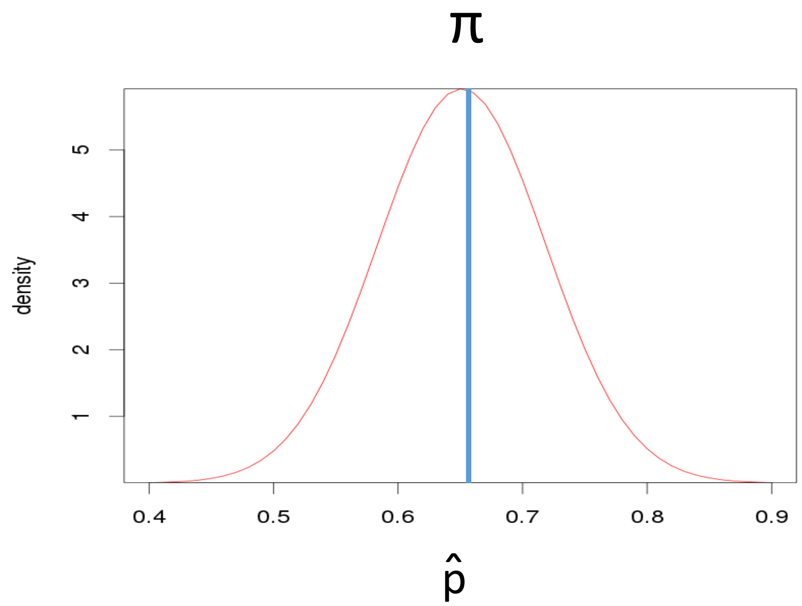
The larger the sample size, the more like a normal distribution it becomes.

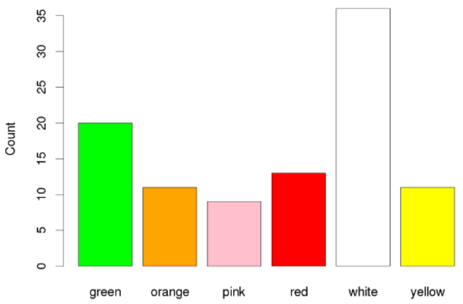
A normal distribution is a good approximation as long as:

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

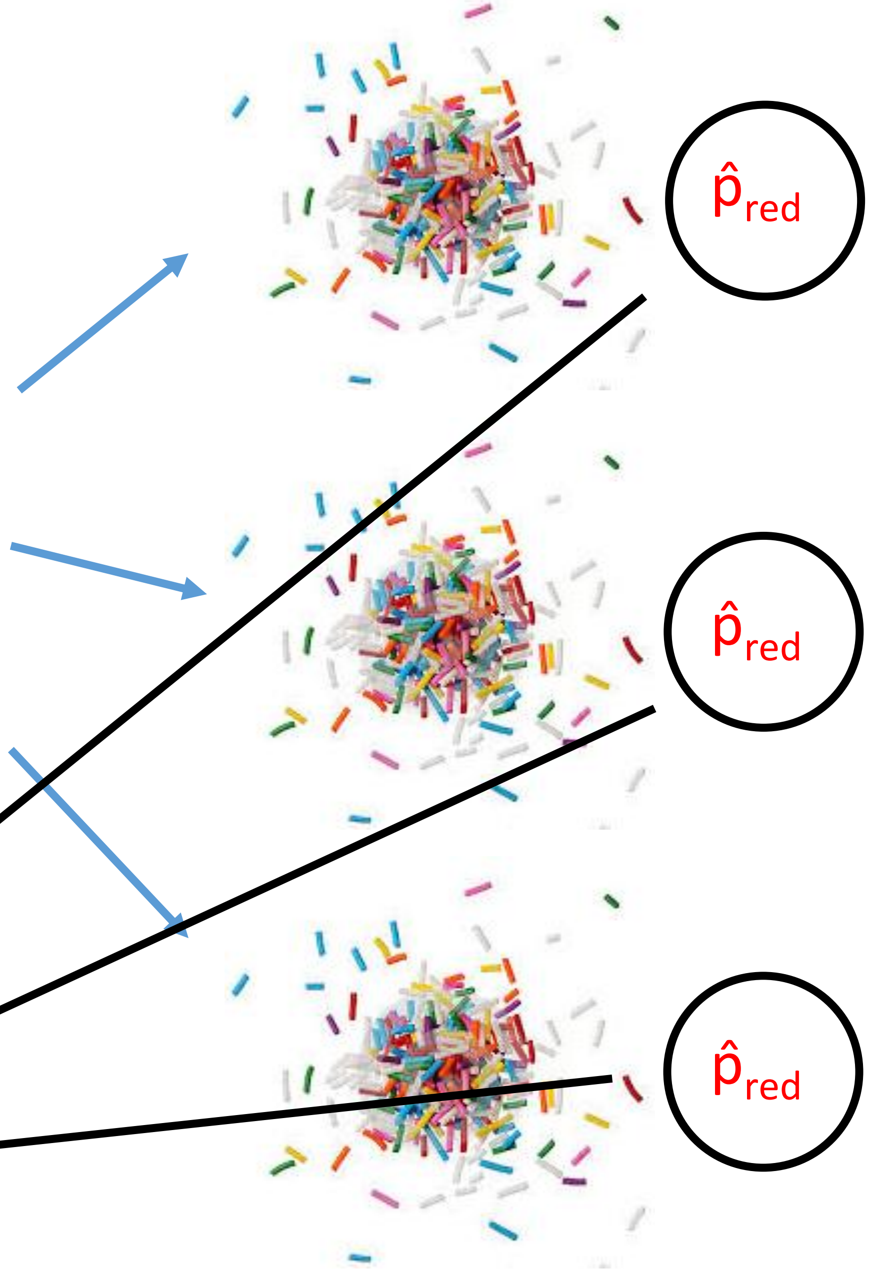
Summary: Central Limit Theorem for Sample Proportions

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

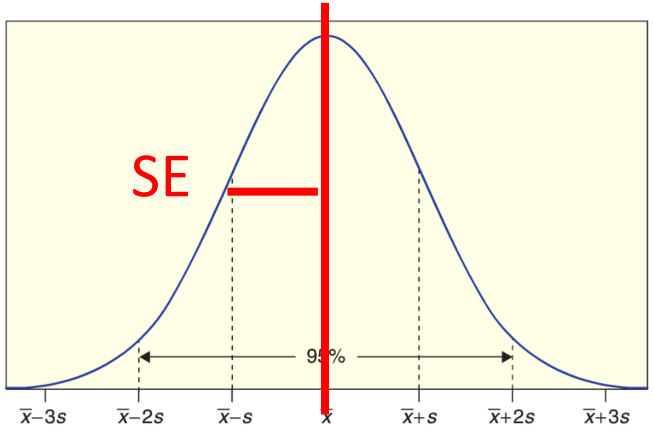




π_{red}



π_{red}



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

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

Sampling distribution!

SE for percentage of houses owned

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

If we randomly selected 50 houses…

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

b) What would this sampling distribution look like?

What if we randomly selected 200 houses?

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

Let's try it in R!

SE for percentage of houses owned

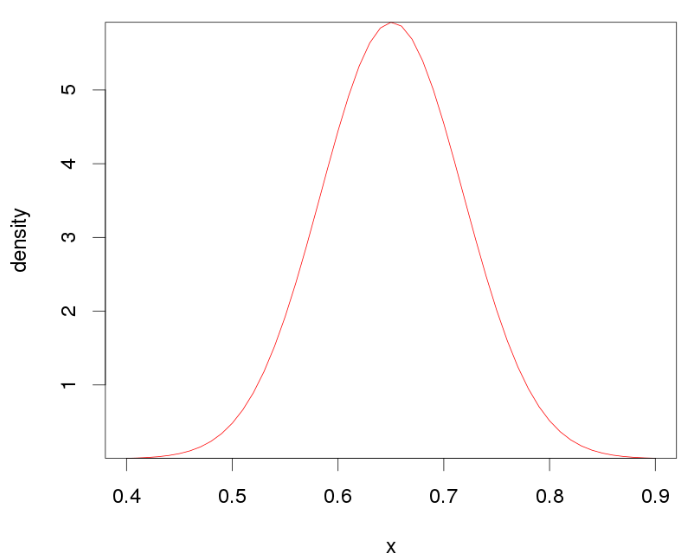
65.1% of all houses are owned

- π = .651
- When n = 50: SE = .0674
- When n = 200: SE = .0337

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

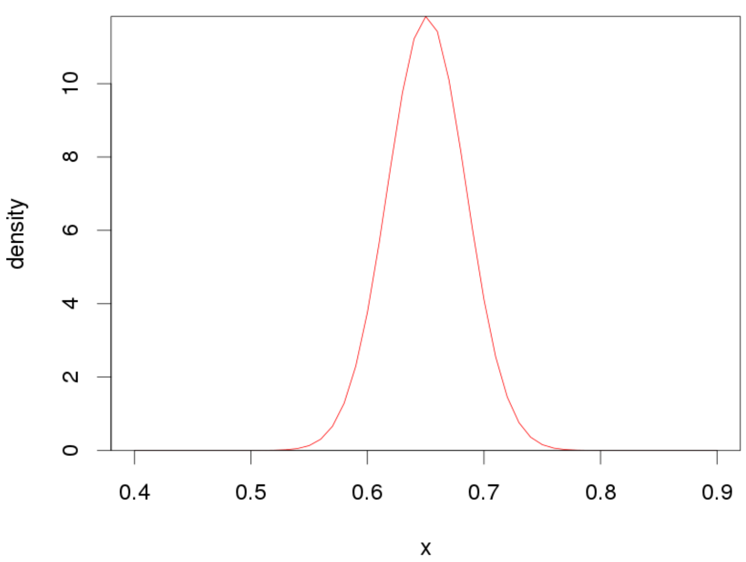
N(.651, .0671)

n = 50



n = 200

N(.651, .0337)




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

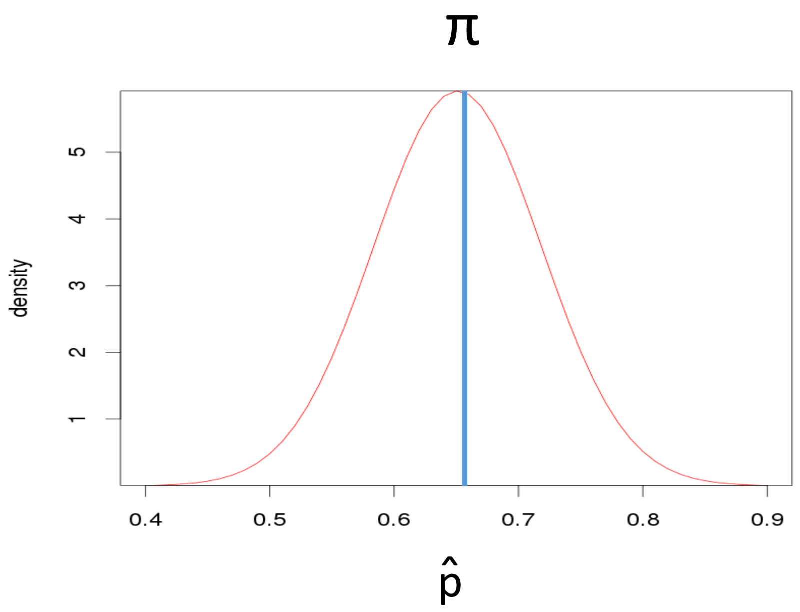
Parametric inference on proportions continued

Summary: Central Limit Theorem for Sample Proportions

We just showed that the sampling distribution of proportions p̂ is normal distributed with a standard error of:

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

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



What would be a problem with using this formula for the SE for inference?

- I.e., what is the problem using this formula for confidence intervals and hypothesis tests?

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

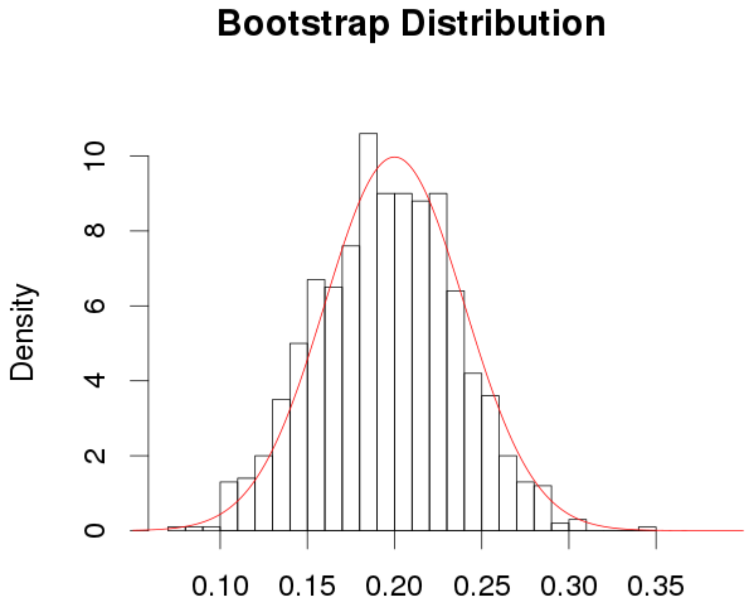
For my green sprinkles I got:

- 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

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

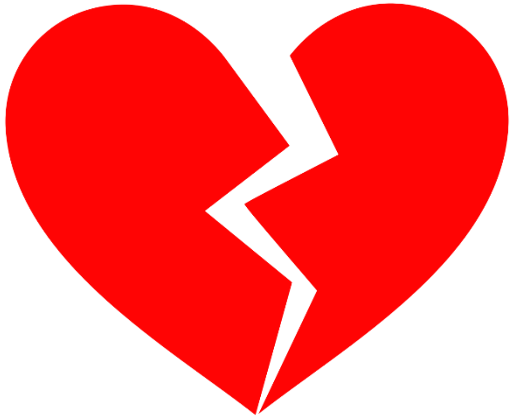
$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Let's try it in R!

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]
```



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: $\pi = \pi_0$

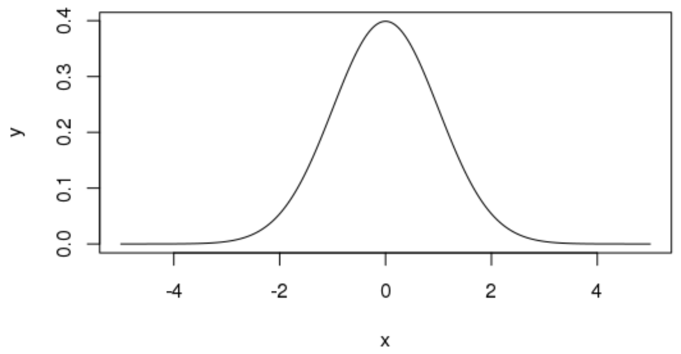
Our formula for z then becomes:

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

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

5. What do you conclude?

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

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

Step 1:

H_0: π = .25

H_A: π > .25

Step 2:

p̂ = .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!