Parametric inference on proportions

Overview

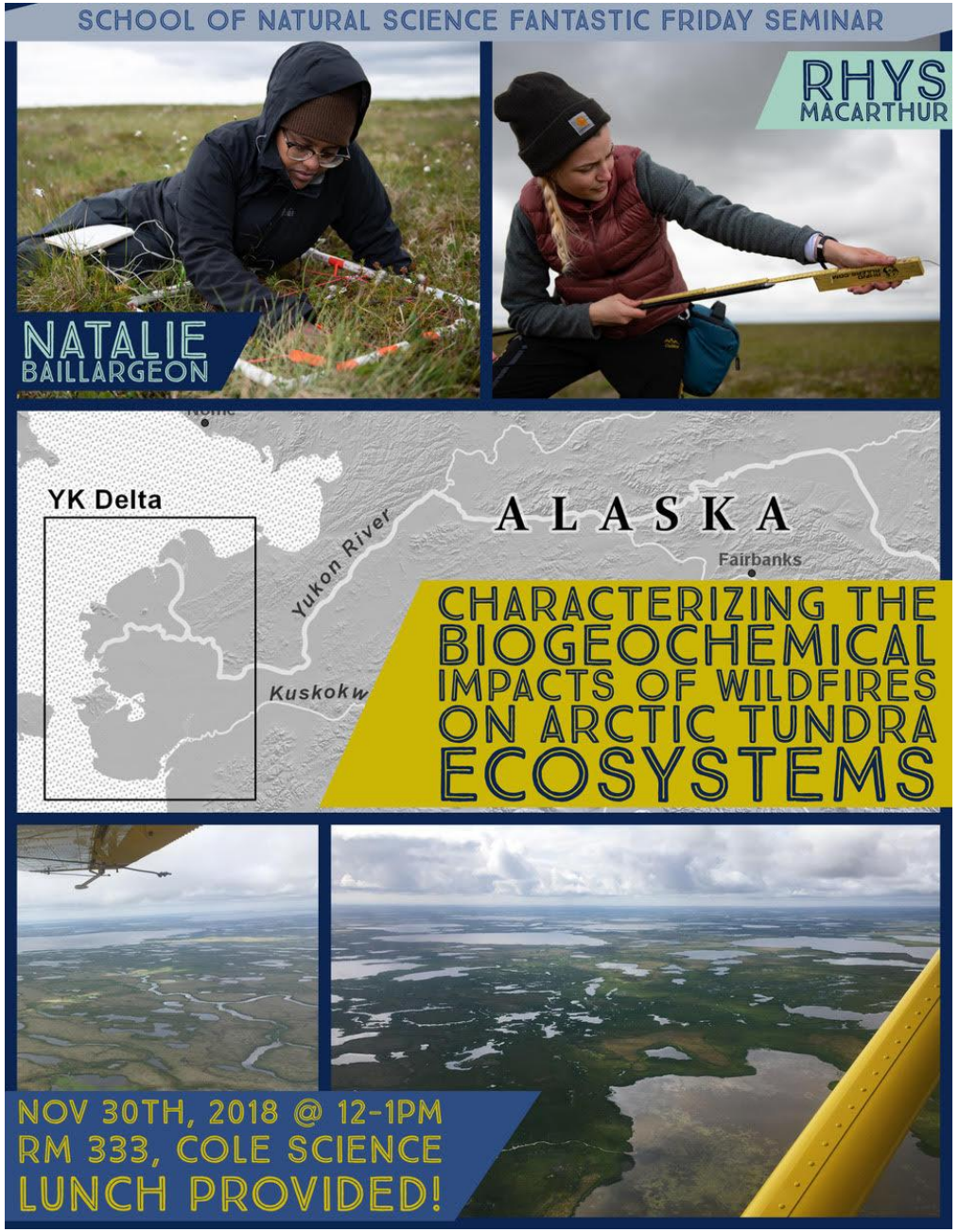
Quick review of Normal distributions

Calculating confidence intervals using normal distributions

Parametric inference on proportions

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

Announcement: NS fantastic Friday talk

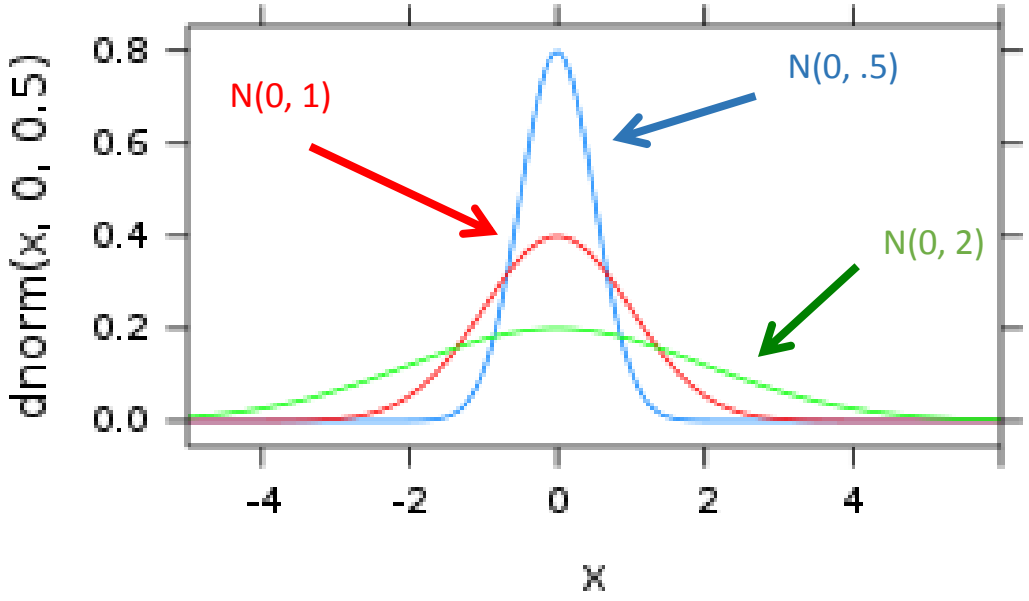


Review of Normal distributions

Normal Density Curve

Normal distributions N(μ, σ) have two parameters:

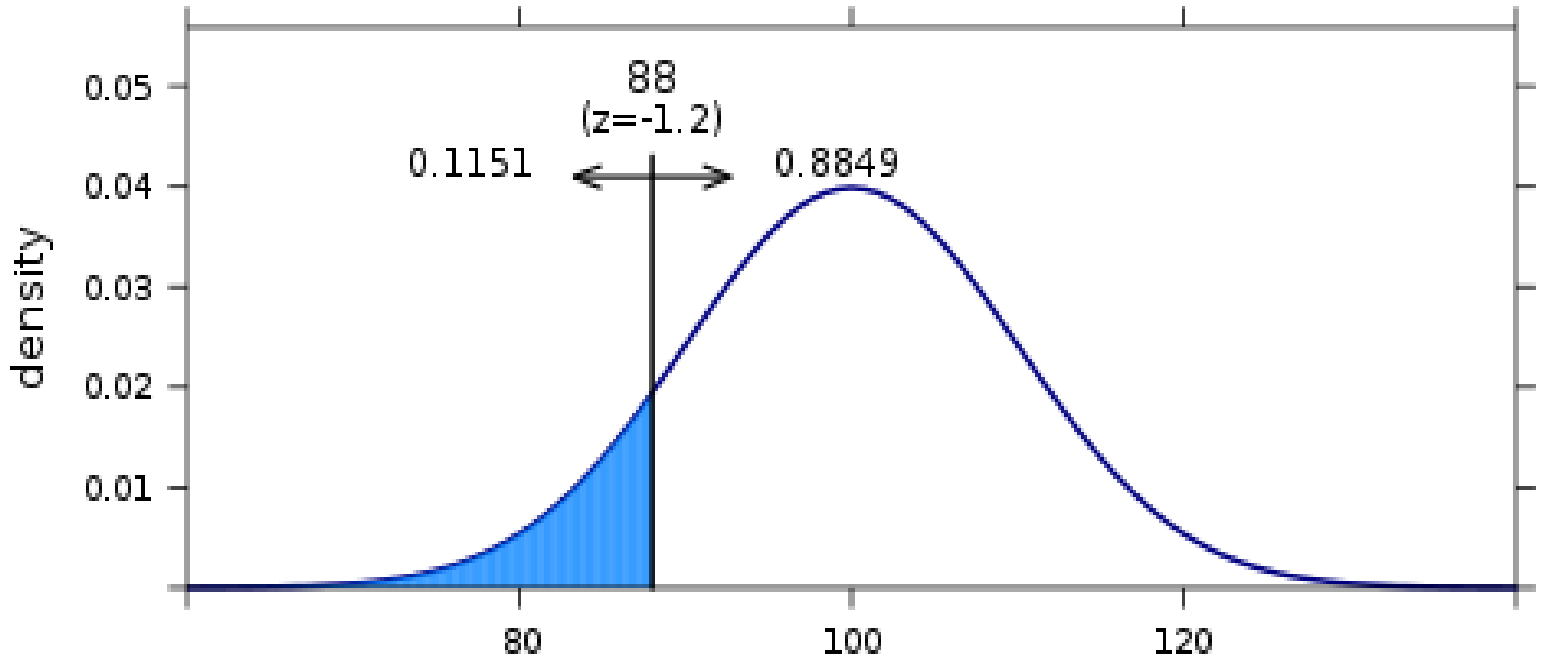
- The mean: μ
- The standard deviation: σ



Calculating probabilities from normal distributions

We can find the probability of getting a value less than or equal to x:

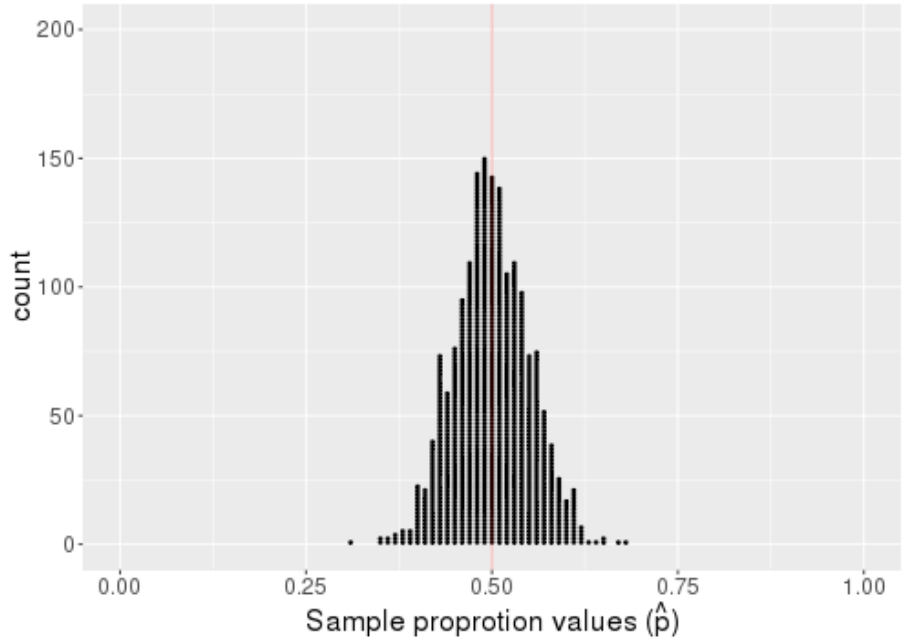
pnorm(x, mu, sigma) # Pr(X ≤ x)



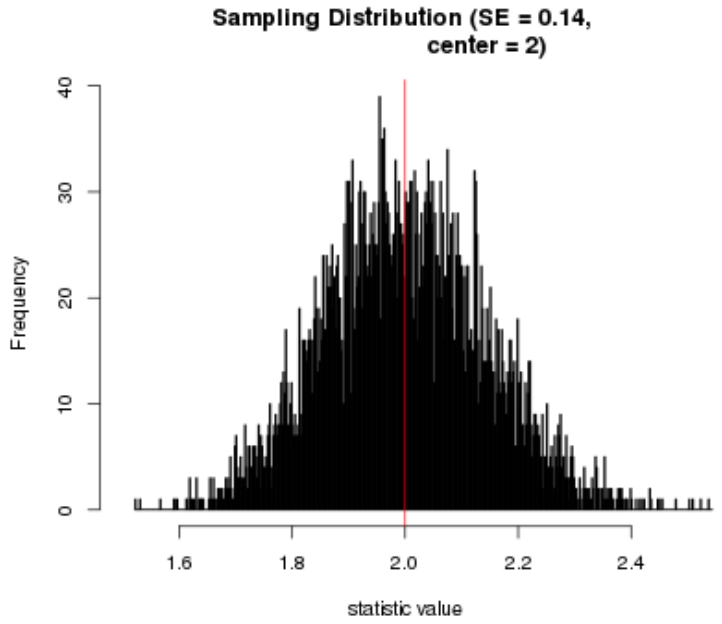
Central limit theorem

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** and is centered at the value of the population parameter

proportion ($\hat{p}$)



mean ($\bar{x}$)



Central limit theorem

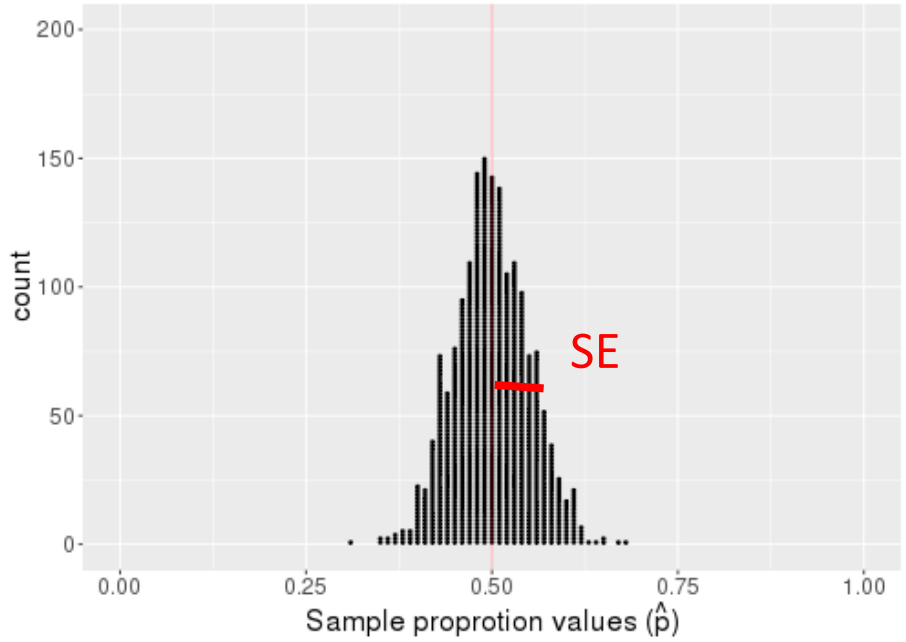
Questions:

1. What is the standard deviation of these sampling distributions called?
2. Suppose we have a $\hat{p}$ or $\bar{x}$ and know the SE, how can we create a 95% CI?

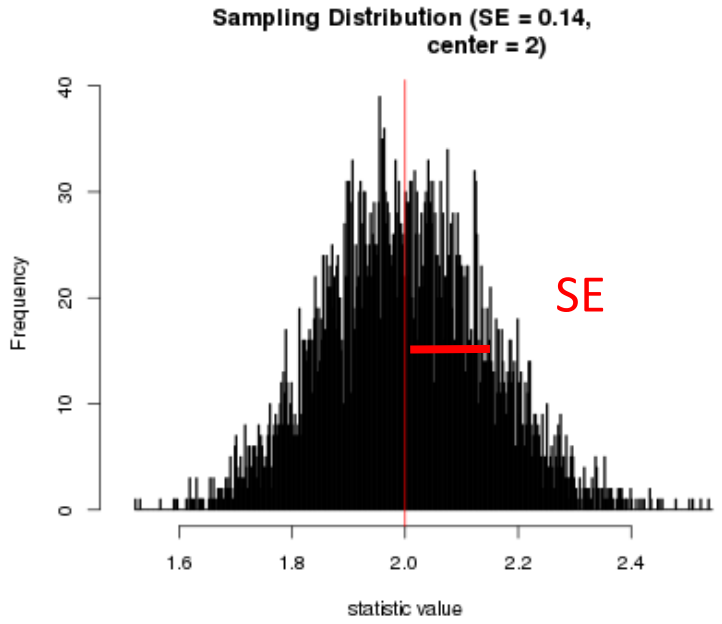
For a proportion π: $CI_{95} = \hat{p} \pm 2 \cdot SE$

For a mean μ: $CI_{95} = \bar{x} \pm 2 \cdot SE$

proportion ($\hat{p}$)

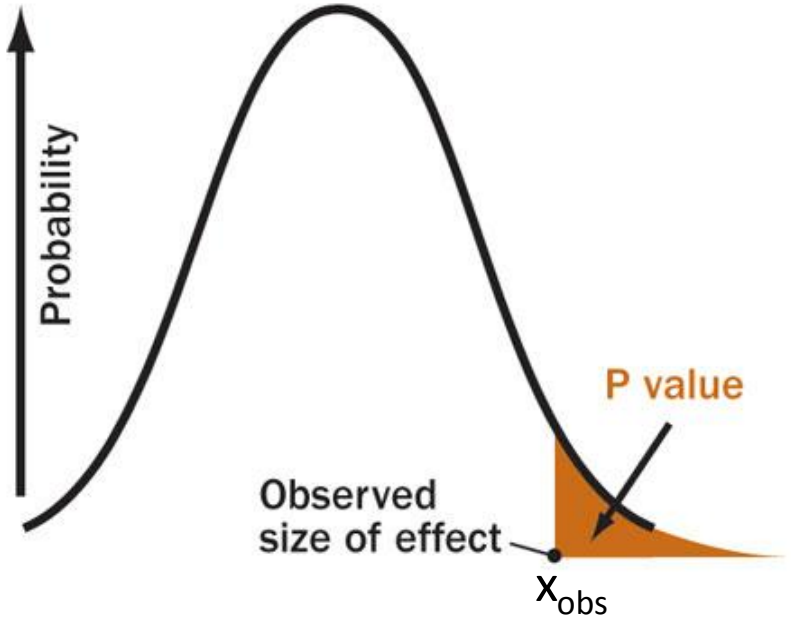


mean ($\bar{x}$)



p-values based on a normal distribution

When a distribution under the **null hypothesis** is normal, we can use the normal density curve to compute p-values rather than a randomization distribution



$$Pr(X \geq x_{obs};\ \mu,\ \sigma)$$

```
pnorm(obs_stat, mu, sigma,
      lower.tail = FALSE)
```

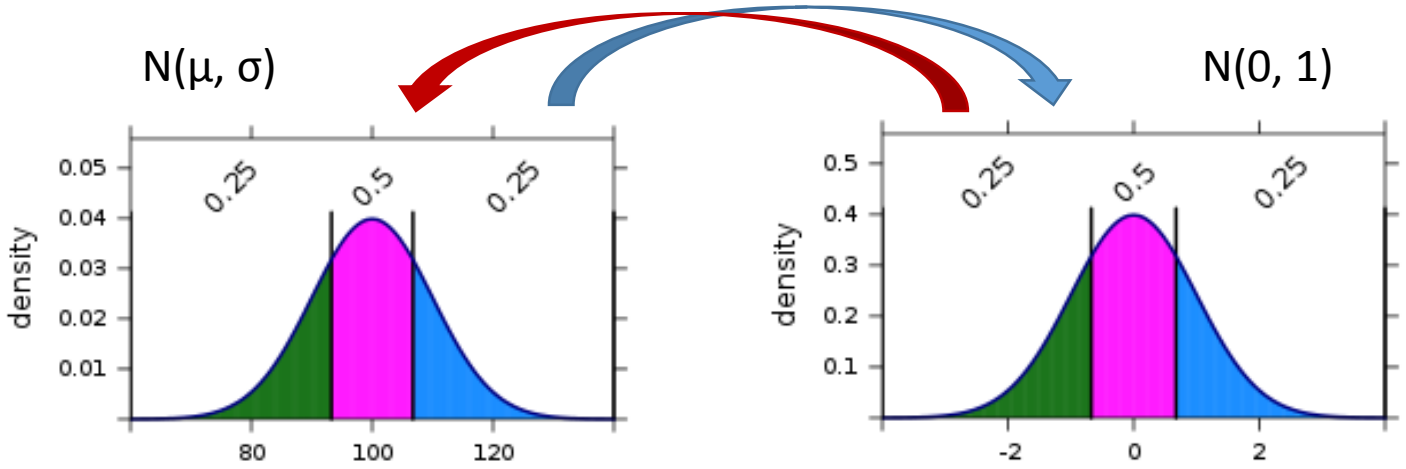
Standard Normal N(0, 1)

We can scale any normal distribution value $X \sim N(\mu, \sigma)$ to a **standard normal** distribution value $Z \sim N(0, 1)$ using:

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

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

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



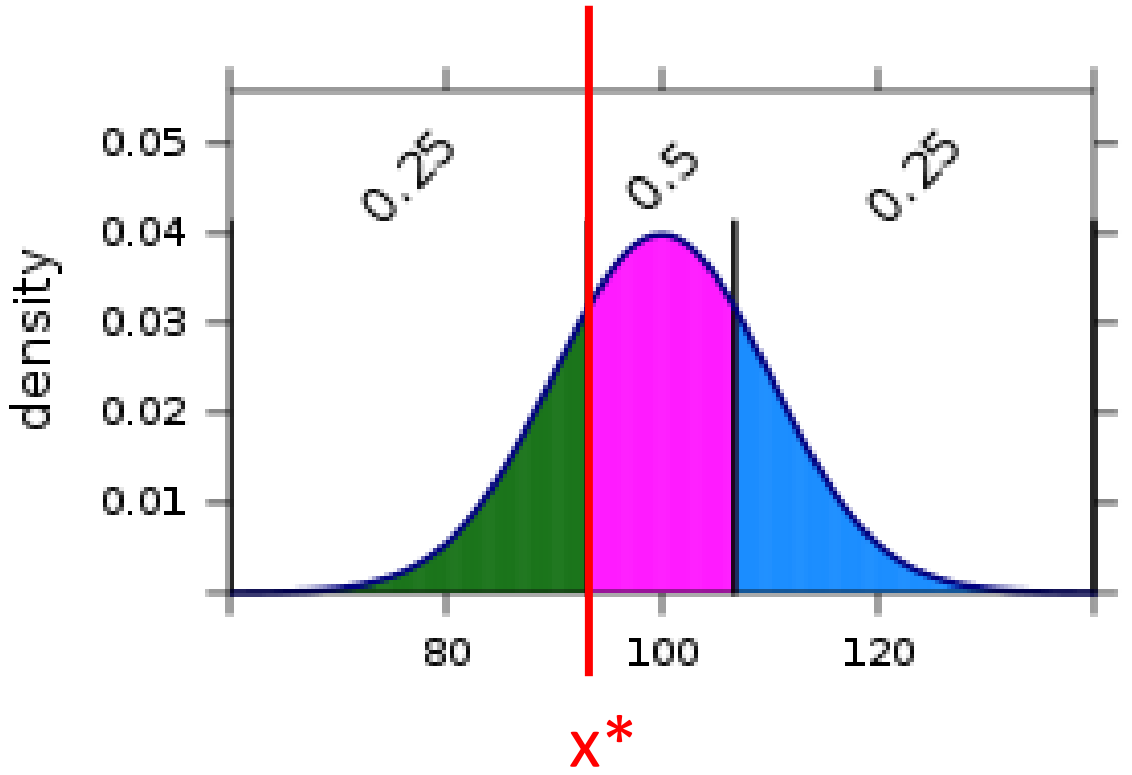
Finding normal probabilities and percentiles

We can find the quantile value from a normal distribution with:

qnorm(q, mu, sigma)

The ‘q’ in qnorm stands for quantile

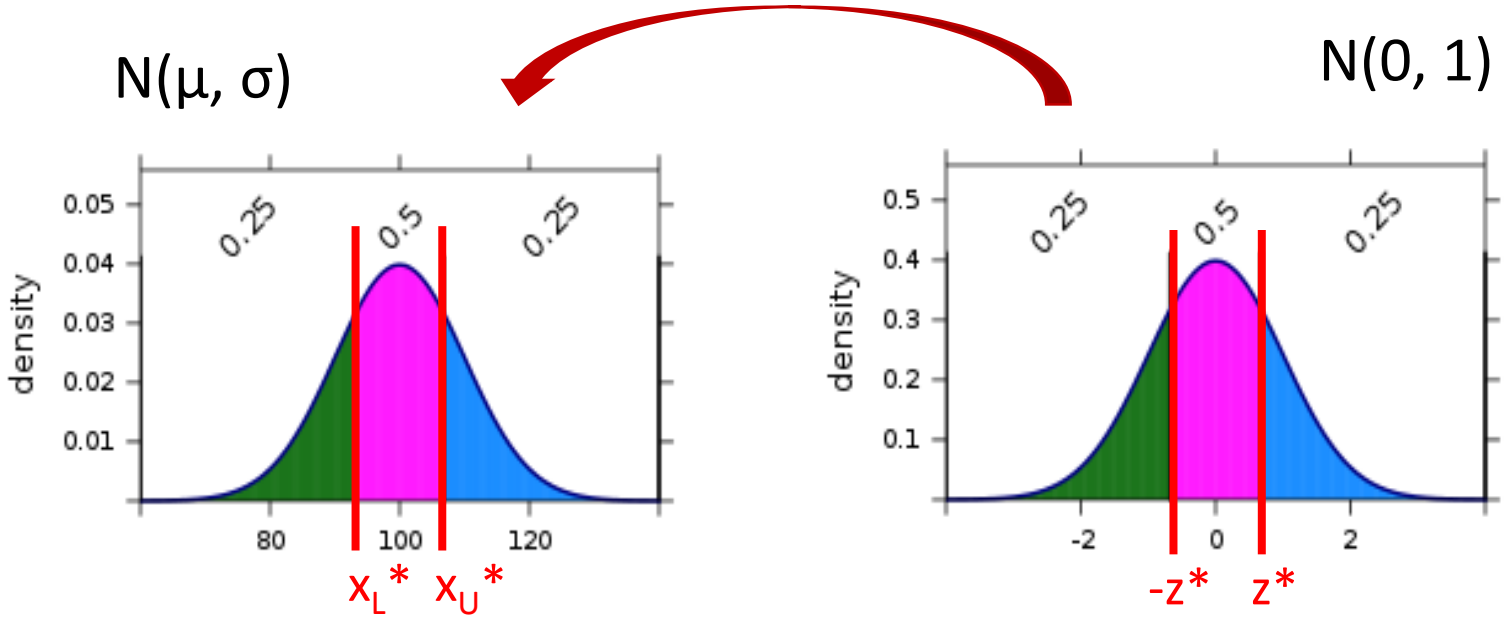
What is the max and min that q can be?



Standard Normal N(0, 1)

It is often convenient to find quantiles on the standard normal distribution $Z \sim N(0, 1)$ and then to transform them to an arbitrary normal distribution $X \sim N(\mu, \sigma)$, using :

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



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:

$$\text{sample statistic} \pm z^* \times \text{SE}$$

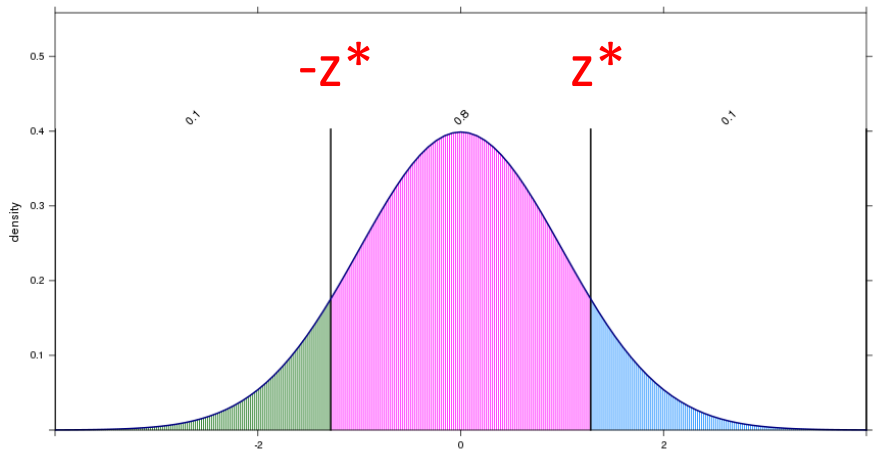
where z* is chosen so that the area between –z* and + z* in the standard normal distribution is the desired confidence level

- i.e., z* is chosen such that say 95% of the distribution is between ± z*

Confidence intervals based on a Normal Distribution

Suppose we are interested in 80% confidence intervals for μ

We calculate the $\pm z_{80}$ that has 80% of the data on N(0, 1)



Let's assume we know the SE but don't know μ. If we have an observed statistic from:

$$x_{obs} \sim N(\mu, SE)$$

We can create an interval that will capture μ 80% of the time using:

$$x_{obs} \pm z_{80} \cdot SE$$

Normal percentiles for common confidence levels

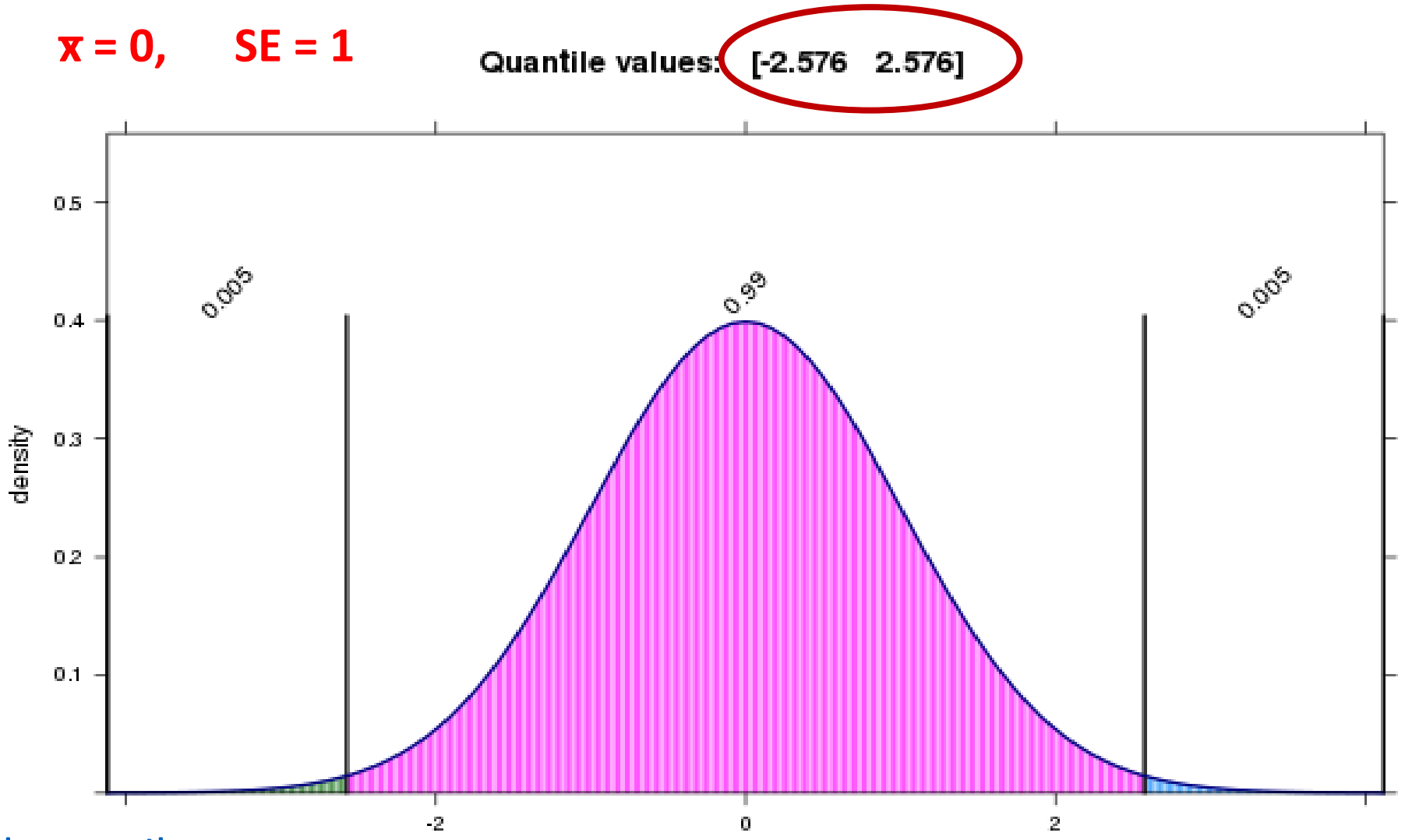
Confidence level	80%	90%	95%	98%	99%
Z*	1.282	1.645	1.960	2.326	2.576

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

Normal quantile app

.99 quantile values

$\bar{x} = 0$, SE = 1



Normal quantile app

What is the most preferred seat?

A survey of 1,000 air travelers found that 60% prefer a window seat, with a bootstrap standard error of SE = 0.015

Use the normal distribution to compute a 90%, 95% and 99% CIs for the proportion of people who prefer a window seat

$$\text{sample statistic} \pm z^* \times \text{SE}$$

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

What is the most preferred seat?

A survey of 1,000 air travelers found that 60% prefer a window seat, with a bootstrap standard error of SE = 0.015.

90% CI = $.6 \pm 1.645 \times .015$ = [.575 .625]

95% CI = $.6 \pm 1.96 \times .015$ = [.571 .629]

99% CI = $.6 \pm 2.576 \times .015$ = [.569 .638]

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

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

Parametric inference on proportions

Review: questions about proportions

1. What symbols have we been using for the parameter and statistic for proportions?
    - What are examples of confidence intervals and hypotheses tests we've run for proportions?

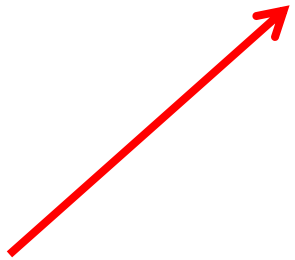
2. What does the shape of a sampling distribution for a proportion look like?

3. Suppose $\pi = .6$, and $n = 100$, can you draw the sampling distribution for $\hat{p}$?
    - If you were given the SE could you do it?

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)

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

Q: How have we calculated SE is previous classes?

A: The bootstrap!

How could we do this for the green sprinkles?

```
bootstrap_dist <- NULL
for (i in 1:10000)  {
    boot_sample <- sample(my_sprinkles, replace = TRUE)
    bootstrap_dist[i]  <-  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

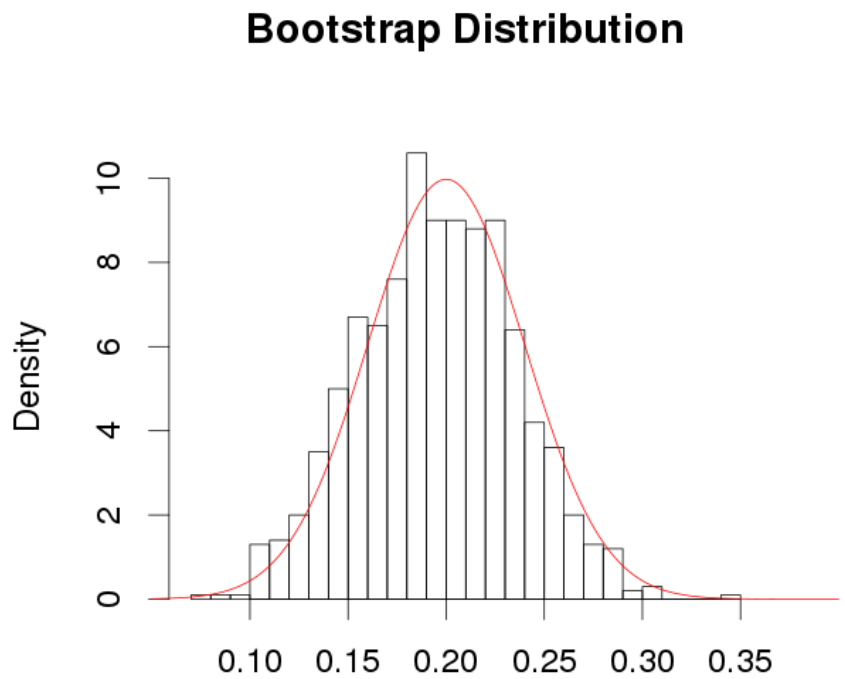
For my green sprinkles I get:

- 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)
```

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 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?

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

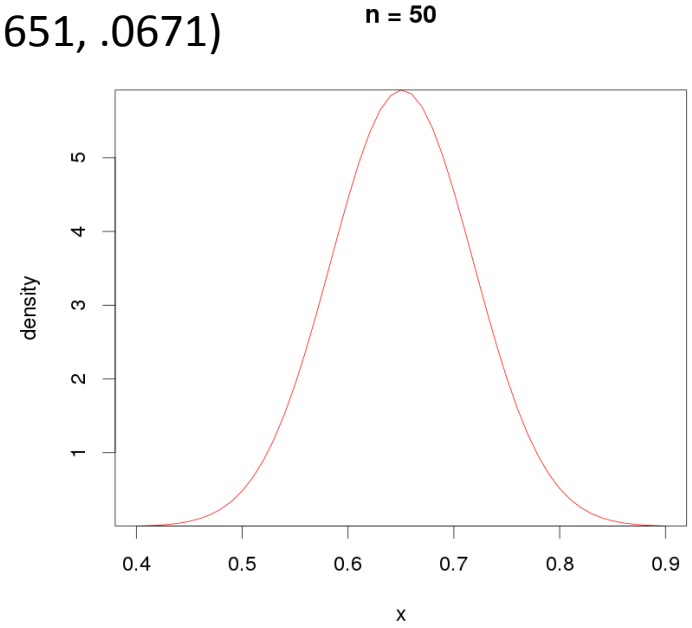
SE for percentage of houses owned

65.1% of all houses are owned

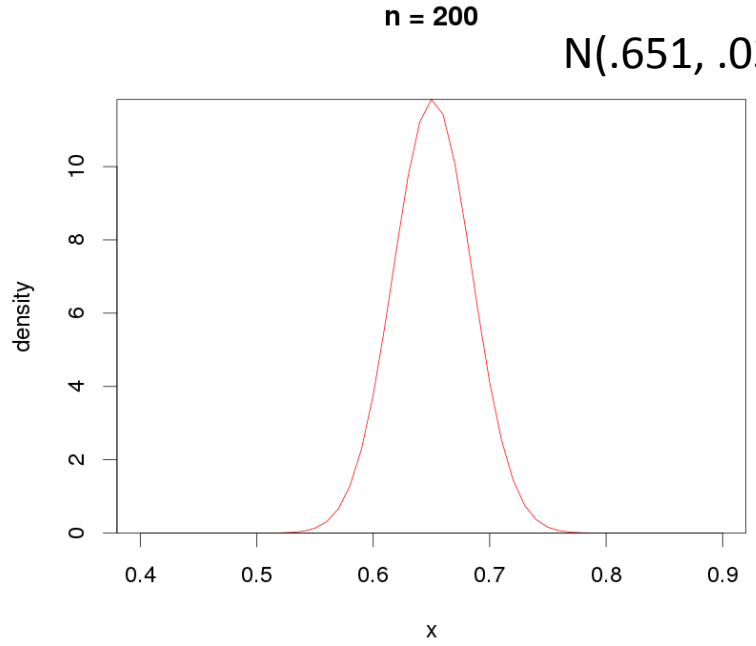
- π = .651
- When n = 50: SE = .0674
- When n = 200: SE = .0337

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

N(.651, .0671)



N(.651, .0337)



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

n = 50

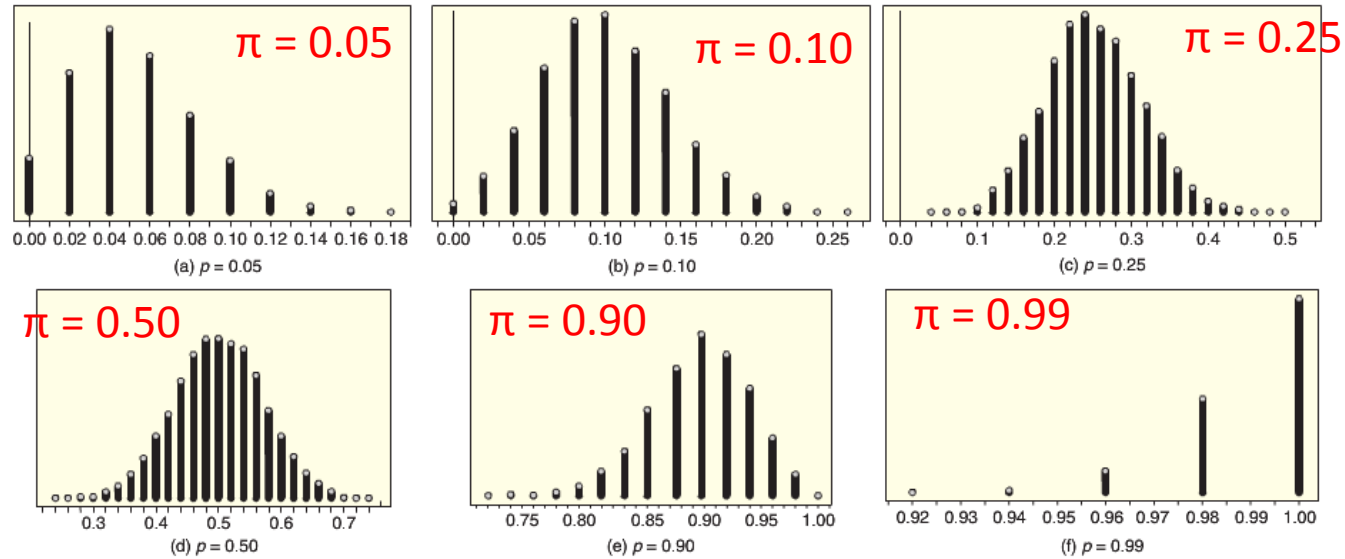


Figure 6.2 *Distributions of sample proportions when* $n = 50$

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

$$n\pi \geq 10 \qquad \text{and} \qquad n(1-\pi) \geq 10$$

Summary: Central Limit Theorem for Sample Proportions

When choosing random samples of size n from a population with a proportion p, 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.

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

We can write this as:

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

Confidence intervals for a single proportion

Provided the sample size is large enough so that $n\pi \geq 10$ and $n(1-\pi) \geq 10$, a confidence interval for a population proportion p can be computed based on a random sample of size n using:

Note we are substituting $\hat{p}$ for π

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

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

My green sprinkles

I counted 100 sprinkles, 20 of which were green

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

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

My green sprinkles

$\hat{p}$ = .20

n = 100

z* = 1.96

SE = .04

CI = 0.1216 to 0.2784

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

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

Test for single proportions

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

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

In the context of proportions we usually state $H_0: \pi = \pi_0$, and the formula for z 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 $\hat{p}$ 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?

H_0: $\pi = .25$ vs. H_A: $\pi > .25$

$\hat{p} = .31$

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?

H_0: $\pi = .25$ vs. H_A: $\pi > .25$

$\hat{p} = .31$

$n = 1000$

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

p-value = 1 - pnorm(z_val, 0, 1)

0.00000589

Indeed, very strong evidence!

Worksheet 11

Lock 5 questions on computing areas/quantiles of normal distributions and doing parametric inference on proportions

```
> source('/home/shared/intro_stats_2016/cs206_functions.R')
> get_worksheet(11)
```