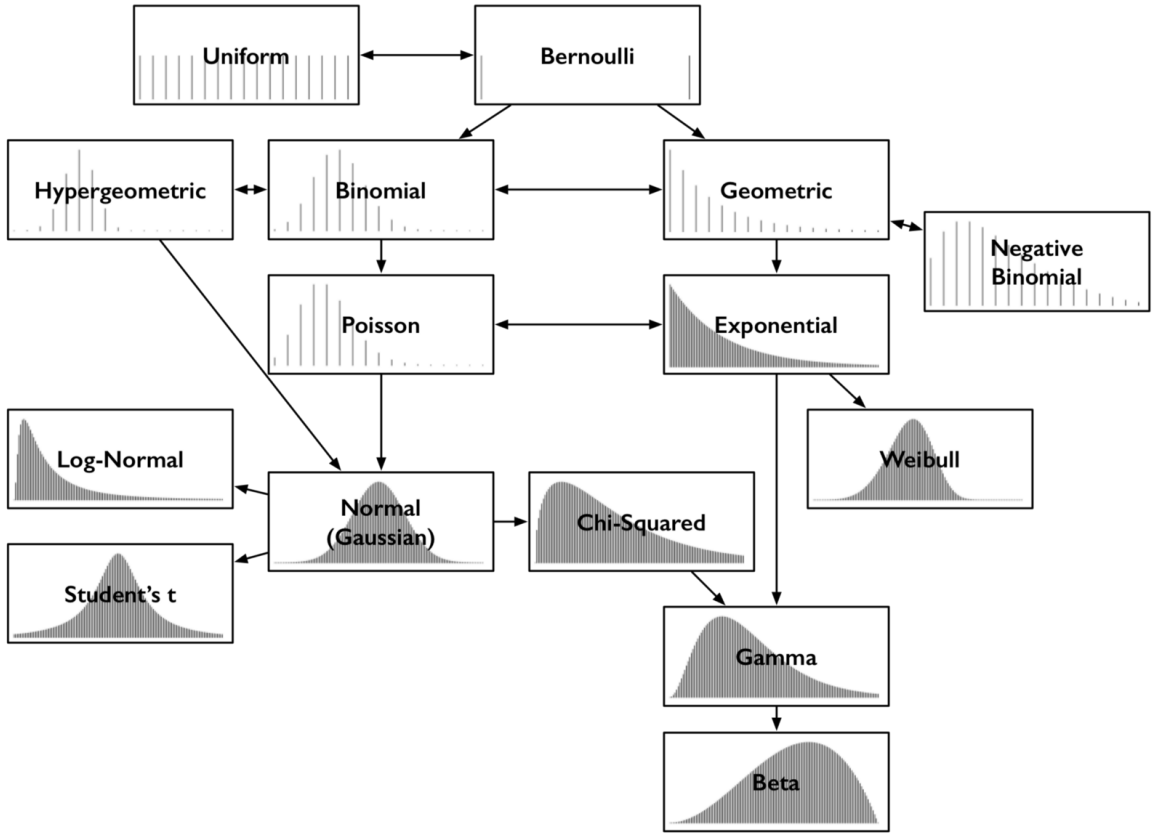
Probability distributions and parametric inference on proportions



Overview

Density curves

The Normal distribution

Confidence intervals and p-values based on the normal distribution

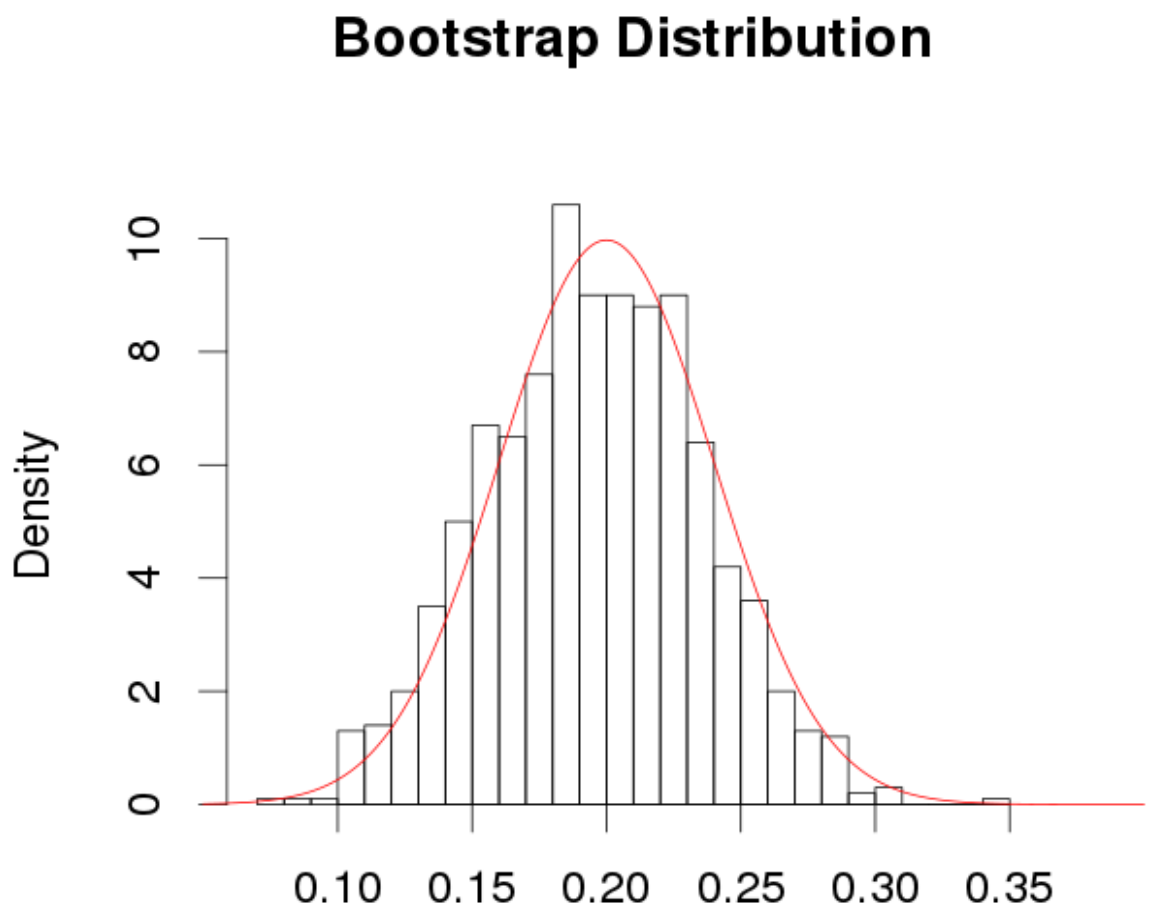
Inference using parametric probability distributions

In the past month we have learned to use computer simulations to create confidence intervals and run hypothesis tests

Now we will use mathematical functions called **probability distributions** to do inference

- e.g. instead of running computer simulations to create null distributions we can use probability distribution functions

Comparing a bootstrap distribution and a probability distribution



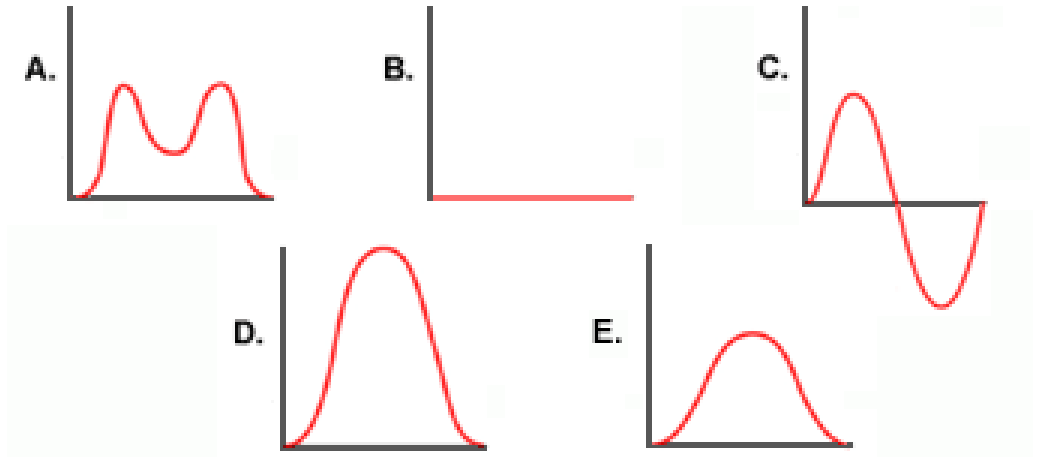
Density Curves

A **density curve** is a mathematical function f(x) that has two important properties:

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

2. The curve is always ≥ 0

Which of these could ***not*** be a density curve?



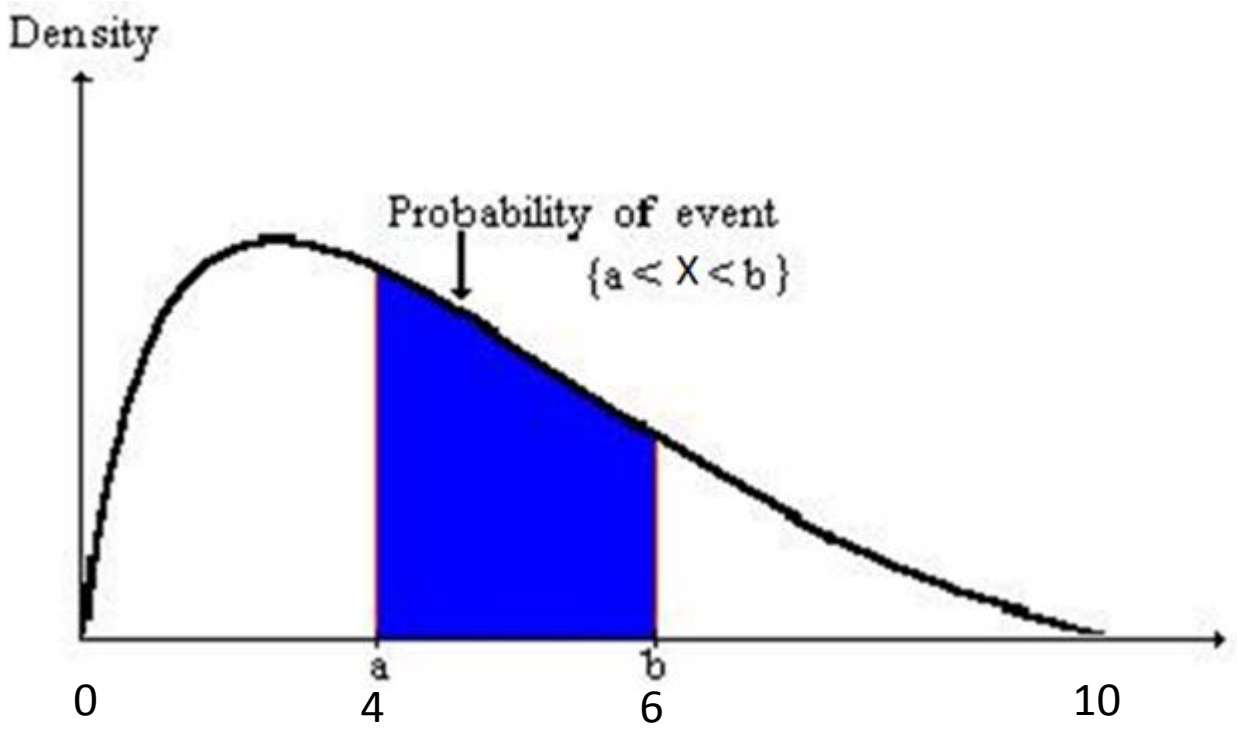
Density Curves

The area under the curve in an interval [a, b] models the probability that a random number X will be in the interval

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

$Pr(4 < X < 6) = .4$

For example…



Normal Density Curve

A normal distribution follows a bell-shaped curve

There are two parameters that characterize normal curves, which are:

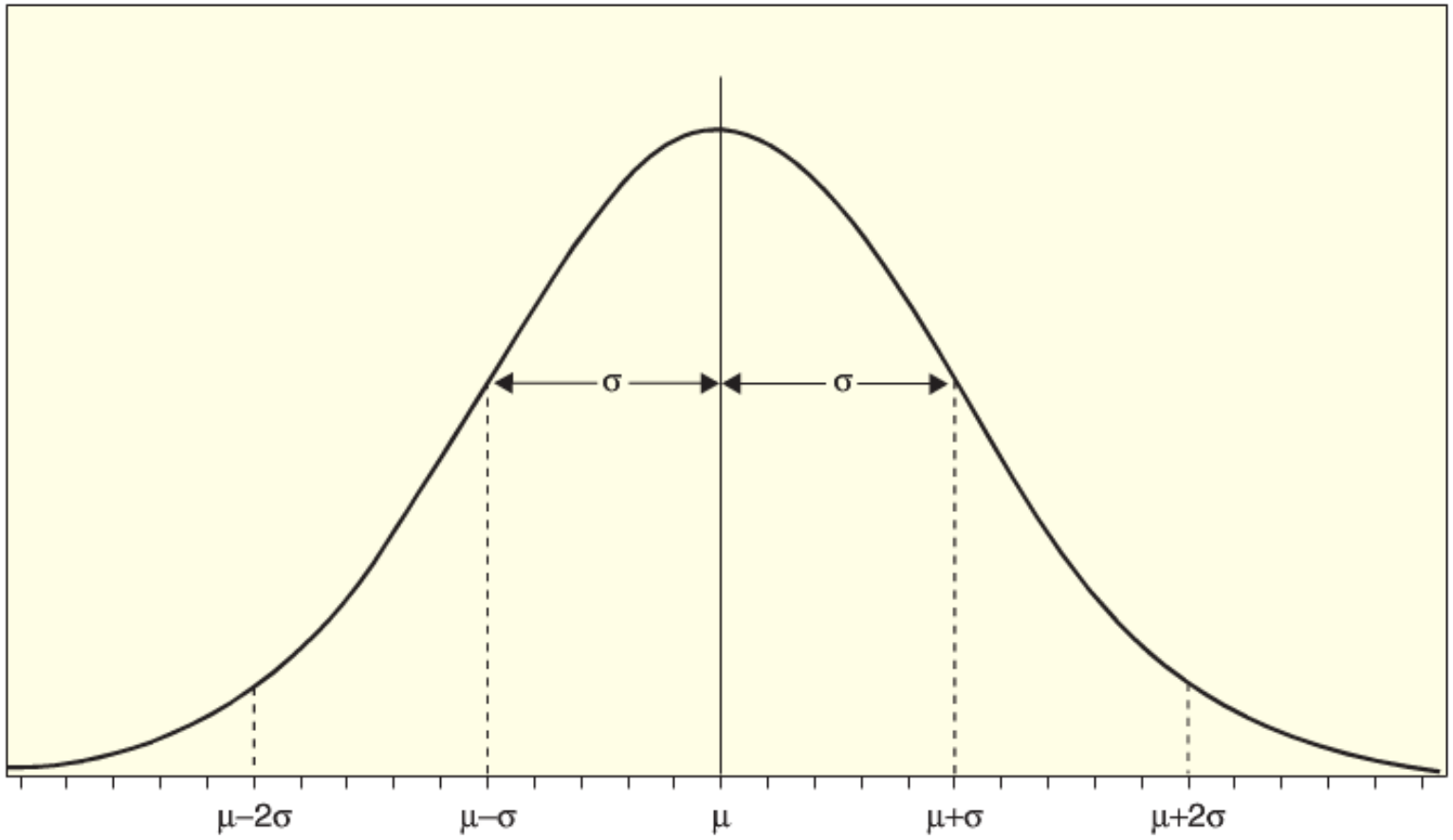
- The mean: μ
- The standard deviation: σ

We use μ and σ because this is often a model for the population

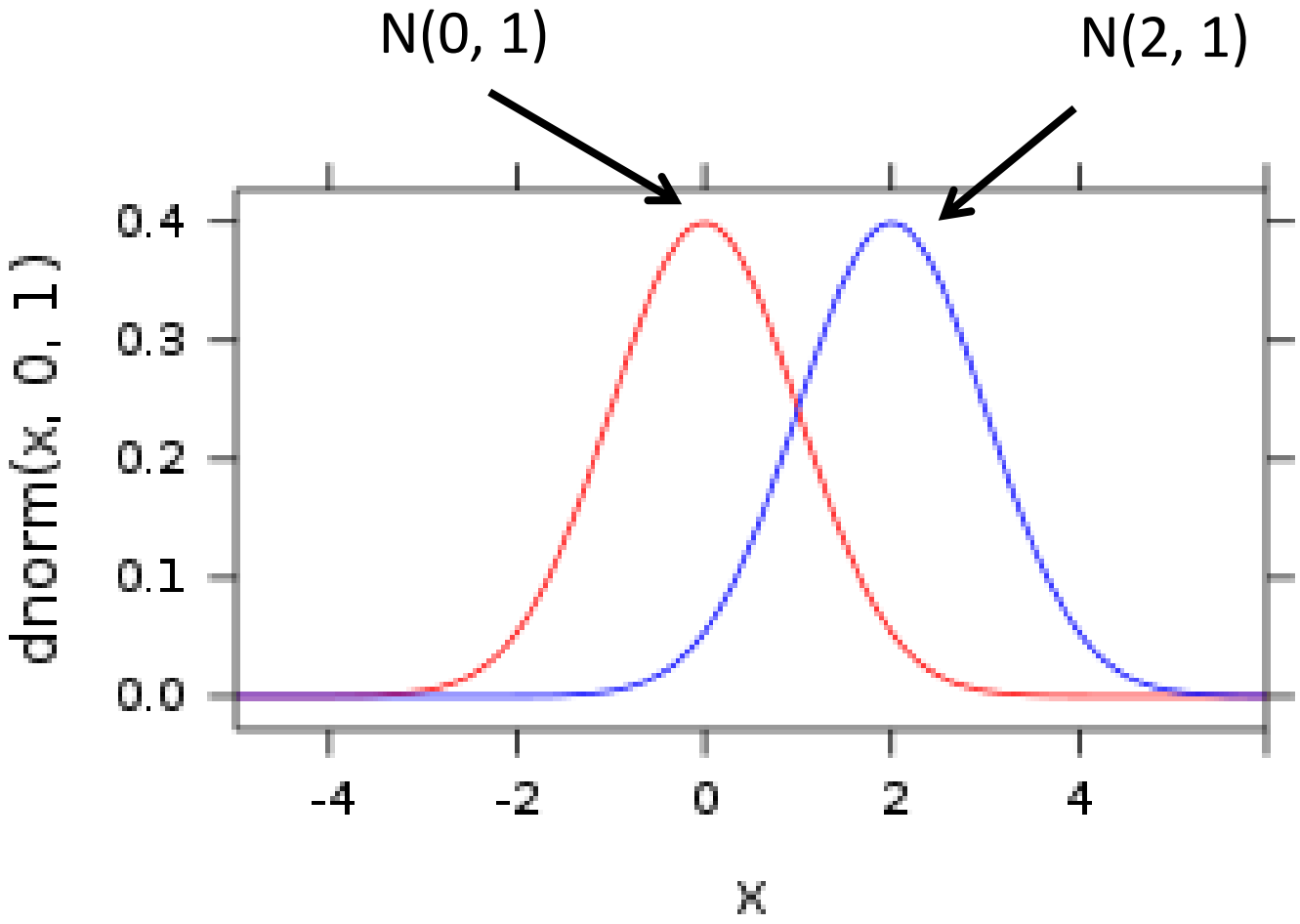
Notation: X ~ N(μ, σ)

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

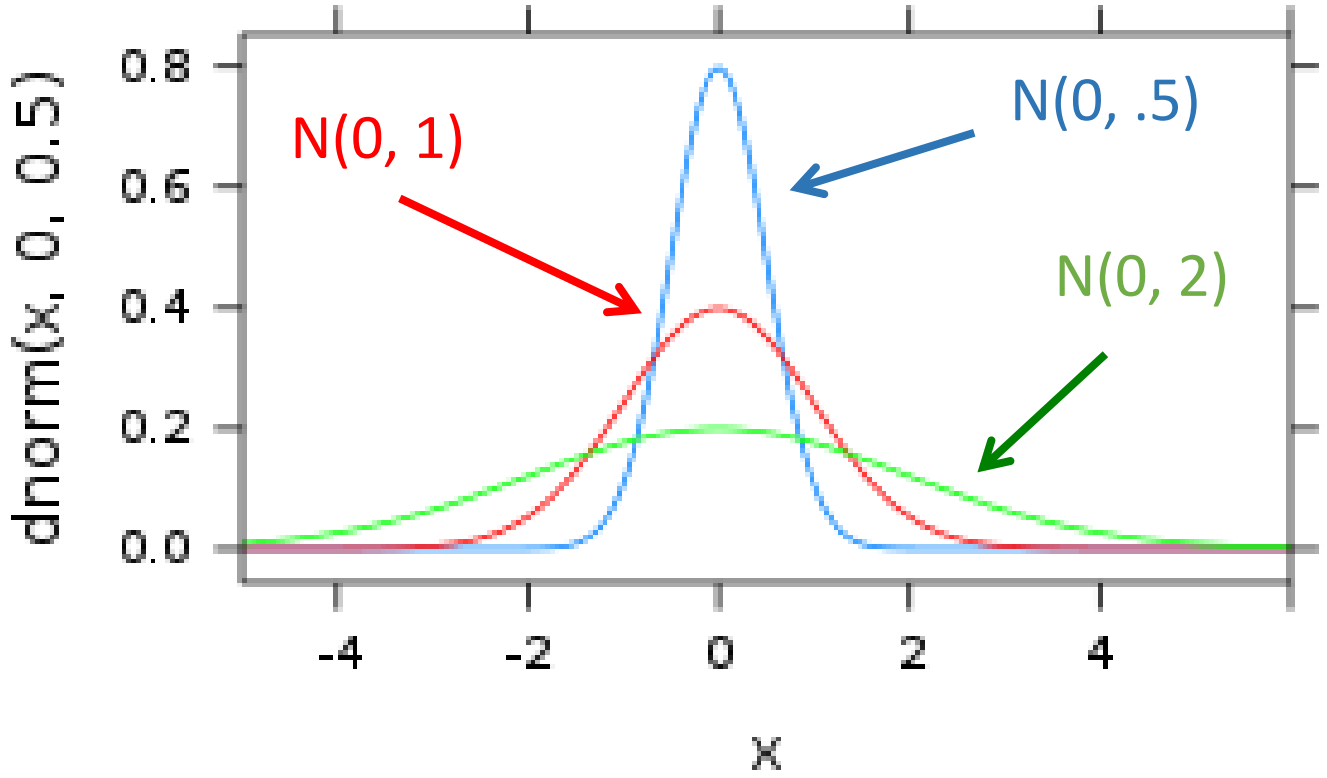
Graph of a Normal Density Curve



Normal curves with different means



Normal curves with different variances



Graphing Normal Curves

IQ scores are defined to have a mean of 100 and a standard deviation of 15

Try drawing this distribution by hand…

We can check the results in R using:

```
x <- 40:150      #  x values to plot points at
density_curve  <-  dnorm(x, mu, sigma)     # y density curve values
plot(x,  density_curve, type = "l")       # plot the x y values
```

The ‘d’ in dnorm stands for density

Finding normal probabilities and percentiles

No simple mathematical formula exists for computing the area under a normal curve

We can use R to find such areas by specifying:

- Mean and standard deviation
- The endpoints of the interval

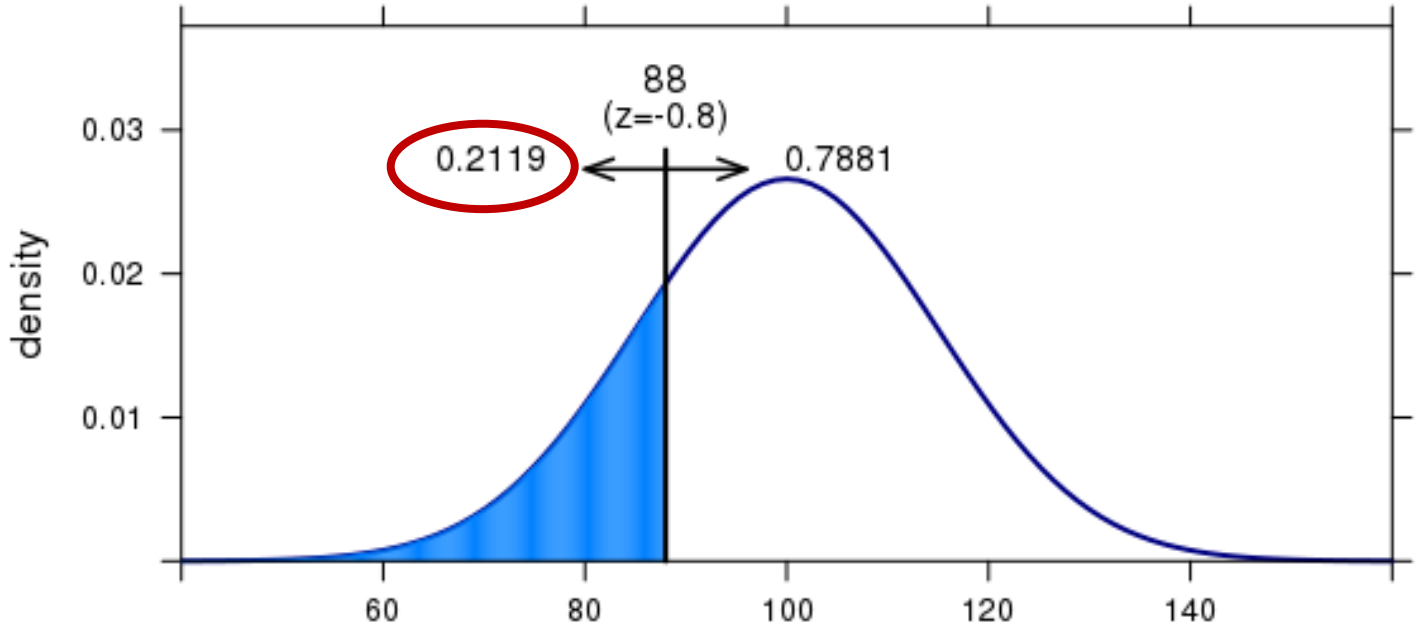
We can use the following to get the probability of Pr(X ≤ x)

> pnorm(x, mu, sigma)

The ‘p’ in pnorm stands for probability

What is the proportion of people who have an IQ less than 88?

> pnorm(88, 100, 15)



Normal area Pr(X ≤ x) app

Normal area Pr(a < X < b) app

Probability for a range of values

1. Can you calculate the proportion of people who have an IQ between 88 and 96?

2. What is probability of someone who has an IQ greater than 96?

Answer 1:

```
> pnorm(96, 100, 15) - pnorm(88, 100, 15)
```

Answer 2:

```
1 - pnorm(96, 100, 15)
```

or

```
pnorm(96, 100, 15, lower.tail = FALSE)
```

Central limit theorem

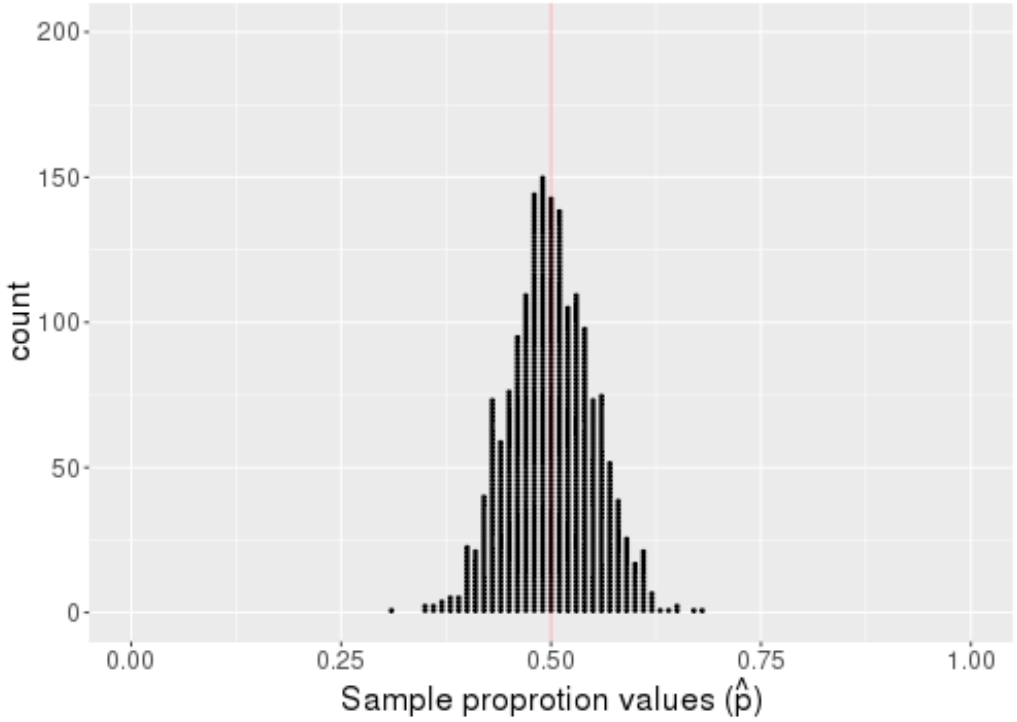
For random samples with a sufficiently large sample size, the distribution of sample statistics for a **mean ($\bar{x}$)** or a **proportion ($\hat{p}$)** is **normally distributed** and is centered at the value of the population parameter

i.e., the sampling distribution of means ($\bar{x}$) and proportions ($\hat{p}$) will be a normal distribution!

- so we don't need to do resampling to get a null distribution!

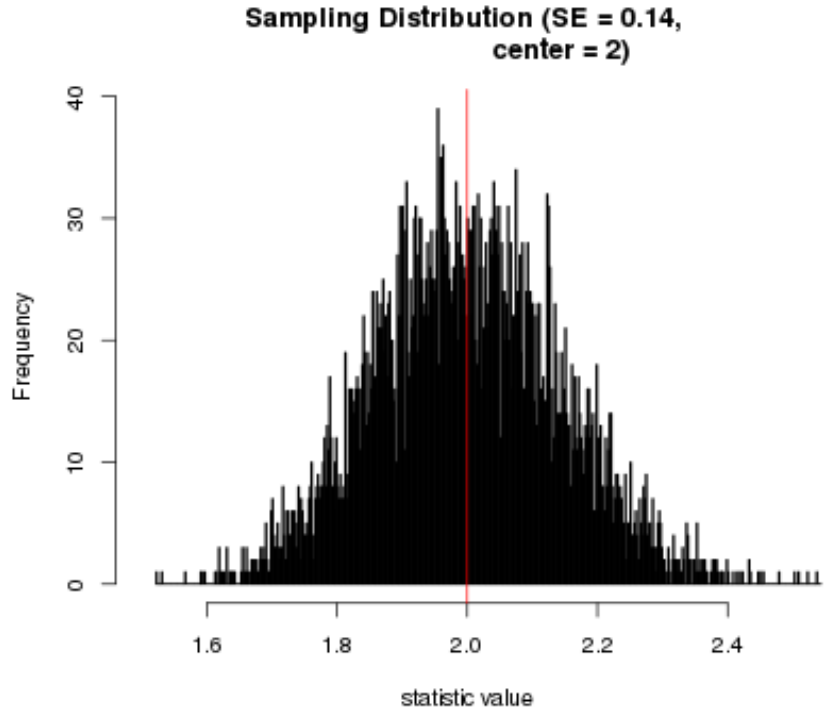
Central limit theorem

proportion ($\hat{p}$)



[Proportion sampling distribution app]

mean ($\bar{x}$)



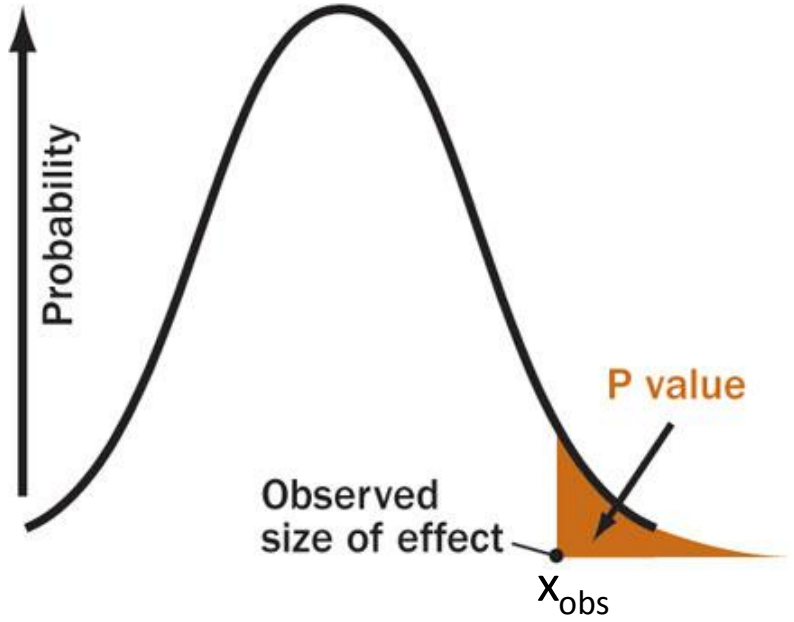
[Sampling/Bootstrap distribution app]

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

Hypothesis tests based on a Normal Distribution

The p-value for the test is the probability that we would get a random statistic X that is greater than our observed test statistic



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

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

Hypothesis tests based on a Normal Distribution

Suppose $param_0$ is the parameter specified by the null hypothesis

- For example: $H_0: \pi = param_0$ $H_A: \pi > param_0$
  or $H_0: \mu = param_0$ $H_A: \mu > param_0$

If the null hypothesis <u>was true</u>, and the null distribution is normal, we have:

- obs_stat ~ $N(param_0, SE)$

<u>To test if the null hypothesis is true</u>, we can get a p-value using:

> 1 - pnorm(obs_stat, $param_0$, SE)

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

1. Start by stating H_0 and H_A

H_0: $\pi = .4$

H_A: $\pi > .4$

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

2. What is the observed statistic?

$\hat{p} = 0.43$

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

2. If $\hat{p}$ comes from a null distribution that is normal, what are μ and σ for this null distribution?

μ = H_0: $\pi = 0.4$

σ = SE = 0.016

Do greater than 40% of Americans go without using cash in a typical week?

Steps: 3-4. What is the probability one would get a statistic as large or larger than $\hat{p} = .43$ from a normal distribution: $N(0.40, 0.016)$?

> pnorm(.43, .4, .016, lower.tail = FALSE)

> 1 - pnorm(.43, .4, .016)

p-value = .0304

Step 5?

Normal area app $Pr(X \leq x)$

Standard Normal N(0, 1)

Since all normal distributions have the same shape, it is often convenient to convert them to a stand scale with:

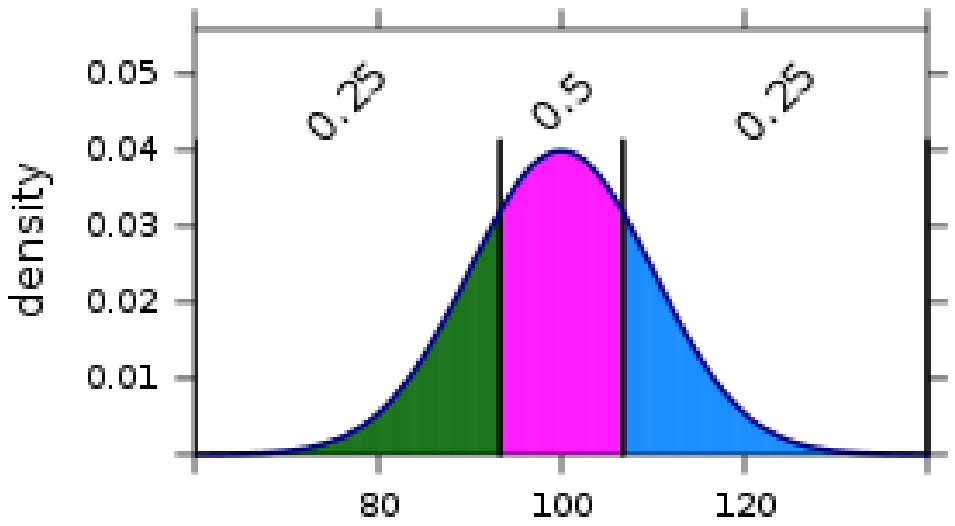
$$\mu = 0, \quad \sigma = 1$$

This is called the **standard normal** distribution:

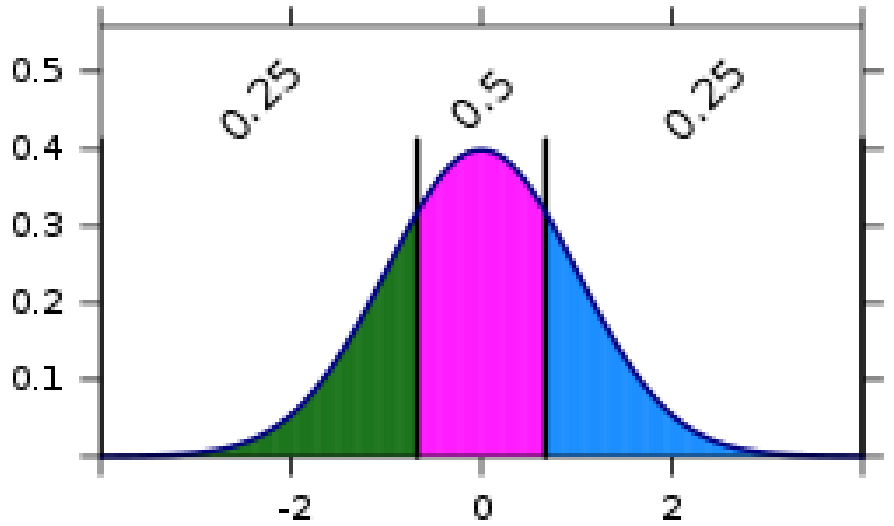
$$Z \sim N(0, 1)$$

IQ scores, X and Z scales

X ~ N(100, 15)



Z ~ N(0, 1)



[Normal quantile app]

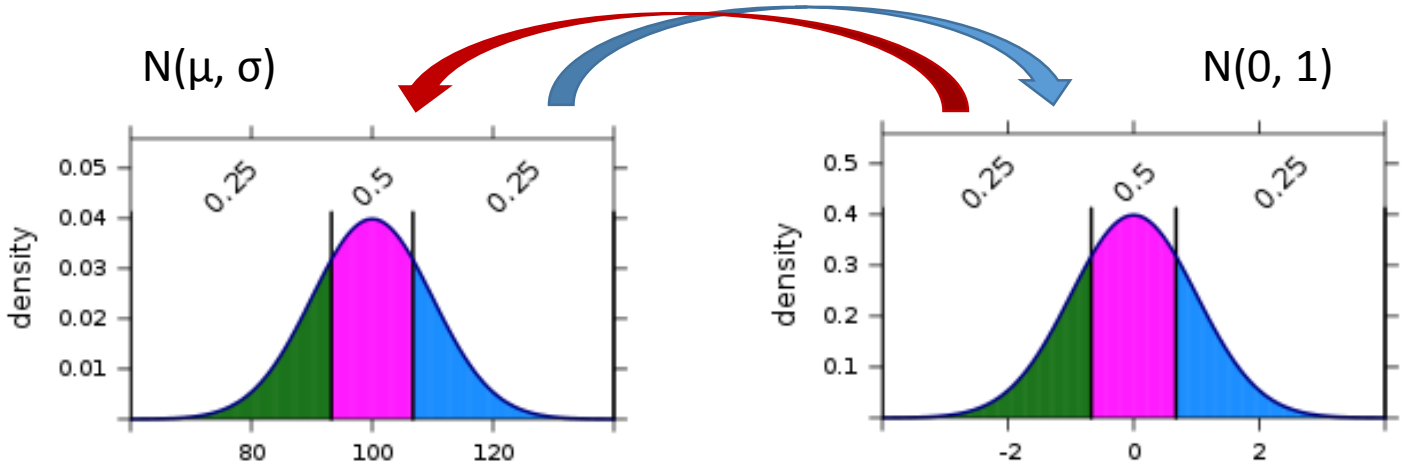
Standard Normal N(0, 1)

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:

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



Converting to the standard normal distribution

1. What is the Z-score of someone who has an IQ score of 112?

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

2. What if someone has an Z-score of 2.2, what is their IQ score?

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

Answer 1: $Z = (112 - 100)/15 = .8$

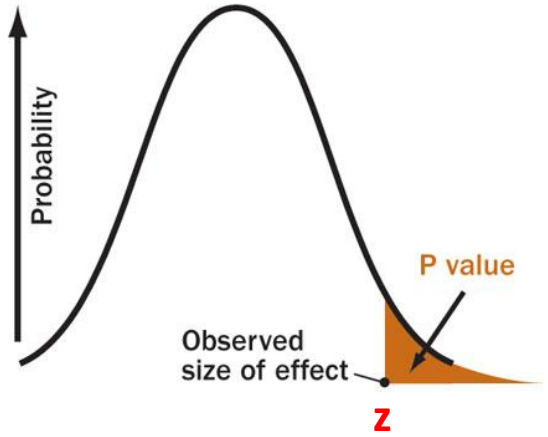
Answer 2: $IQ = 100 + 2.2 * 15 = 133$

Hypothesis tests based on a Normal Distribution

When the distribution of a statistic under H_0 is normal, it is often convenient to use a standardized 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



$Pr(Z \geq z)$

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

Hypothesis tests based on a Normal Distribution

To repeat what was on the last slide: we can transform our obs_stat to a z-statistic that comes from a standard normal distribution N(0, 1) using:

$$z = \frac{stat_{obs} \text{ - } param_0}{SE}$$

The p-value is then the probability of obtaining a value from a standard normal distribution beyond this z statistic

> pnorm(z, 0, 1) if H_A: $\mu < param_0$

> 1 - pnorm(z, 0, 1) if H_A: $\mu > param_0$

> 2 * (1 - pnorm(abs(z), 0, 1)) if H_A: $\mu \neq param_0$

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

1. Start by stating H_0 and H_A

H_0: $\pi = .4$

H_A: $\pi > .4$

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

2. Can you compute the z statistic?

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

Do greater than 40% of Americans go without using cash in a typical week?

A survey of 1,000 Americans reported that 43% said they went an entire week without using cash, with a SE = 0.016

Assuming the distribution of the statistic is normal, calculate whether the proportion of all Americans going a week without using cash is greater than 40%

2. Can you compute the z statistic?

$$z = (.43 - .4)/.016 = 1.875$$

Do greater than 40% of Americans go without using cash in a typical week?

Steps: 3-4. What is the probability one would get a z-statistic as larger or larger than 1.875 from a standard normal distribution?

> pnorm(1.875, 0, 1, lower.tail = FALSE)

> 1 – pnorm(1.875, 0, 1)

p-value = .0304

Step 5?

Normal area app Pr(X ≤ x)

Lock 5 problems

First edition: 5.2, 5.40, 5.56, 5.58