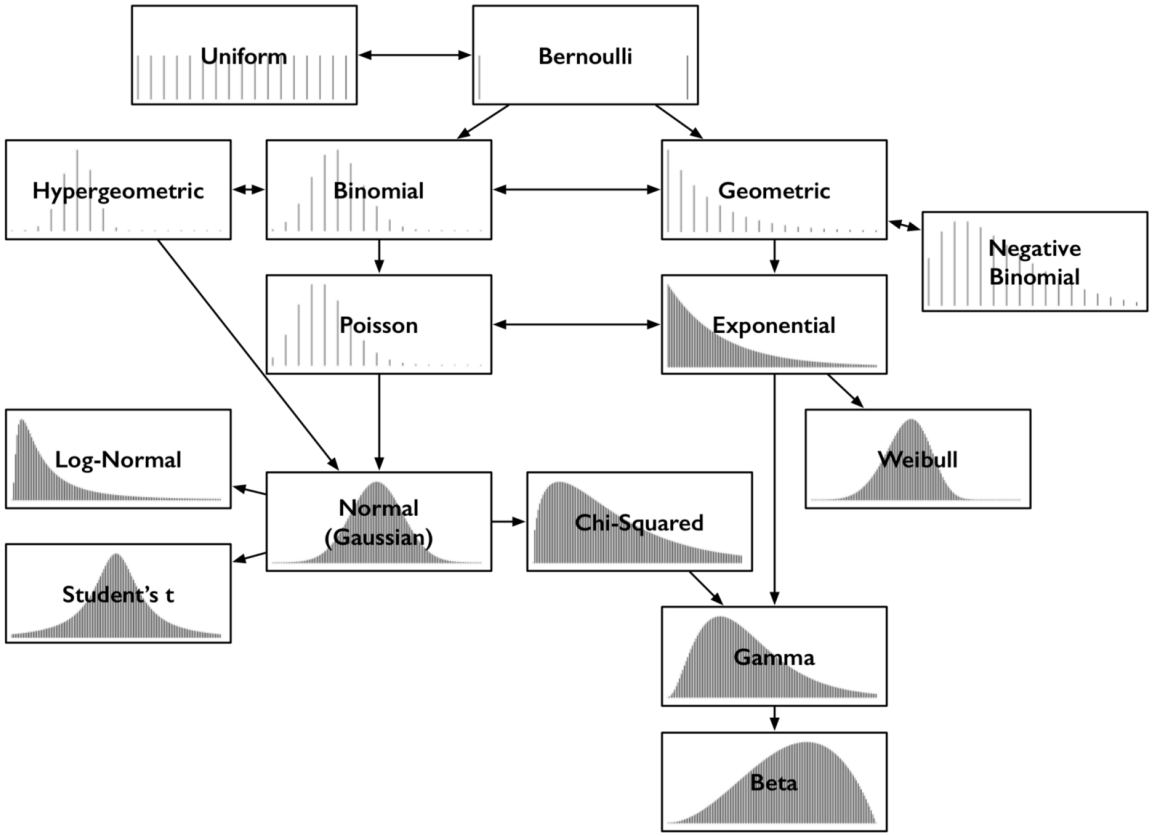
# Probability distributions



# Overview

Review of draft lottery questions on worksheet 10

Connections between confidence intervals and hypothesis tests

Review of theories of hypothesis tests

Density curves

The Normal distribution

# Final project: analyze your own data set

Final project report: a 5-10 page R Markdown document:

```
> source('/home/shared/intro_stats/cs206_functions.R')
> get_worksheet("final")
```

A one paragraph final project proposal is due today

- What question you will answer
- Where you will get the data

# Questions about worksheet 10?

## 1969 Draft Lottery

| sequential_date | draft_number |
|---|---|
| 1 | 305 |
| 2 | 159 |
| 3 | 251 |
| 4 | 215 |
| 5 | 101 |
| 6 | 224 |
| 7 | 306 |
| 8 | 199 |
| 9 | 194 |
| 10 | 325 |
| 11 | 329 |
| 12 | 221 |
| 13 | 318 |
| 14 | 238 |
| 15 | 17 |
| 16 | 121 |
| 17 | 235 |

# 1969 Vietnam Draft

In a perfectly fair, random lottery, what should be the value of the correlation coefficient between *draft number* and *sequential date of birthday*?

**Question**: was the 1970's Vietnam Draft lottery fair?

How many people think it was fair?

- Why?

# 1969 Vietnam Draft

1. State the null and alternative hypothesis in symbols and in words:

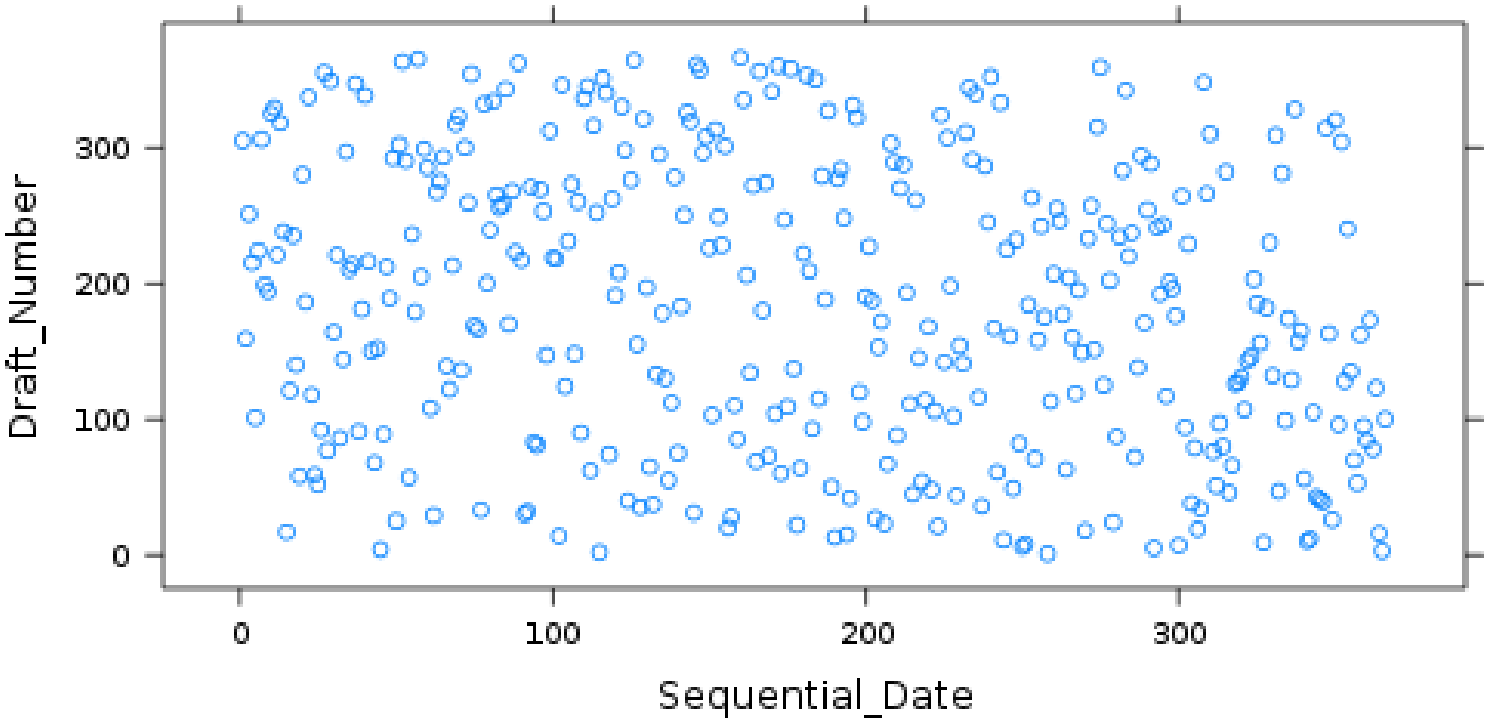
   $H_0$: $\rho = 0$ correlation between draft number and date is 0

Which of these alternative hypotheses would be best to use?

a. $H_A$: $\rho < 0$
b. $H_A$: $\rho > 0$ 
c. $H_A$: $\rho \neq 0$

Please use LaTeX for symbols in your R Markdown documents

# Do draft number and date appear to be related?

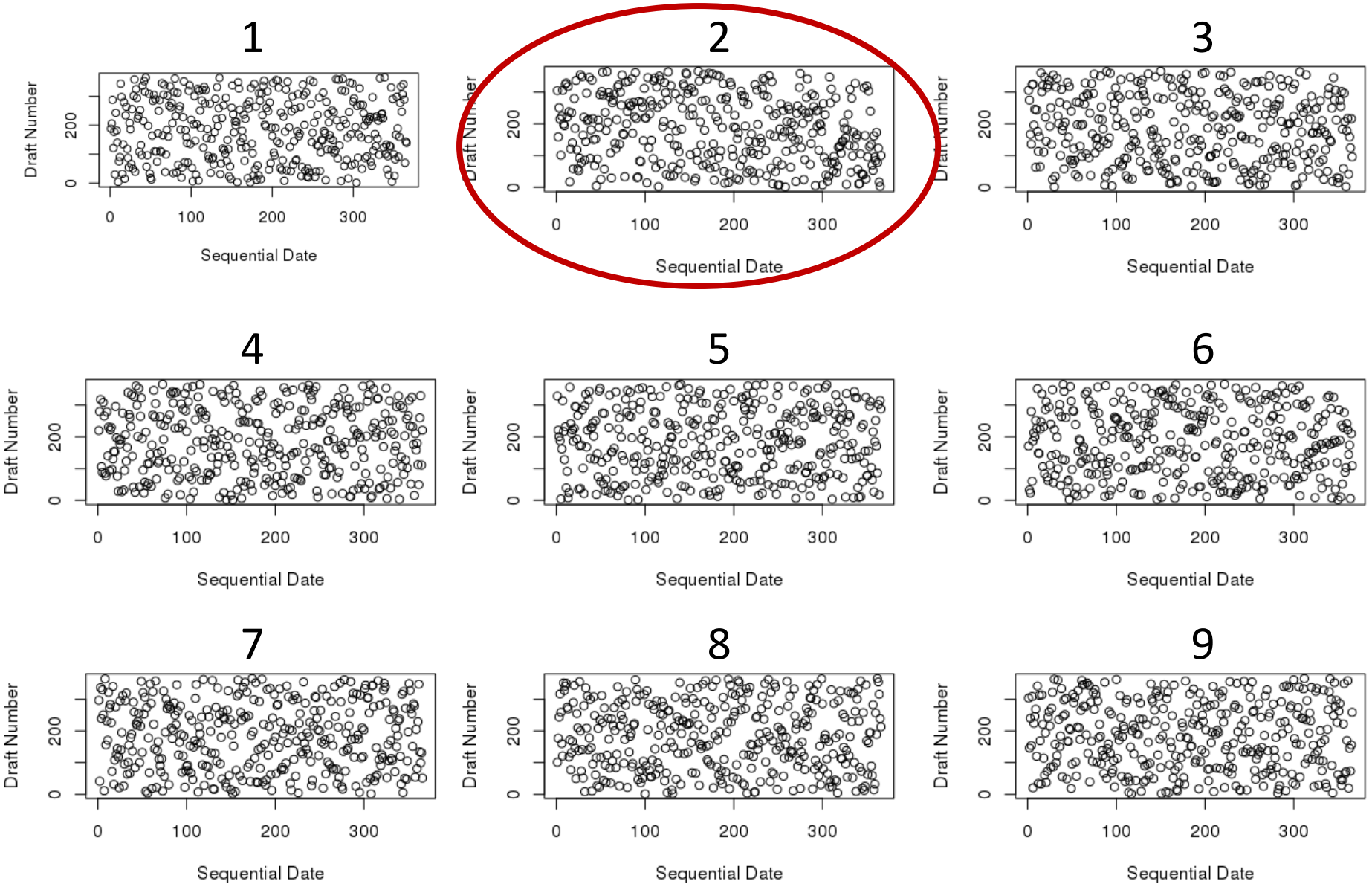



2. observed statistic: r = -0.226

Q: Do people born in February have earlier draft numbers than in October?

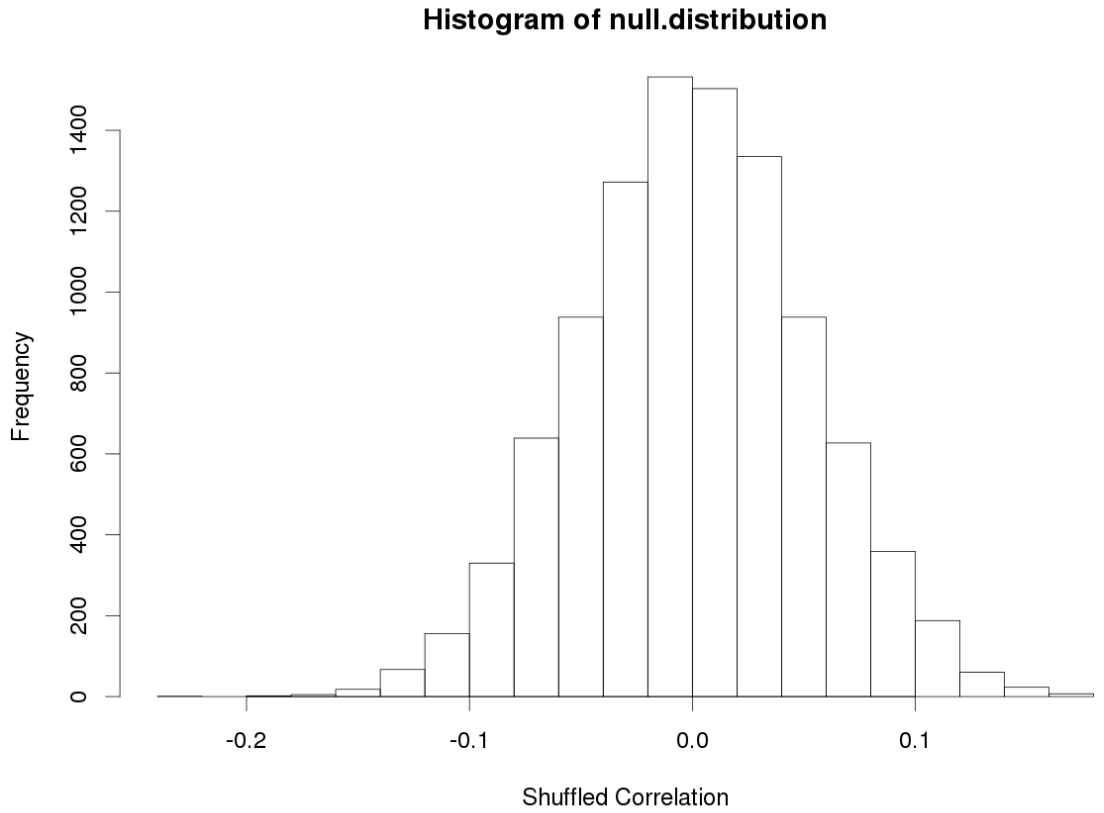
- A: No, the later people were born in the year, the lower their draft number

# Visual hypothesis test – which is the real scatter plot?

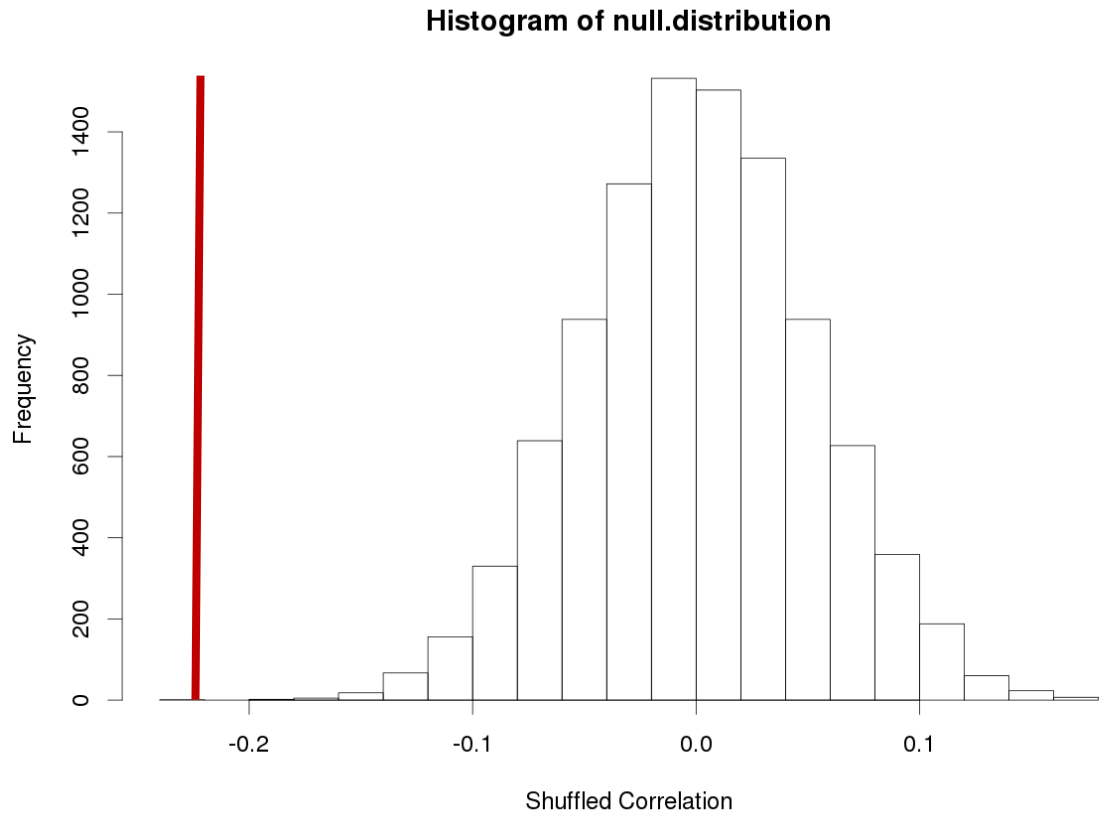


# 3. Null distribution

How did we create the null distribution?



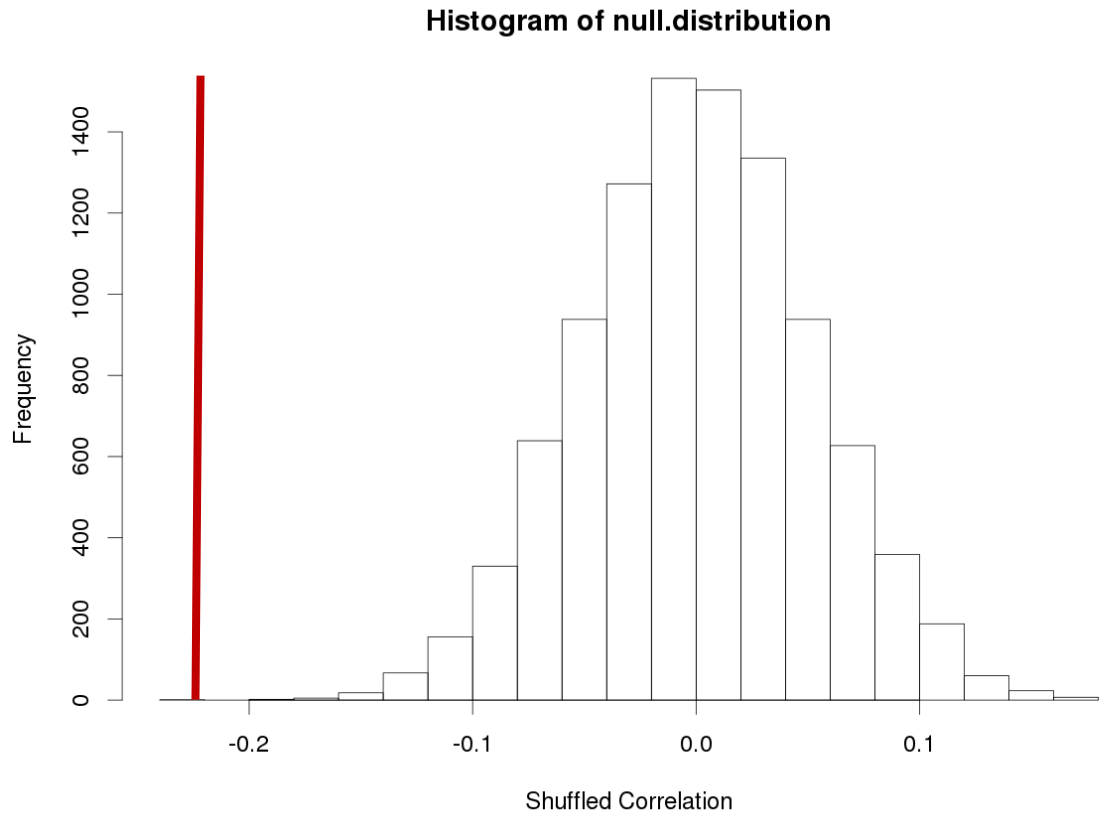
# 4. Calculate p-value




r = -0.226

abline(v = obs_stat, col = "red")

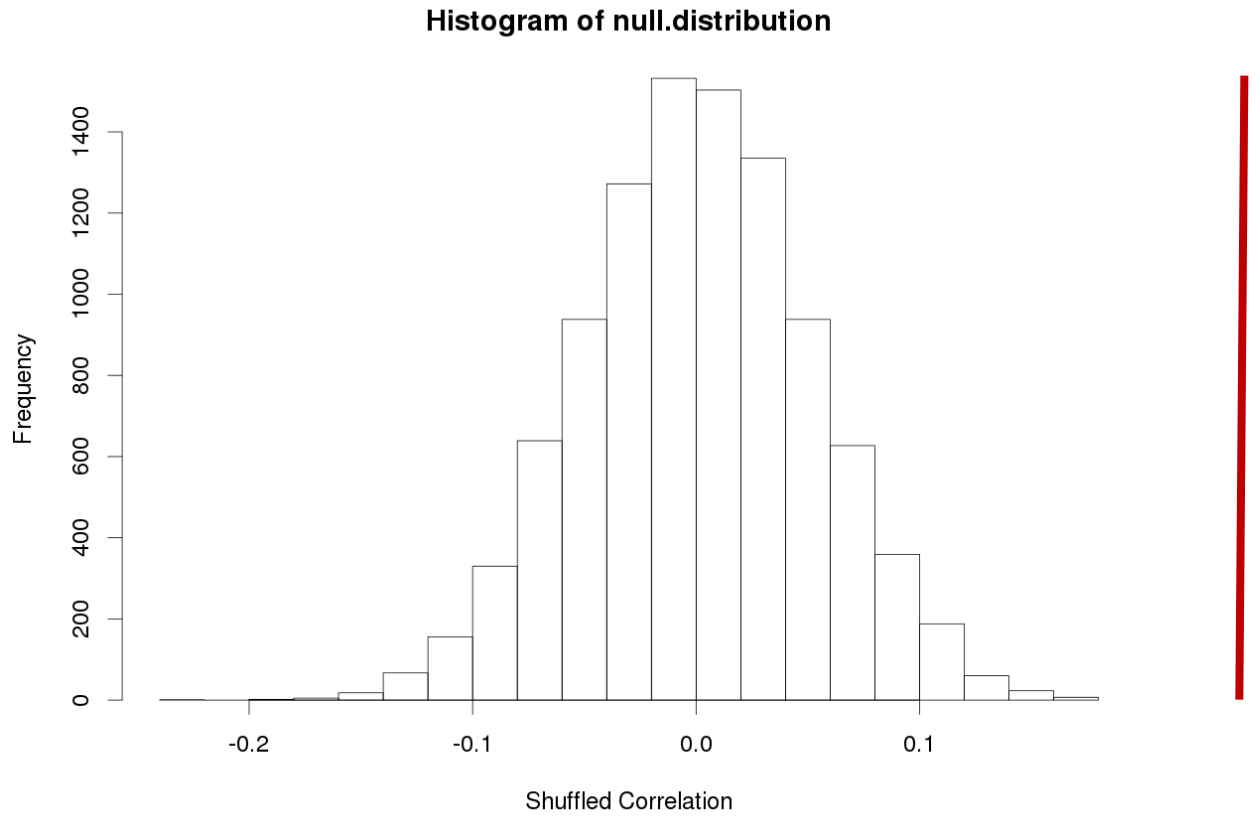
# 4. Calculate p-value




r = -0.226

num_points_left_tail <- sum(null_distribution **<=** obs_stat)

# 4. Calculate p-value




r = -0.226

num_points_right_tail <- sum(null_distribution **>=** abs(obs_stat))

# 4. Calculate p-value

```
num_points_both_tails <-
    num_points_left_tail +
    num_points_right_tail

p_value <- num_points_both_tails/10000

> p_value
[1] 0
```

# 5. Draw a conclusion

Null and alternative hypotheses:

- ~~$H_0$: $\rho = 0$ correlation between draft number and date is 0~~
- $H_A$: $\rho \neq 0$ correlation between draft number and date is not 0

What does a p-value of 0 tell us?

**Question**: was the 1970's Vietnam Draft lottery fair?

# 1969 Vietnam Draft

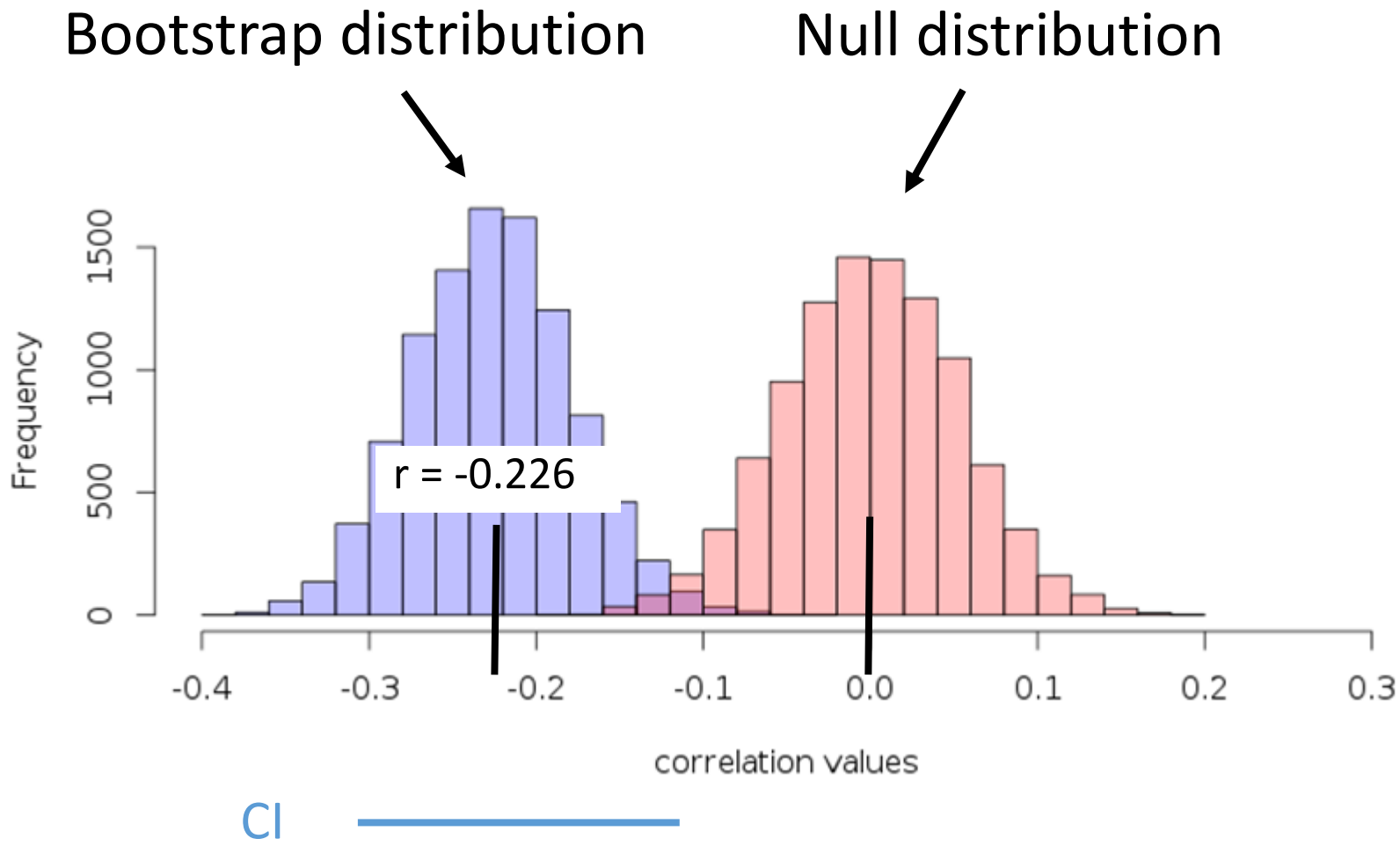
The results show that the draft lottery does not appear to be completely random (very small p-value) with people born later in the year more likely to be drafted.

An explanation for this non-randomness is due to the fact that the capsules that contained the dates were put in a box month by month, January through December, and subsequent mixing efforts were insufficient to overcome this sequencing.

# Draft lottery data

As we saw, for the draft lottery we hypothesized that $H_0 : \rho = 0$

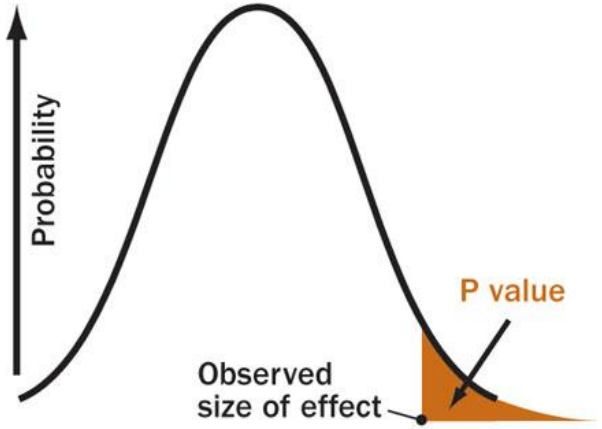
A 95% CI for $\rho$ [-.321 -.131]




The fact that the 95% confidence interval [-.321 -.131] does not contain the null hypothesis parameter ($H_0$: $\rho = 0$) means that a hypothesis test will reject at ($\alpha = 0.05$)

# Review: Two theories of hypothesis testing

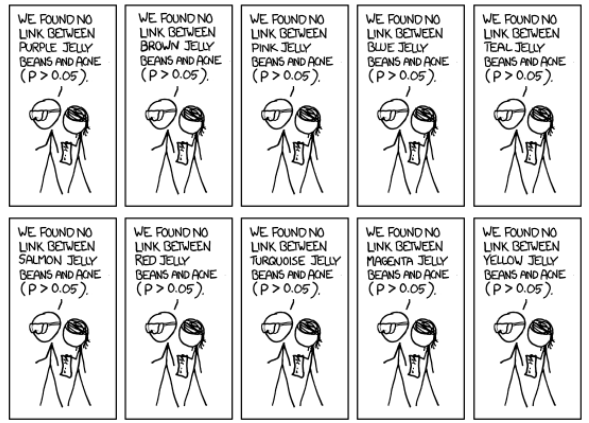
1. **Significance testing** of Ronald Fisher
   - p-value as strength of evidence against the null hypothesis
2. **Hypothesis testing** of Jezy Neyman and Egon Pearson
   - Make a formal decision of whether to reject $H_0$

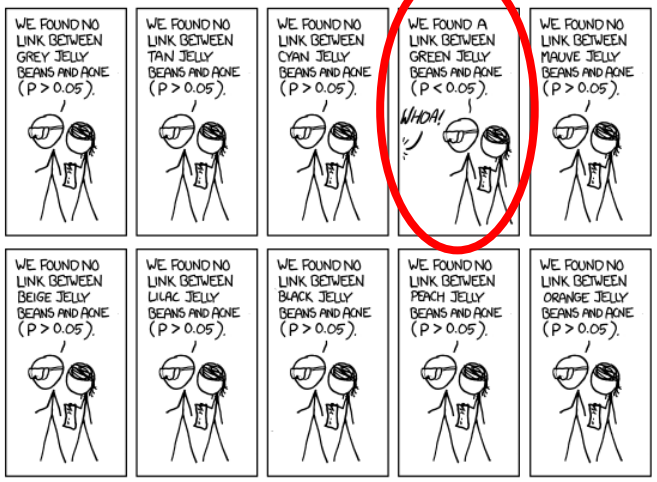


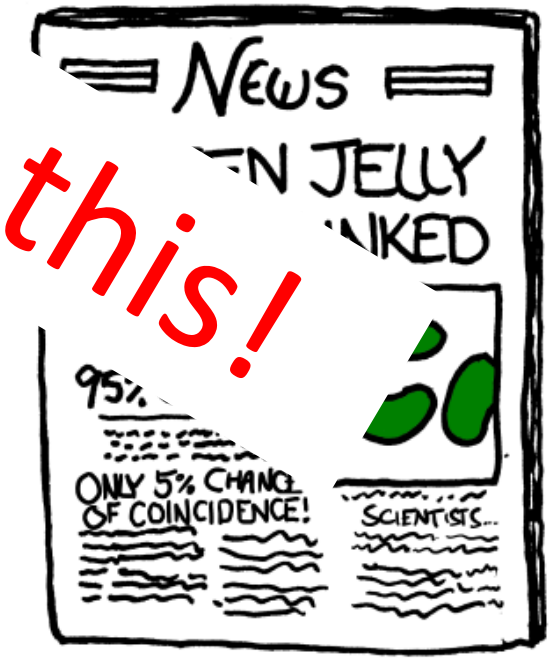

**Significance testing**


**Hypothesis testing**

# Multiple hypothesis tests


Don't ever do this!

# Bayesian analyses

P-value:

$Pr( data \mid H_0)$

Posterior probability:

$Pr( H_0 \mid data)$

Reverend Thomas Bayes

(not actual picture)

Take more advanced Statistics classes to learn more!

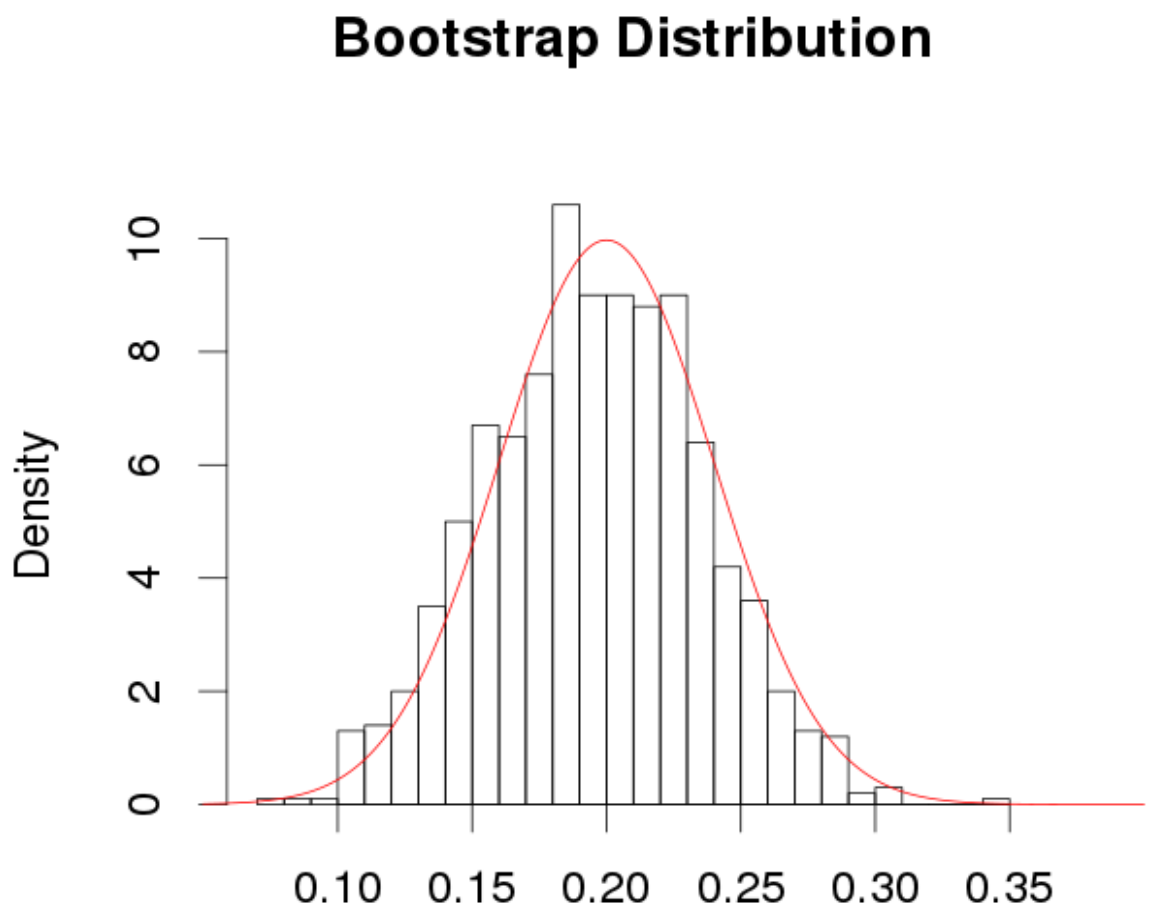
# Inference using parametric probability distributions

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 just mathematical probability distributions

# Comparing bootstrap distribution and a probability distribution



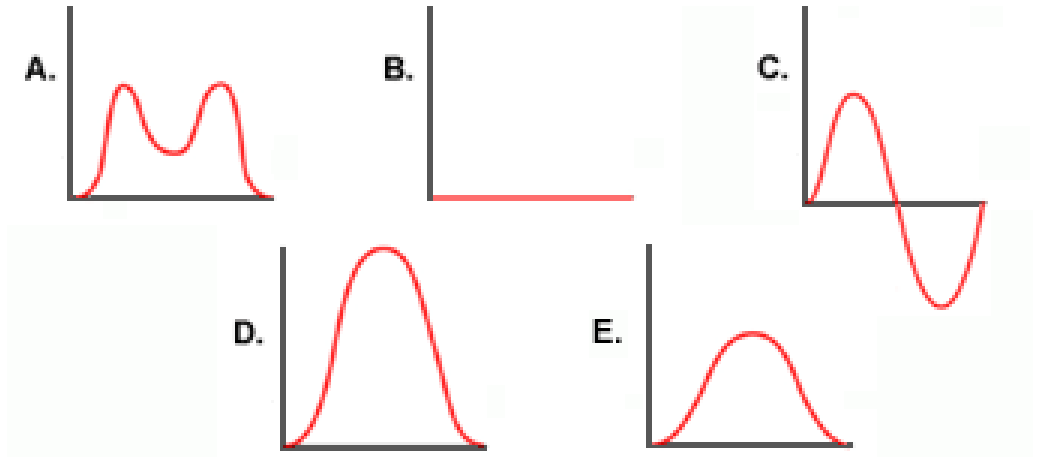
# Density Curves

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

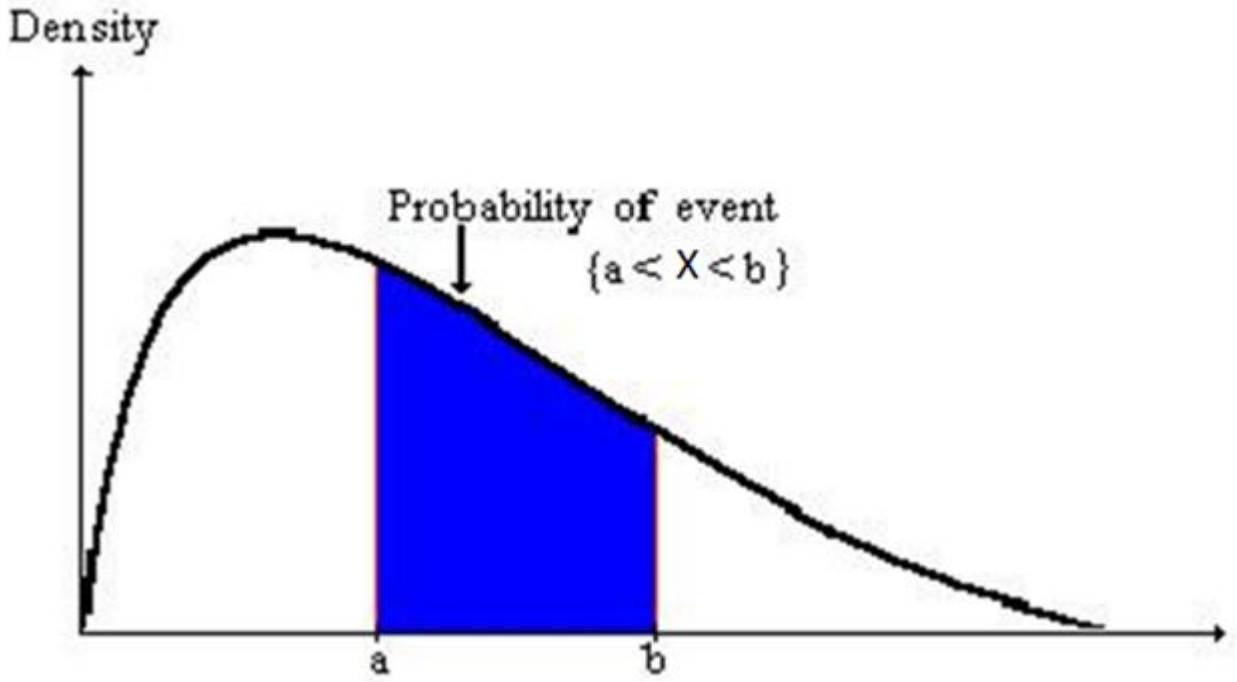
Which of these could ***not*** be a density curve?



# 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



# Normal Density Curve

A normal distribution follows a bell-shaped curve

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

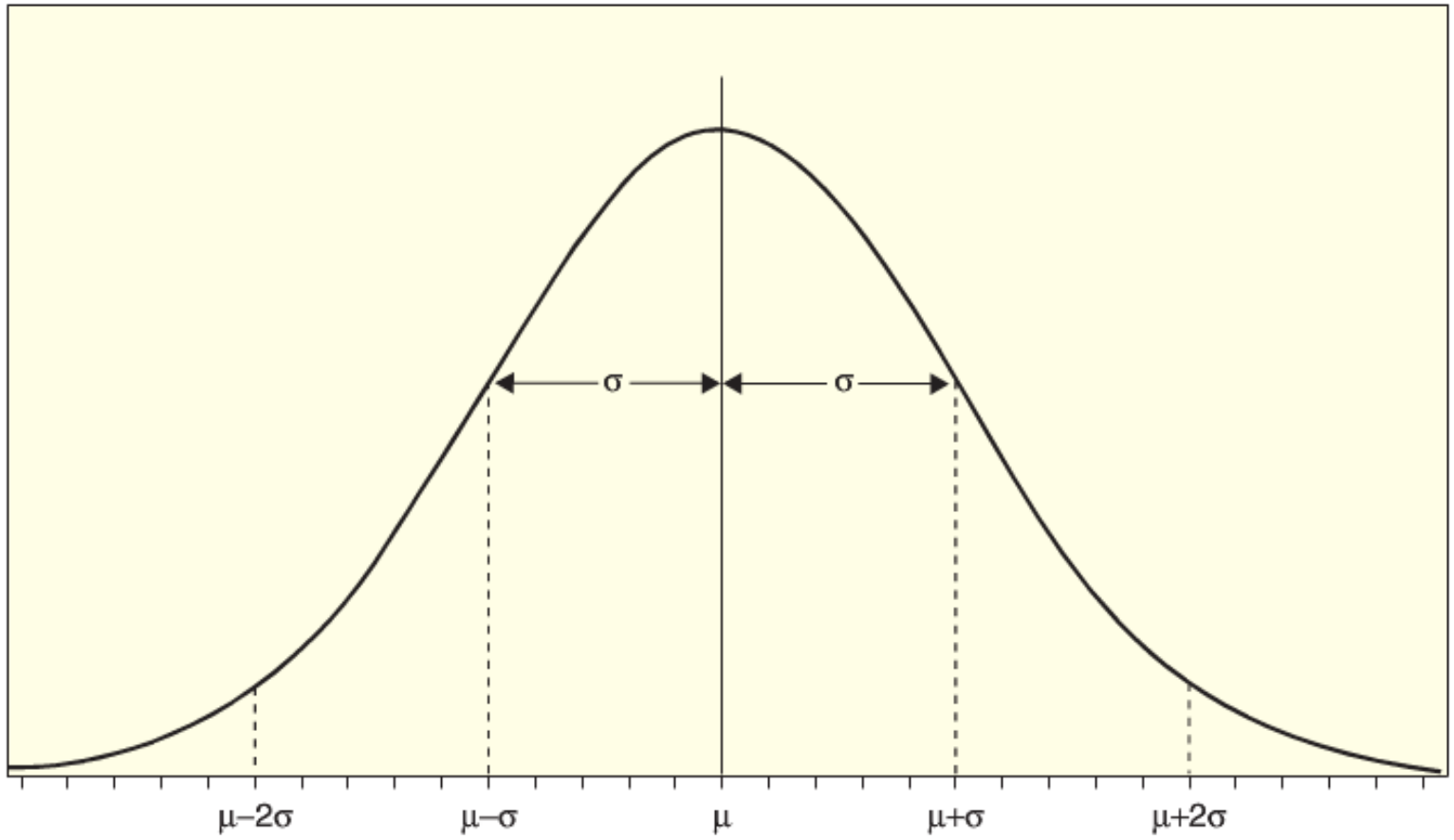
- The mean: μ
- The standard deviation: σ

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

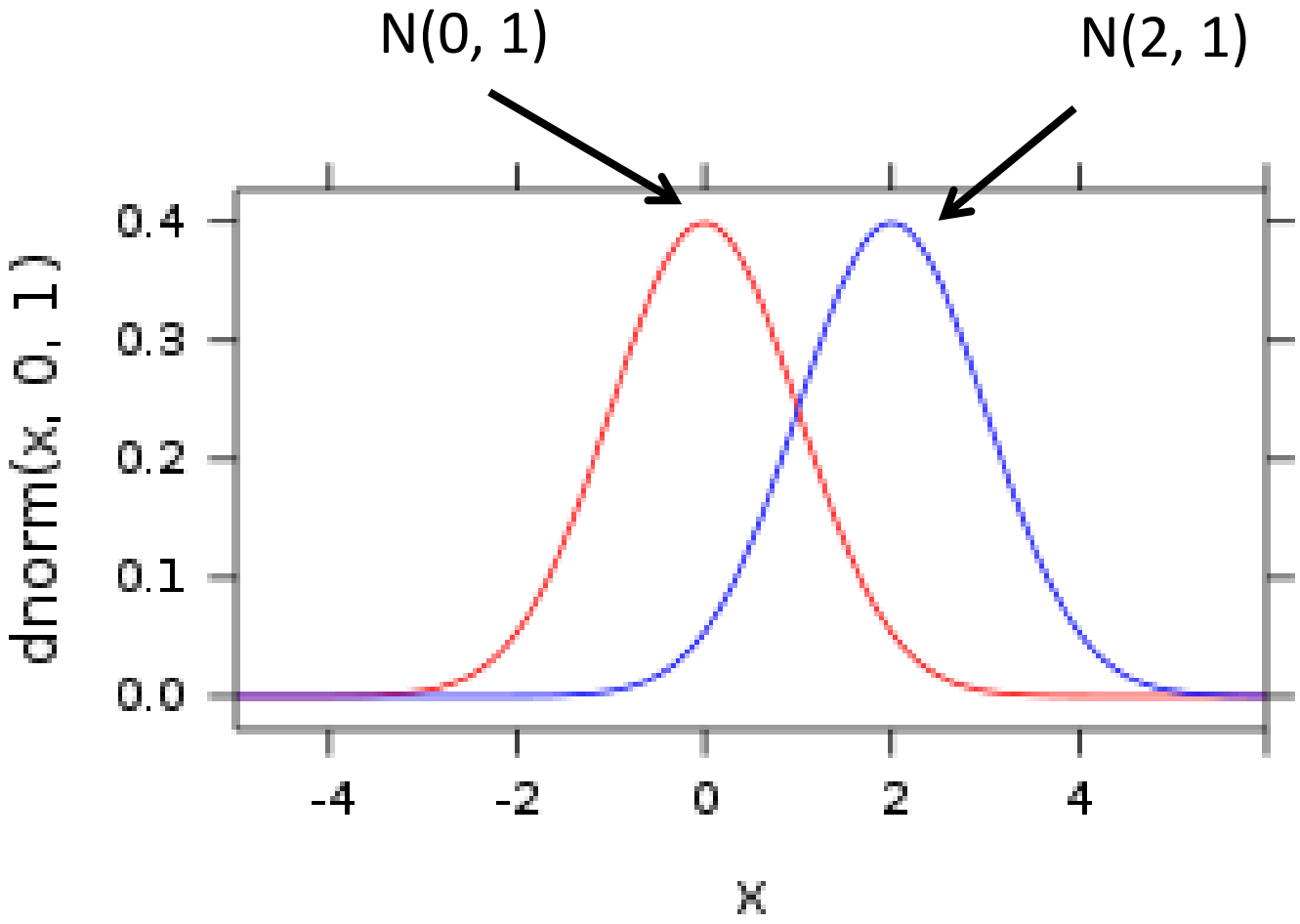
Notation: X ~ N(μ, σ)

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

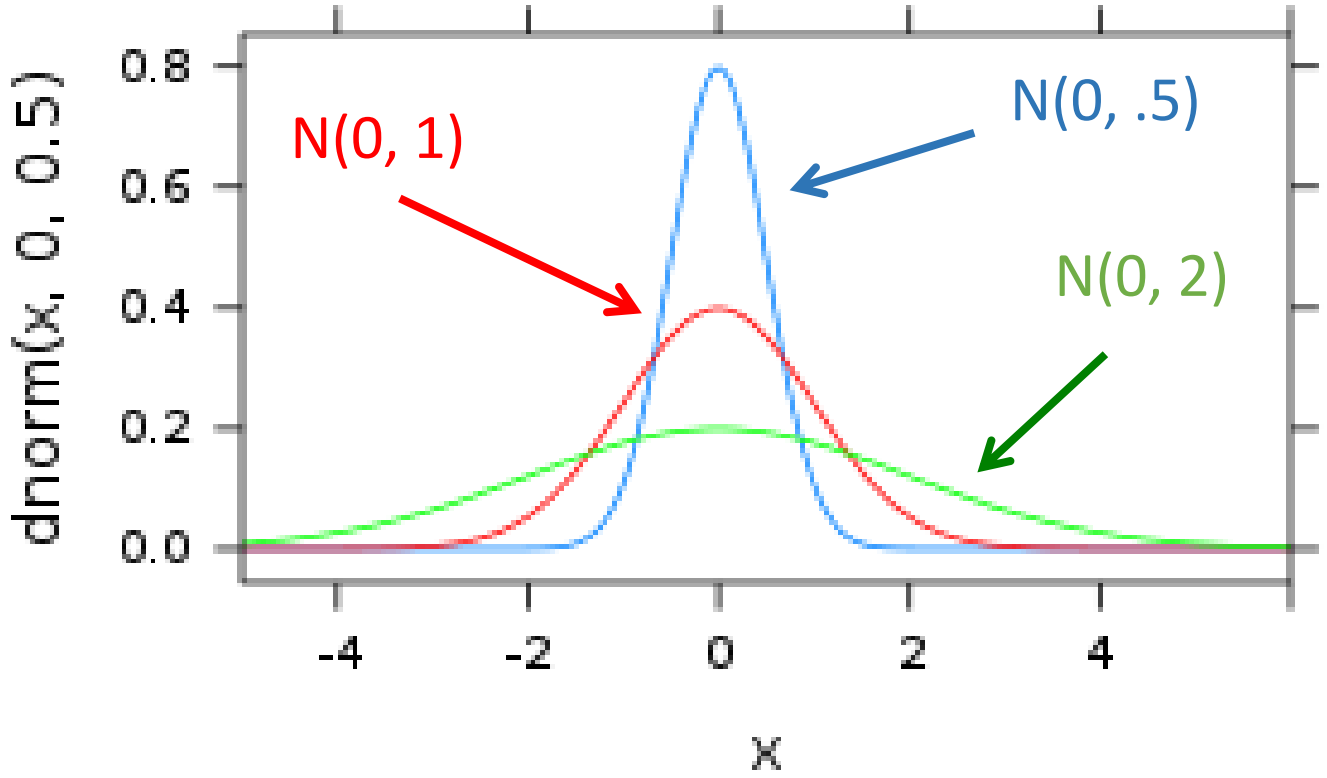
# Graph of a Normal Density Curve



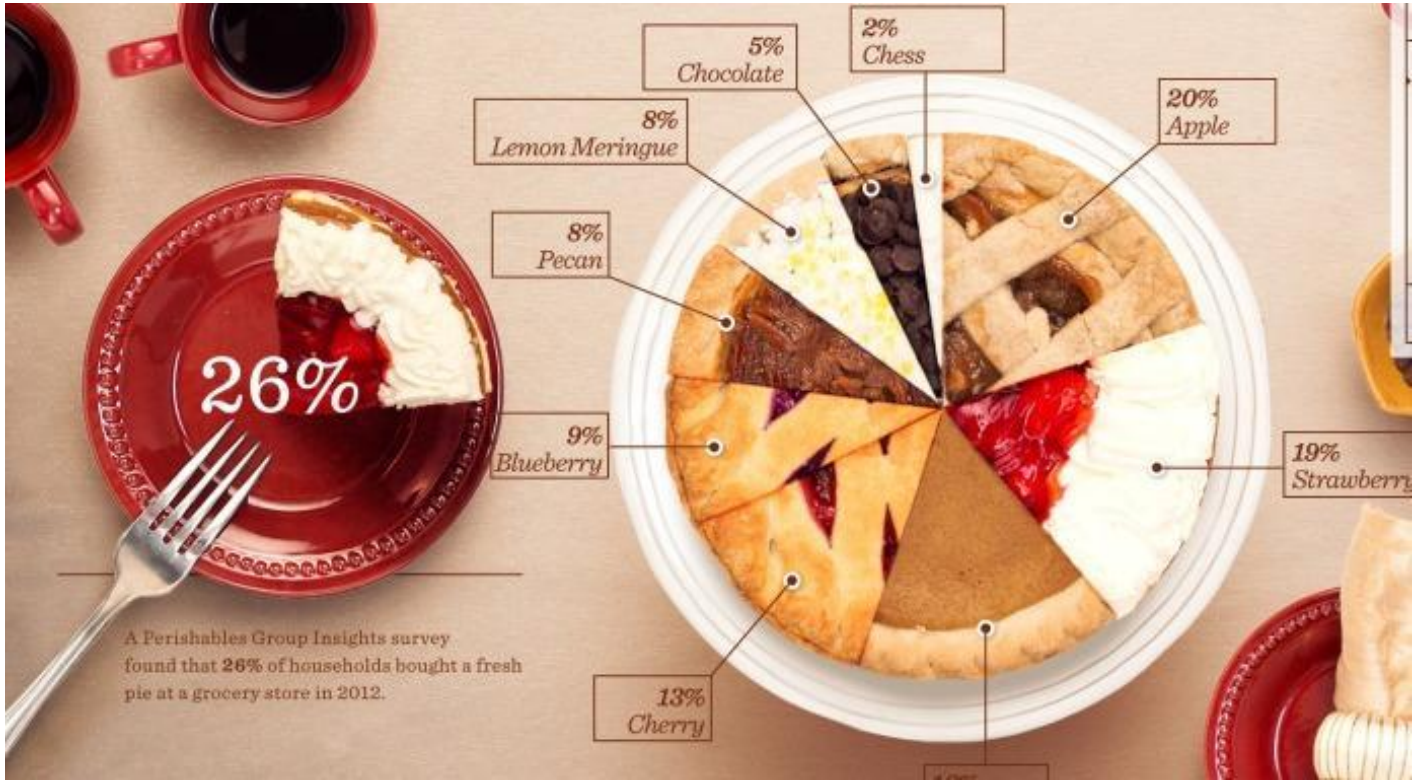
# Normal curves with different means



# Normal curves with different variances



# Have a good Thanksgiving!




Thanksgiving worksheet: resubmit Draft Lottery question from worksheet 10