# Hypothesis tests for more means and confidence intervals

# Outline for today

Review: parametric hypothesis tests for two means

Tests for more than two means

Confidence intervals

# Announcement: Final projects

All final project presentations are due at 11:59pm on Sunday April 30th

- A week from this Sunday

5 minute presentations: 3 slides

Final project written reports are due the Sunday after the last day of class (May 7th)

# Project presentation template

## Basic presentation template:

1. The question am I addressing is _____
   - It is important because _____
2. The way I am addressing the question is _____
3. The results I have show _____
4. I conclude that _____

## Example:

1. AL players hit more home runs than NL players
   - Due to larger stadiums in NL
   - Of interest b/c Red Sox are buying a NL player
2. Looking at players HR rates who were traded from the NL to the AL
3. Players on average show more HRs after a trade (plot and stats)
   - (perhaps controlled for age and other factors)
4. Hypothesis supported, worth factoring this in when buying players

# Review: parametric hypothesis tests for 2 means

Students pulse rates were measured during lecture and while taking an examine to see if pulse rate goes up while taking an examine

The data for the 10 students are:

| Student | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quiz | 75 | 52 | 52 | 80 | 56 | 90 | 76 | 71 | 70 | 66 | 68.8 | 12.5 |
| Lecture | 73 | 53 | 47 | 88 | 55 | 70 | 61 | 75 | 61 | 78 | 66.1 | 12.8 |

**Question:** does pulse rate increase while taking an exam?

How can we go about addressing this question?

# Does taking an examine increase your pulse rate?

Students pulse rates were measured during lecture and while taking an examine to see if pulse rate goes up while taking an examine

The data for the 10 students are:

| Student | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quiz | 75 | 52 | 52 | 80 | 56 | 90 | 76 | 71 | 70 | 66 | 68.8 | 12.5 |
| Lecture | 73 | 53 | 47 | 88 | 55 | 70 | 61 | 75 | 61 | 78 | 66.1 | 12.8 |

**Question:** does pulse rate increase while taking an exam?

load('/home/shared/baseball_stats_2017/QuizPulse10.rda')

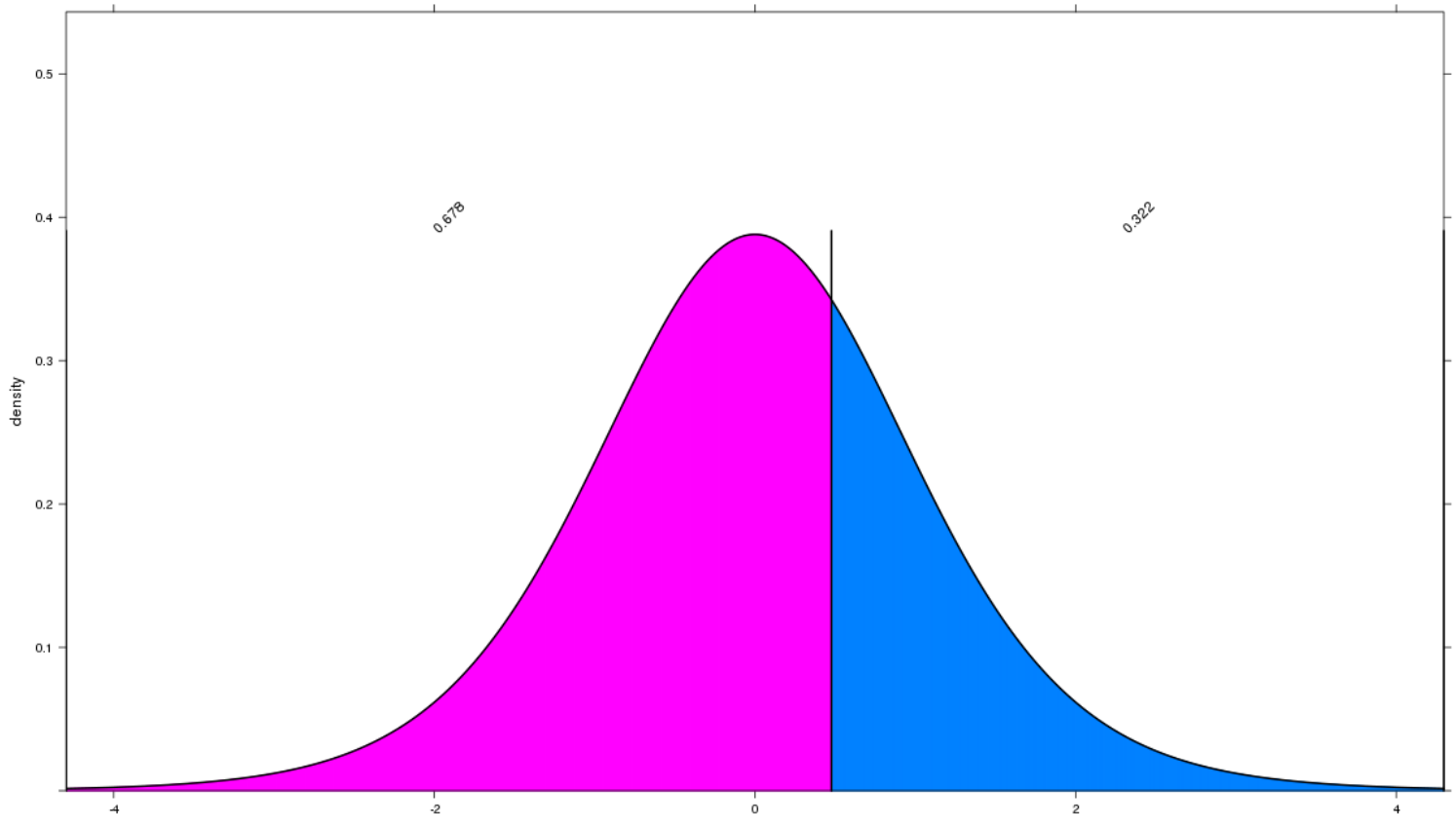
$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

pt(t.stat, df = (n – 1), lower.tail = FALSE)

# Answer in R

```
quiz <- QuizPulse10$Quiz
lecture <- QuizPulse10$Lecture

mean.diff <- mean(quiz) - mean(lecture)
pool.sd <- sqrt(var(quiz)/10 + var(lecture)/10)

t.stat <- mean.diff/pool.sd

pt(t.stat, df = 9, lower.tail = FALSE)
```

# Quiz vs. lecture null distribution



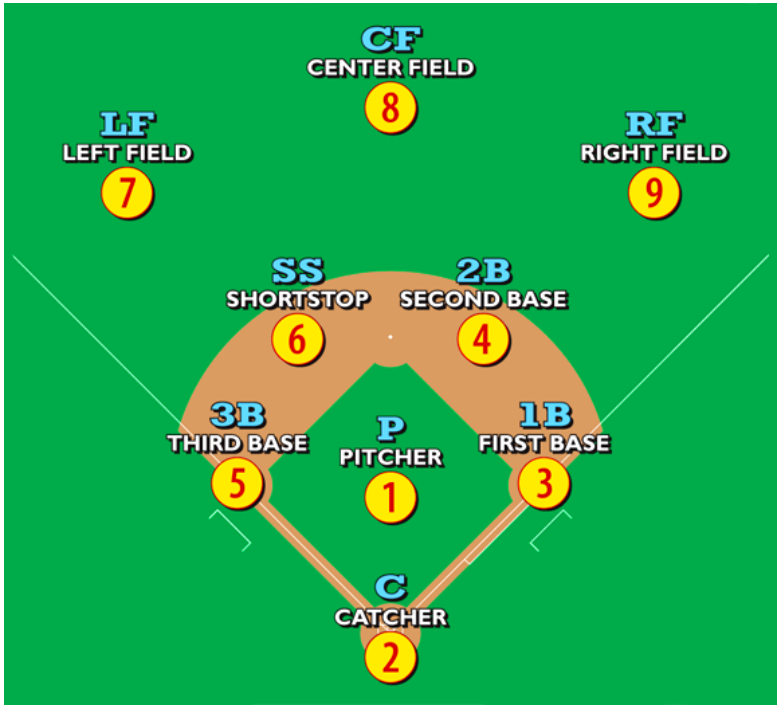

p-value = .322

# Comparing more than two means

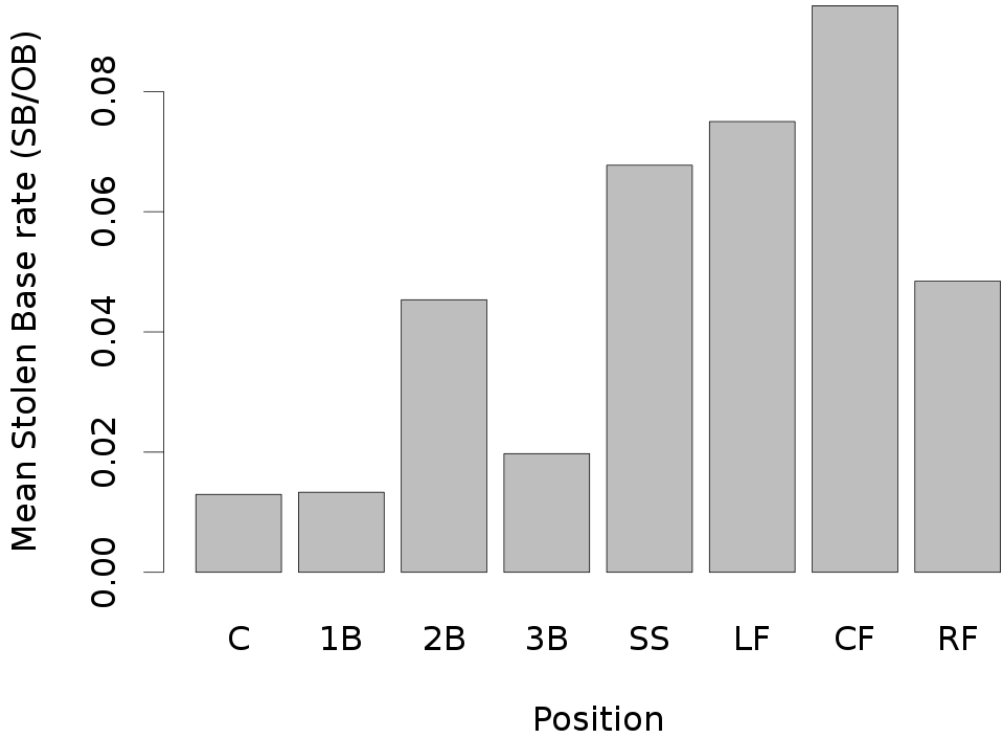
**Question**: Does the stolen base rate vary based on player's fielding position?

- For example, on average, do center fielders steal more bases than catchers?
- Stolen base rate = SB/OB

When analyzing this data, what is a good first thing to do?

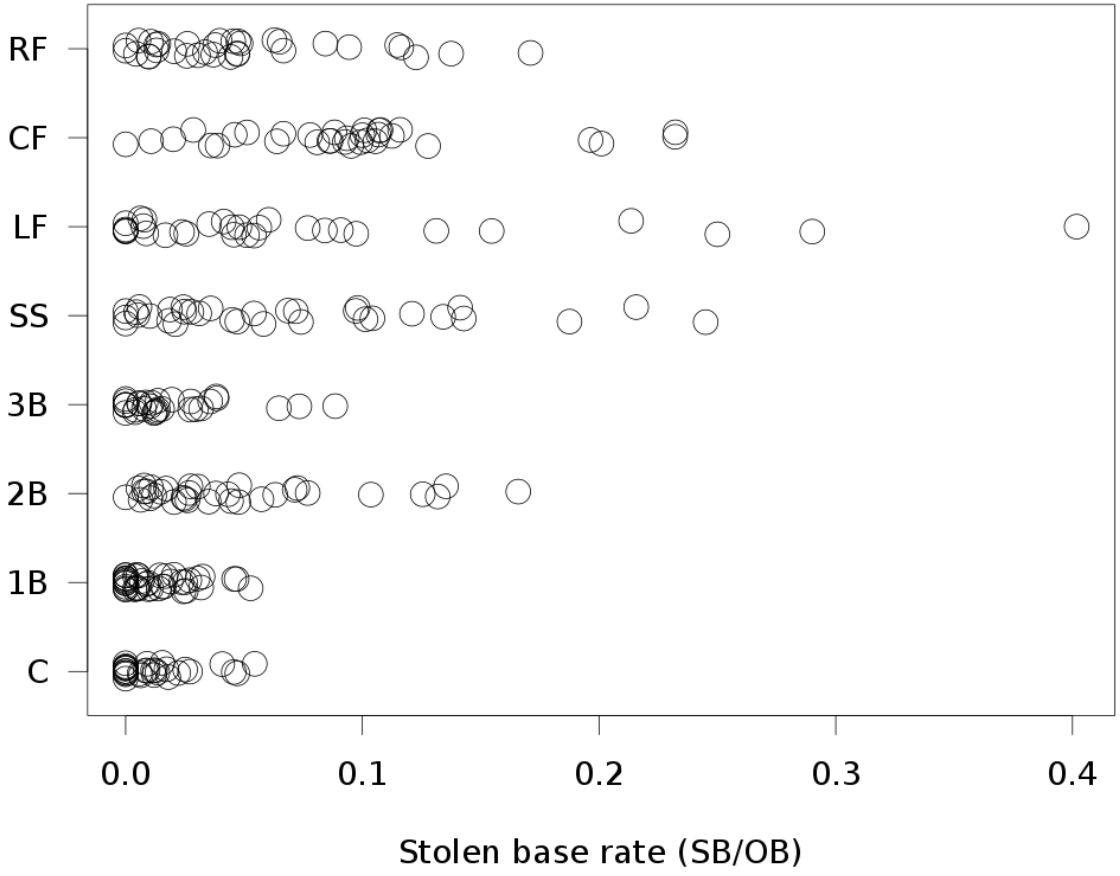


# Comparing multiple means




Data is from 2013 for all players who had over 300 PA.

# Comparing multiple means



# Comparing multiple means

Let‘s do a hypothesis test to tell if all the means are different

What are the null and alternative hypotheses?

- $H_0$: $\mu_c = \mu_{1B} = \mu_{1B} = \ldots = \mu_{RF}$
- $H_A$ : $\mu_i \neq \mu_j$ for some i and j

What is the observed statistic?

# Comparing multiple means

There are many possible statistics we could use. A few choices are:

1. Group range statistic:
   - max $\bar{x}$ - min $\bar{x}$

2. Mean absolute difference (MAD):
   - $(|\bar{x}_c - \bar{x}_{1B}| + |\bar{x}_c - \bar{x}_{2B}| + \ldots + |\bar{x}_c - \bar{x}_{Lf}| + \ldots + |\bar{x}_{Rf} - \bar{x}_{Lf}|)/28$

3. F statistic:

$$F = \frac{between\text{-}position\ \ variability}{within\text{-}position\ \ variability}$$

# Comparing multiple means

Let's use the mean absolute difference (MAD):

- $(|\bar{x}_c - \bar{x}_{1B}| + |\bar{x}_c - \bar{x}_{2B}| + \ldots + |\bar{x}_c - \bar{x}_{Lf}| + \ldots + |\bar{x}_{Rf} - \bar{x}_{Lf}|)/28$

Suppose we have the values from all players in a data frame position.statistic.df

| | playerID | position | SB.rate |
|---|---|---|---|
| **1** | ackledu01 | secondBase | 0.0148 |
| **2** | adamsma01 | firstBase | 0.0000 |
| **3** | alonsyo01 | firstBase | 0.0469 |
| **4** | altuvjo01 | secondBase | 0.1659 |
| **…** | **…** | **…** | **…** |

What do we do as step 2?

- MAD = .0366

# Comparing multiple means

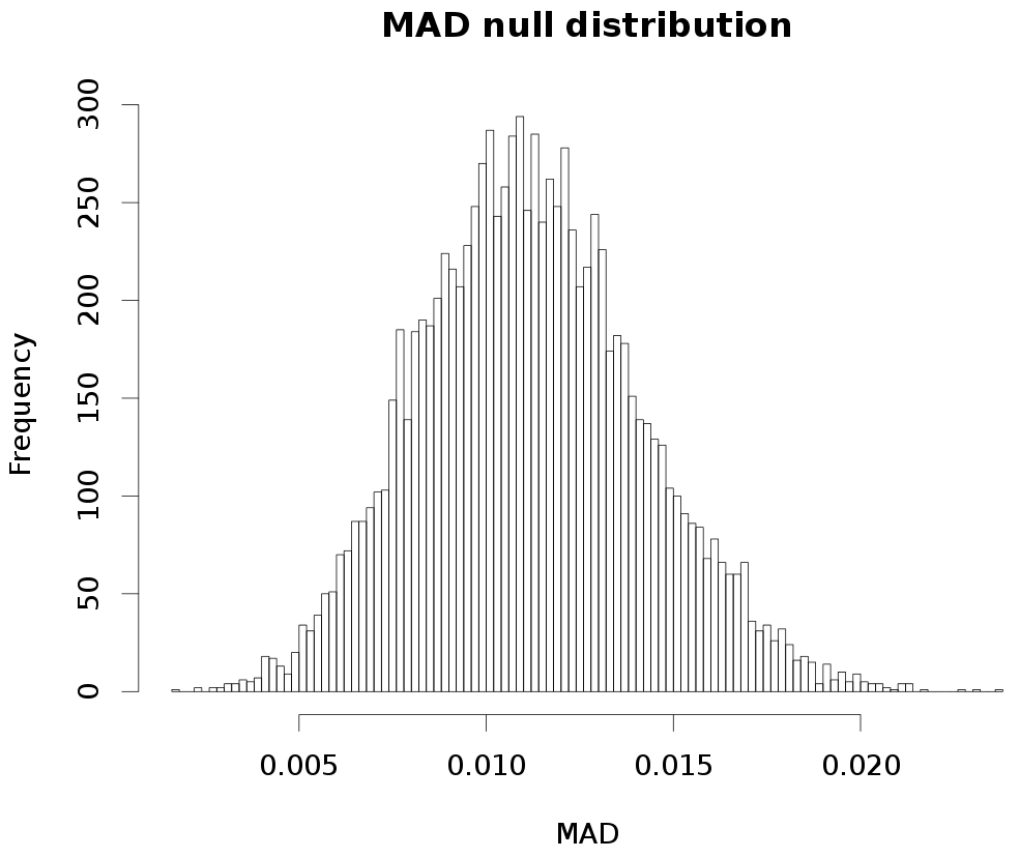
How can we create a null distribution?

| | playerID | position | SB.rate |
|---|---|---|---|
| **1** | ackledu01 | secondBase | 0.0148 |
| **2** | adamsma01 | firstBase | 0.0000 |
| **3** | alonsyo01 | firstBase | 0.0469 |
| **4** | altuvjo01 | secondBase | 0.1659 |
| **…** | … | … | … |

1. Shuffle the position labels
2. Recalculate MAD statistic with shuffled labels
3. Repeat 10,000 times

```
load("/home/shared/baseball_stats_2017/player_position_stats_300PA_2013.rda")
source("/home/shared/baseball_stats_2017/get_MAD_statistic.R")
get_MAD_statistic(position_stats_df$position, position_stats_df$SB.rate)
```

# Comparing multiple means




What is our p-value?

- Observed statistic was .0366, sooo….

What is our conclusion?

# Comparing multiple means

```
# get our observed statistic
obs.statistic <- get_MAD_statistic(position_stats_df$position, position_stats_df$SB.rate)
null.dist <- NULL

for (i in 1:10000) {


    # shuffle the position labels
    position =     shuffle(position_stats_df$position)

   # get our statistic with the labels shuffled
  null.dist[i] <-  get_MAD_statistic(position, position_stats_df$SB.rate)
}


p-value <- sum(null.dist >= obs.statistic)/10000
```

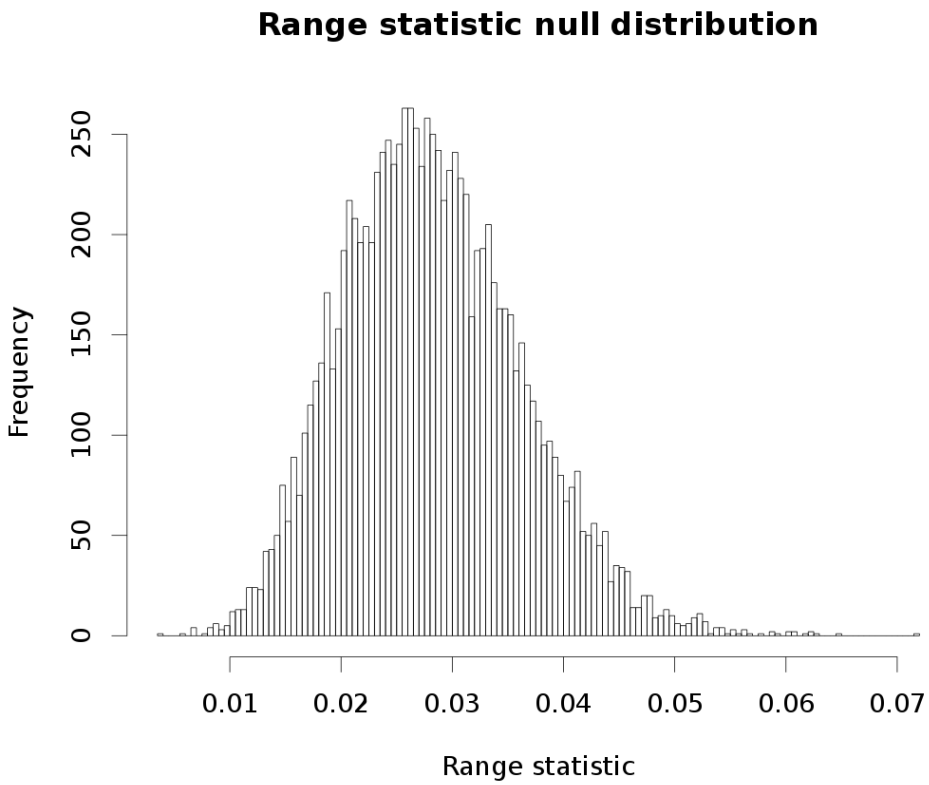
# Calculating MAD in R

```
get.MAD.statistic <- function(position.statistic.df) {

  position.statistic.df <- group_by(position.statistic.df, position)      # get the mean statistic value for each position
  mean.df <- summarize(position.statistic.df, group.means = mean(statistic.value))
  group.means <- mean.df$group.means
  num.groups <- length(group.means)

  # calculate the MAD statistic
  MAD <- 0
   for (i in 1:(num.groups -1))  {
        for (j in (i + 1):num.groups){
             MAD <- MAD + abs(group.means[i] - group.means[j])
         }
    }

    MAD <- MAD/choose(num.groups, 2)
}
```

# Comparing multiple means

<u>Using the group range statistic:</u> max $\bar{x}$ - min $\bar{x}$

- Observed statistic: .0813



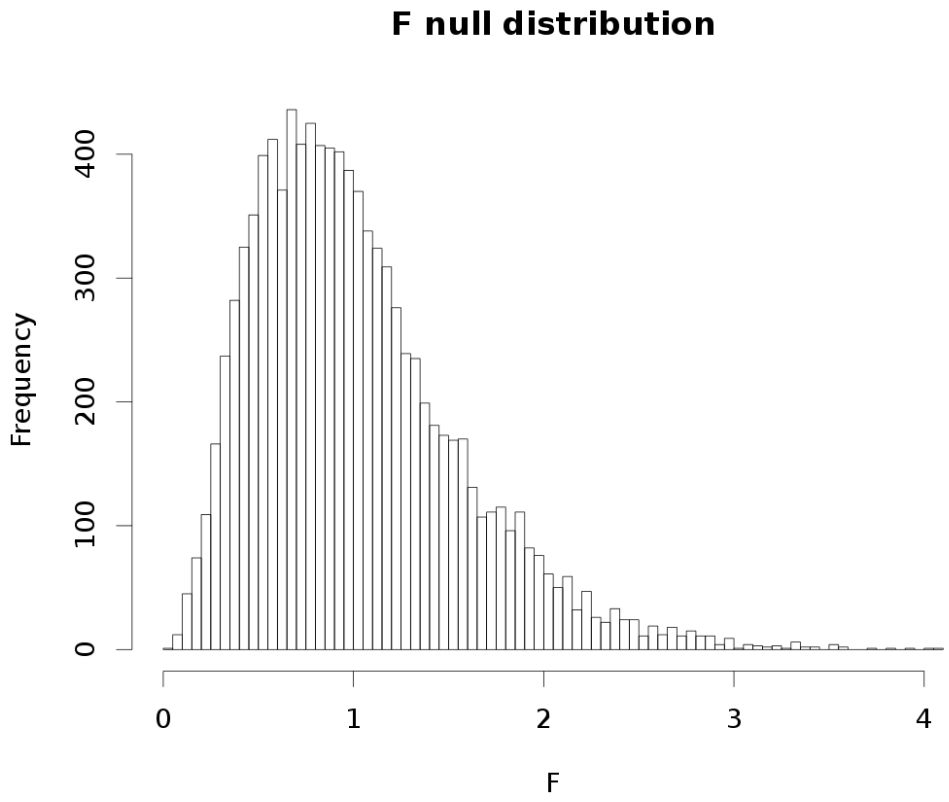

p-value and conclusion?

# Comparing multiple means

F statistic:

- Observed statistic: 12.30

$$F = \frac{\frac{1}{7}\sum_{pos=2}^{9} n_{pos}\,(\overline{x_{pos}} - \overline{x_{tot}})^2}{\frac{1}{N-8}\sum_{pos=2}^{9}\sum_{i=1}^{n_{pos}}(x_{pos,i} - \overline{x_{pos}})^2}$$




P-value and conclusion?

# Parametric test for multiple means

There is also a parametric hypothesis test for multiple means, which is called an **analysis of variance** (ANOVA)

What are the null and alternative hypotheses for an ANOVA?

- $H_0$: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
- $H_A$ : $\mu_i \neq \mu_j$ for some i and j

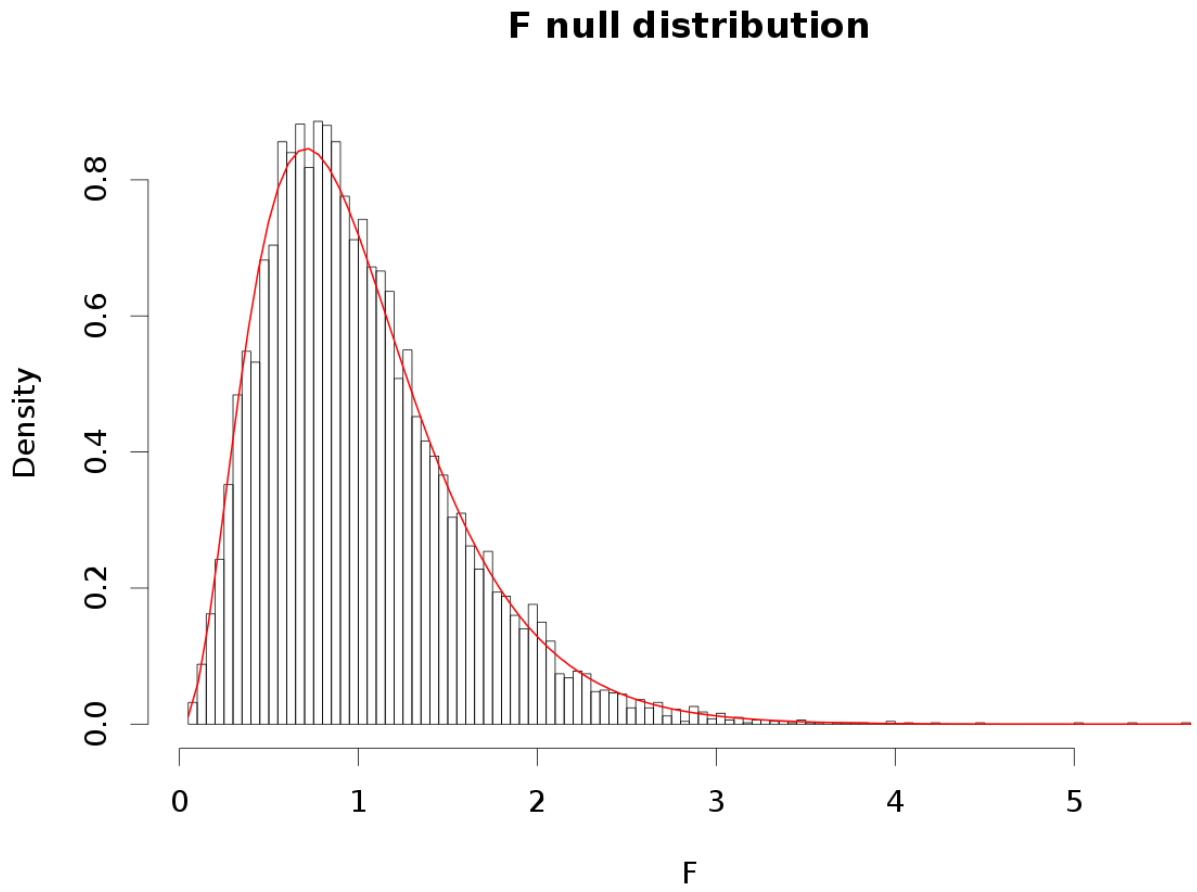
For an ANOVA F-statistic, and if $H_0$ is true, the null distribution is an F-distribution

- The F-distribution has parameters (called "degrees of freedom) that depend on the number of means being tested (k) and the total number of data points

How can we get a p-value?

- We can assess the probability of our observed F-statistic is likely to come from the F-distribution

# Parametric hypothesis tests for more than two means




Where is 12.3 on this plot?

# Confidence Intervals

A **confidence interval** is an interval computed (from sample data) by a method that will capture a *population parameter* a specified proportion of times

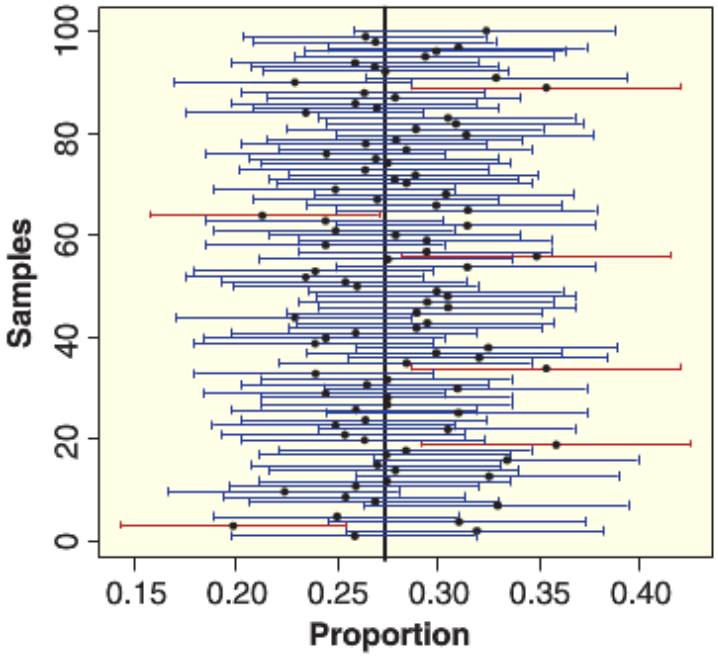
For example, suppose we want to estimate the proportion of soccer games that Paul will guess correct $\pi$

- Note that $\pi$ is a parameter
  - i.e., it says how many games Paul would guess correctly if he made an infinite number of guesses
- We could use the observed number of games Paul actually guessed correctly ($\hat{p} = 11/13$) as a **point estimate** of $\pi$
  - This estimate is not likely completely accurate
- Instead what we will do is come up with a range of values that is likely to contain $\pi$
  - i.e., we will construct a range of values in such a way that it will contain true $\pi$'s most of the time

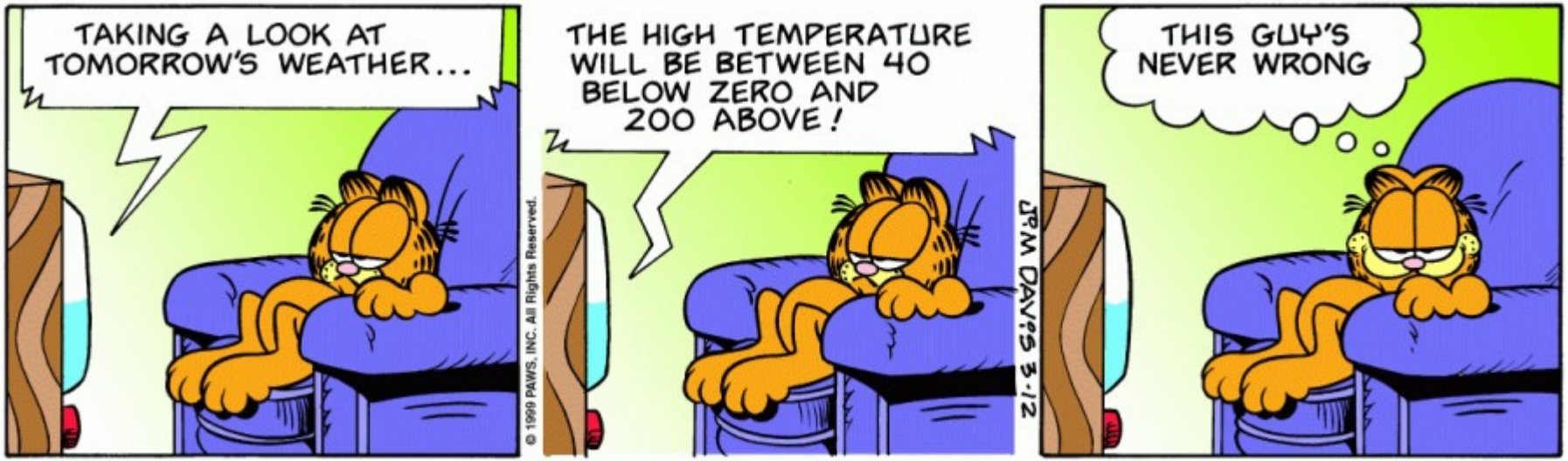
# Confidence Intervals

Stated another way: If we apply this procedure 100 times, we will capture the population parameter say 95% of those times

- This would be a 95% confidence interval



# Wits and wagers


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JIM DAVIS 3-12

# Think ring toss…

Parameter exists in the ideal world

We toss intervals at it

95% of those intervals capture the parameter