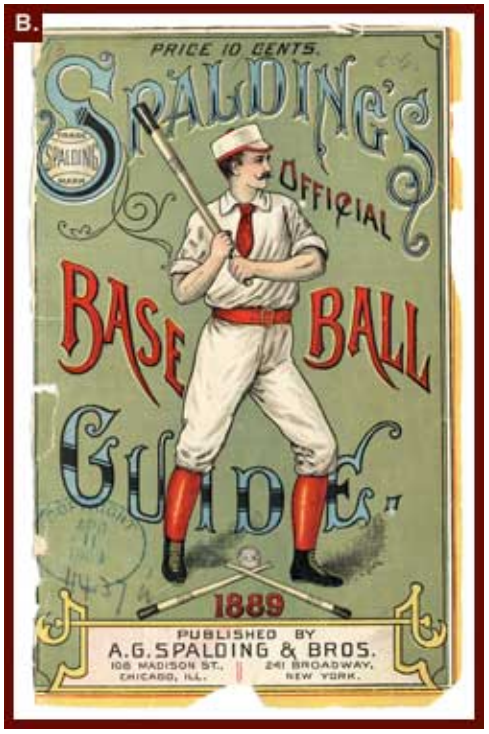
Review

Early statistics to sabermetrics

Who is Henry Chadwick?

- (1824-1908) created the first boxes core in the 1859 issue of Clipper.



BOSTON.	T.	R.	1B.	PO.	A.	E.
G. Wright, s. s	6	4	4	1	5	2
Leonard, 2 b.	6	3	3	4	4	3
O'Rourke, lb	6	2	3	9	0	1
Murnan, l. f.	6	1	0	3	1	0
Schafer, 2d b	6	3	3	3	1	2
McGinley, c. f	6	0	0	0	0	1
Manning, r. f.	6	0	2	2	0	0
Morrill, c.	6	2	2	4	1	2
Josephs, p.	5	4	4	1	1	2
Totals.	53	19	21	27	13	13

ATHLETIC.	T.	R.	1B.	PO.	A.	E.
Force, s. s.	5	1	2	1	3	2
Eggler, c. f.	5	3	3	0	0	0
Fisler, r. f.	5	0	1	2	0	0
Meyerle, 3db	5	1	2	2	3	3
Sutton, 1st b.	5	1	2	10	0	0
Coons, c.	5	1	0	1	1	3
Hall, l. f.	5	1	3	5	0	0
Fowser, 2d b.	6	1	2	6	7	5
Knight, p.	5	2	2	0	1	2
Totals.	46	11	17	27	15	15

Boston..... 9 1 3 3 4 1 0 2 5—19
Athletic..... 1 0 0 0 3 3 2 2 0—11

Runs earned—Boston, 4; Athletic, 5. Home-run—Hall, 1. Total bases on hits—Boson, 22; Athletic, 20. First base by errors—Boston, 8; Athletic, 5. Umpire, George White of Lowell, Mass. Time 2h. 47m.

	playerID	yearID	stint	teamID	lgID	G	G_batting	AB	R	H
1	aardsda01	2004	1	SFN	NL	11	11	0	0	0
2	aardsda01	2006	1	CHN	NL	45	43	2	0	0
3	aardsda01	2007	1	CHA	AL	25	2	0	0	0
4	aardsda01	2008	1	BOS	AL	47	5	1	0	0
5	aardsda01	2009	1	SEA	AL	73	3	0	0	0
6	aardsda01	2010	1	SEA	AL	53	4	0	0	0
7	aardsda01	2012	1	NYA	AL	1	NA	NA	NA	NA
8	aaronha01	1954	1	ML1	NL	122	122	468	58	131

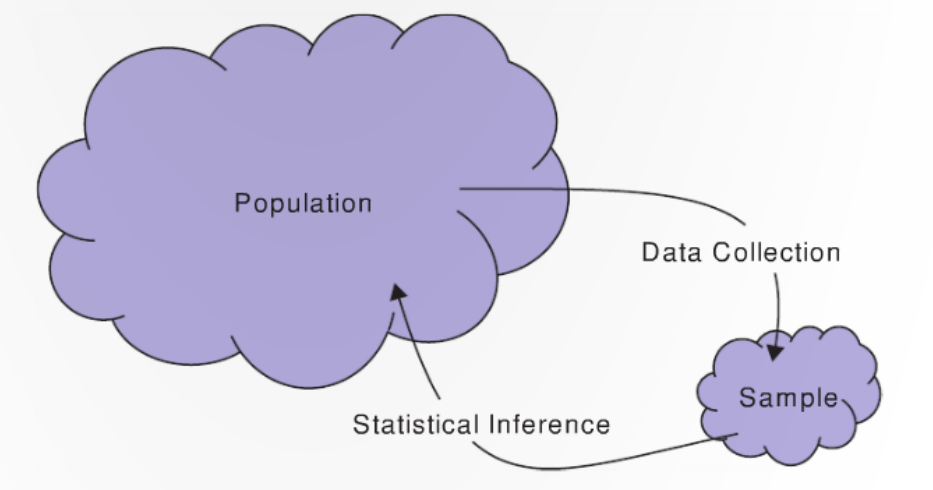
Sample from a Population

Population: all objects of interest (or an infinite processes)

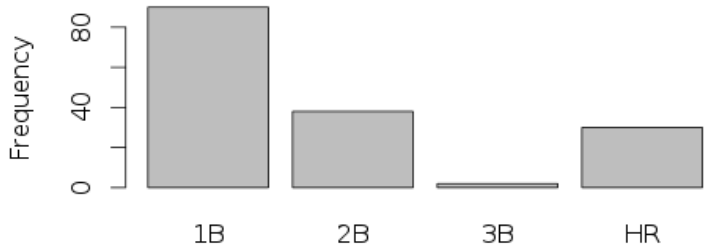
Sample: A subset of the population

statistic: A number that is computed from data in a sample

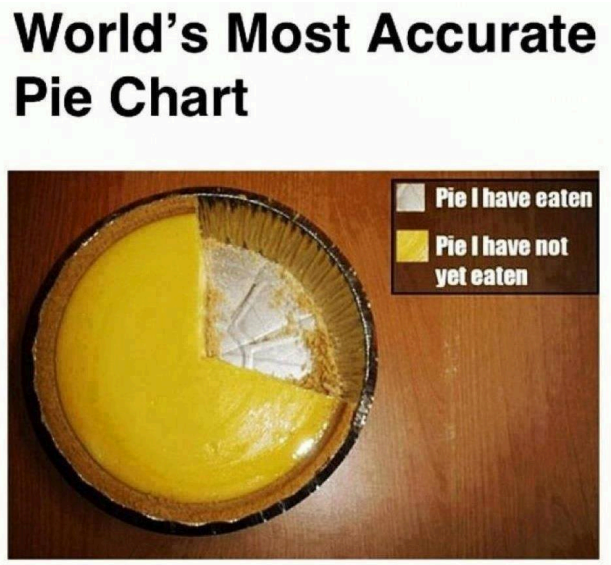
Statistical inference: Making judgments about the population using data from the sample.



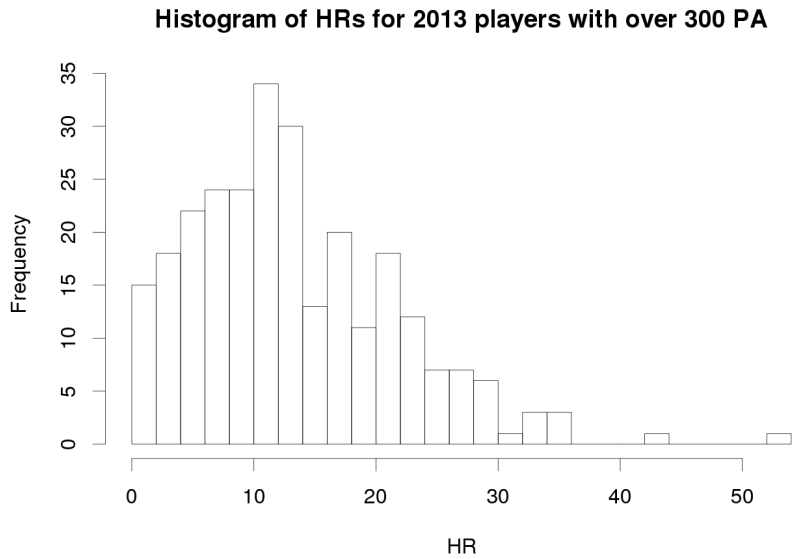
Plotting data



`R: barplot(x)`



`R: pie(x)`

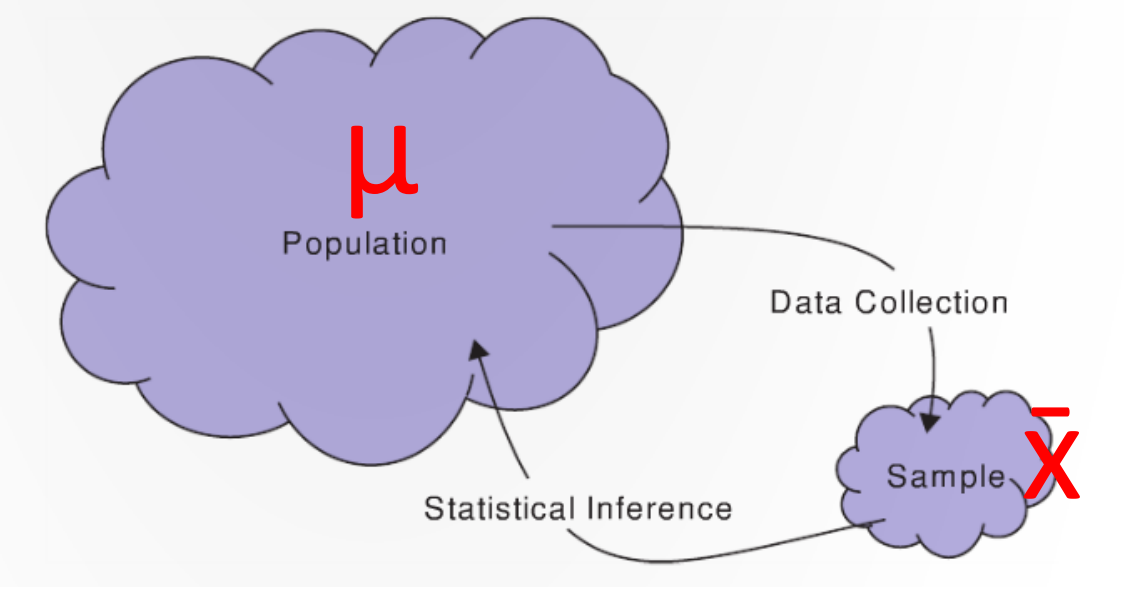


`R: hist(x)`

Sample vs. Population mean

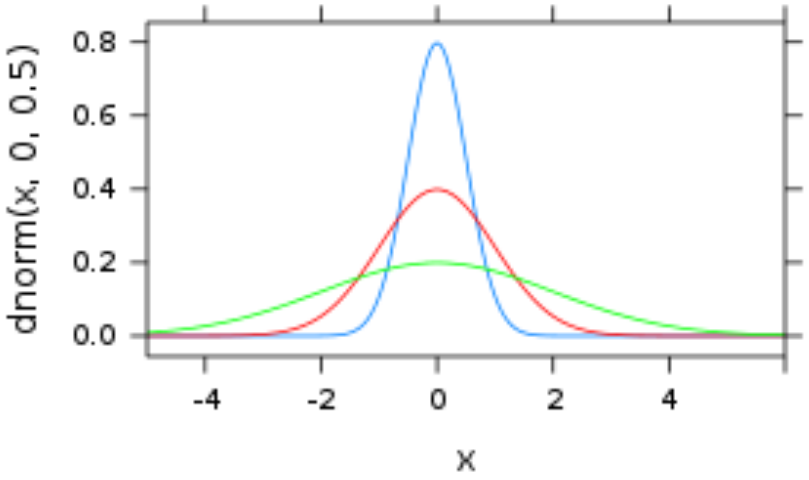
The mean for a ***sample*** is denoted $\bar{x}$ (pronounced "x-bar")

The mean for a ***population*** is denoted μ, which is the Greek letter "mu"

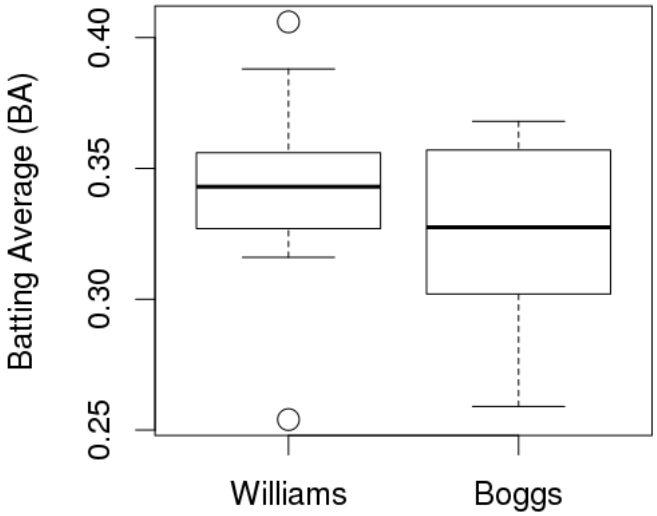


$$\bar{x} = \frac{1}{n}\sum_i^n x_i$$

Describing variability



Same mean, different standard deviation

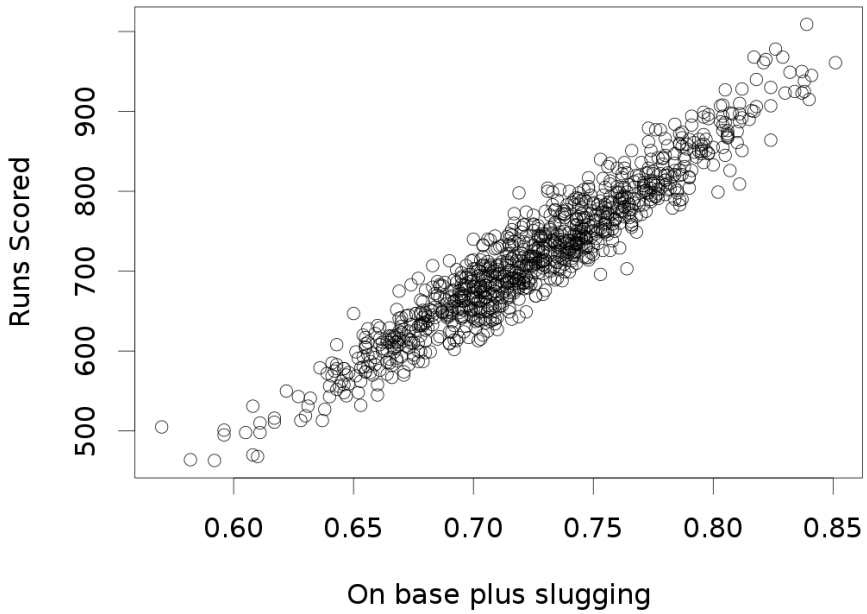


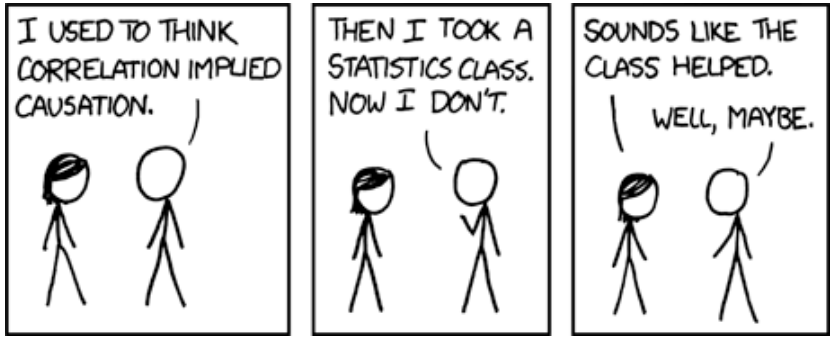
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

$$z_i = \frac{x_i - \bar{x}}{s}$$

Correlation

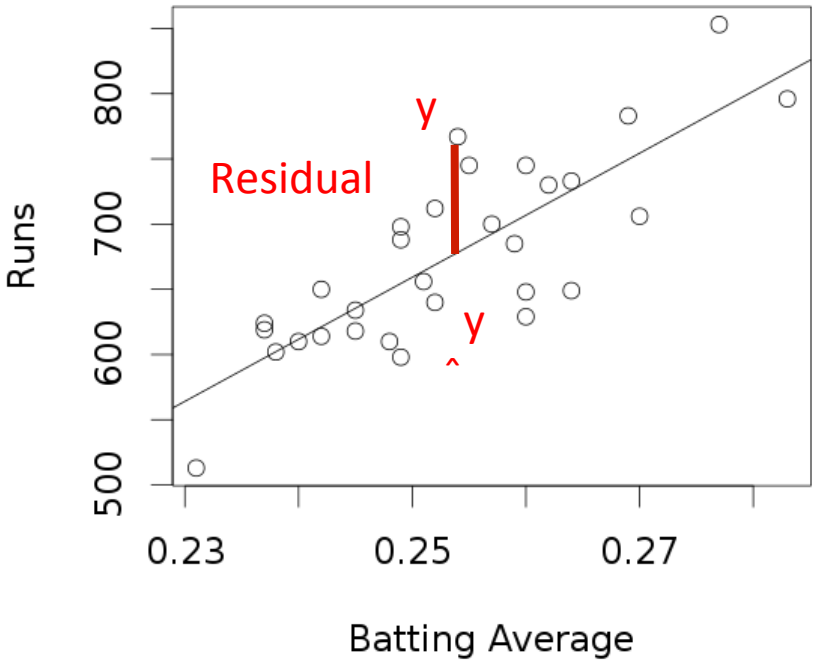
What is the correlation value (r) in this plot?

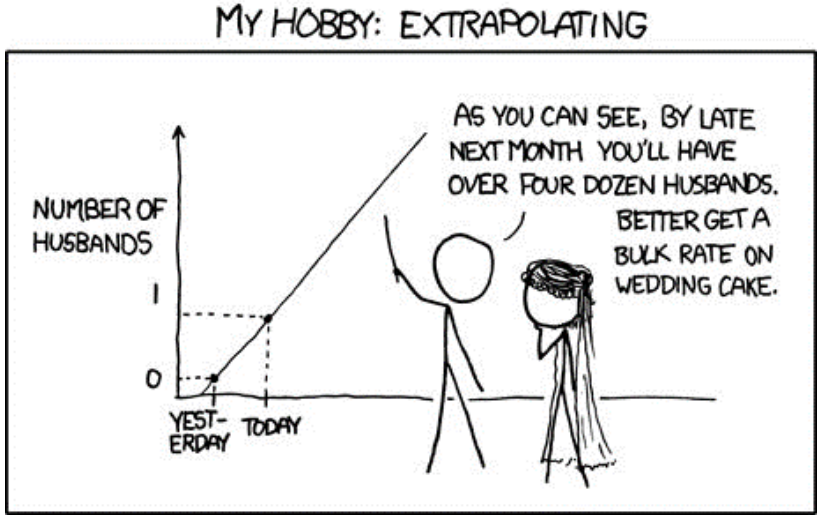




$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

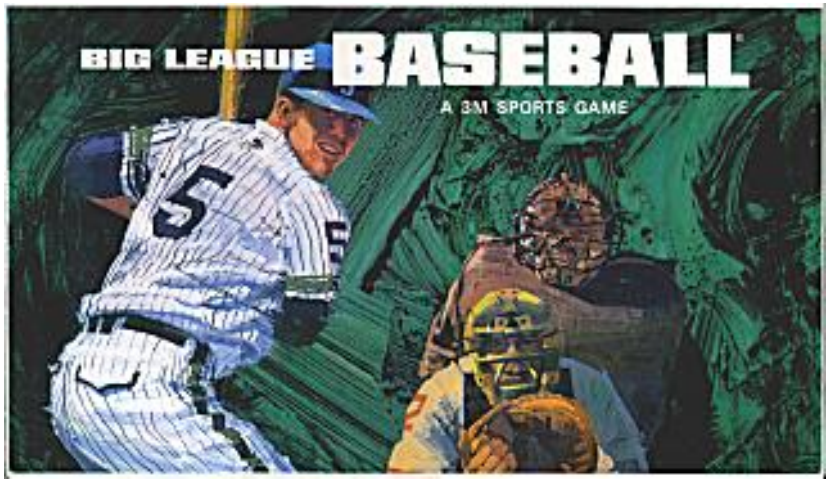
Regression





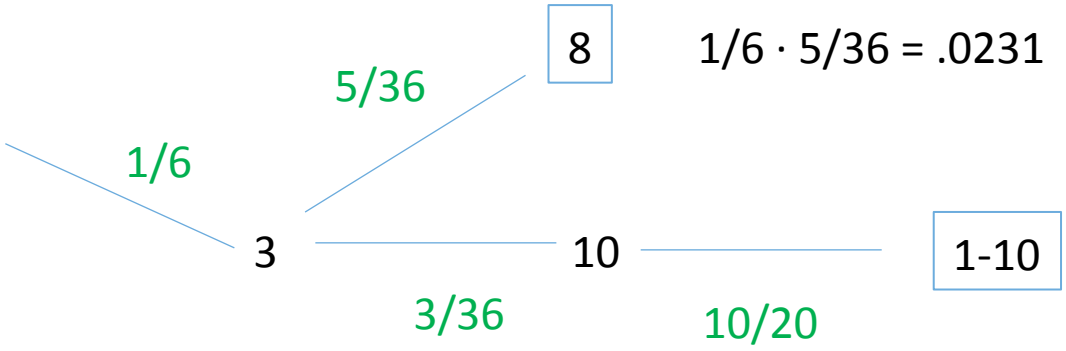
$$MSE = \frac{1}{n}\sum_i^n (\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_i^n (a + b \cdot x_i - y_i)^2$$

Probability



What 2 probability rules did we discuss?

1) Additive rule
2) Multiplicative rules



$$Pr(X = k) = f(k; n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

Statistical inference

What is statistical inference?

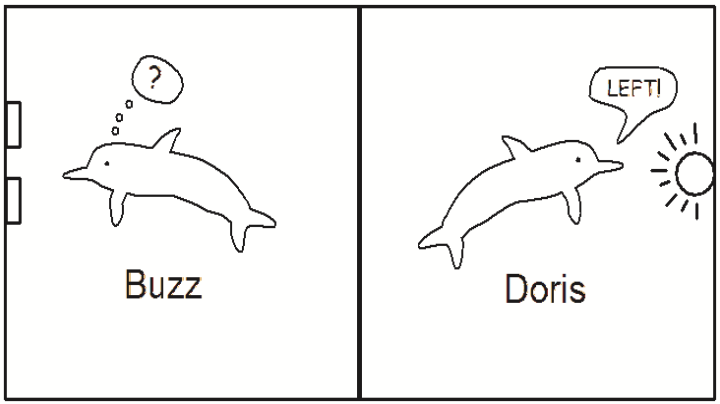
Use sample of data to deduce properties of an underlying population, or stochastic process

- E.g., looking at a player's **performance** to estimate their **ability**

Hit, Out, Hit, Out, Out, Out, …

Estimate π_{hit}

Review: hypothesis tests for a single proportion



What are the steps in hypothesis testing?

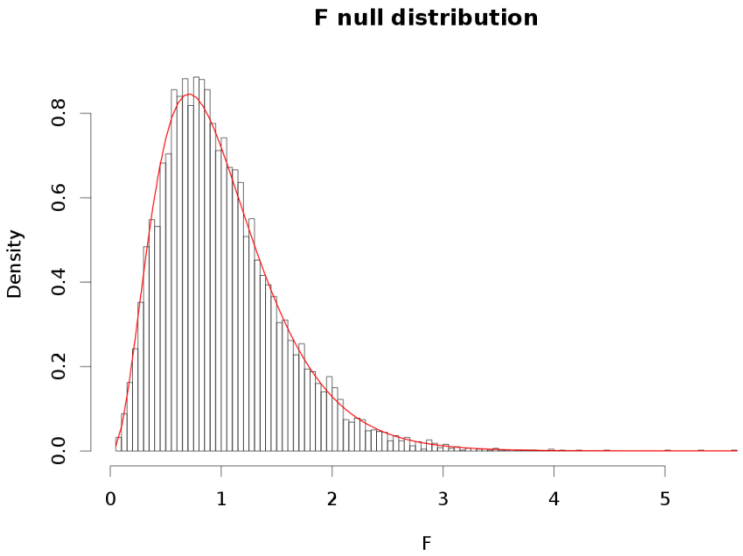
1. State the null and alternative hypothesis
   - H_0: $\pi = .25$
   - H_A: $\pi > .25$

2. Calculate the observed statistic

3. Create a null distribution

4. Create a p-value by calculating the probability of getting a statistics as great or greater than the observed from the null distribution

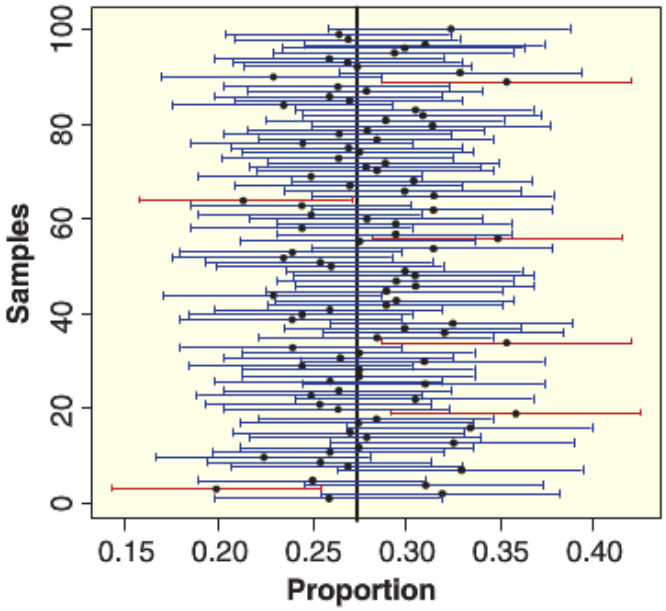
More hypothesis tests



Question: Does this steroid work?

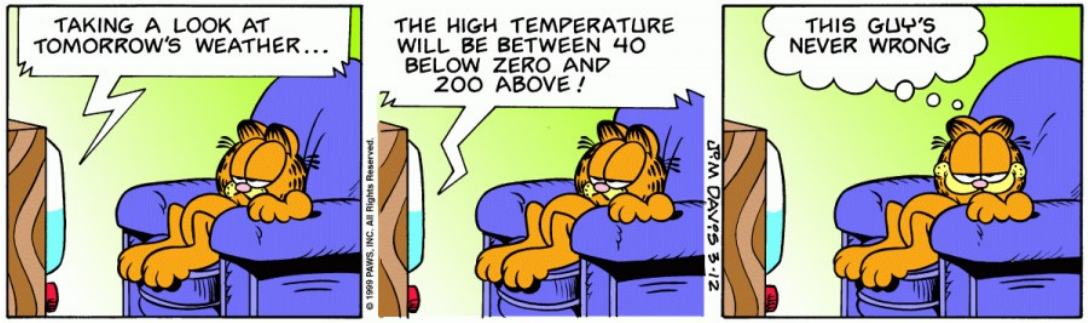
Permutation test: computer randomized methods)
Parametric tests: use probability theory to get p-values

Confidence Intervals



A 95% confidence interval is: $\hat{p} \pm 2 \cdot SE$ where

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



We learned a lot of R

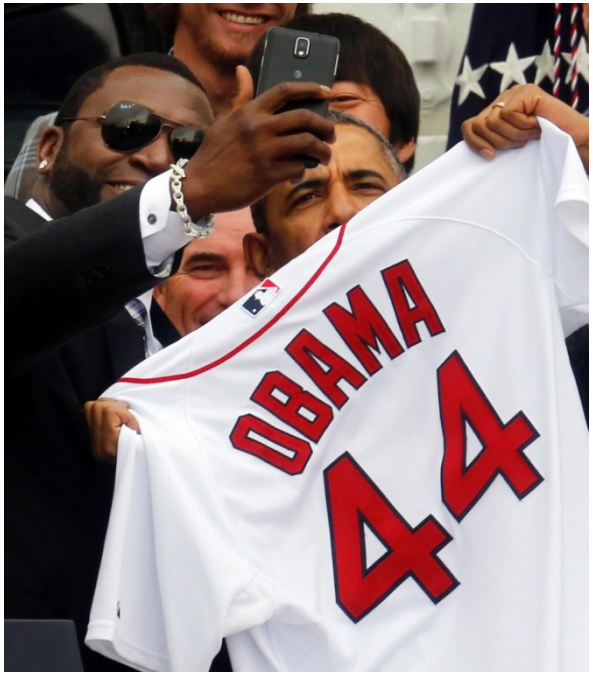
Q: What programming language do the pirates use?

A: Arrrr

Q: Worst joke of the semester?

A: ???

We had fun!



Right?

Have a good summer!

If you're looking for something fun to do over the summer…

Challenge questions!

The challenge questions can be downloaded by typing:

> get.worksheet(11)