Hypothesis tests for comparing two proportions and two means

Outline for today

Better know a player Masanori Murakami

Review of hypothesis tests for two proportions

Hypothesis tests for two means

Worksheet 9

Better know a player

Masanori Murakami

MLBAM @ 2014 MIT Sloan Analytics Conference

Big Data Baseball Page 174

Analyzing a catch by Jason Heyward

Steps for doing a hypothesis test

1. State the null and alternative hypothesis

2. Calculate the observed statistic

3. Create a null distribution
   - Typical statistics you would expect to get if the null hypothesis was true

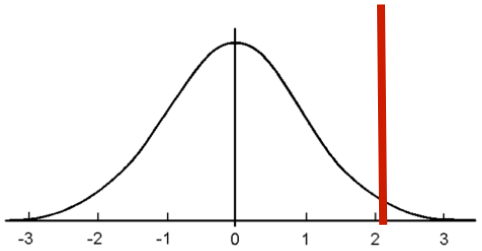
4. Create a p-value
   - calculate the probability of getting a statistics as great or greater than the observed from the null distribution

Five steps of hypothesis testing

1. Assume innocence: H_0 is true
   - State H_0 and H_A

2. Gather evidence
   - Calculate the observed statistic

3. Create a distribution of what evidence would look like if H_0 is true
   - Null distribution



4. Assess the probability that the observed evidence would come from the null distribution
   - p-value

5. Make a judgement
   - Assess whether the results are statistically significant

Testing whether two proportions differ

Question: Is Alex Rodriguez's OBP ***ability*** higher than Mario Mendoza's?

- i.e., if they had infinite plate appearances, who is better?

A-Rod and Mendoza's ***performance*** through the 2014 season is:

	Mendoza	**A-Rod**
On-base	345	4348
Plate appearances (minus SH)	1407	11328
OBP	0.245	0.384

Is A-Rod's OBP ability better than Mendoza's

We can do a hypothesis test whether A-Rod's OBP ability is better than Mendoza's ability

1. State the null and alternative hypotheses in symbols and words

- H_0: $\pi_{ARod} = \pi_M$ or $\pi_{ARod} - \pi_M = 0$
- H_A: $\pi_{ARod} > \pi_M$ or $\pi_{ARod} - \pi_M > 0$

What do we do next?

2. Observed statistic is: .384 - .245 = .139

Now what do we do?

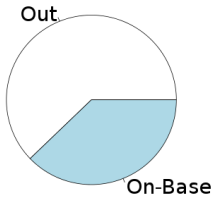
Is A-Rod's OBP ability better than Mendoza's

How can we create a ***null distribution*** of this statistic that is consistent with the null hypothesis?

We can use simulations!

	Mendoza	A-Rod
On-base	345	4348
Plate appearances (minus SH)	1407	11328
OBP	0.245	0.384

Total On-base = 4693 Total PA = 12736 Total OBP = .378

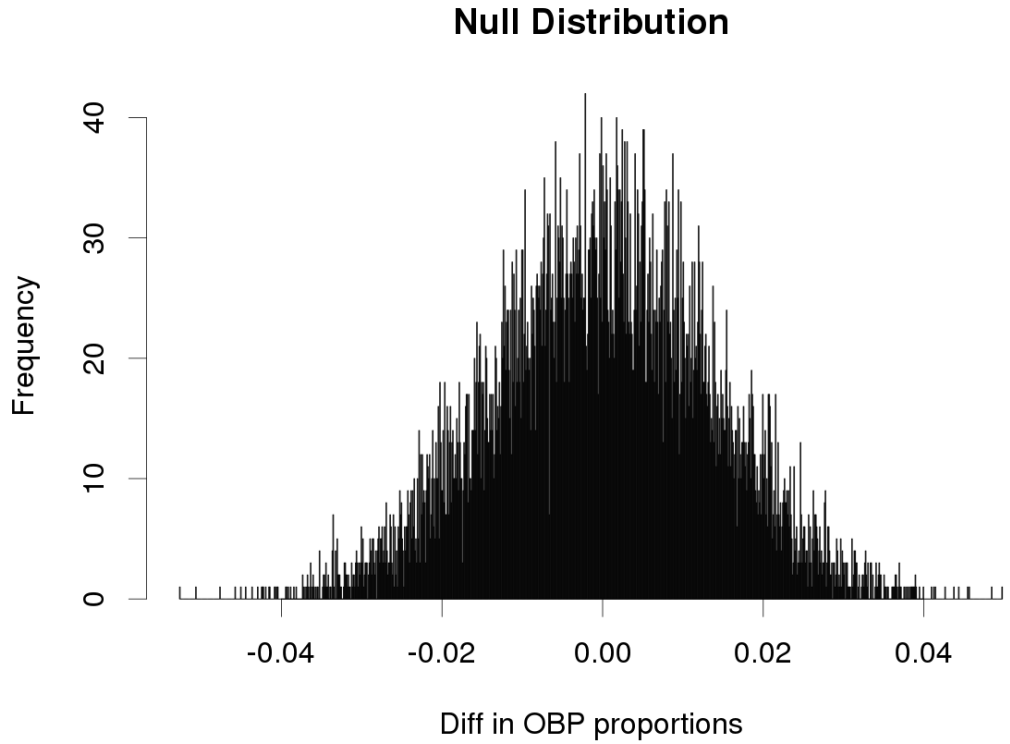


<u>Simulation:</u>

1. Flip a coin 1407 times for Mendoza and 12505 for A-Rod
2. Compute difference in proportion of on-base events
3. Repeat 10,000 times

Is A-Rod's OBP ability better than Mendoza's

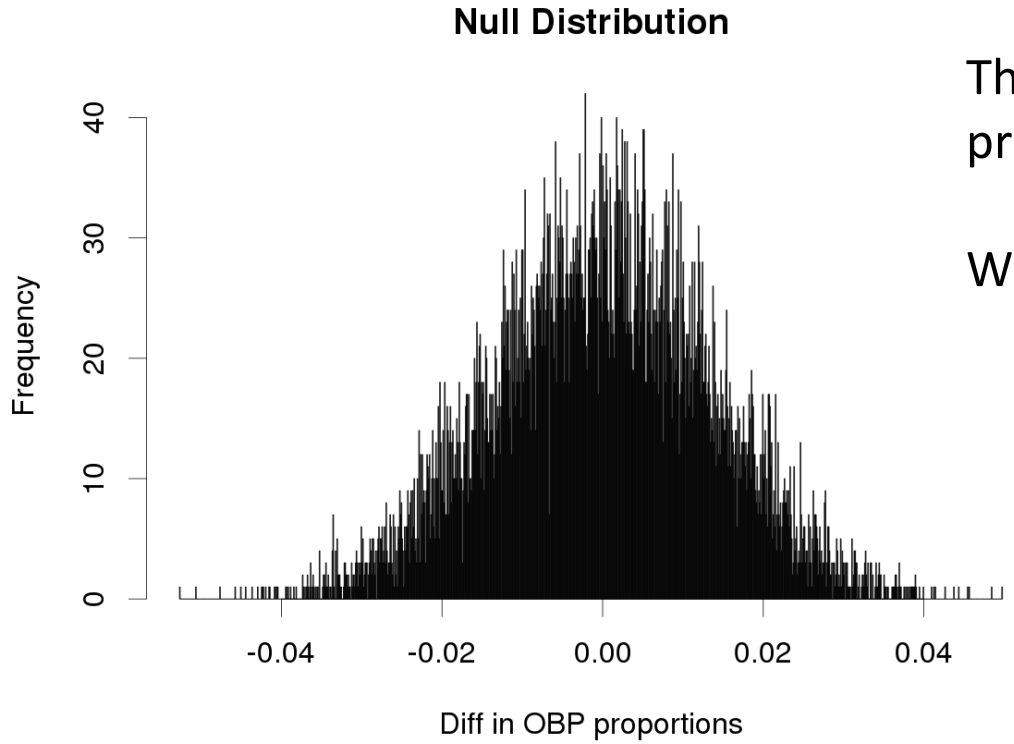
Results of the simulation…



What do we do next?

Is A-Rod's OBP ability better than Mendoza's

Results of the simulation…



The observed difference in proportions was . 139

Where is this .139 on this plot?

Is A-Rod's OBP ability better than Mendoza's

0 of the 10,000 different in simulated proportions were greater than the observed different in OBP proprotions

so the p-value is…

p-value = 0

Conclusion?

Creating the null (randomized) distribution in R

	Mendoza	A-Rod
On-base	345	4348
Plate appearances (minus SH)	1407	11328
OBP	.245	.384

```
ARod.PA <-  11328
Mendoza.PA  <- 1407

Total.PA <- 11328  +  1407          # total PA for both players
Total.OB <- 345 + 4348              # total times on-base for both players

total.OBP <- Total.OB/Total.PA      # OBP if both players had the same ability

obs.diff.OBP <- .384  - .245        # observed statistic of interest
```

Creating the null distribution in R

```
sim.ARod.OBP <- rbinom(10000, ARod.PA, total.OBP)/ARod.PA
sim.Mendoza.OBP <- rbinom(10000, Mendoza.PA, total.OBP)/Mendoza.PA

null.dist.vec  <-  sim.ARod.OBP - sim. Mendoza.OBP
p.value   <-   sum(null.dist.vec >=   obs.diff.OBP)/10000
```

Comparing two means

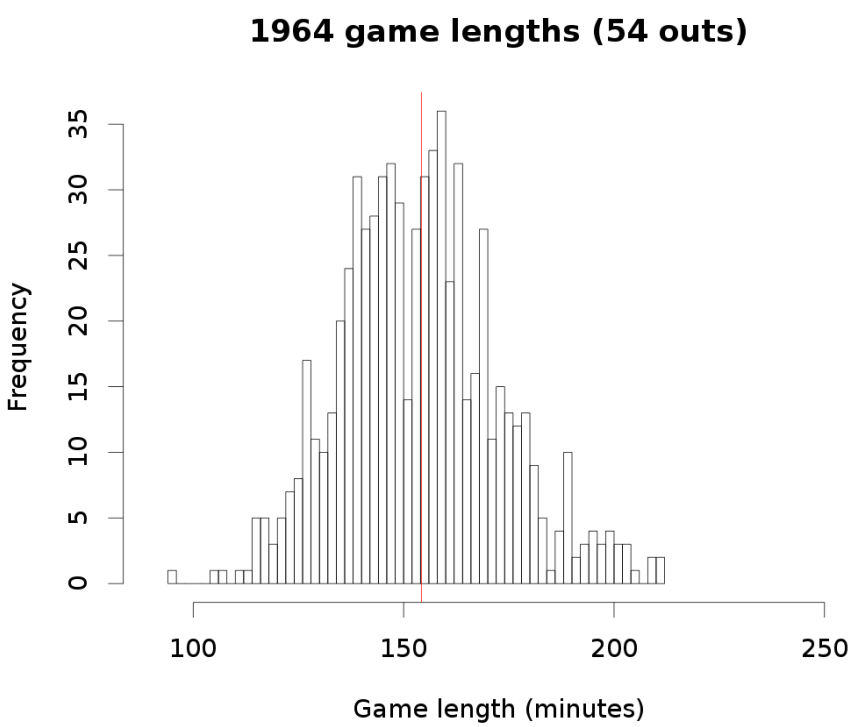
Have baseball games gotten longer in the past 50 years?

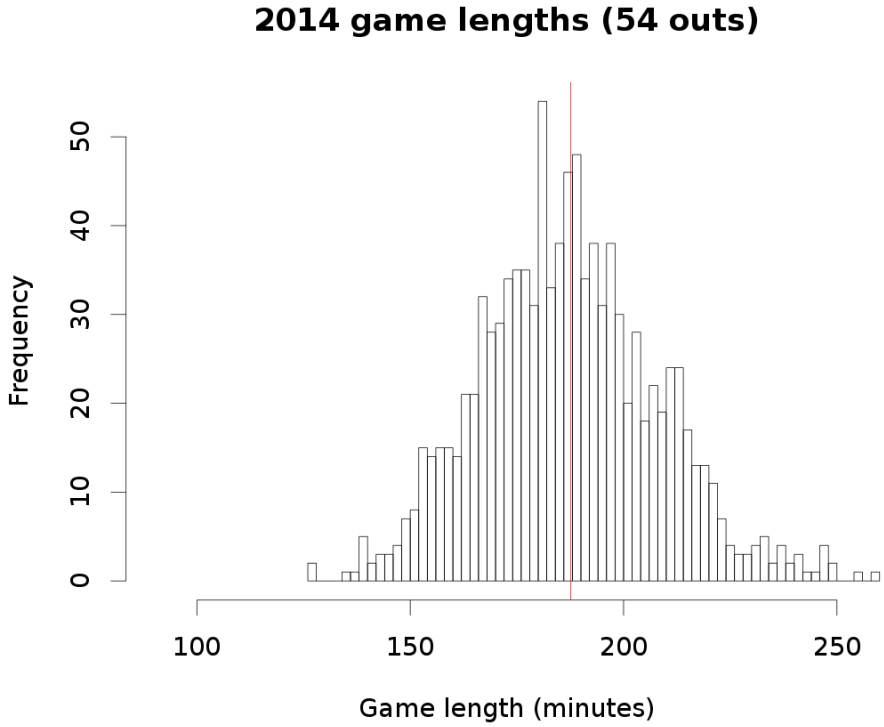
How could we examine this?

- Compare mean lengths of games in 1964 to those in 2014
- See Moodle:
  - load("/home/shared/all.game.logs.Rda")
  - game.logs$year <- substr(game.logs$Date, 1, 4)
  - game.logs54 <- filter(game.logs, LengthInOuts == 54)
  - game.logs.54.out.1964 <- filter(game.logs54, year == 1964)
  - game.logs.54.out.2014 <- filter(game.logs54, year == 2014)

What would be a good first thing to do?

Plot the data





Average game length 1964 is: $\bar{x}_{1964}$ = 154.21 minutes

- (based on **n = 684** games with 54 outs)

Average game length 2014 is: $\bar{x}_{2014}$ = 187.64 minutes

- (based on **n = 1021** games with 54 outs)

1. Null and Alternative Hypotheses

1a. State the null and alternative hypotheses in words

- **Null hypothesis**: Baseball games are the same length in 1964 as they are in 2014
- **Alternative hypothesis**: Baseball games are longer in 2014 than in 1964

1b. State the null and alternative hypotheses using symbols

- H_0: $\mu_{2014} = \mu_{1964}$ or $\mu_{2014} - \mu_{1964} = 0$
- H_A: $\mu_{2014} > \mu_{1964}$ or $\mu_{2014} - \mu_{1964} > 0$

What do we do next?

- 2. Compute the statistic of interest

What is the statistic of interest?

2. Compute the statistic of interest

Average game length 1964 is: $\bar{x}_{1964}$ = 154.21 minutes

- (based on **n = 684** games with 54 outs)

Average game length 2014 is: $\bar{x}_{2014}$ = 187.64 minutes

- (based on **n = 1021** games with 54 outs)

So the statistic of interest is…?

- observed.stat <- 187.64 - 154.21 = 33.42 minutes

What do we do next?

- 3. Create a null distribution

Hypothesis tests for two means

3. Calculate the null distribution

- How can we create a null distribution???

One way: under the null hypothesis all games lengths from 1964 and 2014 are equally likely

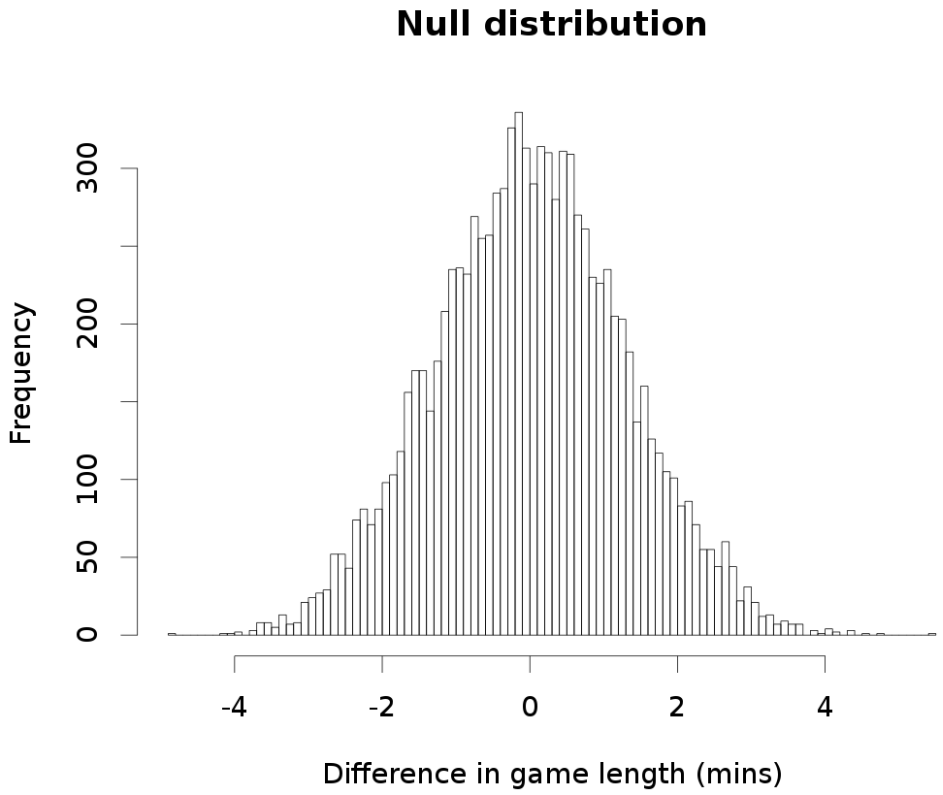
Thus combine all the games lengths from the 1964 and 2014 seasons into one vector

We can then randomly select **684** games to simulate the 1964 season and take the remaining **1021** to simulate the 2014 season

The difference in these means of these 684 and 1021 games gives us one point in the null distribution

If we repeat this 10,000 times we will get a full null distribution

Hypothesis tests for two means



Do the results seem statistically significant?

- Observed difference of 33 minutes is not even close to being on this figure
- Conclusions?

Implementing this permutation test in R

```
# game durations for 1964 and 2014
 game.duration.1964 <- game.logs.54.outs.1964$Duration
 game.duration.2014 <- game.logs.54.outs.2014$Duration

# number of games in 1964 and 2014
   num.games.1964 <- length(game.duration.1964)
   num.games.2014 <- length(game.duration.2014)

# the observed statistic
   obs.diff <- mean(game.duration.2014) - mean(game.duration.1964)

# combine data from both seasons together
combined.durations <- c(game.duration.1964, game.duration.2014)
```

Implementing this permutation test in R

```
null.dist <- NULL

for (i in 1:10000)  {

      # shuffle the combined game durations
      shuffled.durations <- sample(combined.durations)

       # get the random durations for 1964 and 2014
       shuff.1964 <- shuffled.durations[1:num.games.1964]
       shuff.2014 <- shuffled.durations[(num.games.1964 +1) :length(shuffled.durations)]

       # calculate the observed statistic under the null hypothesis
      null.dist[i]  <- mean(shuff.2014) - mean(shuff.1964)
}

hist(null.dist, n = 100, main = 'Null Dist',  xlab = 'Mean diff'))       # display the null distribution
p.value <- sum(null.dist >= obs.diff)/10000
```

Worksheet 9

```
> source('/home/shared/baseball_stats/baseball_class_functions.R')

> get.worksheet(9)
```