Statistical Methods in Sports Analytics

DATA 299

Based on Ch. 4 in *Analytic Method in Sports, 2nd ed.*, Severini

Hitchman

Spring '25

Statistics model Reality

Our perspective in this chapter:

- One goal of analytic methods is to use data to draw conclusions about the process generating the data (Severini, p. 89)
- Statistics are simplifications of much more complex realities (Bill James, *The New Historical Baseball Abstract* (2001))

Random Variables in Sports

- We want to investigate the true ability of an NFL running back in a particular season.
- We decide to measure his true ability via a (theoretical) random variable Y representing yards gained on a carry that season.
- Actual Data: Let $Y_1, Y_2, \ldots Y_n$ denote the yards he gained on each of his n rushing attempts for the season.
- View $Y_1, Y_2, \ldots Y_n$ as an independent random sample drawn from pop'n of all (theoretical) rushing attempts for the season, as represented by the r.v. Y.

Estimating Y from Y_i

- Because true ability Y is unknown (we're trying to gauge this ability!) we look to the Y_i
- Sometimes estimation can be made on the analogy principle: To estimate a characteristic of Y we use the corresponding characteristic of the data:
  - We estimate $E(Y)$ by computing the mean of the data
  - We estimate standard deviation of Y, σ_Y, by computing the sample standard deviation.

Example: Jerome Ford

- Jerome Ford: running back (RB) for the Cleveland Browns
- Consider the random variable $Y =$ yards gained on a carry by Jerome Ford in 2024.
- We claim the (unknown) distribution of Y gives a measure of his true ability as an RB.
- In 2024 he had 104 carries for a total of 565 yards, for an average of 5.43 yards per carry.
- We let $Y_1, Y_2, \ldots, Y_{104}$ denote his rushing totals on these 104 carries.
- If we were able to replay the NFL season, we would not expect his average yards per carry to be exactly 5.43... but it will likely be close!
- How close?
- Let's simulate!

Intermission: The `nflfastR` package in R

The R package `nflfastR` provides us with an extremely rich NFL play-by-play database. It is free, popular, and allows you to answer a wide variety of questions, so we begin our football analytics journey here. It is far from the only source of football data out there - there's also Pro-Football-Reference, Open Source Football, the cfbfastR package for college football, among others.

Let's install `nflfastR` and familiarize ourselves with its pbp data sets.

```
install.packages("nflfastR")
```

```
library(nflfastR)
```

```
pbp <- load_pbp(2024)
```

Here's a beginner's guide to nflfastR

Gathering Jerome Ford's Carries

```
library(tidyverse)

Ford <- pbp |>
 filter(
  play_type == "run" &
  posteam == "CLE" &
  str_detect( desc, "J.Ford" )
 )

sum(Ford$yards_gained)

#check this seems to account for every carry on the season
```

Margin of Error

Margin of error is a statistical measure of what *close to* means in the context of "if we were able to replay the season, Ford's ability staying the same, his average yards per carry would likely be close to 5.43."

Quantifying Variation

If we were to magically replay the season many, many times, and Ford still gets 104 carries each time, how much variation in his season average yards per carry should we expect?

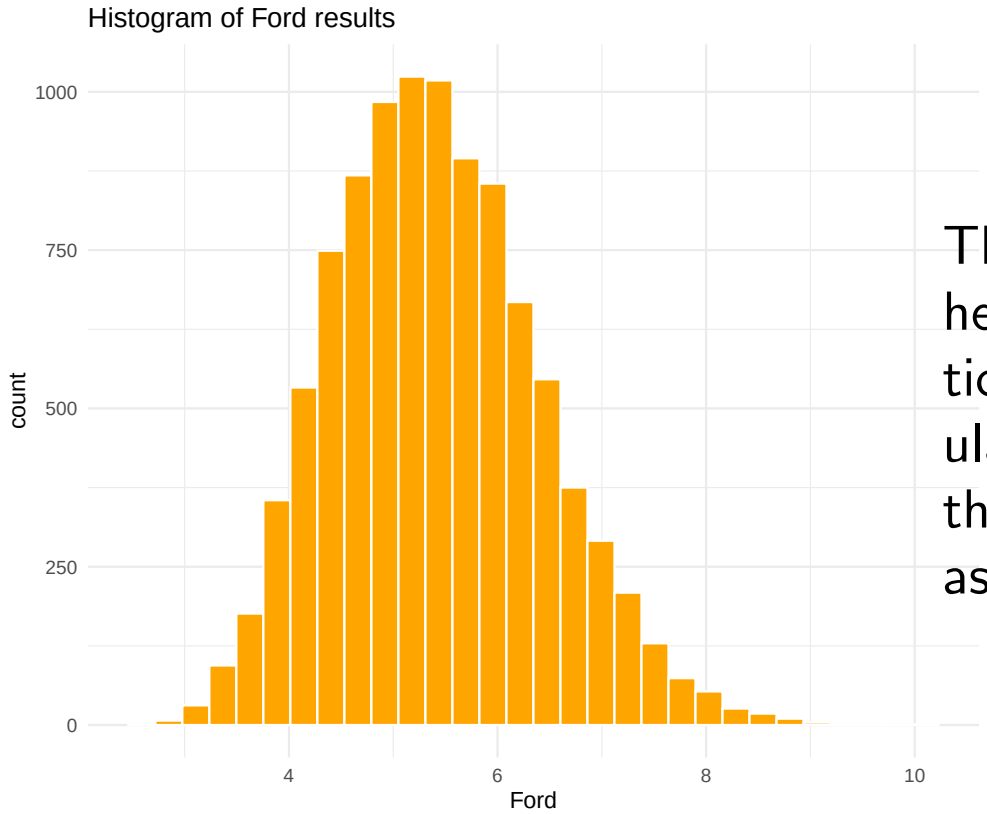
Two approaches:

- simulation
- theory (central limit theorem)

Quantify variation via simulation

The code below creates a *results* vector to store 10000 simulated Ford seasons, each season consisting of 104 carries, each carry randomly selected from his actual 104 carries! We'll do this in R.

```
results <- c()
for (i in 1:10000)
  results[i] <- mean(sample(Y,104,replace=TRUE))
```

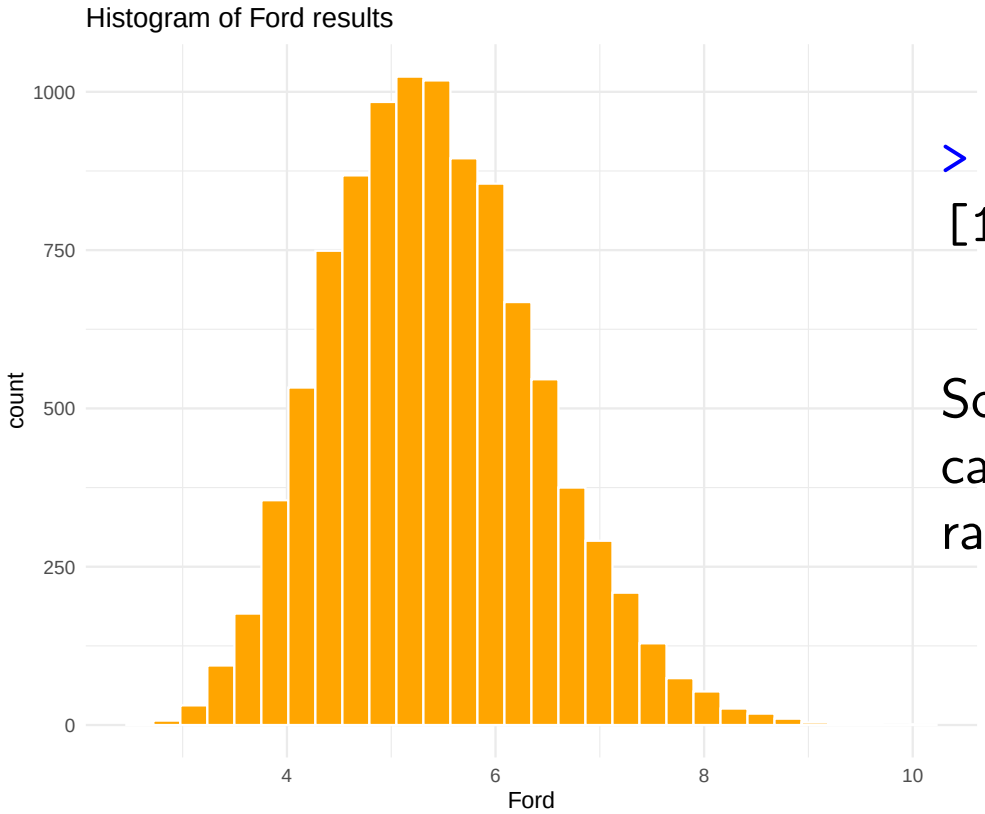


The (bell-shaped!) distribution of results helps us see the magnitude of the variation in yearly averages upon repeated simulation. We can **quantify variation** with the *margin of error*, defined in Section 4.2 as `2*sd(results)`.

Quantify variation via simulation

The code below creates a *results* vector to store 10000 simulated Ford seasons, each season consisting of 104 carries, each carry randomly selected from his actual 104 carries! We'll do this in R.

```
results <- c()
for (i in 1:10000)
  results[i] <- mean(sample(Y,104,replace=TRUE))
```




```
> 2*sd(results)
[1] 2.03088
```

So with high probability Ford's yards-per-carry value would fall somewhere in the range

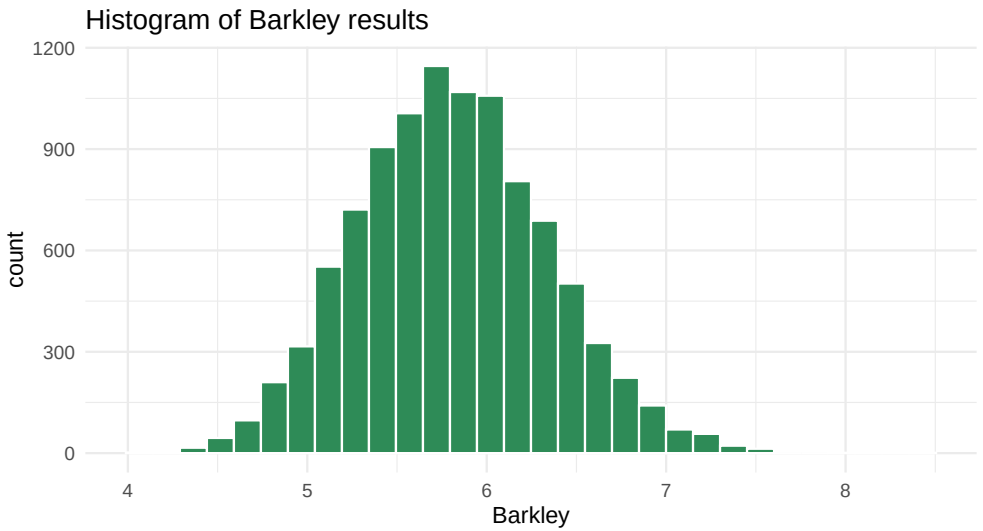
$$5.43 \pm 2.03 = (3.40, 7.46).$$

Using Data to estimate true ability

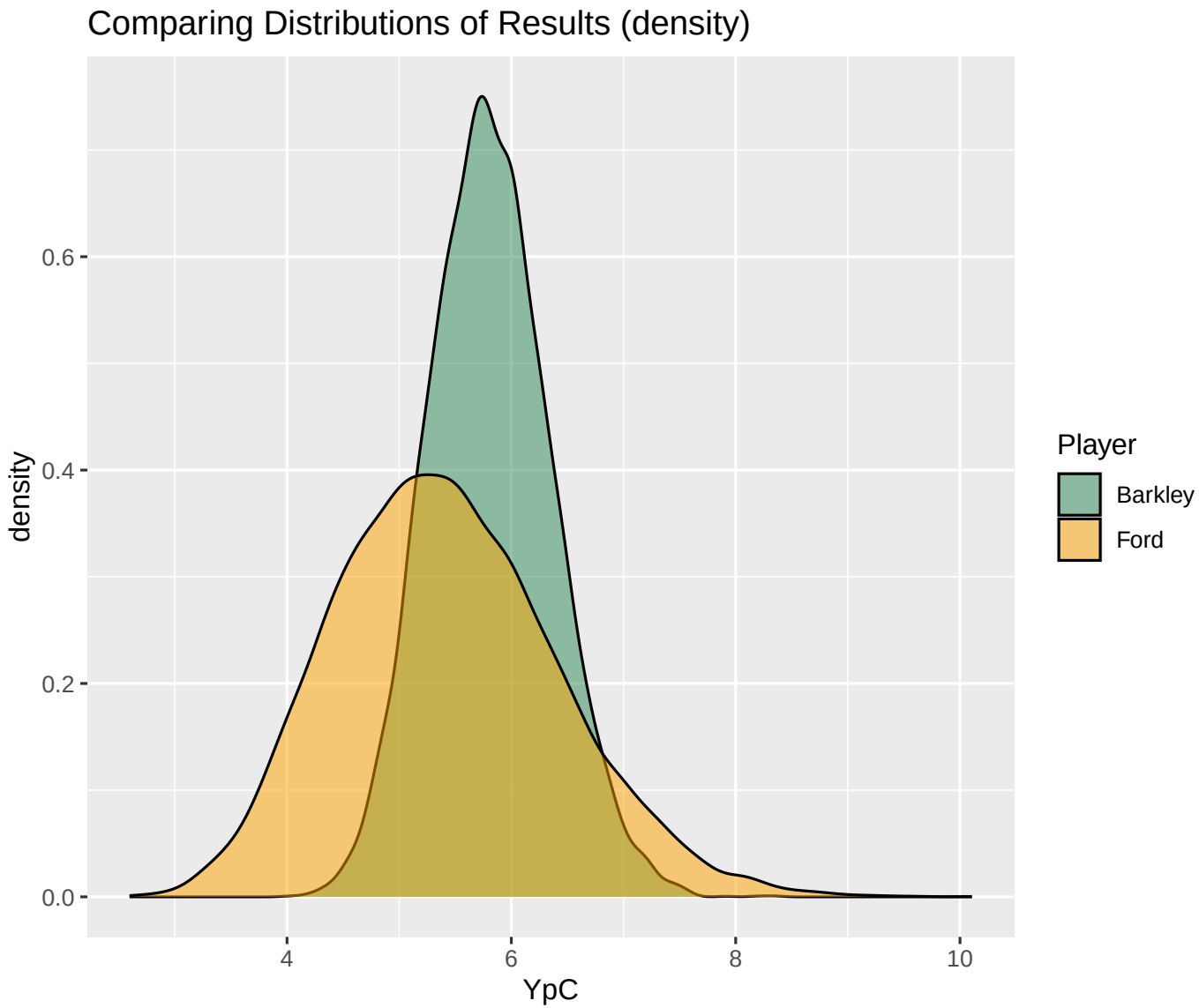
- Margin of error also addresses this question: how close to we expect Ford's hypothetical population mean value (which represents his actual ability) to be to the observed value from the season (5.43 yards-per-carry).
- The interval (3.40,7.46) gives a range of values such that we are "reasonably certain" that Ford's true average yards per carry lies in that range.
- This seems like a big range... in large part because he didn't have a full season's worth of carries ($n = 104$).

Example: Saquon Barkley

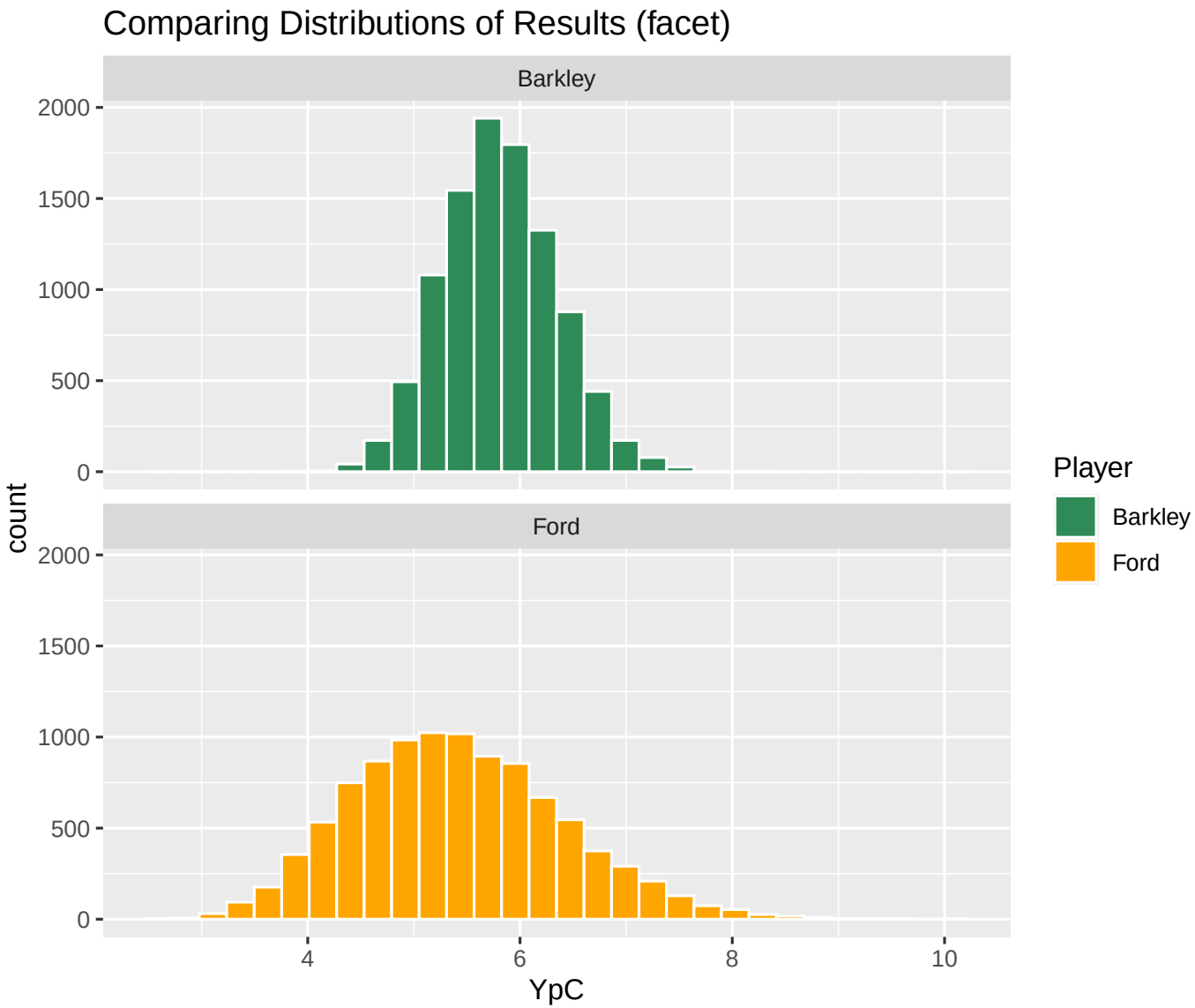
- 2024 Statistics: $n = 345$ carries for 2005 yards, for a yards-per-carry value of 5.81.
- Simulating 10000 seasons of 345 carries per season, drawn from actual rushing results from 2024, we obtain results distribution at right.
- MOE = 2*sd(results) = 1.07



Comparing the Two Running Backs



Alternate Visual



Intermission: Underlying Probability Theory

- Suppose we draw a random sample of size n from a population, and calculate the sample mean.
- If we draw a second random sample of size n, we should expect a different sample mean.
- The *sampling distribution for the sample mean* is a theoretical distribution of all possible values for the sample mean we can obtain from different random samples.
- We can use a sampling distribution to understand variation due to sampling.
- Let's do a quick demo in R

Approximate Sampling Distributions via Simulation

```
#create synthetic population
pop <- rnorm(10000,70,8)

#create vector for storing sample means
samp_dist <- c()

#a for loop to  sample 1000 from pop and record the sample mean
for (i in 1:1000){
     samp_dist[i] <- mean(sample(pop,1000))
}

# peek at our approx. to sampling distribution
hist(samp_dist)
```

Central Limit Theorem

Central Limit Theorem

When we collect a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, the sampling distribution of the sample mean will be nearly normal with

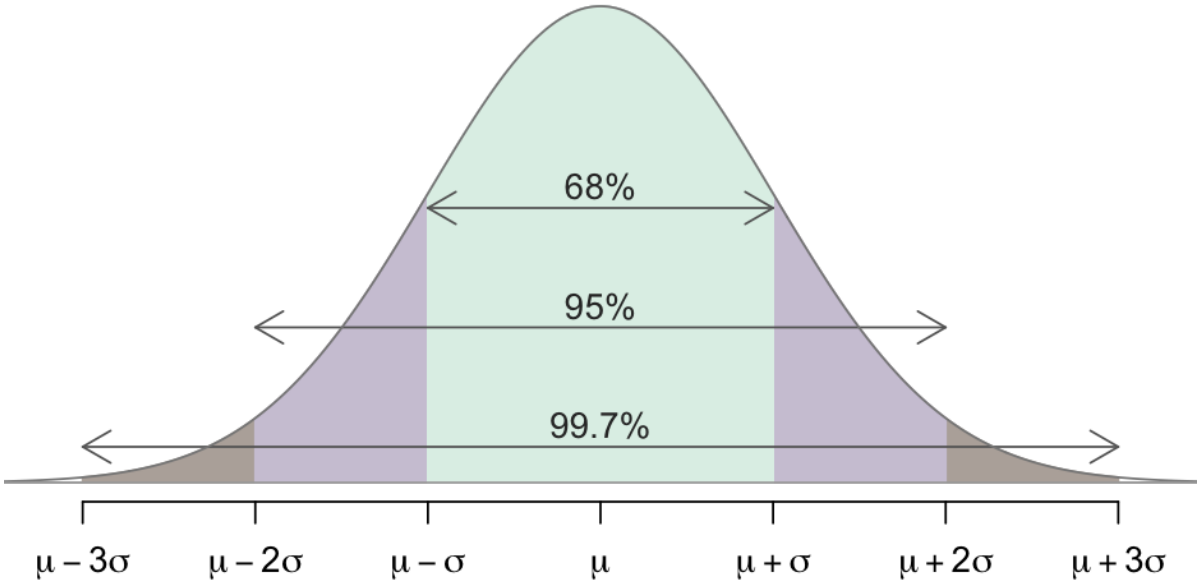
$$\text{mean} = \mu \qquad\qquad \text{standard deviation} = \frac{\sigma}{\sqrt{n}}$$

▸ Shiny App CLT Visual!

There are caveats! Are your data strongly skewed? Are there big outliers?

Normal Distribution: 68-95-99.7 Rule

Recall that about 95% of any normal distribution is within 2 standard deviations of the mean



Returning to Jerome Ford, NFL Running Back

- In his 2024 season he had $n = 104$ carries for an average of 5.43 yards per carry. We treat this as a random sample, drawn from a population of all possible carries.
- We assume the mean μ of this (theoretical) population is a measure of his true ability as a runner, and σ is the population standard deviation.
- Thanks to `nflfastR` play-by-play we can calculate the sample standard deviation of his carries. It is $s = 10.31$.
- We know from the CLT that our one sample mean (5.43) lives in a sampling distribution that is roughly bell-shaped centered at μ with standard deviation $\sigma/\sqrt{n}$. For large sample sizes, $\sigma \approx s$.
- Using the 95% rule, and substituting s for σ, about 95% of all possible sample means will be within $2 * s/\sqrt{n}$ of μ.

Here

$$2 * s/\sqrt{n} = 2 * 10.31/\sqrt{104} = 2.02.$$

Central Limit Theorem for proportions

Central limit theorem for proportions

If certain conditions are met, sample proportions will be nearly normally distributed with mean equal to the population proportion, p, and standard error equal to $\sqrt{\frac{p\ (1-p)}{n}}$.

$$\hat{p} \sim N\left(p, \sqrt{\frac{p\ (1-p)}{n}}\right)$$

▸ Shiny App CLT for Proportions

Example: Estimating variation in a proportion

In 2024 QB Geno Smith of the Seahawks attempted 578 passes with a 70.4% completion rate.

- We might treat these results as a random sample of $n = 578$ Bernoulli Trials (2 options: success and failure), where the (unknown) theoretical probability of success on any given trial is p.
- The sample proportion is $\hat{p} = .704$. How close to p is $\hat{p}$ likely to be?

Example: Estimating variation in a proportion

2024 QB Geno Smith: $n = 578$ passing attempts, with $\hat{p} = .704$.

- The CLT for proportions says that the sampling distribution for $\hat{p}$ is roughly normal, centered at the unknown p, with standard deviation approximated by the quantity

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

- So the margin of error here is about

$$2 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

and plugging in $\hat{p} = .704$ and $n = 578$ we obtain

$$\text{MOE} = 0.038.$$

- We can be reasonably certain that Smith's hypothetical true passing percentage for 2024 is in the interval

$$70.4\% \pm 3.8\% = (66.6\%, 74.2\%).$$

Estimating variation in a categorical variable with more than two possible outcomes

- Suppose an event has m possible outcomes $A_1, A_2, \ldots, A_m$, and we are interested in a statistic that assigns weights w_j to the A_j.
- Suppose we make n observations. Let p_j be the proportion of time A_j occurs.
- The average value of the statistic of interest is then

$$w_1 p_1 + w_2 p_2 + \cdots w_m p_m.$$

- For the statistic $T = w_1 p_1 + w_2 p_2 + \cdots w_m p_m$, we may use

$$\text{MOE} = 2 \cdot \frac{\sqrt{(w_1^2 p_1 + w_2^2 p_2 + \cdots + w_m^2 p_m) - T^2}}{\sqrt{n}}.$$

Example: Variation in Slugging Percentage

In baseball, we may view the event of an at-bat as having 5 possible outcomes: 'single', 'double', 'triple', 'home run', and 'other'.
For instance, in 2023, Corbin Carroll's had $n =$ with these results:

AB (n)	1B	2B	3B	HR	Other
565	96	30	10	25	404

A hitter's *slugging percentage* is calculated by dividing the weighted sum

$$1 \cdot 1\text{B} + 2 \cdot 2\text{B} + 3 \cdot 3\text{B} + 4 \cdot \text{HR} + 0 \cdot \text{Other}$$

by the total number of at bats.

Example: Variation in Slugging Percentage

Corbin Carroll in 2023:

AB (n)	1B	2B	3B	HR	Other
565	96	30	10	25	404

Carroll's p_j values are
$p_1 = 96/565; p_2 = 30/565; p_3 = 10/565; p_4 = 25/565;$ and $p_5 = 404/565.$
His slugging percentage is thus

$$\sum_{j=1}^{5} w_j \cdot p_j = (1 \cdot p_1) + (2 \cdot p_2) + (3 \cdot p_3) + (4 \cdot p_4) + (0) = 0.506.$$

And the margin of error here is

$$\text{MOE} = 2 \cdot \frac{\sqrt{(1(96/565) + 4(30/565) + 9(10/565) + 16(25/565) - (.506)^2}}{\sqrt{565}} = 0.084$$

Carroll's true slugging ability might then be described as

$$.506 \pm .084 \quad \text{or} \quad (.422, .590).$$

Comparing Two Players

We have addressed how to take into account the inherent variability in sports data when calculating a player statistic.
Here we consider how to find the Margin of error for the difference between two players.
In Section 4.6 we find this rule:

Margin of Error for the difference between two independent measurements

Add the squares of the two margin of errors for the individuals, and take the square root!

Comparing Two Players

Section 4.6: How to find the Margin of error for the difference between two players.

Margin of Error for a difference between two independent samples

Add the squares of the two margin of errors for the individuals, and take the square root!

For our two running backs:

Player	n	YpC	MOE
Ford	104	5.43	2.03
Barkley	345	5.81	1.07

Comparing the difference in their yards per carry values: $5.81 - 5.43 = 0.38$. Is this a statistically significant difference?
Well, the margin of error for this difference is given by

$$\sqrt{2.03^2 + 1.07^2} = 5.22,$$

which is very large, producing the interval (-4.84, 5.60), which we are fairly certain contains the true difference. Hmm... an undiscerning interval, suggesting no statistically significant difference here.

Statistical Significance

The point of looking at variation inherent in sports data is for us to be able to distinguish between a difference in outcomes that might be reasonably attributed to chance, and a difference that is likely due to performance ability.

Example: Was Peak Babe Ruth actually better at getting on base than Peak Joe DiMaggio, or might their season data differences be reasonably explained by chance?

The Babe vs The Yankee Clipper

Ruth's Peak Seasons: 1920-1927
DiMaggio's Peak Seasons: 1937-1942

Player	N	OBP
Joe DiMaggio	3741	0.412
Babe Ruth	4900	0.498

Ruth's MOE: $\sqrt{\frac{.498(1-.498)}{4900}} = 0.00714$;

DiMaggio's MOE: $\sqrt{\frac{.412(1-.412)}{3741}} = 0.00805$;

The Babe vs The Yankee Clipper

Player	N	OBP	MOE
DiMaggio ('37 - '42)	3741	0.412	0.00714
Babe Ruth ('20 - '27)	4900	0.498	0.00805

So we believe with high probability (95% confidence!) that the difference in on-base percentage between Peak Ruth and Peak DiMaggio is in this interval:

$$(.498 - .412) \pm \sqrt{0.00714^2 + .00805^2}$$

$$.086 \pm .011$$

$$(.075, .097)$$

Conclusion?
Since this entire interval is greater than 0, we have statistically significant evidence that Peak Ruth was better at getting on base than Peak DiMaggio. The observed difference (.498 vs .412) is very likely NOT due to chance variation inherent in sports data.