Probability for Sports Analytics

DATA 299

Based on Ch. 3 in *Analytic Method in Sports, 2nd ed.*, Severini

Hitchman

Spring '25

Probability topics we cover here

1. Random processes, events, and the probability of an event
2. Random variables
3. Probability Distributions
4. Expected Value
5. Binomial distributions
6. Normal distributions
7. Conditional Probability
8. Total Law, Bayes' Rule

Random Processes

We can think of a **random process** as a repeatable process that produces a result whose outcome is unknown ahead of time, but has a predictable set of possible outcomes.

- Roll a 6-sided die.
- Community fun run. Record the mile pace for all runners.
- Swing at a pitch!
- Take a penalty kick!

Sample Space, Events, Probability

We associate to a random process the following terms and ideas:

- **Sample Space** - the set of possible outcomes of the random process
- **Events** - we call any subset of the sample space an **event**
- The **probability of an event** associated to a random process to be the proportion of times the event would occur in the long run, as the random process is repeated a very, very, very large number of times.
- A **probability distribution** associated to a random process is a table or function (having certain features) with which one can compute the probability of events.

Let's discuss these terms in the context of the first two random processes mentioned on the previous slide

Example 1: Roll a 6-sided die

Roll a 6 sided die.

- **Sample Space**: $S = \{1, 2, 3, 4, 5, 6\}$
- **Some Events**:
  - $A = \{1, 3, 5\}$ (roll an odd number!)
  - $B = \{2, 3, 5\}$ (roll a prime number!)
  - $C = \{2\}$ (roll an even prime number!)
- **Some Probabilities**: If we assume the die is fair, so each number has the same chance of coming up (1/6 chance), then
  - $P(A) = 3/6 = 0.5$
  - $P(B) = 3/6 = 0.5$
- **Probability distribution**: Our calculations above relied upon having a fair die (each value has the same chance of being rolled). This assumption defines the probability distribution! Here it is, in table form:

x	1	2	3	4	5	6
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6

Example 2: Fun-run mile pace

Measure the time it takes each runner to finish

Everyone runs at their own pace!

- **Sample Space**: Let S to be the set of possible times
- **Some Events**:
  - A might denote the event that a randomly chosen runner has a mile pace under 6 minutes.
  - B might denote the event that a randomly chosen runner has a mile pace between 7 and 8 minutes.

Some Probabilities: Without knowledge of the **population distribution**, we can't estimate these probabilities, but if we have a frequency distribution, we can estimate these probabilities.

Random Variable

- A **random variable is a variable (commonly X) used to indicate an outcome of a random process if the outcomes are numerical.**
- **Perhaps X represents the total number of home runs hit in MLB on any given year.**
- **Or perhaps X represents my time running the mile.**
- **Or perhaps X represents my score in a game of skee ball.**

Discrete vs Continuous Random variables

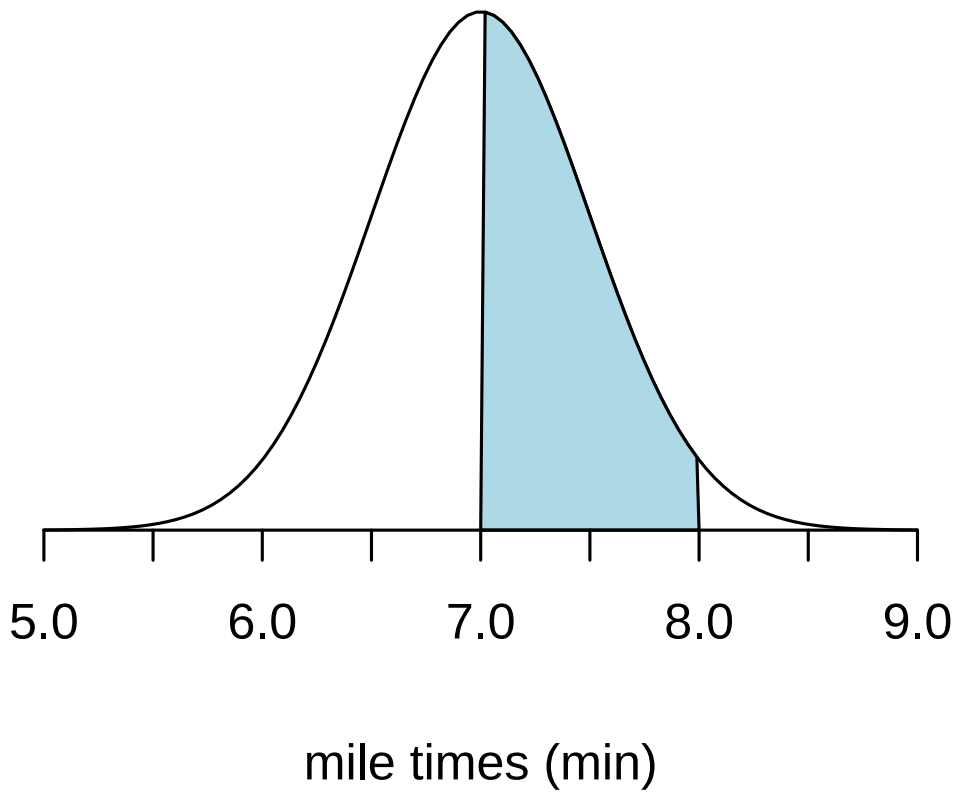
We encounter two different flavors of probability distributions:

- **discrete distributions**, where the sample space has a finite set of outcomes (usually whole numbers)
- **continuous distributions**, where the sample space consists of an interval of possible values on the number line

The dice example above is discrete, and its probability distribution is visualized with a table or a bar plot, or expressed via a function.

Density Curves for Continuous distributions

The run a mile example is continuous, and its probability density is visualized with a density plot. Probabilities would then correspond to areas under this density curve.



A valid probability model

- For a discrete distribution, a valid probability distribution has these two features:
  - $0 \leq p(x) \leq 1$ for each x in the sample space;
  - The sum of all the $p(x)$ must equal 1.
- For a continuous distribution, a valid probability density function has these two features:
  - $f(x) \geq 0$ for each x in the sample space;
  - the area under the density function f equals 1.

NY Yankees Home Runs from Ch. 2 Homework

In Homework 3 for this course you created a frequency distribution for the number of home runs hit by the NY Yankees in each of their games in 2018. The underlying random variable here is $X =$ the number of home runs hit in a game.

HR	0	1	2	3	4	5	6
N	31	53	41	21	12	3	1

We can convert this table to proportions by dividing each count by 162 (total number of games)

x	0	1	2	3	4	5	6
$p(x)$	.191	.327	.253	.130	.074	.019	.006

NY Yankees Home Runs from Ch. 2 Homework

From this table, how can we compute the average number of HR hit per game by the Yankees that season?

HR	0	1	2	3	4	5	6
N	31	53	41	21	12	3	1

NY Yankees Home Runs from Ch. 2 Homework

From this table, which resembles a probability model, how can we compute the average number of HR hit per game by the Yankees that season?

x	0	1	2	3	4	5	6
$p(x)$	.191	.327	.253	.130	.074	.019	.006

Expected Value of a random variable X

If X is a random variable with probability function $p(x)$ we define the **expected value** of X, denoted $E(X)$, to be the sum

$$E(X) = \sum_{\text{all } x} x \cdot p(x).$$

The expected value $E(X)$ is also called the mean of X, and is often denoted as μ_X, or, simply μ.

Expected Value of a random variable X

Example

After several years of experience, I modeled my skee ball ability at *Lakefair* as follows:

x	0	10	20	30	40	50
$p(x)$	.1	.2	.1	.1	.4	.1

- What's the probability I score at least 30 points on any given roll?
- What's the expected value of X?
- In a game with 6 rolls, what's my expected total score?

Binomial Distribution

Scene: A random process has two possible outcomes: "success" and "failure", and the probability of success on any given trial is always the same, call it p. Suppose we want to repeat the process n times and let X denote the number of successes in n trials. Then X has a **binomial distribution with parameters n and p.** Some examples:

- X counts the number of free throws I make in 25 attempts, assuming $p = .7$ is the probability of making the free throw on any given trial. Then X is binomial(25,.7).
- X counts the number of pars I make on the 14 par-4 holes of a golf course, assuming $p = .85$ is the probability of making par on any one of them. Then X is binomial(14,.85)

Binomial Distribution

The probability function for a binomial(n,p) distribution is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, 1, 2, \ldots, n$, where

$$\binom{n}{x} = \frac{n!}{(n-x)!x!}.$$

Expected Value of a binomial(n,p) distribution

If X is binomial on n trials with probability of success p, then $E(X) = np$.

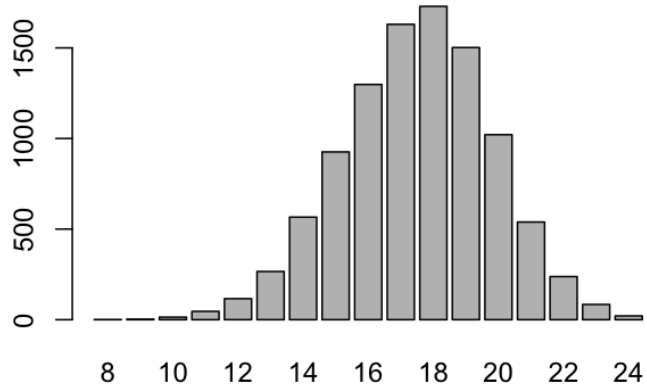
Generate a random sample from a binomial dist'n using `rbinom()`

```
v = rbinom(1000,25,.7)
```

Shooting free throws

Draw a random sample of size 10000 from a `binomial(25,.7)` distribution, to simulate shooting 25 free throws 10000 times (recording each time how many I make). Use this sample to estimate the probability that I make at least 23 free throws.

```
trials = 10000
s = rbinom(trials, 25, .7)
plot(table(s))
```



How likely is it to make at least 23 shots out of 25 for a 70% shooter?

```
sum(s >= 23)/trials
.0105
```

Hitting home runs

Suppose a major league hitter has probability of 0.05 of hitting a home run on any plate appearance (an assumption about as useful as assuming a human is a perfect cylinder). Suppose further that this player plays 18 seasons, getting 600 plate appearances each season. How likely is such a hitter to reach 600 career home runs?

Do you see a binomial distribution lurking here? Can we simulate such a player's career home run total?

Estimating HR career totals with a binomial distribution

Do you see a binomial distribution lurking in this career home runs example? Can we simulate such a player's career home run total?

- Yes!
- `rbinom(18, 600, 0.05)`
- This code will generate a random sample of size 18 (one for each season) drawn from a binomial distribution on $n = 600$ trials (the PA), with probability of success (HR) on any single trial equal to $p = .05$.
- Let's look at this in R.

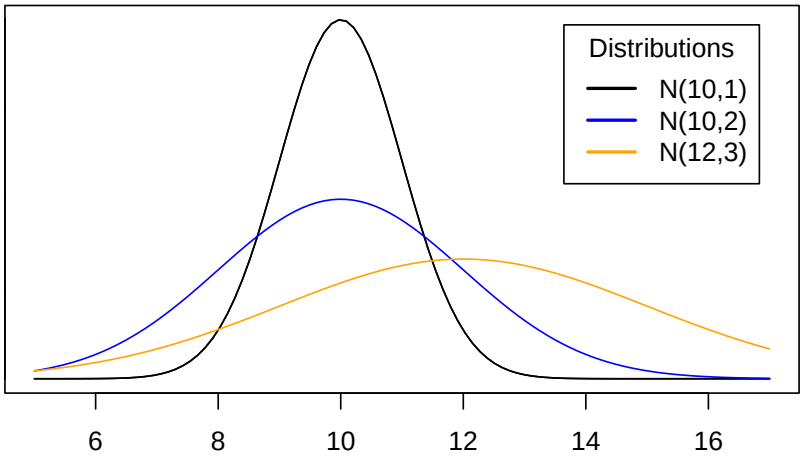
Repeated Sampling to estimate probability

How likely is such a hitter to reach 600 career home runs?

- With a `for` loop we can easily simulate the career HR totals over many, many simulated careers, and record the results from each simulated career in a vector called `results`!
- We can use these simulated career results to estimate the probability that such a player will end their career with at least 600 home runs.
- Let's do this in R!

Normal Distribution

- the classic bell curve
- a continuous distribution, defined for $-\infty < x < \infty$
- defined by two parameters, μ and σ
- Notation: $N(\mu, \sigma)$.



Normal Distribution

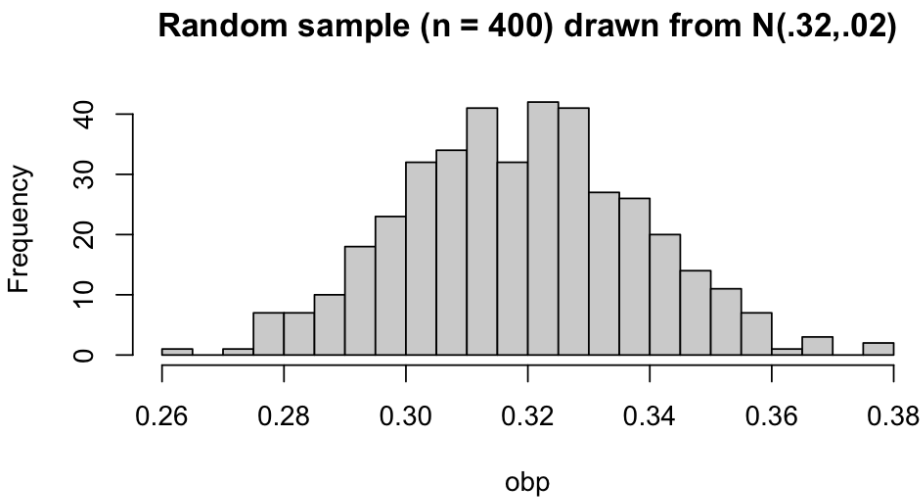
A normal distribution

- provides a great model for physical measurements.
- is central to probability theory thanks to the **central limit theorem**: the distribution of sample means obtained from (large) samples drawn from most distributions will be bell-shaped.

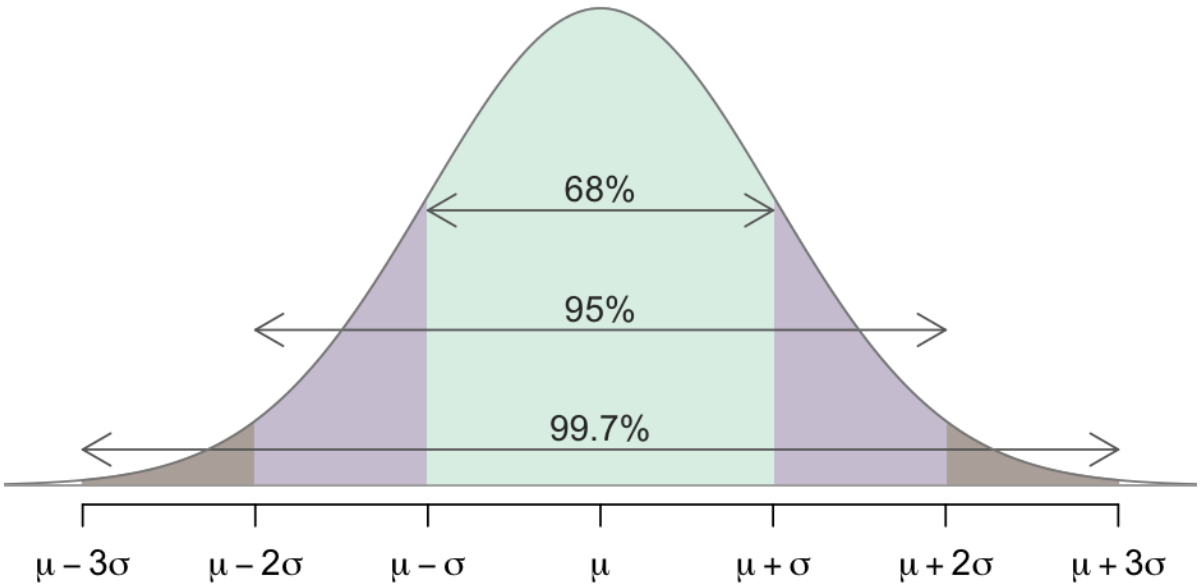
Normal Distribution

Example: Let's say one season the on base percentage among qualified batters in MLB is normally distributed with mean 0.320 and standard deviation 0.020. Make a histogram of a random sample of $n = 400$ "players" drawn from this distribution

```
obp <- rnorm(400, 0.320, 0.020)
hist(obp,breaks=20)
```



Normal Distribution: 68-95-99.7 Rule



So in a season in which OBP follows a N(0.320, 0.020) distribution the middle 68% of hitters have an OBP between 0.300 and 0.340; 95% between 0.280 and 0.360; and essentially all between 0.260 and 0.380.

Z-scores

If X is a single observation drawn from a normal distribution $N(\mu, \sigma)$ then its Z-score is the value

$$Z = \frac{X - \mu}{\sigma}.$$

The Z-score is a unitless measure of how many standard deviations a value is above or below the mean of the bell-curve.

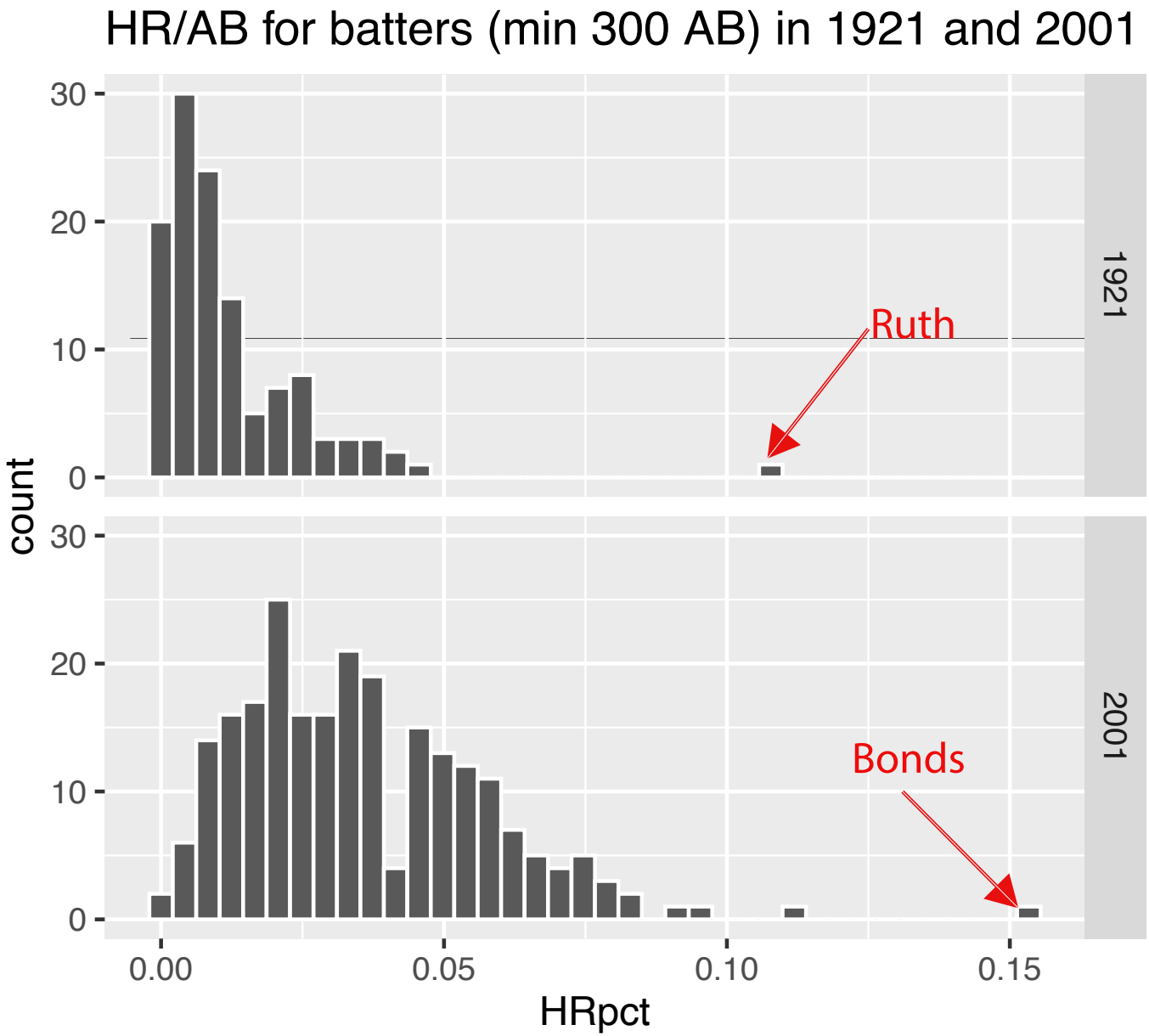
Z-scores

Which is a better performance relative to the field?
$X = 220$ in a distribution that is $N(200, 12)$ or
$X = 210$ in a distribution that is $N(200, 5)$?

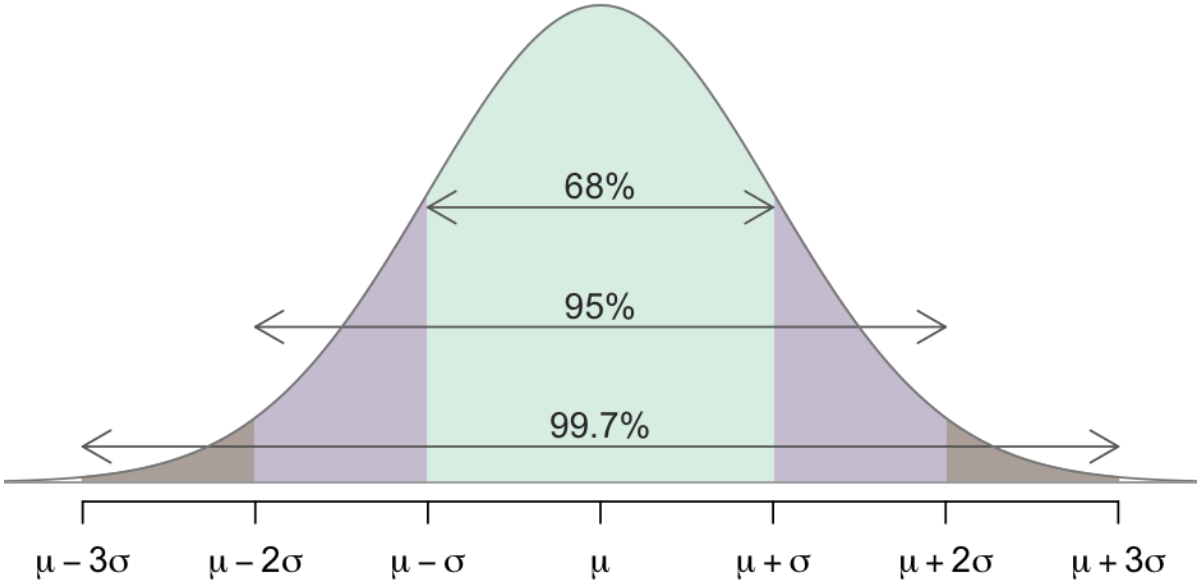
Z-scores

- When Babe Ruth hit 59 HR in 1921, his HRpct (HR/AB) was 0.109.
- When Barry Bonds hit 73 HR in 2001 his HRpct was 0.153.
- Which performance was better relative to the field?

Z-scores


HR/AB for batters (min 300 AB) in 1921 and 2001
count
30
20
10
0
1921
Ruth
30
20
10
0
2001
Bonds
0.00
0.05
0.10
0.15
HRpct

Normal Distribution: 68-95-99.7 Rule



Suppose X is an observation drawn from any normal distribution. If its Z score is...

- $Z = -2$, then X lies at the 2.5th percentile
- $Z = -1$, then X lies at the 16th percentile,
- $Z = 0$, then X lies at the 50th percentile,
- $Z = 1$, then X lies at the 86th percentile,
- $Z = 2$, then X lies at the 97.5th percentile

Conditional Probability

Scene

You have two events A and B associated with a random process. Then we might be interested in the following probabilities

- $P(A)$, the probability that event A occurs.
- $P(B)$, the probability that event B occurs.
- $P(A \cap B)$, the probability that both A and B occur.
- $P(A \cup B)$, the probability that either A or B occurs (or possibly both).
- $P(A \mid B)$, the probability that A occurs given that event B has occurred.

The last probability, $P(A \mid B)$ is called the **conditional probability** that A occurs given that B has occurred.

Conditional Probability

General Formula for Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

so long as $P(B) > 0$.

Example: We roll a 6 sided die. Let A be the event we roll an odd number, and B the event that we roll a 5 or 6.

- What is $P(A|B)$?
- What is $P(B|A)$?

Conditional Probability and Independence

Independent Events

A and B are called **independent events** if $P(A \mid B) = P(A)$ (i.e., the probability that A occurs doesn't change if B has occurred.)

Dice example again: Here $A = \{1, 3, 5\}$, and $B = \{5, 6\}$. On the previous slide we showed:

- $P(A) = 3/6 = 1/2$.
- $P(B) = 2/6 = 1/3$.
- $P(A \cap B) = 1/6$.
- $P(A \mid B) = 1/2$.
- $P(B \mid A) = 1/3$.

Conclusion: Since $P(A \mid B) = P(A)$ in this example, A and B are independent events.

Conditional Probability and Independence

Independent Events

If A and B are independent events then $P(A \cap B) = P(A)P(B)$.

- In the NCAA basketball tournament, there are 4 first round games between a 5 seed and a 12 seed. Suppose that in each game, the 5 seed has an 75% chance of winning. What is the probability that there is no upset in any of the 5 vs 12 seed matches?
- Steph Curry as a career free throw percentage of 91%. The all-time record for consecutive free throws made is 97, set by Michael Williams in 1993. What is the probability that Steph Curry makes his next 98 free throws?

Bag of Taffy

A bag of Taffy has three flavors: vanilla (20%), root beer (30%), and peppermint (50%). 5% of the vanilla pieces are sticky to the touch, 10% of the root beer pieces are sticky to the touch, and 4% of the peppermint pieces are sticky to the touch. If you pull a candy out of the bag at random, what is the probability that it is sticky to the touch?

Total Law of Probability

A collection of subsets $B_1, B_2, \ldots, B_n$ is a partition of a space S if the union of all the B_i equals S, and the B_i are pairwise disjoint. In such a setting, for any event A,

$$P(A) = P(A \mid B_1) \cdot P(B_1) \; + \; P(A \mid B_2) \cdot P(B_2) \; + \; \cdots \; + \; P(A \mid B_n) \cdot P(B_n).$$

Bag of Taffy part Deux

A bag of Taffy has three flavors: vanilla (20%), root beer (30%), and peppermint (50%). 5% of the vanilla pieces are sticky to the touch, 10% of the root beer pieces are sticky to the touch, and 4% of the peppermint pieces are sticky to the touch. If you pull a candy out of the bag at random with your eyes closed and find that it is sticky to the touch, what is the probability that it is peppermint?

Bayes Rule

Suppose $B_1, B_2, \ldots, B_n$ is a partition of a space S and A is any event. Then for any particular set B_k in the partition,

$$P(B_k \mid A) = \frac{P(A \mid B_k) \cdot P(B_j)}{P(A)}.$$

▸ Testing Positive for a Deadly Disease