Statistical inference on proportions

Outline for today

Review of the binomial, normal distributions

Bayesian inference

Hypothesis tests for proportions

Announcement

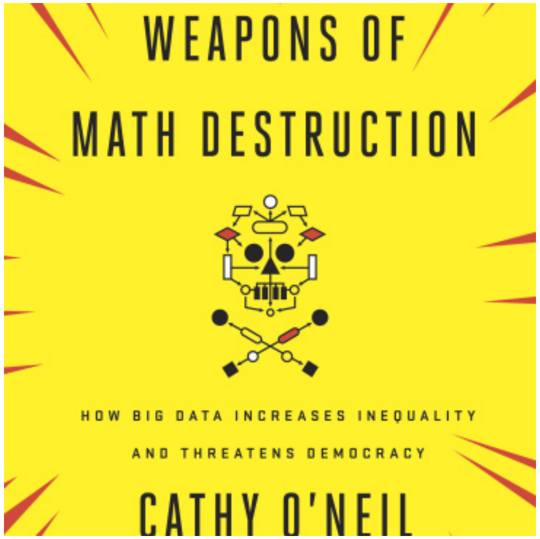
Opening day was yesterday

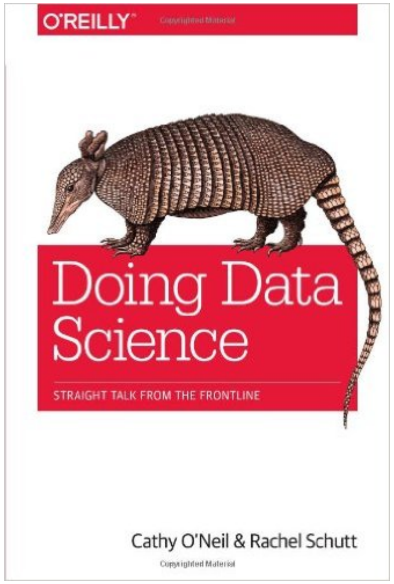
The Red Sox beat the Pirates 5-3

Off to a good start!

Announcement 2

Cathy O'Neil speaking on "Weapons of Math Destruction"





<u>When</u>: Saturday April 8th, at 10:30am

<u>Where</u>: Hooker Auditorium, Clapp Laboratory, Mount Holyoke

Questions about worksheet 7?

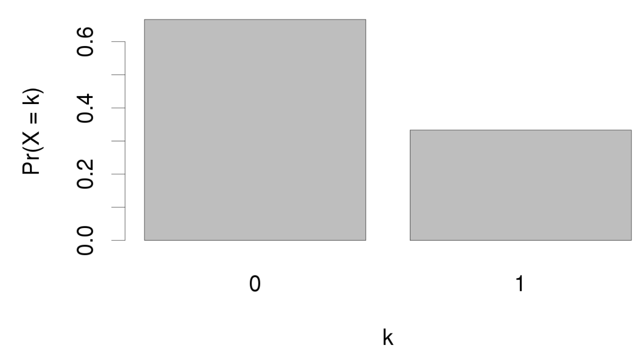
Review: Discrete probability models

Discrete distributions: random variable X takes on discrete numbers

- Examples?

Bernoulli Distribution:

- What does it model?
  - Probability of success on a single trial (coin flip)
- What is the sample space?
  - X is in {0, 1}
- What are the parameters?
  - π: probability of success
    - i.e., Pr(X = 1)



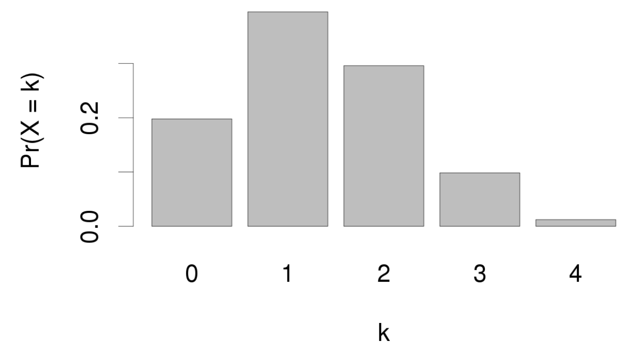
Review: Binomial distribution

What does it model?

- It models the probability of ***k*** success out of ***n*** trials

What are the parameters?

- n: number of trials
- π: probability of success on each trial
- k: number of successes



$$Pr(a \le X \le b) = \sum_{k=a}^{b} f(k; ...)$$

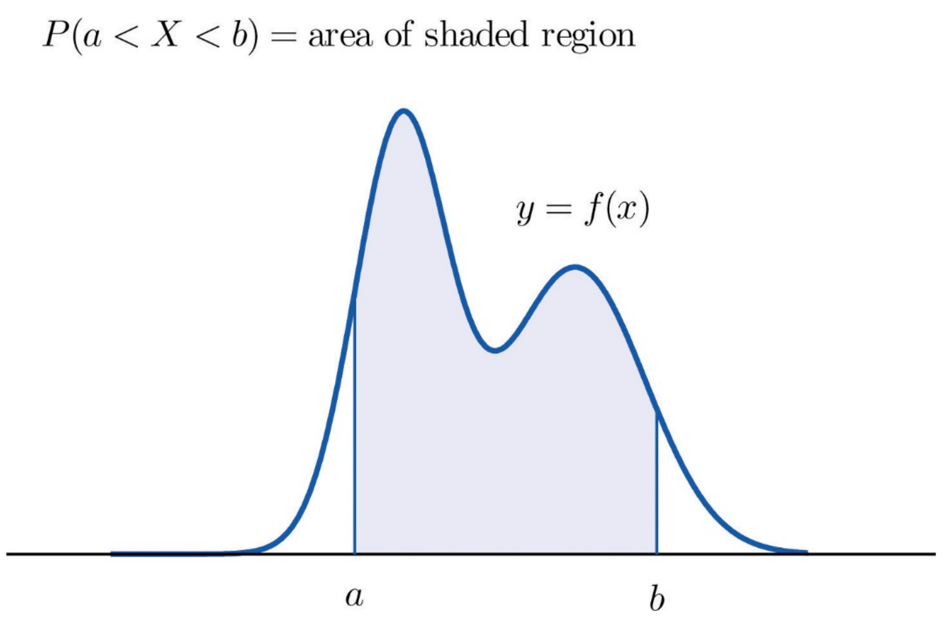
In R:

- Pr(X = k): dbinom(x = k, size = n, prob = pi)
- Pr(X ≤ a): pbinom(a, size = n, prob = pi)

Continuous probability distributions

Continuous distributions have **probability density functions** that can be used to calculate the probability that an even is in some interval [a b]

$$Pr(a \leq X \leq b) = \int_a^b f(x; ...)dx$$



Normal Density Curve

A **normal distribution** follows a bell-shaped curve

- Useful for describing the mean of many random outcomes

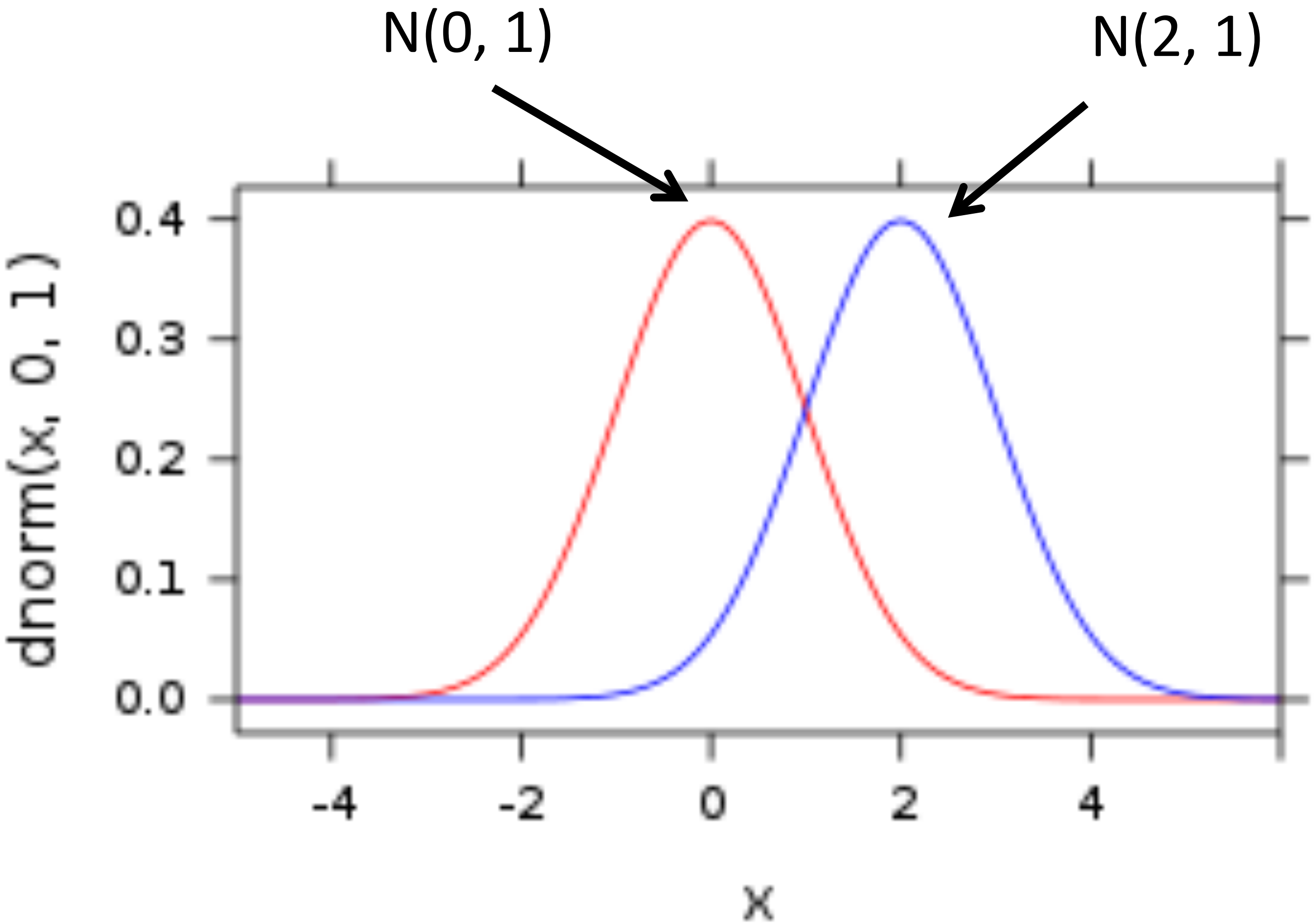
There are two parameters that characterize normal curves, which are:

- The mean: μ
- The standard deviation: σ

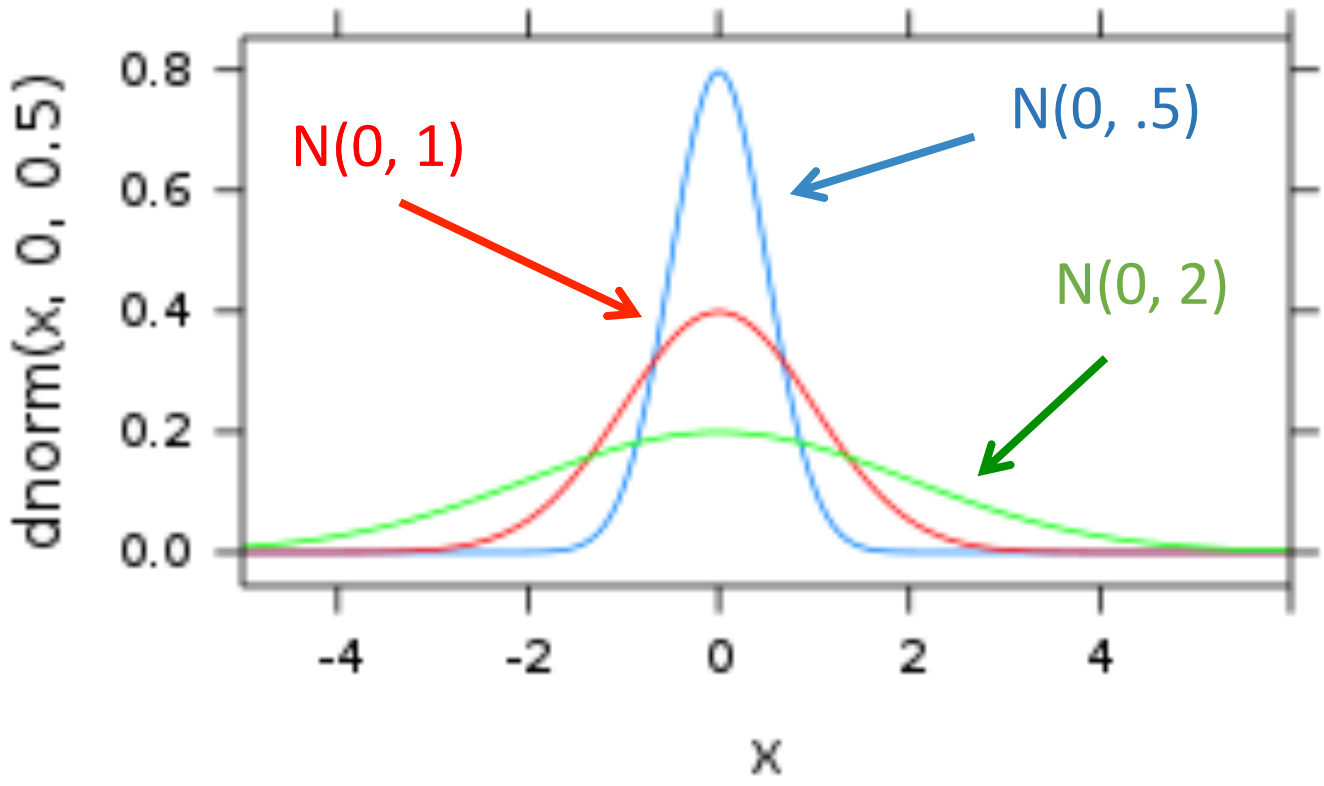
Notation: X ~ N(μ, σ)

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

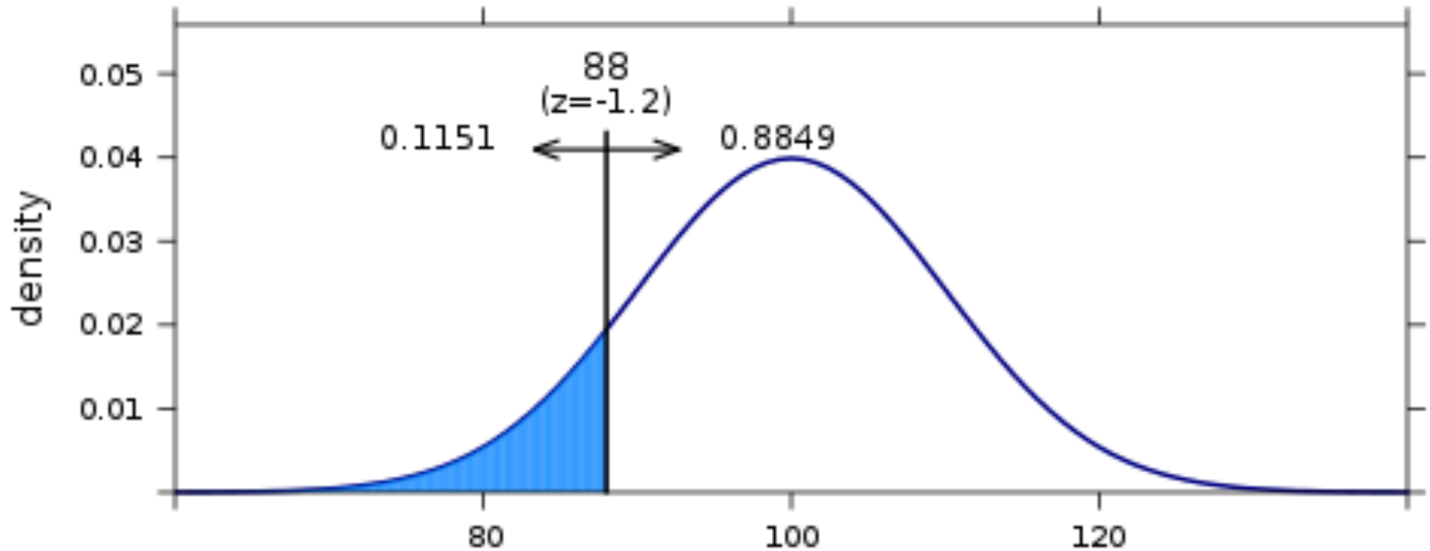
Normal curves with different means



Normal curves with different variances



Normal probability



Example: IQ scores are normally distributed with μ = 100, sd = 10

What is the probability someone would have an IQ less than 88?

$$Pr(X < 88) =$$

```
R: pnorm(88, 100, 10)
```

Why the normal distribution?

Central limit theorem: the mean ($\bar{x}$) of a number of random variables independently drawn from the same distribution is normal

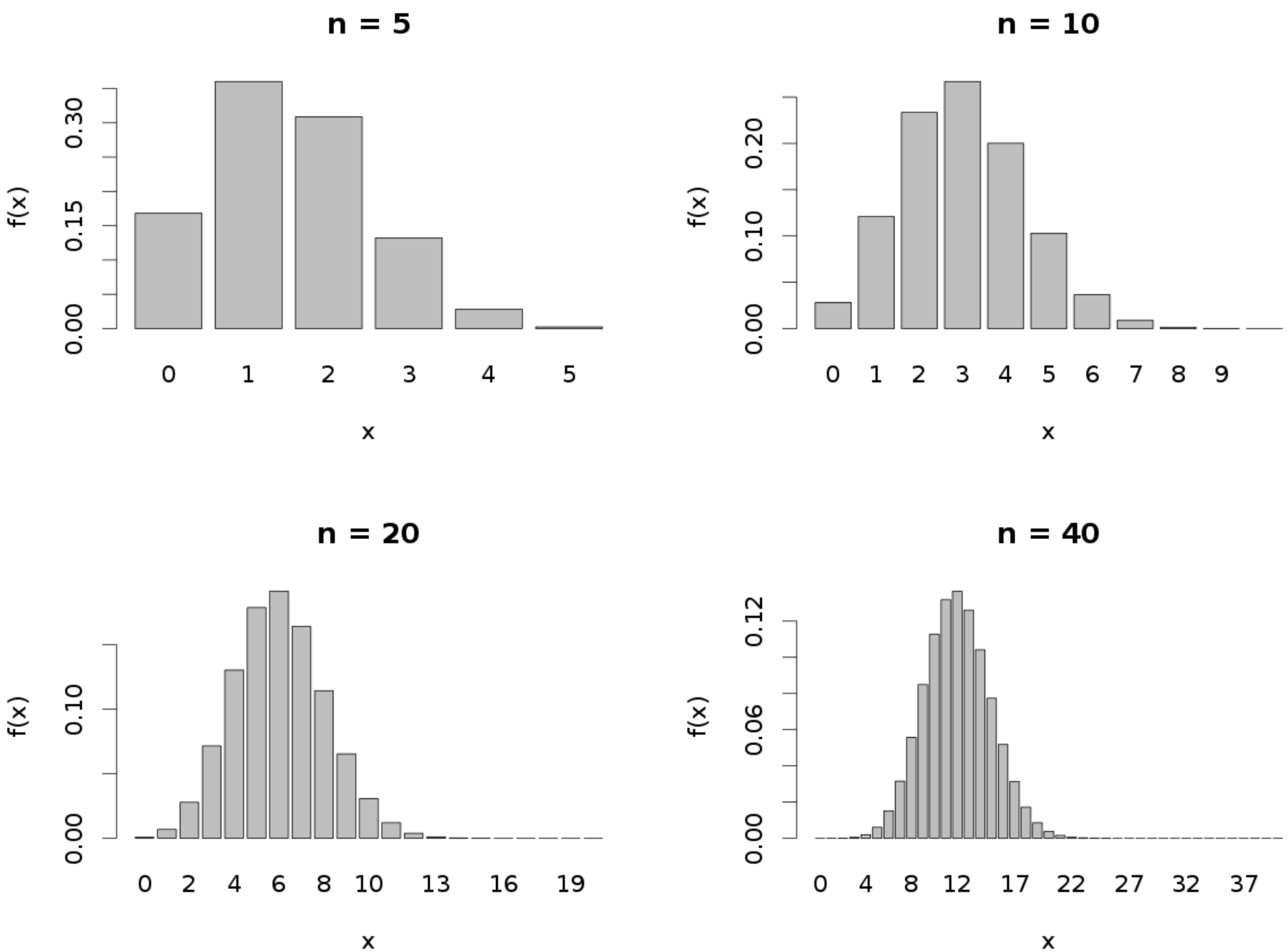
Example: Binomial distribution comes normal as n increases

Binomial distribution is the sum of Bernoulli outcomes

- e.g., binomial distribution says if I toss a coin ***n*** times, what is the probability I get ***k*** heads

By the central limit theorem, it should become like a normal distribution as ***n*** increases

Example: Binomial distribution comes normal as n increases (here $\pi = .3$)

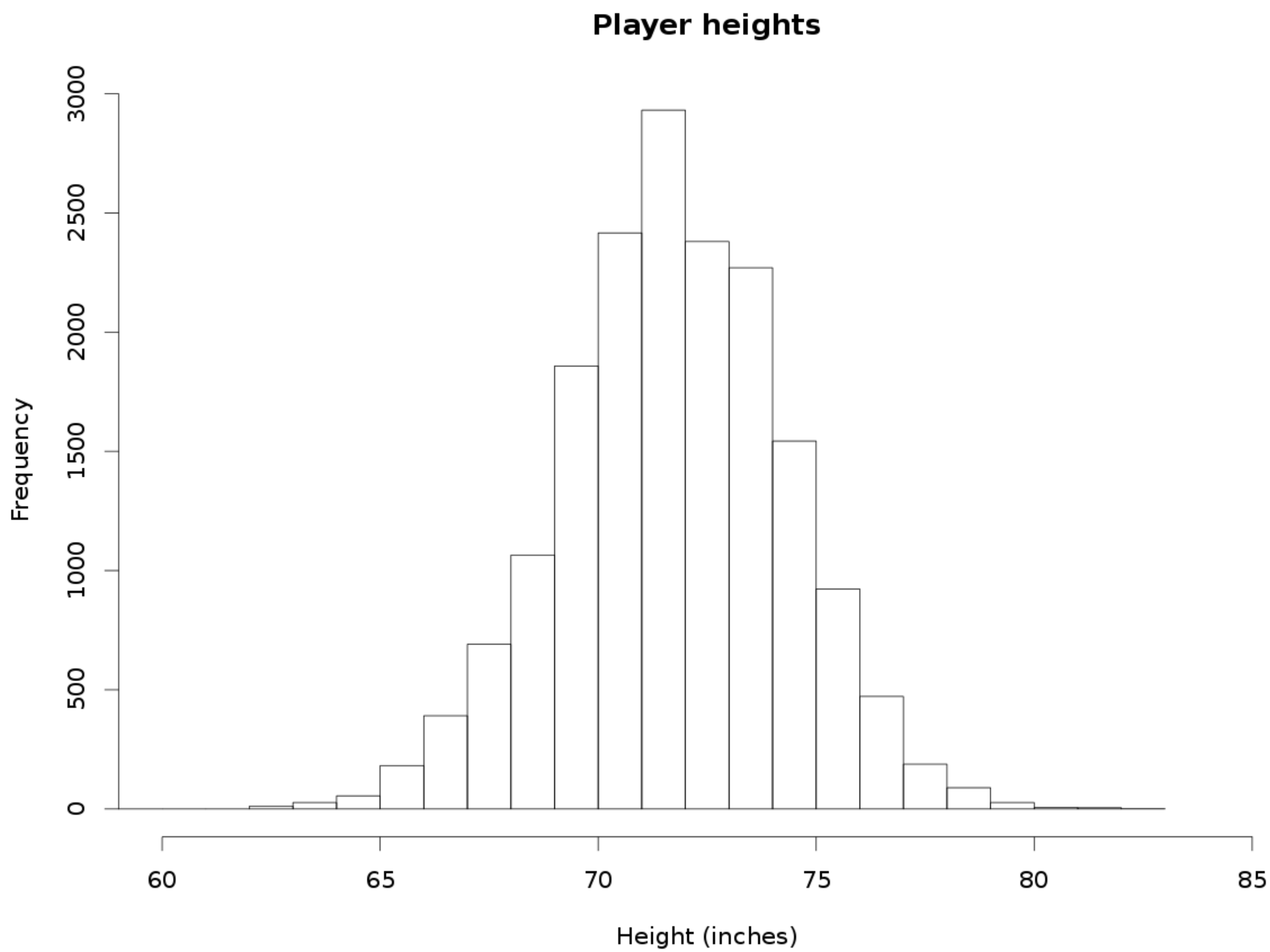


Examples of normal distributions in baseball?

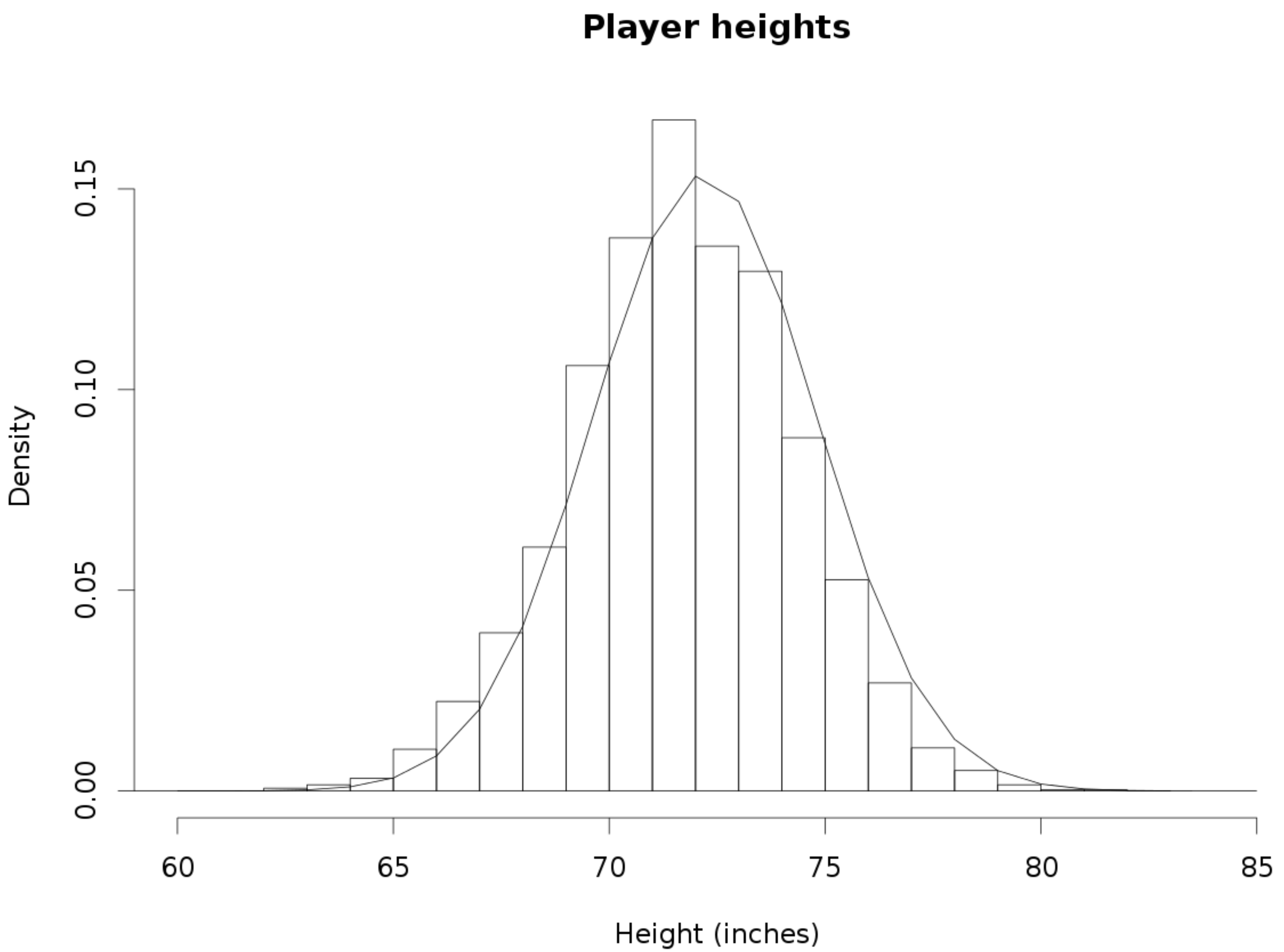
Can you think of any examples?

- Height of players?
- Weight of players?
- Length of games?

Examining player heights…



Examining player heights…



Any questions about probability?

On to statistical inference!

Statistical inference

Statistical inference: use sample of data to deduce properties of an underlying population, or stochastic process

In the context of baseball this usually means: looking at a player's **performance** to tell something about the player's **ability**

- **Ability**: innate talent
- **Performance**: outcomes from playing a number of games

Statistical inference

We've seen many cases of simulating a player's performance based on pre-specified abilities (probability)

- E.g., we can simulate a .333 OBP by rolling a die (or using an R function) to generate random data consistent with a .333 OBP

Hit, Out, Hit, Out, Out, Out, …

With statistical inference we go in the <u>other direction</u>: we take a collection of outcomes and estimate the probability model parameters

Hit, Out, Hit, Out, Out, Out, …

Estimate π_{hit}

Bayesian Inference: Determining the ability of a player

Suppose there are 3 players with different true OBP abilities:

- Player H_1's true OBP is .200 H_1: $\pi = .200$
- Player H_2's true OBP is .333 H_2: $\pi = .333$
- Player H_3's true OBP is .500 H_3: $\pi = .500$

One player is selected at random and we observe the player for 10 plate appearances

Can we tell whether it was player H_1, H_2, H_3, who was picked?

Let's simulate this with a 4, 6, and 10 sided die

- 1 or 2 is on base
- Higher numbers are outs
- I will roll the die 10 times…

Bayesian Inference: Determining the ability of a player

Suppose we got 5 on base events out of 5 plate appearances

Question: What die was chosen?

- i.e., what value is π?

Determining the ability of a player

Here are the simulation results from 1000 simulations of rolling the different dice 10 times:

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.200	95	271	315	199	80	35	5	0	0	0	0
0.333	17	81	206	270	226	121	59	16	4	0	0
0.500	7	44	111	206	252	205	122	47	6	0	0

Total number of simulations that produced 5 hits = 35 + 121 + 205 = 361

$\Pr(\pi = .200 \mid 5 \text{ hits}) = 35/361 = .097$

$\Pr(\pi = .333 \mid 5 \text{ hits}) = 121/361 = .335$

$\Pr(\pi = .500 \mid 5 \text{ hits}) = 205/361 = .568$

Determining the ability of a player

Results based on the binomial distribution

- What are n, k and π here?

	0	1	2	3	4	5	6	7	8	9	10
0.200	0.11	0.27	0.3	0.2	0.09	0.03	0.01	0	0	0	0
0.333	0.02	0.09	0.2	0.26	0.23	0.14	0.06	0.02	0	0	0
0.500	0	0.01	0.04	0.12	0.21	0.25	0.21	0.12	0.04	0.01	0
sum	0.13	0.37	0.54	0.58	0.53	0.42	0.28	0.14	0.04	0.01	0

Binomial distribution results normalized

	0	1	2	3	4	5	6	7	8	9	10
0.200	0.85	0.73	0.56	0.34	0.17	0.07	0.04	0.00	0.00	0.00	
0.333	0.15	0.24	0.37	0.45	0.43	0.33	0.21	0.14	0.00	0.00	
0.500	0.00	0.03	0.07	0.21	0.40	0.60	0.75	0.86	1.00	1.00	

Bayesian inference

Gives a probability distribution over ability (parameters) given that we have observed some performance (data)

$$Pr(H_i|data) = \frac{Pr(data|H_i) \cdot Pr(H_i)}{Pr(data)}$$

Hypothesis tests

Often we want to test if a parameter is equal to a particular value

e.g., we might want to test if $\pi = .350$

To test if a parameter is equal to a particular value we can use hypothesis tests

Paul the Octopus

In the 2010 World Cup, Paul the Octopus (in a German aquarium) became famous for correctly predicting 11 out of 13 soccer games

Question: is Paul psychic?

Paul the Octopus

Question: If Paul was psychic, what proportion of games would we expect him to guess correctly?

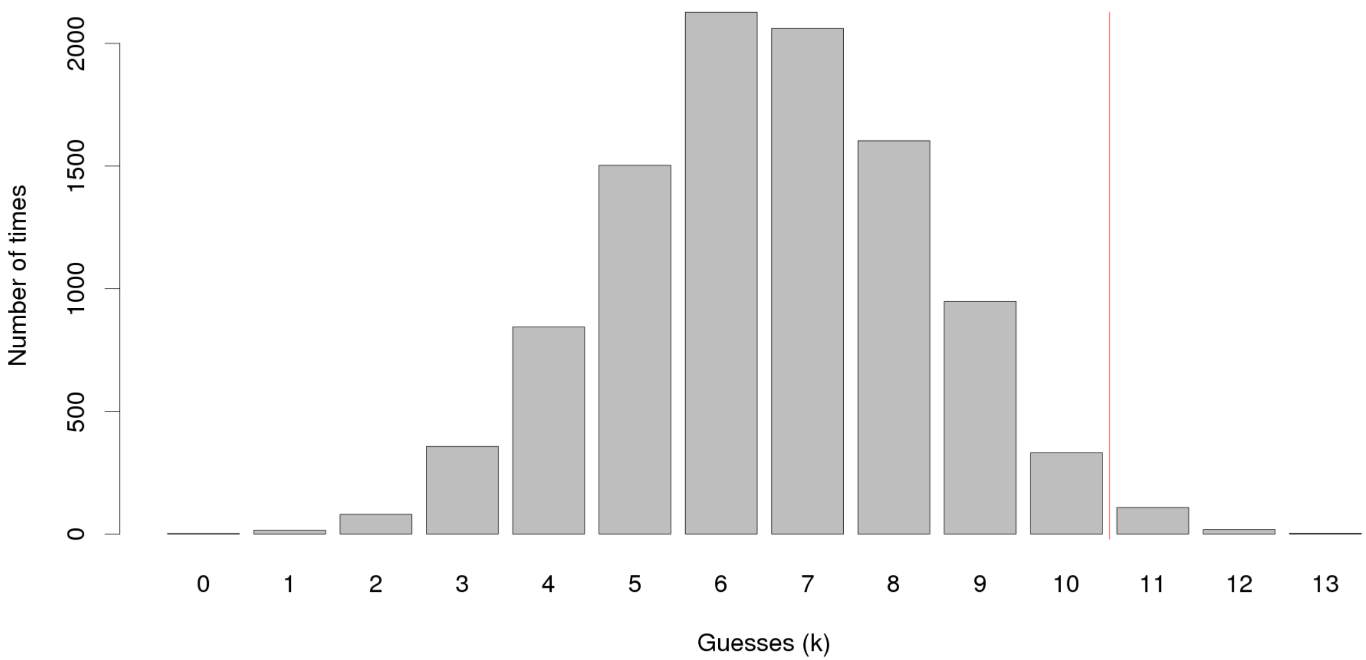
- Answer: $\pi = .5$

Question: How could we calculate the probability Paul would guess 11 or more games correctly?

- Answer 1: We could flip a fair coin 13 times and see how many times we get 11 or more heads. Then repeat this process 10,000 times.

- We can do this simulation in R using the following commands:
  - `> simulated.correct.guesses <- rbinom(10000, 13, .5)`
  - `> num.sims.as.good.as.paul <- sum(simulated.correct.guesses >= 11)`
  - `> proportion.as.good.as.paul <- num.sims.as.good.as.paul/10000`

Paul the Octopus



What are some answers for how often a randomly guessing octopus would guess 11 out of 13 games correct?

I got: 129 of 10,000 simulated trials have 11 or more correct guesses

- If Paul was guessing, he would only get 11 right 129/10,000 = 1.2% of the time

Paul the Octopus

Do you think Paul is psychic?

Paul the Octopus: second approach

Question: If Paul was psychic, what proportion of games would he have guessed correctly?

- Answer: $\pi = .5$

Question: How can we calculate the probability he would guess 11 games correctly?

- Answer 2: We could use the binomial distribution to the probability of getting 11 or more heads

- We can do this in R using:
  - > sum(dbinom(11:13, 13, .5)) # sum Pr(X = 11) + Pr(X = 12) + Pr(X = 13)
  - > 1 – pbinom(10, 13, .5) # equivalently: 1 – Pr(X ≤ 10)