SEIS 631
Foundations of Data Analysis

Question

Roughly 20% of undergraduates at a university are vegetarian. What is the probability that, among a random sample of 3 undergraduates, at least one is vegetarian?

(a) 1 - 0.2 x 3

(b) $1 - 0.2^3$

(c) 0.8^3

(d) 1 - 0.8 x 3

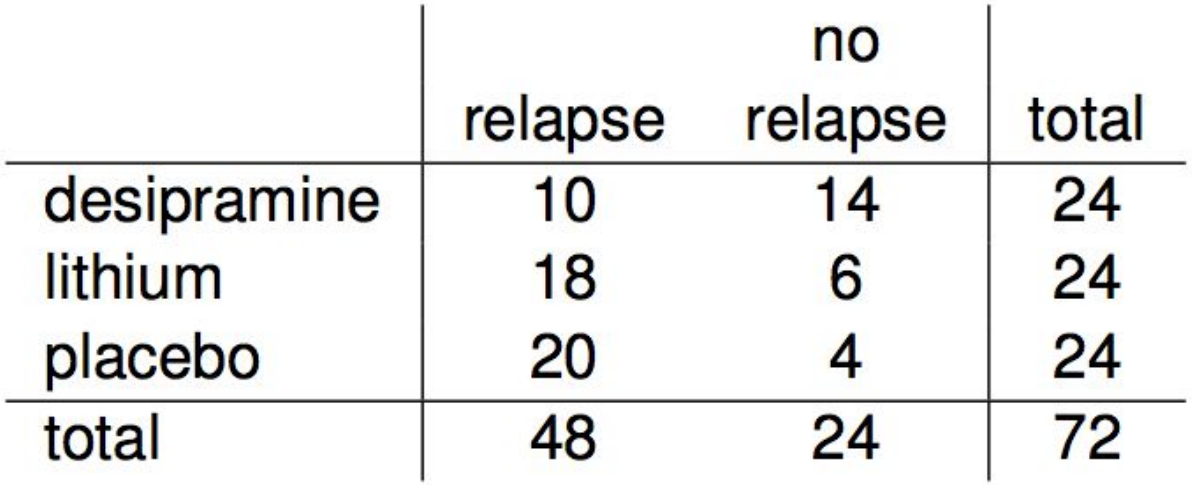
(e) $1 - 0.8^3$

P(At Least 1 Vegetarian) = 1 - P(none vegetarian)
= 1 - P(not vegetarian)^3
= 1 - 0.8^3
= 0.488

Applying Conditional Probability

Relapse

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below.

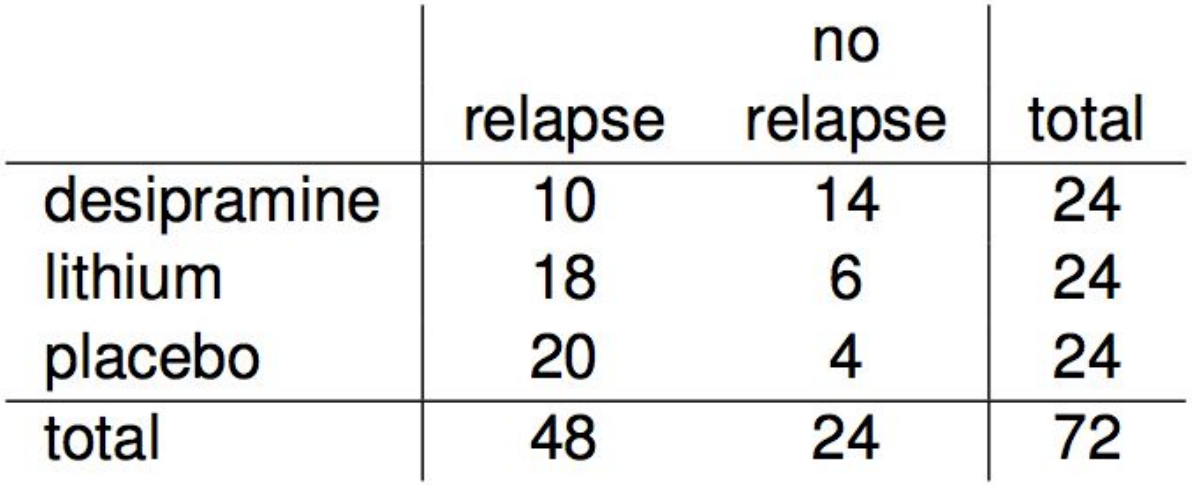


	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

Marginal probability

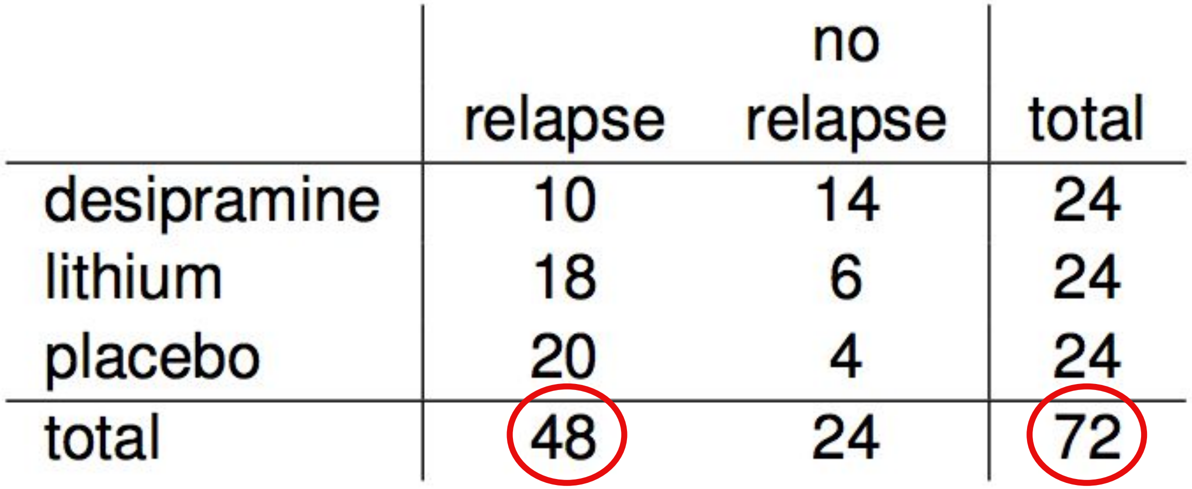
What is the probability that a patient relapsed?



	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Marginal probability

What is the probability that a patient relapsed?

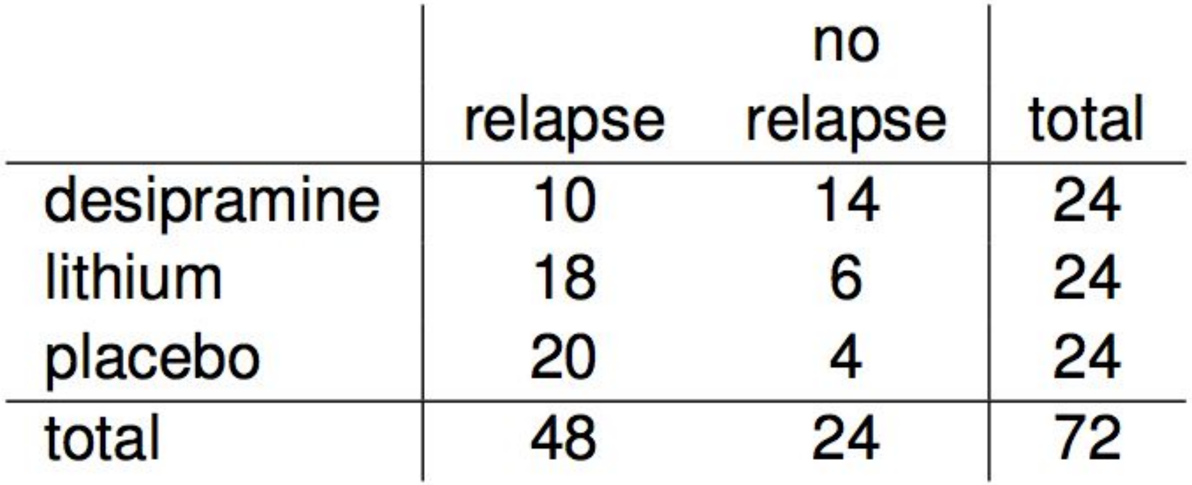


	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(relapse) = \frac{48}{72} \approx 0.67$$

Joint probability

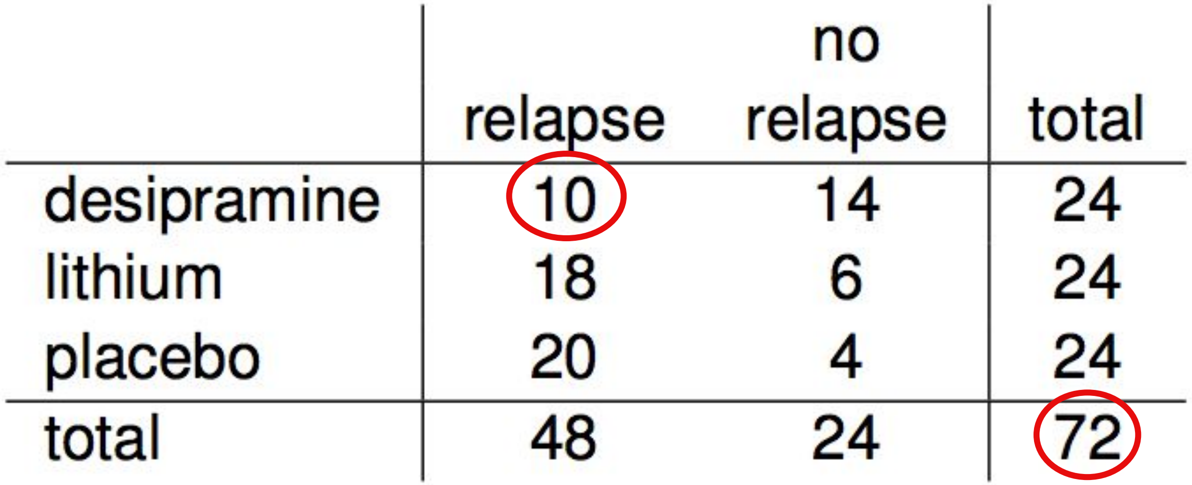
What is the probability that a patient received the antidepressant (desipramine) and relapsed?



		no	
	relapse	relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Joint probability

What is the probability that a patient received the antidepressant (desipramine) and relapsed?



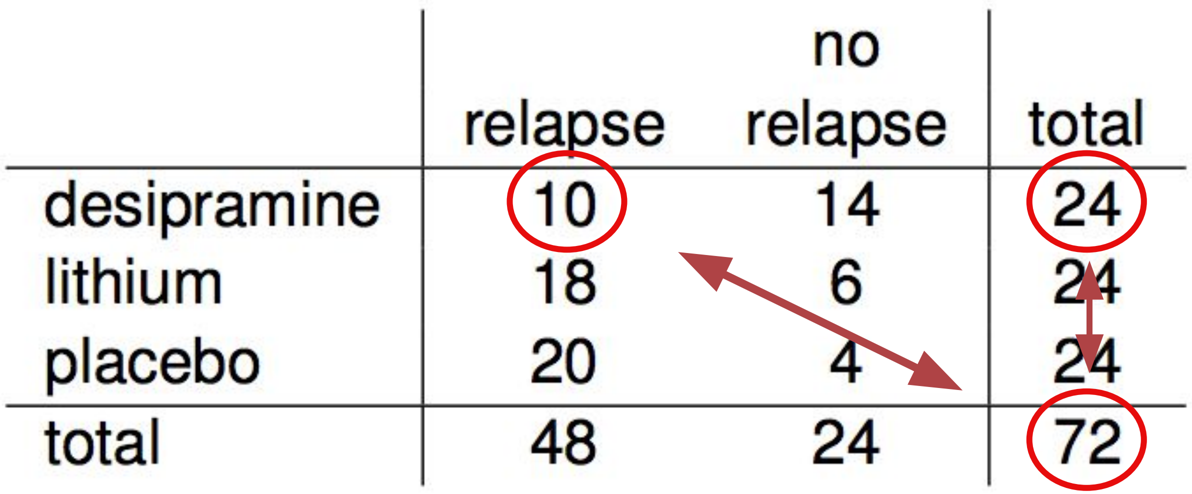
	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

$$P(\text{relapse and desipramine})$$

$$= \frac{10}{72} \approx 0.14$$

Conditional Probability

Given that a person was given desipramine, what is the probability that they relapsed?
In other words, what is the probability of relapse *conditional on desipramine*?



	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Method 1: Bayes Theorem

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{A}\ \mathit{and}\ \mathrm{B})}{P(\mathrm{B})}$$

$$P(\mathrm{rel}|\mathrm{desip}) = \frac{P(\mathrm{rel\ and\ desip})}{P(\mathrm{desip})}$$

$$= \frac{10/72}{24/72}$$

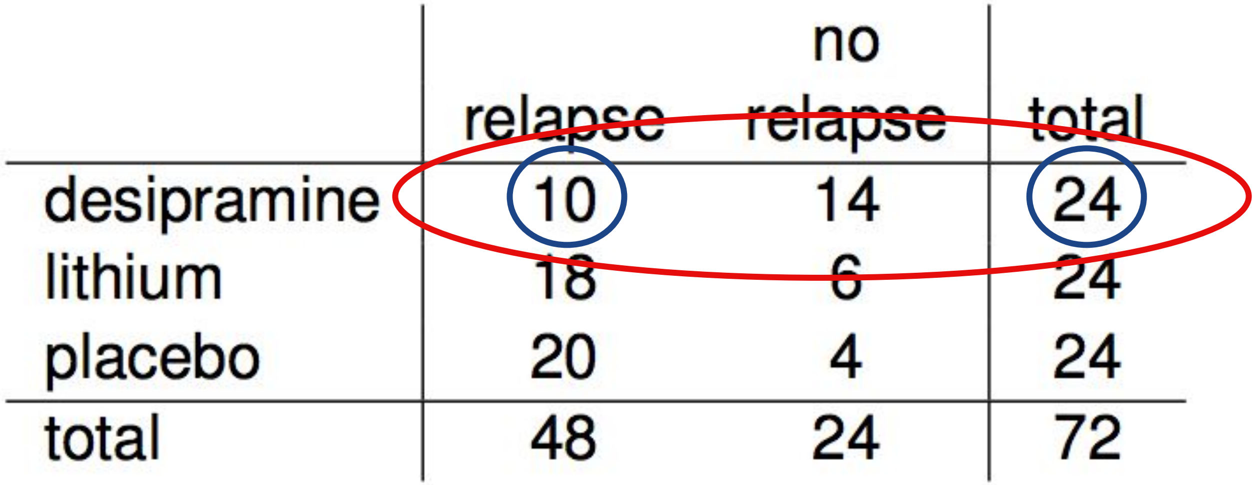
$$= \frac{10}{24}$$

$$\approx 0.42$$

Conditional Probability

Given that a person was given desipramine, what is the probability that they relapsed?
In other words, what is the probability of relapse *conditional on desipramine*?

Method 2: Change your perspective



	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

"Given the person took desipramine"

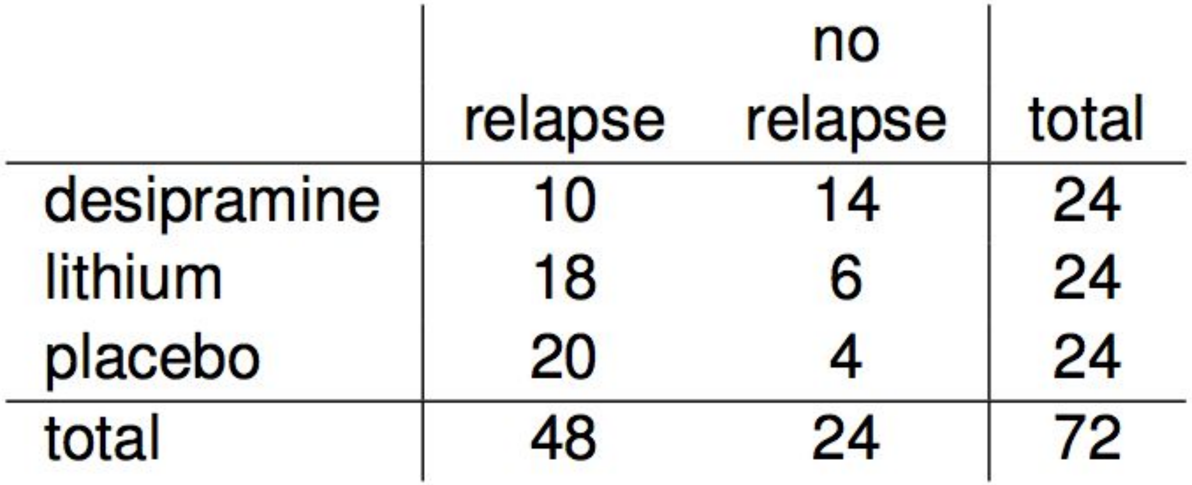
Focus only on the desipramine row

$$P(\text{rel}|\text{desip}) = \frac{10}{24}$$

$$\approx 0.42$$

Comparing the drugs

How do the conditional probabilities of relapse compare for each drug?



	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

P(relapse | desipramine) = 10 / 24 ~ 0.42

P(relapse | lithium) = 18 / 24 ~ 0.75

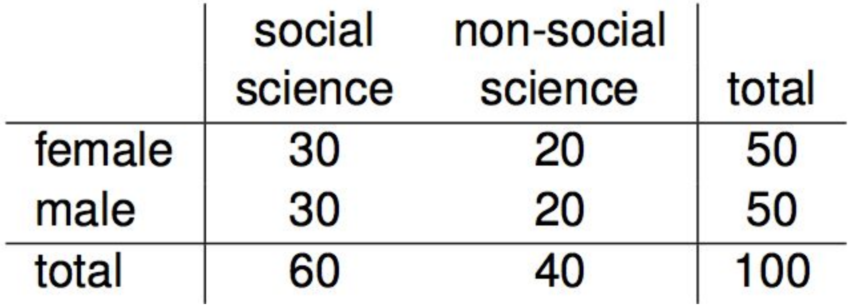
P(relapse | placebo) = 20 / 24 ~ 0.83

What does this tell us?

Desipramine has a much lower probability of relapse which is evidence that it is likely a more effective treatment for cocaine addiction.

Example

Consider the following (hypothetical) distribution of gender and major of students in an introductory statistics class:



	social science	non-social science	total
female	30	20	50
male	30	20	50
total	60	40	100

- The probability that a randomly selected student is a social science major is P(SS) = 60 / 100 = 0.6.
- The probability that a randomly selected student is a social science major given that they are female is P(SS|F) = 30 / 50 = 0.6
- Similarly, P(SS|M) = 30/50 = 0.6
- P(SS) = P(SS|M) = P(SS|F) → Gender and Major are Independent

Probability Trees

Inverting Probabilities, i.e. P(A|B) ----> P(B|A)

Example

Suppose 13% of students earned an A on the midterm. Of those students who earned an A on the midterm, 47% received an A on the final, and 11% of the students who earned lower than an A on the midterm received an A on the final. You randomly pick up a final exam and notice the student received an A. What is the probability that this student earned an A on the midterm?

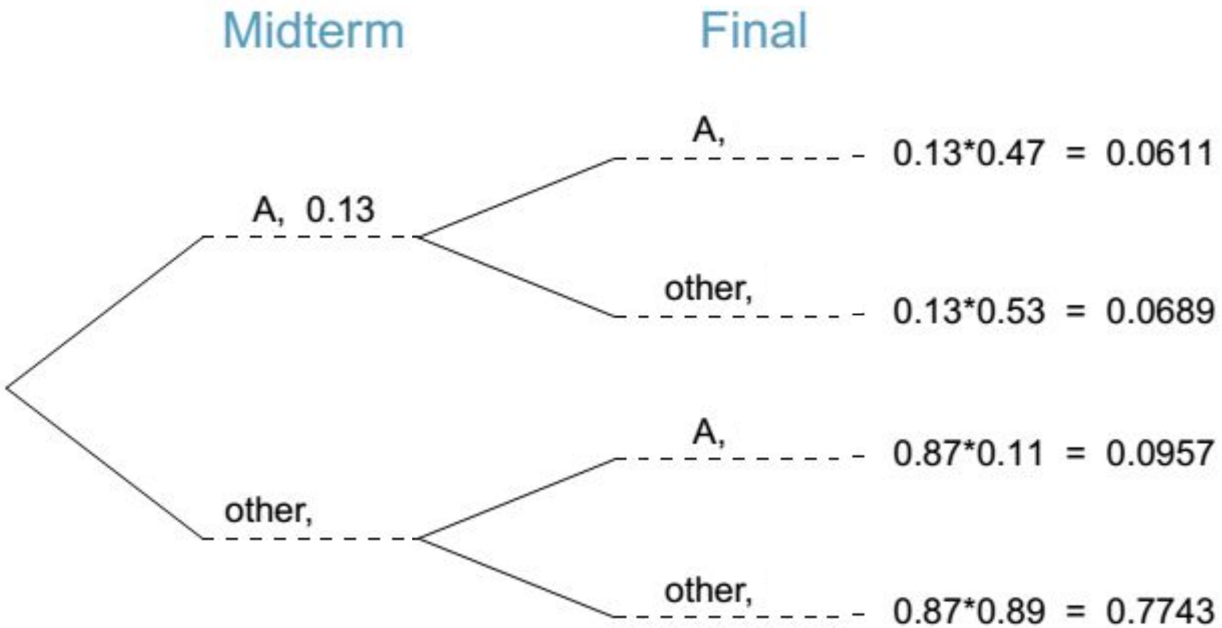
Bayes Theorem

P(midterm=A) = 0.13
P(final=A | midterm=A) = 0.47
P(final=A | midterm=other) = 0.11

$$P(X|Y) = \frac{P(X \text{ and } Y)}{P(Y)} = \frac{P(Y|X) \cdot P(X)}{P(\mathrm{Y})}$$

P(midterm=A|final=A) = ?

$$\textit{We need to calculate } P\,(\textbf{midterm} = \textbf{A} \mid \textbf{final} = \textbf{A}) \text{ and } P\,(\textbf{final} = \textbf{A})$$



P(midterm=A) = 0.13
P(final=A | midterm=A) = 0.47
P(final=A | midterm=other) = 0.11

P(midterm=A|final=A) = ?

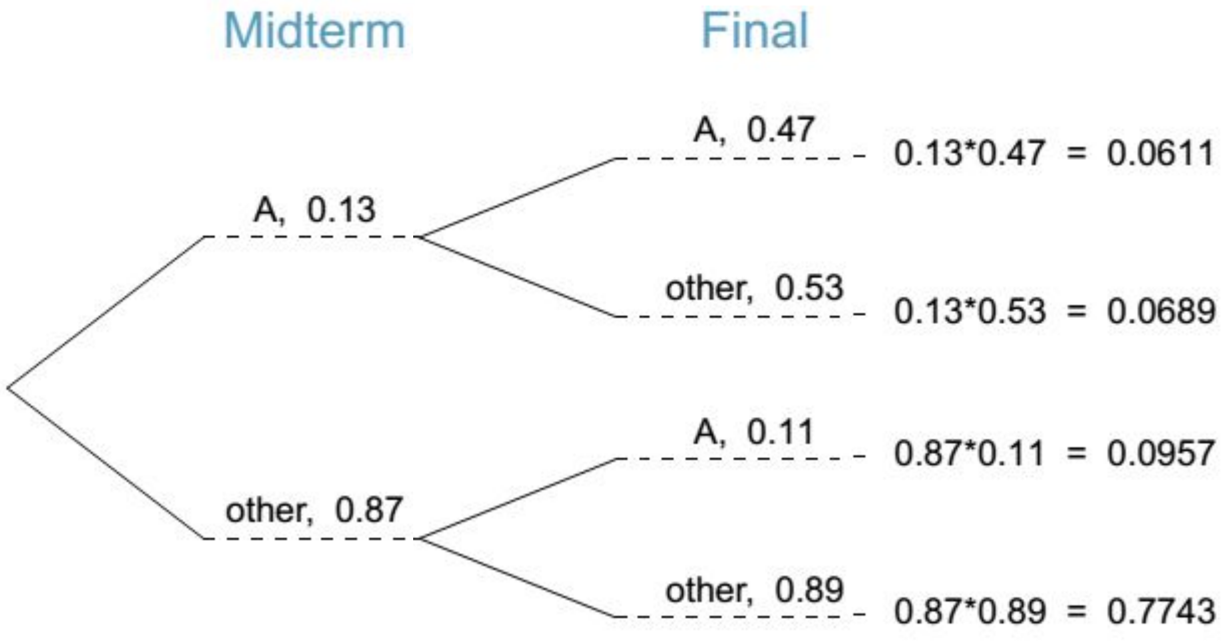
$$P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) = 0.0611$$

$$P(\underline{\texttt{final} = \texttt{A}}) = P(\texttt{midterm} = \texttt{other} \text{ and } \underline{\texttt{final} = \texttt{A}}) + P(\texttt{midterm} = \texttt{A} \text{ and } \underline{\texttt{final} = \texttt{A}}) \quad = 0.0957 + 0.0611 = 0.1568$$

$$\begin{aligned} P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A}) &= \frac{P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A})}{P(\texttt{final} = \texttt{A})} \\ &= \frac{0.0611}{0.1568} = 0.3897 \end{aligned}$$

(posterior probability)

OpenIntro Statistics, Third Edition, by Diez, Barr, Cetinkaya-Rundel



P(midterm=A) = 0.13
P(final=A | midterm=A) = 0.47
P(final=A | midterm=other) = 0.11

P(midterm=A|final=A) = ?

$$P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A}) = 0.0611$$

$$P(\underline{\texttt{final} = \texttt{A}}) = P(\texttt{midterm} = \texttt{other} \text{ and } \underline{\texttt{final} = \texttt{A}}) + P(\texttt{midterm} = \texttt{A} \text{ and } \underline{\texttt{final} = \texttt{A}}) \quad = 0.0957 + 0.0611 = 0.1568$$

$$\begin{aligned} P(\texttt{midterm} = \texttt{A} | \texttt{final} = \texttt{A}) \quad &= \quad \frac{P(\texttt{midterm} = \texttt{A} \text{ and } \texttt{final} = \texttt{A})}{P(\texttt{final} = \texttt{A})} \\ &= \frac{0.0611}{0.1568} = 0.3897 \end{aligned}$$

(posterior probability)

OpenIntro Statistics, Third Edition, by Diez, Barr, Cetinkaya-Rundel

Question

As of June 22 at 6pm, 45.3% of Americans (all, not just eligible) have gotten a COVID-19 vaccine. If we randomly select 5 Americans, what is the probability that at least 1 is NOT vaccinated?

P(at least 1 is not vaccinated) = 1 - P(none are NOT vaccinated)

= 1 - P(all are vaccinated)

= 1 - P(vaccinated)^5

= 1 - 0.453^5

= 0.9809

If we select 5 Americans at random, the probability that at least one of them is vaccinated is about 98%

Source: https://www.beckershospitalreview.com/public-health/states-ranked-by-percentage-of-population-vaccinated-march-15.html

Question

You roll 5 dice (like in Yahtzee). Find the following probabilities.

1. The probability of rolling a 5 on all of them?

   $(\frac{1}{6})^5$

2. The probability of rolling less than (but not equal to) 5 on all of them?

   $(4/6)^5 = (\frac{2}{3})^5 = 0.13$

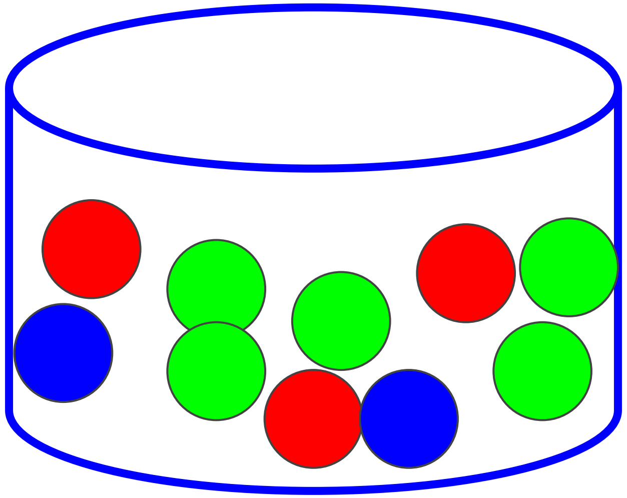
3. The probability of rolling at least one 6.

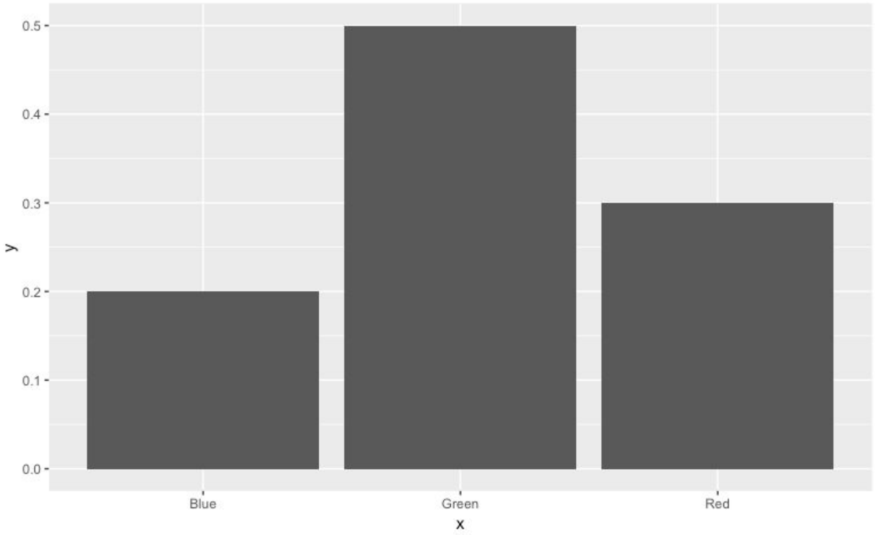
   $P(\text{at least 1 six}) = 1 - P(\text{no sixes}) = 1 - P(\text{not six})^5 = 1 - (\frac{5}{6})^5 = 0.598$

Probability Distributions

Probability Distributions

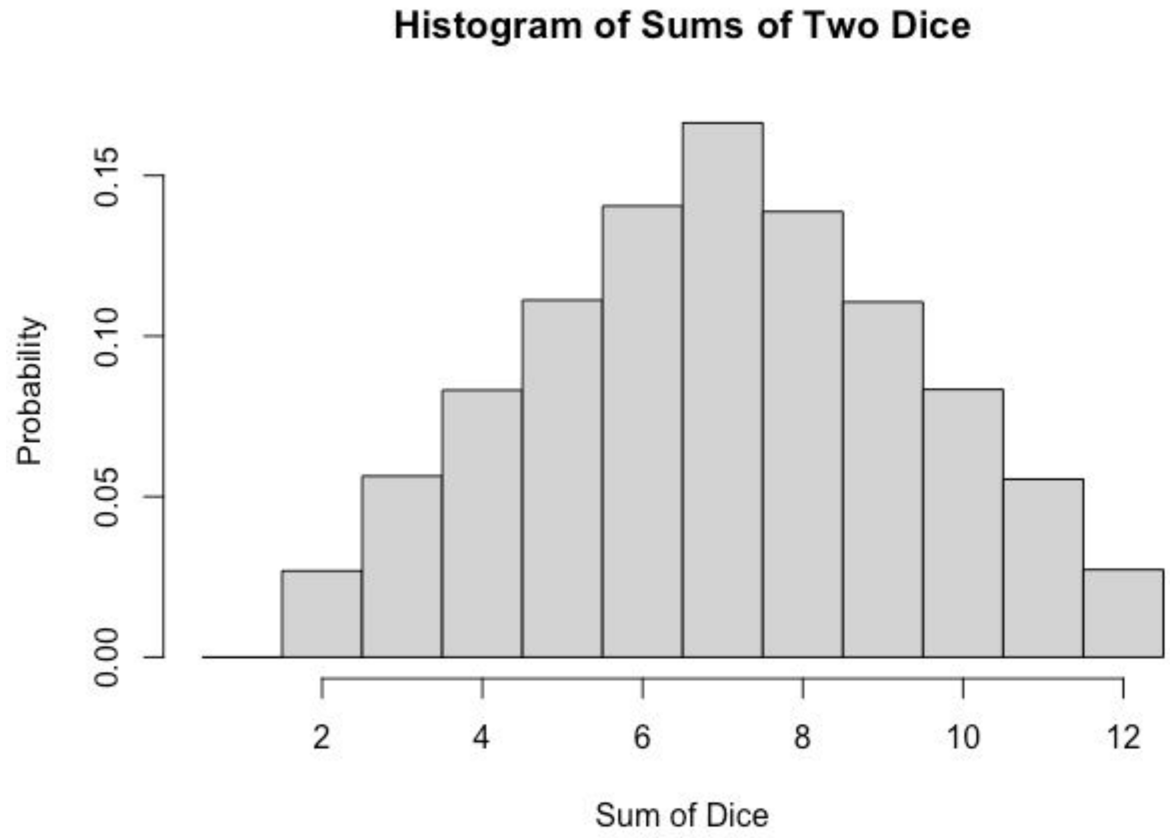
- Similar to sample space: A list of all the possible outcomes, but it also contains the PROBABILITY of each outcome:
- If I draw one marble from the jar at random
  - S = {R, G, B}
  - The probability distribution is:





Event	Red	Green	Blue
Probability	0.3	0.5	0.2

Probability Distribution for the Sum of Two Dice

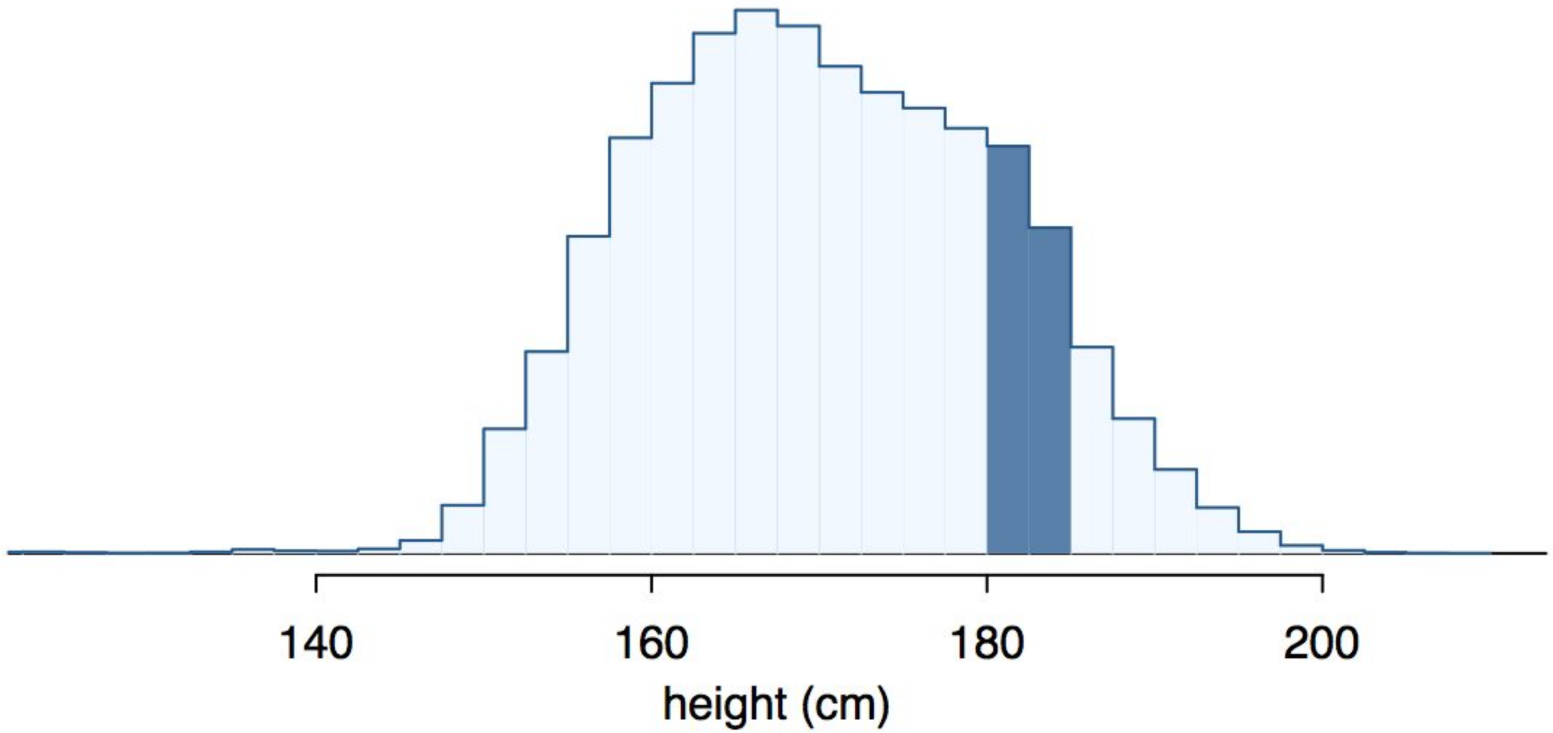


Continuous Distributions

Slides adapted from Mine Çetinkaya-Rundel of OpenIntro

Continuous distributions

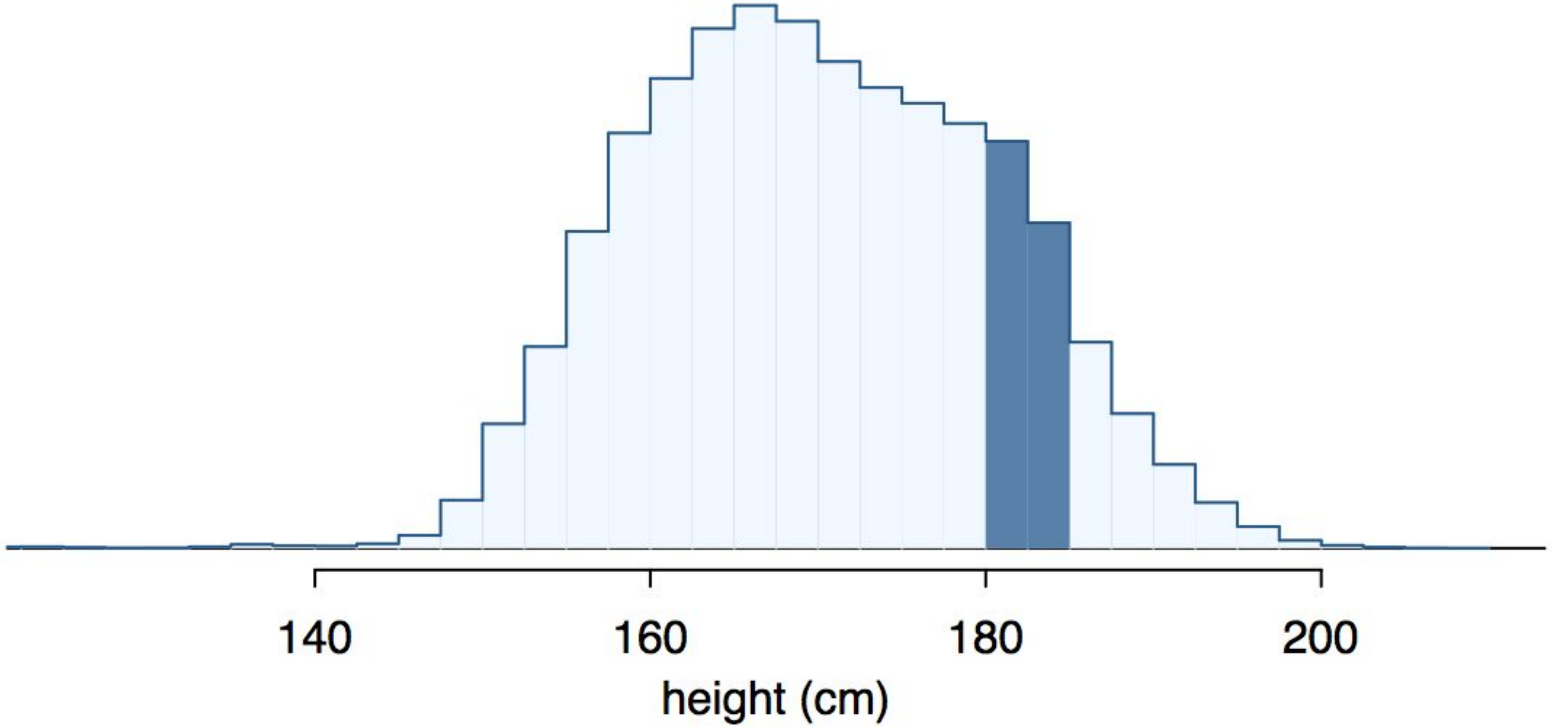
- Below is a proportional histogram of the distribution of heights of US adults.
- The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



Continuous distributions

- If the heights of the bins are scaled properly (i.e. so the total area = 1 or 100%)...

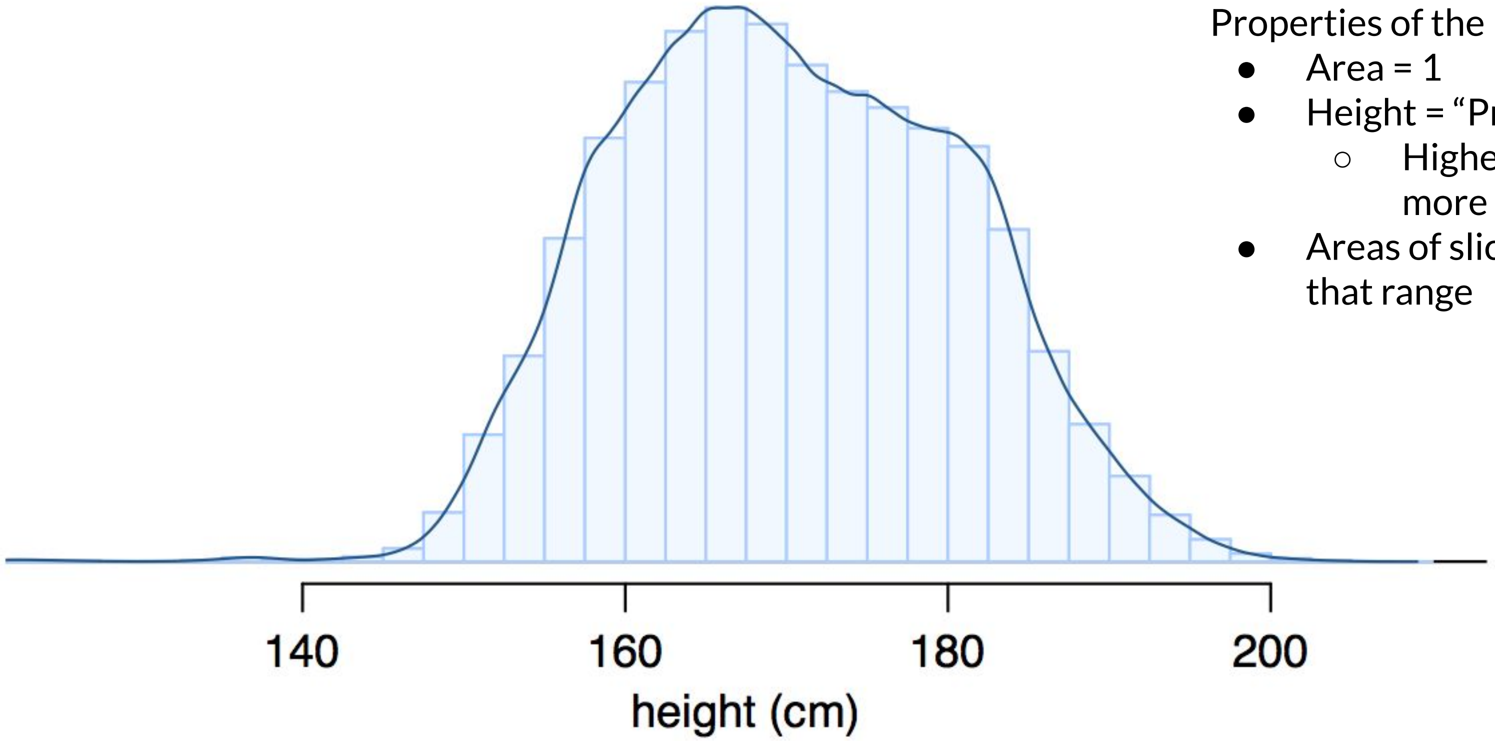
Area ⟷ Probability



From histograms to continuous distributions

If we imagine shrinking the bin width, but constantly re-scaling so the area remains 1, the histogram becomes a smooth curve.

Probability Density Function (PDF)

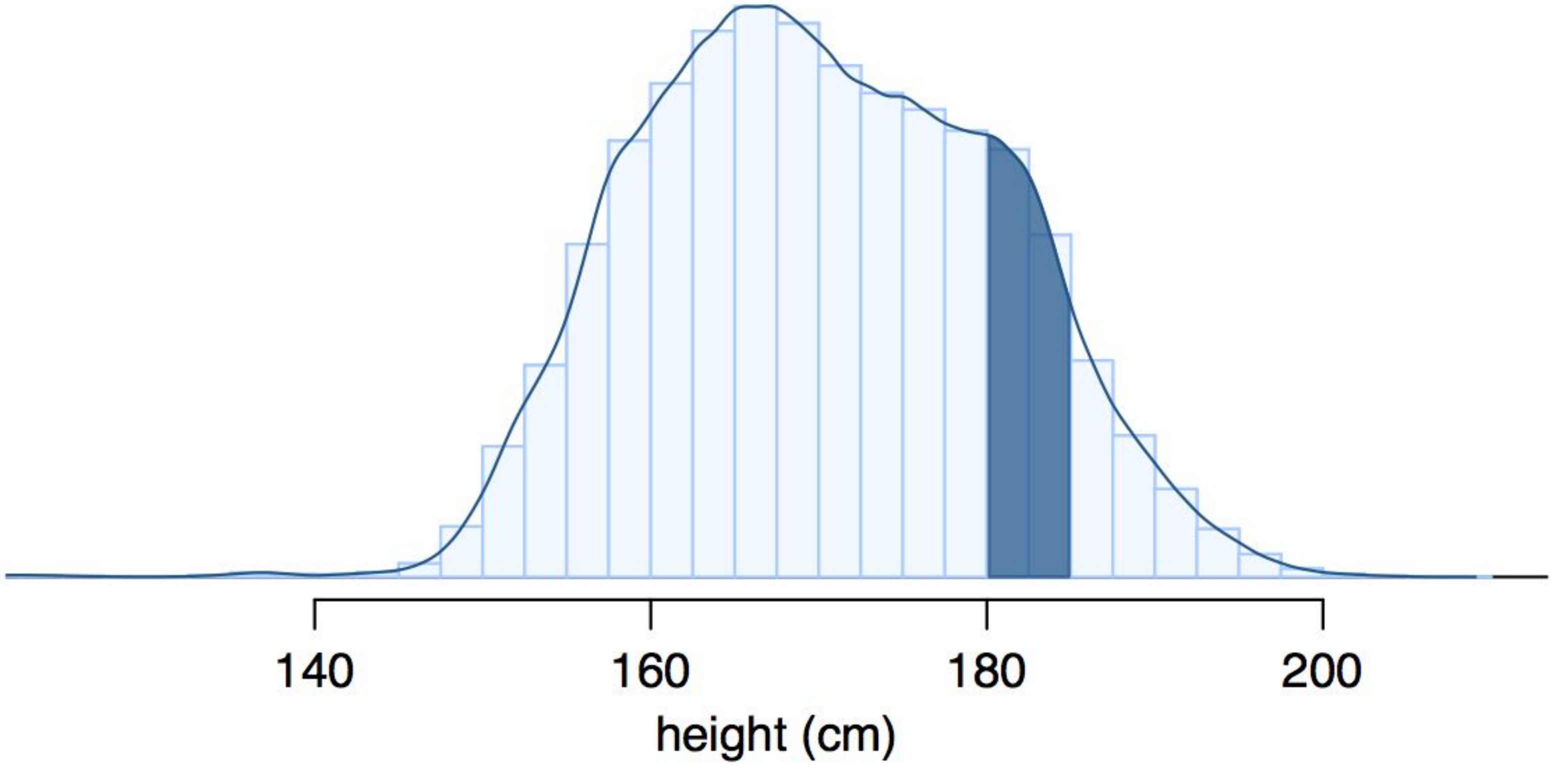


Properties of the PDF

- Area = 1
- Height = “Probability Density”
  - Higher → more dense, a.k.a more likely
- Areas of slices = probability of being in that range

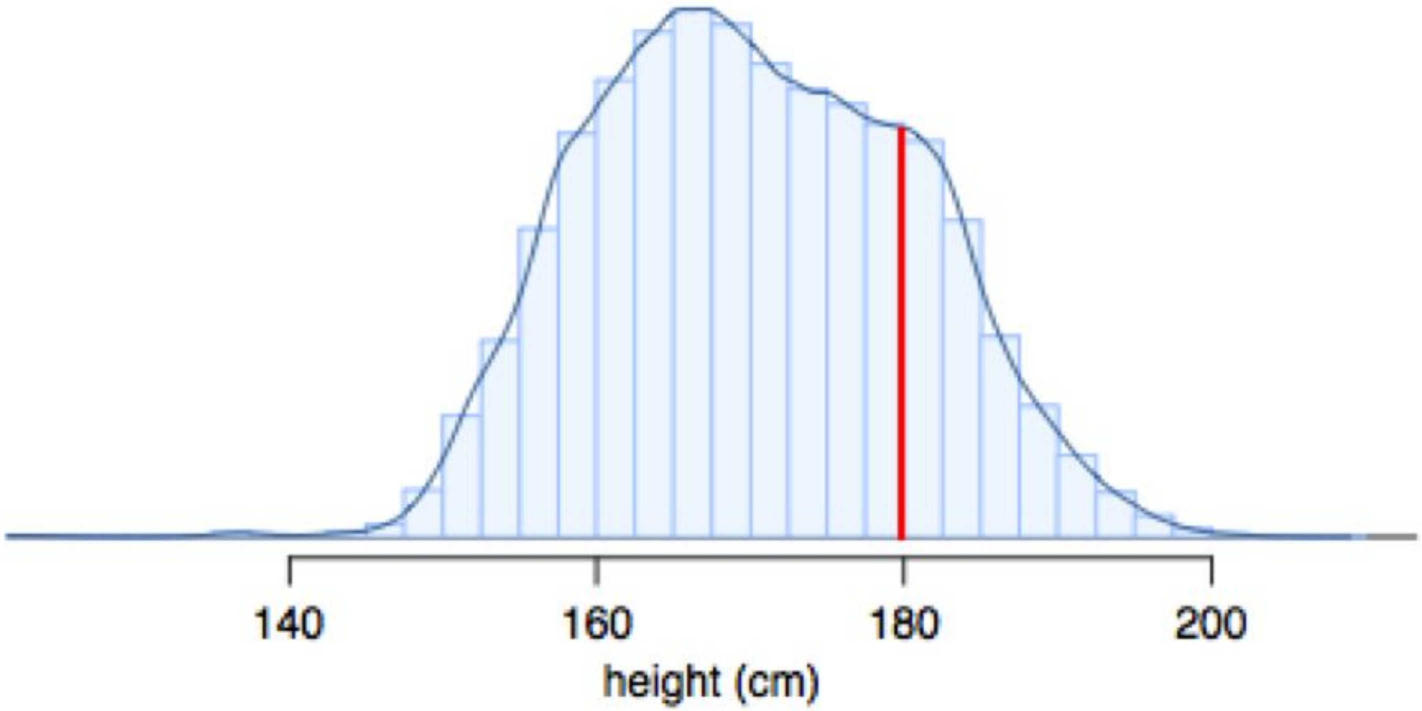
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



By definition...

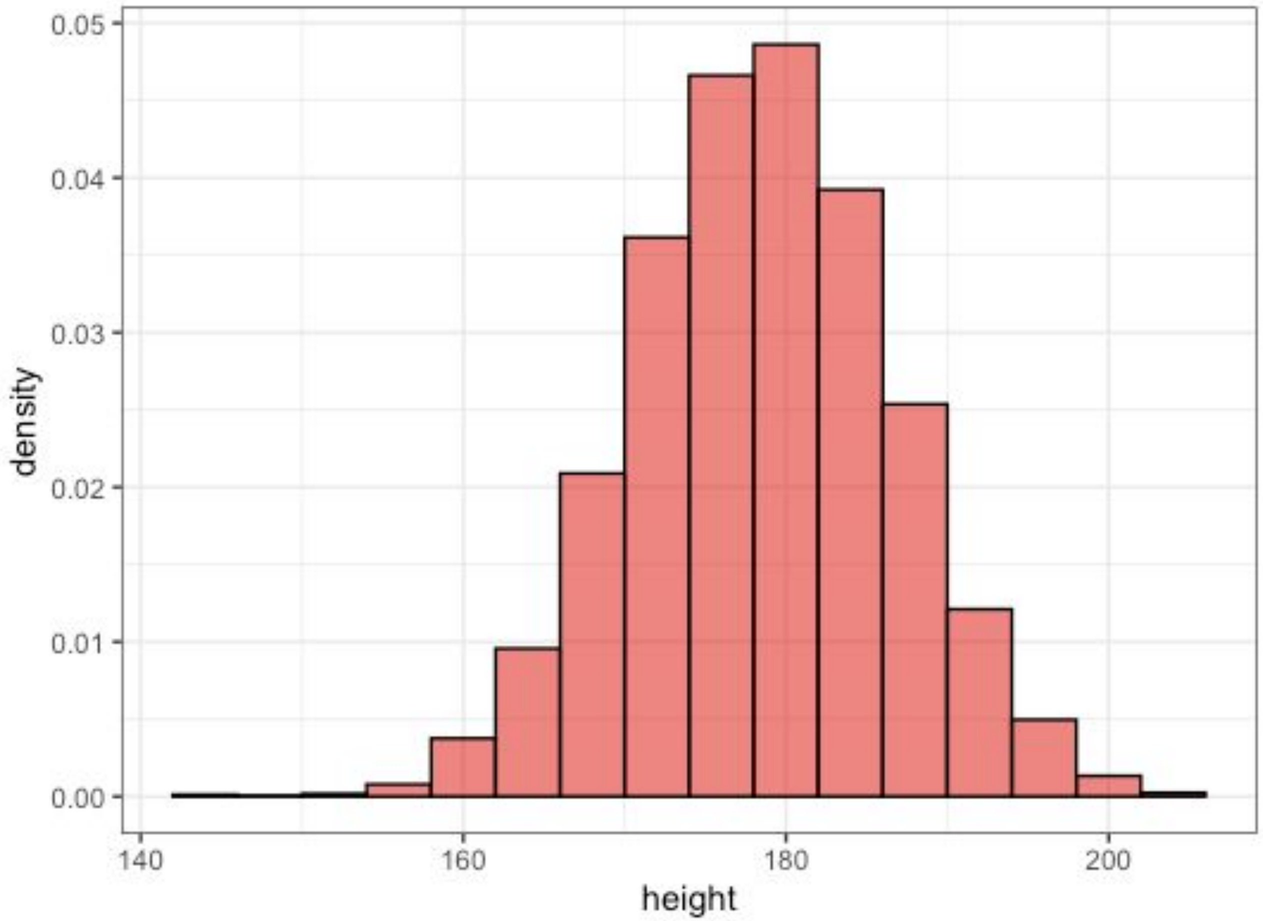
Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



Normal distribution

Slides developed by Mine Çetinkaya-Rundel of OpenIntro

How would you describe this distribution?

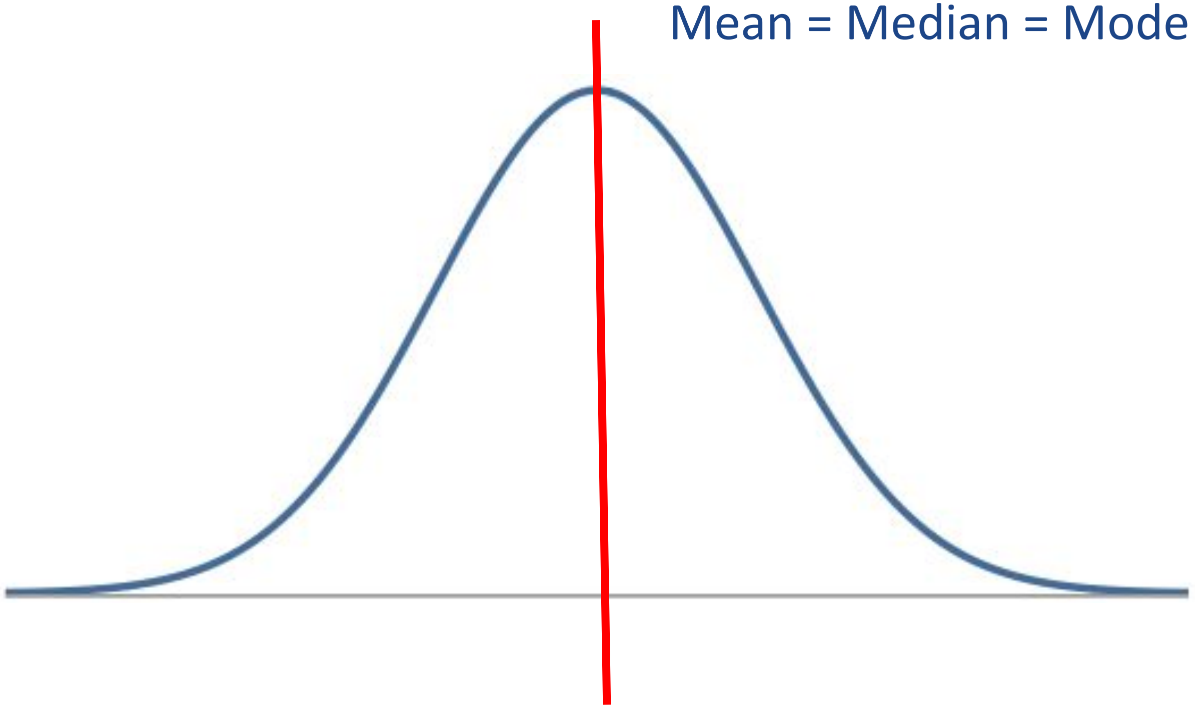


- Symmetric
- Unimodal
- “Bell” shaped

Many distributions take this shape. It’s an example of a very special distribution.

Normal Distribution

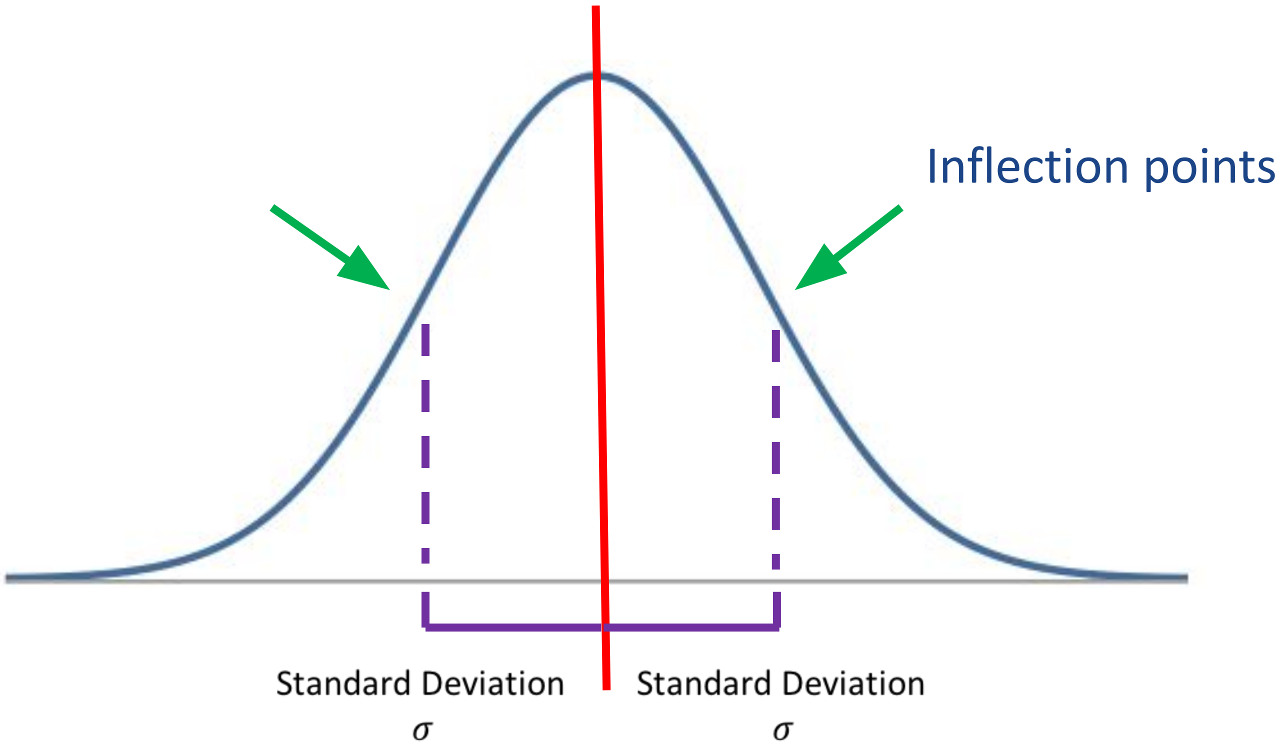
- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as N(μ, σ) → Normal with mean μ and standard deviation σ



Mean = Median = Mode

Normal Distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as N(μ, σ) → Normal with mean μ and standard deviation σ

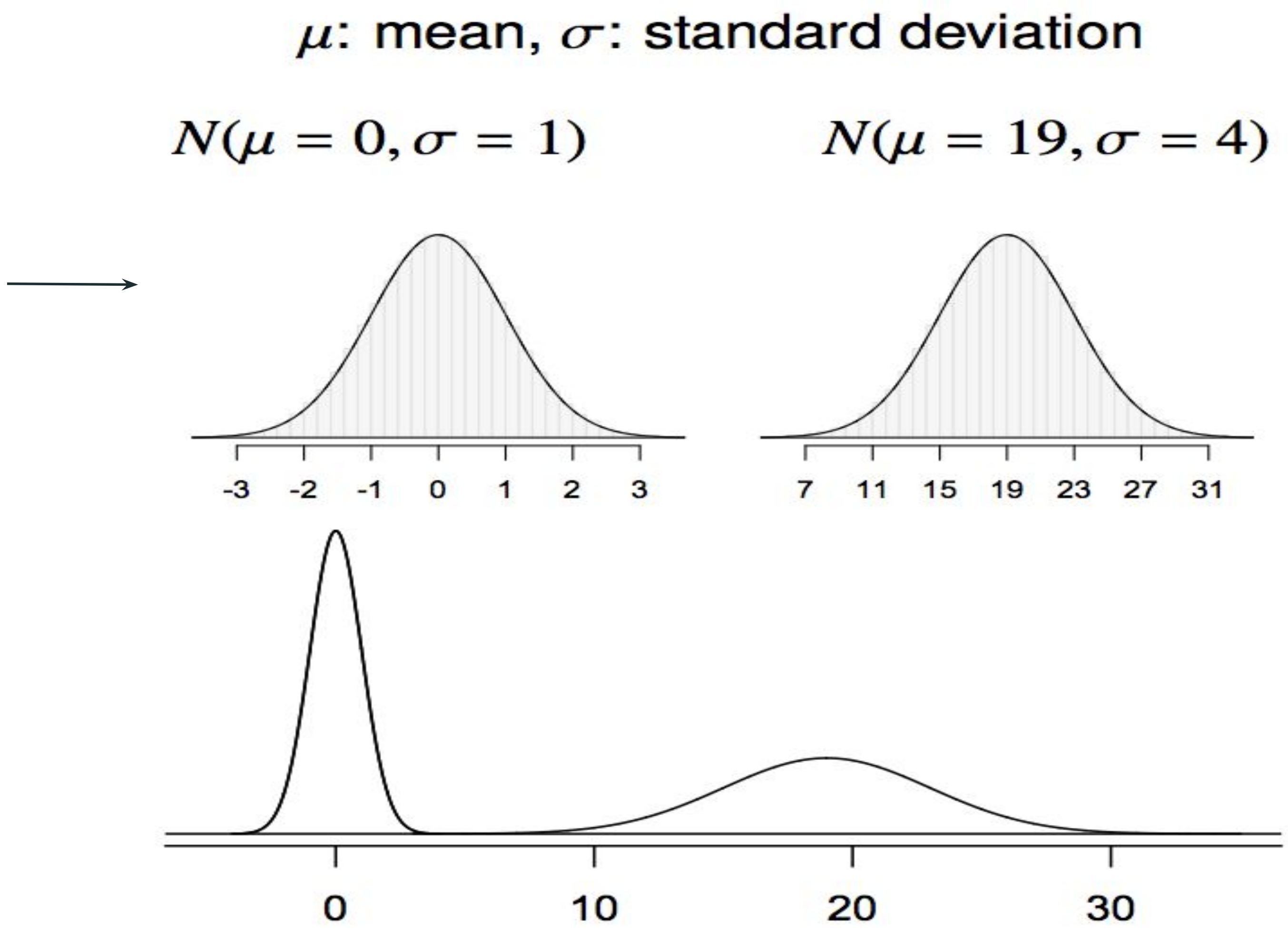


Not All Bell Shaped Distributions are “Normal”

- R Demo

Normal distributions with different parameters

Standard Normal

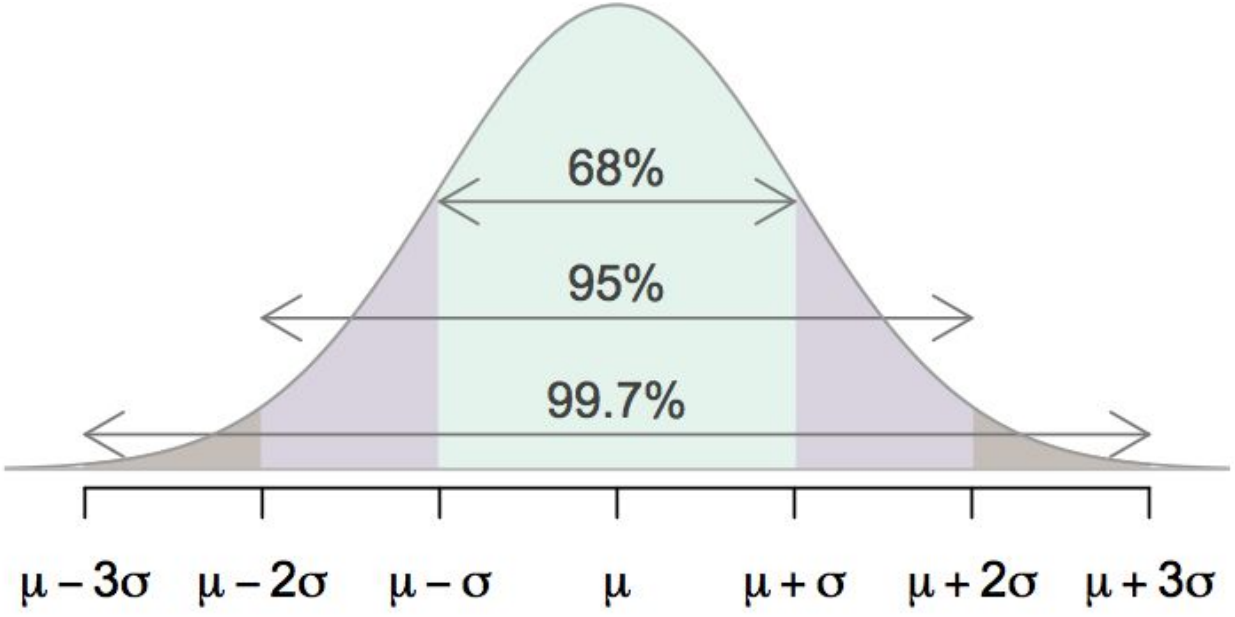


68-95-99.7 Rule

For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

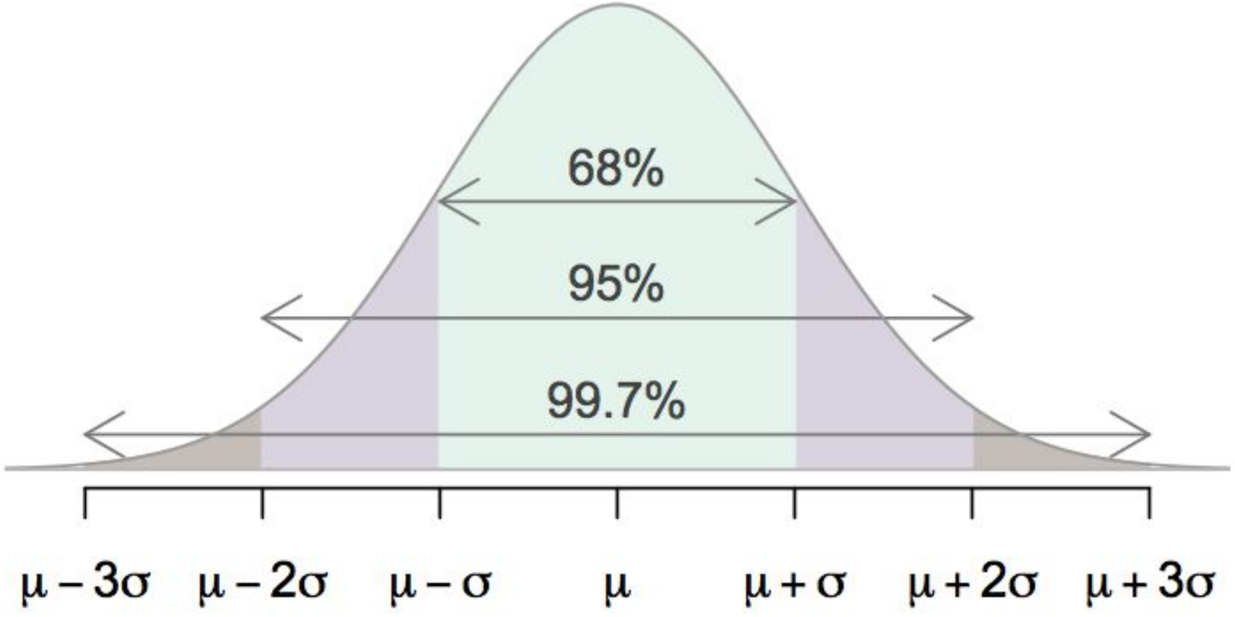
It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



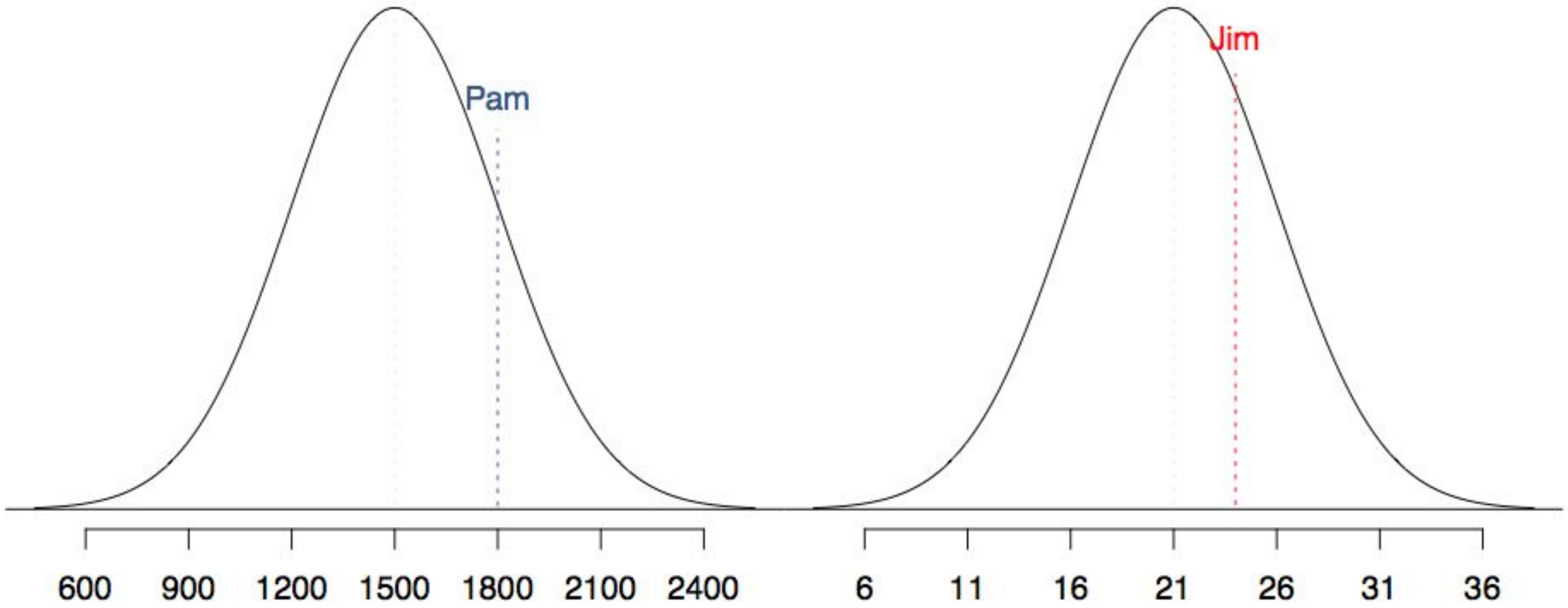
Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ~68% of students score between 1200 and 1800 on the SAT.
- ~95% of students score between 900 and 2100 on the SAT.
- ~99.7% of students score between 600 and 2400 on the SAT.



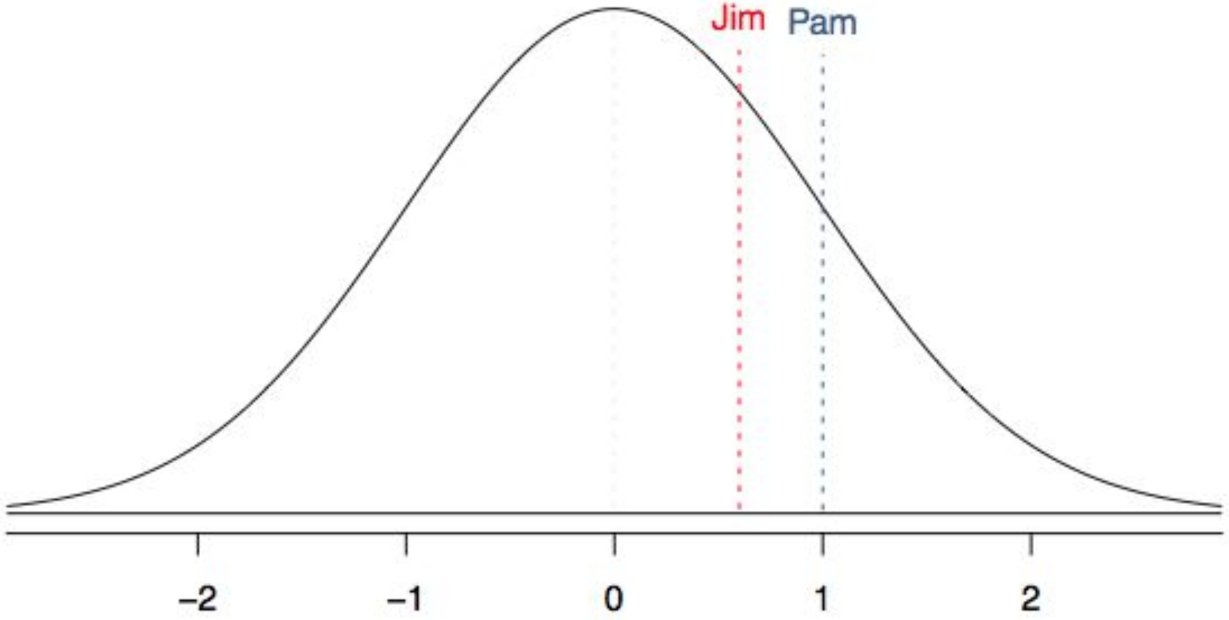
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is (24 - 21) / 5 = 0.6 standard deviations above the mean.



Standardizing with Z scores (cont.)

These are called standardized scores, or Z scores.

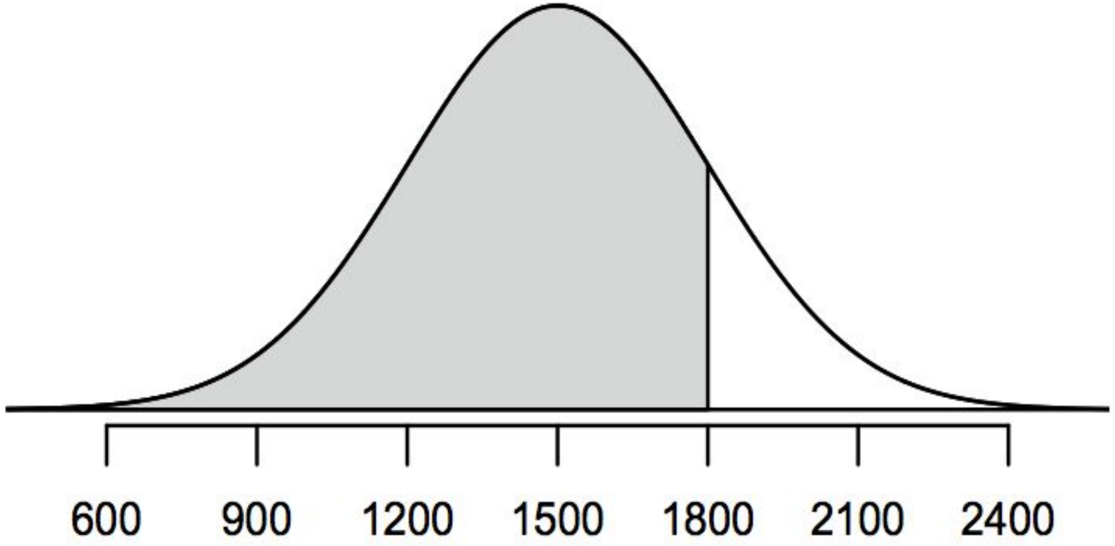
- Z score of an observation is the number of standard deviations it falls above or below the mean.

  $$Z = (\text{observation} - \text{mean}) / SD$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

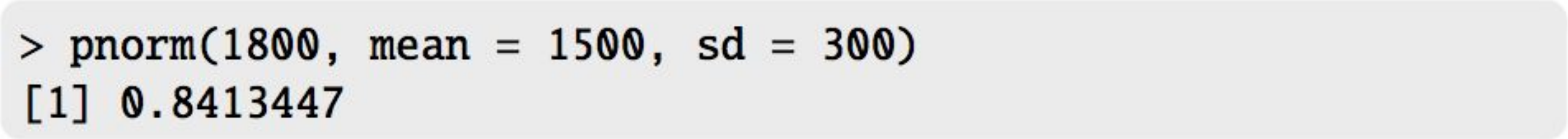
Percentiles

- Percentile is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.



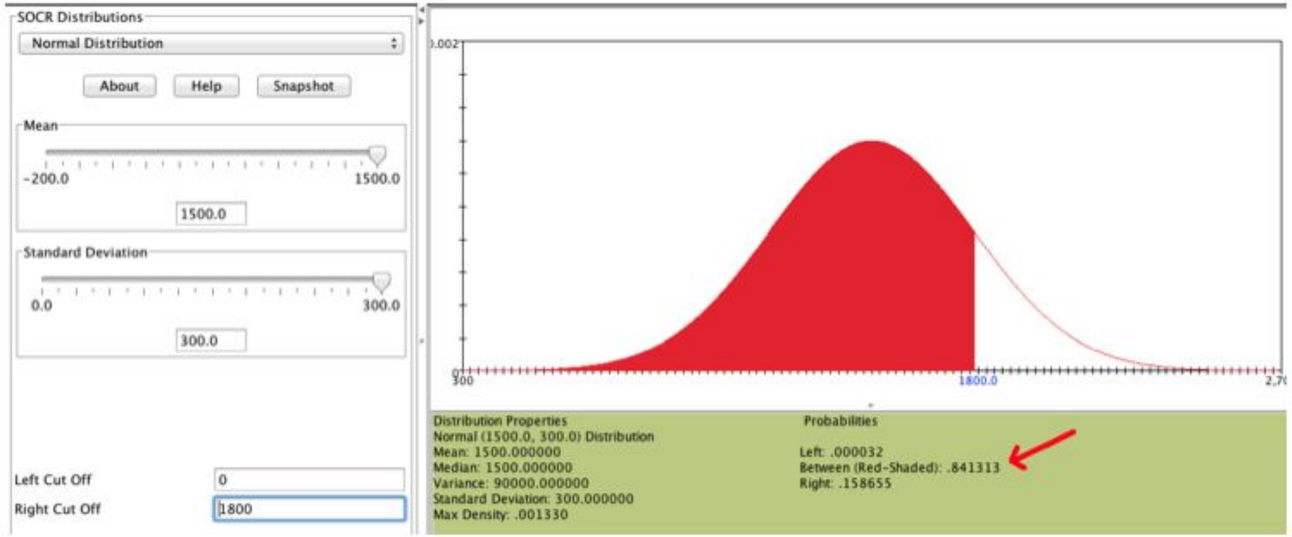
Calculating percentiles -- using computation

There are many ways to compute percentiles/areas under the curve. R:



```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

Applet: www.socr.ucla.edu/htmls/SOCR_Distributions.html



Mean: 1500.0

Standard Deviation: 300.0

Left Cut Off: 0

Right Cut Off: 1800

Distribution Properties
Normal (1500.0, 300.0) Distribution
Mean: 1500.000000
Median: 1500.000000
Variance: 90000.000000
Standard Deviation: 300.000000
Max Density: .001330

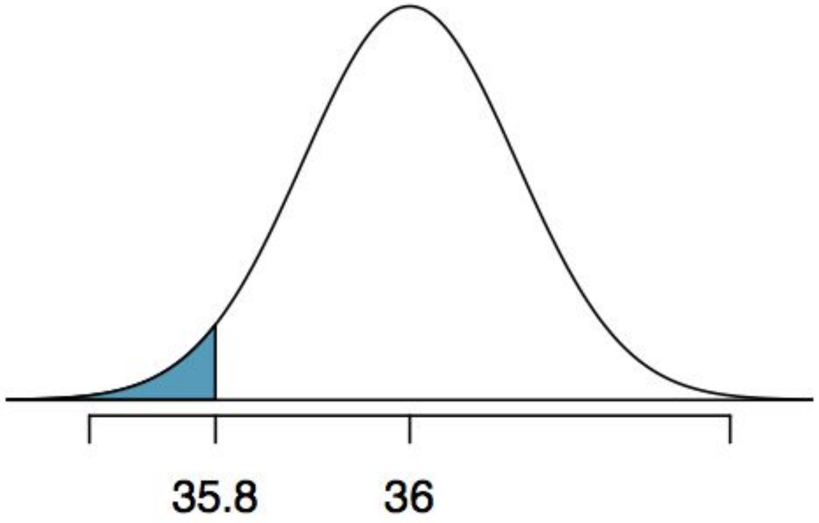
Probabilities

Left: .000032
Between (Red-Shaded): .841313
Right: .158655

Example

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

- *Let X = amount of ketchup in a bottle: X ~ N(μ = 36, σ = 0.11)*



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

```
>pnorm(35.8, 36, 0.11)
[1] 0.03451817
```

Question

What percent of bottles pass the quality control inspection?

(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%

Question

What percent of bottles pass the quality control inspection?

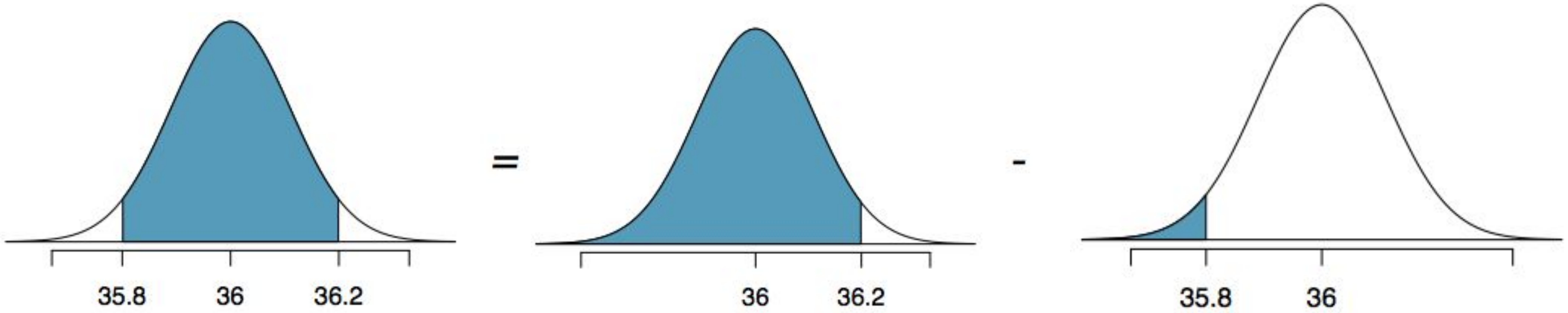
(a) 1.82%

(b) 3.44%

(c) 6.88%

(d) 93.12%

(e) 96.56%



$$Z_{35.8} = \frac{35.8 - 36}{0.11} = -1.82$$

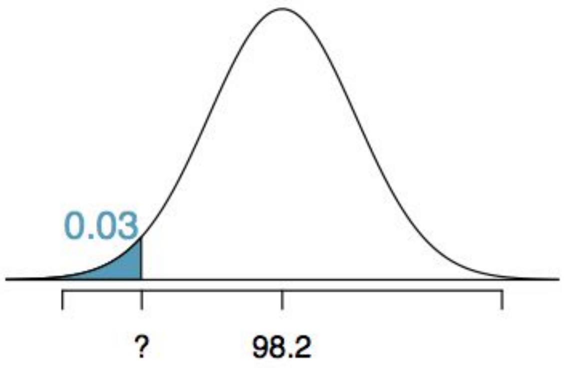
$$Z_{36.2} = \frac{36.2 - 36}{0.11} = 1.82$$

$$P(35.8 < X < 36.2) = P(-1.82 < Z < 1.82) = 0.9656 - 0.0344 = 0.9312$$

```
>pnorm(36.2,36,0.11) - pnrom(35.8, 36, 0.11)
>pnorm(1.82) - pnorm(-1.82)
```

Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2ºF and standard deviation 0.73ºF. What is the cutoff for the lowest 3% of human body temperatures?



R Function: qnorm( )
qnorm(0.03, 98.2, 0.73)

```
> qnorm(0.03, 98.2, 0.73)
[1] 96.82702
> qnorm(0.03)
[1] -1.880794
>
```

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88$$

$$x = (-1.88 \times 0.73) + 98.2 = 96.8^\circ F$$

Mackowiak, Wasserman, and Levine (1992), A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.

Question

Body temperatures of healthy humans are distributed nearly normally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F

Question

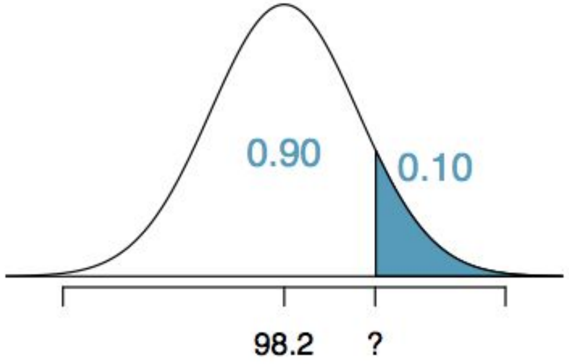
Body temperatures of healthy humans are distributed nearly normally with mean 98.2^{o}F and standard deviation 0.73^{o}F. What is the cutoff for the highest 10% of human body temperatures?

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F



R Function: qnorm( ) (final answer)

qnorm(0.9, 98.2, 0.73) or

qnorm(0.1, 98.2, 0.73, lower.tail = FALSE)

$$Z = \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28$$

$$x = (1.28 \times 0.73) + 98.2 = 99.1$$

To get z-score:

qnorm(0.9) or

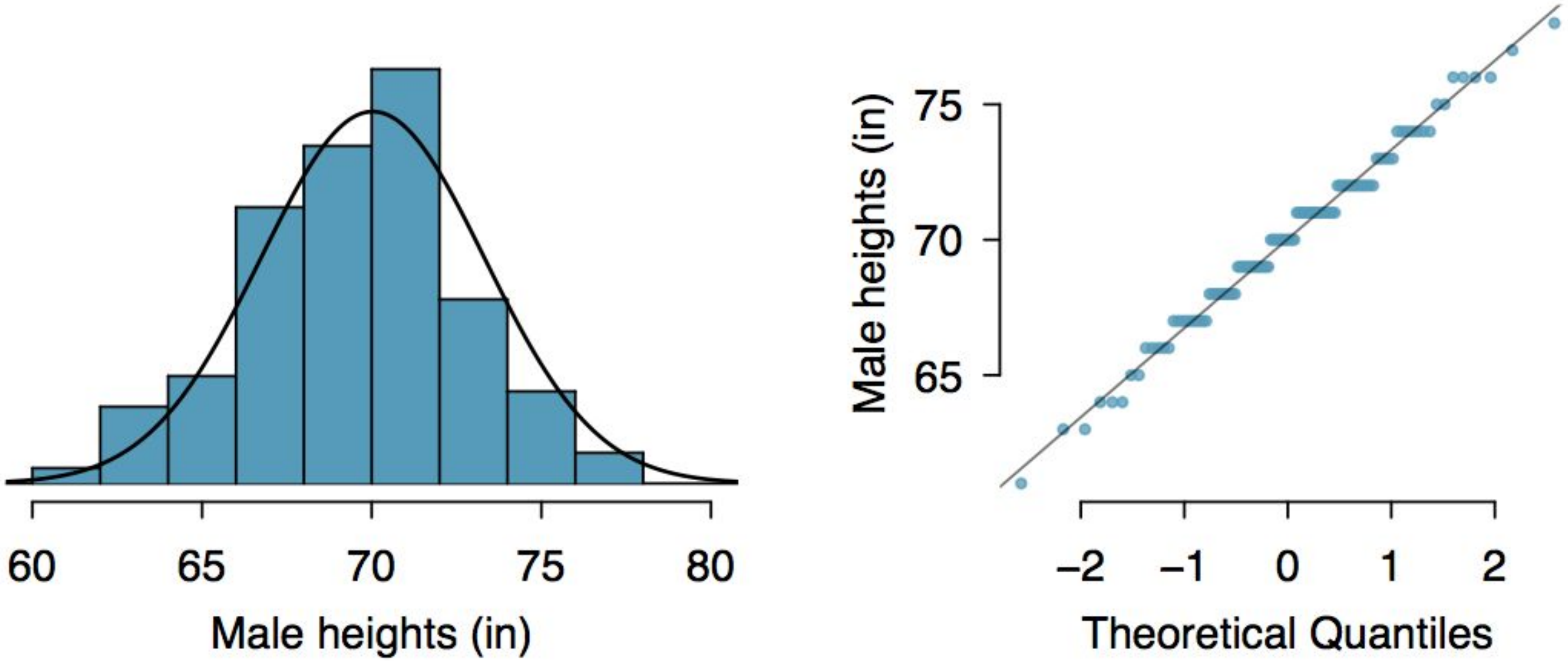
qnorm(0.1, lower.tail = FALSE)

Evaluating the normal distribution

Slides developed by Mine Çetinkaya-Rundel of OpenIntro

Normal probability plot

A histogram and normal probability plot of a sample of 100 male heights.

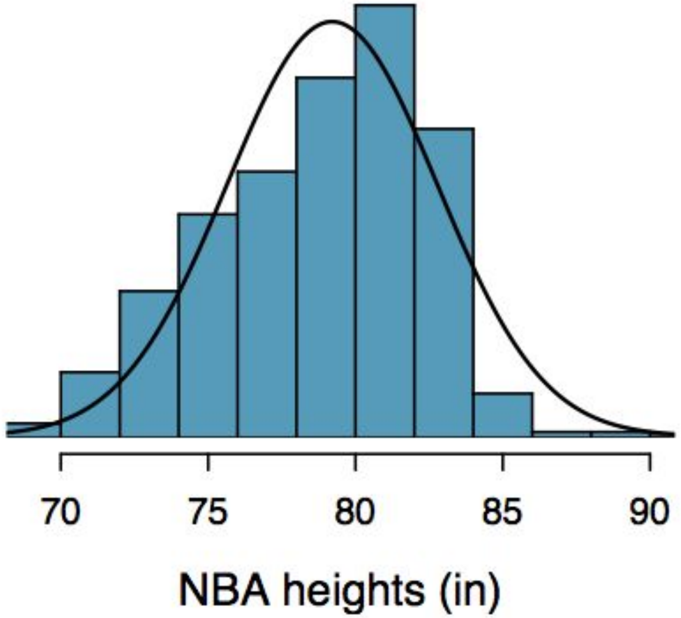


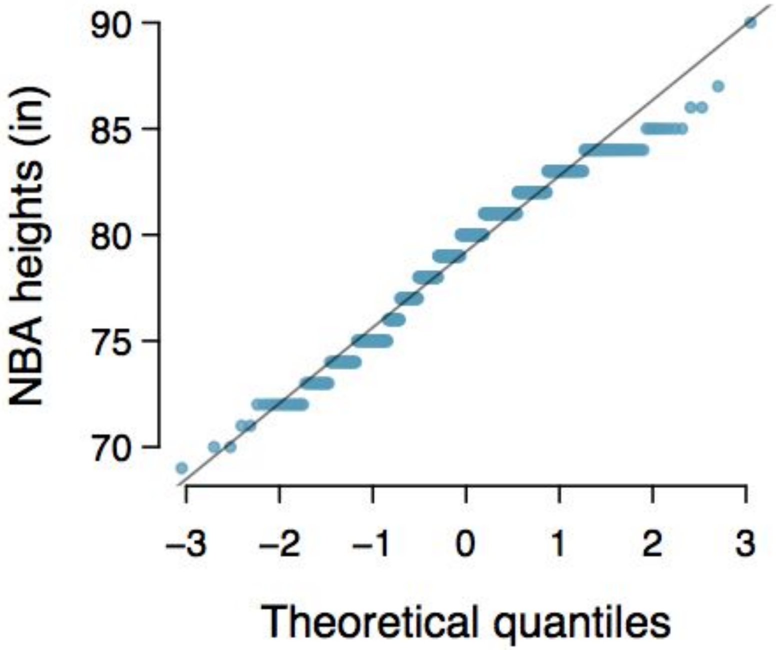
Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

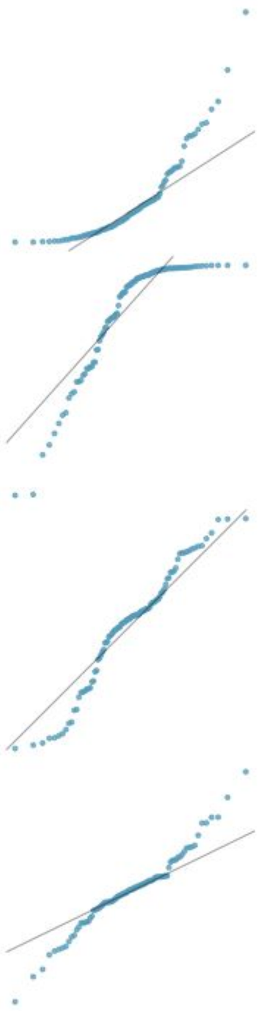
Question

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?





Normal probability plot and skewness

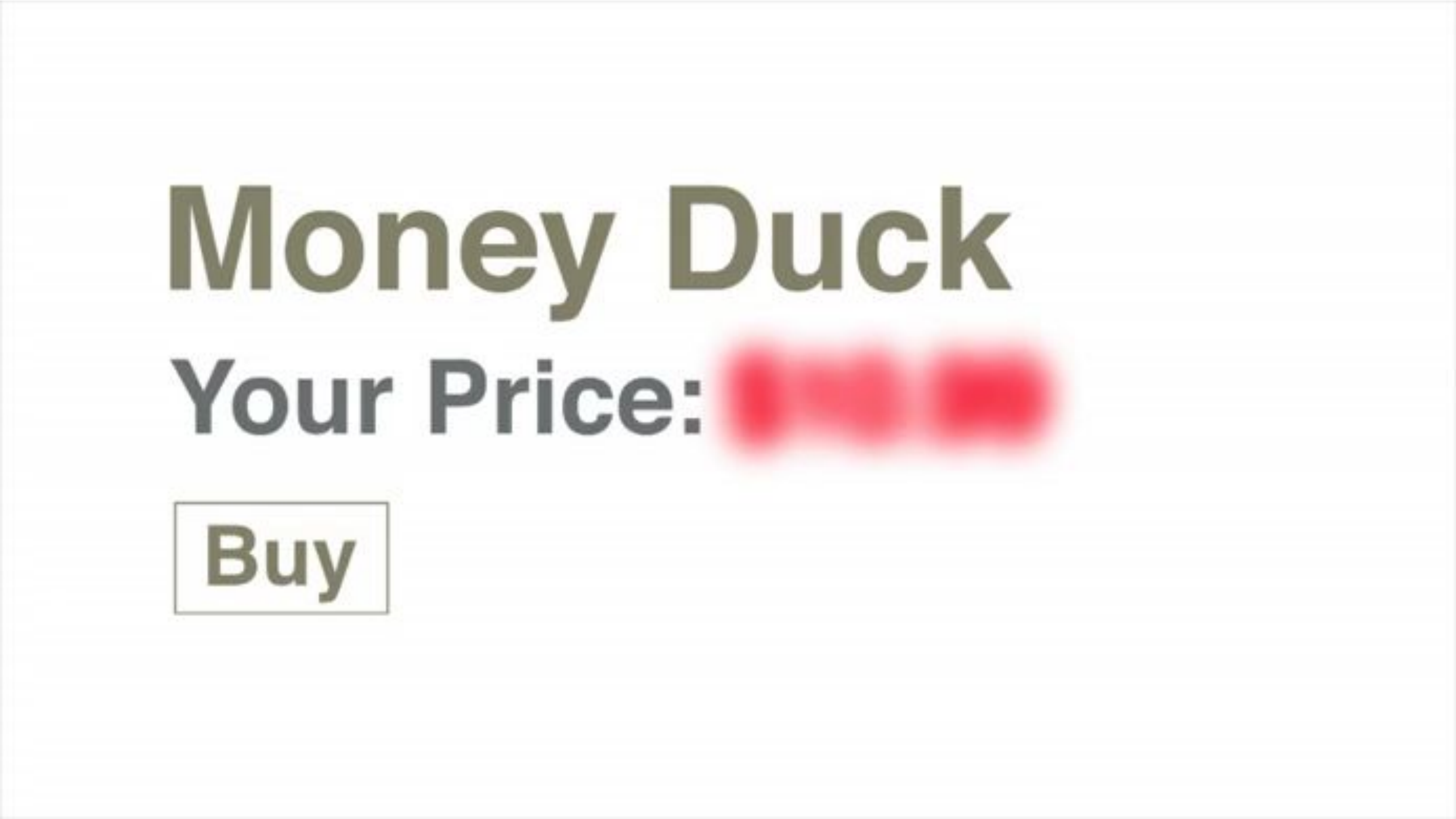


Right skew - Points bend up and to the left of the line.

Left skew - Points bend down and to the right of the line.

Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.

Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

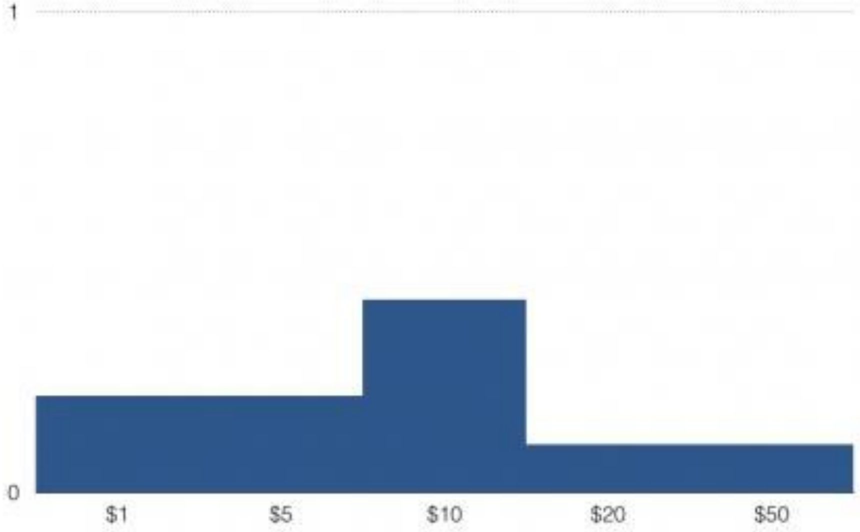


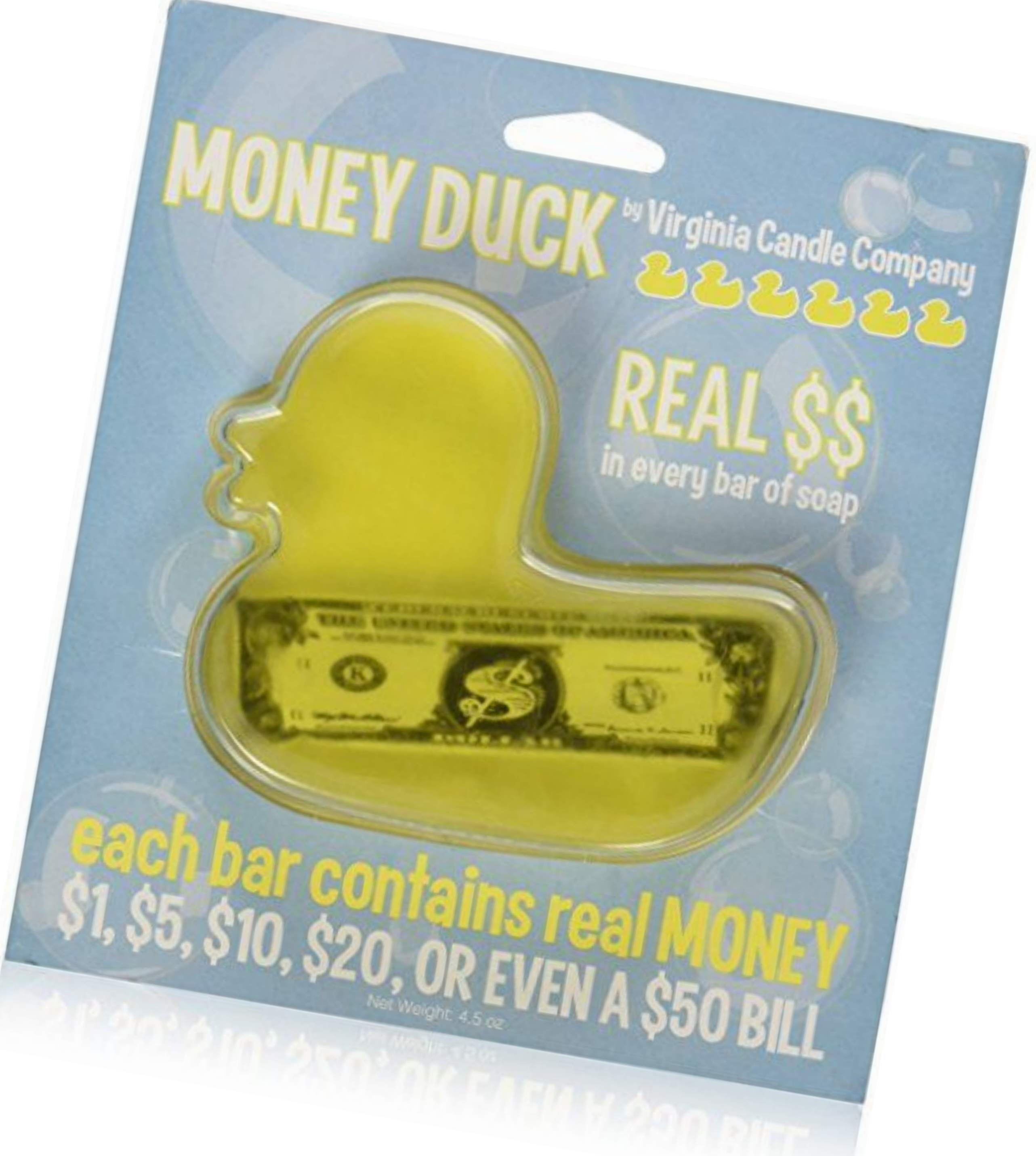
Money Duck

Your Price: [redacted]

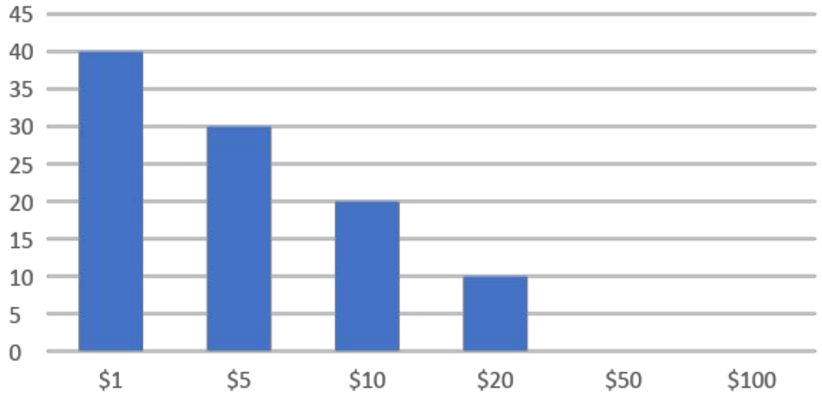
Money Duck

- How much would you pay?

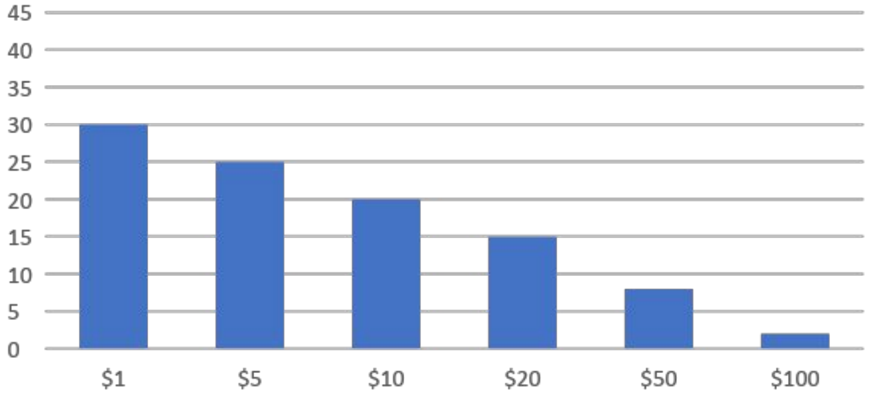



MONEY DUCK
by Virginia Candle Company
REAL $$
in every bar of soap
each bar contains real MONEY
$1, $5, $10, $20, OR EVEN A $50 BILL
Net Weight: 4.5 oz

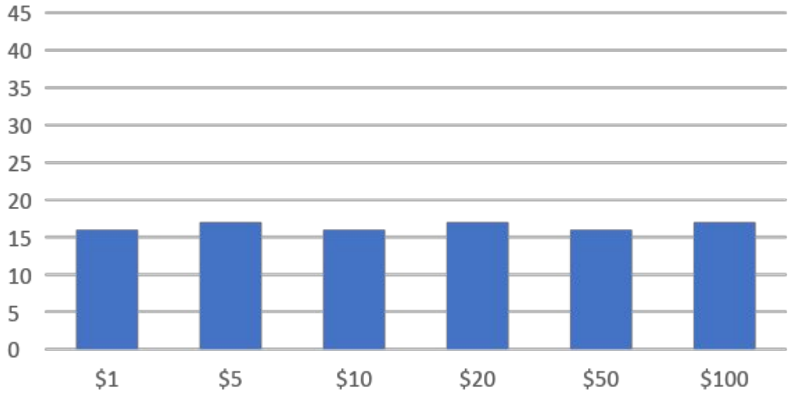
Which would you rather play?



A

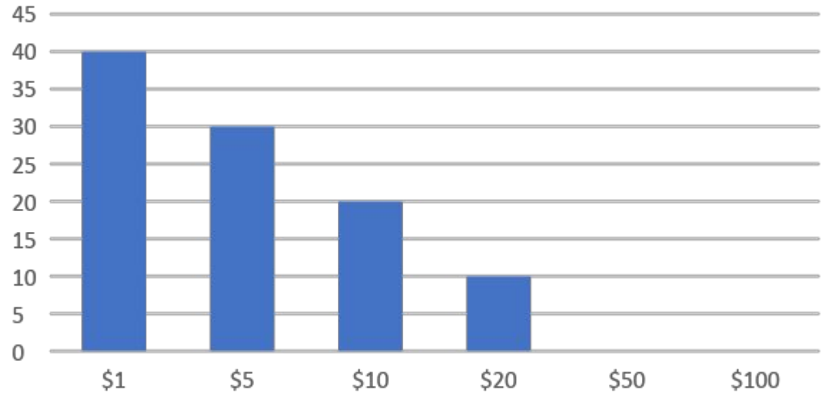


C



B

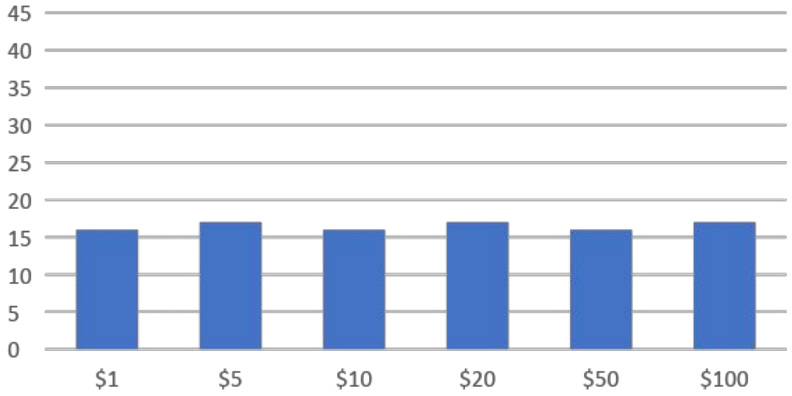
Which would you rather *pay* to play?



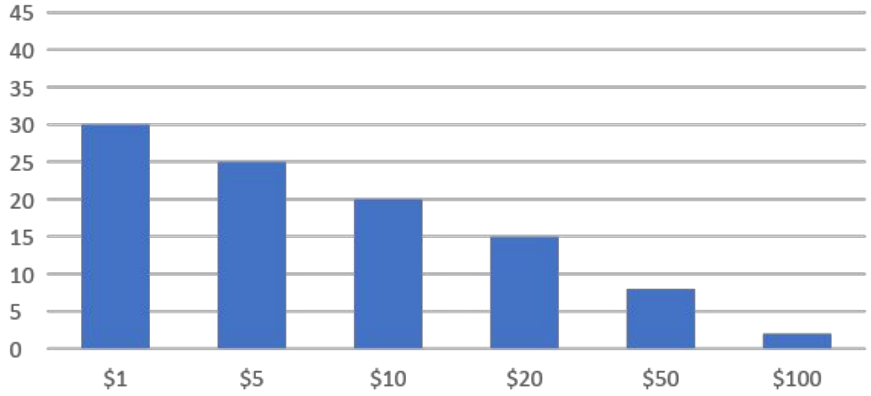
$6.00

A

$30.00



B



$13.00

C

Expected Value

- The expected value of some event is the average of the values you get if you repeat the event many, many times.

- Expected Values are very similar to means (averages) in a lot of situations

Expected Value

Calculated slightly differently depending on the situation.

- Example: The average weight of an orange is 75g. If I randomly pick an orange, what's the expected value of it's weight?
  - E(weight of orange) = 75 g (the weight of an average orange)
- Example: Roll a die. What's the expected value of the roll?
  - E(Dice Roll) = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5 (average value of the faces)
- Example: Money Duck
  - E(Money Duck) = 0.2 x $1 + 0.2 x $5 + 0.4 x $10 + 0.1 x $20 + 0.1 x $50 = $12.2

Expected value and its variability: Normal Distribution

In the normal distribution the *expected value* is the **mean**.

$$E(x) = \mu$$

The *variation* we expect in the value is the **standard deviation**

$$Var(x) = \sigma$$

Expected value: Binomial Distribution

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, 100 x 0.262 = 26.2.
- Or more formally, $\mu = np$ = 100 x 0.262 = 26.2.
- But this doesn't mean in every random sample of 100 people exactly 26.2 will be obese. In fact, that's not even possible. In some samples this value will be less, and in others more. How much would we expect this value to vary?

Expected value and its variability: Binomial Distribution

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

Going back to the obesity rate:

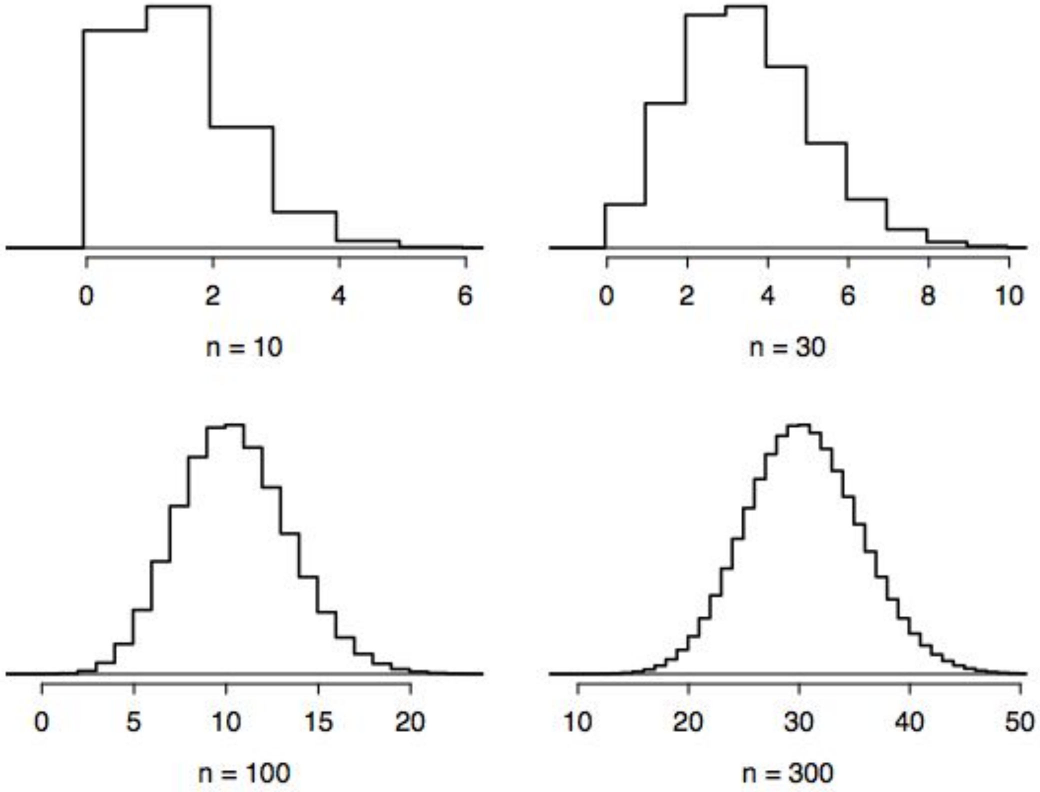
$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

Note: Mean and standard deviation of a binomial might not always be whole numbers, and that is alright, these values represent what we would expect to see on average.

Approximating Binomial with Normal

Distributions of number of successes



Hollow histograms of samples from the binomial model where p = 0.10 and n = 10, 30, 100, and 300. What happens as n increases?

An analysis of Facebook users

A recent study found that ``Facebook users get more than they give". For example:

1. 40% of Facebook users in our sample made a friend request, but 63% received at least one request
2. Users in our sample pressed the like button next to friends' content an average of 14 times, but had their content ``liked" an average of 20 times
3. Users sent 9 personal messages, but received 12
4. 12% of users tagged a friend in a photo, but 35% were themselves tagged in a photo

Source: Power users contribute much more content than the typical user.
http://www.pewinternet.org/Reports/2012/Facebook-users/Summary.aspx

Example

This study also found that approximately 25% of Facebook users are considered ***power users***. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that **$n = 245$, $p = 0.25$**, and we are asked for the probability **$P(K \geq 70)$**. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$$P(K \geq 70) = P(K = 70) + P(K = 71) + P(K = 72) + \dots + P(K = 245)$$

Using R

We can calculate this directly using R

The probability of exactly 70:

```
> dbinom(70,size=245,p=0.25))

[1] 0.02509227
```

The probability of at least 70:

```
> sum(dbinom(70:245,size=245,p=0.25))

[1] 0.112763

> pbinom(69,245,.25,lower.tail=FALSE)
[1] 0.112763
```

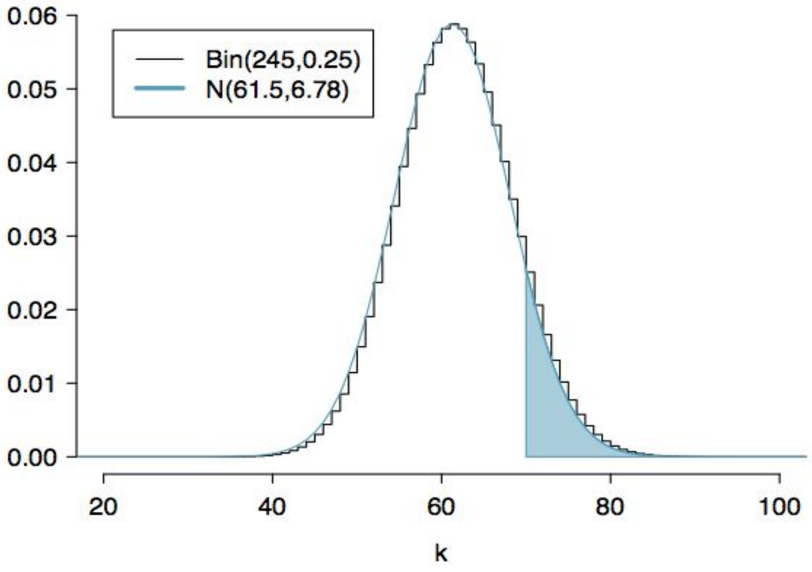
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

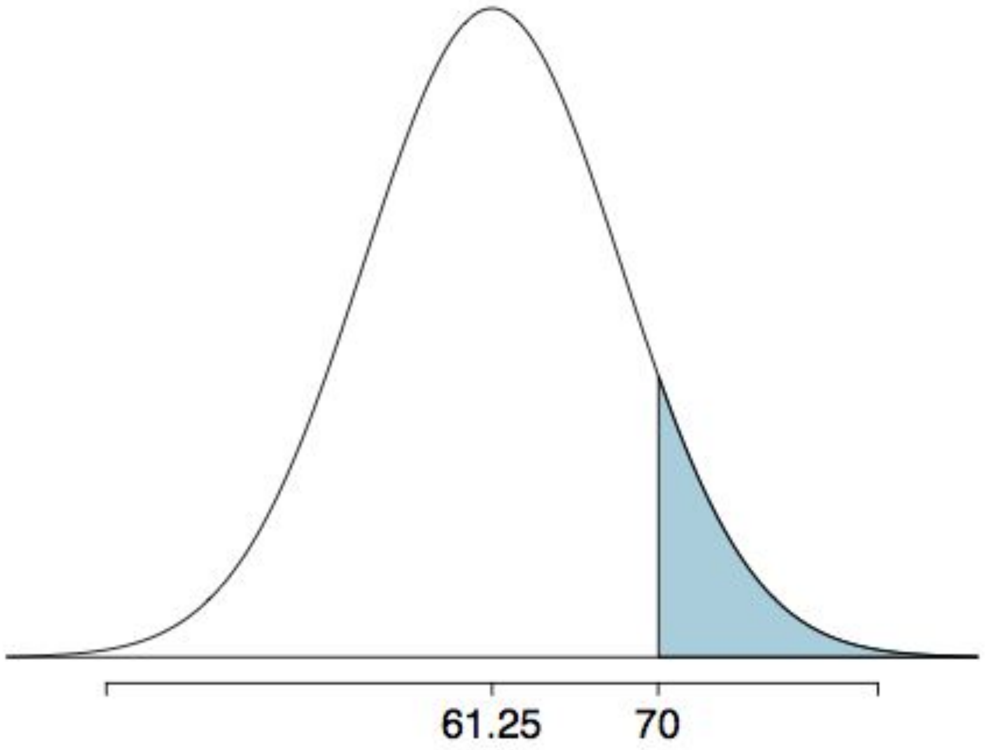
- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

Using R:

```
    pnorm(1.29, lower.tail = FALSE)
Or  1 - pnorm(1.29)
Or  pnorm(70, 61.25, 6.78, lower.tail = FALSE)
Or  1 - pnorm(70, 61.25, 6.78)
```

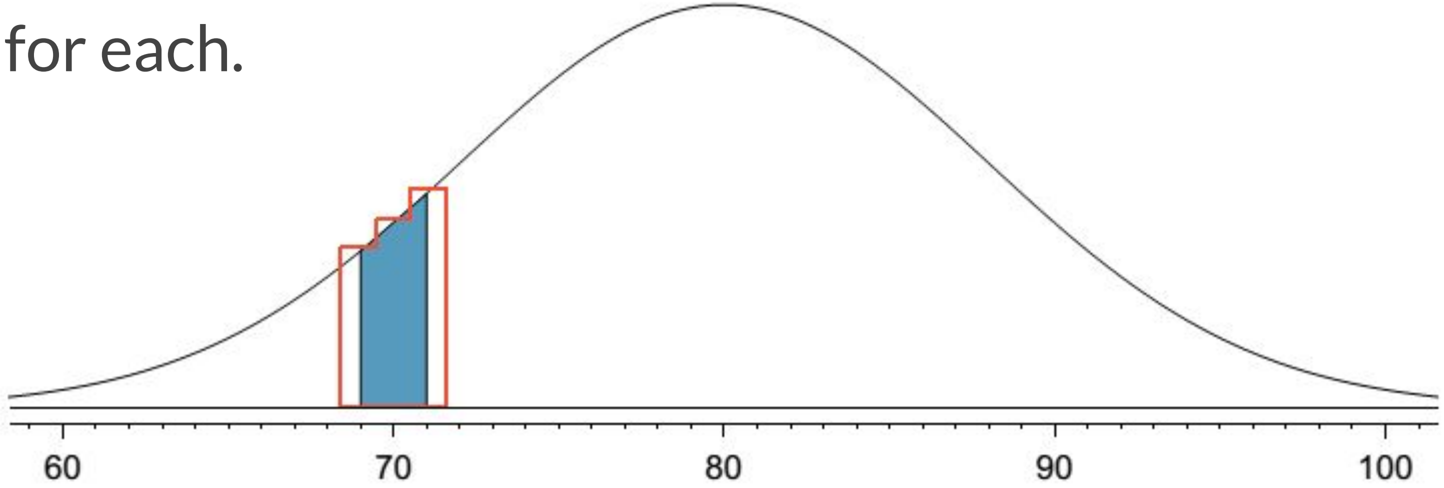
Each of them = 0.0985

$$P(X > 70) = P(Z > 1.29) = 0.0985$$

9.8% vs. 11% is not that great.

We can improve the accuracy by making an adjustment:

- The normal distribution is continuous but binomial trials are discrete.
- X ≥ 70 for a *Normal Distribution* includes everything even slightly larger that 70 (e.g. 70.1, 70.01, etc) but for the binomial distribution it only includes 70,71, 72, 73, etc.
- Compare X > 69 and X ≥ 70 for each.



Adjusting the Normal Approximation

To adjust we lower the lower bound by 0.5 and raise the upper bound by 0.5.

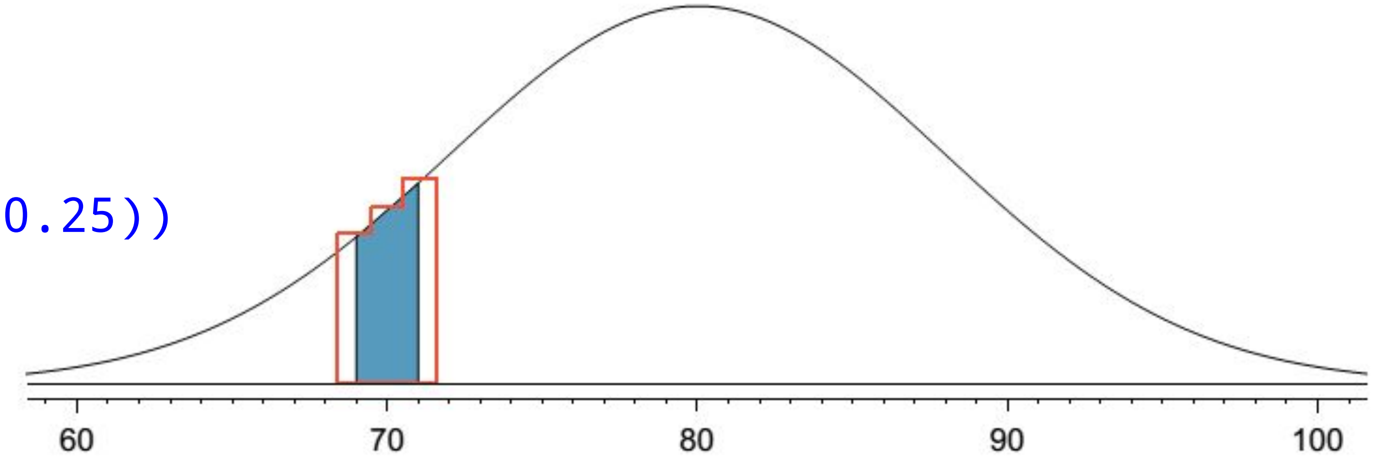
Instead of `pnorm(70, 61.25, 6.78, lower.tail = FALSE) = 0.0985`

we use `pnorm(69.5, 61.25, 6.78, lower.tail = FALSE) = 0.1118`

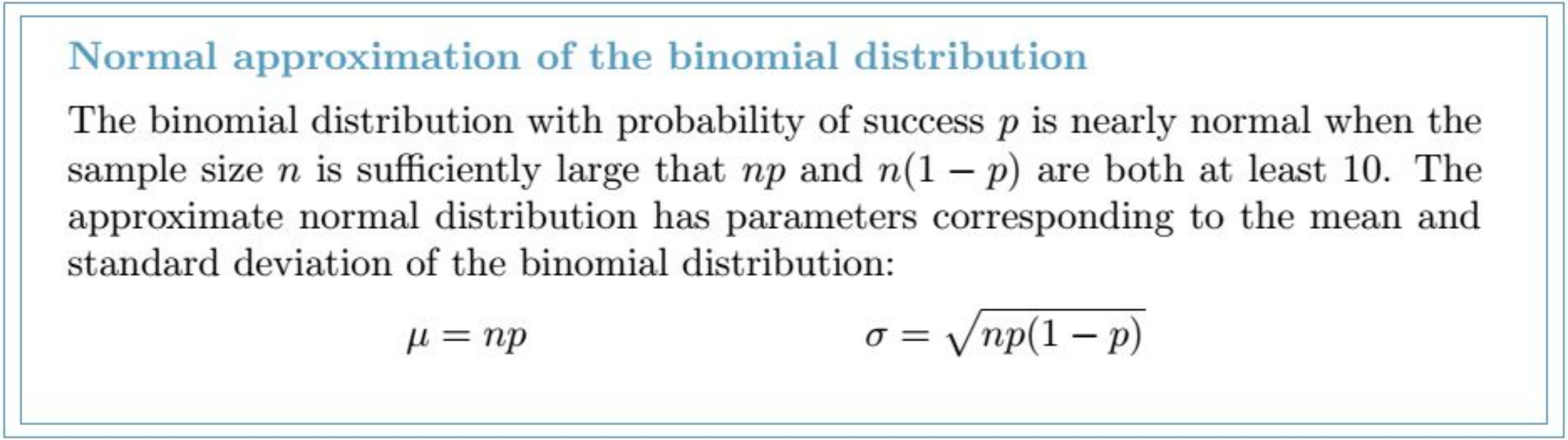
Compare this to the value of

```
> sum(dbinom(70:245,size=245,p=0.25))

[1] 0.112763
```



Normal Approximation of Binomial - Condition



Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad\qquad \sigma = \sqrt{np(1-p)}$$

i.e. $np \geq 10$ *and* $n(1\text{-}p) \geq 10$

Example

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

A. n = 100, p = 0.95

B. n = 25, p = 0.45

C. n = 150, p = 0.05

D. n = 500, p = 0.015

Example

Below are four pairs of Binomial distribution parameters. Which distribution can be approximated by the normal distribution?

A. $n = 100, p = 0.95$

B. $n = 25, p = 0.45 \rightarrow 25 \times 0.45 = 11.25, 25 \times 0.55 = 13.75$

C. $n = 150, p = 0.05$

D. $n = 500, p = 0.015$

Question

What is the minimum n required for a binomial distribution with p=0.3 to follow a normal distribution?

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing exactly 75 smokers in a sample of 400 people?

```
> dbinom(75,size=400,p=0.2)
[1] 0.04185711
```

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing 70 or fewer smokers in a sample of 400 people?

```
> sum(dbinom(0:70,size=400,p=0.2))
[1] 0.1163917
```

Example

If the true proportion of smokers in a community is 20%, what is the probability of observing 70 or fewer smokers in a sample of 400 people?

Conditions: $np = 400 \times 0.20 = 80$ $\qquad n(1-p) = 400 \times 0.8 = 320$

With these conditions checked, we may use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model:

$$\mu = np = 80 \qquad \sigma = \sqrt{np(1-p)} = 8$$

Use the normal model $N(\mu = 80, \sigma = 8)$ to estimate the probability of observing 70 or fewer smokers.

Approximating a Binomial with a Normal Dist.

- Even with the adjustment it’s still not a great approximation.

- Technology allows us to make the calculations directly.

- We can approximate a Binomial distribution with a Normal Distribution, but we shouldn’t.

Hypothesis Testing

Swimming with Dolphins

- Does swimming with Dolphins help depression?
    - Antonioli & Reveley, British Medical Journal
- 30 participants. 13 showed improvement
- How many of the improvers were in the dolphin group?

Swimming with Dolphins

- What are the possible outcomes?
- Null hypothesis, H_0, is the hypothesis that things *aren't different*, or *haven't changed*

H_0=Swimming with dolphins doesn't improve depression
H_a=Swimming with dolphins does improve depression

Confidence Intervals
Or
Compatibility Intervals

Confidence intervals

A plausible range of values for the population parameter is called a confidence interval.

Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

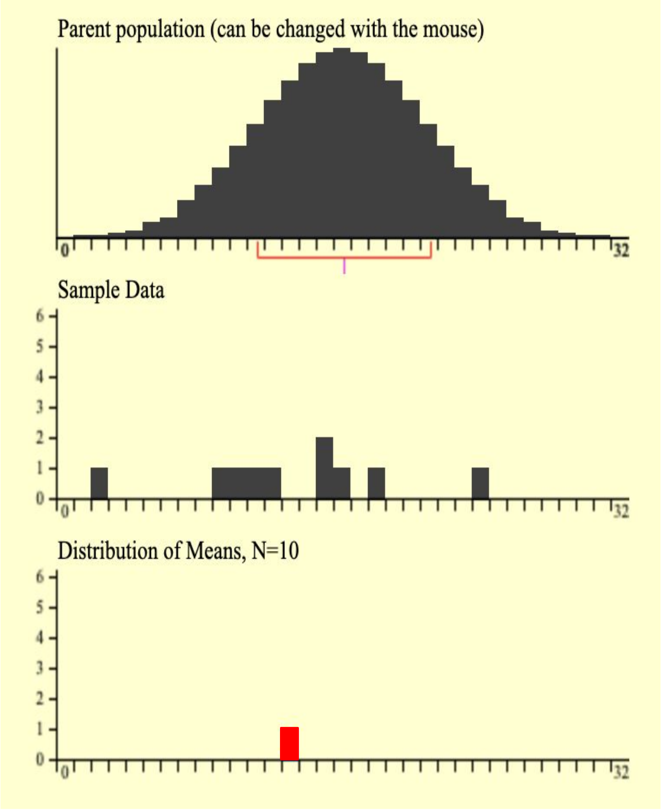
We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

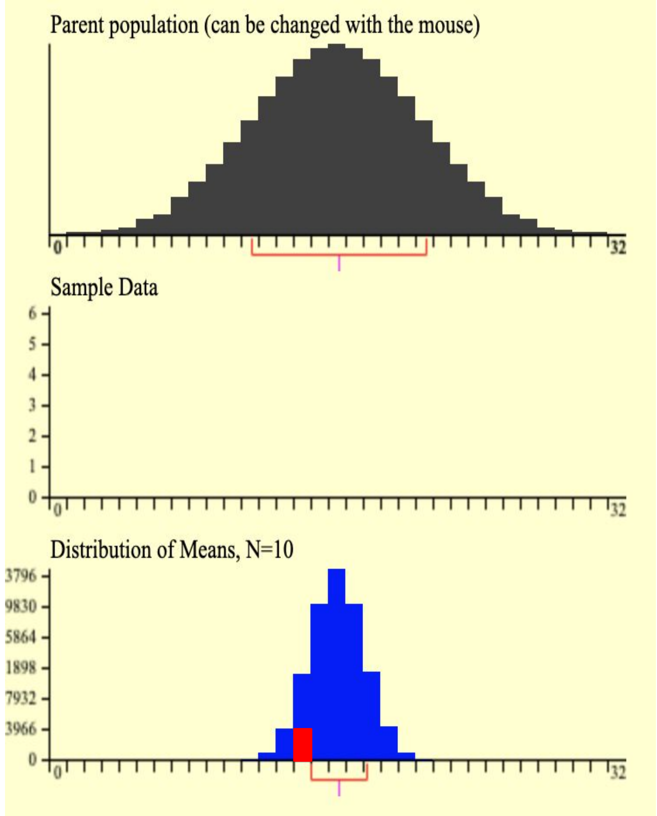
Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462)
and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Confidence Intervals

Reporting Just A Point Estimate


Parent population (can be changed with the mouse)
0
32
Sample Data
6
5
4
3
2
1
0
0
32
Distribution of Means, N=10
6
5
4
3
2
1
0
0
32

Reporting a Confidence Interval


Parent population (can be changed with the mouse)
0
32
Sample Data
6
5
4
3
2
1
0
0
32
Distribution of Means, N=10
3796
9830
5864
1898
7932
3966
0
0
32

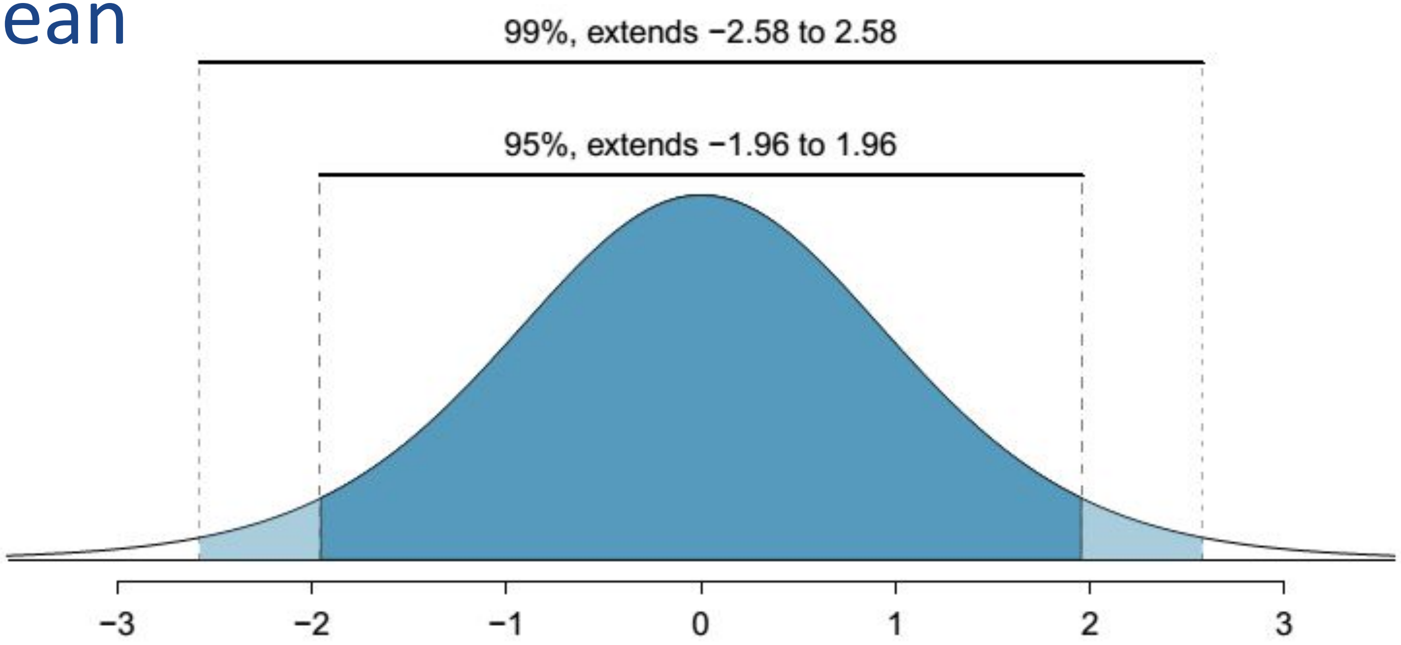
Confidence Intervals

CLT says that

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

Approximate 95% CI = x ± 2SE

95% of sample will have a mean in this range.



Question

One of the earliest examples of behavioral asymmetry is a preference in humans for turning the head to the right, rather than to the left, during the final weeks of gestation and for the first 6 months after birth. This is thought to influence subsequent development of perceptual and motor preferences. A study of 124 couples found that 64.5% turned their heads to the right when kissing. The standard error associated with this estimate is roughly 4%. Which of the below is *false*?

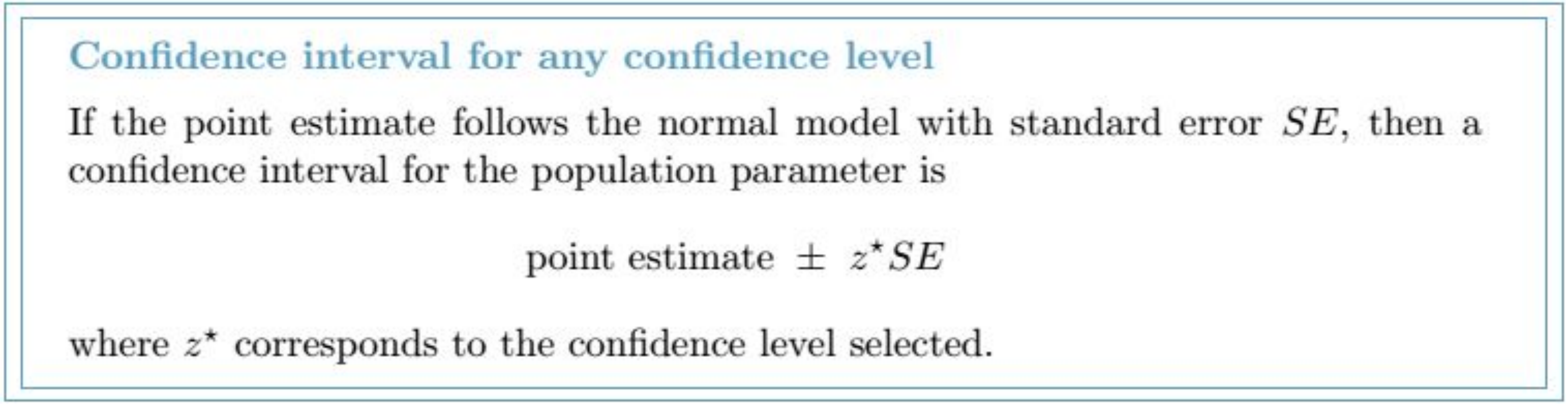
a) The 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 4%.

b) A higher sample size would yield a lower standard error.

c) The margin of error for a 95% confidence interval for the percentage of kissers who turn their heads to the right is roughly 8%.

d) The 99.7% confidence interval for the percentage of kissers who turn their heads to the right is roughly 64.5% ± 12%.

Güntürkün, O (2003) Adult persistence of head-turning asymmetry. *Nature*. Vol 421.

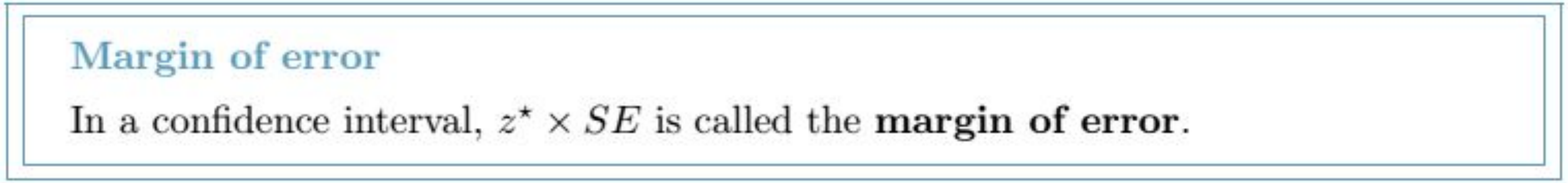


Confidence interval for any confidence level

If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z^{\star}SE$$

where $z^{\star}$ corresponds to the confidence level selected.



Margin of error

In a confidence interval, $z^{\star} \times SE$ is called the **margin of error**.

Conditions:

1. **Independence**: Sampled observations must be independent.
   - Random sample/assignment
   - If sampling without replacement, then n < 10% of the population
2. **Sample size and skew**: n > 30, larger if the population distribution is skewed.

Changing the confidence level

point estimate ± z x SE*

- In a confidence interval, *z* x SE* is called the margin of error, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust z* in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, z* = 1.96.
- However, using the standard normal (z) distribution, it is possible to find the appropriate z* for any confidence level.

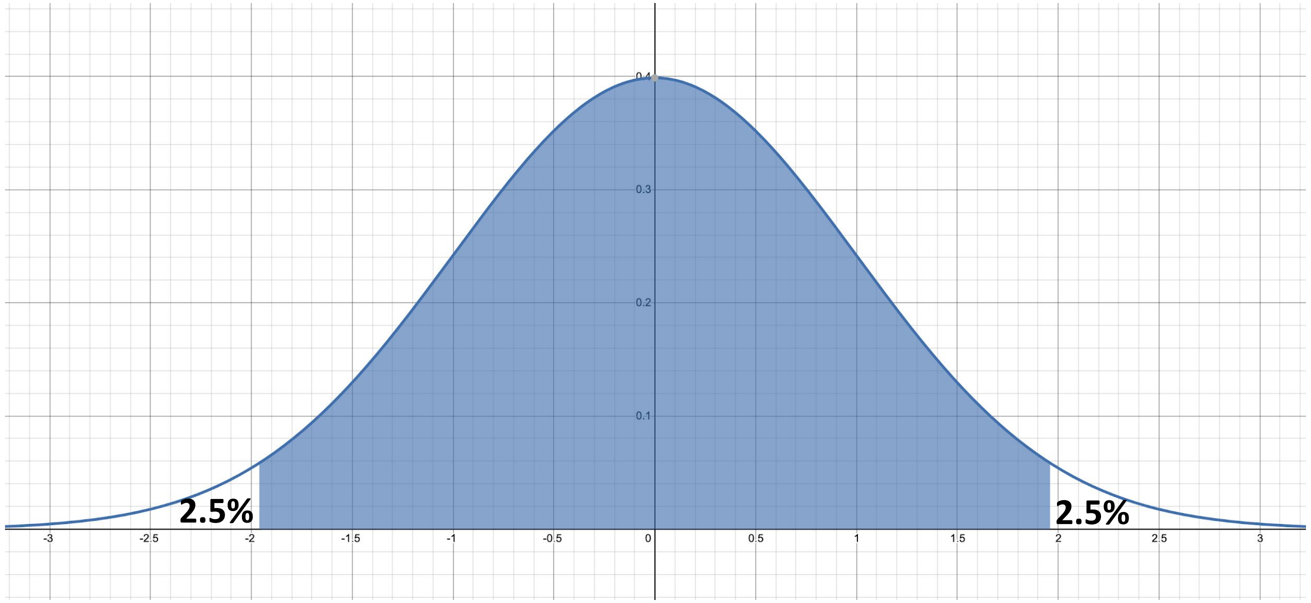
Finding the Critical Value

$$\text{point estimate} \pm z^{\star} SE$$

- For 95% CI, the sum of two tails is 5%
    - One tail is 2.5%

```
> qnorm(0.025) [1]
-1.959964
```

Z* = 1.96



Question

Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65

Question

Which of the below Z scores is the appropriate z* when calculating a 98% confidence interval?

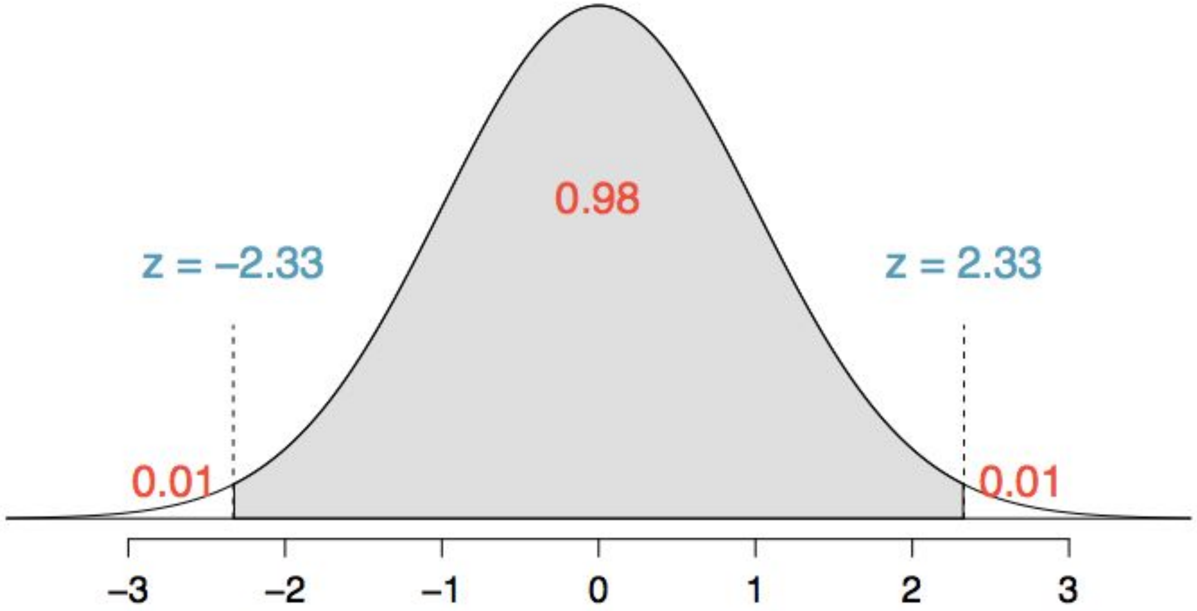
(a) Z = 2.05

(b) Z = 1.96

(c) Z = 2.33

(d) Z = -2.33

(e) Z = -1.65



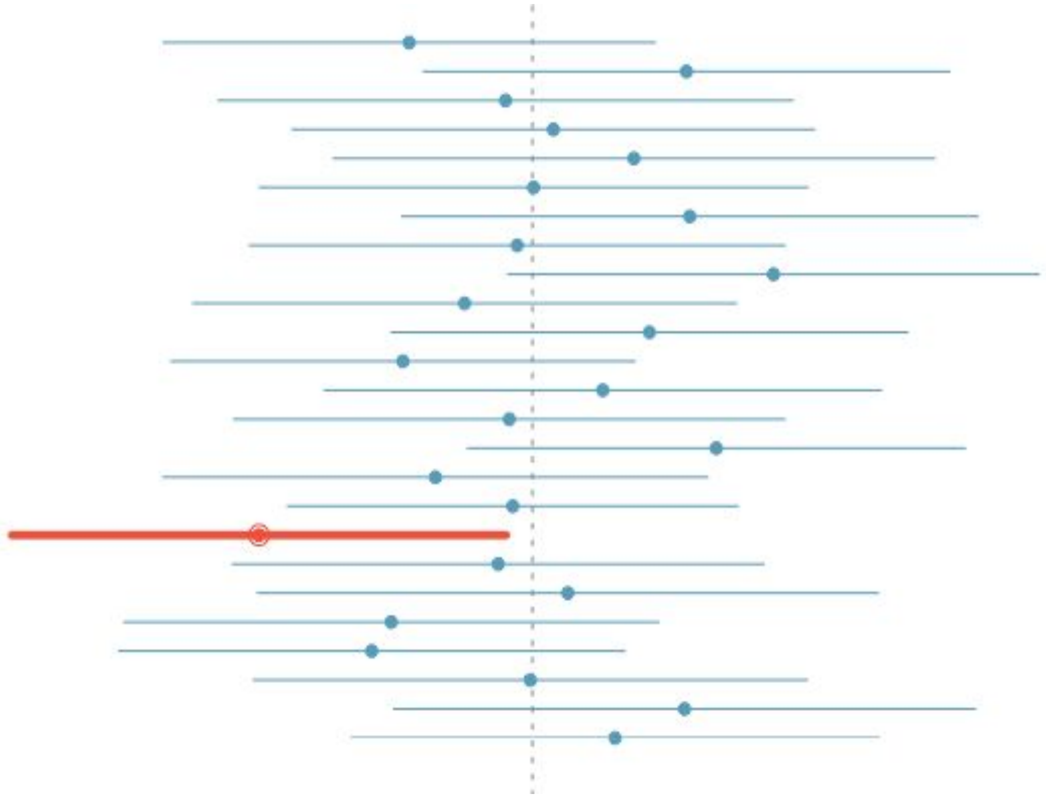
Accuracy vs Precision of CI

What does 95% confident mean?

Suppose we took many samples and built a confidence interval from each sample using the equation *point estimate ± 2 SE*.

Then about 95% of those intervals would contain the true population mean (μ).

The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



What does a 95% Compatibility Window Mean?

- The parameters in the Window are compatible with the data at a 95% confidence level.
- It’s reasonable to expect data like ours given values within the compatibility window.
- Values for the true parameter in this window are consistent with our data.

How to Interpret Confidence Intervals

<u>Correct:</u> We are XX% confident that the true population parameter lies within (our interval)

Things to remember to include:

<u>Our confidence level</u>: ex. 99%, 95%, etc

<u>What the population parameter we are estimating</u>: ex. True average height of US men, true proportion of smokers in MN, etc.

<u>Lower and upper limits of our interval</u>: Ex. Lies between 5.4 and 6.6, lies betweend 0.4 and 0.9, etc.

<u>Incorrect</u>: Our confidence interval captures the true population parameter with a probability of 0.95

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Can you see any drawbacks to using a wider interval?

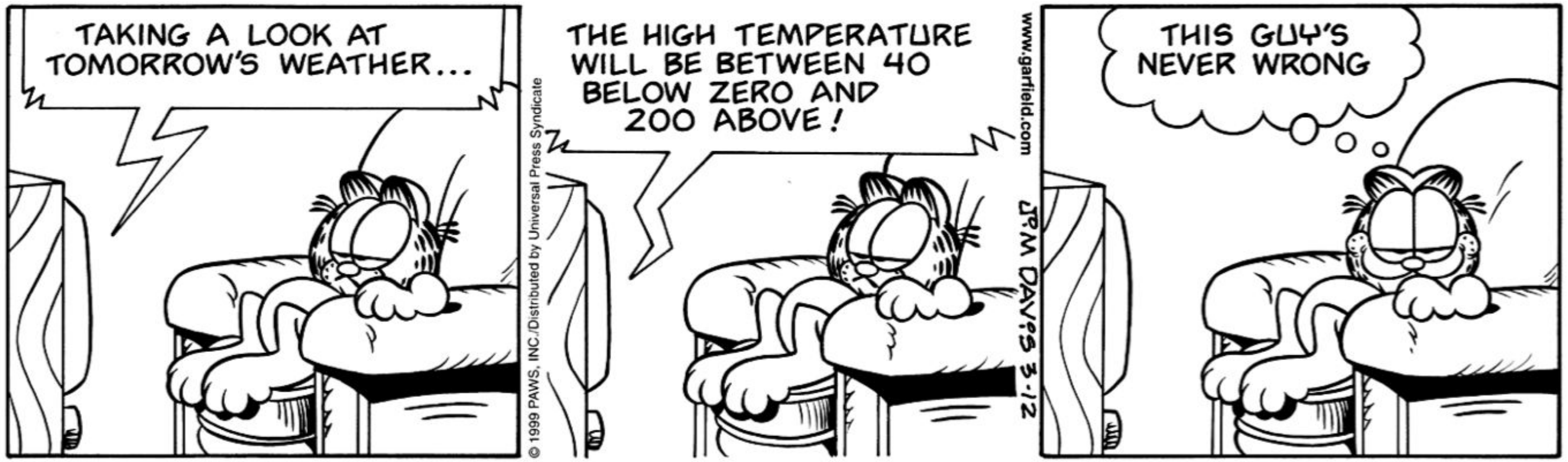
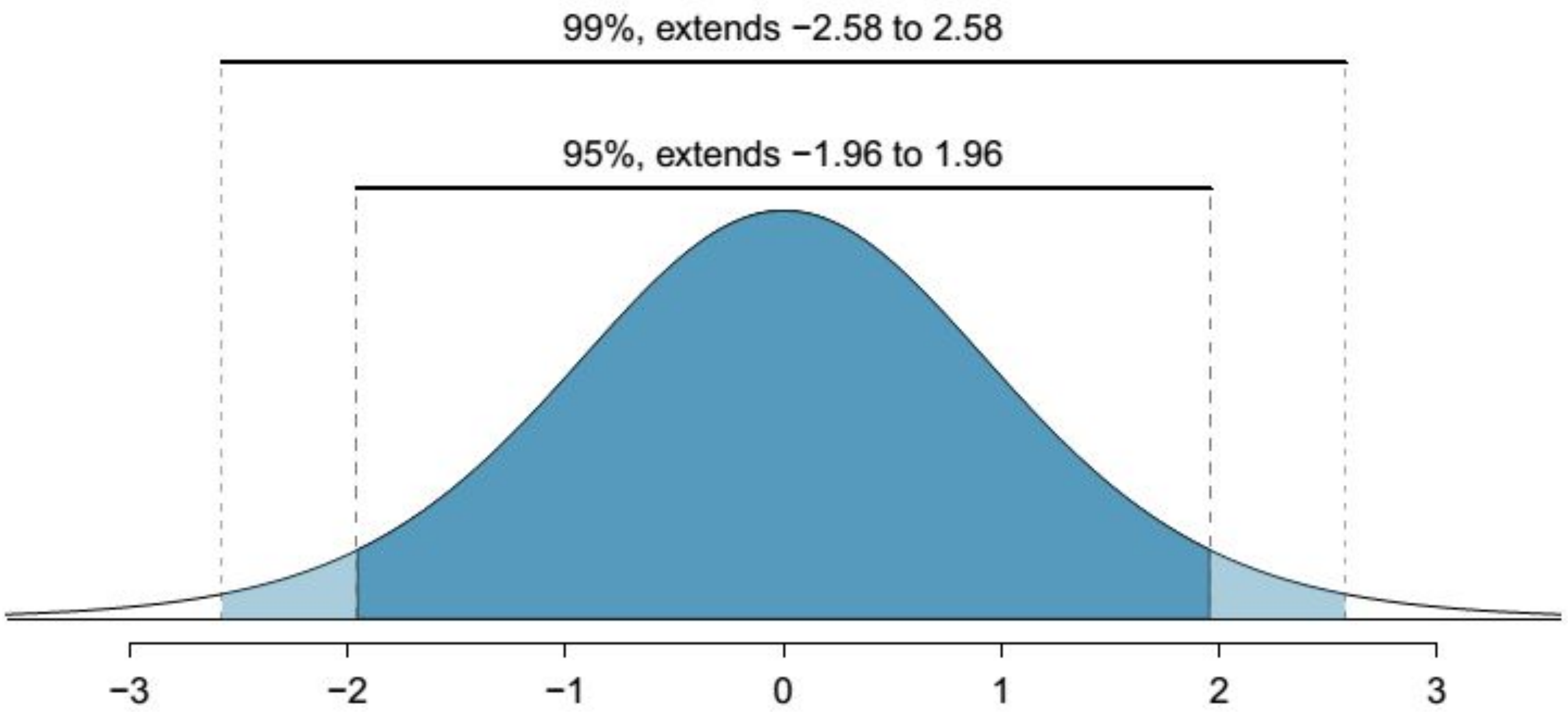
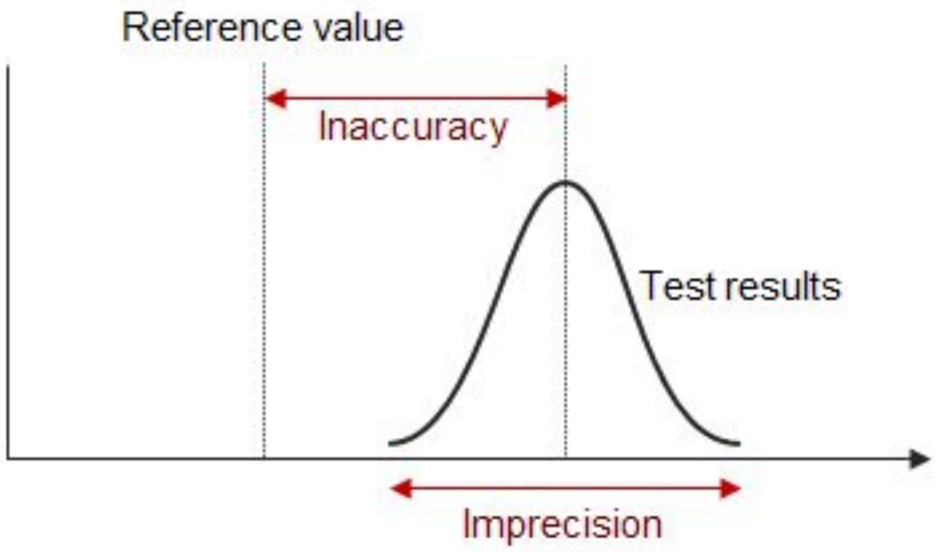

TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC./Distributed by Universal Press Syndicate
www.garfield.com
JPM DAVIS 3-12

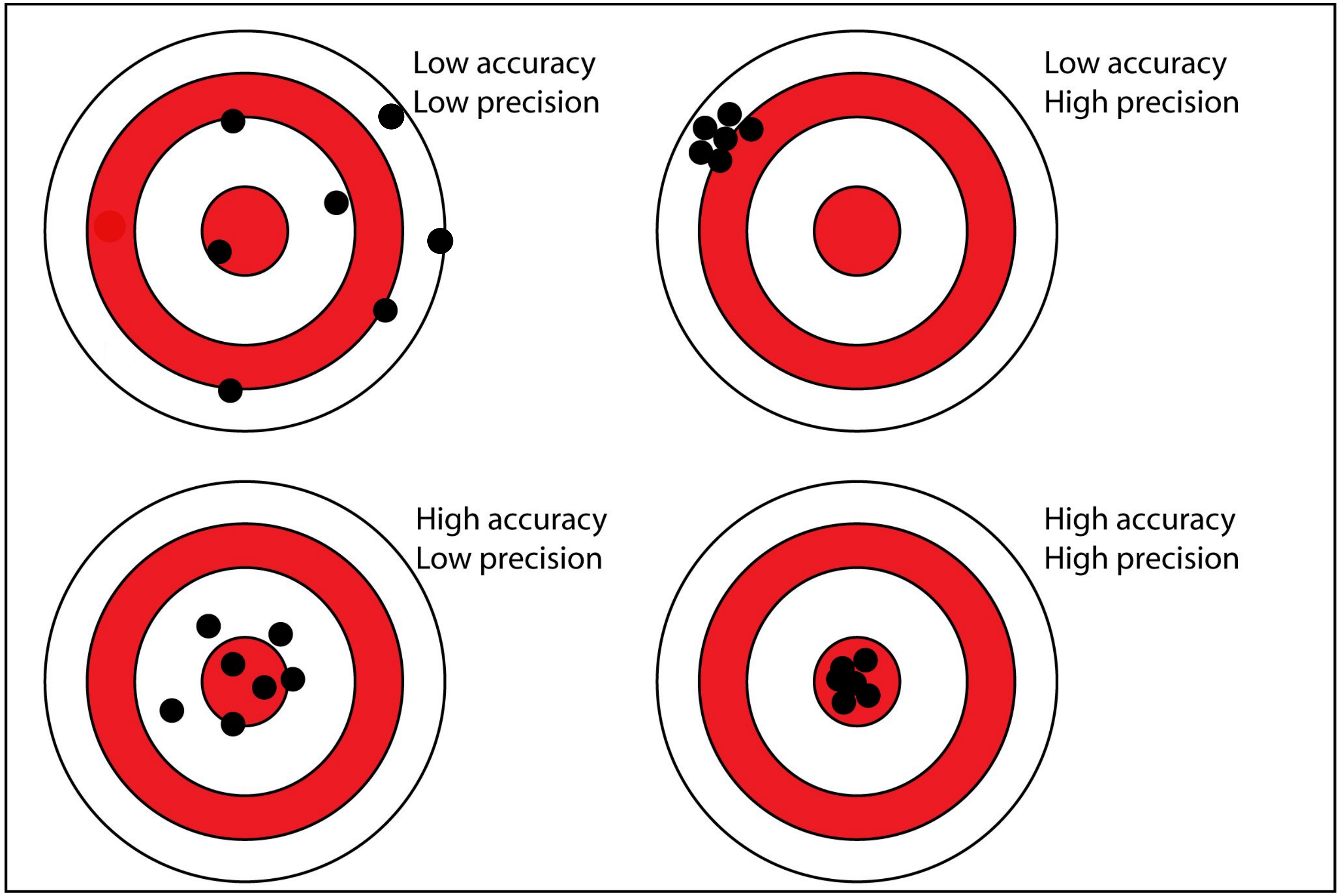
Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif



As Confidence Level goes up, the width goes up

- As Confidence Level goes up:
    - Width Increases
    - Accuracy Increases
    - Precision Decreases





Question

The mean of the sampling distribution is 5.75, and the standard deviation of the sampling distribution (also called the standard error) is 0.75. Which of the following is the most reasonable guess for the 95% confidence interval for the true average number of basketball games attended by students?

a) 5.75 ± 0.75

b) 5.75 ± 2 x 0.75

c) 5.75 ± 3 x 0.75

d) cannot tell from the information given

Sample Size

Slides adapted from work by Mine Çetinkaya-Rundel of OpenIntro

Finding a sample size for a certain margin of error

A group of researchers wants to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this particular medication during pregnancy. Previous studies suggest that the standard deviation of IQ scores of three-year-old children is 18 points. How many such children should the researchers sample in order to obtain a 96% confidence interval with a margin of error less than or equal to 4 points?

We know that the critical value associated with the 96% confidence level:

z* = 2.05.

$$4 \geq 2.05 * 18/\sqrt{n} \rightarrow n \geq (2.05 * 18/4)^2 = 85.1$$

The minimum number of children required to attain the desired margin of error is 85.1. Since we can't sample 0.1 of a child, we must sample at least 86 children (round up, since rounding down to 85 would yield a slightly larger margin of error than desired).

Example

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

Conditions:

- Random Sample and 50 < 10% of all College Students.
- We can assume that the number of exclusive relationships one student in the sample has been in, is independent of another. So we have independent observations.
- Sample size is greater than 30, and the distribution of the sample is not so skewed. We can assume, that the sampling distribution of average number of exclusive relationships from samples of size 50 will be nearly normal.

$$\bar{x} = 3.2 \qquad s = 1.74$$

95% confidence interval is defined as

point estimate ± 1.96 SE

$$SE = s / \sqrt{n} = 1.74 / \sqrt{50} \approx 0.246$$

$$\bar{x} \pm 1.96\ SE \longrightarrow 3.2 \pm 1.96 \text{ x } 0.246$$

$$\longrightarrow 3.2 \pm 0.48$$

$$\longrightarrow (2.72, 3.68)$$

We are 95% confident that college students on average have been in 2.72 to 3.68 exclusive relationships.

Question

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.