

Statistics is the science of data, and is used to evaluate claims.

Ex: I make 80% of free throws I shoot.

Chapter 1: Picturing Distributions with Graphs

Def: An individual is an object described by data.

Ex: Person, city, animal, company.

Def: A variable is a characteristic of an individual.

Ex: Age, population, species, profit.

Ex: We randomly select 4 people in the US and ask them to report their age and gender. We also ask them what state they're living in.

State	Age	Reported Gender
Kentucky	61	Female
Florida	27	Female
Wisconsin	27	Male
California	33	Female

catagorical → *quantitative* → ↗ *catagorical*
4 individuals and 3 variables measured for
each individual

Def: A variable is quantitative if it
takes numerical values and arithmetic

makes sense.

Def: A variable is categorical if it is not quantitative.

Now we ask for zip codes

State	Age	Reported Gender	Zip
Kentucky	61	Female	41375
Florida	27	Female	93402
Wisconsin	27	Male	97403
California	33	Female	49102

categorical!

Ex: A study classifies bison in Yellowstone as young or adult. State the

individuals, variables, and the type of variable.

Bison, age, categorical

Def: The distribution of a variable is the information of both its possible values and how often they occur.

Student ID	Hair color
003	Red
005	Brown
035	Brown
089	Black

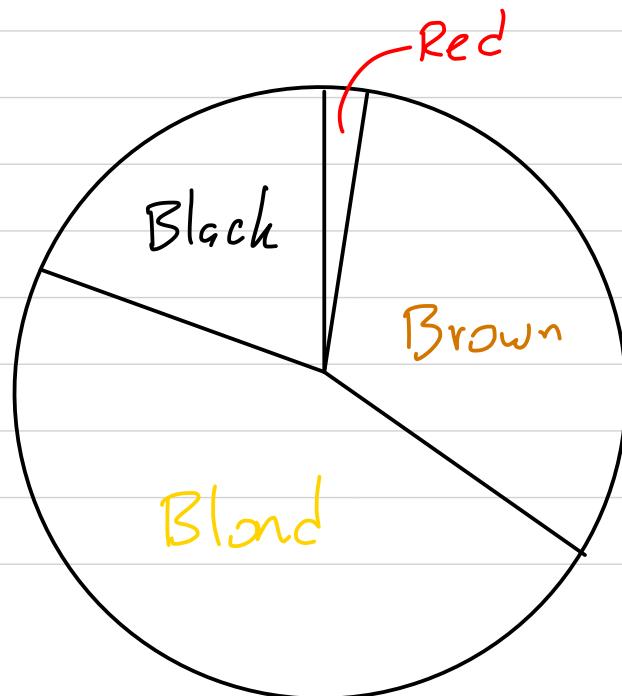
not a
distribution

Hair color	% of students w/ this color
Red	2 %
Brown	35 %
Blond	43 %
Black	20 %

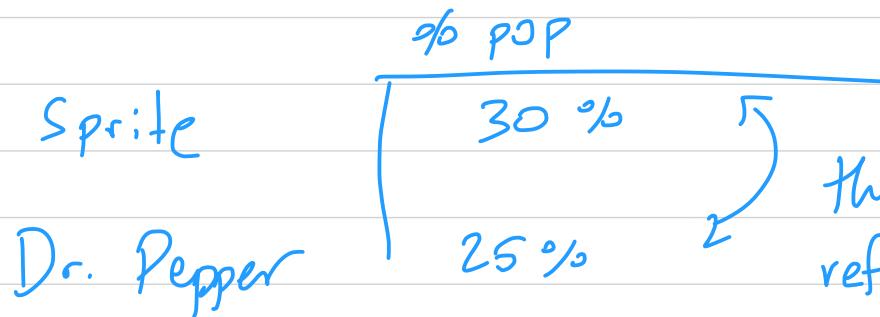
distribution



Pie chart



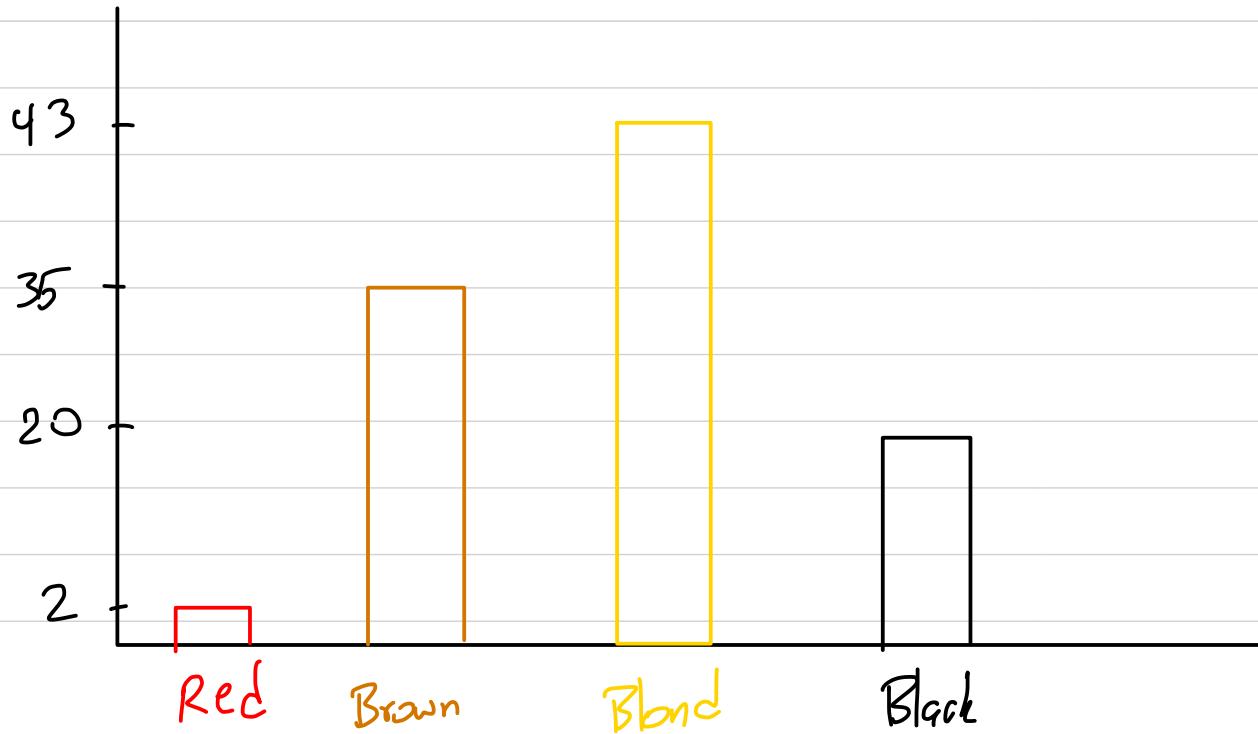
Comment: Only use pie charts when the values the variable can take are mutually exclusive — i.e. every individual has at most one value. Hair color is mutually exclusive since you can have at most one. A survey asking which types of soda you'd had in the past month would not be mutually exclusive since you could have had more than one type.



this doesn't reflect the people who have had both

Hair color	% of students w/ this color
Red	2 %
Brown	35 %
Blond	43 %
Black	20 %

Bar graph:



<u>Ex</u>	Music source	% of 12-24 year olds who have used it
	Radio	72
	YouTube	77
	iTunes	47

Don't use a pie chart, because the different music sources aren't mutually exclusive!

Histograms: when given a sample of individuals, you can make a histogram by dividing the data into ranges (called classes) and counting the number of individuals in each class. Then we make a bar graph of the result. This

roughly approximates the distribution.

Ex: We get a set of ages:

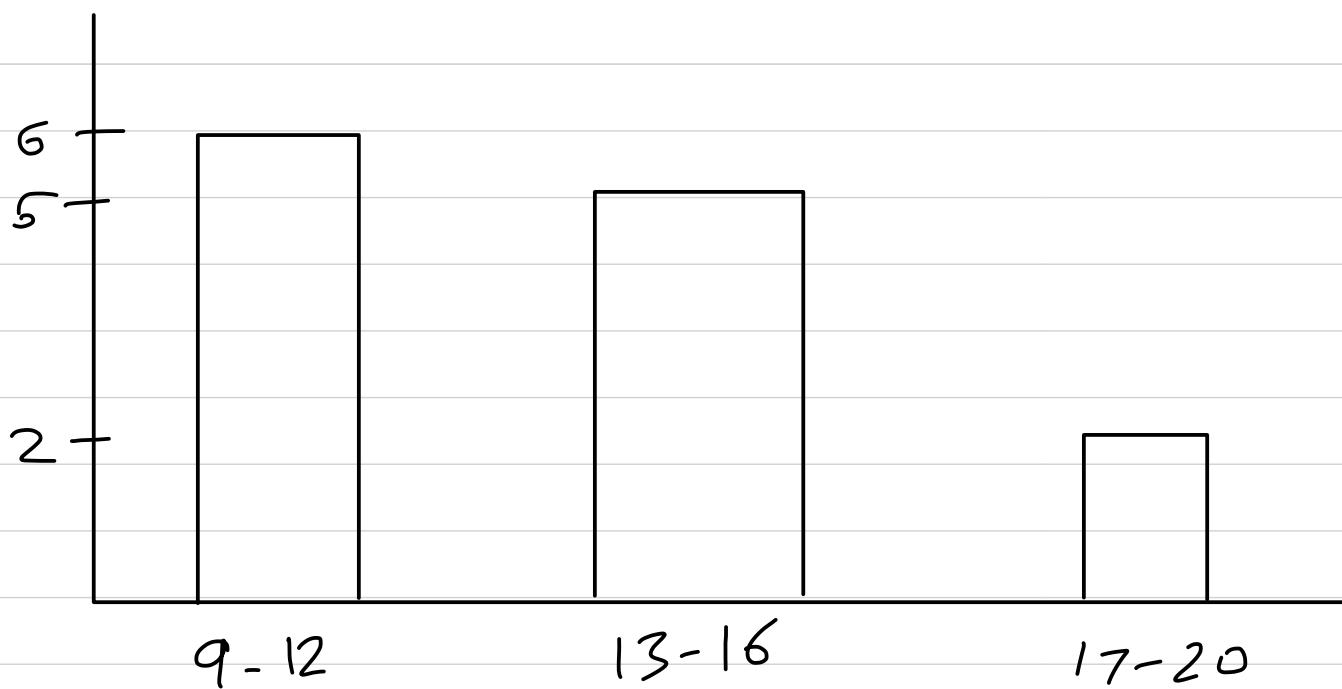
9 10 10 11 12 12 14 15 15

16 16 18 20

classes: 9-12, 13-16, 17-20

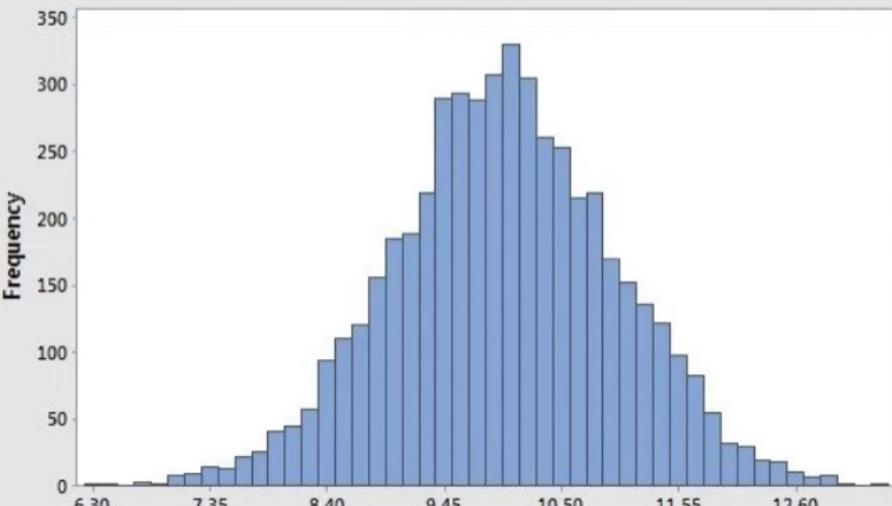
~~~~~

6 5 2

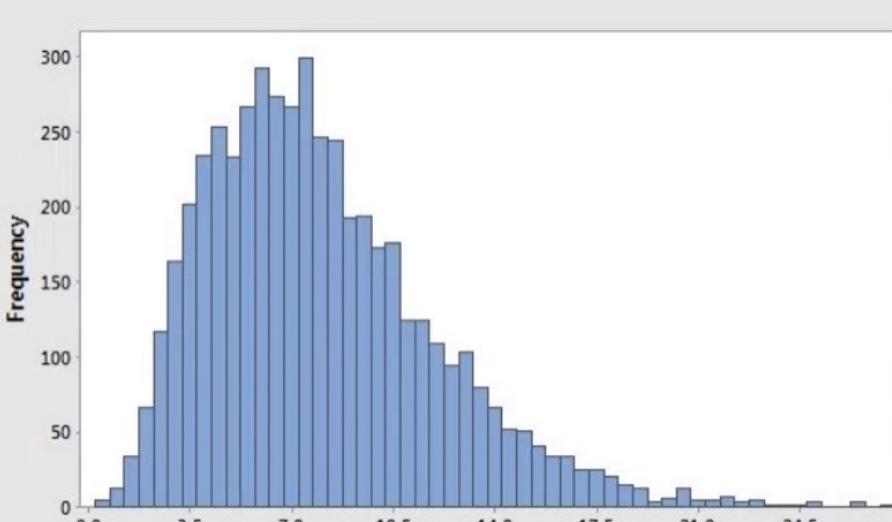


A **symmetric** distribution.

Ex: Heights of young women, Lengths of bird bills



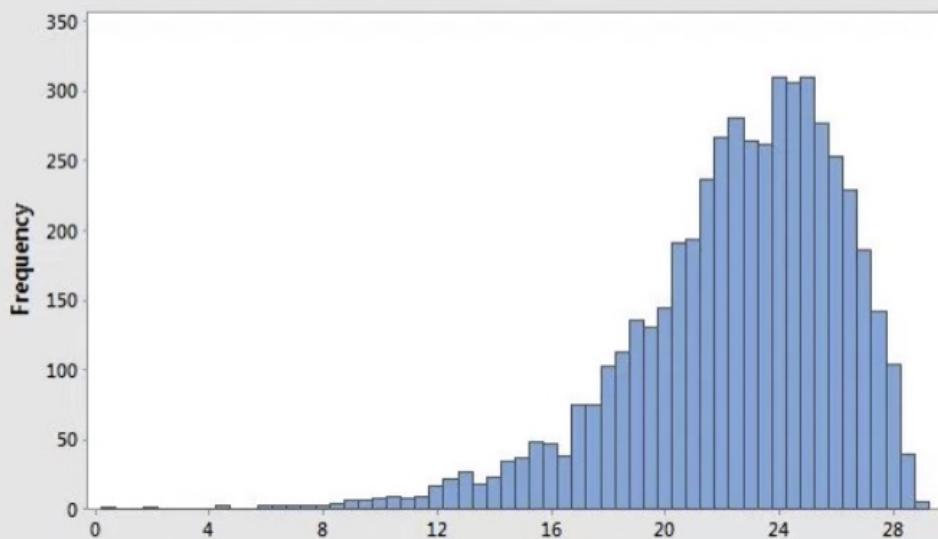
Right-skewed



Ex: incomes

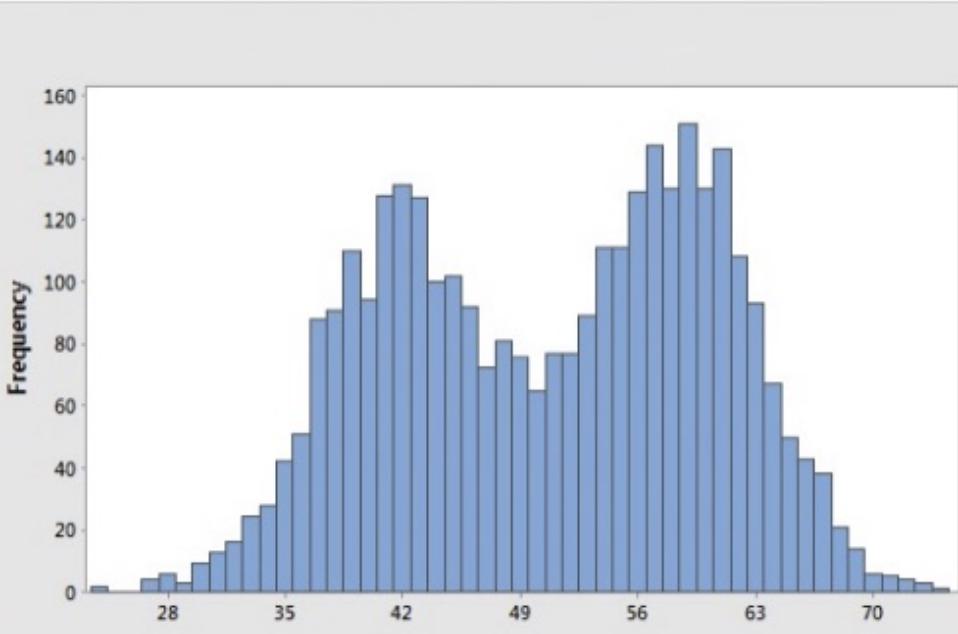
**A left-skewed distribution.**

Ex: Grades on an easy test



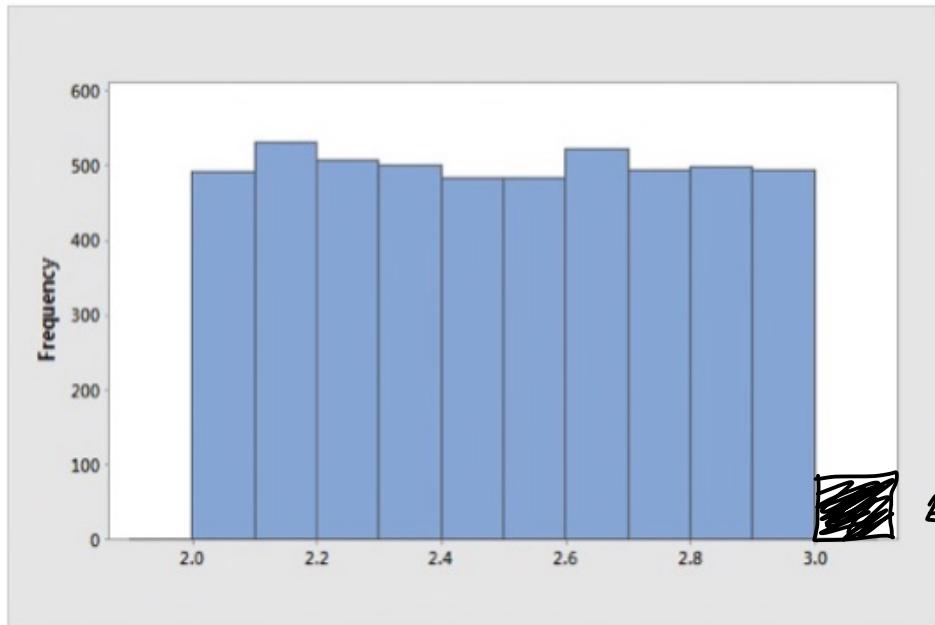
**A bimodal distribution.**

Ex: Exam scores when one group studied and another didn't



An approximately uniform distribution.

Ex: Rolling a die



Def: The center of the distribution is the mean or median. The variability is roughly how spread out the distribution is. Outliers are individuals who don't fit the pattern.

Def: Given a set of data, we can form a stem-and-leaf plot : take all of the numbers and split them into the last digit and all the other digits. Then write the second piece (i.e. the prefix) and all the final digits with that prefix.

Ex: 9 10 10 11 12 12 14 15 15  
16 16 18 20

|   |  |             |
|---|--|-------------|
| 0 |  | 9           |
| 1 |  | 00122455668 |
| 2 |  | 0           |

Ex :  $\boxed{5, 13, 18, 32, 91}$        $\boxed{\underline{40, 45, 19, 60}}$

|   |    |              |
|---|----|--------------|
| 0 | 5  |              |
| 1 | 38 | $\leftarrow$ |
| 3 | 2  |              |
| 9 | 1  |              |

Webwork + Textbook :

|    |   |    |
|----|---|----|
| 9  | 0 | 5  |
| 9  | 1 | 38 |
| 9  | 2 |    |
| 9  | 3 | 2  |
| 05 | 4 |    |
| 0  | 5 |    |
| 0  | 6 |    |
| 0  | 7 |    |
| 0  | 8 |    |
| 0  | 9 | 1  |

Comment: we can also split the stems:

|   |  |                       |
|---|--|-----------------------|
| 0 |  | 9                     |
| 1 |  | 0 0 1 2 2 4 5 5 6 6 8 |
| 2 |  | 0                     |

||

|   |  |             |
|---|--|-------------|
| 0 |  | 9           |
| 0 |  |             |
| 1 |  | 0 0 1 2 2   |
| 1 |  | 4 5 5 6 6 8 |
| 2 |  | 0           |
| 2 |  |             |



## Chapter 2 : Describing Distributions with Numbers

Ex: A list of travel fines to work  
in North Carolina:

30, 20, 10, 40, 25, 20, 10, 60,  
15, 40, 5, 30, 12, 10, 10

How to calculate center? One way  
is taking the average.

Def Given a set of data  $x_1, \dots, x_n$ ,  
the mean of the data is

$$\bar{x} = \frac{x_1 + \dots + x_n}{n}.$$

Ex:  $\bar{x} = \frac{30 + 20 + 10 + \dots + 12 + 10 + 10}{15} = 22.5$

$\nwarrow$  15 samples

Ex: 5, 10, 15, 200 ← Right-skewed

$$\bar{x} = \frac{5 + 10 + 15 + 200}{4} = \frac{230}{4} = 57.5$$

Comment: In a skewed distribution, the mean is drawn toward the skew (i.e. the tail). We say the mean is not a resistant measure of center.

Def: Let  $x_1, \dots, x_n$  be a set of data.

The median is  $M$ , defined by:

① if  $n$  is odd, then  $M$  is the data point such that as many  $x_i$  are greater than  $M$  as are less than  $M$

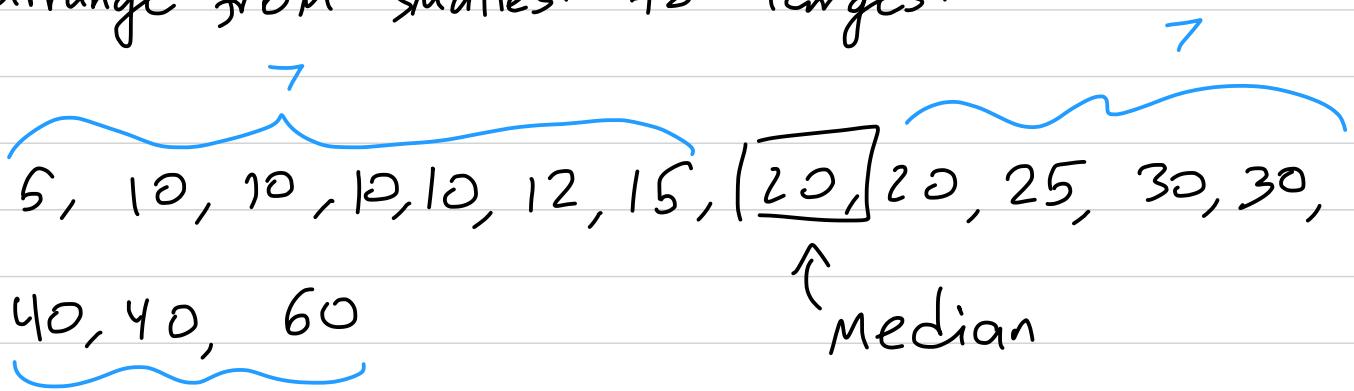
② if  $n$  is even,  $M$  is the average of the two numbers with

as many  $x_i$  greater than them

as there are  $x_i$  less than them

Ex: 30, 20, 10, 40, 25, 20, 10, 60,  
15, 40, 5, 30, 12, 10, 10

First arrange from smallest to largest



15 data points, which is odd, so we  
want the number "in the middle"

Ex: 5, 10, 15, 200

$$\text{average is } \frac{10+15}{2} = 12.5$$

median : 12.5

Comment: The median is a resistant measure of center.

Ex: you roll a die. If you roll a 1-5, you get nothing. If you roll a 6, you get \$100. What should you expect to get on average from rolling 6 times?

0 0 0 0 0  100

median: 0

mean:  $\frac{100}{6}$  ← this is better for our purposes!

How do we measure variability?

Start small: min and max

Ex · 5, 10, 10, 10, 10, 12, 15, 20, 20, 25, 30, 30,  
40, 40, 60

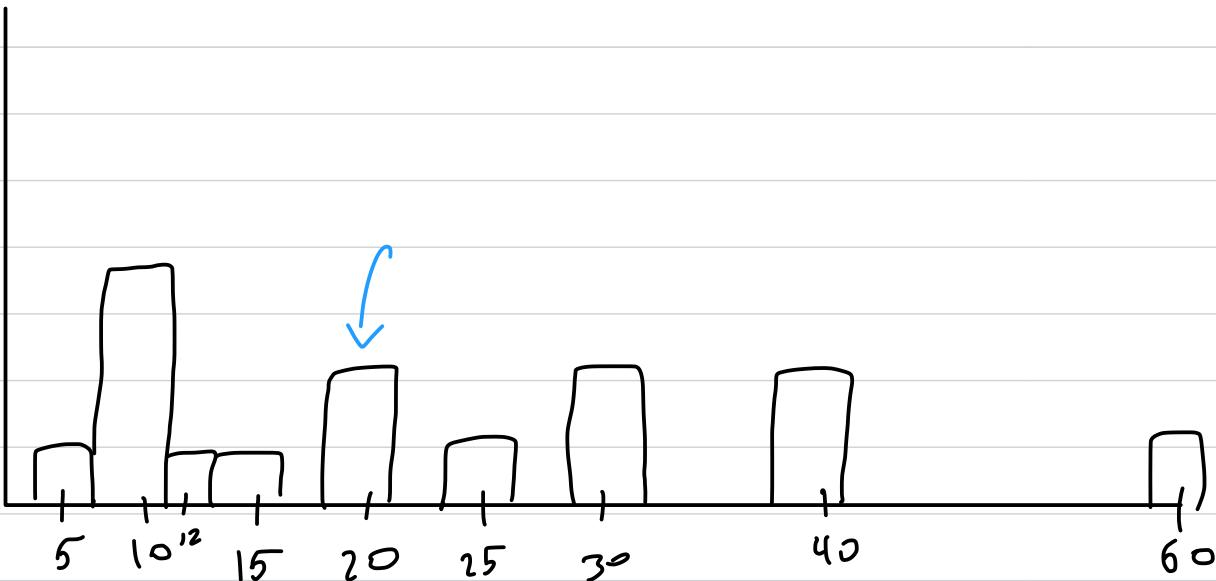
5, 60

Better: min, median, max

5, 20, 60



gap indicates that  
this is a right-skewed distribution



Def: The first and third quartiles,  $Q_1$ , and  $Q_3$ , are the medians of the two halves of the data, not including the median of the whole data.

|                              |                      |
|------------------------------|----------------------|
| $5, 10, 10, 10, 10, 12, 15,$ | $20, 20, 25, 30, 30$ |
| $40, 40, 160$                |                      |

$$Q_1 = 10$$

(you could say that  $Q_2 = 20$ )

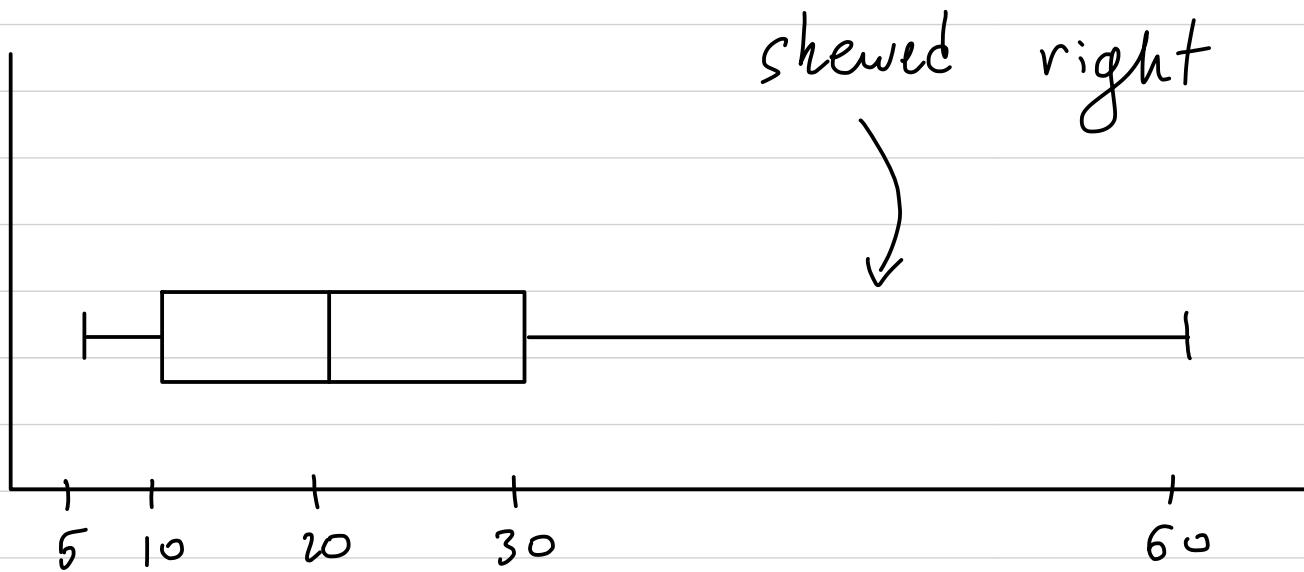
$$Q_3 = 30$$

Def: The 5-number summary of a set of data is min,  $Q_1$ , median,  $Q_3$ , max

Ex:  $5, 10, 20, 30, 60$

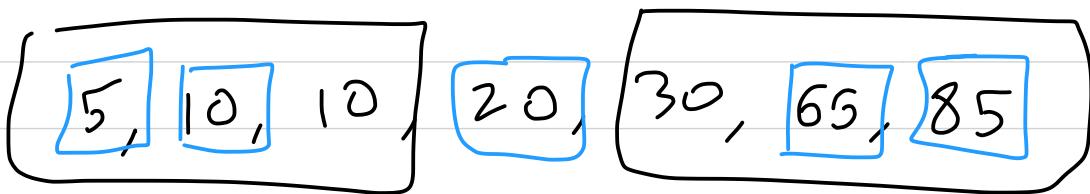
All 4 gaps have the same # of data points.

Box plots:



Ex: Draw a box plot of

10, 30, 5, 85, 65, 20, 10.





Def: The interquartile range, or IQR, is given by  $IQR = Q_3 - Q_1$ ,

Def: An outlier in a data set is any point more than  $1.5 \text{ IQR}$  above  $Q_3$  or below  $Q_1$ .

Ex: 10, 30, 5, 1000, 65, 20, 10.

5-num: 5, 10, 20, 65, 1000

$$Q_1 = 10$$

$$Q_3 = 65$$

$$\text{IQR} = 65 - 10 = 55$$

$$1.5 \text{ IQR} = 82.5$$

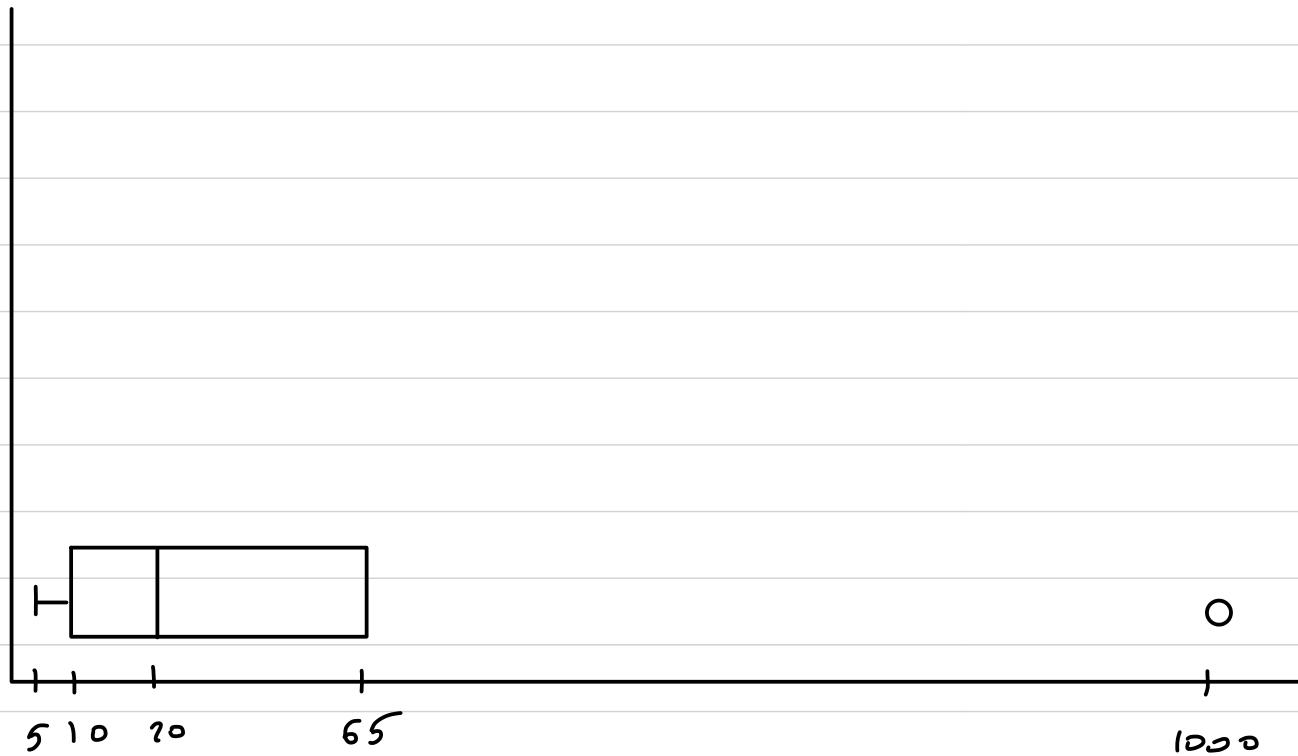
$$Q_3 + 82.5 = 147.25$$

$$Q_1 - 82.5 = -72.5$$

outliers are anything not between  
-72.5 and 147.25. So 1000 is an  
outlier.

Represent outliers by modifying the

box-plot : make the whiskers only reach the non-outliers.



The 5-number summary is a resistant measure of variability (but it's a little lacking)

How do we get a nonresistant measure of variability? Naive approach: take average distance to the mean

$x_1, \dots, x_n$  mean:  $\bar{x}$

$$\frac{(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x})}{n}$$

This actually always is zero!

$$= \frac{x_1 + x_2 + \dots + x_n - n \bar{x}}{n}$$

$$= \frac{\overbrace{x_1 + x_2 + \dots + x_n}^n - \frac{n \bar{x}}{n}}{n}$$

$\bar{x} - \bar{x}$

$$= 0$$

We can fix this issue by making the

distance to the mean always be positive:

Def: Let  $x_1, \dots, x_n$  be data with mean  $\bar{x}$ . The variance is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$



Bessel's correction

(eliminates bias)

don't worry about this yet

Def: The standard deviation is  $s$ .

Ex: SAT math scores at Georgia

# southern High school

490    580    450    570    650

$x_1$      $x_2$      $x_3$      $x_4$      $x_5$

$$\bar{x} = \frac{490 + 580 + 450 + 570 + 650}{5} = 548$$

$$(490 - 548)^2 + (580 - 548)^2 + (450 - 548)^2 +$$

$$(570 - 548)^2 + (650 - 548)^2$$

$$s^2 = \frac{(490 - 548)^2 + (580 - 548)^2 + (450 - 548)^2 + (570 - 548)^2 + (650 - 548)^2}{5 - 1}$$

$$= 6220$$

$s = 78.87$  ← standard deviation  
think of as: the average

distance to mean among this

data is 78.87

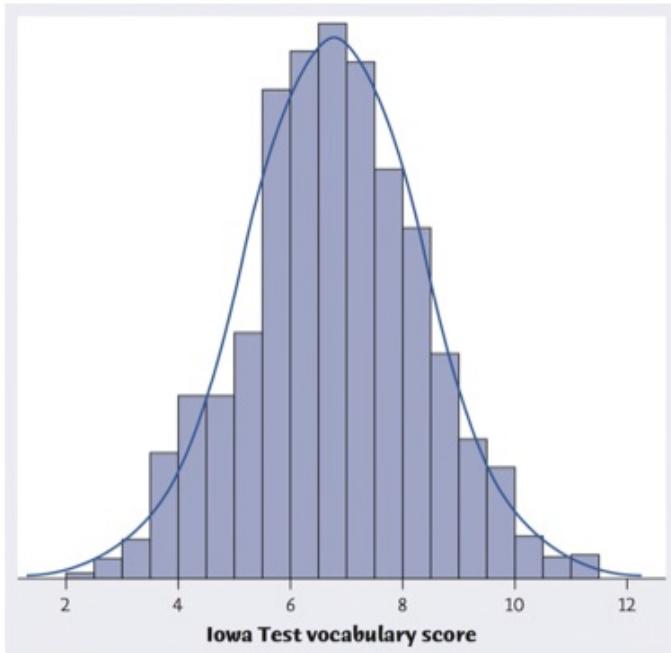
$s$  is a good measure of variability,  
but only when  $\bar{x}$  is a good measure of  
center.

Note:  $\bar{x}$  and  $s$  do not give a  
complete description of data

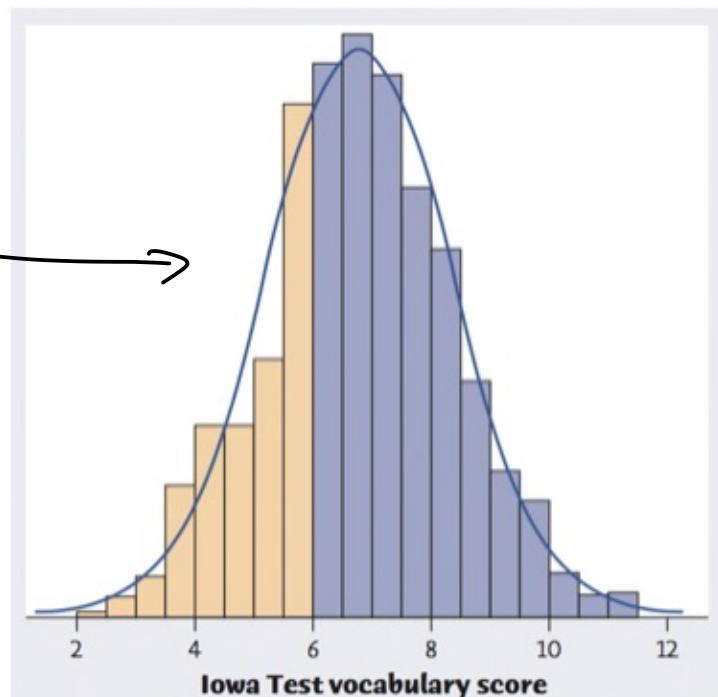


## Chapter 3: Normal Distributions

Ex 947 students in Gary, IN took the Iowa test. Here is a histogram of their scores. The histogram is roughly symmetric, has no large gaps or outliers



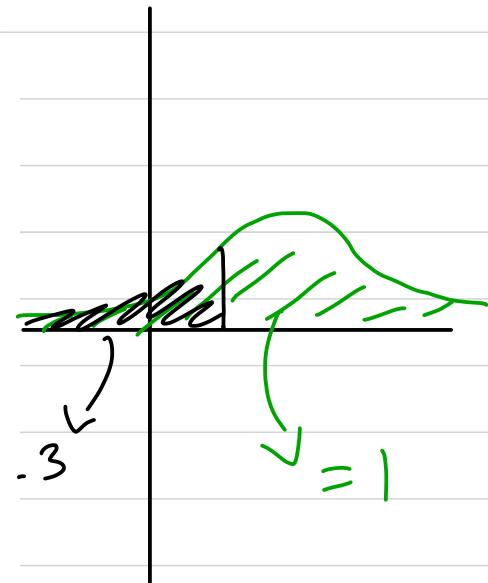
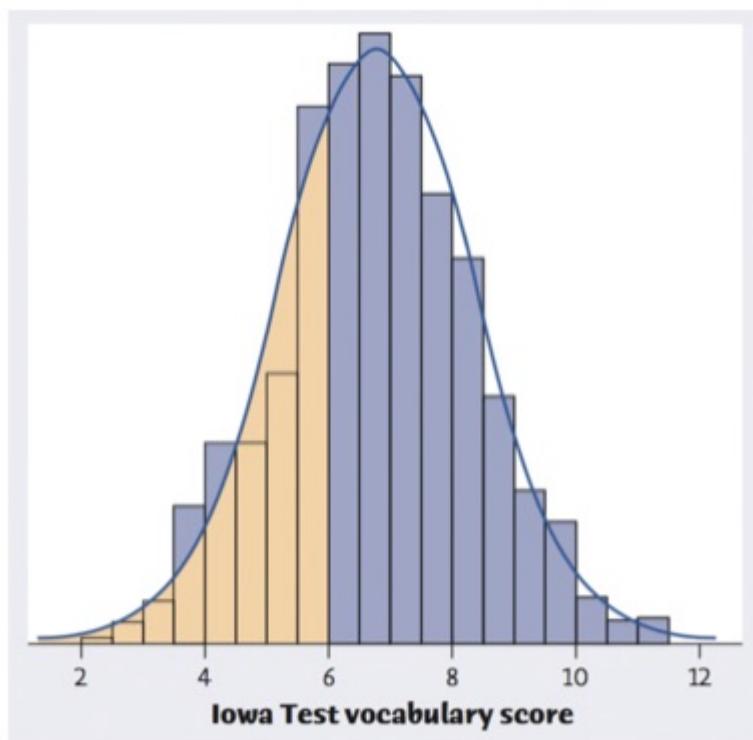
What is the proportion of students who scored 6 or lower?



Curve:  
approximate  
distribution

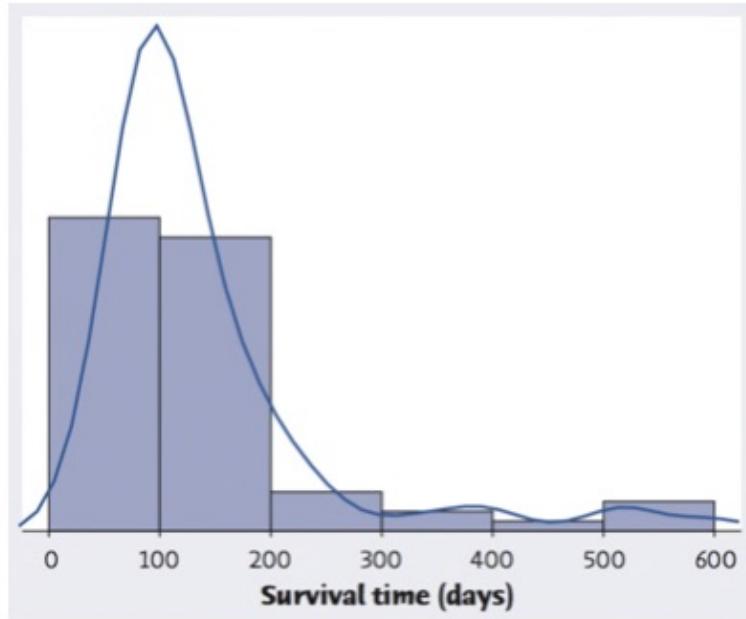
bars:  
observation

We want to find this proportion via the bell curve and not the histogram. Want the area under the curve less than 6

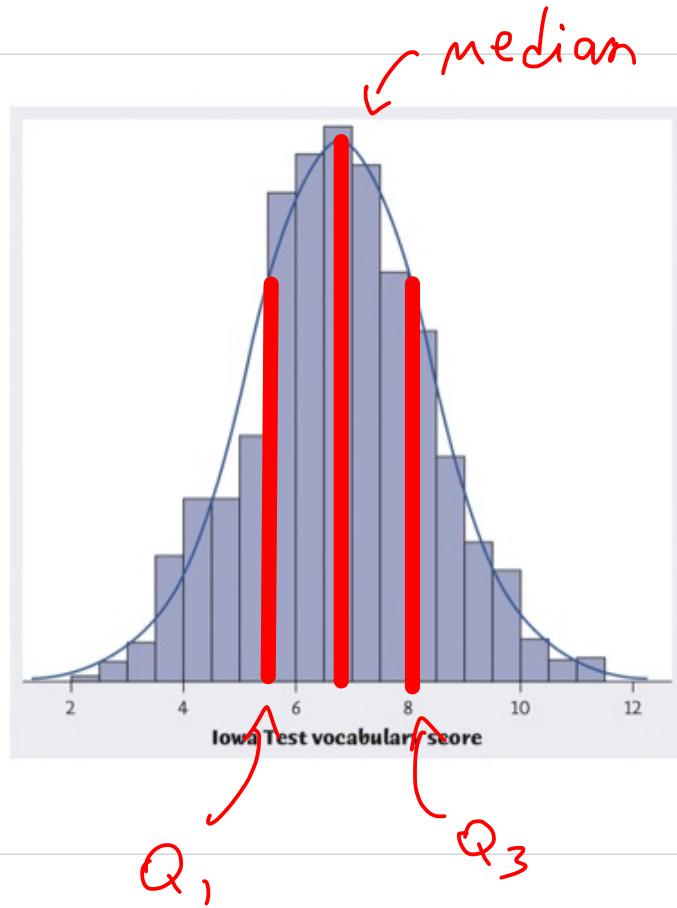


If we rescale the bell curve to have area 1, then the orange area will already be the proportion we want.

This defines something called a density curve: it's positive and has area 1. They come in many shapes: here's one that approximates a skewed dist:



Median + quartiles of density curves: just split the area into quarters.



no max b/c  
there is an asymptote at 0

Think of the mean as a weighted average:  
It's the "balance point" of the density curve

For symmetric distributions, mean = median

Notational convention:

Observation  $\xrightarrow{\text{sample}}$

$\bar{x}$  : mean

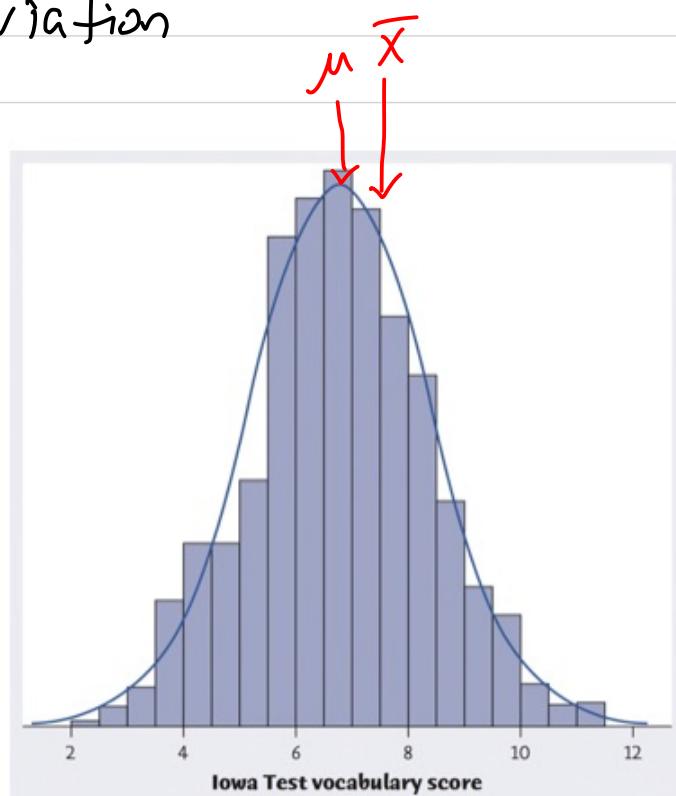
$s$  : standard deviation

$\xrightarrow{\text{population}}$

Distribution

$\mu (\mu_0)$  : mean

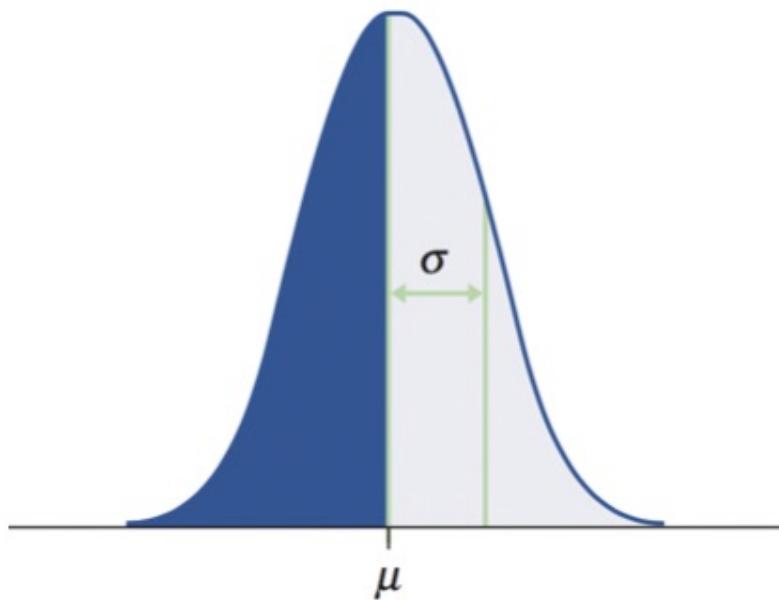
$\sigma$  (sigma) : std dev



It turns out that normal distributions are completely determined by  $\mu$  and  $\sigma$ .

Def: A normal distribution is one whose density curve is symmetric, single-peaked, and bell-shaped.

Eyeball  $\sigma$ : it's where the curve changes concavity: imagine skiing down the curve. The point when the slope stops getting steeper is the inflection point, and it's where  $\sigma$

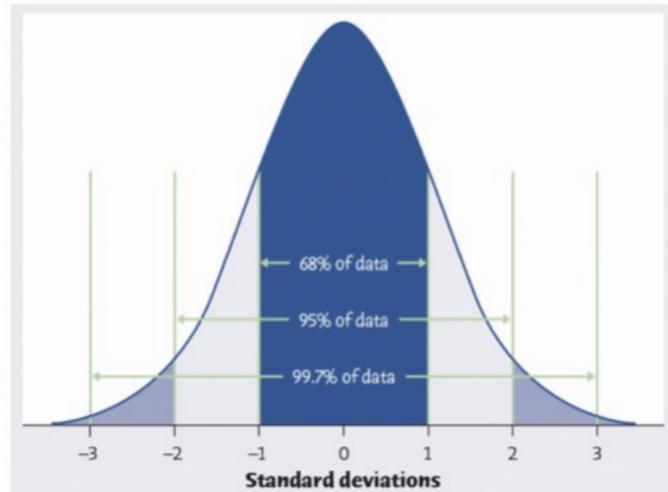


is.

Prop: In a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ ,

- ① 68% of the observations lie within  $\sigma$  of  $\mu$ .
- ② 95% of the observations lie within  $2\sigma$  of  $\mu$ .
- ③ 99.7% of the observations lie within  $3\sigma$  of  $\mu$ .

This is called the 68-95-99.7 rule,



$$\begin{array}{l} \mu = 0 \\ \sigma = 1 \end{array}$$

Ex: The # of heads obtained by flipping a coin 1000 times is roughly given by a normal distribution.

$$\mu = 500$$

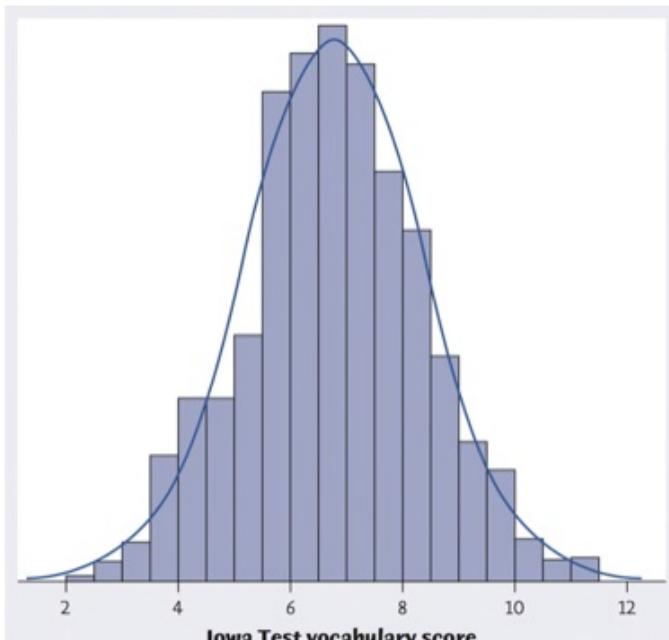
can't get to  
this yet  $\rightarrow \sigma = \sqrt{250} \approx 15.8$

If you run many trials of flipping a coin 1000 times, roughly 68% of those trials will have between  $500 - 15.8$

and  $500 + 15.8$  heads.  
 $515.8$

184.2

Ex:



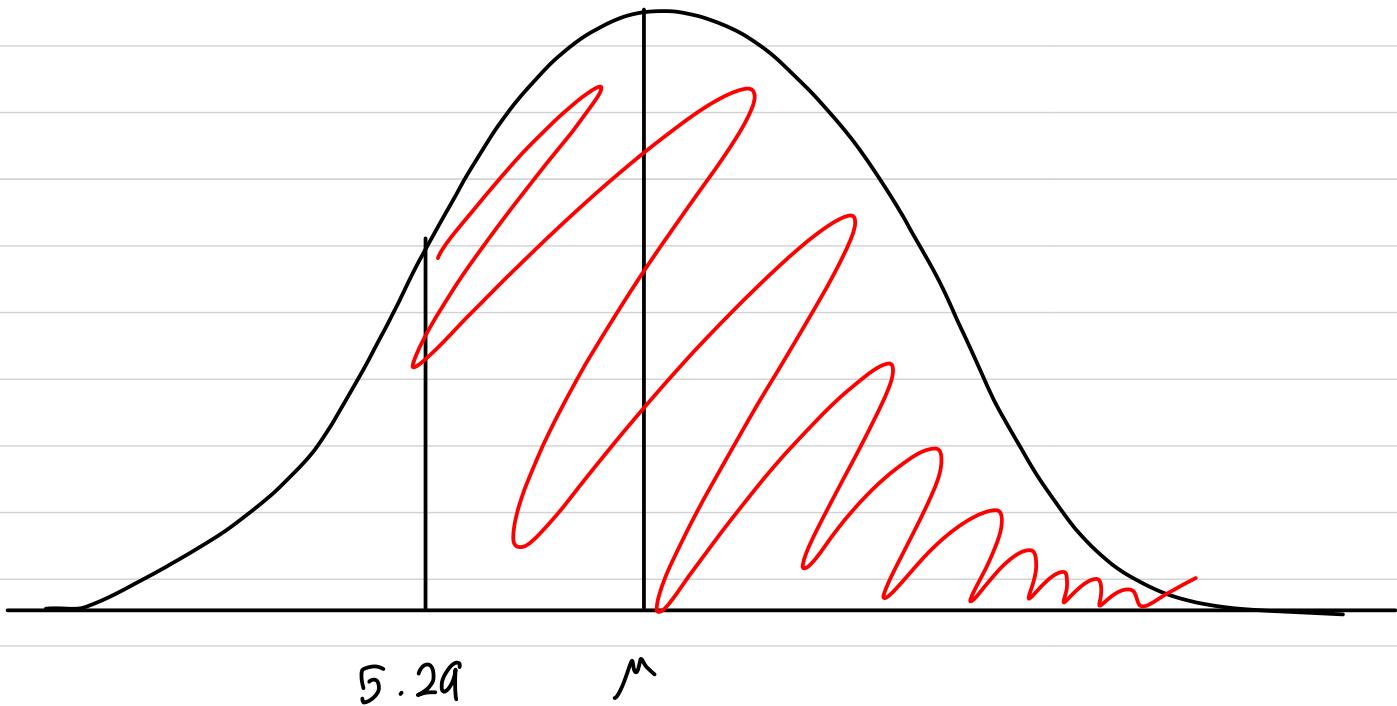
$$\mu = 6.84 \quad \sigma = 1.55 \quad \left. \right\} \text{found by a computer}$$

We expect 95% of the scores to be  
 within 3.74 and 9.94  
 $\uparrow$                              $\uparrow$   
 $6.84 - (2)(1.55)$                $6.84 + (2)(1.55)$

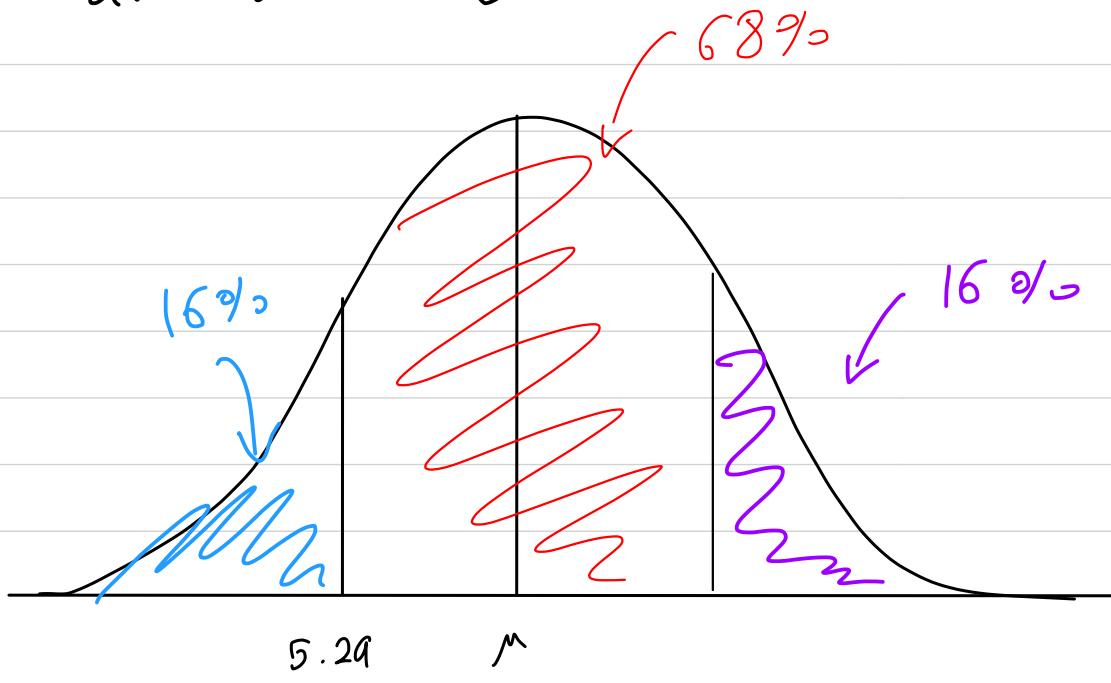
obtained by counting bars  
 Actually: 900 between 3.74 and 9.94,  
 so  $900/974 = .9504 = 95.04\%$ . Really  
 accurate!

Since the distribution is symmetric, we'd  
 expect as many students to be above  
 9.94 as below 3.74. In reality, it's 27  
 and 20.

Ex :



In this distribution, 5.29 is one standard deviation below the mean. What percent of scores are above 5.29?



Since the bell curve is symmetric, the 32% that isn't red is split equally between the left and right tails. In particular, 16% is contained in the right one. So, a total of  $68\% + 16\% = 84\%$  is above 5.29.

Def: We denote a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  by  $N(\mu, \sigma)$ .

Ex: the distribution of test scores from the past example is approximately  $N(6.84, 1.55)$ .

Comment: All normal distributions are the

same up to  $\mu$  and  $\sigma$ .

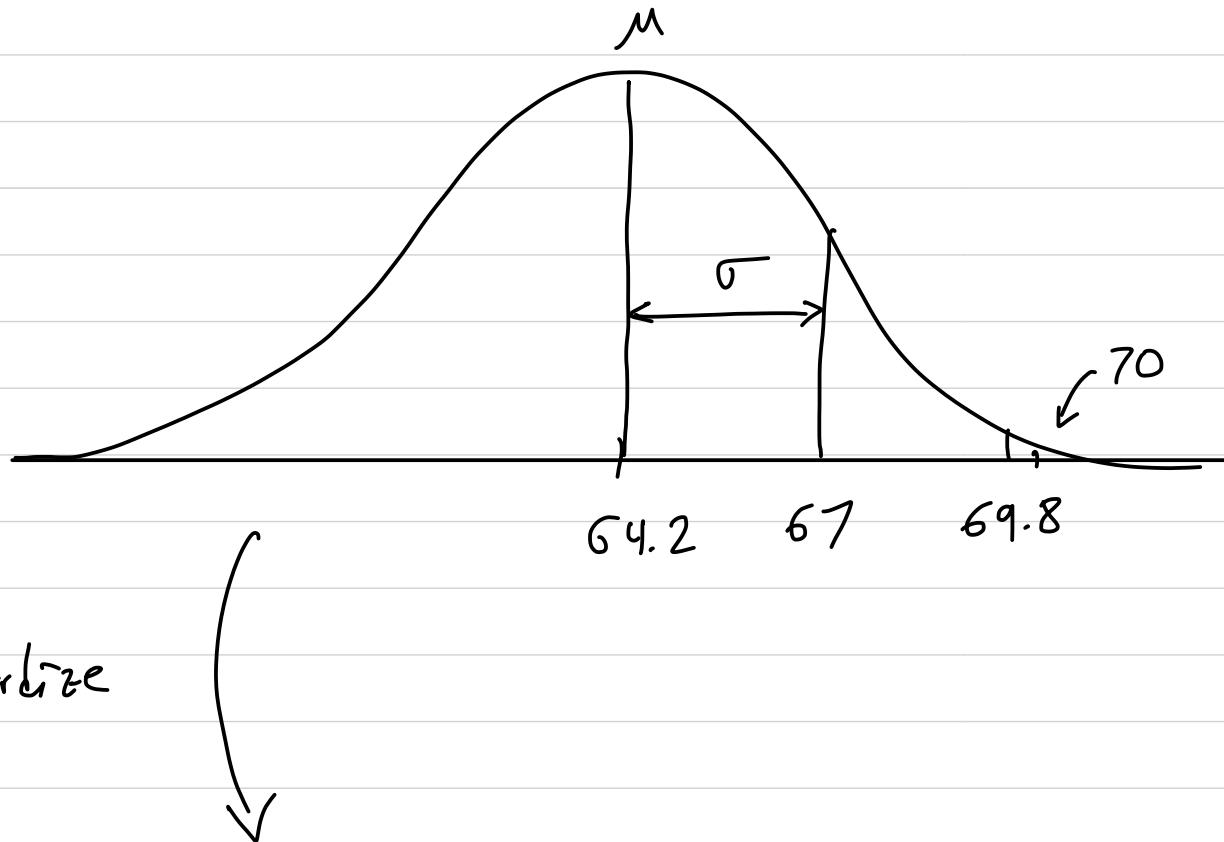
Def: A normal distribution is standardized if we measure in units of  $\sigma$  and center at  $\mu$ . To standardize a value from a normal distribution, first subtract  $\mu$  and then divide by  $\sigma$ . So if  $x$  is an observation from a distribution of type  $N(\mu, \sigma)$ , the standardized value of  $x$  is

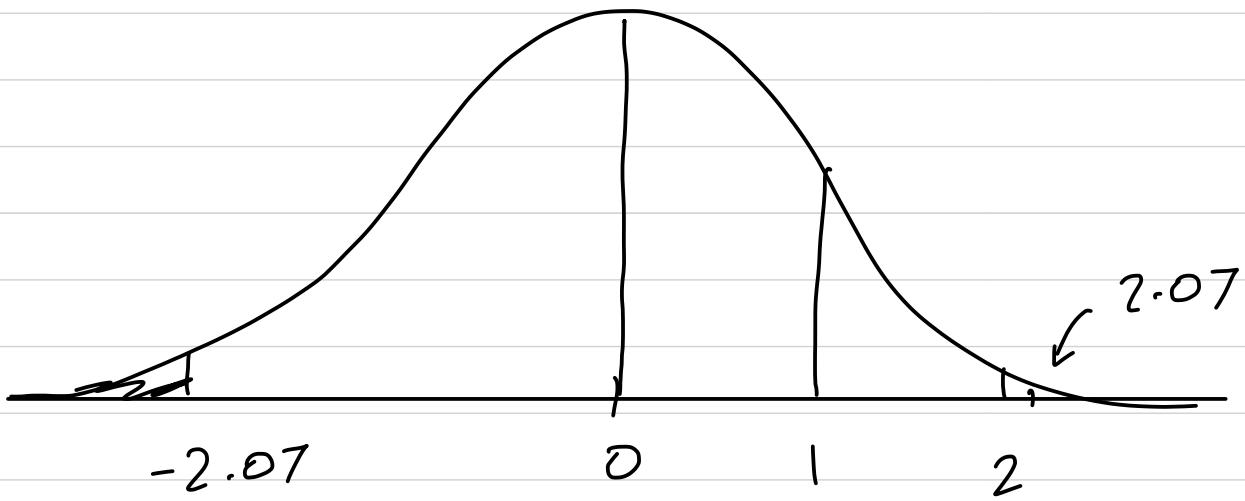
$$z = \frac{x - \mu}{\sigma}.$$

We call this value a  $z$ -score. It tells us how many standard deviations away from the mean  $x$  was.

Ex: the heights of women between 20 and 29 in the US is approximately  $N(64.2, 2.8)$  in inches.

A woman between 20 and 29 in the US with a height of 70 inches has a z-score of  $z = \frac{70 - 64.2}{2.8} = 2.07$ , so her height is 2.07 standard deviations above the mean.

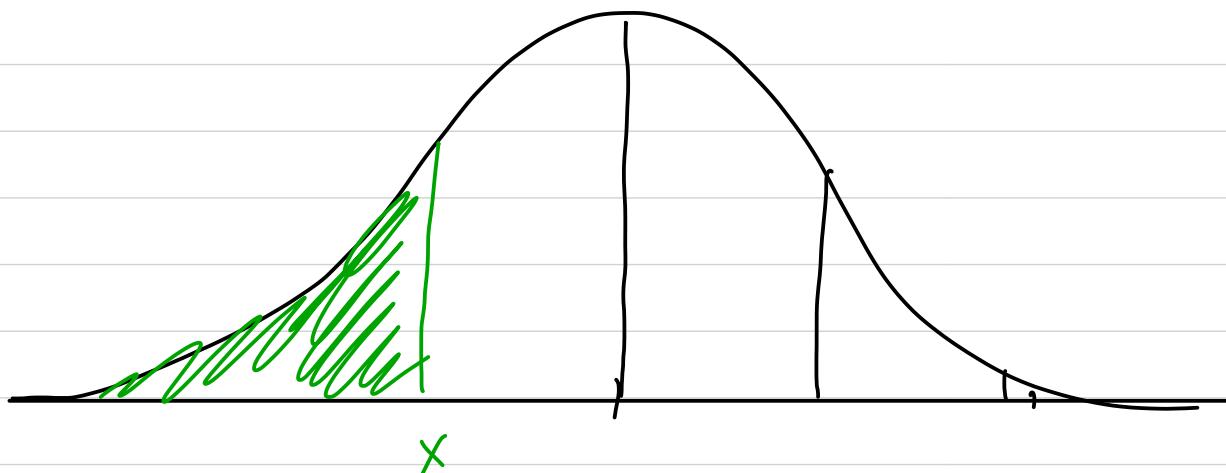




Why do we care?  $z$ -scores let us compare across different bell curves. For example, men between 20 and 29 in US will have heights that follow a bell curve with a higher mean, but if we use the  $z$ -score, we can compare a given man's height to a woman's as a measure of how tall they are relative to their populations.

Although we don't have a formula for an arbitrary area under the standard bell curve, we can use a calculator or table to get a certain kind of area, called a cumulative proportion.

Def: A cumulative proportion for a value  $x$  is the proportion of observations in the distribution that are  $\leq x$ .



Key: standardizing doesn't change the proportion

**TABLE A Standard normal cumulative proportions**

| z    | .00   | .01   | .02   | .03   | .04   | .05   | .06   | .07   | .08   | .09   |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| -3.4 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0003 | .0002 |
| -3.3 | .0005 | .0005 | .0005 | .0004 | .0004 | .0004 | .0004 | .0004 | .0004 | .0003 |
| -3.2 | .0007 | .0007 | .0006 | .0006 | .0006 | .0006 | .0006 | .0005 | .0005 | .0005 |
| -3.1 | .0010 | .0009 | .0009 | .0009 | .0008 | .0008 | .0008 | .0008 | .0007 | .0007 |
| -3.0 | .0013 | .0013 | .0013 | .0012 | .0012 | .0011 | .0011 | .0011 | .0010 | .0010 |
| -2.9 | .0019 | .0018 | .0018 | .0017 | .0016 | .0016 | .0015 | .0015 | .0014 | .0014 |
| -2.8 | .0026 | .0025 | .0024 | .0023 | .0023 | .0022 | .0021 | .0021 | .0020 | .0019 |
| -2.7 | .0035 | .0034 | .0033 | .0032 | .0031 | .0030 | .0029 | .0028 | .0027 | .0026 |
| -2.6 | .0047 | .0045 | .0044 | .0043 | .0041 | .0040 | .0039 | .0038 | .0037 | .0036 |
| -2.5 | .0062 | .0060 | .0059 | .0057 | .0055 | .0054 | .0052 | .0051 | .0049 | .0048 |
| -2.4 | .0082 | .0080 | .0078 | .0075 | .0073 | .0071 | .0069 | .0068 | .0066 | .0064 |
| -2.3 | .0107 | .0104 | .0102 | .0099 | .0096 | .0094 | .0091 | .0089 | .0087 | .0084 |
| -2.2 | .0139 | .0136 | .0132 | .0129 | .0125 | .0122 | .0119 | .0116 | .0113 | .0110 |
| -2.1 | .0179 | .0174 | .0170 | .0166 | .0162 | .0158 | .0154 | .0150 | .0146 | .0143 |
| -2.0 | .0228 | .0222 | .0217 | .0212 | .0207 | .0202 | .0197 | .0192 | .0188 | .0183 |
| -1.9 | .0287 | .0281 | .0274 | .0268 | .0262 | .0256 | .0250 | .0244 | .0239 | .0233 |
| -1.8 | .0359 | .0351 | .0344 | .0336 | .0329 | .0322 | .0314 | .0307 | .0301 | .0294 |
| -1.7 | .0446 | .0436 | .0427 | .0418 | .0409 | .0401 | .0392 | .0384 | .0375 | .0367 |
| -1.6 | .0548 | .0537 | .0526 | .0516 | .0505 | .0495 | .0485 | .0475 | .0465 | .0455 |
| -1.5 | .0668 | .0655 | .0643 | .0630 | .0618 | .0606 | .0594 | .0582 | .0571 | .0559 |
| -1.4 | .0808 | .0793 | .0778 | .0764 | .0749 | .0735 | .0721 | .0708 | .0694 | .0681 |
| -1.3 | .0968 | .0951 | .0934 | .0918 | .0901 | .0885 | .0869 | .0853 | .0838 | .0823 |
| -1.2 | .1151 | .1131 | .1112 | .1093 | .1075 | .1056 | .1038 | .1020 | .1003 | .0985 |
| -1.1 | .1357 | .1335 | .1314 | .1292 | .1271 | .1251 | .1230 | .1210 | .1190 | .1170 |
| -1.0 | .1587 | .1562 | .1539 | .1515 | .1492 | .1469 | .1446 | .1423 | .1401 | .1379 |
| -0.9 | .1841 | .1814 | .1788 | .1762 | .1736 | .1711 | .1685 | .1660 | .1635 | .1611 |
| -0.8 | .2119 | .2090 | .2061 | .2033 | .2005 | .1977 | .1949 | .1922 | .1894 | .1867 |
| -0.7 | .2420 | .2389 | .2358 | .2327 | .2296 | .2266 | .2236 | .2206 | .2177 | .2148 |
| -0.6 | .2743 | .2709 | .2676 | .2643 | .2611 | .2578 | .2546 | .2514 | .2483 | .2451 |
| -0.5 | .3085 | .3050 | .3015 | .2981 | .2946 | .2912 | .2877 | .2843 | .2810 | .2776 |
| -0.4 | .3446 | .3409 | .3372 | .3336 | .3300 | .3264 | .3228 | .3192 | .3156 | .3121 |
| -0.3 | .3821 | .3783 | .3745 | .3707 | .3669 | .3632 | .3594 | .3557 | .3520 | .3483 |
| -0.2 | .4207 | .4168 | .4129 | .4090 | .4052 | .4013 | .3974 | .3936 | .3897 | .3859 |
| -0.1 | .4602 | .4562 | .4522 | .4483 | .4443 | .4404 | .4364 | .4325 | .4286 | .4247 |
| -0.0 | .5000 | .4960 | .4920 | .4880 | .4840 | .4801 | .4761 | .4721 | .4681 | .4641 |

.0192 means that 1.92% of women have a z-score below -2.07  
(so 1.92% have one above 2.07)