



Statistics for Health Care

("Statistics in Medicine" Christen Sainani @ Stanford University)

Unit 1

Overview/Teasers



First rules of statistics...

Use common sense!

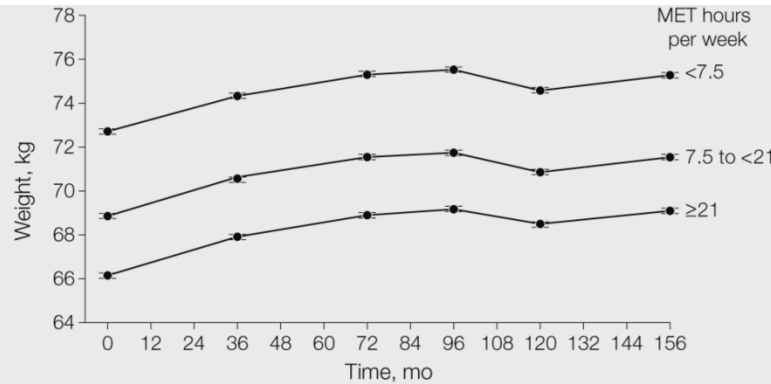
Draw lots of pictures!



What's wrong with this?

- Study with sample size of 10 ($N=10$)
- Results: "Objective scoring by blinded investigators indicated that the treatment resulted in improvement in all (100%) of the subjects. Of patients showing overall improvement, 78% were graded as having either excellent or moderate improvement."

Take-home message?



No. by MET hours
per week

<7.5	16856	15634	15153	15661	13779	13353
7.5 to <21	9819	9171	9005	9260	8336	8106
≥21	7404	6924	6808	6992	6264	6107

Do the three groups differ meaningfully in weight change over time?



Preview: Unit 1

- How to think about, look at, and describe data



Teaser 1, Unit 1

- Hypothetical randomized trial comparing two diets:
- Those on diet 1 ($n=10$) lost an average of 34.5 lbs.
- Those on diet 2 ($n=10$) lost an average of 18.5 lbs.
- Conclusion: diet 1 is better?



Teaser 2, Unit 1

- “400 shades of lipstick found to contain lead”, FDA says” *Washington Post*, Feb. 14, 2012
- “What’s in Your Lipstick? FDA Finds Lead in 400 Shades,” *Time.com* February 15, 2012
- How worried should women who use lipstick be?



Statistics for Health Care

Introduction to Data



Example Data

- Data compiled from previous Stanford students (anonymous, non-identifiable)
- Sample size = 50

Example Data Set

MEDSTAT.MOCKDATA																			
26	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
56	ID	coffee	Varsity	Milk	exercise	wakeup	politics	fruitveg	obama	bushjr	clinton	bushsr	regan	carter	Mathle				
1	1	16	0	6	2.0	7.5	38	4	40	70	70	50	70	49					
2	2	0	1	0	0.0	9.5	60	3	72	24	76	19	64	62					
3	3	2	1	0	3.0	7.0	95	7	85	0	92	10	20	40					
4	4	0	1	0	3.0	7.5	80	3	80	8	86	.	25	51					
5	5	2	0	0	2.0	9.0	72	10	65	18	72	37	39	48					
6	6	12	1	0	3.0	7.5	70	6	80	20	75	30	35	60					
7	7	16	1	6	2.5	7.5	80	3	80	20	80	40	45	50					
8	8	4	1	2	10.0	6.5	75	8	80	0	.	75	20	60					
9	9	12	1	1	3.5	7.5	80	1	80	5	90	20	21	65					
10	10	4	0	4	3.5	9.0	85	4	65	10	80	.	25	.					
11	11	0	0	12	1.0	7.5	.	2	72	24					
12	12	6	1	6	7.0	8.5	80	3	78	15	78	23	23	78					
13	13	32	0	4	4.5	6.0	84	5	100	11	69	28	38	60					
14	14	16	1	4	5.0	6.5	80	3	75	30	72	40	.	.					
15	15	12	1	8	4.0	6.0	75	4	88	16	90	.	73	.					
16	16	3	0	6	2.5	6.5	75	5	95	0	.	15	50	51					
17	17	8	0	0	3.0	7.0	50	3	70	45	70	49	49	51					
18	18	8	1	4	4.0	7.0	80	3	90	2	100	20	10	80					
19	19	40	0	4	3.0	6.0	80	6	80	20	80	35	35	50					
20	20	16	1	0	6.0	6.5	71	4	74	32					
21	21	4	1	4	0.5	7.5	75	2	75	10	65	5	5	65					
22	22	2	0	8	2.0	7.0	66	3	80	7	88	37	44	.					
23	23	0	1	16	7.0	7.5	65	1	71	0	72	57	.	.					
24	24	16	1	10	6.0	7.5	.	2					
25	25	10	1	3	4.5	6.0	85	4	80	10	90	20	25	90					
26	26	16	1	6	3.0	7.5	78	2	82	0	93	0	0	.					
27	27	12	1	5	2.0	7.5	64	4	66	37	61	.	.	.					
28	28	12	1	12	12.0	7.0	67	4	71	14	73	12	18	.					

MEDSTAT.MOCKDATA

	26	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
56	ID	coffee	Varsity	Milk	exercise	wakeup	politics	fruitveg	obama	bushjr	clinton	bushsr	regan	carter	Mathle			
1	1	16	0	6	2.0	7.5	38	4	40	70	70	50	70	49				
2	2	0	1	0	0.0	9.5	60	3	72	24	76	19	64	62				
3	3	2	1	0	3.0	7.0	95	7	85	0	92	10	20	40				
4	4	0	1	0	3.0	7.5	80	3	80	8	86	.	25	51				
5	5	2	0	0	2.0	9.0	72	10	65	18	72	37	39	48				
6	6	12	1	0	3.0	7.5	70	6	80	20	75	30	35	60				
7	7	16	1	6	2.5	7.5	80	3	80	20	80	40	45	50				
8	8	4	1	2	10.0	6.5	75	8	80	0	.	75	20	60				
9	9	12	1	1	3.5	7.5	80	1	80	5	90	20	21	65				
10	10	4	0	4	3.5	9.0	85	4	65	10	80	.	25	.				
11	11	0	0	12	1.0	7.5	.	2	72	24				
12	12	6	1	6	7.0	8.5	80	3	78	15	78	23	23	78				
13	13	32	0	4	4.5	6.0	84	5	100	11	69	28	38	60				
14	14	16	1	4	5.0	6.5	80	3	75	30	72	40	.	.				
15	15	12	1	8	4.0	6.0	75	4	88	16	90	.	73	.				
16	16	3	0	6	2.5	6.5	75	5	95	0	.	15	50	51				
17	17	8	0	0	3.0	7.0	50	3	70	45	70	49	49	51				
18	18	8	1	4	4.0	7.0	80	3	90	2	100	20	10	80				
19	19	40	0	4	3.0	6.0	80	6	80	20	80	35	35	50				
20	20	16	1	0	6.0	6.5	71	4	74	32				
21	21	4	1	4	0.5	7.5	75	2	75	10	65	5	5	65				
22	22	2	0	8	2.0	7.0	66	3	80	7	88	37	44	.				
23	23	0	1	16	7.0	7.5	65	1	71	0	72	57	.	.				
24	24	16	1	10	6.0	7.5	.	2				
25	25	10	1	3	4.5	6.0	85	4	80	10	90	20	25	90				
26	26	16	1	6	3.0	7.5	78	2	82	0	93	0	0	.				

Each row stores the data for 1 student (1 observation).

Each column stores the values for 1 variable (e.g., ounces of coffee per day).

MEDSTAT.MOCKDATA

	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
56	ID	coffee	Varsity	Milk	exercise	wakeup	politics	fruitveg	obama	bushjr	clinton	bushsr	regan	carter	Mathl	
1	1	16	0	6	2.0	7.5	38	4	40	70	70	50	70	49		
2	2	0	1	0	0.0	9.5	60	3	72	24	76	19	64	62		
3	3	2	1	0	3.0	7.0	95	7	85	0	92	10	20	40		
4	4	0	1	0	3.0	7.5	80	3	80	8	86	.	25	51		
5	5	2	0	0	2.0	9.0	72	10	65	18	72	37	39	48		
6	6	12	1	0	3.0	7.5	70	6	80	20	75	30	35	60		
7	7	16	1	6	2.5	7.5	80	3	80	20	80	40	45	50		
8	8	4	1	2	10.0	6.5	75	8	80	0	.	75	20	60		
9	9	12	1	1	3.5	7.5	80	1	80	5	90	20	21	65		
10	10	4	0	4	3.5	9.0	85	4	65	10	80	.	25	.		
11	11	0	0	12	1.0	7.5	.	2	72	24		
12	12	6	1	6	7.0	8.5	80	3	78	15	78	23	23	78		
13	13	32	0	4	4.5	6.0	84	5	100	11	69	28	38	60		
14	14	16	1	4	5.0	6.5	80	3	75	30	72	40	.	.		
15	15	12	1	8	4.0	6.0	75	4	88	16	90	.	73	.		
16	16	3	0	6	2.5	6.5	75	5	95	0	.	15	50	51		
17	17	8	0	0	3.0	7.0	50	3	70	45	70	49	49	51		
18	18	8	1	4	4.0	7.0	80	3	90	2	100	20	10	80		
19	19	40	0	4	3.0	6.0	80	6	80	20	80	35	35	50		
20	20	16	1	0	6.0	6.5	71	4	74	32		
21	21	4	1	4	0.5	7.5	75	2	75	10	65	5	5	65		
22	22	2	0	8	2.0	7.0	66	3	80	7	88	37	44	.		
23	23	0	1	16	7.0	7.5	65	1	71	0	72	57	.	.		
24	24	16	1	10	6.0	7.5	.	2		
25	25	10	1	3	4.5	6.0	85	4	80	10	90	20	25	90		
26	26	16	1	6	3.0	7.5	78	2	82	0	93	0	0	.		
27	27	12	1	5	2.0	7.5	64	4	66	37	61	.	.	.		
28	28	12	1	12	12.0	7.0	67	4	71	14	73	12	18	.		
29	29	20	0	0	2.0	5.5	39	3	32	29	61	39	76	47		
30	30	2	1	50	7.0	5.0	27	5	59	96	92	90	11	12		
31	31	12	0	6	2.5	7.0	99	1	100	10	80	1	1	8		
32	32	0	0	16	8.0	7.0	82	1	90	0	86	1	4	8		
33	33	12	1	0	3.0	9.0	72	5	82	22	72	3	4	4		

**Missing
Data!**

MEDSTAT.MOCKDATA																	
	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int	Int
56	ID	coffee	Varsity	Milk	exercise	wakeup	politics	fruitveg	obama	bushjr	clinton	bushsr	regan	carter	Mathl		
1	1	16	0	6	2.0	7.5	38	4	40	70	70	50	70	49			
2	2	0	1	0	0.0	9.5	60	3	72	24	76	19	64	62			
3	3	2	1	0	3.0	7.0	95	7	85	0	92	10	20	40			
4	4	0	1	0	3.0	7.5	80	3	80	8	86	.	25	51			
5	5	2	0	0	2.0	9.0	72	10	65	18	72	37	39	48			
6	6	12	1	0	3.0	7.5	70	6	80	20	75	30	35	60			
7	7	16	1	6	2.5	7.5	80	3	80	20	80	40	45	50			
8	8	4	1	2	10.0	6.5	75	8	80	0	.	75	20	60			
9	9	12	1	1	3.5	7.5	80	1	80	5	90	20	21	65			
10	10	4	0	4	3.5	9.0	85	4	65	10	80	.	25	.			
11	11	0	0	12	1.0	7.5	.	2	72	24			
12	12	6	1	6	7.0	8.5	80	3	78	15	78	23	23	78			
13	13	32	0	4	4.5	6.0	84	5	100	11	69	28	38	60			
14	14	16	1	4	5.0	6.5	80	3	75	30	72	40	.	.			
15	15	12	1	8	4.0	6.0	75	4	88	16	90	.	73	.			
16	16	3	0	6	2.5	6.5	75	5	95	0	.	15	50	51			
17	17	8	0	0	3.0	7.0	50	3	70	45	70	49	49	51			
18	18	8	1	4	4.0	7.0	80	3	90	2	100	20	10	80			
19	19	40	0	4	3.0	6.0	80	6	80	20	80	35	35	50			
20	20	16	1	0	6.0	6.5	71	4	74	32			
21	21	4	1	4	0.5	7.5	75	2	75	10	65	5	5	65			
22	22	2	0	8	2.0	7.0	66	3	80	7	88	37	44	.			
23	23	0	1	16	7.0	7.5	65	1	71	0	72	57	.	.			
24	24	16	1	10	6.0	7.5	.	2			
25	25	10	1	3	4.5	6.0	85	4	80	10	90	20	25	90			
26	26	16	1	6	3.0	7.5	78	2	82	0	93	0	0	.			
27	27	12	1	5	2.0	7.5	64	4	66	37	61	.	.	.			
28	28	12	1	12	12.0	7.0	67	4	71	14	73	12	18	.			
29	29	20	0	0	2.0	5.5	39	3	32	29	61	39	76	47			



Statistics for Health Care

Module 2: Types of Data



Types of data

- Quantitative
- Categorical (binary, nominal or ordinal)
- Time-to-event



Quantitative variable

- Numerical data that you can add, subtract, multiply, and divide
- Examples:
 - Age
 - Blood pressure
 - BMI
 - Pulse
- Examples from our example data:
 - Optimism on a 0 to 100 scale
 - Exercise in hours per week
 - Coffee drinking in ounces per day



Quantitative variable

- Continuous vs. Discrete
 - Continuous: can theoretically take on any value within a given range (e.g., height=68.99955... inches)
 - Discrete: can only take on certain values (e.g., count data)



Categorical Variables

- Binary = two categories
 - Dead/alive
 - Treatment/placebo
 - Disease/no disease
 - Exposed/Unexposed
 - Heads/Tails
 - Example data: played varsity sports in high school (yes/no)



Categorical Variables

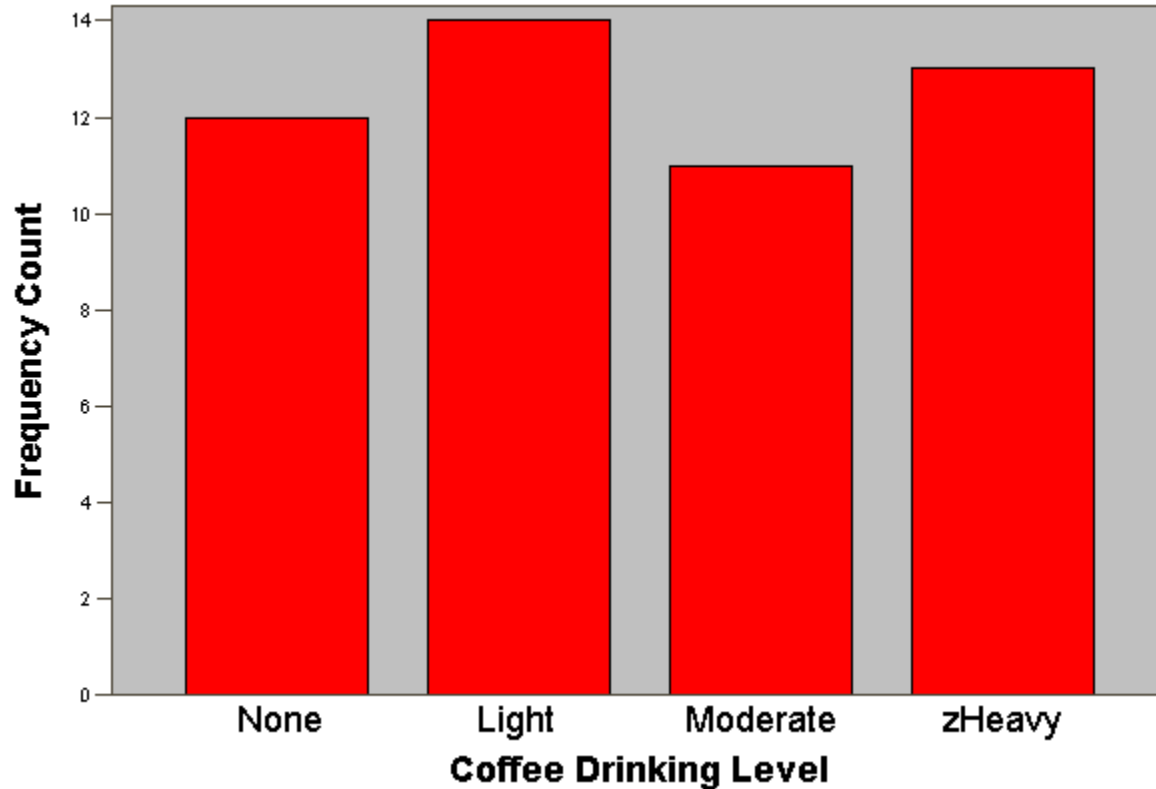
- Nominal = unordered categories
 - The blood type of a patient (O, A, B, AB)
 - Marital status
 - Occupation



Categorical Variables

- Ordinal = Ordered categories
 - Staging in breast cancer as I, II, III, or IV
 - Birth order—1st, 2nd, 3rd, etc.
 - Letter grades (A, B, C, D, F)
 - Ratings on a Likert scale (e.g., strongly agree, agree, neutral, disagree, strongly disagree)
 - Age in categories (10-20, 20-30, etc.)
 - Example data: non-drinker, light drinker, moderate drinker, and heavy drinker of coffee

Coffee Drinking Categories (Ordinal)





Time-to-event variables

- The time it takes for an event to occur, if it occurs at all
- Hybrid variable—has a continuous part (time) and a binary part (event: yes/no)
- Only encountered in studies that follow participants over time—such as cohort studies and randomized trials
- Examples:
 - Time to death
 - Time to heart attack
 - Time to chronic kidney disease



Statistics for Health Care

Module 3: Looking at Data



Always Plot Your Data!

- ✓ Are there “outliers”? some things that abnormal
- ✓ Are there data points that don't make sense?
- ✓ How are the data distributed?

Are there points that don't make sense?





How are the data distributed?

Categorical data:

- What are the N's and percents in each category?

Quantitative data:

- What's the shape of the distribution (e.g., is it normally distributed or skewed)?
- Where is the center of the data?
- What is the spread/variability of the data?



Frequency Plots (univariate)

Categorical variables

- Bar Chart

Quantitative/continuous variables

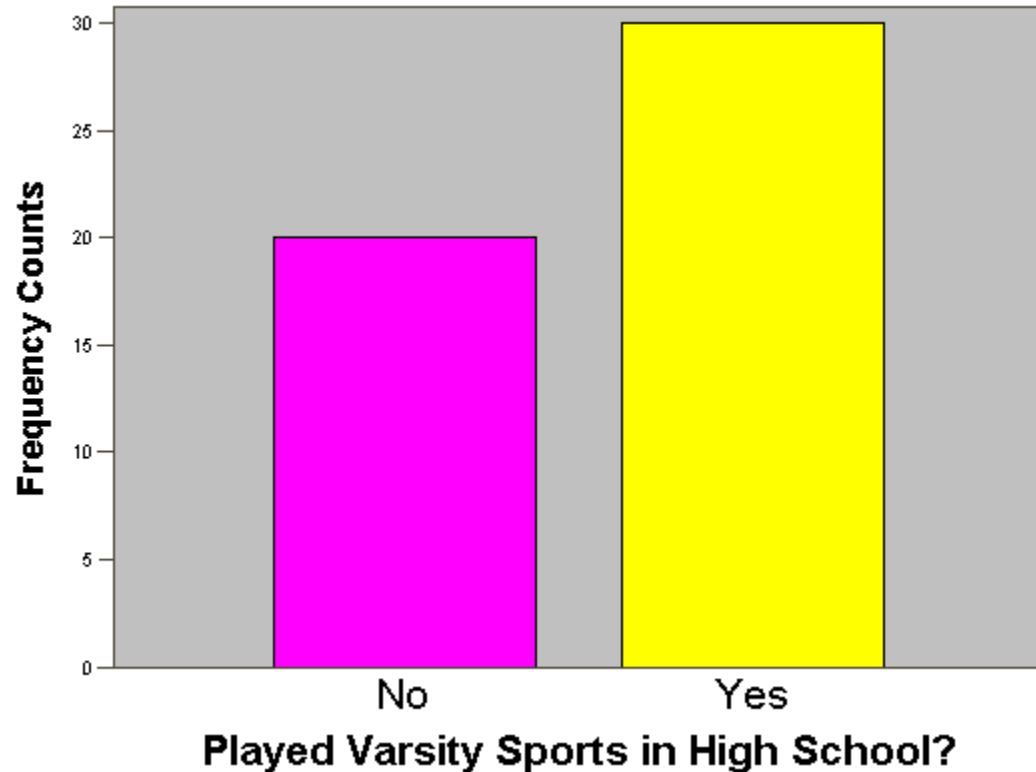
- Box Plot
- Histogram



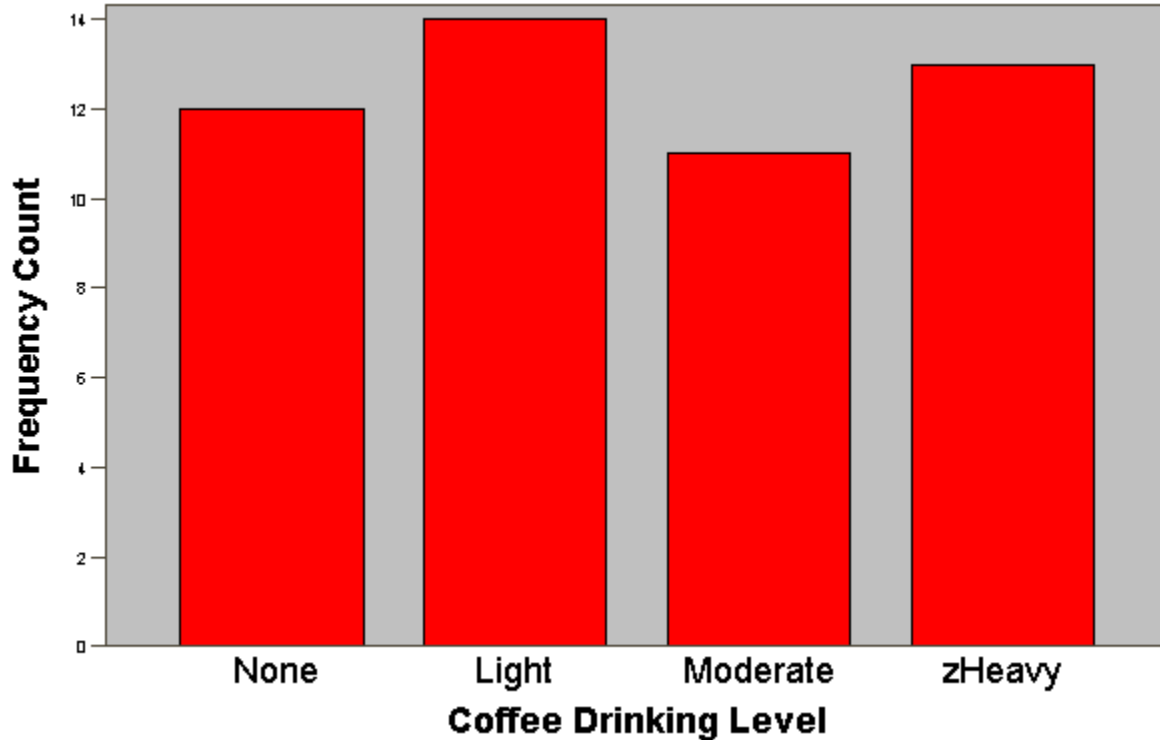
Bar Chart

- Used for categorical variables to show frequency or proportion in each category.

Bar Chart: categorical variables



Bar Chart: categorical variables

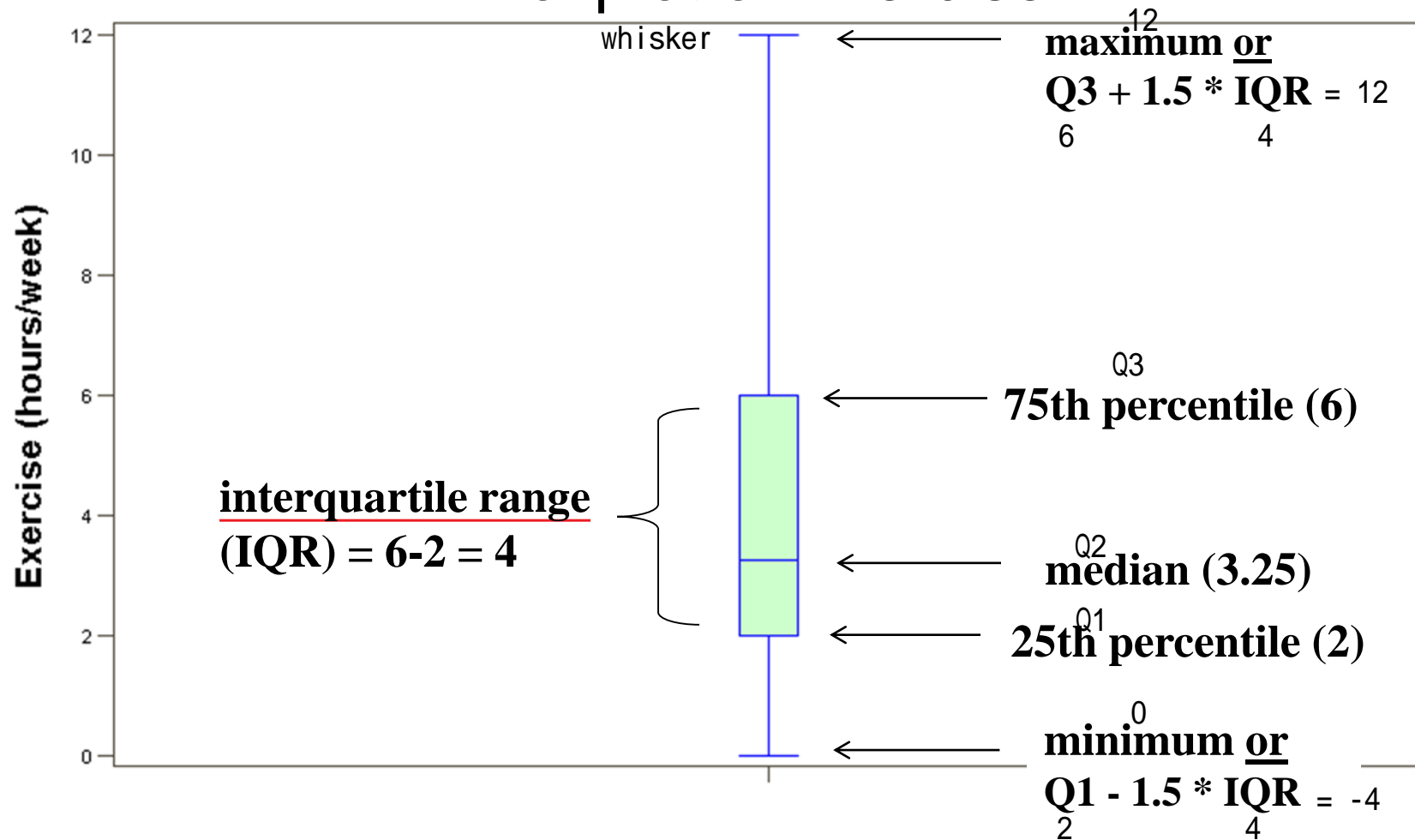




Box plot and histograms: for quantitative variables

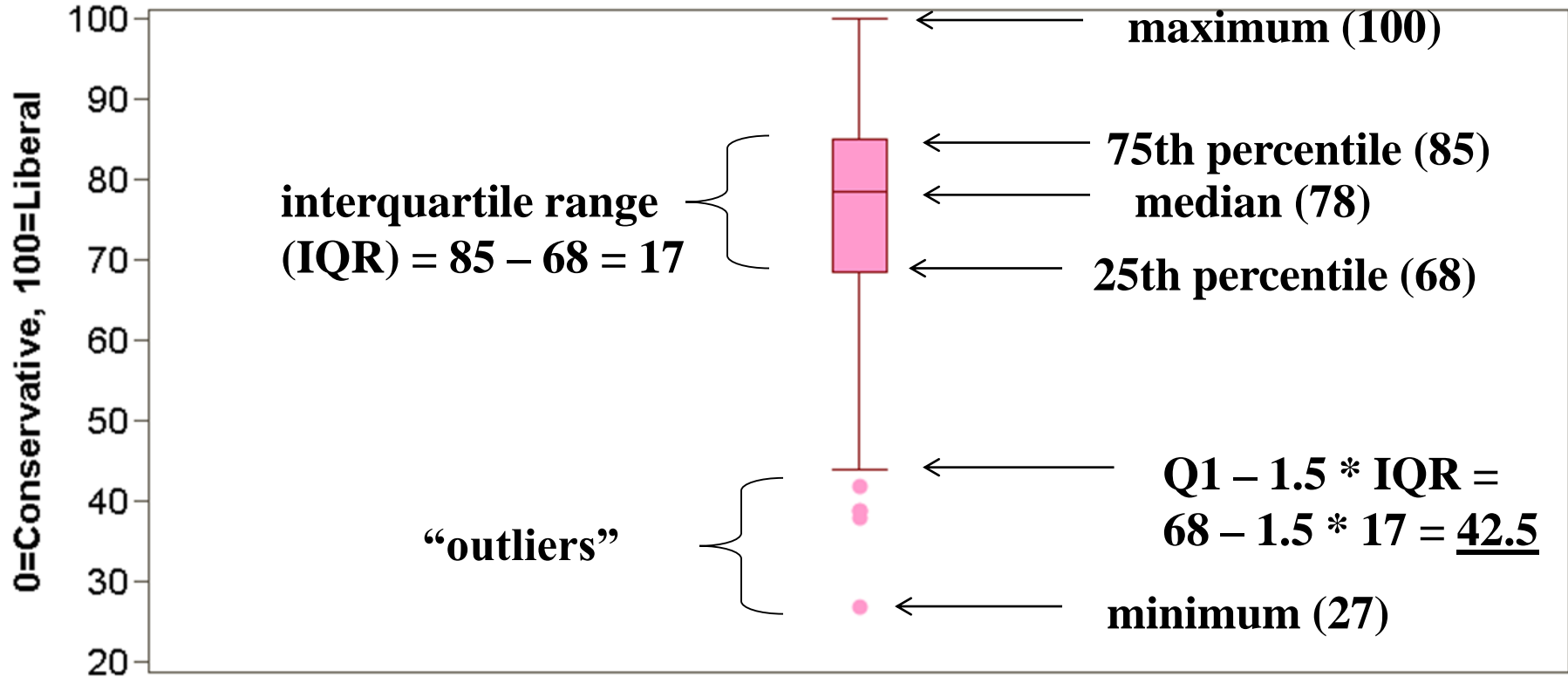
- To show the distribution (shape, center, range, variation) of quantitative variables.

Boxplot of Exercise



Boxplot of Political Bent

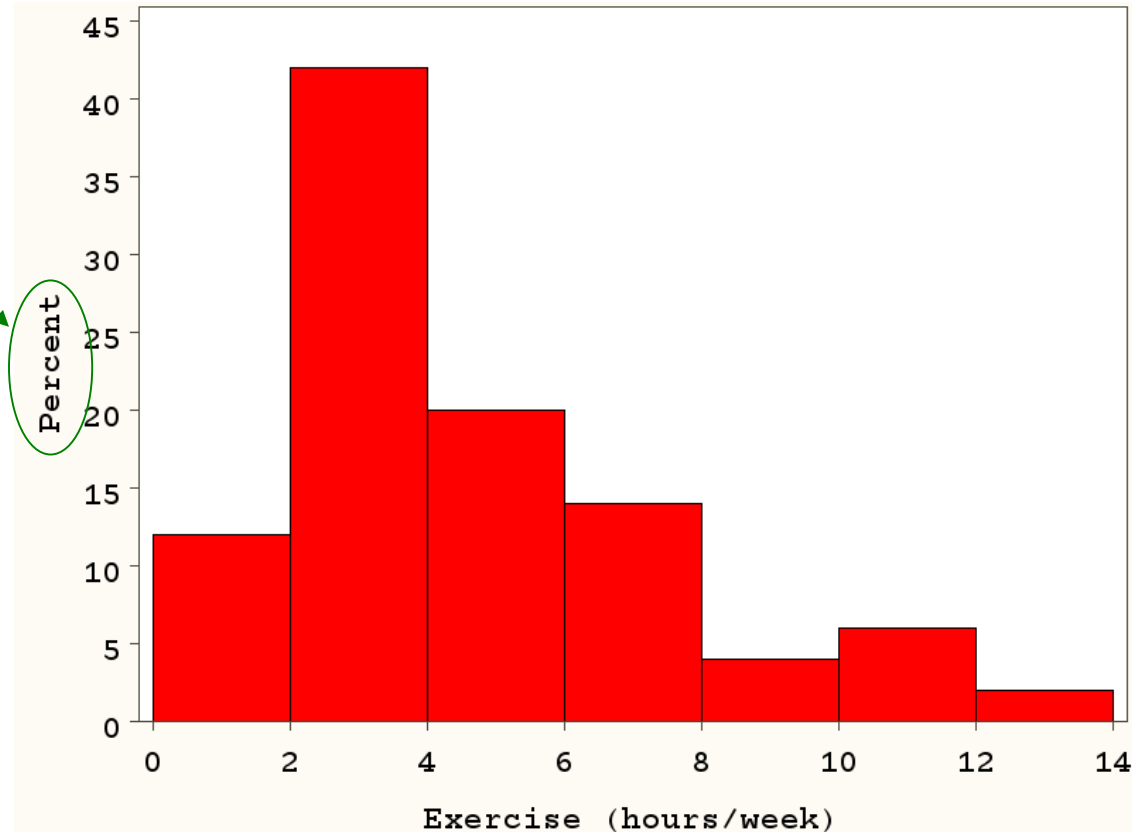
(0=Most Conservative, 100=Most Liberal)



Histogram of Exercise

Bins of size = 2 hours/week

Y-axis: The percent of observations that fall within each bin.

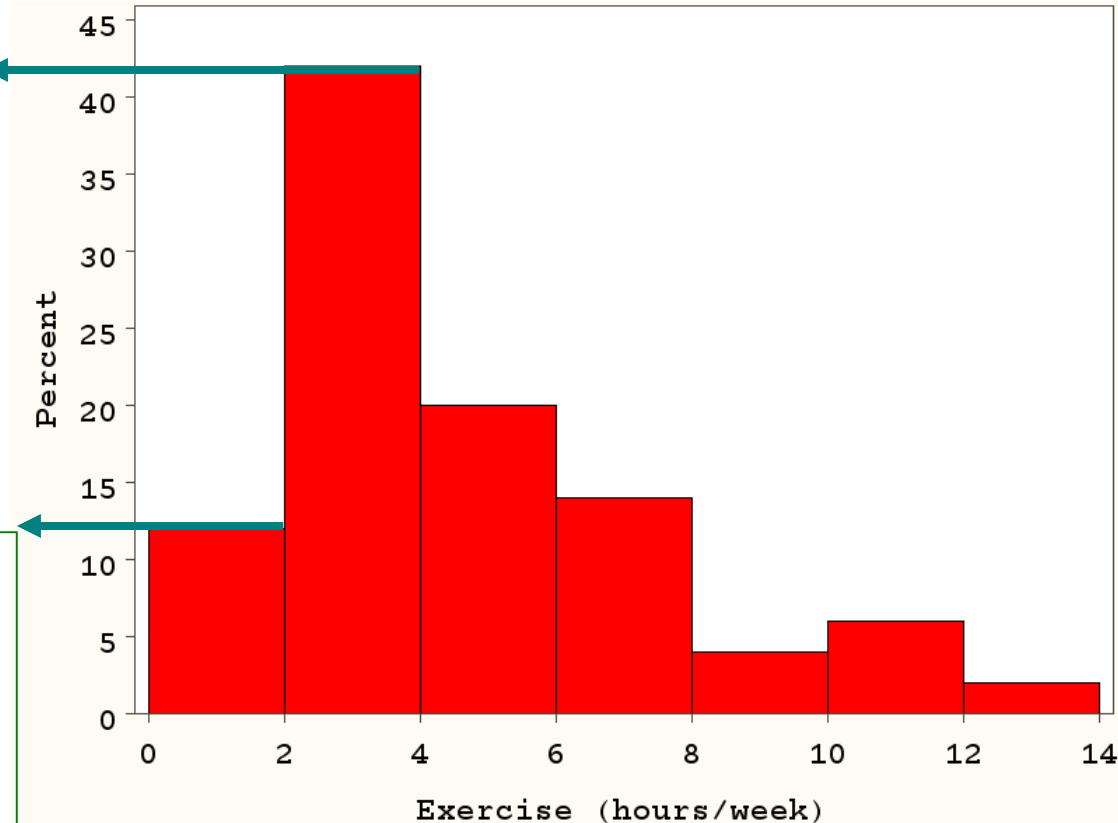


Histogram of Exercise

Bins of size = 2 hours/week

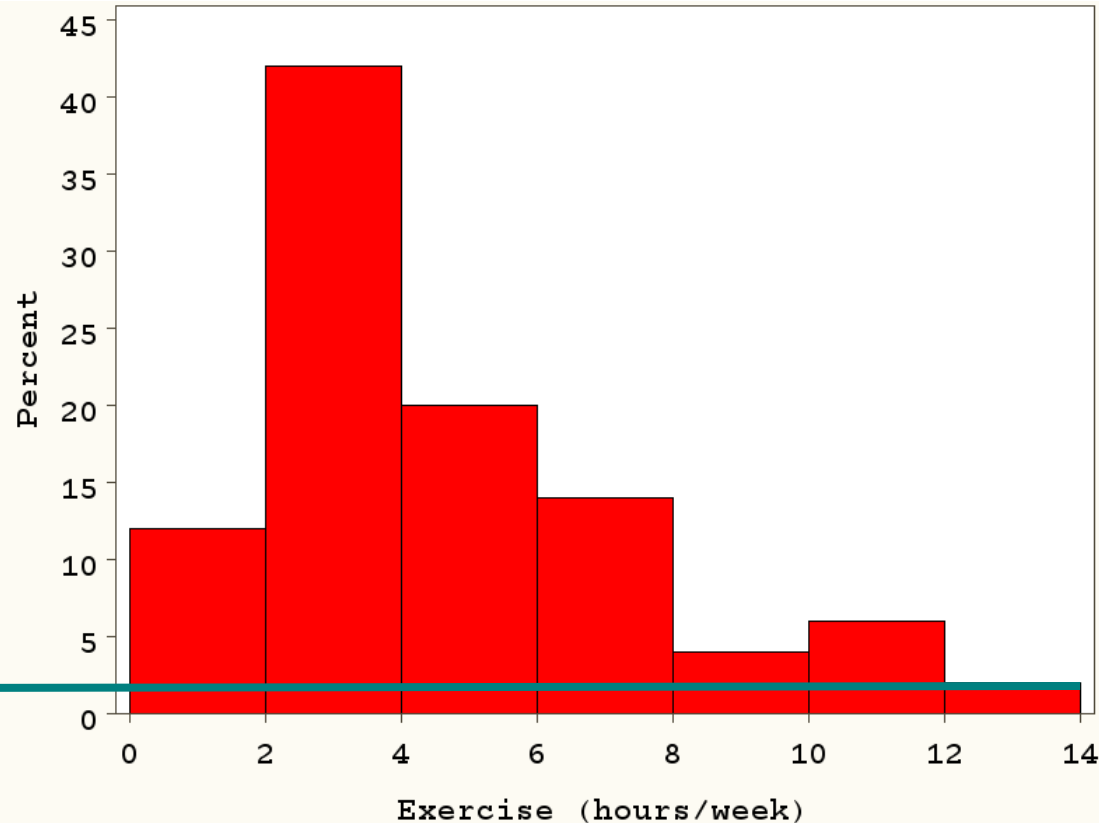
42% of students (n=21) exercise between 2 and 3.999... hours per week.

12% of students (n=6) exercise between 0 and 1.999... hours per week.



Histogram of Exercise

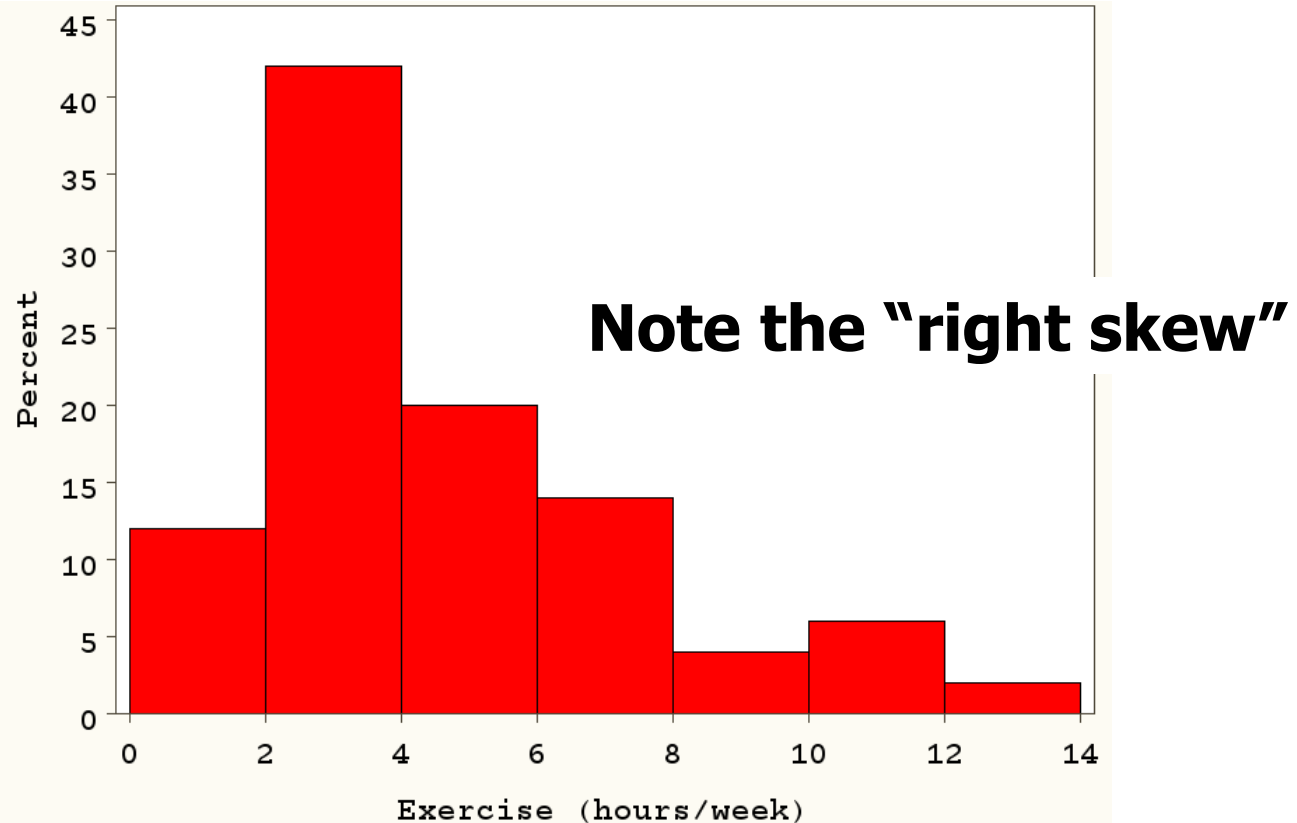
Bins of size = 2 hours/week



**2% of students
(n=1) exercise
 ≥ 12 hr/wk**

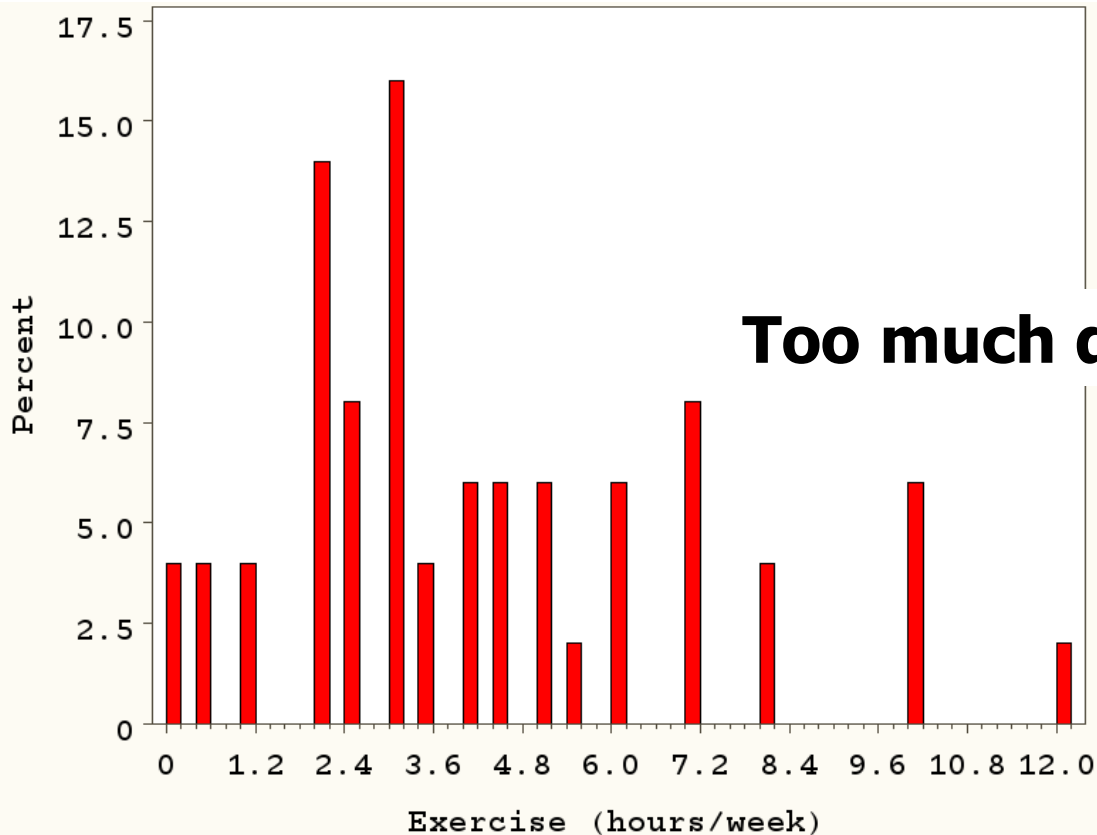
Histogram of Exercise

Bins of size = 2 hours/week



Histogram of Exercise

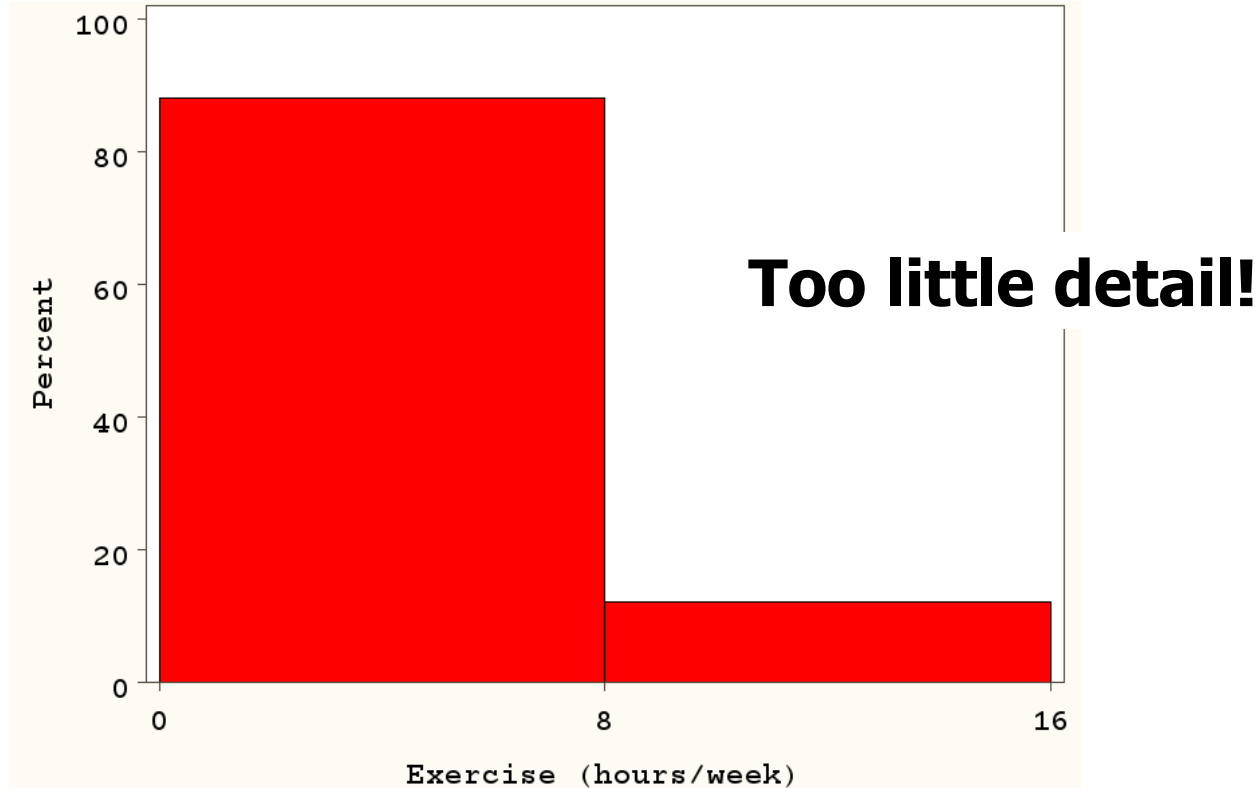
Bins of size = 0.2 hours/week



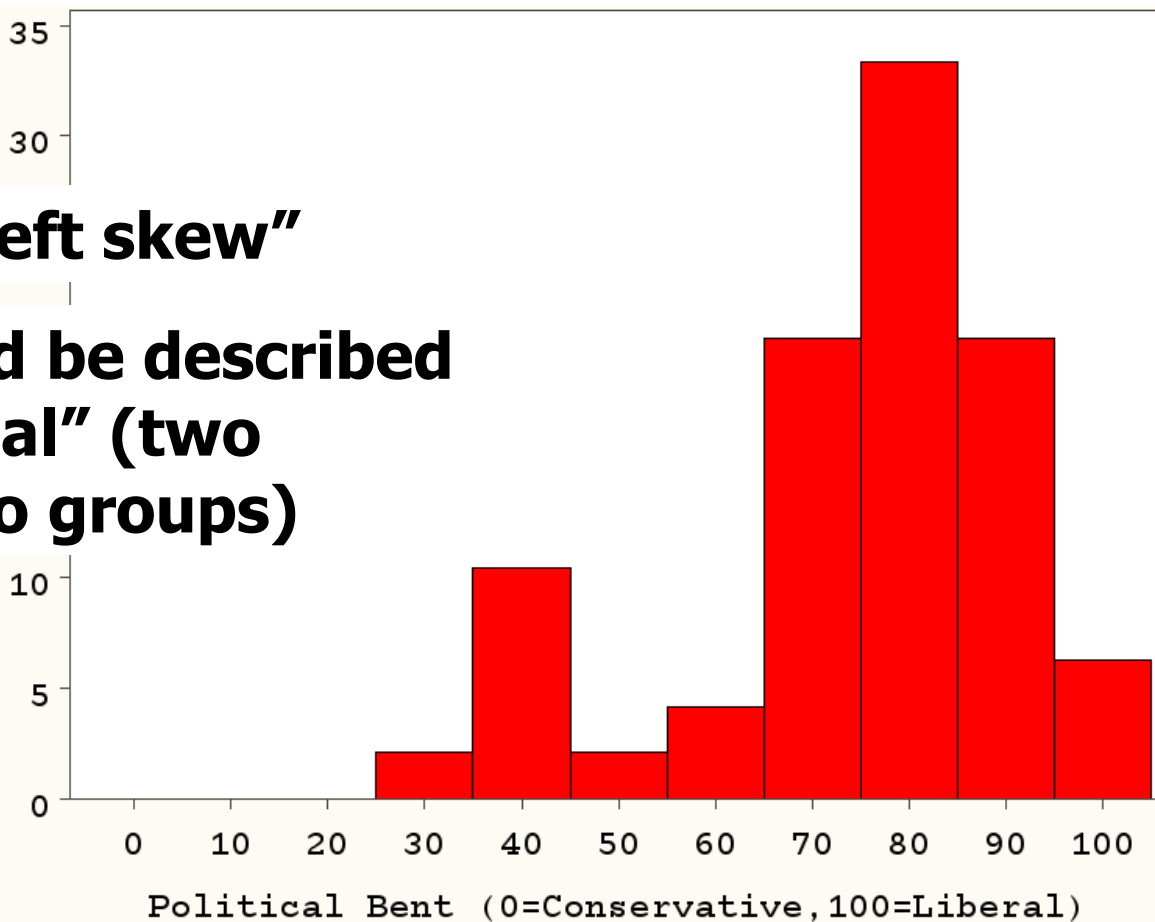
Too much detail!

Histogram of Exercise

Bins of size = 8 hours/week



Histogram of Political Bent



Note the "left skew"

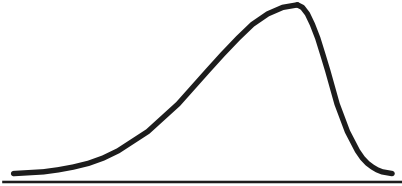
**Also, could be described
as "bimodal" (two
peaks, two groups)**



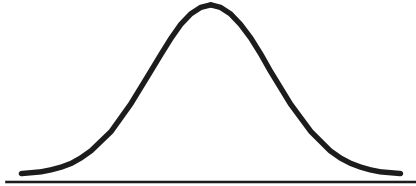
Shape of a Distribution

- Left-skewed/right-skewed/symmetric

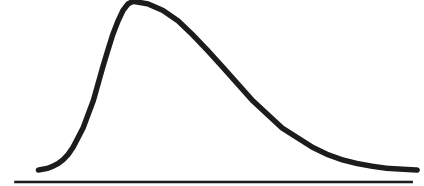
Left-Skewed



Symmetric



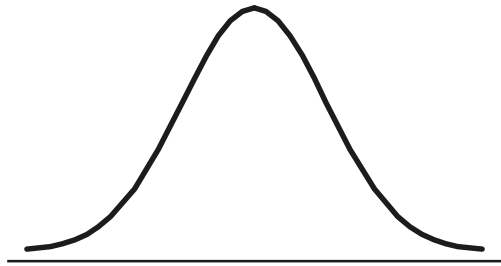
Right-Skewed



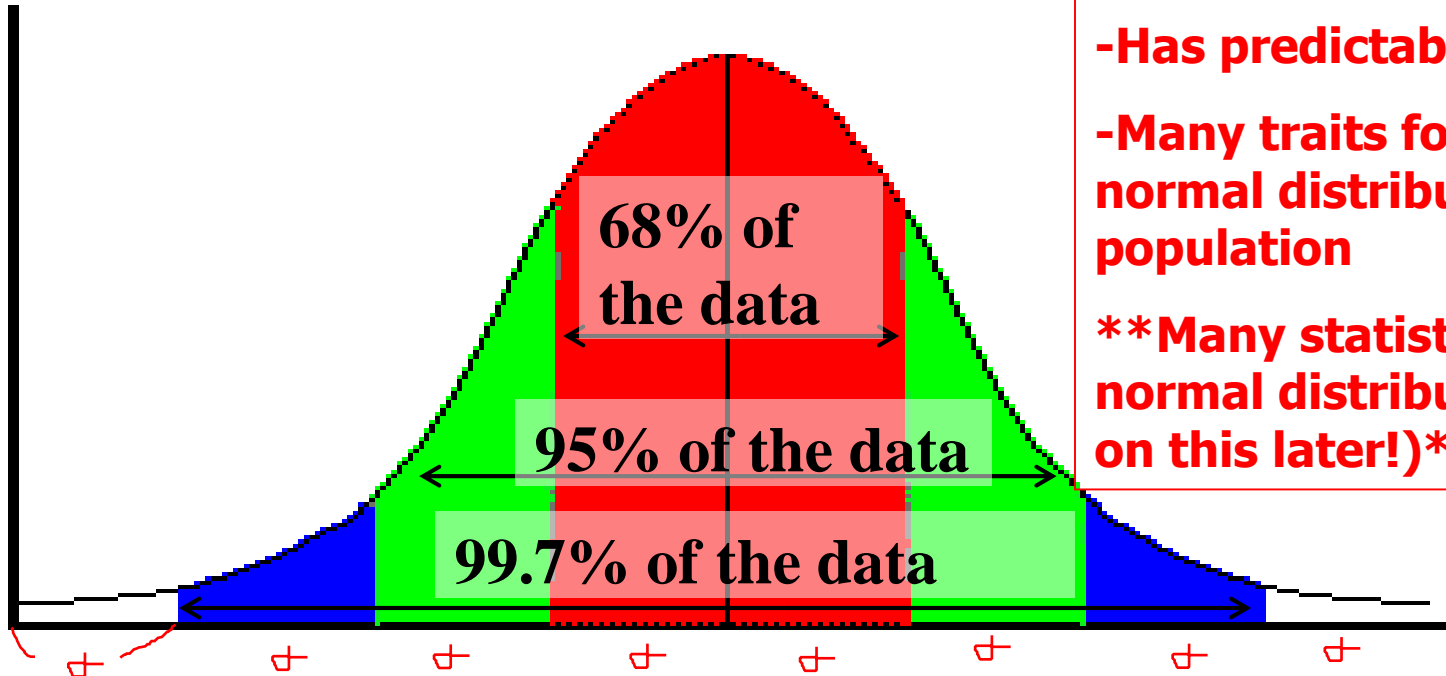


Shape of a Distribution

- Symmetric
 - Bell curve (Gaussian “normal distribution”)



Normal distribution (bell curve)

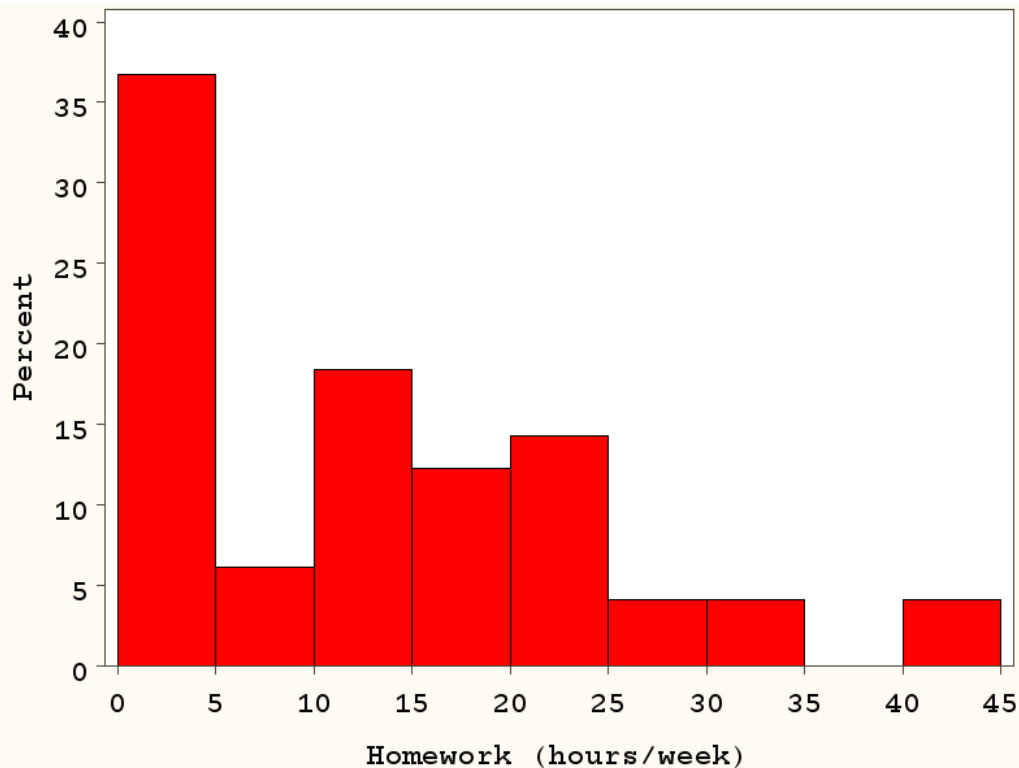


Useful for many reasons:

- Has predictable behavior
- Many traits follow a normal distribution in the population

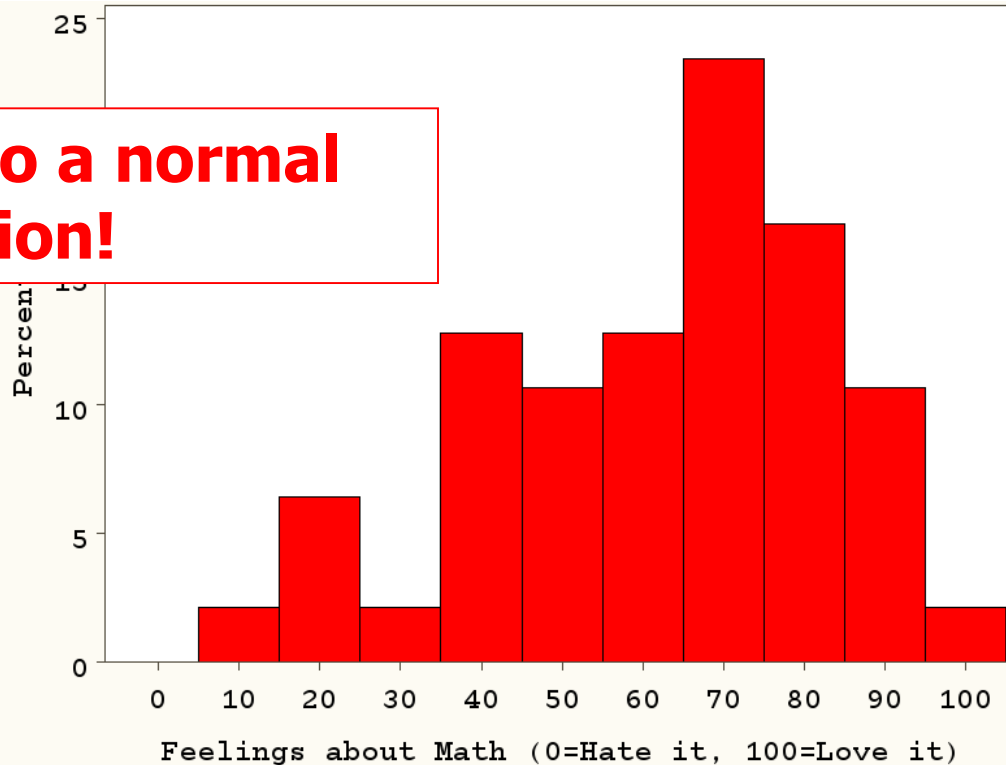
****Many statistics follow a normal distribution (more on this later!)****

Homework (hours/week)...



Feelings about math (0=lowest, 100=highest)

Closest to a normal distribution!





Statistics for Health Care

Describing Quantitative Data:
Where is the center?



Measures of “central tendency”

- Mean
- Median



Mean

- Mean – the average; the balancing point

calculation: the sum of values divided by the sample size

In math
shorthand:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



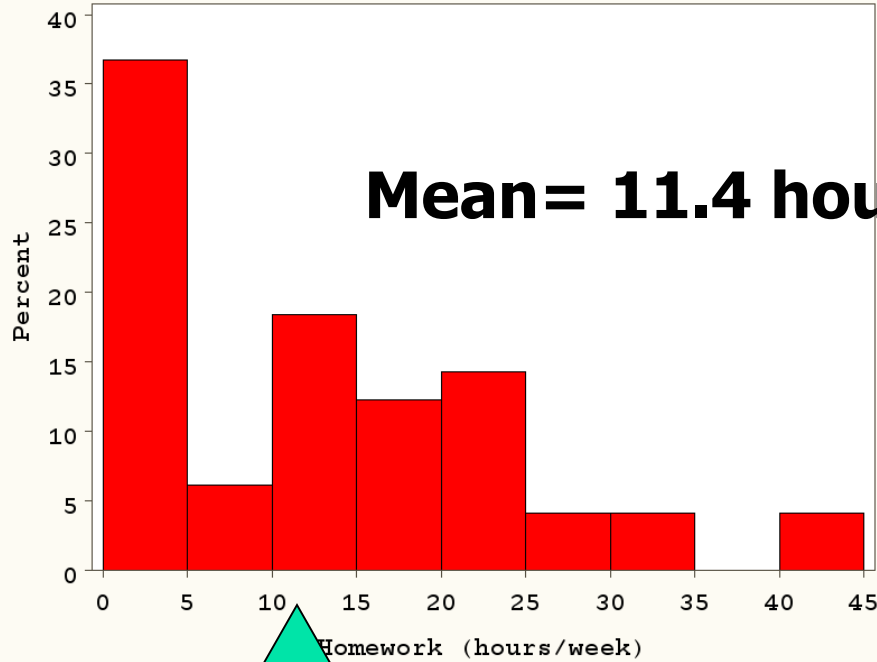
Mean: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

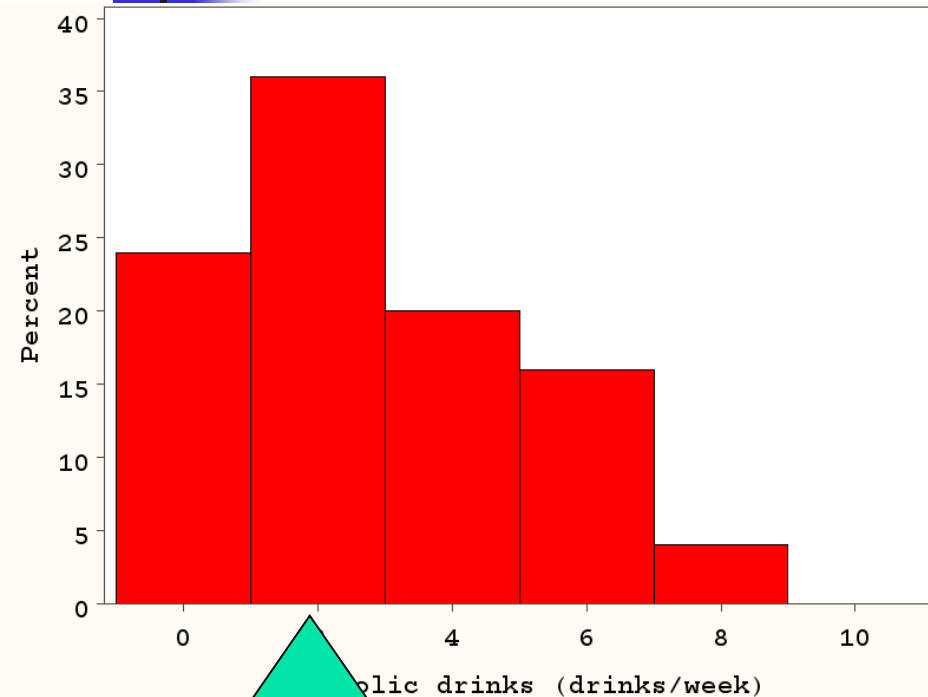
$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{17 + 19 + 21 + 22 + 23 + 23 + 23 + 38}{8} = 23.25$$

Mean of homework



The balancing point

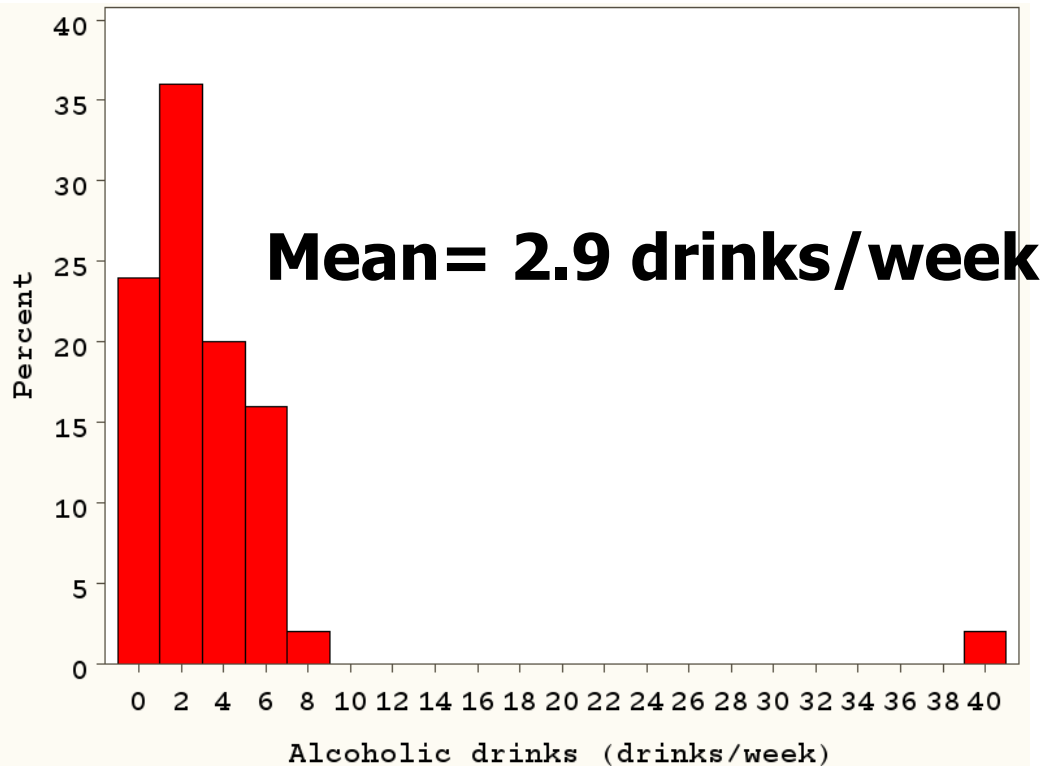
The mean is affected by extreme values...



Mean = 2.3 drinks/week

The balancing point

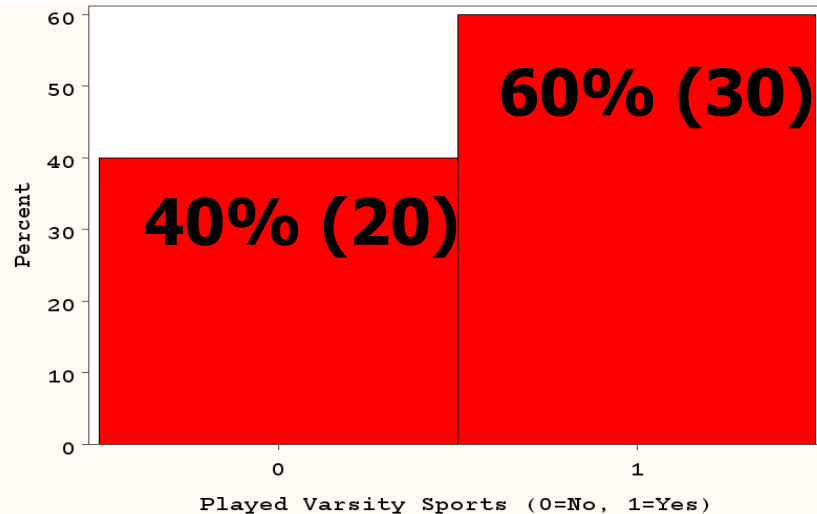
The mean is affected by extreme values...



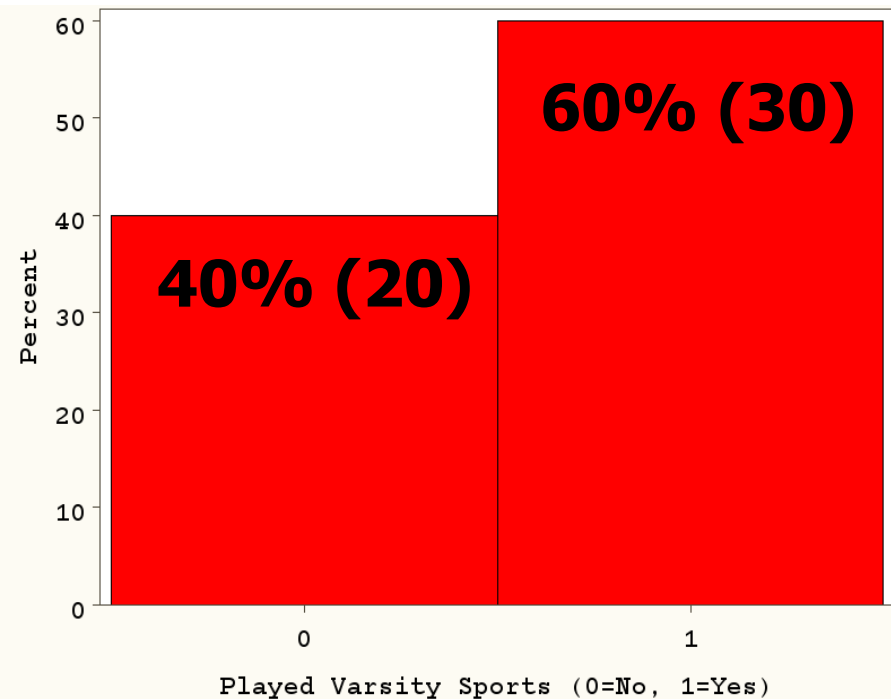
Does a binary variable have a mean?

Yes! If coded as a 0/1 variable...

**Example: Played Varsity Sports in High School
(0=no, 1=yes)**



Does a binary variable have a mean?



$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} = \frac{30 * 1 + 20 * 0}{50} = \frac{30}{50} = .60$$



Central Tendency

- Median – the exact middle value

Calculation:

- If there are an odd number of observations, find the middle value
- If there are an even number of observations, find the middle two values and average them.



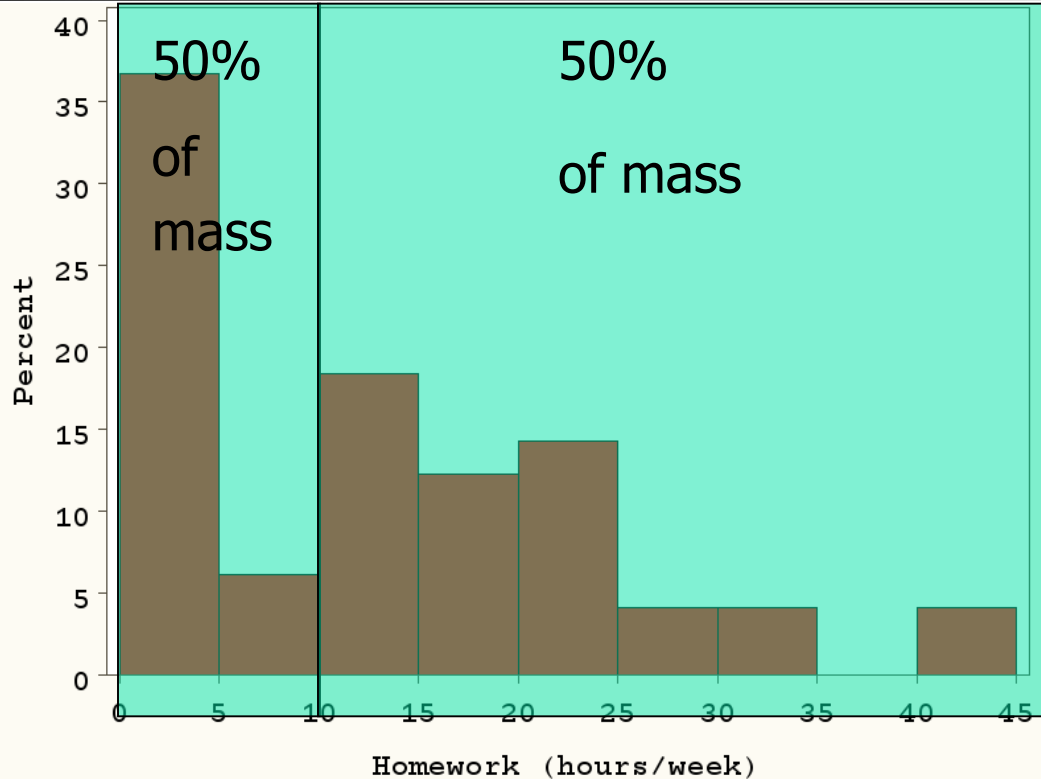
Median: example

Some data:

Age of participants: 17 19 21 22 23 23 23 38

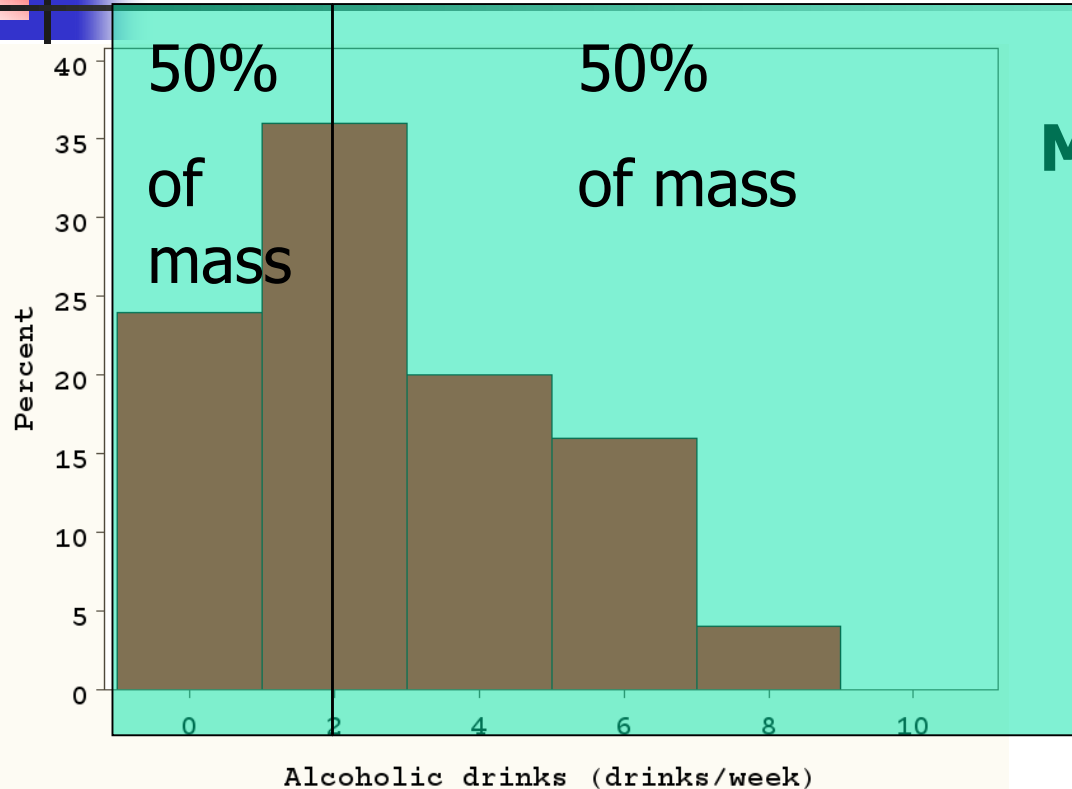
$$\text{Median} = (22+23)/2 = 22.5$$

Median of homework



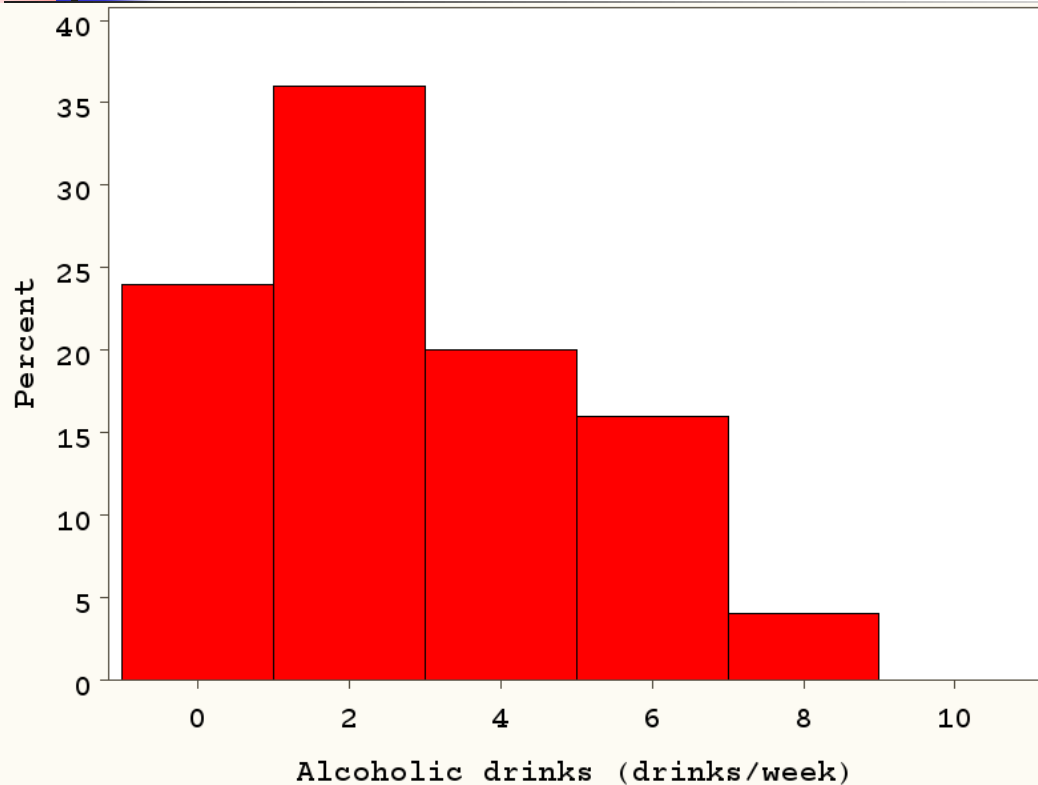
**Median = 10
hours/week**

Median of alcohol drinking

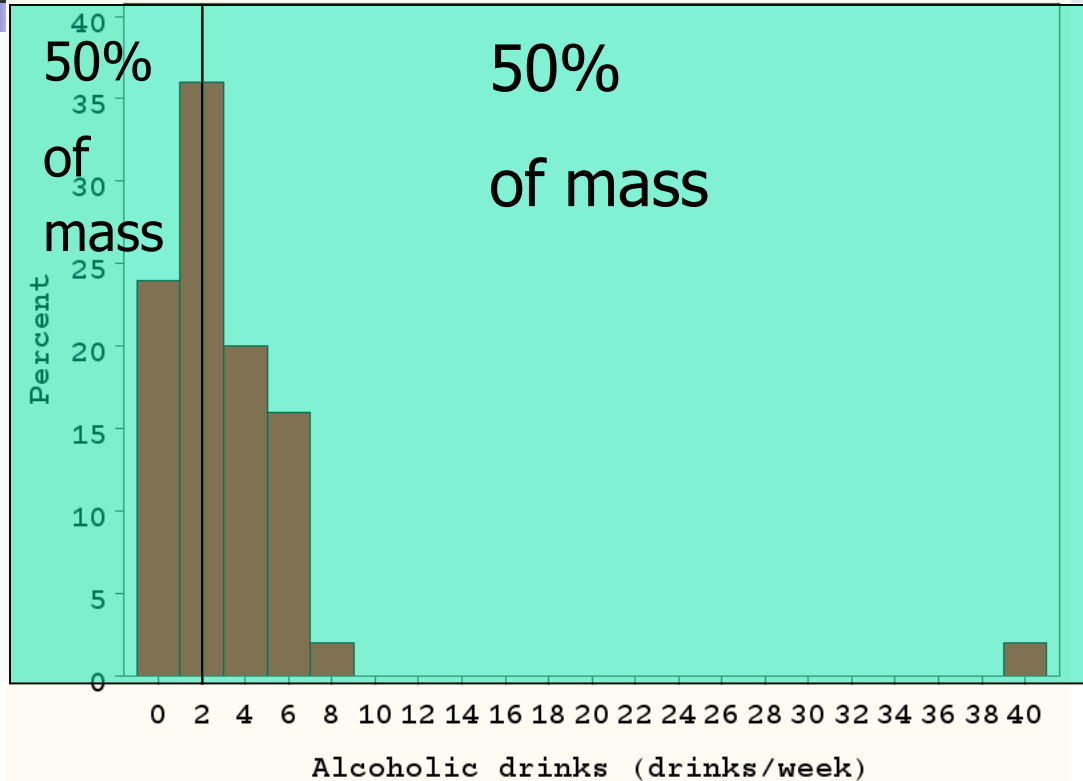


Median = 2.0 drinks/wk

The median is NOT affected by extreme values...



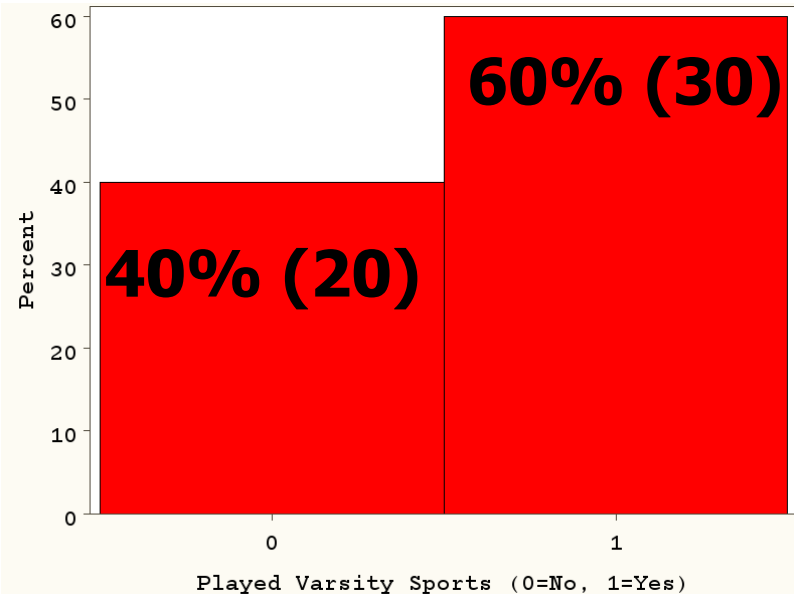
The median is NOT affected by extreme values...



**Median = 2.0
drinks/week**

Does Varsity Sports (binary variable) have a median?

- Yes, if you line up the 0's and 1's, the middle number is 1.






Should I present means or medians?

- For skewed data, the median is preferred because the mean can be highly misleading...

Hypothetical example: means vs. medians...



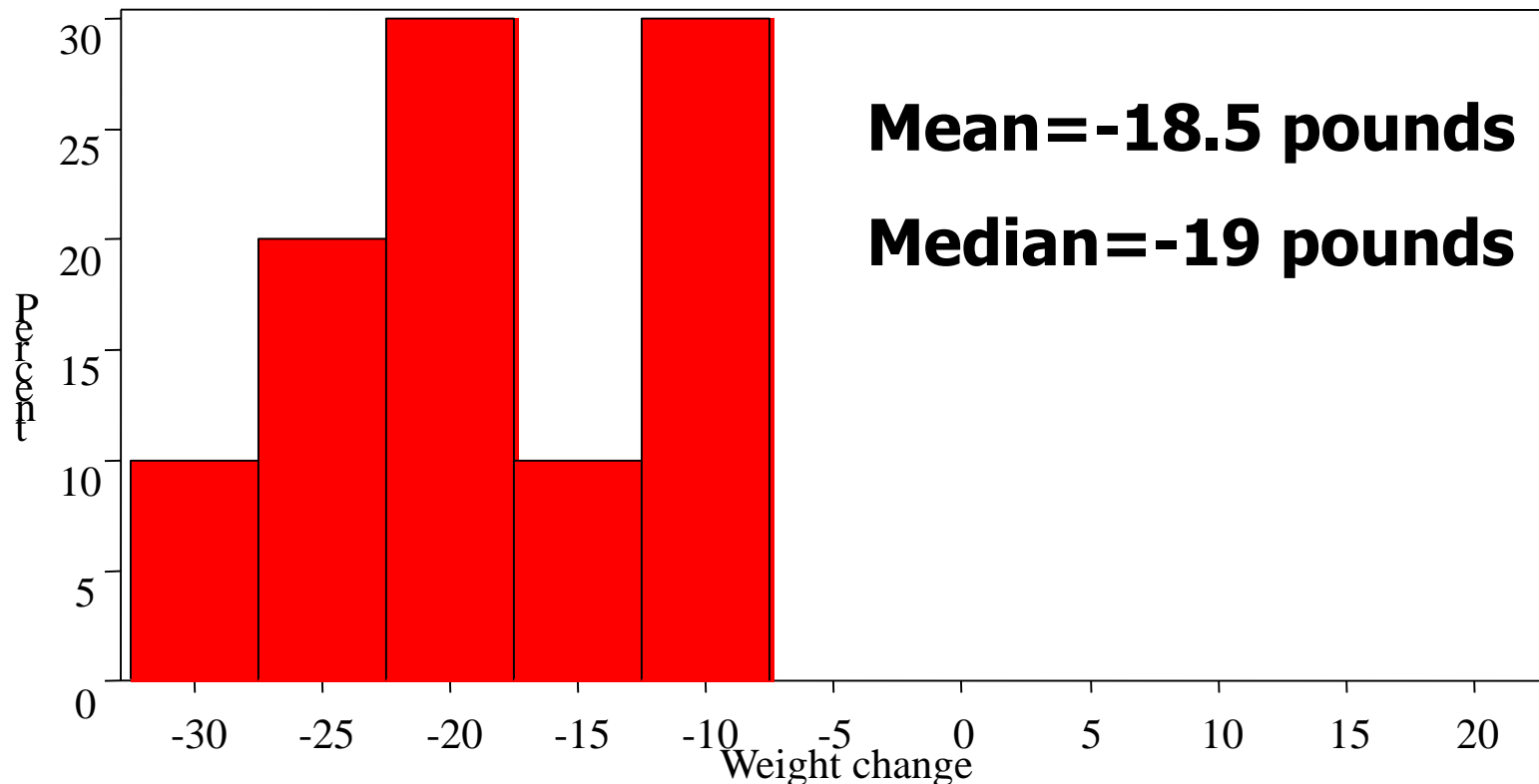
10 dieters following diet 1 vs. 10 dieters following diet 2

Group 1 ($n=10$) loses an average of 34.5 lbs.

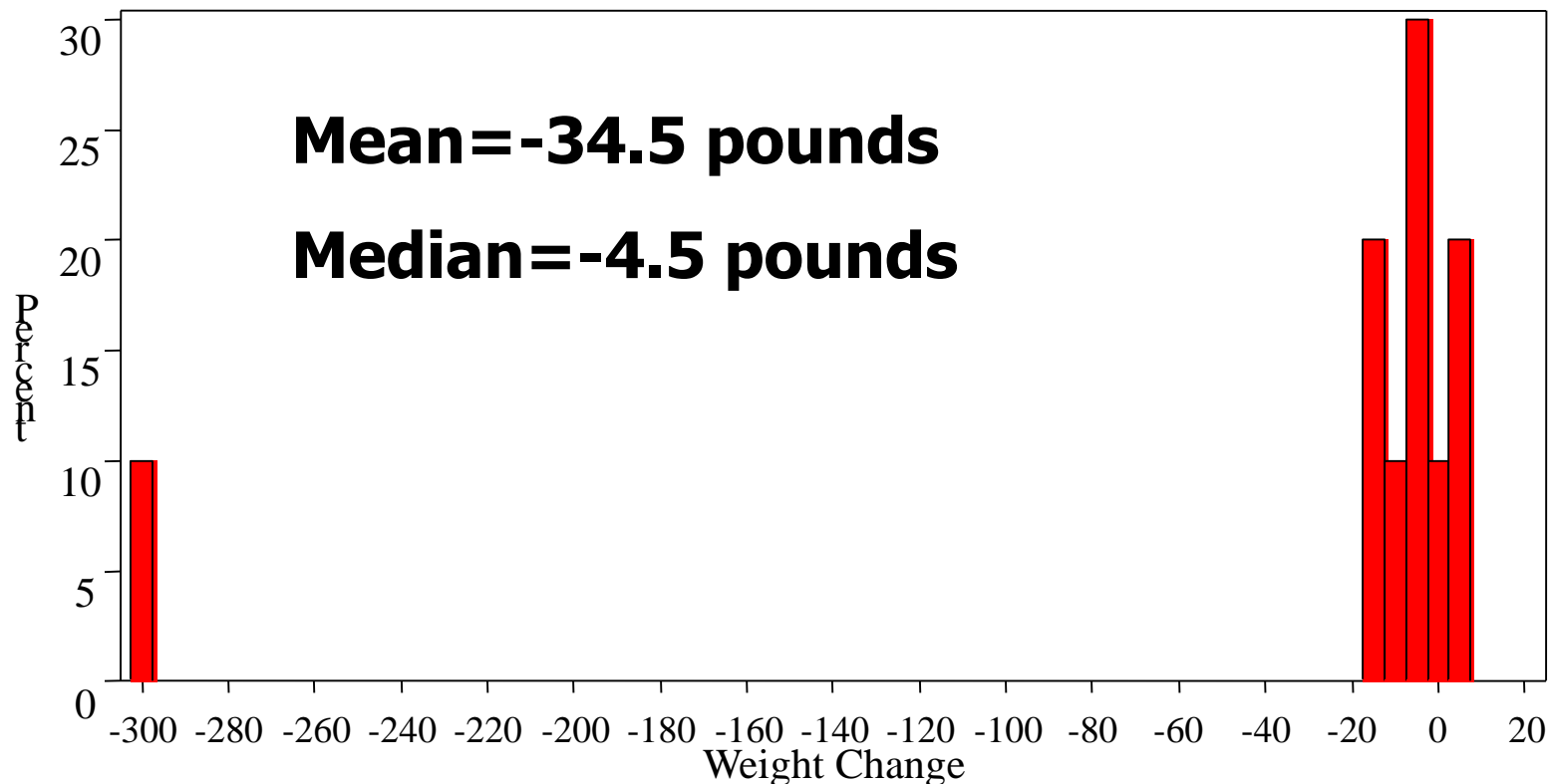
Group 2 ($n=10$) loses an average of 18.5 lbs.

Conclusion: diet 1 is better?

Histogram, diet 2...



Histogram, diet 1...





The data...

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30



Compare medians via a “non-parametric test”

We need to compare medians (ranked data) rather than means; requires a “non-parametric test”

Apply the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) as follows...



Rank the data...

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Ranks: 1 2 3 4 5 6 9 11 12 20

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30

Ranks: 7 8 10 13 14 15 16 17 18 19



Sum the ranks...

Diet 1, change in weight (lbs):

+4, +3, 0, -3, -4, -5, -11, -14, -15, -300

Ranks: 1 2 3 4 5 6 9 11 12 20

Sum of the ranks: $1+2+3+4+5+6+9+11+12+20 = 73$

Diet 2, change in weight (lbs)

-8, -10, -12, -16, -18, -20, -21, -24, -26, -30

Ranks: 7 8 10 13 14 15 16 17 18 19

Sum of the ranks: $7+8+10+13+14+15+16+17+18+19 = 137$

**Diet 2 is
superior to
Diet 1,
 $p=.018$.**



Statistics for Health Care

Module 5:

Describing Quantitative Data:
What is the variability in the data?



Measures of Variability

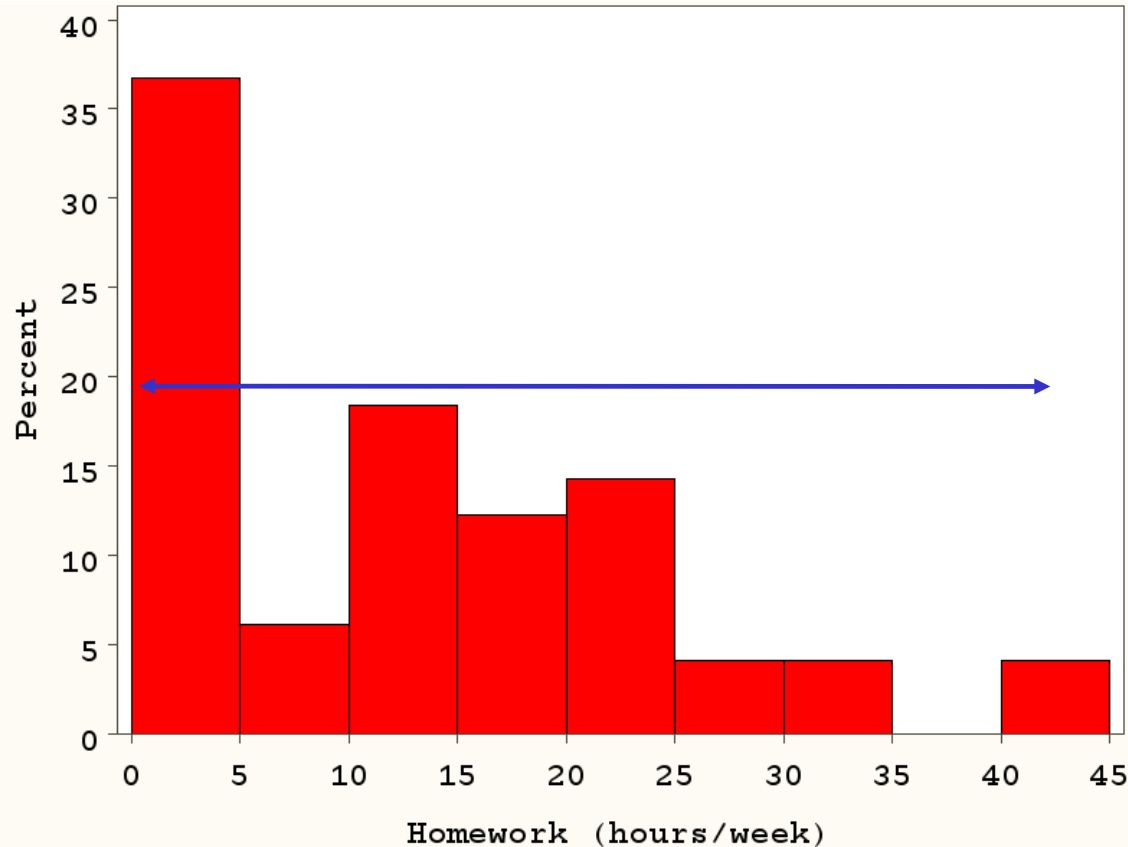
- Range
- Standard deviation/Variance
- Percentiles
- Inter-quartile range (IQR)



Range

- Difference between the largest and the smallest observations.
-

Range of homework: 40 hours – 0 hours = 40 hours/wk





Standard deviation

- Challenge: devise a statistic that gives the average distance from the mean.

- Distance from the mean: $x_i - \bar{X}$

- Average distance from the mean??:
$$\frac{\sum_i^n (x_i - \bar{X})}{n} ?$$



Standard deviation

- But this won't work!

$$\frac{\sum_i^n (x_i - \bar{X})}{n} = 0$$



How can I get rid of negatives?

- Absolute values?
 - Too messy mathematically!
- Squaring eliminates negatives!

$$S^2 = \frac{\sum_i^n (x_i - \bar{X})^2}{n}$$



Variance

- Average squared distance from the mean:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n-1}$$

degree of freedom

We lose a “degree of freedom because we have already estimated the mean.



Standard Deviation

- Gets back to the units of the original data
- Roughly, the average spread around the mean.

$$S = \sqrt{\frac{\sum_i^n (x_i - \bar{X})^2}{n-1}}$$



The standard deviation is affected by extreme values

- Because of the squaring, values farther from the mean contribute more to the standard deviation than values closer to the mean:

$$\bar{X} = 5$$

$$(6 - 5)^2 = 1$$

$$(10 - 5)^2 = 25$$



Calculation Example: Standard Deviation

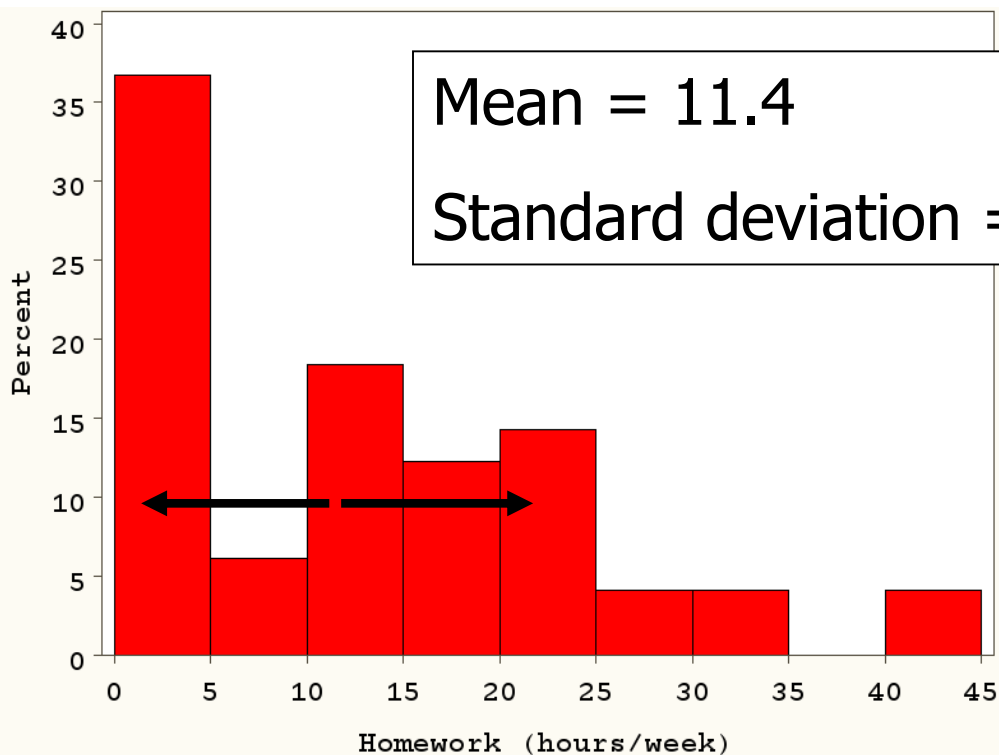
Age data (n=8) : 17 19 21 22 23 23 23 38

$n = 8$

Mean = 23.25

$$S = \sqrt{\frac{(17 - 23.25)^2 + (19 - 23.25)^2 + \dots + (38 - 23.25)^2}{8 - 1}}$$
$$= \sqrt{\frac{280}{7}} = 6.3$$

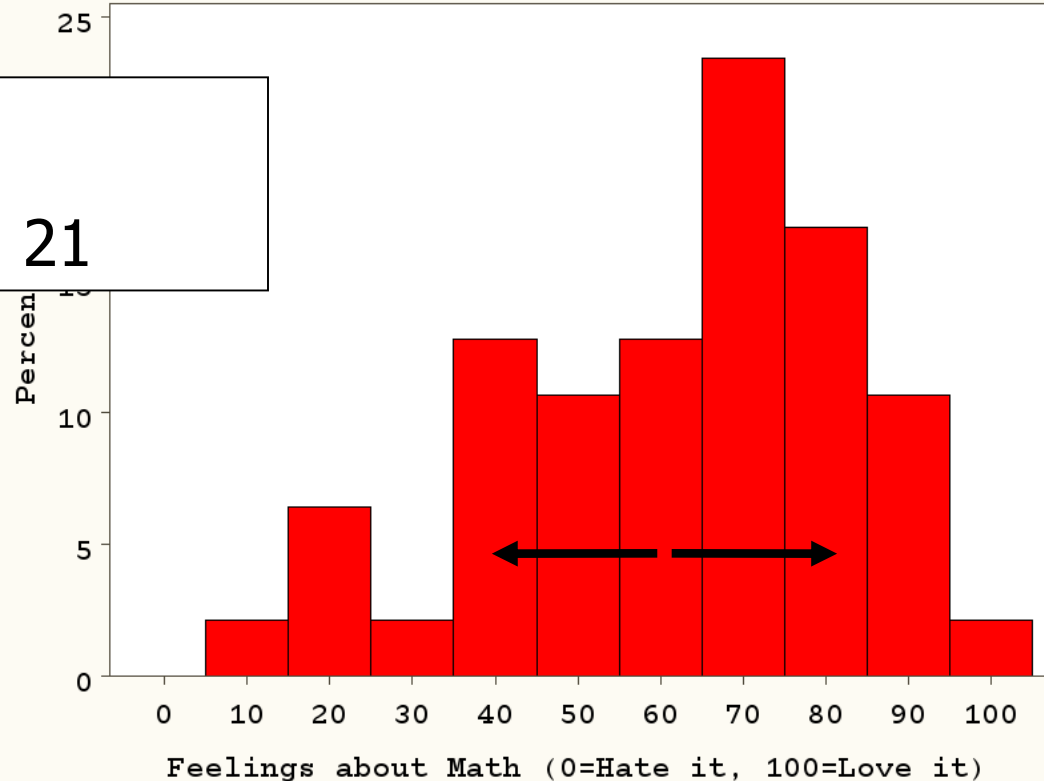
Homework (hours/week)



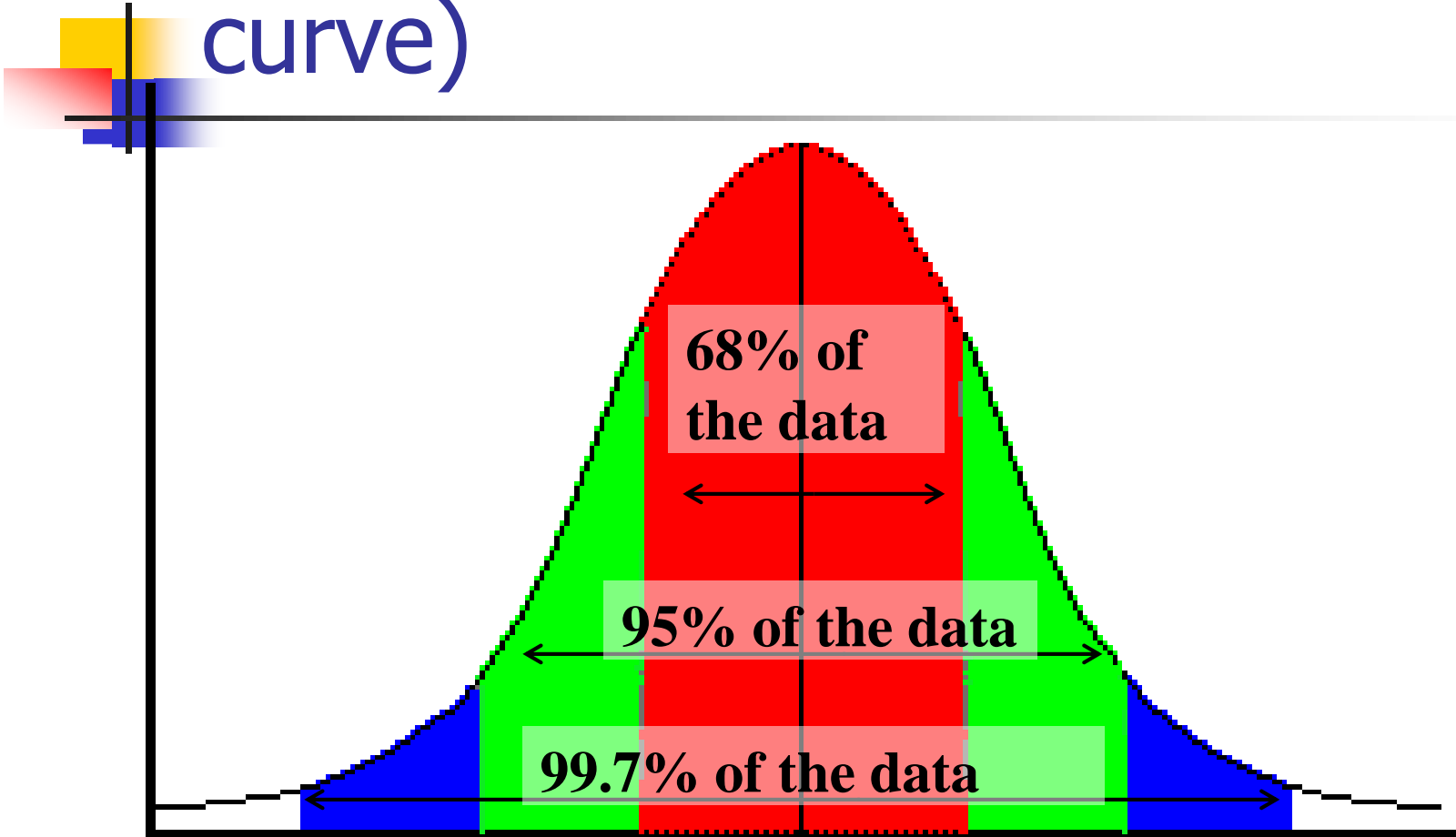
Feelings about math (0=lowest, 100=highest)

Mean = 61

Standard deviation = 21



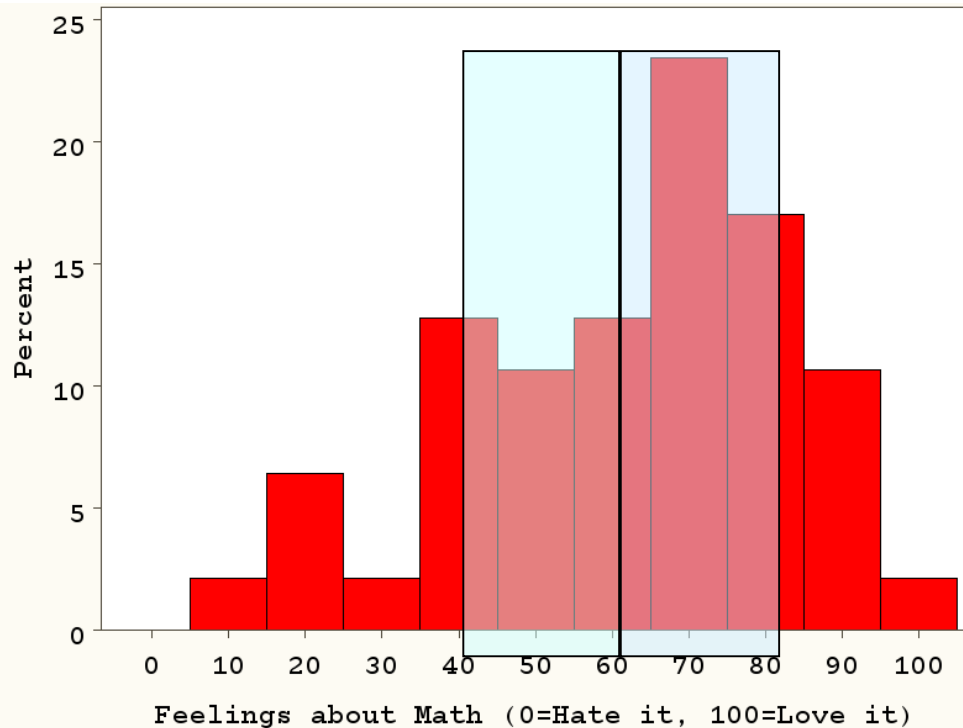
68-95-99.7 rule (for a perfect bell curve)



Feelings about math (0=lowest, 100=highest)

Mean +/- 1 std =
40 – 82

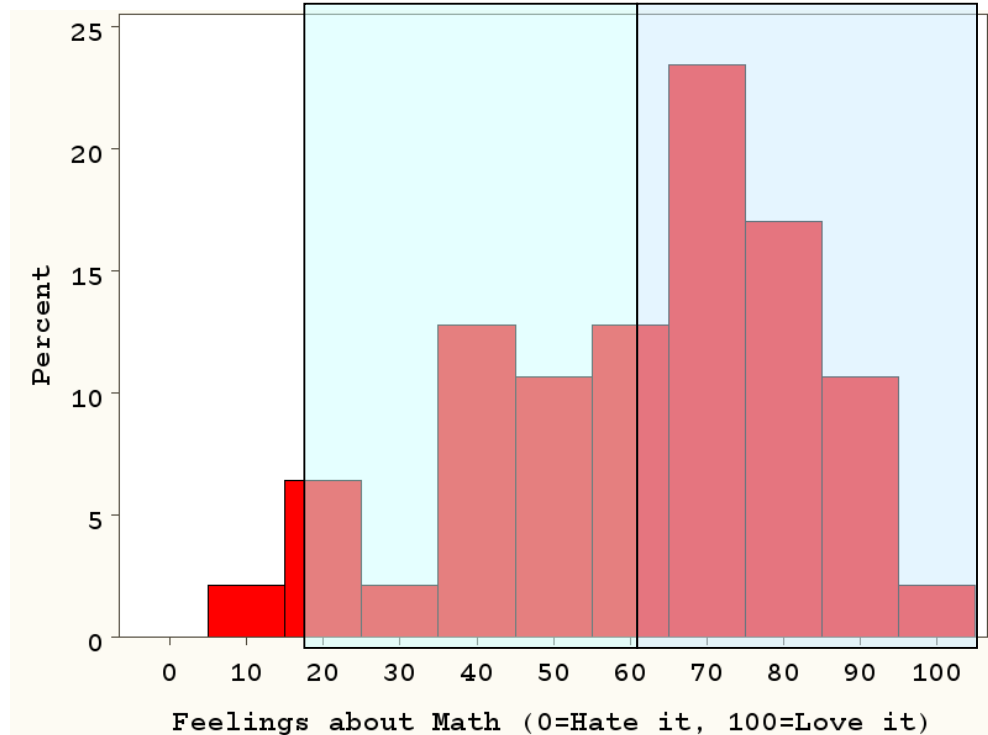
Percent
between 40
and 82 =
 $34/47 = 72\%$



Feelings about math (0=lowest, 100=highest)

Mean +/- 2 std =
19 – 100

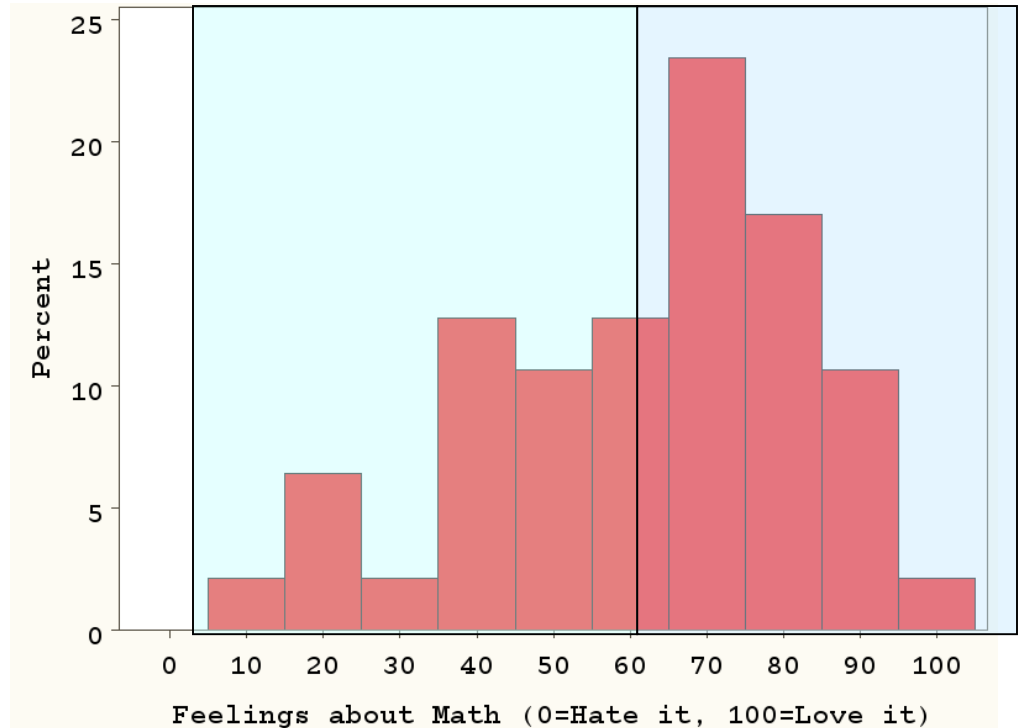
Percent
between 19
and 100 =
 $46/47 = 98\%$



Feelings about math (0=lowest, 100=highest)

Mean +/- 3 std =
0 – 100

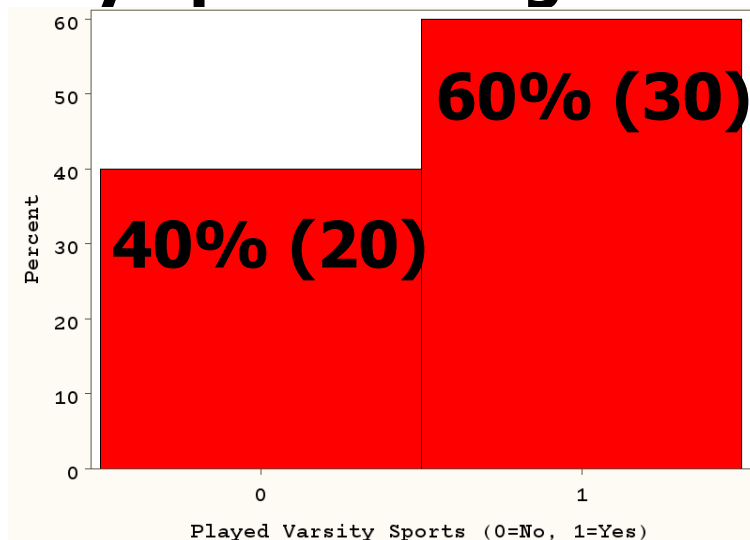
100% of the
data!



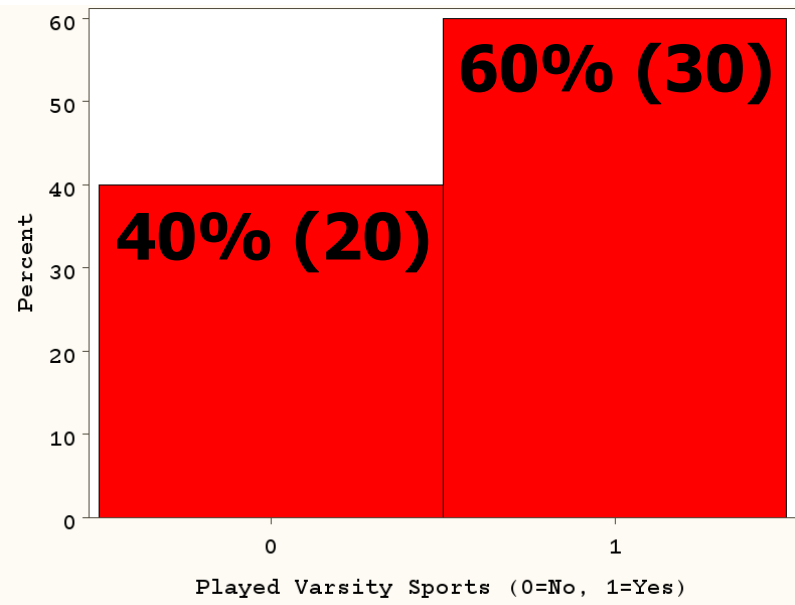
Does a binary variable have a standard deviation?

Yes! If coded as a 0/1 variable...

**Example: Played Varsity Sports in High School
(0=no, 1=yes)**



Does a binary variable have a standard deviation?

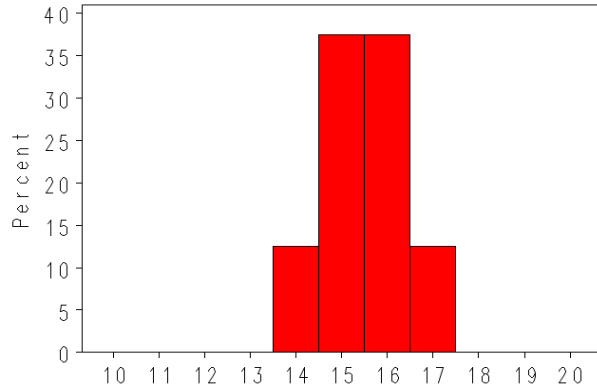


$$S = \sqrt{\frac{30 * (1 - .60)^2 + 20 * (0 - .60)^2}{50 - 1}}$$
$$= \sqrt{\frac{30(.16) + 20(.36)}{49}} = \sqrt{\frac{12}{49}} = .49$$

Understanding Standard Deviation:

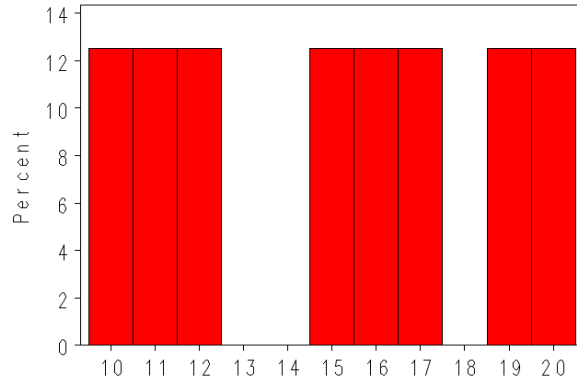
Mean = 15
S = 0.9

data=B



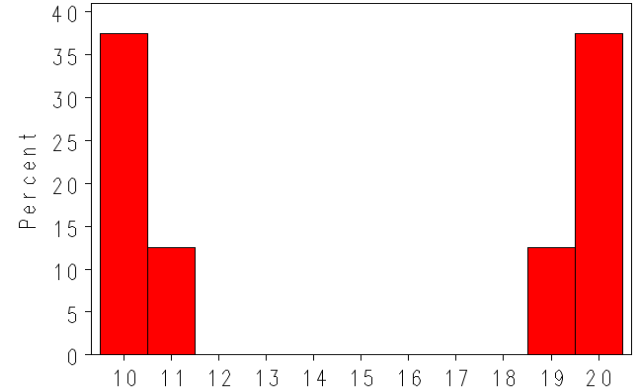
Mean = 15
S = 3.7

data=A



Mean = 15
S = 5.1

data=C



Standard deviations vs. standard errors



- Standard deviation measures the variability of a trait.
- Standard error measures *the variability of a statistic*, which is a theoretical construct! (much more on this later!)



Percentiles

- Based on ranking the data
 - The 90th percentile is the value for which 90% of observations are lower
 - The 50th percentile is the median
 - The 10th percentile is the value for which 10% of observations are lower
- Percentiles are not affected by extreme values (unlike standard deviations)

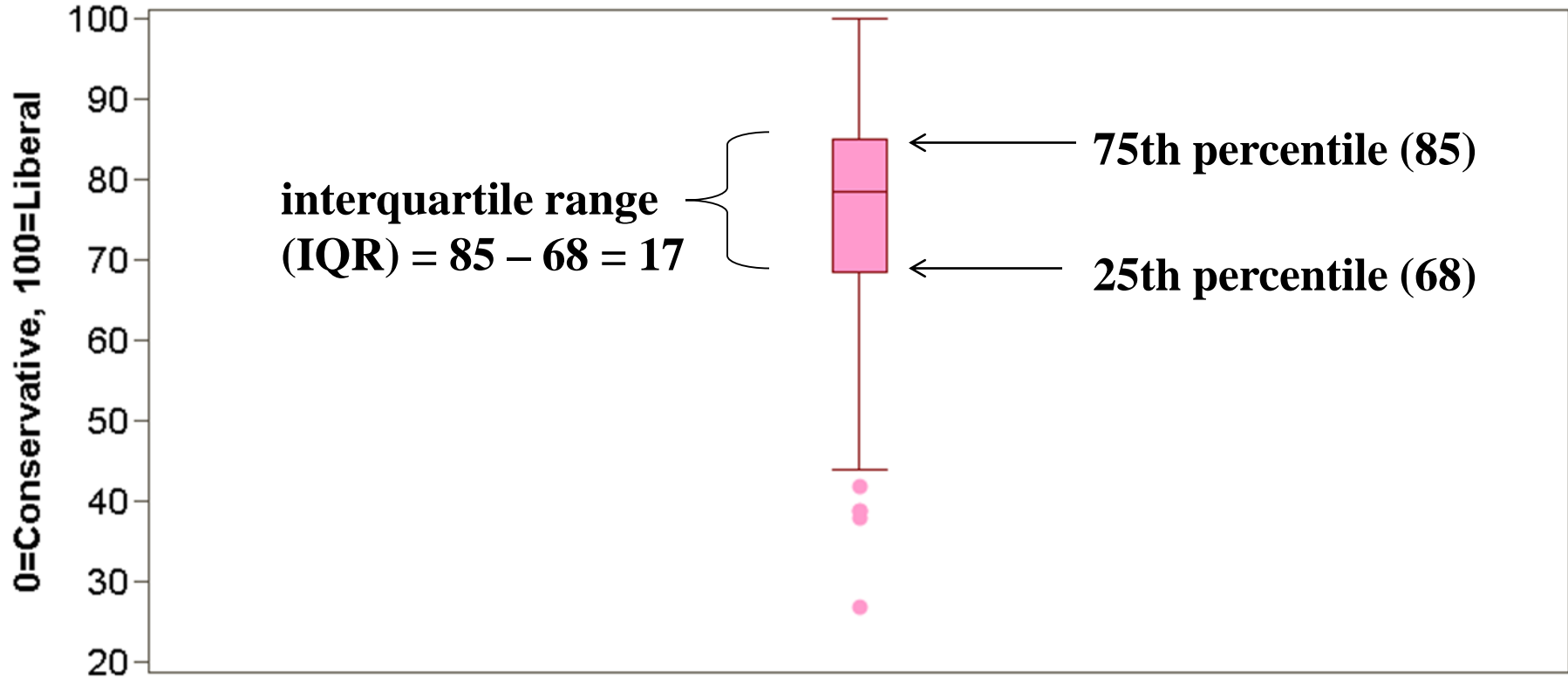


Interquartile Range (IQR)

- Interquartile range = 3rd quartile – 1st quartile
- The middle 50% of the data.
- Interquartile range is not affected by outliers.

Boxplot of Political Bent

(0=Most Conservative, 100=Most Liberal)





Symbols

- S^2 = Sample variance
- S = Sample standard deviation
- σ^2 = Population (true or theoretical) variance
- $\underline{\sigma}$ = Population standard deviation
- \bar{X} = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)



Statistics for Health Care

Module 6:

Exploring real data: Lead in lipstick



2007 Headlines

- “Lipsticks Contain Excessive Lead, Tests Reveal”
- “One third of lipsticks on the market contain high lead”

Link to example news coverage:

<http://www.reuters.com/article/2007/10/11/us-lipstick-lead-idUSN1140964520071011>



2007 report by a consumer advocacy group...

- “One-third of the lipsticks tested contained an amount of lead that exceeded the U.S. Food and Drug Administration’s 0.1 ppm limit for lead in candy—a standard established to protect children from ingesting lead.”



2007 report by a consumer advocacy group...

- “One-third of the lipsticks tested contained an amount of lead that **exceeded the U.S. Food and Drug Administration’s 0.1 ppm limit for lead in candy**—a standard established to protect children from ingesting lead.”

1 ppm = 1 part per million = 1 microgram/gram



Recent Headlines

- “400 shades of lipstick found to contain lead”, FDA says” *Washington Post*, Feb. 14, 2012
- “What’s in Your Lipstick? FDA Finds Lead in 400 Shades,” *Time* February 15, 2012

Link to example news coverage:

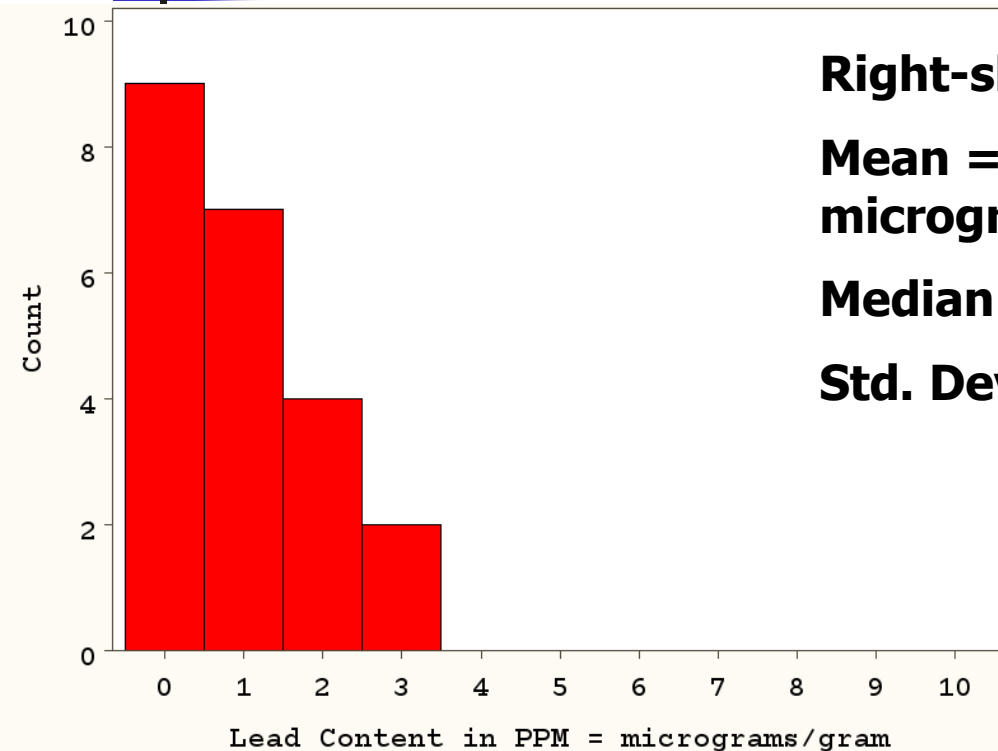
<http://healthland.time.com/2012/02/15/whats-in-your-lipstick-fda-finds-lead-in-400-shades/>



How worried should women be?

- What is the dose of lead in lipstick?
- How much lipstick are women exposed to?
- How much lipstick do women ingest?

Distribution of lead in lipstick (FDA 2009, n=22)



Right-skewed!

**Mean = 1.07
micrograms/gram**

Median = 0.73

Std. Dev = 0.96

max = 3.06

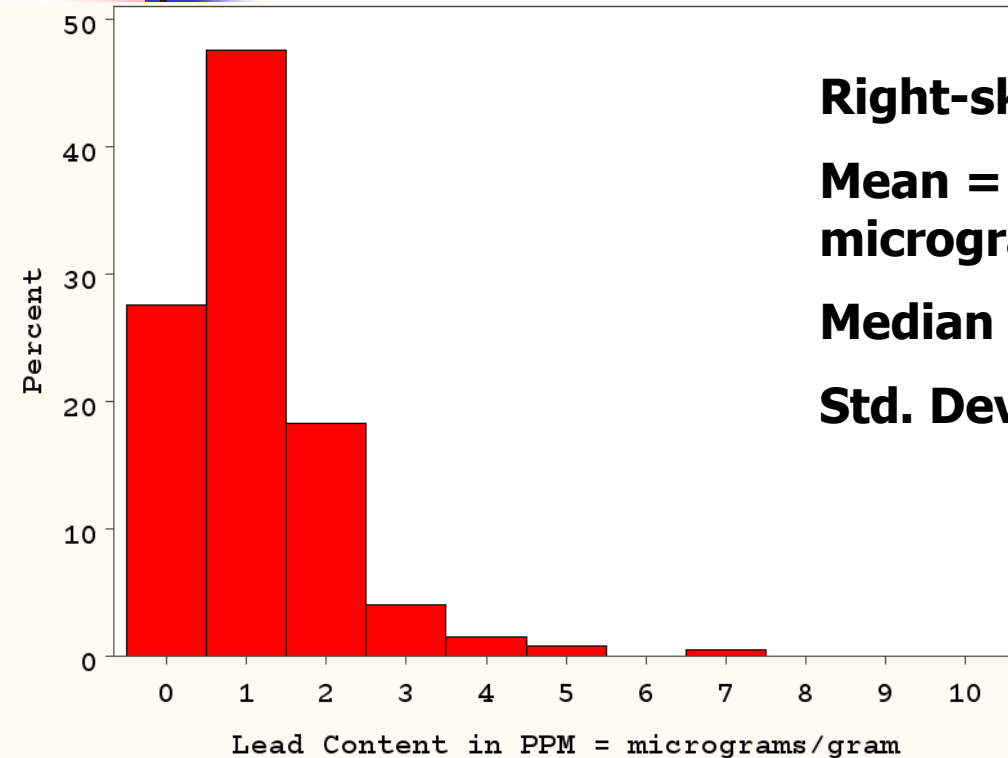
99th percentile : 3.06

95th percentile: 3.05

90th percentile: 2.38

75th percentile: 1.76

Distribution of lead in lipstick (FDA 2012, n=400)



Right-skewed!

**Mean = 1.11
micrograms/gram**

Median = 0.89

Std. Dev = 0.97

max = 7.19

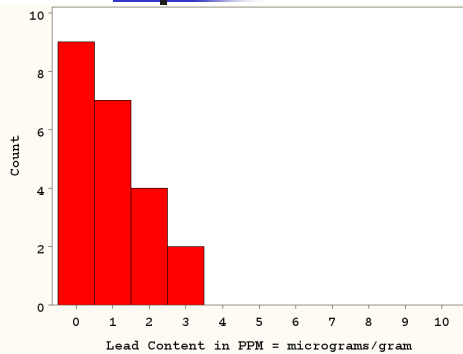
99th percentile : 4.91

95th percentile: 2.76

90th percentile: 2.23

75th percentile: 1.50

FDA 2009 (n=22) vs. FDA 2012 (n=400)



2009 (n=22)

**Mean = 1.07
micrograms/gram**

Median = 0.73

Std. Dev = 0.96

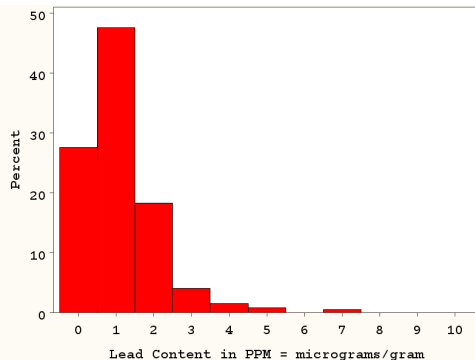
max = 3.06

99th percentile : 3.06

95th percentile: 3.05

90th percentile: 2.38

75th percentile: 1.76



2012 (n=400)

**Mean = 1.11
micrograms/gram**

Median = 0.89

Std. Dev = 0.97

max = 7.19 Did lead increase?

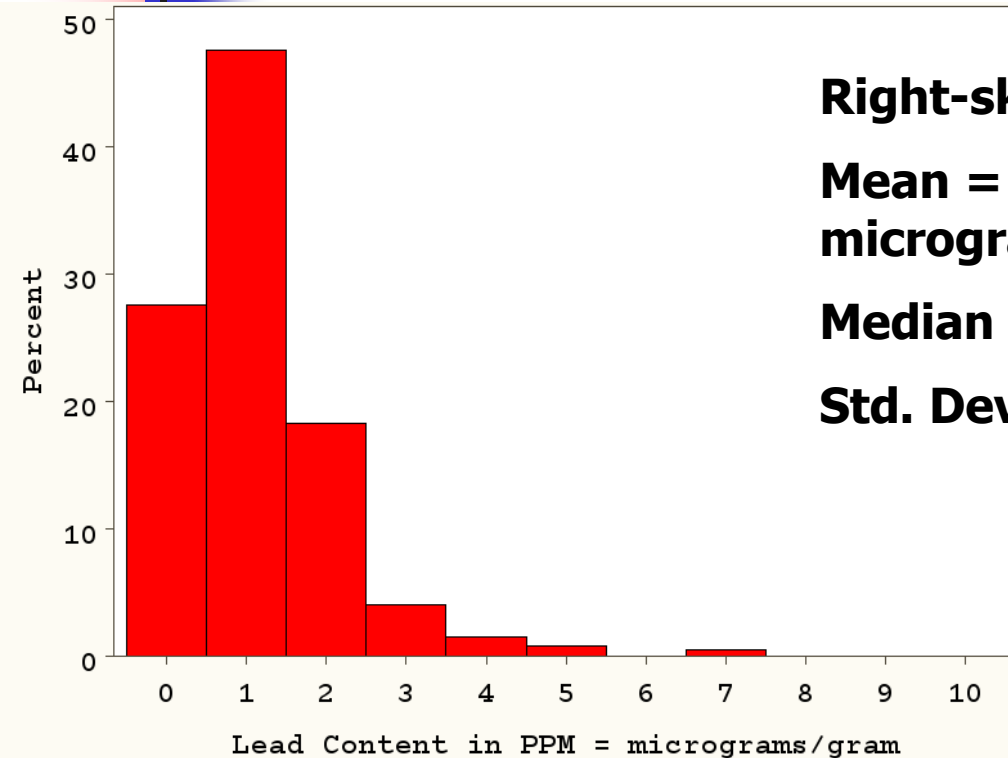
99th percentile : 4.91

95th percentile: 2.76

90th percentile: 2.23

75th percentile: 1.50

Distribution of lead in lipstick (FDA 2012, n=400)



Right-skewed!

**Mean = 1.11
micrograms/gram**

Median = 0.89

Std. Dev = 0.97

max = 7.19

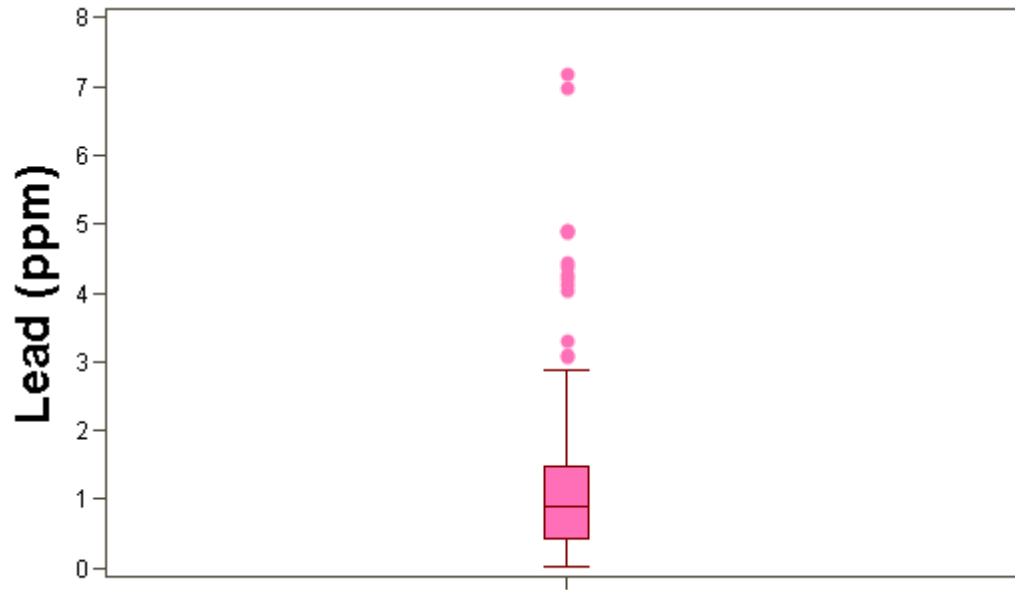
99th percentile : 4.91

95th percentile: 2.76

90th percentile: 2.23

75th percentile: 1.50

Distribution of lead in lipstick (n=400 samples, FDA 2012)

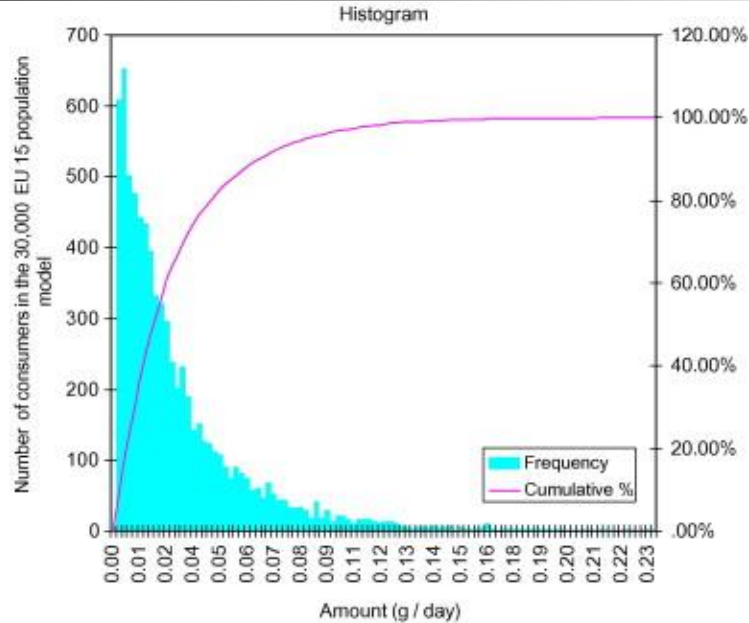


FDA data available at:

<http://www.fda.gov/Cosmetics/ProductandIngredientSafety/ProductInformation/ucm137224.htm#expanalyses>

Data on lipstick exposure

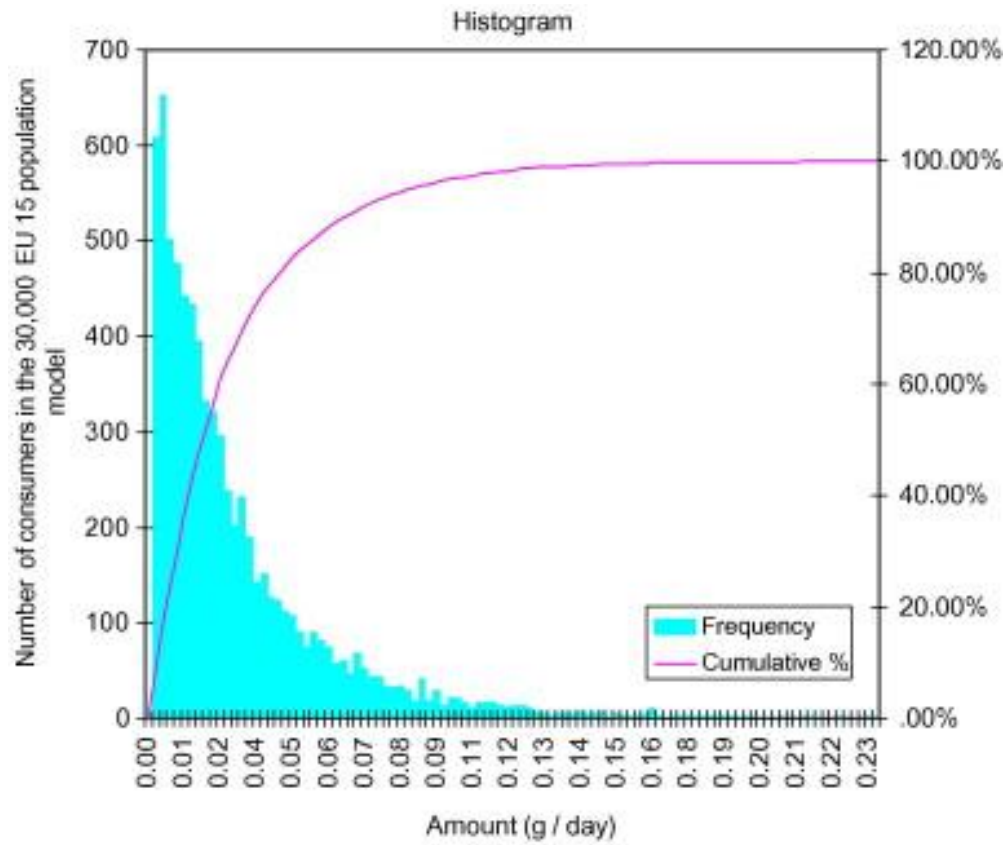
Fig. 6 Lipstick exposure for women in grams/day.



Value	Amount (mg / day)
mean	24.61
std	24.05
median	17.11
minimum	0.13
maximum	217.53
p01	0.57
p02.5	1.00
p05	1.68
p10	2.95
p20	5.69
p30	9.20
p40	12.93
p50	17.11
p60	22.37
p70	29.43
p80	39.70
p90	56.53
p92	61.66
p94	68.29
p95	72.51
p96	77.78
p97.5	89.08
p98	94.46
p99	110.98
p99.5	126.71
p99.9	160.06

Percentiles
in mg/day

Distribution of lipstick exposure:



Value	Amount (mg / day)
mean	24.61
std	24.05
median	17.11
minimum	0.13
maximum	217.53
p01	0.57
p02.5	1.00
p05	1.68
p10	2.95
p20	5.69
p30	9.20
p40	12.93
p50	17.11
p60	22.37
p70	29.43
p80	39.70
p90	56.53
p92	61.66
p94	68.29
p95	72.51
p96	77.78
p97.5	89.08
p98	94.46
p99	110.98
p99.5	126.71
p99.9	160.06

Percentiles
in mg/day



Highest use (1 in 30,000 women)

- 1 in 30,000 women uses 218 milligrams of lipstick per day.
- 1 tube of lipstick contains 4000 milligrams.
- $4000 \text{ mg/tube} \div 218 \text{ mg/day} = 18 \text{ days per tube.}$
- The heaviest user goes through an entire tube of lipstick in 18 days.



Exercise

Assuming that women ingest 50% of the lipstick they apply daily, calculate:

1. What is the typical lead exposure to lipstick for women, in micrograms (mcg) of lead (based on medians)?
2. What is the highest daily lead exposure to lipstick for women, in mcg of lead?

Lead in lipstick:

Median = 0.89
micrograms/gram

Maximum = 7.19 mcg/g

Daily lipstick usage:

Median = 17.11 milligrams

Maximum = 217.53 mg



Typical user

Daily exposure:

Daily ingestion:



Typical user

Daily exposure:

$$0.89 \text{ mcg/g} \times 17.11 \text{ mg} \times 1 \text{ g/1000 mg} =$$

0.0152 mcg

Daily ingestion:

$$0.0152 \text{ mcg/2} = \mathbf{0.0076 \text{ mcg}}$$



Highest user

Daily exposure:

$$7.19 \text{ mcg/g} \times 217.53 \text{ mg} \times 1 \text{ g/1000 mg} =$$

1.56 mcg

Daily ingestion:

$$1.56 \text{ mcg/2} = \mathbf{0.78 \text{ mcg}}$$

**Frequency of usage this high:
 $1/30,000 \times 1/400 =$
1 woman in 12 million**

To put these numbers in perspective:

PTDI

- **“Provisional tolerable daily intake” for an adult is 75 micrograms/day**
- **$0.0076 \text{ mcg} / 75 \text{ mcg} = 0.02\%$ of your PTDI**
- **$0.78 \text{ mcg} / 75 \text{ mcg} = 1\%$ of your PTDI (1 in 12 million women)**
- **Average American consumes 1 to 4 mcg of lead per day from food alone.**

US FDA report: Total Diet Study Statistics on Element Results. December 14, 2010.

<http://www.fda.gov/downloads/Food/FoodSafety/FoodContaminantsAdulteration/TotalDietStudy/UCM184301.pdf>



Comparison with candy:

Median level of lead in milk chocolate = 0.016 mcg/g (FDA limit = 0.1 mcg/g)

Comparing concentrations of lead in lipstick and chocolate:

0.016 mcg/g << 0.89 mcg/g << 7.19 mcg/g

US FDA report: Total Diet Study Statistics on Element Results. December 14, 2010.

<http://www.fda.gov/downloads/Food/FoodSafety/FoodContaminantsAdulteration/TotalDietStudy/UCM184301.pdf>



Comparison with candy:

1 bar of chocolate has about 43 grams

Exposure from 1 chocolate bar:

$$0.016 \text{ mcg/g} \times 43 \text{ g} = \mathbf{0.69 \text{ mcg}}$$

Average American consumes 13.7 grams/day (11 pounds per year)

Typical daily exposure from chocolate:

$$0.016 \text{ mcg/g} \times 13.7 \text{ g} = \mathbf{0.22 \text{ mcg}}$$



It all comes down to dose!

Typical daily exposure from chocolate (0.22 mcg) is *29 times* the typical exposure from lipstick (0.0076 mcg)

And extreme daily exposure to lead from lipstick (0.78 mcg) is similar to exposure from daily consumption of an average chocolate bar (0.69 mcg)