GRAPHIC ANALYSIS

BAR, PIE, LINE, PICTO...

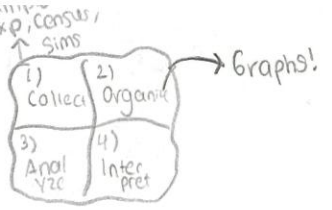
HI
LOW
ANY TREND?
WHY?

HISTOGRAM, STEM/LEAF, BOX/WHISKER, DOTPLOT:

CENTER (I.D. WHICH CTR)
SPREAD (WHICH SPREAD + # + WORD)
SHAPE
OUTLIERS? (undr 5%, or Lg gap)

SCATTERPLOT:

+/- ASSOCIATION
TREND
INS/OUTS
CORRELATION (# & WORD)



Basic Graphs

Visual Representations of Data

- all: keep it simple, ppl only look for 7 seconds
  - Starts w/ a title, ends w/ analysis (looks for highs/lows/trends, why?)

Bar Graphs

- good for differences/ comparing subjects
- categorical data on x-axis
- vertical or horizontal
- keep increments same & width
- can have double/ triple bar graphs
- "squiggle" one time mulligan, break in the axis to jump to data @ higher pts (does this help zoom in or does it exaggerate?
- numbers go inside the bars or nowhere
- Pareto Chart: vertical bar graph arranged tall to small (descending order)
- Ogive: a line graph, building to a limit (100%) it is cumulative always go up, never down

Pie Charts

- % out 100 parts of a whole
- circle/disc slices
- numbers inside sectors / key
- total cannot > 100%

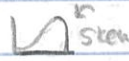
Picto Graphs

- Symbols repeated
- must have a key
- maintain same sized symbols + same spacing

Line Graphs

- Change tracking
- subject tracked repeatedly
- can have multiple on the same graph
- Time Plots: involve time, time on x-axis
- Look for trends!

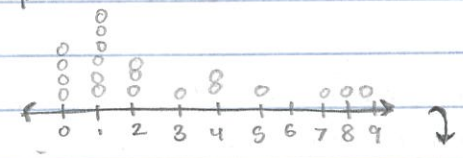
Histogram

- **Shapes**
  - Rectangular  uniform (20%, 20%, 20%, etc)
  - Bimodal  ups and downs, peaks
  - Skewed : tail is the skew
     skewed right  skewed left 
  - Symmetric :  bimodal + symmetric  uniform + symmetric
    - left + right look alike (split down middle)
  - Normal Curve: (also symmetric) 
- **Analysis** (for Histograms, Dot plots, Stem-Leaf plots, box + whisker plot)
  - CSSO (Center, Shape, Spread, outline)

Dot Plots

- newer but not modern technical
- mostly Quantitative data
- good w/ mid size am of data
- if high spread of data, bad
- analysis is easy, original data is visible (advantage over histo) ← exact data visible

Example: Books read last summer



Ana:

C: 1.5 (median)

S: skewed right

S: "ch 3"

O: none

why?

Stats HW 2.2
#1-3, 5, 6, 9, 11, 12

Dasha Heydari
Period 4

10

(1) a) Yes, I think that respondents could select more than one response because all the percents (79, 56, 37, 25) do not add up to 100%, which is what should be the sum if only one response was available. Together, all the data add up to 197%, which is greater than 100%. This could be explained by multiple response options being available for selection.

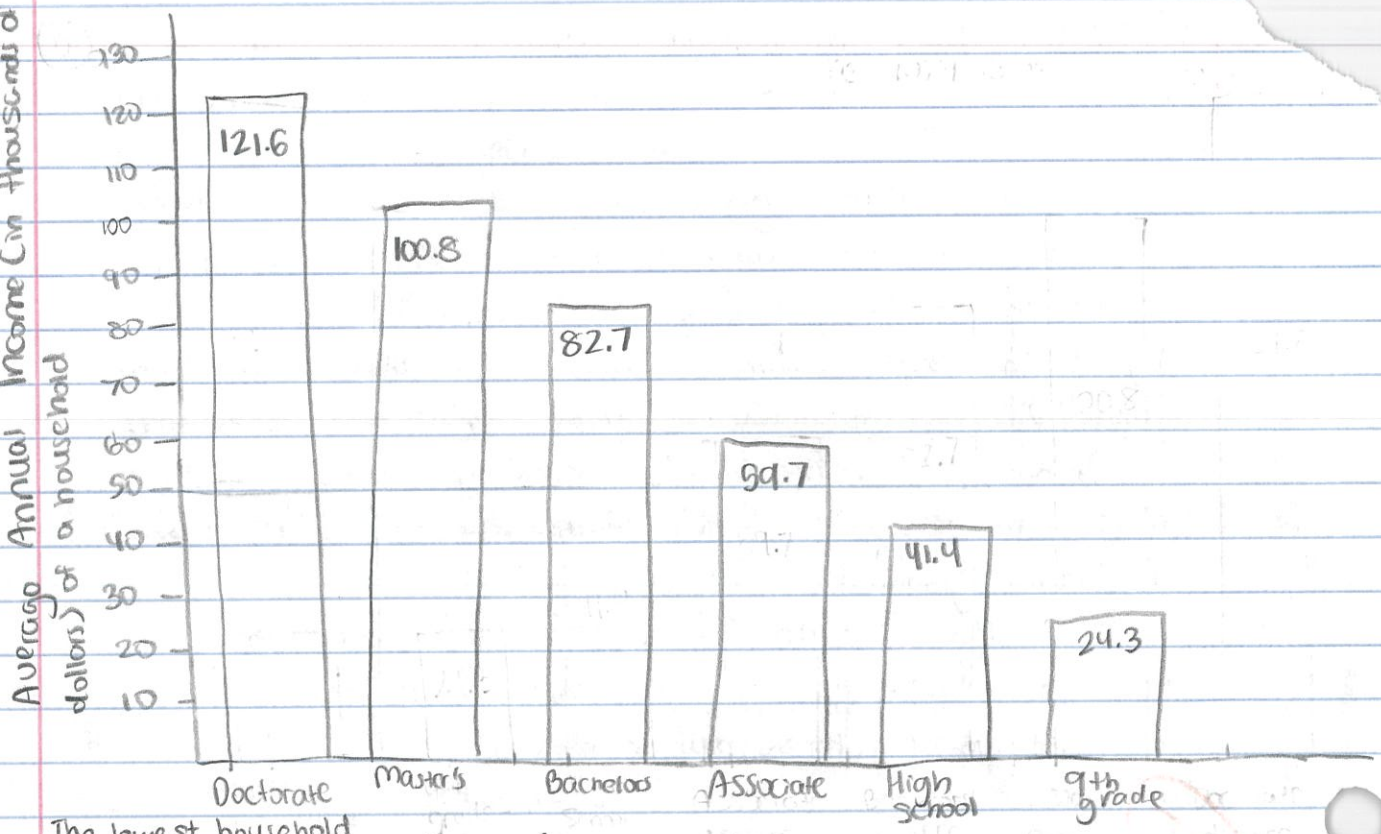
b) No, this same information could not be displayed in a circle graph / pie chart because circle graphs must add up to 100%, not more or less. Because of reasons listed above (the data values add up to 197%), this information could not be represented in a circle graph.

c) NO, graph (a) isn't a Pareto chart, because even though values are ligned up left to right in decreasing order, it is not vertical (Pareto must be vertical)

(2) No, this is not a proper bar graph because the bars are not uniformly the same width, as some are much wider, and others are much skinnier. Bars on bar graphs must be uniformly wide and spaced, which is not being followed in chart b). The y-values should be all inside or outside the bars, not both.

(3) A Pareto Chart would be most useful for representing the data. Visualy, one would be able to see the conditions valued most in decreasing order, so it is easy to see the items in orda of importance to most employees, as well as a visual representaton of how they look compared to one another.

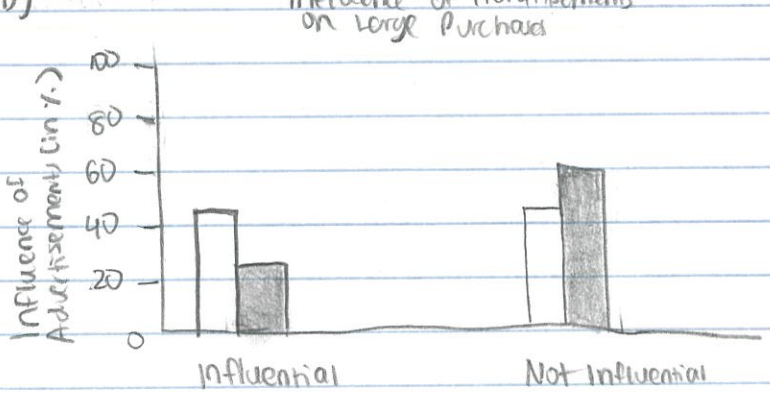
(5) Highest Level of Education and Average Annual Household Income (in dollars)



The lowest household income comes from people with lowest levels of education and the highest household income comes from people with highest levels of education. The overall trend is that as education level increases, the household income also increases — positive correlation. This can be bc more education allows for more high paying jobs.

(6) a) I would change the scaling of the first graph (18-34) to match the scaling of the second graph. In the first graph, the scaling is very misleading, as it starts off at 43% and ends at 46%, so a one percent difference looks vast. If both graphs are scaled at 0-100% or 0-80%, like graph 2, Jenna can more easily and accurately compare the graphs.

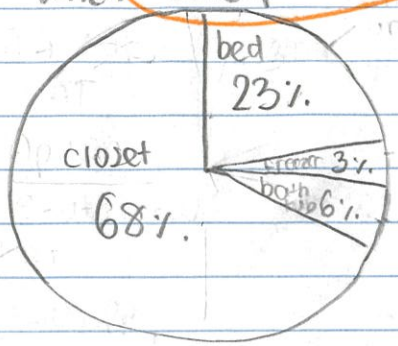
b)



Overall, the 18-34 year olds found the ads to be more influential than not influential, and the 45-54 years olds found the ads to be less influential than influential. Older people found ads less influential than younger ones. This is likely because of older people being wiser and having more skepticism about ads.

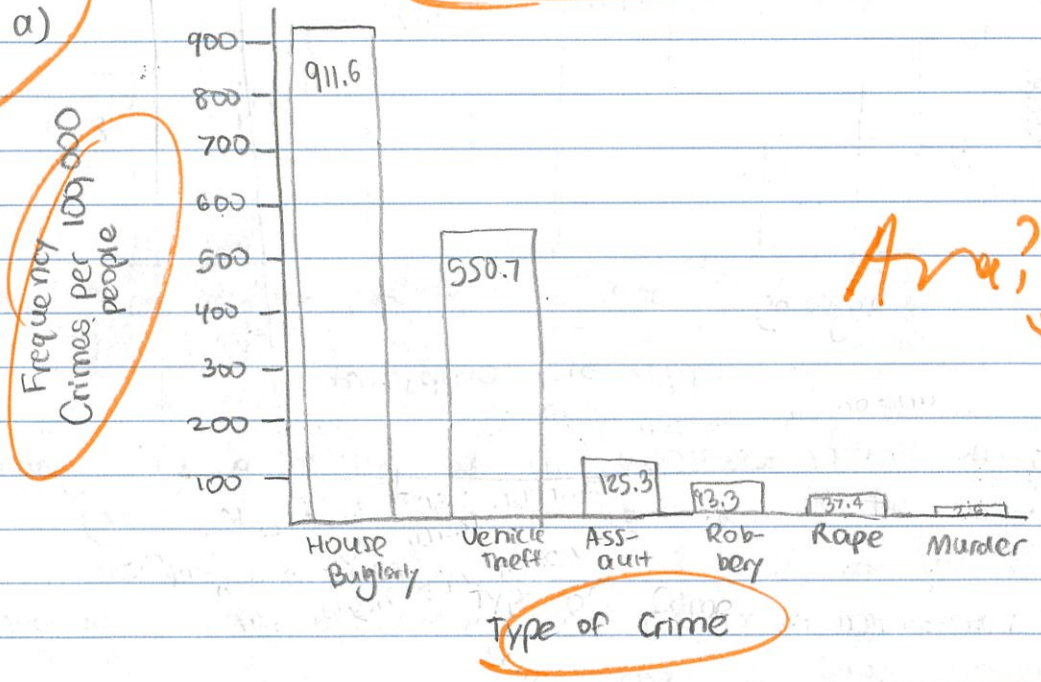
House Keeping Secrets: Where People Hide Messes When Unexpected Company comes

(9)



The closet is the most common place to hide the mess, and the freezer is the least common. The overall trend is that the places in the bedroom are more common hiding places, whereas those outside the bedroom (kitchen, bathroom) are less common. This might be because closets and bed are closer and more spacious and private

(11) a)



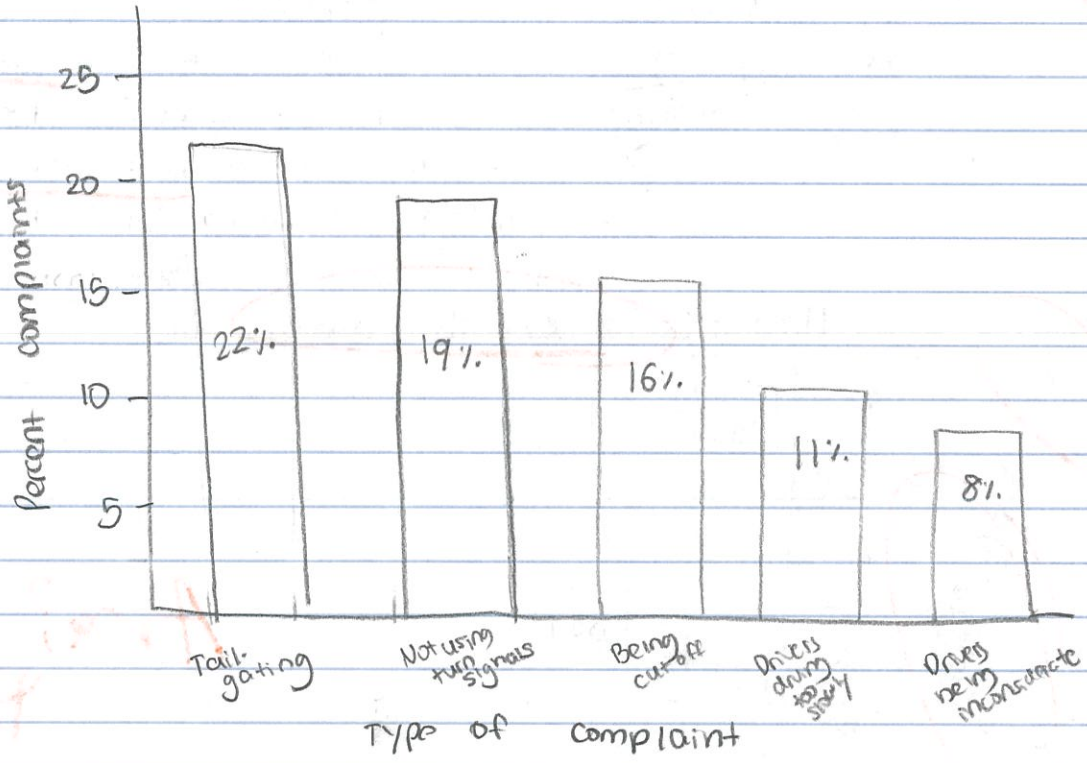
Any?

put alongside the graph.

b) A circle graph is not appropriate to use because the numbers do not add up to a total of 100%, and they are different crimes that would not be represented. Also, someone could commit two or more crimes which could not be represented on a pie chart.

The highest crime rate is house burglaries and the lowest crime rate is murder. The overall trend is that the stealing of objects (burglary, theft) has a generally higher crime rate than those pertaining to humans (rape, murder, assault.) This might be because punishments are less severe for crimes not committed against humans

(12) Percentage of Drivers Complaints



No, this information could not be put in a circle graph as all the numbers do not add up to 100 (they add up to 65), so other types of compaints are not recorded and it is unclear if more than one option could be chosen.

The highest driver complaint is tailgating and the lowest driver compaint is drivers being inconsiderate. The overall trend is that dangerous and illegal activities (no signals, tailgating, cut off) have more compliants than those due to least dangerous activities (too slow, inconsiderate) that are more annoying but not illegal.

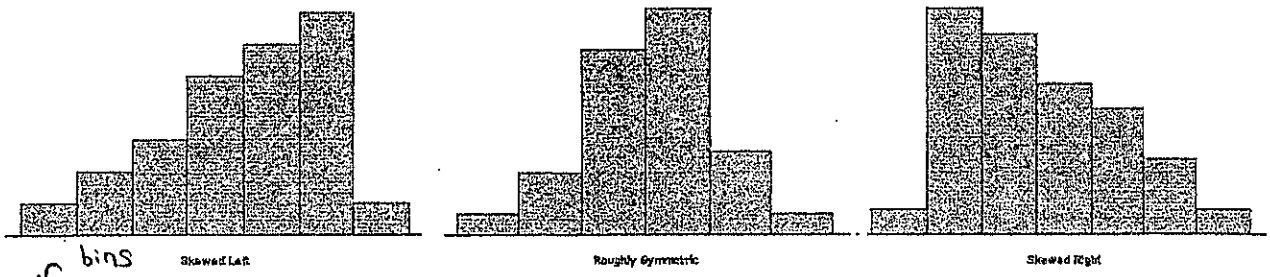
KEY TERMS

A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.

quantitative
used when a lot of data



bins
miss out on specific data, use ranges (bins)

Figure 3.10. Shapes of histograms.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. The video opens by describing a study of lightning strikes in Colorado. What variable does the first histogram display?

The lightning flashes in Colorado, and the time of the first strike

2. In this lightning histogram, what does the horizontal scale represent? What does the vertical scale represent?

Horizontal: Time of First Flash

Vertical: Percent of Days

3. Was the overall shape of this histogram symmetric, skewed, or neither?

Symmetric

4. Why were a few values in the second lightning histogram called outliers?

outliers:

- Points that stand out from the rest of the graph / the normal distribution
- Occured on days with larger weather systems

5. When you choose the classes for a histogram, what property must the classes have if the histogram is to be correct?

- The width / distance of intervals should all be the same
-

6. What happens to a histogram if you use too many classes? What happens if you use too few?

Too few will clump too much data together and will make it not-so informative / hard to differentiate.

Too many will be unwieldy and will be less informative, get lost in the data.

Histograms

- only quantitative (no categorical data)
- bars touch
- grouped into classes / bins
- width of bar has meaning
- useful when Large amounts of data (BIG data)
- recognizing patterns
- deviations + specific data hidden ∴ (since its all grouped)
  -  ↗ can't pick out specific data from the bins (eg, 25-34 is a bin, what is 21 yr old data?, u can't know)
- chosing # of classes / bars / bins
  - ask yourself: would info be hidden / lost? will nothing stand out? play around w/ whats appropriate.
- W = ? (width)
  - $W = \frac{\text{max} - \text{min}}{\#\text{classes}}$ (eg: $W = \frac{47-15}{10}$, $W = 3.2 \rightarrow W = 4$)
  - always round UP (even if 3.005)
- dont start w/ 0, start with smallest data value (squiggle)
- class limits !

↗ useful for making histogram

Ex Frequency Table

10 bars

- $W: \frac{hi-lo}{\#b}$

↳ $\frac{44-14}{10} = 3.0$

3 W

lower class limit

class limits	Tally	Freq	Rel Freq
14-16	𝍸 |	6	$\frac{6}{150} = 4\%$
17-19		22	$\frac{22}{150}$ 14.67
20-22		29	$\frac{29}{150}$ 19.33%
23-25		31	$\frac{31}{150}$ 20.67%
26-28		21	$\frac{21}{150}$ 14%
...		15	$\frac{15}{150}$ 10%
41-44			

W=4? not 3 !

one time, on last one, can grab the furthest data and include (so that you dont have to make a separate last bin)

Stem and Leaf Plots

- Small to medium amts of data
- not too spread out data

1) Find max + min
2) Draw a vertical line splitting data in 2 parts (stem | leaf)
3) Let the data flow (from stems, out to leaves)
4) Rake up /organize

Eg:

Stem	Leaves
0	3 0
1	2 7 8 9
2	0 1 1 1

} have key, tittle, anaysis (& have units)

Max: 37
Min: 01

tens Stems	ones leaves
0	7 2 4 2 1 2 3
1	5 3 1 7
2	1 3
3	7
4	

→

P4 ~off

0	1 2 2 2 3 4 7
1	1 3 5 7
2	13
3	7

hap

2|3 = 23 sec

Ana:
Ctr: 9 (med),
Sh: Skew right (high)
Spd: ch3
Our: 37

2 Lines Per Stem:

0-4 0 2 23421
5-9 . 7
1 3 1
. 5 7
2 1
.
3
. 7

→

0	1 2 2 2 3 4
.	7
1	1 3
.	5 7
2	1
.	
3	
.	7

Stem & Leaf (back to back)

① max 35 min 3

②

M		F
73	0	5 5 5 5 7 5 7 8 8 5
050	1	0 3 5
070	2	5 5 5 8
5	3	5

↓

M		F
73	0	5 5 5 5 5 5 7 7 8 8
500	1	0 3 5
7500	2	5 5 5 8
5	3	5

Key:

0|2 = 20 min male

2|5 = 25 min Female

7|2|5

= 25 min female, 27 min male

Ana:

Both sides!

M
- C : 17.5 min
- S : skewed small F
- S
- O : 35 min

M
- C : 17.5 min
- S : skewed large / high
- S
- O : 35 min

1.4 Stats College Board

Categorical Data:

- frequency table
- bar graphs
  - Relative frequency and frequency table: (Diff)
  - Label axes, scale them
  - equal width + gap
  - y-axis represents freq/rel freq
- pie chart
  - use a key
  - add up to 100

Multiple Sets of Data

1.4

- compare w/ freq table (best is rel freq)
  - when comparing, if diff #s of data, find rel freq to compare, not raw data
- bar graphs!
- (not pie chart)

VS

1.5

Quantitative Variables

- takes numerical data
- countable number of values (w/ gaps) = discrete (eg # of siblings can't be 2.5 → gaps) counting!
- infinitely many values but cannot be counted (no gaps) = continuous (eg height) measuring! (not countable)
- Can be represented w/
  - Dotplot: can see all data in graph and shape of distribution
    - not good w/ a lot of data
  - Stem & Leaf: values
    - have stem + leaves
    - can have split stem (row1 0-4, row2 5-9)
    - can see all values + shape of distribution
    - not good w/ a lot of data
  - Histogram: easier w/ a lot of data, + can see shape
    - can't see individual data

1.6

Distribution of Quantitative Data

1) Shape:
   - symmetric
   - skewed (left + right)
   - unimodal (one peak)
   - bimodal (2 peaks)
   - uniform (about same values)
2) Center
   - mode / median / mean
3) Variability / Spread
   - close or spread out?
4) Unusual features
   - outliers
   - gaps / clusters

The Averages

Mean:

- "average"
- Sigma $\frac{\Sigma x}{n}$ (Sum ÷ # of values) $\frac{\Sigma x}{N}$ $\frac{\Sigma x}{n}$
- $\bar{x}$ for sample
- μ (mu) for a population
- size matters
- Freds test scores:

$\mu = \frac{\Sigma x}{n}$ $\frac{367}{5}$

73.40

rounding rules →

0.____
1.____
2+.__

- class siblings ↓ N = 1.7407
- disadvantage: is easily affected by outliers, pulled toward outlier

Median:

- #s ordered from small → large
- position # numeric values
- resistant to outliers
- odd # of #s = Clear
- even # of #s = halfway
- you can also miss big data as well

(think)

1. 13 13 17 18 21 21

2. 13 13 17 18 32 43 21

Median is 17 for both sets, but is it really representative for the second?

↓ need a measure for variation

mode:

- categorical and quantitative
- most occurring values
- overlooked but often most appropriate
- eg hat sizes
- not affected by outliers
- can be bimodal, but can also have no mode

Trimmed Mean:

- mean that resists extremes
- eliminates pull of outliers
- eg: 10% of n (20)
  - .1 x 29 = 2.9 (always rounds up)
  - ↳ 3

this means, trim off 3 highest and 3 lowest values

- 10% trim means 10% below and above

Measures of Variation

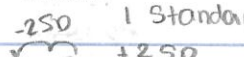
Spread, dispersion

- r
- s or σ (Standard of deviation)
- s^2 or σ^2 (variance, diff than variation)
- CV (coefficient of variation)
- IQR
- Chebyshev's Theorem
- Empirical Rule

Chebychev's Theorem

$1 - \frac{1}{k^2}$ has to at least be more than

1 Standard deviation (2SD, 3SD)

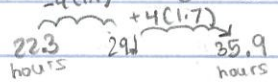


2SD: $1 - \frac{1}{2^2} \rightarrow 1 - \frac{1}{4} = 75\%$

according to theorem, 75% falls within 2SD

3SD: $1 - \frac{1}{3^2}$, $1 - \frac{1}{9}$, 88.9% falls within

Ex working backwards



One sentence interpretation

Standard Deviation

sample ↘

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

how much the data differs from the mean

Example: 2, 4, 6, 8, 10

$\frac{\sum x}{n} = \frac{30}{5} = 6$

$x - \bar{x} \rightarrow -4, -2, 0, 2, 4$

$(x - \bar{x})^2 \rightarrow 16, 4, 0, 4, 16$

$\sum (x - \bar{x})^2 \rightarrow 40$

$\frac{\sum (x - \bar{x})^2}{n-1} \rightarrow 10 = s^2$

$\sqrt{10} \rightarrow 3.16$

- one disadvantage is that it relies on units of msrmt, difficult to compare across situations
- sltn.. CV

Coefficient of Variation

- % of ST Dev to the mean

$\frac{s}{\bar{x}}$, $\frac{\sigma}{\mu}$

Eg: $\bar{x}$ is 6.4 f CV: $\frac{12.10}{6.4}$

s = 12.10 f = 1.89 → 189%.

- can compare across random situations, units of msrmt cancel, pure percentages

- Categories of CVs
  - 0-10%. "banded", clustered
  - 13-33% low
  - 35-67% moderate
  - 70-95% high
  - 100+%. Data gone wild, inconsistent data

Study Guide for Variance (102-110)

- One number may not represent an entire set of numbers well, so a cross reference is measures of variation
- Name 3 measures of variation: SD, VAR, CV, chebysheus
- Advantages & disadvantages of the range:
  - it tells diff between largest and smallest values, v easy to get
  - does not tell us how much the other values vary from one another or from the mean
- The measurement that helps us see how the data is different from the mean is standard deviation
- Why do we divide by n-1 sometimes and N other times? (in standard dev)
  - dividing by N is population (σ)
  - dividing by n-1 is sample (s)
- Obtuse ollie's "new formula" is wrong
  - using median (5.5), but standard deviation only works w/ mean
- Compare + contrast formulas for mean + SD for sample, and population

mean:	SD:
$\bar{x} = \frac{\Sigma x}{n}$	$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$
$\mu = \frac{\Sigma x}{N}$	$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$

Trend + Variation

- dog + owner
- trend = overall, slow
- variations: ups + downs around the trend

Percentile

- ranking, doesn't necessarily indicate score (eg 47/100 can still be 99 percentile)
- splits data into one hundredths
- quartiles more valuable

IQR, 5 Number Sum, and ...

Quartiles:

- Splits data into 4ths: Q1, Q2, Q3
- Q2 is median, P_{50}

1 7 10 14 18 20 21 21 22 22 22 23 23 24 24 24 25 25 25 25 26 26 27 27 28 28 28 30

- Q1 / P25 / LQ / 21
- Q2 / median / 24 / P_{50}
- Q3 / P75 / UQ / 25.5

IQR Interquartile Range

* Q3 - Q1

(eg: 25.5 - 21 = 4.5)

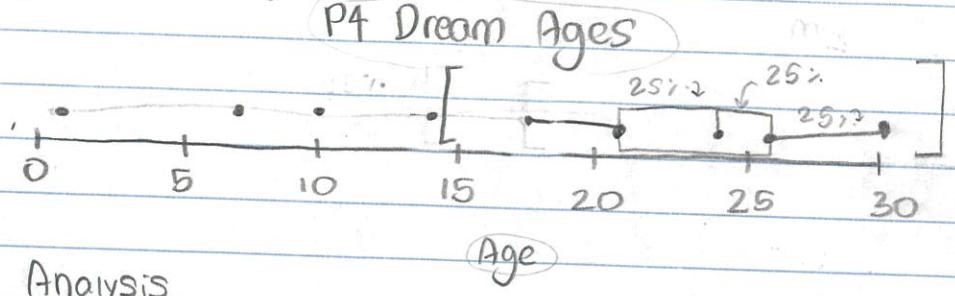
- like a trimmed mean, but for spread
- eliminates 25% off each side
- insulated against outliers

Five Number Summary

- Min, Max, Q1, Q2, and Q3

Box & Whisker Plots

Eg: Ages Pausing ☑



Analysis

use IQR for box + whisker

Ctr: 24 (median)

Shp: Skewed low/left

Spd: IQR = 4.5, "clumped"

Outlier: 1, 7, 10, 14

Outlier Formula

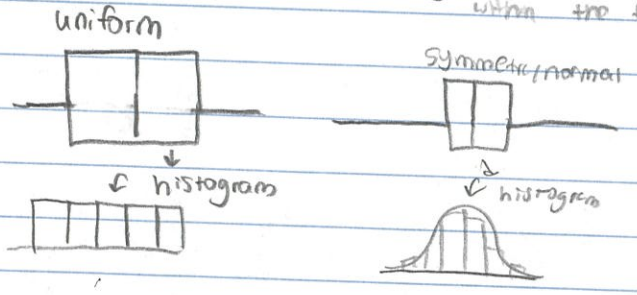
$Q_3 + 1.5(IQR)$

$Q_1 - 1.5(IQR)$

- make fences,
  - take away the line and instead put dots for values excluded
  - don't extend line all the way to fence, go to smallest value within the fence

$25.6 + 1.5(4.5) = 32.25$!

$21 - 1.5(4.5) = 14.25$



Stats HW P135 + 138

Dasha Heydari

10-

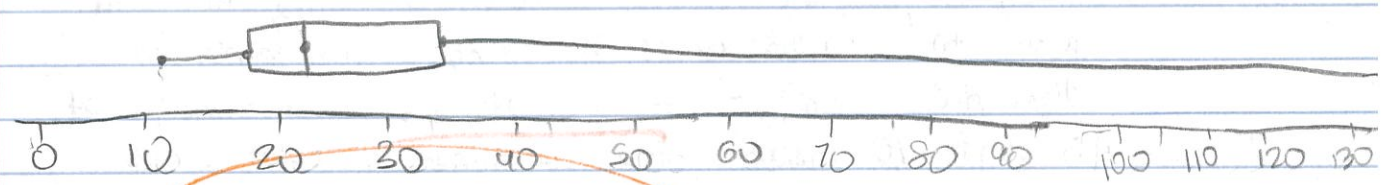
Old Faithful:

Before 1959, the behavior of the geyser was very symmetric, and shaped like a bell curve / normal distribution. The interval with the highest frequency was at 70 minutes. The rest of the data is quite symmetric as it falls to the right relatively symmetric, but there are more points on the left end of the data, towards lower eruption times, so the data could be skewed somewhat low. It does have a single mode frequency, however, at about 70 minutes. Although the mean stayed the same between the years, the data after the earthquake is no longer symmetric. The data does seem to be bimodal, with the modes at about 45 and 70 minutes. The 45 minute mode is shorter, though, with less frequency than 70 minutes. The graph with the higher standard deviation is the graph after the earthquake. The data is spread out widely, whereas the graph from before has a higher concentration of data around the mean, meaning it has a lower standard deviation. The graph after the earthquake will also have a larger coefficient of variation. If actualy at Yellowstone, it would be hard to predict when the next eruption occurs because the data is so wiadly varied and unpredictable.

Hawaii Winds

a) Mean: 33.17
Median: 21
Mode: 18
Range: 138 - 11 = 127
Variance: 845.06
SD: 29.07

b) Min=11 Q3=34
Q1=16 Max=138
Q2=21

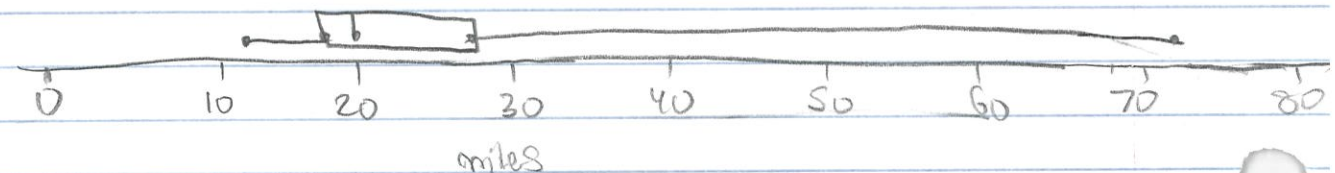


Ana
C: 21 (Median)
Shp: Skewed high
Spd: 18 (IQR) = "?"
Outliers: 138, 113, 105, 100

c) Mean: 25.48 Mode: 18 Variance: 208.22
Median: 20 Range: 72-11=61 SD: 14.43
Min: 11
Q1: 16 IQR: 12
Q2: 20
Q3: 28 Max: 72

Total Air Movement

Ana:
C: 20 (media
Shp: skewed hig
Spd: 12 (IQR)
Outlier: 72



The average most affected is the mean, since it is affected by outliers.
The plot is still skewed high but much less so.
The Standard deviation decreased massively

Dasha
Heydon

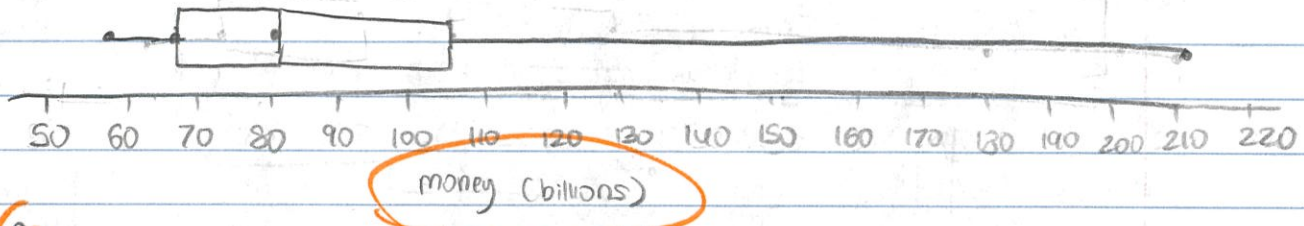
9+

Stats 3.3 (3, 4, 7, 8, 10-12) + rich people

Rich people

Q2: 80.6 (P_{50}, median)
Q1: 66.2 (P_{25}, LQ)
Q3: 105 (P_{75}, UQ)
min: 57.6
max: 211

Rich people :



Ana:
Ctr: 80.6 (median)
Shp: Skewed high
Spd: IQR: 38.8 (big spread)
Outliers: 180, 211 billion

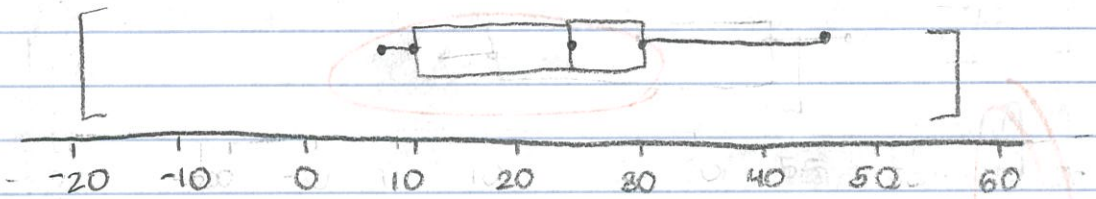
(3) No as a score of 82 might not be in the 70th percentile to know percentile you must know all the other data. how? xplain.

(4) Timothy performed better as his percentile was 72, whereas although Clayton scored higher, his percentile was lower. Since percentile is comparing to other data, with respect to the other data / students, Timothy did better.

(7) Min: 7 Q2: 23 Max: 42
Q1: 9.5 Q3: 28.5 IQR: 28.5 − 9.5 = 19

outlier limit: 57, −19 (so none)

Length of nurses positions (in months)

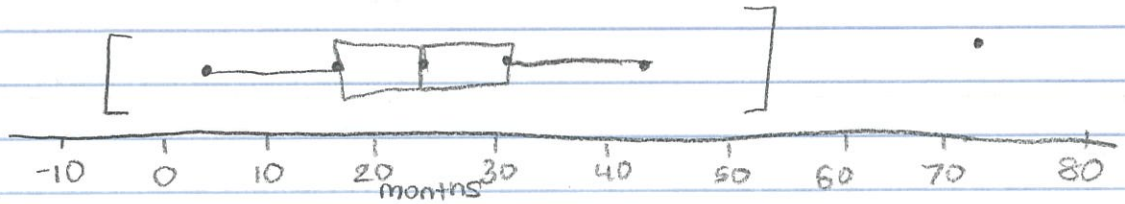


Ana
C: 23 (median)
Spd: 19 (IQR) (medium)
Shp: skewed high
outliers: none

(8) Min: 3 Q3: 30 IQR: 30 − 16 = 14
Q1: 16
Q2: 23 Max: 72

outlier limit: −5 and 51
outliers: 72

Length of nurse posters (in months)

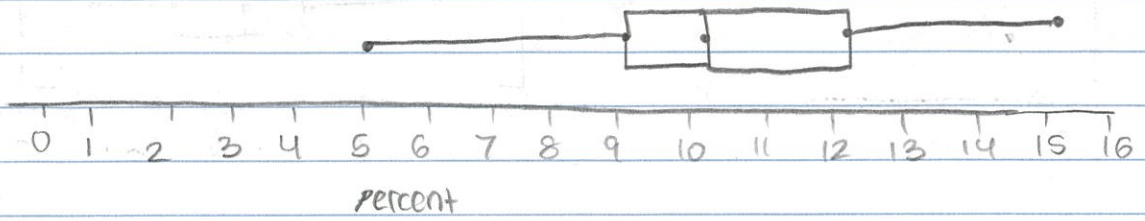


Ana:
C: 23
S: 14 (IQR) (medium / low)
Sh: symmetric / normal
O: 72

Comparing: The medians for both are the same, however, the data in this problem is more symmetric and problem 7) is skewed high. The middle half of the data in 7) has a higher spread but in 8) it is closer together / less spread.

(10) Min: 5 Q3: 12
Q1: 9 Max: 15
Q2: 10 IQR: 3

Percent of HS dropouts by state



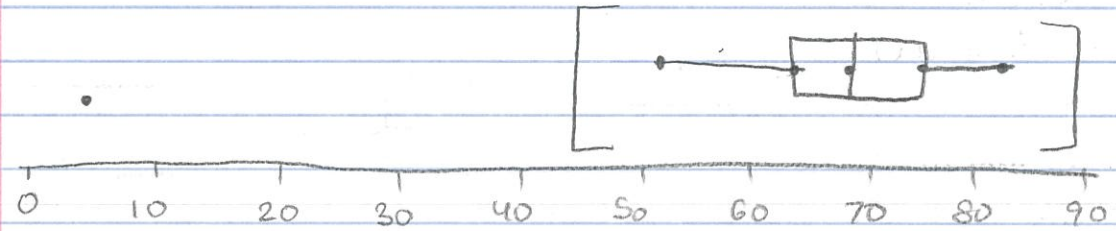
Ana:
Ctr: 10 (median)
Shp: relatively normal/symmetric
Spd: 3 (IQR) (low!)
Outl: none

b) First quartile (between min and Q1).

(11) a) California (lowest) and Pennsylvania (highest)
b) Pennsylvania
c) California has smallest range, Texas smallest IQR
d) a = Texas
b = Pennsylvania
c = California

Min: 4 Q1: 61.5 Q3: 71.5
Max: 80 Q2: 65.5 IQR: 10

(12)



b) IQR = 10

c) Lower limit: 61.5 − 1.5(10) = 46.5
Upper limit: 71.5 + 1.5(10) = 86.5

d) Yes, 4 is an outlier.

Ana:
Ctr: 65.5 (median)
Shp: Symmetric / normal curve
Spread: 10 (IQR)
Outliers: 4

Dasha
Heydari

KEY TERMS

A **five-number summary** of a set of data consists of the following:

minimum, first quartile (Q_1), median, third quartile (Q_3), maximum.

The **first quartile**, Q_1, is the one-quarter point in an ordered set of data. To compute Q_1, calculate the median of the lower half of the ordered data. The **third quartile**, Q_3, is the three-quarter point in an ordered set of data. To compute Q_3, calculate the median of the upper half of the ordered data.

A basic **boxplot** (or **box-and-whisker plot**) is a graphical representation of the five-number summary. A modified boxplot indicates outliers and adjusts the whiskers.

The **interquartile range** or **IQR** measures the spread of the middle half of the data:

$$IQR = Q_3 - Q_1$$

The **range** measures the spread of the data from its extremes:

range = maximum – minimum

10

THE VIDEO: BOXPLOTS

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What *variable* is used to compare different brands of hot dogs?

Calories

2. What name do we give to the value for which one-quarter of the data values falls at or below it?

Median Q1

3. What numbers make up a five-number summary?

Q1, Q2, Q3, max, and min

4. How do you calculate the interquartile range?

Q3-Q1

5. Boxplots show that poultry hot dogs as a group differ from all-beef hot dogs. Compare the distribution of calories between the two types of hot dogs.

✓ Beef is mostly normal/uniformly shaped. All poultry dogs are less in calories than half of the beef dogs. Half of the poultry is less calories than all the beef. Poultry skewed high

Scatter Plot

- Bivariate Data
- + Association: linear, direct relationship ↗
- − Association: linear, indirect relationship ↘
- No association, no any clear pattern, random
- Residual is distance from the dot (x,y) to the line of best fit (predicted x,y); if a big residual = outlier!
- linear vs curved
- explanatory variable x-axis, response variable y-axis
- can introduce categorical variables (although are quantitative) using color to distinguish. add dimensions to data.
- influential follow pattern, just gapped
- lobf (line of best fit) is the least squares regression line
  ↳ $y = a + bx$
  - b = slope
  - a = y-int

Analysis:

- +/− Association
- trends (sentence)
- ins/outliers (in follows pattern but gapped)
- Correlation (r)

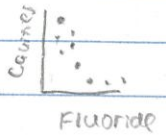
number and "word"

} ATIC

Eg Ana:

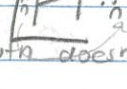
A − Negative curved association

T − As fluoride lvls increase, we observe the number of cavities drop



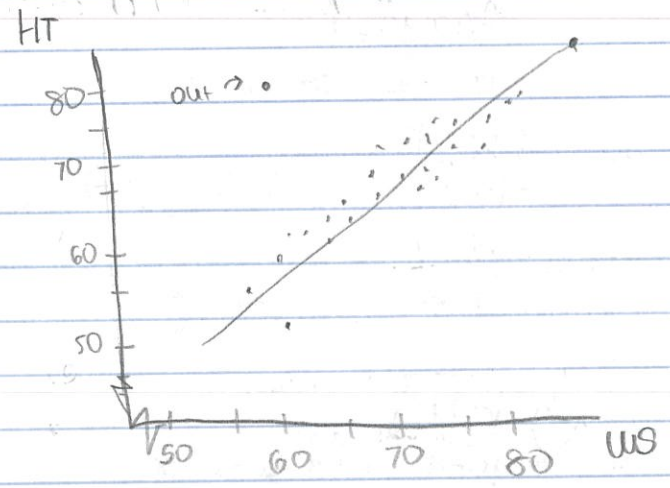
Line of Best Fit *

- results in smallest sum if square all the residuals and add them
- $\hat{y} = bx + a$ (b = slope, a = y-int) ^ = hat - prediction
- $b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$
- $a = \bar{y} - b\bar{x}$
- talls about it! interpret in contex!
- extrapolation (doesn't grow forever, lines off @ a certain point)
  - interpolation
  - Sometimes linear growth doesn't go forever
  - So LoBF only fits data that behaves in a linear way



- outliers affect the slope!
  - mark two points used to make line, in a special way (they r not part of the data)

P4 squares



Ana:

- + Linear Association
- Shorter people have shorter arms, and taller ppl have longer arms. WS & HT are proportional
- Outliers: (60, 80)
  Ins: (81, 80) following pattern but gapped
- r = not yet...

(3) "Famous" cars

$\hat{y} = 6.56 + 1.01x$

for every ad, u sell a car (ea) → $b = 1.01$

zero ads = 6.56 cars sold any way → $a = 6.56$

if $x = 12$, $y_p = 18.68$

Correlation Coefficient (r)

- ranges: $-1 \leq r \leq 1$
- tells us where the dots are in relation to line of best fit, how clustered to line?

★
- $r = +1$, perfect linear positive correlation (no residuals, all on line)
- $r = 0$, no correlation
- $r = -1$, perfect negative linear correlation (no residuals, all on line)

- doesn't have to do with slope of line
- only works for linear correlation
  - if not a line, you would just say $r \approx 0$
  - if slope negative, r negative. if b positive, r positive.

- need a word -
.969 = "SSNLC"

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

formula same as b, except for the right denom portion (same thing w/ y)

Coefficient of Determination (r^2)

- always positive, boundary is $0 \leq r^2 < 1$
- need a full sentence - exactly like this! not just a word

$r^2 =$ (eg) 93.9 %, of the variation in the y (# of mugging in the park) is explained by the Least Square) regression line and the variation in x (# of police officer in the park.)

↳ you can add that 6.1 % is unexplained by this.

conjectures : weather, time of year, park closure

AP STATISTICS: AGAINST ALL ODDS
Video 8 Worksheet

Name Dasha Heydari

DESCRIBING RELATIONSHIPS

1. What is a plot of quantitative variables? Scatterplot

2. What is the x-variable called in studies? explanatory variable the y-variable? response

3. What is a variable that records into which of several categories a case falls? categorical variables

4. How do categorical variables enrich a scatterplot? they separate the data and provide insight

5. What type of smoothing is found by slicing the scatterplot vertically, calculating the median within each slice, and connecting these medians by a straight line? median trace

6. What example in the video illustrates the use of a median trace? military draft

7. What is the best fitting line that fits data by minimizing the sum of the squares of the residuals?
line of best fit

8. What example is used to illustrate the use of the least squares regression line? lean body weight + metabolism

9. In the equation $\hat{y} = a + bx$ ($\hat{y}$ prediction), what is the formula for b? $\frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

   What is b in the equation? slope What is the formula for a? $a = \bar{y} - b\bar{x}$

   What does y represent? value of y predicted by x x? any data point

   What is a in the equation? y-intercept

10. Even though you can fit a regression line to any set of data, when is the line valid?
data that behave in a roughly linear way.

11. What are points with unusually large residuals? outliers

12. What are points that deviate strongly in the x-direction? influential observation

AP STATISTICS: AGAINST ALL ODDS
Video 9 Worksheet

Name _____

CORRELATION

1. What is the measure of the strength and direction of the linear relationship between quantitative variables?

   r → correlation coefficient

2. What values does *r* vary between? -1 and 1

3. What indicates a perfect positive correlation? 1 a perfect negative correlation? -1

4. What study in the video illustrates the use of correlation? twin study

   Which characteristics showed a strong correlation? height

   Which characteristics showed a moderately strong correlation? personality

5. In the formula for *r*, what do $\frac{x-\bar{x}}{s_x}$ and $\frac{y-\bar{y}}{s_y}$ do? z score

   Why does the formula divide by n – 1? _____

   When is *r* positive? positive correlation

   When is *r* negative? negative correlation

6. What kind of relationships does *r* measure? linear

7. What describes the amount of variation in y described by the linear relationship with x? _____

8. What example in the video uses the squared correlation coefficient? _____

r ≠ b
dont have to do w/ each other, can have high slope but not high correlation, as well as low slope & high correlation

outliers affect r, influential influences slope

Residuals / Residual Plots

- $y - \hat{y}$ = actual - predicted
  - + result: model underestimated
  - − result: model overestimate
- residual plot: plot of all these ↑ values, not y values
- linear = good fit when
  - residual plot is random
  - centered at 0
  - no clear patterns
- linear = bad fit
  - definite pattern of a curve
  - residuals aren't "noise", there is a curved trend that model didn't capture
- help see best model for fit

LSRL

Removing points

- removing points near the mean does not change slope (or y int) of LSRL (low leverage points)
- removing high-leverage points (unusually smaller or higher than $\bar{x}$ / far away $\bar{x}$) will cause the y-int and slope to change, large effect
- removing an outlier w/ a high residual that isn't a high-leverage point (x value close to $\bar{x}$) will change correlation(r)

Summary:

- outliers change correlation
- high leverage change slope / y-int
- or both & high lvg and outlier: change slope / y int and correlation

Dasha Heydari ?

WHAT WAS THE MOST POPULAR SONG OF 2013? 10

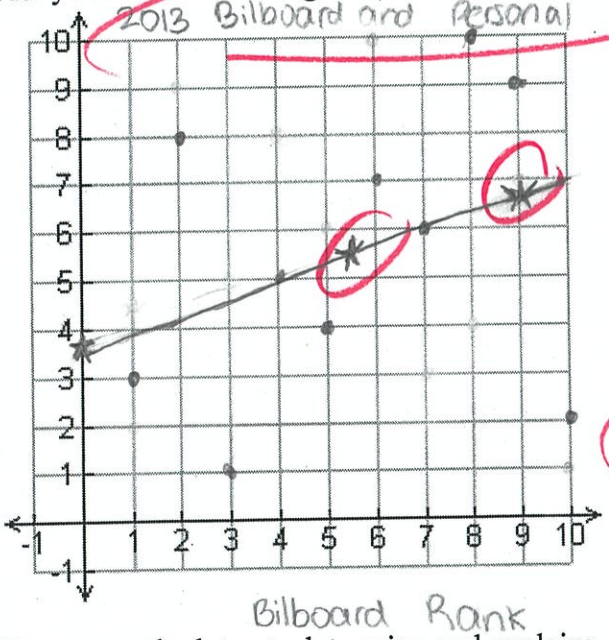
We use the top songs on Billboard's year-end charts to explore this question.

A. Can't Hold Us
B. Cruise
C. Thrift Shop
D. Just Give Me a Reason
E. Harlem Shake
F. Radioactive
G. Mirrors
H. Blurred Lines
I. Roar
J. When I Was Your Man

Rank the songs from 1 for first place to 10 for tenth place. When you have finished your rankings, Mr. Micek will give you the Billboard rankings.

Song	X Billboard Rank	Y Your Rank
A.	5	4
B.	9	9
C.	1	3
D.	7	6
E.	4	5
F.	3	1
G.	6	7
H.	2	8
I.	10	2
J.	8	10

Plot your results on the grid. (x-axis is billboard rank, y-axis is your rank)



Ana

- Very weak positive linear association
- Trend: As the billboard rating increases, my rank also increases (somewhat)
- Ins: none
- Outs: (10,2) + others
- correlation: r = 0.2970 "very weak positive linear correlation"

$\hat{y} = a + bx$

$\hat{y} = 3.87 + 0.2970x$

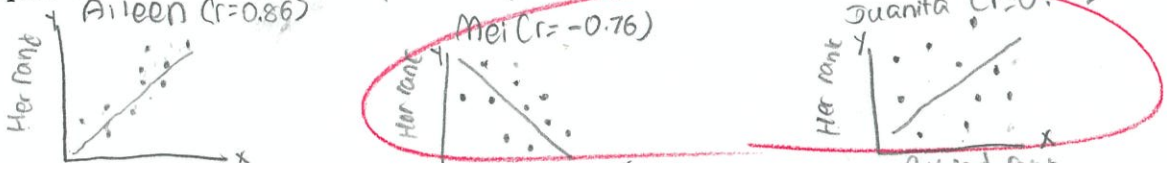
	X	Y
$(\bar{x}, \bar{y})$	5.5	5.5
	9.0	6.543
	0.0	3.87

Use your calculator to determine and explain:

R = 0.2970 "weak positive linear correlation"

R^2 = 8.82 % 8.82% of the variation in y (your rank) is explained by the least squares regression line and the variation in x (Billboard rank.)

Extension: You know that Aileen had a measure of 0.86 while Mei had -0.76 and Juanita had 0.24, sketch a plot of their rankings. 3 separate plots



$y - y_1 =$ residual

2) A residual is the diff between an observed & predicted value / vert distance (sum = 0 always)

3) An outlier is a point w/ a large residual, fall outside LSRL pattern. (points falls away from pattern)

4) What is a residual plot? plot of residual points. definite pattern = nonlinear model better

5) An influential observation is (point whose removal would sharply change the LSRL) than LSRL. Point w/ extreme x-value, may have a small residual, but can have ↑ effect on LSRL than outlier w/ avg x-values

6) 4.10, y (24, 76000 influential?): (x-value ↑ than normal. changes slope of the line drastically)

7) One type of transformation could be done w/ x w/ log y, x^2, etc etc

Residuals:

- $y - \hat{y}$ (or $\frac{y - y_p}{y - y_1}$)
- vertical distance from point to the line
- above line = + res
  below line = - res
- outliers have ↑ res
- influential obs/a point whos removal would sharply change the LSRL. point w/ extreme x-value may have a small residual.. but can ↑ effect on LSRL than outlier w/ avg x-values

→ **Plots:** (any!)

- if u see pattern, linear is not the best fit
- random pattern, or close to 0, means linear is good (clse to 0 = low residuals)

residuals

go to stat → calc → 8 → 2nd stat → 7 → resids

Sto → L#, then u can graph w/ og x list & L# chosen as scatterplot