Histograms 10/2/12

~ defined: special type of bar graph in which the bars touch
~ width of bar has meaning (age range, distance, time)
~ data grouped into classes
~ useful with large amounts of data
~ shape tells you info
~ deviations hard to point out
~ exact data is hidden
~ How to choose classes
  ~ would info be hidden/lost?
  ~ will anything stand out?
  ~ 5-10 bars ... Mr. N likes 7
~ width
  ~ W = $\frac{\text{large} - \text{small}}{\#\text{ of bars}}$ = answer (round up!)
~ basically colums to organize raw data before graphically displaying them: Frequency Table
~ Class/groups, tallies, frequencies, midpoints, relative frequency
~ first class begins w smallest, next is small + W
~ Types of Histograms
  ~ Normal (symmetric)
  ~ Rectangular
  ~ Bimodal
  ~ Skewed left/right (tailing off side is in name
~ Analysis (for histograms, dot plots, stem and leaf, box and whisker, etc)
  ~ center
  ~ shape
  ~ spread
  ~ outliers
~ Ex: high = 44, low = 14 $W = \frac{44-14}{10} = 3$

Classes	Marks	Tally	FRQ (%)	REL. FRQ (%)
14-16				
+3 17-19				
⋮				
41-44 ← include last one, to make 10 classes				

Section 2.2 10/2/12

~ Histograms
  ~ Bars are same width and always touch
  ~ width represents a quantitative value
  ~ height indicates frequency

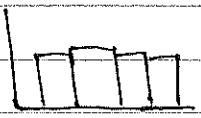
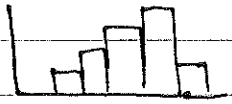
~ Classes
  ~ Number: 5-15
  ~ width: $W = \frac{\text{Largest} - \text{smallest}}{\#\text{ of classes}}$ round up!

~ Midpoint
  ~ center of class or "class mark"
  ~ MP = $\frac{\text{Lower class limit} + \text{upper class limit}}{2}$

~ Class Boundary
  ~ Halfway point between upper limit of one class and lower limit of next class
  ~ used as endpoints for bars on histogram (in some conventions...)

~ Frequency Tables
  ~ Determine class width → create classes → tally data → obtain class frequency (total the tallies) → compute midpoint → determine class boundaries
  ~ Relative Frequency table: do above plus calculate relative frequency

$$RF = \frac{f}{n} = \frac{\text{class frequency}}{\text{total frequency}}$$

~ Distributions
  ~ symmetrical 
  ~ uniform 
  ~ skewed left 
  ~ right 
  ~ bimodal 

For Stats class:
~ don't use Boundary, use limit
~ include tallies on frequency table, no boundary, no MP

Graph Analysis 10/4/12

~ Circle, line, bar, and Pictographs
  ~ trend
  ~ unusual?
  ~ highs and lows
  ~ Why?
~ Dot Plot, Histogram, stemplot, boxplot
  ~ Center
  ~ Shape
  ~ Spread
  ~ Outliers

Stem and Leaf Plots

~ easily constructed
~ shows original, specific data
~ shows shape and distribution
~ allows analysts to view...

~ small quantities of data
~ How to:
  1) Find smallest and largest value
  2) Decide on stems and write them vertically w a line to right
  3) separate values into stem and leaves. Put leaves on right
  4) On a new plot, arrange leaves from smallest to largest
  5) Have a Key (ex: 2|3 = 23), Title, make sure #s line up
~ Analysis
  ~ Center: median, be sure to indicate what kind; ex: 22 (median)
  ~ Shape: turn graph to side → OMG it's a histogram
  ~ Spread: where is the bulk; number and word
    (ex: 12-24 (range of 16); moderate)
  ~ outliers: what is far away from the bulk
~ Example:
  1) Domestic (mpg): 23, 24, 15
     Foreign (mpg): 33, 35, 27

If you have a lot of data:

Domestic		Foreign
5	1	
4 3	2	7
	3	3 5
3|2=23	4	

3|3=33

CSSO CSSO

Dom.		FOREIGN
	1	
5	•	
4 3	2	
	•	7
	3	3
	•	5
	4	
	•	

1 → 10-14
• → 15-19

CSSO CSSO

2) From 2.3 #8

1	5
2	
3	
4	9
5	4
6	
7	
8	5
9	0 5
10	0 2 2 6
11	
12	3 6
13	6 0 9
14	4 9
15	0 4 9
16	6 3
17	5
18	5
19	
20	
21	
22	
23	5

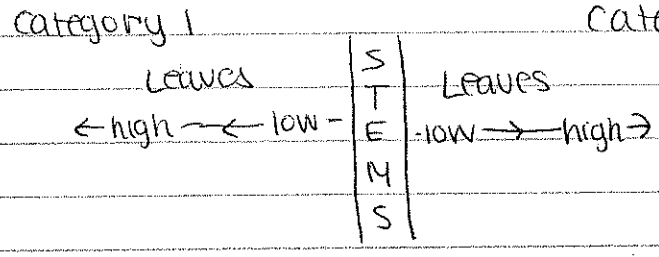
23|5 = 23.5

3) How OFF are you?

women	man
36	28
4	8
12	20
3	40
2	20
13	12
20	15
18	5
	10

Section 2.3

- Exploratory Data Analysis (EDA)
  - useful for detecting patterns and extreme data values
- Stem and Leaf Display
  - one EDA technique
  - organizes and groups data, but allows us to see original data too
  - How to: (see class notes)
  - Back-to-Back Stem and Leaf Plots



Against All Odds video 2 Notes 10/8/12

- Histogram (Lightning Example)
  - Distribution: overall pattern (first flash b/w 11-12 noon, symmetric)
  - outliers: data that is away from the majority
  - Good for large quantities of data
  - Always look for the big picture first
- Shape, center, Distribution/Deviation
  - symmetry, skewed left/right, bimodal
  - Center: where is the tallest bar, median
  - Outlier: you want to be able to explain why? ***
- spread
  - If spread is too much, graph is less useful → split it up
- Stemplot
  - Center: median
  - outlier: why?
  - keep exact data, good for small quantities of data
  - CSSO, then conjecture why

Semester One Project (Due Oct 29)

- Procedure
- Script
  - so that each partner says same thing
  - have the person step away
- Visual, Auditory, Written (for assurance about bag and incentive)
- BIG BAG. With surveys in it always! and incentive!
- Sample size
  - 100 or more
  - MUST GO OFF CAMPUS AS WELL (at Medea, check-in day before, approval
  - subgroups must be 25 or greater
- Survey
  - Instructions : remind of bag and incentive
  - 12-15 questions, one page
  - Read tutorials
  - 5 peer reviews
  - 13 Kids to convenience sample as Practice / test run
  - Run it by Mr. Micek

Measures of central tendencies 10/10/12

- The Averages
  - Mean
  - Median
  - Mode
  - Trimmed
- Arithmetic Mean
  - usual meaning behind "average"
  - mean = $\frac{\text{sum}}{\#\text{ of items}}$
  - population : μ
  - sample : $\bar{x}$
  - outliers can have large impact. Pulls mean toward outliers
- Median
  - middle value when ordered small to large
  - position, not numeric values → resistant to outliers

↓

~ ex: 13 15 17) (18 19 20 median = 17.5

13 15 17) (18 320 450 median = 17.5 ← Should it be?

↑

~ Mode
- ~ most often occuring value
- ~ overlooked, but often more appropriate
  - ~ ex: average hat size
- ~ can be bimodal or trimodal or multimodal; or no mode
- ~ outliers have no affect

~ Trimmed Means
- ~ mean that resists extremes; eliminates pull of extremely low or high values
- ~ n% trimmed mean → lop of n% off each end (ROUND UP!) → find mean as normal

Bird Sneak

~ mean, median, mode, 5% Trimmed mean (show work)

~ Pattern?

~ use any wild animal (use same one every time, 10 reps)

~ Prediction, Procedure, Data, Graphs (w/ CSSO), Conclude

Measures of Variation

10/12/12

~ Population Parameters: μ, σ^2, σ

~ Sample stats: $\bar{x}, s^2, s$

~ Standard deviation

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

~ Variance
- ~ standard deviation squared

~ examples

~ Temp. in last 10 years

-8 x
3 x
25 x
17 x
-2 x
10 x
-12 x
21 x
4
6 x

$\bar{x} = 6.4$

median = 5

Range = 25 − (−12) = 37

$s^2 = 146.49$

$s = 12.10$

~ Ruby weight

19.8
43.8
36.1
52.4
63.1
20.7
46.3

range = 43.3

$\mu = 40.31$

$\sigma = 14.82$

Sections 3.1 and 3.2

- Mode: most common value
- Median: odd (middle number once small→large); even (average of the two middle numbers once small→large)
- Mean: sum / # of entries
  - sample mean: $\bar{x}$
  - population mean: μ
- Trimmed Mean: order small→large, take x% of front and back, take mean of the rest

- Measures of Variation
  - Range: large − small
  - Standard Deviation: how data differs from the mean; small SD = small spread
    - sample standard deviation
      - $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$ ★ $\sum(x-\bar{x})^2 = SS_x$ = sum of squares
    - Population standard deviation
      - $\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$ * uses "N" not "n-1" b/c samples often don't have extremes. "n-1" → s bigger → accounts for this
  - variance: standard deviation squared
    - sample variance
      - $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
    - Population variance
      - $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$
- Coefficient of Variation: expresses standard deviation as a percent of sample or population mean (good for comparison b/c no units)
  - Sample: $CV = \frac{s}{\bar{x}} \times 100$
  - Population: $CV = \frac{\sigma}{\mu} \times 100$
- Chebyshev's Theorum: for any set of data and for any constant $K > 1$, the proportion of data that must lie with K standard deviations on either side of the mean is at least $1 - \frac{1}{K^2}$

~ for any set of data:
  ~ at least 75% of data fall into interval $\mu-2\sigma$ to $\mu+2\sigma$
  ~ at least 88.9% of data fall into interval $\mu-3\sigma$ to $\mu+3\sigma$
  ~ at least 93.8% of data fall into interval $\mu-4\sigma$ to $\mu+4\sigma$

~ Spread (OSSO)
  ~ use coefficient of variation

~ CV	Word	
↓10-13%	Tight	within each range use "moderately" or "very"
15-30%	Low	
33-67%	Medium	
70-95%	High	
97%↑	Wild	

  ~ ex) spread: 10.28% CV, Tight

~ Grouped data (ex: frequency table)
  ~ $\bar{x} = \frac{\sum xf}{n}$ ✱ x = midpoint of the class
  $n = \sum f$
  ~ $s = \sqrt{\frac{\sum (x-\bar{x})^2 f}{n-1}}$ f = # of entries in the class
  ~ $s = \sqrt{\frac{SS_x}{n-1}}$
  $SS_x = \sum x^2 f - \frac{(\sum xf)^2}{n}$

Box and Misker Plots 10/18/12

~ Percentile
  ~ like mile markers; 10th percentile = above/equal to 10% of people
  ~ There is no 100th percentile, highest is 99th percentile
  ~ ex: P. 112 #3
  Not fair. You would fire 70% of staff. If all scores were very high, 82% could be in the bottom

~ Quartiles
  ~ special percentiles that splits data into fourths
  ~ 1st Quartile = LQ = Q_1 = 25th percentile
  2nd Quartile = median = Q_2 = 50th percentile
  3rd Quartile = UQ = Q_3 = 75th percentile

Section 3.3

~ Percentile
  ~ for whole numbers P (where $1 \leq P \leq 99$), the P^{th} percentile is a value such that P% of the data fall at or below it and (100-P%) of the data fall at or above it.

~ Quartiles

~ 25% | 25% | 25% | 25%

Q_1 (LQ, 25%ile) | Q_2 (median, 50%ile) | Q_3 (UQ, 75%ile)

  ~ To compute:
    1) order small → large
    2) Find Median - This is Q_2
    3) Q_1 is median of data below Q_2
    4) Q_3 is median of data above Q_2
  ~ Inter Quartile Range = $Q_3 - Q_1$

~ 5 Number summary
  ~ min, Q_1, Q_2, Q_3, max

~ Box and whisker
  ~ find the 5 Number summary & IQR
  ~ Draw a box around IQR
  ~ Draw a line through median
  ~ Draw whiskers out to min & max
  ~ TITLE & AXES LABLED
  ~ CSSO!!