Part I: Organizing Data
Chapter 1: Exploring Data
1.1 Displaying Distributions w/ Graphs - Day 1

Statistics is the science of data.

Individuals - objects described by the set of data.
Ex: people, animals or things.
Variable - any characteristic of an individual.

Categorical variable - places an individual into one of several groups or categories. (Gender; school Dist.)

Quantitative Var - takes numerical values. (Age, Grade).

Statement: 82% of the population are blonde.

When you encounter a new set of data ask:
Who - Who are the individuals?
What - What are the variables? What are the units?
When - When was the data taken? (1940, 2018)
Where - Where was the data taken from?
Why - Why did you collect the data?
How - How was the data collected?

Distribution - what value the variable takes + how often it takes those values.

How to construct a bar graph - pg 8
Bar Graphs show counts.
Pie Charts show part of the whole

Figure 1.1 - Pg 9 (Which is Better?)

Activity: Take Pulse Rate 3x
AVG
Any unusual Pulse Rate?

Pulse Rate AVG FEM

AVG=

Pulse Rate AVG MALE

AVG=

Save Data for Back-to-Back Stemplots.

Questions:
1) What individual do the data describe?
2) How many variables are there? What are they?
3) In what units is each var. recorded?
4) Are the vars. Quant. or Cat.?

M/F in class Pulse Rate Gender

Pulse - Quant
Gender - Cat.

1.1 HW (1.2 - 1.5) (Tomorrow to Pg 26)

1.1 Displaying Distributions w/ Graphs – Day 2

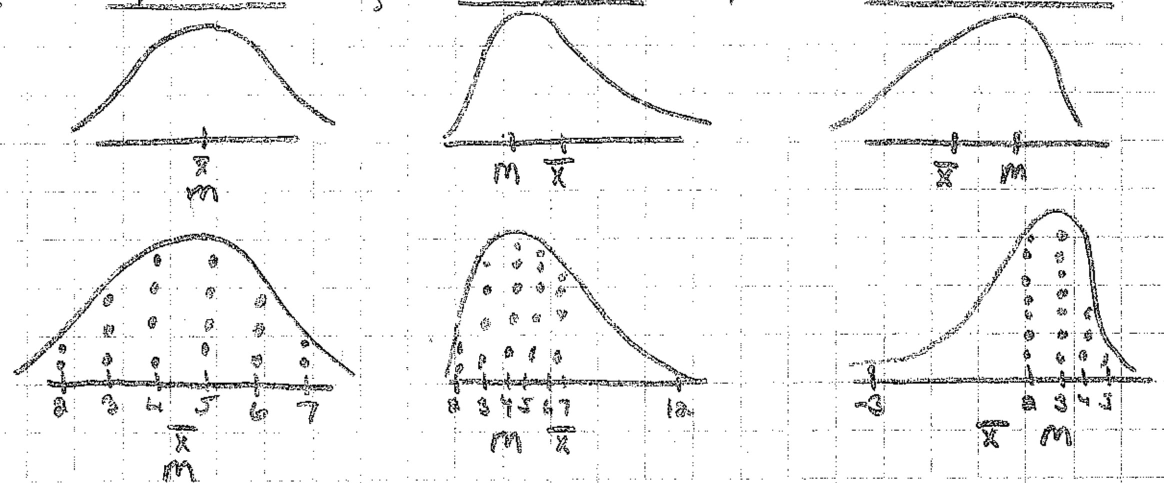
Obj: You will display data using dotplots, stemplots, + Histograms.

Discuss **Dotplot** – Pg 11 (2008, Section II)
2012

When Describing a Graph State:

Must Include:

- Center – Median (½ obs. above + Below) – Best (m)
  Mean (AVG) – sometimes not possible ($\bar{x}$)
- Shape – Symmetric, Skewed Rt, Skewed Left



- Spread – Variability (Give smallest + Largest values)
  Range (Largest – Smallest) ⇒ NOT Required
- Outlier – An observation that falls outside the overall pattern of the graph.

Gaps – when several consecutive measurable outcomes have no observations.

Stemplot –

Title

Stem	Leaf
1	0 1 1 4 5
2	0 3
3	1 4 6
4	2 3 7 8 8 9

Key: 4|2 means a person chewed 42 pieces of gum / week.

★ 5 stems are a good min (Avoid skyscrapers)

★ Round if necessary

Example 1.5

Histograms – Do example 1.6 ⇒ Determine: Center, Shape, Spread, outliers } 1-Var Stats

HW: 1.9 1.10 plus 2k 1.16 1.17 (Turn to pg 33)

Obj: You will construct an OGIVE and a Time Plot.

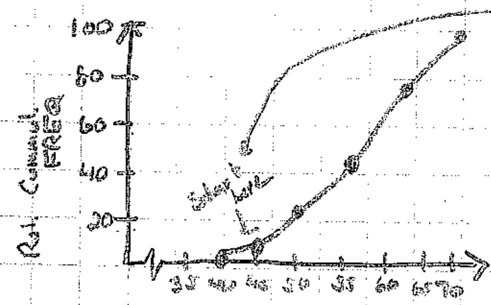
OGIVE (Relative cummulative frequency graph)
- used when you are interested in describing the relative position of an individual within a distribution.

You were in the 80th Percentile – Meaning?
- 80% of the students who took the test scored less than or equal to your score.
- The other 20% scored higher
- Think of a ladder – 80% up the ladder...

Example 1.8

Class	FREQ	Rel. FREQ	Cummul. FREQ	Rel. Cumm Freq	
40-44	2	2/43 = .047 = 4.7%	2	2/43 = 4.7%	(Before 45 yrs
45-49	6	6/43 = 14%	8	8/43 = 18.6%	(" 50 yrs)
50-54	13	30%	21	48.8%	(" 55 yrs)
55-59	12	27.9%	33	76.7%	(" 60 yrs)
60-64	7	16.3%	40	93.0%	(" 65 yrs)
65-69	3	7.0%	43	100.0%	(" 70 yrs)

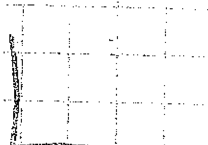
This % Before This age



Locate Bill Clinton – Percentile?
Locate 60th Percentile – Age?
Locate Q_1, M, Q_3 – Age?

2nd Stat Plot: [Scatter Plot | Mod Box Plot] [Time/Line | Box Plot] [Hist | Norm Prob]

Time Plot – Plots each obs. against the time it was measured.
- change over time
- Look for: overall pattern (up, Down) – A.K.A "Trend"
  Seasonal variation – Look @ Example 1.9



Store exercise 1.21 yr in L2, Deaths in L3 window [1935, 2000] [105, 280].

Discuss:
- Then Δ X MAX to 2040
- Then Δ X MIN to 1960, Xmax to 2000
- Then Δ X max to 2050, Xmin to 1935

HW: 1.22
Quiz Tomorrow

1.2 Describing Distributions w/ Numbers – Day 1

Obj: You will describe distributions by using the mean, median, quartiles and IQR (Interquartile Range)

I. Mean

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ } $\boxed{\bar{x} = \frac{1}{n}\sum x_i}$ ← on F.S.

Arithmetic Avg., not the # at a Typical variable!

Roger Maris' HR's:

1st 5 years: 13, 23, 26, 16, 33 $\bar{x} = 22.2$
1st 6 years: 13, 23, 26, 16, 33, 61 $\bar{x} = 28.7$
What if?: 13, 23, 26, 16, 33, 100 $\bar{x} = 35.2$

The mean is nonresistant because it is sensitive to extreme observations (May or May not be outliers)

II Median (m)

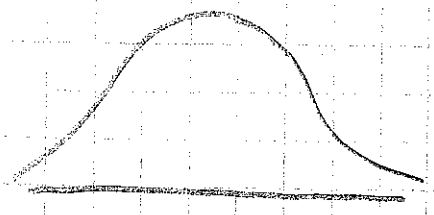
Roger Maris' HR's:

1st 5 years: 13, 23, 26, 16, 13 $m = 23$
1st 6 years: 13, 23, 26, 16, 13, 61 $m = 24.5$
What if?: 13, 23, 26, 16, 13, 100 $m = 24.5$

The median is resistant to extreme obs.

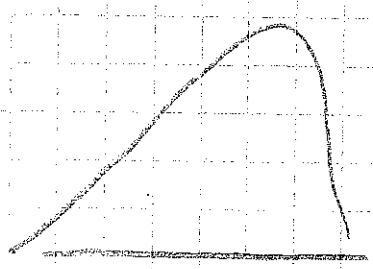
So... use the mean for All years.

or... use the median if you want the # of HR's in a typical year.



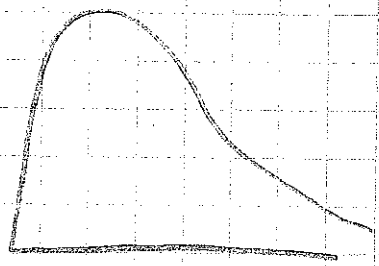
Symmetric

$\bar{x}$
m
mode



Skewed Left

$\bar{x}$ m mode



Skewed Right

mode m $\bar{x}$

III The Quartiles Q_1 and Q_3

Barry Bonds' HR's (In increasing order)
16, 19, 24, 25, 25, 33, 33, 34, 34, 37, 37, 40, 42, 46, 49, 73

What is the spread or variability of HR's?
What is the Range? (Low - High)
* But the #'s you used may be outliers!!

$M = 34$ } Do these #'s tell you anything about his
$\bar{x} = 35.4$ } yearly HR's?

The Quartiles improve the description.

Q_1 (The first Quartile) is 1/4 of the way up the list.
- Larger than 25% of the obs.
- The median of the observations to the left of the overall M.

Q_3 (The Third Quartile) is 3/4 of the way up the list.
- Larger than 75% of the observations.
- The median of the obs. to the right of the overall M.

16 19 24 25 | 25 33 33 34 | 34 37 37 40 | 42 46 49 73
(Q_1 = 25, M = 34, Q_3 = 42)
25% 25% 25% 25%

IQR - Interquartile Range ($Q_3 - Q_1$)
- The "middle half" of the data.

An obs. is not unusually high (upper 25%) or unusually low (lower 25%) if it falls between Q_1 & Q_3

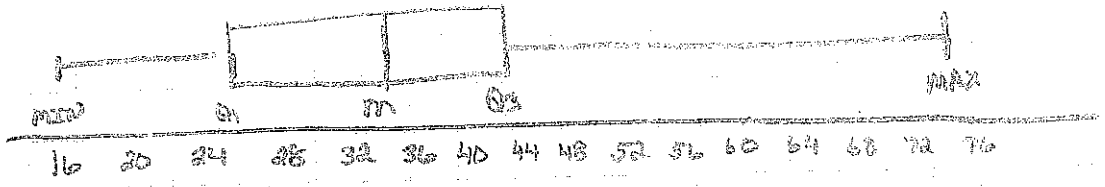
Outlier if: $< Q_1 - (1.5 \times IQR)$
$> Q_3 + (1.5 \times IQR)$

Lower cutoff
$< Q_1 - (1.5 \times IQR)$
$< 25 - (1.5 \times (41-25))$
< 1
Any #'s < 1? NO

Upper cutoff
$> Q_3 + (1.5 \times IQR)$
$> 41 + (1.5 \times (41-25))$
> 65
Any #'s > 65? Yes!! ∴ 73 is an outlier (Low HR)

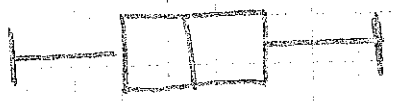
TL: Five-Number Summary and Box Plots

Bonds HR's in L1 → 1-Var STATS

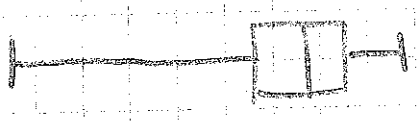


Barry Bonds' HR's

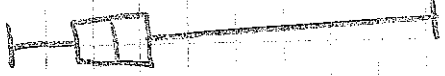
Symmetric



Skewed Left



Skewed Right



Modified Box Plot: Plots outliers as isolated points. (Reg Box Plot hides outliers)

Bonds HR's in L1
Ruths HR's in L2 (54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22)

- Box Plots (A window to suit) } Discuss
- Mod. Box Plots

HW: (1.34, 1.35, 1.37, 1.39) (Tomorrow to Pg 58)

1.2 Describing Dist. w/ Numbers – Day 2

OBS: You will use standard deviations to describe dist.

Act. Based Stats: Boxplots + Histograms (Matching Graphs to var)

IV Standard Deviation – Measures how far an observation is from the mean.

x_i | $\bar{x}$ | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$

$\sum(x_i - \bar{x})$ → Sum of deviations is always = 0

→ ∴ Square the deviations $(x_i - \bar{x})^2$
- obs. far from $\bar{x}$ will have large pos. squared deviations.
- But units will be squared ..... so →

s^2 = Variance = $\frac{\sum (x_i - \bar{x})^2}{n-1}$ → why n-1? You are intro error by not using σ, ∴ St. Dev inc; so... Reduce error by ÷ n-1

s = St. Deviation = $\sqrt{\text{Variance}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ * on F.S.

Example 1.14 Then show with x_i: 1, 3, 4, 9, 13

		Spread	$\bar{x}$	s_x	
L_1:	1, 2, 3, 4, 5, 6, 7, 8, 9	1–9 (8)	5	2.582	
L_2:	L_1 + 10	11–19 (8)	15	2.582	$\bar{x}$ Δ's, σ no Δ +/−
L_3:	L_1 − 10	−1–−9 (8)	−5	2.582	
L_4:	L_1 × 10	10–90 (80)	50	25.82	$\bar{x}$ Δ's, σ Δ's +/÷ +/÷ → High side
L_5:	L_1 ÷ 10	.1–.9 (.8)	.5	.2582	
L_6:	1, 2, 3, 4, 5, 6, 7, 8, 9, (9)	1–9 (8)	5.4	2.782	Away from $\bar{x}$, $\bar{x}$ + σ inc
L_7:	1, 2, 3, 4, 5, 6, (6), 7, 8, 9	1–9 (8)	5.1	2.4495	Close to $\bar{x}$; $\bar{x}$ inc, σ dec → high side

Degrees of Freedom (df): Test Scores

If you want a 90% test average for 1st Quarter (5 tests) and... $T_1 = 88$, $T_2 = 93$, $T_3 = 85$, $T_4 = 92$, $T_5 = ?$

n = 5 df: n − 1 = 5 − 1 = 4 df

- You are free to score "what you want" on $T_1 \rightarrow T_4$, But then you must receive a certain score such that $\bar{x} = 90$.

STOP ⟹

Properties of s

- s measures spread about the mean and should only be used when the mean is chosen as the measure of center.
- $s = 0$ when spread = 0; otherwise $s > 0$.
- s is nonresistant, outliers can make s very large.

5 # Summary - used for skewed distributions.
$\bar{x}$ with s - used for roughly symm. or symm. dist.

If you are comparing 2 graphs and one is skewed & the other is symmetric, use the 5 # summary for both.

1.43

Activity Based Stats: Histograms & $(\bar{x}, m, s)$

Side by side Bargraph (Pg 57) - used when comparing Cat. Data.
Back to Back stemplots (Pg 58) - used when comparing Quant. Data.

HW: 1.44-1.46, Read Ex 1.16 & 1.17, 1.50