

The Four Parts to Statistical Thinking

Data Analysis

exploring } graphs
organizing } numerical summaries
describing }

Data Production

producing data to get clear answers to specific questions
select samples
design experiments

Probability

used to describe chance, variation and risk
helps separate truth from distractions in data

Statistical Inference

draw conclusions about the big picture
understanding that variation is everywhere
conclusions are uncertain

Chapter 1: Exploring Data

The BIG ideas:

- exploratory data analysis
- describe what you see
- organize the data
 - categorical or quantitative?
 - examine each variable alone
 - study relationships among the variables

ALWAYS begin with a plot of your data...

Categorical (Qualitative):

- pie charts
- bar graphs, segmented bar, ribbon

Quantitative:

- dotplots
- stemplots
- histograms

Numeric Summaries

1.1 Displaying Distributions with Graphs

Categorical - (objects are classified):

Data can be:

nominal

ordinal

binary

Hair color

- blonde, brown, red, black, etc.

Race

- Caucasian, African-American, Asian, etc.

Smoking Status

- smoker, non-smoker

Political beliefs

- liberal, moderate, conservative

Class

- freshman, sophomore, junior, senior

Quantitative - (objects are measured):

Data can be

discrete

continuous

Examples:

Cholesterol level

Height

Age

SAT score

Number of students late for class

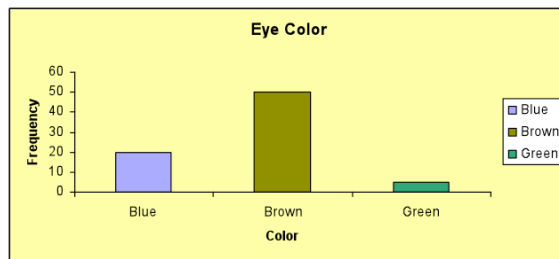
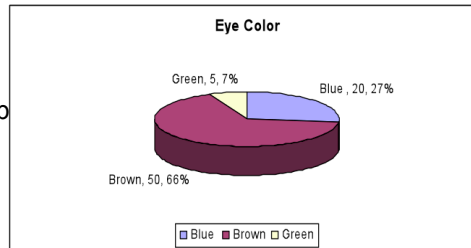
Time spent on homework

Example of single variable categorical data:

Eye Color	BLUE	BROWN	GREEN
Frequency (COUNTS)	20	50	5
Relative Frequency	$20/75 = .27$	$50/75 = .66$	$5/75 = .07$

Pie Chart:

- % or counts must add up to total of data (100%)
- labeled and titled



Bar Graph:

- separate bars for each category
- 2 axes - clearly labeled & scaled
- can show frequency (counts) or relative frequency (%)
- labeled and titled

Graphs for Quantitative Data (these two are good for smaller data sets):

dotplot: Labeled & titled

horizontal line (like a number line)

a mark above each number to represent each observation



Stemplot: labeled & titled

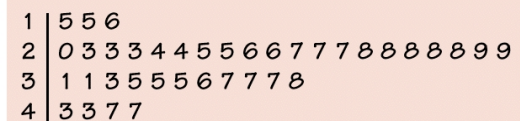
stems represent bulk of number, while leaves represent last digit

can split stems under certain conditions

must include a key (2|0 means 20 mg of caffeine/8 oz)

no fixed number of stems, but 5 is a good minimum

can round the data to make a stemplot more suitable



(a)

Remember your SOCS



S - Shape

symmetry: symmetric, left (negative) skewed, right (positive) skewed
 modes (peaks): unimodal, bimodal, uniformly distributed (no peaks)

O - Outlier

an individual observation that falls outside the overall pattern
 (we will learn a mathematical definition for this in the next section)

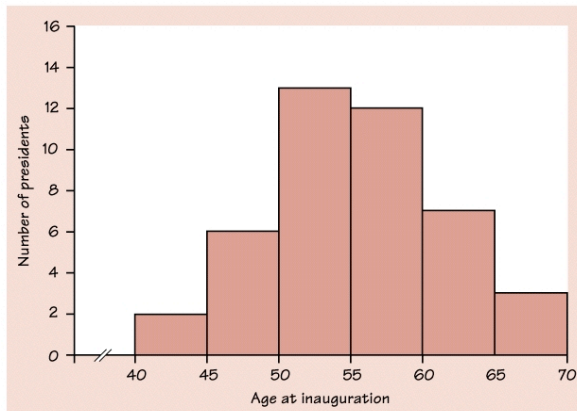
C - Center

The place in the graph where half of the data falls on either side

S - Spread

How far apart the low and high values in the data are

the histogram: (histogram:quantitative as bar graph:categorical)

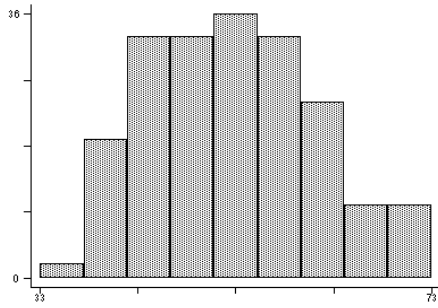


Things to note:

- the double hash mark shows a break in the axis
- minimum of 5 bins is good
- all bins equal width
- describe with SOCS
 - S - roughly symmetric and unimodal
 - O - no apparent outliers
 - C - typical age is about 55 (for now look for the place where 1/2 of data is on either side)
 - S - ages vary from 42 to 69

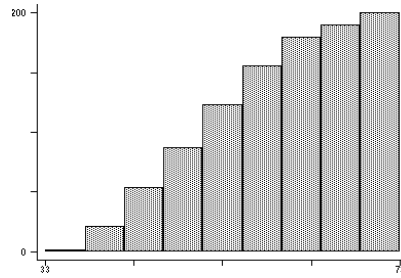
Frequency Histogram

(vertical axis in counts or frequency)



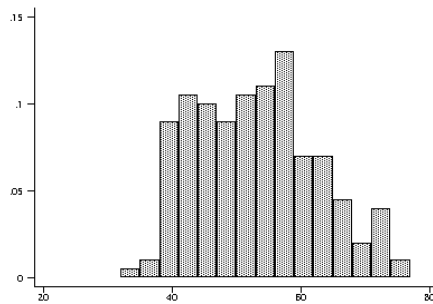
Cumulative Frequency Histogram

(change vertical axis to % and it becomes a Relative Cumulative Frequency Histogram)



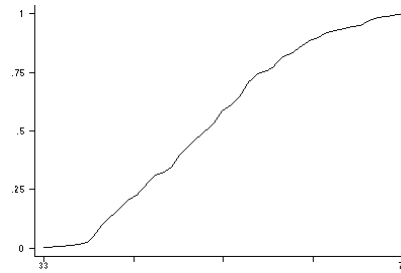
Relative Frequency Histogram

(vertical axis in %)



Relative Cumulative Frequency Graph (ogive)

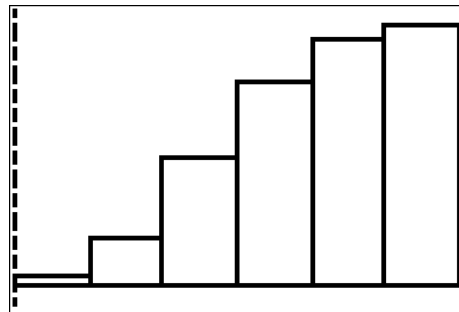
(can be mapped on top of a relative cumulative frequency histogram)
(vertical axis in %)



Cumulative Frequency - adding up the counts as you move through a data distribution

Cumulative Frequency Histogram (using PREZ data)

Class	Frequency	Cumulative Frequency
40 - 44	2	2
45 - 49	6	8
50 - 54	13	21
55 - 59	12	33
60 - 64	7	40
65 - 69	3	43



You can also add up the percents and graph the same histogram with **Relative cumulative frequency** along the vertical axis (should always go to 100%)

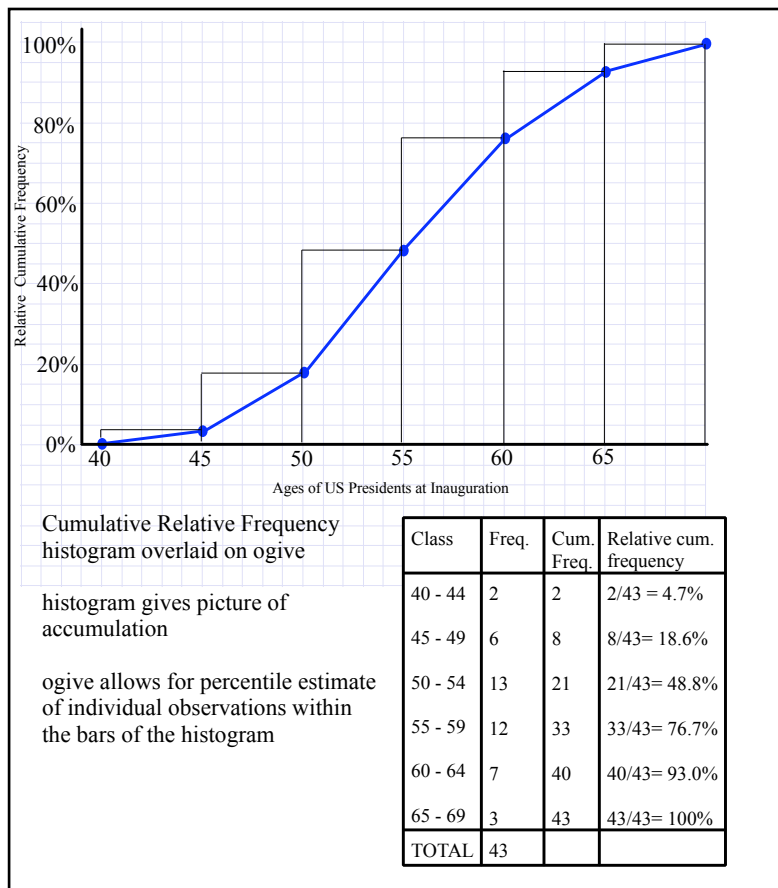
Sometimes we want a clearer picture of the relative standing of an individual observation so we construct an ogive or relative cumulative frequency graph.

first, though - need to understand percentile:

relative position of an individual observation in a distribution...

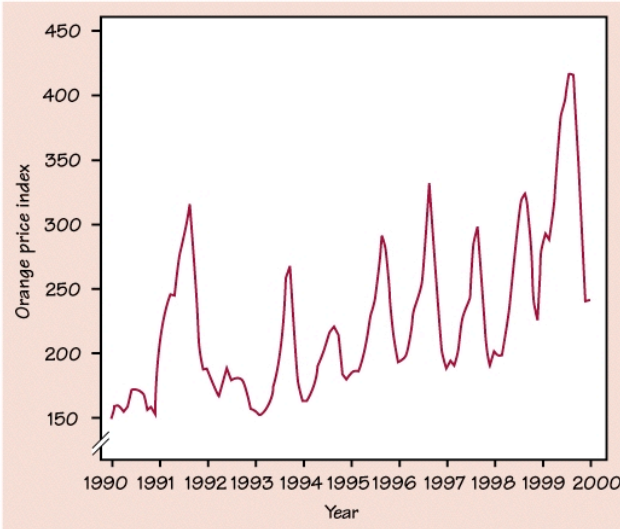
75th percentile - means the value of interest is at the spot in the distribution where 75% of the data were the same or lower

a well constructed ogive can help us determine percentile...



Time Plots: graph of data over time

- good for spotting overall and seasonal trends



fresh orange prices from 1990 to 2000

What is happening overall?
Is it a constant change?

When making a time plot:

- decide how to scale time axis (quarters of a year?)
- plot pairs of data as points on a graph
- connect (straight lines are fine)

1.2 Describing Distributions with Numbers

Numerical summary should include **center** and **spread**

Two ways to measure **Center**:

Mean (x-bar) \bar{x}

- use "x" to denote an observation
- \bar{x} to denote the mean or average of the values in a quantitative data set
- mean is not resistant

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

or

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

Median (M)

- midpoint of a distribution
- half smaller, half larger
- mean is resistant

1. sort list in ascending order
2. if n is odd, median is center observation
3. if n is even, median is the average between the center pair of observations

Find mean and median:

Hank Aaron's home runs, by year, through his last year with Atlanta:

13 27 26 44 30 39 40 34 45 44 24
32 44 39 29 44 38 47 34 40 20

Enter into L₁ and store as AARON for future use

SortA(L₁) STAT | 2 | L₁

Find median by counting observations and finding the halfway point

Now find mean with your calculator:

2nd | STAT | MATH | 3 | L₁

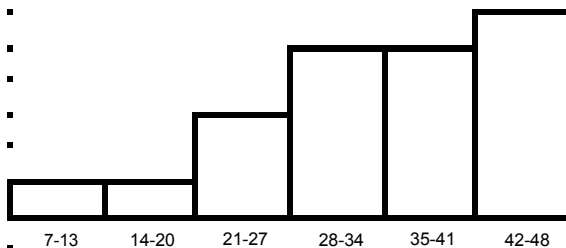
Repeat for Barry Bonds for 1986 - 2001

16 25 24 19 33 25 34 46 37 33 42
40 37 34 49 73

Enter in L₁ and store as BONDS

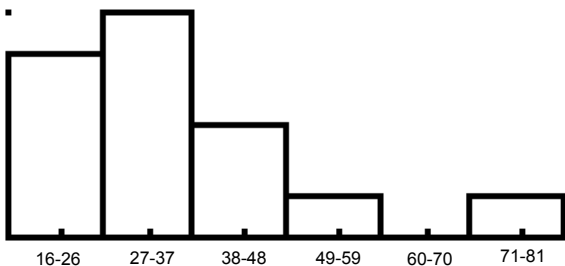
find median and mean

Hank Aaron's HR, by year



- enter data into L1
- 2nd|Y=|ZOOM|9|
- WINDOW X[7,49], Y[-2,9],
- GRAPH

Barry Bonds' HR, by year



X[16,82], Y[-2,9]

Comparing mean and median

Symmetric Distribution

- exactly symmetric: then mean = median
- roughly symmetric: then mean and median are close

Skewed Distribution

- skewed left: mean is lower (closer to left tail) than median
- skewed right: mean is higher (closer to right tail) than median



Spread:

spread is thought of as full range of data; low to high values

range is high value - low value, but what if one is an outlier?...

There are two other tools used for measuring spread:

Quartiles and Standard Deviation

Quartiles:

Data can be divided by number of observations into four groups. Three quartiles divide these groups:

Q1 - the 25th percentile (the median of the data left of the median)

Q3 - the 75th percentile (the median of the data right of the median)

the median M is Q2 (the 50th percentile)

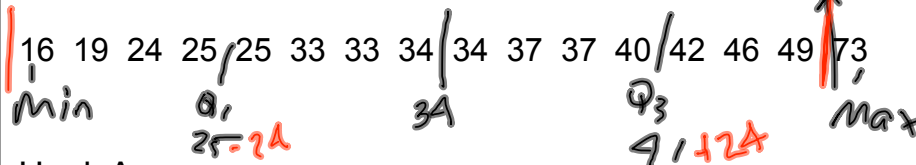
Interquartile range (another good measure of spread)

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Look at the two home run sets:

find M, Q1, Q3 and IQR

Barry Bonds:



Hank Aaron:

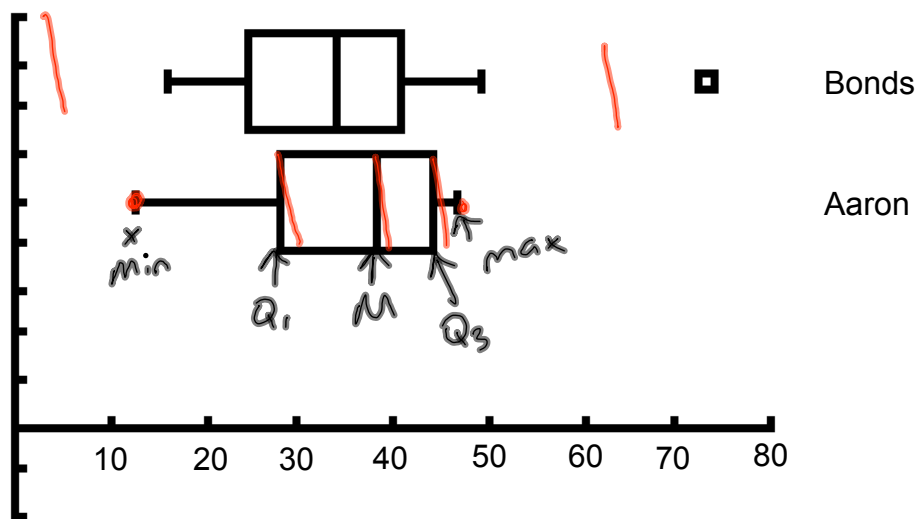
$16 \times 1.5 = 24$

13 20 24 26 27 29 30 32 34 34 38 39 39 40 40 44 44 44 44 45 47

Outliers: the numerical approach

- find IQR $Q_3 - Q_1$ $41 - 25 = 16$
 - multiply IQR by 1.5 $16 \cdot 1.5 = 24$
 - subtract result from Q1 (lower cutoff / fence) $25 - 24 = 1$
 - add to Q3 (upper cutoff / fence) $41 + 24 = 65$
- 24 is the key number

Now make a boxplot of each set of data on the same axis:



The Five-Number Summary: (pretty good description of center & spread)

Min Q1 Median Q3 Max

* 1-VAR stats STAT/CALC/1

With these five numbers, we can graph a boxplot

Standard Deviation:

Mean & Standard Deviation are the most common measures of center and spread

To calculate standard deviation (s), we have to start with the idea of variance (s^2):

We want to know how spread out each observation is from the mean of the data set. If we measure the distance alone, the positives and the negatives will essentially cancel each other out so a method was created to square each distance and find the average (the variance) and then take the square root of that variance to get the standard deviation. The standard deviation is a measure of the average distance of each data observation from its mean.

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

*n-1
degrees of
freedom*

OR

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

Which summary is best?

The **Standard Deviation** † *Center*

- is useful for symmetric distributions
(skewed distributions will have different spread on either side of the mean)
- is only valuable when paired with mean (since it is relative to mean)

The **Five Number Summary**

- is useful for skewed distributions
- or distributions with outliers

ALWAYS plot the data and study the SOCS of the distribution.

Example: Calculate the standard deviation

Record High temperatures for Orlando, FL:

Mar 33.3 Apr 35 May 37.2 Jun 37.7 Jul 37.7 Aug 37.7

start with mean: $36.4 (\bar{x})$

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

$$s^2 = \frac{(33.3 - 36.4)^2 + (35 - 36.4)^2 + (37.2 - 36.4)^2 + (37.7 - 36.4)^2 + \dots}{6 - 1}$$

$$s^2 = \frac{9.61 + 1.96 + .64 + 1.69 + 1.69 + 1.69}{5}$$

$$s^2 = \frac{17.28}{5} = 3.46$$

$$s = 1.86$$

then take the square root to find standard deviation:

Linear Transformations:

changing a distribution by:

- addition/subtraction of a constant to each observation
- multiplication by a constant to each observation
- or both

$$x_{new} = a + bx$$

big ideas:

- adding or subtracting "a" shifts the whole distribution up or down
 - does change center (mean and median)
 - doesn't change spread
 - doesn't change shape
- multiplying by "b" spreads the distribution out or tightens it
 - does change center
 - does change spread
 - doesn't change shape

change Orlando record temps from Celcius to Fahrenheit:

$$F = \frac{9}{5}C + 32$$

C	F
33.3	91.9
35	95
37.2	99
37.7	99.9
37.7	99.9
37.7	99.9

Notice that the center changes by the scalar (b) and the constant (a)

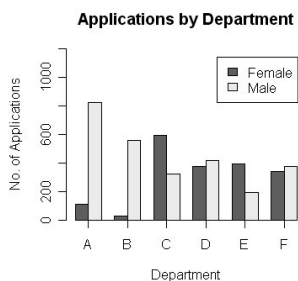
The spread changes by only $(9/5) * (4.4)$

4.4 is the original spread

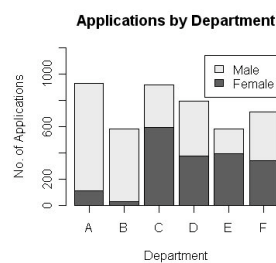
Comparing Distributions:

For Categorical:

Side by side bar graph

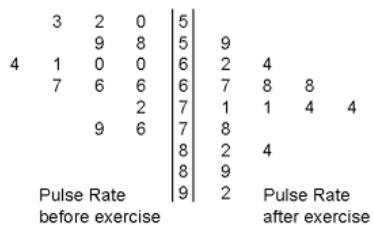


Segmented bar graph



For Quantitative:

Back to back stemplots



Side by side boxplots

