Stat 202-2015S W2-Tues

Review The distribution of a variable tells us what values it takes and how often it takes these values.

Exploratory Data Analysis

1) Study each variable by itself
then 2) Study relationships among variables

Each step has the same substeps
(a) start with a graph or graphs
then (b) move on to numerical summaries.

Graphs for single categorical variables

* pie chart } shows distribution } what value variable takes and how often it takes
* bar graph }

IQ: these variables can you see this?

Graphs for quantitative variables

* stemplots - homework 2
* histograms - new today

Stemplots - gives a quick picture of the shape of a distribution while including actual numerical values in the graph

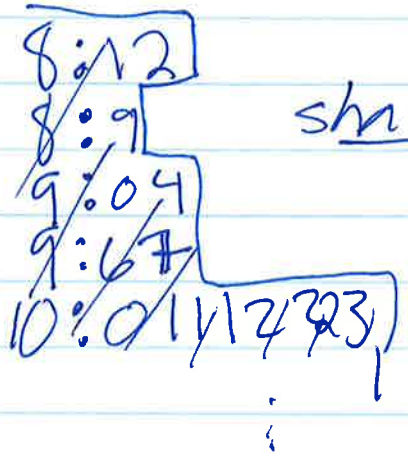
What values variable take and how often it takes these values, IQ: Can you see this?

Stat 202 - 20155 - W2 - Tues (pg 2)

Load Iq data set
Plot Histogram (bin width 5) / Plot Stemplot (value above bar)

What is the relationship between a stemplot and a histogram?

You can make a histogram from a stemplot as follows



Histograms are conventionally drawn vertically not horizontally

Do histograms with software.

When do you prefer histograms over Stemplots?

- When you have lots of data (10K data points impractical with stemplot)

This histogram has a bin width of 5
That means

First bar is 80 - 85 | really - 84.9999
Second bar is 85 - 90 | - 89.9999
etc

this is the width of the bins

For stemplots the bin width is determined by the spacing of the digits

The counts of observations in bins are called Frequencies

80 - 85	2
85 - 90	1
90 - 95	2
95 - 100	2
100 - 105	8
⋮	

this is called a frequency table

A histogram is essentially a bar graph of the frequency table EXCEPT that customarily we do not put spaces between bars in a histogram. (we do for bar graphs)

Stat 202-20155-W2-Tues (Pg 4)

The height of a bar can be (in StatCrunch)

frequency

relative frequency $\left(\frac{\text{Frequency}}{\text{total \# of data points}}\right)$

density $\left(\frac{\text{relative frequency}}{\text{bin width}}\right)$

For these choices shape of histogram doesn't change, only vertical scale

Choice of width

no one right choice

too large (E.g. 150) "Skyscraper" (if bins are drawn thin)

too small (eg 1) "pancake" (if not integers may be only one observation, in each bin, most have none)

neither of these results is desirable

Use your best judgment

Note: changing width of bins changes shape of histogram

Stat 202 2015S-W2-TUES

Another difference between bar graphs and histograms –

A bar graph need not have any measurement scale on the horizontal axis (if the variable is categorical). It simply identifies the items being compared.

A histogram's horizontal scale indicates possible values of a quantitative variable to be covered – no space between bars indicates all values covered. Horizontal scale of a histogram usually has units.

Look at call center

Stat 202-20155 - W2 - Tues (pg 6)

How do you analyze a data set like this

(1) • Understand the background of the data set

(cases, variables, units of measurement

(calls, length of call, seconds)

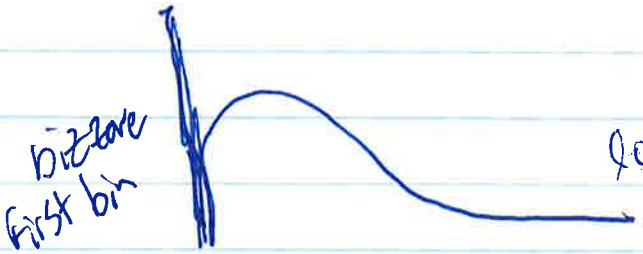
(2) • Plot data

(3) • Look for an over all pattern

bin width 10

where length ≤ 100 (≤ 1000 | (≤ 5000) (≤ 2000)

• Look for striking deviations from the pattern



explain 1st bin

Stat 202-20155-W2-Tues

- Describe overall pattern in terms of
  - shape
  - center
  - spread

Shape • number of peaks or modes
• symmetric or skewed

IQ had one mode (unimodal)
Call center had two (bimodal)

IQ was symmetric about midpoint
Call center wasn't

Symmetric meaning values smaller and larger than midpoint mirror images on histogram.

Skewed to right – Right (or upper) tail much longer than left (or lower)

Skewed to left – analogous

Tail – Extreme values are said to be within tails of distribution

Stat 202 - 2015S - W2 - Tues (P98)

Center - midpoint half the values above half values below,
Also (better) called Median,

Can compute median with Statcrunch

Spread - difference between min and max values
called range in Statcrunch

Can compute same way.