Data Types

MULTI-variate (multiple variables)

Uni-variate (one variable)

Bi-variate (two variable)

CATEGORICAL **(Qualitative)** Sex (M/F), Eye Color (Gr, Bl, Br, Hz), State (Al, Az, Ca)

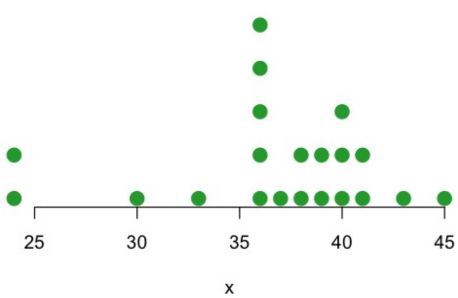
NUMERIC **(Quantitative)** **Discreet** (Whole Number, Student Grade on 5Q Test, Times to pass Drv Test)
Continuous (Uncountable Number, SPLdb, Time between 1st&2nd Place)

(B/B) Stem and Leaf Plot
(Putting Data into Bins and Stacks

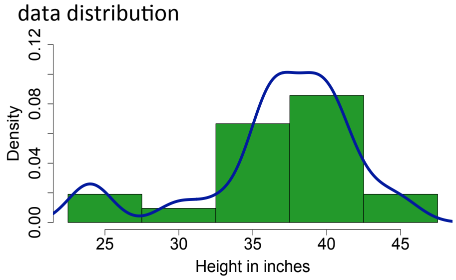
38 24 40 36 36 41 38
24 40 41 45 37 36 36
39 40 36 43 33 39 30

4 4	**2**	44
998876666630	**3**	036666678899
531100	**4**	0001135

Strip Chart Instead of stems **x-axis**. Rather than bins, identical items **stacked**. **Dots** than digits



Histogram (Good for Large Data sets)
Density estimate can capture what happens with more data, displaying smooth curve showing the shape of the data distribution



Descriptive Statistics

Central Tendency Measures Relates to the way in which quantitative data is clustered around some value. A measure of central tendency is a way of specifying - central value.

MEAN is often used to report central tendencies, it is not a robust statistic, meaning that it is greatly influenced by outliers. Notably, for skewed distributions, the mean may not accord with one's notion of "middle", and robust statistics such as the median may be a better description of central tendency.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{or} \quad AM = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

Median is described as the numerical value separating the higher half of a sample, a population, or a probability distribution, from the lower half

$$Q_2 = \begin{cases} x_{\frac{n+1}{2}} & \texttt{if } n \texttt{ is odd} \\ Mean(x_{\frac{n}{2}}, x_{\frac{n}{2}+1}) & \texttt{if } n \texttt{ is even} \end{cases}$$

Mid-Range of a set of statistical data values is the arithmetic mean of the maximum and minimum values in a data set

$$M = \frac{\max x + \min x}{2}$$

Variance of a random variable or distribution is the expectation, or mean, of the squared *deviation* of that variable from its expected value or mean. Thus the variance is a measure of the amount of variation of the values of that variable, taking account of all possible values and their probabilities or weightings (not just the extremes which give the range).

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

Deviation xi = xi - x

Standard deviation is a widely used measure of variability or diversity used in statistics and probability theory. It shows how much variation or "dispersion" there is from the average (mean, or expected value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data points are spread out over a large range of values.

$$\sqrt{s^2}$$

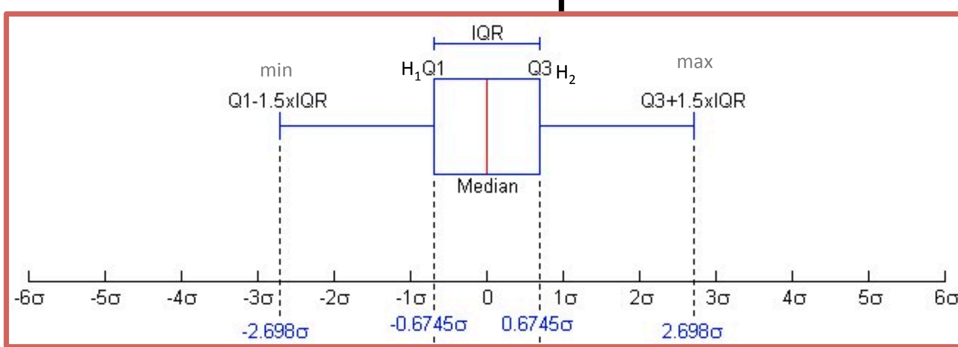
Interquartile range (IQR), also called the midspread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles.

IQR = Q3 – Q1

Unlike (total) range, the interquartile range is a robust statistic, having a breakdown point of 25%, and is thus often preferred to the total range. The IQR is used to build box plots, simple graphical representations of a probability distribution. For a symmetric distribution (so the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD). The median is the corresponding measure of central tendency.

Z-scores indicates how many standard deviations an observation or datum is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing or normalizing

$$z_i = \frac{(x_i - \bar{x})}{s}$$



One Categorical Variable results in a..

Frequency Table!

Clear	Partly Cloudy	Cloudy
11	11	9

And…Relative Frequency Table!

Clear	Partly Cloudy	Cloudy
.36	.36	.3

(Divide each count by relative sample, (in this case 3))

Two or more result in a….

Sex	Computer
M	PC
F	Mac
F	PC
M	PC
F	PC
F	Mac
M	Mac
M	PC

$\Longrightarrow$

Contingency Table! (Joint Frequencies…)

		Computer	
		PC	Mac
Sex	M	3	1
	F	2	2

Marginal Frequencies for Computer

		Computer		Marginal
		PC	Mac	
Sex	M	3	1	4
	F	2	2	4
Marginal		5	3	

Marginal Frequencies, (for SEX)

		Computer		Marginal
		PC	Mac	
Sex	M	3 +	1 =	4
	F	2 +	2 =	4

Calculating Joint and Marginal Proportions…

1st…		Computer		Marginal
		PC	Mac	
Sex	M	3	1	4
	F	2	2	4
Marginal		5	3	$n = 8$

Divide each frequency by n (here $n = 8$):

2nd…		Computer		Marginal
		PC	Mac	
Sex	M	$3/n$	$1/n$	$4/n$
	F	$2/n$	$2/n$	$4/n$
Marginal		$5/n$	$3/n$	n/n

- Each color represents a different distribution.
  - Lavender: Joint distribution of Sex and Computer
  - Pink: Marginal distribution of Sex
  - Lime Green: Marginal distribution of Computer

Finally..		Computer		Marginal
		PC	Mac	
Sex	M	0.375	0.125	0.500
	F	0.250	0.250	0.500
Marginal		0.625	0.375	1.000

- Notice that each distribution sums to 1.

Conditioning on Sex…

1st…		Computer		Marginal
		PC	Mac	
Sex	M	3/4	1/4	4/4
	F	2/4	2/4	4/4
Marginal		5/8	3/8	8/8

The resulting conditional proportions make up *two different* conditional distributions: one for each sex.

Finally…		Computer		Marginal
		PC	Mac	
Sex	M	0.75	0.25	1.00
	F	0.50	0.50	1.00
Marginal		0.625	0.375	1.00

Notice that, as before, *each distribution* sums to one (but now each row is its own distribution: we consider males and females separately).

Divide each frequency by the total *for that computer*:

Conditioning on Computer…

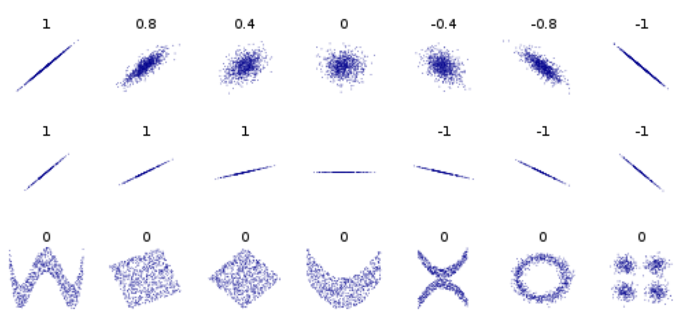
1st…		Computer		Marginal
		PC	Mac	
Sex	M	3/5	1/3	4/8
	F	2/5	2/3	4/8
Marginal		5/5	3/3	8/8

Finally…		Computer		Marginal
		PC	Mac	
Sex	M	0.60	0.33	0.50
	F	0.40	0.67	0.50
Marginal		1.00	1.00	1.00

- Notice that the rows do *not* form distributions: the marginal distribution of sex is a *weighted average* of the conditional distributions (what are the weights?)

We define **Pearson's Correlation Coefficient** as:

$$r = \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i}$$



- We can define **Spearman's Rank Correlation** in the exact same way as Pearson's, just using the ranks instead of the values:

$$\rho = \frac{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})(\text{Rank}(y_i) - \frac{n+1}{2})}{\sqrt{\sum_{i=1}^{n}(\text{Rank}(x_i) - \frac{n+1}{2})^2 \sum_{i=1}^{n}(\text{Rank}(y_i) - \frac{n+1}{2})^2}}$$