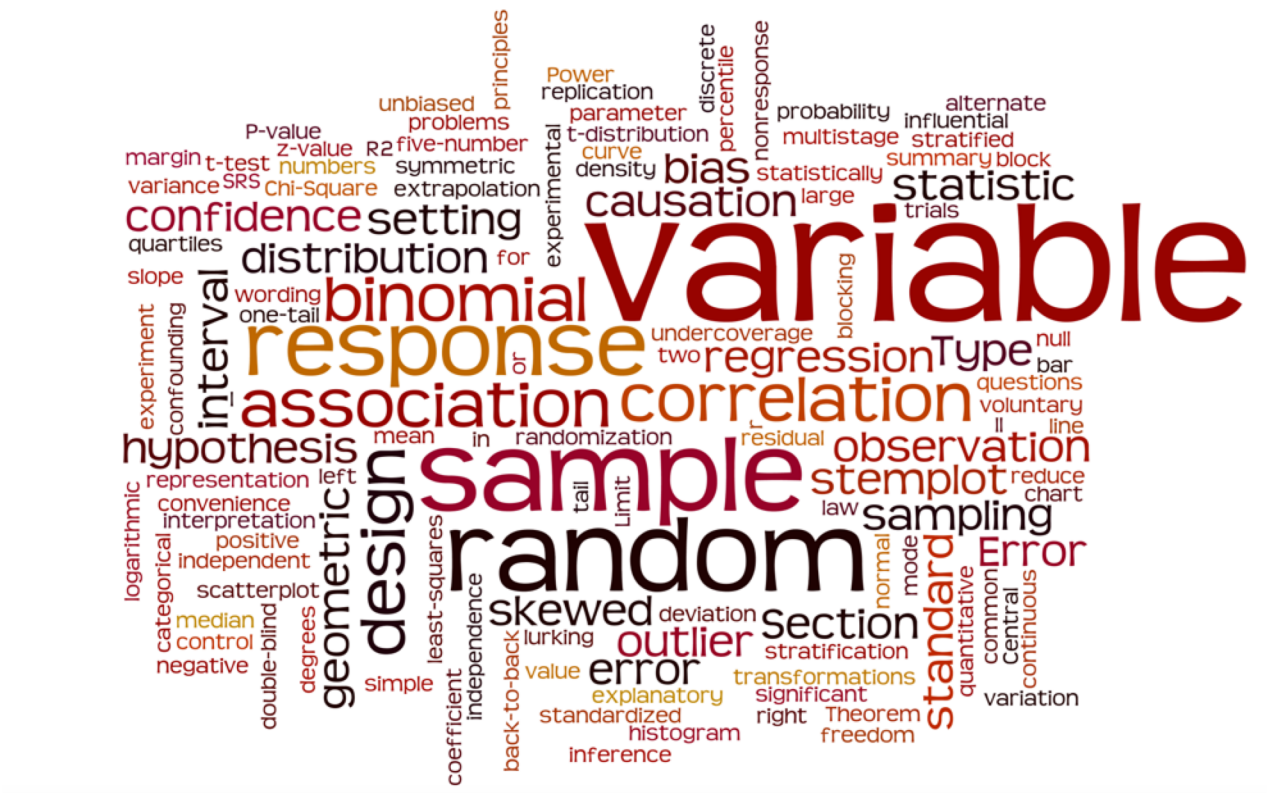
Introduction to Quantitative Methods



Kristin Eccles

How to evaluate these variables?

Table 1. Canadian forestry statistics expressed as area by year.
Source: http://cfs.nrcan.gc.ca/statsprofile/forest/ca

Year	Certified (millions of ha)	Harvested (millions of ha)	Planted (millions of ha)	Seeded (millions of ha)	Insect defoliation (millions ha)	Burned (millions of ha)	Drought Year
2001	17.17	1.02	0.47	0.02	22.6	0.63	No
2002	28.22	1	0.45	0.02	20.66	2.77	Yes
2003	58.39	1	0.44	0.05	20.47	1.74	No
2004	86.47	1.01	0.43	0.02	13.12	3.18	Yes
2005	119.77	1.06	0.45	0.02	16.32	1.67	Yes
2006	123.75	0.83	0.47	0.02	19.94	2.26	Yes
2007	137.89	0.79	0.48	0.03	18.68	1.54	No
2008	145.75	0.69	0.44	0.02	14	1.72	Yes
2009	142.78	0.61	0.4	0.02	15.21	0.78	No
2010	149.84	0.69	0.38	0.01	12.82	3.05	Yes
2011	150.57	0.64	0.37	0.01	9.19	2.43	Yes

....For comparison – Ottawa is 0.277 million hectares

Outline

Descriptive Statistics	Metric
Central Tendency	• Mean • Median • Mode
Variation	• Range, Quartiles, IQR, Quantiles • Boxplots and 6-number summaries • Deviation / squared deviation • Sum of squares • Variance, standard deviation • Degrees of freedom • Coefficient of Variation
Distributions	• Skewedness • Kurtosis

What is central tendency?

- Metrics that represents the center or typical value of a frequency distribution
- Completed for each individual variable

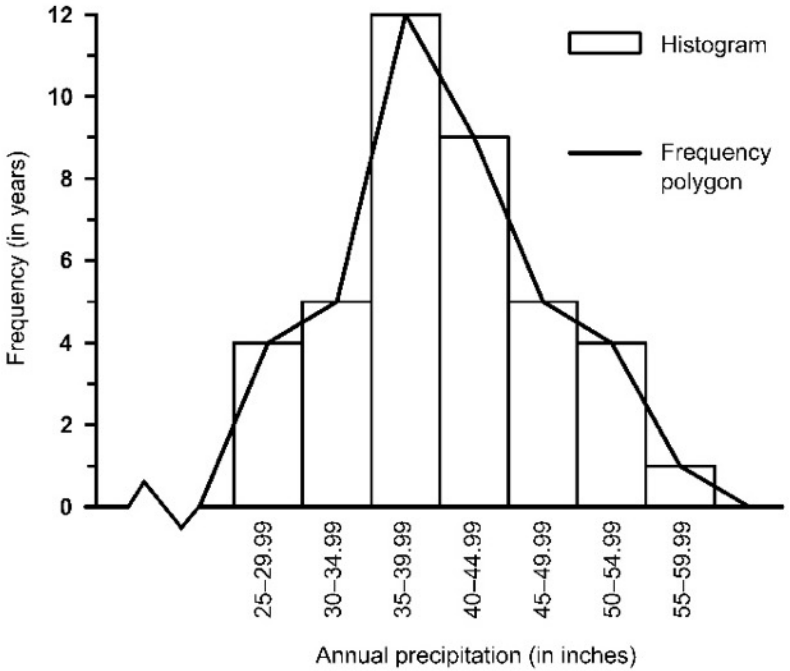
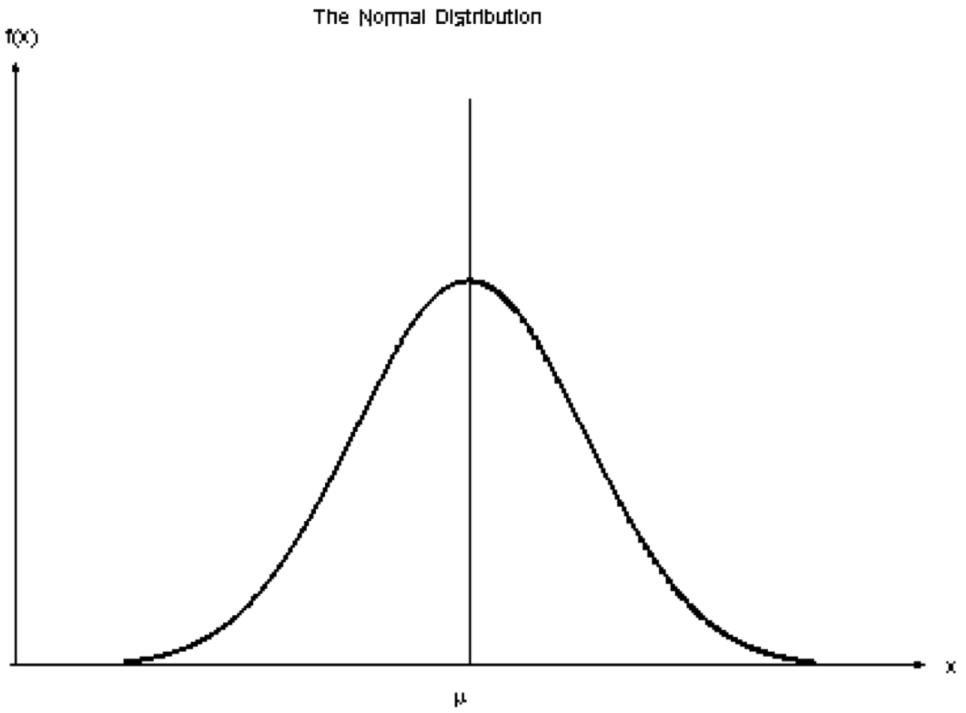


FIGURE 3.1
Histogram and Frequency Polygon for Washington, DC 40-Year Annual Precipitation Data
Source: National Climatic Data Center (NCDC)



AKA Gaussian Distribution, Bell Curve

Central tendency

How do we measure central tendency?

- Arithmetic **mean** (or simply the mean)
- **Median**
- **Mode** (or modal average)

Why do we care?

- The distribution of the data determines what statistical tests we can use (parametric= normal, non-parametric= non-normal)

Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad = \quad \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

Sum all values, then divide by the number of observations

Formula:

- *x* represents a variable
- x_i is an individual value (from *x*)
- *x-bar* is the 'mean' of *x*
- *n* is the number of observations (the sample size)
- Σ (=Sigma) is the 'sum of' is the sum of *x* from first value (*i=1*) to last (*i=n*)

Mean

Population Mean

$$\mu = \frac{\sum x}{N}$$

Sample Mean

$$\bar{X} = \frac{\sum x}{n}$$

μ= Greek letter mu

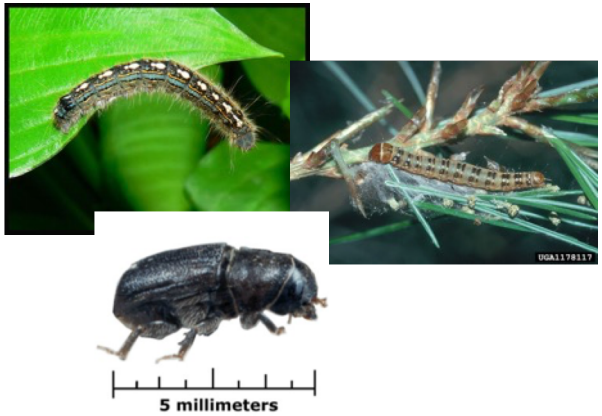
Median

$$\tilde{x} = \left((n+1) \div 2\right)^{th} sorted\ value$$

- The middle value of a set of ordered data.
- Half the observations are above & half are below the median
- Formula:
  - median = $\tilde{x}$ (tilde over *x*)
  - *n* = number of values
- What if n is an even number?
  - compute the mean of the *middle*

Mode

- The most frequently occurring value
  - Look through and count (easiest to sort the data first)
  - **Only meaningful with discrete variables (insect?)**
- But for continuous variables:
  - Can identify the modal class – the bin in the histogram that has the highest frequency



Calculate mean, median & mode

- Survey response 'Do you like statistics?' 1=strongly disagree,... 5=strongly agree
  - Data: 1,2,2,3,4,1,3,3,5
    - Mean: 2.66
    - Median: 3
    - Mode: 3

  Variable type: _____?

- City populations in millions
  - Data: 5.11, 1.03, 1.13, 1.08, 0.72, 0.69
    - Mean: 1.63
    - Median: 1.05
    - Mode: 5.11, 1.03, 1.13, 1.08, 0.72, 0.69

  Variable type: _____?

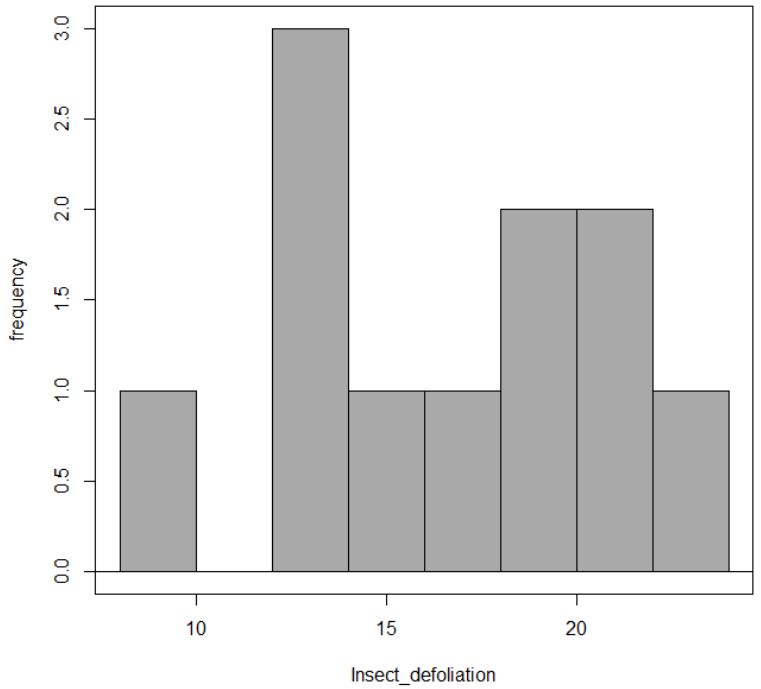
- Favourite colour: (1=blue, 2=red, 3=green, 4=other)
  - Data:1,4,3,1,2,1,2,3,4,4,1
    - Mean: 2.36
    - Median: 2
    - Mode: 1

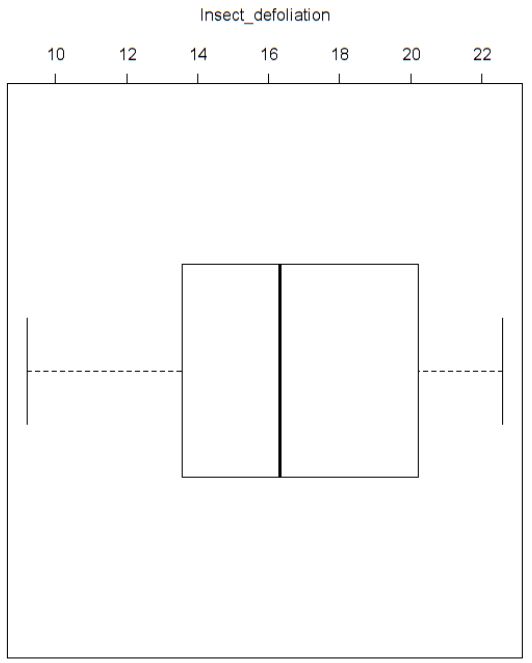
  Variable type: _____?

- Does it make sense to use all 3 averages in all situations?

Descriptive Statistics in R

```
> summary(Dataset)
      Year        Certified       Harvested        Planted          Seeded        Insect_defoliation     Burned       Drought
 Min.   :2001   Min.   : 17.17   Min.   :0.6100   Min.   :0.3700   Min.   :0.01000   Min.   : 9.19   Min.   :0.630   No :4
 1st Qu.:2004   1st Qu.: 72.43   1st Qu.:0.6900   1st Qu.:0.4150   1st Qu.:0.02000   1st Qu.:13.56   1st Qu.:1.605   Yes:7
 Median :2006   Median :123.75   Median :0.8300   Median :0.4400   Median :0.02000   Median :16.32   Median :1.740
 Mean   :2006   Mean   :105.51   Mean   :0.8491   Mean   :0.4345   Mean   :0.02182   Mean   :16.64   Mean   :1.979
 3rd Qu.:2008   3rd Qu.:144.26   3rd Qu.:1.0050   3rd Qu.:0.4600   3rd Qu.:0.02000   3rd Qu.:20.20   3rd Qu.:2.600
 Max.   :2011   Max.   :150.57   Max.   :1.0600   Max.   :0.4800   Max.   :0.05000   Max.   :22.60   Max.   :3.180
```





Which average to use?

- Mean
  - Used most often
  - Best for interval/ ratio level data
  - Basis of more advanced statistics
  - Sensitive to outliers
- Median
  - Not sensitive to outliers
  - Not used as much as mean but
  - Should be used variables where there are extreme values (where data is skewed)
    - E.g. income
- Mode
  - Not sensitive to outliers
  - But not in widespread use
  - More useful for nominal data

Quantifying Variation

Data range

- Gives an idea of the spread (or dispersion or variation) in the data
- Maximum value
  - 1.06
- Minimum value
  - 0.61
- Range
  - = max. value – min. value
  - = 1.06 – 0.61
  - = 0.45

Year	Harvested
2001	1.02
2002	1
2003	1
2004	1.01
2005	1.06
2006	0.83
2007	0.79
2008	0.69
2009	0.61
2010	0.69
2011	0.64

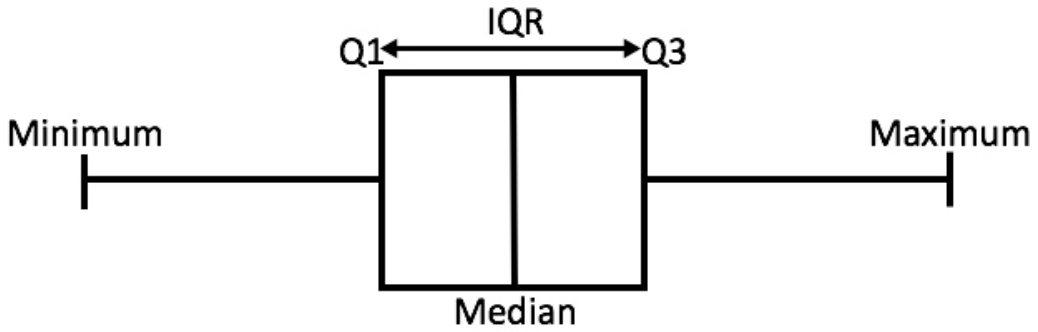
Sort →

Year	Harvested
2009	0.61
2011	0.64
2010	0.69
2008	0.69
2007	0.79
2006	0.83
2003	1.00
2002	1.00
2004	1.01
2001	1.02
2005	1.06

Quartiles

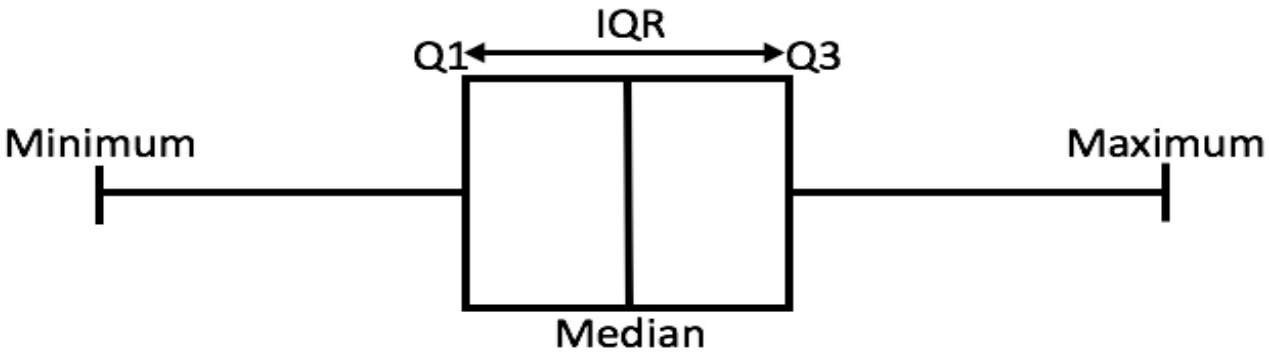
- The range is a measure of spread that uses the most extreme values
- What about sorting the data and looking for:
  - Min., Median and Max **AND**
  - the value ½ way between median and min (¼ along)
  - the value ½ way between max and median (¾ along)

- These are known as the quartiles:
  - Q1 =First quartile: 25% of the data are below Q1
  - Q2 =Second quartile: 50% of the data are below Q2
  - Q3 =Third quartile: 75% of the data are below Q3



Interquartile range

- The range between Q3 and Q1 is the interquartile range (IQR)
- A measure of dispersion (or spread) but not as extreme as the range – more like 'mid-spread'
- 50% of the (the least extreme) data are in the IQR
- For example... area harvested:
  - Max. = 1.06
  - Min = 0.61
  - Range = 0.45
  - Q1 = 0.69
  - Q2 (Median) = 0.83
  - Q3 = 1.01
  - IQR = 0.32



Quantiles

Extend the concept of quartiles to:

Quintiles (divide data into 5 parts)

Deciles (divide data into 10 parts)

Percentiles (divide data into 100 parts)

Quantiles (divide data into ___ parts)

Quartiles, and all the above are **quantiles**

$$\boldsymbol{rank = quantile \times n + \frac{1}{2}}$$

Where *quantile* is expressed as a fraction and *n* is the number of observations
Take the value at nearest rank

Area harvested:

90th percentile (9th decile) is 1.02

40th percentile (4th decile) is 0.79

Harvest in 2011: 0.64 million ha = 2nd/11 years = 18th percentile

Harvest in 2001: 1.02 million ha = 10th/11 years = 91st percentile

Boxplot

A graph showing the central tendency and the dispersion of the data in a given variable

Box is Q1 to Q3 with a line for median

Whiskers are placed at:

Q3 + (IQR x hinge value)

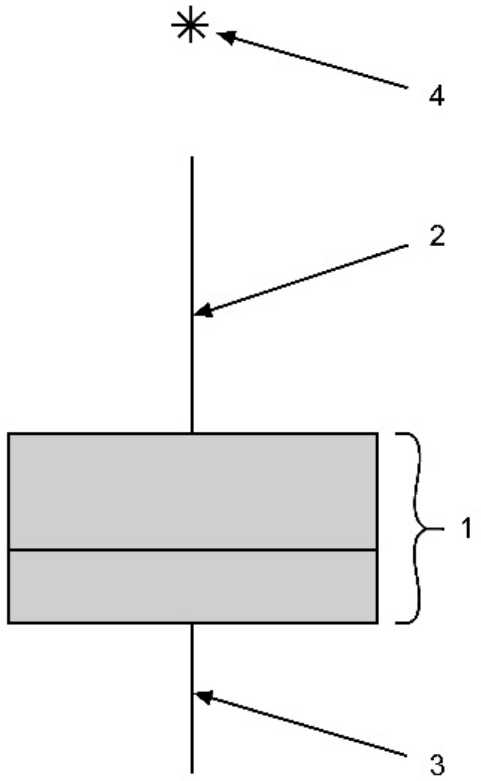
Q1 – (IQR x hinge value)

The **hinge value** is often 1.5 (but you could chose a different value)

BUT: Whiskers are ***never*** **more extreme than max./min.** (so place the whiskers at max/min in that case)

Outliers (extreme values relative to the others) are indicated by symbols (any value that may be more extreme than the whiskers)

Boxplot



1 - Interquartile range box

- Top line - Q3 (third quartile) 75% of the data are less than or equal to this value.
- Middle line - Q2 (median) 50% of the data are less than or equal to this value.
- Bottom line - Q1 (first quartile) 25% of the data are less than or equal to this value.

2 - Upper whisker

Extends to the maximum data point within 1.5 box heights form the top of the box.

3 - Lower whisker

Extends to the minimum data point within 1.5 box heights from the bottom of the box.

4 - Outlier (✳)

Observation that is beyond the upper or lower whisker.

FIGURE 3.5
Generalized Diagram of a Boxplot

5 or 6 number summary

Boxplots are a visual version of a table such as:

Table 1. Descriptive statistics for forestry variables. Source: Canadian Forestry Service

Variable	Min.	Q1	Median	Mean	Q3	Max.
Harvested area (millions of ha)	0.61	3.25	0.83	0.85	1.01	1.06
Planted area (millions of ha)	0.37	0.40	0.44	0.43	0.47	0.48

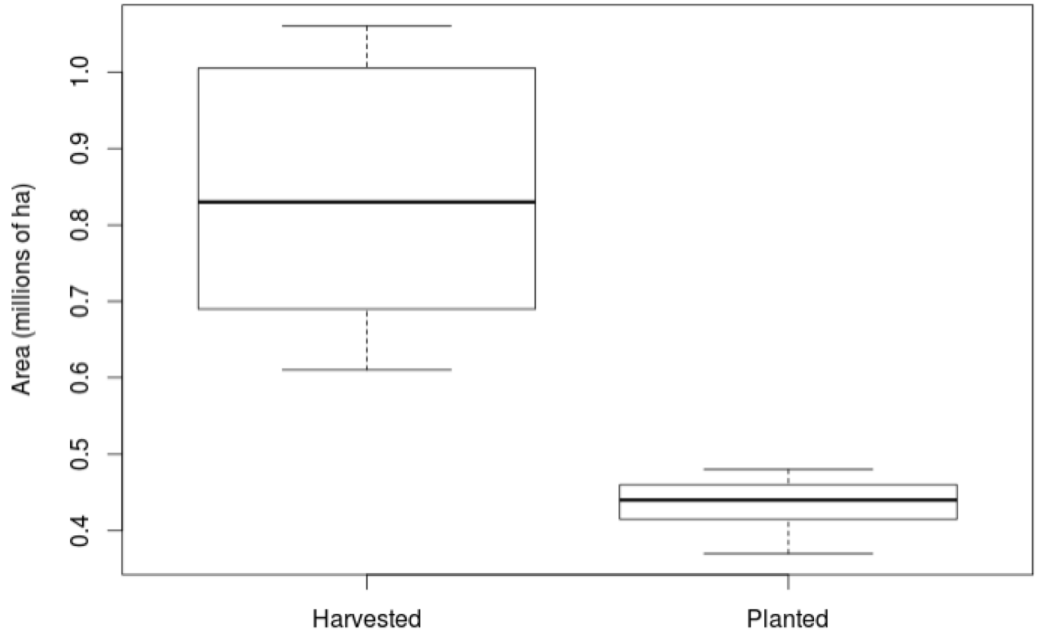


Figure 1. Boxplot of area harvested and planted in Canada from 2001 to 2011. Source: Canadian Forestry Service

Deviation

- How much does a value [interval/ratio] deviate from the 'norm'?
- If the norm is defined as the mean, then we should write:

$$d_i = x_i - \bar{x}$$

  - Where di = deviation
  - x_i is the value in question
  - X-bar= _____?

- Great for comparing values to the mean:
  - Higher value than average: deviation is positive
  - Lower value than average: deviation is negative
- Not great to show the overall variation (spread) of the variable since
  - Sum of all deviations in a variable is 0
  - Average of all deviations in a variable is 0

Squared deviations / sum of squares

- So how to describe the deviation of an entire variable?
- Could square each deviation to remove negative numbers, like:

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

- These new squared deviations can be summed:

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

- This is known as the sum of squares

Variance

- Sum of squares can be large or small depending:
  - On the variation (aka spread, dispersion, etc.)
  - On the number of observations in the variable

- Eliminate the second factor by dividing by the number of observations (n)... actually by n-1

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

- Known as the variance (s^2)
  - Essentially the mean of the squared deviations

- Why n-1 and not n?

Degrees of freedom

- Why n-1 and not n?
- Briefly and intuitively:
  - Variables with fewer observations are error-prone
  - Small sample size
  - This means variance is underestimated in small samples
- Correct this with n-1
  - Increases s^2 at small n
  - Yet doesn't really influence anything at large n
- Formally:
  - Degrees of freedom is the number of values that are free to vary in the calculation

Standard deviation

- Variance describes the amount of variation (dispersion/spread) by looking at all the observations, not just min/max, Q1 and Q3 etc.
- Used in advanced stats
- But... looking at the spread in units^2 is not intuitive
- So, take the square root of the variance:

Sample

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

Population

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

- This is known as the standard deviation (s)
- A measure of the average deviation about the mean
- Note: s = √s2 OR Standard deviation is the square root of variance

Coefficient of Variation (CV)

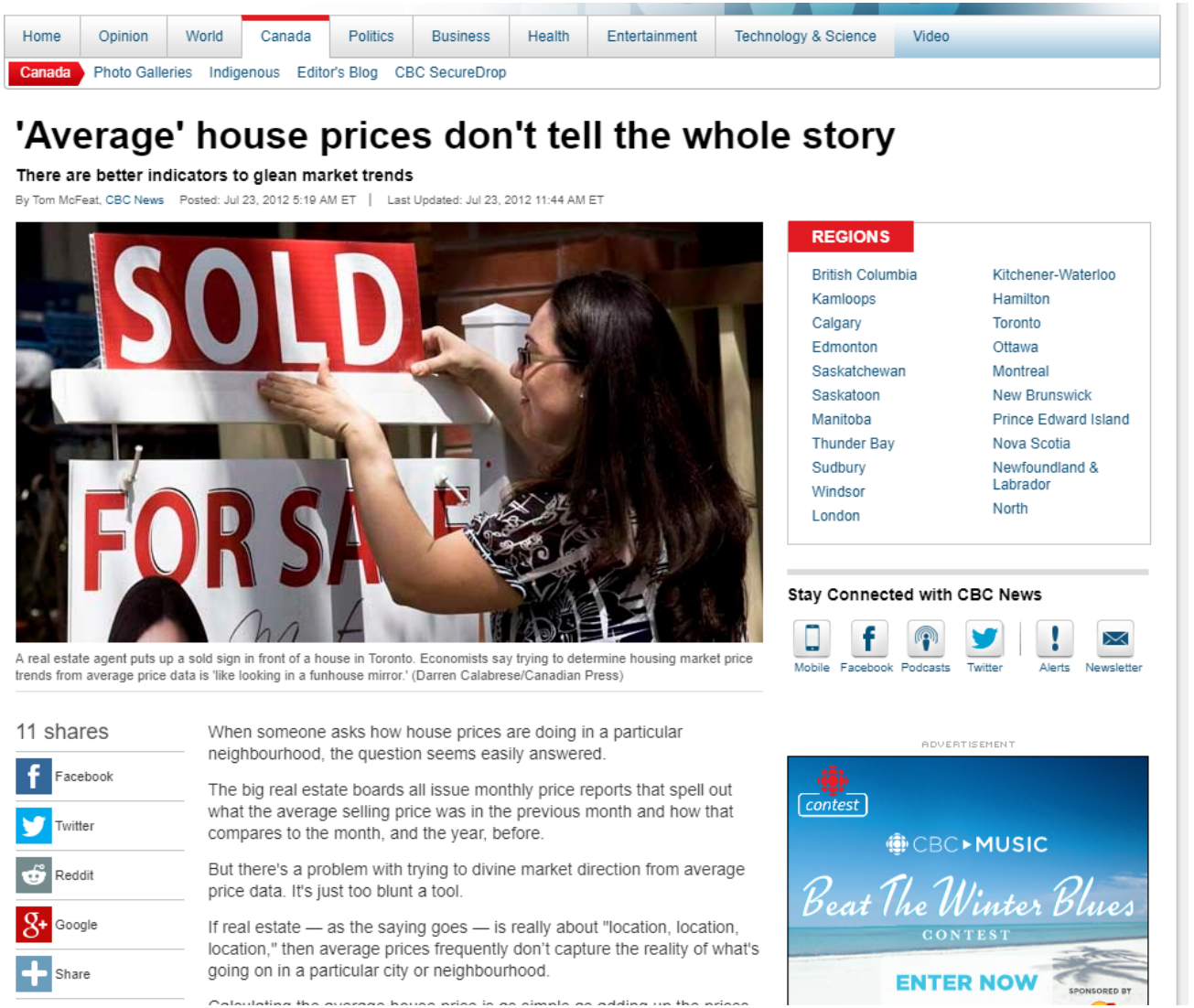
- Standard deviation and variance are absolute measures
- The value depends on the size and magnitude of the units from which they were calculated
- They are not comparable across
- Coefficient of variation is a relative measure of dispersion
- Can be expressed as a proportion OR a percentage (x100)
- Only appropriate for ratio data

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$



Home | Opinion | World | Canada | Politics | Business | Health | Entertainment | Technology & Science | Video

Canada | Photo Galleries | Indigenous | Editor's Blog | CBC SecureDrop

'Average' house prices don't tell the whole story

There are better indicators to glean market trends

By Tom McFeat, CBC News Posted: Jul 23, 2012 5:19 AM ET | Last Updated: Jul 23, 2012 11:44 AM ET

A real estate agent puts up a sold sign in front of a house in Toronto. Economists say trying to determine housing market price trends from average price data is 'like looking in a funhouse mirror.' (Darren Calabrese/Canadian Press)

When someone asks how house prices are doing in a particular neighbourhood, the question seems easily answered.

The big real estate boards all issue monthly price reports that spell out what the average selling price was in the previous month and how that compares to the month, and the year, before.

But there's a problem with trying to divine market direction from average price data. It's just too blunt a tool.

If real estate — as the saying goes — is really about "location, location, location," then average prices frequently don't capture the reality of what's going on in a particular city or neighbourhood.

Cautionary tale – use the correct measure of central tendency:

http://www.cbc.ca/news/canada/average-house-prices-don-t-tell-the-whole-story-1.1215736

Quantifying Shape

Histograms

- Plot a histogram to see the shape of the distribution
- Remember these rules:
  - Define classes or bins
  - Should be equal in size
  - Not too many or too few (see next slide)
  - Count how many observations fall into each class
  - Plot as a vertical bar chart
    - x-axis classes
    - y-axis freq. counts

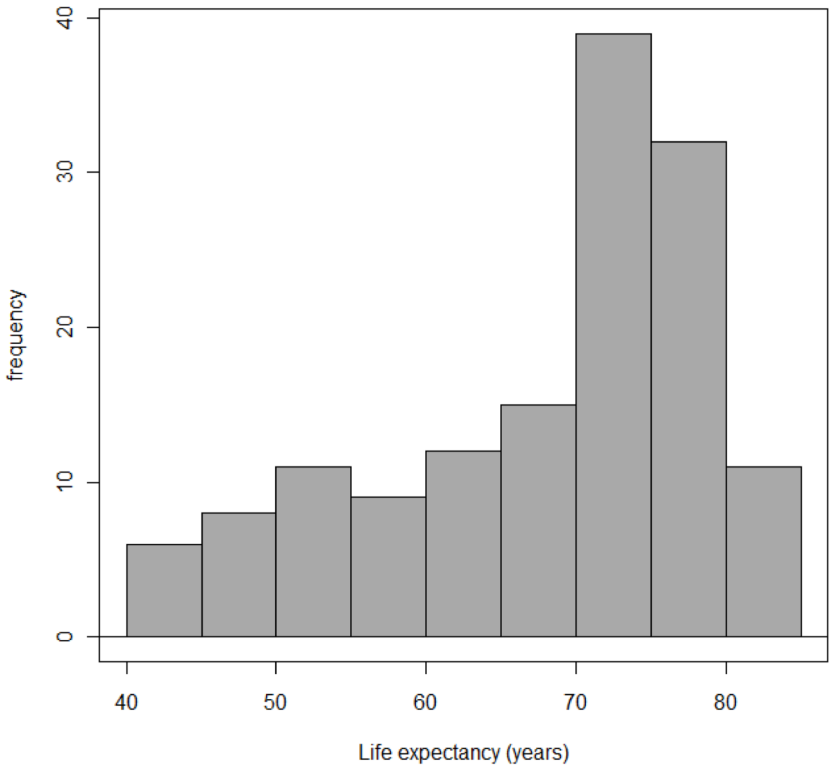
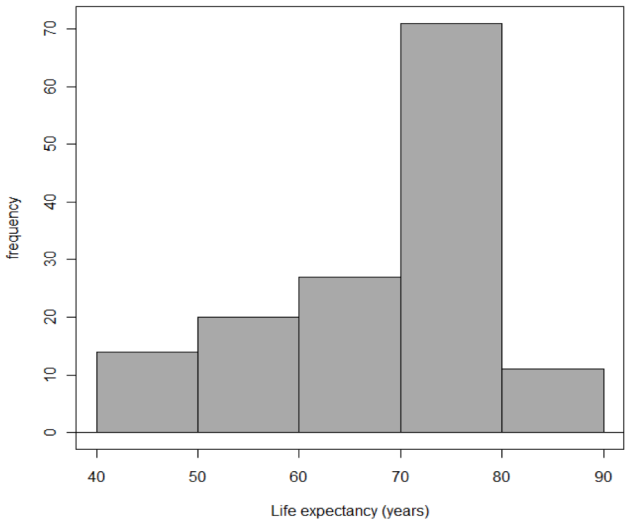


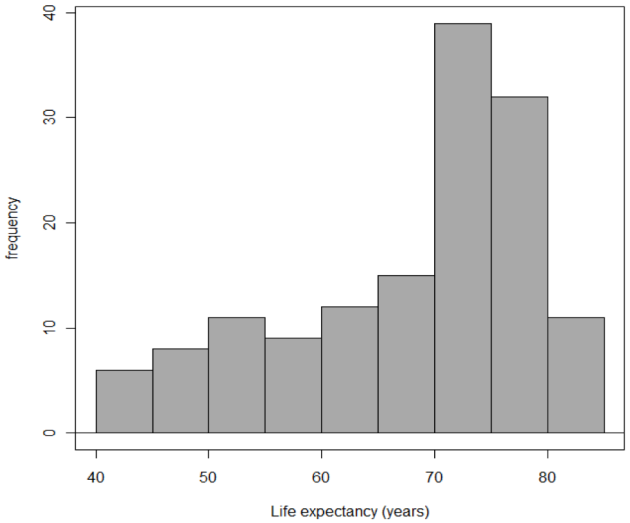
Figure 2. Histogram of life expectancy.
Source: www.happyplanetindex.org

Number of classes

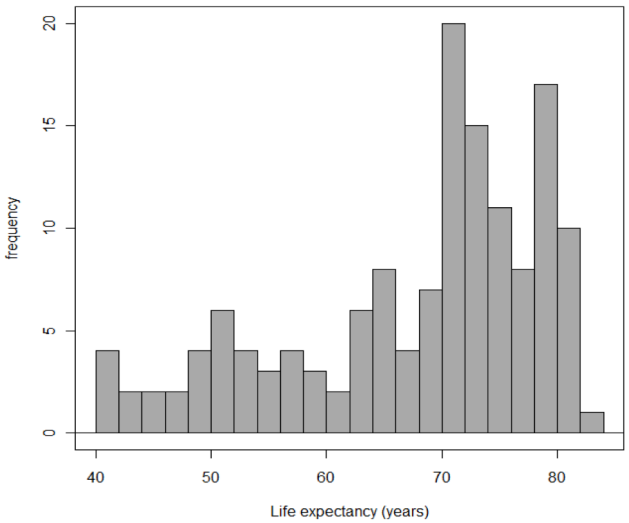
Too few (5)



About right (10)



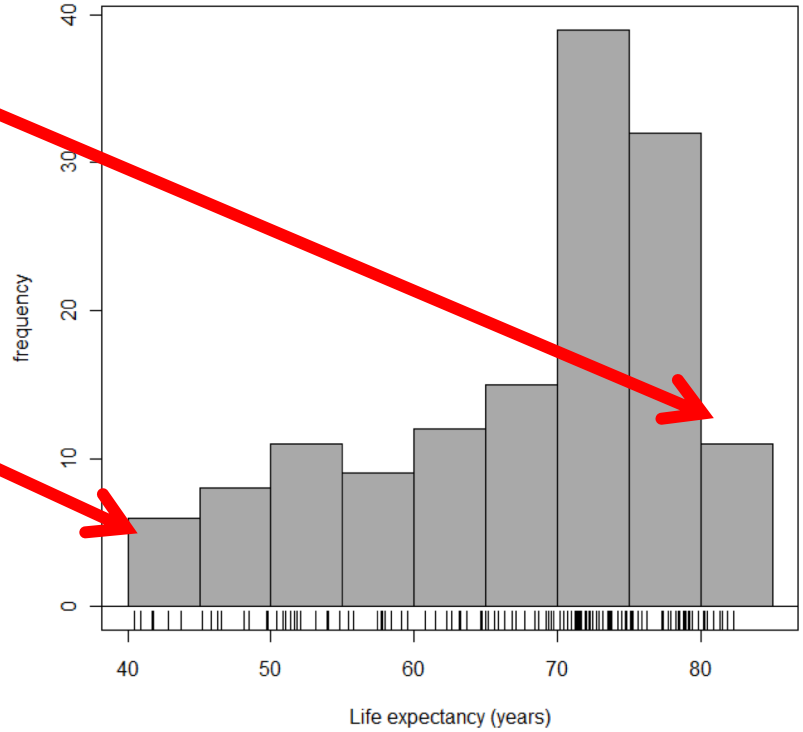
Too many (20)



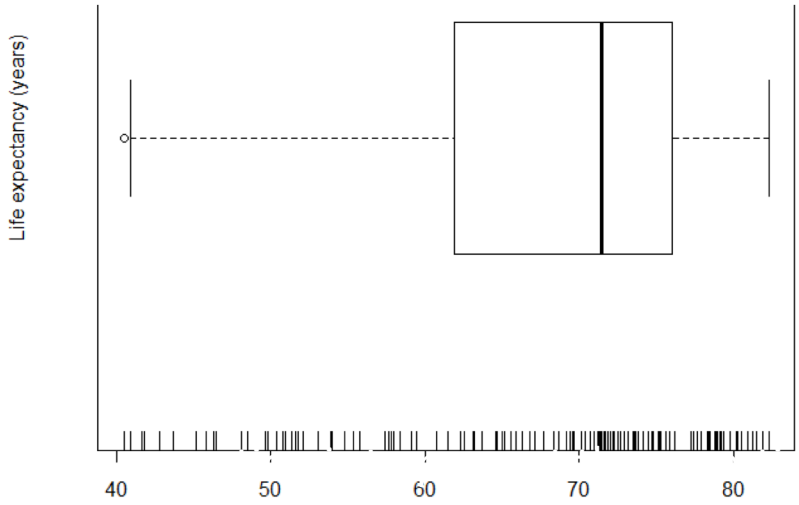
Best practice:

- # of classes approx. = √n (square root of the number of observations)
- Sturges Rule is another good one ($k = 1 + 3.322(\log_{10} n)$ (default in R)

- Histogram



- Boxplot (horizontal here for comparison to the histogram)



- Note the **rug plots** (ticks indicate individual observations)

Why do mean, median, and mode matter?

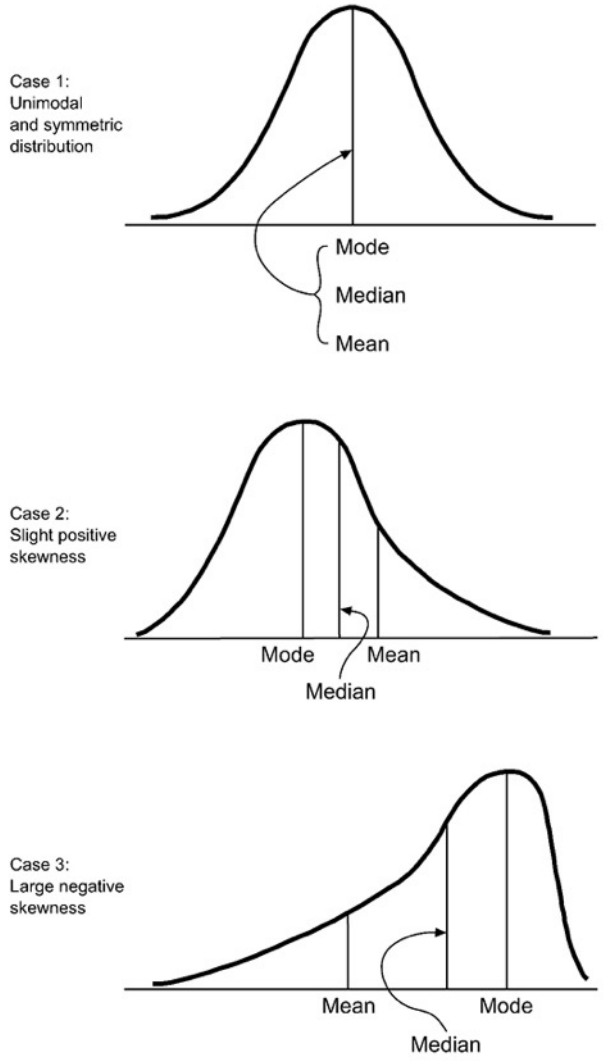


FIGURE 3.3
Measures of Central Tendency Placed on Symmetric and Asymmetric (Skewed) Frequency Distributions

- Give us information on how "normal" the distribution of our data is:
- How broad or narrow is a given distribution? (variation/dispersion/spread)
- Is the shape symmetrical? Is it 'peaked' or 'flat'?

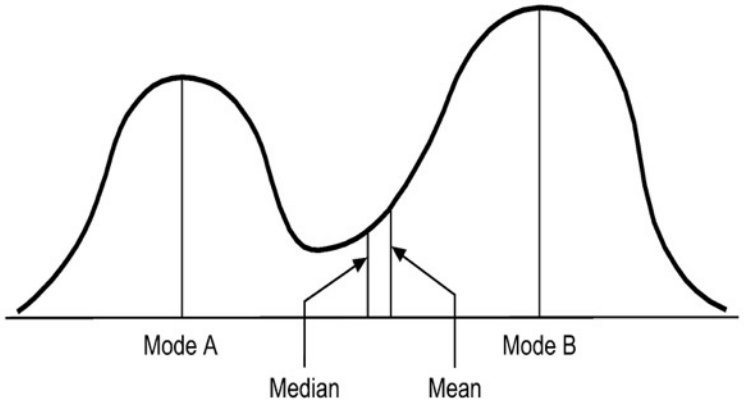


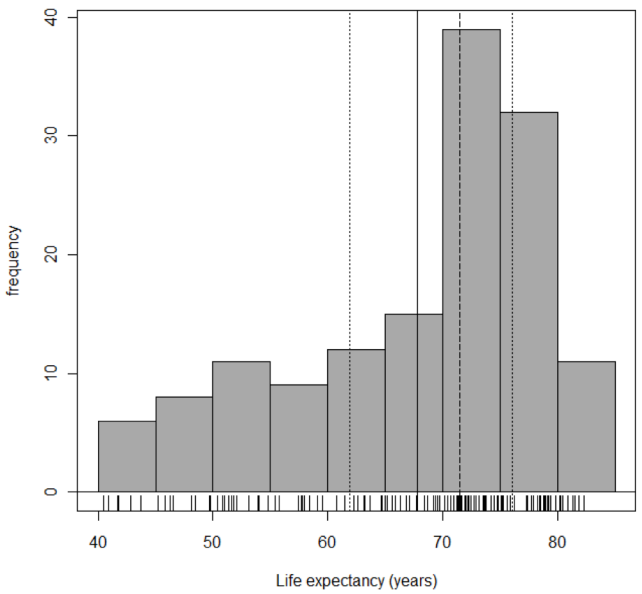
FIGURE 3.4
Measures of Central Tendency Placed on Bimodal Frequency Distribution

Skewness

Negative skew

Long tail to left

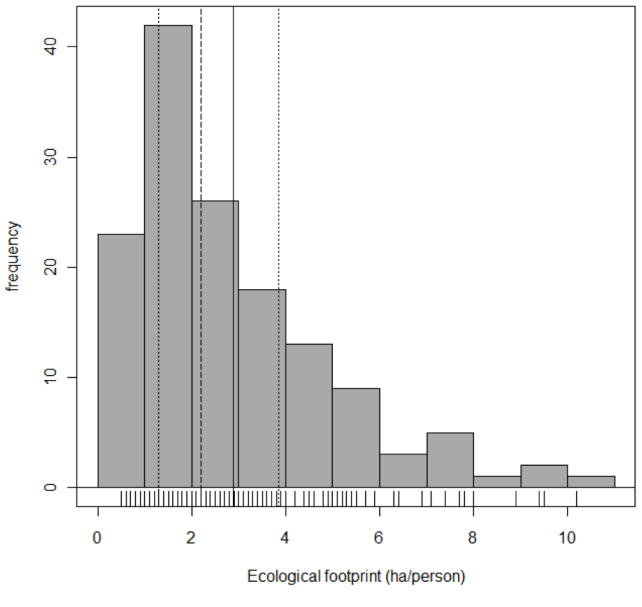
Mean < median



Positive skew

Long tail to right

Mean > median



Mean: solid line, Median: dashed line, Q1 and Q2: dotted line.

Source: www.happyplanetindex.org

Skewness

The degree that a distribution is skewed:

- Sk is negative, negative skew
- Sk = 0, not skewed
- Sk is positive, positive skew

$$Sk = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)s^3}$$

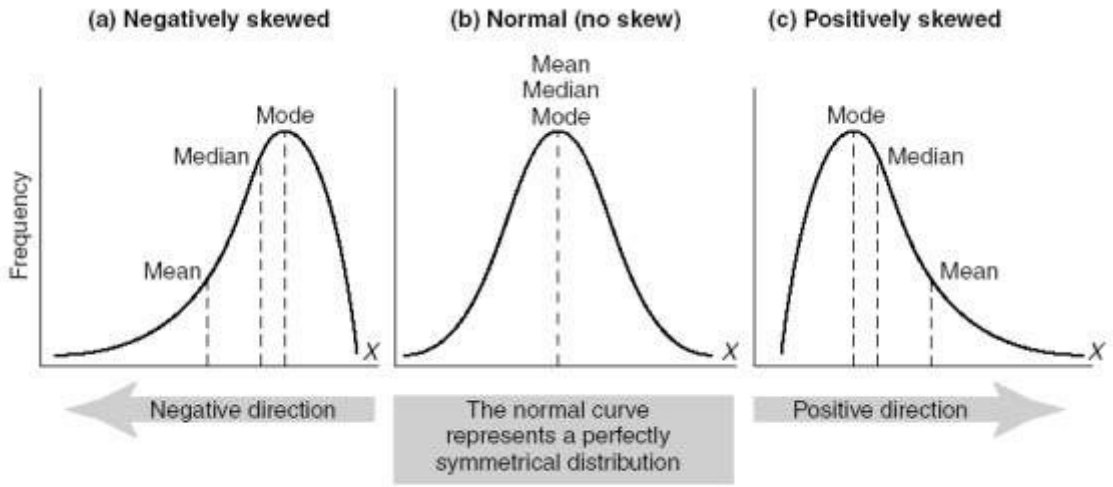


FIGURE 15.6 Examples of normal and skewed distributions

Kurtosis

Kurtosis - How peaked or flat the distribution is:

Peaked = **leptokurtic** (positive Ku)

Not peaked or flat = **mesokurtic** (Ku=0)

Flat = **platykurtic** (negative Ku)

$$Ku = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4} - 3$$

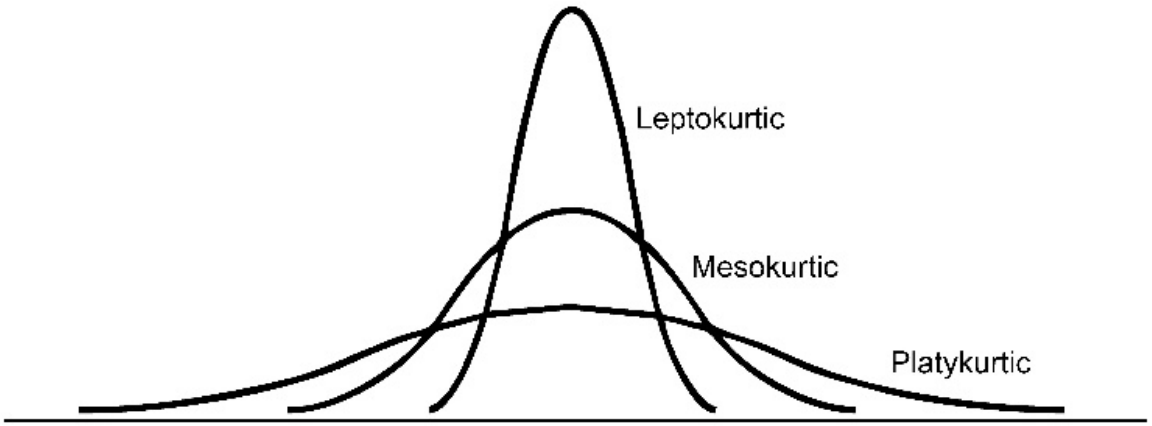


FIGURE 3.10
Different Levels of Kurtosis

Descriptive Statistics in R

```
> library(psych)

> describe(Dataset)
                   vars  n    mean    sd  median trimmed   mad     min     max  range  skew kurtosis    se
Year                  1 11 2006.00  3.32 2006.00 2006.00  4.45 2001.00 2011.00  10.00  0.00    -1.53  1.00
Certified             2 11  105.51 49.95  123.75  110.32 38.68   17.17  150.57 133.40 -0.67    -1.31 15.06
Harvested             3 11    0.85  0.17    0.83    0.85  0.25    0.61    1.06   0.45 -0.10    -1.89  0.05
Planted               4 11    0.43  0.04    0.44    0.44  0.04    0.37    0.48   0.11 -0.52    -1.24  0.01
Seeded                5 11    0.02  0.01    0.02    0.02  0.00    0.01    0.05   0.04  1.42     1.53  0.00
Insect_defoliation    6 11   16.64  4.16   16.32   16.80  5.19    9.19   22.60  13.41 -0.20    -1.37  1.25
Burned                7 11    1.98  0.85    1.74    2.00  1.02    0.63    3.18   2.55 -0.11    -1.36  0.26
Drought*              8 11    1.64  0.50    2.00    1.67  0.00    1.00    2.00   1.00 -0.49    -1.91  0.15
```

Do you have a moment?

- The moments of a distribution are used to characterize it:
- 1st moment = Mean
- 2nd moment = Standard deviation
- 3rd moment = Skewness
- 4th moment = Kurtosis

- Note that the formula for skewness and kurtosis may vary when comparing moments between studies

Formula Summary

Descriptor	Formula
Mean (μ, x-bar)	$\bar{X} = \frac{\sum x}{n}$
Median (x tidle)	$\tilde{x} = \left((n+1) \div 2\right)^{th} sorted\ value$
Mode	No formula-number that occurs most often
Range	max. value – min. value
Deviation (d)	$d_i = x_i - \bar{x}$
Squared Deviations/ sum of squares	$(x_i - \bar{x})^2 \ / \ \sum_{i=1}^{n} (x_i - \bar{x})^2$
Variance (s^2)	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Standard Deviation (s, σ)	$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$
Coefficient of Variation (CV)	$CV = \frac{s}{\bar{x}} * 100\%$

Summary

- Describing data means giving an idea of the variable's:
- Measures of central tendency:
- Measures of variation/dispersion
- Metrics describing the shape of the data
- Other resources:
- http://dev1.ed-projects.nyu.edu/statistics/descriptive-statistics/