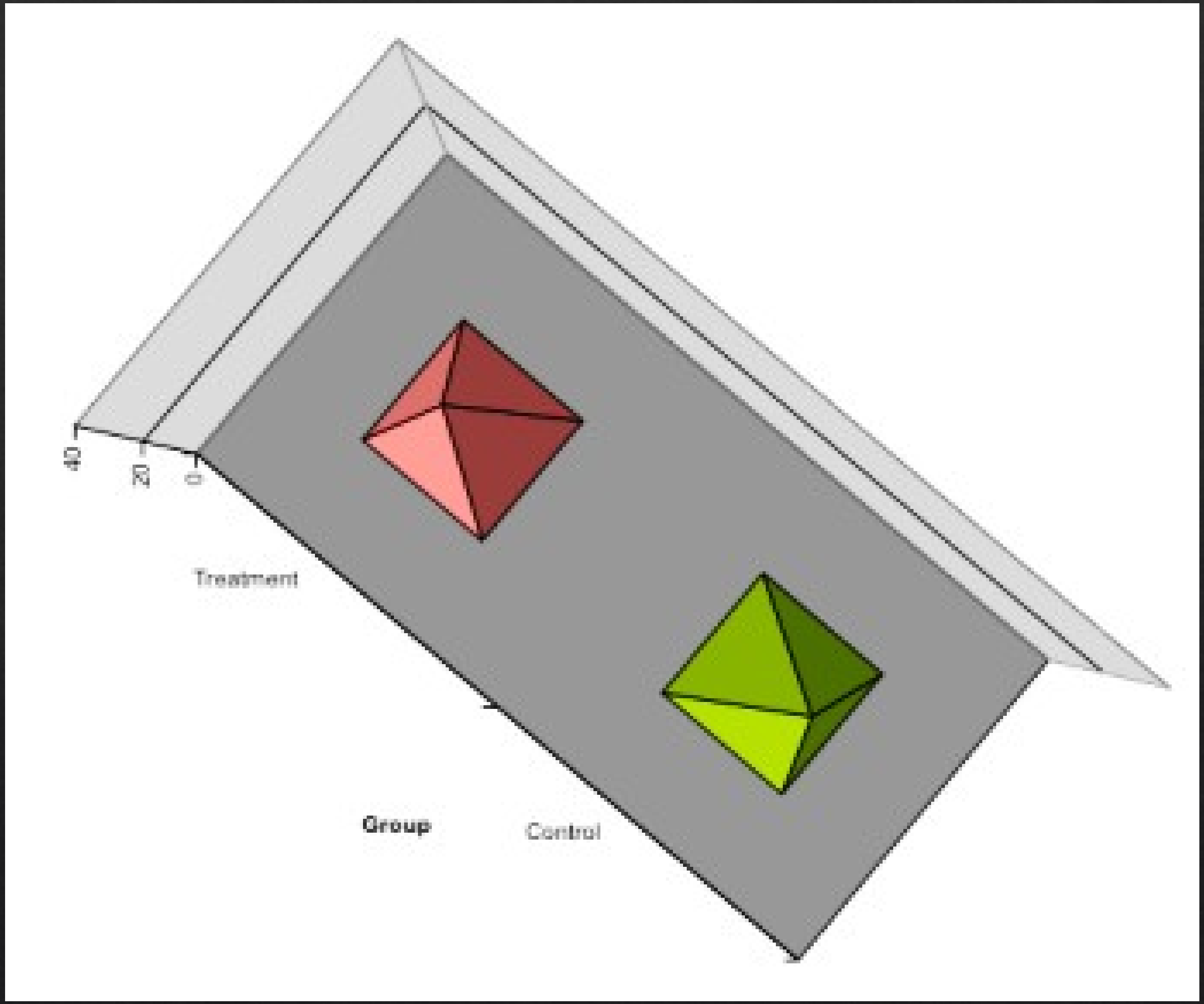
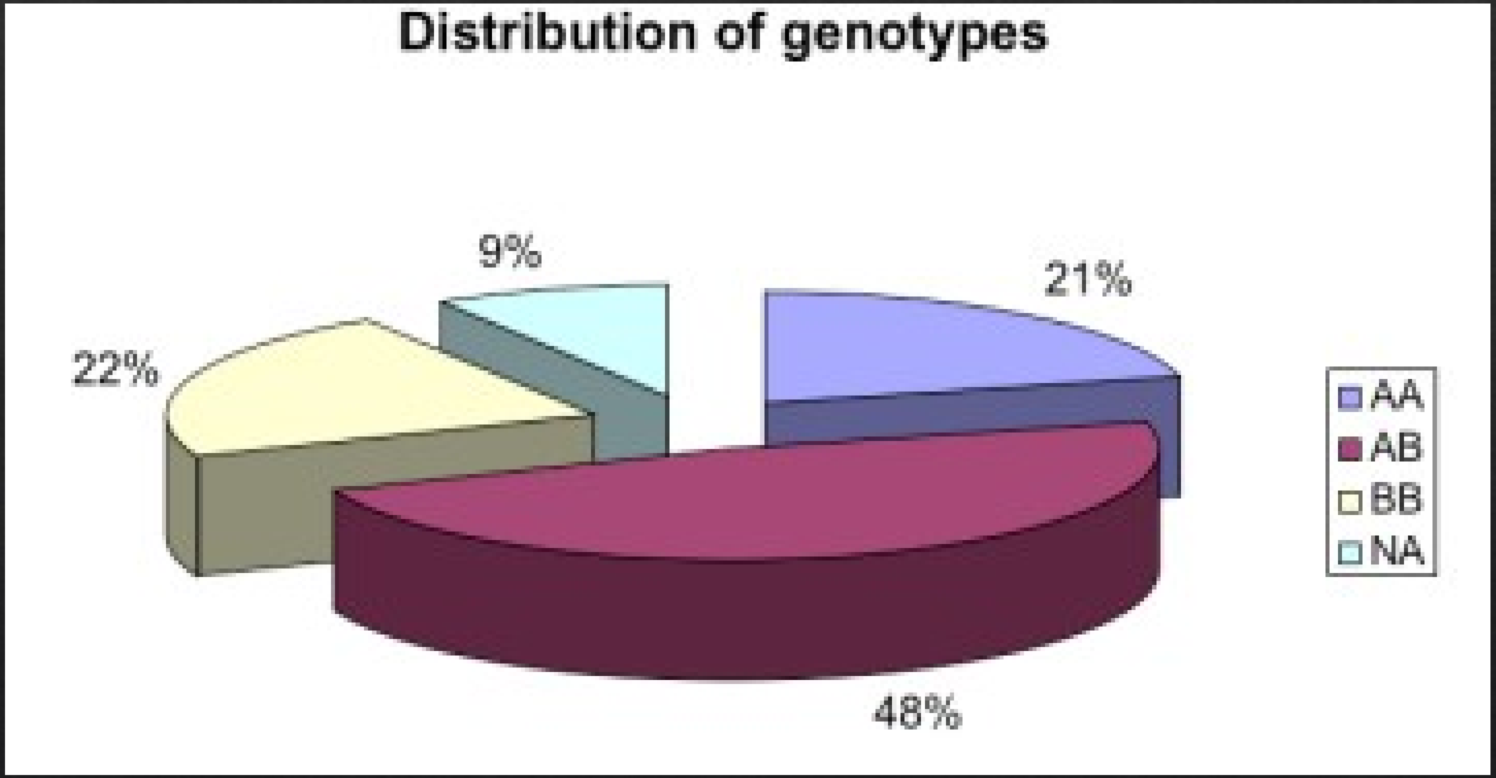
Charts and Statistical Visualization

Debajyoti Mondal
University of Saskatchewan

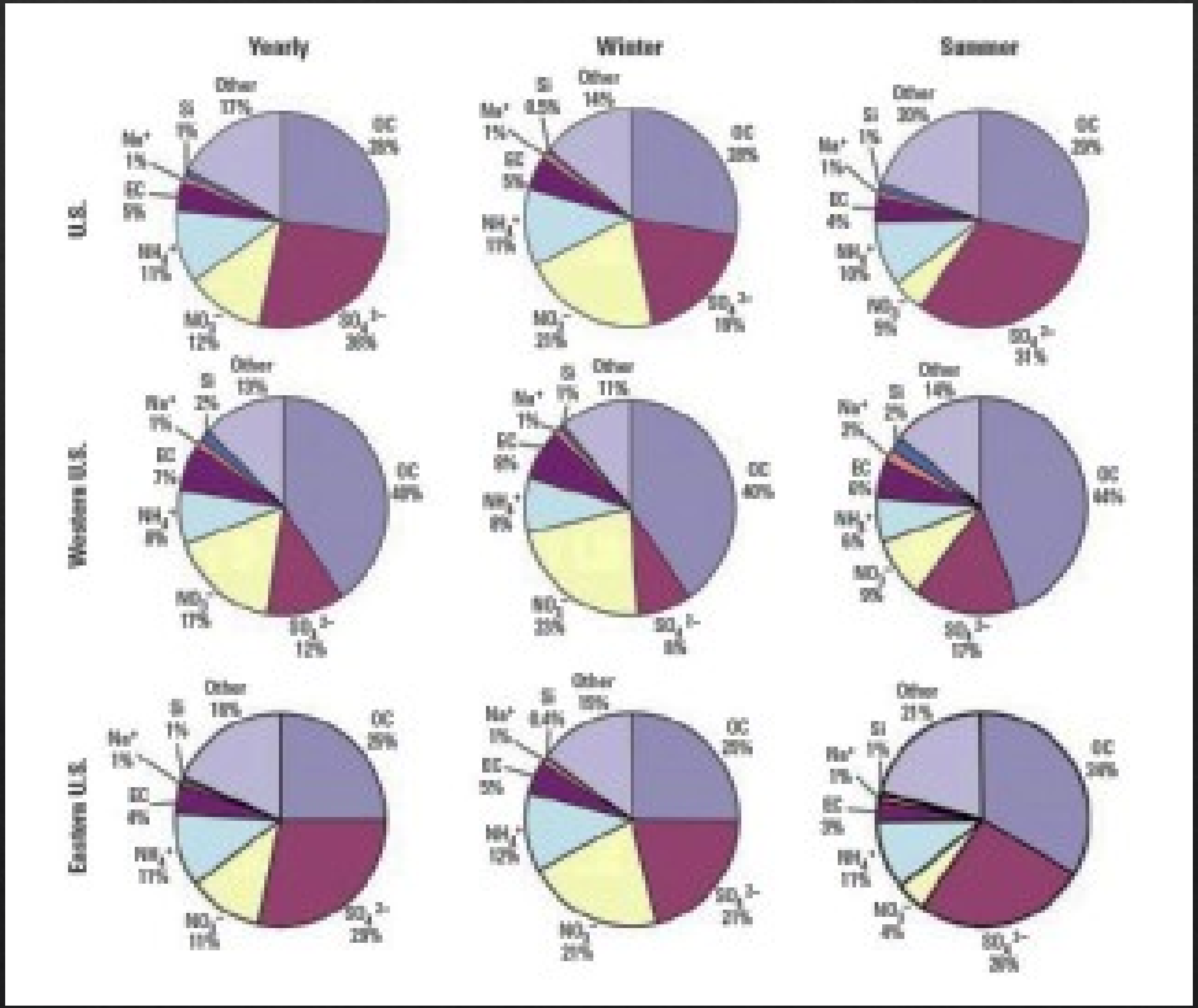
Example of Bad Visualizations



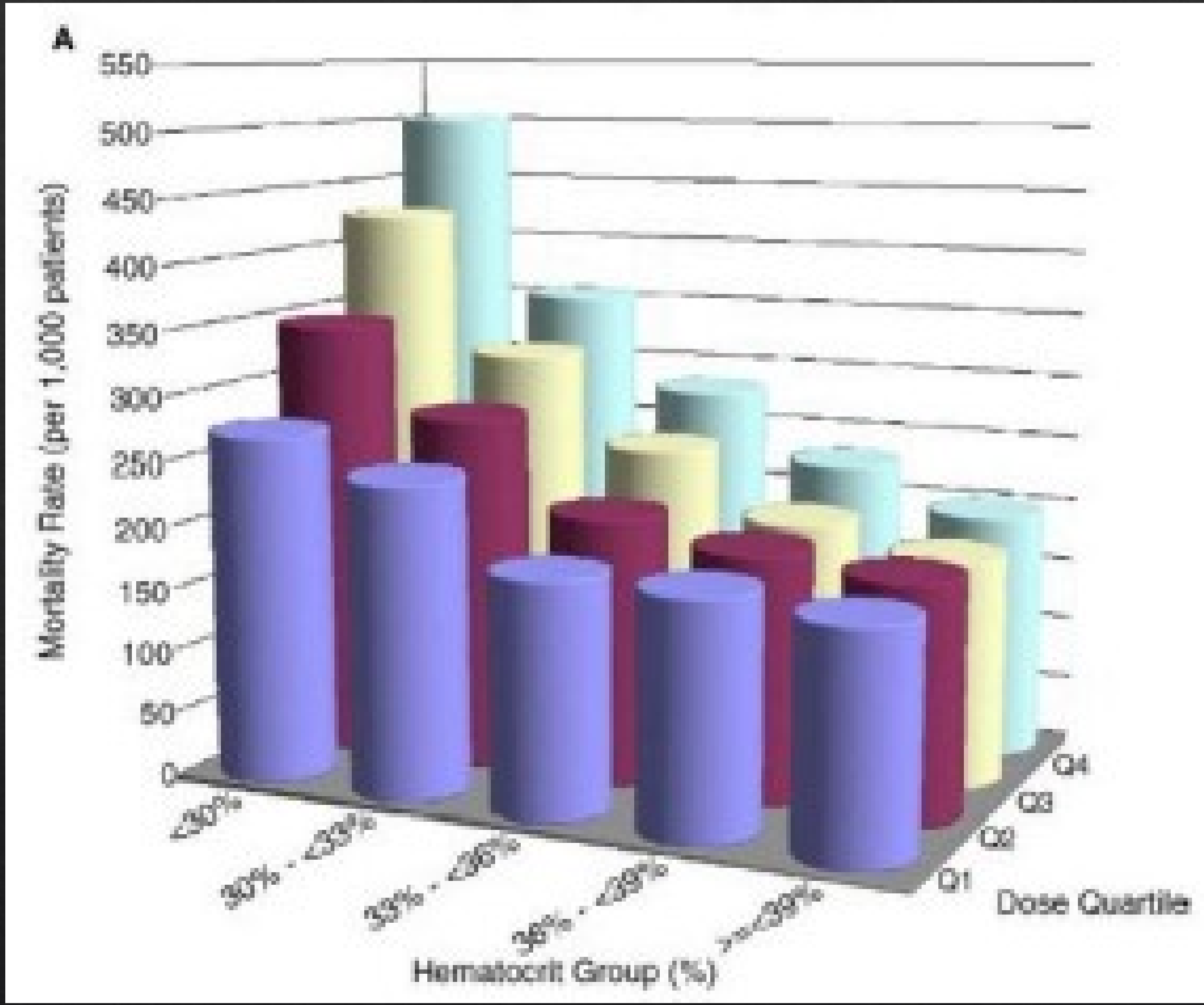
Example of Bad Visualizations



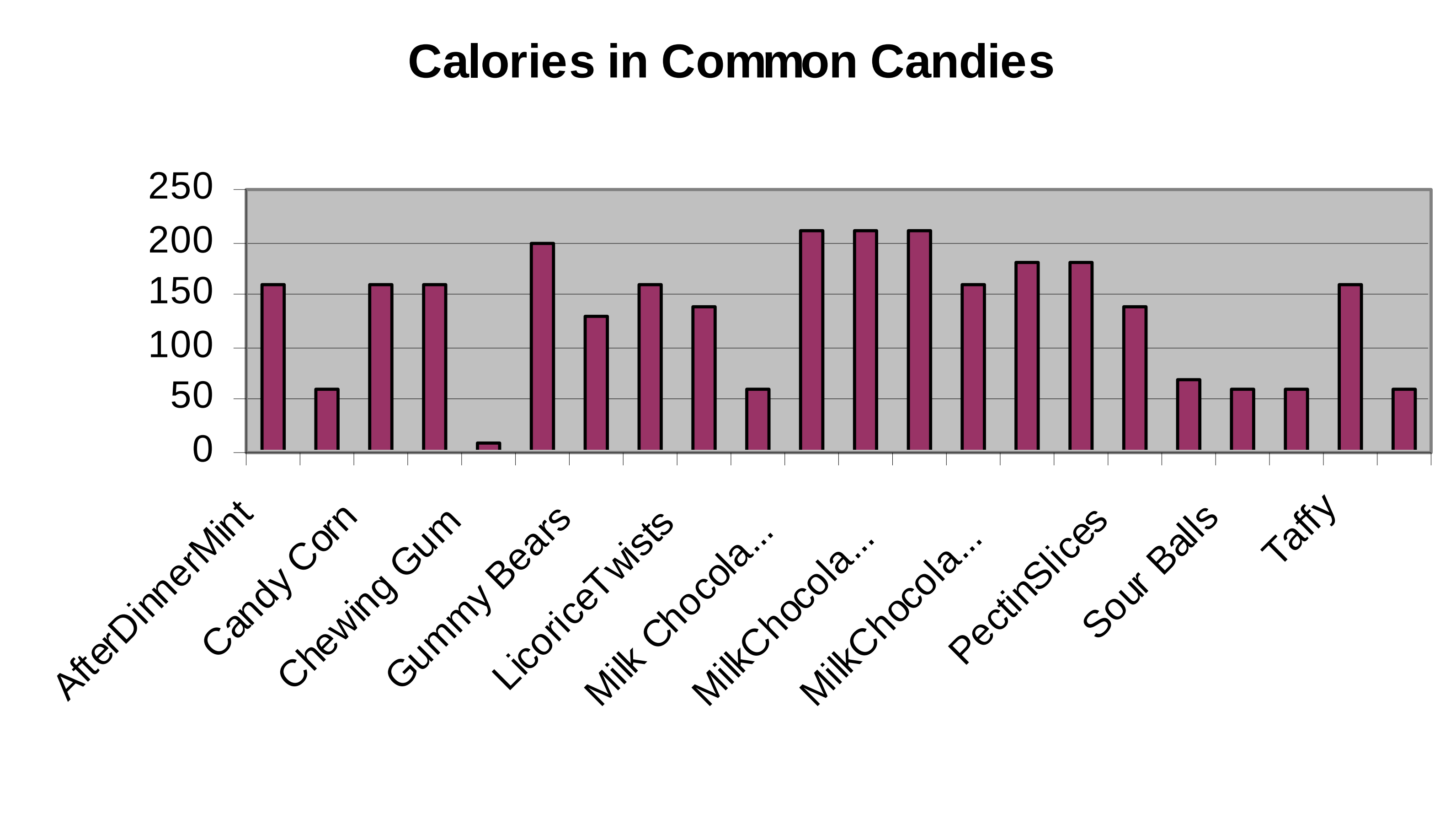
Example of Bad Visualizations



Example of Bad Visualizations



Bar Chart

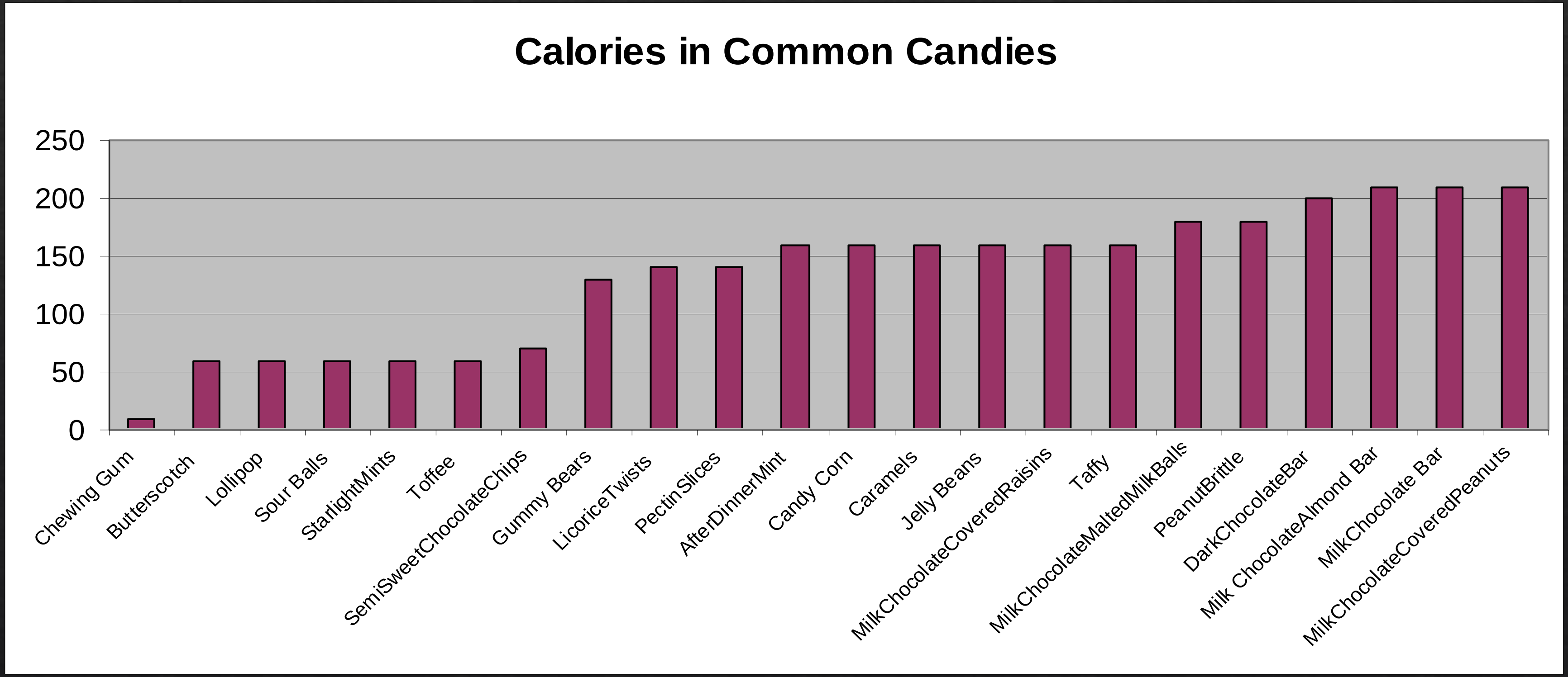


What are the problems with this graph?

Curtsey: STA6166-2-6

Bar Chart

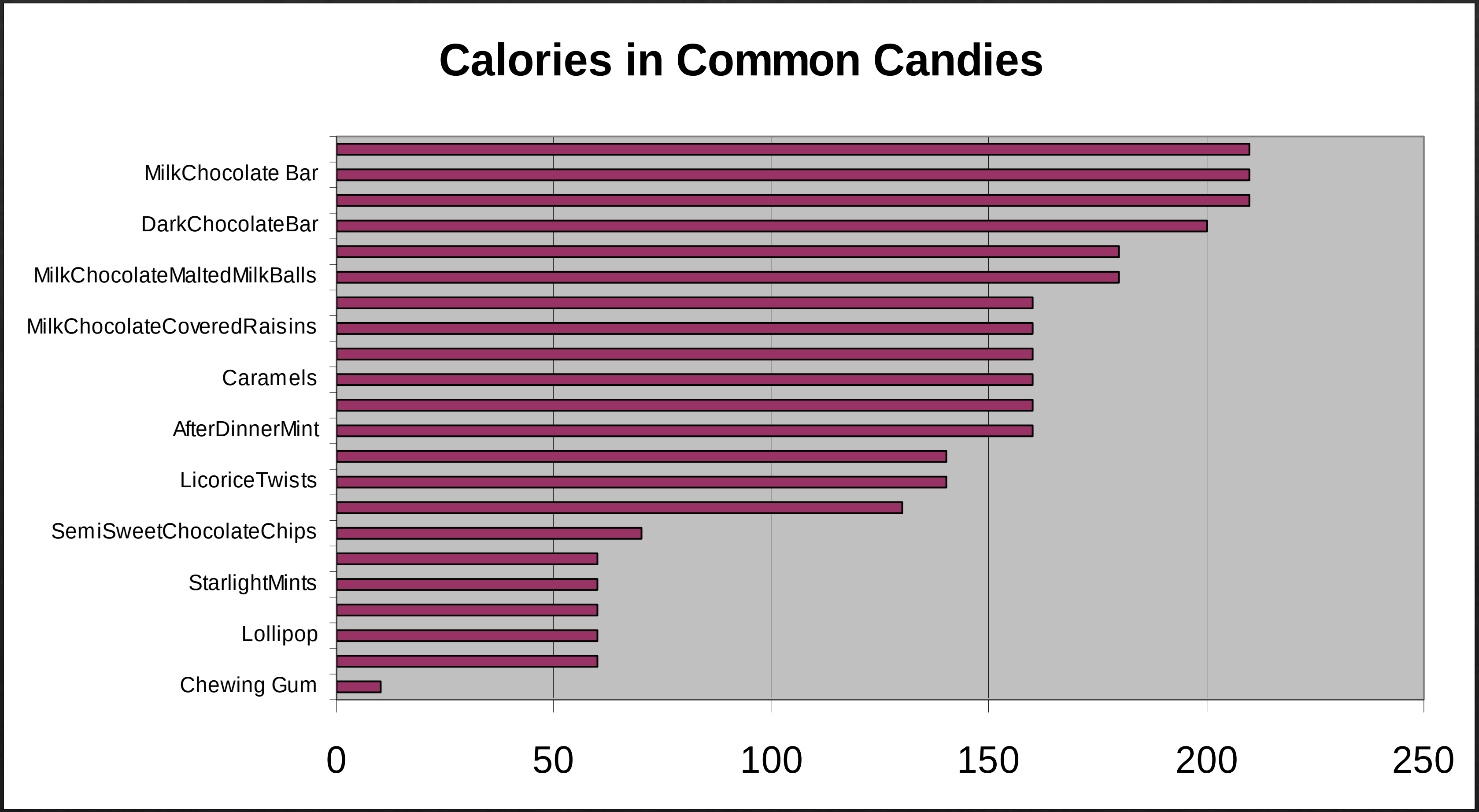
Sorting and expanding the scale of the graph allows all labels to be seen as well as displaying a characteristic of the data.



Curtsey: STA6166-2-7

Bar Chart

A vertical display allows better comparison of calorie amounts.



Curtsey: STA6166-2-8

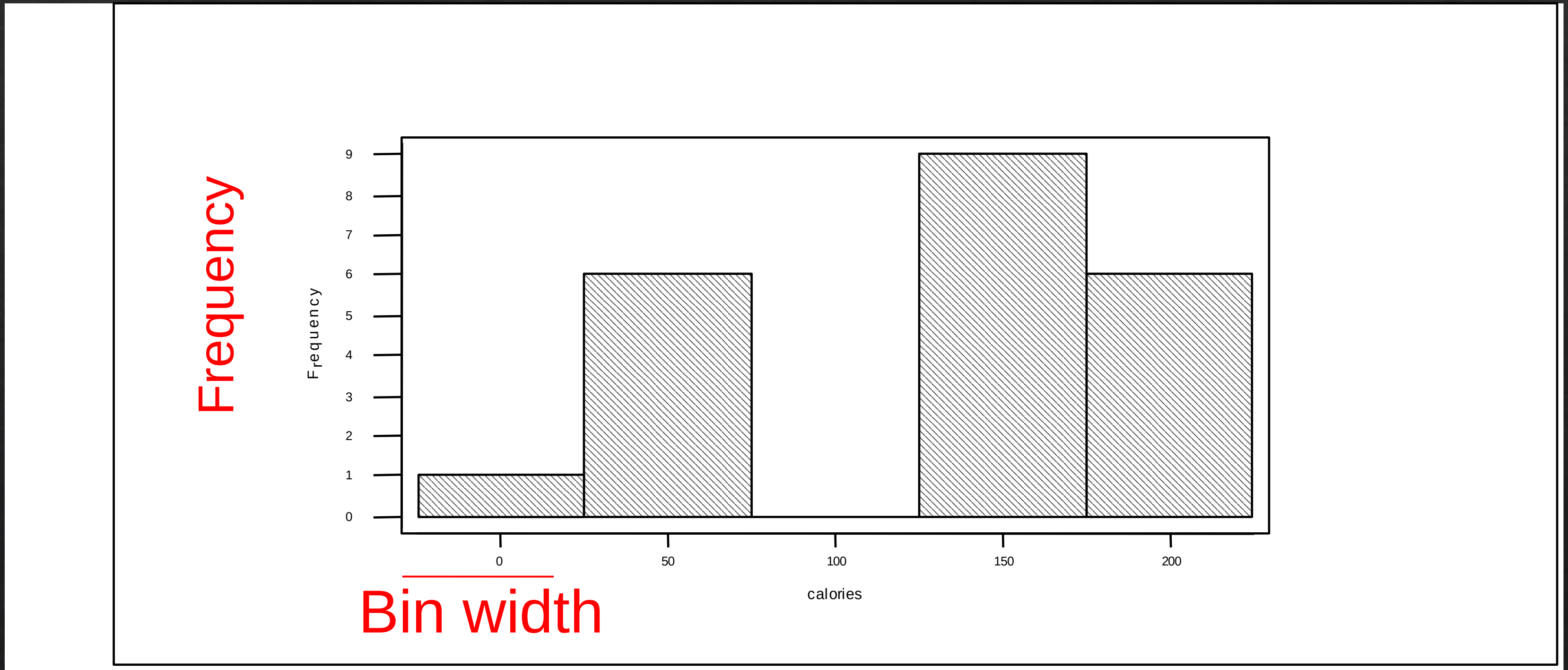
Histogram

One axis representing the range of the variable, and the other axis representing the data density at positions within the range

- Most commonly-used for univariate data
- For relatively continuous data, observations are grouped into mutually exclusive categories, called “bins”
- Used to visualize distribution (shape, center, range, variation)
- A major challenge is to come up with an appropriate binning process

Frequency Histogram

A graphical presentation of the frequency table where the relative areas/height of the bars are in proportion to the frequencies.

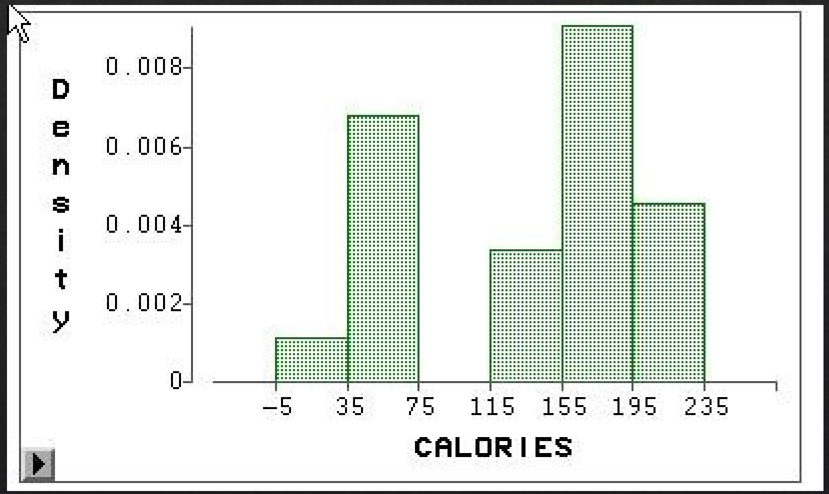


Density Histogram

A density histogram (or simply a histogram) is constructed just like a frequency histogram, but now the total area of the bars sums to one.

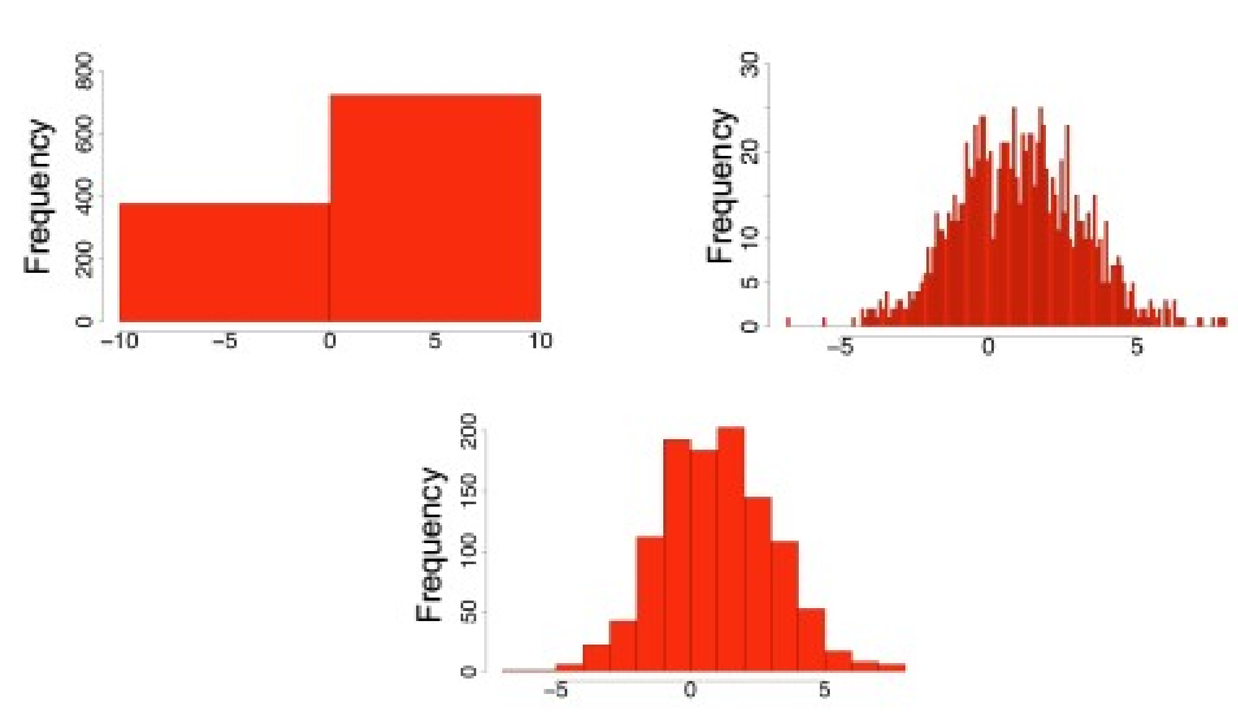
This is accomplished by rescaling the vertical axis. Instead of frequencies, the vertical axis records the rescaled value of the density.

Histograms have important ties to probability.



Sum of shaded area is equal to one.

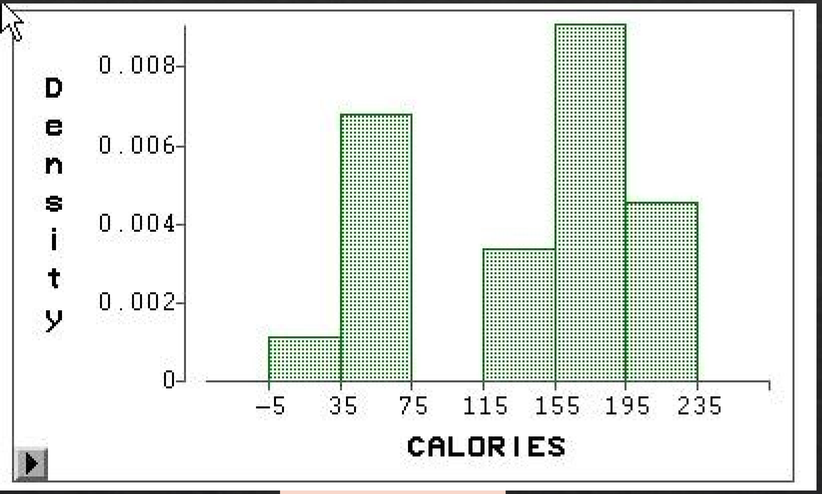
Number of Bins for Histograms



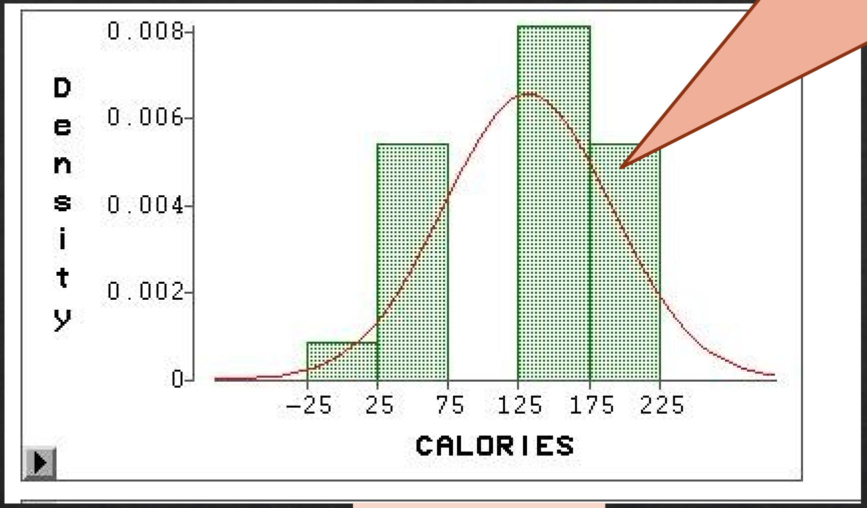
How we view the "distribution" of a dataset can depend on how much data we have and how it is binned.

Number of Bins for Histograms

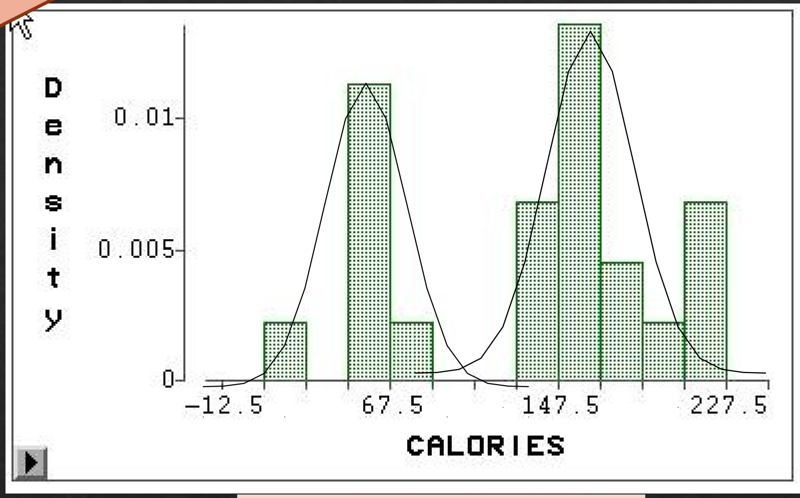
Smoothed histogram or density curve



Six bins



Five bins

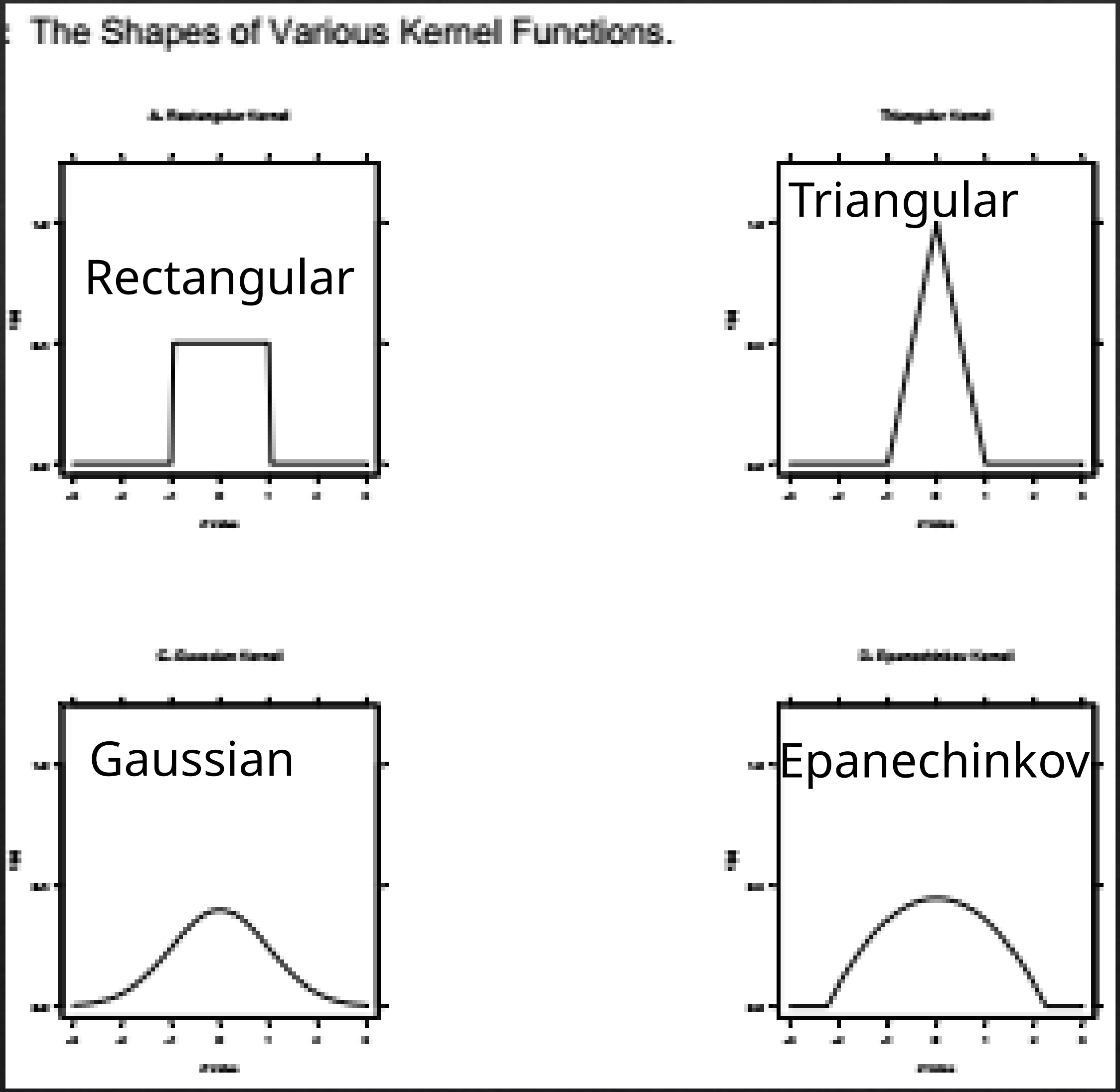


Eleven bins

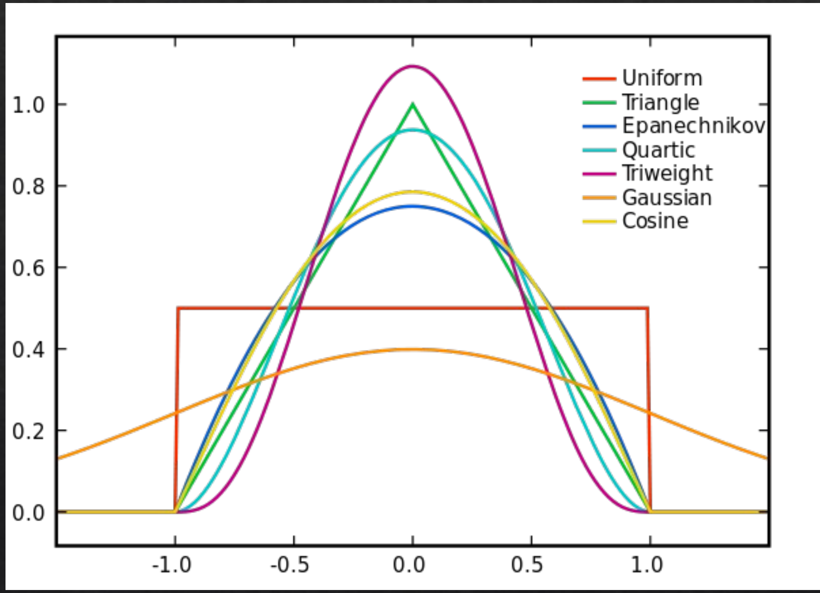
How we view the "distribution" of a dataset can depend on how much data we have and how it is binned.

Smoothed Histogram

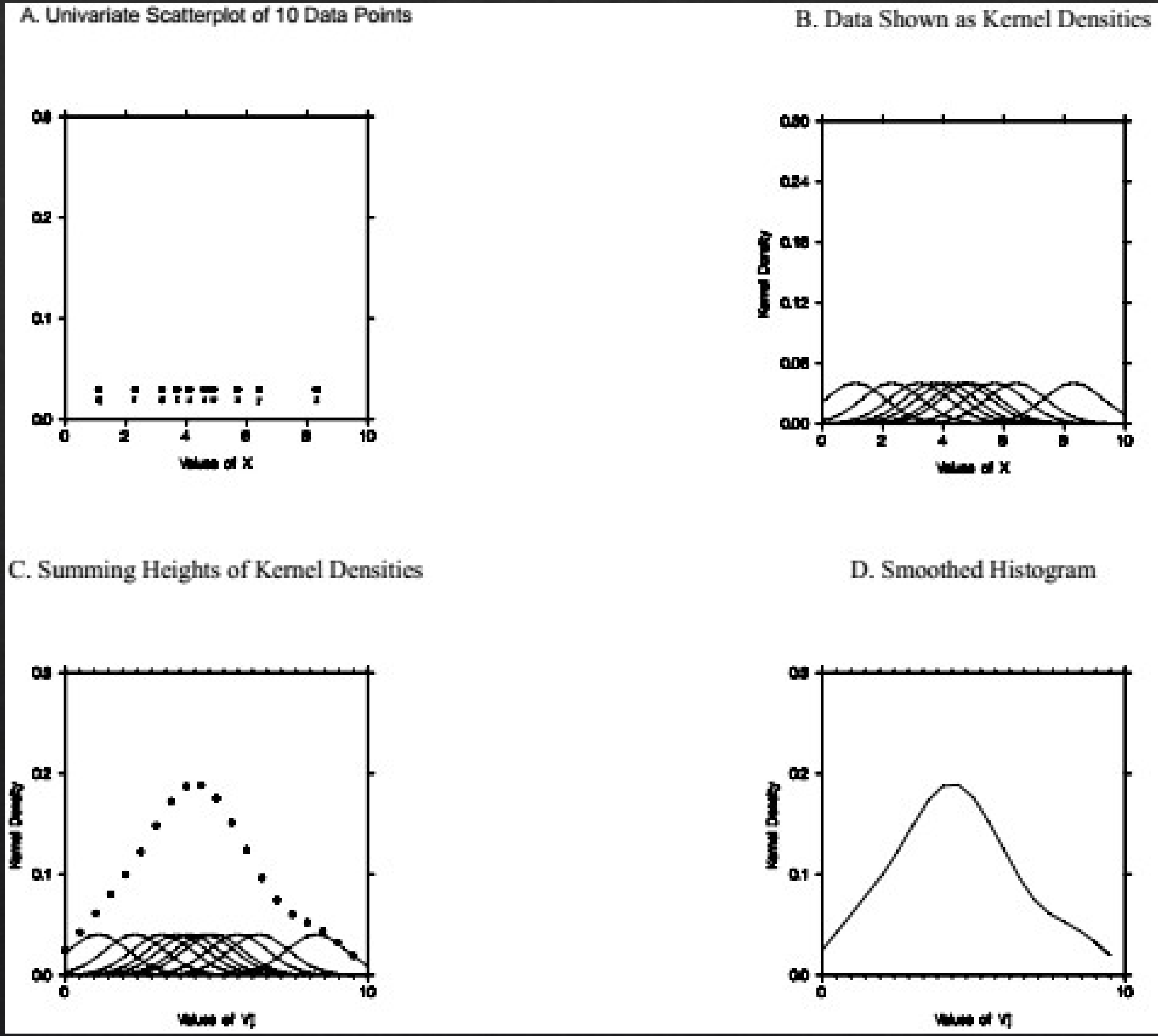
- A continuous function of the original data values
- Construction of a smoothed histogram requires a kernel function



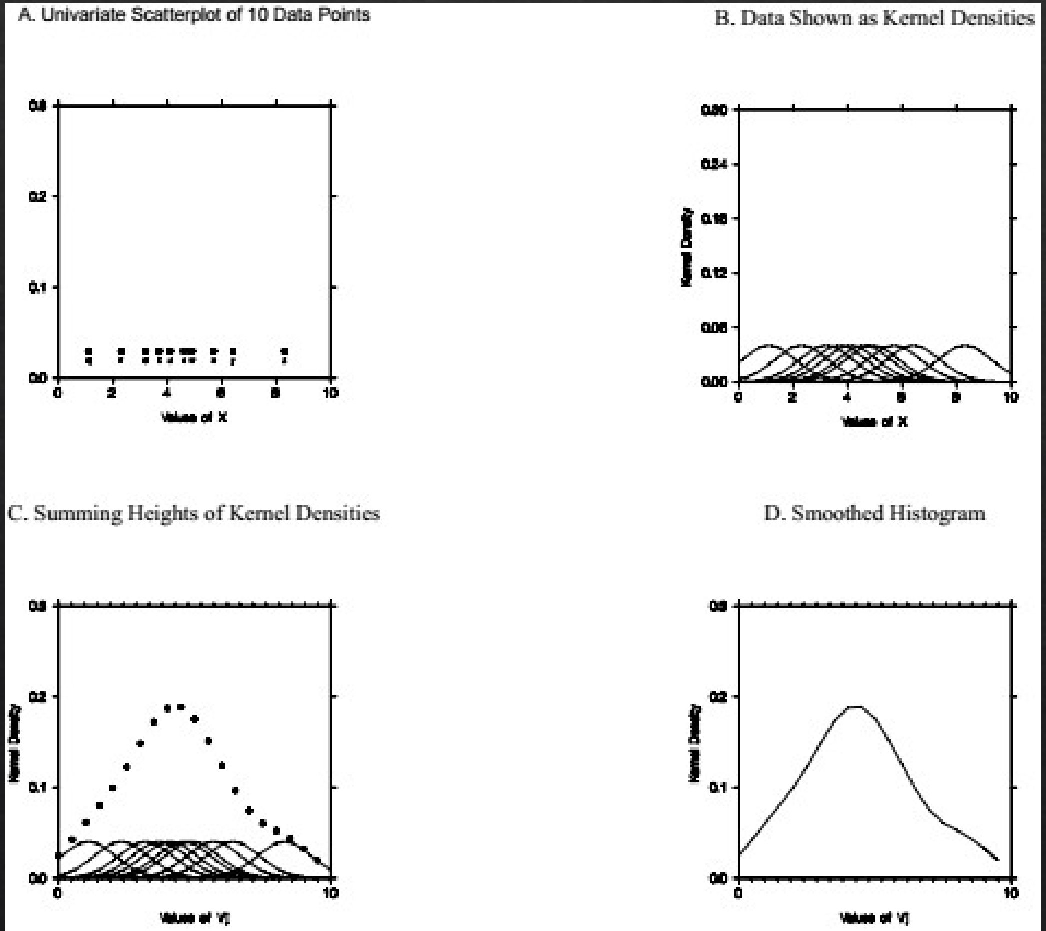
Smoothed Histogram

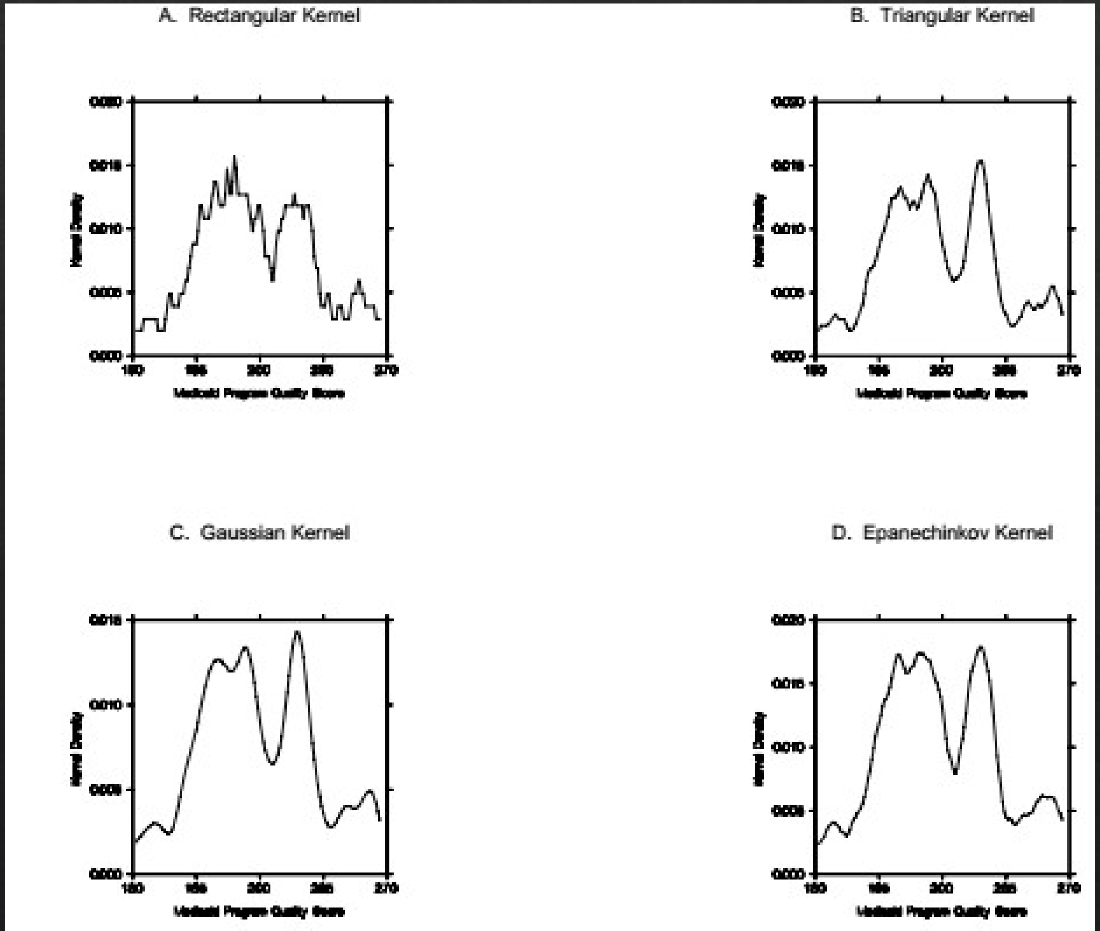


https://en.wikipedia.org/wiki/
Kernel_(statistics)#Kernel_functions_in_common_use



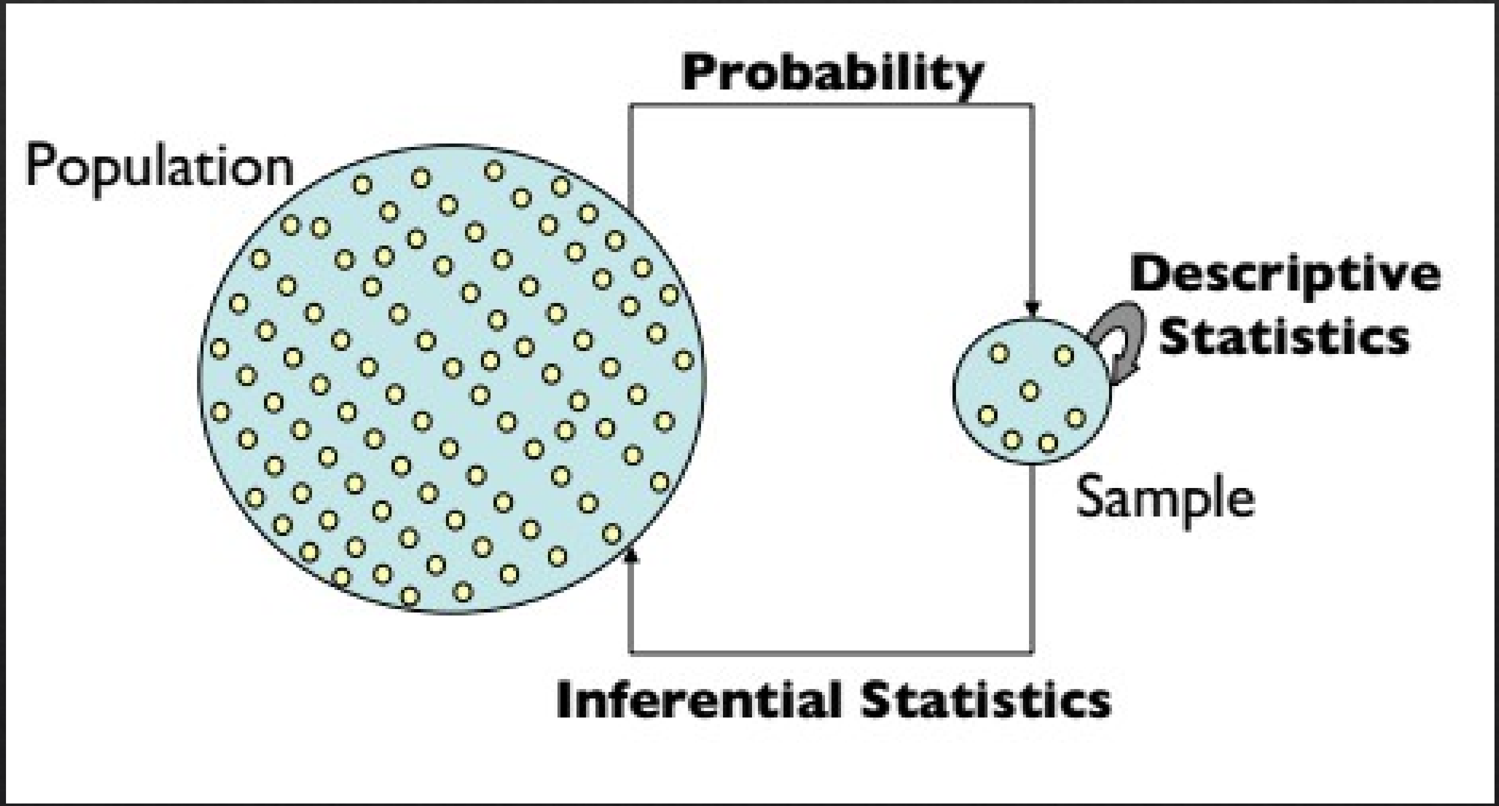
Smoothed Histogram





To effectively use charts we need to know some statistics

“Central Dogma” of Statistics



Basic Statistics

A statistic is a function of the sample data.

We will learn some “descriptive statistics”. These statistics address specific aspects of the distribution of the data.

- What is the range of the data?
- When we sort the data, what number might we see in the “middle” of the range of values?
- What number tells us over what sub-range do we find the bulk of the data ?

Extremes

If we sort the data we can immediately identify the extremes.

Extremes

- *Minimum(calories) = 10*
- *Maximum(calories) = 210*

The minimum and maximum are “statistics”.

10	60	60	60	60	60	70	130	140	140	160	160	160	160	160	160	180	180	200	210	210	210

Range

Range: the difference between the largest and smallest measurements of a variable.

Extremes

- *Minimum(calories) = 10*
- *Maximum(calories) = 210*

→ *Range = 210-10 = 200*

Tells us something about the spread of the data.

Range

Range: the difference between the largest and smallest measurements of a variable.

Extremes

- *Minimum(calories) = 10*
- *Maximum(calories) = 210*

Range = 210-10 = 200

Tells us something about the spread of the data.

Midrange = minimum + (Range/2)
=10 + 200/2
=110

Is it a “good” measure of the center of the data?

Measures of Central Tendency

Estimate the value that is in the center of the “distribution” of the data .

Median = middle value in the sorted list of n numbers: at position (n+1)/2
= unique value at (n+1)/2 if n is an odd number or
= average of the values at n/2 and n/2+1 if n is even
= (160 + 160)/2 = 160

Measures of Central Tendency

Estimate the value that is in the center of the "distribution" of the data .

Mean = sum of all values divided by number of values (average)
= (10 + 60 + 60 + 60 + … + 210 + 210)/22
= 133.6

Trimmed mean = mean of data where some fraction of the smallest and largest data values are not considered. Usually the smallest 5% and largest 5% values (rounded to nearest integer) of data are removed for this computation.
= 136.0 (with 10% trimmed, 5% each tail).

Measures of Central Tendency

Estimate the value that is in the center of the "distribution" of the data .

Weighted means:

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

<u>Other Types</u>

Geometric:

$$\bar{x} = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Harmonic:

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

Variance and Standard Deviation

Variance: The sum of squared deviations of measurements from their mean divided by n-1.

Sample Mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

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

Standard Deviation: The square root of the varian

$$s = \sqrt{s^2}$$

These measure the spread of the data.

Why squared deviation?

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

Which one is correct?

A. Square is not necessary - taking absolute values would also work

B. Squares are necessary because we divide by (n-1) instead of n

C. Square increases the contribution to the variance is as you go farther from the mean.

D. Any power of value at least 2 works, but square works best

Why Standard Deviation?

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

$$s = \sqrt{s^2}$$

Why Standard Deviation?

Variance is somewhat arbitrary, But if you “standardize” that value, you could talk about any variance (i.e. deviation) in equivalent terms

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

$$s = \sqrt{s^2}$$

Square root – now the value is in the units we started with

Why Standard Deviation?

Variance is somewhat arbitrary, But if you "standardize" that value, you could talk about any variance (i.e. deviation) in equivalent terms

Note use of μ (mu) to represent "mean".

Note use of σ (sigma) to represent "standard deviation."

At least	within
$(1 - 1/1^2) = 0\%$ .... k=1	$(\mu \pm 1\sigma)$
$(1 - 1/2^2) = 75\%$ .... k=2	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$ ....k=3	$(\mu \pm 3\sigma)$

Regardless of how the data are distributed, a certain percentage of values must fall within *k* standard deviations from the mean:

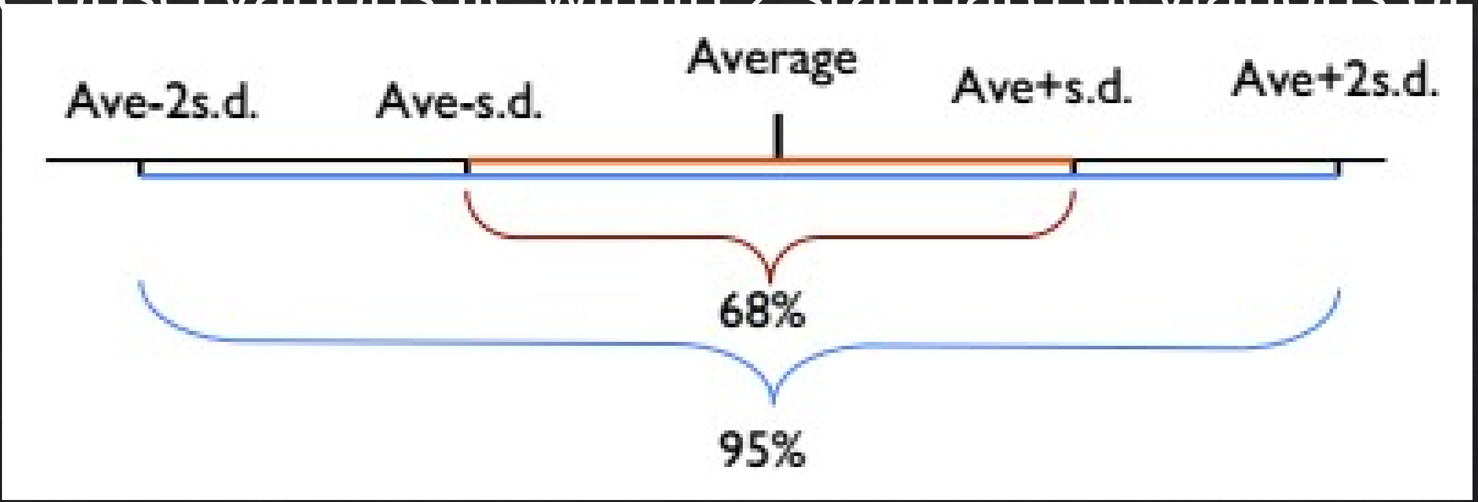
Courtesy: Joshua Akey

Often We Can Do Better

- For many lists of observations – especially if their histogram is bell-shaped

- Roughly 68% of the observations lie within 1 standard deviation of the average

- 95% of the observations lie within 2 standard deviations of the average



Courtesy: Joshua Akey

Quartiles

Suppose we divide the sorted data into four equal parts. The values which separate the four parts are known as the quartiles.

Because the sample size, does not always divide easily by 4, we do some estimating of these quartiles by linear interpolation between values.

Here n=22, (n+1)/4=23/4=5.75, hence Q1 is three quarters between the 5th and 6th observations in the sorted list. The 5th value is 60 and the 6th value is 60, thus
Q1 = 60 + .75(60-60)=60.

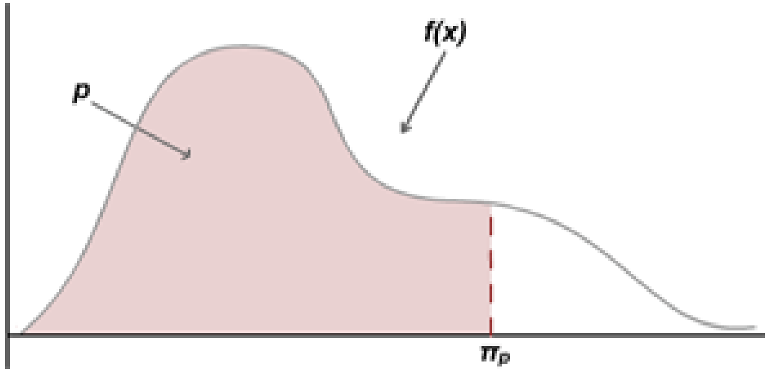
For Q2, (n+1)/2 = 23/2 = 11.5, e.g. half way between the 11th and 12th obs.
Q2 = 160 + .5(160-160) = 160.

For Q3, 3(n+1)/4 = 3(23)/4 = 69/4 = 17.25, a quarter of the way between the 17th and 18th observations.
Q3 = 180 + .25(180-180) = 180

10	60	60	60	60	60	70	130	140	140	160	160	160	160	160	160	180	180	200	210	210	210

Percentiles

Definition. If X is a continuous random variable, then the **$(100p)^{th}$ percentile** is a number π_p such that the area under $f(x)$ and to the left of π_p is p.



That is, p is the integral of $f(x)$ from $-\infty$ to π_p:

$$p = \int_{-\infty}^{\pi_p} f(x)dx = F(\pi_p)$$

We assign these detailed calculations to software packages…

Some percentiles are given special names:

- The 25th percentile, $\pi_{0.25}$, is called the **first quartile** (denoted q_1).
- The 50th percentile, $\pi_{0.50}$, is called the **median** (denoted m) or the **second quartile** (denoted q_2).
- The 75th percentile, $\pi_{0.75}$, is called the **third quartile** (denoted q_3).

Interquartile Range (IQR)

Difference between the third quartile (Q3) and the first quartile (Q1).

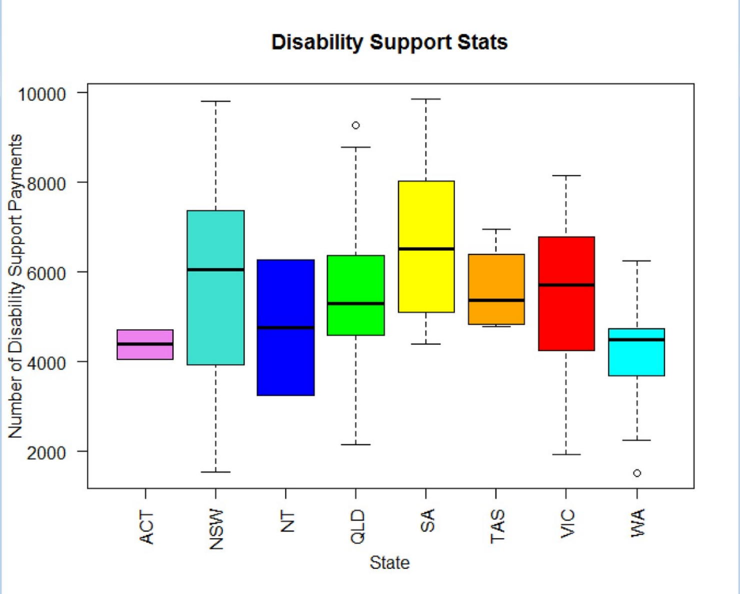
Quartiles:
Q1 = 25th = 60
Q2 = 50th = median = 160
Q3 = 75th = 180

- IQR = Q3-Q1 = 180 - 60 = 120

- The first quartile, Q1, is the value for which 25% of the observations are smaller and 75% are larger
- Q2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Ready for Box Plots!

John Tukey - 1977



Box Plot for Calories

A visualization of most of the basic statistics.