Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr

http://avesis.yildiz.edu.tr/zeyneb/

Data Exploration (cont'd)

Sample Mean

- Sample mean is sensitive to very large or very small values, which might be outliers (unusual values).
- For instance, suppose that we have measured the resting heart rate (in beats per minute) for five people.

$$x = \{74, 80, 79, 85, 81\}, \qquad \bar{x} = \frac{74 + 80 + 79 + 85 + 81}{5} = 79.8.$$

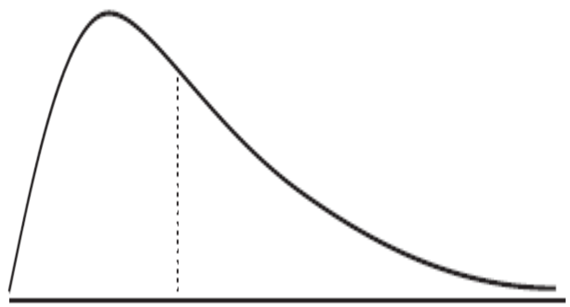
- In this case, the sample mean is 79.8, which seems to be a good representative of the data.
- Now suppose that the heart rate for the first individual is recorded as 47 instead of 74.

$$x = \{47, 80, 79, 85, 81\}, \qquad \bar{x} = \frac{47 + 80 + 79 + 85 + 81}{5} = 74.4.$$

- Now, the sample mean does not capture the central tendency.

Median

- Sometimes known as the mid-point.
  - It is used to represent the average when the data are not symmetrical (skewed distribution)



- The median value of a group of observations or samples, x_i, is the middle observation when samples, x_i, are listed in descending order.

- Note that if the number of samples, n, is odd, the median will be the middle observation.
- If the sample size, n, is even, then the median equals the average of two middle observations.
- Compared with the sample mean, the sample median is less susceptible to outliers.

Median

- Compared with the sample mean, the sample median is less susceptible to outliers.
- For instance, consider the resting heart rate mentioned in slide 77;
- The sample medians (denoted x~) are

$$
\begin{aligned}
x &= \{74, 79, 80, 81, 85\}, \qquad \tilde{x} = 80; \\
x &= \{47, 79, 80, 81, 85\}, \qquad \tilde{x} = 80.
\end{aligned}
$$

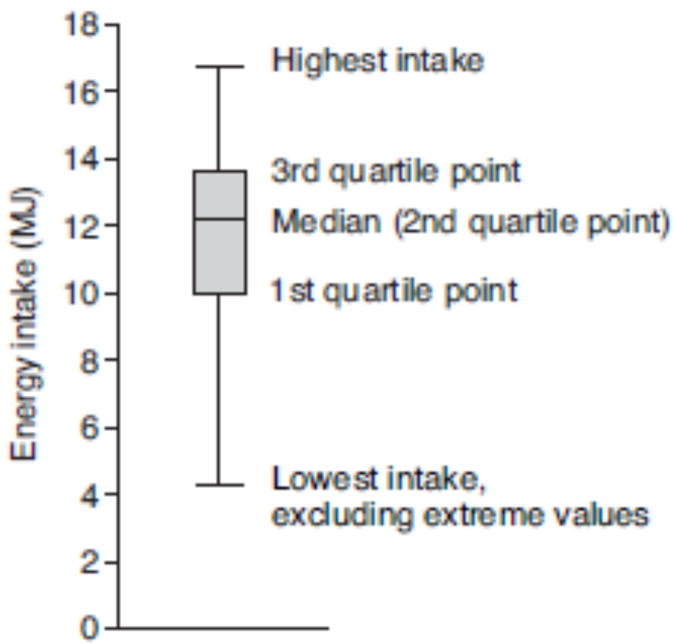
- So, the median is more robust against outliers.

Median

- The median may be given with its inter-quartile range (IQR).
- The 1st quartile point has the 1⁄4 of the data below it
- The 3rd quartile point has the 3⁄4 of the sample below it
- The IQR contains the middle 1⁄2 of the sample
- This can be shown in a “box and whisker” plot.

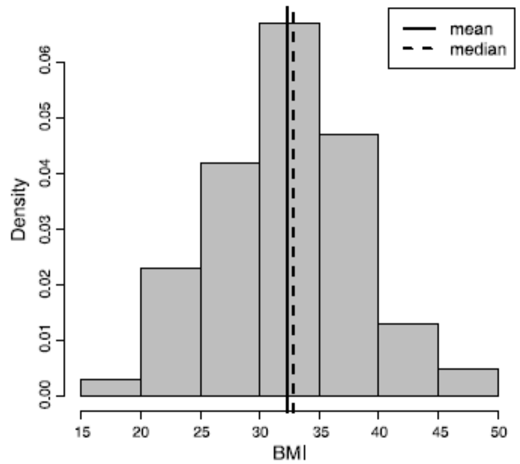
Median (example)

- A dietician measured the energy intake over 24 hours of 50 patients on a variety of wards. One ward had two patients that were "nil by mouth". The median was 12.2 megajoules, IQR 9.9 to 13.6. The lowest intake was 0, the highest was 16.7.
- This distribution is represented by the box and whisker plot below.

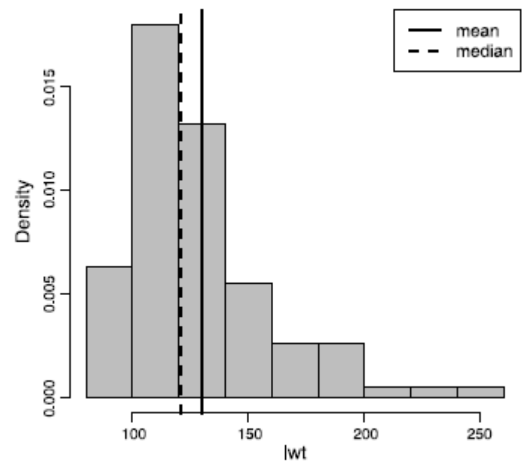


- Box and whisker plot of energy intake of 50 patients over 24 hours.
- The ends of the whiskers represent the maximum and minimum values, excluding extreme results like those of the two "nil by mouth" patients.

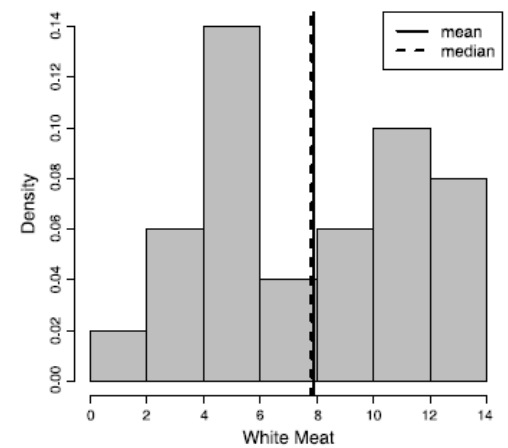
Sample Mean and Median



Histogram of *bmi*.
in the *Pima.tr* data set.



Histogram of *lwt*.
in the *birthwt* data set.



Histogram of *WhiteMeat*
in the *Protein* data set.

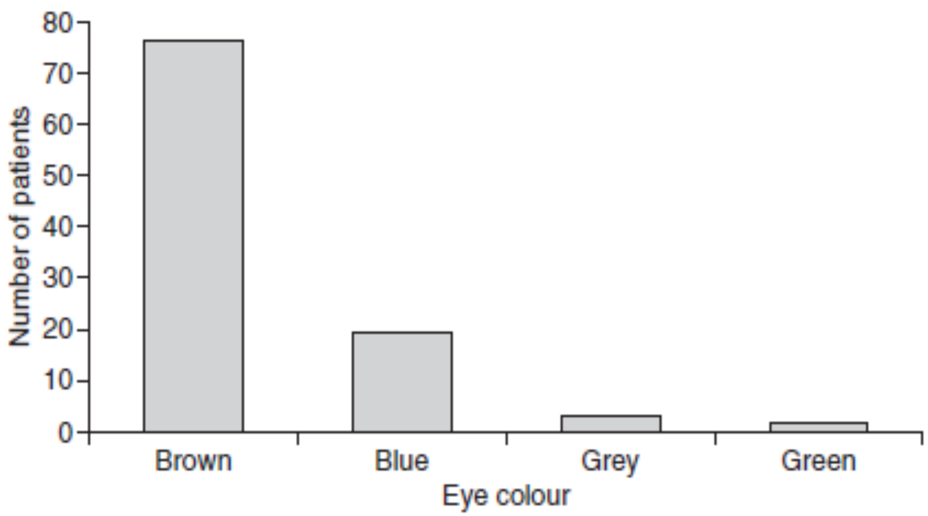
The mean and median are nearly equal since the histogram is Symmetric.

The mean is shifted to the right of the median. Because the histogram is skewed to the right.

Neither mean nor median is a good measurement for central tendency since the histogram is bimodal.

Mode

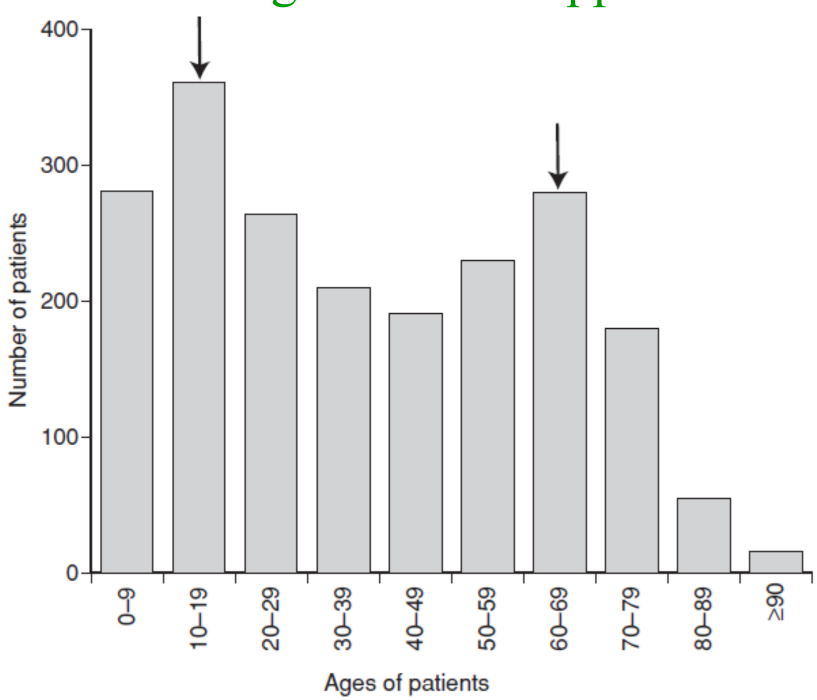
- the most common of a set of events
  - used when we need a label for the most frequently occurring event
    - Example: An eye clinic sister noted the eye colour of 100 consecutive patients. The results are shown below



- Graph of eye colour of patients attending an eye clinic.
- In this case the mode is brown, the commonest eye colour.

Mode

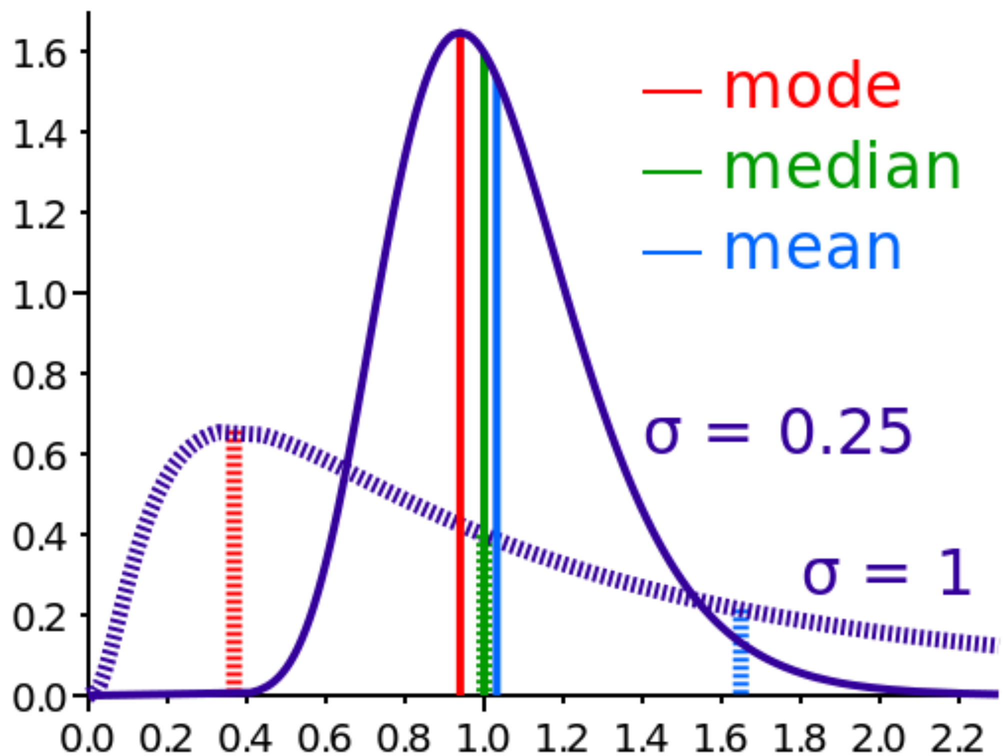
- You may see reference to a bi-modal distribution.
  - Generally when this is mentioned in papers it is as a concept rather than from calculating the actual values,
    - e.g. “The data appear to follow a bi-modal distribution”.



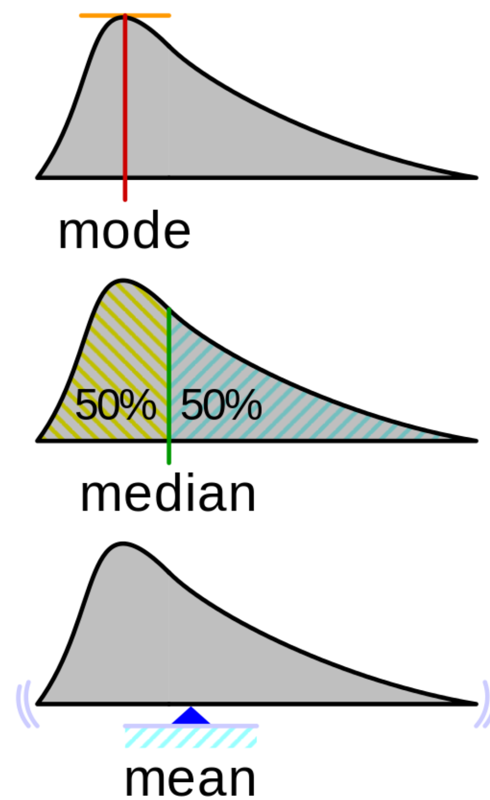
- Graph of ages of patients with asthma in a practice
  - The arrows point to the modes at ages 10–19 and 60–69.
- Bi-modal data may suggest that two populations are present that are mixed together,
  - so an average is not a suitable measure for the distribution.

Mean, Median, Mode

- Comparison of the arithmetic mean, median and mode of two skewed (log-normal) distributions.



- Geometric visualisation of the mode, median and mean of an arbitrary probability density function.



An applicaton of mean: moving AVERAGE filter

- Highlights trends in a signal (smoothing)

$$x_n : n = 1, \ldots, N$$

$$y_n = \sum_{j=-k}^{k} w_j x_{n+j} \quad : \quad n = k+1, \ldots, N-k,$$

k: pozitif integer, w_j: weights, $\Sigma\, w_j = 1$

- Algorithm for the 1st order MA filter

```
for n=1:N
        y(n)=0.5*(x(n)+x(n+1));
end
```

- Example (2 point moving AVERAGE filter)

$x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$

$y=([x_1+x_2]/2, [x_2+x_3]/2, [x_3+x_4]/2, \ldots)$

Complementary procedure: moving DIFFERENCE filter

- Removes trends from a signal (sharpening)
- 1st order differencing

$$\mathrm{D}y_t = y_t - y_{t-1}$$

- Higher order differences (2nd order)

$$\mathrm{D}^2 y_t = \mathrm{D}(\mathrm{D}y_t) = \mathrm{D}y_t - \mathrm{D}y_{t-1} = y_t - 2y_{t-1} + y_{t-2}$$

- Example (1st order moving DIFFERENCE filter)

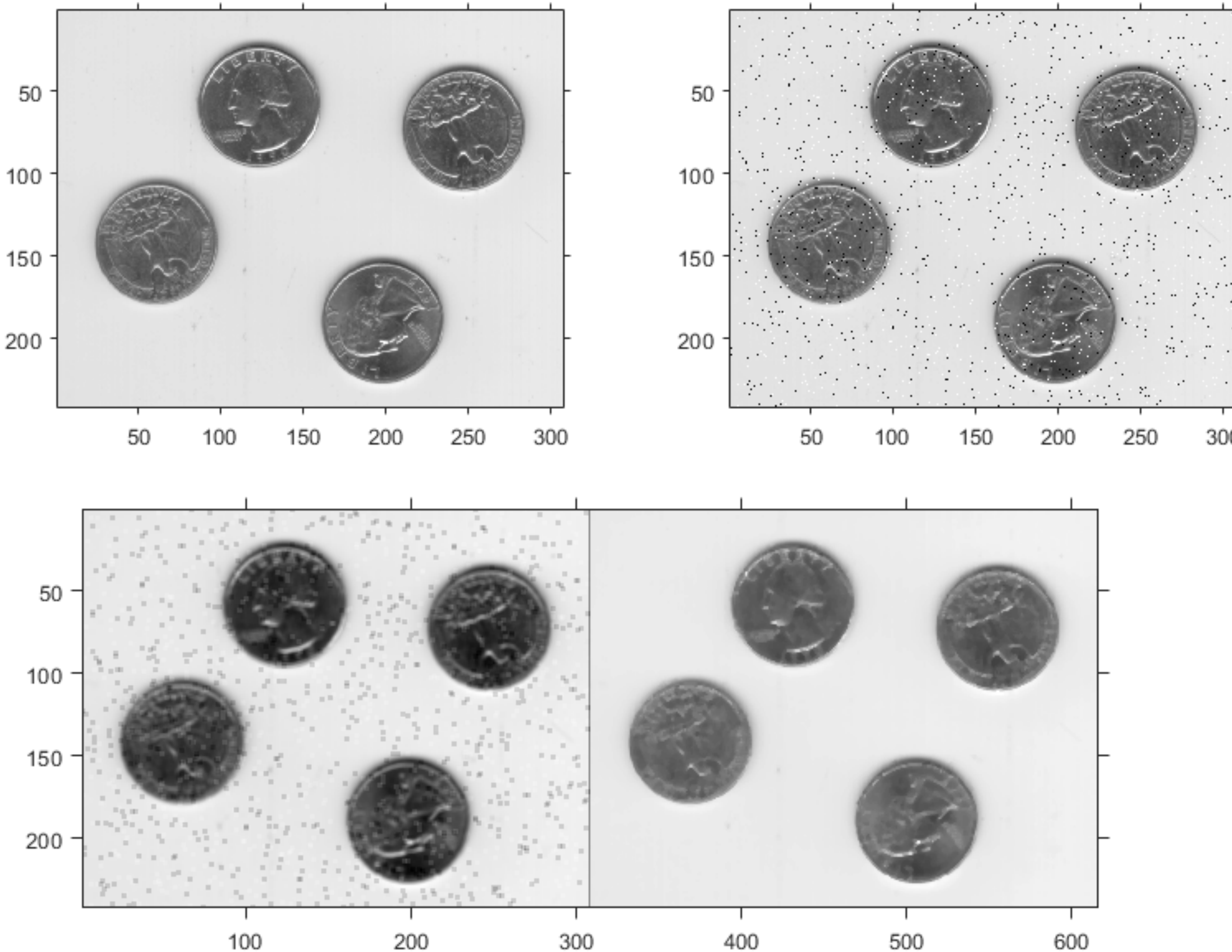
$$x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$$
$$y=([x_2-x_1], [x_3-x_2], [x_4-x_3], \ldots)$$

Moving median filtering

- Useful in impulsive noise removal (image processing, sliding median filtering)
- Example:
  - 3 point moving median filtering

$$x=(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, \ldots)$$
$$y=(\text{med}[x_1,x_2,x_3], \text{med}[x_2,x_3,x_4], \text{med}[x_3,x_4,x_5], \ldots)$$

  - If a window with even number of samples are selected median is average of two mid-point samples

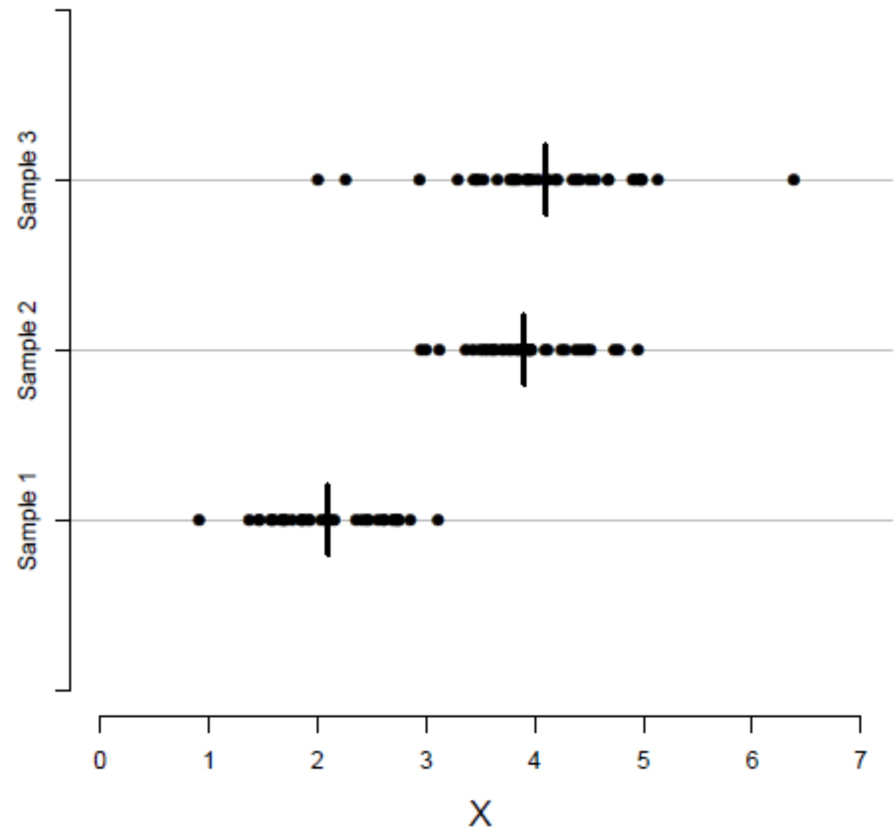

50
100
150
200
50 100 150 200 250 300
50
100
150
200
50 100 150 200 250 300
50
100
150
200
100 200 300 400 500 600

Measures of Variability

- When summarizing the variability of a population or process, we typically ask,
  - "How far from the center (sample mean) do the samples (data) lie?"
- To answer this question, we typically use the following estimates that represent the spread of the sample data:
  - sample variance,
  - sample standard deviation.
  - interquartile ranges,

Variance and standard deviation

- Consider Sample 2 and Sample 3.
- The two samples have similar locations, but Sample 3 is more dispersed than Sample 2.
- The deviations (differences) of observations from the center (e.g., mean) tend to be larger in Sample 3 compared to Sample 2.



Variance and standard deviation

- Two common summary statistics for measuring dispersion are the sample variance and sample standard deviation.
- These two summary statistics are based on the deviation of observed values from the mean as the center of the distribution.
- For each observation, the deviation from the mean is calculated as

$$x_i - \bar{x}$$

Variance and standard deviation

- The sample variance is a common measure of dispersion based on the squared deviations.

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

- The square root of the variance is called the sample standard deviation.

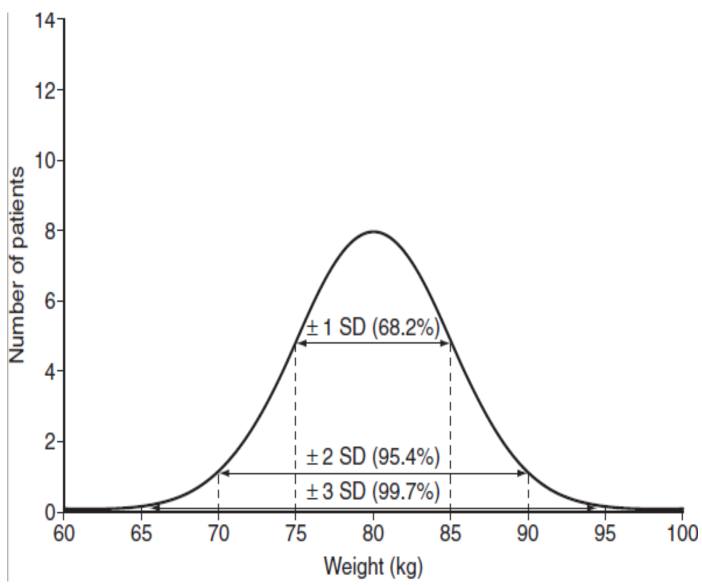
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

Measures of Variability

- Standard deviation (SD) is used for data which are "normally distributed",
  - to provide information on how much the data vary around their mean.
- SD indicates how much a set of values is spread around the average.
  - A range of one SD above and below the mean (abbreviated to ± 1 SD) includes 68.2% of the values.
  - ± 2 SD includes 95.4% of the data.
  - ± 3 SD includes 99.7%.

Measures of Variability

- Example 1:
- Let us say that a group of patients enrolling for a trial had a normal distribution for weight. The mean weight of the patients was 80 kg. For this group, the SD was calculated to be 5 kg.
- normal distribution of weights of patients enrolling in a trial with mean 80 kg, SD 5 kg.



- 1 SD below the average is 80 – 5 = 75 kg.
- 1 SD above the average is 80 + 5 = 85 kg.
- ± 1 SD will include 68.2% of the subjects, so 68.2% of patients will weigh between 75 and 85 kg.
- 95.4% will weigh between 70 and 90 kg (± 2 SD).
- 99.7% of patients will weigh between 65 and 95 kg (± 3 SD)

Variance and standard deviation

- Example 2

Patient A			Patient B		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
95	-1	1	85	-11	121
98	2	4	106	10	100
96	0	0	88	-8	65
95	-1	1	105	9	81
96	0	0	96	0	0
Σ	0	6	Σ	0	366

Patient A:

$s^2 = 6/4 = 1.5$

$s = \sqrt{1.5} = 1.22$

Patient B:

$s^2 = 366/4 = 91.5$

$s = \sqrt{91.5} = 9.56$

Variance and standard deviation

- some properties that can help you when interpreting a standard deviation:
  - The standard deviation can never be a negative number.
  - The smallest possible value for the standard deviation is 0
    - (when every number in the data set is exactly the same).
  - Standard deviation is affected by outliers, as it's based on distance from the mean, which is affected by outliers.
  - The standard deviation has the same units as the original data, while variance is in square units.

Measures of Variability

- It is important to note that for normal distributions (symmetrical histograms),
  - sample mean and sample deviation are the only parameters needed to describe the statistics of the underlying phenomenon.
- Thus, if one were to compare two or more normally distributed populations,
  - one only needs to test the equivalence of the means and variances of those populations.

Quantile

- comes from the word quantity
- A quantile is where a sample is divided into equal-sized, adjacent, subgroups
  - (quantile is also called a fractile)
- It can also refer to dividing a probability distribution into areas of equal probability
- Quartiles are also quantiles;
  - they divide the distribution into four equal parts.
- Percentiles are quantiles;
  - they divide a distribution into 100 equal parts
- Deciles are quantiles;
  - they divide a distribution into 10 equal parts.

Percentiles

- the most common way to report relative standing of a number within a data set
- A percentile is the percentage of individuals in the data set who are below where your particular number is located.
  - For example,
  - if your exam score is at the 90th percentile, that means
    - 90% of the people taking the exam with you scored lower than you did
    - 10 percent scored higher than you did

Percentiles

- Steps to calculate the k^{th} percentile (where k is any number between 1 and one 100):
  1. Order all the numbers in the data set from smallest to largest.
  2. Multiply k percent times the total number of numbers, n.

  3a. If your result from Step 2 is a whole number, go to Step 4.

  If the result from Step 2 is not a whole number, round it up to the nearest whole number and go to Step 3b.

Percentiles

3b. Count the numbers in your data set from left to right (from the smallest to the largest number) until you reach the value from Step 3a.

This corresponding number in your data set is the k^{th} percentile.

4. Count the numbers in your data set from left to right until you reach that whole number.

The k^{th} percentile is the average of that corresponding number in your data set and the next number in your data set.

Percentiles - example

- Suppose 25 test scores, in order from lowest to highest:

  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99.

- To find the 90^{th} percentile for these scores
  - multiply 90% by the total number of scores,
    - $90\% \times 25 = 0.90 \times 25 = 22.5$ (step 2).
    - This is not a whole number;
  - Step 3a says round up to the nearest whole number, 23, then go to step 3b

Percentiles - example

- Counting from left to right
  - you go until you find the 23^{rd} number in the data set.
- That number is 98,
  - which is the 90^{th} percentile for this data set.

- To find the 20^{th} percentile,
  - take $0.20 * 25 = 5$;
    - this is a whole number so proceed to Step 4, which tells us the 20^{th} percentile is the average of the 5^{th} and 6^{th} numbers in the ordered data set (62 and 66).
  - 20^{th} percentile then becomes $(66 + 62) / 2 = 64$
- The median is the 50^{th} percentile,
  - the point in the data where 50% of the data fall below that point and 50% fall above it.
    - The median for the test scores example is the 13^{th} number, 77.

Percentiles

- A percentile is **not** a percent;
  - a percentile is a number that is a certain percentage of the way through the data set,
    - when the data set is ordered.
- Suppose your score on the GRE was reported to be the 80^{th} percentile.
  - This does not mean you scored 80% of the questions correctly.
  - It means that 80% of the students' scores were lower than yours, and 20% of the students' scores were higher than yours.

Quartile

- For sampled data, the median is also known as
  - the 2nd quartile, Q2.
- Given Q2, we can find the 1st quartile, Q1,
  - by simply taking the median value of those samples that lie below the 2nd quartile.
- We can find the 3d quartile, Q3,
  - by taking the median value of those samples that lie above the 2nd quartile.
- Quartiles can also be found in terms of percentiles:
  - 1st quartile is 25th percentile
  - 2nd quartile is 50th percentile
  - 3rd quartile is 75th percentile

Quartile

- Considering the following (25) test scores

  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99

- Q1 (25^{th} percentile)

  $0.25 * 25 = 6.25 \rightarrow$ (round up) $\rightarrow 7$ Q1 = 68

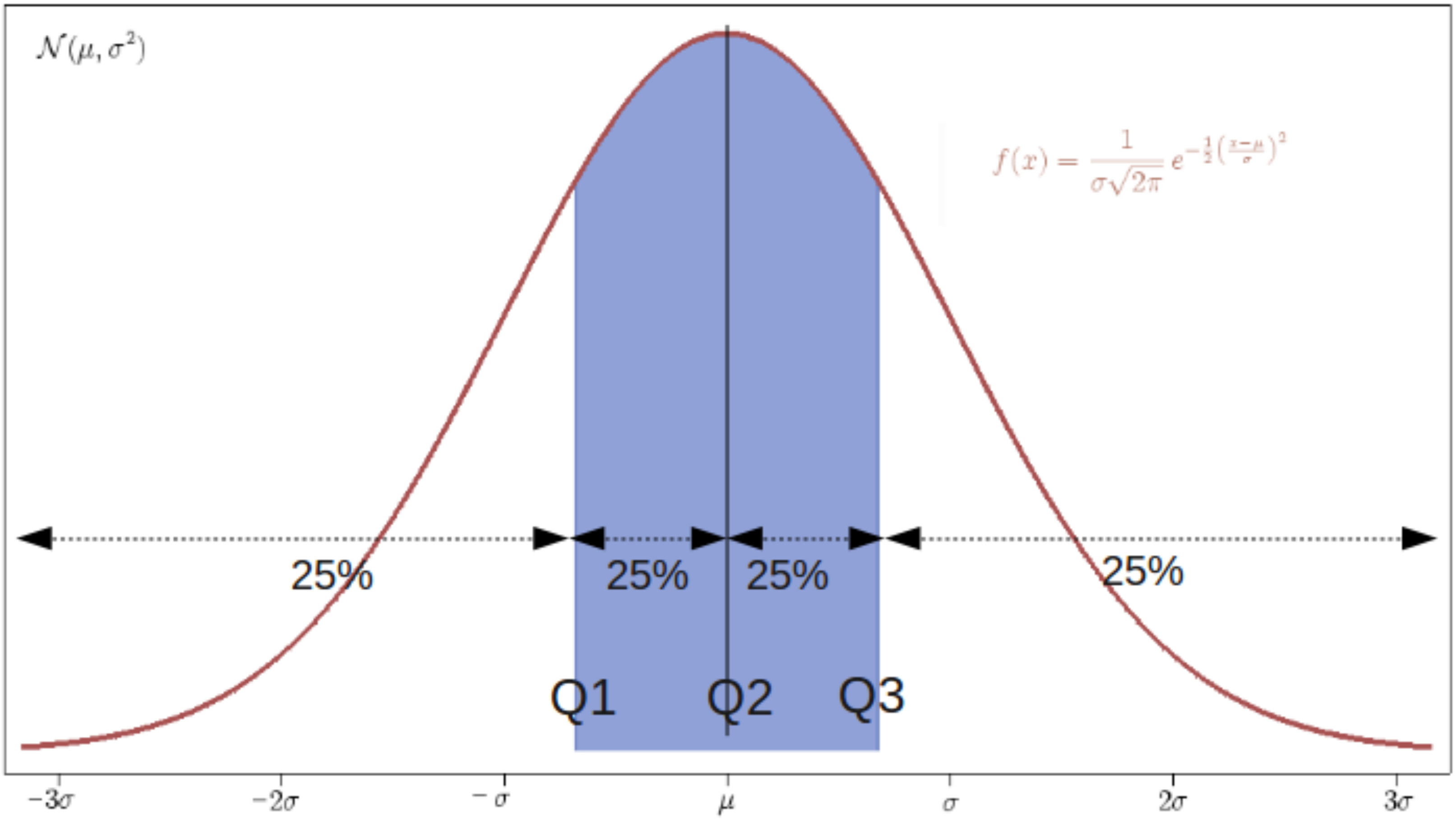
- Q2 (50^{th} percentile)

  $0.50 * 25 = 12.5 \rightarrow$ (round up) $\rightarrow 13$ Q2 = 77

- Q3 (75^{th} percentile)

  $0.75 * 25 = 18.75 \rightarrow$ (round up) $\rightarrow 19$ Q3 = 89

Measures of Variability



Five-number summary

- The minimum (min), which is the smallest value of the variable in our sample, is in fact the 0 quantile.
- On the other hand, the maximum (max), which is the largest value of the variable in our sample, is the 1 quantile.
- The minimum and maximum along with quartiles (Q1, Q2, and Q3) are known as five-number summary.
- These are usually presented in the increasing order:
  - min, 1st quartile, median, 3rd quartile, max
  - min, 25th percentile, median, 75th percentile, max
- This way, the five-number summary provides
  - 0, 0.25, 0.50, 0.75, and 1 quantiles

Five-number summary

- The five-number summary can be used to derive two measures of dispersion:
  - the range
    - the difference between the maximum observed value and the minimum observed value.
  - the interquartile range (IQR)
    - the difference between the third quartile (Q3) and the first quartile (Q1).

$$IQR = Q3 - Q1$$

Measures of Variability – example 1

- As an illustration, we have the following samples:

  99, 99, 56, 61, 62, 66, 68, 98, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 69, 54, 43
- list these samples in ascending order,

  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
- the median value (Q2) for these samples is 77 (13^{th} sample).
- The 1^{st} quartile, Q1, can be found by taking the median of the following samples,

  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77
  - which is 68
- The 3^{rd} quartile, Q3, may be found by taking the median value of the following samples:

  77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
  - which is 89.
- Thus, the interquartile range, (Q1 = 68; Q2 = 77; Q3 = 89)

  $Q3 - Q1 = 89 - 68 = 21$

Measures of Variability – example 1

- Using percentiles;
  - list the samples in ascending order,

    43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
- Q1 (25^{th} percentile)

  0.25 * 25 = 6.25 ➔ (round up) ➔ 7 Q1 = 68
- Q2 (50^{th} percentile)

  0.50 * 25 = 12.5 ➔ (round up) ➔ 13 Q2 = 77
- Q1 (75^{th} percentile)

  0.75 * 25 = 18.75 ➔ (round up) ➔ 19 Q3 = 89

- In this case, the interquartile range, (Q1 = 68; Q2 = 77; Q3 = 89)

  $Q3 - Q1 = 89 - 68 = 21$

Measures of Variability – example 1

- Alternative calculation;
  - Use the following formula to estimate the ith observation:

    i^{th} observation = q (n + 1)
  - where q is the quantile, n is the number of items in a data set
- list the samples in ascending order;

  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 9
- Q1 (25^{th} percentile)

  0.25 * (25 +1) = 6. 5 ➔ (round down) ➔ 6 Q1 = 66
- Q2 (50^{th} percentile)

  0.50 * (25 +1) = 13 ➔ 13 Q2 = 77
- Q1 (75^{th} percentile)

  0.75 * (25 +1) = 19.5 ➔ (round down) ➔ 19 Q3 = 89

- In this case, the interquartile range, (Q1 = 66; Q2 = 77; Q3 = 89)

  Q3 − Q1 = 89 − 66 = 23

Measures of Variability – example 2

- As an illustration, we have the following samples:

  1, 3, 3, 2, 5, 1, 1, 4, 3, 2.
- list these samples in descending order,

  5, 4, 3, 3, 3, 2, 2, 1, 1, 1
- the median value (Q2) for these samples is 2.5
- The 1st quartile, Q1, can be found by taking the median of the following samples,

  2.5, 2, 2, 1, 1, 1
  - which is 1.5
- The 3rd quartile, Q3, may be found by taking the median value of the following samples:

  5, 4, 3, 3, 3, 2.5
  - which is 3.
- Thus, the interquartile range, (Q1 = 1.5; Q2 = 2.5; Q3 = 3)

$$Q3 - Q1 = 3 - 1.5 = 1.5$$

Measures of Variability – example 2

- Using percentiles;
  - list the samples in ascending order,

    1, 1, 1, 2, 2, 3, 3, 3, 4, 5
- Q1 (25th percentile)

  0.25 *10 = 2.5 ➔ (round up) ➔ 3   Q1 = 1
- Q2 (50th percentile)

  0.50 *10 = 5 ➔ 5   Q2 = (2+3)/2 = 2.5
- Q1 (75th percentile)

  0.75 *10 = 7.5 ➔ (round up) ➔ 8   Q3 = 3

- In this case, the interquartile range, (Q1 = 1; Q2 = 2.5; Q3 = 3)

  Q3 − Q1 = 3 − 1 = 2

Measures of Variability – example 2

- Alternative calculation;
  - Use the following formula to estimate the i^{th} observation:

    i^{th} observation = $q\ (n + 1)$
  - where q is the quantile, n is the number of items in a data set
- list the samples in ascending order; 1, 1, 1, 2, 2, 3, 3, 3, 4, 5

- Q1 (25^{th} percentile)

  $0.25 * (10 + 1) = 2.75$ ➔ (round down) ➔ 2   Q1 = 1
- Q2 (50^{th} percentile)

  $0.50 * (10 + 1) = 5.5$ ➔ (round down) ➔ 5   Q2 = 2
- Q1 (75^{th} percentile)

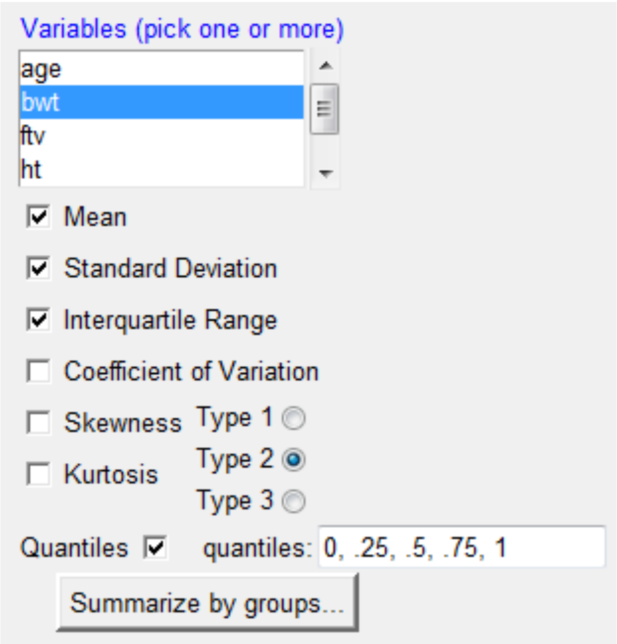
  $0.75 * (10 + 1) = 8.25$ ➔ (round down) ➔ 8   Q3 = 3

- In this case, the interquartile range, (Q1 = 1; Q2 = 2; Q3 = 3)

  $Q3 - Q1 = 3 - 1 = 2$

Five-number summary

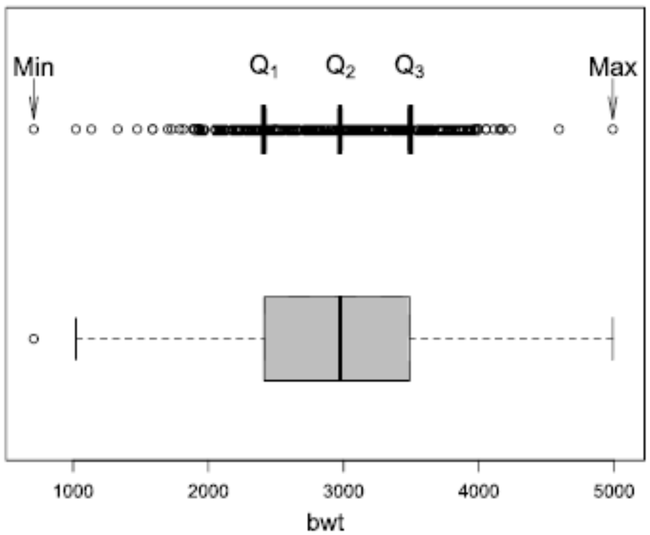
- We can use R-Commander to obtain the five-number summary along with mean and standard deviation.
- Make sure *birthwt* is the active data set.
  - Click *Statistics → Summaries → Numerical summaries*.
  - Now select *bwt*.
    - Make sure *Mean, Standard Deviation*, *Interquantile* and *Quantiles* are checked.
  - The resulting summary statistics are:

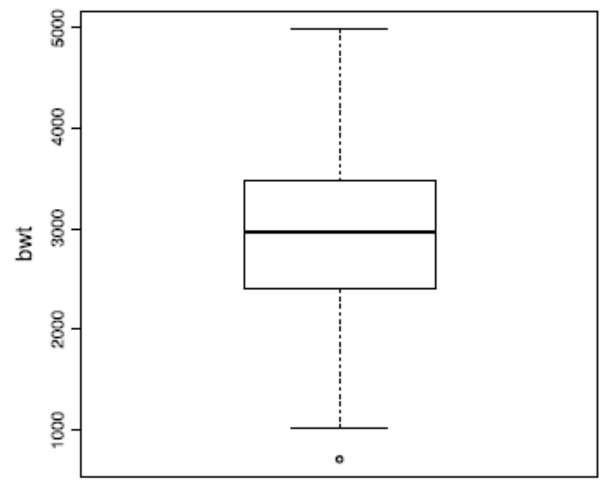



```
> numSummary(Dataset[,"bwt"], statistics=c("mean", "sd", "IQR", "quantiles"),
+   quantiles=c(0,.25,.5,.75,1))
     mean       sd  IQR  0%  25%  50%  75% 100%   n
 2944.656 729.0224 1061 709 2414 2977 3475 4990 189
```

Boxplot

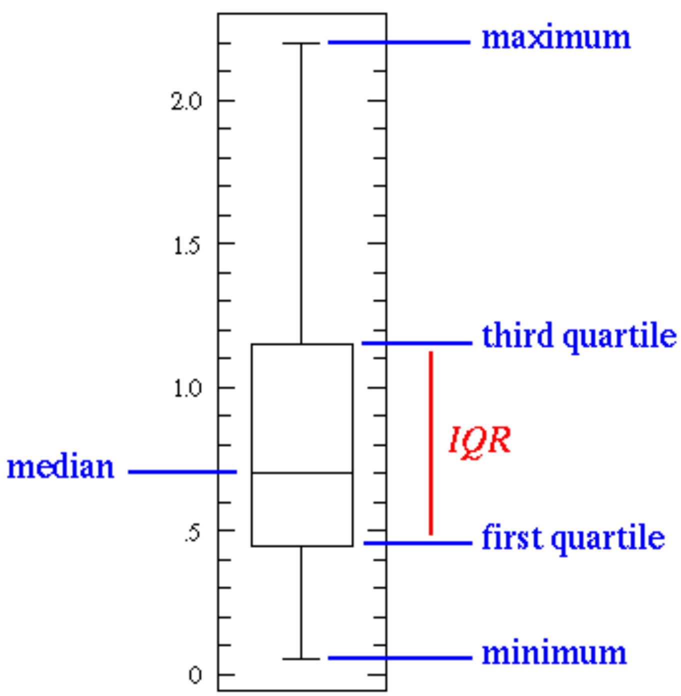
- To visualize the five-number summary, the range and the IQR,
  - we often use a boxplot
    - a.k.a. box and whisker plot





- Very often, boxplots are drawn vertically.
- To create a boxplot for *bwt* in R-Commander,
  - make sure *birthwt* is the active dataset,
  - click *Graphs* → *Boxplot*, and select *bwt*.

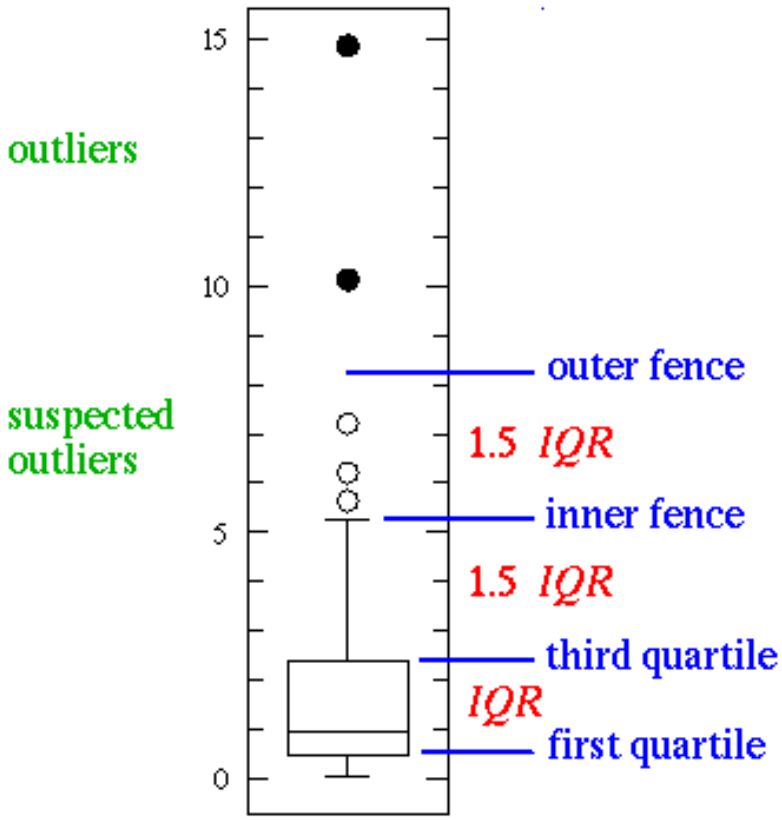
Boxplot



- This simplest possible box plot displays the full range of variation (from min to max), the likely range of variation (the IQR), and a typical value (the median).
- Not uncommonly real datasets will display surprisingly high maximums or surprisingly low minimums called outliers.

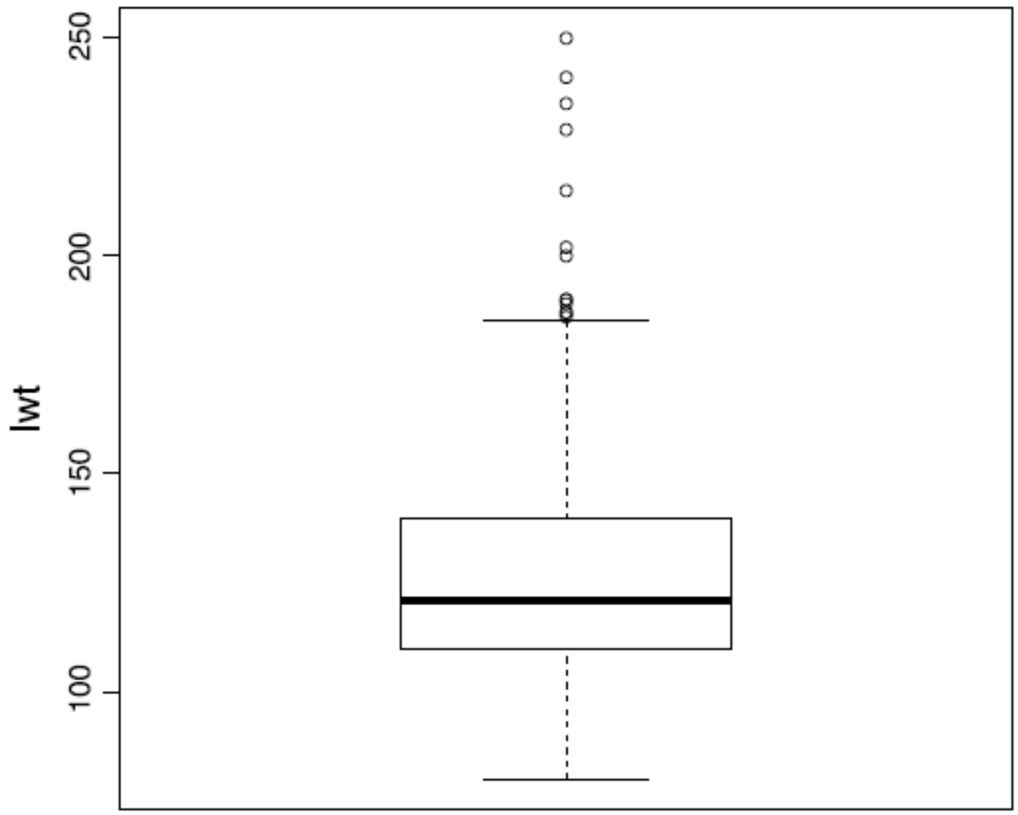
- John Tukey has provided a precise definition for two types of outliers:
  - Outliers are either 3×IQR or more above the third quartile or 3×IQR or more below the first quartile.
  - Suspected outliers are slightly more central versions of outliers:
    - either 1.5×IQR or more above the third quartile
      - (Q3 + 1.5 x IQR)
    - or 1.5×IQR or more below the first quartile
      - (Q1-1.5 x IQR)

Boxplot



- If either type of outlier is present
  - the whisker on the appropriate side is taken to 1.5×IQR from the quartile (the "inner fence") rather than the max or min,
- individual outlying data points are displayed as
  - unfilled circles for suspected outliers
  - or filled circles for outliers.
- The "outer fence" is 3×IQR from the quartile.

Boxplot



- Vertical boxplot of *lwt*.
- This plot reveals that the variable *lwt* is right-skewed and there are several possible outliers,
  - whose values are beyond the whisker on the top of the box

Data Preprocessing

- We refer to data in their original form (i.e., collected by researchers) as the raw data.
- Before using the original data for analysis, we should thoroughly check them for missing values and possible outliers.
- We refer to the process of preparing the raw data for analysis as data preprocessing.
- The data set we have been using so far (*Pima.tr*) was obtained after removing these observations from *Pima.tr2*.

Missing Data

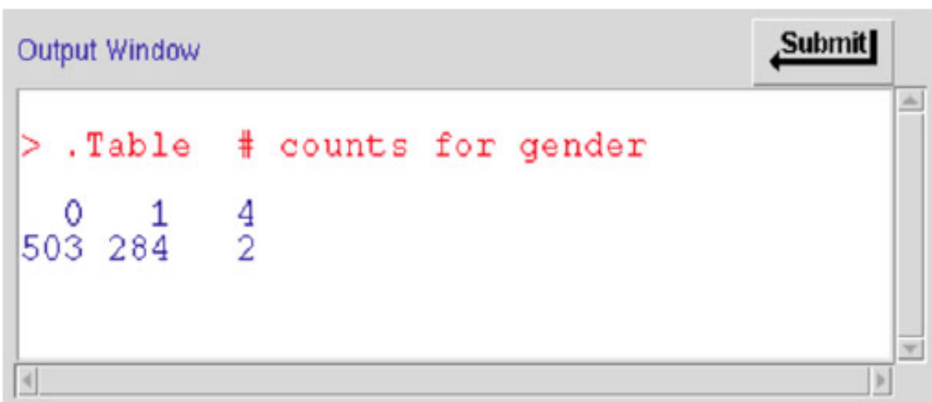
Pima.tr2

	npreg	glu	bp	skin	bmi	ped	age	type
193	1	128	48	45	40.5	0.613	24	Yes
194	2	112	68	22	34.1	0.315	26	No
195	1	140	74	26	24.1	0.828	23	No
196	2	141	58	34	25.4	0.699	24	No
197	7	129	68	49	38.5	0.439	43	Yes
198	0	106	70	37	39.4	0.605	22	No
199	1	118	58	36	33.3	0.261	23	No
200	8	155	62	26	34.0	0.543	46	Yes
201	2	134	70	NA	28.9	0.542	23	Yes
202	10	75	82	NA	33.3	0.263	38	No
203	0	146	70	NA	37.9	0.334	28	Yes
204	1	180	NA	NA	43.3	0.282	41	Yes
205	5	104	74	NA	28.8	0.153	48	No
206	9	164	78	NA	32.8	0.148	45	Yes
207	1	80	55	NA	19.1	0.258	21	No
208	4	171	72	NA	43.6	0.479	26	Yes

- Here, missing values are denoted NA (Not Available)
- In general, it is up to the researcher to decide whether to remove the observations with missing values or impute (guess) the missing values in order to keep the observations.
- To remove all observations with missing values
  - click *Data* → *Active data set* → *Remove cases with missing data.*
- To remove individual observations,
  - click *Data*→*Active data set* → *Remove row(s) from active data* and enter the *row numbers* for observations you want to remove.

Outliers

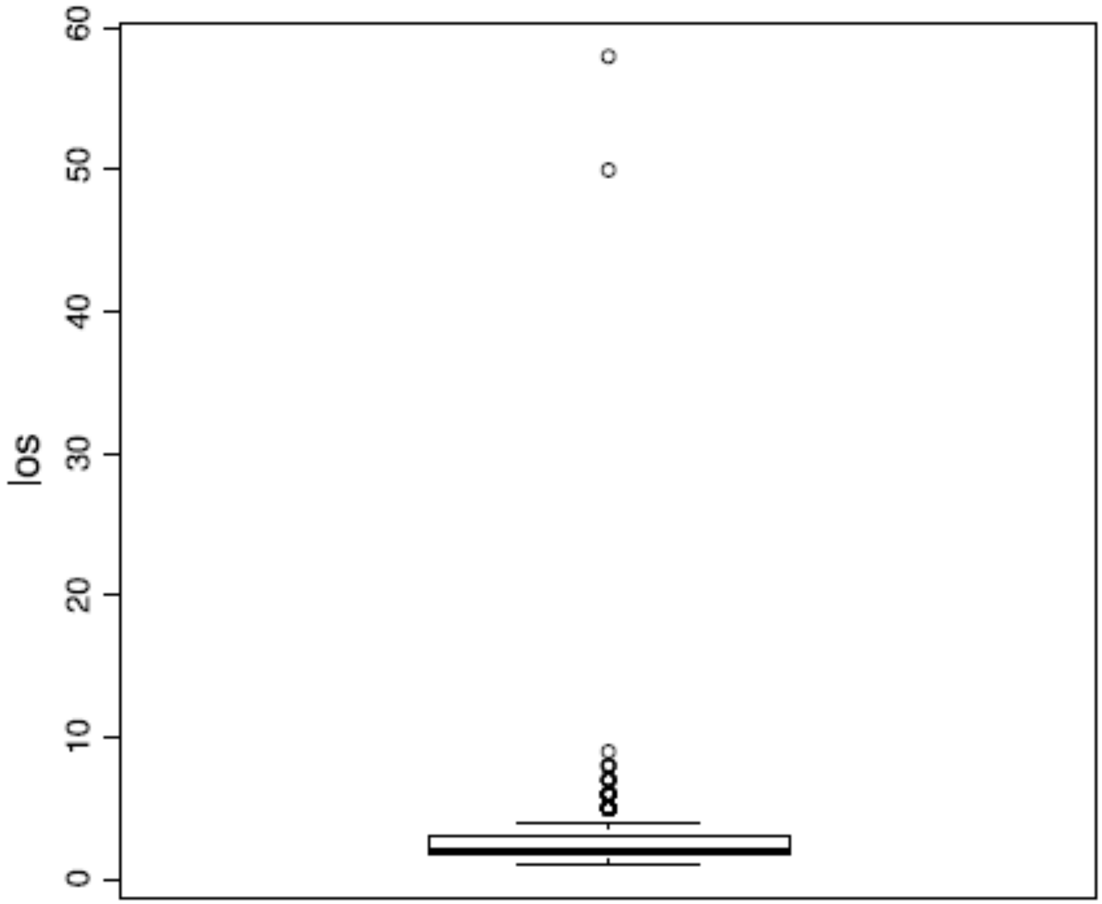
- Sometimes, an observed value of a variable is suspicious since it does not follow the overall patterns presented by the rest of the data.
  - We refer to such observations as outliers.
- For analyzing such data, we could use statistical methods that are more robust against outliers (e.g., median, IQR).
- Frequency table for gender from the *AsthmaLOS* data set.



- The value of gender for two observations are entered as “4”, while gender can only take 0 or 1

Data Set *AsthmaLOS*

- ***los*: length of stay in hospital (in days).
- ***hospital.id:*** hospital ID.
- ***insurer***: the insurer, which is either 0 or 1.
- ***age***: the age of the patient.
- ***gender***: the gender of the patient; 1 for female, and 0 for male.
- ***race***: the race of the patient; 1 for white, 2 for Hispanic, 3 for African-American, 4 for Asian/Pacific Islander, 5 for others.
- ***bed.size***: the number of beds in the hospital; 1 means 1 to 99, 2 means 100 to 249, 3 means 250 to 400, 4 means 401 to 650.
- ***owner.type***: the hospital owner; 1 for public, 2 for private.
- ***complication***: if there were any treatment complication; 0 means there were no complications, 1 means there were some complications.

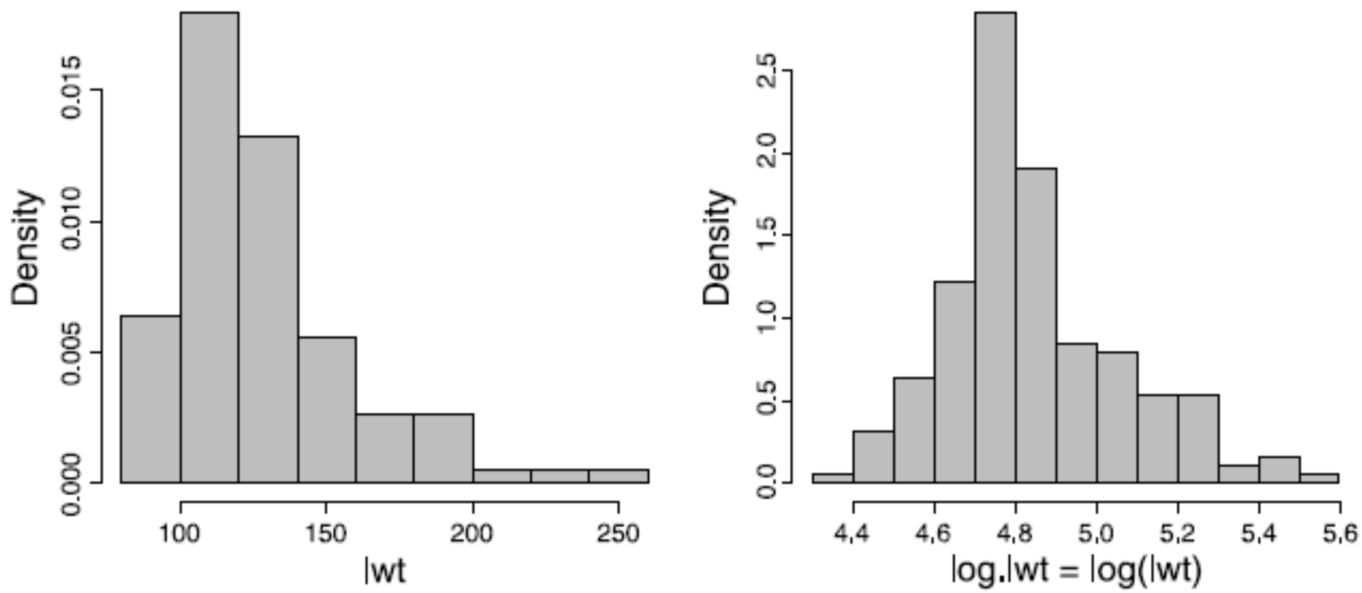


The boxplot of *los* with two extremely large values

Data Transformation

- We rely on data transformation techniques (i.e., applying a function to the variable)
  - to reduce the influence of extreme values in our analysis.
- Two of the most commonly used transformation functions for this purpose are
  - logarithm
  - square root.
- To use log-transformation,
  - Select the *birthwt* dataset.
  - click *Data→ Manage variables in active data set → Compute new variable*.
  - Under *New variable name*, enter *log.lwt*, and under *Expression to compute*, enter *log*(*lwt*)

Data Transformation



- *Left panel*: Histogram of variable *lwt* in the *birthwt* data set.
- *Right panel*: Histogram of log-transformation of variable *lwt*

Data Transformation

- The reasons for data transformation:
  - to make the distribution of the data normal,
    - this fulfills one of the assumptions of conducting a parametric means comparison.
  - to create more informative graphs of the data,
  - better outlier identification (or getting outliers in line)
  - increasing the sensitivity of statistical tests

Data Transformation

- A data transformation is defined to be a process in which the measurements on the original scale are systematically converted to a new scale of measurement.
- Transformations involve applying a mathematical function to each data point.
- A transformation is needed when the data is excessively skewed positively or negatively.

Some data transformations

- Different types of data are often better analyzed with different transformations: examples include:

  arcsine transformation $p' = \arcsin(\sqrt{p})$ (only for proportions);

  square root transformation $y' = \sqrt{y}$, often used for count data (the text suggests $\sqrt{y + 0.5}$);

  reciprocal transformation $y' = 1/y$, sometimes useful for ratios or strongly right-skewed data—even more extreme than ln;

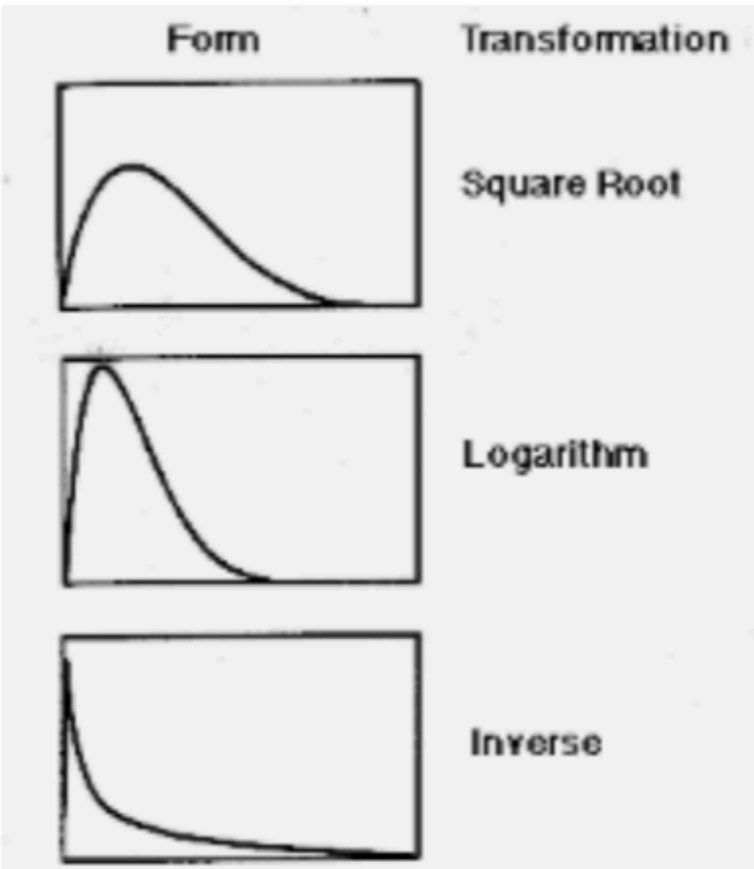
  square transformation $y' = y^2$, sometimes helps with left-skewed data;

  exponential transformation $y' = e^y$, sometimes helps with left-skewed data.

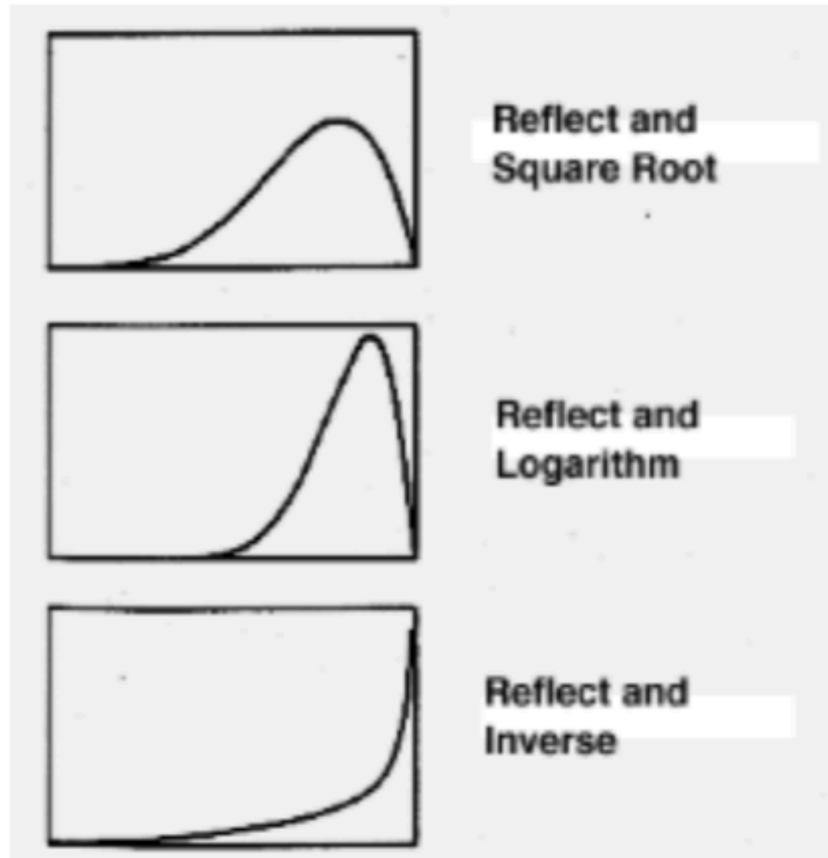
Data Transformation

- The figure below suggests the type of transformation that can be applied depending upon the degree of skewness.

Positively skewed data



Negatively skewed data



Data Transformation

- Logarithms:
  - Growth rates are often exponential and log transforms will often normalize them.
  - Log transforms are particularly appropriate if the variance increases with the mean.
- Reciprocal:
  - If a log transform does not normalize your data you could try a reciprocal (1/x) transformation.
    - This is often used for enzyme reaction rate data.

Data Transformation

- Square root:
  - used when the data are counts, e.g. blood cells on a haemocytometer or woodlice in a garden.
    - Carrying out a square root transform will convert data with a Poisson distribution to a normal distribution.
- Arcsine:
  - a.k.a. the angular transformation
  - especially useful for percentages and proportions which are not normally distributed.

Data Transformation

- Tabachnick and Fidell (2007) and Howell (2007) suggest to use the following guidelines when transforming data:

If your data distribution is…	**Try this transformation method**
Moderately positive skewness	Square-Root NEWX = SQRT(X)
Substantially positive skewness	Logarithmic (Log 10) NEWX = LG10(X)
Substantially positive skewness (with zero values)	Logarithmic (Log 10) NEWX = LG10(X + C)
Moderately negative skewness	Square-Root NEWX = SQRT(K – X)
Substantially negative skewness	Logarithmic (Log 10) NEWX = LG10(K – X)

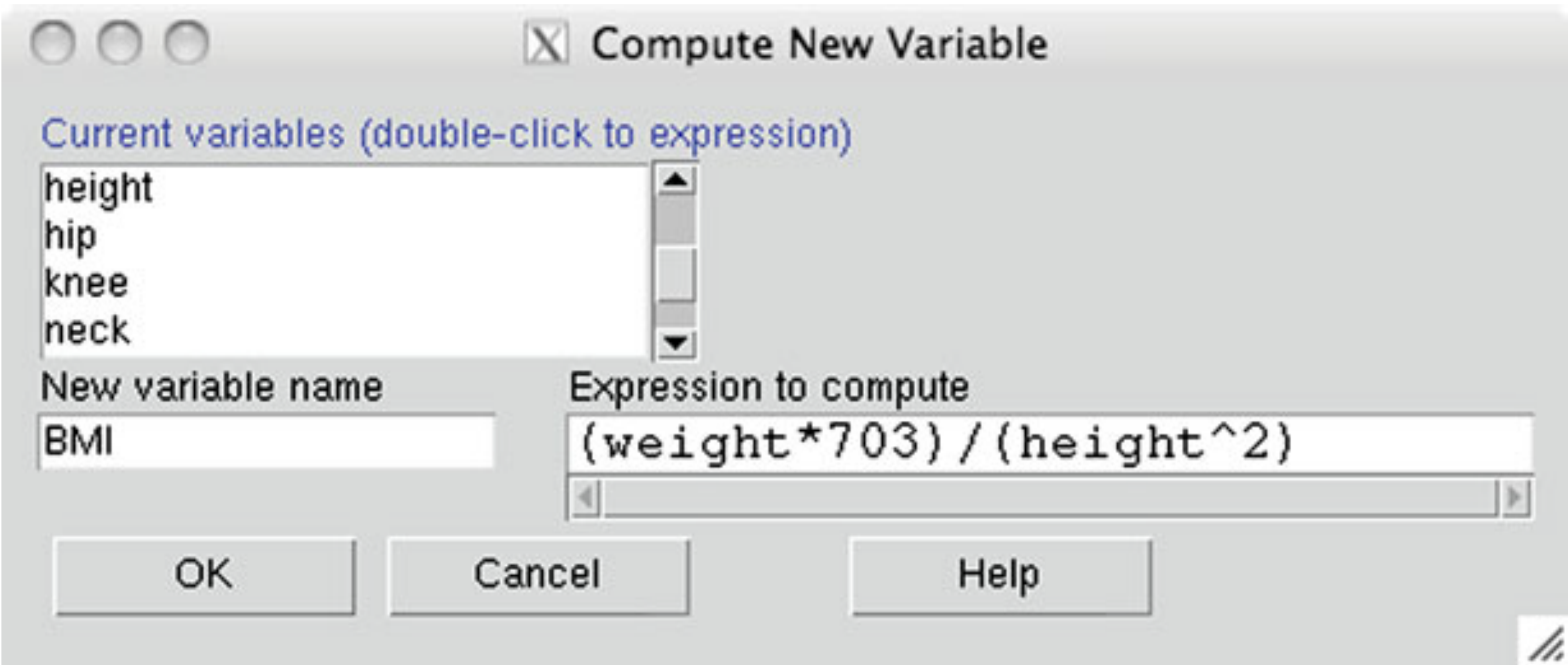
- C = a constant added to each score so that the smallest score is 1.
- K = a constant from which each score is subtracted

Howell, D. C. (2007). Statistical methods for psychology (6th ed.). Belmont, CA: Thomson Wadsworth.
Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston: Allyn and Bacon

Creating New Variable

- We can create a new variable based on two or more existing variables.
- Consider the *bodyfat* data set, which includes weight and height.
- To create BMI,
  - click *Data → Manage variables in active data set → Compute new variable.*
  - Under *New variable name*, enter *BMI*, and under *Expression to compute*, enter
    - (weight * 703)/(height^2)

- Creating a new variable *BMI* based on weight and height for each person in the *bodyfat* data set



Creating New Variable

- This will create a new variable called *BMI*.
- We can now investigate the linear relationship between this variable and percent body fat by calculating their sample correlation coefficient.
- Pearson's correlation coefficient between *siri* and *BMI* is 0.72,
  - which indicates a strong positive linear relationship as expected.

Creating Categories for Numerical Variables

- This could help us to see the patterns more clearly and identify relationships more easily.
- Histograms are created by dividing the range of a numerical variable into intervals.
- Instead of using arbitrary intervals, we might prefer to group the values in a meaningful way.

BMI	Weight Status
Below 18.5	Underweight
18.5–24.9	Normal
25.0–29.9	Overweight
30.0 and Above	Obese

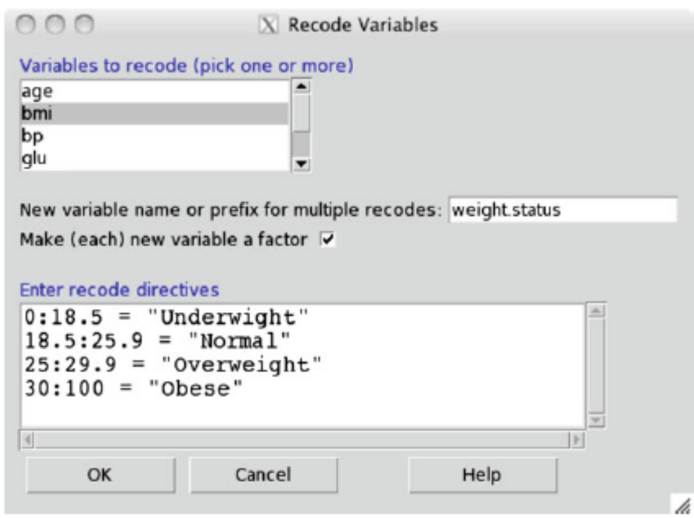
- Standard weight status based on *BMI* according to CDC (Centers for Disease Control and Prevention)

Creating Categories for Numerical Variables

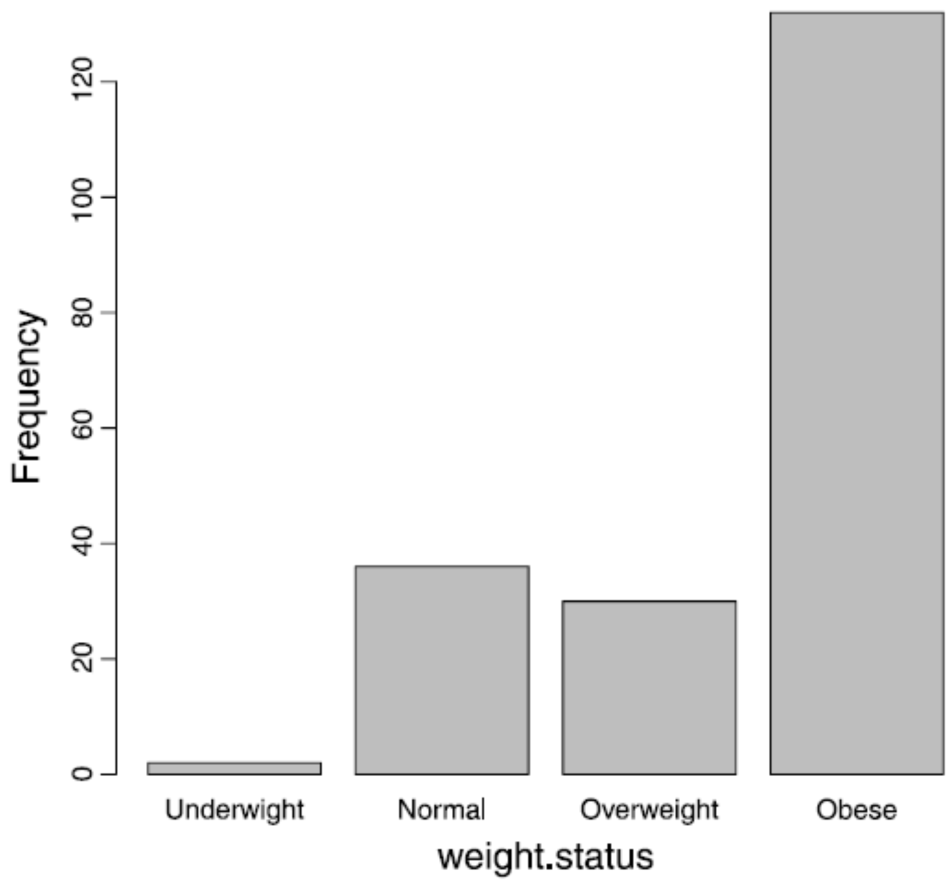
- In R-Commander, let us divide subjects based on their *bmi* (from the *Pima.tr*) into four groups:

  Underweight, Normal, Overweight, and Obese.

  – Click *Data → Manage variables in active data set → Recode variables.*
- To specify the order of categories in R-Commander,
  – click *Data → Manage variables in active data set → Reorder factor levels*. Then select *weight.status*.

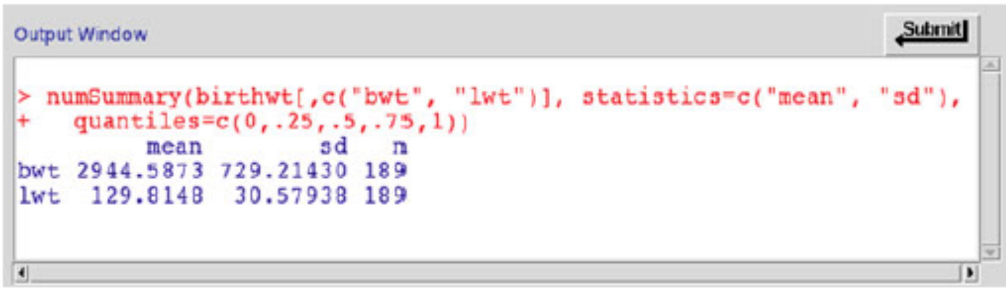


Creating Categories for Numerical Variables

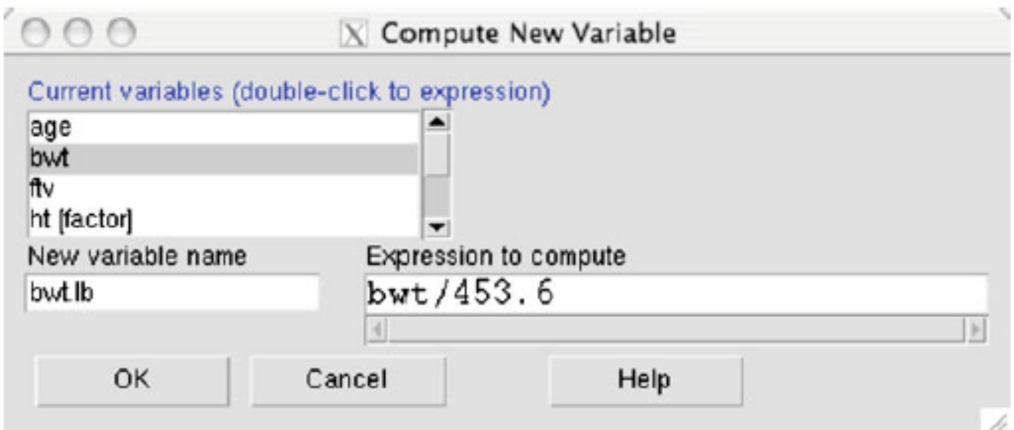


- The bar graph for *bmi* after converting the numerical variable to a categorical variable (*weight.status*)

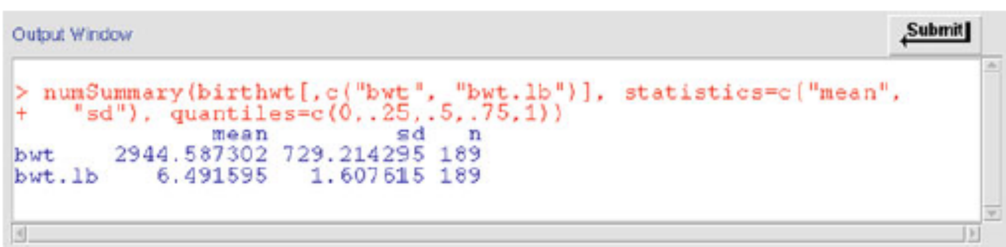
Creating a new variable and obtaining its summary statistics



- Summary statistics for *bwt* and *lwt* from the birthwt data set



- Creating a new variable *bwt.lb* (birth weight in pounds) and obtaining its summary statistics



- Creating a new variable *bwt.lb* (birth weight in pounds) and obtaining its summary statistics

Coefficient of Variation

- In general, the coefficient of variation is used to compare variables in terms of their dispersion when the means are substantially different
  - possibly as the result of having different measurement units.
- To quantify dispersion independently from units, we use the coefficient of variation,
  - which is the standard deviation divided by the sample mean
    - assuming that the mean is a positive number:

$$CV = \frac{s}{\bar{x}}$$

Coefficient of Variation

- The coefficient of variation
  - for *bwt* (birth weight in grams) is
    - 729.2/ 2944.6 = 0.25
  - for *bwt.lb* (birth weight in pounds) is
    - 1.6/6.5 = 0.25.
  - for *lwt* (weight in pounds) is
    - 30.6/129.8 = 0.24
- Comparing this coefficient of variation suggests that the two variables have roughly the same dispersion in terms of CV.