Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Data Exploration

Data Visualization and Summary Statistics

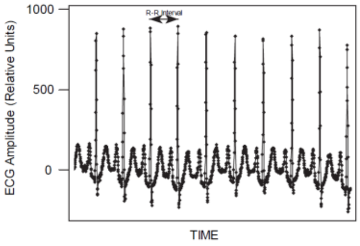
- Preliminary steps before analysis:
  - defining the scientific question we try to answer,
  - selecting a set of representative members from the population of interest
  - collecting data (either through observational studies or randomized experiments),
- Analysis usually begins with data exploration.
  - We start by focusing on data exploration techniques for one variable at a time.

Data Visualization and Summary Statistics

- Objective is
  - to develop a high-level understanding of the data,
  - learn about the possible values for each characteristic,
  - find out how a characteristic varies among individuals in our sample.
- Basicaly, we want to learn about the distribution of variables.
  - Recall that for a variable, the distribution shows
    - the possible values,
    - the chance of observing those values,
    - how often we expect to see them in a random sample from the population.

Data Visualization and Summary Statistics

- Example of an ECG recording



- R-R interval is defined as the time interval between successive R waves of the QRS complex,

Data Visualization and Summary Statistics

- A normally functioning heart exhibits considerable variability in beat-to-beat intervals.
  - variability reflects the body's continual effort to maintain homeostasis
    - so that the body may continue to perform its most essential functions and supply the body with the oxygen and nutrients required to function normally.
- It has been demonstrated through biomedical research that there is a loss of heart rate variability associated with some diseases,
  - such as diabetes and ischemic heart disease.

Data Visualization and Summary Statistics

- Researchers seek to determine
  - if this difference in variability between normal subjects and subjects with heart disease is significant
    - meaning, it is due to some underlying change in biology and not simply a result of chance
  - whether it might be used to predict the progression of the disease.
- One will note that the probability model changes as a consequence of changes in the underlying biological function or process.

Data Visualization and Summary Statistics

- To make sound decisions in the context of the uncertainty with some level of confidence,
  - we need to assume some probability models for the populations from which the samples have been collected.
- Once we have assumed an underlying model,
  - we can select the appropriate statistical tests for comparing two or more populations
  - then we can use these tests to draw conclusions about our hypotheses for which we collected the data in the first place

Data Visualization and Summary Statistics

- The data exploration methods allow us to reduce the amount of information so that we can focus on the key aspects of the data.
- We do this by using data visualization techniques and summary statistics.
- The visualization techniques and summary statistics we use for a variable depend on its type
  - Recall that we can classify them into two general groups:
    - Numerical (quantitative) variables
      - discrete, continuous
    - Categorical variables
      - nominal, ordinal

Graphical summarization of data

- Before blindly applying the statistical analysis, it is always good to look at the raw data,
  - usually in a graphical form,

  and then use graphical methods to summarize the data in an easy to interpret format.
  - A Picture is worth a thousand word
- The types of graphical displays that are most frequently used by engineers
  - scatterplots, time series, box-and-whisker plots, and histograms.

Data Visualization and Summary Statistics

- As a computational tool, R will be used.
  - RStudio is an IDE for R.
    - It is available in two formats:
      - RStudio Desktop is a regular desktop application
      - RStudio Server runs on a remote server and allows accessing RStudio using a web browser.
  - **R-Commander** can also be used
    - It allows us to do basic statistical analysis without necessarily learning the programing language of R.
- First, you must download and install R
  - Go to http://www.r-project.org/
  - Click on the download R link
  - Then select a location closest to you.
  - Click on your operating system

Data Visualization and Summary Statistics

- You can download R-Commander from the command line by following these steps:
  - Once you have installed R, open it by double-clicking on the icon.
  - A window called “R Console” will open.
  - Make sure you have a working internet connection. Then, at the prompt (the > symbol), type the following command exactly and then press enter :
    - > install.packages("Rcmdr", dependencies = TRUE)
  - R may respond by asking you to select a mirror site and listing them in a pop-up box. Choose a nearby location.
  - Depending on your connection speed, the installation may take awhile.
- If R is not already open, open it by clicking on its icon.
- To open R-Commander, at the prompt enter the following command
  - > library(Rcmdr)

What is R?

- a language and environment for statistical computing and graphics
- provides a wide variety of statistical (linear and non linear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible
- available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.
- compiles and runs on a wide variety of UNIX platforms and similar systems, Windows and MacOS.

The R environment

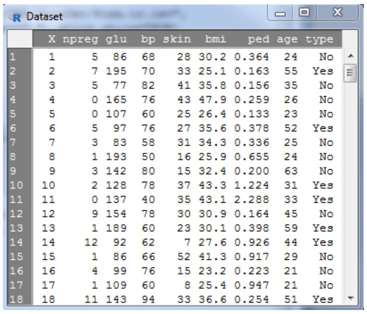
- R is an integrated suite of software facilities for data manipulation, calculation and graphical display.
- It includes
  - an effective data handling and storage facility,
  - a suite of operators for calculations on arrays, in particular matrices,
  - a large, coherent, integrated collection of intermediate tools for data analysis,
  - graphical facilities for data analysis and display either on-screen or on hardcopy,
  - a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Data Visualization and Summary Statistics

- You can download R-Commander from the command line by following these steps:
  - Once you have installed R, open it by double-clicking on the icon.
  - A window called "R Console" will open.
  - Make sure you have a working internet connection. Then, at the prompt (the > symbol), type the following command exactly and then press enter :
    - > install.packages("Rcmdr", dependencies = TRUE)
  - R may respond by asking you to select a mirror site and listing them in a pop-up box. Choose a nearby location.
  - Depending on your connection speed, the installation may take awhile.
- If R is not already open, open it by clicking on its icon.
- To open R-Commander, at the prompt enter the following command
  - > library(Rcmdr)

Data Visualization and Summary Statistics

- Example Data set : *Pima.tr*



	X	npreg	glu	bp	skin	bmi	ped	age	type
1	1	5	86	68	28	30.2	0.364	24	No
2	2	7	195	70	33	25.1	0.163	55	Yes
3	3	5	77	82	41	35.8	0.156	35	No
4	4	0	165	76	43	47.9	0.259	26	No
5	5	0	107	60	25	26.4	0.133	23	No
6	6	5	97	76	27	35.6	0.378	52	Yes
7	7	3	83	58	31	34.3	0.336	25	No
8	8	1	193	50	16	25.9	0.655	24	No
9	9	3	142	80	15	32.4	0.200	63	No
10	10	2	128	78	37	43.3	1.224	31	Yes
11	11	0	137	40	35	43.1	2.288	33	Yes
12	12	9	154	78	30	30.9	0.164	45	No
13	13	1	189	60	23	30.1	0.398	59	Yes
14	14	12	92	62	7	27.6	0.926	44	Yes
15	15	1	86	66	52	41.3	0.917	29	No
16	16	4	99	76	15	23.2	0.223	21	No
17	17	1	109	60	8	25.4	0.947	21	No
18	18	11	143	94	33	36.6	0.254	51	Yes

Data Visualization and Summary Statistics

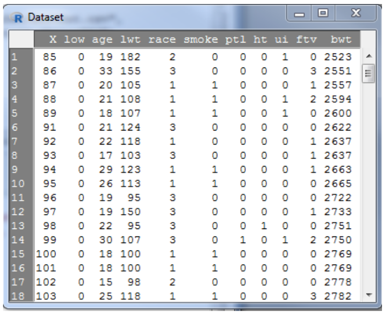
- Example Data set : *Pima.tr*
  - Pima Indians Diabetes Data Set
  - Attribute Information:
    - npreg: Number of times pregnant.
    - glu: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
    - bp: Diastolic blood pressure (mm Hg) .
    - skin: Triceps skin fold thickness (mm).
    - bmi: Body mass index (weight in kg/(height in m)^2).
    - ped: Diabetes pedigree function.
    - age: Age in years.
    - type: Class variable (disease status)
      - Yes for diabetic, No for nondiabetic

Data Visualization and Summary Statistics

- Categorical variables are either nominal or ordinal, depending on the extent of information the numerical coding provides.
- For nominal variables, the numbers are simply labels, which are chosen arbitrarily.
  - Therefore, they do not provide any information.
- The *type* variable in *Pima.tr* is nominal.
- For ordinal variables, although the numbers do not have their usual meaning, they preserve a rank ordering.
  - Therefore, they provide information about the ordering of categories.

Data Visualization and Summary Statistics

- Example Data set : *birthwt*



Dataset

	X	low	age	lwt	race	smoke	ptl	ht	ui	ftv	bwt
1	85	0	19	182	2	0	0	0	1	0	2523
2	86	0	33	155	3	0	0	0	0	3	2551
3	87	0	20	105	1	1	0	0	0	1	2557
4	88	0	21	108	1	1	0	0	1	2	2594
5	89	0	18	107	1	1	0	0	1	0	2600
6	91	0	21	124	3	0	0	0	0	0	2622
7	92	0	22	118	1	0	0	0	0	1	2637
8	93	0	17	103	3	0	0	0	0	1	2637
9	94	0	29	123	1	1	0	0	0	1	2663
10	95	0	26	113	1	1	0	0	0	0	2665
11	96	0	19	95	3	0	0	0	0	0	2722
12	97	0	19	150	3	0	0	0	0	1	2733
13	98	0	22	95	3	0	0	1	0	0	2751
14	99	0	30	107	3	0	1	0	1	2	2750
15	100	0	18	100	1	1	0	0	0	0	2769
16	101	0	18	100	1	1	0	0	0	0	2769
17	102	0	15	98	2	0	0	0	0	0	2778
18	103	0	25	118	1	1	0	0	0	3	2782

Data Visualization and Summary Statistics

- Example Data set : *birthwt*
  - the birth weight of 189 newborn babies along with some characteristics
  - Attribute Information:
    - **low:** indicator of birth weight less than 2.5 kg (0 = normal birth weight, 1 = low birth weight).
    - **age:** mother's age in years.
    - **lwt:** mother's weight in pounds at last menstrual period.
    - **race:** mother's race (1 = white, 2 = African-American, 3 = other).
    - **smoke:** smoking status during pregnancy (0 = not smoking, 1 = smoking).
    - **ptl:** number of previous premature labors.
    - **ht:** history of hypertension (0 = no, 1 = yes).
    - **ui:** presence of uterine irritability (0 = no, 1 = yes).
    - **ftv:** number of physician visits during the first trimester.
    - **bwt:** birth weight in grams.

Data Visualization and Summary Statistics

- In the *birthwt* data set, variables age, lwt, ptl, ftv, and bwt are numerical variables.
  - Among these variables, ptl and ftv are count variables.
  - The variables low, race, smoke, ht, and ui are all categorical.
- Note that all categorical variables are coded with numerical values.
- In these situations, R and R-Commander cannot automatically recognize them as categorical variables.
  - In fact, they are considered as numerical variables by default.
- Therefore, we need to convert them to categorical variables.
  - To do this, make sure *birthwt* is the active data set, then
  - click on
    - *Data→ Manage variables in active data set → Convert numeric variables to factors*

Data Visualization and Summary Statistics

- Many data set can be downloaded from the following site:

- https://vincentarelbundock.github.io/Rdatasets/datasets.html

Exploring Categorical Variables

- Consider the *type* variable in *Pima.tr* data set.
- A simple way for summarizing the data is to create a table that shows the number of times each category has been observed.
- **The number of times a specific category is observed is called frequency.**
  - We denote the frequency for category *c* by n_c.
- The sum of the frequencies for all catagories is equal to the total sample size

$$\sum_c n_c = n$$

Exploring Categorical Variables

- In R-Commander, to obtain the frequencies for the *type* variable,
  - click *Statistics → Summaries → Frequency distributions* and select *type* as the *Variable*.
- Frequency table for the *type* variable in the Pima.tr data set:

Type	Frequency
No	132
Yes	68
Total	200

For the *type* variable we have

$$\sum_c n_c = n_1 + n_2 = 132 + 68 = 200$$

Relative Frequency and Percentage

- The relative frequency is the sample proportion for each possible category.
- It is obtained by dividing the frequencies n_c by the total number of observations n:

$$p_c = \frac{n_c}{n}$$

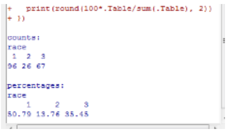
- Relative frequencies are sometimes presented as percentages after multiplying proportions p_c by 100.

Relative Frequency and Percentage

- Consider the *race* variable in the *birthwt* data set.
  - The frequencies are n_1 = 96, n_2 = 26, and n_3 = 67 for “White”, “African-American”, and “Other” categories, respectively.
  - The sum of these frequencies is equal to the sample size n = 189.
- The relative frequencies and percentages for the *race* variable in *birthwt* data set are
  - p_1 = 96/189 = 0.508 = 50.8%,
  - p_2 = 26/189 = 0.138 = 13.8%,
  - p_3 = 67/189 = 0.354 = 35.4%.
- Therefore, 50.8% of the women in the sample were white, 13.8% were African-American, and the remaining 35.4% were from other races.

Relative Frequency and Percentage

- In R-Commander, make sure *birthwt* is the active data set,
  - then click
    - *Statistics* → *Summaries* → *Frequency distributions*, and select *race* as the Variable.
- The frequencies and percentages are given in the *Output* window.
- For *race*, the category “1” (i.e., white women) has the highest frequency.



```
+    print(round(100*.Table/sum(.Table), 2))
+ })

counts:
race
 1  2  3
96 26 67

percentages:
race
    1     2     3
50.79 13.76 35.45
```

- In this case, we say that the mode of the variable *race* is “1”.
- **For a categorical variable, the mode is the most common value,**
  - **i.e., the value with the highest frequency.**

Relative Frequency and Percentage

- Since the relative frequencies are proportions of the sample size, their sum is 1,

$$\sum_c p_c = 1$$

where p_c is the relative frequency of category c.

- For the *race* variable, we have

$$\sum_c p_c = 0.508 + 0.138 + 0.354 = 1$$

- Similarly, the sum of the percentages for different categories is 100%.

Bar Graph

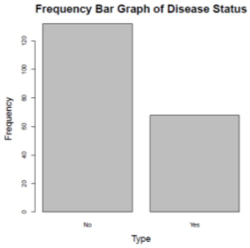
- For categorical variables, **bar graphs** are one of the simplest ways of visualizing the data.
- Using a bar graph, we can visualize the possible values (categories) a categorical variable can take,
  - as well as the number of times each category has been observed in our sample.
- The height of each bar in this graph shows the number of times the corresponding category has been observed.
  - {Create a bar graph for *type* by clicking *Graphs*→*Bar graph* and then selecting *type* as the *Variable*.}

Bar graphs and frequencies

- Frequency table for the *type* variable in the *Pima.tr* data set

Type	Frequency
No	132
Yes	68
Total	200

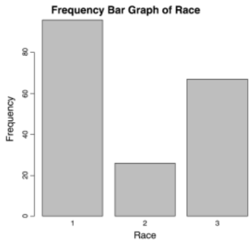
- Bar graph for the *type* variable
- Overall, bar graphs show us how the observed values of a categorical variable in our sample are distributed



Bar graphs and frequencies

- Frequency table for the *race* variable in the *birthwt* data set

Race	Frequency	Relative frequency
White	96	0.508
African-American	26	0.138
Other	67	0.354
Total	189	1



- Bar graph for mother's race in the *birthwt* data set,
- where 1, 2, and 3 represent the categories "white", "African-American", and "other", respectively

Bar Graph

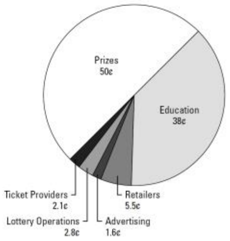
- Checklist for evaluating bar graphs:
  - Check the units on the y-axis.
    - Make sure they are evenly spaced.
  - Be aware of the scale of the bar graph (the units in which bar heights are represented).
    - Using a smaller scale you can make differences look more dramatic.
      - (for example, each half inch of height representing 10 units versus 50)
  - In the case where the bars represent percents and not counts, make sure to ask for the total number of individuals summarized by the bar graph if it is not listed.

Pie chart

- A pie chart takes categorical data and shows the percentage of individuals that fall into each category.
- The sum of all the slices of the pie should be 100% or close to it (with a bit of round-off error).
- Because a pie chart is a circle, categories can easily be compared and contrasted to one another.

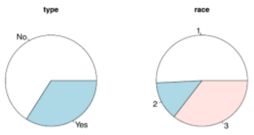
Pie chart

- An example:
  - The Florida lottery uses a pie chart to report where the money goes when a lottery ticket is purchased.



Pie chart

- We can use a pie chart to visualize the relative frequencies of different categories for a categorical variable.



- Pie charts for the *type* variable from *Pima.tr* and the *race* variable from *birthwt*, where 1, 2, and 3 represent the categories "white", "African-American", and "other", respectively

- In a pie chart, the area of a circle is divided into sectors, each representing one of the possible categories of the variable.
- The area of each sector *c* is proportional to its frequency.
- To create pie charts in R-Commander, click *Graphs*→*Pie chart*.

Pie chart

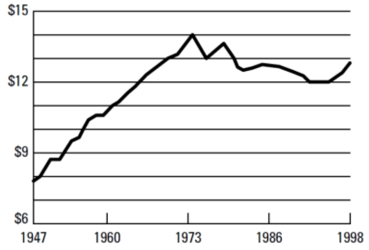
- To evaluate a pie chart for statistical correctness:
  - Check to be sure the percentages add up to 100% or close to it
    - any round-off error should be very small
  - Beware of slices of the pie called "other" that are larger than many of the other slices.
    - This shows a lack of detail in the information gathered.
  - A pie chart only shows the percentage in each group, not the number in each group.
    - Always ask for or look for a report of the total size of the data set.

Time Charts (line graphs)

- a data display whose main point is to examine trends over time.
- Typically a time chart has
  - some unit of time on the horizontal axis
    - (year, day, month, and so on)
  - a measured quantity on the vertical axis
    - (average income, birth rate, total sales, etc.).
- At each time period, the amount is shown as a dot, and the dots connect to form the time chart.

Time Charts (line graphs)

- Average hourly wage for production workers, 1947–1998



Time Charts (line graphs)

- A time chart can present information in a misleading way, for example
  - charting the number of crimes over time, rather than the crime rate (crimes per capita).
    - Because the population size of a city changes over time, crime rate is the appropriate measure.
- Make sure you understand what statistics are being presented and examine them for fairness and appropriateness.

Time Charts (line graphs)

- Checklist for evaluating time charts:
  - Examine the scale on the vertical (quantity) axis as well as the horizontal (timeline) axis;
    - results can be made to look more or less dramatic than they actually are simply by changing the scale.
  - Take into account the units used in the chart and be sure they are appropriate for comparison over time
    - (for example, are dollar amounts adjusted for inflation?).
  - Watch for gaps in the timeline on a time chart.
    - Connecting the dots across a short period of time is better than connecting across a long time.

- Further reading on visualization of categorical data
  - http://www.sciencedirect.com/science/book/9780122990458
  - http://www.datavis.ca/books/vcd/vcdstory.pdf
  - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.140.599&rep=rep1&type=pdf
  - https://www.jstatsoft.org/article/view/v053i07/v53i07.pdf
  - http://vis.berkeley.edu/courses/cs294-10-fa08/wiki/images/9/99/Seth-FinalPaper.pdf

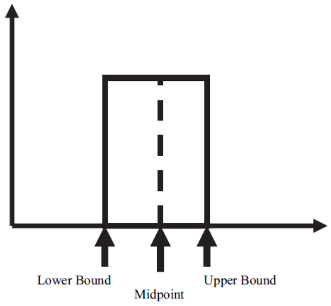
Exploring Numerical Variables

- For numerical variables, we are especially interested in two key aspects of the distribution:
  - its **location**
    - refers to the central tendency of values, that is, the point around which most values are gathered.
  - its **spread**
    - refers to the dispersion of possible values, that is, how scattered the values are around the location.

Histograms

- defined as a frequency distribution commonly used to visualize numerical variables.
- A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of nonoverlapping intervals (bins), usually of equal width.
- Given N samples or measurements, x_i ranging from X_{min} to X_{max}, the samples are binned into bins
- Typically, the number of bins is on the order of 7–14, depending on the nature of the data.
  - In addition, we typically expect to have at least three samples per bin.
    - Sturgess’rule may also be used to estimate the number of bins and is given by $k = 1 + 3.3 \log(n)$.
      - where k is the number of bins and n is the number of samples.

Histograms

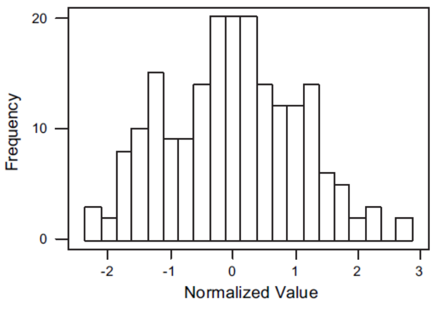


- One bin of a histogram plot
- The bin is defined by
  - a lower bound,
  - a midpoint,
  - an upper bound

Histograms

- constructed by plotting the number of samples in each bin.
  - horizontal axis,
    - the sample value,
  - the vertical axis,
    - the number of occurrences of samples falling within a bin
- Next slide illustrates a histogram for 1000 samples drawn from a normal distribution with mean (μ) = 0 and standard deviation (σ) = 1.0.

Histograms

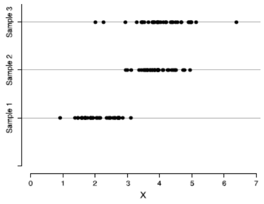


Histograms

- Two useful measures in describing a histogram:
  - the absolute frequency in one or more bins
    - f_i = absolute frequency in ith bin
  - the relative frequency in one or more bins
    - f_i /n = relative frequency in ith bin,
      - where n is the total number of samples being summarized in the histogram
- The histogram can exhibit several shapes
  - symmetric, skewed, or bimodal.

Histograms - example

- As a running example, consider a numerical variable, X, for which three sets (samples) of observations denoted as Sample 1, Sample 2, and Sample 3 have been collected.
- Dot plots for these three sets of observations:



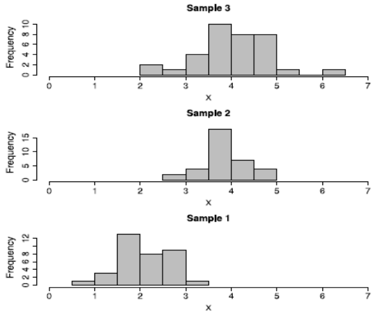
Observations in *Sample* 1 are gathered around 2,

Observations in *Sample* 2 and *Sample* 3 are gathered around 4.

Observations in *Sample* 3 are more dispersed compared to those in *Sample* 1 and *Sample* 2

Histograms - example

- Histograms of the three samples.



Histograms - example

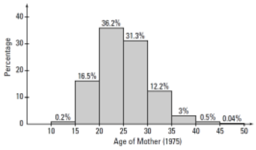
- The following table shows the number of live births in Colorado by age of mother for selected years from 1975–2000.

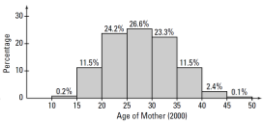
Year	*Total births*	*10–14*	*15–19*	*20–24*	*25–29*	*30–34*	*35–39*	*40–44*	*45–49*
1975	40,148	88	6,627	14,533	12,565	4,885	1,211	222	16
1980	49,716	57	6,530	16,642	16,081	8,349	1,842	198	12
1985	55,115	90	5,634	16,242	18,065	11,231	3,464	370	13
1990	53,491	91	5,975	13,118	16,352	12,444	4,772	717	15
1995	54,310	134	6,462	12,935	14,286	13,186	6,184	1,071	38
2000	65,429	117	7,546	15,865	17,408	15,275	7,546	1,545	93

- The numerical variable age is broken down into categories of 5-year groupings.

Histograms - example

- Relative frequency histograms comparing 1975 and 2000





- You can see more older mothers in 2000 than in 1975.

Histograms

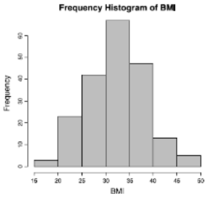
- The bar height for each interval could be set to its relative frequency $p_c = n_c/n$, or the percentage p_c *x100*, of observations that fall into that interval.
- For histograms, however, it is more common to use the density instead of the relative frequency or percentage.
  - **The density is the relative frequency for a unit interval.**
    - **It is obtained by dividing the relative frequency by the interval width:**

$$f_c = p_c/w_c$$

  - Here, $p_c = n_c/n$ is the relative frequency with n_c as the frequency of interval c and n as the total sample size.
  - The width of interval c is denoted w_c .
- To create the *density histogram* in R-Commander, click *Graphs* → *Histogram*, select a variable, and choose *Densities* for the *Axis Scaling*.

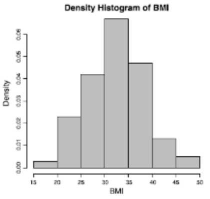
Histograms

- The frequency histogram for the numerical variable *bmi* in the *Pima.tr* data set.



- The height of the rectangles represent the frequency of the interval and sum to the total sample size *n*.
- Here, the values of the variable are divided into seven bins

- The density histogram for *bmi* from the *Pima.tr* data set.



- Here, the scale on the y-axis is density (not frequency).
- Once again, the values of *bmi* are divided into seven bins of width $w = 5$

Histograms

- Assuming the frequency histogram for variable *bmi*, let us calculate the density of the interval [30, 35], which is the 4th interval.
  - There are $n_4 = 67$ observations in this interval.
  - Therefore, the relative frequency is $p_4 = 67/200 = 0.335$.
  - The interval width is $w_4 = 5$.
  - The density for this interval is therefore $f_4 = 0.335/5 = 0.067$
- To create the density histogram for *bmi* in R-Commander, click
  - *Graphs* → *Histogram*, select *bmi* as the Variable, and choose *Densities for the Axis Scaling*

Histograms

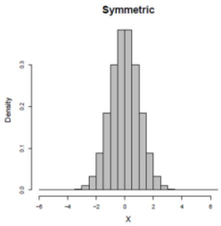
- The height of each bar in density histogram shows the density of the corresponding interval (as opposed to its frequency).
- For each interval *c*, the area of the corresponding bar in the density histogram is calculated as follows (hight×width):

$$a_c = f_c \times w_c = (p_c / w_c) \times w_c = p_c$$

- Therefore, the area of each bar (rectangle) is the relative frequency for the corresponding interval.
  - Since the sum of relative frequencies is 1, the total area of bars in a density histogram is 1.

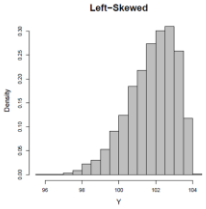
Histograms

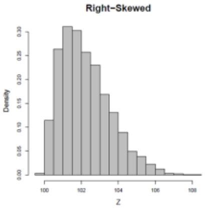
- When creating a histogram, it is important to choose an appropriate value for *w* (*Number of Bins*) .
- Besides the location and spread of a distribution, the shape of a histogram also shows us how the observed values spread around the location.
- We say the following histogram is symmetric around its location (here, zero) since the densities are the [almost] same for any two intervals that are equally distant from the center.



Histograms

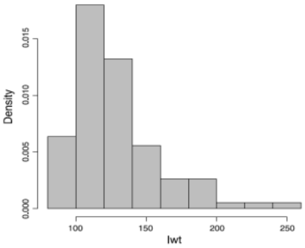
- In many situations, we find that a histogram is stretched to the left or right.
- We call such histograms skewed.
  - More specifically, we call them left-skewed if they are stretched to the left, or right-skewed if they are stretched to the right.





Histograms

- As an example, histogram of variable *lwt* in the *birthwt* data set is right-skewed

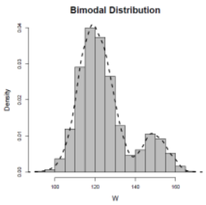


Histograms

- The histograms in previous slides, whether symmetric or skewed, have one thing in common
  - they all have one peak (or mode).
- We call such histograms (and their corresponding distributions) unimodal.
- Sometimes histograms have multiple modes.
  - The bimodal histogram appears to be a combination of two unimodal histograms.
    - Indeed, in many situations bimodal histograms (and multimodal histograms in general) indicate that the underlying population is not homogeneous and may include two (or more in case of multimodal histograms) subpopulations.

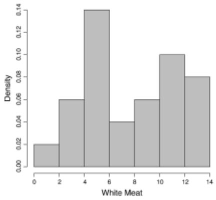
Histograms

- Histogram of a bimodal distribution



- A *smooth curve* is superimposed so that the two peaks are more evident

- Histogram of protein consumption in 25 European countries for white meat in *Protein* data set.



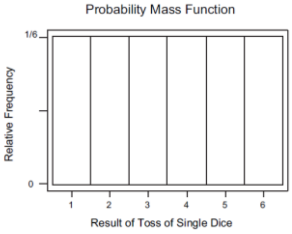
- The histogram is bimodal, which indicates that the sample might be comprised of two subgroups

Histograms

- The histogram is important because it serves as
  - a rough estimate of the true probability density function or
  - probability distribution of the underlying random process from which the samples are being collected.
- The probability density function or probability distribution is a function that quantifies the probability of a random event, *x*, occurring.
  - When the underlying random event is discrete in nature, we refer to the probability density function as the probability mass function

Histograms

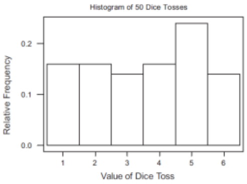
- The probability density function for a discrete random variable (probability mass function).

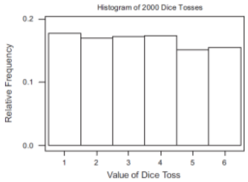


- In this case, the random variable is the value of a toss of a single dice.
  - Note that each of the six possible outcomes has a probability of occurrence of 1 of 6.
- This probability density function is also known as a uniform probability distribution.

Histograms

- Histograms representing the outcomes of experiments in which a single dice is tossed 50 and 2000 times, respectively





  - Note that as the sample size increases, the histogram approaches the true probability distribution (uniform probability distribution)

Histograms

- Engineers are trying to make decisions about populations or processes to which they have limited access.
- Thus, they design experiments and collect samples that they think will fairly represent the underlying population or process.
- Regardless of what type of statistical analysis will result from the investigation or study, all statistical analysis should follow the same general approach:
  - Measure a limited number of representative samples from a larger population.
  - Estimate the true statistics of larger population from the sample statistics.

Histograms

- Once the researcher has estimated the sample statistics from the sample population,
  - he or she will try to draw conclusions about the larger (true) population.
- The most important question to ask when reviewing the statistics and conclusions drawn from the sample population is
  - how well the sample population represents the larger, underlying population.

Histograms

- Checklist for evaluating a histogram:
  - Examine the scale used for the vertical axis and beware of results that appear exaggerated or played down through the use of inappropriate scales.
  - Check out the units on the vertical axis to see whether the histogram reports frequencies (numbers) or relative frequencies (percentages), and then take this into account when evaluating the information.
  - Look at the scale used for the groupings of the numerical variable (on the horizontal axis).
    - If the range for each group is very small, the data may look overly volatile.
    - If the ranges are very large, the data may appear to be smoother than they really are.

Measures of Central Tendency

- Histograms are useful for visualizing numerical data and identifying their location and spread.
- However, we typically use descriptive or summary statistics for more precise specification of the
  - central tendency
  - dispersion

  of observed values.

Measures of Central Tendency

- A central tendency is a central or typical value for a probability distribution.
  - also called a center or location of the distribution.
- Measures of central tendency are often called averages.
- There are several measures that reflect the central tendency
  - sample mean,
  - sample median,
  - sample mode.

Mean

- In mathematics, mean has several different definitions depending on contex.
- In probability and statistics
  - mean and expected value are synonymous
- In case of a discrete probability distribution of random variable *x*,
  - the mean is equal to the sum over every possible value weighted by the probability of that value

$$\mu = \sum xP(x)$$

Mean

- For a data set, the terms
  - arithmetic mean,
  - mathematical expectation,
  - sometimes average

  are used synonymously to refer to a central value of a discrete set of numbers
  - specifically, the sum of the values divided by the number of values.
- If the data set were based on a series of observations obtained by sampling from a statistical population,
  - the arithmetic mean is termed as the sample mean to distinguish it from the population mean

Mean

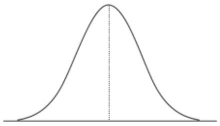
- Outside of probability and statistics, a wide range of other notions of mean are often used in geometry and analysis:
  - Pythagorean means
    - Arithmetic mean, Geometric mean, Harmonic mean
  - Generalized means
    - Power mean,
      - a.k.a generalized mean, Hölder mean, mean of degree (or order or power) *p*
    - *f*-mean
  - Weighted arithmetic mean
  - Truncated mean
  - Interquartile mean
  - Fréchet mean
  - …

Mean

- Arithmetic mean (or simply mean) of a sample x_1, x_2…, x_n, usually denoted by $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- It is used when the spread of the data is fairly similar on each side of the mid point
  - when the data are "normally distributed".
    - If a value is a lot smaller or larger than the others, "skewing" the data, the mean will then not give a good picture of the typical value.



Mean

- Geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product, e.g. rates of growth

$$\bar{x} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_\mathrm{n}}$$

- Harmonic mean is an average which is useful for sets of numbers that are defined in relation to some unit, for example speed

$$\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_\mathrm{n}}}$$

Mean

- The relationship between Arithmetic mean, Geometric mean, and Harmonic mean:
  Arithmetic mean × Harmonic mean = Geometric mean^2
- Arithmetic mean, Geometric mean, and Harmonic mean satisfy the following inequalities:
  Arithmetic mean ≥ Geometric mean ≥ Harmonic mean
  - Equality holds if and only if all the elements of the given sample are equal
- The arithmetic mean is best used in situations where:
  - the data are not skewed (no extreme outliers)
  - the individual data points are not dependent on each other
- The geometric mean should be used whenever the data are inter-related
- The harmonic mean is best to use when there is:
  - A large population where the majority of the values are distributed uniformly but where there are a few outliers with significantly higher values

Mean

- Weighted arithmetic mean is used if one wants to combine average values from samples of the same population with different sample sizes

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

  - The weights w_i represent the sizes of the different samples.
  - In other applications, they represent a measure for the reliability of the influence upon the mean by respective values.

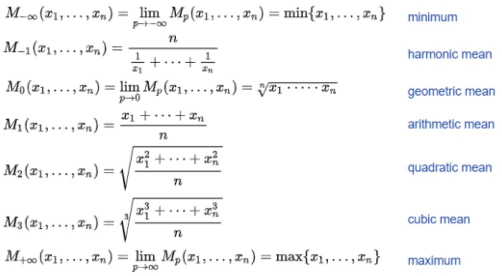
Mean

- A power mean is a mean of the form

$$M_p = \left(\frac{1}{n} \sum_{k=1}^{n} x_k^p \right)^{1/p}$$

$M_{-\infty}$	minimum
M_{-1}	harmonic mean
M_0	geometric mean
M_1	arithmetic mean
M_2	root-mean-square
M_∞	maximum

Mean



$$M_{-\infty}(x_1,\ldots,x_n) = \lim_{p\to-\infty} M_p(x_1,\ldots,x_n) = \min\{x_1,\ldots,x_n\}$$ minimum

$$M_{-1}(x_1,\ldots,x_n) = \frac{n}{\frac{1}{x_1}+\cdots+\frac{1}{x_n}}$$ harmonic mean

$$M_0(x_1,\ldots,x_n) = \lim_{p\to0} M_p(x_1,\ldots,x_n) = \sqrt[n]{x_1\cdot\cdots\cdot x_n}$$ geometric mean

$$M_1(x_1,\ldots,x_n) = \frac{x_1+\cdots+x_n}{n}$$ arithmetic mean

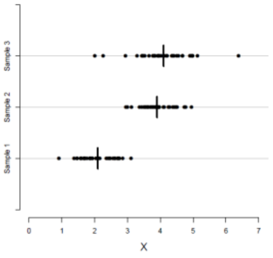
$$M_2(x_1,\ldots,x_n) = \sqrt{\frac{x_1^2+\cdots+x_n^2}{n}}$$ quadratic mean

$$M_3(x_1,\ldots,x_n) = \sqrt[3]{\frac{x_1^3+\cdots+x_n^3}{n}}$$ cubic mean

$$M_{+\infty}(x_1,\ldots,x_n) = \lim_{p\to\infty} M_p(x_1,\ldots,x_n) = \max\{x_1,\ldots,x_n\}$$ maximum

Sample Mean

- Plotting the three samples along with their means (*short vertical lines*)
- For Sample 1, Sample 2, and Sample 3, the means are 2.1, 3.9, and 4.1, respectively.



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.

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

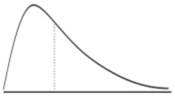
- In this case, the sample mean is 79.8, which seems to be a good representative of the data.
- Now suppose that the heart rate for the first individual is recorded as 47 instead of 74.

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

- 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 61;
- The sample medians (denoted $\tilde{x}$) are

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

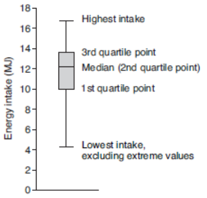
- So, the median is more robust against outliers.

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.

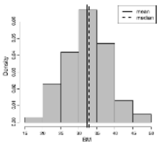
Median (example)

- 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.

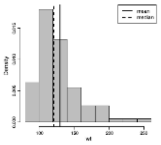


- Box and whisker plot of energy intake of 50 patients over 24 hours.
- The ends of the whiskers represent the maximum and minimum values, excluding extreme results like those of the two "nil by mouth" patients.

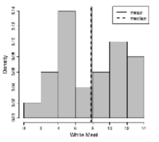
Sample Mean and Median



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.

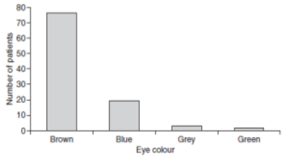
The mean and median are nearly equal since the histogram is Symmetric.

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

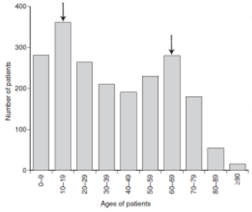
- the most common of a set of events
  - used when we need a label for the most frequently occurring event
    - Example: An eye clinic sister noted the eye colour of 100 consecutive patients. The results are shown below



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

Mode

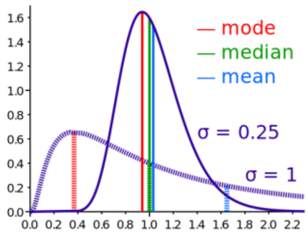
- You may see reference to a bi-modal distribution.
  - Generally when this is mentioned in papers it is as a concept rather than from calculating the actual values,
    - e.g. “The data appear to follow a bi-modal distribution”.



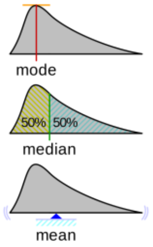
- Graph of ages of patients with asthma in a practice
  - The arrows point to the modes at ages 10–19 and 60–69.
- Bi-modal data may suggest that two populations are present that are mixed together,
  - so an average is not a suitable measure for the distribution.

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)

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

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

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)

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

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

Complementary procedure: moving DIFFERENCE filter

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

$Dy_t = y_t - y_{t-1}$

- Higher order differences (2nd order)

$D^2y_t = D(Dy_t) = Dy_t - Dy_{t-1} = y_t - 2y_{t-1} + y_{t-2}$

- Example (1st order moving DIFFERENCE filter)

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

$y=([x_2-x_1], [x_3-x_2], [x_4-x_3], \dots)$

Moving median filtering

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

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

y=(med[x_1,x_2,x_3], med[x_2,x_3,x_4], med[x_3,x_4,x_5], …)

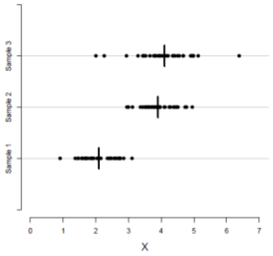
- If a window with even number of samples are selected median is average of two mid-point samples

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.

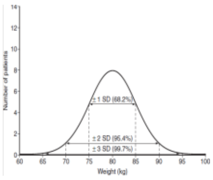
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

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

$s^2 = 6/4 = 1.5$
$s = \sqrt{1.5} = 1.22$

$s^2 = 366/4 = 91.5$
$s = \sqrt{91.5} = 9.56$

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 90th percentile for these scores
  - multiply 90% by the total number of scores,
    - 90% × 25 = 0.90 × 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 23rd number in the data set.
- That number is 98,
  - which is the 90th percentile for this data set.

- To find the 20th percentile,
  - take 0.20 ∗ 25 = 5;
    - this is a whole number so proceed to Step 4, which tells us the 20th percentile is the average of the 5th and 6th numbers in the ordered data set (62 and 66).
  - 20th percentile then becomes (66 + 62) / 2 = 64
- The median is the 50th 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 13th 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 80th 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

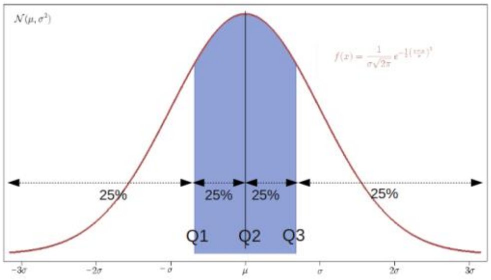
  43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99
- Q1 (25th percentile)

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

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

  0.75 * 25 = 18.75 ➔ (round up) ➔ 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 (13th sample).
- The 1st 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 3rd 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 (25th percentile)

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

  0.50 * 25 = 12.5 ➔ (round up) ➔ 13 Q2 = 77
- Q1 (75th 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:

    ith 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 (25th percentile)

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

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

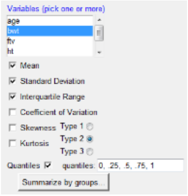
Measures of Variability – example 2

- Alternative calculation;
  – Use the following formula to estimate the ith observation:
  ith 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 (25th percentile)
  0.25 * (10 +1) = 2.75 ➔ (round down) ➔ 2   Q1 = 1
- Q2 (50th percentile)
  0.50 * (10 +1) = 5.5 ➔ (round down) ➔ 5   Q2 = 2
- Q1 (75th 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:



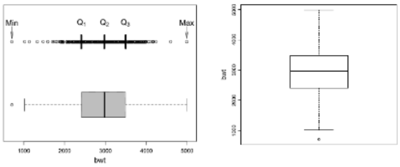
Variables (pick one or more)
age
bwt
ftv
ht

☑ Mean
☑ Standard Deviation
☑ Interquartile Range
□ Coefficient of Variation
□ Skewness
□ Kurtosis
Type 1 ○
Type 2 ●
Type 3 ○
Quantiles ☑ quantiles: 0, .25, .5, .75, 1

```
> 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
```

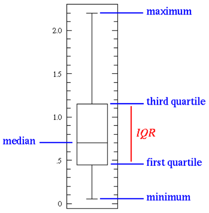
Boxplot

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



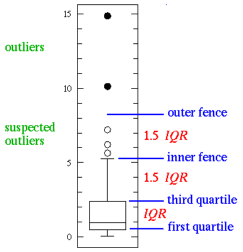
- Very often, boxplots are drawn vertically.
- To create a boxplot for *bwt* in R-Commander,
  – make sure *birthwt* is the active dataset,
  – click *Graphs → Boxplot*, and select *bwt*.

Boxplot



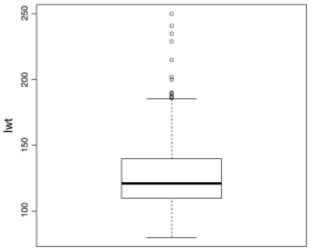
- This simplest possible box plot displays the full range of variation (from min to max), the likely range of variation (the IQR), and a typical value (the median).
- Not uncommonly real datasets will display surprisingly high maximums or surprisingly low minimums called outliers.
- John Tukey has provided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

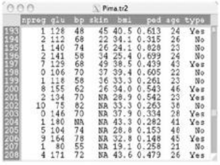


- Vertical boxplot of *lwt*.
- This plot reveals that the variable *lwt* is right-skewed and there are several possible outliers,
  - whose values are beyond the whisker on the top of the box

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



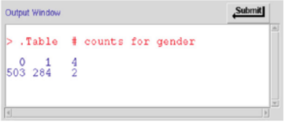
```
    npreg glu  bp skin  bmi   ped age type
193     1 128  48   45 40.5 0.613  24  Yes
194     2 112  68   22 34.1 0.315  26   No
195     1 140  74   26 24.1 0.828  23   No
196     2 141  58   34 25.4 0.699  24   No
197     7 129  68   49 38.5 0.439  43  Yes
198     0 106  70   37 39.4 0.605  22   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.



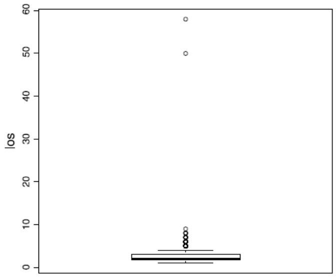
```
> .Table  # counts for gender

  0   1   4
503 284   2
```

- 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.

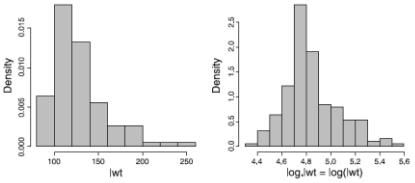


The boxplot of *los* with two extremely large values

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,
  - 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*

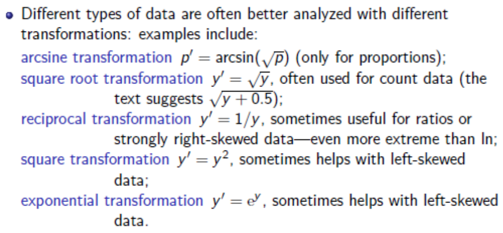
Data Transformation

- 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;

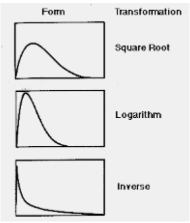
  square transformation $y' = y^2$, sometimes helps with left-skewed data;

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

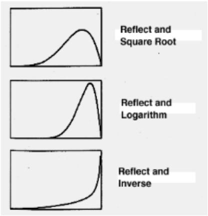
Data Transformation

- 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)

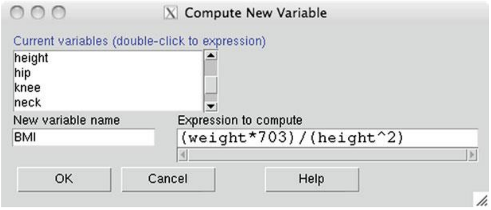
- C = a constant added to each score so that the smallest score is 1.
- K = a constant from which each score is subtracted

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


Compute New Variable

Current variables (double-click to expression)

height
hip
knee
neck

New variable name: BMI

Expression to compute: (weight*703)/(height^2)

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 Catagories 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.

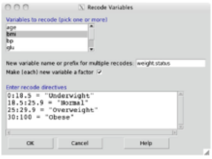
BMI	Weight Status
Below 18.5	Underweight
18.5–24.9	Normal
25.0–29.9	Overweight
30.0 and Above	Obese

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

Creating Catagories 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.*

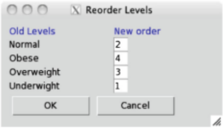


New variable name or prefix for multiple recodes: weight.status

Make (each) new variable a factor ☑

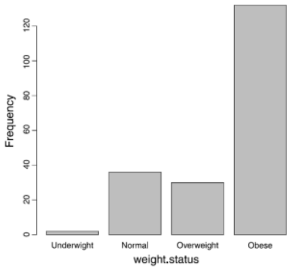
Enter recode directives

```
0:18.5 = "Underwight"
18.5:25.9 = "Normal"
25:29.9 = "Overweight"
30:100 = "Obese"
```



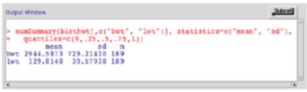
Old Levels	New order
Normal	2
Obese	4
Overweight	3
Underweight	1

Creating Catagories for Numerical Variables



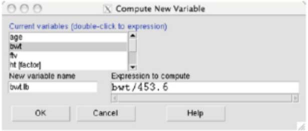
- The bar graph for *bmi* after converting the numerical variable to a categorical variable

Creating Catagories for Numerical Variables



```
> numSummary(birthwt[,c("bwt", "lwt")], statistics=c("mean", "sd"),
+    quantiles=c(0,.25,.5,.75,1))
         mean        sd   n
bwt 2944.5873 729.21430 189
lwt  129.8148  30.57938 189
```

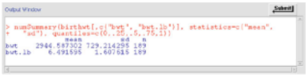
- Summary statistics for *bwt* and *lwt* from the birthwt data set



New variable name: bwt.lb

Expression to compute: bwt/453.6

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



```
> numSummary(birthwt[,c("bwt", "bwt.lb")], statistics=c("mean",
+    "sd"), quantiles=c(0,.25,.5,.75,1))
             mean         sd   n
bwt    2944.587302 729.214295 189
bwt.lb    6.491595   1.607615 189
```

- 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.

Scaling and Shifting Variables

- In general, when we multiply the observed values of a variable by a constant *a*, its mean, standard deviation, and variance are multiplied by *a*, |*a*|, and a^2, respectively.
  - That is, if $y = ax$, then
    - $\bar{y}=a\bar{x}$, $s_y=|a|s_x$, $s_y^2=a^2s_x^2$
- The coefficient of variation is not affected.

$$CV_y = \frac{s_y}{\bar{y}} = \frac{a s_x}{a\bar{x}} = \frac{s_x}{\bar{x}} = CV_x$$

Scaling and Shifting Variables

- If we shift the observed values by *b*, i.e., $y = x + b$, then

  $\bar{y}=\bar{x}+b$, $s_y=s_x$, $s_y^2=s_x^2$

- If we multiply the observed values by the constant *a* and then add the constant *b* to the result, i.e., $y = ax + b$, then

  $\bar{y}=a\bar{x}+b$, $s_y=|a|s_x$, $s_y^2=a^2s_x^2$

- the coefficient of variation will change. If $y = ax + b$ (assuming $a > 0$ and $b = 0$), then

$$CV_y = \frac{s_y}{\bar{y}} = \frac{a s_x}{a\bar{x}+b} \neq \frac{s_x}{\bar{x}}.$$

Variable Standardization

- Variable standardization is a common *linear* transformation,
  - where we subtract the sample mean $\bar{x}$ from the observed values and divide the result by the sample standard deviation *s*,
    - in order to shift the mean to zero and make the standard deviation 1:

$$y_i = \frac{x_i - \bar{x}}{s}.$$

- Using such transformation is especially common in regression analysis and clustering.
- Subtracting $\bar{x}$ from the observations shifts the sample mean to zero.
  - This, however, does not change the standard deviation.
- Dividing by *s*, on the other hand, changes the sample standard deviation to 1

Data Exploration with R Programming

- Load *Pima.tr* data set, which is available from MASS package

  ```
  > library(MASS)
  > data(Pima.tr)
  ```

  > data("Pima.tr") is also valid
- The *head()* function shows only the first part of the data set.

  ```
  > head(Pima.tr)
  ```
- Use the *help()* function to view description on the data available in the package

  ```
  > help(Pima.tr)
  ```
- Use *table()* function to obtain the frequencies for the catagorical variable

  ```
  > type.freq <- table(Pima.tr$type)
  > type.freq
      No   Yes
      132  68
  ```

  Note that the $ symbol is being used to access the type variable in the Pima.tr data set.

Data Exploration with R Programming

- Now, use the *type.freq* table to create the bar graph.

  ```
  > barplot(type.freq, xlab = "Type", ylab = "Frequency",  main = "Frequency Bar Graph of Type")
  ```

  The first parameter to the *barplot()* function is the frequency table. The options *xlab* and *ylab* label the *x* and *n* axes, respectively. Likewise, the *main* option puts a title on the plot.
- The relative frequency can be calculated as

  ```
  > n <- sum(type.freq)
  > type.rel.freq <- type.freq/n
  > round(type.rel.freq, 2)
  > round(type.rel.freq, 2) * 100
  ```

Data Exploration with R Programming

- **If the levels of a categorical variable in the data set is coded as numbers, we need to convert the type of variable to *factor* using the *factor()* function, so that R recognizes it as categorical.**
- You can use the function *is.factor()* to examine whether a variable is a factor.

```
> data(birthwt)
> is.factor(birthwt$smoke)
      [1] FALSE
> birthwt$smoke <- factor(birthwt$smoke)
> is.factor(birthwt$smoke)
      [1] TRUE
> table(birthwt$smoke)
       0     1
       115   74
```

Data Exploration with R Programming

- To create a *frequency* histogram for age, use the *hist()* function with the freq option set to “TRUE” (which is the default):

```
> hist(Pima.tr$age, freq = TRUE, xlab = "Age", ylab =
"Frequency", col = "grey", main = "Frequency Histogram of
Age")
```

- Then create a *density* histogram of age by setting the freq option to “FALSE”:

```
> hist(Pima.tr$age, freq = FALSE,xlab = "Age", ylab =
"Density", col = "grey", main = "Density Histogram of Age")
```

Data Exploration with R Programming

- We can obtain the mean and median of numerical data with the mean() and median() functions.
- Find these statistics for numerical variables in Pima.tr:

```
> mean(Pima.tr$npreg)
      [1] 3.57
> median(Pima.tr$bmi)
      [1] 32.8
```

- The quantile() function with the probs option returns the specified quantiles:

```
> quantile(Pima.tr$bmi, probs = c(0.1, 0.25, 0.5, 0.9))
      10%     25%     50%     90%
      24.200  27.575  32.800  39.400
```

- The five-number summary along with the mean can simply be obtained with the summary() function:

```
> summary(Pima.tr$bmi)
      Min.    1st Qu.  Median  Mean   3rd Qu.  Max.
      18.20   27.58    32.80   32.31  36.50    47.90
```

Data Exploration with R Programming

- We can present the five-number summary visually with a boxplot:

```
> boxplot(Pima.tr$bmi, ylab = "BMI")
```

- While the default is to create vertical boxplots, we can also create horizontal boxplots by specifying the horizontal option to true:

```
> boxplot(Pima.tr$bmi, ylab = "BMI", horizontal = TRUE)
```

- Find the interquartile range (IQR) with the IQR() function:

```
> IQR(Pima.tr$bmi)
      [1] 8.925
```

- The smallest and largest observations can be obtained with the range() function
  - the functions min() and max() could also be applied):

```
> minMax <- range(Pima.tr$bmi)
> minMax
      [1] 18.2 47.9
```

Data Exploration with R Programming

- The variance and standard deviation are also easily calculated with var() and sd():

```
> var(Pima.tr$bmi)
      [1] 37.5795
> sd(Pima.tr$bmi)
      [1] 6.130212
```

Data Exploration with R Programming

- Creating Categories for Numerical Variables:
  - To create a categorical variable weight.status based on the *bmi* variable in *Pima.tr*, we can go through each observation one by one and assign each observation to one of the four categories:
    - “Underweight”,
    - “Normal”,
    - “Overweight”,
    - “Obese”.
  - To do this, we can use loops and conditional statements
  - First, we start by creating an empty vector of size 200 within the *Pima.tr* data frame:

```
> Pima.tr$weight.status <- rep(NA, 200)
```

Data Exploration with R Programming

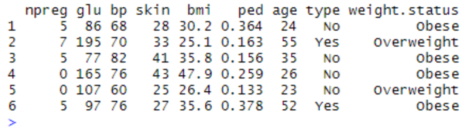
- Next, we set the values of weight.status for all observations by using ifelse() statements within a for() loop:

```
> for (i in 1:200) {
        if (Pima.tr$bmi[i] < 18.5) {
                Pima.tr$weight.status[i] <- "Underweight"
        }
        else if (Pima.tr$bmi[i] >= 18.5 &
                Pima.tr$bmi[i] < 24.9) {
                Pima.tr$weight.status[i] <- "Normal"
         }
        else if (Pima.tr$bmi[i] >= 24.9 &
                Pima.tr$bmi[i] < 29.9) {
                Pima.tr$weight.status[i] <- "Overweight"
        }
        else {
                Pima.tr$weight.status[i] <- "Obese"
        }
  }
```

Data Exploration with R Programming

- Here, the loop counter goes from 1 to 200.
- Use the head() function to view the result:

```
> head(Pima.tr)
```



```
  npreg glu bp skin  bmi   ped age type weight.status
1     5  86 68   28 30.2 0.364  24   No         Obese
2     7 195 70   33 25.1 0.163  55  Yes    Overweight
3     5  77 82   41 35.8 0.156  35   No         Obese
4     0 165 76   43 47.9 0.259  26   No         Obese
5     0 107 60   25 26.4 0.133  23   No    Overweight
6     5  97 76   27 35.6 0.378  52  Yes         Obese
>
```

Data Exploration with R Programming

- Before using the newly created variable weight.status in statistical analysis, its type should be converted to factor.

  > Pima.tr$weight.status <- factor(Pima.tr$weight.status)
- While the above code makes weight.status a factor variable, it does not take into account the ordering of levels.
- The levels are ordered alphabetically and can be examined using the levels() function:

```
> levels(Pima.tr$weight.status)
[1] "Normal"          "Obese"
[3] "Overweight"      "Underweight"
```

Data Exploration with R Programming

- The right ordering can be provided when the factor() function is used to convert the variable:

```
> Pima.tr$weight.status <-
factor(Pima.tr$weight.status,  levels =
c("Underweight", "Normal",  "Overweight",
"Obese"))
> levels(Pima.tr$weight.status)
[1] "Underweight" "Normal"
[3] "Overweight" "Obese"
```

Handling Missing Data in R

- To find missing values of a variable,
  - the is.na() function can be used
    - It returns “TRUE” when the value is missing and “FALSE” otherwise,
  - Consider the Pima.tr2 data set from the MASS library
    - the Pima.tr data set is obtained from Pima.tr2 by removing observations with missing values):

```
> data(Pima.tr2)
> is.na(Pima.tr2$bp)
```

- To obtain the indices of observations whose values are missing, we can use the which() function along with the is.na() function.

  > which(is.na(Pima.tr2$bp))

Handling Missing Data in R

- The complete.cases() function returns a logical vector indicating which cases (observations) in the data set are complete

  > complete.cases(Pima.tr2)
- To remove cases with missing values, the na.omit() function can be used:

  > Pima.complete <- na.omit(Pima.tr2)
  - Here, the newly created Pima.complete data set includes only the complete cases from Pima.tr2.