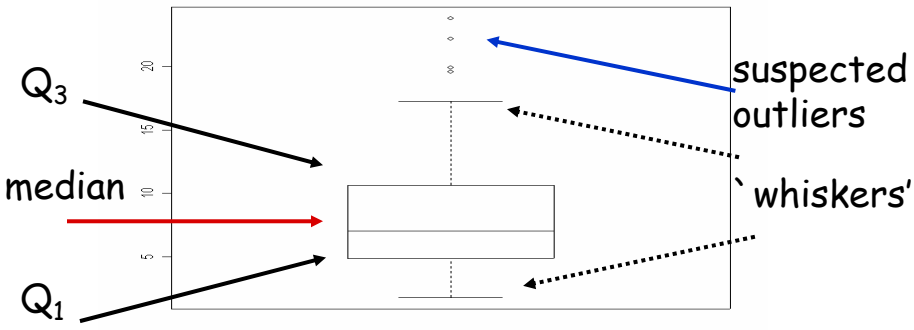
Statistics and Probability

Graphical and Numerical Data Summaries



http://www.isrec.isb-sib.ch/~darlene/geneve/

Outline

- Variables
- R statistical software
- Graphical data summaries
- Numerical data summaries
- Boxplot
- R session management

Variables (I)

- Statisticians call characteristics which can differ across individuals *variables*
- Types of variables
  - *Categorical* (also called *qualitative*)
    - Examples: eye color, favorite television program
  - *Numerical* (also called *quantitative*)
    - Examples: height, number of children, fluorescence intensity

Variables (II)

- Categorical variables may be
  - *Nominal* – the categories have names, but no ordering (e.g. eye color)
  - *Ordinal* – categories have an ordering (e.g. `Always', `Sometimes', 'Never')
- Numerical variables may be
  - *Discrete* – possible values can differ only by fixed amounts (most commonly counting values)
  - *Continuous* – can take on any value within a range (e.g. any positive value)

Sample surveys

- *Surveys* are carried out with the aim of learning about characteristics (or *parameters*) of a *target population*, the group of interest
- The survey may select *all* population members (*census*) or only *a part* of the population (*sample*)
- Typically studies sample individuals (rather than obtain a census) because of time, cost, and other practical constraints

Sampling variability

- Sample from a pop. in order to estimate the (pop.) mean of some (numerical) variable of interest (*e.g.* weight, height, number of children, *etc.*)
- We would use the *sample mean* as our guess for the unknown value of the population mean
- Our sample mean is very unlikely to be exactly equal to the (unknown) population mean - *chance variation* in sampling
- => Useful to quantify the *likely size* of this chance variation

The research process

- Scientific question of interest
- Decision on what data to collect (and how)
- Collection and analysis of data
- Conclusions, generalization
- Communication and dissemination of results

Generic Question: Does a 'treatment' have an 'effect'?

Examples:

- Does smoking cause cancer, heart disease, etc?
- Does oat bran lower cholesterol?
- Does echinacea prevent illness?
- Does exercise slow the aging process?

Addressing the question

- A basic means to address this type of question involves comparing two groups of *study subjects*
  - *Control* group: provides a baseline for comparison
  - *Treatment* group: group receiving the 'treatment'

Types of studies

- *Controlled experiment*: subjects assigned to groups by the investigator
  - randomization: protects against bias in assignment to groups
  - blind, double-blind: protects against bias in outcome assessment/measurement
  - placebo: fake 'treatment'
- *Observational study*: subjects 'assign' themselves to groups
  - confounder: associated with both group membership and the outcome of interest

A few comments

- With a well-planned and carried out controlled experiment, it is possible to infer *causality*
- This is *not* possible with observational studies due to the presence of confounders
- With confounding, it is not possible to tell whether the observed difference between groups is due to the treatment or to the confounder
- Not always possible to carry out an experiment, due to *practical* and *ethical* reasons

Exploratory data analysis

- Also called *descriptive statistics*, this term is used to describe the process of 'looking at the data' prior to formal analysis
- In this phase of analysis, data are examined for quality and 'cleaned' as well as displayed to provide an overall impression of results
- We will look at two types of summaries:
  - Graphical summaries
  - Numerical summaries
- Necessary to use *statistical software*

Why R?

- Powerful, flexible, and extensible statistical computing language and environment
- Wide range of built-in statistical functions and add-on packages available, including a growing number specifically for biological data analysis
- High quality, customizable graphics capabilities
- Available for Unix/Linux, Windows, Mac
- All this and … R is free!

R: finding help

- Comprehensive documentation is available at the CRAN (Comprehensive R Archive Network) web site, located at

  http://cran.r-project.org/

- Many books for users of S and S+ (Splus)
- *Introductory Statistics with R*, by Peter Dalgaard, a member of the R core team
- **`help, help.search`** facilities within R

R as a giant calculator

- Arithmetic operations: `+, -, *, /, ^`

```
> 2+2
[1] 4
> pi
[1] 3.141593
> 27*4.2
[1] 113.4
```

```
> 15 - 12
[1] 3
> 5^2
[1] 25
> 400/57
[1] 7.017544
```


R: storing results

- Assignment operator: `<-` (two symbols)

```
> radius <- 12
> pi * radius^2
[1] 452.3893
```

R: assignment caveats

- Variable names can consist of letters, digits and periods (dot), but have a few limitations:
  - The name must NOT start with a digit
  - The name must NOT start with a period
- Names are case-sensitive: *WEIGHT*, *Weight*, and *weight* all refer to different variables
- Names which are already used by the system should be avoided (easier to do when you know more about the system, but it's best to avoid single letter names)

R: creating data

- R has a number of functions to create data vectors, including:

```
> c( )
> seq( )
> rep( )
```

- Example:

```
> weight <- c(60,72,57,90,95,72)
> height <- c(1.75, 1.80, 1.65, 1.90,
  1.74, 1.91)
```

R: calculations with vectors

- Calculations with vectors of the same length work just like calculations on numbers – all operations are performed component-wise
- Example: BMI (from Dalgaard's book)

```
> bmi <- weight/height^2
> bmi    # Type this in R to see
  the computed values
```


R: simulating data

- R can also be used to generate (pseudo-) random numbers from a number of distributions (for example, the normal distribution)
- These functions look like **`rnorm()`**, **`rbinom()`**, etc.
- This is a useful facility for learning about sampling variability
- It is also useful for quickly getting a bunch of numbers to use as practice data

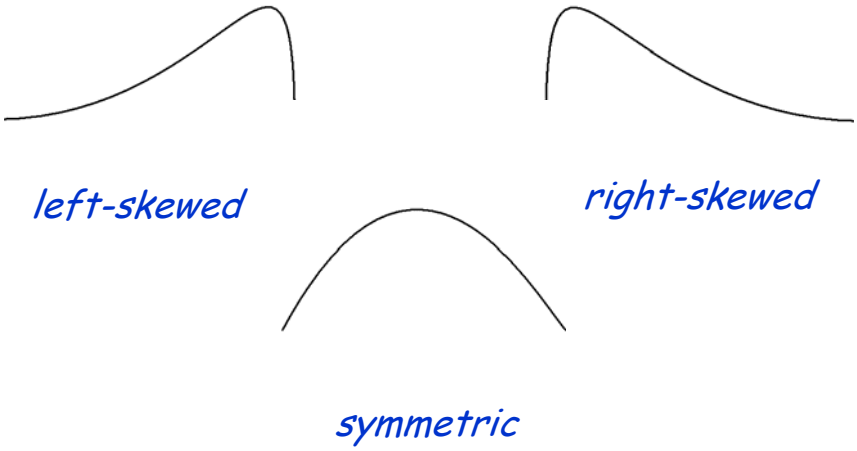
Graphical data summaries

- For a single categorical variable:
  - Bar plot, dot plot (not covered here)
- For a single numerical variable:
  - Histogram (next)
  - Boxplot (a little later)
- For two numerical variables:
  - Scatterplot (in a few weeks)

Histogram

- A *histogram* is a special kind of bar plot
- It allows you to visualize the *distribution* of values for a numerical variable
- When drawn with a *density scale:*
  - the *AREA* (NOT height) of each bar is the proportion of observations in the interval
  - the *HEIGHT* represents density, or 'crowding'
  - the *TOTAL AREA* is 100% (or 1)

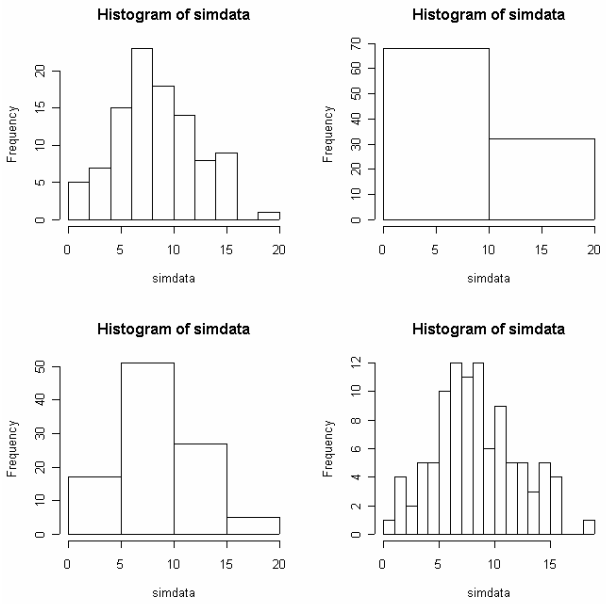
Some general histogram forms



R: making a histogram

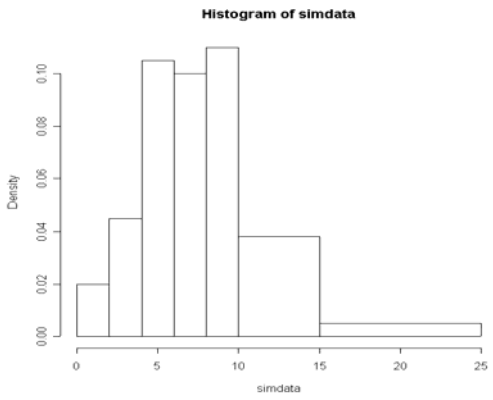
- Type **?hist** to view the help file
  - Note some important arguments, esp **breaks**
- Simulate some data, make histograms varying the number of bars (also called 'bins' or 'cells'):

```
> par(mfrow=c(2,2))    # set up mult. plots
> simdata <-rchisq(100,8)
> hist(simdata)  # default number of bins
> hist(simdata,breaks=2)  # etc,4,20
```



R: setting your own breakpoints

```
> bps <- c(0,2,4,6,8,10,15,25)
> hist(simdata,breaks=bps)
```



(BREAK)

Numerical Summaries

- Categorical/Qualitative variables
  - frequency table (not covered here)
- Numerical/Quantitative variables
  - *center*
  - *spread*

Measures of center: Mean

- The *mean* value of a variable is obtained by computing the *total* of the values divided by the *number of values n*

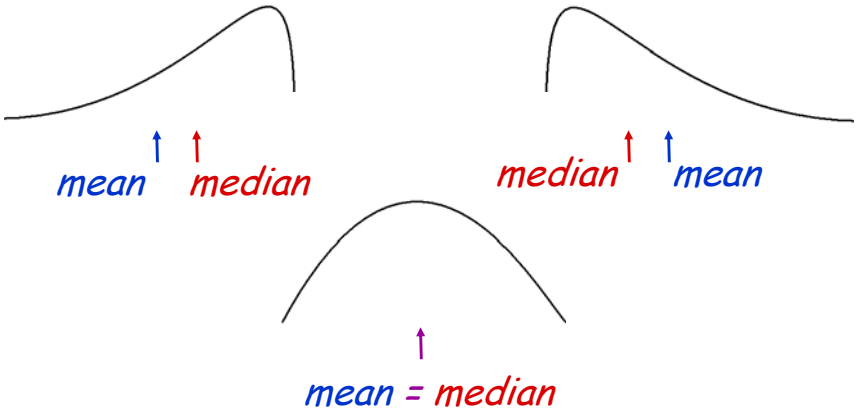
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- Appropriate for distributions that are fairly *symmetrical*
- It is sensitive to presence of *outliers*, since all values contribute equally
- The mean is the 'balance-point' for a histogram
- In R: `> mean(z1)`

Measures of center: Median

- The *median* value of a variable is the 'middlemost number': that is, the number having 50% (half) of the values smaller than it (and the other half bigger)
- It is NOT sensitive to presence of outliers, since it 'ignores' almost all of the data values
- The median is thus usually a more appropriate summary for *skewed distributions*
- In R: `> median(z1)`

Relative location of mean and median



Measures of spread: SD

- The *standard deviation (SD)* of a variable is the square root of the average* of squared deviations from the mean (*for uninteresting technical reasons, instead of dividing by the number of values n, usually divide by n-1)
- The *SD* is an appropriate measure of spread when center is measured with the *mean*
- In R: `> sd(z1)`

Calculating SD

- Although you will virtually never compute the SD by hand in practice, it is instructive to do it once to see how it measures spread
- The *SD* is the *RMS (root-mean-square)* of the *deviations from the average*
- To get the RMS of a list of numbers, do what is says, but in reverse:
  1. *SQUARE* each number in the list
  2. Take the *MEAN* of these squares
  3. Take the square *ROOT* of this mean

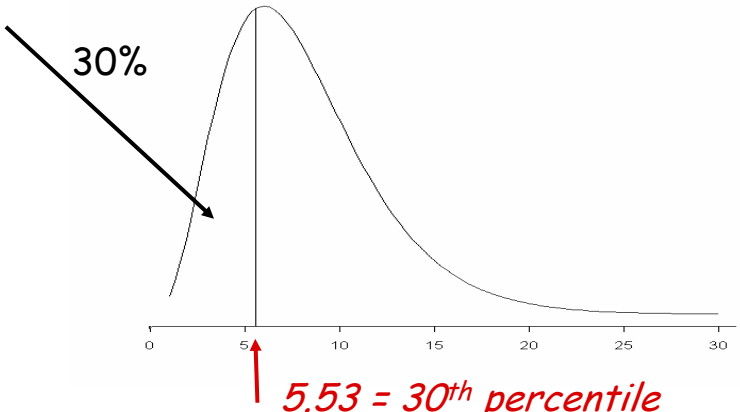
Worked example

Number	Deviation	Squared Dev
4	4 – 7.6 = -3.6	$(-3.6)^2 = 12.96$
14	6.4	40.96
6	-1.6	2.56
9	1.4	1.96
5	-2.6	6.76
sum = 38		sum = 65.2
average = 7.6		average = 13.04

$$SD = \sqrt{13.04} = 3.6$$

Quantiles

- The p^{th} *quantile* is the number that has the proportion p of the data values smaller than it



Measures of spread: IQR

- The 25th (Q_1), 50th (median), and 75th (Q_3) percentiles divide the data into 4 equal parts; these special percentiles are called *quartiles*
- The *interquartile range (IQR)* of a variable is the distance between Q_1 and Q_3:

$$IQR = Q_3 - Q_1$$

- The *IQR* is one way to measure spread when center is measured with the *median*
- In R: `> IQR(z1)  # note CAPITALS here`

Measures of spread: MAD

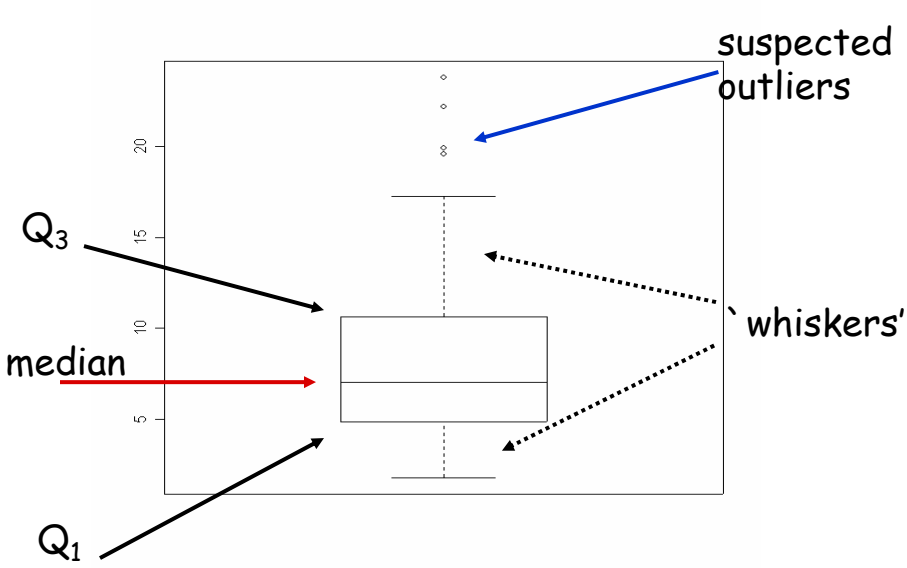
- The *median absolute deviation (MAD)* of a variable is obtained by
  1. getting the absolute values of the deviations between data values and the median, and then
  2. taking the median of those absolute deviations.
- MAD is a more *robust* measure of spread than the SD
- The *MAD* is another way (besides IQR) to measure spread when center is measured with the *median*
- In R: `> mad(z1)`

Five-number summary and boxplot

- An overall summary of the distribution of variable values is given by the five values:

  Min, Q_1, Median, Q_3, and Max
- In R, this summary can be obtained with the function **`quantile()`** (or the function **`summary()`**, which also includes the mean)
- A *boxplot* provides a visual summary of this five-number summary

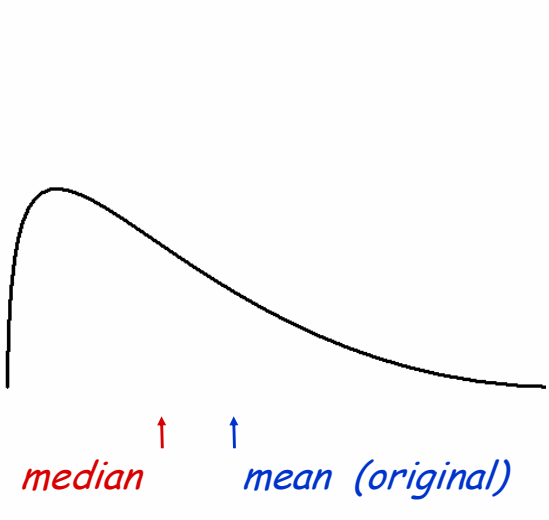
Boxplot of simdata



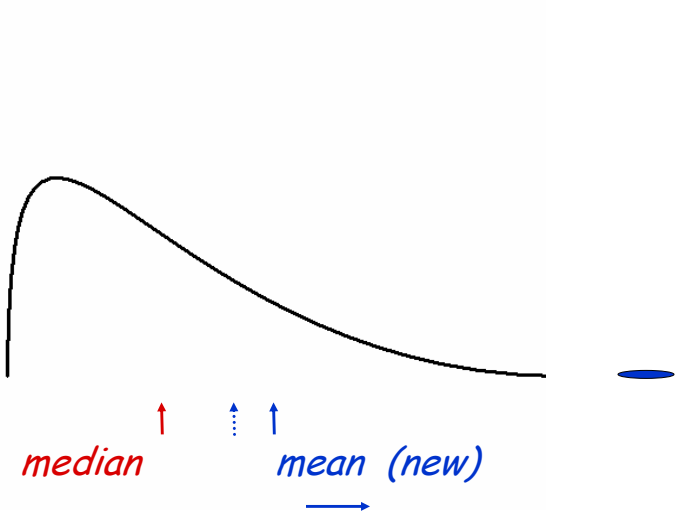
Robustness and resistance

- These concepts refer to *lack of sensitivity* to assumed distributions and effects of a small number of values or outliers
- These qualities are *desirable*: you don't want inferences to be strongly influenced by only a small part of the data set
- The mean is very sensitive to outlying values, the median is very resistant

Robustness of mean, median (1)



Robustness of mean, median (2)



R: session management

- Your R objects are stored in a *workspace*
- To list the objects in your workspace: `> ls()`
- To remove objects you no longer need:
  `> rm(weight, height, bmi)`
- To remove ALL objects in your workspace:
  `> rm(list=ls())` or use Remove all objects in the Misc menu
- To save your workspace to a file, you may type
  `> save.image()`
- The default workspace file is called .RData

R: saving your work and quitting

- You may also save your command history
- When you have finished your R session, you can quit by typing the R command `> q()` or by clicking on the X to close the window
- Don't forget the parentheses!
- You will be asked if you want to save the workspace image; generally, you will say 'yes' if you want R to save the data there for you