群聊：统计学本科教学 2023

该二维码7天内(2月21日前)有效，重新进入将更新

Biological statistics

Li, Xinhai (李欣海)

Ph.D., Associate Professor
Institute of Zoology, Chinese Academy of Sciences
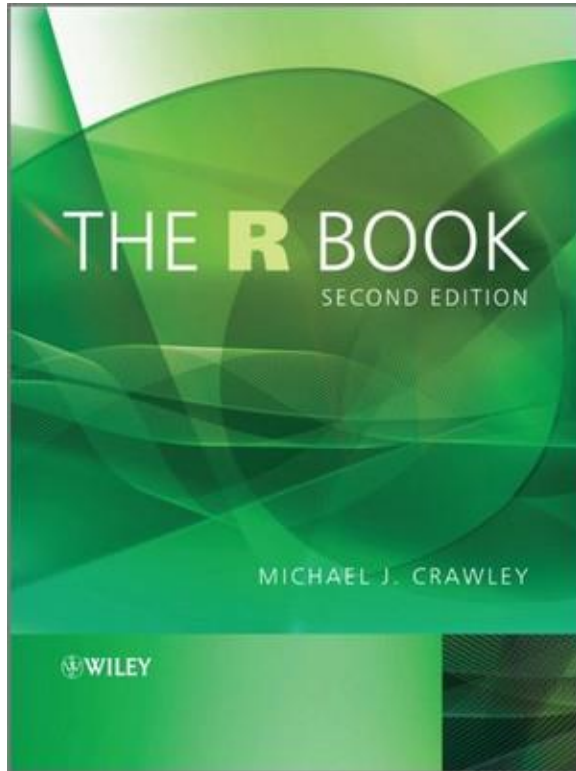1-5 Beichen West Road, Beijing 100101, China

李欣海
北京 朝阳

Phone: 86-10-64807898 (office)
Email: lixh@ioz.ac.cn
WeChat: Xinhai__Li (double underlines)
Homepage: http://people.ucas.edu.cn/~LiXinhai?language=en
Blog: http://blog.sciencenet.cn/u/lixinhai (in Chinese)
Microblog: http://weibo.com/lixinhaiblog (李欣海微博, in Chinese)
ResearchGate: https://www.researchgate.net/profile/Xinhai-Li-3
ORCID: 0000-0003-4514-0149
ResearcherID: G-9111-2011
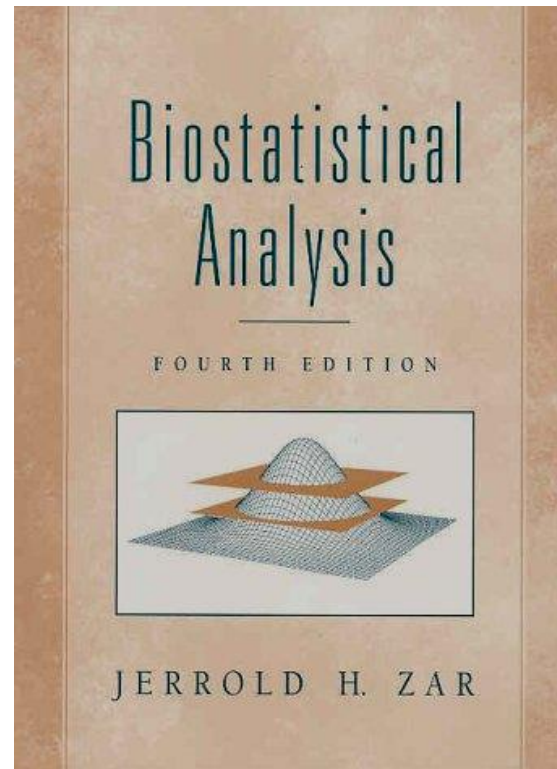
扫一扫上面的二维码图案，加我微信

How to learn statistics in this class

- No preview needed before the class

- Focus on listening and thinking at class (2×2 hours / week)

    – Don't take notes (It distracts your attention)

- Intensive review (~1-2 hours) after each class

    – Google/Bing your questions (using the key words I provided)

- Do the homework (~1 hour / class)

Key references

Text books

Zuur A, E. N.Ieno, and E. Meesters. 2009. A Beginner's Guide to R. Springer. 216 pp.
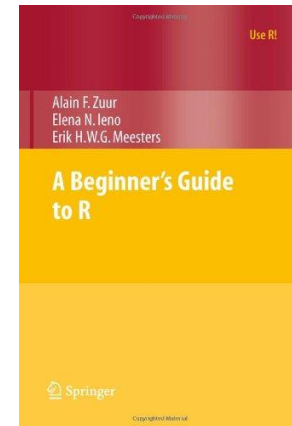

Crawley M. J. 2012. **The R book**. John Wiley & Sons, Chichester, UK, 1076 pp.

Zar, J. H. 1999. **Biostatistical Analysis**. Fourth Edition. Prentice Hall, New Jersey, 663 pp.
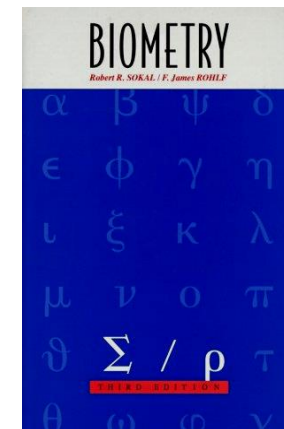
Sokal, R. R. and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. Third Edition. W. H. Freeman and Co., New York. 887 pp.

3

What is statistics?

- Statistics is the science of collection, analysis, interpretation, and presentation of data.

- Descriptive statistics are numerical estimates that organize, sum up or present the data.

- Inferential statistics is the process of inferring from a sample to the population.

Statistical errors in publications

Nuzzo, Regina. 2014. Scientific method: Statistical errors. **Nature** 506: 150-152

"*P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume."

Statistical errors in publications

Underwood (1981) found statistical errors in **78%** of the papers he surveyed in marine ecology.

Hurlbert (1984) reported that in two separate surveys **26%** and **48%** of the ecological papers surveyed showed the statistical error of pseudoreplication.

Charles J. Krebs. 1999. Ecological Methodology, 2nd ed. Addison-Wesley Educational Publishers, Inc.

50% of medical literature have statistical flaws (Altman et al. 1991).

Serious statistical errors were found in **40%** of 164 articles published in a psychiatry journal (McGuigan 1995)" (Ercan et al. 2007).

Ilker Ercan, Berna Yazıcı, Yaning Yang, Guven Özkaya, Sengul Cangur, Bulent Ediz, Ismet Kan. Misusage Of Statistics In Medical Research. Eur J Gen Med 2007; 4(3):128-134

Contents $y = x_1 + x_2 + \ldots + x_n + \varepsilon$

- Brief history, data description, and descriptive statistics (2h)

- Probability theory and important distributions (2h)

- Hypothesis testing (4h)

- Analysis of variance (ANOVA) (4h)

- Simple linear regression and correlation (2h)

- Analysis of covariance (ANCOVA) (2h)

- Nonparametric statistics (2h)

- Multiple correlation and regression (2h)

- Cluster analysis and discriminant analysis (2h)

- Ordination (2h)

- Generalized linear model (2h)

- Sample survey (2h)

- Bayesian method (2h)

- Machine learning (2h)

- Reports and practices (4h)

Statistical software R   http://cran.r-project.org

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R is based on S, a commercial language. In 1995, to use S for free, **R**oss Ihaka and **R**obert Gentleman (at the Department of Statistics of the University of Auckland in Auckland, New Zealand) designed a software, named R.

Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive.

It has 12237 packages in March, 2018.
http://cran.r-project.org/web/packages/

Citation
R Development Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0. http://www.R-project.org.

Number of Active CRAN Packages

(Hornik 2012)

8
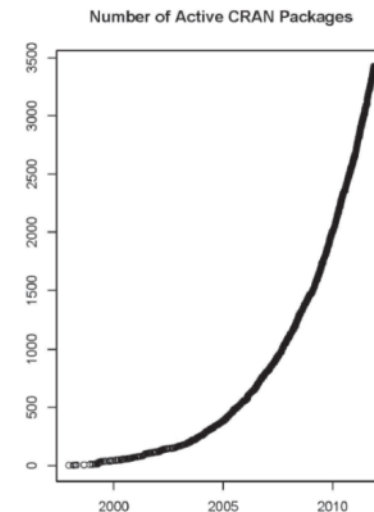
TIOBE
the software quality company   **indicator of the popularity of programming languages**

About us   Join TIOBE   News   Coding Standards   TIOBE Index   Contact

Products ⌄   Quality Models ⌄   Markets ⌄   Schedule a demo

Feb 2023	Feb 2022	Change		Programming Language	Ratings	Change
1	1			Python	15.49%	+0.16%
2	2			C	15.39%	+1.31%
3	4	⌃		C++	13.94%	+5.93%
4	3	⌄		Java	13.21%	+1.07%
5	5			C#	6.38%	+1.01%
6	6			Visual Basic	4.14%	-1.09%
7	7			JavaScript	2.52%	+0.70%
8	10	⌃		SQL	2.12%	+0.58%
9	9			Assembly language	1.38%	-0.21%
10	8	⌄		PHP	1.29%	-0.49%
11	11			Go	1.11%	-0.12%
12	13	⌃		R	1.08%	-0.04%
13	14	⌃		MATLAB	0.99%	-0.04%
14	15	⌃		Delphi/Object Pascal	0.95%	+0.05%
15	12	⌄		Swift	0.93%	-0.25%
16	16			Ruby	0.83%	-0.06%
17	19	⌃		Perl	0.79%	-0.01%
18	22	⌃		Scratch	0.76%	+0.13%
19	17	⌄		Classic Visual Basic	0.74%	-0.09%
20	24	⌃		Rust	0.70%	+0.16%

9

TOOLBOX

PROGRAMMING TOOLS: ADVENTURES WITH R

A guide to the popular, free statistics and visualization software that gives scientists control of their own data analysis.
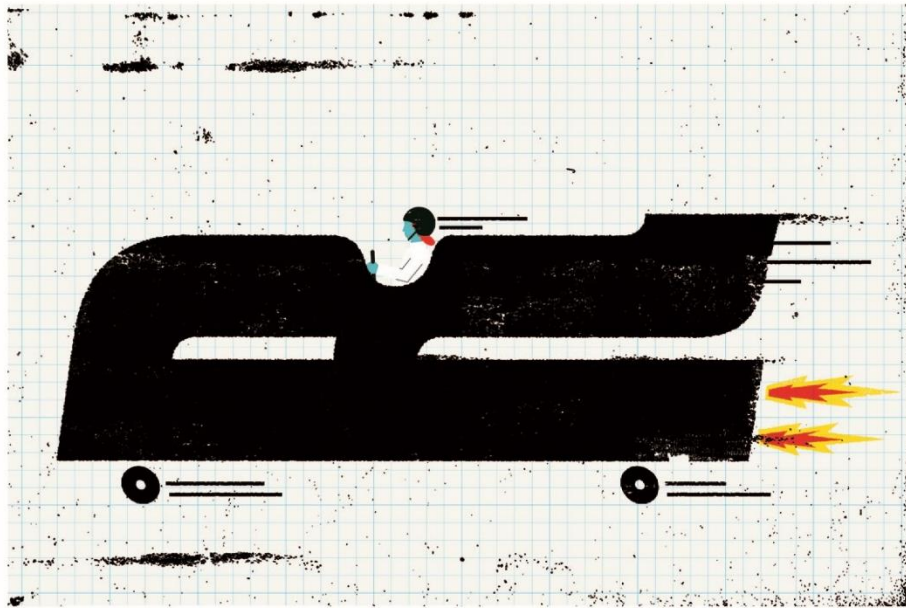
BY SYLVIA TIPPMANN

For years, geneticist Helene Royo used commercial software to analyse her work. She would extract DNA from the developing sperm cells of mice, send it for analysis and then fire up a package called GeneSpring to study the results. "As a scientist, I wanted to understand everything I was doing," she says. "But this kind of analysis didn't allow that: I just pressed buttons and got answers." And as Royo's studies comparing genetic activity on different chromosomes became more involved, she realized that the commercial tool could not keep up

with her data-processing demands.

With the results of her first genomic sequencing experiments in hand at the start of a new postdoc, Royo had a choice: pass the sequences over to the experts or learn to analyse the data herself. She took the plunge, and began learning how to parse data in the free, open-source software package R. It helped that the centre she had joined — the Friedrich Miescher Institute for Biomedical Research in Basel, Switzerland — ran regular courses on the software. But she was also following a wider trend: for many academics seeking to wean themselves off commercial software, R is the data-analysis tool of choice.

Besides being free, R is popular partly because it presents different faces to different users. It is, first and foremost, a programming language — requiring input through a command line, which may seem forbidding to non-coders. But beginners can surf over the complexities and call up preset software packages, which come ready-made with commands for statistical analysis and data visualization. These packages create a welcoming middle ground between the comfort of commercial 'black-box' solutions and the expert world of code. "R made it very easy," says Rojo. "It did everything for me."

That, indeed, is what R's developers ▶

Programming tools: Adventures with R

Tippmann, Sylvia. 2015. **Nature**: 517, 109-110.

R is the popular, free statistics and visualization software that gives scientists control of their own data analysis.

Although most people like click-and-drop interfaces, programming is usually needed by scientists. Among all the languages, R is one of the easiest.
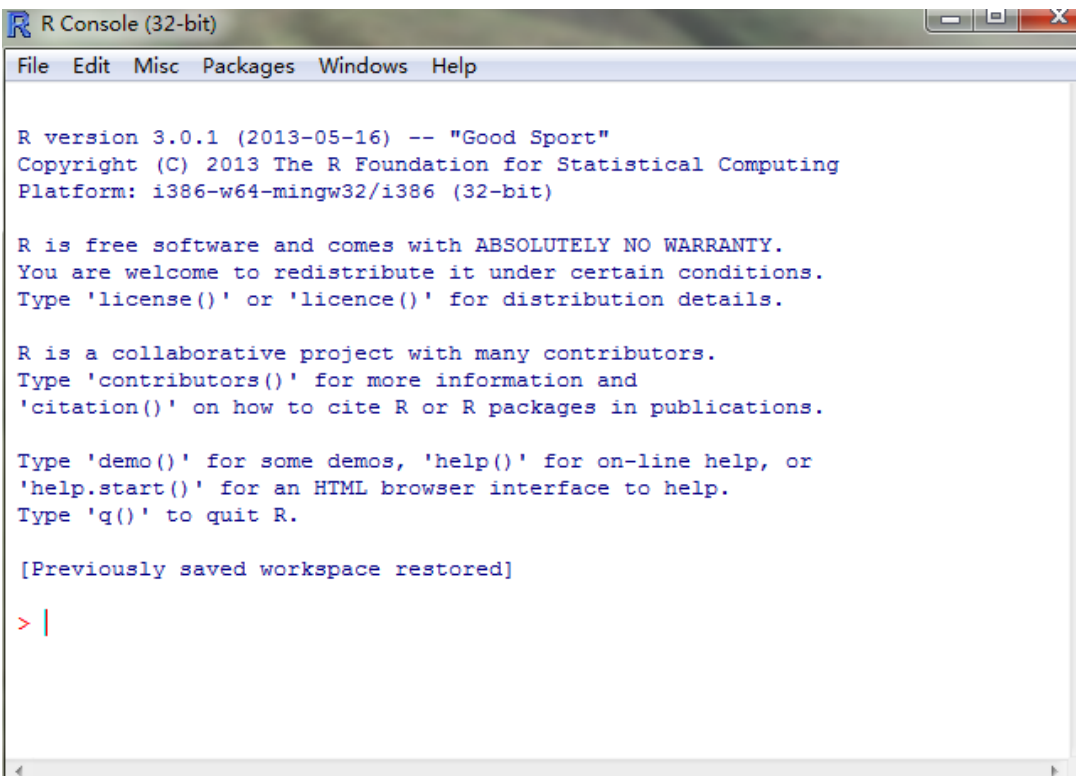
With over 6000 packages, R is too powerful.

Installing R–Base System

1. Go to http://CRAN.R-project.org

2. Choose your computer from the list (*Linux, MacOS X, or Windows*)

3. Click on Base

4. Click on R-4.1.2-win64.exe (for Windows)

5. Install it to D:***\\R

R script

http://cran.r-project.org

```
R Console (32-bit)
File  Edit  Misc  Packages  Windows  Help

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> |
```

To create a new R script , you may:
choose File > New script

print the current working directory
> getwd()

list the objects in the workspace
> ls()

change the working directory
> setwd("d:/models")

load the package "raster"
> library(raster)

to get help with the package
> help(raster)

calculation
> 99^3

12

R Interfaces

- **RStudio**
- RWinEdt
- Tinn-R
- JGR (Java Gui for R)
- Emacs + ESS
- Rattle
- AKward
- Playwith (for graphics)

https://rstudio.com/products/rstudio/
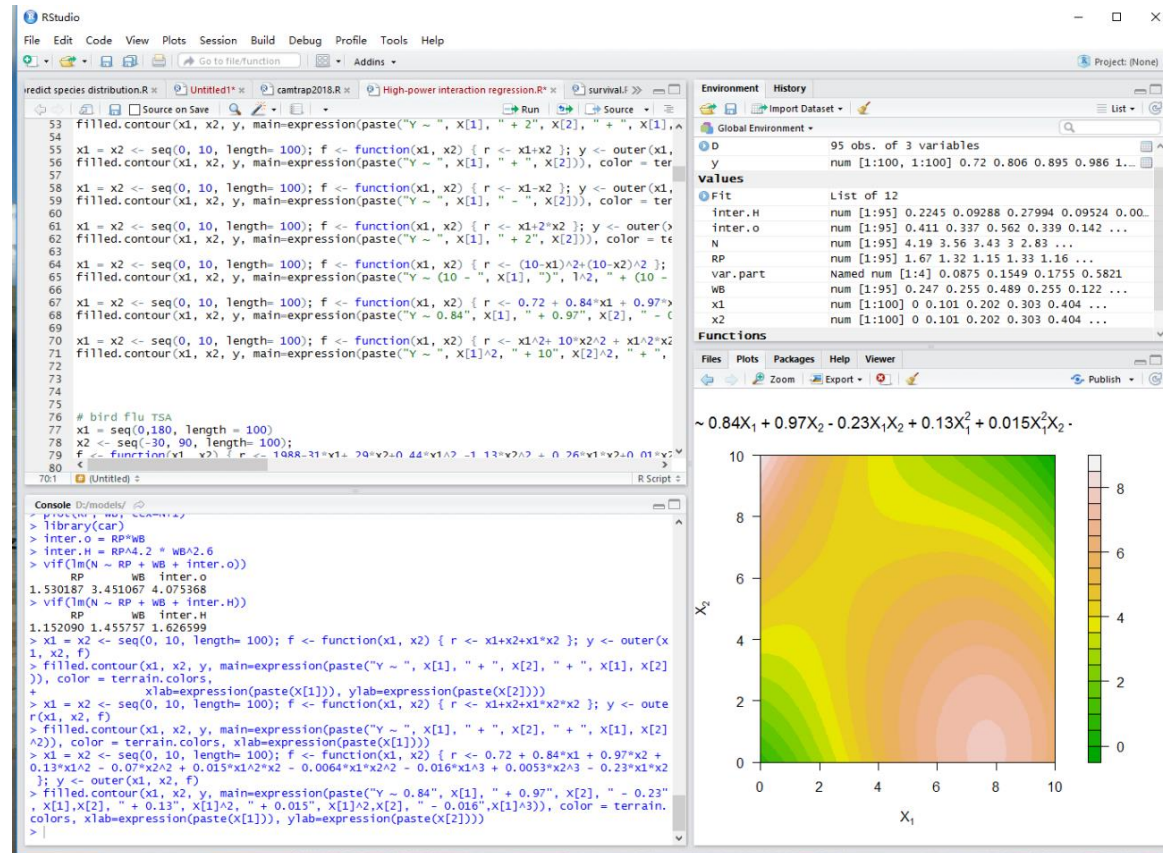
R code - calculation

9+2

[1] 11

6+2^2

[1] 10

(3+4)^2

[1] 49

sqrt(2)

[1] 1.414214

log(2)

[1] 0.6931472

x = 51

y = 10

z <- x+y

z

[1] 61

Today's contents
Introduction to biological statistics

- History

- Data in biology

- Descriptive statistics

History of statistics

- 1500s, the arithmetic mean, the median

- 1600s, life table

- 1700s, probability theory, normal distribution

- 1800s, least squares, standard deviation, correlation, regression

- 1900s, design of experiments, hypothesis test, maximum likelihood; categorical / multivariate / time-series / survival analysis

- Today, computer-intensive methods

Most important people

- John Graunt (1620-1674, British) and William Petty (1623-1687, British): developed early human statistical and census methods that later provided a framework for modern demography based on life table, mean value, census, longevity, and mortality.

- Blaise Pascal (1623-1662, French) and Pierre de Fermat (1601-1665, French), Jacques Bernoulli (1654-1705, Swiss): probability theory (binomial coefficients, mathematical expectation, the law of large numbers).

- Abraham de Moivre (棣莫弗)(1667-1754, French): defined "Independent Event"; provided Binomial Distribution, approximated the normal distribution though the expansion of the binomial distribution.

- Carl Friedrich Gauss (1777-1855, Germany): least square, normal distribution.

- Adolphe Quetelet (凯特勒) (1796-1874, Belgium): significance of constancy of large numbers (rate of criminal events).

- Florence Nightingale (1820-1910, British): graphic presentation of statistics.

17
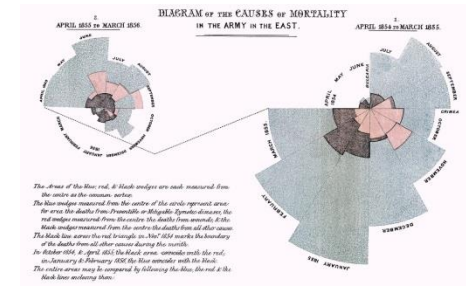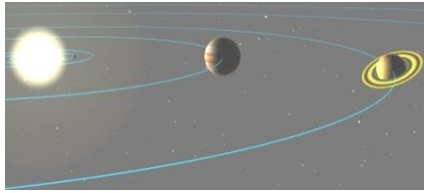
Emergence of statistics in 1800's

- Laplace wrote a book describing how to compute the future positions of planets and comets on the basis of a few observations from earth.

- Napoleon: "I find no mention of God in your treatise, Mr. Laplace."

- Laplace replied: "I had no need for that hypothesis."

- The observations of planets and comets from this earthly platform did not fit the predicted positions exactly. Laplace and his fellow scientists attributed this to errors in the observations, sometimes due to perturbations in the earth's atmosphere, other times due to human error.

- By the end of the nineteenth century, the errors had mounted instead of diminishing. As measurements became more and more precise, more and more error cropped up.

18

Gaps between Darwinism and genetics in early 1900's

Core Evolution Concepts

Population: Organisms that share a common gene pool (Species = actually or potentially interbreeding organisms)

Variation: Modifications of forms are produced by chance via mutations, genetic coding errors of individual organisms

Natural Selection: Reproduction & survival of organisms whose heritable traits are better suited to existing environmental conditions

Retention: Persistence within a population of the selected variation(s) over successive generations

Mendel's law of segregation

By carrying out the monohybrid crosses, Mendel determined that the 2 alleles for each character segregate during gamete production.

19
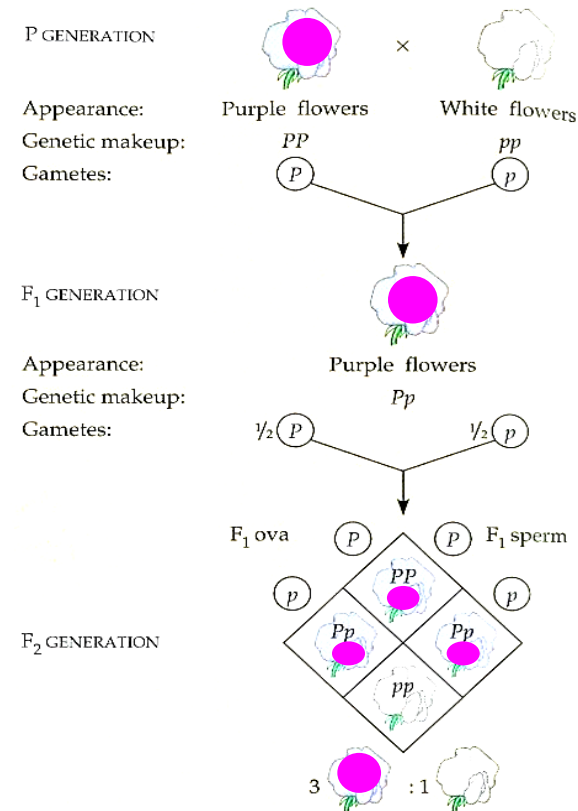
Neo-Darwinian modern evolutionary synthesis in 1930's

- **Ronald A. Fisher** (1890-1962, British) developed several basic statistical methods in support of his work *The Genetical Theory of Natural Selection.*

- **Sewall G. Wright** (1889-1988, American) computed the distribution of gene frequencies among populations as a result of the interaction of natural selection, mutation, migration and genetic drift, proposed the inbreeding coefficient.

- **John B. S. Haldane** (霍尔丹1892-1964, British**)** used maximum likelihood for estimation of human linkage maps, and pioneering methods for estimating human mutation rates.

Francis Galton

- Francis Galton (1822-1911, British) (father of biometry and eugenics): regression, correlation

  - African Explorer and elected Fellow in the Royal Geographic Society

  - Creator of the first weather maps and establisher of the meteorological theory of anticyclones

  - Coined term "eugenics" and phrase "nature versus nurture"

  - Developed statistical concepts of correlation and regression

  - Discovered that fingerprints were an index of personal identity and persuaded Scotland Yard to adopt a fingerprinting system

  - First to utilize the survey as a method for data collection

  - Produced over 340 papers and books throughout his lifetime

  - Knighted in 1909

Galton, F. (1869/1892/1962). *Hereditary Genius: An Inquiry into its Laws and Consequences*. Macmillan/Fontana, London.
Galton, F. (1883/1907/1973). *Inquiries into Human Faculty and its Development*. AMS Press, New York.
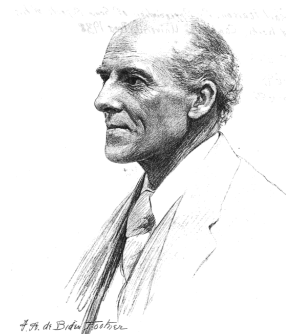
Karl Pearson

- Karl Pearson (1857-1936, British): continued in the tradition of Galton and laid the foundation for much of descriptive statistics.

    – In 1884, Pearson became Professor of Applied Mathematics and Mechanics at University College London.

    – In 1901 Pearson, Weldon and Galton founded Biometrika, a "Journal for the Statistical Study of Biological Problems".

    – In 1907, Pearson took over a research unit founded by Galton and reconstituted it as the Francis Galton Laboratory of National Eugenics.

    – In 1911, Pearson founded the world's first university statistics department at University College London.

    ✓ method of moments
    ✓ chi-square
    ✓ correlation

22

Ronald A. Fisher

http://en.wikipedia.org/wiki/Image:RonaldFisher.jpg

- Sir Ronald Aylmer Fisher, (1890 –1962), an English statistician, evolutionary biologist, and geneticist.

- He was described by Anders Hald as "a genius who almost single-handedly created the foundations for modern statistical science"[1] and Richard Dawkins described him as "the greatest biologist since Darwin".[2] (from Wikipedia)

  – In 1933 he became a Professor of Eugenics at University College London
  – In 1943 he was offered the Balfour Chair of Genetics at Cambridge University

✓ Analysis of variance
✓ Maximum likelihood
✓ Fisher information

Fisher, R.A. 1925. Statistical Methods for Research Workers
Fisher, R.A. 1935. The design of experiments

[1] Hald, Anders (1998). A History of Mathematical Statistics. New York: Wiley.

[2] Dawkins, Richard (1995). River out of Eden.

23

Society and publications in early years

- In 1901, Pearson, Weldon and Galton founded **Biometrika**, a "Journal for the Statistical Study of Biological Problems".

- Until the 1940s, the application of statistics to biological questions began to have a profound impact on the scientific community.

- The biometrics section of the American Statistical Association to publish the **Biometrics Bulletin**, in 1945.

- In 1947, International Biometric Society (IBS) was established. Shortly thereafter, the IBS began publishing **Biometrics**.

Current statistical journals (top 25 from 179)

Rank	Title	SJR	SJR Quartile	H index	Total Docs. (2015)	Cites / Doc. (2years)	Country
1	Annals of Mathematics	10.358	Q1	78	43	3.04	United States
2	Journal of the Royal Statistical Society. Series B	7.429	Q1	96	61	4.28	United Kingdom
3	Annals of Statistics	6.653	Q1	113	72	2.91	United States
4	Vital and health statistics. Series 10	6.119	Q1	31	0	8.86	United States
5	Bioinformatics	4.643	Q1	271	867	4.98	United Kingdom
6	Statistical Methods in Medical Research	3.774	Q1	58	55	3	United Kingdom
7	Annals of Probability	3.519	Q1	58	81	1.72	United States
8	Journal of the American Statistical Association	3.447	Q1	133	154	1.85	United States
9	Finance and Stochastics	3.019	Q1	30	33	1.75	Germany
10	Journal of Statistical Software	2.97	Q1	76	91	2.47	United States
11	Probability Surveys	2.906	Q1	25	3	2.09	United States
12	Probability Theory and Related Fields	2.896	Q1	53	63	1.67	United States
13	Biometrika	2.801	Q1	90	75	1.06	United Kingdom
14	Annals of Applied Probability	2.685	Q1	57	96	1.88	United States
15	Journal of Business and Economic Statistics	2.566	Q1	71	48	1.49	United States
16	Multivariate Behavioral Research	2.431	Q1	53	42	1.31	United States
17	British Journal of Mathematical and Statistical Psychology	2.38	Q1	36	23	3.39	United States
18	Statistical Science	2.366	Q1	72	33	2.31	United States
19	Journal of Computational and Graphical Statistics	2.321	Q1	61	59	1.58	United States
20	Statistica Sinica	2.292	Q1	57	24	0.83	Taiwan
21	Bernoulli	2.12	Q1	44	91	1.38	Netherlands
22	Extremes	2.044	Q1	21	26	1.42	Netherlands
23	Statistics and Computing	1.993	Q1	50	82	1.35	Netherlands
24	Biostatistics	1.955	Q1	55	48	1.92	United Kingdom
25	Biometrics	1.906	Q1	96	186	1.23	United Kingdom

A story of statistics in industry

- In 1980, the NBC television network aired a documentary entitled "If Japan Can, Why Can't We?"

  – The documentary was really a description of the influence one man had on Japanese industry, **W. Edwards Deming**.

- Deming's major point about quality control is that the output of a production line is variable, because that is the nature of all human activity. What the customer wants is not a perfect product but a reliable product.

W. Edwards Deming

W. Edwards Deming (1990-1993)

Dr. William Edwards Deming was an American engineer, statistician, professor, lecturer, and management consultant. He helped develop the sampling techniques still used by the U.S. Department of the Census and the Bureau of Labor Statistics, and helped hasten Japan's recovery after the Second World War.

Quotes

• "A bad system will beat a good person every time."
• "It is not necessary to change. Survival is not mandatory."
• "Quality is pride of workmanship."

Out of the Crisis	Quality, productivity, and competitive pos...	The New Economics for Industry, Gover...	Some Theory of Sampling	Statistical adjustment of data	The best of Deming	Sample design in business research	The Essential Deming: Leadership Prin...	The world of W. Edwards Deming
★★★★☆	★★★★☆	★★★★½	★★★★☆	★★★½☆	★★★★☆	★★★★☆	★★★★☆	★★★★☆
Feb 02, 1982	1982	1993	Sep 16, 2010	1984	Nov 09, 1994	Dec 1960	Oct 12, 2012	Jun 01, 1988
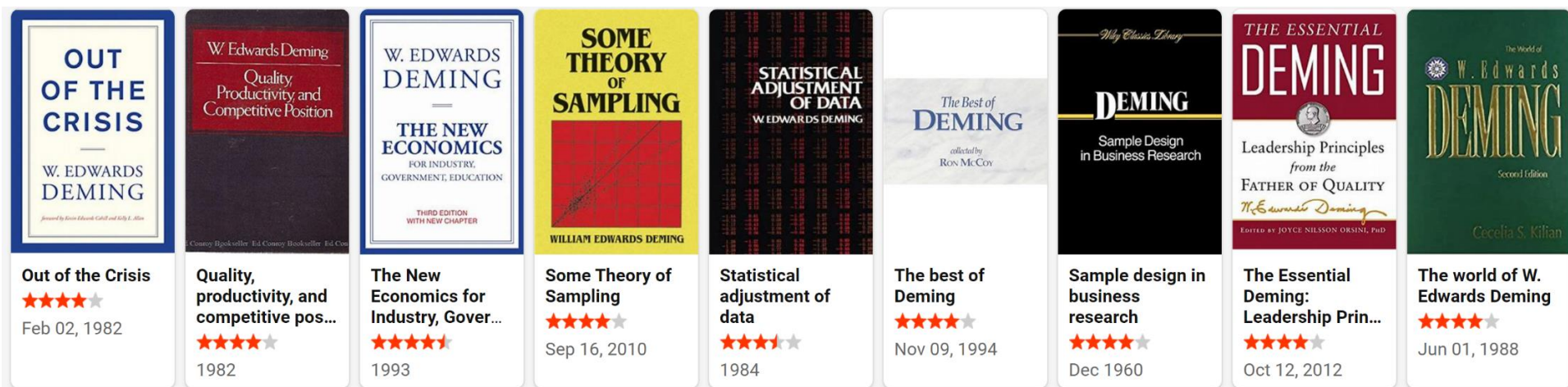
A story about statistics and industry:
Deming's quality control

- Deming proposed that the production line be seen as a stream of activities that start with raw material and end with finished product.

- Each activity can be measured, so each activity has its own variability due to environmental causes.

- Instead of waiting for the final product to exceed arbitrary limits of variability, the managers should be looking at the variability of each of these activities.

- The most variable of the activities is the one that should be addressed. Once that variability is reduced, there will be another activity that is "most variable," and it should then be addressed.

- Thus, quality control becomes a continuous process, where the most variable aspect of the production line is constantly being worked on.

Data

- Datum is one observation about the variable being measured.

- Data are a collection of observations.

- A **population** consists of all subjects about whom the study is being conducted.

- A **sample** is a sub-group of population being examined.

Variables

Quantitative   Qualitative

Ordinal   Interval or ratio

- **Nominal variable**
  - classification data, e.g., male/female, 0/1, etc

- **Ordinal variable**
  - ordered but differences between values are not important
  - e.g., Likert scales, rank on a scale of 1..5 (degree of satisfaction); restaurant ratings

- **Interval scale variable**
  - ordered, constant scale, but no natural zero
  - differences make sense, but ratios do not (e.g.,
  - 30^0-20^0 = 20^0-10^0, but $20^0/10^0$ is not twice as hot!
  - e.g., temperature (C,F), dates

- **Ratio scale variable**
  - ordered, constant scale, natural zero
  - e.g., height, weight, age, length

30
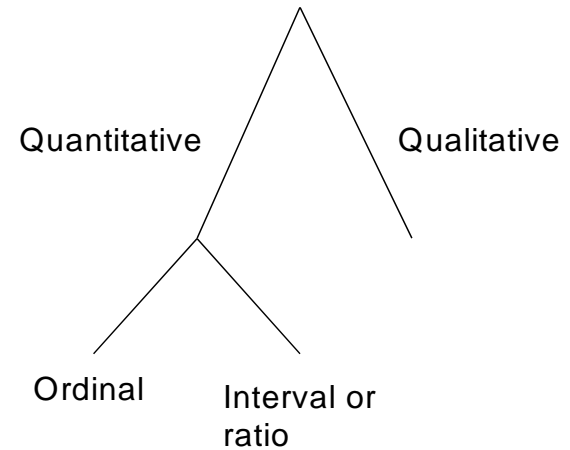
Derived variables

– Ratio

Sex ratio

– Index

S&P 500 index (stock market)

– Rate

Growth rate

Parameters vs. Statistics

- A parameter is a numerical quantity measuring some aspect of a population of scores.
    - For example, the mean is a measure of central tendency
    - Usually use Greek letters

- A statistic computed in samples is used to estimate parameters

Quantity	Parameter	Statistic
Mean	μ	M
Standard deviation	σ	s
Proportion	π	p
Correlation	ρ	r

32

Accuracy and precision of data
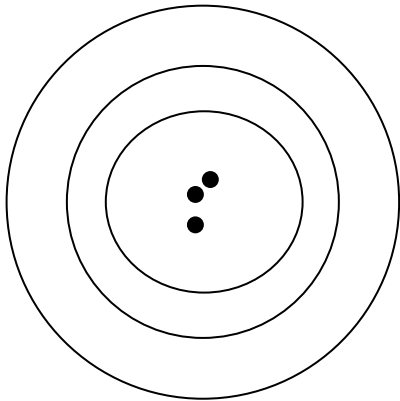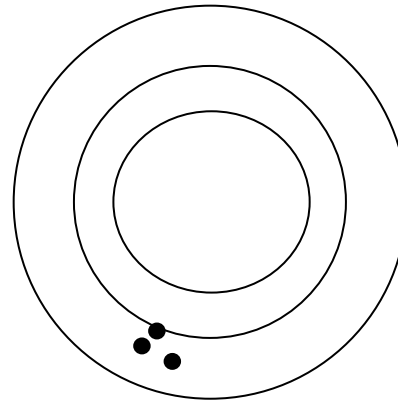
Accuracy   Precision   Inaccuracy
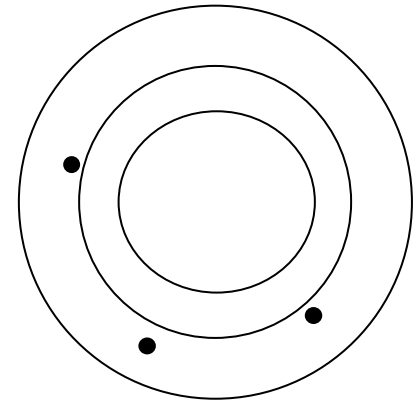
Accuracy of data

Mean square error
for estimating population mean (μ)
using sample mean (m)

$$MSE(M)$$

$$= E[(M - \mu)^2]$$

$$= Var(M) + [E(M) - \mu]^2$$

precision   bias

34
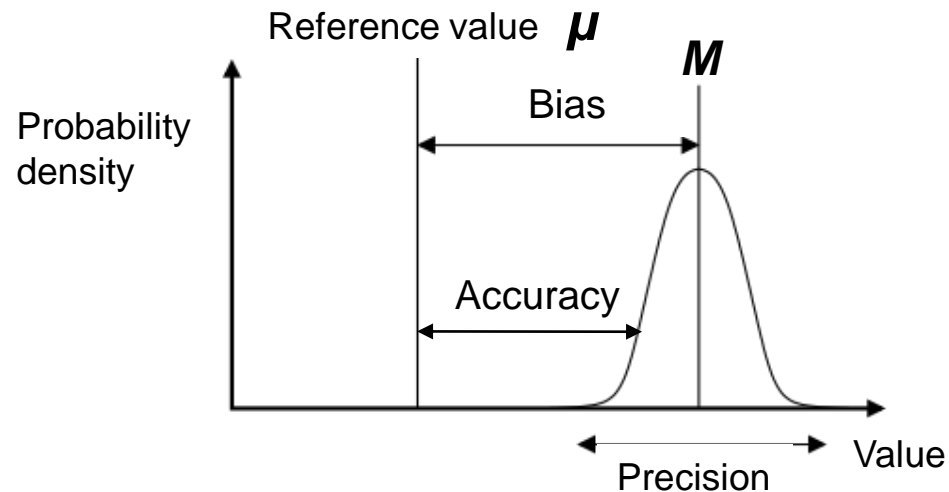
Summarizing data

- Frequency distribution
- Cumulative distributions
- Relative frequency distribution
- Percent frequency distribution
- Bar graph
- Histogram
- Pie chart
- Dot plot

Frequency distribution for qualitative data

A frequency distribution is a tabular summary of data showing the frequency (or number) of items in each of several nonoverlapping classes.

The objective is to provide insights about the data that cannot be quickly obtained by looking only at the original data.

Frequency distribution

An investigator estimated habitat quality for a species:

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	**20**

An example for quantitative data:
Length of first stage juvenile sturgeon (mm)

Total length

73	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	91

Frequency distribution

- ## Guidelines for selecting number of classes

Use between 5 and 20 classes

Data sets with a larger number of elements usually require a larger number of classes

Smaller data sets usually require fewer classes

Use classes of equal width

Approximate class width =

$$\frac{\text{Largest data value} - \text{smallest data value}}{\text{Number of classes}}$$

Frequency distribution

For sturgeon, if we choose six classes:

Approximate Class Width = $(109 - 52)/6 = 9.5 \cong 10$

Length (mm)	Frequency
50-59	2
60-69	13
70-79	16
80-89	7
90-99	7
100-109	5
Total	50

Relative frequency distribution

The relative frequency of a class is the fraction or proportion of the total number of data items belonging to the class.

A relative frequency distribution is a tabular summary of a set of data showing the relative frequency for each class.

Percent frequency distribution
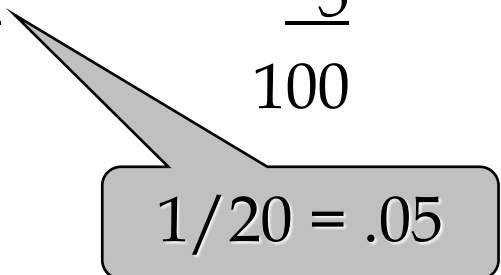
The percent frequency of a class is the relative frequency multiplied by 100.

A percent frequency distribution is a tabular summary of a set of data showing the percent frequency for each class.

Relative frequency and percent frequency distributions

Habitat quality for a species
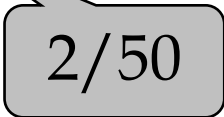
Rating	Relative Frequency	Percent Frequency
Poor	.10	10
Below Average	.15	15
Average	.25	25
Above Average	.45	45
Excellent	.05	5
Total	1.00	100

$1/20 = .05$

43

Relative frequency and percent frequency distributions

Fish length

Length	Relative Frequency	Percent Frequency
50-59	.04	4
60-69	.26	26
70-79	.32	32
80-89	.14	14
90-99	.14	14
100-109	.10	10
Total	1.00	100

2/50

44

Relative frequency and percent frequency distributions

Insights gained from the percent frequency distribution

- Only 4% of the fish are in the 50-59mm class.

- 30% of the fish are under 70mm.

- The greatest percentage (32% or almost one-third) of the fish are in the 70-79mm class.

- 10% of the fish are 100mm or more.

Probability density function (PDF)

A probability density function (pdf) is a function that represents a probability distribution in terms of integrals.

Formally, a probability distribution has density *f(x)*, such that the probability of the interval [a, b] is given by

$$\int_a^b f(x)dx$$

Intuitively, if a probability distribution has density $f(x)$, then the infinitesimal interval [x, $x +$ dx] has probability $f(x)$ dx.
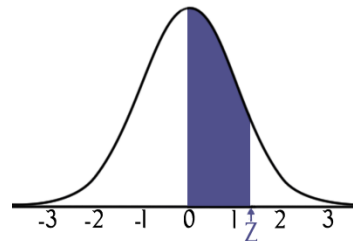
x = seq(-5, 5, by=0.1)
plot(dnorm(x, mean = 0,sd = 1), type='l')

The total area under the graph is 1 $\int_{-\infty}^{\infty} f(x)dx = 1$

46

Cumulative distributions

Cumulative frequency distribution - shows the number of items with values less than or equal to the upper limit of each class..

Cumulative relative/ percent frequency distribution
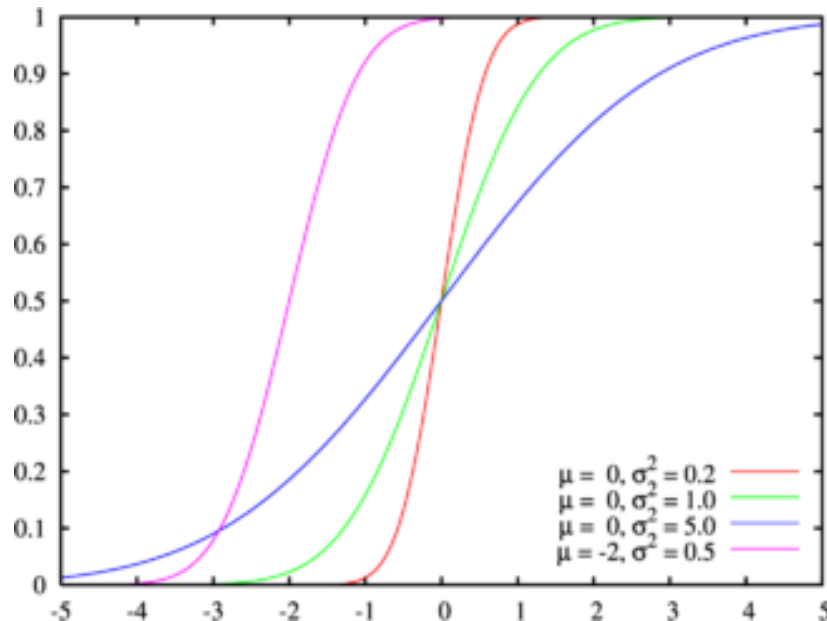
R code

```
x = seq(-5,5,by=0.1)
plot(pnorm(x, mean=0,sd=1),type='l')
```

47

Cumulative distributions

Fish length

Length	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
≤ 59	2	.04	4
≤ 69	15	.30	30
≤ 79	31	.62	62
≤ 89	38	.76	76
≤ 99	45	.90	90
≤ 109	50	1.00	100

2 + 13

15 / 50

48

```
barplot(c(10,26,4,42,19,3))
```

Bar graph
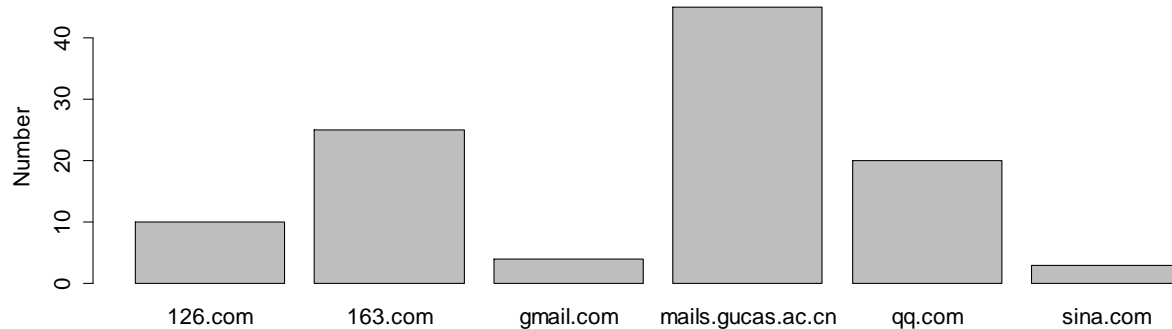
- A bar graph is a graphical device for depicting qualitative data.

- Specify the labels that are used for each of the classes on one axis (usually the horizontal axis).

- A frequency, relative frequency, or percent frequency scale can be used for the other axis (usually the vertical axis).

- Use a bar of fixed width drawn above each class label.

- The bars are separated to emphasize the fact that each class is a separate category.

49

Histogram

- Another common graphical presentation of quantitative data is a histogram.

- The variable of interest is placed on the horizontal axis.

- A rectangle is drawn above each class interval with its height corresponding to the interval's frequency, relative frequency, or percent frequency.

- Unlike a bar graph, a histogram has no natural separation between rectangles of adjacent classes.

Histogram of rnorm(100)

```
R script
```
hist(rnorm(100), nclass=9, col="grey")

50
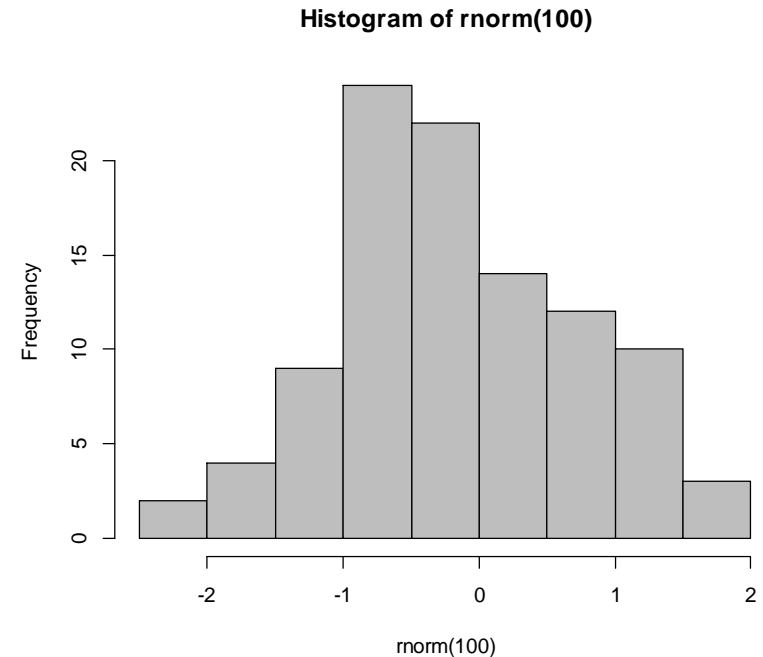
Our class

students <- read.csv('D:/text/statistics/2021/undergraduate/name list.csv', header=T)
family.name <- substr(students$Name,1,1)
length(unique(family.name)) #24
table(family.name)

曾	陈	董	苟	郭	黄	纪	兰	李	梁	廖	刘	马	孟	潘	田	王	文	肖	徐	杨	羿	张	朱
2	4	1	1	1	1	1	1	3	2	1	3	3	1	1	1	3	2	1	2	3	1	6	2

phone <- substr(students$Phone,1,3)
barplot(table(phone),
        xlab = 'Phone',
        ylab = "Frequency")
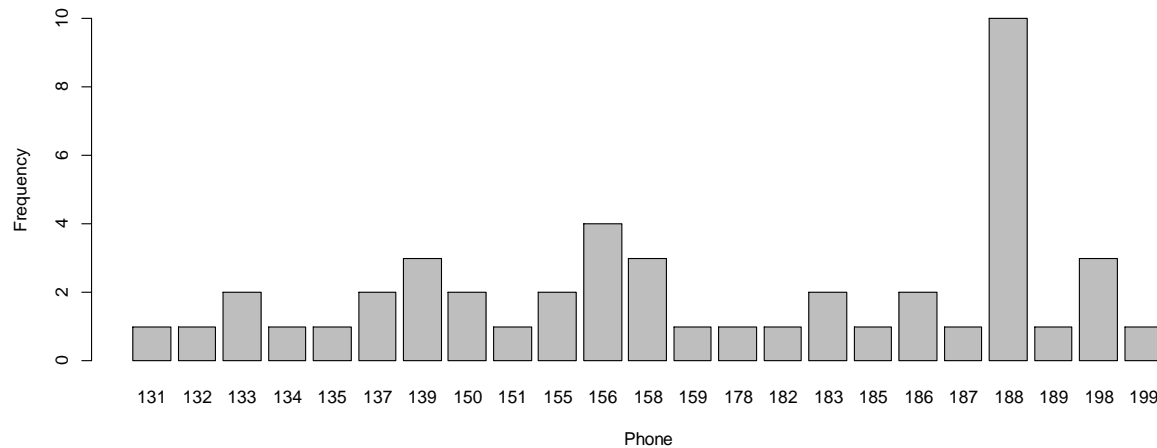
barplot(table(students$Major),
        xlab = 'Major',
        ylab = "Frequency")

51
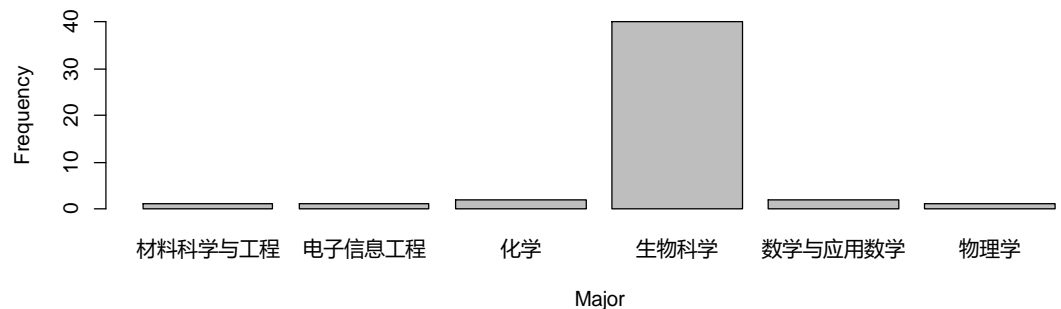
Pie Chart

R code
```
X <- sample(1:100,6,replace=TRUE)
names(x) <- c('A','B','C','D','E','F')
pie(x)
```
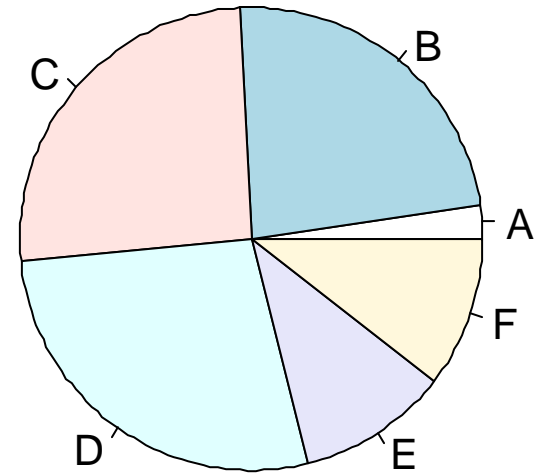
- The pie chart is a commonly used graphical device for presenting relative frequency distributions for qualitative data.

- First draw a circle; then use the relative frequencies to subdivide the circle into sectors that correspond to the relative frequency for each class.

- Since there are 360 degrees in a circle, a class with a relative frequency of .25 would consume .25(360) = 90 degrees of the circle.

Dot Plot

- One of the simplest graphical summaries of data is a dot plot.

- A horizontal axis shows the range of data values.

- Then each data value is represented by a dot placed above the axis.

Fish length

Length (mm)
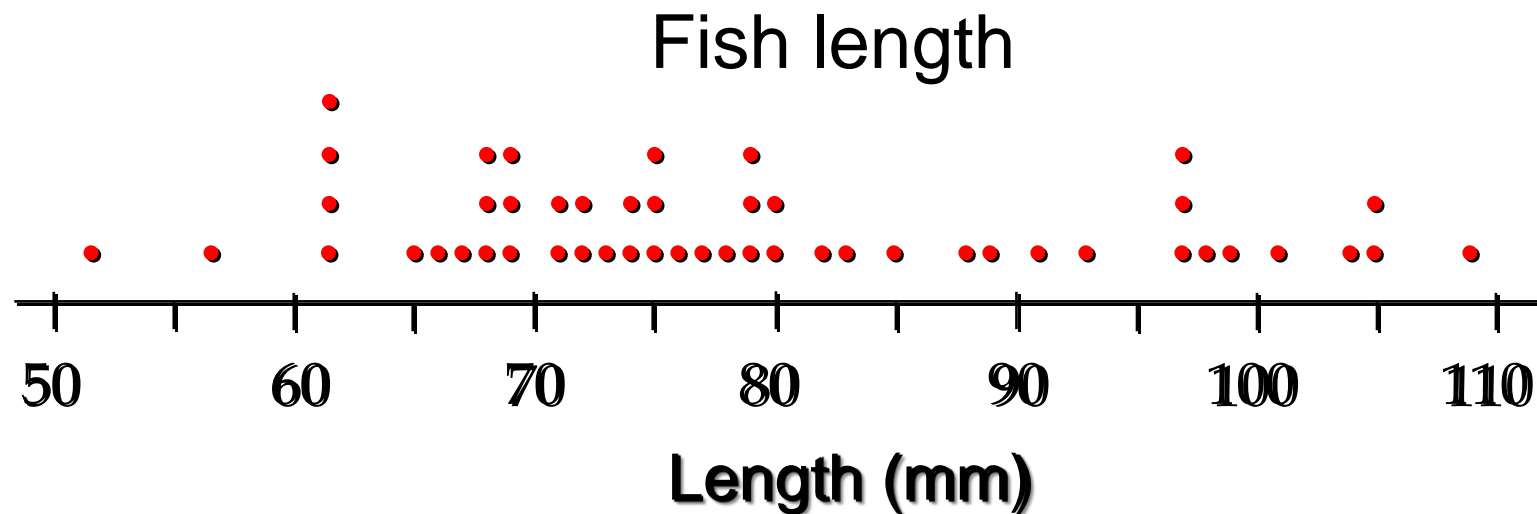
Leaf Unit = 0.1

8	6 8
9	1 4
10	2
11	0 7

Stem-and-Leaf display

- A stem-and-leaf display shows both the rank order and shape of the distribution of the data.

- It is similar to a histogram on its side, but it has the advantage of showing the actual data values.

- The first digits of each data item are arranged to the left of a vertical line.

- To the right of the vertical line we record the last digit for each item in rank order.

- Each line in the display is referred to as a stem.

- Each digit on a stem is a leaf.

Example: Leaf Unit = 0.1

If we have data with values such as

8.1   11.7   9.4   9.1   10.2   11.0   8.8

a stem-and-leaf display of these data will be

Leaf Unit = 0.1

```
 8 | 1  8
 9 | 1  4
10 | 2
11 | 0  7
```

Example: Leaf Unit = 10

If we have data with values such as

1909 1717 1874 1791 1682 1910 1838

a stem-and-leaf display of these data will be

Leaf Unit = 10

16	8
17	1 9
18	7 3
19	0 1

The 82 in 1682 is rounded down to 80 and is represented as an 8.

Descriptive statistics

- Are the scores generally high or generally low?

- Where the center of the distribution tends to be located

- Three measures of central tendency

  - Mode
  - Median
  - Mean

Mode

- The most frequently occurring score

- Report mode when using nominal scale, the most frequently occurring category

- Based on the simple frequency of each score

- If you have a rectangular distribution, do not report the mode

- Unimodal, bimodal, multimodal, antimode

Example of Mode

Value
10
5
5
1
7
2
6
7
0
4

- In this case the data have tow modes:

- 5 and 7

- Both measurements are repeated twice

x = c(10,5,5,1,7,2,6,7,0,4)
frq = table(x)
names(frq)[frq == max(frq)]

Example of Mode

Value
3
5
1
1
4
7
3
8
3

- Mode: 3

- Notice that it is possible for a data not to have any mode.

Median   median(x)

- Score at the 50th percentile

- For normal distribution the **median** is the same as the **mode**

- Arrange scores from lowest to highest, if odd number of scores the Median is the one in the middle, if even number of scores then average the two scores in the middle

- Used when have ordinal scale and when the distribution is skewed

Example of Median

Value	Value Ranked
3	0
5	1
5	2
1	3
7	4
2	5
6	5
7	6
0	7
4	7

- **Median: (4+5)/2 = 4.5**

- Notice that only the two central values are used in the computation.

- The median is not sensible to extreme values

Mean   **mean(x)**

- Score at the exact mathematical center of distribution (average)

- Used with interval and ratio scales, and when have a symmetrical and unimodal distribution

- Not accurate when distribution is skewed because it is pulled towards the tail

$$\overline{X} = \frac{\sum\limits_{i=1}^{n} x_i}{n} \approx \mu$$

Uses of the mean

- Describes scores; predict scores

- Deviation of mean gives us the error of our estimate of the score, with total error equal to zero

- Describe the population mean (μ) which is a parameter

Deviations around the mean

- The score minus the mean

- Include plus or minus sign

- Sum of deviations of the mean always equals zero $\Sigma(X-M)=0$

Range   **range(x)[2] - range(x)[1]**

- Report the maximum difference between the lowest and highest

- Semi-interquartile range used with the median: one half the distance between the scores at the 25th and 75th percentile

Boxplot

A box plot or boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles.

Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles.
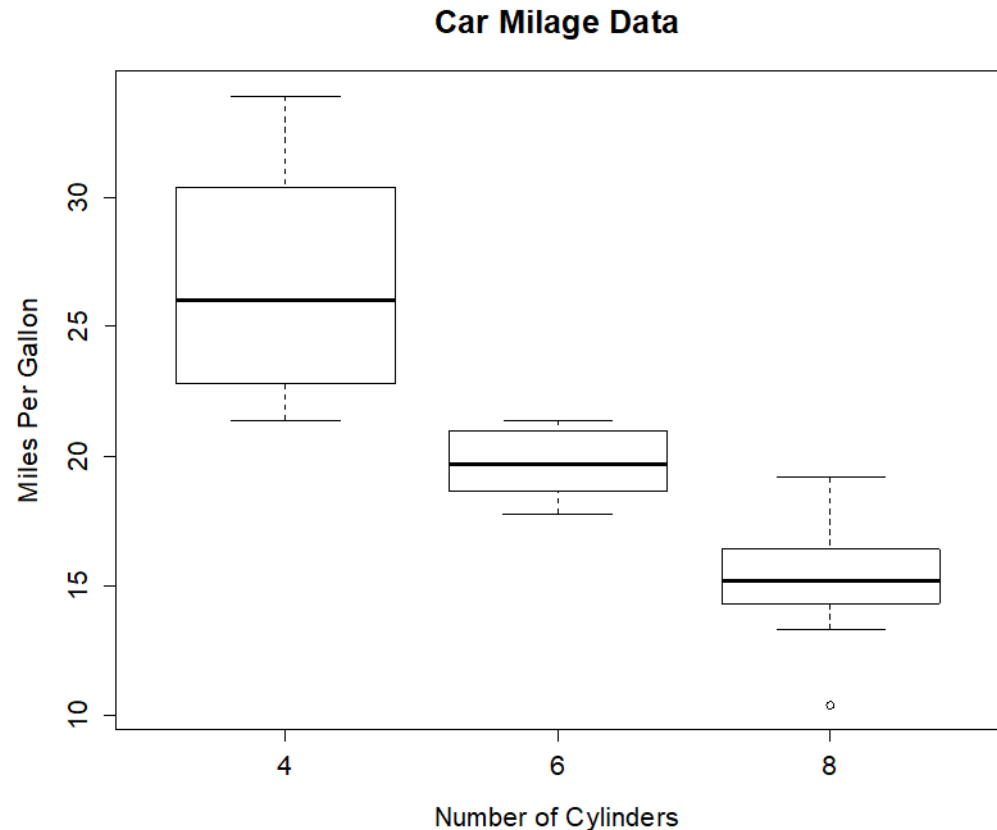
Car Milage Data

Boxplot of MPG by Car Cylinders
boxplot(mpg~cyl, data=mtcars, main="Car Milage Data",
     xlab="Number of Cylinders", ylab="Miles Per Gallon")

67

Measures of variability

- Extent to which the scores differ from each other or how spread out the scores are

- Tells us how accurately the measure of central tendency describes the distribution

- Shape of the distribution

Why do we care about variability?

- Where would you rather vacation, Kuming, where the mean temperature is 24 degrees, or Ulumiqi where the mean temperature is also 24 degrees?

◊ Kuming temperature range:

- ➢ day = 26
- ➢ night = 22

◊ Ulumiqi temperature range:

- ➢ day = 40
- ➢ night = 8

Variance   var(x)

- Uses the deviation from the mean

- Remember, the sum of the deviations always equals zero, so you have to square each of the deviations

- S^2_X= sum of squared deviations divided by the number of scores

- Provides information about the relative variability

Some limits of variance

- It isn't the average deviation

- Interpretation doesn't make sense because:

    – Number is too large

    – And it is a squared value

The standard deviation (SD)   sd(x)

- Take the square root of the variance

- S_X

- Uses the same units of measurement as the raw scores

- How much scores deviate below and above the mean

The standard deviation (SD)

- ## Standard deviation ~ the mean of deviations from the mean

σ (lowercase sigma) is the population standard deviation.

S the sample standard deviation

$\hat{s}$ (s-hat) is the sample estimate of σ

The deviation (definitional) formula for the population standard deviation

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

- The larger the standard deviation the more variability there is in the scores

- The standard deviation is somewhat less sensitive to extreme outliers than the range (as N increases)

The deviation (definitional) formula for the sample standard deviation

$$S = \sqrt{\frac{\sum\left(X_i - \overline{X}\right)^2}{N}}$$

- What's the difference between this formula and the population standard deviation?

- In the first case, all the *x*s represent the entire population. In the second case, the *X*s represent a sample.

Standard deviation: example

X	$\left(X - \overline{X}\right)$	$\left(X - \overline{X}\right)^2$
21	-5.8	33.64
25	-1.8	3.24
24	-2.8	7.84
30	3.2	10.24
34	7.2	51.84
Mean **26.8**	**0**	**21.36**

$$S = \sqrt{\frac{106.8}{5}} = \sqrt{21.36} = 4.62$$

Calculating S using the raw-score formula

$$S = \sqrt{\frac{\sum X^2 - \frac{\left(\sum X\right)^2}{N}}{N}}$$

To calculate ΣX^2 you square all the scores first and then sum them

To calculate $(\Sigma X)^2$ you sum all the scores first and then square them

Population and sample variance and standard deviation

- When we have data from the entire population we use μ (not x bar) to compute σ_X using the same formula

- Variance and standard deviations of the sample are biased estimates of the population

Estimating the population standard deviation from a sample

$$S = \sqrt{\frac{\sum(X_i - \overline{X})^2}{N}}$$

- S, the sample standard deviation, is usually a little smaller than the population standard deviation. Why?

- The sample mean minimizes the sum of squared deviations (SS). Therefore, if the sample mean differs at all from the population mean, then the SS from the sample will be an understimate of the SS from the population

- Therefore, statisticians alter the formula of the sample standard deviation by subtracting 1 from N

Formulas for s-hat (estimated)

Definitional
formula:

$$\hat{s} = \sqrt{\frac{\sum\left(X - \bar{X}\right)^2}{N - 1}}$$

Raw-score
formula:

$$\hat{s} = \sqrt{\frac{\sum X^2 - \dfrac{\left(\sum X\right)^2}{N}}{N - 1}}$$
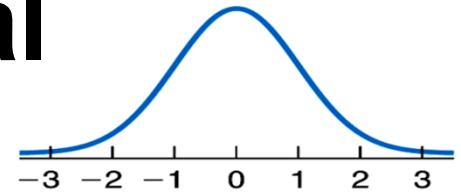
The estimated variance

The standard deviation squares

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} \qquad \hat{s}^2 = \frac{\sum (X - \overline{X})^2}{n - 1}$$

The variance is not a very useful descriptive statistic, but it is very important value you will use in other techniques (e.g., the analysis of variance)

For a standard normal distribution

- Sample mean is a good estimate of population mean

- The estimate of the population variance and standard deviation tells us how spread out the scores are

- 68% of the scores are within +1 and −1 S_X

Standard error

The standard error of a sample of sample size n is the sample's standard deviation divided by $\sqrt{n}$.

It therefore estimates the **standard deviation of the sample mean** based on the population mean.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

se = function(x) sqrt(var(x) / length(x))
se(x)

83

Coefficient of variation

The coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution. It is defined as the ratio of the standard deviation σ to the mean μ :
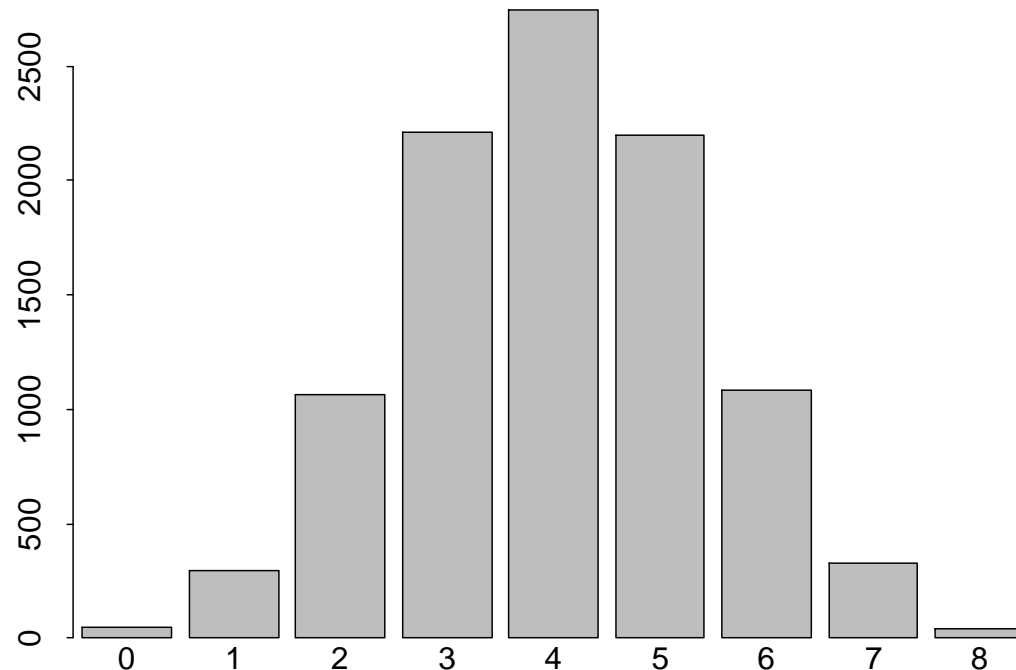
$$CV = \frac{\sigma}{\mu} \times 100$$

CV = function(x) sd(x) / mean(x)
CV(x)

Skewness
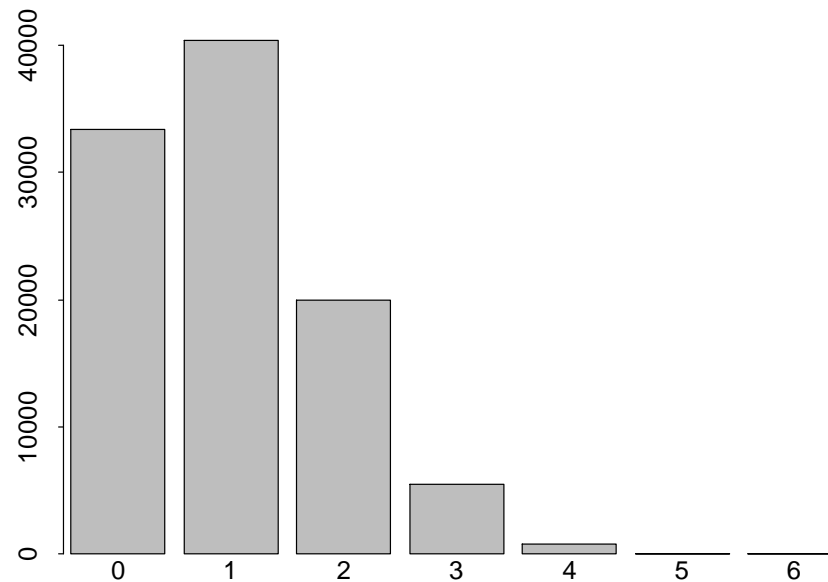## **Symmetrical distribution**

• Symmetric

   – Left tail is the mirror image of the right tail

   – Examples: heights and weights of people

Skewness
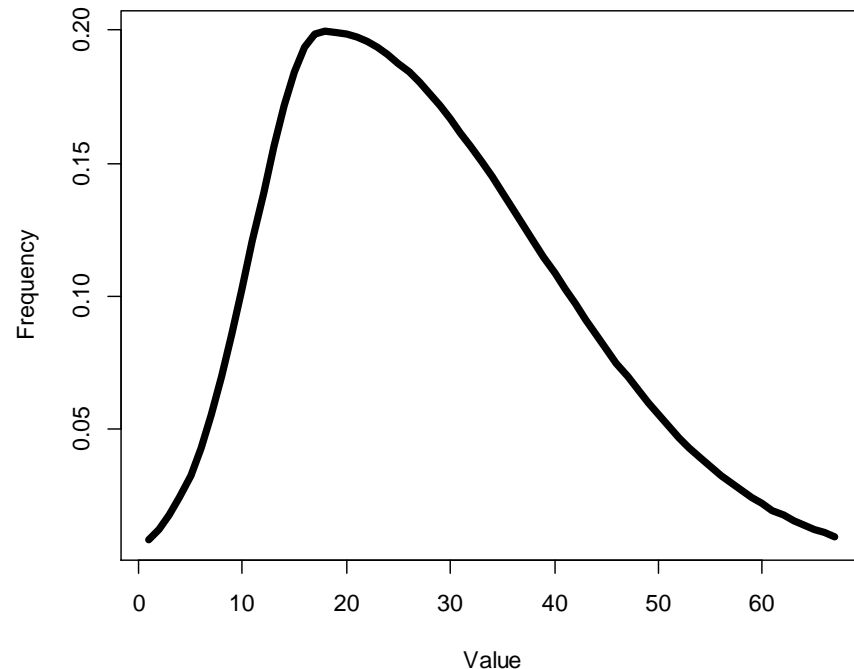## **Asymmetrical distribution**

- Moderately skewed right
  - A longer tail to the right
  - Example: income

Skewness
Asymmetrical distribution

- ## Skewed right
  - ### – A longer tail to the right

- ## Income
- ## Populations of countries

curve = dnorm(seq(-5, 5, by = .1), 0, 2)
curve = curve[c(seq(1, 50, by = 3), 51:100)]
plot(curve, xlab = 'Value', ylab = 'Frequency', type='l')

87

Skewness

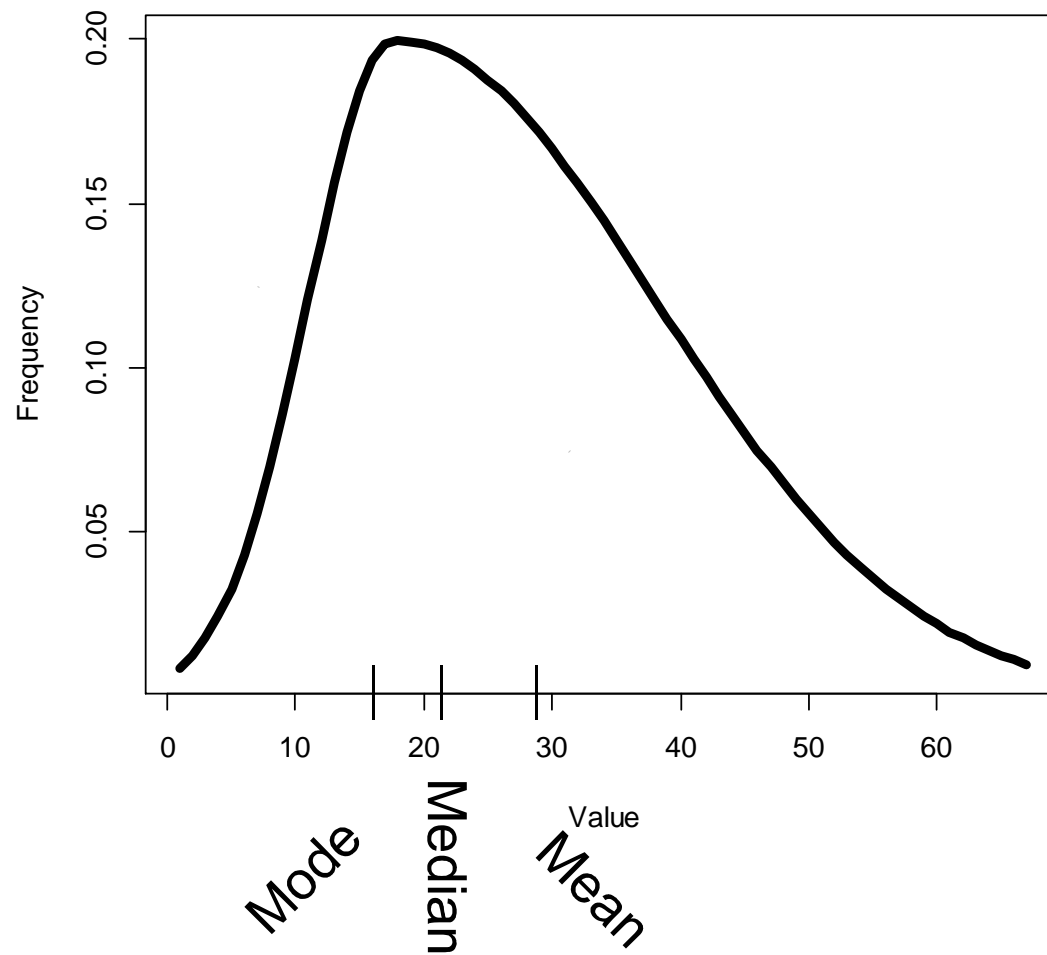A Measure of skewness based on the 3rd moment about the Mean

$$skewness = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{(N-1)s^3}$$

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sum_{i=1}^{n}(x_i - \overline{x})^{3/2}} \times \frac{(n-1)^{3/2}}{n-2} \approx$$

$$(mean - \mathrm{mod}\,e)/s \approx 3 \times (mean - median)/s$$

skewness = sum((x-mean(x))^3/sqrt(var(x))^3) / (length(x)-1)

Skewness

Kurtosis

- Measures of Kurtosis

  - Kurtosis is a measure of the flatness or peakedness of a Distribution

    - Normal Kurtosis - Mesokurtic

    - Flat Kurtosis - Platokurtic

    - Peaked Kurtosis - Leptokurtic

  - A Measure of Kurtosis based on the 4th moment about the Mean

Kurtosis

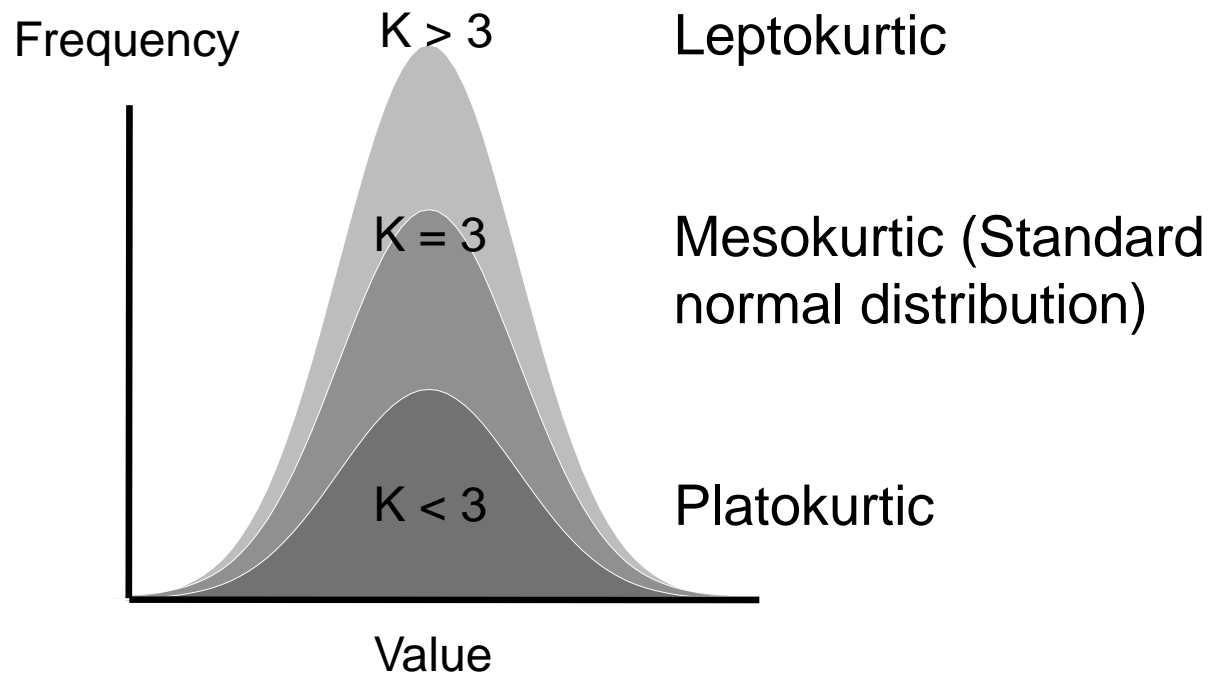$$\text{kurtosis} = \frac{\sum\limits_{i=1}^{N}(\chi_i - \overline{\chi})^4}{(N-1)s^4} - 3$$

If less than 0 = Platokurtic
More than 0 = Leptokurtic
If 0 then = Mesokurtic

kurtosis = sum((x-mean(x))^4/var(x)^2) /(length(x)-1) -3

Kurtosis

Frequency

K > 3   Leptokurtic

K = 3   Mesokurtic (Standard normal distribution)

K < 3   Platokurtic

Value

Describing data

	Statistic (mean based)	Statistic (non-mean based)
Center	Mean	Mode, median
Spread	Variance, SD (standard deviation), SE, CV	Range, Interquartile range
Skew	Skewness	--
Peaked	Kurtosis	--

R code

```r
x = rnorm(100)
mean(x)
sd(x)
var(x)
min(x)
max(x)
median(x)
range(x)
quantile(x)
summary(x)
```

Assignment

Be familiar with the following terms:
*   Probability density function (PDF)
*   Deviation
*   Variance
*   Standard deviation
*   Standard error
*   Range
*   Mode
*   Quantile
*   Coefficient of variation

Download and install R on your laptop
Plot histograms using
hist(rnorm(100), nclass=6)