

UNIVERSITY OF LIÈGE

Université
de Liège



Statistics – Theory

PROF. DR. DR. K. VAN STEEN

July 2012

Contents

1	Introduction	2
1.1	The Birth of Probability and Statistics	2
1.2	Statistical Modeling under Uncertainties: From Data to Knowledge	3
1.3	Course Outline	4
1.4	Motivating Examples	5
2	Samples, random sampling and sample geometry	8
2.1	Introduction: No statistics without Data!	8
2.2	Populations and samples	9
2.3	Sampling schemes	12
2.3.1	Non-probability sampling	12
2.3.2	Probability sampling	13
2.4	Sampling Challenges	14
2.5	Distribution of a sample	15
2.6	Statistics	15
2.7	Sample Geometry	18
2.7.1	The geometry of the sample	18
2.7.1.1	The mean	18
2.7.1.2	Variance and correlation	19
2.7.2	Expected values of the sample mean and the covariance matrix	20
2.7.3	Generalized variance	21
2.8	Resampling	27
2.9	The Importance of Study Design	28
3	Exploratory Data Analysis	30
3.1	Typical data format and the types of EDA	30
3.2	Univariate non-graphical EDA	31
3.2.1	Categorical data	32
3.2.2	Characteristics of quantitative data	32
3.2.3	Central tendency	33
3.2.4	Spread	35
3.2.5	Skewness and kurtosis	37
3.3	Univariate graphical EDA	38
3.3.1	Histograms	38
3.3.2	Stem-and-leaf plots	44
3.3.3	Boxplots	44
3.3.4	Quantile-normal plots	47

3.4	Multivariate non-graphical EDA	50
3.4.1	Cross-tabulation	51
3.4.2	Correlation for categorical data	52
3.4.3	Univariate statistics by category	53
3.4.4	Correlation and covariance	53
3.4.5	Covariance and correlation matrices	55
3.5	Multivariate graphical EDA	56
3.5.1	Univariate graphs by category	56
3.5.2	Scatterplots	56
3.6	A note on degrees of freedom	58
4	Estimation	59
4.1	Introduction	59
4.2	Statistical philosophies	60
4.3	The frequentist approach to estimation	62
4.4	Estimation by the method of moments	63
4.4.1	Traditional methods of moments	64
4.4.2	Generalized methods of moments	67
4.5	Properties of an estimator	67
4.5.1	Unbiasedness	67
4.5.2	Trading off Bias and Variance	67
4.5.2.1	Mean-Squared Error	67
4.5.2.2	Minimum-Variance Unbiased	69
4.5.3	Efficiency	73
4.5.4	Consistency	73
4.5.5	Loss and Risk Functions	74
4.6	Sufficiency	75
4.7	The Likelihood approach	77
4.7.1	Maximum likelihood estimation	77
4.7.2	Properties of MLE	79
4.7.3	The Invariance principle	80
4.8	Properties of Sample Mean and Sample Variance	90
4.9	Multi-parameter Estimation	92
4.10	Newton-Raphson optimization	95
4.10.1	One-paramter scenario	95
4.10.2	Two-paramter scenario	96
4.10.3	Initial values	96
4.10.4	Fisher's method of scoring	97
4.10.5	The method of profiling	97
4.10.6	Reparameterization	98
4.10.7	The step-halving scheme	99
4.11	Bayesian estimation	99
4.11.1	Bayes' theorem for random variables	99
4.11.2	Post 'is' prior \times likelihood	100
5	Confidence intervals	103
5.1	Introduction	103

5.2	Exact confidence intervals	106
5.3	Pivotal quantities for use with normal data	110
5.4	Approximate confidence intervals	114
5.5	Bootstrap confidence intervals	115
5.5.1	The empirical cumulative distribution function	115
6	The Theory of hypothesis testing	120
6.1	Introduction	120
6.2	Terminology and notation	122
6.2.1	Hypotheses	122
6.2.2	Tests of hypotheses	122
6.2.3	Size and power of tests	123
6.3	Examples	123
6.4	One-sided and two-sided Tests	126
6.4.1	Case (a): Alternative is one-sided	127
6.4.2	Case (b): Two-sided Alternative	128
6.4.3	Two approaches to hypothesis testing	129
6.5	Two-sample problems	131
6.6	Connection between hypothesis testing and CI's	133
6.7	Summary	134
6.8	Non-parametric hypothesis testing	135
6.8.1	Kolmogorov-Smirnov (KS)	136
6.8.2	Asymptotic distribution	137
6.8.3	Bootstrap Hypothesis tests	138
6.9	The general testing problem	140
6.10	Hypothesis testing for normal data	141
6.11	Generally applicable test procedures	146
6.12	The Neyman-Pearson lemma	148
6.13	Goodness of fit tests	151
6.14	The χ^2 test for contingency tables	153
7	Chi-square Distribution	155
7.1	Distribution of S^2	155
7.2	Chi-Square Distribution	157
7.3	Independence of \bar{X} and S^2	162
7.4	Confidence intervals for σ^2	162
7.5	Testing hypotheses about σ^2	164
7.6	χ^2 and $\text{Inv-}\chi^2$ distributions in Bayesian inference	166
7.6.1	Non-informative priors	166
7.7	The posterior distribution of the Normal variance	167
7.7.1	Inverse Chi-squared distribution	168
7.8	Relationship between χ^2_ν and $\text{Inv-}\chi^2_\nu$	168
7.8.1	Gamma and Inverse Gamma	168
7.8.2	Chi-squared and Inverse Chi-squared	168
7.8.3	Simulating Inverse Gamma and $\text{Inv-}\chi^2$ random variables	169
8	Analysis of Count Data	171

8.1	Introduction	171
8.2	Goodness-of-Fit tests	171
8.3	Contingency tables	179
	8.3.1 Method	179
8.4	Special Case: 2×2 Contingency table	182
8.5	Fisher's exact test	184
8.6	Parametric Bootstrap- X^2	186
9	Simple Linear Regression	190
9.1	Introduction	190
9.2	Estimation of α and β	191
9.3	Estimation of σ^2	196
9.4	Inference about $\hat{\alpha}$, β and μ_Y	197
9.5	Correlation	202
10	Multiple Regression	205
10.1	The model	205
10.2	Least squares estimator of β : a vector differentiation approach	206
10.3	Least squares estimator of β : an algebraic approach	207
10.4	Elementary properties of $\hat{\beta}$	209
10.5	Estimation of σ^2	211
10.6	Distribution theory for $\hat{\beta}$ and MSE	213
10.7	A fundamental decomposition for the total sum of squares	216
10.8	F -test	217
10.9	Inferences about regression parameters	222
	10.9.1 Confidence limits	222
	10.9.2 Test	223
	10.9.3 Confidence region	223
10.10	Inferences about mean responses	223
	10.10.1 Interval estimation of $E(Y_h)$	223
	10.10.2 Confidence region for regression surface	224
	10.10.3 Simultaneous confidence intervals for several mean responses	225
10.11	Predictions of new observations	225
	10.11.1 Prediction of a new observation	225
	10.11.2 Prediction of g new observations	226
	References	I

This booklet is based on

- Chapters 6.1 - 6.2 of [12] (which in part lead to Chapter 2, in particular, sections 2.1, 2.2)
- [4] for Section 2.3
- [3] for Sections 2.4 and 2.9
- [17] and [2] for Section 2.8
- Chapters 1–4 of [8] (no bib citation in the text indicates that the associated text is, by default, coming from [8]),
- Chapter 3 of [10] (which leads to Chapter 2),
- Chapter 1 of [9] (which leads to Chapter 10 and to Appendix ??),
- Chapters 1, 2 (except section 2.8), 3, 6 and 8 from [5], that lead to the following sections: 4.1, 2.6, 4.4, 4.8, 4.7.2, 4.11, 5.5, 5.1, 6.2, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8, Chapters 7, 8, 9, and those where this is explicitly mentioned.

CHAPTER

1

INTRODUCTION

1.1 The Birth of Probability and Statistics

The original idea of “statistics” was the collection of information about and for the “state”. The word statistics derives directly, not from any classical Greek or Latin roots, but from the Italian word for state.

The birth of statistics occurred in mid-17th century. A commoner, named John Graunt, who was a native of London, began reviewing a weekly church publication issued by the local parish clerk that listed the number of births, christenings, and deaths in each parish. These so called Bills of Mortality also listed the causes of death. Graunt who was a shopkeeper organized this data in the form we call descriptive statistics, which was published as Natural and Political Observations Made upon the Bills of Mortality. Shortly thereafter he was elected as a member of Royal Society. Thus, statistics has to borrow some concepts from sociology, such as the concept of Population. It has been argued that since statistics usually involves the study of human behavior, it cannot claim the precision of the physical sciences.

Probability has much longer history. Probability is derived from the verb to probe meaning to “find out” what is not too easily accessible or understandable. The word “proof” has the same origin that provides necessary details to understand what is claimed to be true. Probability originated from the study of games of chance and gambling during the 16th century. Probability theory was a branch of mathematics studied by Blaise Pascal and Pierre de Fermat in the seventeenth century. Currently in 21st century, probabilistic modeling is used to control the flow of traffic through a highway system, a telephone interchange, or a computer processor; find the genetic makeup of individuals or populations; quality control; insurance; investment; and other sectors of business and industry.

New and ever growing diverse fields of human activities are using statistics; however, it seems that this field itself remains obscure to the public. Professor Bradley Efron expressed this fact nicely: During the 20th Century statistical thinking and methodology have become the scientific

framework for literally dozens of fields including education, agriculture, economics, biology, and medicine, and with increasing influence recently on the hard sciences such as astronomy, geology, and physics. In other words, we have grown from a small obscure field into a big obscure field.

1.2 Statistical Modeling under Uncertainties: From Data to Knowledge

In this diverse world of ours, no two things are exactly the same. A statistician is interested in both the differences and the similarities; i.e., both departures and patterns.

The actuarial tables published by insurance companies reflect their statistical analysis of the average life expectancy of men and women at any given age. From these numbers, the insurance companies then calculate the appropriate premiums for a particular individual to purchase a given amount of insurance.

Exploratory analysis of data makes use of numerical and graphical techniques to study patterns and departures from patterns. The widely used descriptive statistical techniques are: Frequency Distribution; Histograms; Boxplot; Scattergrams; and diagnostic plots.

In examining distribution of data, you should be able to detect important characteristics, such as shape, location, variability, and unusual values. From careful observations of patterns in data, you can generate conjectures about relationships among variables.

Data must be collected according to a well-developed plan if valid information on a conjecture is to be obtained. The plan must identify important variables related to the conjecture, and specify how they are to be measured. From the data collection plan, a statistical model can be formulated from which inferences can be drawn.

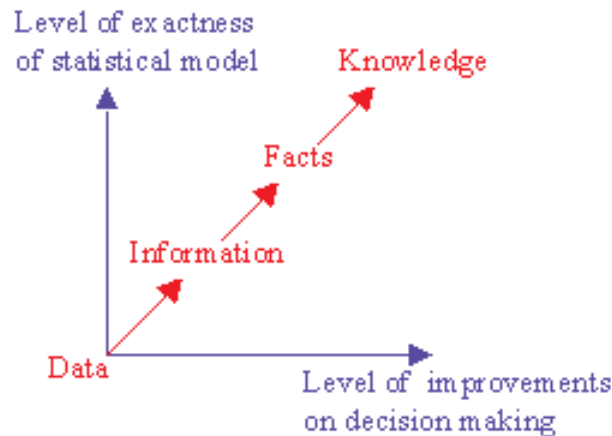
Frequently, for example the marketing managers or clinical trials investigators are faced with the question, What Sample Size Do I Need? This is an important and common statistical decision, which should be given due consideration, since an inadequate sample size invariably leads to wasted resources.

The notion of how one variable may be associated with another permeates almost all of statistics, from simple comparisons of proportions through linear regression. The difference between association and causation must accompany this conceptual development. As an example of statistical modeling with managerial implications, such as “what-if” analysis, consider regression analysis. Regression analysis is a powerful technique for studying relationship between dependent variables (i.e., output, performance measure) and independent variables (i.e., inputs, factors, decision variables). Summarizing relationships among the variables by the most appropriate equation (i.e., modeling) allows us to predict or identify the most influential factors and study their impacts on the output for any changes in their current values.

Statistical models are currently used in various fields of business and science. However, the terminology differs from field to field. For example, the fitting of models to data, called calibration, history matching, and data assimilation, are all synonymous with parameter estimation.

Knowledge is what we know well. Information is the communication of knowledge. In every knowledge exchange, there is a sender and a receiver. The sender makes common what is private, does the informing, the communicating. Information can be classified as explicit and tacit forms. The explicit information can be explained in structured form, while tacit information is inconsistent and fuzzy to explain. Data are only crude information and not

knowledge by themselves. The sequence from data to knowledge is: from Data to Information, from Information to Facts, and finally, from Facts to Knowledge. Data becomes information, when it becomes relevant to your decision problem. Information becomes fact, when the data can support it. Facts are what the data reveals. However the decisive instrumental (i.e., applied) knowledge is expressed together with some statistical degree of confidence. Statistical inference aims at determining whether any statistical significance can be attached to that results after due allowance is made for any random variation as a source of error. Intelligent and critical inferences cannot be made by those who do not understand the purpose, the conditions, and applicability of the various techniques for judging significance.



1.3 Course Outline

Four main topics will be covered :

Sampling Statistical inference, samples, populations, the role of probability. Sampling procedures, data collection methods, types of statistical studies and designs. Measures of location and variability. Types of data. Statistical modeling techniques. Exploratory data analysis: scientific inspection and graphical diagnostics, examining univariate distributions, examining relationships, regression diagnostics, multivariate data displays, data mining.

Estimation Unbiasedness, mean square error, consistency, relative efficiency, sufficiency, minimum variance. Fisher's information for a function of a parameter, Cramér-Rao lower bound, efficiency. Fitting standard distributions to discrete and continuous data. Method of moments. Maximum likelihood estimation: finding estimators analytically and numerically, invariance, censored data. Random intervals and sets. Use of pivotal quantities. Relationship between tests and confidence intervals. Use of asymptotic results.

Hypothesis testing Simple and composite hypotheses, types of error, power, operating characteristic curves, p -value. Neuman-Pearson method. Generalised likelihood ratio test. Use of asymptotic results to construct tests. Central limit theorem, asymptotic distributions of maximum likelihood estimator and generalised ratio test statistic. Sample size calculations

Statistical modeling

Analysis of Count Data Quantifying associations. Comparing proportions, confidence intervals for Relative Risks and Odds Ratios. Type of chi-squared tests. Logit transformations and logistic regression. Association and causation.

Regression Analysis Quantifying linear relationships. Model components, assumptions and assumption checking. Parameter estimation techniques: least-squares estimation, robust estimation, Kalman filters. Interpretation of model parameters. Centering. Hypothesis testing and confidence intervals of regression coefficients. Multiple covariates and confounding.

1.4 Motivating Examples

Example 1.1 (Radioactive decay).

Let X denote the number of particles that will be emitted from a radioactive source in the next one minute period. We know that X will turn out to be equal to one of the non-negative integers but, apart from that, we know nothing about which of the possible values are more or less likely to occur. The quantity X is said to be a random variable.

Suppose we are told that the random variable X has a Poisson distribution with parameter $\theta = 2$. Then, if x is some non-negative integer, we know that the probability that the random variable X takes the value x is given by the formula

$$P(X = x) = \frac{\theta^x \exp(-\theta)}{x!}$$

where $\theta = 2$. So, for instance, the probability that X takes the value $x = 4$ is

$$P(X = 4) = \frac{2^4 \exp(-2)}{4!} = 0.0902 .$$

We have here a probability model for the random variable X . Note that we are using upper case letters for random variables and lower case letters for the values taken by random variables. We shall persist with this convention throughout the course.

Suppose we are told that the random variable X has a Poisson distribution with parameter θ where θ is some unspecified positive number. Then, if x is some non-negative integer, we know that the probability that the random variable X takes the value x is given by the formula

$$P(X = x|\theta) = \frac{\theta^x \exp(-\theta)}{x!}, \quad (1.4.1)$$

for $\theta \in \mathbb{R}^+$. However, we cannot calculate probabilities such as the probability that X takes the value $x = 4$ without knowing the value of θ .

Suppose that, in order to learn something about the value of θ , we decide to measure the value of X for each of the next 5 one minute time periods. Let us use the notation X_1 to denote the number of particles emitted in the first period, X_2 to denote the number emitted in the second period and so forth. We shall end up with data consisting of a random vector $\mathbf{X} = (X_1, X_2, \dots, X_5)$. Consider $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (2, 1, 0, 3, 4)$. Then \mathbf{x} is a possible value for the random vector \mathbf{X} . We know that the probability that X_1 takes the value $x_1 = 2$ is given by the formula

$$P(X = 2|\theta) = \frac{\theta^2 \exp(-\theta)}{2!}$$

and similarly that the probability that X_2 takes the value $x_2 = 1$ is given by

$$P(X = 1|\theta) = \frac{\theta \exp(-\theta)}{1!}$$

and so on. However, what about the probability that \mathbf{X} takes the value \mathbf{x} ? In order for this probability to be specified we need to know something about the joint distribution of the random variables X_1, X_2, \dots, X_5 . A simple assumption to make is that the random variables X_1, X_2, \dots, X_5 are mutually independent. (Note that this assumption may not be correct since X_2 may tend to be more similar to X_1 than it would be to X_5 .) However, with this assumption we can say that the probability that \mathbf{X} takes the value \mathbf{x} is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= \prod_{i=1}^5 \frac{\theta^{x_i} \exp(-\theta)}{x_i!}, \\ &= \frac{\theta^2 \exp(-\theta)}{2!} \times \frac{\theta^1 \exp(-\theta)}{1!} \times \frac{\theta^0 \exp(-\theta)}{0!} \\ &\quad \times \frac{\theta^3 \exp(-\theta)}{3!} \times \frac{\theta^4 \exp(-\theta)}{4!}, \\ &= \frac{\theta^{10} \exp(-5\theta)}{288}. \end{aligned}$$

In general, if $\mathbf{X} = (x_1, x_2, x_3, x_4, x_5)$ is any vector of 5 non-negative integers, then the probability that \mathbf{X} takes the value \mathbf{x} is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= \prod_{i=1}^5 \frac{\theta^{x_i} \exp(-\theta)}{x_i!}, \\ &= \frac{\theta^{\sum_{i=1}^5 x_i} \exp(-5\theta)}{\prod_{i=1}^5 x_i!}. \end{aligned}$$

We have here a probability model for the random vector \mathbf{X} .

Our plan is to use the value \mathbf{x} of \mathbf{X} that we actually observe to learn something about the value of θ . The why's, ways and means to accomplish such a task will be the core business of this course.

Example 1.2 (Tuberculosis).

Suppose we are going to examine n people and record a value 1 for people who have been exposed to the tuberculosis virus and a value 0 for people who have not been so exposed. The data will consist of a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ where $X_i = 1$ if the i th person has been exposed to the TB virus and $X_i = 0$ otherwise.

A Bernoulli random variable X has probability mass function

$$P(X = x|\theta) = \theta^x(1 - \theta)^{1-x}, \quad (1.4.2)$$

for $x = 0, 1$ and $\theta \in (0, 1)$. A possible model would be to assume that X_1, X_2, \dots, X_n behave like n independent Bernoulli random variables each of which has the same (unknown) probability θ of taking the value 1.

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be a particular vector of zeros and ones. Then the model implies that the probability that the random vector \mathbf{X} takes the value \mathbf{x} is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|\theta) &= \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}. \end{aligned}$$

Once again our plan is to use the value \mathbf{x} of \mathbf{X} that we actually observe to learn something about the value of θ .

Example 1.3 (Viagra).

A chemical compound Y is used in the manufacture of Viagra. Suppose that we are going to measure the micrograms of Y in a sample of n pills. The data will consist of a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ where X_i is the chemical content of Y for the i th pill.

A possible model would be to assume that X_1, X_2, \dots, X_n behave like n independent random variables each having a $\mathcal{N}(\mu, \sigma^2)$ density with unknown mean parameter $\mu \in \mathbb{R}$, (really, here $\mu \in \mathbb{R}^+$) and known variance parameter $\sigma^2 < \infty$. Each X_i has density

$$f_{X_i}(x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}.$$

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be a particular vector of real numbers. Then the model implies the joint density

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Once again our plan is to use the value \mathbf{x} of \mathbf{X} that we actually observe to learn something about the value of μ .

Example 1.4 (Blood pressure).

We wish to test a new device for measuring blood pressure. We are going to try it out on n people and record the difference between the value returned by the device and the true value as recorded by standard techniques. The data will consist of a random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ where X_i is the difference for the i th person. A possible model would be to assume that X_1, X_2, \dots, X_n behave like n independent random variables each having a $\mathcal{N}(0, \sigma^2)$ density where σ^2 is some unknown positive real number. Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be a particular vector of real numbers. Then the model implies that the probability that the random vector \mathbf{X} takes the value \mathbf{x} is given by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\} \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left\{-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right\}. \end{aligned}$$

Once again our plan is to use the value \mathbf{x} of \mathbf{X} that we actually observe to learn something about the value of σ^2 . Knowledge of σ is useful since it allows us to make statements such as that 95% of errors will be less than $1.96 \times \sigma$ in magnitude. \square

CHAPTER

2

SAMPLES, RANDOM SAMPLING AND SAMPLE GEOMETRY

2.1 Introduction: No statistics without Data!

Progress in science is often ascribed to experimentation. The research worker performs an experiment and obtains some data. On the basis of data, certain conclusions are drawn. The conclusions usually go beyond the materials and operations of the particular experiment. In other words, the scientist may generalize from a particular experiment to the class of all similar experiments. This sort of extension from the particular to the general is called *inductive* inference. It is one way in which new knowledge is found.

Inductive inference is well known to be a hazardous process. In fact, it is a theorem of logic that in inductive inference uncertainty is present. One simply cannot make absolutely certain generalizations. However, uncertain inferences can be made, and the degree of uncertainty can be measured if the experiment has been performed in accordance with certain principles. One function of statistics is the provision of techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences. Uncertainty is measured in terms of probability, and that is the reason you have devoted so much time to the theory of probability (last year).

As an illustration, suppose we have a storage bin that contains 10 million flower seeds which we know will each produce either white or red flowers. The information which we want is: How many of these 10 million seeds will produce white flowers? The only way in which we can be absolutely sure that this question is answered correctly is to plant every seed and observe the number producing white flowers. This is not feasible since we want to sell the seeds! Even if we did not want to sell the seeds, we would prefer to obtain an answer without expending so much effort. Without planting each seed and observing the color of flower that each produces

we cannot be certain of the number of seeds producing white flowers. However, another thought which occurs is: Can we plant a few of the seeds and, on the basis of the colors of these few flowers make a statement as to how many of the 10 million flower seeds will produce white flowers? The answer is that we cannot make an exact prediction as to how many white flowers the seeds will produce, but we can make a probabilistic statement if we select the few seeds in a certain fashion.

There is another sort of inference though, namely *deductive inference*. As an illustration of deductive inference, consider the following two statements. If we accept these two statements, then we are forced to the conclusion that one of the angles of triangle T equals 90 degrees. This example of deductive inference clearly shows that it can be described as a method of deriving information from accepted facts [statements (i) and (ii)].

- (i) One of the interior angles of each right triangle equals 90 degrees
- (ii) The triangle T is a right triangle

While conclusions which are reached by inductive inference are only probable, those reached by deductive inference are conclusive. While deductive inference is extremely important, much of the new knowledge in the real world comes about by the process of inductive inference. In the science of mathematics, deductive inference is used to prove theorems, while in the empirical sciences inductive inference is used to find new knowledge.

In general:

Inference Inference studies the way in which data we observe should influence our beliefs about and practices in the real world.

Statistical inference Statistical inference considers how inference should proceed when the data is subject to random fluctuation.

The concept of probability is used to describe the random mechanism which gave rise to the data. This involves the use of probability models.

The incentive for contemplating a probability model is that through it we may achieve an economy of thought in the description of events enabling us to enunciate laws and relations of more than immediate validity and relevance. A probability model is usually completely specified apart from the values of a few unknown quantities called parameters. We then try to discover to what extent the data can inform us about the values of the parameters.

Statistical inference assumes that the data is given and that the probability model is a correct description of the random mechanism which generated the data.

2.2 Populations and samples

We have seen that a central problem in discovering new knowledge in the real world consists of observing a few of the elements under discussion and on the basis of these few we make a statement about the totality of elements. We shall now investigate this procedure in more detail.

Definition 2.1 (Target population). The totality of elements which are under discussion and about which information is desired will be called the target population.

The target population in the example of the flower seeds before is formed by the 10 million seeds. In general, the important thing is that the target population must be capable of being quite well defined; it may be real or hypothetical. Since we want to make inferences regarding the entire target population on the basis of only a selective set of elements, the question arises as to how the sample of the population should be selected. We have stated before that we could make probabilistic statements about the population if the sample is selected “in a certain fashion”. Of particular importance is the case of a simple random sample, usually called a random sample, which can be defined for any population which has a density. That is, we assume that each element in our population has some numerical value associated with it and that the distribution of these numerical values is given by a density. For such a population we technically define a random sample as follows (see also 2.3):

Definition 2.2 (Random sample). Let the random variables X_1, X_2, \dots, X_n have a joint density f_{X_1, X_2, \dots, X_n} that factors as

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n),$$

where $f(\cdot)$ is the (common) density of each X_i . Then X_1, X_2, \dots, X_n is defined to be a random sample of size n from a population with density $f(\cdot)$

Note that a random variable is technically defined as:

Definition 2.3 (Random variable). For a given probability space $(\Omega, \mathcal{A}, P(\cdot))$, a random variable, denoted by X or $X(\cdot)$, is a function with domain Ω and counterdomain the real line. The function X must be such that the set defined by $\{\omega : X(\omega) \leq r\}$ belongs to \mathcal{A} for every real number r .

In the above definition, Ω represents the sample space, this is the totality of possible outcomes of a conceptual experiment of interest and \mathcal{A} is a set of subsets of Ω , called the event space. The event space \mathcal{A} is assumed to be a (Boolean) algebra (explaining the use of the symbol \mathcal{A} , meaning that the collection of events \mathcal{A} satisfies the following properties:

- (i) The *universum* $\Omega \in \mathcal{A}$
- (ii) If $A \in \mathcal{A}$ then $\Omega - A = \bar{A} \in \mathcal{A}$
- (iii) If A_1 and $A_2 \in \mathcal{A}$, then $A_1 \cup A_2 \in \mathcal{A}$

The probability function $P(\cdot)$ is a set function having domain \mathcal{A} and counterdomain the interval $[0, 1]$. Probability functions allow to compute the probability of certain ‘events’ and satisfy the defining properties or axioms:

- (i) $P(A) \geq 0$ for all $A \in \mathcal{A}$
- (ii) $P(\Omega) = 1$
- (iii) If A_1, A_2, \dots is a sequence of mutually exclusive events in \mathcal{A} (i.e., $A_i \cap A - j \neq \phi$ for $i \neq j; i, j = 1, 2, \dots$) and if $A_1, A_2, \dots = \bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Definition 2.4 (Cumulative distribution function). Any function $F(\cdot)$ with domain the real line and counterdomain $[0, 1]$ satisfying the following 3 properties is defined to be a *cumulative distribution function*:

- (i) $F(-\infty) \equiv \lim_{x \rightarrow -\infty} F(x) = 0$ and $F(\infty) \equiv \lim_{x \rightarrow \infty} F(x) = 1$
- (ii) $F(\cdot)$ is a monotone, nondecreasing function ($F(a) \leq F(b)$ for any $a < b$)
- (iii) $F(\cdot)$ is continuous from the right; that is $\lim_{0 < h \rightarrow 0} F(x + h) = F(x)$

So the cumulative distribution function describes the distribution of values of a random variable. For instance, when X is the random variable of interest, the associated cumulative distribution function is sometimes denoted as $F_X(\cdot)$ (F_X) (instead of simply $F(\cdot)$ or F), to avoid confusion. For two distinct classes of random variables, the distribution of values can be described more simply by using *density functions*. These two classes are distinguished by the words ‘discrete’ (the range of the random variable is countable) and ‘continuous’ (the range of the random variable encompasses a continuum of values) and the associated density functions are respectively called *discrete density function* and *probability density function*. A cumulative distribution function is uniquely defined for each random variable. Although a density function can be obtained from a cumulative distribution function (and vice versa), it has the additional advantage that we can speak of density functions without reference to random variables.

In the example of the 10 million flower seeds, each seed is an element of the population we wish to sample and will produce a white or red flower. So strictly speaking, there is not a numerical value associated with each element of the population. However, when we associate for instance number 1 with white and number 0 with red, then there is a numerical value associated with each element of the population, and we can discuss whether a particular sample is random or not. The random variable X_i is then 1 or 0 depending on whether the i -th seed sampled produces a white or red flower, $i = 1, \dots, n$. If the sampling is performed in such a way that the random variables X_1, X_2, \dots, X_n are independent and have the same density (cfr i.i.d.), then, according to the previous definition of a random sample, the sample is random.

An important part of the definition of a random sample is therefore the meaning of a random variable. Recall from the probability class that a random variable is in essence a function with domain the sample space (loosely speaking: totality of possible outcomes of an experiment) and counterdomain the real line. The random variable X_i is therefore a representation for the numerical values that the i th item (or element) sampled can assume. After the sample is observed, the actual values of X_1, X_2, \dots, X_n are known, and are usually denoted with small letters x_1, x_2, \dots, x_n .

In practice, taking a sample from a target population may not be possible, and instead a sample needs to be taken from another, related population. To distinguish between the two populations, we define sampled population:

Definition 2.5 (Sampled population). Let X_1, X_2, \dots, X_n be a random sample from a population with density $f(\cdot)$, then this population is called the sampled population.

For instance, an electrical engineer is studying the safety of nuclear power plants throughout Europe. He has at his disposal five power plants scattered all over Europe that he can visit and of which he can assess their safety. The sampled population consists of the safety outcomes on the five power plants, whereas the target population consists of the safety outcomes for every power plant in Europe.

Another example to show the difference: suppose that a sociologist desires to study the social habits of 20-year-old students in Belgium. He draws a sample from the 20-year-old students at ULg to make his study. In this case, the target population is the 20-year-old students in

Belgium, and the sampled population is the 20-year-old students at ULg which he sampled. He can draw valid relative-frequency-probabilistic conclusions about his sampled population, but he must use his personal judgment to extrapolate to the target population, and the reliability of the extrapolation cannot be measured in relative-frequency-probability terms.

Note that when a series of experiments or observations can be made under rather uniform conditions, then a number p can be postulated as the probability of an event A happening, and p can be approximated by the *relative frequency* of the event A in a series of experiments.

It is essential to understand the difference between target and sampled population: Valid probability statements can be made about sampled populations on the basis of random samples, but statements about the target populations are not valid in a relative-frequency-probability sense unless the target population is also the sampled population.

2.3 Sampling schemes

In general, two different sampling techniques can be adopted: probability or non-probability based. Probability sampling is a sampling technique where the samples are gathered in a process that gives all the individuals in the population equal chances of being selected. This is in contrast to non-probability sampling techniques, where the samples are gathered in a process that does not give all the individuals in the population equal chances of being selected.

2.3.1 Non-probability sampling

Reliance On Available Subjects Relying on available subjects, such as stopping people on a street corner as they pass by, is one method of sampling, although it is extremely risky and comes with many cautions. This method, sometimes referred to as a convenience sample, does not allow the researcher to have any control over the representativeness of the sample. It is only justified if the researcher wants to study the characteristics of people passing by the street corner at a certain point in time or if other sampling methods are not possible. The researcher must also take caution to not use results from a convenience sample to generalize to a wider population.

Purposive or Judgmental Sample A purposive, or judgmental, sample is one that is selected based on the knowledge of a population and the purpose of the study. For example, if a researcher is studying the nature of school spirit as exhibited at a school pep rally, he or she might interview people who did not appear to be caught up in the emotions of the crowd or students who did not attend the rally at all. In this case, the researcher is using a purposive sample because those being interviewed fit a specific purpose or description.

Snowball Sample A snowball sample is appropriate to use in research when the members of a population are difficult to locate, such as homeless individuals, migrant workers, or undocumented immigrants. A snowball sample is one in which the researcher collects data on the few members of the target population he or she can locate, then asks those individuals to provide information needed to locate other members of that population whom they know. For example, if a researcher wishes to interview undocumented immigrants from Mexico, he or she might interview a few undocumented individuals that he or she knows

or can locate and would then rely on those subjects to help locate more undocumented individuals. This process continues until the researcher has all the interviews he or she needs or until all contacts have been exhausted.

Quota Sample A quota sample is one in which units are selected into a sample on the basis of pre-specified characteristics so that the total sample has the same distribution of characteristics assumed to exist in the population being studied. For example, if you are a researcher conducting a national quota sample, you might need to know what proportion of the population is male and what proportion is female as well as what proportions of each gender fall into different age categories, race or ethnic categories, educational categories, etc. The researcher would then collect a sample with the same proportions as the national population.

2.3.2 Probability sampling

Simple Random Sample The simple random sample is the basic sampling method assumed in statistical methods and computations. To collect a simple random sample, each unit of the target population is assigned a number. A set of random numbers is then generated and the units having those numbers are included in the sample. For example, let's say you have a population of 1,000 people and you wish to choose a simple random sample of 50 people. First, each person is numbered 1 through 1,000. Then, you generate a list of 50 random numbers (typically with a computer program) and those individuals assigned those numbers are the ones you include in the sample.

Systematic Sample In a systematic sample, the elements of the population are put into a list and then every k th element in the list is chosen (systematically) for inclusion in the sample. For example, if the population of study contained 2,000 students at a high school and the researcher wanted a sample of 100 students, the students would be put into list form and then every 20th student would be selected for inclusion in the sample. To ensure against any possible human bias in this method, the researcher should select the first individual at random. This is technically called a systematic sample with a random start.

Stratified Sample A stratified sample is a sampling technique in which the researcher divided the entire target population into different subgroups, or strata, and then randomly selects the final subjects proportionally from the different strata. This type of sampling is used when the researcher wants to highlight specific subgroups within the population. For example, to obtain a stratified sample of university students, the researcher would first organize the population by college class and then select appropriate numbers of freshmen, sophomores, juniors, and seniors. This ensures that the researcher has adequate amounts of subjects from each class in the final sample.

Cluster Sample Cluster sampling may be used when it is either impossible or impractical to compile an exhaustive list of the elements that make up the target population. Usually, however, the population elements are already grouped into subpopulations and lists of those subpopulations already exist or can be created. For example, let's say the target population in a study was church members in the United States. There is no list of all church members in the country. The researcher could, however, create a list of churches in

the United States, choose a sample of churches, and then obtain lists of members from those churches.

Capture Recapture Sampling is a sampling technique used to estimate the number of individuals in a population. We capture a first sample from the population and mark the individuals captured. If the individuals in a certain population are clearly identified, there is no need for any marking and consequently, we simply register these initially captured individuals. After an appropriate waiting time, a second sample from the population is selected independently from the initial sample. If the second sample is representative of the population, then the proportion of marked individuals in the second capture should be the same as the proportion of marked individuals in the population. From this assumption, we can estimate the number of individuals from a population. This procedure has been used not only to estimate the abundance of animals such as birds, fish, insects, and mice, among others, but also the number of minority individuals, such as the homeless in a city, for possible adjustments of undercounts in a census.

2.4 Sampling Challenges

Because researchers can seldom study the entire population, they must choose a subset of the population, which can result in several types of error. Sometimes, there are discrepancies between the sample and the population on a certain parameter that are due to random differences. This is known as *sampling error* and can occur through no fault of the researcher.

Far more problematic is *systematic error*, which refers to a difference between the sample and the population that is due to a systematic difference between the two rather than random chance alone. The response rate problem refers to the fact that the sample can become self-selecting, and that there may be something about people who choose to participate in the study that affects one of the variables of interest. For example, in the context of giving eye care to patients, we may experience this kind of error if we simply sample those who choose to come to an eye clinic for a free eye exam as our experimental group and those who have poor eyesight but do not seek eye care as our control group. It is very possible in this situation that the people who actively seek help happen to be more proactive than those who do not. Because these two groups vary systematically on an attribute that is not the dependent variable (economic productivity), it is very possible that it is this difference in personality trait and not the independent variable (if they received corrective lenses or not) that produces any effects that the researcher observes on the dependent variable. This would be considered a failure in internal validity.

Another type of systematic sampling error is coverage error, which refers to the fact that sometimes researchers mistakenly restrict their sampling frame to a subset of the population of interest. This means that the sample they are studying varies systematically from the population for which they wish to generalize their results. For example, a researcher may seek to generalize the results to the population of developing countries, yet may have a coverage error by sampling only heavily urban areas. This leaves out all of the more rural populations in developing countries, which have very different characteristics than the urban populations on several parameters. Thus, the researcher could not appropriately generalize the results to the broader population and would therefore have to restrict the conclusions to populations in urban areas of developing countries ([14]).

First and foremost, a researcher must think very carefully about the population that will be included in the study and how to sample that population. Errors in sampling can often be avoided by good planning and careful consideration. However, in order to improve a sampling frame, a researcher can always seek more participants. The more participants a study has, the less likely the study is to suffer from sampling error. In the case of the response rate problem, the researcher can actively work on increasing the response rate, or can try to determine if there is in fact a difference between those who partake in the study and those who do not. The most important thing for a researcher to remember is to eliminate any and all variables that the researcher cannot control. While this is nearly impossible in field research, the closer a researcher comes to isolating the variable of interest, the better the results ([3]).

2.5 Distribution of a sample

Definition 2.6 (Distribution of sample). Let X_1, X_2, \dots, X_n denote a sample of size n . The distribution of the sample X_1, X_2, \dots, X_n is defined to be the joint distribution of X_1, X_2, \dots, X_n .

Hence, if X_1, X_2, \dots, X_n is a random sample of size n from $f(\cdot)$ then the distribution of the random sample X_1, X_2, \dots, X_n , defined as the joint distribution of X_1, X_2, \dots, X_n , is given by $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n)$, and X_1, X_2, \dots, X_n are stochastically independent.

Remark: Note that our definition of random sampling has automatically ruled out sampling from a finite population without replacement since, then, the results of the drawings are not independent...

2.6 Statistics

We will introduce the technical meaning of the word **statistic** and look at some commonly used statistics.

Definition 2.7. Any function of the elements of a random sample, which does not depend on unknown parameters, is called a **statistic**.

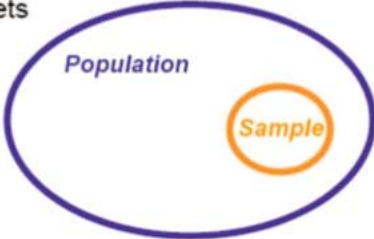
Strictly speaking, $H(X_1, X_2, \dots, X_n)$ is a statistic and $H(x_1, x_2, \dots, x_n)$ is the observed value of the statistic. Note that the former is a random variable, often called an **estimator of θ** , while $H(x_1, x_2, \dots, x_n)$ is called an **estimate of θ** . However, the word *estimate* is sometimes used for both random variable and its observed value.

In what follows, we give some classical examples of statistics. Suppose that we have a random sample X_1, X_2, \dots, X_n from a distribution with (population) mean μ and variance σ^2 .

1. $\bar{X} = \sum_{i=1}^n X_i/n$ is called the **sample mean**.
2. $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is called the **sample variance** (sometimes denoted as S_{n-1}^2).
3. $S = \sqrt{S^2}$ is called the **sample standard deviation**.
4. $M_r = \sum_{i=1}^n X_i^r/n$ is called the **r th sample moment** about the origin.

5. Suppose that the random variables X_1, \dots, X_n are ordered and re-written as $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. The vector $(X_{(1)}, \dots, X_{(n)})$ is called the **ordered sample**.
- (a) $X_{(1)}$ is called the **minimum of the sample**, sometimes written X_{\min} or $\min(X_i)$.
 - (b) $X_{(n)}$ is called the **maximum of the sample**, sometimes written X_{\max} or $\max(X_i)$.
 - (c) $X_{\max} - X_{\min} = R$ is called the **sample range**.
 - (d) The **sample median** is $X_{(\frac{n+1}{2})}$ if n is odd, and $\frac{1}{2} (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})$ if n is even.

Remark: Note the difference between the concepts ‘parameter’ and ‘statistic’.

<p>Population: The entire group of individuals in which we are interested but can't usually assess directly.</p> <p>Example: All humans, all working-age people in California, all crickets</p>	<p>Sample: The part of the population we actually examine and for which we do have data.</p> <p>How well the sample represents the population depends on the sample design.</p>
	
<p>A parameter is a number describing a characteristic of the population.</p>	<p>A statistic is a number describing a characteristic of a sample.</p>

Definition 2.8. The **standard error** is the standard deviation of the sampling distribution of a statistic. The standard error of the mean (i.e., the sample mean is considered here as the statistic of interest) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population at large.

A special type of statistics are *sufficient statistics*. These are functions of the sample at hand that tell us just as much about the parameter we are interested (for instance, a parameter θ) in as the entire sample itself. Hence, by using it, no information about θ will be lost. It would be sufficient for estimation purposes (i.e., estimating the parameter θ), which explains the name. A more formal definition will follow in Chapter 4 on Estimation.

Computer Exercise 2.1. Generate a random sample of size 100 from a normal distribution with mean 10 and standard deviation 3. Use R to find the value of the (sample) mean, variance, standard deviation, minimum, maximum, range, median, and M_2 , the statistics defined above. Repeat for a sample of size 100 from an exponential distribution with parameter 1.

Solution of Computer Exercise 2.1.

<pre> #_____ SampleStats.R _____ # Generate the normal random sample rn <- rnorm(n=100,mean=10,sd= 3) print(summary(rn)) cat("mean = ",mean(rn),"\n") cat("var = ",var(rn),"\n") cat("sd = ",sd(rn),"\n") cat("range = ",range(rn),"\n") cat("median = ",median(rn),"\n") cat("Second Moment = ",mean(rn^2),"\n") </pre>	<pre> > source("SampleStats.R") Min. 1st Qu. Median Mean 3rd Qu. Max. 1.8 7.9 9.4 9.8 12.0 18.2 mean = 9.9 var = 9.5 sd = 3.1 range = 1.8 18 median = 9.4 Second Moment = 106 </pre>
---	--

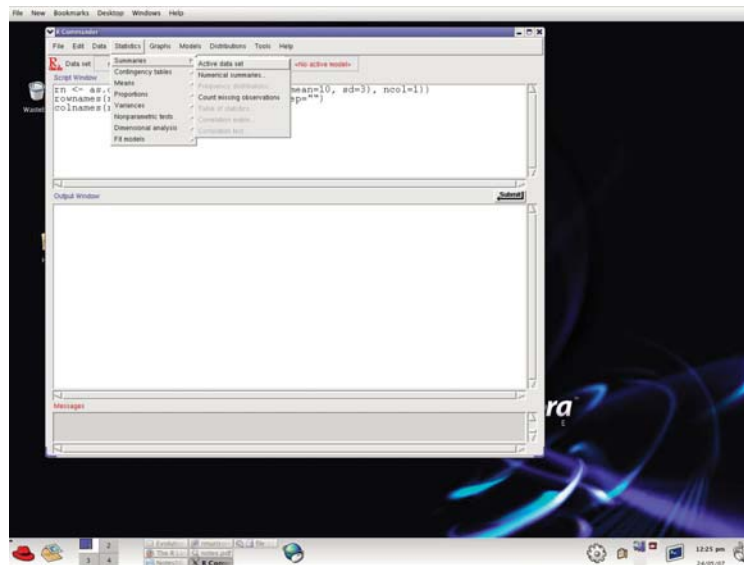


Figure 2.6.1: The summary statistics are output in the Output window.

The exponential random sample is generated using: `re <- rexp(n=100, rate=1)` or the Rcmdr menus `Distributions → Exponential distribution → Sample from an exponential distribution`.

The mean and variance often involve relatively simple functions of the parameter(s), so when considering statistics, it is helpful to note which statistics you'd expect to give information about the mean and which ones give information about the variance. Intuitively, \bar{X} and the sample median give information about μ whereas S^2 gives information about σ^2 . Before going into the geometry of a sample, let us recapitulate some facts about \bar{X} , when X_1, X_2, \dots, X_n is a random sample, all having the same distribution with mean μ and variance σ^2 :

1. It is a random variable with $E(\bar{X}) = \mu$, $Var(\bar{X}) = \sigma^2/n$, where $\mu = E(X_i)$, $\sigma^2 = Var(X_i)$.
2. If X_1, X_2, \dots, X_n is from a **normal** distribution, then \bar{X} is also normally distributed.
3. For large n and *any* distribution of the X_i for which a mean (μ) and a variance (σ^2) exist, \bar{X} is distributed approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ (by the Central Limit Theorem).

2.7 Sample Geometry

2.7.1 The geometry of the sample

A single multivariate observation is the collection of measurements on p different variables, taken on the same item or trial. For each item (individual, experimental unit) p variables or characters are recorded, each of which is indicated by:

$$x_{jk} \equiv \begin{cases} j & \text{th item} \\ k & \text{th variable} \end{cases}.$$

All available data (on all available n samples) can be displayed in matrix form as follows:

$$\mathbf{X} = \begin{matrix} & \text{variable 1} & \cdots & \text{variable } p \\ \text{item 1} & \left(\begin{matrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ \text{item } j & \begin{matrix} x_{j1} & \cdots & x_{jp} \\ \vdots & & \\ \text{item } n & \begin{matrix} x_{n1} & \cdots & x_{np} \end{matrix} \end{matrix} \end{matrix} \right)$$

Obviously, the data can be plotted in two different ways. For instance, for the p -dimensional scatterplot (see also Chapter 3), the rows of \mathbf{X} represent the n points in p -dimensional space. This space is called the *sample space*, in contrast to the n -dimensional *variable space*.

Variable space. We look at the variables, we have p vectors of size $(n \times 1)$.

Sample space. We look at items, we have n vectors of size $(p \times 1)$.

$$x_j^T = (x_{j1}, \dots, x_{jp})$$

The scatter of n points in p -dimensional space provides information on their locations and variability. If the points are regarded as solid spheres, the sample mean vector, \bar{x} is the center of balance. Variability occurs in more than one direction and it is quantified by the sample variance-covariance matrix \mathbf{S}_n . A single numerical measure of variability is provided by the determinant of the sample variance-covariance matrix. Although it is hard to visualize anything in dimensions greater than 3, consideration of the data in as n points in p dimensions provides insights that are not readily available from algebraic expressions.

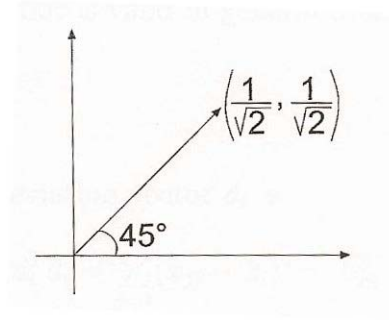
2.7.1.1 The mean

It is possible to give a geometrical interpretation of the process of finding a sample mean. Let us introduce the following notations:

$$\mathbf{1}_n^T = (1, \dots, 1) \quad (n \text{ times})$$

$$y_i^T = (x_{1i}, x_{2i}, \dots, x_{ni}) \quad (n \text{ measurements on the } i\text{th variable})$$

In the n -dimensional space $\frac{1}{\sqrt{n}}\mathbf{1}_n$ is a vector of unit length having equal angles with each of the n coordinate axes.



Example 2.1 ($n = 2$).

Length of vector: $\left(\left(\frac{1}{\sqrt{2}} \right)^2 + \left(\frac{1}{\sqrt{2}} \right)^2 \right)^{\frac{1}{2}} = 1$.

The projection of a vector y on a vector a of unit length is

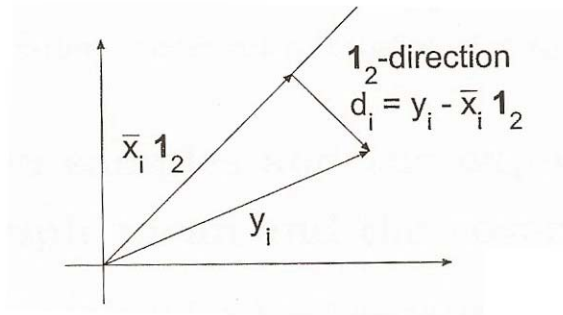
$$(y^T a)a$$

Apply this to y_i and $\frac{1}{\sqrt{n}}\mathbf{1}_n$:

$$\begin{aligned} \left(y_i^T \frac{1}{\sqrt{n}}\mathbf{1}_n \right) \frac{1}{\sqrt{n}}\mathbf{1}_n &= \frac{x_{1i} + \dots + x_{ni}}{n} \mathbf{1}_n \\ &= \bar{x}_i \mathbf{1}_n \end{aligned}$$

Thus, the sample mean for the i th variable \bar{x}_i corresponds to the multiple of $\mathbf{1}_n$ required to give the projection of the vector y_i onto the line determined by $\mathbf{1}_n$

Furthermore, each vector can be decomposed into a mean and a deviation component



$$y_i = \bar{x}_i \mathbf{1}_2 + d_i$$

The vector d_i is called the deviation (or *residual*) vector or the y_i -vector corrected for the mean. The picture is in two dimensions but is valid in general dimensional n .

2.7.1.2 Variance and correlation

The squared length of the deviation vector d_i is

$$d_i^T d_i = \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 = L_{d_i}^2 = n \tilde{s}_{ii} = n \tilde{s}_i^2$$

or a sum of squared deviations.

For any two deviation vectors d_i and d_k we have

$$d_i^T d_k = L_{d_i} L_{d_k} \cos(\theta_{ik}) = n \widetilde{s}_{ik}$$

with θ_{ik} the angle between d_i and d_k . This implies that the deviation vectors comprise information about S_n .

Since

$$d_i^T d_k = \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

we obtain

$$r_{ik} = \cos(\theta_{ik}) = \frac{\sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \sum_{j=1}^n (x_{jk} - \bar{x}_k)^2}}$$

With

$$\begin{aligned} \widetilde{s}_{ii} &= \widetilde{s}_i^2 = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2 \\ \widetilde{s}_{ik} &= \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \end{aligned}$$

we have

$$r_{ik} = \cos(\theta_{ik}) = \frac{\widetilde{s}_{ik}}{\widetilde{s}_i \widetilde{s}_k}$$

So the cosine of the angle determined by the deviation vectors d_j and d_k is nothing else but the *sample correlation coefficient*. Note that the **correlation coefficient** r_{ik} satisfies $-1 \leq r_{ik} \leq 1$ (it is a cosine!!)

2.7.2 Expected values of the sample mean and the covariance matrix

In order to study the sampling variability of statistics like the mean and variance-covariance with the ultimate aim of making inferences, we need to make assumptions about the variables whose observed values constitute the data \mathbf{X} . In the previous section we looked at X as a descriptive data scheme. We now look at X as a single trial from a random experiment.

X_{jk} is a random variable

$X_j^T = (X_{j1}, \dots, X_{jp})$ is a random vector

$$X = \begin{bmatrix} X_1^T \\ \vdots \\ X_j^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{j1} & \cdots & X_{jp} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

X is a random sample if X_1^T, \dots, X_n^T represent **independent** random vectors from a **common joint distribution** (density) $f(x_1, \dots, x_p)$.

The mean and the variance covariance are given by:

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix} \quad \text{a } (p \times 1)\text{-vector}$$

$$\widetilde{S}_n = (\widetilde{s}_{ij}) \quad \text{a } (p \times p)\text{-matrix}$$

If the aforementioned joint distribution has mean vector μ and variance-covariance matrix Σ , then the vector \bar{X} will be an unbiased estimator of μ (i.e., $E(\bar{X}) = \mu$) and its variance-covariance matrix will be $1/n\Sigma$ (i.e., $\text{Cov}(\bar{X}) = 1/n\Sigma$). A proof is left as exercise. Note that the population variance-covariance is divided by the sample size and not $n - 1$.

Summarizing the main properties of \bar{X} and \widetilde{S}_n :

- $E(\bar{X}) = \mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$, $\mu_i = E(X_{ji})$

- $\text{Cov}(\bar{X}) = \frac{1}{n}\Sigma = \frac{1}{n}(\sigma_{ij})$ where

$$\begin{aligned} \sigma_{ij} &= \text{Cov}(X_{\ell i}, X_{\ell j}) \\ &= E[(X_{\ell i} - \mu_i)(X_{\ell j} - \mu_j)] \end{aligned}$$

- $E(\widetilde{S}_n) = \frac{n-1}{n}\Sigma$ and therefore $S = \frac{n}{n-1}\widetilde{S}_n$ is unbiased, i.e.,

- $E(S) = \Sigma$ and therefore S is the quantity to use

2.7.3 Generalized variance

With

$$S = (s_{ik})$$

where $s_{ik} = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$, the **generalized sample variance** is defined as

$$|S| = \det(S)$$

We use $S = \frac{n}{n-1}S_n$ instead of S_n because $E(S_n) = \frac{n-1}{n}\Sigma$, whereas $E(S) = \Sigma$.

The generalized sample variance provides one way of writing the information on all variances and covariances as a single number. It can be interpreted as follows:

The p -vectors d_1, \dots, d_p span a volume (see Fig. 3.6 in Johnson et al. (2002)). It can be shown that

$$|S| = (n - 1)^{-p} \times (\text{volume})^2$$

We demonstrate this for $p = 2$

$$\begin{aligned} |S| &= \det \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix} = s_{11}s_{22} - s_{12}^2 \\ &= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 \\ &= s_{11}s_{22}(1 - \cos^2(\theta_{12})) \\ &= \frac{L_{d_1}^2 L_{d_2}^2}{(n - 1)^2} \sin^2(\theta_{12}) \\ &= \frac{(\text{area trapezoid})^2}{(n - 1)^2} \end{aligned}$$

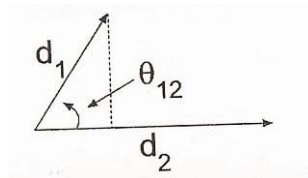


Figure 2.7.1: area trapezoid = $L_{d_1} L_{d_2} \sin(\theta_{12})$

Properties of generalized sample variance. The quantity of $|S|$ increases with

- length of d_i
- size of the angle between d_i and d_k

But (as one can expect) S matrices that show a different correlation structure can have the same value for $|S|$. This is clear from Fig. 2.7.2 and the discussion that follows.

Picture (a)

$$\begin{aligned} S &= \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \\ |S| &= 25 - 16 = 9 \end{aligned}$$

Eigenvalues

$$\begin{aligned} |S - \lambda I_2| &= 0 \\ (5 - \lambda)^2 - 16 &= 0 \\ \lambda^2 - 10\lambda + 9 &= 0 \\ (\lambda - 9)(\lambda - 1) &= 0 \end{aligned}$$

Although the generalized variance has one intuitively pleasing geometrical interpretation, it suffers from a basic weakness as a descriptive summary of the sample covariance matrix S , as the following example shows.

Example 2.2 (Interpreting the generalized variance).

Fig. 2.7.2 gives three scatter plots with very different patterns of correlation. All three data sets have $\bar{x}^T = [1, 2]$ and the covariance matrices are

$$S = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, r = .8 \quad S = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, r = 0 \quad S = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, r = -.8$$

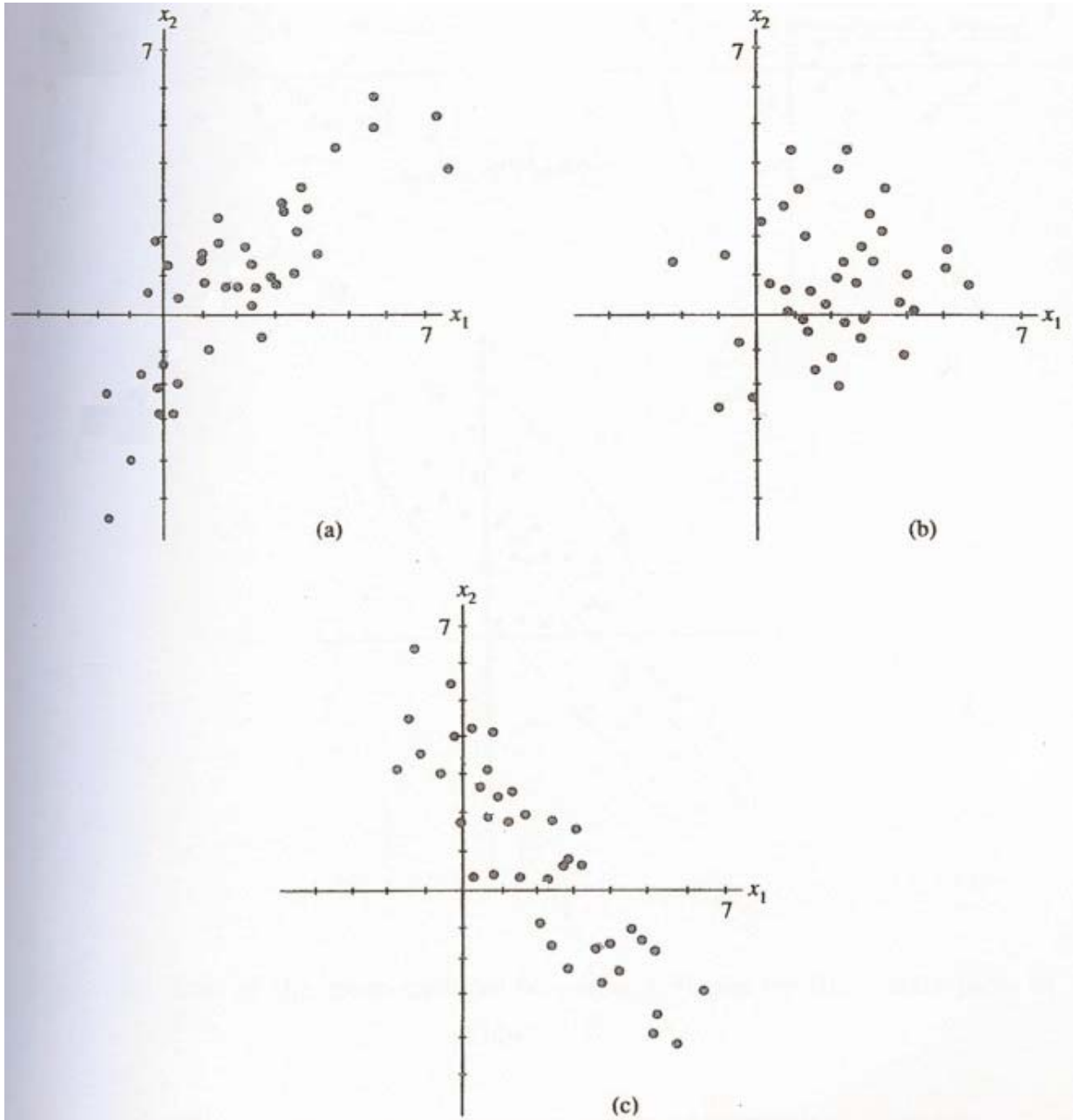


Figure 2.7.2: Scatter plots with three different orientations.

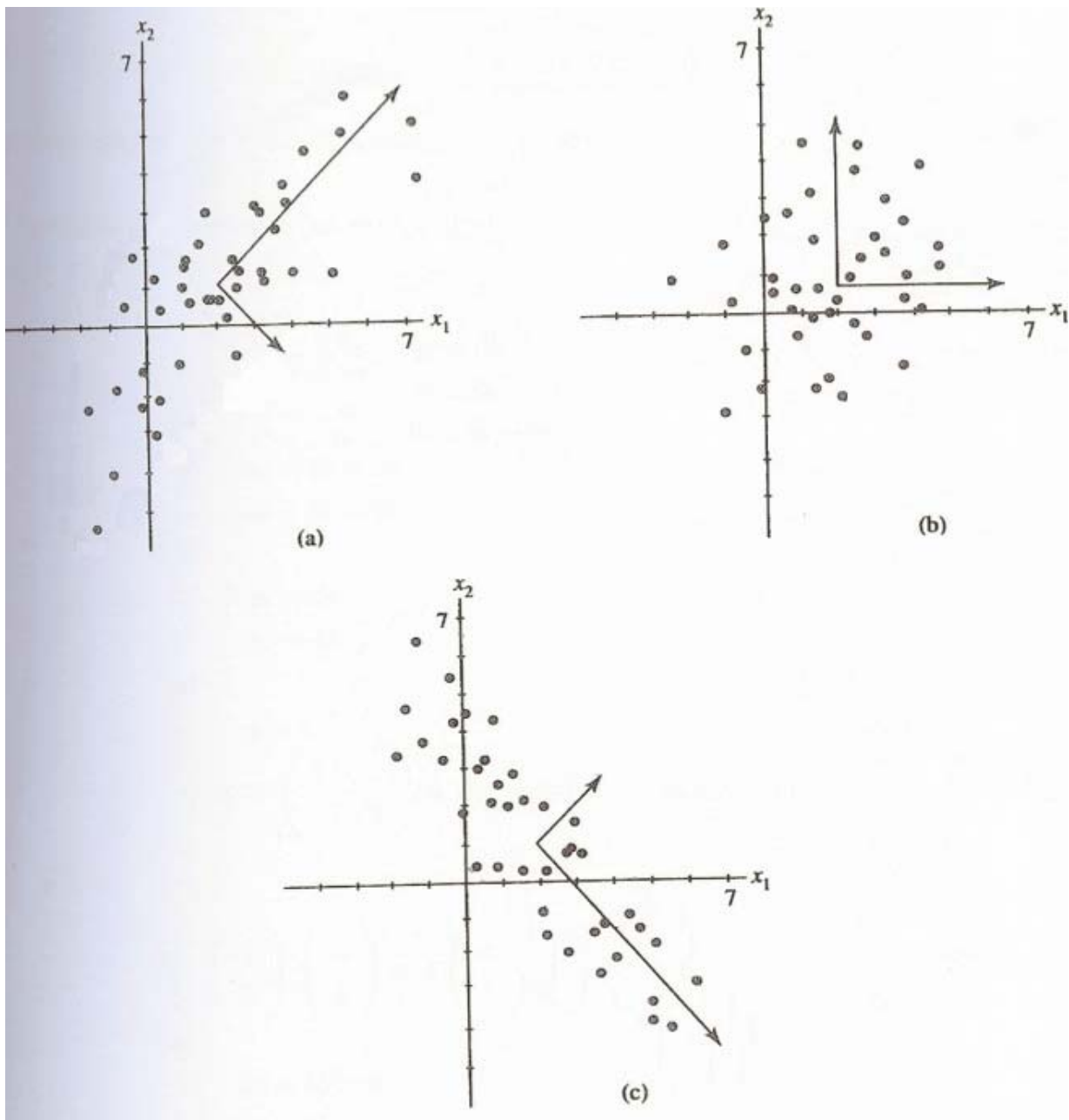


Figure 2.7.3: Axes of the mean-centered 95% ellipses for the scatter plots in Fig. 2.7.2.

Note: Here the generalized variance $|S|$ gives the same value, $|S| = 9$, for all three patterns. But generalized variance does not contain any information on the orientation of the patterns. Generalized variance is easier to interpret when the two or more samples (patterns) being compared have nearly the same orientations.

Notice that our three patterns of scatter appear to cover approximately the same area. The ellipses that summarize the variability

$$(x - \bar{x})^T S^{-1} (x - \bar{x}) \leq c^2$$

do have exactly the same area, since all have $|S| = 9$.

Discussion on picture (a) continued

$$\lambda_1 = 9$$

$$\begin{aligned} \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} &= 9 \begin{pmatrix} a \\ b \end{pmatrix} \\ \Leftrightarrow \begin{cases} 5a + 4b &= 9a \\ 4a + 5b &= 9b \end{cases} \\ \Leftrightarrow \begin{cases} 4a &= 4b \\ 4a &= 4b \end{cases} \\ \Leftrightarrow a &= b \\ \Rightarrow e_1 &= \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \text{ is the unit-length eigenvector} \end{aligned}$$

$$\lambda_2 = 1$$

$$\begin{aligned} \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} &= 1 \begin{pmatrix} a \\ b \end{pmatrix} \\ \Leftrightarrow \begin{cases} 5a + 4b &= a \\ 4a + 5b &= b \end{cases} \\ \Leftrightarrow \begin{cases} 4a &= -4b \\ 4a &= -4b \end{cases} \\ \Leftrightarrow a &= -b \\ \Rightarrow e_2 &= \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} \text{ is the unit-length eigenvector} \end{aligned}$$

Note. Since S is a positive definite matrix we can give the **spectral decomposition** of S in terms of (λ_1, e_1) and (λ_2, e_2) :

$$\begin{aligned} S &= \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T \\ &= 9 \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + 1 \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \\ &= \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} \end{aligned}$$

For each of the scatter plots in Fig. 2.7.2 we can look at the mean centered ellipsoid (Mahalanobis distance)

$$(x - \bar{x})^T S^{-1} (x - \bar{x}) = c^2$$

- Note that if λ is an eigenvalue of S , λ^{-1} is an eigenvalue of S^{-1}

$$\begin{aligned} Se &= \lambda e \\ \Leftrightarrow S^{-1}Se &= \lambda S^{-1}e \\ \Leftrightarrow \frac{1}{\lambda}e &= S^{-1}e \end{aligned}$$

- From these equivalences it also follows that S and S^{-1} have the same eigenvectors

The axes of the ellipsoid are

$$\begin{aligned} c\sqrt{\lambda_1}e_1 \\ c\sqrt{\lambda_2}e_2 \end{aligned}$$

(see Fig. 2.7.3)

Note. It holds in general that a matrix is positive definite iff all eigenvalues are (strictly) positive.

Problem 2.1. *Given*

1. $S = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$
2. $S = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}$

For the matrices S given in 1 and 2 obtain

1. the generalized variance
2. the spectral decomposition
3. the axes of the ellipsoid

$$(x - \bar{x})^T S^{-1} (x - \bar{x}) = 4$$

4. show that $|S| = \lambda_1 \lambda_2$.

Check your findings by looking at Fig. 2.7.2 and 2.7.3.

Note. Problem 2.1 4 is an interesting result. This property holds for a general positive definite ($p \times p$)-matrix

$$|S| = \lambda_1 \lambda_2 \dots \lambda_p$$

It provides a way for a multi-number summary of the variance-covariance matrix. Since axes of ellipsoid are of the form

$$c\sqrt{\lambda}e$$

and since e is a unit-length vector λ that indicates how much variability (noise) we have in the e direction, the conclusion is clear: the knowledge of eigenvalues is a key issue if interest is in understanding the variability in the data.

Generalized sample variance: zero

- $|S| = 0$ iff d_1, \dots, d_p are linearly dependent
- d_1, \dots, d_p linearly dependent

$$\Leftrightarrow (d_1 | d_2 | \dots | d_p) \begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix} = 0$$

$$\Leftrightarrow a_1 d_1 + \dots + a_p d_p = 0$$

$$\Leftrightarrow (X - \mathbf{1}\bar{x}^T)a = 0 \quad \text{where we note that } \mathbf{1}_n \bar{x}^T \text{ is a } (n \times 1) \times (1 \times p) = (n \times p) \text{ matrix.}$$

$$\Leftrightarrow (n-1)Sa = (X - \mathbf{1}_n \bar{x}^T)^T \underbrace{(X - \mathbf{1}_n \bar{x}^T)a}_0 = 0$$

$$\Leftrightarrow \text{the columns of } S \text{ are dependent}$$

$$\Leftrightarrow |S| = 0$$

In such cases some variables are linearly dependent on others and can be removed: we call this the **degenerate case**.

Property

- For $n \leq p$: $|S| = 0$
- For $n > p$:

$$|S| > 0 \Leftrightarrow \forall \ell : \text{Var}(\ell^T X) = \text{Var}\left(\sum_{j=1}^p \ell_j X_j\right) > 0$$

- The generalized variance of the *standardized* variables $(x_{ij} - \bar{x}_i)/\sqrt{s_{ii}}$ is R , since

$$R = (r_{ij}) = \frac{s_{ij}}{s_i s_j}$$

- From the definition of R , it easily follows that

$$|S| = s_1^2 \dots s_p^2 |R|$$

Problem 2.2. For $p = 3$ prove that

$$|S| = s_1^2 s_2^2 s_3^2 |R|.$$

2.8 Resampling

We have indicated before that by considering a sample as a population, general population results carry over to samples. This is the basis for commonly used resampling techniques.

Bootstrapping is a generic methodology, whose implementation involves a simple yet powerful principle: creating many repeated data samples from the single one in hand, and making

inference from those samples. During the Monte Carlo sampling, the probability that a data point is picked is $1/N$ irrespective of whether it has been picked before. In the statistics literature this is called picking from a set “with replacement”. The name ‘bootstrapping’ derives from the expression pulling oneself up by one’s bootstraps, meaning no outside help (additional data, parametric assumptions) is involved. It seems almost magical, like getting something for nothing. Bootstrapping can be used for many statistical purposes besides variable selection. One of the most common is to compute confidence intervals, as we will see later.

There exist several resampling strategies. Whereas bootstrap uses N_{boot} data sets each containing N points obtained by repeated random (i.e. Monte Carlo) sampling of the original set of N points, *jackknife* (sampling) considers N new data sets, each of containing all the original data points minus 1.

Note that the aforementioned techniques involve taking samples from data directly, without making assumptions about underlying distributions. This is different from taking samples from a (particular) distribution. The latter can be achieved by relying on so-called *inverse transform sampling* (see last year). Inverse transform sampling (also known as inversion sampling, the inverse probability integral transform, the inverse transformation method) is a basic method for pseudo-random number sampling, i.e. for generating sample numbers at random from any probability distribution given its cumulative distribution function, at least when the latter is relatively simple and can easily be inverted. It is built on the principle that if Y has a uniform distribution on $[0, 1]$ and if X has a cumulative distribution F_X , then the cumulative distribution function of the random variable $F_X^{-1}(Y)$ is F_X . When samples need to be taken from relatively complex distributions, other strategies need to be adopted, including *rejection sampling* and *importance sampling*.

2.9 The Importance of Study Design

Designing a new study is an art. To design effective studies, you must have well-developed criteria and procedures to help you choose between design alternatives. But the development of such criteria is only the first step in developing skill in the art of study design. There are not a limited set of design alternatives between which you can choose. There are an unlimited variety of design possibilities and study objectives. The artful study designer will develop a unique approach for every new study. The development of that unique approach requires making judgments about how well the objectives can be achieved under different design possibilities and constraints. These judgements will influence your choice of objectives.

Careful study design is essential to carry out high quality research that yields valid results. Planning an effective study involves defining the target population, specifying an intervention, selecting appropriate outcome measures, and choosing the appropriate design. Randomized controlled trials, observational cohort studies, and case-control studies all have advantages and disadvantages. When planning or evaluating a study, threats to its validity, including chance, bias, and confounding must be considered.

Study designs can be classified in two big groups: 1) observational studies and 2) experimental studies. Although the randomized controlled trial may be considered the most valid design, well-designed observational studies often yield similar results. If you want to read more about this, please refer to [13].

Cross-sectional studies are simple in design and are aimed at finding out the prevalence of a phenomenon, problem, attitude or issue by taking a snap-shot or cross-section of the population. This obtains an overall picture as it stands at the time of the study. For example, a cross-sectional design would be used to assess demographic characteristics or community attitudes. These studies usually involve one contact with the study population and are relatively cheap to undertake.

Pre-test/post-test studies measure the change in a situation, phenomenon, problem or attitude. Such studies are often used to measure the efficacy of a program. These studies can be seen as a variation of the cross-sectional design as they involve two sets of cross-sectional data collection on the same population to determine if a change has occurred.

Retrospective studies investigate a phenomenon or issue that has occurred in the past. Such studies most often involve secondary data collection, based upon data available from previous studies or databases. For example, a retrospective study would be needed to examine the relationship between levels of unemployment and street crime in Belgium over the past 100 years.

Prospective studies seek to estimate the likelihood of an event or problem in the future. Thus, these studies attempt to predict what the outcome of an event is to be. General science experiments are often classified as prospective studies because the experimenter must wait until the experiment runs its course in order to examine the effects. Randomized controlled trials are always prospective studies and often involve following a cohort of individuals to determine the relationship between various variables.

Researchers constantly face the design dilemma that deals with reconciling the ideal question one would pose and the data that can be collected or accessed.

CHAPTER

3

EXPLORATORY DATA ANALYSIS

Exploratory data analysis or “EDA” is a critical first step in analyzing the data from an experiment. Here are the main reasons we use EDA:

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis.

3.1 Typical data format and the types of EDA

The data from an experiment are generally collected into a rectangular array (e.g., spreadsheet or database), most commonly with one row per experimental subject and one column for each subject identifier, outcome variable, and explanatory variable. Each column contains the numeric values for a particular quantitative variable or the levels for a categorical variable. (Some more complicated experiments require a more complex data layout.)

People are not very good at looking at a column of numbers or a whole spreadsheet and then determining important characteristics of the data. They find looking at numbers to be

tedious, boring, and/or overwhelming. Exploratory data analysis techniques have been devised as an aid in this situation. Most of these techniques work in part by hiding certain aspects of the data while making other aspects more clear.

Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).

Non-graphical methods generally involve calculation of summary statistics, while graphical methods obviously summarize the data in a diagrammatic or pictorial way. Univariate methods look at one variable (data column) at a time, while multivariate methods look at two or more variables at a time to explore relationships. Usually our multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables. *It is almost always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.*

Beyond the four categories created by the above cross-classification, each of the categories of EDA have further divisions based on the role (outcome or explanatory) and type (categorical or quantitative) of the variable(s) being examined.

Although there are guidelines about which EDA techniques are useful in what circumstances, there is an important degree of looseness and art to EDA. Competence and confidence come with practice, experience, and close observation of others. Also, EDA need not be restricted to techniques you have seen before; sometimes you need to invent a new way of looking at your data.

<p>The four types of EDA are univariate non-graphical, multivariate non-graphical, univariate graphical and multivariate graphical.</p>
--

This chapter first discusses the non-graphical and graphical methods for looking at single variables, then moves on to looking at multiple variables at once, mostly to investigate the relationships between the variables.

3.2 Univariate non-graphical EDA

The data that come from making a particular measurement on all of the subjects in a sample represent our observations for a single characteristic such as age, gender, speed at a task, or response to a stimulus. We should think of these measurements as representing a “sample distribution” of the variable, which in turn more or less represents the “population distribution” of the variable. The usual goal of univariate non-graphical EDA is to better appreciate the “sample distribution” and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis.

3.2.1 Categorical data

The characteristics of interest for a *categorical* variable are simply the range of values and the frequency (or relative frequency) of occurrence for each value. (For ordinal variables it is sometimes appropriate to treat them as quantitative variables using the techniques in the second part of this section.) Therefore the only useful univariate non-graphical techniques for categorical variables is some form of **tabulation** of the frequencies, usually along with calculation of the fraction (or percent) of data that falls in each category. For example if we categorize subjects by College at Carnegie Mellon University as H&SS, MCS, SCS and “other”, then there is a true population of all students enrolled in the 2007 Fall semester. If we take a random sample of 20 students for the purposes of performing a memory experiment, we could list the sample “measurements” as H&SS, H&SS, MCS, other, other, SCS, MCS, other, H&SS, MCS, SCS, SCS, other, MCS, MCS, H&SS, MCS, other, H&SS, SCS. Our EDA would look like this:

Statistic/College	H&SS	MCS	SCS	other	Total
Count	5	6	4	5	20
Proportion	0.25	0.30	0.20	0.25	1.00
Percent	25%	30%	20%	25%	100%

Note that it is useful to have the total count (frequency) to verify that we have an observation for each subject that we recruited. (Losing data is a common mistake, and EDA is very helpful for finding mistakes.) Also, we should expect that the proportions add up to 1.00 (or 100%) if we are calculating them correctly (count/total). Once you get used to it, you won’t need both proportion (relative frequency) and percent, because they will be interchangeable in your mind.

A simple tabulation of the frequency of each category is the best univariate non-graphical EDA for categorical data.

3.2.2 Characteristics of quantitative data

Univariate EDA for quantitative variable is a way to make preliminary assessments about the population distribution of the variable using the data of the observed sample.

The characteristics of the population distribution of a *quantitative* variable are its center, spread, modality (number of peaks in the pdf), shape (including “heaviness of the tails”), and outliers. Our observed data represent just one sample out of an infinite number of possible samples. *The characteristics of our randomly observed sample are not inherently interesting, except to the degree that they represent the population that it came from.*

What we observe in the **sample** of measurements for a particular variable that we select for our particular experiment is the “sample distribution”. We need to recognize that this would be different each time we might repeat the same experiment, due to selection of a different random sample, a different treatment randomization, and different random (incompletely controlled) experimental conditions. In addition we can calculate “sample statistics” from the data, such as

sample mean, sample variance, sample standard deviation, sample skewness and sample kurtosis. These again would vary for each repetition of the experiment, so they don't represent any deep truth, but rather represent some uncertain information about the underlying population distribution and its parameters, which are what we really care about.

Many of the sample's distributional characteristics are seen qualitatively in the univariate graphical EDA technique of a histogram. In most situations it is worthwhile to think of univariate non-graphical EDA as telling you about aspects of the histogram of the distribution of the variable of interest. Again, these aspects are quantitative, but because they refer to just one of many possible samples from a population, they are best thought of as random (non-fixed) estimates of the fixed, unknown parameters of the distribution of the population of interest.

If the quantitative variable does not have too many distinct values, a tabulation, as we used for categorical data, will be a worthwhile univariate, non-graphical technique. But mostly, for quantitative variables we are concerned here with the quantitative numeric (non-graphical) measures which are the various **sample statistics**. In fact, sample statistics are generally thought of as estimates of the corresponding population parameters.

Figure 3.2.1 shows a histogram of a sample of size 200 from an infinite population characterized by distribution **C**. Remember that in that section we examined the parameters that characterize theoretical (population) distributions. Now we are interested in learning what we can (but not everything, because parameters are “secrets of nature”) about these parameters from measurements on a (random) sample of subjects out of that population.

The bi-modality is visible, as is an **outlier** at $X = -2$. There is no generally recognized formal definition for outlier, but roughly it means values that are outside of the areas of a distribution that would commonly occur. This can also be thought of as sample data values which correspond to areas of the population pdf (or pmf) with low density (or probability). The definition of “outlier” for standard boxplots is described below. Another common definition of “outlier” consider any point more than a fixed number of standard deviations from the mean to be an “outlier”, but these and other definitions are arbitrary and vary from situation to situation.

For quantitative variables (and possibly for ordinal variables) it is worthwhile looking at the central tendency, spread, skewness, and kurtosis of the data for a particular variable from an experiment. *But for categorical variables, none of these make any sense.*

3.2.3 Central tendency

The **central tendency** or “location” of a distribution has to do with typical or middle values. The common, useful measures of central tendency are the statistics called (arithmetic) mean, median, and sometimes mode. Occasionally other means such as geometric, harmonic, truncated, or Winsorized means are used as measures of centrality. While most authors use the term “average” as a synonym for arithmetic mean, some use average in a broader sense to also include geometric, harmonic, and other means.

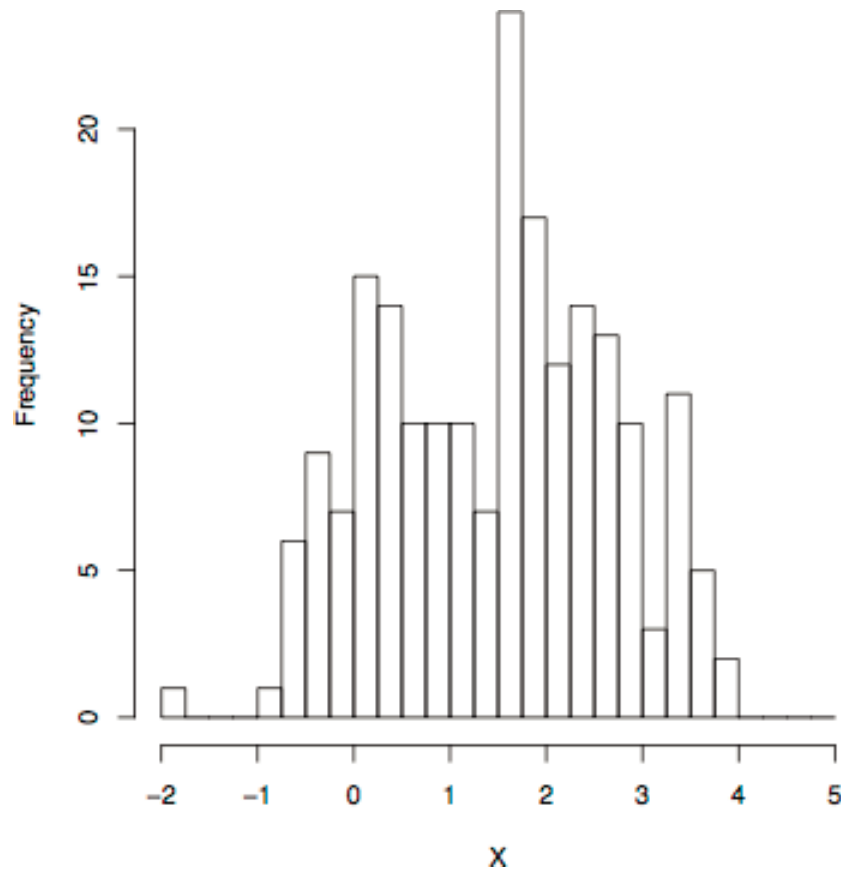


Figure 3.2.1: Histogram from distribution C

Assuming that we have n data values labeled x_1 through x_n , the formula for calculating the sample (arithmetic) **mean** is

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

The arithmetic mean is simply the sum of all of the data values divided by the number of values. It can be thought of as how much each subject gets in a “fair” re-division of whatever the data are measuring. For instance, the mean amount of money that a group of people have is the amount each would get if all of the money were put in one “pot”, and then the money was redistributed to all people evenly. I hope you can see that this is the same as “summing then dividing by n ”.

For any symmetrically shaped distribution (i.e., one with a symmetric histogram or pdf or pmf) the mean is the point around which the symmetry holds. For non-symmetric distributions, the mean is the “balance point”: if the histogram is cut out of some homogeneous stiff material such as cardboard, it will balance on a fulcrum placed at the mean.

For many descriptive quantities, there are both a sample and a population version. For a fixed finite population or for a theoretic infinite population described by a pmf or pdf, there is a single population mean which is a fixed, often unknown, value called the mean **parameter**. On the other hand, the “sample mean” will vary from sample to sample as different samples are

taken, and so is a random variable. The probability distribution of the sample mean is referred to as its **sampling distribution**. This term expresses the idea that any experiment could (at least theoretically, given enough resources) be repeated many times and various statistics such as the sample mean can be calculated each time. Often we can use probability theory to work out the exact distribution of the sample statistic, at least under certain assumptions.

The **median** is another measure of central tendency. The sample median is the middle value after all of the values are put in an ordered list. If there are an even number of values, take the average of the two middle values. (If there are ties at the middle, some special adjustments are made by the statistical software we will use. In unusual situations for discrete random variables, there may not be a unique median.)

For symmetric distributions, the mean and the median coincide. For unimodal skewed (asymmetric) distributions, the mean is farther in the direction of the “pulled out tail” of the distribution than the median is. Therefore, for many cases of skewed distributions, the median is preferred as a measure of central tendency. For example, according to the US Census Bureau 2004 Economic Survey, the median income of US families, which represents the income above and below which half of families fall, was \$43,318. This seems a better measure of central tendency than the mean of \$60,828, which indicates how much each family would have if we all shared equally. And the difference between these two numbers is quite substantial. Nevertheless, both numbers are “correct”, as long as you understand their meanings.

The median has a very special property called **robustness**. A sample statistic is “robust” if moving some data tends not to change the value of the statistic. The median is highly robust, because you can move nearly all of the upper half and/or lower half of the data values any distance away from the median without changing the median. More practically, a few very high values or very low values usually have no effect on the median.

A rarely used measure of central tendency is the **mode**, which is the most likely or frequently occurring value. More commonly we simply use the term “mode” when describing whether a distribution has a single peak (unimodal) or two or more peaks (bimodal or multi-modal). In symmetric, unimodal distributions, the mode equals both the mean and the median. In unimodal, skewed distributions the mode is on the other side of the median from the mean. In multi-modal distributions there is either no unique highest mode, or the highest mode may well be unrepresentative of the central tendency.

The most common measure of central tendency is the mean. For skewed distribution or when there is concern about outliers, the median may be preferred.

3.2.4 Spread

Several statistics are commonly used as a measure of the **spread** of a distribution, including variance, standard deviation, and interquartile range. Spread is an indicator of how far away from the center we are still likely to find data values.

The **variance** is a standard measure of spread. It is calculated for a list of numbers, e.g., the n observations of a particular measurement labeled x_1 through x_n , based on the n **sample deviations** (or just “deviations”). Then for any data value, x_i , the corresponding deviation is $(x_i - \bar{x})$, which is the signed ($-$ for lower and $+$ for higher) distance of the data value from the mean of all of the n data values. It is not hard to prove that the sum of all of the deviations of a sample is zero.

The variance of a population is defined as the mean squared deviation. The sample formula for the variance of observed data conventionally has $n - 1$ in the denominator instead of n to achieve the property of “unbiasedness”, which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here, σ^2). The most commonly used symbol for sample variance is s^2 , and the formula is

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

which is essentially the average of the squared deviations, except for dividing by $n - 1$ instead of n . This is a measure of spread, because the bigger the deviations from the mean, the bigger the variance gets. (In most cases, squaring is better than taking the absolute value because it puts special emphasis on highly deviant values.) As usual, a sample statistic like s^2 is best thought of as a characteristic of a particular sample (thus varying from sample to sample) which is used as an estimate of the single, fixed, true corresponding parameter value from the population, namely σ^2 .

Another (equivalent) way to write the variance formula, which is particularly useful for thinking about ANOVA is

$$s^2 = \frac{SS}{df}$$

where SS is “sum of squared deviations”, often loosely called “sum of squares”, and df is “degrees of freedom” .

Because of the square, variances are always non-negative, and they have the somewhat unusual property of having squared units compared to the original data. So if the random variable of interest is a temperature in degrees, the variance has units “degrees squared”, and if the variable is area in square kilometers, the variance is in units of “kilometers to the fourth power”.

Variances have the very important property that they are additive for any number of different independent sources of variation. For example, the variance of a measurement which has subject-to-subject variability, environmental variability, and quality-of-measurement variability is equal to the sum of the three variances. This property is not shared by the “standard deviation”.

The **standard deviation** is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable. The sample standard deviation is usually represented by the symbol s . For a theoretical Gaussian distribution, we learned in the previous chapter that mean plus or minus 1, 2 or 3 standard deviations holds

68.3, 95.4 and 99.7% of the probability respectively, and this should be approximately true for real data from a Normal distribution.

The variance and standard deviation are two useful measures of spread. The variance is the mean of the squares of the individual deviations. The standard deviation is the square root of the variance. For Normally distributed data, approximately 95% of the values lie within 2 sd of the mean.

A third measure of spread is the **interquartile range**. To define IQR, we first need to define the concepts of **quartiles**. The quartiles of a population or a sample are the three values which divide the distribution or observed data into even fourths. So one quarter of the data fall below the first quartile, usually written Q_1 ; one half fall below the second quartile (Q_2); and three fourths fall below the third quartile (Q_3). The astute reader will realize that half of the values fall above Q_2 , one quarter fall above Q_3 , and also that Q_2 is a synonym for the median. Once the quartiles are defined, it is easy to define the IQR as $IQR = Q_3 - Q_1$. By definition, half of the values (and specifically the middle half) fall within an interval whose width equals the IQR. If the data are more spread out, then the IQR tends to increase, and vice versa.

The IQR is a more robust measure of spread than the variance or standard deviation. Any number of values in the top or bottom quarters of the data can be moved any distance from the median without affecting the IQR at all. More practically, a few extreme outliers have little or no effect on the IQR.

In contrast to the IQR, the **range** of the data is not very robust at all. The range of a sample is the distance from the minimum value to the maximum value: $\text{range} = \text{maximum} - \text{minimum}$. If you collect repeated samples from a population, the minimum, maximum and range tend to change drastically from sample to sample, while the variance and standard deviation change less, and the IQR least of all. The minimum and maximum of a sample may be useful for detecting outliers, especially if you know something about the possible reasonable values for your variable. They often (but certainly not always) can detect data entry errors such as typing a digit twice or transposing digits (e.g., entering 211 instead of 21 and entering 19 instead of 91 for data that represents ages of senior citizens.)

The IQR has one more property worth knowing: for normally distributed data *only*, the IQR approximately equals $4/3$ times the standard deviation. This means that for Gaussian distributions, you can approximate the sd from the IQR by calculating $3/4$ of the IQR.

The interquartile range (IQR) is a robust measure of spread.

3.2.5 Skewness and kurtosis

Two additional useful univariate descriptors are the skewness and kurtosis of a distribution. Skewness is a measure of asymmetry. Kurtosis is a measure of “peakedness” relative to a Gaussian shape. Sample estimates of skewness and kurtosis are taken as estimates of the corresponding population parameters. If the sample skewness and kurtosis are calculated along with their standard errors, we can roughly make conclusions according to the following table

where e is an estimate of skewness and u is an estimate of kurtosis, and $SE(e)$ and $SE(u)$ are the corresponding standard errors.

Skewness (e) or kurtosis (u)	Conclusion
$-2SE(e) < e < 2SE(e)$	not skewed
$e \leq -2SE(e)$	negative skew
$e \geq 2SE(e)$	positive skew
$-2SE(u) < u < 2SE(u)$	not kurtotic
$u \leq -2SE(u)$	negative kurtosis
$u \geq 2SE(u)$	positive kurtosis

For a positive skew, values far above the mode are more common than values far below, and the reverse is true for a negative skew. When a sample (or distribution) has positive kurtosis, then compared to a Gaussian distribution with the same variance or standard deviation, values far from the mean (or median or mode) are more likely, and the shape of the histogram is peaked in the middle, but with fatter tails. For a negative kurtosis, the peak is sometimes described as having “broader shoulders” than a Gaussian shape, and the tails are thinner, so that extreme values are less likely.

Skewness is a measure of asymmetry. Kurtosis is a more subtle measure of peakedness compared to a Gaussian distribution.

3.3 Univariate graphical EDA

If we are focusing on data from observation of a single variable on n subjects, i.e., a sample of size n , then in addition to looking at the various sample statistics discussed in the previous section, we also need to look graphically at the distribution of the sample. Non-graphical and graphical methods complement each other. While the non-graphical methods are quantitative and objective, they do not give a full picture of the data; therefore, graphical methods, which are more qualitative and involve a degree of subjective analysis, are also required.

3.3.1 Histograms

The only one of these techniques that makes sense for categorical data is the histogram (basically just a barplot of the tabulation of the data). A pie chart is equivalent, but not often used. The concepts of central tendency, spread and skew have no meaning for nominal categorical data. For ordinal categorical data, it sometimes makes sense to treat the data as quantitative for EDA purposes; you need to use your judgment here.

The most basic graph is the **histogram**, which is a barplot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Typically the bars run vertically with the count (or proportion) axis running vertically. To manually construct

a histogram, define the range of data for each bar (called a **bin**), count how many cases fall in each bin, and draw the bars high enough to indicate the count. For the simple data set found in EDA1.dat¹ the histogram is shown in figure 3.3.1. Besides getting the general impression of the shape of the distribution, you can read off facts like “there are two cases with data values between 1 and 2” and “there are 9 cases with data values between 2 and 3”. Generally values that fall exactly on the boundary between two bins are put in the lower bin, but this rule is not always followed.

Generally you will choose between about 5 and 30 bins, depending on the amount of data and the shape of the distribution. Of course you need to see the histogram to know the shape of the distribution, so this may be an iterative process. It is often worthwhile to try a few different bin sizes/numbers because, especially with small samples, there may sometimes be a different shape to the histogram when the bin size changes. But usually the difference is small. Figure 3.3.2 shows three histograms of the same sample from a bimodal population using three different bin widths (5, 2 and 1). If you want to try on your own, the data are in EDA2.dat². The top panel appears to show a unimodal distribution. The middle panel correctly shows the bimodality. The bottom panel incorrectly suggests many modes. There is some art to choosing bin widths, and although often the automatic choices of a program like SPSS are pretty good, they are certainly not always adequate.

It is very instructive to look at multiple samples from the same population to get a feel for the variation that will be found in histograms. Figure 3.3.3 shows histograms from multiple samples of size 50 from the same population as figure 3.3.2, while 3.3.4 shows samples of size 100. Notice that the variability is quite high, especially for the smaller sample size, and that an incorrect impression (particularly of unimodality) is quite possible, just by the bad luck of taking a particular sample.

With practice, histograms are one of the best ways to quickly learn a lot about your data, including central tendency, spread, modality, shape and outliers.

¹File available at <http://www.stat.cmu.edu/~hseltman/309/Book/data/EDA1.dat>

²File available at <http://www.stat.cmu.edu/~hseltman/309/Book/data/EDA2.dat>

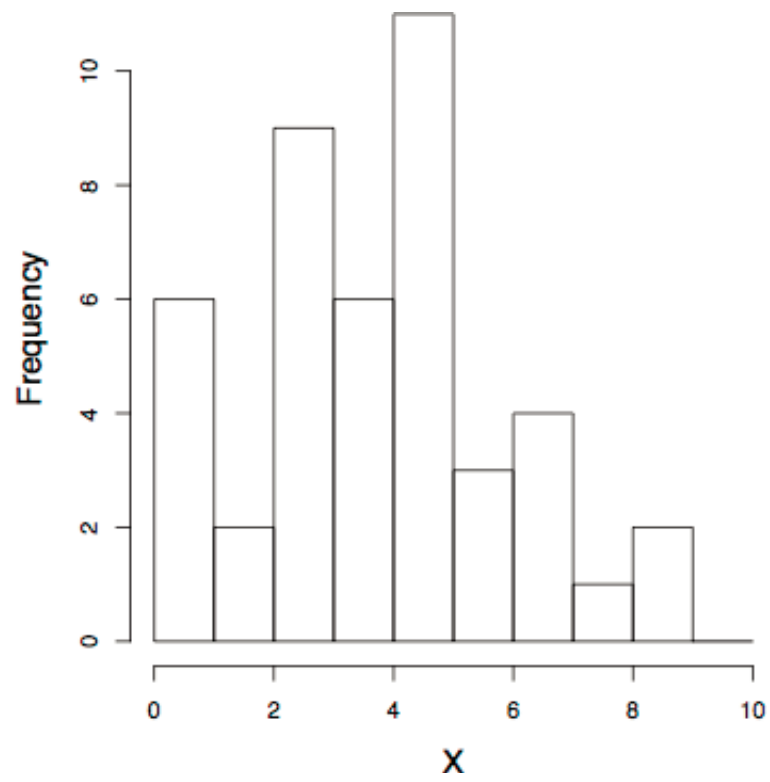


Figure 3.3.1: Histogram of EDA1.dat

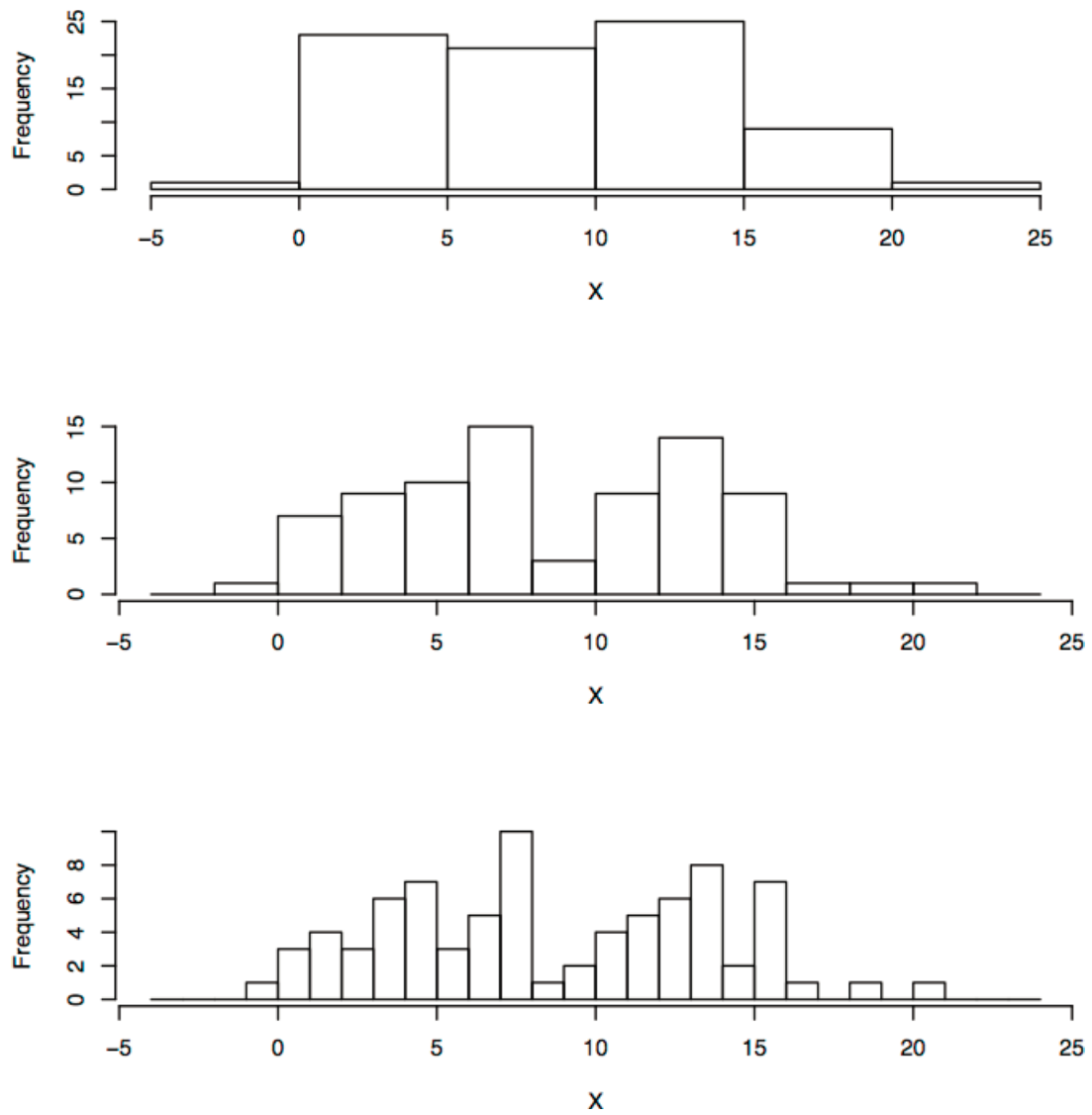


Figure 3.3.2: Histograms of EDA2.dat with different bin widths

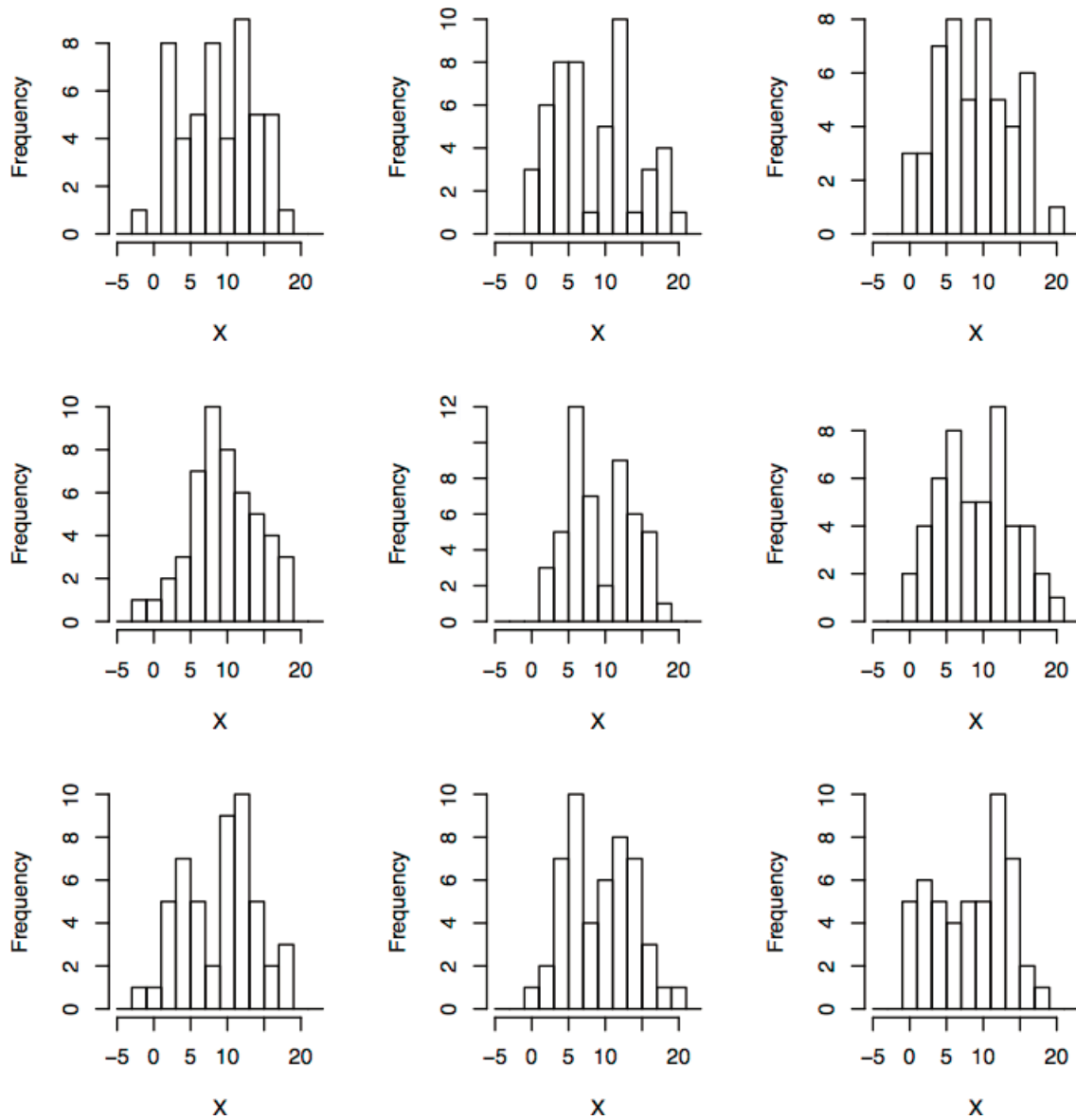


Figure 3.3.3: Histograms of multiple samples of size 50

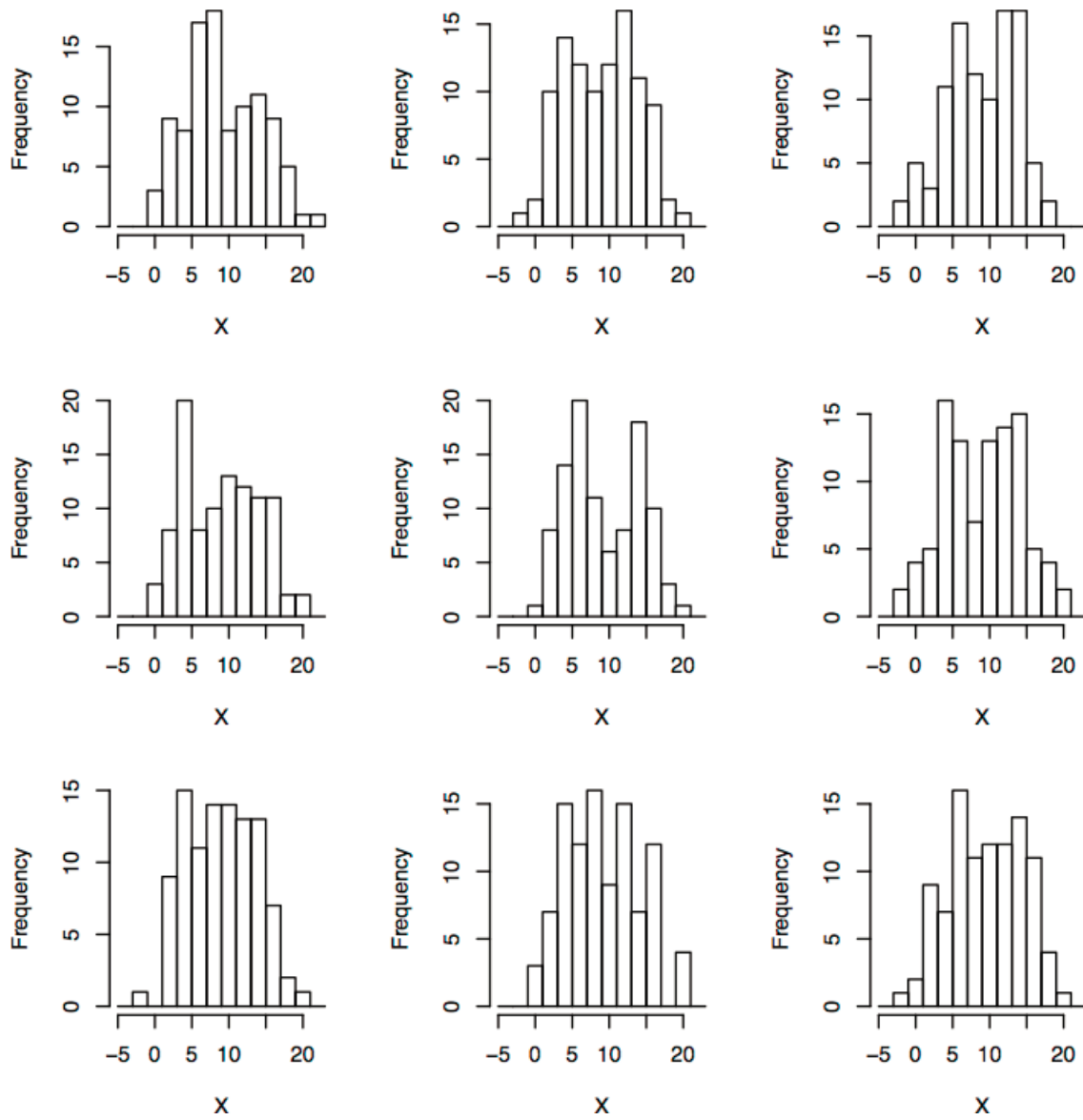


Figure 3.3.4: Histograms of multiple samples of size 100

3.3.2 Stem-and-leaf plots

A simple substitute for a histogram is a stem and leaf plot. A stem and leaf plot is sometimes easier to make by hand than a histogram, and it tends not to hide any information. Nevertheless, a histogram is generally considered better for appreciating the shape of a sample distribution than is the stem and leaf plot. Here is a stem and leaf plot for the data of figure 3.3.1:

The decimal place is at the ‘|’.

```
1|000000
2|00
3|000000000
4|000000
5|00000000000
6|000
7|0000
8|0
9|00
```

Because this particular stem and leaf plot has the decimal place at the stem, each of the 0’s in the first line represent 1.0, and each zero in the second line represents 2.0, etc. So we can see that there are six 1’s, two 2’s etc. in our data.

A stem and leaf plot shows all data values and the shape of the distribution.

3.3.3 Boxplots

Another very useful univariate graphical technique is the **boxplot**. The boxplot will be described here in its vertical format, which is the most common, but a horizontal format also is possible. An example of a boxplot is shown in figure 3.3.5, which again represents the data in EDA1.dat.

Boxplots are very good at presenting information about the central tendency, symmetry and skew, as well as outliers, although they can be misleading about aspects such as multimodality. One of the best uses of boxplots is in the form of side-by-side boxplots (see multivariate graphical analysis below).

Figure 3.3.6 is an annotated version of figure 3.3.5. Here you can see that the boxplot consists of a rectangular box bounded above and below by “hinges” that represent the quartiles Q3 and Q1 respectively, and with a horizontal “median” line through it. You can also see the upper and lower “whiskers”, and a point marking an “outlier”. The vertical axis is in the units of the quantitative variable.

Let’s assume that the subjects for this experiment are hens and the data represent the number of eggs that each hen laid during the experiment. We can read certain information directly off of the graph. The median (**not mean!**) is 4 eggs, so no more than half of the hens laid more than 4 eggs and no more than half of the hens laid less than 4 eggs. (This is based on

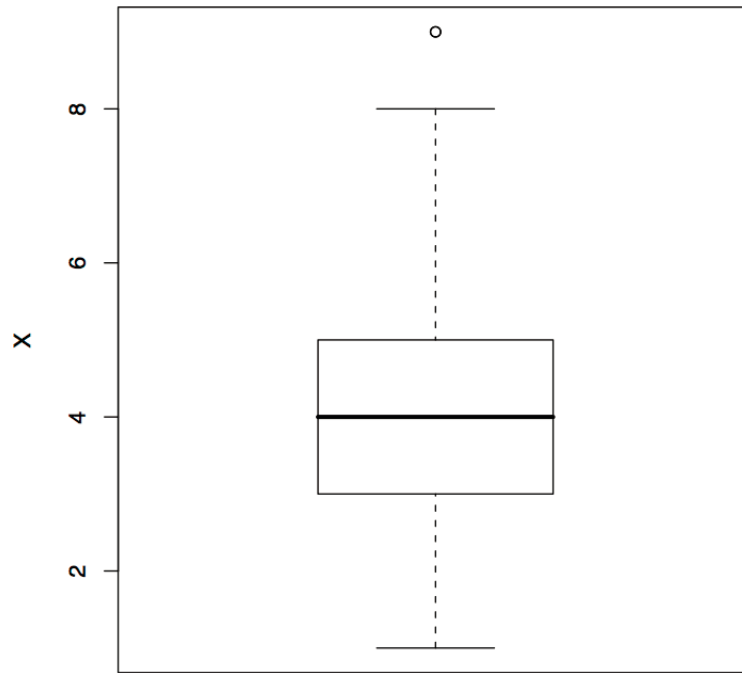


Figure 3.3.5: A boxplot of the data from EDA1.dat.

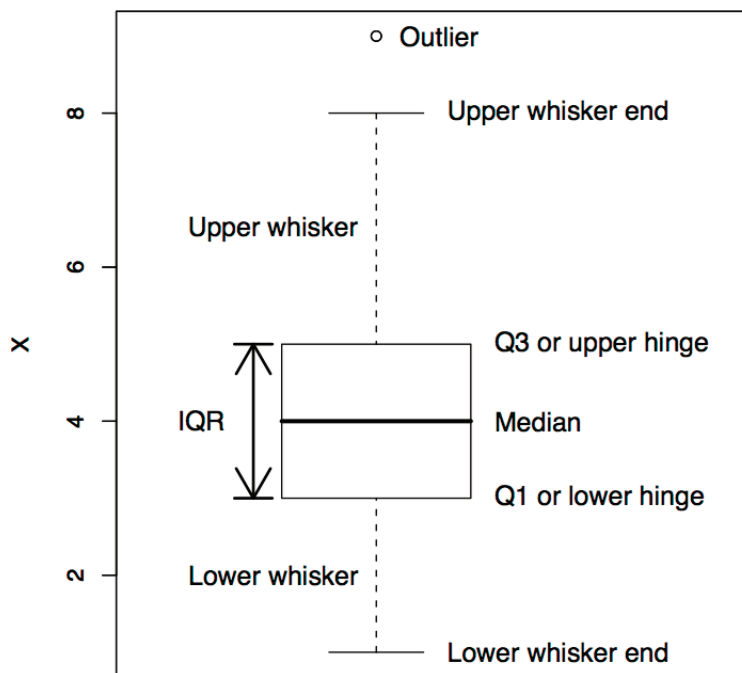


Figure 3.3.6: Annotated boxplot.

the technical definition of median; we would usually claim that half of the hens lay more or half less than 4, knowing that this may be only approximately correct.) We can also state that one quarter of the hens lay less than 1 egg and one quarter lay more than 5 eggs (again, this may not be exactly correct, particularly for small samples or a small number of different possible values). This leaves half of the hens, called the “central half”, to lay between 3 and 5 eggs, so the interquartile range (IQR) is $Q3-Q1=5-3=2$.

The interpretation of the whiskers and outliers is just a bit more complicated. Any data value more than 1.5 IQRs beyond its corresponding hinge in either direction is considered an “outlier” and is individually plotted. Sometimes values beyond 3.0 IQRs are considered “extreme outliers” and are plotted with a different symbol. In this boxplot, a single outlier is plotted corresponding to 9 eggs laid, although we know from figure 3.3.1 that there are actually two hens that laid 9 eggs. This demonstrates a general problem with plotting whole number data, namely that multiple points may be superimposed, giving a wrong impression. (Jittering, circle plots, and starplots are examples of ways to correct this problem.) This is one reason why, e.g., combining a tabulation and/or a histogram with a boxplot is better than either alone.

Each whisker is drawn out to the most extreme data point that is less than 1.5 IQRs beyond the corresponding hinge. Therefore, the whisker ends correspond to the minimum and maximum values of the data *excluding* the “outliers”.

Important: The term “outlier” is not well defined in statistics, and the definition varies depending on the purpose and situation. The “outliers” identified by a boxplot, which could be called “boxplot outliers” are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This *does not* by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. It is also important to realize that the number of boxplot outliers depends strongly on the size of the sample. In fact, for data that is perfectly Normally distributed, we expect 0.70 percent (or about 1 in 150 cases) to be “boxplot outliers”, with approximately half in either direction.

The boxplot information described above could be appreciated almost as easily if given in non-graphical format. The boxplot is useful because, with practice, all of the above and more can be appreciated at a quick glance. The additional things you should notice on the plot are the symmetry of the distribution and possible evidence of “fat tails”. Symmetry is appreciated by noticing if the median is in the center of the box and if the whiskers are the same length as each other. For this purpose, as usual, the smaller the dataset the more variability you will see from sample to sample, particularly for the whiskers. In a skewed distribution we expect to see the median pushed in the direction of the shorter whisker. If the longer whisker is the top one, then the distribution is positively skewed (or skewed to the right, because higher values are on the right in a histogram). If the lower whisker is longer, the distribution is negatively skewed (or left skewed.) In cases where the median is closer to the longer whisker it is hard to draw a conclusion.

The term **fat tails** is used to describe the situation where a histogram has a lot of values far from the mean relative to a Gaussian distribution. This corresponds to positive kurtosis. In a boxplot, many outliers (more than the 1/150 expected for a Normal distribution) suggests fat tails (positive kurtosis), or possibly many data entry errors. Also, short whiskers suggest negative kurtosis, at least if the sample size is large.

Boxplots are excellent EDA plots because they rely on robust statistics like median and IQR rather than more sensitive ones such as mean and standard deviation. With boxplots it is easy to compare distributions (usually for one variable at different levels of another; see multivariate graphical EDA, below) with a high degree of reliability because of the use of these robust statistics.

It is worth noting that some (few) programs produce boxplots that do not conform to the definitions given here.

Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

3.3.4 Quantile-normal plots

The final univariate graphical EDA technique is the most complicated. It is called the **quantile-normal** or **QN plot** or more generally the **quantile-quantile** or **QQ plot**. It is used to see how well a particular sample follows a particular theoretical distribution. Although it can be used for any theoretical distribution, we will limit our attention to seeing how well a sample of data of size n matches a Gaussian distribution with mean and variance equal to the sample mean and variance. By examining the quantile-normal plot we can detect left or right skew, positive or negative kurtosis, and bimodality.

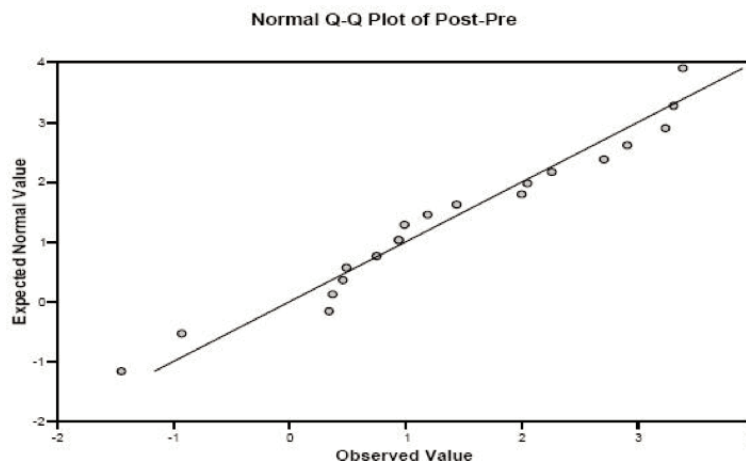


Figure 3.3.7: A quantile-normal plot.

The example shown in figure 3.3.7 shows 20 data points that are approximately normally distributed. **Do not confuse a quantile-normal plot with a simple scatter plot of two variables.** The title and axis labels are strong indicators that this is a quantile-normal plot. For many computer programs, the word “quantile” is also in the axis labels.

Many statistical tests have the assumption that the outcome for any fixed set of values of the explanatory variables is approximately normally distributed, and that is why QN plots are useful: if the assumption is grossly violated, the p-value and confidence intervals of those tests are wrong. As we will see in the ANOVA and regression chapters, the most important situation

where we use a QN plot is not for EDA, but for examining something called “residuals”. For basic interpretation of the QN plot you just need to be able to distinguish the two situations of “OK” (points fall randomly around the line) versus “non-normality” (points follow a strong curved pattern rather than following the line).

If you are still curious, here is a description of how the QN plot is created. Understanding this will help to understand the interpretation, but is not required in this course. Note that some programs swap the x and y axes from the way described here, but the interpretation is similar for all versions of QN plots. Consider the 20 values observed in this study. They happen to have an observed mean of 1.37 and a standard deviation of 1.36. Ideally, 20 random values drawn from a distribution that has a true mean of 1.37 and sd of 1.36 have a perfect bell-shaped distribution and will be spaced so that there is equal area (probability) in the area around each value in the bell curve.

In figure 3.3.8 the dotted lines divide the bell curve up into 20 equally probable zones, and the 20 points are at the probability mid-points of each zone. These 20 points, which are more tightly packed near the middle than in the ends, are used as the “Expected Normal Values” in the QN plot of our actual data.

In summary, the sorted actual data values are plotted against “Expected Normal Values”, and some kind of diagonal line is added to help direct the eye towards a perfect straight line on the quantile-normal plot that represents a perfect bell shape for the observed data.

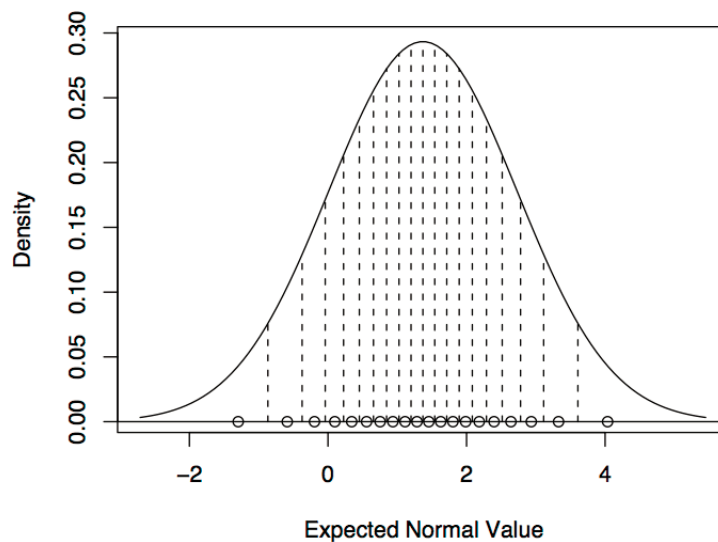


Figure 3.3.8: A way to think about QN plots.

The interpretation of the QN plot is given here. If the axes are reversed in the computer package you are using, you will need to correspondingly change your interpretation. If all of the points fall on or nearly on the diagonal line (with a random pattern), this tells us that a histogram of the variable will show a bell shaped (Normal or Gaussian) distribution.

Figure 3.3.9 shows all of the points basically on the reference line, but there are several vertical bands of points. Because the x-axis is “observed values”, these bands indicate ties, i.e., multiple points with the same values. And all of the observed values are at whole numbers. So either the data are rounded or we are looking at a discrete quantitative (counting) variable. Either way, the data appear to be nearly normally distributed.

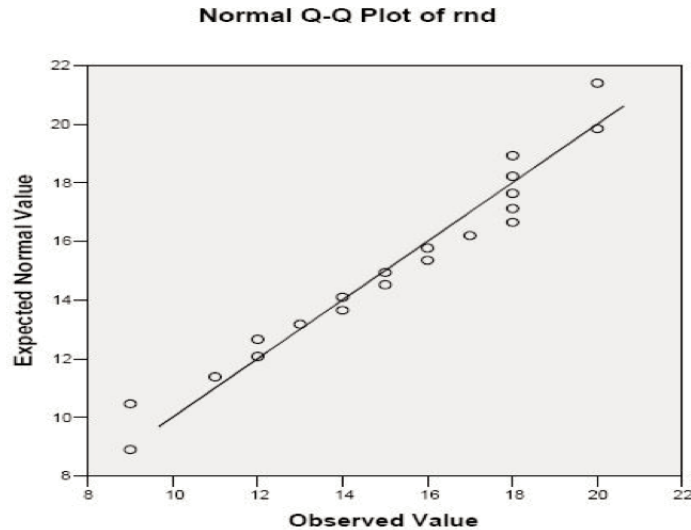


Figure 3.3.9: Quantile-normal plot with ties.

In figure 3.3.10 note that we have many points in a row that are on the same side of the line (rather than just bouncing around to either side), and that suggests that there is a real (non-random) deviation from Normality. The best way to think about these QN plots is to look at the low and high ranges of the Expected Normal Values. In each area, see how the observed values deviate from what is expected, i.e., in which “x” (Observed Value) direction the points appear to have moved relative to the “perfect normal” line. Here we observe values that are too high in both the low and high ranges. So compared to a perfect bell shape, this distribution is pulled asymmetrically towards higher values, which indicates positive skew.

Also note that if you just *shift* a distribution to the right (without disturbing its symmetry) rather than skewing it, it will maintain its perfect bell shape, and the points remain on the diagonal reference line of the quantile-normal curve.

Of course, we can also have a distribution that is skewed to the left, in which case the high and low range points are shifted (in the Observed Value direction) towards lower than expected values. In figure 3.3.11 the high end points are shifted too high and the low end points are shifted too low. These data show a positive kurtosis (fat tails). The opposite pattern is a negative kurtosis in which the tails are too “thin” to be bell shaped.

In figure 3.3.12 there is a single point that is off the reference line, i.e. shifted to the right of where it should be. (Remember that the pattern of locations on the Expected Normal Value axis is fixed for any sample size, and only the position on the Observed axis varies depending on the observed data.) This pattern shows nearly Gaussian data with one “high outlier”.

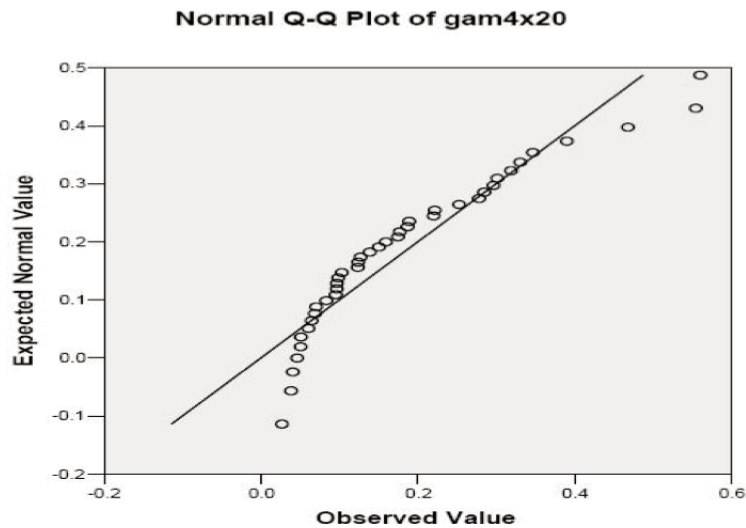


Figure 3.3.10: Quantile-normal plot with ties.

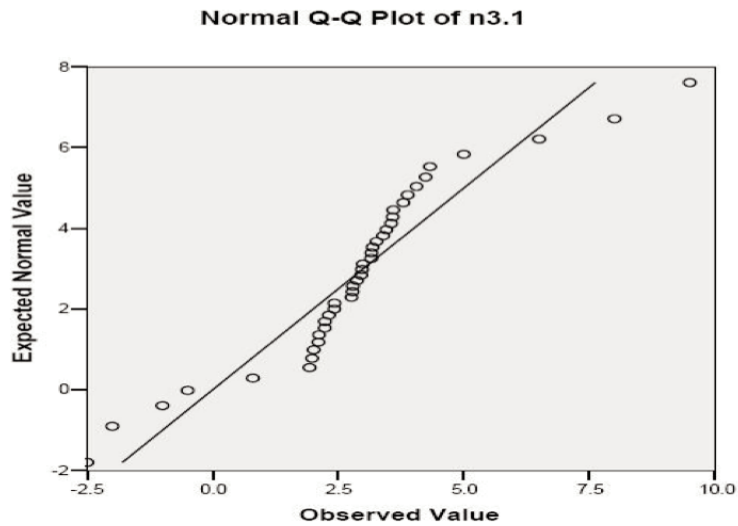


Figure 3.3.11: Quantile-normal plot showing fat tails.

Finally, figure 3.3.13 looks a bit similar to the “skew left” pattern, but the most extreme points tend to return to the reference line. This pattern is seen in bi-modal data, e.g. this is what we would see if we would mix strength measurements from controls and muscular dystrophy patients.

Quantile-Normal plots allow detection of non-normality and diagnosis of skeweness and kurtosis.

3.4 Multivariate non-graphical EDA

Multivariate non-graphical EDA techniques generally show the relationship between two or more variables in the form of either cross-tabulation or statistics.

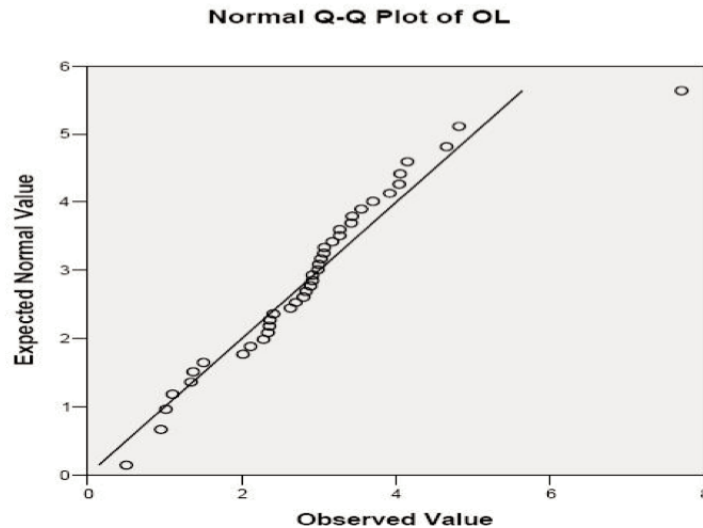


Figure 3.3.12: Quantile-normal plot showing a high outlier.

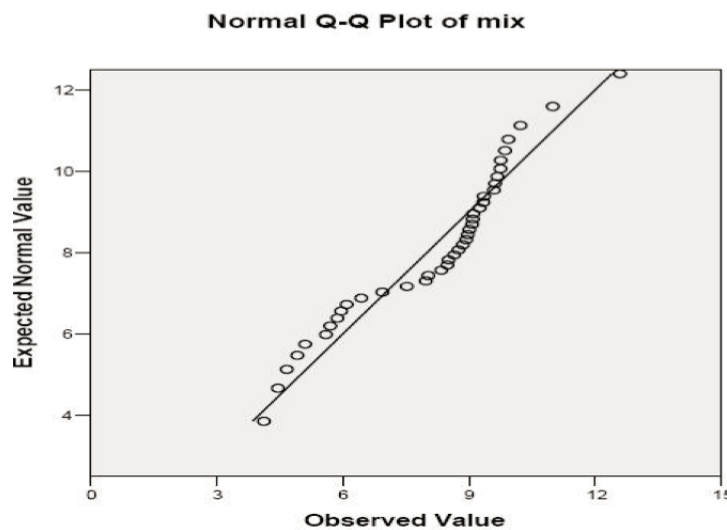


Figure 3.3.13: Quantile-normal plot showing bimodality.

3.4.1 Cross-tabulation

For categorical data (and quantitative data with only a few different values) an extension of tabulation called **cross-tabulation** is very useful. For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels. The two variables might be both explanatory, both outcome, or one of each. Depending on the goals, row percentages (which add to 100% for each row), column percentages (which add to 100% for each column) and/or cell percentages (which add to 100% over all cells) are also useful.

Here is an example of a cross-tabulation. Consider the data in table 3.1. For each subject we observe sex and age as categorical variables.

Table 3.2 shows the cross-tabulation.

We can easily see that the total number of young females is 2, and we can calculate, e.g., the corresponding cell percentage is $\frac{2}{11} \times 100 = 18.2\%$, the row percentage is $\frac{2}{5} \times 100 = 40.0\%$, and the column percentage is $\frac{2}{7} \times 100 = 28.6\%$.

Cross-tabulation can be extended to three (and sometimes more) variables by making separate two-way tables for two variables at each level of a third variable.

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

Table 3.1: Sample Data for Cross-tabulation

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

Table 3.2: Cross-tabulation of Sample Data

For example, we could make separate age by gender tables for each education level.

Cross-tabulation is the basic bivariate non-graphical EDA technique.

3.4.2 Correlation for categorical data

Another statistic that can be calculated for two categorical variables is their correlation. But there are many forms of correlation for categorical variables, and that material is currently beyond the scope of this book.

3.4.3 Univariate statistics by category

For one categorical variable (usually explanatory) and one quantitative variable (usually outcome), it is common to produce some of the standard univariate non- graphical statistics for the quantitative variables separately for each level of the categorical variable, and then compare the statistics across levels of the categorical variable. Comparing the means is an informal version of ANOVA. Comparing medians is a robust informal version of one-way ANOVA. Comparing measures of spread is a good informal test of the assumption of equal variances needed for valid analysis of variance.

Especially for a categorical explanatory variable and a quantitative outcome variable, it is useful to produce a variety of univariate statistics for the quantitative variable at each level of the categorical variable.

3.4.4 Correlation and covariance

For two quantitative variables, the basic statistics of interest are the sample co- variance and/or sample correlation, which correspond to and are estimates of the corresponding population parameters. The sample covariance is a measure of how much two variables “co-vary”, i.e., how much (and in what direction) should we expect one variable to change when the other changes.

Sample covariance is calculated by computing (signed) deviations of each measurement from the average of all measurements for that variable. Then the deviations for the two measurements are multiplied together separately for each subject. Finally these values are averaged (actually summed and divided by $n-1$, to keep the statistic unbiased). Note that the units on sample covariance are the products of the units of the two variables.

Positive covariance values suggest that when one measurement is above the mean the other will probably also be above the mean, and vice versa. Negative covariances suggest that when one variable is above its mean, the other is below its mean. And covariances near zero suggest that the two variables vary independently of each other.

Technically, independence implies zero correlation, but the reverse is not necessarily true.

Covariances tend to be hard to interpret, so we often use correlation instead. The correlation has the nice property that it is always between -1 and $+1$, with -1 being a “perfect” negative linear correlation, $+1$ being a perfect positive linear correlation and 0 indicating that X and Y are uncorrelated. The symbol r or $r_{x,y}$ is often used for sample correlations.

Subject ID	Age	Strength
GW	38	20
JA	62	15
TJ	22	30
JMA	38	21
JMO	45	18
JQA	69	12
AJ	75	14
MVB	38	28
WHH	80	9
JT	21	22
JKP	51	20

Table 3.3: Covariance Sample Data

The general formula for sample covariance is

$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

It is worth noting that $Cov(X, Y) = Var(X)$..

If you want to see a “manual example” of calculation of sample covariance and correlation consider an example using the data in table ?. For each subject we observe age and a strength measure.

Table ? shows the calculation of covariance. The mean age is 50 and the mean strength is 19, so we calculate the deviation for age-50 and deviation for strength and strength-19. Then we find the product of the deviation and add them up. This total is 1106 and since $n=11$, the covariance of x and y is $-\frac{1106}{10} = -110.6$. The fact that the covariance is negative indicates that as age goes up strength tends to go down (and vice versa).

The formula for the sample correlation is

$$cor(X, Y) = \frac{Cov(X, Y)}{s_x s_y}$$

where s_x is the standard deviation of X and s_y is the standard deviation of Y .

In this example, $s_x = 18.96$, $s_y = 6.39$, so $r = \frac{-110.6}{18.96 \cdot 6.39} = -0.913$. This is a strong negative correlation.

Subject ID	Age	Strength	Age-50	Str-19	Product
GW	38	20	-12	+1	-12
JA	62	15	+12	-4	-48
TJ	22	30	-28	+11	-308
JMA	38	21	-12	+2	-24
JMO	45	18	-5	-1	+5
JQA	69	12	+19	-7	-133
AJ	75	14	+25	-5	-125
MVB	38	28	-12	+ 9	-108
WHH	80	9	+30	-10	-300
JT	21	22	-18	+ 3	-54
JKP	51	20	+1	+1	+1
Total			0	0	-1106

Table 3.4: Covariance Calculation

3.4.5 Covariance and correlation matrices

When we have many quantitative variables the most common non-graphical EDA technique is to calculate all of the pairwise covariances and/or correlations and assemble them into a matrix. Note that the covariance of X with X is the variance of X and the correlation of X with X is 1.0. For example the covariance matrix of table 3.5 tells us that the variances of X , Y , and Z are 5, 7, and 4 respectively, the covariance of X and Y is 1.77, the covariance of X and Z is -2.24, and the covariance of Y and Z is 3.17.

Similarly the correlation matrix in figure 3.6 tells us that the correlation of X and Y is 0.3, the correlation of X and Z is -0.5. and the correlation of Y and Z is 0.6.

	X	Y	Z
X	5.00	1.77	-2.24
Y	1.77	7.0	3.17
Z	-2.24	3.17	4.0

Table 3.5: A covariance Matrix

	X	Y	Z
X	1.0	0.3	-0.5
Y	0.3	1.0	0.6
Z	-0.5	0.6	1.0

Table 3.6: A Correlation Matrix

The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

3.5 Multivariate graphical EDA

There are few useful techniques for graphical EDA of two categorical random variables. The only one used commonly is a grouped barplot with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable.

3.5.1 Univariate graphs by category

When we have one categorical (usually explanatory) and one quantitative (usually outcome) variable, graphical EDA usually takes the form of “conditioning” on the categorical random variable. This simply indicates that we focus on all of the subjects with a particular level of the categorical random variable, then make plots of the quantitative variable for those subjects. We repeat this for each level of the categorical variable, then compare the plots. The most commonly used of these are **side-by-side boxplots**, as in figure 3.5.1. Here we see the data from EDA3.dat³, which consists of strength data for each of three age groups. You can see the downward trend in the median as the ages increase. The spreads (IQRs) are similar for the three groups. And all three groups are roughly symmetrical with one high strength outlier in the youngest age group.

Side-by-side boxplots are the best graphical EDA technique for examining the relationship between a categorical variable and a quantitative variable, as well as the distribution of the quantitative variable at each level of the categorical variable.

3.5.2 Scatterplots

For two quantitative variables, the basic graphical EDA technique is the scatterplot which has one variable on the x-axis, one on the y-axis and a point for each case in your dataset. *If one variable is explanatory and the other is outcome, it is a very, very strong convention to put the outcome on the y (vertical) axis.*

One or two additional categorical variables can be accommodated on the scatterplot by encoding the additional information in the symbol type and/or color.

An example is shown in figure 3.5.2. Age vs. strength is shown, and different colors and symbols are used to code political party and gender.

³File available at <http://www.stat.cmu.edu/~hseltman/309/Book/data/EDA3.dat>

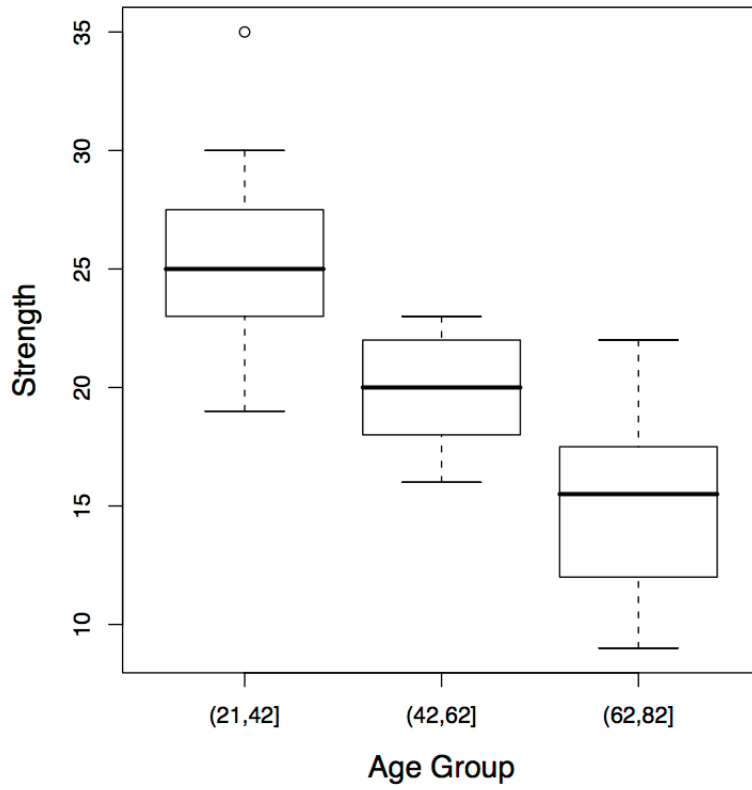


Figure 3.5.1: Side-by-side Boxplot of EDA3.dat.

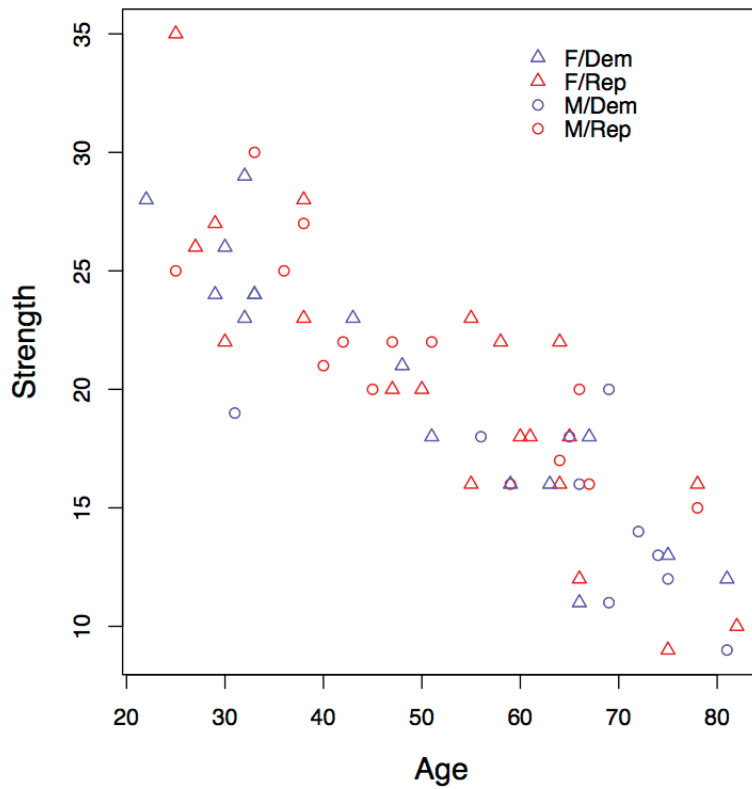


Figure 3.5.2: Scatterplot with two additional variables.

In a nutshell: You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science – it is a very important art!

3.6 A note on degrees of freedom

Degrees of freedom are numbers that characterize specific distributions in a family of distributions. Often we find that a certain family of distributions is needed in a some general situation, and then we need to calculate the degrees of freedom to know which specific distribution within the family is appropriate.

The most common situation is when we have a particular statistic and want to know its sampling distribution. If the sampling distribution falls in the “t” family as when performing a t-test, or in the “F” family when performing an ANOVA, or in several other families, we need to find the number of degrees of freedom to figure out which particular member of the family actually represents the desired sampling distribution. One way to think about degrees of freedom for a statistic is that they represent the number of independent pieces of information that go into the calculation of the statistic,

Consider 5 numbers with a mean of 10. To calculate the variance of these numbers we need to sum the squared deviations (from the mean). It really doesn’t matter whether the mean is 10 or any other number: as long as all five deviations are the same, the variance will be the same. This make sense because variance is a pure measure of spread, not affected by central tendency. But by mathematically rearranging the definition of mean, it is not too hard to show that the sum of the deviations (not squared) is always zero. Therefore, the first four deviations can (freely) be any numbers, but then the last one is forced to be the number that makes the deviations add to zero, and we are not free to choose it. It is in this sense that five numbers used for calculating a variance or standard deviation have only four degrees of freedom (or independent useful pieces of information). In general, a variance or standard deviation calculated from n data values and one mean has $n - 1$ df.

Another example is the “pooled” variance from k independent groups. If the sizes of the groups are n_1 through n_k , then each of the k individual variance estimates is based on deviations from a different mean, and each has one less degree of freedom than its sample size, e.g., $n_i - 1$ for group i . We also say that each numerator of a variance estimate, e.g., SS_i , has $n_i - 1$ df. The pooled estimate of variance is

$$s_{\text{pooled}}^2 = \frac{SS_1 + \cdots + SS_k}{df_1 + \cdots + df_k}$$

and we say that both the numerator SS and the entire pooled variance has $df_1 + \cdots + df_k$ degrees of freedom, which suggests how many independent pieces of information are available for the calculation.