

Chapter 1

INTRODUCTION

1.1. Outline

The subject of the following ten chapters can be divided into two main parts:

- Theory of Probability: Chapters 2–4
- Statistics: Chapters 5–11

The theory of probability is needed only to provide the necessary tools for statistics, which forms the main body of the course.

Chapters 5 and 6 define two general approaches to the choice of estimators: the *information* approach and the *decision theory* approach. The former consists essentially in maximizing the amount of information in the estimate, whereas the latter is based on minimizing the loss involved in making the wrong decision about the parameter value. In the limit of large data samples, the two approaches are equivalent, but where they differ we will try to point out the distinction.

Estimation of parameters is divided into three chapters, 7 and 8 dealing with point estimation, theory and practice, and 9 dealing with interval estimation. Tests of hypotheses are divided into general testing, Chapter 10, and goodness-of-fit tests, Chapter 11.

Our reference policy is as follows. We quote literature when we have omitted the proof of an important result, or when we want to give hints for further

reading. We do not usually attempt to give credit to original results. In the text the references take the form

[First author, Volume*, Chapter*, page*]

* if necessary.

At the end of the course before the subject index, the literature references are ordered alphabetically after the first author.

1.2. Language

Statistics, like any other branch of learning, has its own terminology which one has to become accustomed to. Certain confusion may, however, arise when the same term has different meaning in statistics and in physics, or when the same concept has different names. In the former case we usually imply the statistical meaning (obliging the physicist to recognize and learn the difference); in the second case we often choose the physical term.

An example of the first kind is the following:

| Physicists say | Statisticians say |
|-------------------|----------------------|
| Determine | Estimate |
| Estimate | Guess |

Thus the word estimate has different meaning in physics and in statistics. We use it as statisticians do. (We use three chapters to explain what statisticians mean thereby).

An example of the second kind is “the demographic approach” to experimental physics. Much of statistics has been developed in connection with population studies (sociology, medicine, agriculture) and at the production line (industrial quality control). Then one is not able to study the whole population, so one “draws a sample”. And the population exists in a real sense.

In experimental physics, the set of all measurements under study corresponds to the “sample”. Increasing the number of measurements, the physicist increases the “size of the sample”, but he never attains the “population”. Thus the “population” is an underlying abstraction which does not exist in any real sense. These “demographic” terms are therefore to some extent inappropriate and unnecessary, and we try to avoid some of them:

| For the “demographic” term | we use the physics term |
|-------------------------------|----------------------------|
| Sample | Data (set) |
| Draw a sample | Observe, measure |
| Sample of size N | N observations |
| Population | Observable space |

Still, one has to be able to distinguish between, say, the mean of the data at hand, and the mean if the data set were infinite. When this distinction is necessary, we use sample mean, sample variance, etc. as contrasted to parent mean, parent variance, etc., or mean and variance of the underlying distribution. Thus

$$\text{Parent mean} = \text{Mean of the underlying distribution} = \text{Population mean.}$$

We avoid the physical term “error”, which is misleading, and use instead “variance of estimate”, “confidence interval” or “interval estimate”. We also try to avoid the words “precision” and “accuracy”, because they are not well defined.

In many books on statistics one finds whole chapters dealing with the “propagation of errors”. Such a term, in our minds, is confusing. The corresponding notion here is “change of variables”.

Other topics which may seem to have got lost, may also sometimes be refound under other names. For instance, the term “regression analysis” is never used, but the techniques are treated under least-squares fits of linear models.

1.3. Two Philosophies

Unfortunately, statisticians do not agree on basic principles. They can crudely be divided into two schools: Bayesian and frequentist (or classical). The name Bayesian derives from the extended use of Bayes theorem in the former group. We try to present the main results from both approaches.

The Bayesian approach is closer to everyday reasoning, where probability is interpreted as a *degree of belief* that something will happen, or that a parameter will have a given value.

The frequentist approach is closer to scientific reasoning, where probability means the relative frequency of something happening. This makes it more objective, since it can be determined independently of the observer, but restricts its application to repeatable phenomena. In particular, one can define

the frequentist probability for observing data (which are random), but not for the true value of a parameter (which is fixed, even if unknown).

In the areas of parameter estimation and hypothesis testing, numerical results tend to be the same for the two approaches in the asymptotic regime, that is, when there are a lot of data, and statistical uncertainties are small compared with the distance to the nearest physical boundary. However, exact results require for each approach information which is not allowed by the other:

- Exact frequentist results require as input the probabilities of observing all data, including both the data actually observed and that which could have been observed (the Monte Carlo). This violates an important principle in Bayesian theory, and is not allowed in the Bayesian method.
- Exact Bayesian results require as input the prior beliefs of the physicist doing the analysis. This is necessarily subjective, and is not allowed in the frequentist method.

In the area of goodness-of-fit testing (testing of a single hypothesis, where no alternative hypothesis is specified) it is essentially impossible to obtain any results in the Bayesian approach, so that is the traditional bastion of classical statistics.

On the other hand, decision theory, because of its fundamentally subjective nature, is the domain of Bayesian methodology.

In this book we largely follow a classical (frequentist) approach because we feel this is more appropriate for the reporting of experimental results. We do however develop also the main ideas of Bayesian statistics and try to point out the differences wherever possible. It can be a great help in understanding each approach to see how the difficult problems are handled in the other approach.

1.4. Notation

Roman letters

| | |
|---------------------------|--|
| $b(\hat{\theta})$ | bias of estimate $\hat{\theta}$ |
| $\text{corr}(X, Y), \rho$ | correlation between random variables X and Y |
| $\text{cov}(X, Y)$ | covariance of random variables X and Y |
| D_N, D_{NM}, D_N^\pm | Kolmogorov statistic |
| $e(X, Y')$ | detection efficiency |
| $E(X)$ | expectation |
| $f(X)$ | probability density function (p.d.f.) |
| $F(X)$ | cumulative distribution of $f(X)$ |

| | |
|--|---|
| $g(X)$ | probability density function |
| $G(X)$ | generating function |
| H, H_i | hypothesis |
| H_0 | hypothesis under test, “null hypothesis” |
| $\mathcal{L}_X(\boldsymbol{\theta}), \mathcal{L}_N(\boldsymbol{\theta})$ | information matrix |
| K | non-centrality parameter |
| $K(t)$ | cumulant generating function |
| K_r | cumulant |
| ℓ | likelihood ratio |
| L_p | norm of power p |
| $L(\mathbf{X} \theta) = L(\theta)$ | likelihood function |
| $L(\theta, d)$ | loss function |
| N | number of random variables (events, experiments) |
| $N(\mu, \sigma^2)$ | Normal distribution of mean μ and variance σ^2 |
| $O(N^{-1})$ | term of order N^{-1} or less |
| $p(\theta), 1 - \beta$ | power of test |
| $P(A)$ | probability that A is true |
| $P(A B)$ | conditional probability that A is true, given B |
| p.d.f. | probability density function |
| Q^2 | quadratic form, covariance form |
| $r(X, X')$ | resolution function |
| $s^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$ | unbiased estimate of variance |
| $S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$ | sample variance |
| S_N | sum of N random variables, distribution function for order statistics |
| t, T | statistic |
| tr | trace (of matrix) |
| $V(X), \sigma_X^2$ | variance |
| $\mathcal{V}(\mathbf{X})$ | covariance matrix |
| $\mathcal{V}_{(rs)}$ | submatrix of \mathcal{V} having r rows and s columns |
| w_α | critical test region of significance α |
| w_i | weight |
| W | space of test statistic |
| W^2 | test statistic |

| | |
|---|--|
| X, X_i | random variables |
| $X_{(N)}^2 = \sum_{i=1}^N X_i^2$ | sum of squares of random variables |
| Y, Z, Y_i, Z_i | random variables |
| $X_\alpha, Y_\alpha, Z_\alpha$ | α -point of $f(X)$, etc. [defined by $F(X_\alpha) = \alpha$] |
| Greek letters | |
| α | confidence level, significance level, loss |
| β | contamination, confidence level |
| γ_1 | skewness |
| γ_2 | kurtosis |
| $\delta(X)$ | δ “function” of Dirac |
| θ, θ_i | theoretical parameter |
| θ_0 | true value of θ , null hypothesis value of θ |
| λ | maximum likelihood ratio |
| λ_α | α -point of Normal distribution [defined by $\Phi(\lambda_\alpha) = \alpha$] |
| μ, μ' | mean value |
| μ_n, μ'_n | moment of order n |
| ν, ν' | degree of freedom |
| ν_n, ν'_n | moment of order n |
| $\pi(\theta)$ | prior density |
| $\pi(\theta) \mathbf{X}$ | posterior density |
| ρ | correlation coefficient |
| σ^2 | variance |
| $\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$ | unbiased estimate of variance, given the mean μ |
| $\chi^2(N)$ | chi-square distribution of N degrees of freedom |
| $\phi(t)$ | characteristic function |
| $\Phi(X)$ | Normal probability integral |
| Ω | space of random variable |
| Symbols | |
| bar (e.g. $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$) | average (e.g. of random variables X_1, \dots, X_N) |
| hat (e.g. $\hat{\theta}$) | estimate (e.g. of parameter θ) |

| | |
|--------------------------------------|--|
| bold (e.g. \mathbf{X}) | vector (e.g. with components X_1, \dots, X_N) |
| tilde (e.g. $\tilde{\mathcal{L}}$) | matrix (e.g. covariance matrix) |
| upper U (e.g. θ^U) | upper confidence bound (e.g. of parameter θ) |
| lower L (e.g. θ_L) | lower confidence bound (e.g. of parameter θ) |
| T (e.g. \mathcal{A}^T) | transpose of matrix (\mathcal{A}) |
| $[\theta_a, \theta_b]$ | interval $\theta_a \leq \theta \leq \theta_b$ |
| $\binom{p}{q} = \frac{p!}{q!(p-q)!}$ | binomial coefficient |