Part IB — Statistics
Definitions

Based on lectures by D. Spiegelhalter
Notes taken by Dexter Chua

Lent 2015

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

Estimation
Review of distribution and density functions, parametric families. Examples: binomial, Poisson, gamma. Sufficiency, minimal sufficiency, the Rao-Blackwell theorem. Maximum likelihood estimation. Confidence intervals. Use of prior distributions and Bayesian inference. [5]

Hypothesis testing
Simple examples of hypothesis testing, null and alternative hypothesis, critical region, size, power, type I and type II errors, Neyman-Pearson lemma. Significance level of outcome. Uniformly most powerful tests. Likelihood ratio, and use of generalised likelihood ratio to construct test statistics for composite hypotheses. Examples, including t-tests and F-tests. Relationship with confidence intervals. Goodness-of-fit tests and contingency tables. [4]

Linear models
Derivation and joint distribution of maximum likelihood estimators, least squares, Gauss-Markov theorem. Testing hypotheses, geometric interpretation. Examples, including simple linear regression and one-way analysis of variance. Use of software. [7]

Contents

0 Introduction

1 Estimation

1.1 Estimators

Definition (Statistic)**.** A *statistic* is an estimate of θ. It is a function T of the data. If we write the data as $\mathbf{x} = (x_1, \cdots, x_n)$, then our estimate is written as $\hat{\theta} = T(\mathbf{x})$. $T(\mathbf{X})$ is an *estimator* of θ.

The distribution of $T = T(\mathbf{X})$ is the *sampling distribution* of the statistic.

Definition (Bias)**.** Let $\hat{\theta} = T(\mathbf{X})$ be an estimator of θ. The *bias* of $\hat{\theta}$ is the difference between its expected value and true value.

$$\mathrm{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta.$$

Note that the subscript $_\theta$ does not represent the random variable, but the thing we want to estimate. This is inconsistent with the use for, say, the probability mass function.

An estimator is *unbiased* if it has no bias, i.e. $\mathbb{E}_\theta(\hat{\theta}) = \theta$.

1.2 Mean squared error

Definition (Mean squared error)**.** The *mean squared error* of an estimator $\hat{\theta}$ is $\mathbb{E}_\theta[(\hat{\theta} - \theta)^2]$.

Sometimes, we use the *root mean squared error*, that is the square root of the above.

1.3 Sufficiency

Definition (Sufficient statistic)**.** A statistic T is *sufficient* for θ if the conditional distribution of $\mathbf{X}$ given T does not depend on θ.

Definition (Minimal sufficiency)**.** A sufficient statistic $T(\mathbf{X})$ is *minimal* if it is a function of every other sufficient statistic, i.e. if $T'(\mathbf{X})$ is also sufficient, then $T'(\mathbf{X}) = T'(\mathbf{Y}) \Rightarrow T(\mathbf{X}) = T(\mathbf{Y})$.

1.4 Likelihood

Definition (Likelihood)**.** For any given $\mathbf{x}$, the *likelihood* of θ is $\mathrm{like}(\theta) = f_{\mathbf{X}}(\mathbf{x} \mid \theta)$, regarded as a function of θ. The *maximum likelihood estimator* (mle) of θ is an estimator that picks the value of θ that maximizes $\mathrm{like}(\theta)$.

1.5 Confidence intervals

Definition. A $100\gamma\%$ $(0 < \gamma < 1)$ *confidence interval* for θ is a random interval $(A(\mathbf{X}), B(\mathbf{X}))$ such that $\mathbb{P}(A(\mathbf{X}) < \theta < B(\mathbf{X})) = \gamma$, no matter what the true value of θ may be.

1.6 Bayesian estimation

Definition (Prior and posterior distribution)**.** The *prior distribution* of θ is the probability distribution of the value of θ before conducting the experiment. We usually write as $\pi(\theta)$.

The *posterior distribution* of θ is the probability distribution of the value of θ given an outcome of the experiment $\mathbf{x}$. We write as $\pi(\theta \mid \mathbf{x})$.

Definition (Bayes estimator)**.** The *Bayes estimator* $\hat{\theta}$ is the estimator that minimises the expected posterior loss.

2 Hypothesis testing

Definition (Simple and composite hypotheses)**.** A *simple hypothesis* H specifies f completely (e.g. $H_0 : \theta = \frac{1}{2}$). Otherwise, H is a *composite hypothesis.*

2.1 Simple hypotheses

Definition (Critical region)**.** For testing H_0 against an alternative hypothesis H_1, a test procedure has to partition $\mathcal{X}^n$ into two disjoint exhaustive regions C and $\bar{C}$, such that if $\mathbf{x} \in C$, then H_0 is rejected, and if $\mathbf{x} \in \bar{C}$, then H_0 is not rejected. C is the *critical region.*

Definition (Type I and II error)**.**

(i) *Type I error*: reject H_0 when H_0 is true.

(ii) *Type II error*: not rejecting H_0 when H_0 is false.

Definition (Size and power)**.** When H_0 and H_1 are both simple, let

$$\alpha = \mathbb{P}(\text{Type I error}) = \mathbb{P}(\mathbf{X} \in C \mid H_0 \text{ is true}).$$

$$\beta = \mathbb{P}(\text{Type II error}) = \mathbb{P}(\mathbf{X} \not\in C \mid H_1 \text{ is true}).$$

The *size* of the test is α, and $1 - \beta$ is the *power* of the test to detect H_1.

Definition (Likelihood)**.** The *likelihood* of a simple hypothesis $H : \theta = \theta^*$ given data $\mathbf{x}$ is

$$L_{\mathbf{x}}(H) = f_{\mathbf{X}}(\mathbf{x} \mid \theta = \theta^*).$$

The *likelihood ratio* of two simple hypotheses H_0, H_1 given data $\mathbf{x}$ is

$$\Lambda_{\mathbf{x}}(H_0; H_1) = \frac{L_{\mathbf{x}}(H_1)}{L_{\mathbf{x}}(H_0)}.$$

A *likelihood ratio test* (LR test) is one where the critical region C is of the form

$$C = \{\mathbf{x} : \Lambda_{\mathbf{x}}(H_0; H_1) > k\}$$

for some k.

Definition (p-value)**.** The quantity p^* is called the *p-value* of our observed data $\mathbf{x}$. For the example above, $z = 0.4$ and so $p^* = 1 - \Phi(0.4) = 0.3446$.

2.2 Composite hypotheses

Definition (Power function)**.** The *power function* is

$$W(\theta) = \mathbb{P}(\mathbf{X} \in C \mid \theta) = \mathbb{P}(\text{reject } H_0 \mid \theta),$$

Definition (Size)**.** The *size* of the test is

$$\alpha = \sup_{\theta \in \Theta_0} W(\theta),$$

Definition (Uniformly most powerful test)**.** A test specified by a critical region C is *uniformly most powerful* (UMP) size α test for test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ if

(i) $\sup_{\theta \in \Theta_0} W(\theta) = \alpha$.

(ii) For any other test C^* with size $\leq \alpha$ and with power function W^*, we have $W(\theta) \geq W^*(\theta)$ for all $\theta \in \Theta_1$.

Note that these may not exist. However, the likelihood ratio test often works.

Definition (Likelihood of a composite hypothesis)**.** The *likelihood* of a composite hypothesis $H : \theta \in \Theta$ given data $\mathbf{x}$ to be

$$L_{\mathbf{x}}(H) = \sup_{\theta \in \Theta} f(\mathbf{x} \mid \theta).$$

2.3 Tests of goodness-of-fit and independence

2.3.1 Goodness-of-fit of a fully-specified null distribution

2.3.2 Pearson's Chi-squared test

2.3.3 Testing independence in contingency tables

Definition (Contingency table)**.** A *contingency table* is a table in which observations or individuals are classified according to one or more criteria.

2.4 Tests of homogeneity, and connections to confidence intervals

2.4.1 Tests of homogeneity

2.4.2 Confidence intervals and hypothesis tests

Definition (Acceptance region)**.** The *acceptance region* A of a test is the complement of the critical region C.

Note that when we say "acceptance", we really mean "non-rejection"! The name is purely for historical reasons.

2.5 Multivariate normal theory

2.5.1 Multivariate normal distribution

Definition (Multivariate normal distribution)**.** $\mathbf{X}$ has a *multivariate normal distribution* if, for every $\mathbf{t} \in \mathbb{R}^n$, the random variable $\mathbf{t}^T\mathbf{X}$ (i.e. $\mathbf{t} \cdot \mathbf{X}$) has a normal distribution. If $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ and $\mathrm{cov}(\mathbf{X}) = \Sigma$, we write $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$.

2.5.2 Normal random samples

2.6 Student's t-distribution

Definition (t-distribution)**.** Suppose that Z and Y are independent, $Z \sim N(0, 1)$ and $Y \sim \chi^2_k$. Then

$$T = \frac{Z}{\sqrt{Y/k}}$$

is said to have a t-distribution on k degrees of freedom, and we write $T \sim t_k$.

Notation. We write $t_k(\alpha)$ be the upper $100\alpha\%$ point of the t_k distribution, so that $\mathbb{P}(T > t_k(\alpha)) = \alpha$.

3 Linear models

3.1 Linear models

Definition (Least squares estimator)**.** In a linear model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the *least squares estimator* $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimizes

$$\begin{aligned} S(\boldsymbol{\beta}) &= \|\mathbf{Y} - X\boldsymbol{\beta}\|^2 \\ &= (\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) \\ &= \sum_{i=1}^n (Y_i - x_{ij}\beta_j)^2 \end{aligned}$$

with implicit summation over j.

If we plot the points on a graph, then the least square estimators minimizes the (square of the) vertical distance between the points and the line.

3.2 Simple linear regression

Definition (Fitted values and residuals)**.** $\hat{\mathbf{Y}} = X\hat{\boldsymbol{\beta}}$ is the *vector of fitted values*. These are what our model says $\mathbf{Y}$ should be.

$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the *vector of residuals*. These are the deviations of our model from reality.

The *residual sum of squares* is

$$\mathrm{RSS} = \|\mathbf{R}\|^2 = \mathbf{R}^T\mathbf{R} = (\mathbf{Y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}).$$

3.3 Linear models with normal assumptions

3.4 The F distribution

Definition (F distribution)**.** Suppose U and V are independent with $U \sim \chi^2_m$ and $V \sim \chi^2_n$. The $X = \frac{U/m}{V/n}$ is said to have an F*-distribution* on m and n degrees of freedom. We write $X \sim F_{m,n}$

3.5 Inference for $\boldsymbol{\beta}$

3.6 Simple linear regression

3.7 Expected response at $\mathbf{x}^*$

3.8 Hypothesis testing

3.8.1 Hypothesis testing

3.8.2 Simple linear regression

3.8.3 One way analysis of variance with equal numbers in each group