Part IB - Statistics

Lectured by D. Spiegelhalter

Lent 2015

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

1 Introduction and probability review

Lecturer apologizes for not turning up the previous lecture

Definition (Statistics)**.** *Statistics* is a set of principle a nd procedures for gaining and processing quantitative evidence in order to help us make judgements and decisions.

Note that we did not mention data. We don't necessarily need data for statistics (even though most often we do).

In this course, we focus on formal *statistical inference*. In the process, we assume

- we have data generated from some unknown probability model
- we aim to use the data to learn about certain properties of the underlying probability model

In particular, we perform parametric inference:

We assume a random variable X takes values in $\mathcal{X}$. We assume its distribution belongs to a family of distribution (e.g. Poisson) indexed by a scalar or vector parameter θ, taking values in some parameter space Θ. We call this a *parametric family*.

For example, we can have $X \sim \text{Poisson}(\mu)$ and $\theta = \mu \in \Theta = (0, \infty)$.

We assume that we already which family it belongs to, and then try to find out θ.

Suppose $X_1, X_2, \cdots, X_n$ are iid with the same distribution as X. Then $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ is a simple random sample (i.e. our data).

We use the observed $\mathbf{X} = \mathbf{x}$ to make inferences about θ, such as

- giving an estimate $\hat{\theta}(\mathbf{x})$ of the true value of θ.
- Giving an interval estimate $(\hat{\theta}_1(\mathbf{x}), \hat{\theta}_2(\mathbf{x}))$ for θ
- testing a hypothesis about θ, e.g. whether $\theta = 0$.

2 Estimation, bias and mean squared error

Suppose that $X_1, \cdots, X_n$ are iid, each with a probability density/mass function $f_X(x|\theta)$. We know f_X but not θ.

Definition (Statistic)**.** A *statistic* is an estimate of θ. It is a function T of a data. If $\mathbf{X} = \mathbf{x} = (x_1, \cdots, x_n)$, 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 its sampling distribution.

Note that capital $\mathbf{X}$ denotes a random variable and $\mathbf{x}$ is an observed value. So $T(\mathbf{X})$ is a random variable and $T(\mathbf{x})$ is a particular value.

Example. Let $X_1, \cdots, X_n$ is an iid $N(\mu, 1)$. A possible estimator for μ is

$$T(\mathbf{X}) = \frac{1}{n}\sum X_i.$$

The for any particular observed sample $\mathbf{x}$, our estimate is

$$T(\mathbf{x}) = \frac{1}{n}\sum x_i.$$

Recall that in general, if $X_i \sim N(\mu_i, \sigma_i^2)$, then $\sum X_i \sim N(\sum \mu_i, \sum \sigma_i^2)$, which can be proved by moment-generating functions.

So we have $T(\mathbf{X}) \sim N(\mu, 1/n)$. Note that by the Central Limit Theorem, even if X_1 were not normal, we still have $T(\mathbf{X}) \sim N(\mu, 1/n)$ for large values of n.

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.

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

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 pmf.

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

To find out $\mathbb{E}_\theta(T)$, we can either find the distribution of T and find its expected value, or evaluate T as a function of $\mathbf{X}$ directly, and find its expected value.

Example. In the above example, $\mathbb{E}_\mu(T) = \mu$. So T is unbiased for μ.

2.1 Mean squared error

In general, we prefer estimators whose sampling distributions "cluster more closely" around the true value of θ.

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.

For an unbiased estimator, the mean squared error is just the variance. In general,

$$\begin{aligned}\mathbb{E}_\theta[(\hat{\theta}-\theta)^2] &= \mathbb{E}_\theta[(\hat{\theta}-E_\theta(\hat{\theta})+\mathbb{E}_\theta(\hat{\theta})-\theta)^2] \\ &= \mathbb{E}_\theta[(\hat{\theta}-E_\theta(\hat{\theta}))^2]+\mathbb{E}_\theta(\hat{\theta})-\theta]^2+2[\mathbb{E}_\theta(\hat{\theta})-\theta]\mathbb{E}_\theta[\hat{\theta}-\mathbb{E}_\theta(\hat{\theta})] \\ &= \text{var}(\hat{\theta})+\text{bias}^2(\hat{\theta}).\end{aligned}$$

Sometimes it can be preferable to have a biased estimator with a low variance - sometimes know as the "bias-variance tradeoff".

For example, suppose $X \sim \text{binomial}(n,\theta)$. The standard estimator is $T_U = \mathbf{X}/n$, which is unbiased. T_U has variance

$$\text{var}_\theta(T_U) = \text{var}_\theta(\mathbf{X})/n^2 = \theta(1-\theta)/n.$$

Consider an alternative estimator

$$T_B = \frac{X+1}{n+2} = w\frac{X}{n}+(1-w)\frac{1}{2},$$

where $w = n/(n+2)$. This can be interpreted to be a weighted average (by the sample size) of the sample mean and $1/2$. We have

$$\mathbb{E}_\theta(T_B)-\theta = \frac{n\theta+1}{n+2}-\theta = (1-w)\left(\frac{1}{2}-\theta\right),$$

and is biased. However,

$$\text{var}_\theta(T_B) = \frac{\text{var}_\theta(\mathbf{X})}{(n+2)^2} = w^2\theta(1-\theta)/n$$

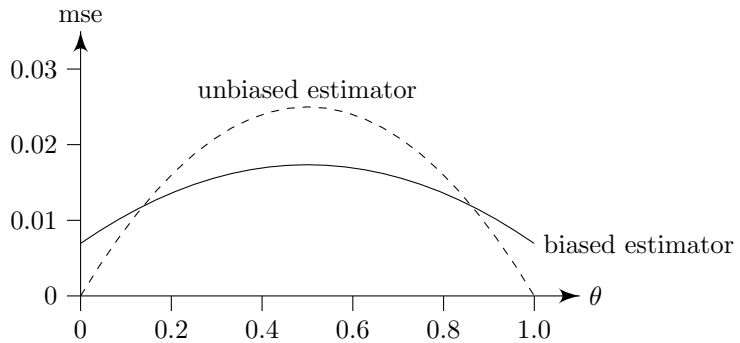
Now

$$\text{mse}(T_U) = \text{var}_\theta(T_U)+\text{bias}^2(T_U) = \theta(1-\theta)/n.$$

and

$$\text{mse}(T_B) = \text{var}_\theta(T_B)+\text{bias}^2(T_B) = w^2\frac{\theta(1-\theta)}{n}+(1-w)^2\left(\frac{1}{2}-\theta\right)^2$$

We can plot the mean squared error for possible values of θ. Here we plot the case where $n = 10$.



We see that this biased estimator has smaller MSE unless θ has extreme values.

Also, suppose $X \sim \text{Poisson}(\lambda)$, and for some reason, you want to estimate $\theta = [P(\mathbf{X} = 0)]^2 = e^{-2\lambda}$. Then any estimator $T(\mathbf{X})$ must satisfy $\mathbb{E}_\theta(T(\mathbf{X})) = \theta$, or equivalently,

$$E_\lambda(T(\mathbf{X})) = e^{-\lambda} \sum_{x=0}^{\infty} T(x) \frac{\lambda^x}{x!} = e^{-2\lambda}.$$

Then the only function T that can satisfy this equation is $T(X) = (-1)^X$.

Then the only unbiased estimator estimate $e^{-2\lambda}$ to be 1 if X is even, -1 if X is odd. CLEARLY crazy.