Notes on *Mathematical Statistics and Data Analysis* 3e by J.A. Rice (2006)

YAP Von Bing, Statistics and Applied Probability, NUS

November 18, 2011

1 Introduction

These notes were the result of a course on Mathematical Statistics for year-two undergraduates at the National University of Singapore offered in Semester 1 (Fall) 2011. Topics covered were

- Selected items in chapters 1-6, on probability
- Chapter 7 Survey sampling: up to, but excluding 7.4 ratio estimation.
- Chapter 8 Estimation: except for 8.6 Bayesian estimation.
- Chapter 9 Testing hypotheses: up to, but excluding 9.7 hanging rootograms.

The materials here are meant to enhance the very readable and useful book. Most of them were implemented in the course, some in a less polished state initially than presented here, but eventually tidied up. The others were not implemented, but I think are good ideas for a future course.

2 Notation

The following are some rules:

1. A random variable is denoted by a capital Roman letter, and the corresponding small letter denotes its realisation. In Example A (page 213), there is “$\bar{X} = 938.5$”. By the rule, “$\bar{x} = 938.5$” is preferable. The phrase “random variable equals constant” is restricted to probability statements, like $P(X = 1)$.

2. An estimator of a parameter $\theta$ is a random variable, denoted by $\hat{\theta}$. On page 262, the parameter is $\lambda_0$, but the estimator is written $\hat{\lambda}$. By the rule, the estimator should be $\hat{\lambda}_0$. A realisation of $\hat{\theta}$ is called an estimate, which is not denoted by $\hat{\theta}$. In Example A (page 261), the estimator of $\lambda$ is $\hat{\lambda} = \bar{X}$. Later, there is “$\hat{\lambda} = 24.9$”. This would be rephrased as something like “24.9 is an estimate of $\lambda$.”
3 Sampling

3.1 With replacement, then without

While simple random sampling from a finite population is a very good starting point, the dependence among the samples gets in the way of variance calculations. This is an obstacle to comprehending the standard error (SE). Going first with sampling with replacement, there are independent and identically distributed (IID) random variables \(X_1, \ldots, X_n\) with expectation \(\mu\) and variance \(\sigma^2\), both unknown constants. The sample mean \(\bar{X}\) is an estimator of \(\mu\), i.e., its realisation \(\bar{x}\) is an estimate of \(\mu\). The actual error in \(\bar{x}\) is unknown, but is around \(SD(\bar{X})\), which is \(\sigma/\sqrt{n}\). This is an important definition:

\[
SE(\bar{x}) := SD(\bar{X})
\]

Now \(\sigma^2\) has to be estimated using the data \(x_1, \ldots, x_n\). A reasonable estimate is \(v\), the realisation of the sample variance

\[
V = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

Since \(E(V) = (n-1)\sigma^2/n\), an unbiased estimator of \(\sigma^2\) is

\[
S^2 = \frac{n}{n-1} V
\]

The approximation of the SE by \(\sqrt{v/n}\) or \(s/\sqrt{n}\) is called the bootstrap. After the above is presented, we may go back to adjust the formulae for sampling without replacement from a population of size \(N\). The estimator is still \(\bar{X}\), but the SE should be multiplied by

\[
\sqrt{1 - \frac{n-1}{N-1}}
\]

which is important if \(n\) is large relative to the population size. The unbiased estimator of \(\sigma^2\) is \((N-1)S^2/N\), which matters only if \(N\) is small.

The random variable \(V\) is usually written as \(\hat{\sigma}^2\). I reserve \(\hat{\sigma}^2\) for a generic estimator of \(\sigma^2\). For example, using the method of moments, \(\hat{\mu} = \bar{X}\) and \(\hat{\sigma}^2 = V\).

3.2 \(S^2\) or \(V\)?

Let \(X_1, \ldots, X_n\) be IID with mean \(\mu\) and variance \(\sigma^2\), both unknown. Consider confidence intervals for \(\mu\). \(S^2\) must be used when the \(t\) distribution applies, i.e., when the \(X\)’s are normally distributed. Otherwise, \(n\) should be large enough for the normal approximation to be good, in which case there is little difference between the two.

For the calculation of SE, \(S^2\) is more accurate.

3.3 Sampling without replacement

Let \(X_1, \ldots, X_n\) be a simple random sample from a population of size \(N\) with variance \(\sigma^2\). Lemma B (page 207) states: if \(i \neq j\), then \(\text{cov}(X_i, X_j) = -\sigma^2/(N-1)\). The proof assumes, without justification, that the \(X\)’s are exchangeable, or more specifically, the conditional distribution of \(X_j\) given \(X_i\) is the same as that of \(X_2\) given \(X_1\).
3.4 Systematic sampling

On page 238, systematic sampling is defined. Essentially, the population is arranged in a rectangle. A row (or column) is chosen at random (equal probability) as the sample. It is said that periodic structure can cause bias. This statement cannot apply to the sample mean, which is unbiased for the population mean. Perhaps the point is: its variance is larger than that for a simple random sample of the same size.

4 Estimation

4.1 Motivation

The main idea is a natural extension from sampling with replacement. IID random variables $X_1, \ldots, X_n$ are used to construct an estimator $\hat{\theta}$, a random variable, for a parameter $\theta$, an unknown constant. Given data, i.e., realisations $x_1, \ldots, x_n$, the corresponding realisation of $\hat{\theta}$ is an estimate of $\theta$. The actual error in the estimate is unknown, but is roughly SD($\hat{\theta}$).

General definition:

$$\text{SE(estimate)} := \text{SD(estimator)}$$

The SE often involves some unknown parameters. Obtaining an approximate SE, by replacing the parameters by their estimates, is called the bootstrap.

4.2 Sample size

It is important to have a strict definition of sample size:

$$\text{sample size} = \text{the number of IID data}$$

The sample size is usually denoted by $n$, but here is a confusing example. Let $X \sim \text{Binomial}(n, p)$. The sample size is 1, not $n$ (Wouldn’t it be so confusing to write “$n$ is 1, not $n$.”?). That it is equivalent to $n$ IID Bernoulli($p$) random variables is the job of sufficiency.

4.3 What does $\theta$ mean?

Often, $\theta$ denotes two things, both in the parameter space $\Theta$: (a) a particular unknown constant, (b) a generic element. This double-think is quite unavoidable in maximum likelihood, but for method of moments (MOM), it can be sidestepped by keeping $\Theta$ implicit. For example, the estimation problem on pages 255-257 can be stated as follows: For $n = 1, 207$, assume that $X_1, \ldots, X_n$ are IID Poisson($\lambda$) random variables, where $\lambda$ is an unknown positive constant. Given data $x_1, \ldots, x_n$, estimate $\lambda$ and find an approximate SE. The MOM estimator $\hat{\lambda}$ is constructed without referring to the parameter space: the positive real numbers.

4.4 Alpha particle emissions (page 256)

$$\frac{10220}{1207} \approx 8.467, \text{ not close to 8.392. In Berk (1966), the first three categories (0, 1, 2 emissions) were reported as 0, 4, 14, and the average was 8.3703. Since 8.3703 \times 1207 = 10102.95, the total number of emissions is most likely 10103. Combined with the reported } s^2 = 8.6592, \text{ the sum of squares is close to 95008.54, likely 95009. Going through all possible counts for 17, 18, 19, 20 emissions, the complete table is reconstructed, shown on the next page.}$$
<table>
<thead>
<tr>
<th>Emissions</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>126</td>
</tr>
<tr>
<td>7</td>
<td>146</td>
</tr>
<tr>
<td>8</td>
<td>164</td>
</tr>
<tr>
<td>9</td>
<td>161</td>
</tr>
<tr>
<td>10</td>
<td>123</td>
</tr>
<tr>
<td>11</td>
<td>101</td>
</tr>
<tr>
<td>12</td>
<td>74</td>
</tr>
<tr>
<td>13</td>
<td>53</td>
</tr>
<tr>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Total 1207

Table 1: Reconstructed Table 10 A of Berk (1966). \( \bar{x} = 8.3703, s^2 = 8.6596. \)

4.5 Rainfall (pages 264-265)

Let \( X_1, \ldots, X_{227} \) be IID gamma(\( \alpha, \lambda \)) random variables, where the shape \( \alpha \) and the rate \( \lambda \) are unknown constants. The MOM estimator of (\( \alpha, \lambda \)) is

\[
(\hat{\alpha}, \hat{\lambda}) = \left( \frac{\bar{X}^2}{V}, \frac{\bar{X}}{V} \right)
\]

The SEs, SD(\( \hat{\alpha} \)) and SD(\( \hat{\lambda} \)), have no known algebraic expressions, but can be estimated by Monte Carlo, if (\( \alpha, \lambda \)) is known. It is unknown, estimated as (0.38,1.67). The bootstrap assumes

\[
SD(\hat{\alpha}) \approx SD(\hat{\alpha}_{0.38}), \quad SD(\hat{\lambda}) \approx SD(\hat{\lambda}_{1.67})
\]

where (0.38,1.67) is the MOM estimator of (0.38,1.67). This odd-looking notation is entirely consistent. To get a realisation of 0.38, generate \( x_1, \ldots, x_{227} \) from the gamma(0.38,1.67) distribution and compute \( \bar{x}^2/v \).

The double-act of Monte Carlo and bootstrap can be used to assess bias. Define

\[
\text{bias}(0.38) = E(\hat{\alpha}) - \alpha
\]

By the bootstrap,

\[
E(\hat{\alpha}) - \alpha \approx E(0.38) - 0.38
\]

which can be estimated by Monte Carlo. The bias in (0.38,1.67) is around (0.10,0.02). The maximum likelihood estimates have smaller bias and SE.
4.6 Maximum likelihood

Unlike MOM, in maximum likelihood (ML) the two meanings of $\theta$ clash, like an actor playing two characters who appear in the same scene. The safer route is to put the subscript 0 to the unknown constant. This is worked out in the Poisson case.

Let $X_1, \ldots, X_n$ be IID Poisson($\lambda_0$) random variables, where $\lambda_0 > 0$ is an unknown constant. The aim is to estimate $\lambda_0$ using realisations $x_1, \ldots, x_n$. Let $\lambda > 0$. The likelihood at $\lambda$ is the probability of the data, supposing that they come from Poisson($\lambda$):

$$L(\lambda) = \Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

On $(0, \infty)$, $L$ has a unique maximum at $\bar{x}$. $\bar{x}$ is the ML estimate of $\lambda_0$. The ML estimator is $\hat{\lambda}_0 = \bar{X}$.

The distinction between $\lambda_0$ and $\lambda$ helps make the motivation of maximum likelihood clear. However, for routine use, $\lambda_0$ is clunky (I have typed tutorial problems where the subscript appears in the correct places; it is hard on the eyes). The following is more economical, at the cost of some mental gymnastics.

Let $X_1, \ldots, X_n$ be IID Poisson($\lambda$) random variables, where $\lambda > 0$ is an unknown constant.

[Switch.] The likelihood function on $(0, \infty)$ is

$$L(\lambda) = \Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

which has a unique maximum at $\bar{x}$. [Switch back.] The ML estimator of $\lambda$ is $\hat{\lambda} = \bar{X}$.

With some practice, it gets more efficient. The random likelihood function is a product of the respective random densities:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!}$$

Since $L$ has a unique maximum at $\bar{X}$, $\hat{\lambda} = \bar{X}$.

The distinction between a generic $\theta$ and a fixed unknown $\theta_0$ is useful in the motivation and in the proof of large-sample properties of the ML estimator (pages 276-279). Otherwise, $\theta_0$ can be quietly dropped without confusion.

4.7 Hardy-Weinberg equilibrium (pages 273-275)

By the Hardy-Weinberg equilibrium, the proportions of the population having 0, 1 and 2 $a$ alleles are respectively $(1 - \theta)^2$, $2 \theta (1 - \theta)$ and $\theta^2$. Hence the number of $a$ alleles in a random individual has the binomial(2,$\theta$) distribution; so the total number of $a$ alleles in a simple random sample of size $n < N$ has the binomial($2n$, $\theta$) distribution.

Let $X_1, X_2, X_3$ be the number of individuals of genotype $AA$, $Aa$ and $aa$ in the sample. The number of $a$ alleles in the sample is $X_2 + 2X_3 \sim$ binomial($2n$, $\theta$), so

$$\text{var} \left( \frac{X_2 + 2X_3}{2n} \right) = \frac{\theta(1 - \theta)}{2n}$$

The Monte Carlo is unnecessary. The variance can also be obtained directly using the expectation and variance of the trinomial distribution, but the binomial route is much shorter.

Since there is no cited work, it is not clear whether the data were a simple random sample of the Chinese population of Hong Kong in 1937. If not, the trinomial distribution is in doubt.
4.8 MOM on the multinomial

Let \( X = (X_1, \ldots, X_r) \) have the multinomial\((n, p)\) distribution, where \( p \) is an unknown fixed probability vector. The first moment \( \mu_1 = np \). Since the sample size is 1, the estimator of \( \mu_1 \) is \( X \). The MOM estimator of \( p \) is \( X/n \). This illustrates the definition of sample size.

Let \( r = 3 \). Under Hardy-Weinberg equilibrium, \( p = ((1 - \theta)^2, 2\theta(1 - \theta), \theta^2) \). There are three different MOM estimators of \( \theta \).

4.9 Confidence interval for normal parameters (page 280)

Let \( x_1, \ldots, x_n \) be realisations from IID random variables \( X_1, \ldots, X_n \) with normal\((\mu, \sigma^2)\) distribution, where \( \mu \) and \( \sigma^2 \) are unknown constants. The \((1 - \alpha)\)-confidence interval for \( \mu \):

\[
\left( \bar{x} - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2), \bar{x} + \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2) \right)
\]

is justified by the fact that

\[
P\left( \bar{X} - \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{n-1}(\alpha/2) \right) = 1 - \alpha
\]

Likewise,

\[
P\left( \frac{nV}{\chi^2_{n-1}(\alpha/2)} \leq \sigma^2 \leq \frac{nV}{\chi^2_{n-1}(1 - \alpha/2)} \right) = 1 - \alpha
\]

justifies the confidence interval for \( \sigma^2 \):

\[
\left( \frac{nV}{\chi^2_{n-1}(\alpha/2)} \chi_{n-1}^2(1 - \alpha/2) \right)
\]

This illustrates the superiority of \( V \) over \( \hat{\sigma}^2 \).

4.10 The Fisher information

Let \( X \) have density \( f(x|\theta) \), \( \theta \in \Theta \), an open subset of \( \mathbb{R}^p \). The Fisher information in \( X \) is defined as

\[
\mathcal{I}(\theta) = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X|\theta) \right]
\]

When the above exists, it equals the more general definition

\[
\mathcal{I}(\theta) = E \left[ \frac{\partial}{\partial \theta} \log f(X|\theta) \right]^2
\]

In all the examples, and most applications, they are equivalent, and the first definition is easier to use.
4.11 Asymptotic normality of ML estimators

Let $\hat{\theta}$ be the ML estimator of $\theta$ based on IID random variables $X_1, \ldots, X_n$ with density $f(x|\theta)$, where $\theta$ is an unknown constant. Let $I(\theta)$ be the Fisher information in $X_1$. Under regularity conditions, as $n \to \infty$,

$$\sqrt{nI(\theta)}(\hat{\theta} - \theta) \to N(0, 1)$$

Note the Fisher information of a single unit of the IID $X$’s, not that of all of them, which is $n$ times larger.

For large $n$, approximately

$$\hat{\theta} \sim N(\theta, \frac{I(\theta)^{-1}}{n})$$

$I(\theta)^{-1}$ plays a similar role as the population variance in sampling.

4.12 Asymptotic normality: multinomial, regression

This is essentially the content of the paragraph before Example C on page 283. Let $X = (X_1, \ldots, X_r)$ have a multinomial($n, p(\theta)$) distribution, where $\theta$ is an unknown constant. Let $I(\theta)$ be the Fisher information. Then as $n \to \infty$,

$$\sqrt{I(\theta)}(\hat{\theta} - \theta) \to N(0, 1)$$

Proof: $X$ has the same distribution as the sum of $n$ IID random vectors with the multinomial(1,$p(\theta)$) distribution. The ML estimator of $\theta$ based on these data is $\hat{\theta}$ (by sufficiency or direct calculation). The Fisher information of a multinomial(1,$p(\theta)$) random vector is $I(\theta)/n$. Now apply the result in the previous section.

If the $X$’s are independent but not identically distributed, like in regression, asymptotic normality of the MLEs does not follow from the previous section. A new definition of sample size and a new theorem are needed.

4.13 Characterisation of sufficiency

Let $X_1, \ldots, X_n$ be IID random variables with density $f(x|\theta)$, where $\theta \in \Theta$. Let $T$ be a function of the $X$’s. The sample space $S$ is a disjoint union of $S_t$ across all possible values $t$, where

$$S_t = \{x : T(x) = t\}$$

The conditional distribution of $X$ given $T = t$ depends on $\theta$ in general. If for every $t$, it is the same for every $\theta \in \Theta$, then $T$ is sufficient for $\theta$. This gives a characterisation: $T$ is sufficient for $\theta$ if and only if there is a function $q(x)$ such that for every $t$ and every $\theta \in \Theta$

$$f(x|t) = q(x), \quad x \in S_t$$

This clarifies the proof of the factorisation theorem (page 307).

Logically, sufficiency is a property of $\Theta$. “$T$ is sufficient for $\Theta$” is more correct.
4.14 Problem 4 (page 313)

Let $X_1, \ldots, X_{10}$ be IID with distribution $P(X = 0) = 2\theta/3$, $P(X = 1) = \theta/3$, $P(X = 2) = 2(1 - \theta)/3$, $P(X = 3) = (1 - \theta)/3$, where $\theta \in (0, 1)$ is an unknown constant. The MOM estimator of $\theta$ is

$$\hat{\theta} = \frac{7}{6} - \frac{\bar{X}}{2}$$

The ML estimate based on the given numerical data can be readily found. That of general $x_1, \ldots, x_{10}$, or the ML estimator, is hard because the distribution of $X_1$ is not in the form $f(x|\theta)$. One possibility is

$$f(x|\theta) = \left[\frac{2}{3}\theta\right]^{y_0} \left[\frac{1}{3}\theta\right]^{y_1} \left[\frac{2}{3}(1 - \theta)\right]^{y_2} \left[\frac{1}{3}(1 - \theta)\right]^{y_3}, \quad y_j = 1_{\{x=j\}}$$

It follows that the ML estimator is

$$\hat{\theta} = \frac{Y}{n}$$

where $Y$ is the number of times 0 or 1 turn up.

This is an interesting example. $\hat{\theta}$ and $\tilde{\theta}$ are different, both are unbiased. $\hat{\theta}$ is efficient, but not $\tilde{\theta}$:

$$\text{var}(\hat{\theta}) = \frac{\theta(1 - \theta)}{n}, \quad \text{var}(\tilde{\theta}) = \frac{1}{18n} + \frac{\theta(1 - \theta)}{n}$$

They illustrate relative efficiency more directly than on pages 299-300, and the sample size interpretation of relative efficiency. $Y$ is sufficient for $\theta$, so $E(\tilde{\theta}|Y)$ has smaller variance than $\hat{\theta}$. Another clue to the superiority of ML: If 0, 1, 2, 3 are replaced by any other four distinct numbers, the ML estimator stays the same, but the MOM estimator will be fooled into adapting. Fundamentally, the distribution is multinomial.

5 Hypothesis tests

Let $X = (X_1, \ldots, X_r)$ be multinomial$(n, p(\theta))$, where $\theta \in \Theta$, an open subset of $\mathbb{R}$. Let $\hat{\theta}$ be the ML estimator of $\theta$ based on $X$. Then the null distribution of the Pearson’s

$$X^2 = \sum_{i=1}^{r} \frac{[X_i - np_i(\hat{\theta})]^2}{np_i(\hat{\theta})}$$

is asymptotically $\chi^2_{r-2}$. Let $X$ be obtained from a Poisson distribution, so that $\theta = \lambda \in (0, \infty)$. The above distribution holds for $\hat{\lambda}$ based on $X$. Example B (page 345) assumes it also holds for $\tilde{\lambda} = \bar{X}$, the ML estimator based on the raw data, before collapsing. This is unsupported. An extreme counter-example: collapse Poisson$(\lambda)$ to the indicator of 0, i.e., the data is binomial$(n, e^{-\lambda})$. Since $r = 2$, theory predicts $X^2 \equiv 0$, as can be verified directly, but plugging $\hat{\lambda} = \bar{X}$ into $X^2$ gives a positive random variable. Using $R$, the correct ML estimate of $\lambda$ in Example B is 2.39, giving $X^2 = 79.2$.

6 Minor typographical errors

- Page A39: In the solution to Problem 55 on page 324, the asymptotic variance has $1 + 2\theta$ in the denominator, not $1 + \theta$. Refer to Fisher (1970) page 313.
- Page A26: Bliss and Fisher (1953) is on pages 176-200.