Sufficiency
(Hogg Chapter Seven)

STAT 406-01: Mathematical Statistics II *

Spring Semester 2016

Contents

1 Sufficiency and the Factorization Theorem **2**

2 The Rao-Blackwell Theorem **2**

3 Recap of Sufficiency **2**
- 3.1 Comment on the Likelihood Principle 2
- 3.2 Sufficiency 3
- 3.3 Minimum-Variance Unbiased Estimators 3
- 3.4 Uniqueness and Completeness 3

4 The Exponential Class of Distributions **4**
- 4.1 Example: Exponential Distribution 5

5 Reparametrization and Other Pitfalls **6**
- 5.1 Exponential Example Continued 6
- 5.2 Change of Parameters 6

6 Generalizations to Multiple Parameters **7**
- 6.1 Joint Sufficient Statistics 7
- 6.2 Example: Bradley-Terry Model 8
- 6.3 Other Extensions 9

7 Minimal Sufficiency and Ancillary Statistics **10**
- 7.1 Minimal Sufficient Statistics 10
- 7.2 Sufficiency and Ancillarity 11
  - 7.2.1 Example: Uniform Sampling Distribution . 11
  - 7.2.2 Location Models 12
  - 7.2.3 Example: Bernoulli Trials 12
- 7.3 Ancillarity and Independence 13
  - 7.3.1 Basu's Theorem 13
  - 7.3.2 Example: Mean of a Normal Distribution . 13
- 7.4 Location and Scale Models Revisited 14

Tuesday 15 March 2016
Guest lecture from Prof. James Marengo
– Read Sections 7.1-7.2 of Hogg

1 Sufficiency and the Factorization Theorem

Thursday 17 March 2016
Guest lecture from Prof. James Marengo
– Read Section 7.3 of Hogg

2 The Rao-Blackwell Theorem

Tuesday 29 March 2016
– Read Section 7.4 of Hogg

3 Recap of Sufficiency

3.1 Comment on the Likelihood Principle

Return to an example Hogg gives at the end of section 7.1: consider an experiment with a series of trials with some unknown probability of success $\theta \in [0, 1]$, in which one success is observed in a set of 10 trials. But consider two different scenarios in which that can occur:

1. The experiment was designed to carry out 10 trials. The observable is then the number of successes Y, which is a binomial random variable $\text{Bin}(10, \theta)$ with pmf $p_Y(y) = \binom{10}{y}\theta^y(1-\theta)^{10-y}$
2. The experiment was designed to stop at the first success. The observable then the number of trials Z, which is a geometric random variable with pmf $p_Z(z) = \theta(1-\theta)^{z-1}$, $z = 1, 2, 3, \ldots$.

In classical statistics, the inference problem is related to the probabilities for outcomes of repeated experiments of the same sort, so it depends not just on what was actually observed, but what the "rules" of the experiment were, and so we might draw different inferences from the same observation under the two scenarios described above. In particular, as Hogg shows, if we define an estimator of θ to be (# of successes)/(# of trials), this is Y/n under the first scenario and $1/Z$ under the second. But while $E(Y/n) = \theta$, $E(1/Z) \neq \theta$, so this estimator would be unbiased under the first description but biased under the second. So if we required our estimator to be unbiased, we would construct a different estimator if we'd decided to stop at the first success rather than to do 10 trials, even though in either case we'd seen the exact same thing.

In this example, the likelihood functions for the two scenarios, evaluated at the actual observed values, are the same up to a constant, i.e., $L_{(a)}(\theta; y = 1) = 10\,\theta\,(1-\theta)^9$ and $L_{(b)}(\theta; z = 10) = \theta\,(1-\theta)^9$. Hogg states something called the *likelihood principle*, which "many statisticians" favor, stating that if two problems have likelihood functions which are the same up to a constant, we should draw the same inferences in them. Most Bayesians go a step further and point out that the posterior pdf $f(\theta|D)$ should depend only on what actually happened, and not on the hypothetical outcomes of imaginary repetitions of the experiment. (This situation is known as "optional stopping" because in frequentist inference, the conclusions can depend on why the experimenter chose to stop the experiment.) Jaynes[1] makes this point by stating that the information about the rules of the experiment is redundant, If we already know

[1] *Probability Theory: The Logic of Science* section 6.9.1

that we've seen 9 failures followed by 1 success, the information that we conducted 10 trials (in the binomial case) or that we stopped at the first success (in the geometric case) tells us nothing we don't already know from the data. He claims that any inference which treats the two situations differently violates the basic principles of logic: the sample space): "Of course, that violates the principle $AA = A$ of elementary logic; it is astonishing that such a thing could be controversial in the 20th century."

3.2 Sufficiency

Definition of sufficiency: $Y_1 = u(\mathbf{X})$ is a sufficient statistic for a parameter θ iff the likelihood for the whole sample is a θ-independent multiple of the likelihood for Y_1:

$$\frac{f_{\mathbf{X}}(\mathbf{x};\theta)}{f_{Y_1}(y_1;\theta)} = H(\mathbf{x}) \tag{3.1}$$

Factorization theorem: under the usual regularity conditions, this is equivalent to the likelihood factoring into a θ-dependent part involving only $u(\mathbf{x})$ and a θ-independent part:

$$f_{\mathbf{X}}(\mathbf{x};\theta) = k_1(u(\mathbf{x});\theta)\, k_2(\mathbf{x}) \tag{3.2}$$

Of course, we've seen this in a Bayesian context; it means the posterior for θ depends only on $u(\mathbf{x})$, since

$$f_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\, f_{\Theta}(\theta) \propto k_1(u(\mathbf{x});\theta)\, f_{\Theta}(\theta) \tag{3.3}$$

3.3 Minimum-Variance Unbiased Estimators

We considered the variance of an unbiased ($E(Y) = \theta$) estimator

$$\mathrm{Var}(Y) = E\left([Y-\theta]^2\right) \tag{3.4}$$

when we studied the Cramér-Rao bound. Thinking of $(Y-\theta)^2$ as a loss function, the variance is seen as a measure of the risk of making a poor estimate, so choosing an estimator which minimizes it is desirable.

Rao-Blackwell theorem: if Y_1 is a sufficient statistic for θ, and Y_2 is an unbiased estimator of θ based on some other combination of the data, we can make another unbiased estimator $\varphi(Y_1) = E(Y_2|Y_1)$ which has an equal or lesser variance, so when looking for the MVUE, we can confine our attention to sufficient statistics.

3.4 Uniqueness and Completeness

The Rao-Blackwell theorem tells us that if there is a sufficient statistic, one function of it is a MVUE. But we could imagine a case where, even though we've constructed an unbiased estimator $\varphi(Y_1)$, there might be another unbiased estimator $\psi(Y_1)$ which could have a lower variance. One thing we do know about such a pair of unbiased estimators is that

$$E(\varphi(Y_1) - \psi(Y_1)) = \theta - \theta = 0 \tag{3.5}$$

In general, just because $E(\varphi(Y_1) - \psi(Y_1)) = 0$, it doesn't necessarily mean that $E(\varphi(Y_1) - \psi(Y_1)) = 0$. But suppose that the distribution function $f_{Y_1}(y_1;\theta)$ has the property that any function $u(Y_1)$ which satisfies $E(u(Y_1)) = 0$ for all θ also satisfies $u(Y_1 = 0$. Then we know the function $\varphi(Y_1)$ is unique, and there's only one unbiased estimator that can be constructed from the sufficient statistic, and this unique estimator is therefore the MVUE.

This property is known as *completeness*. We say that $f(x;\theta)$, $\theta \in \Omega$ is a complete family of distributions if, for any $u(x)$ satisfying $E(u(X)) = 0$ for all θ, $P(u(X) = 0) = 1$. This seems like a very specialized sort of distribution, but it's more

common than you'd imagine. For example, consider a Poisson distribution

$$p(x;\theta) = \begin{cases} \frac{e^{-\theta}\theta^x}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

where $0 \leq \theta < \infty$. If we have a function $u(x)$ such that $E\left(u(X)\right) = 0$ for all θ, that means

$$0 = \sum_{x=0}^{\infty} u(x)\frac{e^{-\theta}\theta^x}{x!} = e^{-\theta}\left(u(0) + u(1)\,\theta + \frac{u(2)}{2}\,\theta^2 + \frac{u(3)}{3!}\,\theta^3 + \cdots\right) \tag{3.7}$$

The prefactor $e^{-\theta}$ is a positive number, so the series in parentheses has to vanish. The only way that can be the case for all θ is if each of the coefficients vanishes separately, i.e.,

$$0 = u(0) = u(1) = u(2) = \cdots \tag{3.8}$$

i.e.,

$$u(x) = 0, \qquad x = 0, 1, 2, \ldots \tag{3.9}$$

but this is precisely the set of values for which $P(X = x)$ is non-zero, so $p(x;\theta)$ is a complete family of distributions.

Thursday 31 March 2016
– Read Section 7.4 of Hogg

4 The Exponential Class of Distributions

We consider now a class of distributions for which we can read off a complete sufficient statistic, as well as several other useful properties. This is known as the regular exponential class, for which the pdf (or pmf) has the form

$$f(x;\theta) = \exp\left[\eta(\theta)K(x) + H(x) + q(\theta)\right] \qquad x \in \mathcal{S}, \quad \gamma < \theta < \delta \tag{4.1}$$

for which

1. The support space $\mathcal{S}$ doesn't depend on θ

2. $\eta(\theta)$ is a nontrivial (not necessarily monotonic) function of θ

3. $K(x)$ is a non-trivial function of x, and if X is a continuous random variable, $K'(x) \neq 0$ and both $K'(x)$ and $H(x)$ are continuous.

Many families of distributions have this form, a number of which are given as examples in Hogg. To pick one which isn't, consider the Beta distribution

$$f(x;\theta) = \frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2}x^{\theta-1}(1-x)^{\theta-1} \qquad 0 < x < 1, \quad 0 < \theta < \infty \tag{4.2}$$

On the support space, we can write the pdf as

$$f(x;\theta) = \exp\left[\theta\ln(x[1-x]) - \ln(x[1-x]) + \ln\frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2}\right] \tag{4.3}$$

This is of the exponential form with $\gamma = 0$, $\delta = \infty$, $\mathcal{S} = (0, 1)$, $\eta(\theta) = \theta$, $K(x) = \ln x + \ln(1-x)$, $H(x) = -\ln x - \ln(1-x)$, and $q(\theta) = \ln\Gamma(2\theta) - 2\ln\Gamma(\theta)$.

Returning to the generic exponential class, we see that the likelihood for a sample of size n can be written as

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \exp\left[\eta(\theta)\sum_{i=1}^{n} K(x) + \sum_{i=1}^{n} H(x) + nq(\theta)\right]$$
$$= \exp\left[\eta(\theta)\sum_{i=1}^{n} K(x) + nq(\theta)\right]\exp\left[\sum_{i=1}^{n} H(x)\right] \tag{4.4}$$

We can see by the factorization theorem that $Y_1 \equiv \sum_{i=1}^{n} K(X_i)$ is a sufficient statistic for the parameter θ, and it's not hard to convince ourselves that the pdf for Y_1 has the form

$$f_{Y_1}(y_1;\theta) \propto \exp\left[\eta(\theta)y_1 + nq(\theta)\right] \tag{4.5}$$

where the proportionality "constant" can depend on y_1 but not θ. We write this more precisely as

$$f_{Y_1}(y_1;\theta) = R(y_1)\exp\left[\eta(\theta)y_1 + nq(\theta)\right] \tag{4.6}$$

You'll show on the homework that

$$E\left(Y_1\right) = -n\frac{q'(\theta)}{\eta'(\theta)} \tag{4.7}$$

and it can also be shown that

$$\mathrm{var}(Y_1) = \frac{n}{[p'(\theta)]^3}[p''(\theta)q'(\theta) - q''(\theta)p'(\theta)] \tag{4.8}$$

So, for the Beta distribution example, we have a sufficient statistic

$$Y_1 = \sum_{i=1}^{n} \ln[X_i(1-X_i)] = \ln\prod_{i=1}^{n} X_i(1-X_i) \tag{4.9}$$

It turns out that the statistic Y_1 is a complete sufficient statistic, i.e., that the family of distributions $f_{Y_1}(y_1;\theta)$ is complete in the sense we defined last time, i.e., the only function $u(y_1)$ which has $E\left(u(Y_1)\right) = 0$ is equal to zero (except possibly at points corresponding to zero probability). The demonstration of this takes

$$E\left(u(Y_1)\right) = e^{nq(\theta)}\int_{\mathcal{S}_{Y_1}} u(y_1)R(y_1)e^{\eta(\theta)y_1}\,dy_1 \tag{4.10}$$

Since $e^{nq(\theta)} > 0$, the only way this can vanish is if the integral does. There is a result from the theory of Laplace transforms that the only way this can vanish for all $\eta(\theta)$ is if $u(y_1)R(y_1) = 0$ for all $y_1 \in \mathcal{S}_{Y_1}$, and since $R(y_1) > 0$ on $\mathcal{S}_{Y_1}$, we must have $u(y_1) = 0$ everywhere.

The consequences of this completeness are that, if we find a function $\varphi(Y_1)$ which is an unbiased estimator, it will be unique and therefore (by the Rao-Blackwell theorem) the minimum-variance unbiased estimator.

4.1 Example: Exponential Distribution

Consider the exponential distribution with rate parameter θ, which has pdf

$$f(x;\theta) = \theta e^{-\theta x} \qquad 0 < x < \infty; \quad 0 < \theta < \infty \tag{4.11}$$

(Note that this is also a Gamma$(1, \frac{1}{\theta})$ distribution. We can write this in the exponential form as

$$f(x;\theta) = \exp\left(-\theta x + \ln\theta\right) \tag{4.12}$$

which has $\eta(\theta) = -\theta$, $K(x) = x$, $H(x) = 0$, and $q(\theta) = \ln\theta$. Thus the complete sufficient statistic is

$$Y_1 = \sum_{i=1}^{n} X_i \tag{4.13}$$

which has expectation value

$$E\left(Y_1\right) = -n\frac{q'(\theta)}{\eta'(\theta)} = -n\frac{1/\theta}{-1} = \frac{n}{\theta} \tag{4.14}$$

If we can make an unbiased estimator of θ out of this, we know it will be the MVUE. It's not immediately obvious how to get this from our knowledge of $E\left(Y_1\right)$, though, which illustrates a drawback to the formalism. In particular,

$$E\left(\frac{n}{Y_1}\right) = n\,E\left(\frac{1}{Y_1}\right) \neq \frac{n}{E\left(Y_1\right)} = \theta \tag{4.15}$$

(Note that if we has defined the parameter to be $\beta = 1/\theta$ instead, we'd have had no such problem, since $E\left(Y_1\right) = n\beta$ so $\frac{1}{n}Y_1 = \overline{X}$ is an unbiased estimator of β.)

Tuesday 5 April 2016
– Read Section 7.6 of Hogg

5 Reparametrization and Other Pitfalls

5.1 Exponential Example Continued

By construction, we can see that $Y_1 \sim \text{Gamma}(n, \frac{1}{\theta})$, so

$$f_{Y_1}(y_1;\theta) = \frac{\theta^n\, y_1^{n-1}}{\Gamma(n)}\, e^{-\theta y_1} \qquad 0 < y_1 < \infty \tag{5.1}$$

Note that for $n > 1$, we can actually calculate

$$\begin{aligned} E\left(\frac{1}{Y_1}\right) &= \int_0^\infty \frac{f_{Y_1}(y_1;\theta)}{y_1}\, dy_1 = \int_0^\infty \frac{\theta^n\, y_1^{n-2}}{\Gamma(n)}\, e^{-\theta y_1}\, dy_1 \\ &= \theta\frac{\Gamma(n-1)}{\Gamma(n)} = \frac{\theta}{n-1} \end{aligned} \tag{5.2}$$

so the unbiased estimator constructed from Γ is actually

$$\varphi(Y_1) = \frac{n-1}{Y_1} \tag{5.3}$$

which we then know is the unique minimum variance unbiased estimator.

5.2 Change of Parameters

For an example that illustrates the pitfalls inherent in seeking an unbiased estimator, consider the Bernoulli distribution $b(1,\theta)$, with pmf

$$\begin{aligned} p(x;\theta) &= \theta^x(1-\theta)^{1-x} = \exp\left(x\ln\theta + (1-x)\ln(1-\theta)\right) \\ &= \exp\left(x\left[\ln\theta - \ln(1-\theta)\right] + \ln(1-\theta)\right) \quad x = 0,1; \quad 0 < \theta < 1 \end{aligned} \tag{5.4}$$

which is of the exponential form with $\eta(\theta) = \ln\frac{\theta}{1-\theta}$, $K(x) = x$, $H(x) = 0$, and $q(\theta) = \ln(1-\theta)$. This means that $Y = \sum_{i=1}^n X_i$ is a complete sufficient statistic for θ, and it's easy to see $\overline{X} = \frac{Y}{n}$ is a MVUE for θ. But what if we choose the parameter instead to be

$$\tau = \frac{\theta}{1-\theta} \qquad 0 < \tau < \infty \tag{5.5}$$

$\overline{X}$ is still a sufficient statistic, but we run into trouble when we try to construct an unbiased estimator from it. The maximum likelihood estimator is

$$\widehat{\tau} = \frac{\overline{X}}{1-\overline{X}} \tag{5.6}$$

but we can see that the expectation value of this is divergent, since there is, for any τ, a non-zero probability $\theta^n = (1+\tau)^{-n}$ that $Y = n$ and thus $\overline{X} = 1$. We can see explicitly that it's

impossible to construct an unbiased estimator of τ out of Y by noting that $Y \sim b(n, \frac{1}{1+\tau})$, i.e.,

$$p(y;\tau) = \binom{n}{y} \frac{\tau^{n-y}}{(1+\tau)^n} \tag{5.7}$$

so

$$E\left(\varphi(Y)\right) = \sum_{y=0}^{n} \binom{n}{y} \frac{\tau^{n-y}}{(1+\tau)^n} \varphi(y) \tag{5.8}$$

Since the numerator and denominator are both polynomials in τ of order n, there's no set of $\varphi(y)$ values which can make the sum equal to τ.

Thursday 7 April 2016 – Review for Prelim Exam Two

The exam covers materials from the weeks 5-9 of the term, i.e., Hogg sections 6.3-6.5 and 7.1-7.5 (and associated topics covered in class through April 5), and problem sets 5-8.

Tuesday 12 April 2016 – Second Prelim Exam

Thursday 14 April 2016 – Read Section 7.7 of Hogg

6 Generalizations to Multiple Parameters

Most of our results and definitions carry over to the case where the single parameter θ is replaced by a vector of parameters

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix} \tag{6.1}$$

It can also happen that the $\boldsymbol{\theta}$ dependence of the likelihood is described not by a single statistic but a vector of statistics

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_1 \\ \vdots \\ Y_m \end{pmatrix} = \begin{pmatrix} u_1(\mathbf{X}) \\ u_2(\mathbf{X}) \\ \vdots \\ u_m(\mathbf{X}) \end{pmatrix} \tag{6.2}$$

where m may or may not be equal to p.

6.1 Joint Sufficient Statistics

We say that $\mathbf{Y} = \mathbf{u}(\mathbf{X})$ are *joint sufficient statistics* for $\boldsymbol{\theta}$ if and only if the ratio

$$\frac{f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta})} = H(\mathbf{x}) \tag{6.3}$$

is independent of $\boldsymbol{\theta}$. There is a corresponding factorization theorem that says that an equivalent condition is that the joint pdf

can be written

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = k_1(\mathbf{u}(\mathbf{x});\boldsymbol{\theta})\, k_2(\mathbf{x}) \tag{6.4}$$

The dataset need not be a sample drawn from a univariate distribution $f(x;\boldsymbol{\theta})$ either. One simple generalization is to work with a probability distribution for a k-dimensional random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \tag{6.5}$$

,and then we could draw a sample of size n from $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta})$, which would have a joint distribution function of

$$f_{\{\mathbf{X}_i\}}(\mathbf{x}_1,\ldots,\mathbf{x}_n;\boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}_1;\boldsymbol{\theta})\, f_{\mathbf{X}}(\mathbf{x}_2;\boldsymbol{\theta}) \cdots f_{\mathbf{X}}(\mathbf{x}_n;\boldsymbol{\theta}) \tag{6.6}$$

Of course, if you wanted to, you could expand the notation and let $\mathbf{X}$ refer to the whole data set again:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1k} & X_{2k} & \cdots & X_{nk} \end{pmatrix} \tag{6.7}$$

In general, $\mathbf{X}$ refers to our data set, whatever indices we find it convenient to label it with.

6.2 Example: Bradley-Terry Model

There are numerous examples of joint sufficient statistics in Hogg, so we look at a slightly more involved one. Consider again the Bradley-Terry model[2], which we looked at back in February when considering maximum likelihood estimation. As a reminder, this model describes "paired comparisons" between objects. (E.g., games between teams or players, taste tests between foods, etc.) Each object has a strength θ_i, and the probability of object i "winning" a given comparison with object j is $\frac{\theta_i}{\theta_i+\theta_j}$. Since the probabilities are unchanged if we multiply all of the strengths by a constant, if we have $p+1$ objects, we specify that $\theta_{p+1} = 1$ and then have p free parameters θ_1, θ_2, $\ldots$, θ_p. We suppose that there are $n_{ij} = n_{ji}$ comparisons conducted between teams i and j and let the observed quantities be the number X_{ij} of comparisons won by object i over object j where $i = 1,2,\ldots,p$ and $j = i+1, i+2, \ldots, p+1$. (We assume that there are no comparisons of an object with itself, and we avoid double-counting by restricting attention to comparisons with $i < j$.) We basically have a binomial experiment for each pair of objects, where $X_{ij} \sim b(n_{ij}, \frac{\theta_i}{\theta_i+\theta_j})$. The sampling distribution is then[3]

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) &= \prod_{i=1}^{p}\prod_{j=i+1}^{p+1} \frac{n_{ij}!}{x_{ij}!(n_{ij}-x_{ij})!} \left(\frac{\theta_i}{\theta_i+\theta_j}\right)^{x_{ij}} \left(\frac{\theta_j}{\theta_i+\theta_j}\right)^{n_{ij}-x_{ij}} \\ &= \left(\prod_{i=1}^{p}\prod_{j=i+1}^{p+1} \frac{n_{ij}!}{x_{ij}!(n_{ij}-x_{ij})!}\right) \frac{\prod_{i=1}^{p}\prod_{j=i+1}^{p+1} \theta_i^{x_{ij}}\theta_j^{n_{ij}-x_{ij}}}{\prod_{i=1}^{p}\prod_{j=i+1}^{p+1}(\theta_i+\theta_j)^{n_{ij}}} \end{aligned} \tag{6.8}$$

The factor in parentheses is independent of the parameters, and will become the $k_2(\mathbf{x})$ in the factorization of the likelihood. The factor in the denominator is independent of the data $\mathbf{x}$, and will become part of the $k_1(\mathbf{y};\boldsymbol{\theta})$, but won't affect the identification

[2] Zermelo, *Mathematische Zeitschrift* **29**, 436 (1929); Bradley and Terry *Biometrika* **39**, 324 (1952)

[3] We could define things so that the observed data were the exact sequence of comparison results rather than the numbers of comparison wins between pairs of objects, which would remove the combinatorical factor, but it would have no effect on the inference since it would multiply the likelihood by a $\boldsymbol{\theta}$-independent factor.

of the joint sufficient statistics $\mathbf{y}$. The numerator is thus "where all the action is". If we define $x_{ji} = n_{ij} - x_{ij}$ we can write it as

$$\prod_{i=1}^{p}\prod_{j=i+1}^{p+1}\theta_i^{x_{ij}}\theta_j^{x_{ji}} = \prod_{i=1}^{p+1}\prod_{j=1}^{p+1}\theta_i^{x_{ij}} = \prod_{i=1}^{p+1}\theta_i^{\sum_{j=1}^{p+1}x_{ij}} \tag{6.9}$$

where we can see the first equality most easily by writing it out, e.g., for $p+1=4$ (keeping in mind that $x_{ii}=0$):

$$\begin{aligned}&([\theta_1^{x_{12}}\theta_2^{x_{21}}][\theta_1^{x_{13}}\theta_3^{x_{31}}][\theta_1^{x_{14}}\theta_4^{x_{41}}])\,([\theta_2^{x_{23}}\theta_3^{x_{32}}][\theta_2^{x_{24}}\theta_4^{x_{42}}])\,([\theta_3^{x_{34}}\theta_4^{x_{43}}])\\ &\quad = \theta_1^{x_{12}+x_{13}+x_{14}}\theta_2^{x_{21}+x_{23}+x_{24}}\theta_3^{x_{31}+x_{32}+x_{34}}\theta_4^{x_{41}+x_{42}+x_{43}}\end{aligned} \tag{6.10}$$

and in the second step we have used the fact that $\theta_{p+1}=1$. We can thus write the $\boldsymbol{\theta}$-dependent part of the likelihood as a function only of the p joint sufficient statistics m

$$y_i = \sum_{j=1}^{p+1} x_{ij} \qquad i = 1,\ldots,p \tag{6.11}$$

which are the total number of comparisons won by each object. Note that even if we defined y_{p+1} it could be determined from the other statistics by

$$\sum_{i=1}^{p+1} y_{p+1} = \sum_{i=1}^{p}\sum_{j=i+1}^{p+1} n_{ij} \tag{6.12}$$

I.e., if you add up the total number of wins for each object, you have to get the total number of comparisons.

These joint sufficient statistics[4] illustrate that any likelihood-based inference of $\boldsymbol{\theta}$, be it frequentist or f f,Bayesian, will depend only on the total number of comparisons won by each object (and the numbers of comparisons performed for each pair). This seems pretty straightforward, but when applied to rating sports teams, it means that some principles often applied will be irrelevant. From the model (which assumes one constant strength for each team), it's apparent that the order of the results can't matter, but this sufficient statistic also makes it clear that the concept of "quality wins" doesn't have independent meaning. Consider the following two scenarios:

1. RIT defeats Air Force (a strong team based on their other results) and loses to Bentley (a weak team); Air Force defeats Bentley as well.
2. RIT loses to Air Force and beats Bentley. Bentley (while still being much weaker overall based on their other game results) upsets Air Force in a game that Air Force won in the other scenario.

Some rating systems might explicitly reward RIT for their "good win" over Air Force in the first scenario, more than they penalize RIT for their "bad loss" against Bentley. But we know that any likelihood-based inference using the Bradley-Terry model would produce identical results in the two scenarios above, assuming that all of the other results besides the three specified were the same.

6.3 Other Extensions

The definition of completeness is extended in the obvious way. We're usually thinking of it for the vector $\mathbf{Y}$ of sufficient statistics, but we can write it more generally for a random vector $\mathbf{V} = \begin{pmatrix} V_1 \\ V_1 \\ \vdots \\ V_m \end{pmatrix}$. The joint pdf $f_{\mathbf{V}}(\mathbf{v};\boldsymbol{\theta})$ is said to be a complete family if the only function $u(\mathbf{v})$ whose expectation value

[4] Davidson and Solomon *Biometrika* **60**, 477 (1973)

is $E\left(u(\mathbf{V})\right) = 0$ for all $\boldsymbol{\theta} \in \Omega$ is $u(\mathbf{v}) = 0$.

Similarly, the Rao-Blackwell and Lehmann-Scheffé theorems still hold. In particular, if $\mathbf{Y}$ is a set of complete sufficient statistics, then a function $\varphi_\alpha(\mathbf{Y})$ which is an unbiased estimator of a parameter θ_α is the minimum variance unbiased estimator of θ_α.

And finally, the concept and properties of the regular exponential class of distributions carries through. If we can write the pdf (or pmf) for $\mathbf{X}$ as

$$f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = \exp\left(\sum_{j=1}^{m} \eta_j(\boldsymbol{\theta}) K_j(\mathbf{x}) + H(\mathbf{x}) + q(\boldsymbol{\theta})\right) \qquad (6.13)$$

Then

$$Y_j = \sum_{i=1}^{n} K_j(\mathbf{X}_i) \qquad j = 1, \ldots, m \qquad (6.14)$$

are complete joint sufficient statistics for the parameters $\boldsymbol{\theta}$.

Tuesday 19 April 2016
– Read Section 7.8 of Hogg

7 Minimal Sufficiency and Ancillary Statistics

Useful additional reading for this topic is Chapter Eight of Jaynes's *Probability Theory – The Logic of Science*, entitled "Sufficiency, Ancillarity and All That". In addition to an interesting perspective, it has lots of wonderfully Jaynesian snark, such as "From his failure to attack it when he was attacking almost every other principle of inference, we may infer that R. A. Fisher probably accepted the likelihood principle, although his own procedures did not respect it."

7.1 Minimal Sufficient Statistics

In our generalization to joint sufficient statistics, we allowed for the possibility that m statistics would provide the necessary information about the dependence of the likelihood on the k parameters, where m need not be equal to k. This can be taken to a trivial extreme, since one can always use the entire sample $X_1, X_2, \ldots, X_n$ as the set of joint sufficient statistics. What we're really interested in is the least information necessary, so basically the smallest set of sufficient statistics. Note, though, that there's more to it than just counting, since a slightly less trivial set of joint sufficient statistics for a sample is the order statistics $Y_1 < Y_2 < \cdots < Y_n$, since the likelihood can be written as

$$L(\boldsymbol{\theta};\mathbf{x}) = f(x_1;\boldsymbol{\theta}) \cdots f(x_n;\boldsymbol{\theta}) = f(y_1;\boldsymbol{\theta}) \cdots f(y_n;\boldsymbol{\theta}) \; ; \qquad (7.1)$$

since each value in the sample appears in the argument of one distribution function, each order statistic also appears once. Although either set has n statistics in it, information about the sequence of values is lost in going from the full sample to the n order statistics.

Hogg tries to provide some principles for deciding when we've found the minimal sufficient set, but unfortunately most of them are tautologies. For instance, if the maximum likelihood estimator $\widehat{\theta}$ is a sufficient statistic, it is minimal. But that's basically trivially true, since if one single statistic is sufficient, it's automatically the minimal set. The only way around this is if we can use a non-invertable function of the sufficient statistic and still have the result be sufficient. For instance, $|Y|$ might be just as good a sufficient statistic as Y. In any event, there is an important property of the mle: it has to be constructed from the sufficient statistics. (That's sort of the point of sufficient statistics in the first place.)

7.2 Sufficiency and Ancillarity

7.2.1 Example: Uniform Sampling Distribution

Hogg has an extended example that's actually really useful for illustrating an important point not illustrated in the text. Consider a sample of size n drawn from a uniform distribution

$$f(x;\theta) = \begin{cases} 1 & \theta - \frac{1}{2} < x < \theta + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{7.2}$$

The likelihood function is 1 if all of the $\{x_i\}$ lie in $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, and 0 if any of them lie outside it. This is easily summarized in terms of the order statistics $y_1 < y_2 < \cdots < y_n$ for the sample:

$$L(\theta;\mathbf{x}) = \begin{cases} 1 & \theta - \frac{1}{2} < y_1 < y_n < \theta + \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{7.3}$$

Since all of the other order statistics fall in between y_1 and y_n by construction, the restrictions on θ in the likelihood are just $y_n - \frac{1}{2} < \theta < y_1 + \frac{1}{2}$, and Y_1 and Y_n are joint sufficient statistics for θ. They turn out to also be minimal sufficient statistics, so in this case the number of sufficient statistics is greater than the number of parameters.

One option for a single estimator of θ is the maximum likelihood estimator, but in this case the likelihood is constant when it's not zero, so any θ between $y_n - \frac{1}{2}$ and $y_1 + \frac{1}{2}$ would work. We can take the midpoint of that interval, though, and define

$$T_1 = \frac{Y_1 + Y_n}{2} \tag{7.4}$$

It's natural to describe the "other" degree of freedom in the sufficient statistics by

$$T_2 = Y_n - Y_1 \tag{7.5}$$

i.e., the range over which the sample is spread. As we'll confirm in a moment, the marginal distribution for T_2 doesn't depend on the parameter θ. Such a statistic is called an *ancillary statistic*.

It's enlightening to examine the joint distribution for T_1 and T_2. The distribution for the order statistics Y_1 and Y_n is

$$\begin{aligned} f_{Y_1Y_n}(y_1, y_n;\theta) &= n(n-1)f(y_1)[F(y_n) - F(y_1)]^{n-2}f(y_n) \\ &= n(n-1)(y_n - y_1)^{n-2}\,, \qquad \theta - \frac{1}{2} < y_1 < y_n < \theta + \frac{1}{2} \end{aligned} \tag{7.6}$$

Since the Jacobian transforming between (y_1, y_n) and (t_1, t_2) has unit determinant, the joint distribution function is

$$\begin{aligned} f_{T_1T_2}(t_1, t_2;\theta) &= n(n-1)t_2^{n-2}\,, \\ & 0 < t_2 < 1; \quad \theta - \frac{1 - t_2}{2} < t_1 < \theta + \frac{1 - t_2}{2} \end{aligned} \tag{7.7}$$

The support space comes from enforcing the conditions

$$\theta - \frac{1}{2} < y_1 = t_1 - \frac{t_2}{2} \tag{7.8a}$$

$$t_1 + \frac{t_2}{2} = y_n < \theta + \frac{1}{2} \tag{7.8b}$$

If we marginalize over t_1, we can get the distribution for T_2:

$$\begin{aligned} f_{T_2}(t_2) &= n(n-1)t_2^{n-2}\int_{\theta - \frac{1-t_2}{2}}^{\theta + \frac{1-t_2}{2}} dt_1 \\ &= n(n-1)t_2^{n-2}(1 - t_2) \qquad 0 < t_2 < 1 \end{aligned} \tag{7.9}$$

which we see is indeed independent of θ. Although $\{T_2\}$ is an ancillary statistic, it is important for inferences about θ, which we see if we write the likelihood function as

$$L(\theta;\mathbf{x}) = \begin{cases} 1 & t_1 - \frac{1-t_2}{2} < \theta < t_1 + \frac{1-t_2}{2} \\ 0 & \text{otherwise} \end{cases} \tag{7.10}$$

So we have T_1 as an estimator of θ, and T_2 as an ancillary statistic which together with T_1 make up the joint sufficient statistics for θ. We say that T_2 is the *ancillary complement* to T_1. In this case t_2 tells us how good an estimate t_1 is for θ. If we constructed a confidence interval at confidence level $1-\alpha$, it would be centered at T_1 and have a width of $(1-\alpha)(1-T_2)$. Qualitatively, the more spread out the sample is, the more it constrains the possible value of θ.

7.2.2 Location Models

This example is a case of what are in general known as *location models* where the random variables in the sample can be written as

$$X_i = \theta + W_i \tag{7.11}$$

and the offsets $\{W_i\}$ are drawn from a distribution which doesn't depend on θ. The behavior of sufficient statistics is different depending on this distribution. For example:

- As we've seen, if the distribution for W_i is uniform with a known width, the sufficient statistics are the order statistics Y_1 and Y_n.
- If the distribution is $N(0,1)$ (or in general normal with a known variance), we know that the mle $\overline{X}$ is a sufficient statistic by itself.
- It turns out that if the offsets are drawn from a Cauchy distribution, you need the whole sample, or at least all n order statistics, to describe the shape of the likelihood function.

7.2.3 Example: Bernoulli Trials

As another example of ancillary statistics, consider the case of Bernoulli trials with a probability of success of θ, where $0 < \theta < 1$. If we have observed a sequence of n trials containing k successes, the likelihood function is

$$L(\theta) \propto \theta^k (1-\theta)^{n-k} \tag{7.12}$$

where there may be a factor depending on k and/or n depending on whether the observable is the sequence of successes and failures, the number of successes in a fixed number of trials (binomial), the number of trials needed to obtain a fixed number of successes (negative binomial), etc. We've said that the number of successes $K = \sum_{i=1}^{n} X_i$ is a sufficient statistic in the context of n being a fixed property of the experiment. And the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{k}{n} \tag{7.13}$$

But in general, both k and n can be thought of as properties of the data, and if we want to characterize our inference of θ, we need to know the number of trials and not just the fraction which were successful. To cast this in a frequentist framework, suppose that we draw a random variable N from any distribution which doesn't depend on θ (it can be Poisson with a fixed mean, but we really don't care about the specifics) and then perform N trials, of which K are successes. The likelihood is

$$L(\theta; n, k) = f_N(n) \binom{n}{k} \theta^k (1-\theta)^{n-k} \tag{7.14}$$

and we see that K and N (or $\widehat{\theta}$ and N) are minimal joint sufficient statistics for θ. The MLE $\widehat{\theta}$ is an estimator for θ, but we need to know N to know the associated uncertainty, even though the distribution of N doesn't depend on θ. N is the ancillary complement to the maximum likelihood estimator $\widehat{\theta}$.

Thursday 21 April 2016
– Read Section 7.9 of Hogg

7.3 Ancillarity and Independence

Suppose $\mathbf{Y} \equiv \{Y_1, Y_2, \ldots, Y_m\}$ is a the vector of m joint sufficient statistics for the parameters $\boldsymbol{\theta}$. We can show that if Z is some other statistic constructed from the data, then the conditional distribution

$$f_{Z|\mathbf{Y}}(z|\mathbf{y}) \tag{7.15}$$

is independent of $\boldsymbol{\theta}$. Note that the Bayesian equivalent of this would be

$$f(z|\mathbf{y}, \boldsymbol{\theta}) = f(z|\mathbf{y}) \ , \tag{7.16}$$

i.e., $\mathbf{y}$ gives you all of the information about the parameters $\boldsymbol{\theta}$ which is contained in the data. This is trivial if $z = u(\mathbf{y})$ is some function of the sufficient statistics, since that means the distribution for Z is degenerate, with 100% probability that $Z = u(\mathbf{Y})$. If z is not a function of $\mathbf{y}$, we can consider a transformation from $x_1, \ldots x_n$ to $v_1 = y_1, \ldots v_m = y_m$, $v_{m+1} = z$, $v_{m+2}, \ldots v_n$. Since the transformation is independent of $\boldsymbol{\theta}$, the Jacobian determinant will not contain $\boldsymbol{\theta}$, and we can write

$$f_{\mathbf{U}}(\mathbf{u}; \boldsymbol{\theta}) = |J(\mathbf{x})| \, f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) \tag{7.17}$$

The conditional distribution $f_{Z|\mathbf{Y}}(z|\mathbf{y})$ is

$$f_{Z|\mathbf{Y}}(z|\mathbf{y}) = \frac{f_{\mathbf{Y}Z}(\mathbf{y}, z; \boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} = \frac{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{U}}(\mathbf{u}; \boldsymbol{\theta}) \, du_{m+2} \cdots du_n}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} \tag{7.18}$$

But

$$\frac{f_{\mathbf{U}}(\mathbf{u}; \boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} = |J(\mathbf{x})| \, \frac{f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta})}{f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})} \tag{7.19}$$

and the fact that $\mathbf{Y}$ are joint sufficient statistics tells us that $f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) / f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$ is a function of $\mathbf{x}$ independent of $\boldsymbol{\theta}$, so $f_{Z|\mathbf{Y}}(z|\mathbf{y})$ is indeed independent of $\boldsymbol{\theta}$.

Note that Z is independent of $\mathbf{Y}$, in the sense of independent random variables, $f_Z(z) = f_{Z|\mathbf{Y}}(z|\mathbf{y})$, which means that Z is an ancillary statistic, since $f_Z(z)$ doesn't depend on $\boldsymbol{\theta}$.

7.3.1 Basu's Theorem

We can consider basically the converse. If Z is an ancillary statistic for $\boldsymbol{\theta}$ and $\mathbf{Y}$ the joint sufficient statistics for $\boldsymbol{\theta}$, under what circumstances does that imply Z and $\mathbf{Y}$ are independent? So we want to compare $f_Z(z)$ to $f_{Z|\mathbf{Y}}(z|\mathbf{y})$. Note that, by marginalization,

$$f_Z(z) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Z|\mathbf{Y}}(z|\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \theta) \, d^m y \tag{7.20}$$

On the other hand,

$$f_Z(z) = f_Z(z) \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{Y}}(\mathbf{y}; \theta) \, d^m y \tag{7.21}$$

since the second factor is just the normalization integral for $f_{\mathbf{Y}}(\mathbf{y}; \theta)$. If we subtract these two expressions, we get

$$0 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [f_{Z|\mathbf{Y}}(z|\mathbf{y}) - f_Z(z)] \, f_{\mathbf{Y}}(\mathbf{y}; \theta) \, d^m y \tag{7.22}$$

Now, if $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ is a complete family of distributions, the only function which integrates to zero when weighted by $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ is identically zero, and thus $f_{Z|\mathbf{Y}}(z|\mathbf{y}) = f_Z(z)$. This is known as Basu's Theorem:

If $\mathbf{Y}$ is are joint complete sufficient statistics for $\boldsymbol{\theta}$, any ancillary statistic Z is independent of $\mathbf{Y}$.

7.3.2 Example: Mean of a Normal Distribution

As an example, consider a sample of size n from a $N(\theta, \sigma^2)$ distribution. We know the sample mean $\overline{X}$ is a complete sufficient

statistic for θ. If we construct the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{7.23}$$

we know by Student's theorem that this is independent of $\overline{X}$, and is equal to $\sigma^2/(n-1)$ times a $\chi^2(n-1)$ random variable. But even if we didn't already know about the independence, it would follow from Basu's theorem, since S^2 is an ancillary statistic for θ (its distribution function doesn't depend on θ). In fact, we see that *any* ancillary statistic for θ will be independent of $\overline{X}$. In particular, any location-invariant statistic like $Y_n - Y_1$ (the spread of the dataset) is an auxiliary statistic for the mean θ and therefore independent of $\overline{X}$.

7.4 Location and Scale Models Revisited

Recall the concept of a location model, where the random variables in the sample can be written

$$X_i = \theta + W_i \tag{7.24}$$

where the probability distribution $f_W(w)$ from which the $\{W_i\}$ are drawn doesn't depend on θ. A statistic $Z = u(X_1, \ldots, X_n)$ is called *location-invariant* if the function $u(x_1, \ldots, x_n)$ is unchanged by adding the same constant d to each of its arguments:

$$u(x_1 + d, x_2 + d, \ldots, x_n + d) = u(x_1, x_2, \ldots, x_n) \tag{7.25}$$

A location-invariant statistic can be written

$$Z = u(X_1, \ldots, X_n) = u(X_1 - \theta, \ldots, X_n - \theta) = u(W_1, \ldots, W_n)\ ; \tag{7.26}$$

since Z is a function only of the $\{W_i\}$, its probability distribution cannot depend on θ, so it is an ancillary statistic.

You could imagine exponentiating the whole picture, so that addition turns into multiplication. We say that we have a scale model when the random variables can be written

$$X_i = \theta W_i \tag{7.27}$$

where again the distribution $f_W(w)$ is independent of the *scale parameter* θ. In that case, a scale-invariant statistic

$$Z = u(X_1, \ldots, X_n) \qquad u(cx_1, \ldots, cx_n) = u(x_1, \ldots, x_n) \tag{7.28}$$

is an ancillary statistic for θ.

Similarly, a location and scale family is one where

$$X_i = \theta_1 + \theta_2 W_i \tag{7.29}$$

with $\{W_i\}$ drawn from a distribution which doesn't depend on θ_1 or θ_2. For example. if $X_i \sim N(\theta_1, \theta_2^2)$, then any statistic constructed from

$$W_i = \frac{X_i - \theta_1}{\theta_2} \tag{7.30}$$

is location- and scale-invariant, and therefore ancillary for θ_1 and θ_2.

Finally, there are of course some parameters which are neither location parameters nor scale parameters. If $\{X_i\}$ is drawn from a distribution with pdf

$$f(x; \theta_1, \theta_2, \theta_3) = \frac{(x-\theta_1)^{\theta_3 - 1}}{\theta_2^{\theta_3}} e^{-(x-\theta_1)/\theta_2} \qquad \theta_1 < x < \infty \tag{7.31}$$

then we can construct

$$W_i = \frac{X_i - \theta_1}{\theta_2} \sim \text{Gamma}(\theta_3, 1) \tag{7.32}$$

we say θ_1 is a location parameter, θ_2 is a scale parameter, and θ_3 is a *shape parameter*.