apts.ac.uk
Academy for PhD Training in Statistics

Statistical Inference

preliminary material

D. R. Cox and D. Firth

Introduction

These notes and exercises are designed to help students to prepare for the first APTS week, in order to get the most out of the intensive module on *Statistical Inference*. Many APTS students will have met all of this material before, as undergraduates or at Masters level; others may have seen only some parts of it. Some of the material is very basic indeed, and is included here only for completeness. The APTS-week lectures themselves will be at a rather higher level, and will assume that students already have a solid grasp of everything that appears here.

Interspersed with the notes are some exercises. The ideal preparation would be to do enough work to allow you to understand the notes in detail and to complete all of the exercises. The amount of work needed is likely to vary from one student to another. Students who find themselves unable to complete all of the exercises in, say, 3 full days of work are advised to spend at least a whole week acquiring/refreshing the necessary background knowledge.

The notes here are brief, and should ideally be supplemented by reading from a good textbook or two. Casella, G. and Berger, R. L., *Statistical Inference* (2nd edn; Duxbury, 2002) is a good text book at about the right level for this preliminary material (there are of course others). For the APTS week itself, the most appropriate single book would be Cox, D. R., *Principles of Statistical Inference* (Cambridge University Press, 2006).

The notes are arranged with plenty of white space, to facilitate annotation by hand as you work through them.

Part I

Some commonly used (univariate) probability models

Distributions in statistics serve two main purposes:

- to describe the assumed behaviour of the observations made in an experiment, survey or other study;
- to calibrate the values of derived statistics used in constructing confidence regions, hypothesis tests, posterior distributions, etc.

Some distributions are much used for both purposes (the normal distribution being the prime example).

In this Part we review some key (families of) distributions used for the first purpose. Distributions used mainly for the second purpose include the χ^2, t and F distributions, which will be briefly reviewed in Part 2.

Some abbreviations that will be used, in connection with the distribution of a random variable Y:

cdf: cumulative distribution function, $F_Y(y) = \mathrm{pr}(Y \leq y)$;

pmf: probability mass function (for discrete random variable), $f_Y(y) = \mathrm{pr}(Y = y)$;

pdf: probability density function (for absolutely continuous random variable), f_Y such that $F_Y(y) = \int_{-\infty}^{y} f_Y(z)dz$;

mgf: moment generating function, $M_Y(t) = E(e^{tY})$, when the expectation exists for t in a neighbourhood of $t = 0$.

Binomial distribution

The distribution of the number of 'successes' in m independent binary 'trials'; or, equivalently, random sampling (with replacement) from a binary population.

The pmf is

$$f_Y(y) = \binom{m}{y} \theta^y (1-\theta)^{m-y} \quad (y = 0, 1, \ldots, m).$$

where θ is the probability of success (assumed constant for all trials).

The mean and variance are $m\theta$ and $m\theta(1-\theta)$, and the mgf is $M_Y(t) = [\theta e^t + (1-\theta)]^m$.

Special case $m = 1$: the *Bernoulli* distribution

When $m = 1$,

$$\begin{aligned} f_Y(y) &= \theta^y (1-\theta)^{1-y} \quad (y = 0, 1) \\ &= \begin{cases} \theta & (y = 1) \\ 1-\theta & (y = 0) \end{cases} \end{aligned}$$

This simple distribution is the *Bernoulli* distribution.

Independent trials with binary outcomes are often referred to as *Bernoulli trials*.

Negative binomial and geometric distributions

The negative binomial is the distribution of the number of Bernoulli trials needed in order to see k successes (for any fixed integer $k > 0$). If Y is the trial at which the kth success occurs, the pmf of Y is

$$f_Y(y) = \binom{y-1}{k-1} \theta^k (1-\theta)^{y-k} \quad (y = k, k+1, \ldots)$$

The name 'negative binomial' comes from noting that if $Z = Y - k$ (the number of failures seen before the kth success),

$$f_Z(z) = (-1)^z \binom{-k}{z} \theta^k (1-\theta)^z \quad (z = 0, 1, 2, \ldots)$$

which looks strikingly similar to the binomial pmf.

The mean and variance of Z are $k(1-\theta)/\theta$ and $k(1-\theta)/\theta^2$ respectively.

The mgf is $M_Z(t) = [\theta/\{1-(1-\theta)e^t\}]^k$.

The *geometric distribution* is the special case with $k=1$; i.e., Z is the number of failures seen before the *first* success.

Importantly, the negative binomial also arises (***exercise***) as the marginal distribution of a random variable Z whose distribution conditional upon a *gamma-distributed* latent variable M is $Z|M \sim \text{Pois}(M)$. This is useful when modelling 'overdispersed' (relative to the Poisson distribution) count data.

Poisson distribution

The distribution of a count of events that occur (separately and independently, by assumption) in time, or space, say, according to a *Poisson process*.

A Poisson rv takes any value in $\{0,1,2,\ldots\}$, and has pmf

$$f_Y(y) = e^{-\mu}\mu^y/y! \quad (y=0,1,2,\ldots).$$

The mean — the expected number of events — is μ. The variance is also μ. The mgf is $M_Y(t) = \exp[\mu(e^t-1)]$.

If Y and Z are independently Poisson distributed with means μ and λ, then $Y+Z \sim \text{Pois}(\lambda+\mu)$.

(***exercise***: prove these last four statements)

Some relationships

The *Poisson* distribution plays a useful approximation role for some of the other main discrete distributions:

- the $\text{Bin}(m,\theta)$ is well approximated by $\text{Pois}(m\theta)$ when θ is small.
- the $\text{NegBin}(k,\theta)$ is well approximated by $\text{Pois}[k(1-\theta)]$ for k large and θ close to 1.

Poisson and binomial: an *exact* relationship

In addition to the approximation of binomial probabilities using the Poisson pmf, mentioned above, we have the following.

Equivalence of binomial and conditional Poisson sampling

If Y and Z are independent Poisson rv's with means λ and μ, then the conditional distribution of Y, given $Y + Z = t$, is Bin$[t,\ \lambda/(\lambda+\mu)]$.

Proof: simply apply the definition of conditional probability, $\mathrm{pr}(Y = y|Y+Z = t) = \mathrm{pr}(Y = y)\,\mathrm{pr}(Z = t-y)/\,\mathrm{pr}(Y+Z = t)$, and use the fact that the Poisson family is closed under independent addition. (***exercise***)

Exponential distribution

The exponential distribution is often used to describe the distribution of measured time intervals ('duration data' or 'waiting-time data'). The pdf is

$$f_Y(y) = \begin{cases} \frac{1}{\mu}\exp(-y/\mu) & (y > 0) \\ 0 & \text{(otherwise)} \end{cases}$$

The mean and variance are μ and μ^2, and the mgf is

$$M_Y(t) = \frac{1}{1 - t\mu} \qquad (t < 1/\mu).$$

(***exercise***: verify these)

Gamma distribution

The gamma family generalizes the exponential. The pdf is

$$f_Y(y) = \begin{cases} \frac{\alpha}{\mu}\frac{1}{\Gamma(\alpha)} z^{\alpha-1} e^{-z} & (z > 0) \\ 0 & \text{(otherwise)}, \end{cases}$$

where $z = \alpha y/\mu$. The mean of Y is μ. The extra parameter $\alpha > 0$ is often called the 'shape' parameter; the exponential distribution is the special case $\alpha = 1$.

The mgf is $M_Y(t) = 1/(1 - \mu t/\alpha)^\alpha$ $(t < \alpha/\mu)$. From this, for example, we see how α generalizes the mean-variance relationship:

$$\mathrm{var}(Y) = \mu^2/\alpha,$$

so the *coefficient of variation*, $\mathrm{sd}(Y)/E(Y)$, is $1/\sqrt{\alpha}$.
(***exercise***: verify these statements)

We will write Gamma(μ, α) as shorthand for the above parameterization of a gamma distribution.

The gamma, like the exponential, is also often used for modelling durations (lengths of time intervals).

From the mgf we see immediately that, when α is a positive integer, the gamma distribution is the distribution of the sum of α independent exponential random variables each having mean μ/α.

Beta distribution

The beta distributions are distributions on the unit interval $(0, 1)$.

The pdf of a beta distribution is

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & (0 < x < 1) \\ 0 & \text{(otherwise)} \end{cases}$$

where $B(\alpha, \beta)$ is the *beta function*,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}.$$

The beta family includes a variety of distributional shapes, including the uniform distribution ($\alpha = \beta = 1$).

The beta distribution has

$$\begin{aligned} \mu = E(X) &= \frac{\alpha}{\alpha + \beta} \\ \text{var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

and the rather less elegant

$$M_X(t) = 1 + \sum_{k=1}^{\infty} \left(\prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r} \right) \frac{t^k}{k!}.$$

The mean is thus determined by the *relative* values of α and β.

The variance is inversely related to the sum $\alpha + \beta$: it can be re-expressed as $\mu(1 - \mu)/(\alpha + \beta + 1)$.

Normal (or Gaussian) distribution

The most-used of all continuous distributions (largely on account of the *Central Limit Theorem*).

The pdf of the $N(\mu, \sigma^2)$ distribution is

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \qquad (-\infty < y < \infty) \\ &= \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) \end{aligned}$$

where $\phi(y) = \exp(-y^2/2)/\sqrt{2\pi}$ is the pdf of the *standard normal* distribution $N(0,1)$.

The parameters μ and σ are respectively *location* and *scale* parameters: for any constants c and d, linear transformation $cY + d$ has the normal distribution with location $c\mu + d$ and scale $c\sigma$.

The mean, variance and mgf are

$$\begin{aligned} E(Y) &= \mu \\ \text{var}(Y) &= \sigma^2 \\ M_Y(t) &= \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right) \end{aligned}$$

(***exercise***: prove these)

The normal cdf

The cdf of the $N(\mu, \sigma^2)$ distribution is

$$F_Y(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt = \Phi\left(\frac{y-\mu}{\sigma}\right),$$

where $\Phi(z) = \int_{-\infty}^{z} \phi(t)dt$ is the cdf of $(Y-\mu)/\sigma$.

Values of $\Phi(z)$ must be read from a table. By symmetry, $\Phi(-z) = 1 - \Phi(z)$.

Some values of Φ worth remembering: $\Phi(1.64) \approx 0.95$, and $\Phi(1.96) \approx 0.975$. The latter, for example, says that

$$\text{pr}(\mu - 1.96\sigma < Y < \mu + 1.96\sigma) = \Phi(1.96) - \Phi(-1.96) \approx 0.95$$

i.e., roughly, about 95% of probability is within 2 standard deviations of the mean.

Transformation: the lognormal distribution

A much-used distribution for modelling positive quantities, in economics in particular, is the *log-normal* distribution.

If $Y \sim N(\mu, \sigma^2)$, then $W = \exp(Y)$ is said to be log-normal with parameters μ and σ.

The pdf is

$$f_W(w) = \frac{1}{w\sigma\sqrt{2\pi}} \exp\left[-\frac{(\log w - \mu)^2}{2\sigma^2}\right] \qquad (w > 0)$$

(***exercise***)

The mean and variance are $E(W) = \exp(\mu + \sigma^2/2)$, $\text{var}(W) = [E(W)]^2[\exp(\sigma^2) - 1]$, and the integral formally defining the mgf does not converge for any real $t \neq 0$.

Connections between distributions: Normal approximation

The normal family can be used — largely on account of the Central Limit Theorem — to approximate various other distributions.

Some prominent examples are:

- approximation of Pois(λ) by $N(\lambda, \lambda)$, for large values of λ
- approximation of Bin(m, θ) by $N[m\theta,\ m\theta(1-\theta)]$, for large m (and θ not too close to 0 or 1).
- approximation of Gamma(μ, α) by $N(\mu,\ \mu^2/\alpha)$, for large values of α.

The Central Limit Theorem tells us that the normal can be used to approximate the distribution of *any* random variable which can be thought of as the sum of a large number of independent, identically distributed components. All of the above examples are of this kind:

- $Y \sim$ Pois(λ) can be thought of as $\sum_1^n Y_i$, where the Y_i are independent Pois(λ/n)
- $Y \sim$ Bin(m, θ) is $\sum_1^m Y_i$ where $Y_i \sim$ Bin$(1, \theta)$ are independent
- $Y \sim$ Gamma(μ, α) can be thought of as $\sum_1^n Y_i$ where the Y_i are independent Gamma$(\mu/n,\ \alpha/n)$.

Normal approximation in practice: continuity correction

When approximating a *discrete* distribution, the normal approximation is much improved by use of a 'continuity correction'.

Example: $Y \sim \text{Bin}(25,\ 0.6)$
The approximating normal distribution is then $N(15,\ 6)$. A binomial probability such as

$$\text{pr}(Y \leq 13) = \sum_{y=0}^{13} \binom{25}{y} (0.6)^y (0.4)^{25-y} = 0.267$$

can then be approximated as

$$\Phi\left(\frac{13-15}{\sqrt{6}}\right) = \Phi(-0.82) = 0.206$$

— but this is not a very good approximation!

Much better is to recognise that $\text{pr}(Y \leq 13)$ is the same as $\text{pr}(Y \leq 13.5)$, and to approximate the latter:

$$\Phi\left(\frac{13.5-15}{\sqrt{6}}\right) = \Phi(-0.61) = 0.271.$$

Some *exact* relationships

The gamma, Poisson and normal families are related to one another also in various *exact* ways.

These include the following important relationships:

- Poisson with gamma
- normal with gamma

The Poisson-gamma relationship

Poisson and gamma (which includes exponential) are closely related when the gamma shape parameter α is an integer.

(This is because waiting times in a *Poisson process* model for randomly occurring events in continuous time are gamma-distributed.)

Specifically, if $Z \sim \text{Gamma}(\alpha, \beta)$, then for any $t > 0$

$$\text{pr}(Z > t) = \text{pr}(Y < \alpha)$$

where $Y \sim \text{Pois}(t/\beta)$.

Special case $\alpha = 1$ (exponential distribution) is most easily shown:

$$\text{pr}(Z > t) = \text{pr}(Y = 0) = \exp(-t/\beta).$$

The normal-gamma (exact) relationship

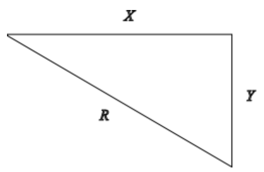
Suppose that $Y \sim N(0, \sigma^2)$, and consider $Z = Y^2$. The pdf of Z is

$$\begin{aligned} f_Z(z) &\propto f_Y(\sqrt{z}) \left| \frac{1}{2\sqrt{z}} \right| \qquad (z > 0) \\ &\propto z^{-1/2} \exp[-z/(2\sigma^2)] \end{aligned}$$

which we recognise as the kernel of the $\text{Gamma}(\sigma^2, \frac{1}{2})$ pdf. Hence Y^2 has this particular gamma distribution.

The distribution of the *standardized* squared normal, Y^2/σ^2, is thus $\text{Gamma}(1, \frac{1}{2})$. This is the *chi-squared distribution with one degree of freedom*.

Continuing a little further with this: suppose X and Y are independent $N(0, \sigma^2)$, and let R be the length of the random vector (X, Y):



$$R = \sqrt{X^2 + Y^2}$$

Then R^2 has an exponential distribution.

Proof: $M_{X^2}(t) = M_{Y^2}(t) = 1/(1 - 2\sigma^2 t)^{1/2}$, so $M_{R^2}(t) = 1/(1 - 2\sigma^2 t)$, which is the mgf of the $\text{Exp}(2\sigma^2)$ distribution.

Part II

Sampling from a normal distribution

Distributions derived from $N(\mu, \sigma^2)$

- the chi-squared distributions, χ^2_n ($n = 1, 2, \ldots$)
- the variance-ratio (or "F") distributions, $F_{m,n}$ ($m, n \in \{1, 2, \ldots\}$)
- the "Student t" distributions, t_n ($n = 1, 2, \ldots$)

The chi-squared distributions

Definition: if $Y_1, \ldots, Y_n$ are independent $N(0, 1)$, then

$$Y = Y_1^2 + \ldots + Y_n^2 \sim \chi^2_n.$$

In words: the sum of n squared, independent standard normal random variables is said to have the *chi-squared distribution with n degrees of freedom*.

All chi-squared distributions have support on $(0, \infty)$ and are skewed to the right (sketch a typical pdf). The cdf is tabluated.

Chi-squared and gamma

We have already seen in Part 1 that

$$Y_1^2 + \ldots + Y_n^2 \sim \text{Gamma}(\mu = n,\ \alpha = \frac{n}{2})$$

— so every chi-squared distribution is of the gamma form.

From this we also have immediately that

- the mean of a χ_n^2 rv is n, and the variance is $2n$;
- the Exponential(μ) distribution is the distribution of $\mu Y/2$ where $Y \sim \chi_2^2$.

The F distributions

Definition: if $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ independently, then

$$R = \frac{X/m}{Y/n} \sim F_{m,n}.$$

In words: the ratio of two independent chi-squared rv's, each scaled to have mean 1, is said to have the F *distribution with degrees of freedom* m *and* n.

Sometimes m is called the *numerator degrees of freedom*, and n the *denominator degrees of freedom*.

Clearly if $R \sim F_{m,n}$ then $1/R \sim F_{n,m}$.

The $F_{m,n}$ cdf's are tabulated.

The t distributions

Definition: if $X \sim N(0, 1)$ and $Y \sim \chi_n^2$ independently, then

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n.$$

In words: the ratio of a standard normal rv to the square root of a scaled chi-squared rv has the *Student* t *distribution with* n *degrees of freedom*.

("Student": W. S. Gosset, 1876–1937)

The cdf's of the t_n distributions are tabulated.

Note that as $n \to \infty$, T converges in distribution to X (by Slutsky's theorem, since Y/n converges in probability to 1). The t distributions are like the normal, but with "fatter tails".

Relationship between t and F

If $T \sim t_n$, then

$$T^2 = \frac{X^2}{Y/n} = \frac{X^2/1}{Y/n} \sim F_{1,n}.$$

So every F distribution with 1 numerator df is the distribution of a *squared* t-distributed rv.

Distribution of $\bar{Y}$ and S^2

Suppose that $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$, and let

$$\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2.$$

Four important things to know:

(a) $\bar{Y}_n \sim N(\mu,\ \sigma^2/n)$
(b) $\bar{Y}_n$ and S_n^2 are *independent*
(c) $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$
(d) $(\bar{Y}_n - \mu)/(S_n/\sqrt{n}) \sim t_{n-1}$

Interpretation/applications of properties (a)–(d)

A very brief overview:

(a) $\bar{Y}_n \sim N(\mu,\ \sigma^2/n)$ can be used for inference on μ when σ is known. In practice this is fairly rare, though: σ is most often *not* known.

(b) Independence of $\bar{Y}$ and S^2: e.g., the sample mean has no predictive power for the average size of (squared) deviations from the sample mean.

(c) $(n-1)S_n^2/\sigma^2 \sim \chi^2_{n-1}$ can be used for inference on σ^2 when μ is unknown — in essence, it is a 'corrected' version of the result that would hold if μ were known, namely $\sum(Y_i-\mu)^2/\sigma^2 \sim \chi^2_n$. The correction is to take account of the use of $\bar{Y}$ in place of μ.

(d) $(\bar{Y}_n-\mu)/(S_n/\sqrt{n}) \sim t_{n-1}$ is the corresponding 'corrected' version of (a) — corrected, that is, for the replacement of σ by S_n. It allows straightforward inference on μ when σ is unknown.

Part III

Likelihood and sufficiency

Likelihood

Consider a statistical model for random vector Y whose distribution depends on an unknown parameter (vector) θ.

Write $f(Y;\theta)$ for the joint pdf or pmf of random vector $Y=(Y_1,\ldots,Y_n)$ when θ is the value of the unknown parameter. Then, given that $Y=y$ is observed, the function of θ defined by

$$L(\theta; y) = f(y;\theta)$$

is the *likelihood function* for θ based on data y.

For any fixed value of θ, say $\theta = \theta_1$, $L(\theta_1; Y)$ is a *statistic* — a scalar-valued transformation of Y.

Note the key distinction between

- f, which is considered as a function of y (and, for example, must sum or integrate to 1)
- L, which is considered as a function of θ.

The purpose of $L(\theta; y)$ is to compare the plausibility of different candidate values of θ, given the observed data y.

If $L(\theta_1; y) > L(\theta_2; y)$, then the data y were more likely to occur under the hypothesis that $\theta = \theta_1$ than under the hypothesis that $\theta = \theta_2$. In that sense, θ_1 is a more plausible value than θ_2 for the unknown parameter θ.

Likelihood ratio

The relative plausibility of candidate parameter values, θ_1 and θ_2 say, may be measured by the *likelihood ratio*,

$$\frac{L(\theta_1; y)}{L(\theta_2; y)}.$$

Interpretation: for example, if $L(\theta_1; y)/L(\theta_2; y) = 10$, then the observed data y were 10 times more likely under truth θ_1 than under truth θ_2.

The use of likelihood *ratios* to compare the plausibility of different θ-values means that any *constant* factor in the likelihood — that is, any factor not depending on θ — can be neglected.

Example: $Y_i \sim \text{Bin}(m_i, \theta)$, independently $(i = 1, \ldots, n)$.

Here

$$\begin{aligned} L(\theta; y) &= \prod_{i=1}^{n} \binom{m_i}{y_i} \theta^{y_i} (1-\theta)^{m_i - y_i} \\ &= \text{constant} \times \left(\frac{\theta}{1-\theta}\right)^{\sum_1^n y_i} (1-\theta)^{\sum_1^n m_i}. \end{aligned}$$

- the binomial coefficients $\binom{m_i}{y_i}$ are not needed, since they do not involve θ
- the (non-constant part of) the likelihood depends on y only through $s(y) = \sum_1^n y_i$.

The function $s(Y) = \sum_{i=1}^n Y_i$ here is a *sufficient statistic* for θ: the value of $s(y)$ is all the knowledge that is needed of y in order to compute the likelihood (ignoring constants).

A note on continuous distributions

For a continuous rv Y the pdf is not invariant to a change of measurement scale. If $Z = g(Y)$, then

$$f_Z(z;\theta) = f_Y[g^{-1}(z);\theta] \left| \frac{dy}{dz} \right| .$$

But the derivative factor here does not involve θ; the likelihood for data y, or for the equivalent data $z = g(y)$, is thus

$$L(\theta; z) = L(\theta; y) \times \text{ constant},$$

i.e., likelihood (unlike probability density) is essentially unaffected by a change of measurement scale.

Log likelihood

In practice, especially when observations are independent, it is usually most convenient to work with the (natural) logarithm of the likelihood,

$$l(\theta) = \log L(\theta),$$

since this converts products into sums, which are easier to handle.

Example: n independent binomials (continued),

$$\begin{aligned} l(\theta) &= \log \left[\text{constant } \times \left(\frac{\theta}{1-\theta} \right)^{\sum_1^n y_i} (1-\theta)^{\sum_1^n m_i} \right] \\ &= \text{constant } + \left(\sum_{i=1}^n y_i \right) \log \left(\frac{\theta}{1-\theta} \right) + \left(\sum_{i=1}^n m_i \right) \log(1-\theta). \end{aligned}$$

In terms of the log-likelihood, then, any two candidate values of θ are compared via the *log-likelihood-ratio*

$$\log \frac{L(\theta_1)}{L(\theta_2)} = l(\theta_1) - l(\theta_2).$$

On the log scale, it is *additive* constants that can be ignored.

Sufficiency

We have introduced the notion of *sufficient statistic* already, informally, as a data summary that provides all that is needed in order to compute the likelihood.

Here we will give a formal definition, and then prove the *factorization theorem*, which

- provides a straightforward way of checking whether a particular statistic is sufficient
- allows a sufficient statistic, to be identified by simple inspection of the likelihood function (as we did in the example of n binomials)

Sufficient statistic: the definition

A statistic $s(Y)$ is said to be a *sufficient statistic for* θ if the conditional distribution of Y, given the value of $s(Y)$, does not depend on θ.

In this precise sense, a sufficient statistic $s(Y)$ carries all of the information about θ that is contained in Y. The notion is that, given the observed value $s(y)$ of $s(Y)$, all further knowledge about y is uninformative about θ.

In particular, this is useful for *data reduction*: e.g., if $s(Y)$ is a *scalar* sufficient statistic, then all of the information in $y = (y_1, \ldots, y_n)$ relating to θ is contained in the single-number summary $s(y)$ (assuming the model is correct).

The factorization theorem

Statistic $s(Y)$ is sufficient for θ if and only if, for all y and θ,

$$f(y;\theta) = g(s(y), \theta)h(y) \qquad (*)$$

for some pair of functions $g(t, \theta)$ and $h(y)$.

Proof: (discrete case)

Suppose that $s(Y)$ is sufficient. Let

$$g(t, \theta) = \mathrm{pr}(s(Y) = t),$$

and

$$h(y) = \mathrm{pr}[Y = y | s(Y) = s(y)]$$

(the latter of which does not involve θ). Then

Then

$$\begin{aligned} f(y;\theta) &= \text{pr}(Y = Y) \\ &= \text{pr}[Y = y \text{ and } s(Y) = s(y)] \\ &= \text{pr}[s(Y) = s(y)]\,\text{pr}[Y = y|s(Y) = s(y)] \\ &= g(s(y),\theta)h(y). \end{aligned}$$

Now suppose that (*) holds. Write $q(t;\theta)$ for the pmf of $s(Y)$. Define the sets $A_t = \{z : s(z) = t\}$. Then

$$\begin{aligned} \text{pr}[Y = y|s(Y) = s(y)] = \frac{f(y;\theta)}{q(s(y);\theta)} &= \frac{g(s(y),\theta)h(y)}{\sum_{A_{s(y)}} g(s(z),\theta)h(z)} \\ &= \frac{g(s(y),\theta)h(y)}{g(s(y),\theta)\sum_{A_{s(y)}} h(z)}, \end{aligned}$$

which is $h(y)/\sum_{A_{s(y)}} h(z)$ and does not involve θ.

An essentially similar argument applies in the continuous case.

Example: $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$, with σ known.

We can write

$$f(y;\mu) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{2\sigma^2}\right)}_{h(y)} \underbrace{\exp\left(-n\frac{(\bar{y} - \mu)^2}{2\sigma^2}\right)}_{g(\bar{y},\mu)}$$

— so $\bar{Y}$ is a sufficient statistic for μ.

(***exercise***: verify this.)

Example: $Y_1, \ldots, Y_n$ iid discrete Uniform random variables on $\{1, 2, \ldots, \theta\}$

(e.g., a town has bus routes numbered $1, \ldots, \theta$, with θ being unknown; data are n bus numbers sampled at random.)

For each Y_i the pmf is

$$f(y;\theta) = \begin{cases} 1/\theta & (y = 1, 2, \ldots, \theta) \\ 0 & (\text{otherwise}) \end{cases}$$

so the joint pmf is

$$f(y,\theta) = \begin{cases} 1/\theta^n & (\text{all } y_i \in \{1, 2, \ldots\} \text{ and } \max(y_i) \le \theta) \\ 0 & (\text{otherwise}) \end{cases}$$

Hence, if we let $s(Y) = \max(Y_i)$, then

...then

$$f(y;\theta) = g(s(y),\theta)h(y),$$

where

$$g(t,\theta) = \begin{cases} 1/\theta^n & (t \le \theta) \\ 0 & (\text{otherwise}) \end{cases}$$

and

$$h(y) = \begin{cases} 1 & (y \in \{1,2\ldots\}) \\ 0 & (\text{otherwise}) \end{cases}$$

Hence $s(Y) = \max(Y_i)$ is a sufficient statistic for θ.

Minimal sufficient statistic

There clearly is no *unique* sufficient statistic in any problem. For if $s(Y)$ is a scalar sufficient statistic, then for example

(i) $r(s(Y))$ is sufficient, for any 1-1 function $r(.)$

(ii) the pair $\{s(Y), Y_1\}$, for example, is sufficient

(iii) the full vector Y is *always* (trivially) sufficient

(***exercise***: use the factorization theorem to check these assertions)

The idea of a *minimal* sufficient statistic is to eliminate redundancy of the kind evident in (ii) or (iii) [but not (i)] above, in order to achieve *maximal* reduction of the data from y to $s(y)$.

Definition

Sufficient statistic $s(Y)$ is said to be *minimal sufficient* if, for any other sufficient statistic $s'(Y)$, $s(Y)$ is a function of $s'(Y)$ [i.e., whenever $s'(y) = s'(z)$, we have that $s(y) = s(z)$].

The definition is clear enough in its meaning, but is not constructive: it does not help us to *find* a minimal sufficient statistic in any given situation.

The following theorem helps:

Theorem (Lehmann and Scheffé)

Suppose that statistic $s(Y)$ is such that for every pair of sample points y and z the ratio

$$\frac{f(y;\theta)}{f(z;\theta)}$$

is constant if and only if

$$s(y) = s(z).$$

Then $s(Y)$ is minimal sufficient.

Proof: omitted. See, e.g., Casella & Berger p281.

Example: $Y_1, \ldots, Y_n$ Uniform on the interval $(\theta,\ \theta + 1)$

The joint pdf of Y is

$$f(y;\theta) = \begin{cases} 1 & (\theta < y_i < \theta + 1 \qquad \forall i) \\ 0 & (\text{otherwise}) \end{cases}$$

which can be usefully re-expressed as

$$f(y;\theta) = \begin{cases} 1 & (\max(y_i) - 1 < \theta < \min(y_i)) \\ 0 & (\text{otherwise}) \end{cases}$$

Thus, for two sample points y and x, $f(y;\theta)/f(z;\theta)$ takes the constant value 1 (for all θ for which the ratio is defined) if and only if both $\min(y_i) = \min(z_i)$ and $\max(y_i) = \max(z_i)$.

Example [Unif($\theta,\ \theta + 1$) continued]

Hence the two-component statistic

$$s(Y) = \{\min(Y_i),\ \max(Y_i)\}$$

is a minimal sufficient statistic for this problem.

Note, then, that the minimal sufficient statistic in a one-parameter problem is not necessarily a scalar.

Part IV

Exponential families

Exponential families

A family of distributions is a set of distributions indexed (smoothly) by a parameter (in general, a vector) θ.

Suppose that θ is d-dimensional, and that the joint pdf (or pmf) of vector rv Y can be written as

$$f(y) = m(y)\exp[s^T(y)\phi - k(\phi)]$$

for some d-dimensional statistic $s(Y)$ and one-one transformation ϕ of θ. Then $S = s(Y)$ is sufficient, and (subject to regularity conditions) the model is a *full exponential family* with *canonical parameter* ϕ.

Example: $Y \sim$ Binomial(m, θ). This is a full, one-parameter exponential family, with $\phi = \log[\theta/(1-\theta)]$. (***exercise***)

Example: $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ (iid). This is a full, 2-parameter exponential family with sufficient statistic $\{\sum Y_i,\ \sum Y_i^2\}$. (***exercise***)

Example: $Y_1, \ldots, Y_n \sim N(\mu, \mu^2)$ (iid) — normal distribution with unit coefficient of variation. The minimal sufficient statistic is still $\{\sum Y_i,\ \sum Y_i^2\}$, but this is only a 1-parameter model. So this is *not* a full exponential family. This is an example of a *curved* exponential family. A curved EF with d-dimensional parameter is derived from a full exponential family of dimension k ($k > d$) by imposing one or more nonlinear constraints on the canonical parameters of the full EF.

Mean parameterization of a full EF:

Since every value of ϕ indexes a distribution, we have that

$$\int m(y)\exp[s^T\phi - k(\phi)]dy = 1$$

and indeed, for any d-vector p,

$$\int m(y)\exp[s^T(\phi + p) - k(\phi + p)]dy = 1.$$

Hence the mgf of S is

$$M_S(p) = E[\exp(p^T S)] = \exp[k(\phi + p) - k(\phi)],$$

from which the *mean parameter* is derived as

$$E(S) = \nabla M_S(p)|_{p=0} = \nabla k(\phi) = \eta, \text{ say.}$$

[The symbol '∇' denotes a vector of partial derivatives, e.g., $(\partial/\partial p_1, \partial/\partial p_2, \ldots)$, or $(\partial/\partial\phi_1, \partial/\partial\phi_2, \ldots)$, as appropriate.]

Maximum likelihood in a full EF:

Subject to regularity conditions, the unique value of ϕ that maximizes $l(\phi; y)$ in a full EF model solves the system of d simultaneous equations

$$\nabla l(\hat{\phi}) = 0,$$

which reduces to

$$s(y) = \eta(\hat{\phi}).$$

In a full EF model, then, the MLE is also a method-of-moments estimator: the observed values of the sufficient statistics $s_1(y), \ldots, s_d(y)$ are equated with their respective expectations $\eta_1(\phi), \ldots, \eta_d(\phi)$.

Part V

Linear models

Normal-theory linear model

If the conditional distribution of n-vector Y, given (full-rank) $n \times p$ covariate matrix or design matrix x, is $N(x\beta,\ \sigma^2 I_n)$, then the least-squares estimator is $\hat{\beta} = (x^T x)^{-1} x^T Y$, and the log likelihood can be written as a function of $\hat{\beta}$ and the residual sum of squares:

$$l(\beta, \sigma; y) = -n \log \sigma - \frac{\|x\hat{\beta} - x\beta\|^2 + \|y - x\hat{\beta}\|^2}{2\sigma^2}$$

(where $\|v\|^2$, for a vector v, means $v^T v$).

(***Exercise***: show this, and interpret it geometrically in terms of the projection of n-vector y onto the linear subspace spanned by the columns of matrix x.)

Hence the least-squares estimates $\hat{\beta}$ and residual sum of squares $\|y - x\hat{\beta}\|^2$ are jointly minimal sufficient for β and σ.

Generalized linear model

The normal-theory linear model, with σ^2 known, is a full p-dimensional exponential family indexed by β. (***exercise***)

More generally, suppose that $Y_1, \ldots, Y_n$ are independent, each with distribution in a 'natural' [i.e., such that $s(y) = y$] exponential family:

$$f(y_i; \phi_i) = m_i(y_i) \exp[y_i \phi_i - k_i(\phi_i)].$$

Examples include binomial, Poisson and gamma [with α known] distributions for Y_i, as well as the normal [with σ known].

Then if $\phi_1, \ldots, \phi_n$ are assumed to be such that $\phi_i = x_i^T \beta$, with x_i a specified p-vector for each i, the resulting model with parameter vector β is a full p-dimensional EF.

Exercise: show that $\{\sum_i Y_i x_{ir} : r = 1, \ldots, p\}$ are sufficient.

Such a model is a *generalized linear model with canonical link*. Examples include *logistic* regression for binary or binomial Y, and *log-linear* models for Poisson-distributed Y.

In practice, not all generalized linear models have canonical link. A more general dependence of ϕ_i on x_i is $h(\phi_i) = x_i^T \beta$, for some specified function $h(.)$. When h is not the identity, the resulting model with parameters β is usually a *curved* exponential family. Probit and complementary log-log models for binary response, and log-linear models for gamma-distributed response, are examples.

Exercise

Suppose that $Y_1, \ldots, Y_n$ are independent Poisson, with

$$\mu_i = E(Y_i) = t_i \beta$$

for specified positive ('exposure') constants $t_1, \ldots, t_n$.

Show that this is a full 1-dimensional exponential family, and find the sufficient statistic.

(As written above, this is a generalized linear model with non-canonical link. But it can be re-expressed as

$$\log(\mu_i) = \log t_i + \gamma,$$

with $\gamma = \log(\beta)$; this is a log-linear model — i.e., it has the canonical link for the Poisson family — involving the constants $\log t_i$ as a so-called 'offset' term.)

Part VI

Bayesian inference

Bayes' theorem

The formula known as *Bayes' theorem* or *Bayes' rule* comes directly from the definition of conditional probability. If events $A_1, A_2, \ldots$ partition the sample space, and B is any event, then for any i

$$\mathrm{pr}(A_i|B) = \frac{\mathrm{pr}(B|A_i)\,\mathrm{pr}(A_i)}{\sum_j \mathrm{pr}(B|A_j)\,\mathrm{pr}(A_j)}.$$

In *Bayesian inference* the probability model includes unknown parameters as random variables, and the events A_i partition the set of possible parameter values.

Bayes' theorem and Bayesian inference

In Bayesian inference, the likelihood combines with a specified *prior distribution* to produce a *posterior distribution*. This comes simply from treating the parameter as a random variable Θ, and applying Bayes' theorem:

$$f_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)f_\Theta(\theta)}{\int f_{Y|\Theta}(y|\phi)f_\Theta(\phi)d\phi}$$

or, with some notational shortcuts,

$$f(\theta|y) \propto L(\theta; y)f(\theta).$$

The posterior density is then used as the basis for (conditional) probability statements about the random variable Θ.

The data y enter a Bayesian analysis only through the likelihood function $L(\theta; y)$.

Conjugate family of priors

The definition and choice of a prior distribution for a Bayesian analysis may raise challenging conceptual and practical issues. Here we merely note one possible simplification which is available in some situations, and which may sometimes be helpful either for mathematical tractability or for interpretation.

For a given likelihood function $L(\theta; y)$, a family of prior distributions which also contains the posterior, whatever the value of y, is said to be *conjugate* to the likelihood.

Example: For the model $Y|\Theta \sim$ Binomial(m, Θ), any prior distribution $f_\Theta(\theta)$ in the family of *beta distributions* leads to a posterior density $f_{\Theta|y}(\theta|y)$ which is also a beta distribution. (***exercise***: show this)

Conjugate prior for full EF model

If the likelihood takes the full exponential family form

$$L(\phi; y) = m(y)\exp[s^T\phi - k(\phi)],$$

then a prior (for canonical parameter ϕ) proportional to

$$\exp[s_0^T\phi - a_0 k(\phi)]$$

leads to a posterior density that is proportional to

$$\exp[(s + s_0)^T\phi - (1 + a_0)k(\phi)],$$

which is in the same family (indexed by s_0, a_0) as the prior.

Exercise: show how this works for the binomial/beta conjugate likelihood/prior pair mentioned above, and how the beta prior might be interpreted in terms of 'pseudo-data' from a (notional) prior experiment with binomial outcome.