Statistical Theory

Prof. Gesine Reinert

November 23, 2009

Aim: To review and extend the main ideas in Statistical Inference, both from a frequentist viewpoint and from a Bayesian viewpoint. This course serves not only as background to other courses, but also it will provide a basis for developing novel inference methods when faced with a new situation which includes uncertainty. Inference here includes estimating parameters and testing hypotheses.

Overview

- Part 1: Frequentist Statistics

  - Chapter 1: *Likelihood, sufficiency and ancillarity.* The Factorization Theorem. Exponential family models.
  - Chapter 2: *Point estimation.* When is an estimator a good estimator? Covering bias and variance, information, efficiency. Methods of estimation: Maximum likelihood estimation, nuisance parameters and profile likelihood; method of moments estimation. Bias and variance approximations via the delta method.
  - Chapter 3: *Hypothesis testing.* Pure significance tests, significance level. Simple hypotheses, Neyman-Pearson Lemma. Tests for composite hypotheses. Sample size calculation. Uniformly most powerful tests, Wald tests, score tests, generalised likelihood ratio tests. Multiple tests, combining independent tests.
  - Chapter 4: *Interval estimation.* Confidence sets and their connection with hypothesis tests. Approximate confidence intervals. Prediction sets.
  - Chapter 5: *Asymptotic theory.* Consistency. Asymptotic normality of maximum likelihood estimates, score tests. Chi-square approximation for generalised likelihood ratio tests. Likelihood confidence regions. Pseudo-likelihood tests.

- Part 2: Bayesian Statistics

  - Chapter 6: *Background.* Interpretations of probability; the Bayesian paradigm: prior distribution, posterior distribution, predictive distribution, credible intervals. Nuisance parameters are easy.

  - Chapter 7: *Bayesian models.* Sufficiency, exchangeability. De Finetti's Theorem and its intepretation in Bayesian statistics.
  - Chapter 8: *Prior distributions.* Conjugate priors. Noninformative priors; Jeffreys priors, maximum entropy priors posterior summaries. If there is time: Bayesian robustness.
  - Chapter 9: *Posterior distributions.* Interval estimates, asymptotics (very short).

- Part 3: Decision-theoretic approach:
  - Chapter 10: *Bayesian inference as a decision problem.* Decision theoretic framework: point estimation, loss function, decision rules. Bayes estimators, Bayes risk. Bayesian testing, Bayes factor. Lindley's paradox. Least favourable Bayesian answers. Comparison with classical hypothesis testing.
  - Chapter 11: *Hierarchical and empirical Bayes methods.* Hierarchical Bayes, empirical Bayes, James-Stein estimators, Bayesian computation.
- Part 4: *Principles of inference.* The likelihood principle. The conditionality principle. The stopping rule principle.

Books

1. Bernardo, J.M. and Smith, A.F.M. (2000) Bayesian Theory. Wiley.
2. Casella, G. and Berger, R.L. (2002) Statistical Inference. Second Edition. Thomson Learning.
3. Cox, D.R. and Hinkley, D.V. (1974) Theoretical Statistics. Chapman and Hall.
4. Garthwaite, P.H., Joliffe, I.T. and Jones, B. (2002) Statistical Inference. Second Edition. Oxford University Press.
5. Leonard, T. and Hsu, J.S.J. (2001) Bayesian Methods. Cambridge University Press.

6. Lindgren, B.W. (1993) Statistical Theory. 4th edition. Chapman and Hall.

7. O'Hagan, A. (1994) Kendall's Advanced Theory of Statistics. Vol 2B, Bayesian Inference. Edward Arnold.

8. Young, G.A. and Smith, R.L. (2005) Essential of Statistical Inference. Cambridge University Press.

Lecture take place Mondays 11-12 and Wednesdays 9-10. There will be four problem sheets. Examples classes are held Thursdays 12-1 in weeks 3, 4, 6, and 8.

While the examples classes will cover problems from the problem sheets, there may not be enough time to cover all the problems. You will benefit most from the examples classes if you (attempt to) solve the problems on the sheet ahead of the examples classes.

You are invited to hand in your work on the respective problem sheets on Tuesdays at 5 pm in weeks 3, 4, 6, and 8. Your marker is Eleni Frangou; there will be a folder at the departmental pigeon holes.

Additional material may be published at `http://stats.ox.ac.uk/~reinert/stattheory/stattheory.htm`.

The lecture notes may cover more material than the lectures.

Part I

Frequentist Statistics

Chapter 1

Likelihood, sufficiency and ancillarity

We start with *data* $x_1, x_2, \ldots, x_n$, which we would like to use to draw inference about a parameter θ.
Model: We assume that $x_1, x_2, \ldots, x_n$ are realisations of some random variables $X_1, X_2, \ldots, X_n$, from a distribution which depends on the parameter θ.
Often we use the model that $X_1, X_2, \ldots, X_n$ independent, identically distributed (*i.i.d.*) from some $f_X(x, \theta)$ (probability density or probability mass function). We then say $x_1, x_2, \ldots, x_n$ *is a random sample of size* n *from* $f_X(x, \theta)$ (or, shorter, from $f(x, \theta)$).

1.1 Likelihood

If $X_1, X_2, \ldots, X_n$ i.i.d. $\sim f(x, \theta)$, then the joint density at $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$f(\mathbf{x}, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

Inference about θ given the data is based on the **Likelihood function** $L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta)$; often abbreviated by $L(\theta)$. In the i.i.d. model, $L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta)$. Often it is more convenient to use the *log likelihood* $\ell(\theta, \mathbf{x}) = \log L(\theta, \mathbf{x})$ (or, shorter, $\ell(\theta)$). Here and in future, the log denotes the natural logarithm; $e^{\log x} = x$.

Example: Normal distribution. Assume that $x_1, \ldots, x_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown parameters, $\mu \in \mathbf{R}, \sigma^2 > 0$. With $\theta = (\mu, \sigma^2)$, the likelihood is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n (2\pi\sigma^2)^{-1/2} \exp\left\{-\tfrac{1}{2\sigma^2}(x_i-\mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right\} \end{aligned}$$

and the log-likelihood is

$$\ell(\theta) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

Example: Poisson distribution. Assume that $x_1, \ldots, x_n$ is a random sample from $Poisson(\theta)$, with unknown $\theta > 0$; then the likelihood is

$$L(\theta) = \prod_{i=1}^n \left(e^{-\theta}\frac{\theta^{x_i}}{x_i!}\right) = e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n (x_i!)^{-1}$$

and the log-likelihood is

$$\ell(\theta) = -n\theta + \log(\theta)\sum_{i=1}^n x_i - \sum_{i=1}^n \log(x_i!).$$

1.2 Sufficiency

Any function of $\mathbf{X}$ is a *statistic*. We often write $T = t(\mathbf{X})$, where t is a function. Some examples are the sample mean, the sample median, and the actual data. Usually we would think of a statistic as being some summary of the data, so smaller in dimension than the original data.
A statistic is *sufficient* for the parameter θ if it contains all information about θ that is available from the data: $\mathcal{L}(\mathbf{X}|T)$, the conditional distribution of $\mathbf{X}$ given T, does not depend on θ.

Factorisation Theorem (Casella + Berger, p.250) A statistic $T = t(\mathbf{X})$ is sufficient for θ if and only if there exists functions $g(t, \theta)$ and $h(\mathbf{x})$ such that for all $\mathbf{x}$ and θ

$$f(\mathbf{x}, \theta) = g(t(\mathbf{x}), \theta)h(\mathbf{x}).$$

Example: Bernoulli distribution. Assume that $X_1, \ldots, X_n$ are i.i.d. Bernoulli trials, $Be(\theta)$, so $f(x, \theta) = \theta^x (1-\theta)^{1-x}$; let $T = \sum_{i=1}^n X_i$ denote number of successes. Recall that $T \sim Bin(n, \theta)$, and so

$$P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \quad t = 0, 1, \ldots, n.$$

Then

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = 0 \text{ for } \sum_{i=1}^n x_i \neq t,$$

and for $\sum_{i=1}^n x_i = t$,

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} \\ &= \frac{\prod_{i=1}^n \left(\theta^{x_i}(1-\theta)^{(1-x_i)}\right)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}. \end{aligned}$$

This expression is independent of θ, so T is sufficient for θ.

Alternatively, the Factorisation Theorem gives

$$\begin{aligned} f(\mathbf{x}, \theta) &= \prod_{i=1}^n \left(\theta^{x_i}(1-\theta)^{(1-x_i)}\right) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= g(t(\mathbf{x}), \theta) h(\mathbf{x}) \end{aligned}$$

with $t = \sum_{i=1}^n x_i$; $g(t, \theta) = \theta^t(1-\theta)^{n-t}$ and $h(\mathbf{x}) = 1$, so $T = t(\mathbf{X})$ is sufficient for θ.

Example: Normal distribution. Assume that $X_1, \ldots, X_n$ are i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$; put $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$, then

$$\begin{aligned} f(\mathbf{x}, \theta) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\} \\ &= \exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\} (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}. \end{aligned}$$

If σ^2 is known: $\theta = \mu$, $t(\mathbf{x}) = \overline{x}$, and $g(t, \mu) = \exp\left\{-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right\}$, so $\overline{X}$ is sufficient.
If σ^2 is unknown: $\theta = (\mu, \sigma^2)$, and $f(\mathbf{x}, \theta) = g(\overline{x}, s^2, \theta)$, so $(\overline{X}, S^2)$ is sufficient.

Example: Poisson distribution. Assume that $x_1, \ldots, x_n$ are a random sample from $Poisson(\theta)$, with unknown $\theta > 0$. Then

$$L(\theta) = e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n (x_i!)^{-1}.$$

and (exercise)

$$\begin{aligned} t(\mathbf{x}) &= \\ g(t, \theta) &= \\ h(\mathbf{x}) &= \end{aligned}$$

Example: order statistics. Let $X_1, \ldots, X_n$ be i.i.d.; the order statistics are the ordered observations $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Then $T = (X_{(1)}, X_{(2)}, \cdots, X_{(n)})$ is sufficient.

1.2.1 Exponential families

Any probability density function $f(x|\theta)$ which is written in the form

$$f(\mathbf{x}, \theta) = \exp\left\{\sum_{i=1}^k c_i\phi_i(\theta)h_i(\mathbf{x}) + c(\theta) + d(\mathbf{x}),\right\}, \quad \mathbf{x} \in \mathcal{X},$$

where $c(\theta)$ is chosen such that $\int f(x, \theta)\, dx = 1$, is said to be in the *k-parameter exponential family.* The family is called *regular* if $\mathcal{X}$ does not depend on θ; otherwise it is called *non-regular.*

Examples include the binomial distribution, Poisson distributions, normal distributions, gamma (including exponential) distributions, and many more.

Example: Binomial (n, θ). For $x = 0, 1, ..., n$,

$$\begin{aligned} f(x; \theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \exp\left\{x\log\left(\frac{\theta}{1-\theta}\right) + \log\left(\binom{n}{x}\right) + n\log(1-\theta)\right\}. \end{aligned}$$

Choose $k = 1$ and

$$\begin{aligned} c_1 &= 1 \\ \phi_1(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \\ h_1(x) &= x \\ c(\theta) &= n\log(1-\theta) \\ d(x) &= \log\left(\binom{n}{x}\right) \\ \mathcal{X} &= \{0,\ldots,n\}. \end{aligned}$$

Fact: In k-parameter exponential family models,

$$t(\mathbf{x}) = (n, \sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(x_j))$$

is sufficient.

1.2.2 Minimal sufficiency

A statistic T which is sufficient for θ is *minimal sufficient* for θ is if can be expressed as a function of any other sufficient statistic. To find a minimal sufficient statistic: Suppose $\frac{f(\mathbf{x},\theta)}{f(\mathbf{y},\theta)}$ is constant in θ if and only if

$$t(\mathbf{x}) = t(\mathbf{y}),$$

then $T = t(\mathbf{X})$ is minimal sufficient (see Casella + Berger p.255).

In order to avoid issues when the density could be zero, it is the case that if for any possible values for $\mathbf{x}$ and $\mathbf{y}$, we have that the equation

$$f(\mathbf{x},\theta) = \phi(\mathbf{x},\mathbf{y})f(\mathbf{y},\theta) \text{ for all } \theta$$

implies that $t(\mathbf{x}) = t(\mathbf{y})$, where ϕ is a function which does not depend on θ, then $T = t(\mathbf{X})$ is minimal sufficent for θ.

Example: Poisson distribution. Suppose that $X_1, \ldots, X_n$ are i.i.d. $Po(\theta)$, then $f(\mathbf{x},\theta) = e^{-n\theta}\theta^{\sum_{i=1}^n x_i}\prod_{i=1}^n (x_i!)^{-1}$ and

$$\frac{f(\mathbf{x},\theta)}{f(\mathbf{y},\theta)} = \theta^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}\prod_{i=1}^n \frac{y_i!}{x_i!},$$

which is constant in θ if and only if

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i;$$

so $T = \sum_{i=1}^{n} X_i$ is minimal sufficient (as is $\overline{X}$). Note: $T = \sum_{i=1}^{n} X_{(i)}$ is a function of the order statistic.

1.3 Ancillary statistic

If $a(\mathbf{X})$ is a statistics whose distribution does not depend on θ, it is called an *ancillary* statistic.

Example: Normal distribution. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$. Then $T = X_2 - X_1 \sim \mathcal{N}(0, 2)$ has a distribution which does not depend on θ; it is ancillary.

When a minimal sufficient statistic T is of larger dimension than θ, then there will often be a component of T whose distribution is independent of θ.

Example: some uniform distribution (*Exercise*). Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{U}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$ then $\left(X_{(1)}, X_{(n)}\right)$ is minimal sufficient for θ, as is

$$(S, A) = \left(\frac{1}{2}(X_{(1)} + X_{(n)}), X_{(n)} - X_{(1)}\right),$$

and the distribution of A is independent of θ, so A is an ancillary statistic. Indeed, A measures the accuracy of S; for example, if $A = 1$ then $S = \theta$ with certainty.

Chapter 2

Point Estimation

Recall that we assume our data $x_1, x_2, \ldots, x_n$ to be realisations of random variables $X_1, X_2, \ldots, X_n$ from $f(\mathbf{x}, \theta)$. Denote the expectation with respect to $f(\mathbf{x}, \theta)$ by E_θ, and the variance by Var_θ.

When we estimate θ by a function $t(x_1, \ldots, x_n)$ of the data, this is called a *point estimate*; $T = t(X_1, \ldots, X_n) = t(\mathbf{X})$ is called an *estimator* (random). For example, the sample mean is an estimator of the mean.

2.1 Properties of estimators

T is *unbiased* for θ if $E_\theta(T) = \theta$ for all θ; otherwise T is *biased*. The *bias* of T is

$$Bias(T) = Bias_\theta(T) = E_\theta(T) - \theta.$$

Example: Sample mean, sample variance. Suppose that $X_1, \ldots, X_n$ are i.i.d. with unknown mean μ; unknown variance σ^2. Consider the estimate of μ given by the sample mean

$$T = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then

$$E_{\mu,\sigma^2}(T) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu,$$

so the sample mean is unbiased. Recall that

$$Var_{\mu,\sigma^2}(T) = Var_{\mu,\sigma^2}(\overline{X}) = E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2)\} = \frac{\sigma^2}{n}.$$

If we estimate σ^2 by

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

then

$$\begin{aligned}
& E_{\mu,\sigma^2}(S^2) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E_{\mu,\sigma^2}\{(X_i-\mu+\mu-\overline{X})^2\} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\big\{E_{\mu,\sigma^2}\{(X_i-\mu)^2\} + 2E_{\mu,\sigma^2}(X_i-\mu)(\mu-\overline{X}) \\
&\quad + E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\}\big\} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\sigma^2 - 2\frac{n}{n-1}E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\} + \frac{n}{n-1}E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\} \\
&= \sigma^2\left(\frac{n}{n-1} - \frac{2}{n-1} + \frac{1}{n-1}\right) = \sigma^2,
\end{aligned}$$

so S^2 is an unbiased estimator of σ^2. *Note*: the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is **not** unbiased.

Another criterion for estimation is a small *mean square error* (MSE); the MSE of an estimator T is defined as

$$MSE(T) = MSE_\theta(T) = E_\theta\{(T-\theta)^2\} = Var_\theta(T) + (Bias_\theta(T))^2.$$

Note: $MSE(T)$ is a function of θ.

Example: $\hat{\sigma}^2$ has smaller MSE than S^2 (see *Casella and Berger, p.304*) but is biased.

If one has two estimators at hand, one being slightly biased but having a smaller MSE than the second one, which is, say, unbiased, then one may

well prefer the slightly biased estimator. Exception: If the estimate is to be combined linearly with other estimates from independent data.

The efficiency of an estimator T is defined as

$$Efficiency_\theta(T) = \frac{\mathrm{Var}_\theta(T_0)}{\mathrm{Var}_\theta(T)},$$

where T_0 has minimum possible variance.

Theorem: Cramér-Rao Inequality, Cramér-Rao lower bound: Under regularity conditions on $f(\mathbf{x}, \theta)$, it holds that for any unbiased T,

$$\mathrm{Var}_\theta(T) \geq (I_n(\theta))^{-1},$$

where

$$I_n(\theta) = E_\theta \left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta} \right)^2 \right]$$

is the *expected Fisher information* of the sample.

Thus, for any unbiased estimator T,

$$Efficiency_\theta(T) = \frac{1}{I_n(\theta)\mathrm{Var}_\theta(T)}.$$

Assume that T is unbiased. T is called *efficient* (or a *minimum variance unbiased estimator*) if it has the minimum possible variance. An unbiased estimator T is efficient if $\mathrm{Var}_\theta(T) = (I_n(\theta))^{-1}$.

Often $T = T(X_1, \ldots, X_n)$ is efficient as $n \to \infty$: then it is called *asymptotically efficient*.

The *regularity* conditions are conditions on the partial derivatives of $f(\mathbf{x}, \theta)$ with respect to θ; and the domain may not depend on θ; for example $\mathcal{U}[0, \theta]$ violates the regularity conditions.

Under more regularity: the first three partial derivatives of $f(\mathbf{x}, \theta)$ with respect to θ are integrable with respect to x; and again the domain may not depend on θ; then

$$I_n(\theta) = E_\theta \left[-\frac{\partial^2 \ell(\theta, \mathbf{X})}{\partial \theta^2} \right].$$

Notation: We shall often omit the subscript in $I_n(\theta)$, when it is clear whether we refer to a sample of size 1 or of size n. For a random sample,

$$I_n(\theta) = nI_1(\theta).$$

Calculation of the expected Fisher information:

$$\begin{aligned} I_n(\theta) &= E_\theta\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right] \\ &= \int f(\mathbf{x}, \theta)\left[\left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] d\mathbf{x} \\ &= \int f(\mathbf{x}, \theta)\left[\frac{1}{f(\mathbf{x}, \theta)}\left(\frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}\right)\right]^2 d\mathbf{x} \\ &= \int \frac{1}{f(\mathbf{x}, \theta)}\left[\left(\frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] d\mathbf{x}. \end{aligned}$$

Example: Normal distribution, known variance For a random sample $\mathbf{X}$ from $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known, and $\theta = \mu$,

$$\begin{aligned} \ell(\theta) &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2; \\ \frac{\partial \ell}{\partial \theta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\overline{x} - \mu) \end{aligned}$$

and

$$\begin{aligned} I_n(\theta) &= E_\theta\left[\left(\frac{\partial \ell(\theta, \mathbf{X}}{\partial \theta}\right)^2\right] \\ &= \frac{n^2}{\sigma^4}E_\theta(\overline{X} - \mu)^2 = \frac{n}{\sigma^2}. \end{aligned}$$

Note: $\mathrm{Var}_\theta(\overline{X}) = \frac{\sigma^2}{n}$, so $\overline{X}$ is an efficient estimator for μ. Also note that

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{n}{\sigma^2};$$

a quicker method yielding the expected Fisher information.

In future we shall often omit the subscript θ in the expectation and in the variance.

Example: Exponential family models in canonical form
Recall that one-parameter (i.e., scalar θ) exponential family density has the form

$$f(x;\theta) = exp\left\{\phi(\theta)h(x) + c(\theta) + d(x)\right\}, \quad x \in \mathcal{X}.$$

Choosing θ and x to make $\phi(\theta) = \theta$ and $h(x) = x$ is called the *canonical form*;

$$f(x;\theta) = \exp\{\theta x + c(\theta) + d(x)\}.$$

For the canonical form

$$EX = \mu(\theta) = -c'(\theta), \text{ and} \qquad \operatorname{Var} X = \sigma^2(\theta) = -c''(\theta).$$

Exercise: Prove the mean and variance results by calculating the moment-generating function $E\exp(tX) = \exp\{c(\theta) - c(t+\theta)\}$. Recall that you obtain mean and variance by differentiating the moment-generating function (how exactly?)

Example: Binomial (n, p). Above we derived the exponential family form with

$$\begin{aligned} c_1 &= 1 \\ \phi_1(p) &= \log\left(\frac{p}{1-p}\right) \\ h_1(x) &= x \\ c(p) &= n\log(1-p) \\ d(x) &= \log\left(\binom{n}{x}\right) \\ \mathcal{X} &= \{0,\ldots,n\}. \end{aligned}$$

To write the density in canonical form we put

$$\theta = \log\left(\frac{p}{1-p}\right)$$

(this transformation is called the *logit* transformation); then

$$p = \frac{e^\theta}{1 + e^\theta}$$

and

$$\begin{aligned} \phi(\theta) &= \theta \\ h(x) &= x \\ c(\theta) &= -n \log\left(1 + e^\theta\right) \\ d(x) &= \log\left(\binom{n}{x}\right) \\ \mathcal{X} &= \{0, \ldots, n\} \end{aligned}$$

gives the canonical form. We calculate the mean

$$-c'(\theta) = n\frac{e^\theta}{1 + e^\theta} = \mu(\theta) = np$$

and the variance

$$\begin{aligned} -c''(\theta) &= n\left\{\frac{e^\theta}{1 + e^\theta} - \frac{e^{2\theta}}{(1 + e^\theta)^2}\right\} \\ &= \sigma^2(\theta) = np(1 - p). \end{aligned}$$

Now suppose $X_1, \ldots, X_n$ are i.i.d., from the canonical density. Then

$$\begin{aligned} \ell(\theta) &= \theta \sum x_i + nc(\theta) + \sum d(x_i), \\ \ell'(\theta) &= \sum x_i + nc'(\theta) = n(\overline{x} + c'(\theta)). \end{aligned}$$

Since $\ell''(\theta) = nc''(\theta)$, we have that $I_n(\theta) = E(-\ell''(\theta)) = -nc''(\theta)$.

Example: Binomial (n, p) and

$$\theta = \log\left(\frac{p}{1 - p}\right)$$

then (exercise)

$$I_1(\theta) =$$

2.2 Maximum Likelihood Estimation

Now θ may be a vector. A *maximum likelihood estimate*, denoted $\hat{\theta}(\mathbf{x})$, is a value of θ at which the likelihood $L(\theta, \mathbf{x})$ is maximal. The estimator $\hat{\theta}(\mathbf{X})$ is called *MLE* (also, $\hat{\theta}(\mathbf{x})$ is sometimes called mle). An mle is a parameter value at which the observed sample is most likely.

Often it is easier to maximise log likelihood: **if** derivatives exist, then set first (partial) derivative(s) with respect to θ to zero, check that second (partial) derivative(s) with respect to θ less than zero.

An mle is a function of a sufficient statistic:

$$L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta) = g(t(\mathbf{x}), \theta)h(\mathbf{x})$$

by the factorisation theorem, and maximizing in θ depends on $\mathbf{x}$ only through $t(\mathbf{x})$.

An mle is usually efficient as $n \to \infty$.

Invariance property: An mle of a function $\phi(\theta)$ is $\phi(\hat{\theta})$ (Casella + Berger p.294). That is, if we define the likelihood induced by ϕ as

$$L^*(\lambda, x) = \sup_{\theta:\phi(\theta)=\lambda} L(\theta, x),$$

then one can calculate that for $\hat{\lambda} = \phi(\hat{\theta})$,

$$L^*(\hat{\lambda}, x) = L(\hat{\theta}, x).$$

Examples: Uniforms, normal

1. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[0, \theta]$:

   $$L(\theta) = \theta^{-n}\mathbf{1}_{[x_{(n)},\infty)}(\theta),$$

   where $x_{(n)} = \max_{1 \le i \le n} x_i$; so $\hat{\theta} = X_{(n)}$

2. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, then any $\theta \in [x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}]$ maximises the likelihood (*Exercise*)

3. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, then (*Exercise*) $\hat{\mu} = \overline{X}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$. So $\hat{\sigma}^2$ is biased, but $Bias(\hat{\sigma}^2) \to 0$ as $n \to \infty$.

Iterative computation of MLEs

Sometimes the likelihood equations are difficult to solve. Suppose $\hat{\theta}^{(1)}$ is an initial approximation for $\hat{\theta}$. Using Taylor expansion gives

$$0 = \ell'(\hat{\theta}) \approx \ell'(\hat{\theta}^{(1)}) + (\hat{\theta} - \hat{\theta}^{(1)})\ell''(\hat{\theta}^{(1)}),$$

so that

$$\hat{\theta} \approx \hat{\theta}^{(1)} - \frac{\ell'(\hat{\theta}^{(1)})}{\ell''(\hat{\theta}^{(1)})}.$$

Iterate this procedure (this is called the *Newton-Raphson method*) to get

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - (\ell''(\hat{\theta}^{(k)}))^{-1}\ell'(\hat{\theta}^{(k)}), \quad k = 2, 3, \ldots;$$

continue the iteration until $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}| < \epsilon$ for some small ϵ.

As $E\left\{-\ell''(\hat{\theta}^{(1)})\right\} = I_n(\hat{\theta}^{(1)})$ we could instead iterate

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + I_n^{-1}(\hat{\theta}^{(k)})\ell'(\hat{\theta}^{(k)}), \quad k = 2, 3, \ldots$$

until $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}| < \epsilon$ for some small ϵ. This is *Fisher's modification of the Newton-Raphson method*.

Repeat with different starting values to reduce the risk of finding just a local maximum.

Example: Suppose that we observe x from the distribution Binomial(n, θ). Then

$$\begin{aligned}
\ell(\theta) &= x\ln(\theta) + (n-x)\ln(1-\theta) + \log\binom{n}{x} \\
\ell'(\theta) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x - n\theta}{\theta(1-\theta)} \\
\ell''(\theta) &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \\
I_1(\theta) &= \frac{n}{\theta(1-\theta)}
\end{aligned}$$

Assume that $n = 5, x = 2, \epsilon = 0.01$ (in practice rather $\epsilon = 10^{-5}$); and start with an initial guess $\hat{\theta}^{(1)} = 0.55$. Then *Newton-Raphson* gives

$$\begin{aligned} \ell'(\hat{\theta}^{(1)}) &\approx -3.03 \\ \hat{\theta}^{(2)} &\approx \hat{\theta}^{(1)} - (\ell''(\hat{\theta}^{(1)}))^{-1}\ell'(\hat{\theta}^{(1)}) \approx 0.40857 \\ \ell'(\hat{\theta}^{(2)}) &\approx -0.1774 \\ \hat{\theta}^{(3)} &\approx \hat{\theta}^{(2)} - (\ell''(\hat{\theta}^{(2)}))^{-1}\ell(\hat{\theta}^{(2)}) \approx 0.39994. \end{aligned}$$

Now $|\hat{\theta}^{(3)} - \hat{\theta}^{(2)}| < 0.01$ so we stop.

Using instead *Fisher scoring* gives

$$I_1^{-1}(\theta)\ell'(\theta) = \frac{x - n\theta}{n} = \frac{x}{n} - \theta$$

and so

$$\theta + I_1^{-1}(\theta)\ell'(\theta) = \frac{x}{n}$$

for all θ. To compare: analytically, $\hat{\theta} = \frac{x}{n} = 0.4$.

2.3 Profile likelihood

Often $\theta = (\psi, \lambda)$, where ψ contains the parameters of interest and λ contains the other unknown parameters: *nuisance parameters*. Let $\hat{\lambda}_\psi$ be the MLE for λ for a given value of ψ. Then the *profile likelihood* for ψ is

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi)$$

(in $L(\psi, \lambda)$ replace λ by $\hat{\lambda}_\psi$); the *profile log-likelihood* is $\ell_P(\psi) = \log[L_P(\psi)]$.

For point estimation, maximizing $L_P(\psi)$ with respect to ψ gives the same estimator $\hat{\psi}$ as maximizing $L(\psi, \lambda)$ with respect to both ψ and λ (but possibly different variances)

Example: Normal distribution. Suppose that $X_1, \ldots, X_n$ are i.i.d. from $\mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown. Given μ, $\hat{\sigma}^2_\mu = (1/n)\sum(x_i - \mu)^2$, and given σ^2, $\hat{\mu}_{\sigma^2} = \overline{x}$. Hence the profile likelihood for μ is

$$\begin{aligned} L_P(\mu) &= (2\pi\hat{\sigma}^2_\mu)^{-n/2} \exp\left\{-\frac{1}{2\hat{\sigma}^2_\mu}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= \left[\frac{2\pi e}{n}\sum(x_i - \mu)^2\right]^{-n/2}, \end{aligned}$$

which gives $\hat{\mu} = \overline{x}$; and the profile likelihood for σ^2 is

$$L_P(\sigma^2) \quad = \quad (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - \overline{x})^2\right\},$$

gives (Exercise)

$$\hat{\sigma}^2_{\mu} =$$

2.4 Method of Moments (M.O.M)

The idea is to match population moments to sample moments in order to obtain estimators. Suppose that $X_1, \ldots, X_n$ are i.i.d. $\sim f(x; \theta_1, \ldots, \theta_p)$. Denote by

$$\mu_k = \mu_k(\theta) = E(X^k)$$

the k^{th} moment and by

$$M_k = \frac{1}{n}\sum(X_i)^k$$

the k^{th} sample moment. In general, $\mu_k = \mu_k(\theta_1, \ldots, \theta_p)$. Solve the equation

$$\mu_k(\theta) = M_k$$

for $k = 1, 2, \ldots,$ until there are sufficient equations to solve for $\theta_1, \ldots, \theta_p$ (usually p equations for the p unknowns). The solutions $\tilde{\theta}_1, \ldots, \tilde{\theta}_p$ are the *method of moments estimators*.

They are often not as efficient as MLEs, but may be easier to calculate. They could be used as initial estimates in an iterative calculation of MLEs.

Example: Normal distribution. Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$; μ and σ^2 are unknown. Then

$$\mu_1 = \mu;\ M_1 = \overline{X}$$

"Solve"

$$\mu = \overline{X}$$

so

$$\tilde{\mu} = \overline{X}.$$

Furthermore

$$\mu_2 = \sigma^2 + \mu^2;\ M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

so solve

$$\sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

which gives

$$\tilde{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(which is not unbiased for σ^2).

Example: Gamma distribution. Suppose that $X_1, \ldots, X_n$ are i.i.d. $\Gamma(\psi, \lambda)$;

$$f(x; \psi, \lambda) = \frac{1}{\Gamma(\psi)}\lambda^{\psi} x^{\psi-1} e^{-\lambda x} \qquad \text{for } x \geq 0.$$

Then $\mu_1 = EX = \psi/\lambda$ and

$$\mu_2 = EX^2 = \psi/\lambda^2 + (\psi/\lambda)^2.$$

Solving

$$M_1 = \psi/\lambda, \qquad M_2 = \psi/\lambda^2 + (\psi/\lambda)^2$$

for ψ and λ gives

$$\tilde{\psi} \;=\; \overline{X}^2 / [n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2], \text{ and } \tilde{\lambda} = \overline{X} / [n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2].$$

2.5 Bias and variance approximations: the delta method

Sometimes T is a function of one or more averages whose means and variances can be calculated exactly; then we may be able to use the following simple approximations for mean and variance of T:

Suppose $T = g(S)$ where $ES = \beta$ and $\operatorname{Var} S = V$. Taylor expansion gives

$$T = g(S) \approx g(\beta) + (S - \beta)g'(\beta).$$

Taking the mean and variance of the r.h.s.:

$$ET \approx g(\beta), \qquad \operatorname{Var} T \approx [g'(\beta)]^2 V.$$

If S is an average so that the central limit theorem (CLT) applies to it, i.e., $S \approx N(\beta, V)$, then

$$T \approx N(g(\beta), [g'(\beta)]^2 V)$$

for large n.

If $V = v(\beta)$, then it is possible to choose g so that T has approximately constant variance in θ: solve $[g'(\beta)]^2 v(\beta) = \text{constant}$.

Example: Exponential distribution. $X_1, \ldots, X_n$ i.i.d. $\sim exp(\frac{1}{\mu})$, mean μ. Then $S = \overline{X}$ has mean μ and variance μ^2/n. If $T = \log \overline{X}$ then $g(\mu) = \log(\mu)$, $g'(\mu) = \mu^{-1}$, and so $\operatorname{Var} T \approx n^{-1}$, which is independent of μ: this is called a *variance stabilization*.

If the Taylor expansion is carried to the second-derivative term, we obtain

$$ET \approx g(\beta) + \frac{1}{2} V g''(\beta).$$

In practice we use numerical estimates for β and V if unknown.

When S, β are vectors (V a matrix), with T still a scalar: Let $\big(g'(\beta)\big)_i = \partial g / \partial \beta_i$ and let $g''(\beta)$ be the matrix of second derivatives, then Taylor expansion gives

$$\operatorname{Var} T \approx [g'(\beta)]^T V g'(\beta)$$

and

$$ET \approx g(\beta) + \frac{1}{2} \operatorname{trace}[g''(\beta) V].$$

2.5.1 Exponential family models in canonical form and asymptotic normality of the MLE

Recall that a one-parameter (i.e., scalar θ) exponential family density in canonical form can be written as

$$f(x; \theta) = \exp\{\theta x + c(\theta) + d(x)\},$$

and $EX = \mu(\theta) = -c'(\theta)$, as well as $\operatorname{Var} X = \sigma^2(\theta) = -c''(\theta)$. Suppose $X_1, \ldots, X_n$ are i.i.d., from the canonical density. Then

$$\ell'(\theta) \;\; = \;\; \sum x_i + nc'(\theta) = n(\overline{x} + c'(\theta)).$$

Since $\mu(\theta) = -c'(\theta)$,

$$\ell'(\theta) = 0 \iff \overline{x} = \mu(\hat{\theta}),$$

and we have already calculated that $I_n(\theta) = E(-\ell''(\theta)) = -nc''(\theta)$. If μ is invertible, then

$$\hat{\theta} = \mu^{-1}(\overline{x}).$$

The CLT applies to $\overline{X}$ so, for large n,

$$\overline{X} \approx \mathcal{N}(\mu(\theta), -c''(\theta)/n).$$

With the delta-method, $S \approx \mathcal{N}(a, b)$ implies that

$$g(S) \approx \mathcal{N}\left(g(a), b[g'(a)]^2\right)$$

for continuous g, and small b. For $S = \overline{X}$, with $g(\cdot) = \mu^{-1}(\cdot)$ we have $g'(\cdot) = (\mu'(\mu^{-1}(\cdot))^{-1}$, thus

$$\hat{\theta} \approx \mathcal{N}\big(\theta, I_n^{-1}(\theta)\big) :$$

the M.L.E. is asymptotically normal.

Note: The approximate variance equals the Cramér-Rao lower bound: *quite generally the MLE is asymptotically efficient.*

Example: Binomial(n, p). With $\theta = \log\left(\frac{p}{1-p}\right)$ we have $\mu(\theta) = n\frac{e^\theta}{1+e^\theta}$, and we calculate

$$\mu^{-1}(t) = \log\left(\frac{\frac{t}{n}}{1 - \frac{t}{n}}\right).$$

Note that here we have a sample, x, of size 1. This gives

$$\hat{\theta} = \log\left(\frac{\frac{x}{n}}{1 - \frac{x}{n}}\right),$$

as expected from the invariance of mle's. We hence know that $\hat{\theta}$ is approximately normally distributed.

2.6 Excursions

2.6.1 Minimum Variance Unbiased Estimation

There is a pretty theory about how to construct minimum variance unbiased estimators (MVUE) based on sufficient statistics. The key underlying result is the *Rao-Blackwell Theorem* (Casella+Berger p.316). We do not have time to go into detail during lectures, but you may like to read up on it.

2.6.2 A more general method of moments

Consider statistics of the form $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$. Find the expected value as a function of θ

$$\frac{1}{n}\sum_{i=1}^{n} Eh(X_i) = r(\theta).$$

Now obtain an estimate for θ by solving $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ for θ.

2.6.3 The delta method and logistic regresstion

For logistic regression, the outcome of an experiment is 0 or 1, and the outcome may depend on some explanatory variables. We are interested in

$$P(Y_i = 1|x) = \pi(x|\beta).$$

The outcome for each experiment is in $[0, 1]$; in order to apply some normal regression model we use the logit transform,

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

which is now spread over the whole real line. The ratio $\frac{p}{1-p}$ is also called the *odds*. A (Generalised linear) model then relates the logit to the regressors in a linear fashion;

$$logit(\pi(x|\beta)) \quad = \quad \log\left(\frac{\pi(x|\beta)}{1-\pi(x|\beta)}\right) = x^T\beta.$$

The coefficients β describe how the odds for π change with change in the explanatory variables. The model can now be treated like an ordinary linear

regression, X is the design matrix, β is the vector of coefficients. Transforming back,

$$P(Y_i = 1|x) \;=\; \exp(x^T\beta) \Big/ \left(1 + \exp(x^T\beta)\right).$$

The invariance property gives that the MLE of $\pi(x|\beta)$, for any x, is $\pi(x|\hat{\beta})$, where $\hat{\beta}$ is the MLE obtained in the ordinary linear regression from a sample of responses $y_1, \ldots, y_n$ with associated covariate vectors $x_1, \ldots, x_n$. We know that $\hat{\beta}$ is approximately normally distributed, and we would like to infer asymptotic normality of $\pi(x|\hat{\beta})$.

(i) If β is scalar: Calculate that

$$\begin{aligned}
\frac{\partial}{\partial\beta}\pi(x_i|\beta) &= \frac{\partial}{\partial\beta}\exp(x_i\beta) \Big/ \left(1 + \exp(x_i\beta)\right) \\
&= x_i e^{x_i\beta}\left(1 + \exp(x_i\beta)\right)^{-1} - \left(1 + \exp(x_i\beta)\right)^{-2} x_i e^{x_i\beta} e^{x_i\beta} \\
&= x_i\pi(x_i|\beta) - x_i(\pi(x_i|\beta))^2 \\
&= x_i\pi(x_i|\beta)(1 - \pi(x_i|\beta))
\end{aligned}$$

and the likelihood is

$$L(\beta) \;=\; \prod_{i=1}^{n}\pi(x_i|\beta) = \prod_{i=1}^{n}\exp(x_i\beta) \Big/ \left(1 + \exp(x_i\beta)\right).$$

Hence the log likelihood has derivative

$$\begin{aligned}
\ell'(\beta) &= \sum_{i=1}^{n}\frac{1}{\pi(x_i|\beta)}x_i\pi(x_i|\beta)(1 - \pi(x_i|\beta)) \\
&= \sum_{i=1}^{n} x_i(1 - \pi(x_i|\beta))
\end{aligned}$$

so that

$$\ell''(\beta) \;=\; -\sum_{i=1}^{n} x_i^2\pi(x_i|\beta))(1 - \pi(x_i|\beta)).$$

Thus $\hat{\beta} \approx \mathcal{N}(\beta, I_n^{-1}(\beta))$ where $I_n(\beta) = \sum x_i^2\pi_i(1 - \pi_i)$ with $\pi_i = \pi(x_i|\beta)$.

So now we know the parameters of the normal distribution which approximates the distribution of $\hat{\beta}$. The delta method with $g(\beta) = e^{\beta x}/(1+e^{\beta x})$, gives

$$g'(\beta) \;\; = \;\; xg(\beta)(1-g(\beta))$$

and hence we conclude that $\pi = \pi(x|\hat{\beta}) \approx \mathcal{N}\big(\pi, \pi^2(1-\pi)^2x^2I^{-1}(\beta)\big)$.

(ii) If β is vector: Similarly it is possible to calculate that $\hat{\beta} \approx \mathcal{N}\big(\beta, I_n^{-1}(\beta)\big)$ where $[I_n(\beta)]_{kl} = E\left(-\partial^2\ell/\partial\beta_k\partial\beta_l\right)$. The vector version of the delta method then gives

$$\pi(x|\hat{\beta}) \approx \mathcal{N}\left(\pi, \pi^2(1-\pi)^2x^TI^{-1}(\beta)x\right)$$

with $\pi = \pi(x|\beta)$ and $I_n(\beta) = X^TRX$. Here X is the design matrix, and

$$R = Diag\left(\pi_i(1-\pi_i),\ i = 1, \ldots, n\right)$$

where $\pi_i = \pi(x_i|\beta)$. Note that this normal approximation is likely to be poor for π near zero or one.

Chapter 3

Hypothesis Testing

3.1 Pure significance tests

We have data $\mathbf{x} = (x_1, \ldots, x_n)$ from $f(\mathbf{x}, \theta)$, and a hypothesis H_0 which restricts $f(\mathbf{x}, \theta)$. We would like to know: Are the data consistent with H_0?

H_0 is called the *null hypothesis*. It is called *simple* if it completely specifies the density of $\mathbf{x}$; it is called *composite* otherwise.

A *pure significance test* is a means of examining whether the data are consistent with H_0 where the only distribution of the data that is explicitly formulated is that under H_0. Suppose that for a test statistic $T = t(\mathbf{X})$, the larger $t(\mathbf{x})$, the more inconsistent the data with H_0. For simple H_0, the *p-value* of $\mathbf{x}$ is then

$$p = P(T \geq t(\mathbf{x})|H_0).$$

Small p indicate more inconsistency with H_0.

For composite H_0: If S is sufficient for θ then the distribution of $\mathbf{X}$ conditional on S is independent of θ; when $S = s$, the *p-value* of $\mathbf{x}$ is

$$p = P(T \geq t(\mathbf{x})|H_0; S = s).$$

Example: Dispersion of Poisson distribution. Let H_0: $X_1, \ldots, X_n$ i.i.d. $\sim Poisson(\mu)$, with unknown μ. Under H_0, $\mathrm{Var}(X_i) = \mathbf{E}(X_i) = \mu$ and so we would expect $T = t(\mathbf{X}) = S^2/\overline{X}$ to be close to 1. The statistic T is also called the *dispersion index*.

We suspect that the X_i's may be over-dispersed, that is, $\text{Var}(X_i) > EX_i$: discrepancy with H_0 would then correspond to large T. Recall that $\overline{X}$ is sufficient for the Poisson distribution; the p-value under the Poisson hypothesis is then $p = P(S^2/\overline{X} \geq t(\mathbf{x})|\overline{X} = \overline{x}; H_0)$, which makes p independent of the unknown μ. Given $\overline{X} = \overline{x}$ and H_0 we have that

$$S^2/\overline{X} \approx \chi^2_{n-1}/(n-1)$$

(see Chapter 5 later) and so the p-value of the test satisfies

$$p \approx P(\chi^2_{n-1}/(n-1) \geq t(\mathbf{x})).$$

Possible alternatives to H_0 guide the choice and interpretation of T. What is a "best" test?

3.2 Simple null and alternative hypotheses: The Neyman-Pearson Lemma

The general setting here is as follows: we have a random sample $X_1, \ldots, X_n$ from $f(x; \theta)$, and two hypotheses:

a *null hypothesis* $H_0 : \theta \in \Theta_0$

an *alternative hypothesis* $H_1 : \theta \in \Theta_1$

where $\Theta_1 = \Theta \setminus \Theta_0$; Θ denotes the whole parameter space. We want to choose a *rejection region* or *critical region* R such that

reject $H_0 \iff \mathbf{X} \in R$.

Now suppose that $H_0 : \theta = \theta_0$, and $H_1 : \theta = \theta_1$ are both simple. The *Type I error* is: reject H_0 when it is true;

$$\alpha = P(\text{reject } H_0|H_0),$$

this is also known as *size* of the test. The *Type II error* is: accept H_0 when it is false;

$$\beta = P(\text{accept } H_0|H_1).$$

The *power* of the test is $1 - \beta = P(\text{accept } H_1 | H_1)$.

Usually we fix α (e.g., $\alpha = 0.05$, 0.01, etc.), and we look for a test which minimizes β: this is called a *most powerful* or *best* test of size α.

Intuitively: we reject H_0 in favour of H_1 if likelihood of θ_1 is much larger than likelihood of θ_0, given the data.

Neyman-Pearson Lemma: (see, e.g., Casella and Berger, p.366). The most powerful test at level α of H_0 versus H_1 has rejection region

$$R = \left\{ \mathbf{x} : \frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq k_\alpha \right\}$$

where the constant k_α is chosen so that

$$P(\mathbf{X} \in R | H_0) = \alpha.$$

This test is called the the *likelihood ratio* (LR) *test*.

Often we simplify the condition $L(\theta_1; \mathbf{x})/L(\theta_0; \mathbf{x}) \geq k_\alpha$ to

$$t(\mathbf{x}) \geq c_\alpha,$$

for some constant c_α and some statistic $t(\mathbf{x})$; determine c_α from the equation

$$P(T \geq c_\alpha | H_0) = \alpha,$$

where $T = t(\mathbf{X})$; then the test is "reject H_0 if and only if $T \geq c_\alpha$". For data $\mathbf{x}$ the p-value is $p = P(T \geq t(\mathbf{x}) | H_0)$.

Example: Normal means, one-sided. Suppose that we have a random sample $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, wioth σ^2 known; let μ_0 and μ_1 be given, and assume that $\mu_1 > \mu_0$. We would like to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$. Using the Neyman-Pearson Lemma, the best test is a likelihood ratio test. To calculate the test, we note that

$$\begin{aligned}
\frac{L(\mu_1; \mathbf{x})}{L(\mu_0; \mathbf{x})} \geq k &\Leftrightarrow \ell(\mu_1; \mathbf{x}) - \ell(\mu_0; \mathbf{x}) \geq \log k \\
&\Leftrightarrow -\sum \left[(x_i - \mu_1)^2 - (x_i - \mu_0)^2 \right] \geq 2\sigma^2 \log k \\
&\Leftrightarrow (\mu_1 - \mu_0)\overline{x} \geq k' \\
&\Leftrightarrow \overline{x} \geq c \qquad (\text{since } \mu_1 > \mu_0),
\end{aligned}$$

where k', c are constants, independent of $\mathbf{x}$. Hence we choose $t(\mathbf{x}) = \overline{x}$, and for size α test we choose c so that $P(\overline{X} \geq c|H_0) = \alpha$; equivalently, such that

$$\mathsf{P}\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{c - \mu_0}{\sigma/\sqrt{n}} \;\middle|\; H_0\right) = \alpha.$$

Hence we want

$$(c - \mu_0)/(\sigma/\sqrt{n}) = z_{1-\alpha},$$

(where $\Phi(z_{1-\alpha}) = 1 - \alpha$ with Φ being standard normal c.d.f.), i.e.

$$c = \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}.$$

So the most powerful test of H_0 versus H_1 at level α becomes "reject H_0 if and only if $\overline{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}$".

Recall the notation for standard normal quantiles: If $Z \sim \mathcal{N}(0,1)$ then

$$\mathsf{P}(Z \leq z_\alpha) = \alpha \text{ and } \mathsf{P}(Z \geq z(\alpha)) = \alpha,$$

and note that $z(\alpha) = z_{1-\alpha}$. Thus

$$\mathsf{P}(Z \geq z_{1-\alpha}) = 1 - (1 - \alpha) = \alpha.$$

Example: Bernoulli, probability of success, one-sided. Assume that $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ) then $L(\theta) = \theta^r(1-\theta)^{n-r}$ where $r = \sum x_i$. We would like to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where θ_0 and θ_1 are known, and $\theta_1 > \theta_0$. Now $\theta_1/\theta_0 > 1$, $(1-\theta_1)/(1-\theta_0) < 1$, and

$$\frac{L(\theta_1; \mathbf{x})}{L(\theta_0; \mathbf{x})} = \left(\frac{\theta_1}{\theta_0}\right)^r \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-r}$$

and so $L(\theta_1; \mathbf{x})/L(\theta_0; \mathbf{x}) \geq k_\alpha \iff r \geq r_\alpha$.

So the best test rejects H_0 for large r. For any given critical value r_c,

$$\alpha = \sum_{j=r_c}^{n} \binom{n}{j} \theta_0^j (1-\theta_0)^{n-j}$$

gives the p-value if we set $r_c = r(\mathbf{x}) = \sum x_i$, the observed value.

Note: The distribution is discrete, so we may not be able to achieve a level α test exactly (unless we use additional randomization). For example, if $R \sim Binomial(10, 0.5)$, then $P(R \geq 9) = 0.011$, and $P(R \geq 8) = 0.055$, so there is no c such that $P(R \geq c) = 0.05$. A solution is to randomize: If $R \geq 9$ reject the null hypothesis, if $R \leq 7$ accept the null hypothesis, and if $R = 8$ flip a (biased) coin to achieve the exact level of 0.05.

3.3 Composite alternative hypotheses

Suppose that θ scalar, $H_0 : \theta = \theta_0$ is simple, and we test against a composite alternative hypotheses; this could be one-sided:

$$H_1^- : \theta < \theta_0 \text{ or } H_1^+ : \theta > \theta_0;$$

or a two-sided alternative $H_1 : \theta \neq \theta_0$. The *power function* of a test depends on the true parameter θ, and is defined as

$$\text{power}(\theta) = P(\mathbf{X} \in R|\theta);$$

the probability of rejecting H_0 as a function of the true value of the parameter θ; it depends on α, the size of the test. Its main uses are comparing alternative tests, and choosing sample size.

3.3.1 Uniformly most powerful tests

A test of size α is *uniformly most powerful* (*UMP*) if its power function is such that

$$\text{power}(\theta) \geq \text{power}'(\theta)$$

for all $\theta \in \Theta_1$, where $\text{power}'(\theta)$ is the power function of any other size-α test.

Consider testing H_0 against H_1^+. For exponential family problems, usually for any $\theta_1 > \theta_0$ the rejection region of the LR test is independent of θ_1. At the same time, the test is most powerful for every single θ_1 which is larger than θ_0. Hence the test derived for one such value of θ_1 is UMP for H_0 versus H_1^+.

Example: normal mean, composite one-sided alternative. Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ are i.i.d., with σ^2 known. We want to test

$H_0 : \mu = \mu_0$ against $H_1^+ : \mu > \mu_0$. First pick an arbitrary $\mu_1 > \mu_0$. We have seen that the most powerful test of $\mu = \mu_0$ versus $\mu = \mu_1$ has a rejection region of the form

$$\overline{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}$$

for a test of size α. This rejection region is independent of μ_1, hence the test which rejects H_0 when $\overline{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n}$ is UMP for H_0 versus H_1^+. The power of the test is

$$\begin{aligned}
\text{power}(\mu) &= \mathsf{P}\big(\overline{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n} \;\big|\; \mu\big) \\
&= \mathsf{P}\big(\overline{X} \geq \mu_0 + \sigma z_{1-\alpha}/\sqrt{n} \;\big|\; \overline{X} \sim N(\mu, \sigma^2/n)\big) \\
&= \mathsf{P}\big(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \;\big|\; \overline{X} \sim N(\mu, \sigma^2/n)\big) \\
&= \mathsf{P}\big(Z \geq z_{1-\alpha} - \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \;\big|\; Z \sim N(0,1)\big) \\
&= 1 - \Phi\big(z_{1-\alpha} - (\mu - \mu_0)\sqrt{n}/\sigma\big).
\end{aligned}$$

The power increases from 0 up to α at $\mu = \mu_0$ and then to 1 as μ increases. The power increases as α increases.

Sample size calculation in the Normal example
Suppose want to be near-certain to reject H_0 when $\mu = \mu_0 + \delta$, say, and have size 0.05. Suppose we want to fix n to force $\text{power}(\mu) = 0.99$ at $\mu = \mu_0 + \delta$:

$$0.99 = 1 - \Phi(1.645 - \delta\sqrt{n}/\sigma)$$

so that $0.01 = \Phi(1.645 - \delta\sqrt{n}/\sigma)$. Solving this equation (use tables) gives $-2.326 = 1.645 - \delta\sqrt{n}/\sigma$, i.e.

$$n = \sigma^2(1.645 + 2.326)^2/\delta^2$$

is the required sample size.

UMP tests are not always available. If not, options include a

1. Wald test
2. locally most powerful test (score test)
3. generalised likelihood ratio test.

3.3.2 Wald tests

The Wald test is directly based on the asymptotic normality of the m.l.e. $\hat{\theta} = \hat{\theta}_n$, often $\hat{\theta} \approx \mathcal{N}\big(\theta, I_n^{-1}(\theta)\big)$ if θ is the true parameter. Also it is often true that asymptotically, we may replace θ by $\hat{\theta}$ in the Fisher information,

$$\hat{\theta} \approx \mathcal{N}\big(\theta, I_n^{-1}(\hat{\theta})\big).$$

So we can construct a test based on

$$W = \sqrt{I_n(\hat{\theta})}(\hat{\theta} - \theta_0) \approx \mathcal{N}(0, 1).$$

If θ is scalar, squaring gives

$$W^2 \approx \chi_1^2,$$

so equivalently we could use a chi-square test.

For higher-dimensional θ we can base a test on the quadratic form

$$(\hat{\theta} - \theta_0)^T I_n(\hat{\theta})(\hat{\theta} - \theta_0)$$

which is approximately chi-square distributed in large samples.

If we would like to test $H_0 : g(\theta) = 0$, where g is a (scalar) differentiable function, then the delta method gives as test statistic

$$W = g(\hat{\theta})\{G(\hat{\theta})(I_n(\hat{\theta}))^{-1}G(\hat{\theta})^T\}^{-1}g(\hat{\theta}),$$

where $G(\theta) = \frac{\partial g(\theta)}{\partial \theta}^T$.

An advantage of the Wald test is that we do not need to evaluate the likelihood under the null hypothesis, which can be awkward if the null hypothesis contains a number of restrictions on a multidimensional parameter. All we need is (an approximation) of $\hat{\theta}$, the maximum-likelihood-estimator. But there is also a disadvantage:

Example: Non-invariance of the Wald test

Suppose that $\hat{\theta}$ is scalar and approximately $\mathcal{N}(\theta, I_n(\theta)^{-1})$-distributed, then for testing $H_0 : \theta = 0$ the Wald statistic becomes

$$\hat{\theta}\sqrt{I_n(\hat{\theta})},$$

which would be approximately standard normal. If instead we tested $H_0 : \theta^3 = 0$, then the delta method with $g(\theta) = \theta^3$, so that $g'(\theta) = 3\theta^2$, gives

$$Var(g(\hat{\theta})) \approx 9\hat{\theta}^4 (I_n(\hat{\theta}))^{-1}$$

and as Wald statistic

$$\frac{\hat{\theta}}{3}\sqrt{I_n(\hat{\theta})}.$$

3.3.3 Locally most powerful test (Score test)

We consider first the problem to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 + \delta$. for some small $\delta > 0$. We have seen that the most powerful test has a rejection region of the form

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \geq k.$$

Taylor expansion gives

$$\ell(\theta_0 + \delta) \approx \ell(\theta_0) + \delta \frac{\partial \ell(\theta_0)}{\partial \theta}$$

i.e.

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \approx \delta \frac{\partial \ell(\theta_0)}{\partial \theta}.$$

So a locally most powerful (LMP) test has as rejection region

$$R = \left\{ \mathbf{x} : \frac{\partial \ell(\theta_0)}{\partial \theta} \geq k_\alpha \right\}.$$

This is also called the *score test*: $\partial \ell / \partial \theta$ is known as the *score function*. Under certain regularity conditions,

$$\mathsf{E}_\theta \left[\frac{\partial \ell}{\partial \theta} \right] = 0, \qquad \mathrm{Var}_\theta \left[\frac{\partial \ell}{\partial \theta} \right] = I_n(\theta).$$

As ℓ is usually a sum of independent components, so is $\partial \ell(\theta_0)/\partial \theta$, and the CLT (Central Limit Theorem) can be applied.

Example: Cauchy parameter. Suppose that $X_1, \ldots, X_n$ is a random sample from Cauchy (θ), having density

$$f(x; \theta) = \left[\pi \big(1 + (x - \theta)^2 \big) \right]^{-1} \quad \text{for } -\infty < x < \infty.$$

Test $H_0 : \theta = \theta_0$ against $H_1^+ : \theta > \theta_0$. Then

$$\frac{\partial \ell(\theta_0; \mathbf{x})}{\partial \theta} = 2 \sum \left\{ \frac{x_i - \theta_0}{1 + (x_i - \theta_0)^2} \right\}.$$

Fact: Under H_0, the expression $\partial \ell(\theta_0; \mathbf{X})/\partial \theta$ has mean 0, variance $I_n(\theta_0) = n/2$. The CLT applies, $\partial \ell(\theta_0; \mathbf{X})/\partial \theta \approx \mathcal{N}(0, n/2)$ under H_0, so for the LMP test,

$$P(\mathcal{N}(0, n/2) \geq k_\alpha) = P\left(\mathcal{N}(0,1) \geq k_\alpha \sqrt{\frac{2}{n}} \right) \approx \alpha.$$

This gives $k_\alpha \approx z_{1-\alpha}\sqrt{n/2}$, and as rejection region with approximate size α

$$R = \left\{ \mathbf{x} : 2 \sum \left(\frac{x_i - \theta_0}{1 + (x_i - \theta_0)^2} \right) > \sqrt{\frac{n}{2}} z_{1-\alpha} \right\}.$$

The score test has the advantage that we only need the likelihood under the null hypothesis. It is also not generally invariant under reparametrisation.

The multidimensional version of the score test is as follows: Let $U = \partial \ell / \partial \theta$ be the score function, then the score statistic is

$$U^T I_n(\theta_0)^{-1} U.$$

Compare with a chi-square distribution.

3.3.4 Generalised likelihood ratio (LR) test

For testing $H_0 : \theta = \theta_0$ against $H_1^+ : \theta > \theta_0$, the generalised likelihood ratio test uses as rejection region

$$R = \left\{ \mathbf{x} : \frac{\max_{\theta \geq \theta_0} L(\theta; \mathbf{x})}{L(\theta_0; \mathbf{x})} \geq k_\alpha \right\}.$$

If L has one mode, at the m.l.e. $\hat{\theta}$, then the likelihood ratio in the definition of R is either 1, if $\hat{\theta} \leq \theta_0$, or $L(\hat{\theta}; \mathbf{x})/L(\theta_0; \mathbf{x})$, if $\hat{\theta} > \theta_0$ (and similarly for H_1^-, with fairly obvious changes of signs and directions of inequalities).

The generalised LRT is invariant to a change in parametrisation.

3.4 Two-sided tests

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. If the one-sided tests of size α have symmetric rejection regions

$$R^+ = \{\mathbf{x} : t > c\} \quad \text{and} \quad R^- = \{\mathbf{x} : t < -c\},$$

then a two-sided test (of size 2α) is to take the rejection region to

$$R = \{\mathbf{x} : |t| > c\};$$

this test has as p-value $p = P(|t(\mathbf{X})| \geq t|H_0)$.

The two-sided (generalised) LR test uses

$$T = 2\log\left[\frac{\max_\theta L(\theta;\mathbf{X})}{L(\theta_0;\mathbf{X})}\right] = 2\log\left[\frac{L(\hat{\theta};\mathbf{X})}{L(\theta_0;\mathbf{X})}\right]$$

and rejects H_0 for large T.

Fact: $T \approx \chi^2_1$ under H_0 (to be seen in Chapter 5).

Where possible, the exact distribution of T or of a statistic equivalent to T should be used.

If θ is a vector: there is no such thing as a one-sided alternative hypothesis. For the alternative $\theta \neq \theta_0$ we use a LR test based on

$$T = 2\log\left[\frac{L(\hat{\theta};\mathbf{X})}{L(\theta_0;\mathbf{X})}\right].$$

Under H_0, $T \approx \chi^2_p$ where $p =$ dimension of θ (see Chapter 5).

For the score test we use as statistic

$$\ell'(\theta_0)^T [I_n(\theta_0)]^{-1} \ell'(\theta_0),$$

where $I_n(\theta)$ is the expected Fisher information matrix:

$$[I_n(\theta)]_{jk} = \mathsf{E}[-\partial^2\ell/\partial\theta_j\partial\theta_k].$$

If the CLT applies to the score function, then this quadratic form is again approximately χ^2_p under H_0 (see Chapter 5).

Example: Pearson's Chi-square statistic. We have a random sample of size n, with p categories; $\mathsf{P}(X_j = i) = \pi_i$, for $i = 1, \ldots, p$, $j = 1, \ldots, n$. As $\sum \pi_i = 1$, we take $\theta = (\pi_1, \ldots, \pi_{p-1})$. The likelihood function is then

$$\prod \pi_i^{n_i}$$

where n_i = # observations in category i (so $\sum n_i = n$). We think of $n_1, \ldots, n_p$ as realisations of random counts $N_1, \ldots, N_p$. The m.l.e. is $\hat{\theta} = n^{-1}(n_1, \ldots, n_{p-1})$. Test $H_0 : \theta = \theta_0$, where $\theta_0 = (\pi_{1,0}, \ldots, \pi_{p-1,0})$, against $H_1 : \theta \neq \theta_0$.

The score vector is vector of partial derivatives of

$$\ell(\theta) = \sum_{i=1}^{p-1} n_i \log \pi_i + n_p \log \left(1 - \sum_{k=1}^{p-1} \pi_k \right)$$

with respect to $\pi_1, \ldots, \pi_{p-1}$:

$$\frac{\partial \ell}{\partial \pi_i} = \frac{n_i}{\pi_i} - \frac{n_p}{1 - \sum_{k=1}^{p-1} \pi_k}.$$

The matrix of second derivatives has entries

$$\frac{\partial^2 \ell}{\partial \pi_i \partial \pi_k} = -\frac{n_i \delta_{ik}}{\pi_i^2} - \frac{n_p}{(1 - \sum_{i=1}^{p-1} \pi_i)^2},$$

where $\delta_{ik} = 1$ if $i = k$, and $\delta_{ik} = 0$ if $i \neq k$. Minus the expectation of this, using $E_{\theta_0}(N_i) = n\pi_i$, gives

$$I_n(\theta) = n Diag(\pi_1^{-1}, \ldots, \pi_{p-1}^{-1}) + n 1 1^T \pi_p^{-1},$$

where 1 is a $(p-1)$-dimensional vector of ones.

Compute

$$\ell'(\theta_0)^T [I_n(\theta_0)]^{-1} \ell'(\theta_0) = \sum_{i=1}^{p} \frac{(n_i - n\pi_{i,0})^2}{n\pi_{i,0}};$$

this statistic is called the *chi-squared statistic*, T say. The CLT for the score vector gives that $T \approx \chi^2_{p-1}$ under H_0.

Note: the form of the chi-squared statistic is

$$\sum (O_i - E_i)^2 / E_i$$

where O_i and E_i refer to observed and expected frequencies in category i: This is known as *Pearson's chi-square statistic*.

3.5 Composite null hypotheses

Let $\theta = (\psi, \lambda)$, where λ is a nuisance parameter. We want a test which does not depend on the unknown value of λ. Extending two of the previous methods:

3.5.1 Generalised likelihood ratio test: Composite null hypothesis

Suppose that we want to test $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$. The (generalised) LR test uses the likelihood ratio statistic

$$T = \frac{\max\limits_{\theta \in \Theta} L(\theta; \mathbf{X})}{\max\limits_{\theta \in \Theta_0} L(\theta; \mathbf{X})}$$

and rejects H_0 for large values of T.

Now $\theta = (\psi, \lambda)$. Assuming that ψ is scalar, test $H_0 : \psi = \psi_0$ against $H_1^+ : \psi > \psi_0$. The LR statistic T is

$$T = \frac{\max\limits_{\psi \geq \psi_0, \lambda} L(\psi, \lambda)}{\max\limits_{\lambda} L(\psi_0, \lambda)} = \frac{\max\limits_{\psi \geq \psi_0} L_P(\psi)}{L_P(\psi_0)},$$

where $L_P(\psi)$ is the profile likelihood for ψ. For H_0 against $H_1 : \psi \neq \psi_0$,

$$T = \frac{\max\limits_{\psi, \lambda} L(\psi, \lambda)}{\max\limits_{\lambda} L(\psi_0, \lambda)} = \frac{L(\hat{\psi}, \hat{\lambda})}{L_P(\psi_0)}.$$

Often (see Chapter 5):

$$2 \log T \approx \chi^2_p$$

where p is the dimension of ψ.

An important requirement for this approach is that the dimension of λ does not depend on n.

Example: Normal distribution and Student t-test. Suppose that $\mathbf{X}$ is a random sample of size n, from $N(\mu, \sigma^2)$, where both μ and σ are unknown; we would like to test $H_0 : \mu = \mu_0$. Ignoring an irrelevant additive constant,

$$\ell(\theta) = -n \log \sigma - \frac{n(\overline{x} - \mu)^2 + (n-1)s^2}{2\sigma^2}.$$

Maximizing this w.r.t. σ with μ fixed gives

$$\ell_P(\mu) = -\frac{n}{2} \log \left(\frac{(n-1)s^2 + n(\overline{x} - \mu)^2}{n} \right).$$

If our alternative is $H_1^+ : \mu > \mu_0$ then we maximize $\ell_P(\mu)$ over $\mu \geq \mu_0$:

if $\overline{x} \leq \mu_0$ then the maximum is at $\mu = \mu_0$; if $\overline{x} > \mu_0$ then the maximum is at $\mu = \overline{x}$. So $\log T = 0$ when $\overline{x} \leq \mu_0$ and is

$$\begin{aligned} & -\frac{n}{2} \log \left(\frac{(n-1)s^2}{n} \right) + \frac{n}{2} \log \left(\frac{(n-1)s^2 + n(\overline{x} - \mu_0)^2}{n} \right) \\ = \quad & \frac{n}{2} \log \left(1 + \frac{n(\overline{x} - \mu_0)^2}{(n-1)s^2} \right) \end{aligned}$$

when $\overline{x} > \mu_0$. Thus the LR rejection region is of the form

$$R = \{\mathbf{x} : t(\mathbf{x}) \geq c_\alpha\},$$

where

$$t(\mathbf{x}) = \frac{\sqrt{n}(\overline{x} - \mu_0)}{s}.$$

This statistic is called *Student-t* statistic. Under H_0, $t(\mathbf{X}) \sim t_{n-1}$, and for a size α test set $c_\alpha = t_{n-1,1-\alpha}$; the p-value is $p = \mathsf{P}(t_{n-1} \geq t(\mathbf{x}))$. Here we use the quantile notation $\mathsf{P}(t_{n-1} \geq t_{n-1,1-\alpha}) = \alpha$.

The two-sided test of H_0 against $H_1 : \mu \neq \mu_0$ is easier, as unconstrained maxima are used. The size α test has rejection region

$$R = \{\mathbf{x} : |t(\mathbf{x})| \geq t_{n-1,1-\alpha/2}\}.$$

3.5.2 Score test: Composite null hypothesis

Now $\theta = (\psi, \lambda)$ with ψ scalar, test $H_0 : \psi = \psi_0$ against $H_1^+ : \psi > \psi_0$ or $H_1^- : \psi < \psi_0$. The score test statistic is

$$T = \frac{\partial \ell(\psi_0, \hat{\lambda}_0; \mathbf{X})}{\partial \psi},$$

where $\hat{\lambda}_0$ is the MLE for λ when H_0 is true. Large positive values of T indicate H_1^+, and large negative values indicate H_1^-. Thus the rejection regions are of the form $T \geq k_\alpha^+$ when testing against H_1^+, and $T \leq k_\alpha^-$ when testing against H_1^-.

Recall the derivation of the score test,

$$\ell(\theta_0 + \delta) - \ell(\theta_0) \approx \delta \frac{\partial \ell(\theta_0)}{\partial \theta} = \delta T.$$

If $\delta > 0$, i.e. for H_1^+, we reject if T is large; if $\delta < 0$, i.e. for H_1^-, we reject if T is small.

Sometimes the exact null distribution of T is available; more often we use that $T \approx$ normal (by CLT, see Chapter 5), zero mean. To find the approximate variance:

1. compute $I_n(\psi_0, \lambda)$
2. invert to I_n^{-1}
3. take the diagonal element corresponding to ψ
4. invert this element
5. replace λ by the null hypothesis MLE $\hat{\lambda}_0$.

Denote the result by v, then $Z = T/\sqrt{v} \approx \mathcal{N}(0, 1)$ under H_0.

A considerable advantage is that the unconstrained MLE $\hat{\psi}$ is not required.

Example: linear or non-linear model? We can extend the linear model $Y_j = (x_j^T \beta) + \epsilon_j$, where $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$, to a non-linear model

$$Y_j = (x_j^T \beta)^\psi + \epsilon_j$$

with the same ϵ's. Test $H_0 : \psi = 1$: usual linear model, against, say, $H_1^- : \psi < 1$. Here our nuisance parameters are $\lambda^T = (\beta^T, \sigma^2)$.

Write $\eta_j = x_j^T \beta$, and denote the usual linear model fitted values by $\hat{\eta}_{j0} = x_j^T \hat{\beta}_0$, where the estimates are obtained under H_0. As $Y_j \sim \mathcal{N}(\eta_j, \sigma^2)$, we have up to an irrelevant additive constant,

$$\ell(\psi, \beta, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum (y_j - \eta_j^\psi)^2,$$

and so

$$\frac{\partial \ell}{\partial \psi} = \frac{1}{\sigma^2} \sum (y_j - \eta_j^\psi) \eta_j^\psi \log \eta_j,$$

yielding that the null MLE's are the usual LSEs (least-square estimates), which are

$$\hat{\beta}_0 = (X^T X)^{-1} X^T Y, \quad \hat{\sigma}^2 = n^{-1} \sum (Y_j - x_j^T \hat{\beta}_0)^2.$$

So the score test statistic becomes

$$T = \frac{1}{\hat{\sigma}^2} \sum (Y_j - \hat{\eta}_{j0})(\hat{\eta}_{j0} \log \hat{\eta}_{j0}).$$

We reject H_0 for large negative values of T.

Compute the approximate null variance (see below):

$$I_n(\psi_0, \beta, \sigma) = \frac{1}{\sigma^2} \begin{pmatrix} \sum u_j^2 & \sum u_j x_j^T & 0 \\ \sum u_j x_j & \sum x_j x_j^T & 0 \\ 0 & 0 & 2n \end{pmatrix}$$

where $u_j = \eta_j \log \eta_j$. The $(1,1)$ element of the inverse of I_n has reciprocal

$$\left(u^T u - u^T X (X^T X)^{-1} X^T u\right) / \sigma^2,$$

where $u^T = (u_1, \ldots, u_n)$. Substitute $\hat{\eta}_{j0}$ for η_j and $\hat{\sigma}^2$ for σ^2 to get v. For the approximate p-value calculate $z = t/\sqrt{v}$ and set $p = \Phi(z)$.

Calculation trick: To compute the $(1,1)$ element of the inverse of I_n above: if

$$A = \begin{pmatrix} a & x^T \\ x & B \end{pmatrix}$$

where a is a scalar, x is an $(n-1)\times 1$ vector and B is an $(n-1)\times(n-1)$ matrix, then $(A^{-1})_{11} = 1/(a - x^T B^{-1} x)$.

Recall also:

$$\frac{\partial}{\partial\psi}\eta^\psi = \frac{\partial}{\partial\psi} e^{\psi \ln\eta} = \ln\eta e^{\psi\ln\eta} = \eta^\psi \ln\eta.$$

For the $(1,1)$-entry of the information matrix, we calculate

$$\frac{\partial^2\ell}{\partial\psi^2} = \frac{1}{\sigma^2}\sum\left\{(-\eta_j^\psi \log\eta_j)\eta_j^\psi \log\eta_j + (y_j - \eta_j^\psi)\eta_j^\psi (\log\eta_j)^2\right\},$$

and as $Y_j \sim \mathcal{N}(\eta_j, \sigma^2)$ we have

$$E\left\{-\frac{\partial^2\ell}{\partial\psi^2}\right\} = \frac{1}{\sigma^2}\sum \eta_j^\psi \log\eta_j \eta_j^\psi \log\eta_j = \frac{1}{\sigma^2}\sum u_j^2,$$

as required. The off-diagonal terms in the information matrix can be calculated in a similar way, using that $\frac{\partial}{\partial\beta}\eta = x_j^T$.

3.6 Multiple tests

When many tests applied to the same data, there is a tendency for some p-values to be small: Suppose $P_1, \ldots, P_m$ are the random P-values for m independent tests at level α (before seeing the data); for each i, suppose that $P(P_i \le \alpha) = \alpha$ if the null hypothesis is true. But then the probability that at least one of the null hypothesis is rejected if m independent tests are carried out is

$$1 - P(\text{ none rejected}) = 1 - (1-\alpha)^m.$$

Example: If $\alpha = 0.05$ and $m = 10$, then

$$P(\text{ at least one rejected } | H_0 \text{ true }) = 0.4012.$$

Thus with high probability at least one "significant" result will be found even when all the null hypotheses are true.

Bonferroni: The Bonferroni inequality gives that

$$P(\min P_i \leq \alpha | H_0) \leq \sum_{i=1}^{m} P(P_i \leq \alpha | H_0) \leq m\alpha.$$

A cautious approach for an overall level α is therefore to declare the most significant of m test results as significant at level p only if $\min p_i \leq p/m$.

Example: If $\alpha = 0.05$ and $m = 10$, then reject only if the p-value is less than 0.005.

3.7 Combining independent tests

Suppose we have k independent experiments/studies for the same null hypothesis. If only the p-values are reported, and if we have continuous distribution, we may use that under H_0 each p-value is $\mathcal{U}[0,1]$ uniformly distributed (Exercise). This gives that

$$-2\sum_{i=1}^{k} \log P_i \sim \chi^2_{2k}$$

(exactly) under H_0, so

$$p_{\text{comb}} = \mathsf{P}(\chi^2_{2k} \geq -2\sum \log p_i).$$

If each test is based on a statistic T such that $T_i \approx \mathcal{N}(0, v_i)$, then the best combination statistic is

$$Z = \sum (T_i/v_i) / \sqrt{\sum v_i^{-1}}.$$

If H_0 is a hypothesis about a common parameter ψ, then the best combination of evidence is

$$\sum \ell_{P,i}(\psi),$$

and the combined test would be derived from this (e.g., an LR or score test).

Advice

Even though a test may initially be focussed on departures in one direction, it is usually a good idea not to totally disregard departures in the other direction, even if they are unexpected.

Warning:

Not rejecting the null hypothesis does not mean that the null hypothesis is true! Rather it means that there is not enough evidence to reject the null hypothesis; the data are consistent with the null hypothesis.

The p-value is **not** the probability that the null hypothesis is true.

3.8 Nonparametric tests

Sometimes we do not have a parametric model available, and the null hypothesis is phrased in terms of arbitrary distributions, for example concerning only the median of the underlying distribution. Such tests are called *nonparametric* or *distribution-free*; treating these would go beyond the scope of these lectures.

Chapter 4

Interval estimation

The goal for interval estimation is to specify the accurary of an estimate. A $1-\alpha$ *confidence set* for a parameter θ is a set $C(\mathbf{X})$ in the parameter space Θ, depending only on $\mathbf{X}$, such that

$$\mathsf{P}_\theta\big(\theta \in C(\mathbf{X})\big) = 1-\alpha.$$

Note: it is not θ that is random, but the set $C(\mathbf{X})$ is.

For a scalar θ we would usually like to find an interval

$$C(\mathbf{X}) = [l(\mathbf{X}), u(\mathbf{X})]$$

so that $\mathsf{P}_\theta\big(\theta \in [l(\mathbf{X}), u(\mathbf{X})]\big) = 1-\alpha$. Then $[l(\mathbf{X}), u(\mathbf{X})]$ is an *interval estimator* or *confidence interval* for θ; and the observed interval $[l(\mathbf{x}), u(\mathbf{x})]$ is an *interval estimate*. If l is $-\infty$ or if u is $+\infty$, then we have a *one-sided estimator/estimate*. If l is $-\infty$, we have an *upper confidence interval*, if u is $+\infty$, we have an *lower confidence interval*.

Example: Normal, unknown mean and variance. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Then $(\overline{X}-\mu)/(S/\sqrt{n}) \sim t_{n-1}$ and so

$$\begin{aligned} 1-\alpha &= \mathsf{P}_{\mu,\sigma^2}\left(\left|\frac{\overline{X}-\mu}{S/\sqrt{n}}\right| \le t_{n-1,1-\alpha/2}\right) \\ &= \mathsf{P}_{\mu,\sigma^2}(\overline{X} - t_{n-1,1-\alpha/2}S/\sqrt{n} \le \mu \le \overline{X} + t_{n-1,1-\alpha/2}S/\sqrt{n}), \end{aligned}$$

and so the (familiar) interval with end points

$$\overline{X} \pm t_{n-1,1-\alpha/2}S/\sqrt{n}$$

is a $1-\alpha$ confidence interval for μ.

4.1 Construction of confidence sets

4.1.1 Pivotal quantities

A *pivotal quantity* (or *pivot*) is a random variable $t(\mathbf{X}, \theta)$ whose distribution is independent of all parameters, and so it has the same distribution for all θ.

Example: $(\overline{X} - \mu)/(S/\sqrt{n})$ in the example above has t_{n-1}-distribution if the random sample comes from $\mathcal{N}(\mu, \sigma^2)$.

We use pivotal quantities to construct confidence sets, as follows. Suppose θ is a scalar. Choose a, b such that

$$\mathsf{P}_\theta(a \leq t(\mathbf{X}, \theta) \leq b) = 1 - \alpha.$$

Manipulate this equation to give $\mathsf{P}_\theta\big(l(\mathbf{X}) \leq \theta \leq u(\mathbf{X})\big) = 1 - \alpha$ (if t is a monotonic function of θ); then $[l(\mathbf{X}), u(\mathbf{X})]$ is a $1 - \alpha$ confidence interval for θ.

Example: Exponential random sample. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with unknown mean μ. Then we know that $n\overline{X}/\mu \sim \text{Gamma}(n, 1)$. If the α-quantile of Gamma$(n, 1)$ is denoted by $g_{n,\alpha}$ then

$$1 - \alpha \;\; = \;\; \mathsf{P}_\mu(n\overline{X}/\mu \geq g_{n,\alpha}) = \mathsf{P}_\mu(\mu \leq n\overline{X}/g_{n,\alpha}).$$

Hence $[0, n\overline{X}/g_{n,\alpha}]$ is a $1 - \alpha$ confidence interval for μ. Alternatively, we say that $n\overline{X}/g_{n,\alpha}$ is the upper $1 - \alpha$ confidence limit for μ.

4.1.2 Confidence sets derived from point estimators

Suppose $\hat{\theta}(\mathbf{X})$ is an estimator for a scalar θ, from a known distribution. Then we can take our confidence interval as

$$[\hat{\theta} - a_{1-\alpha}, \hat{\theta} + b_{1-\alpha}]$$

where $a_{1-\alpha}$ and $b_{1-\alpha}$ are chosen suitably.

If $\hat{\theta} \sim N(\theta, v)$, perhaps approximately, then for a symmetric interval choose

$$a_{1-\alpha} = b_{1-\alpha} = z_{1-\alpha/2}\sqrt{v}.$$

Note: $[\hat{\theta} - a_{1-\alpha}, \hat{\theta} + b_{1-\alpha}]$ is not immediately a confidence interval for θ if v depends on θ: in that case replace $v(\theta)$ by $v(\hat{\theta})$, which is a further approximation.

4.1.3 Approximate confidence intervals

Sometimes we do not have an exact distribution available, but normal approximation is known to hold.

Example: asymptotic normality of m.l.e. . We have seen that, under regularity, $\hat{\theta} \approx \mathcal{N}\big(\theta, I^{-1}(\theta)\big)$. If θ is scalar, then (under regularity)

$$\hat{\theta} \pm z_{1-\alpha/2}/\sqrt{I_n(\hat{\theta})}$$

is an approximate $1 - \alpha$ confidence interval for θ.

Sometimes we can improve the accuracy by applying (monotone) transformation of the estimator, using the delta method, and inverting the transformation to get the final result.

As a guide line for transformations, in general a normal approximation should be used on a scale where a quantity ranges over $(-\infty, \infty)$.

Example: Bivariate normal distribution. Let (X_i, Y_i), $i = 1, \ldots, n$, be a random sample from a bivariate normal distribution, with unknown mean vector and covariance matrix. The parameter of interest is ρ, the bivariate normal correlation. The MLE for ρ is the sample correlation

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}},$$

whose range is $[-1, 1]$. For large n,

$$R \approx N(\rho, (1 - \rho^2)^2/n),$$

using the expected Fisher information matrix to obtain an approximate variance (see the section on asymptotic theory).

But the distribution of R is very skewed, the approximation is poor unless n is very large. For a variable whose range is $(-\infty, \infty)$, we use the tranformation

$$Z = \frac{1}{2} \log[(1 + R)/(1 - R)];$$

this transformation is called the *Fisher z transformation*. By the delta method,

$$Z \approx N(\zeta, 1/n)$$

where $\zeta = \frac{1}{2}\log[(1+\rho)/(1-\rho)]$. So a $1-\alpha$ confidence interval for ρ can be calculated as follows: for ζ compute the interval limits $Z \pm z_{1-\alpha/2}/\sqrt{n}$, then transform these to the ρ scale using the inverse transformation $\rho = (e^{2\zeta}-1)/(e^{2\zeta}+1)$.

4.1.4 Confidence intervals derived from hypothesis tests

Define $C(\mathbf{X})$ to be the set of values of θ_0 for which H_0 would not be rejected in size-α tests of $H_0 : \theta = \theta_0$. Here the form of the $1-\alpha$ confidence set obtained depends on the alternative hypotheses.
Example: to produce an interval with finite upper and lower limits use $H_1 : \theta \neq \theta_0$; to find an upper confidence limit use $H_1^- : \theta < \theta_0$.

Example: Normal, known variance, unknown mean. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, where σ^2 known. For $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ the usual test has an acceptance region of the form

$$\left| \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \right| \leq z_{1-\alpha/2}.$$

So the values of μ_0 for which H_0 is accepted are those in the interval

$$[\overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}];$$

this interval is a $100(1-\alpha)\%$ confidence interval for μ.

For $H_0 : \mu = \mu_0$ versus $H_1^- : \mu < \mu_0$ the UMP test accepts H_0 if

$$\overline{X} \geq \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}$$

i.e., if

$$\mu_0 \leq \overline{X} + z_{1-\alpha}\sigma/\sqrt{n}.$$

So an upper $1-\alpha$ confidence limit for μ is $\overline{X} + z_{1-\alpha}\sigma/\sqrt{n}$.

4.2 Hypothesis test from confidence regions

Conversely, we can also construct tests based on confidence interval:

For $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, if $C(\mathbf{X})$ is $100(1-\alpha)\%$ two-sided confidence region for θ, then for a size α test reject H_0 if $\theta_0 \neq C(\mathbf{X})$: The confidence region is the acceptance region for the test.

If θ is a scalar: For $H_0 : \theta = \theta_0$ against $H_1^- : \theta < \theta_0$, if $C(\mathbf{X})$ is $100(1-\alpha)\%$ upper confidence region for θ, then for a size α test reject H_0 if $\theta_0 \neq C(\mathbf{X})$.

Example: Normal, known variance. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d., where σ^2 is known. For $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ the usual $100(1-\alpha)\%$ confidence region is

$$[\overline{X} - z_{1-\alpha/2}\sigma/\sqrt{n}, \overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}],$$

so reject H_0 if

$$\left| \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2}.$$

To test $H_0 : \mu = \mu_0$ versus $H_1^- : \mu < \mu_0$: an upper $100(1-\alpha)\%$ confidence region is $\overline{X} + z_{1-\alpha/2}\sigma/\sqrt{n}$, so reject H_0 if

$$\mu_0 > \overline{X} + z_{1-\alpha}\sigma/\sqrt{n}$$

i.e. if

$$\overline{X} < \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}.$$

We can also construct approximate hypothesis test based on approximate confidence intervals. For example, we use the asymptotic normality of m.l.e. to derive a Wald test.

4.3 Prediction Sets

What is a set of plausible values for a future data value? A $1-\alpha$ *prediction set* for an unobserved random variable X_{n+1} based on the observed data $\mathbf{X} = (X_1, \ldots, X_n)$ is a random set $P(\mathbf{X})$ for which

$$\mathsf{P}(X_{n+1} \in P(\mathbf{X})) = 1 - \alpha.$$

Sometimes such a set can be derived by finding a *prediction pivot* $t(\mathbf{X}, X_{n+1})$ whose distribution does not depend on θ. If a set R is such that $\mathsf{P}(t(\mathbf{X}, X_{n+1}) \in R) = 1 - \alpha$, then a $1 - \alpha$ prediction set is

$$P(\mathbf{X}) = \{X_{n+1} : t(\mathbf{X}, X_{n+1}) \in R\}.$$

Example: Normal, unknown mean and variance. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d., where both μ and σ^2 are unknown. A possible prediction pivot is

$$t(\mathbf{X}, X_{n+1}) = \frac{X_{n+1} - \overline{X}}{S\sqrt{1 + \frac{1}{n}}}.$$

As $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ and $X_{n+1} \sim N(\mu, \sigma^2)$ is independent of $\overline{X}$, it follows that $X_{n+1} - \overline{X} \sim N(0, \sigma^2(1 + 1/n))$, and so $t(\mathbf{X}, X_{n+1})$ has t_{n-1} distribution. Hence a $1 - \alpha$ prediction interval is

$$\begin{aligned}
&\{X_{n+1} : |t(\mathbf{X}, X_{n+1})| \leq t_{n-1,1-\alpha/2}\} \\
&= \left\{X_{n+1} : \overline{X} - S\sqrt{1 + \frac{1}{n}} t_{n-1,1-\alpha/2} \leq X_{n+1} \leq \overline{X} + S\sqrt{1 + \frac{1}{n}} t_{n-1,1-\alpha/2}\right\}.
\end{aligned}$$

Chapter 5

Asymptotic Theory

What happens as $n \to \infty$?

Let $\theta = (\theta_1, \ldots, \theta_p)$ be the parameter of interest, let $\ell(\theta)$ be the log-likelihood. Then $\ell'(\theta)$ is a vector, with jth component $\partial\ell/\partial\theta_j$, and $I_n(\theta)$ is the Fisher information matrix, whose (j,k) entry is $E_\theta\left(-\partial^2\ell/\partial\theta_j\partial\theta_k\right)$.

5.1 Consistency

A sequence of estimators T_n for θ, where $T_n = t_n(X_1, \ldots, X_n)$, is said to be *consistent* if, for any $\epsilon > 0$,

$$P_\theta(|T_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

In that case we also say that T_n converges to θ *in probability*.

Example: the sample mean. Let $\overline{X}_n$ be and i.i.d. sample of size n, with finite variance, mean θ then, by the weak law of large numbers, $\overline{X}_n$ is consistent for θ.

Recall: The weak law of large numbers states: Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$, and let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,

$$P(|\overline{X}_n - \mu| > \epsilon) \to 0 \text{ as } n \to \infty.$$

A sufficient condition for consistency is that $Bias(T_n) \to 0$ and $\mathrm{Var}(T_n) \to 0$ as $n \to \infty$. (Use the Chebyshev inequality to show this fact).

Subject to regularity conditions, MLEs are consistent.

5.2 Distribution of MLEs

Assume that $X_1, \ldots, X_n$ are i.i.d. where θ scalar, and $\hat{\theta} = \hat{\theta}(\mathbf{X})$ is the m.l.e.; assume that $\hat{\theta}$ exists and is unique. In regular problems, $\hat{\theta}$ is solution to the likelihood equation $\ell'(\theta) = 0$. Then Taylor expansion gives

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta) + (\hat{\theta} - \theta)\ell''(\theta)$$

and so

$$\frac{-\ell''(\theta)}{I_n(\theta)}(\hat{\theta} - \theta) \approx \frac{\ell'(\theta)}{I_n(\theta)}. \tag{5.1}$$

For the left hand side of (5.1):

$$-\ell''(\theta)/I_n(\theta) = \sum Y_i/(n\mu)$$

where

$$Y_i = \partial^2/\partial\theta^2\{\log f(X_i;\theta)\}$$

and $\mu = \mathsf{E}(Y_i)$. The weak law of large numbers gives that

$$-\ell''(\theta)/I_n(\theta) \to 1$$

in probability, as $n \to \infty$. So

$$\hat{\theta} - \theta \approx \frac{\ell'(\theta)}{I_n(\theta)}.$$

For the right hand side of (5.1),

$$\ell'(\theta) = \sum \partial/\partial\theta\{\log f(X_i;\theta)\}$$

is the sum of i.i.d. random variables. By the CLT, $\ell'(\theta)$ is approximately normal with mean $\mathsf{E}[\ell'(\theta)] = 0$ and variance $\mathrm{Var}(\ell'(\theta)) = I_n(\theta)$, and hence $\ell'(\theta) \approx N(0, I_n(\theta))$ or

$$\ell'(\theta)/I(\theta) \approx N(0, [I_n(\theta)]^{-1}). \tag{5.2}$$

Combining:

$$\hat{\theta} - \theta \approx N(0, [I_n(\theta)]^{-1}).$$

Result:

$$\hat{\theta} \approx N(\theta, [I_n(\theta)]^{-1}) \tag{5.3}$$

is the approximate distribution of the MLE.

The above argument generalises immediately to θ being a vector: if θ has p components, say, then $\hat{\theta}$ is approximately multivariate normal in p-dimensions with mean vector θ and covariance matrix $[I_n(\theta)]^{-1}$.

In practice we often use $I_n(\hat{\theta})$ in place of $I_n(\theta)$.

A corresponding normal approximation applies to any monotone transformation of $\hat{\theta}$ by the delta method, as seen before.

Back to our tests:
1. Wald test
2. Score test (LMP test)
3. Generalised LR test.

A normal approximation for the Wald test follows immediately from (5.3).

5.3 Normal approximation for the LMP/score test

Test $H_0 : \theta = \theta_0$ against $H_1^+ : \theta > \theta_0$ (where θ is a scalar:) We reject H_0 if $\ell'(\theta)$ is large (in contrast, for H_0 versus $H_1^- : \theta < \theta_0$, small values of $\ell'(\theta)$ would indicate H_1^-). The score test statistic is $\ell'(\theta)/\sqrt{I_n(\theta)}$. From (5.2) we obtain immediately that

$$\ell'(\theta)/\sqrt{I_n(\theta)} \approx \mathcal{N}(0, 1).$$

To find an (approximate) rejection region for the test: use the normal approximation at $\theta = \theta_0$, since the rejection region is calculated under the assumption that H_0 is true.

5.4 Chi-square approximation for the generalised likelihood ratio test

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, where θ is scalar. Reject H_0 if $L(\hat{\theta}; \mathbf{X})/L(\theta; \mathbf{X})$ is large; equivalently, reject for large

$$2 \log LR = 2[\ell(\hat{\theta}) - \ell(\theta)].$$

We use Taylor expansion around $\hat{\theta}$:

$$\ell(\hat{\theta}) - \ell(\theta) \approx -(\theta - \hat{\theta})\ell'(\hat{\theta}) - \tfrac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}).$$

Setting $\ell'(\hat{\theta}) = 0$, we obtain

$$\ell(\hat{\theta}) - \ell(\theta) \approx -\tfrac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}).$$

By the consistency of $\hat{\theta}$, we may approximate

$$\ell''(\hat{\theta}) \approx -I_n(\theta)$$

to get

$$2[\ell(\hat{\theta}) - \ell(\theta)] \approx (\theta - \hat{\theta})^2 I_n(\theta) = \left(\frac{\theta - \hat{\theta}}{\sqrt{I_n^{-1}(\theta)}} \right)^2 .$$

From (5.2), the asymptotic normality of $\hat{\theta}$, and as χ_1^2 variable is the square of a $N(0, 1)$ variable, we obtain that

$$2 \log LR = 2[\ell(\hat{\theta}) - \ell(\theta)] \approx \chi_1^2.$$

We can calculate a rejection region for the test of H_0 versus H_1 under this approximation.

For $\theta = (\theta_1, \ldots, \theta_p)$, testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, the dimension of the normal limit for $\hat{\theta}$ is p, hence the degrees of freedom of the related chi-squared variables are also p:

$$\ell'(\theta)^T [I_n(\theta)]^{-1} \ell'(\theta) \approx \chi_p^2$$

and

$$2 \log LR = 2[\ell(\hat{\theta}) - \ell(\theta)] \approx \chi_p^2.$$

5.5 Profile likelihood

Now $\theta = (\psi, \lambda)$, and $\hat{\lambda}_\psi$ is the MLE of λ when ψ fixed. Recall that the profile log-likelihood is given by $\ell_P(\psi) = \ell(\psi, \hat{\lambda}_\psi)$.

5.5.1 One-sided score test

We test $H_0 : \psi = \psi_0$ against $H_1^+ : \psi > \psi_0$; we reject H_0 based on large values of the score function $T = \ell'_\psi(\psi, \hat{\lambda}_\psi)$. Again T has approximate mean zero. For the **approximate variance** of T, we expand

$$T \approx \ell'_\psi(\psi, \lambda) + (\hat{\lambda}_\psi - \lambda)\ell''_{\psi,\lambda}(\psi, \lambda).$$

From (5.1),

$$\hat{\theta} - \theta \approx I_n^{-1}\ell'.$$

We write this as

$$\begin{pmatrix} \hat{\psi} - \psi \\ \hat{\lambda} - \lambda \end{pmatrix} \approx \begin{pmatrix} I_{\psi,\psi} & I_{\psi,\lambda} \\ I_{\psi,\lambda} & I_{\lambda,\lambda} \end{pmatrix}^{-1} \begin{pmatrix} \ell'_\psi \\ \ell'_\lambda \end{pmatrix}.$$

Here $\ell'_\psi = \partial\ell/\partial\psi$, $\ell'_\lambda = \partial\ell/\partial\lambda$, $\ell''_{\psi,\lambda} = \partial^2\ell/\partial\psi\partial\lambda$, $I_{\psi,\psi} = \mathsf{E}[-\ell''_{\psi,\psi}]$ etc. Now substitute $\hat{\lambda}_\psi - \lambda \approx I^{-1}_{\lambda,\lambda}\ell'_\lambda$ and put

$$\ell''_{\psi,\lambda} \approx -I_{\psi,\lambda}.$$

Calculate

$$V(T) \approx I_{\psi,\psi} + (I^{-1}_{\lambda,\lambda})^2 I^2_{\psi,\lambda} I_{\lambda,\lambda} - 2I^{-1}_{\lambda,\lambda} I_{\psi,\lambda} I_{\psi,\lambda}$$

to get

$$T \approx \ell'_\psi - I^{-1}_{\lambda,\lambda} I_{\psi,\lambda}\ell'_\lambda \approx N(0, 1/I_n^{\psi,\psi}),$$

where $I_n^{\psi,\psi} = (I_{\psi,\psi} - I^2_{\psi,\lambda} I^{-1}_{\lambda,\lambda})^{-1}$ is the top left element of I_n^{-1}. Estimate the Fisher information by substituting the null hypothesis values. Finally calculate the practical standardized form of T as

$$Z = \frac{T}{\sqrt{\mathrm{Var}(T)}} \approx \ell'_\psi(\psi, \hat{\lambda}_\psi)[I_n^{\psi,\psi}(\psi, \hat{\lambda}_\psi)]^{1/2} \approx N(0, 1).$$

Similar results for vector-valued ψ and vector-valued λ hold, with obvious modifications, *provided that the dimension of λ is fixed* (i.e., independent of the sample size n).

5.5.2 Two-sided likelihood ratio tests

Assume that ψ and λ are scalars. We use similar arguments as above, including Taylor expansion, for

$$2 \log LR \;\; = \;\; 2 \left[\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi) \right]$$

to obtain

$$2 \log LR \approx (\hat{\psi} - \psi)^2 / I_n^{\psi,\psi} \approx \chi_1^2, \tag{5.4}$$

where $I_n^{\psi,\psi} = (I_{\psi,\psi} - I_{\psi,\lambda}^2 I_{\lambda,\lambda}^{-1})^{-1}$ is the top left element of I_n^{-1}. The chi-squared approximation above follows from $\hat{\psi} - \psi \approx$ normal.

(Details can be found in the additional material at the end of this section.)

In general, if θ is p-dimensional, then $2 \log LR \approx \chi_p^2$.

Note: This result applies to the comparison of nested models, i.e., where one model is a special case of the other, but it does *not* apply to the comparison of non-nested models.

5.6 Connections with deviance

In GLM's, the *deviance* is usually $2 \log LR$ for two nested models, one the saturated model with a separate parameter for every response and the other the GLM (linear regression, log-linear model, etc.) For normal linear models the deviance equals the RSS. The general chi-squared result above need not apply to the deviance, because λ has dimension $n-p$ where p is the dimension of the GLM.

But the result does apply to deviance differences: Compare the GLM fit with p parameters (comprising $\theta = (\psi, \lambda)$) to a special case with only q ($< p$) parameters (i.e., with ψ omitted), then $2 \log LR$ for that comparison is the deviance difference, and in the null case (special case correct) $\approx \chi_{p-q}^2$.

5.7 Confidence regions

We can construct confidence regions based on the asymptotic normal distributions of the score statistic and the MLE, or on the chi-square approximation to the likelihood ratio statistic, or to the profile likelihood ratio statistic. These are equivalent in the limit $n \to \infty$, but they may display slightly different behaviour for finite samples.

Example: Wald-type interval. Based on the asymptotic normality of a p-dimensional $\hat{\theta}$, an approximate $1-\alpha$ confidence region is

$$\{\theta : (\hat{\theta} - \theta)^T I_n(\hat{\theta})(\hat{\theta} - \theta) \leq \chi^2_{p,1-\alpha}\}.$$

As an alternative to using $I_n(\hat{\theta})$ we could use $J_n(\hat{\theta})$, the *observed information* or *observed precision* evaluated at $\hat{\theta}$, where $[J_n(\theta)]_{jk} = -\partial^2\ell/\partial\theta_j\partial\theta_k$.

An advantage of the first type of region is that all values of θ inside the confidence region have higher likelihood than all values of θ outside the region.

Example: normal sample, known variance. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d, with σ^2 known. The log LR difference is

$$\begin{aligned} \ell(\hat{\mu}; \mathbf{x}) - \ell(\mu; \mathbf{x}) &= -\frac{1}{2\sigma^2}\left[\sum (x_i - \overline{x})^2 - \sum (x_i - \mu)^2\right] \\ &= \frac{n(\overline{x} - \mu)^2}{2\sigma^2}, \end{aligned}$$

so an approximate confidence interval is given by the values of μ satisfying

$$\frac{n(\overline{X} - \mu)^2}{2\sigma^2} \leq \tfrac{1}{2}\chi^2_{1,1-\alpha} \text{ or } \left|\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{1-\alpha/2},$$

which gives the same interval as in Chapter 4. In this case the approximate χ^2 result is, in fact, exact.

5.8 Additional material: Derivation of (5.4)

Assume ψ and λ scalars, then

$$\begin{pmatrix} \hat{\psi} - \psi \\ \hat{\lambda} - \lambda \end{pmatrix} \approx \begin{pmatrix} I_{\psi,\psi} & I_{\psi,\lambda} \\ I_{\psi,\lambda} & I_{\lambda,\lambda} \end{pmatrix}^{-1} \begin{pmatrix} \ell'_\psi \\ \ell'_\lambda \end{pmatrix}.$$

Similarly we have that

$$\hat{\lambda}_\psi - \lambda \approx I^{-1}_{\lambda,\lambda}\ell'_\lambda.$$

As

$$\ell'_\lambda \approx I_{\psi,\lambda}(\hat{\psi} - \psi) + I_{\lambda,\lambda}(\hat{\lambda} - \lambda),$$

we obtain

$$\hat{\lambda}_\psi - \lambda \;\; \approx \;\; \hat{\lambda} - \lambda + I_{\psi,\lambda}I^{-1}_{\lambda,\lambda}(\hat{\psi} - \psi).$$

Taylor expansion gives

$$\begin{aligned} 2\log LR &= 2\left[\ell(\hat{\psi},\hat{\lambda}) - \ell(\psi,\hat{\lambda}_\psi)\right] \\ &= 2\left[\ell(\hat{\psi},\hat{\lambda}) - \ell(\psi,\lambda)\right] - 2\left[\ell(\psi,\hat{\lambda}_\psi) - \ell(\psi,\lambda)\right] \\ &\approx (\psi - \hat{\psi}, \lambda - \hat{\lambda})I_n(\psi - \hat{\psi}, \lambda - \hat{\lambda})^T - (0, \lambda - \hat{\lambda}_\psi)I_n(0, \lambda - \hat{\lambda}_\psi)^T. \end{aligned}$$

Substituting for $\hat{\lambda}_\psi - \lambda$ gives

$$2\log LR \approx (\hat{\psi} - \psi)^2 / I_n^{\psi,\psi} \approx \chi^2_1,$$

where $I_n^{\psi,\psi} = (I_{\psi,\psi} - I^2_{\psi,\lambda}I^{-1}_{\lambda,\lambda})^{-1}$ is the top left element of I_n^{-1}. This is what we wanted to show.

Part II

Bayesian Statistics

Chapter 6

Bayesian Statistics: Background

Frequency interpretation of probability

In the frequency interpretation of probability, the probability of an event is limiting proportion of times the event occurs in an infinite sequence of independent repetitions of the experiment. This interpretation assumes that an experiment can be repeated!

Problems with this interpretation:

Independence is defined in terms of probabilities; if probabilities are defined in terms of independent events, this leads to a circular definition.

How can we check whether experiments were independent, without doing more experiments?

In practice we have only ever a finite number of experiments.

Subjective probability

Let $P(A)$ denote your personal probability of an event A; this is a numerical measure of the strength of your degree of belief that A will occur, in the light of available information.

Your personal probabilities may be associated with a much wider class of events than those to which the frequency interpretation pertains. For example:

- non-repeatable experiments (e.g. that England will win the World Cup next time)

- propositions about nature (e.g. that this surgical procedure results in increased life expectancy) .

All subjective probabilities are conditional, and may be revised in the light of additional information. Subjective probabilities are assessments in the light of incomplete information; they may even refer to events in the past.

Axiomatic development

Coherence:

Coherence states that a system of beliefs should avoid internal inconsistencies. Basically, a quantitative, coherent belief system must behave as if it was governed by a subjective probability distribution. In particular this assumes that all events of interest can be compared.

Note: different individuals may assign different probabilities to the same event, even if they have identical background information.

(See Chapters 2 and 3 in *Bernardo and Smith* for fuller treatment of foundational issues.)

Bayes Theorem

Let $B_1, B_2, \ldots, B_k$ be a set of mutually exclusive and exhaustive events. For any event A with $P(A) > 0$,

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}. \end{aligned}$$

Equivalently we write

$$P(B_i|A) \propto P(A|B_i)P(B_i).$$

Terminology

$P(B_i)$ is the *prior probability* of B_i;

$P(A|B_i)$ is the *likelihood* of A given B_i;

$P(B_i|A)$ is the *posterior probability* of B_i;

$P(A)$ is the *predictive probability* of A implied by the likelihoods and the prior probabilities.

Example: Two events. Assume we have two events B_1, B_2, then

$$\frac{P(B_1|A)}{P(B_2|A)} = \frac{P(B_1)}{P(B_2)} \times \frac{P(A|B_1)}{P(A|B_2)}.$$

If the data is relatively more probable under B_1 than under B_2, our belief in B_1 compared to B_2 is increased, and conversely.

If $B_2 = B_1^c$:

$$\frac{P(B_1|A)}{P(B_1^c|A)} = \frac{P(B_1)}{P(B_1^c)} \times \frac{P(A|B_1)}{P(A|B_1^c)}.$$

It follows that: **posterior odds = prior odds $\times$ likelihood ratio.**

Parametric models

A Bayesian statistical model consists of

1.) A parametric statistical model $f(x|\theta)$ for the data x, where $\theta \in \Theta$ a parameter; x may be multidimensional. - Note that we write $f(x|\theta)$ instead of $f(x, \theta)$ do emphasise the conditional character of the model. 2.) A prior distribution $\pi(\theta)$ on the parameter.

Note: The parameter θ is now treated as random!

The *posterior distribution* of θ given x is

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}$$

Shorter, we write: $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ or **posterior $\propto$ prior $\times$ likelihood** .

Nuisance parameters

Let $\theta = (\psi, \lambda)$, where λ is a nuisance parameter. Then $\pi(\theta|x) = \pi((\psi, \lambda)|x)$. We calculate the *marginal posterior* of ψ:

$$\pi(\psi|x) = \int \pi(\psi, \lambda|x)d\lambda$$

and continue inference with this marginal posterior. Thus we just integrate out the nuisance parameter.

Prediction

The *(prior) predictive distribution* of x on the basis π is

$$p(x) = \int f(x|\theta)\pi(\theta)d\theta.$$

Suppose data x_1 is available, and we want to predict additional data:

$$\begin{aligned} p(x_2|x_1) &= \frac{p(x_2, x_1)}{p(x_1)} \\ &= \frac{\int f(x_2, x_1|\theta)\pi(\theta)d\theta}{\int f(x_1|\theta)\pi(\theta)d\theta} \\ &= \int f(x_2|\theta, x_1)\frac{f(x_1|\theta)\pi(\theta)}{\int f(x_1|\theta)\pi(\theta)d\theta}d\theta \\ &= \int f(x_2|\theta)\frac{f(x_1|\theta)\pi(\theta)}{\int f(x_1|\theta)\pi(\theta)d\theta}d\theta \\ &= \int f(x_2|\theta)\pi(\theta|x_1)d\theta. \end{aligned}$$

Note that x_2 and x_1 are assumed conditionally independent given θ. They are **not**, in general, unconditionally independent.

Example

(*Bayes*) A billard ball W is rolled from left to right on a line of length 1 with a uniform probability of stopping anywhere on the line. It stops at p. A second ball O is then rolled n times under the same assumptions, and X denotes the number of times that the ball O stopped on the left of W. Given X, what can be said about p?

Our prior is $\theta(p) = 1$ for $0 \leq p \leq 1$; our model is

$$P(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, \ldots, n$

We calculate the predictive distribution, for $x = 0, \ldots, n$

$$\begin{aligned} P(X = x) &= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= \binom{n}{x} B(x+1, n-x+1) \\ &= \binom{n}{x} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{n+1}, \end{aligned}$$

where B is the beta function,

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

We calculate the posterior distribution:

$$\begin{aligned} \pi(p|x) &\propto 1 \times \binom{n}{x} p^x (1-p)^{n-x} \\ &\propto p^x (1-p)^{n-x}, \end{aligned}$$

so

$$\pi(p|x) = \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)};$$

this is the $Beta(x+1, n-x+1)$-distribution.

In particular the posterior mean is

$$\begin{aligned} E(p|x) &= \int_0^1 p \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)} dp \\ &= \frac{B(x+2, n-x+1)}{B(x+1, n-x+1)} \\ &= \frac{x+1}{n+2}. \end{aligned}$$

(For comparison: the mle is $\frac{x}{n}$.)

Further, $P(O$ stops to the left of W on the next roll $|x)$ is Bernoulli-distributed with probability of success $E(p|x) = \frac{x+1}{n+2}$.

Example: exponential model, exponential prior

Let $X_1, \ldots, X_n$ be a random sample with density $f(x|\theta) = \theta e^{-\theta x}$ for $x \geq 0$, and assume $\pi(\theta) = \mu e^{-\mu\theta}$ for $\theta \geq 0$; and some known μ. Then

$$f(x_1, \ldots, x_n|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}$$

and hence the posterior distribution is

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto \theta^n e^{-\theta \sum_{i=1}^n x_i} \mu e^{-\mu\theta} \\ &\propto \theta^n e^{-\theta(\sum_{i=1}^n x_i + \mu)}, \end{aligned}$$

which we recognize as $Gamma(n+1, \mu + \sum_{i=1}^n x_i)$.

Example: normal model, normal prior

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\theta, \sigma^2)$, where σ^2 is known, and assume that the prior is normal, $\pi(\theta) \sim \mathcal{N}(\mu, \tau^2)$, where μ, τ^2 is known. Then

$$f(x_1, \ldots, x_n|\theta) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{ -\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \theta)^2}{\sigma^2} \right\},$$

so we calculate for the posterior that

$$\begin{aligned} \pi(\theta|\mathbf{x}) &\propto exp\left\{ -\frac{1}{2} \left(\sum_{i=1}^n \frac{(x_i - \theta)^2}{\sigma^2} + \frac{(\theta - \mu)^2}{\tau^2} \right) \right\} \\ &=: e^{-\frac{1}{2}M}. \end{aligned}$$

We can calculate (Exercise)

$$\begin{aligned} M &= a\left(\theta - \frac{b}{a}\right)^2 - \frac{b^2}{a} + c, \\ a &= \frac{n}{\sigma^2} + \frac{1}{\tau^2}, \\ b &= \frac{1}{\sigma^2} \sum x_i + \frac{\mu}{\tau^2}, \\ c &= \frac{1}{\sigma^2} \sum x_i^2 + \frac{\mu^2}{\tau^2}. \end{aligned}$$

So it follows that the posterior is normal,

$$\pi(\theta|x) \sim \mathcal{N}\left(\frac{b}{a}, \frac{1}{a}\right)$$

Exercise: the predictive distribution for x is $\mathcal{N}(\mu, \sigma^2 + \tau^2)$.

Note: The posterior mean for θ is

$$\mu_1 = \frac{\frac{1}{\sigma^2}\sum x_i + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}.$$

If τ^2 is very large compared to σ^2, then the posterior mean is approximately $\overline{x}$. If σ^2/n is very large compared to τ^2, then the posterior mean is approximately μ.

The posterior variance for θ is

$$\phi = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} < \min\left\{\frac{\sigma^2}{n}, \tau^2\right\},$$

which is smaller than the original variances.

Credible intervals

A $(1-\alpha)$ *(posterior) credible interval* is an interval of $\theta-$values within which $1-\alpha$ of the posterior probability lies. In the above example:

$$P\left(-z_{\alpha/2} < \frac{\theta - \mu_1}{\sqrt{\phi}} < z_{\alpha/2}\right) = 1 - \alpha$$

is a $(1-\alpha)$ *(posterior) credible interval* for θ.

The equality is correct conditionally on x, but the rhs does not depend on x, so the equality is also unconditionally correct.

If $\tau^2 \to \infty$ then $\phi - \frac{\sigma^2}{n} \to 0$, and $\mu_1 - \overline{x} \to 0$, and

$$P\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \theta < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

which gives the usual $100(1-\alpha)\%$ confidence interval in frequentist statistics.

Note: The interpretation of credible intervals is different to confidence intervals: In frequentist statistics, $\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ applies before $\mathbf{x}$ is observed; the randomness relates to the distribution of $\mathbf{x}$, whereas in Bayesian statistics, the credible interval is conditional on the observed $\mathbf{x}$; the randomness relates to the distribution of θ.

Chapter 7

Bayesian Models

Sufficiency

A sufficient statistic captures all the useful information in the data.
Definition: A statistic $t = t(x_1, \ldots, x_n)$ is *parametric sufficient* for θ if

$$\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x})).$$

Note: then we also have

$$p(x_{new}|\mathbf{x}) = p(x_{new}|t(\mathbf{x})).$$

Factorization Theorem: $t(\mathbf{x})$ is parametric sufficient if and only if

$$\pi(\theta|\mathbf{x}) = \frac{h(t(\mathbf{x}), \theta)\pi(\theta)}{\int h(t(\mathbf{x}), \theta)\pi(\theta)d\theta}$$

for some function h.

Recall: For classical suffiency, the factorization theorem gave as necessary and sufficient condition that

$$f(\mathbf{x}, \theta) = h(t(\mathbf{x}), \theta)g(\mathbf{x}).$$

Theorem: Classical sufficiency is equivalent to parametric sufficiency.

To see this: assume classical sufficiency, then

$$\begin{aligned}\pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)d\theta} \\ &= \frac{h(t(\mathbf{x}),\theta)g(\mathbf{x})\pi(\theta)}{\int h(t(\mathbf{x}),\theta)g(\mathbf{x})\pi(\theta)d\theta} \\ &= \frac{h(t(\mathbf{x}),\theta)\pi(\theta)}{\int h(t(\mathbf{x}),\theta)\pi(\theta)d\theta}\end{aligned}$$

depends on $\mathbf{x}$ only through $t(\mathbf{x})$, so $\pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x}))$.

Conversely, assume parametric sufficiency, then

$$\frac{f(\mathbf{x}|\theta)\pi(\theta)}{f(\mathbf{x})} = \pi(\theta|\mathbf{x}) = \pi(\theta|t(\mathbf{x})) = \frac{f(t(\mathbf{x})|\theta)\pi(\theta)}{f(t(\mathbf{x}))}$$

and so

$$f(\mathbf{x}|\theta) = \frac{f(t(\mathbf{x})|\theta)}{f(t(\mathbf{x}))} f(\mathbf{x})$$

which implies classical sufficiency.

Example: A *k-parameter exponential family* is given by

$$f(x|\theta) = \exp\left\{\sum_{i=1}^{k} c_i\phi_i(\theta)h_i(x) + c(\theta) + d(x)\right\}, \quad x \in \mathcal{X}$$

where

$$e^{-c(\theta)} = \int exp\left\{\sum_{i=1}^{k} c_i\phi_i(\theta)h_i(x) + d(x)\right\} dx < \infty.$$

The family is called *regular* if $\mathcal{X}$ does not depend on θ; otherwise it is called *non-regular*.

Fact: In k-parameter exponential family models,

$$t(\mathbf{x}) = (n, \sum_{j=1}^{n} h_1(x_j), \ldots, \sum_{j=1}^{n} h_k(x_j))$$

is sufficient.

Exchangeability

$X_1, \ldots, X_n$ are *(finitely) exchangeable* if

$$P(X_1 \in E_1, \ldots, X_n \in E_n) \;\; = \;\; P(X_{\sigma(i)} \in E_1, \ldots, X_{\sigma(n)} \in E_n)$$

for any permutation σ of $\{1, 2, \ldots, n\}$, and any (measureable) sets $E_1, \ldots, E_n$. An infinite sequence $X_1, X_2, \ldots$ is *exchangeable* if every finite sequence is (finitely) exchangeable

Intuitively: a random sequence is exchangeable if the random quantities do not arise, for example, in a time ordered way.

Every independent sequence is exchangeable, but NOT every exchangeable sequence is independent.

Example: A simple random sample from a finite population (sampling without replacement) is exchangeable, but not independent.

Theorem (de Finetti). If $X_1, X_2, \ldots$ is *exchangeable*, with probability measure P, then there exists a prior measure Q on the set of all distributions (on the real line) such that, for any n, the joint distribution function of $X_1, \ldots X_n$ has the form

$$\int \prod_{i=1}^{n} F(x_i) dQ(F),$$

where $F_1, \ldots, F_n$ are distributions, and

$$Q(E) = \lim_{n \to \infty} P(F_n \in E),$$

where $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$ is the empirical c.d.f..

Thus each exchangeable sequence arises from a 2-stage randomization:
(a) pick F according to Q;
(b) conditional on F, the observations are i.i.d. .

De Finetti's Theorem tells us that subjective beliefs which are consistent with (infinite) exchangeability must be of the form
(a) There are beliefs (a priori) on the "parameter" F; representing your expectations for the behaviour of $X_1, X_2, \ldots$;

(b) conditional on F the observations are i.i.d. .

In Bayesian statistics, we can think of the prior distribution on the parameter θ as such an F. If a sample is i.i.d. given θ, then the sample is exchangeable.

Example. Suppose $X_1, X_2, \ldots$ are exchangeable 0-1 variables. Then the distribution of X_i is uniquely defined by $p = P(X_i = 1)$; the set of all probability distributions on $\{0, 1\}$ is equivalent to the interval $[0, 1]$. The measure Q puts a probability on $[0, 1]$; de Finetti gives

$$p(x_1, \ldots, x_n) = \int p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i} dQ(p).$$

Not all finitely exchangeable sequences can be imbedded in an infinite exchangeable sequence.

Exercise: X_1, X_2 such that

$$P(X_i = 1, X_2 = 0) = P(X_1 = 0, X_2 = 1) = \frac{1}{2}$$

cannot be embedded in an exchangeable sequence X_1, X_2, X_3.

For further reading on de Finetti's Theorem, see Steffen Lauritzen's graduate lecture at `www.stats.ox.ac.uk/~steffen/teaching/grad/definetti.pdf`.

Chapter 8

Prior Distributions

Let Θ be a parameter space. How do we assign prior probabilities on Θ? Recall: we need a coherent belief system.

In a *discrete parameter space* we assign subjective probabilities to each element of the parameter space, which is in principle straightforward.

In a continuous parameter space:

1. Histogram approach:
Say, Θ is an interval of $\mathbf{R}$. We can discretize Θ, assess the total mass assigned to each subinterval, and smoothen the histogram.

2. Relative likelihood approach:
Again, say, Θ is an interval of $\mathbf{R}$. We assess the relative likelihood that θ will take specific values. This relative likelihood is proportional to the prior density. If Θ is unbounded, normalization can be an issue.

3. Particular functional forms: conjugate priors. A family $\mathcal{F}$ of prior distributions for θ is *closed under sampling* from a model $f(x|\theta)$ if for every prior distribution $\pi(\theta) \in \mathcal{F}$, the posterior $\pi(\theta|x)$ is also in $\mathcal{F}$.

When this happens, the common parametric form of the prior and posterior are called a *conjugate prior family* for the problem. Then we also say that the family $\mathcal{F}$ of prior distributions is *conjugate* to this class of models $\{f(x|\theta), \theta \in \Theta\}$.

Often we abuse notation and call an element in the family of conjugate priors a *conjugate prior* itself.

Example: We have seen already: if $X \sim \mathcal{N}(\theta, \sigma^2)$ with σ^2 known; and if $\theta \sim \mathcal{N}(\mu, \tau^2)$, then the posterior for θ is also normal. Thus the family of normal distributions forms a conjugate prior family for this normal model.

Example: Regular k-parameter exponential family. If

$$f(x|\theta) = \exp\{\sum_{i=1}^{k} c_i \phi_i(\theta) h_i(x) + c(\theta) + d(x)\}, \quad x \in \mathcal{X}$$

then the family of priors of the form

$$\pi(\theta|\tau) = (K(\tau))^{-1} exp\{\sum_{i=1}^{k} c_i \tau_i \phi_i(\theta) + \tau_0 c(\theta)\},$$

where $\tau = (\tau_0, \ldots, \tau_k)$ is such that

$$K(\tau) = \int_{\Theta} e^{\tau_0 c(\theta)} exp\{\sum_{i=1}^{k} c_i \tau_i \phi_i(\theta)\} d\theta < \infty,$$

is a conjugate prior family. The parameters τ are called *hyperparameters*.

Example: Bernoulli distribution: Beta prior

$$f(x|\theta) = \theta^x (1-\theta)^{1-x} = (1-\theta) exp\left\{x \log\left(\frac{\theta}{1-\theta}\right)\right\}$$

so $k = 1$ and

$$\begin{aligned} d(x) &= 0, \quad c(\theta) = \log(1-\theta), \quad h_1(x) = x, \\ \phi_1(\theta) &= \log\left(\frac{\theta}{1-\theta}\right), \\ \pi(\theta|\tau) &\propto (1-\theta)^{\tau_0} exp\left\{\tau_1 \log\left(\frac{\theta}{1-\theta}\right)\right\} = \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1} (1-\theta)^{\tau_0 - \tau_1}. \end{aligned}$$

This density will have a finite integral if and only if $\tau_1 > -1$ and $\tau_0 - \tau_1 > -1$, in which case it is the $Beta(\tau_1 + 1, \tau_0 - \tau_1 + 1)$-distribution. Thus the family of Beta distributions forms a conjugate prior family for the Bernoulli distribution.

Example: Poisson distribution: Gamma prior. Here again $k = 1$,

$$\begin{aligned} d(x) &= -\log((x!)), \quad c(\theta) = -\theta \\ h_1(x) &= x, \quad \phi_1(\theta) = \log\theta, \end{aligned}$$

and an element of the family of conjugate priors is given by

$$\pi(\theta|\tau) = \frac{1}{K(\tau_0, \tau_1)} \theta^{\tau_1} e^{-\theta\tau_0}.$$

The density will have a finite integral if and only if $\tau_1 > -1$ and $\tau_0 > 0$, in which case it is the $Gamma(\tau_1 + 1, \tau_0)$ distribution.

Example: Normal, unknown variance: normal-gamma prior. For the normal distribution with mean μ, we let the *precision* be $\lambda = \sigma^{-2}$, then $\theta = (\mu, \lambda)$

$$\begin{aligned} d(x) &= -\frac{1}{2}\log(2\pi), \quad c(\mu, \lambda) = -\frac{\lambda\mu^2}{2} + \log(\sqrt{\lambda}) \\ (h_1(x), h_2(x)) &= (x, x^2), \quad (\phi_1(\mu, \lambda), \phi_2(\mu, \lambda)) = (\mu\lambda, -\frac{1}{2}\lambda) \end{aligned}$$

and an element of the family of conjugate priors is given by

$$\begin{aligned} &\pi(\mu, \lambda|\tau_0, \tau_1, \tau_2) \\ &\quad \propto \lambda^{\frac{\tau_0}{2}} exp\left\{-\frac{1}{2}\lambda\mu^2\tau_0 + \lambda\mu\tau_1 - \frac{1}{2}\lambda\tau_2\right\} \\ &\quad \propto \lambda^{\frac{\tau_0+1}{2}-1} exp\left\{-\frac{1}{2}\left(\tau_2 - \frac{\tau_1^2}{\tau_0}\right)\lambda\right\} \lambda^{\frac{1}{2}} exp\left\{-\frac{\lambda\tau_0}{2}\left(\mu - \frac{\tau_1}{\tau_0}\right)^2\right\}. \end{aligned}$$

The density can be interpreted as follows: Use a $Gamma((\tau_0 + 1)/2, (\tau_2 - \tau_1^2/\tau_0)/2)$ prior for λ. Conditional on λ, we have a normal $\mathcal{N}(\tau_1/\tau_0, 1/(\lambda\tau_0))$ for μ. This is called a *normal-gamma distribution* for (μ, λ); it will have a finite integral if and only if $\tau_2 > \tau_1^2/\tau_0$ and $\tau_0 > 0$.

Fact: In **regular k-parameter exponential family models**, the family of conjugate priors is closed under sampling. Moreover, if $\pi(\theta|\tau_0, \ldots, \tau_k)$ is in the above conjugate prior family, then

$$\pi(\theta|x_1, \ldots, x_n, \tau_0, \ldots, \tau_k) = \pi(\theta|\tau_0 + n, \tau_1 + H_1(x), \ldots, \tau_k + H_k(x)),$$

where

$$H_i(x) = \sum_{j=1}^{n} h_i(x_j).$$

Recall that $(n, H_1(x), \ldots, H_k(x))$ is a sufficient statistic. Note that indeed the posterior is of the same parametric form as the prior.

The predictive density for future observations $\mathbf{y} = y_1, \ldots, y_m$ is

$$\begin{aligned} & p(\mathbf{y}|x_1, \ldots, x_n, \tau_0, \ldots, \tau_k) \\ &= p(\mathbf{y}|\tau_0 + n, \tau_1 + H_1(x), \ldots, \tau_k + H_k(x)) \\ &= \frac{K(\tau_0 + n + m, \tau_1 + H_1(x, y), \ldots, \tau_k + H_k(x, y))}{K(\tau_0 + n, \tau_1 + H_1(x), \ldots, \tau_k + H_k(x))} e^{\sum_{\ell-1}^{m} d(y_\ell)}, \end{aligned}$$

where

$$H_i(x, y) = \sum_{j=1}^{n} h_i(x_j) + \sum_{\ell=1}^{m} h_i(y_\ell).$$

This form is particularly helpful for inference: The effect of the data $x_1, \ldots, x_n$ is that the labelling parameters of the posterior are changed from those of the prior, $(\tau_0, \ldots, \tau_k)$ by simply adding the sufficient statistics

$$(t_0, \ldots, t_k) = (n, \sum_{i=1}^{n} h_1(x_j), \ldots, \sum_{i=1}^{n} h_k(x_j))$$

to give the parameters $(\tau_0 + t_0, \ldots, \tau_k + t_k)$ for the posterior.

Mixtures of priors from this conjugate prior family also lead to a simple analysis (Exercise).

Noninformative priors

Often we would like a prior that favours no particular values of the parameter over others. If Θ finite with $|\Theta| = n$, then we just put mass $\frac{1}{n}$ at each parameter value. If Θ is infinite, there are several ways in which one may seek a noninformative prior.

Improper priors are priors which do not integrate to 1. They are interpreted in the sense that posterior $\propto$ prior $\times$ likelihood. Often they arise as

natural limits of proper priors. They have to be handled carefully in order not to create paradoxes!

Example: Binomial, Haldane's prior. Let $X \sim Bin(n,p)$, with n fixed, and let π be a prior on p. *Haldane's prior* is

$$\pi(p) \propto \frac{1}{p(1-p)};$$

we have that $\int_0^1 \pi(p)\,dp = \infty$. This prior gives as marginal density

$$p(x) \propto \int_0^1 (p(1-p))^{-1} \binom{n}{x} p^x (1-p)^{n-x} dp,$$

which is not defined for $x = 0$ or $x = n$. For all other values of x we obtain the $Beta(x, n-x)$-distribution. The posterior mean is

$$\frac{x}{x+n-x} = \frac{x}{n}.$$

For $x = 0, n$: we could think of

$$\pi \approx \pi_{\alpha,\beta} \sim Beta(\alpha, \beta),$$

where $\alpha, \beta > 0$ are small. Then the posterior is $Beta(\alpha + x, \beta + n - x)$. Now let $\alpha, \beta \downarrow 0$: then $\pi_{\alpha,\beta}$ converges to π. Note that the posterior mean converges to x/n for *all* values of x.

Noninformative priors for location parameters

Let $\Theta, \mathcal{X}$ be subsets of Euclidean space. Suppose that $f(x|\theta)$ is of the form $f(x-\theta)$: this is called a *location density*, θ is called *location parameter*.

For a noninformative prior for θ: suppose that we observe $y = x + c$, where c fixed. If $\eta = \theta + c$ then y has density $f(y|\eta) = f(y - \eta)$, and so a noninformative priors should be the same (as we assume that we have the same parameter space).

Call π the prior for the (x, θ)-problem, and π^* the prior for the (y, η)-problem. Then we want that for any (measureable) set A

$$\int_A \pi(\theta)d\theta \;=\; \int_A \pi^*(\theta)d\theta = \int_{A-c} \pi(\theta)d\theta = \int_A \pi(\theta - c)d\theta,$$

yielding

$$\pi(\theta) = \pi(\theta - c)$$

for all θ. Hence $\pi(\theta) = \pi(0)$ constant; usually we choose $\pi(\theta) = 1$ for all θ. This is an improper prior.

Noninformative priors for scale parameters

A (one-dimensional) *scale density* is a density of the form

$$f(x|\sigma) = \frac{1}{\sigma} f\left(\frac{x}{\sigma}\right)$$

for $\sigma > 0$; σ is called *scale parameter*. Consider $y = cx$, for $c > 0$; put $\eta = c\sigma$, then y has density $f(y|\eta) = \frac{1}{\eta} f\left(\frac{x}{\sigma}\right)$. Noninformative priors for σ and η should be the same (assuming the same parameter space), so we want that for any (measureable) set A

$$\int_A \pi(\sigma)d\sigma \;\; = \;\; \int_{c^{-1}A} \pi(\sigma)d\sigma = \int_A \pi(c^{-1}\sigma)c^{-1}d\sigma,$$

yielding

$$\pi(\sigma) = c^{-1}\pi(c^{-1}\sigma)$$

for all $\sigma > 0$; $\pi(c) = c^{-1}\pi(1)$; hence $\pi(\sigma) \propto \frac{1}{\sigma}$. Usually we choose $\pi(\sigma) = \frac{1}{\sigma}$. This is an improper prior.

Jeffreys Priors

Example: Binomial model. Let $X \sim Bin(n, p)$.A plausible noninformative prior would be $p \sim U[0, 1]$, but then $\sqrt{p}$ has higher density near 1 than near 0. Thus "ignorance" about p seems to lead to "knowledge" about $\sqrt{p}$, which is paradoxical. We would like the prior to be invariant under reparametrization.

Recall: under regularity conditions, the *Fisher information* equals

$$I(\theta) - E_\theta\left(\frac{\partial^2 log f(x|\theta)}{\partial^2 \theta}\right).$$

Under the same regularity assumptions, we define the *Jeffreys prior* as

$$\pi(\theta) \propto I(\theta)^{\frac{1}{2}}.$$

Here $I(\theta) = I_1(\theta)$. This prior may or may not be improper.

Reparametrization: Let h be monotone and differentiable. The chain rule gives

$$I(\theta) = I(h(\theta)) \left(\frac{\partial h}{\partial \theta} \right)^2 .$$

For the Jeffreys prior $\pi(\theta)$, we have

$$\pi(h(\theta)) \;\; = \;\; \pi(\theta) \left| \frac{\partial h}{\partial \theta} \right|^{-1} \propto I(\theta)^{\frac{1}{2}} \left| \frac{\partial h}{\partial \theta} \right|^{-1} = I(h(\theta))^{\frac{1}{2}};$$

thus the prior is indeed invariant under reparametrization.

The Jeffreys prior favours values of θ for which $I(\theta)$ is large. Hence minimizes the effect of the prior distribution relative to the information in the data.

Exercise: The above non-informative priors for scale and location correspond to Jeffreys priors.

Example: Binomial model, Jeffreys prior. Let $X \sim Bin(n, p)$; where n is known, so that $f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$. Then

$$\frac{\partial^2 log f(x|p)}{\partial^2 p} \;\; = \;\; -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

Take expectation and multiply by minus 1: $I(p) = \frac{n}{p} + \frac{n}{(1-p)} = \frac{n}{p(1-p)}$ Thus the Jeffreys prior is

$$\pi(p) \propto (p(1-p))^{-\frac{1}{2}},$$

which we recognize as the $Beta(1/2, 1/2)$-distribution. This is a proper prior.

If θ is multivariate, the Jeffreys prior is

$$\pi(\theta) \propto [detI(\theta)]^{\frac{1}{2}},$$

which is still invariant under reparametrization.

Example: Normal model, Jeffreys prior. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma)$. We abbreviate

$$\psi(x, \mu, \sigma) = -\log \sigma - \frac{(x-\mu)^2}{2\sigma^2};$$

then

$$\begin{aligned} I(\theta) &= -E_\theta \begin{pmatrix} \frac{\partial^2}{\partial\mu^2}\psi(x,\mu,\sigma) & \frac{\partial^2}{\partial\mu\partial\sigma}\psi(x,\mu,\sigma) \\ \frac{\partial^2}{\partial\mu\partial\sigma}\psi(x,\mu,\sigma) & \frac{\partial^2}{\partial\sigma^2}\psi(x,\mu,\sigma) \end{pmatrix} \\ &= -E_\theta \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{2(x-\mu)}{\sigma^3} \\ -\frac{2(x-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \end{aligned}$$

So

$$\pi(\theta) \propto \left(\frac{1}{\sigma^2} \times \frac{2}{\sigma^2}\right)^{\frac{1}{2}} \propto \frac{1}{\sigma^2}$$

is the Jeffreys prior.

Note: $\mathcal{N}(\mu, \sigma^2)$ is a location-scale density, so we could take a uniform prior for μ, $1/\sigma$-prior for σ; this would yield

$$\pi(\theta) = \frac{1}{\sigma}.$$

Thie prior is *not* equal to the Jeffreys prior.

Maximum Entropy Priors

Assume first that Θ is discrete. The *entropy* of π is defined as

$$\mathcal{E}(\pi) = -\sum_{\Theta} \pi(\theta_i) log(\pi(\theta_i))$$

(where $0log(0) = 0$). It measures the amount of uncertainty in an observation.

If Θ is finite, with n elements, then $\mathcal{E}(\pi)$ is largest for the uniform distribution, and smallest if $\pi(\theta_i) = 1$ for some $\theta_i \in \Theta$.

Suppose we are looking for a prior π, taking partial information in terms of functions $g_1, \ldots, g_m$ into account, where this partial information can be written as

$$E_\pi g_k(\theta) = \mu_k, \quad k = 1, \ldots, m.$$

We would like to choose the distribution with the maximum entropy under these constraints: This distribution is

$$\tilde{\pi}(\theta_i) = \frac{exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_i)\right)}{\sum_i exp\left(\sum_{k=1}^m \lambda_k g_k(\theta_i)\right)},$$

where the λ_i are determined by the constraints.

Example: prior information on the mean. Let $\Theta = \{0, 1, 2, \ldots\}$. The prior mean of θ is thought to be 5. We write this as a constraint: $m = 1, g_1(\theta) = \theta, \mu_1 = 5$. So

$$\tilde{\pi}(\theta) \;=\; \frac{e^{\lambda_1 \theta}}{\sum_{j=0}^{\infty} e^{\lambda_1 j}} = \left(e^{\lambda_1}\right)^{\theta} \left(1 - e^{\lambda_1}\right)$$

is the maximum-entropy prior. We recognize it as the distribution of (geometric - 1). Its mean is $e^{-\lambda_1} - 1$, so setting the mean equal to 5 yields $e^{\lambda_1} = \frac{1}{6}$, and $\lambda_1 = -\log 6$.

If Θ *is continuous*, then $\pi(\theta)$ is a density. The *entropy* of π *relative* to a particular reference distribution with density π_0 is defined as

$$\mathcal{E}(\pi) \;=\; -E_\pi \left(\log \frac{\pi(\theta)}{\pi_0(\theta)} \right) = - \int_\Theta \pi(\theta) log \frac{\pi(\theta)}{\pi_0(\theta)} d\theta.$$

The case that Θ is discrete corresponds to π_0 being discrete uniform.

How do we choose π_0? We would like to choose the "natural" invariant noninformative prior. Assume that we have partial information in terms of functions $g_1, \ldots, g_m$:

$$\int g_k(\theta) \pi(\theta) d\theta = \mu_k, \quad k = 1, \ldots, m.$$

We choose the distribution with the maximum entropy under these constraints (when it exists):

$$\tilde{\pi}(\theta) = \frac{\pi_0(\theta) exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right)}{\int_\Theta \pi_0(\theta) exp\left(\sum_{k=1}^m \lambda_k g_k(\theta)\right) d\theta},$$

where the λ_i are determined by the constraints.

Example: location parameter, known mean, known variance. Let $\Theta = \mathbf{R}$, and let θ be location parameter; we choose as reference prior $\pi_0(\theta) = 1$. Suppose that mean and variance are known:

$$g_1(\theta) = \theta, \mu_1 = \mu; \quad g_2(\theta) = (\theta - \mu)^2, \mu_2 = \sigma^2.$$

Then we choose

$$\begin{aligned}\tilde{\pi}(\theta) &= \frac{exp\,(\lambda_1\theta + \lambda_2(\theta-\mu)^2)}{\int_{-\infty}^{\infty} exp\,(\lambda_1\theta + \lambda_2(\theta-\mu)^2)\,d\theta} \\ &\propto exp(\lambda_1\theta + \lambda_2\theta^2) \propto exp(\lambda_2(\theta-\alpha)^2),\end{aligned}$$

for a suitable α (here the λ's may not be the same). So $\tilde{\pi}$ is normal; the constraints give $\tilde{\pi}$ is $\mathcal{N}(\mu, \sigma^2)$.

Example: location parameter, known mean. Suppose that in the previous example only the prior mean, not the prior variance, is specified; so

$$\tilde{\pi}(\theta) = \frac{exp\,(\lambda_1\theta)}{\int_{-\infty}^{\infty} exp\,(\lambda_1\theta)\,d\theta},$$

and the integral is infinite, so the distribution does not exist.

Additional Material: Bayesian Robustness

To check how much the conclusions change for different priors, we can carry out a sensitivity analysis.

Example: Normal or Cauchy?
Suppose $\Theta = \mathbf{R}$, and $X \sim \mathcal{N}(\theta, 1)$, where θ is known to be either standard normal or standard Cauchy. We calculate the posterior means under both models:

obs. x	post. mean ($\mathcal{N}$)	post. mean (C)
0	0	0
1	0.69	0.55
2	1.37	1.28
4.5	3.09	4.01
10	6.87	9.80

For small x the posterior mean does not change very much, but for large x it does.

Example: Normal model; normal or contaminated class of priors. Assume that $X \sim \mathcal{N}(\theta, \sigma^2)$, where σ^2 known, and $\pi_0 \sim \mathcal{N}(\mu, \tau^2)$. An alternative class of priors is Γ, constructed as follows: Let

$$Q = \{q_k; q_k \sim \mathcal{U}|(\mu-k, \mu+k)\},$$

then put

$$\Gamma = \{\pi : \pi = (1-\epsilon)\pi_0 + \epsilon q, \text{ some } q \in Q\}.$$

The Γ is called an *ϵ-contamination class of priors*. Let $C = (c_1, c_2)$ be an interval. Put $P_0 = P(\theta \in C|x, \pi_0)$ and $Q_k = P(\theta \in C|x, q_k)$. Then (by Bayes' rule) for $\pi \in \Gamma$ we have

$$P(\theta \in C|x) \;\; = \;\; \lambda_k(x)P_0 + (1-\lambda_k(x))Q_k,$$

where

$$\lambda_k(x) = \left(1 + \frac{\epsilon}{1-\epsilon} \times \frac{p(x|q_k)}{p(x|\pi_0)}\right)^{-1},$$

and $p(x|q)$ is the predictive density of x when the prior is q.

The predictive density $p(x|\pi_0)$ is $\mathcal{N}(\mu, \sigma^2 + \tau^2)$,

$$p(x|q_k) = \int_{\mu-k}^{\mu+k} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)\frac{1}{2k}d\theta$$

and

$$Q_k = \frac{1}{p(x|q_k)}\int_{c^*}^{c^{**}} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)\frac{1}{2k}d\theta$$

where $c^* = \max\{c, \mu - k\}$ and $c^{**} = \min\{c_2, \mu + k\}$ (ϕ is the standard normal density).

Numerical example: Let $\epsilon = 0.1, \sigma^2 = 1, \tau^2 = 2, \mu = 0, x = 1$, and $C = (-0.93, 2.27)$ is the 95% credible region for π_0. Then we calculate

$$\inf_{\pi\in\Gamma} P(\theta \in C|x, \pi) = 0.945,$$

achieved at $k = 3.4$, and

$$\sup_{\pi\in\Gamma} P(\theta \in C|x, \pi) = 0.956,$$

achieved at $k = 0.93$. So in this sense the inference is very robust.

Chapter 9

Posterior Distributions

Point estimates

Some natural summaries of distributions are the mean, the median, and the mode. The mode is most likely value of θ, so it is the *maximum Bayesian likelihood estimator*.

For summarizing spread, we use the inter-quartile range (IQR), the variance, etc.

Interval estimates

If π is the density for a parameter $\theta \in \Theta$, then a region $C \subset \Theta$ such that

$$\int_C \pi(\theta) d\theta = 1 - \alpha$$

is called a $100(1-\alpha)\%$ *credible region* for θ with respect to π.

If C is an interval: it is called a *credible interval*. Here we take credible regions with respect to the posterior distribution; we abbreviate the posterior by π, abusing notation.

A $100(1-\alpha)\%$ credible region is not unique. Often it is natural to give the smallest $100(1-\alpha)\%$ credible region, especially when the region is an interval. We say that $C \subset \Theta$ is a $100(1-\alpha)\%$ *highest probability density region* (HPD) with respect to π if

(i) $\int_C \pi(\theta) d\theta = 1 - \alpha$; and

(ii) $\pi(\theta_1) \geq \pi(\theta_2)$ for all $\theta_1 \in C, \theta_2 \notin C$ except possibly for a subset of Θ having π-probability 0.

A $100(1-\alpha)\%$ HPD has minimum volume over all $100(1-\alpha)\%$ credible regions.

The full posterior distribution itself is often more informative than credible regions, unless the problem is very complex.

Asymptotics

When n is large, then under suitable regularity conditions, the posterior is approximately normal, with mean the m.l.e. $\hat{\theta}$, and variance $(nI_1(\hat{\theta}))^{-1}$.

This asymptotics requires that the prior is non-zero in a region surrounding the m.l.e..

Since, under regularity, the m.l.e. is consistent; if the data are i.i.d. from $f(x|\theta_0)$ and if the prior is non-zero around θ_0, then the posterior will become more and more concentrated around θ_0. In this sense Bayesian estimation is automatically consistent.

Part III

A Decision-Theoretic Approach and Bayesian testing

Chapter 10

Bayesian Inference as a Decision Problem

10.1 The decision-theoretic set-up

In Decision Theory we choose between various possible *actions* after observing data. We denote by Θ the set of all possible states of nature (values of parameter); $\mathcal{D}$ is the set of all possible decisions (*actions*). With a decision and a state of nature comes an associated loss. A *loss function* is any function

$$L : \Theta \times \mathcal{D} \to [0, \infty)$$

$L(\theta, d)$ gives the cost (penalty) associated with decision d if the true state of the world is θ. We use the notation $f(x, \theta)$ for the sampling distribution, for a sample $x \in \mathcal{X}$; $\pi(\theta)$ denotes a prior distribution, and $L(\theta, d)$ a loss function. Often the decision d is to evaluate or estimate a function $h(\theta)$ as accurately as possible.

For *Point estimation:* $h(\theta) = \theta$ and $\mathcal{D} = \Theta$; finally $L(\theta, d)$ loss in reporting d when θ is true.

For *Hypothesis testing*: for testing $H_0 : \theta \in \Theta_0$, the decision set it

$$\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\},$$

and

$$h(\theta) \;=\; \begin{cases} 1 & \text{if } \theta \in \Theta_0 \\ 0 & \text{otherwise .} \end{cases}$$

The general loss function is

$$\begin{aligned} L(\theta,\ \text{accept } H_0) &= \begin{cases} \ell_{00} & \text{if } \theta \in \Theta_0 \\ \ell_{01} & \text{otherwise} \end{cases} \\ L(\theta,\ \text{reject } H_0) &= \begin{cases} \ell_{10} & \text{if } \theta \in \Theta_0 \\ \ell_{11} & \text{otherwise .} \end{cases} \end{aligned}$$

Note: ℓ_{01} is the Type II-error, (accept H_0 although false), ℓ_{10} is the Type I-error (reject H_0 although true).

A *Decision rule:* $\delta : \mathcal{X} \to \mathcal{D}$ maps observations to decisions. We aim to choose δ such that the incurred loss is small. In general there is no δ that uniformly mimimizes $L(\theta, \delta(x))$.

Bayesian setting
For a prior π and data $x \in \mathcal{X}$, the *posterior expected loss* of a decision is

$$\rho(\pi, d|x) = \int_\Theta L(\theta, d)\pi(\theta|x)d\theta,$$

which is a function of x. For a prior π, the *integrated risk* of a decision rule δ is

$$r(\pi, \delta) = \int_\Theta \int_\mathcal{X} L(\theta, \delta(x) f(x|\theta) dx \pi(\theta) d\theta,$$

which is a real number. We prefer δ_1 to δ_2 if and only if $r(\pi, \delta_1) < r(\pi, \delta_2)$.

Proposition. An estimator minimizing $r(\pi, \delta)$ can be obtained by selecting, for every $x \in \mathcal{X}$, the value $\delta(x)$ that minimizes $\rho(\pi, \delta|x)$.

Proof (additional material)

$$\begin{aligned} r(\pi, \delta) &= \int_\Theta \int_\mathcal{X} L(\theta, \delta(x)) f(x|\theta) dx \pi(\theta) d\theta \\ &= \int_\mathcal{X} \int_\Theta L(\theta, \delta(x)) \pi(\theta|x) p(x) d\theta dx \\ &= \int_\mathcal{X} \rho(\pi, \delta|x) p(x) dx \end{aligned}$$

(Recall that $p(x) = \int_\Theta f(x|\theta)\pi(\theta)d\theta$.)

A *Bayes estimator* associated with prior π, loss L, is any estimator δ^π which minimizes $r(\pi, \delta)$: For every $x \in \mathcal{X}$ it is

$$\delta^\pi = arg \min_d \rho(\pi, d|x);$$

then $r(\pi) = r(\pi, \delta^\pi)$ is called *Bayes risk*. This is valid for proper priors, and for improper priors if $r(\pi) < \infty$. If $r(\pi) = \infty$ one can define a *generalised Bayes estimator* as the minimizer, for every x, of $\rho(\pi, d|x)$.

Fact: For strictly convex loss functions, Bayes estimators are unique.

Some common loss functions

Loss functions are part of the problem specification. The *Squared error loss:* $L(\theta, d) = (\theta - d)^2$ is convex, and penalises large deviations heavily.

Proposition The Bayes estimator δ^π associated with prior π under squared error loss is the posterior mean,

$$\delta^\pi(x) \quad = \quad E^\pi(\theta|x) = \frac{\int_\Theta \theta f(x|\theta)\pi(\theta)d\theta}{\int_\Theta f(x|\theta)\pi(\theta)d\theta}.$$

Reason: for any random variable Y, $E((Y-a)^2)$ is minimized by $a = EY$.

The *Absolute error loss* is $L(\theta, d) = |\theta - d|$.

Proposition: The posterior median is a Bayes estimator under absolute error loss.

10.2 Bayesian testing

Suppose that we want to $H_0 : \theta \in \Theta_0$, so that

$$\mathcal{D} = \{\text{accept } H_0, \text{ reject } H_0\} = \{1, 0\},$$

where 1 stands for acceptance. We choose as loss function

$$L(\theta, \phi) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \phi = 1 \\ a_0 & \text{if } \theta \in \Theta_0, \phi = 0 \\ 0 & \text{if } \theta \notin \Theta_0, \phi = 0 \\ a_1 & \text{if } \theta \notin \Theta_0, \phi = 1. \end{cases}$$

Proposition Under this loss function, the Bayes decision rule associated with a prior distribution π is

$$\phi^\pi(x) = \begin{cases} 1 & \text{if } P^\pi(\theta \in \Theta_0|x) > \frac{a_1}{a_0+a_1} \\ 0 & \text{otherwise .} \end{cases}$$

Note the special case: If $a_0 = a_1$, then we accept H_0 if $P^\pi(\theta \in \Theta_0|x) > \frac{1}{2}$.

Proof (additional material) The posterior expected loss is

$$\begin{aligned} \rho(\pi, \phi|x) &= a_0 P^\pi(\theta \in \Theta_0|x)\mathbf{1}(\phi(x) = 0) \\ &\quad + a_1 P^\pi(\theta \notin \Theta_0|x)\mathbf{1}(\phi(x) = 1) \\ &= a_0 P^\pi(\theta \in \Theta_0|x) + \mathbf{1}(\phi(x) = 1) \\ &\quad (a_1 - (a_0 + a_1)P^\pi(\theta \in \Theta_0|x)) , \end{aligned}$$

and $a_1 - (a_0 + a_1)P^\pi(\theta \in \Theta_0|x) < 0$ if and only if $P^\pi(\theta \in \Theta_0|x) > \frac{a_1}{a_0+a_1}$.

Example: $X \sim Bin(n, \theta)$, $\Theta_0 = [0, 1/2), \pi(\theta) = 1$

$$\begin{aligned} P^\pi\left(\theta < \frac{1}{2}|x\right) &= \frac{\int_0^{\frac{1}{2}} \theta^x(1-\theta)^{n-x}d\theta}{\int_0^1 \theta^x(1-\theta)^{n-x}d\theta} \\ &= \frac{\left(\frac{1}{2}\right)^{n+1}}{B(x+1, n-x+1)} \left\{ \frac{1}{x+1} + \ldots + \frac{(n-x)!x!}{(n+1)!} \right\} \end{aligned}$$

This expression can be evaluated for particular n and x, and compared with the acceptance level $\frac{a_1}{a_0+a_1}$.

Example: $X \sim \mathcal{N}(\theta, \sigma^2)$, with σ^2 known, and $\theta \sim \mathcal{N}(\mu, \tau^2)$. Then we have already calculated that $\pi(\theta|x) \sim \mathcal{N}(\mu(x), w^2)$ with

$$\mu(x) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2} \quad \text{and } w^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

For testing $H_0 : \theta < 0$ we calculate

$$P^\pi(\theta < 0|x) = P^\pi\left(\frac{\theta - \mu(x)}{w} < -\frac{\mu(x)}{w}\right) = \Phi\left(-\frac{\mu(x)}{w}\right).$$

Let z_{a_0,a_1} be the $\frac{a_1}{a_0+a_1}$ quantile: then we accept H_0 if $-\mu(x) > z_{a_0,a_1} w$, or, equivalently, if

$$x < -\frac{\sigma^2}{\tau^2}\mu - \left(1 + \frac{\sigma^2}{\tau^2}\right) z_{a_0,a_1} w.$$

For $\sigma^2 = 1, \mu = 0, \tau^2 \to \infty$: we accept H_0 if $x < -z_{a_0,a_1}$

Compare to the frequentist test: Accept H_0 if $x < z_{1-\alpha} = -z_\alpha$. This corresponds to

$$\frac{a_0}{a_1} = \frac{1}{\alpha} - 1.$$

So $\frac{a_0}{a_1} = 19$ for $\alpha = 0.05$; and $\frac{a_0}{a_1} = 99$ for $\alpha = 0.01$.

Note:

1) If the prior probability of H_0 is 0, then so will be posterior probability.

2) Testing $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ often really means testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, which is natural to test in a Bayesian setting.

Definition: The *Bayes factor* for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is

$$B^\pi(x) = \frac{P^\pi(\theta \in \Theta_0|x)/P^\pi(\theta \in \Theta_1|x)}{P^\pi(\theta \in \Theta_0)/P^\pi(\theta \in \Theta_1)}.$$

The Bayes factor measures the extent to which the data x will change the odds of Θ_0 relative to Θ_1. If $B^\pi(x) > 1$ the data adds support to H_0. If $B^\pi(x) < 1$ the data adds support to H_1. If $B^\pi(x) = 1$ the data does not help to distinguish between H_0 and H_1.

Note: the Bayes factor still depends on the prior π.

Special case: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, then

$$B^\pi(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}$$

which is the likelihood ratio.

More generally,

$$\begin{aligned} B^\pi(x) &= \frac{\int_{\Theta_0}\pi(\theta)f(x|\theta)d\theta}{\int_{\Theta_1}\pi(\theta)f(x|\theta)d\theta}\Big/\frac{\int_{\Theta_0}\pi(\theta)d\theta}{\int_{\Theta_1}\pi(\theta)d\theta} \\ &= \frac{\int_{\Theta_0}\pi(\theta)f(x|\theta)/P^\pi(\theta\in\Theta_0)d\theta}{\int_{\Theta_1}\pi(\theta)f(x|\theta))/P^\pi(\theta\in\Theta_1)d\theta} \\ &= \frac{p(x|\theta\in\Theta_0)}{p(x|\theta\in\Theta_1)} \end{aligned}$$

is the ratio of how likely the data is under H_0 and how likely the data is under H_1.

Compare: the frequentist likelihood ratio is

$$\Lambda(x) = \frac{\sup_{\theta\in\Theta_0} f(x|\theta)}{\sup_{\theta\in\Theta_1} f(x|\theta)}.$$

Note: with ϕ^π from the Proposition, and $\rho_0 = P^\pi(\theta\in\Theta_0), \rho_1 = P^\pi(\theta\in\Theta_1)$, we obtain

$$B^\pi(x) = \frac{P^\pi(\theta\in\Theta_0|x)/(1-P^\pi(\theta\in\Theta_0|x))}{\rho_0/\rho_1}$$

and so

$$\phi^\pi(x) = 1 \iff B^\pi(x) > \frac{a_1}{a_0}\Big/\frac{\rho_0}{\rho_1}.$$

Also, by inverting the equality it follows that

$$P^\pi(\theta\in\Theta_0|x) = \left(1+\frac{\rho_1}{\rho_0}(B^\pi(x))^{-1}\right)^{-1}.$$

Example: $X\sim Bin(n,p)$, $H_0: p=1/2$, $H_1: p\neq 1/2$ Choose as prior an atom of size ρ_0 at $1/2$, otherwise uniform on $[0,1]$. Then

$$B^\pi(x) = \frac{p(x|p=1/2)}{p(x|p\in\Theta_1)} = \frac{\binom{n}{x}2^{-n}}{\binom{n}{x}B(x+1,n-x+1)}.$$

So

$$P\left(p=\frac{1}{2}|x\right)=\left(1+\frac{(1-\rho_0)}{\rho_0}\frac{x!(n-x)!}{(n-1)!}2^n\right)^{-1}.$$

If $\rho_0=1/2, n=5, x=3$, then $B^\pi(x)=\frac{15}{8}>1$, and

$$P\left(p=\frac{1}{2}|x\right)=\left(1+\frac{2}{120}2^5\right)^{-1}=\frac{15}{23}.$$

The data adds support to H_0, the posterior probability of H_0 is $15/23>1/2$.

Alternatively had we chosen as prior an atom of size ρ_0 at $1/2$, otherwise $Beta(1/2,1/2)$, then this prior favours 0 and 1; for n=10 we would obtain

x	$P(p=\frac{1}{2}\|x)$
0	0.005
1	0.095
2	0.374
3	0.642
4	0.769
5	0.803

Example: $X\sim\mathcal{N}(\theta,\sigma^2)$, σ^2 known, $H_0:\theta=0$ Choose as prior: mass ρ_0 at $\theta=0$, otherwise $\sim\mathcal{N}(0,\tau^2)$. Then

$$\begin{aligned}(B^\pi)^{-1} &= \frac{p(x|\theta\neq 0)}{p(x|\theta=0)}\\ &= \frac{(\sigma^2+\tau^2)^{-1/2}\exp\{-x^2/(2(\sigma^2+\tau^2))\}}{\sigma^{-1}\exp\{-x^2/(2\sigma^2)\}}\end{aligned}$$

and

$$P(\theta=0|x) = \left(1+\frac{1-\rho_0}{\rho_0}\sqrt{\frac{\sigma^2}{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2(\sigma^2+\tau^2)}\right)\right)^{-1}.$$

Example: $\rho_0=1/2, \tau=\sigma$, put $z=x/\sigma$

x	$P(\theta=0\|z)$
0	0.586
0.68	0.557
1.28	0.484
1.96	0.351

For $\tau = 10\sigma$ (a more diffusive prior)

x	$P(\theta = 0\|z)$
0	0.768
0.68	0.729
1.28	0.612
1.96	0.366

so x gives stronger support for H_0 than under tighter prior.

Note: For x fixed, $\tau^2 \to \infty$, $\rho_0 > 0$, we have

$$P(\theta = 0|x) \to 1.$$

For a noninformative prior $\pi(\theta) \propto 1$ we have that

$$p(x|\pi(\theta) \;=\; \int (2\pi\sigma^2)^{-1/2} e^{-\frac{(x-\theta)^2}{2\sigma^2}} d\theta = (2\pi\sigma^2)^{-1/2} \int e^{-\frac{(\theta-x)^2}{2\sigma^2}} d\theta = 1$$

and so

$$P(\theta = 0|x) = \left(1 + \frac{1-\rho_0}{\rho_0}\sqrt{2\pi}\exp(x^2/2)\right)^{-1}$$

which is not equal to 1.

Lindley's paradox: Suppose that $\overline{X} \sim \mathcal{N}(\theta, \sigma^2/n)$, $H_0 : \theta = 0$, n is fixed. If $\frac{\overline{x}}{(\sigma/\sqrt{n})}$ is large enough to reject H_0 in classical test, then for large enough τ^2 the Bayes factor will be larger than 1, indicating support for H_0.

If σ^2, τ^2 are fixed, $n \to \infty$ such that $\frac{\overline{x}}{(\sigma/\sqrt{n})} = k_\alpha$ fixed, is just significant at level α in classical test, then $B^\pi(\overline{x}) \to \infty$.

Results which are just significant at some fixed level in the classical test will, for large n, actually be much more likely under H_0 than under H_1.

A very diffusive prior proclaims great scepticism, which may overwhelm the contrary evidence of the observations.

10.3 Least favourable Bayesian answers

Suppose that we want to test $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, and the prior probability on H_0 is $\rho_0 = 1/2$. What is the prior g in H_1, which is, after

observing x, least favourable to H_0? Let G be a family of priors on H_1; put

$$\underline{B}(x, G) = \inf_{g \in G} \frac{f(x|\theta_0)}{\int_\Theta f(x|\theta) g(\theta) d\theta}$$

and

$$\begin{aligned} \underline{P}(x, G) &= \frac{f(x|\theta_0)}{f(x|\theta_0) + \sup_{g \in G} \int_\Theta f(x|\theta) g(\theta) d\theta} \\ &= \left(1 + \frac{1}{\underline{B}(x, G)}\right)^{-1} \end{aligned}$$

A Bayesian prior $g \in G$ on H_0 will then have posterior probability at least $\underline{P}(x, G)$ on H_0 (for $\rho_0 = 1/2$). If $\hat{\theta}$ is the m.l.e. of θ, G_A the set of all prior distributions, then

$$\underline{B}(x, G_A) = \frac{f(x|\theta_0)}{f(x|\hat{\theta}(x))}$$

and

$$\underline{P}(x, G_A) = \left(1 + \frac{f(x|\hat{\theta}(x))}{f(x|\theta_0}\right)^{-1}.$$

Other natural families are G_S, the set of distributions symmetric around θ_0, and G_{SU}, the set of unimodal distributions symmetric around θ_0.

Example: Normal, unit variance. Let $X \sim \mathcal{N}(\theta, 1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$. Then

p-value	$\underline{P}(x, G_A)$	$\underline{P}(x, G_{SU})$
0.1	0.205	0.392
0.01	0.035	0.109

The Bayesian approach will typically reject H_0 less frequently.

10.4 Comparison with frequentist hypothesis testing

In frequentist hypothesis setting, there is an asymmetry between H_0, H_1: we fix type I error, then minimize the type II error. UMP tests do not

always exist. Furthermore the concept of p-values can be confusing: they have no intrinsic optimality, the space of p-values lacks a decision-theoretic foundation, They are routinely misinterpreted, and they do not take the type II error into account. Confidence regions are a pre-data measure, and can often have very different post data coverage probabilities.

Example. Consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$. Consider repetitions in which one uses the most powerful test with level $\alpha = 0.01$. In frequentist tests: only 1% of the true H_0 will be rejected. But this does not say anything about the proportion of errors made when rejecting!

Example. Suppose in a test probability of type II error is 0.99, and θ_0 and θ_1 occur equally often, then about half of the rejections of H_0 will be in error.

Example: $X \sim \mathcal{N}(\theta, 1/2)$, $H_0 : \theta = -1$, $H_1 : \theta = 1$. We observe $x = 0$: the UMP test has p-value 0.072, but the p-value for the test of H_1 against H_0 takes exactly the same value.

Example: $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\theta, \sigma^2)$, both θ, σ^2 are unknown. The interval

$$C = \left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

for $n = 2, \alpha = 0.5$: has pre-data coverage probability 0.5. However, *Brown (Ann.Math.Stat. 38, 1967, 1068-1071)* showed that

$$P(\theta \in C | |\overline{x}|/s < 1 + \sqrt{2}) > 2/3.$$

The Bayesian approach compares the "probability" of the actual data under the two hypotheses.

Chapter 11

Hierarchical and empirical Bayesian methods

A *hierarchical* Bayesian model consists of modelling a parameter θ through randomness at different levels; for example,

$$\theta|\beta \sim \pi_1(\theta|\beta), \text{ where } \beta \sim \pi_2(\beta);$$

so that then $\pi(\theta) = \int \pi_1(\theta|\beta)\pi_2(\beta)d\beta$.

When dealing with complicated posterior distributions, rather than evaluating the integrals, we might use simulation to approximate the integrals. For simulation in hierarchical models, we simulate first from β, then, given β, we simulate from θ. We hope that the distribution of β is easy to simulate, and also that the conditional distribution of θ given β is easy to simulate. This approach is particularly useful for MCMC (Markov chain Monte Carlo) methods, e.g.: see next term.

Let $x \sim f(x|\theta)$. The *empirical Bayes* method chooses a convenient prior family $\pi(\theta|\lambda)$ (typically conjugate), where λ is a hyperparameter, so that

$$p(x|\lambda) = \int f(x|\theta)\pi(\theta|\lambda)d\theta.$$

Rather than specifying λ, we estimate λ by $\hat{\lambda}$, for example by frequentist methods, based on $p(x|\lambda)$, and we substitute $\hat{\lambda}$ for λ;

$$\pi(\theta|x, \hat{\lambda})$$

is called a *pseudo-posterior*. We plug it into Bayes' Theorem for inference.

The empirical Bayes approach is neither fully Bayesian nor fully frequentist. It depends on $\hat{\lambda}$; different $\hat{\lambda}$ will lead to different procedures. If $\hat{\lambda}$ is consistent, then asymptotically it will lead to a coherent Bayesian analysis. It often outperforms classical estimators in empirical terms.

Example: James-Stein estimators Let $X_i \sim \mathcal{N}(\theta_i, 1)$ be independent given θ_i, $i = 1, \ldots, p$,where $p \geq 3$. In vector notation: $\mathbf{X} \sim \mathcal{N}(\theta, I_p)$. Here the vector θ is random; assume that we have realizations θ_i, $i = 1, \ldots, p$. The obvious estimate for θ_i is $\hat{\theta}_i = x_i$, leading to

$$\hat{\theta} = \mathbf{X}.$$

Assume that $\theta_i \sim \mathcal{N}(0, \tau^2)$, then $p(\mathbf{x}|\tau^2) = \mathcal{N}(0, (1+\tau^2)I_p)$, and the posterior for θ given the data is

$$\theta|\mathbf{x} \sim \mathcal{N}\left(\frac{\tau^2}{1+\tau^2}\mathbf{x}, \frac{1}{1+\tau^2}I_p\right).$$

Under quadratic loss, the Bayes estimator $\delta(\mathbf{x})$ of θ is the posterior mean

$$\frac{\tau^2}{1+\tau^2}\mathbf{x}.$$

In the empirical Bayes approach, we would use the m.l.e. for τ^2, which is

$$\hat{\tau^2} = \left(\frac{\| \mathbf{x} \|^2}{p} - 1\right)\mathbf{1}(\| \mathbf{x} \|^2 > p),$$

where $\| \mathbf{x} \|^2 = \sum_i x_i^2$. The empirical Bayes estimator is the estimated posterior mean,

$$\delta^{EB}(x) = \frac{\hat{\tau}^2}{1+\hat{\tau}^2}\mathbf{x} = \left(1 - \frac{p}{\| x \|^2}\right)^+ \mathbf{x}$$

is the *truncated James-Stein estimator*. It can can be shown to outperform the estimator $\delta(\mathbf{x}) = \mathbf{x}$.

Alternatively, the best unbiased estimator of $1/(1+\tau^2)$ is $\frac{p-2}{\|\mathbf{x}\|^2}$, giving

$$\delta^{EB}(\mathbf{x}) = \left(1 - \frac{p}{\| \mathbf{x} \|^2}\right)\mathbf{x}.$$

This is the *James-Stein estimator*. It can be shown that under quadratic loss function the James-Stein estimator outperforms $\delta(\mathbf{x}) = \mathbf{x}$.

Note: both estimators tend to "shrink" towards 0. It is now known to be a very general phenomenon that when comparing three or more populations, the sample mean is not the best estimator. *Shrinkage* estimators are an active area of research.

Bayesian computation of posterior probabilities can be very computer-intensive; see the MCMC and Applied Bayesian Statistics course.

Part IV

Principles of Inference

The Likelihood Principle

The Likelihood Principle states that the information brought by an observation x about θ is entirely contained in the likelihood function $L(\theta|x)$. From this follows that, if if x_1 and x_2 are two observations with likelihoods $L_1(\theta|x)$ and $L_2(\theta|x)$, and if

$$L_1(\theta|x) = c(x_1, x_2)L_2(\theta|x)$$

then x_1 and x_2 must lead to identical inferences.

Example. We know that $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$. If $f_1(x|\theta) \propto f_2(x|\theta)$ as a function of θ, then they have the same posterior, so they lead to the same Bayesian inference.

Example: Binomial versus negative binomial. (a) Let $X \sim Bin(n, \theta)$ be the number of successes in n independent trials, with p.m.f.

$$f_1(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

then

$$\pi(\theta|x) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\pi(\theta) \propto \theta^x(1-\theta)^{n-x}\pi(\theta).$$

(b) Let $N \sim NegBin(x, \theta)$ be the number of independent trials until x successes, with p.m.f.

$$f_2(n|\theta) = \binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}$$

and

$$\pi(\theta|x) \propto \theta^x(1-\theta)^{n-x}\pi(\theta).$$

Bayesian inference about θ does not depend on whether a binomial or a negative binomial sampling scheme was used.

M.l.e.'s satisfy the likelihood principle, but many frequentist procedures do not!

Example: $Bin(n, \theta)$-sampling. We observe $(x_1, \ldots, x_n) = (0, \ldots, 0, 1)$. An unbiased estimate for θ is $\hat{\theta} = 1/n$. If instead we view n as geometric(θ), then the only unbiased estimator for θ is $\hat{\theta} = \mathbf{1}(n = 1)$.

Unbiasedness typically violates the likelihood principle: it involves integrals over the sample space, so it depends on the value of $f(x|\theta)$ for values of x other than the observed value.

Example: (a) We observe a Binomial randome variable, n=12; we observe 9 heads, 3 tails. Suppose that we want to test $H_0 : \theta = 1/2$ against $H_1 : \theta > 1/2$. We can calculate that the UMP test has $P(X \geq 9) = 0.075$. (b) If instead we continue tossing until 3 tails recorded, and observe that $N = 12$ tosses are needed, then the underlying distribution is negative binomial, and $P(N \geq 12) = 0.0325$.

The conditionality perspective

Example (*Cox 1958*). A scientist wants to measure a physical quantity θ. Machine 1 gives measurements $X_1 \sim \mathcal{N}(\theta, 1)$, but is often busy. Machine 2 gives measurements $X_1 \sim \mathcal{N}(\theta, 100)$. The availability of machine 1 is beyond the scientist's control, independent of object to be measured. Assume that on any given occasion machine 1 is available with probability 1/2; if available, the scientist chooses machine 1. A standard 95% confidence interval is about $(x - 16.4, x + 16.4)$ because of the possibility that machine 2 was used.

Conditionality Principle: If two experiments on the parameter θ are available, and if one of these two experiments is selected with probability 1/2, then the resulting inference on θ should only depend on the selected experiment.

The conditionality principle is satisfied in Bayesian analysis. In the frequentist approach, we could condition on an ancillary statistic, but such statistic is not always available.

A related principle is the *Stopping rule principle (SRP):* A *it stopping rule* is a random variable that tells when to stop the experiment; this random variable depends only on the outcome of the first n experiments (does not look into the future). The *stopping rule principle* states that if a sequence of experiments is directed by a stopping rule, then, given the resulting sample, the inference about θ should not depend on the nature of the stopping rule.

The likelihood principle implies the SRP. The SRP is satisfied in Bayesian inference, but it is not always satisfied in frequentist analysis.