Rao-Blackwell Theorem (GJJ 2.5.2) Let $X_1, \ldots, X_n$ be a random sample of observations from $f(x;\theta)$. Suppose that T is a sufficient statistic for θ and that $\widehat{\theta}$ is any unbiased estimator for θ. Define $\widehat{\theta}_T = \mathbb{E}[\hat{\theta} \mid T]$. Then

1. $\widehat{\theta}_T$ is a function of T alone;
2. $\mathbb{E}[\widehat{\theta}_T] = \theta$; (partition theorem for expectation)
3. Var$(\widehat{\theta}_T) \leq$ Var$(\widehat{\theta})$ (proof on the board and eg Davison/SM).

Corollary If an MVUE $\widehat{\theta}$ for θ exists, then there is a function $\widehat{\theta}_T$ of the minimal sufficient statistic T for θ which is an MVUE.

Proof: T is sufficient so (2.) $\widehat{\theta}_T$ is an unbiased estimator which is (1.) a function of T alone. By (3.) Var$(\widehat{\theta}_T) \leq$ Var$(\widehat{\theta})$, but $\widehat{\theta}$ is already minimum variance, so Var$(\widehat{\theta}_T)$ is also.

Complete Sufficient Statistic

Let $T(X_1, \ldots, X_n)$ be a sufficient statistic for θ. The statistic T is complete if, whenever $h(T)$ is a function of T for which $\mathbb{E}[h(T)] = 0$ for all θ, then $h(T) \equiv 0$ almost everywhere.

Suppose $h = h(T)$ with T complete and sufficient for θ, and $h(T)$ unbiased for θ. Then $h(T)$ is the unique function of T which is an unbiased estimator of θ.

Proof If there were two such unbiased estimators $h_1(T), h_2(T)$, then $\mathbb{E}[h_1(T) - h_2(T)] = \theta - \theta = 0$ for all θ, so $h_1(T) = h_2(T)$ almost everywhere.

Sufficient condition for estimator to be MVUE

An unbiased estimator with efficiency 1 (ie variance at the CRB) is clearly MVUE (subject to regularity conditions). What if we have an unbiased estimator with efficiency less than one. Could it be MVUE?

Suppose $h = h(T)$ with T complete and minimal sufficient for θ, and $h(T)$ unbiased for θ. If an MVUE for θ exists then $h(T)$ is a MVUE.

Proof: if an MVUE exists then there is a function of T which is an MVUE, by the RB corollary. But $h(T)$ is the only function of T which is unbiased for θ. So h must be the function of T which an MVUE.

Method of moments

Generate estimators by equating observed statistics with their expected values

$$\mathbb{E}_\theta\{t(X_1, \ldots, X_n)\} = t(x_1, \ldots, x_n).$$

Example: Uniform iid sample $X = (X_1, X_2, ..., X_n)$ on $(0, \theta)$.

$$f(x; \theta) = \theta^{-n},\ 0 < x_1, \ldots, x_n < \theta.$$

Let $X_{(n)} = \max_i X_i$. In this case the moment relation

$$\mathbb{E}_\theta \left[X_{(n)}\right] = x_{(n)}$$

leads to an unbiased sufficient statistic, $\widehat{\theta} = \frac{n+1}{n} X_{(n)}$.

The distribution of $X_{(n)} = \max_i X_i$ is obtained from the CDF

$$P(X_{(n)} \leq y) = \left(\frac{y}{\theta}\right)^n$$

(the probability all iid X_i fall in $(0, y)$) so

$$f_{X_{(n)}}(y; \theta) = \frac{ny^{n-1}}{\theta^n},\ 0 < y < \theta$$

and

$$\mathbb{E}_\theta\left[X_{(n)}\right] = \frac{n}{n+1}\theta,\ \text{so}\ \widehat{\theta} = \frac{n+1}{n}X_{(n)}$$

is unbiased.

Now check $\widehat{\theta}$ is sufficient. The distribution of $X \mid X_{(n)} = y$ is

$$\begin{aligned} f(x|X_{(n)} = y;\theta) &= \frac{f(x;\theta)}{f_{X_{(n)}}(y;\theta)} && (1) \\ &= ny^{n-1} && (2) \end{aligned}$$

which does not depend on θ.

Finally, $\widehat{\theta} = \widehat{\theta}(X_{(n)})$ is complete. If $\mathbb{E}_\theta[\widehat{\theta}] = 0$ for all $\theta > 0$ then

$$\int_0^\theta h(\widehat{\theta}(y)) \frac{ny^{n-1}}{\theta^n} dy = 0$$

for all $\theta > 0$ and hence

$$\int_0^\theta h(\widehat{\theta}(y)) y^{n-1} dy = 0.$$

Differentiate wrt θ, and conclude $h(\widehat{\theta}) = 0$ identically for all X. It follows that $\widehat{\theta}$ is the unique unbiased estimator based on $X_{(n)}$. It is not hard to show (using PS1 Q6a) that $\widehat{\theta}$ is minimal sufficiant, and hence, if a MVUE exists, it equals $\widehat{\theta}$. We cant simply use the CRB to show a MVUE exists, as the problem is non-regular.

Exercise $l(\theta; x) = -n\log(\theta)$ so

$$\mathbb{E}\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = n^2/\theta^2$$

and the CRB would be $\text{Var}(\widehat{\theta}) \geq \theta^2/n^2$. However, you can check (using $f_{X_{(n)}}(y;\theta)$ above) that

$$\text{Var}(\widehat{\theta}) = \frac{\theta^2}{n(n+1)}$$

which is smaller than θ^2/n^2 for any $n \geq 1$. This is not a contradiction, as $f(x;\theta)$ doesn't satisfy the regularity conditions (limits of x depend on θ).

Example A sample from $N(\mu, \sigma^2)$. The moment relations

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\mu, \qquad \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] = n(\mu^2 + \sigma^2)$$

lead to the estimating equations

$$\widehat{\mu} = \bar{X}, \qquad \widehat{\mu}^2 + \widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} X_i^2.$$

We have seen that these are just the equations for the MLE's in this 2-dimensional exponential linear family, and solve to give the MLE

$$\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Method of moments asymptotics

Distribution of $\widehat{\theta}$ as $n \to \infty$. Suppose the dimension of θ is $d = 1$ and the moment equation is $\bar{H} = k(\theta)$, with a solution $\widehat{\theta}$, where $\bar{H} = n^{-1} \sum_{i=1}^{n} h(x_i)$.

Theorem. As $n \to \infty$, $\widehat{\theta}$ is asymptotically normally distributed with mean θ and variance $n^{-1}\sigma_h^2/(k'(\theta))^2$, where

$$\sigma_h^2 = \int h(x)^2 f(x;\theta) dx - k(\theta)^2$$

Proof. $\bar{H} = k(\theta)$, so by the Central Limit Theorem,

$$\frac{(\bar{H} - k(\theta))\sqrt{n}}{\sigma_h} \xrightarrow{\mathcal{D}} N(0,1)$$

Approximately

$$\bar{H} = k(\widehat{\theta}) = k(\theta) + (\widehat{\theta} - \theta)k'(\theta)$$

Rearrange to get

$$\frac{\sqrt{n}(\widehat{\theta} - \theta)k'(\theta)}{\sigma_h} = \frac{(\bar{H} - k(\theta))\sqrt{n}}{\sigma_h} \xrightarrow{\mathcal{D}} N(0,1)$$

so

$$\widehat{\theta} \approx N\left(\theta, \frac{\sigma_h^2}{n(k'(\theta))^2}\right)$$

as $n \longrightarrow \infty$.

Example

$$f(x;\theta) = \theta \exp(-\theta x),\ x > 0$$

$$\mathbb{E}[X] = \theta^{-1},\ \widehat{\theta} = \bar{X}^{-1}$$

That is

$$k(\theta) = \theta^{-1},\ \bar{H} = \bar{X},\ h(X) = X,\ \mathbb{E}\big[h(X)\big] = \theta^{-1}$$

$$\mathrm{Var}\big[h(X)\big] = \sigma_h^2 = \theta^{-2}, k'(\theta) = -\theta^{-2}$$

Now

$$\widehat{\theta} \approx N\left(\theta, \frac{\sigma_h^2}{n(k'(\theta))^2}\right)$$

so

$$\widehat{\theta} \approx N(\theta, n^{-1}\theta^2)$$

Exponential families

MLE are moment estimates because we are solving

$$\sum_{i=1}^{n} t(X_i) = -nD'(\theta)$$

for $\widehat{\theta}$ and we showed earlier that

$$\mathbb{E}[t(X)] = -D'(\theta)$$

Now

$$k(\theta) = -D'(\theta), \; k'(\theta) = -D''(\theta).$$

$$\text{Var}(\widehat{\theta}) \approx \frac{\text{Var}(t(X))}{nD''(\theta)^2} = \frac{-D''(\theta)}{nD''(\theta)^2} = I_\theta^{-1}$$

Bayesian Inference (preamble)

Ideas of Probability

The statistical methods you have seen to this point are called classical or Frequentist. The probability for an event is defined as the proportion of successes in an infinite number of repeatable trials.

By contrast, in Subjective Bayesian inference (see Wednesday), probability is a measure of the strength of belief. This includes the Frequentist definition as a special case.

What is a classical confidence interval?

It is a set-valued function $C(X) \subseteq \Theta$ of the data X which covers the parameter $\theta \in C(X)$ a fraction $1-\alpha$ of repeated draws of X taken under the null H_0. This is not the same as the statement that, given data $X = x$, the interval $C(x)$ covers θ with probability $1-\alpha$.

Example 1: (GJJ Ex 5.7) Suppose $X_1, X_2 \sim U(\theta - 1/2, \theta + 1/2)$ so that $X_{(1)}$ and $X_{(2)}$ are the order statistics. Then $C(X) = [X_{(1)}, X_{(2)}]$ is a $\alpha = 1/2$ level CI for θ. Suppose in your data $X = x$, $x_{(2)} - x_{(1)} > 1/2$ (this happens in an eighth of data sets). Then $\theta \in [x_{(1)}, x_{(2)}]$ with probability one.

Example 2: see AC Davison, Statistical Models, Problem 8.9.8 for examples where $C(X) = \emptyset, \Theta$.

What is a p-value?

It appears that there are two kinds of hypothesis tests, those where we specify just H_0 (‘pure tests for significance’), and those where we specify both H_0 and H_1 (as in a likelihood ratio test).

In fact, in a pure test we must choose a test statistic $T(x)$, and define the p-value for data $T(x) = t$ as

$$p\text{-value} = P(T(X) \text{ at least as extreme as } t|H_0).$$

The choice of $T(X)$ amounts to a statement about the direction of likely departures from the null, which requires some consideration of alternative models.

A p-value is not $P(H_0|T(X) = t)$.

The likelihood principle (Davison/SM 11.1.2)

If experiments E_1, E_2 give data y_1, y_2 informing θ, and $L(\theta; y_i, E_i)$ gives the likelihood for θ from data $(E_i, y_i), i = 1, 2$, and

$$L(\theta; y_1, E_1) \propto_\theta L(\theta; y_2, E_2) \qquad \text{varying } \theta$$

then the two experiments lead to identical conclusions about θ.

MLE's respect the LP. Significance tests do not.

Example A Bernoulli trial succeeds with probability p.

E_1 fix n_1 Bernoulli trials, count number y_1 of successes

E_2 count number n_2 Bernoulli trials to get fixed number y_2 successes

$$\begin{aligned} L(p; y_1, E_1) &= \binom{n_1}{y_1} p^{y_1}(1-p)^{n_1-y_1} \\ L(p; y_2, E_2) &= \binom{n_2-1}{y_2-1} p^{y_2}(1-p)^{n_2-y_2} \end{aligned}$$

If $n_1 = n_2 = n$, $y_1 = y_2 = y$ then $L(p; y_1, E_1) \propto_p L(p; y_2, E_2)$.

But significance tests contradict: eg, $H_0 : p = 1/2$ against $H_1 : p < 1/2$ and suppose $n = 12$ and $y = 3$. The p-value based on E_1 is

$$P\left(Y \le y | \theta = \frac{1}{2}\right) = \sum_{k=0}^{y} \binom{n}{k} 2^{-k} (= 0.073)$$

while the p-value based on E_2 is

$$P\left(N \ge n | \theta = \frac{1}{2}\right) = \sum_{k=n}^{\infty} \binom{k-1}{y-1} 2^{-k} (= 0.033)$$

so different conclusions at significance level 0.05.

Bayesian Statistics

Bayes Theorem. Suppose $\mathcal{H}_1, \ldots, \mathcal{H}_k$ are a partition of Θ. Let ϑ be a random variable taking values in Θ and let $H_i = \{\vartheta \in \mathcal{H}_i\}$. Then

$$P(H_i \mid D) = \frac{P(H_i)P(D \mid H_i)}{P(D)},$$

where

$$P(D) = \sum_{j=1}^{k} P(H_j)P(D \mid H_i).$$

This simply follows from the definition of conditional probability, and the partition theorem for probability.

$H_i = \{\vartheta \in \mathcal{H}_i\}$, $\mathcal{H}_1, \ldots, \mathcal{H}_k$ a partition of Θ

$$P(H_i \mid D) = \frac{P(H_i)P(D \mid H_i)}{P(D)}$$

Bayesian Inference. Let ϑ be the unknown true value of the parameter θ, so that H_i is a hypothesis $\{\vartheta \in \mathcal{H}_i\}$. Prior to receiving the data D the probability that H_i is true is $P(H_i)$. Post data-arrival, the probability is updated to $P(H_i \mid D)$.

$P(H_i)$ - Prior probabilities
$P(H_i \mid D)$ - Posterior probabilities

The quantity $P(H_i \mid D)$ is simply the probability that hypothesis H_i is true, given data D and prior probabilities $P(H_j), j = 1, 2, ..., k$ for the various alternatives.

One strength of Bayesian inference is the simple way it summarizes evidence for the parameter.

One weakness is that it is often hard to write down representative prior probabilities for complex hypotheses.

$P(D \mid H_i)$ and $P(D)$ are likelihoods, marginal likelihoods or prior predictive distributions depending on the context. We will return to this later.

Example 1. In a population of m individuals N are red and $m - N$ are blue, with unknown true value N unknown. A sample of k individuals is taken, with replacement, and X are red. Before knowing that $X = x$, all possible values of the parameter $N = n, n = 0, 1, 2, ..., m$ are equally likely, so our state of knowledge about N is represented by the prior probability distribution $P(N = n) = \pi(n)$,

$$\pi(n) = \frac{1}{m+1}, \; n = 0, 1, \ldots, m.$$

The distribution of X given $N = n$ is Binomial$(k, n/m)$. This is the likelihood

$$P(X = x|N = n) = L(n; x)$$

$$L(n; x) = \binom{k}{x}(n/m)^x(1 - n/m)^{k-x}.$$

The posterior probability distribution, $P(N=n|X=x)$ say, is

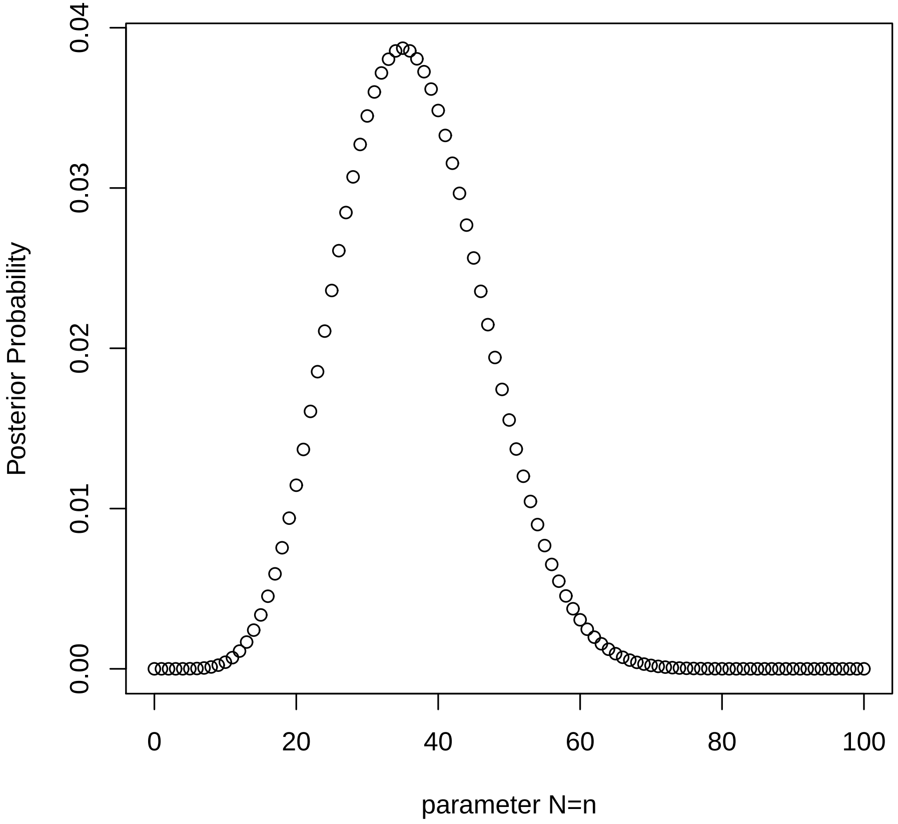
$$P(N=n|X=x)=\frac{P(X=x|N=n)P(N=n)}{\sum_{i=0}^{m}P(X=x|N=i)P(N=i)}$$

since $\mathcal{H}_i=\{i\}, i=0,1,...,m$ are a partition of parameter space $\{0,1,...,m\}$. Let $P(N=n|X=x)=\pi(n\mid x)$. Then

$$\begin{aligned}\pi(n\mid x) &= \frac{L(n;x)\pi(n)}{\sum_{i=0}^{m}L(i;x)\pi(i)}\\ &= \frac{n^x(m-n)^{k-x}}{\sum_{i=0}^{m}i^x(m-i)^{k-x}}.\end{aligned}$$

The posterior distribution contains all information about the parameter (which in this example is discrete).

Example, continued: suppose there are $m = 100$ individuals. We sample $k = 20$ individuals uniformly at random and find $X = 7$ are red. Then $\pi(n \mid x) \propto n^7(100 - n)^{20-7}$.



Natural to summarize $\pi(n|x)$ by a point estimate for N such as

posterior mean $\mathbb{E}_{N|x}(N)$

posterior mode $\arg\max_n \pi(n|x)$ (MAP or ‘maximum a posteriori’ estimate)

posterior std $\sigma_{n|x} = \sqrt{\text{Var}(N)}$ (a measure of uncertainty).

These are useful if the posterior is approximately normal. We will return to this in a more formal way.

Here $\mathbb{E}_{N|x}(N) \simeq 36.4$, the mode is 36, and $\sigma_{n|x} \simeq 10.0$.

Example, continued: we would like to measure the strength of evidence for the statement that the majority are red.

$$\begin{aligned} P(\{N \geq 50\} \mid X = 7\} &= \sum_{n=50}^{100} \pi(n|x) \\ &= \frac{\sum_{n=50}^{100} n^7 (100-n)^{13}}{\sum_{i=0}^{100} i^7 (100-i)^{13}} \\ &\simeq 0.10 \end{aligned}$$

This is the probability that the unknown true number red is 50 or more, given that we observed 7 red in twenty, and given that the number red was equally likely to be any value from 0 to 100, a priori. This is just what a p-value is not.

Parametric models. A family of distributions $f(x \mid \theta)$ and prior distribution $\pi(\theta)$ for ϑ. The posterior distribution of ϑ at $\vartheta = \theta$, given x, is

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

Thus

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta).$$

The same form for θ continuous ($\pi(\theta \mid x)$ a pdf) or discrete ($\pi(\theta \mid x)$ a pmf) posterior $\propto$ prior $\times$ likelihood

Likelihood principle Notice that, if we base all inference on the posterior distribution, then we respect the likelihood principle. If two likelihood functions are proportional, then any constant cancels top and bottom in Bayes rule, and the two posterior distributions are the same.

Example 2. $X \sim \text{Bin}(n, \vartheta)$ for known n and unknown ϑ. Suppose our prior knowledge about ϑ is represented by a Beta distribution on $(0, 1)$, and θ is a trial value for ϑ.

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \; 0 < \theta < 1.$$

and

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \; x = 0, \ldots, n$$

The posterior

$$\pi(\theta \mid x) \propto \theta^{a+x-1}(1-\theta)^{n-x+b-1}$$

which has the same form as the prior with updated parameters a, b replaced by $a + x, b + n - x$, so

$$\pi(\theta \mid x) = \frac{\theta^{a+x-1}(1-\theta)^{n-x+b-1}}{B(a+x, b+n-x)}$$

For a Beta distribution with parameters a, b

$$\mu = \frac{a}{a+b}, \; \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

The posterior mean and variance are

$$\frac{a+X}{a+b+n}, \; \frac{(a+X)(b+n-X)}{(a+b+n)^2(a+b+n+1)}$$

For large n, the posterior mean and variance are approximately

$$\frac{X}{n}, \; \frac{X(n-X)}{n^3}$$

In classical statistics

$$\widehat{\theta} = \frac{X}{n}, \; \frac{\widehat{\theta}(1-\widehat{\theta})}{n} = \frac{X(n-X)}{n^3}$$

In the example above the parametric form of the posterior is like the prior. $f(x \mid \theta)$ and $\pi(\theta)$ form a conjugate prior family.