A9: Statistics

Neil Laws

Hilary Term 2022 (version of 09-12-2021)

I plan to use some slides in lectures and the slides will be available separately, they are not part of this document.

Sections 1–3 of these notes are based on material in previous notes written by Dan Lunn and Simon Myers.

Please send any comments or corrections to: `neil.laws@stats.ox.ac.uk`.

HT 2022 updates:

None so far. If you spot any errors please let me know.

Contents

1 Estimation

1.1 Starting point

Assume the random variable X belongs to a family of distributions indexed by a scalar or vector parameter θ, where θ takes values in some parameter space Θ, i.e. we have a *parametric family.*

E.g. $X \sim \text{Poisson}(\lambda)$. Then $\theta = \lambda \in \Theta = (0, \infty)$.

E.g. $X \sim N(\mu, \sigma^2)$. Then $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

Suppose we have data $\mathbf{x} = (x_1, \ldots, x_n)$ (numerical values). We regard these data as the observed values of iid random variables (RVs) $X_1, \ldots, X_n$ with the same distribution as X. So $\mathbf{X} = (X_1, \ldots, X_n)$ is a *random sample.*

Having observed $\mathbf{X} = \mathbf{x}$, what can we infer/say about θ? E.g. we might wish to

- make a point estimate $t(\mathbf{x})$ of the true value of θ
- construct an interval estimate $(a(\mathbf{x}), b(\mathbf{x}))$ for θ (a confidence interval)
- test a hypothesis about θ, e.g. test $H : \theta = 0$, do the data provide evidence against H?

The first two thirds of the course (approx) will consider the frequentist approach to questions like these. The last third will look at the Bayesian approach.

Notation

If X is a discrete RV, let $f(x; \theta) = P(X = x)$ be the probability mass function (pmf) of X.

If X is a continuous RV, let $f(x; \theta)$ be the probability density function (pdf) of X.

That is, since the distribution of X depends on θ we are writing the pmf/pdf as $f(x; \theta)$.

Write $f(\mathbf{x}; \theta)$ for the joint pmf/pdf of $\mathbf{X} = (X_1, \ldots, X_n)$. Since the X_i are assumed independent we have

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

Occasionally in this course the X_i may not be iid, in which case $f(\mathbf{x}; \theta)$ will still denote the joint pmf/pdf but may not have such a simple form.

Example 1.1 (discrete RV)**.** Let $X_i \sim \text{Poisson}(\theta)$. Then

$$f(x; \theta) = \frac{e^{-\theta}\theta^x}{x!} \quad \text{for } x = 0, 1, \ldots$$

and so

$$f(\mathbf{x}; \theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_i x_i!}.$$

Example 1.2 (continuous RV)**.** Let X_i be "exponential with parameter (or rate) θ", i.e. the pdf is

$$f(x; \theta) = \theta e^{-\theta x} \quad \text{for } x \geqslant 0.$$

Then

$$f(\mathbf{x};\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_i x_i}.$$

Note: $E(X_i) = 1/\theta$. Sometimes we let $\mu = 1/\theta$ and talk about X_i being "exponential with mean μ", so with pdf

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu} \quad \text{for } x \geqslant 0.$$

Note: to change the parameter from θ to μ, all we have to do is replace the constant θ by the constant $1/\mu$ in the pdf.

Example 1.3 (expectation/variance of sums of RVs)**.** Let $a_1, \ldots, a_n$ be constants. Recall that:

(i)
$$E\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i E(X_i)$$
is true whether or not the X_i are independent.

(ii) If the X_i are independent, then
$$\operatorname{var}\left(\sum_{i=1}^{n} a_i X_i \right) = \sum_{i=1}^{n} a_i^2 \operatorname{var}(X_i).$$

Exercise. Suppose $X_1, \ldots, X_n$ are iid with $E(X_i) = \mu$ and $\operatorname{var}(X_i) = \sigma^2$. As usual, let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Show that $E(\overline{X}) = \mu$ and $\operatorname{var}(\overline{X}) = \sigma^2/n$.

Estimators

An *estimator* of θ is any function $t(\mathbf{X})$ that we might use to estimate θ. Note: the function t is not allowed to depend on θ.

The corresponding *estimate* is $t(\mathbf{x})$.

An estimator $t(\mathbf{X})$, which we can think of as a rule for constructing an estimate, is a RV.

An estimate $t(\mathbf{x})$ is just a number, the numerical value of the estimator for a particular set of data.

The estimator $T = t(\mathbf{X})$ is said to be *unbiased for* θ if $E(T) = \theta$ for all θ.

Example 1.4. Suppose the X_i are iid, with $E(X_i) = \mu$ and $\operatorname{var}(X_i) = \sigma^2$.

(i) We might consider $T_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ as an estimator of μ.

Then T_1 is unbiased for μ since $E(T_1) = \mu$.

(ii) We might consider $T_2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$ as an estimator of σ^2.

Then T_2 is biased for σ^2 since $E(T_2) = \frac{n-1}{n}\sigma^2 < \sigma^2$.

In order to have an unbiased estimator, we usually use $S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$ as an estimator of σ^2 (since $E(S^2) = \frac{n}{n-1} E(T_2) = \sigma^2$).

Likelihood

The likelihood for θ, based on $\mathbf{x}$, is

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta)$$

where L is regarded as a function of θ, for a fixed $\mathbf{x}$. We regard information about θ as being contained in L. The idea is that L will be larger for values of θ near the true value of θ which generated the data.

We often write $L(\theta)$ for $L(\theta; \mathbf{x})$. The log-likelihood is $\ell(\theta) = \log L(\theta)$. Or we might sometimes use $\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x})$ if we want to include the dependence on $\mathbf{x}$. Here $\log = \log_e = \ln$.

When the X_i are iid from $f(x; \theta)$ we have

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

and e.g. when $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$ this becomes

$$L(\theta) = \frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_i x_i!}.$$

Here $\overset{\text{iid}}{\sim}$ means "are independent and identically distributed as".

Maximum likelihood

The value of θ which maximises L (or equivalently ℓ) is denoted by $\widehat{\theta}(\mathbf{x})$, or just $\widehat{\theta}$, and is called the maximum likelihood estimate of θ.

The maximum likelihood estimator (MLE) is $\widehat{\theta}(\mathbf{X})$.

For the Poisson example:

$$\ell(\theta) = -n\theta + \sum_i x_i \log\theta - \log(\prod_i x_i!)$$

$$\ell'(\theta) = -n + \frac{\sum_i x_i}{\theta}.$$

So $\ell'(\theta) = 0 \iff \theta = \frac{\sum_i x_i}{n} = \overline{x}$. This is a maximum since $\ell''(\theta) = -\frac{\sum_i x_i}{\theta^2} < 0$.

So the MLE of θ is $\widehat{\theta} = \overline{X}$.

Exercise. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Geometric}(p)$ so that $P(X_i = x) = (1-p)^{x-1}p$ for $x = 1, 2, \ldots$. Let $\theta = 1/p$.

Find (i) the MLE of θ, (ii) the MLE of p. Show that (i) is unbiased but that (ii) is biased.

Example 1.5. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. So here we have a parameter vector $\theta = (\mu, \sigma^2)$.

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

and

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

Differentiating,

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$
$$\frac{\partial \ell}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Solving $\frac{\partial \ell}{\partial \mu} = 0$ and $\frac{\partial \ell}{\partial (\sigma^2)} = 0$ we find

$$\widehat{\mu} = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

1.2 Delta method

Suppose $X_1, \ldots, X_n$ are iid (but *not* necessarily normally distributed) with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$.

Recall (from Prelims, more in Part A Probability) that, by the Central Limit Theorem (CLT),

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{D}}{\approx} N(0, 1)$$

for large n, with this approximation becoming exact in the limit $n \to \infty$. Here $\overset{\text{D}}{\approx}$ means "has approximately the same distribution as".

We often need the asymptotic (i.e. large n) distribution of $g(\overline{X})$, for some function g. E.g. we may have $\widehat{\theta} = 1/\overline{X}$ and we may want the large sample distribution of $\widehat{\theta}$.

Taylor series expansion of $g(\overline{X})$ about $g(\mu)$ gives

$$g(\overline{X}) \approx g(\mu) + (\overline{X} - \mu)g'(\mu). \tag{1.1}$$

By the weak/strong law of large numbers we know that $\overline{X}$ approaches μ as $n \to \infty$. Hence we take the Taylor expansion of $g(\overline{X})$ about $g(\mu)$. The first term $g(\mu)$ is a term of order 1, it is just a constant independent of n. The next term is $(\overline{X} - \mu)g'(\mu)$: by the CLT the $\overline{X} - \mu$ part is of order $n^{-1/2}$, and $g'(\mu)$ is a constant, an order 1 term, so this whole term is of order $n^{-1/2}$ – so is small compared to the initial $g(\mu)$ term. Further terms in the expansion will be smaller still and we omit them.

Taking the expectation, and the variance, of each side of (1.1),

$$E[g(\overline{X})] \approx g(\mu) + g'(\mu)E[(\overline{X} - \mu)] = g(\mu)$$
$$\text{var}[g(\overline{X})] \approx \text{var}[g'(\mu)(\overline{X} - \mu)] = g'(\mu)^2\,\text{var}(\overline{X}) = \frac{g'(\mu)^2\sigma^2}{n}$$

since $E(\overline{X}) = \mu$ and $\text{var}(\overline{X}) = \sigma^2/n$.

Also using (1.1), we see that $g(\overline{X})$ is approximately normal (since it is a linear function of $\overline{X}$, and $\overline{X}$ is approximately normal), hence

$$g(\overline{X}) \overset{\text{D}}{\approx} N\left(g(\mu), \frac{g'(\mu)^2\sigma^2}{n}\right). \tag{1.2}$$

We say that this is the *asymptotic distribution* of $g(\overline{X})$, and we call $g(\mu)$ the *asymptotic mean* and $g'(\mu)^2\sigma^2/n$ the *asymptotic variance*.

The above process is known as the *delta method*.

Example 1.6. Suppose $X_1, \ldots, X_n$ are iid exponential with parameter λ, so with pdf $f(x;\lambda) = \lambda e^{-\lambda x}$, $x \geqslant 0$. Here $\mu = E(X_i) = 1/\lambda$ and $\sigma^2 = \text{var}(X_i) = 1/\lambda^2$.

Let $g(\overline{X}) = \log(\overline{X})$. Then with $g(u) = \log u$ we have $g'(u)^2 = 1/u^2$ and the mean and variance in (1.2) are

$$g(\mu) = \log\mu = \log\frac{1}{\lambda} = -\log\lambda$$
$$g'(\mu)^2\frac{\sigma^2}{n} = \frac{1}{\mu^2}\cdot\frac{\sigma^2}{n} = \lambda^2\frac{1}{\lambda^2 n} = \frac{1}{n}.$$

Hence $g(\overline{X}) = \log\overline{X} \overset{\text{D}}{\approx} N(-\log(\lambda), 1/n)$.

The delta method is not restricted to functions of $\overline{X}$. Suppose we have some estimator T, and that we are interested in the estimator $g(T)$. Let $E(T) = \mu_T$ and $\text{var}(T) = \sigma_T^2$. Then Taylor series expansion of $g(T)$ about $g(\mu_T)$ gives

$$g(T) \approx g(\mu_T) + (T - \mu_T)g'(\mu_T).$$

So again taking expectations, and variances, we get $E[g(T)] \approx g(\mu_T)$ and $\text{var}[g(T)] \approx g'(\mu_T)^2\sigma_T^2$, and if T is approximately normal then $g(T)$ is also approximately normal.

An example where T is not $\overline{X}$ is where T is an order statistic – see the next section.

1.3 Order statistics

The *order statistics* of data $x_1, \ldots, x_n$ are their values in increasing order, which we denote $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$.

The *sample median* is

$$m = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ even} \end{cases}$$

and $\frac{1}{2}$ of the sample is less than the sample median.

Similarly the *lower quartile* has $\frac{1}{4}$ of the sample is less than it, and the *upper quartile* has $\frac{3}{4}$ of the sample is less than it. The lower/upper quartiles can be defined in terms of $x_{(\lfloor n/4\rfloor)}, x_{(\lfloor n/4\rfloor+1)}, \ldots$, using interpolation.

The *inter-quartile range* (IQR) is defined by

$$\text{IQR} = \text{upper quartile} - \text{lower quartile}.$$

SLIDES. Boxplot slides go here.

Definition. The rth *order statistic* of the random sample $X_1, \ldots, X_n$ is the RV $X_{(r)}$ where

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$$

is the ordered sample.

We now assume that the X_i are from a continuous distribution so that $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ with probability 1. So we have

$$\begin{aligned} X_{(1)} &= \min_{1 \leqslant i \leqslant n} \{X_i\} \\ X_{(2)} &= \text{second smallest } X_i \\ &\vdots \\ X_{(n)} &= \max_{1 \leqslant i \leqslant n} \{X_i\}. \end{aligned}$$

The median (a RV) is

$$M = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ odd} \\ \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ even.} \end{cases}$$

Distribution of $X_{(r)}$

Now assume $X_1, \ldots, X_n$ are iid continuous RVs, each having cdf F and pdf f. How do we find the distribution $X_{(r)}$?

First we do the case $r = n$: the cdf of $X_{(n)}$ is

$$\begin{aligned} F_{(n)}(x) &= P(X_{(n)} \leqslant x) \\ &= P(X_1 \leqslant x, \ldots, X_n \leqslant x) \\ &= P(X_1 \leqslant x) \ldots P(X_n \leqslant x) \quad \text{by independence} \\ &= F(x)^n \quad \text{as each } X_i \text{ has cdf } F. \end{aligned}$$

Then, differentiating, the pdf of $X_{(n)}$ is

$$f_{(n)}(x) = F'_{(n)}(x) = nF(x)^{n-1} f(x).$$

Next we do the case $r = 1$: the cdf of $X_{(1)}$ is

$$\begin{aligned} F_{(1)}(x) &= P(X_{(1)} \leqslant x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - P(X_1 > x, \ldots, X_n > x) \\ &= 1 - P(X_1 > x) \ldots P(X_n > x) \quad \text{by independence} \\ &= 1 - [1 - F(x)]^n. \end{aligned}$$

So the pdf of $X_{(1)}$ is

$$f_{(1)}(x) = F'_{(1)}(x) = n[1 - F(x)]^{n-1} f(x).$$

The following theorem gives the general result.

Theorem 1.7. *The pdf of $X_{(r)}$ is given by*

$$f_{(r)}(x) = \frac{n!}{(r-1)!\,(n-r)!} F(x)^{r-1}[1-F(x)]^{n-r} f(x).$$

Proof. By induction. We have shown the result for $r = 1$ (and $r = n$) above, so assume it is true at r.

For all r, the cdf $F_{(r)}$ of $X_{(r)}$ is given by

$$F_{(r)}(x) = P(X_{(r)} \leqslant x) = \sum_{j=r}^{n} \binom{n}{j} F(x)^j [1-F(x)]^{n-j}$$

i.e. the probability that at least r of the X_i are $\leqslant x$.

Hence

$$F_{(r)}(x) - F_{(r+1)}(x) = \binom{n}{r} F(x)^r [1-F(x)]^{n-r}.$$

Differentiating,

$$\begin{aligned} f_{(r+1)}(x) &= f_{(r)}(x) - \binom{n}{r} F(x)^{r-1}[1-F(x)]^{n-r-1}[r - nF(x)]f(x) \\ &= \binom{n}{r} F(x)^r [1-F(x)]^{n-r-1}(n-r)f(x) \quad \text{using the inductive hypothesis} \\ &= \frac{n!}{r!\,(n-(r+1))!} F(x)^{(r+1)-1}[1-F(x)]^{n-(r+1)} f(x) \end{aligned}$$

so the result follows by induction. □

Heuristic method to find $f_{(r)}$

Divide $(-\infty, \infty)$ into 3 parts:

$(-\infty, x)$	the probability of X_1 being in this interval is $F(x)$
$[x, x+\delta x)$	the probability of X_1 being in this interval is approx $f(x)\,\delta x$
$[x+\delta x, \infty)$	the probability of X_1 being in this interval is approx $1 - F(x)$.

For $X_{(r)}$ to be in $[x, x+\delta x)$ we need

$$\begin{aligned} r-1 &\text{ of the } X_i \text{ in } (-\infty, x) \\ 1 \quad &\text{of the } X_i \text{ in } [x, x+\delta x) \\ n-r &\text{ of the } X_i \text{ in } [x+\delta x, \infty). \end{aligned}$$

Approx, this has probability

$$\frac{n!}{(r-1)!\,1!\,(n-r)!} F(x)^{r-1} \cdot f(x)\,\delta x \cdot [1-F(x)]^{n-r}.$$

Omitting the δx gives the density $f_{(r)}(x)$ (i.e. divide by δx and let $\delta x \to 0$).

1.4 Q-Q plots

Q-Q plot is short for "quantile-quantile plot." Q-Q plots are sometimes called probability plots. A Q-Q plot can be used to examine if it is plausible, i.e. if it is reasonable to assume, that a set of data comes from a certain distribution.

For a distribution with cdf F and pdf f, the pth *quantile* (where $0 \leqslant p \leqslant 1$) is the value x_p such that

$$\int_{-\infty}^{x_p} f(u)\,du = p.$$

So $x_p = F^{-1}(p)$. The name "Q-Q plot" comes from the fact that the plot compares quantile values.

Lemma 1.8. *Suppose X is a continuous RV taking values in (a, b), with strictly increasing cdf $F(x)$ for $x \in (a, b)$. Let $Y = F(X)$. Then $Y \sim U(0, 1)$.*

[Proof: Prelims/a question on Sheet 1.]

The transformation $F(X)$, sometimes written $F_X(X)$ to emphasise that F_x is the cdf of X, is called the *probability integral transform* of X.

Let $U \sim U(0, 1)$. We can write the result of the lemma as

$$F(X) \sim U. \tag{1.3}$$

In (1.3), $\sim$ means "has the same distribution as". Applying F^{-1} to both sides of (1.3), we obtain

$$X \sim F^{-1}(U).$$

Lemma 1.9. *If $U_{(1)}, \ldots, U_{(n)}$ are the order statistics of a random sample of size n from a $U(0, 1)$ distribution, then*

(i) $E(U_{(r)}) = \dfrac{r}{n+1}$

(ii) $\operatorname{var}(U_{(r)}) = \dfrac{r}{(n+1)(n+2)}\left(1 - \dfrac{r}{n+1}\right)$.

[Proof: a question on Sheet 1.]

Note that $\operatorname{var}(U_{(r)}) = \frac{1}{n+2} p_r(1 - p_r)$ where $p_r = \frac{r}{n+1} \in [0, 1]$. We know that $p(1 - p)$ is maximised over $p \in [0, 1]$ at $p = \frac{1}{2}$, and hence $\operatorname{var}(U_{(r)}) \leqslant \frac{1}{n+2} \cdot \frac{1}{2} \cdot \frac{1}{2}$. So this variance is of order n^{-1} at most.

The question we are interested in is: *is it reasonable to assume that data $x_1, \ldots, x_n$ are a random sample from F?*

By Lemma 1.8, we can generate a random sample from F by first taking $U_1, \ldots, U_n \overset{\text{iid}}{\sim} U(0, 1)$, and then setting

$$X_k = F^{-1}(U_k), \quad k = 1, \ldots, n.$$

The order statistics are then $X_{(k)} = F^{-1}(U_{(k)})$, $k = 1, \ldots, n$.

If F is indeed a reasonable distribution to assume for data $x_1, \ldots, x_n$, then we expect $x_{(k)}$ to be fairly close to $E(X_{(k)})$. Now

$$\begin{aligned} E(X_{(k)}) &= E[F^{-1}(U_{(k)})] \\ &\approx F^{-1}(E[U_{(k)}]) \quad \text{by the delta method} \\ &= F^{-1}(k/(n+1)) \quad \text{by Lemma 1.9(i).} \end{aligned}$$

In using the delta method here we are using the fact, from Lemma 1.9(ii), that $\text{var}(U_{(k)})$ is small when n is large, so the delta method approximation is fairly good provided n is fairly large.

So we expect $x_{(k)}$ to be fairly close to $F^{-1}(k/(n+1))$.

In a Q-Q plot we plot the values of $x_{(k)}$ against $F^{-1}(k/(n+1))$, for $k = 1, \ldots, n$. So, roughly, a Q-Q plot is a plot of observed values of RVs $(x_{(k)})$ against their expectations $(F^{-1}(k/(n+1)))$.

If the plotted points are a good approximation to the line $y = x$ then it is reasonable to assume the data are a random sample from F.

Note: usually (as above) the data $x_{(k)}$ are on the vertical axis; but occasionally the $F^{-1}(k/(n+1))$ values are plotted on the vertical axis against $x_{(k)}$ on the horizontal axis.

SLIDES. Comparing $N(0,1)$ and t distribution slides go here.

In practice F usually depends on an unknown parameter θ, so F and F^{-1} are unknown. How do we handle this?

The starting point for all of the following Q-Q plots is the approximation we have obtained above, written in the form

$$F(x_{(k)}) \approx \frac{k}{n+1}$$

where we are assuming that $X_1, \ldots, X_n$ are iid from F.

Normal Q-Q plot

If data $x_1, \ldots, x_n$ are from a $N(\mu, \sigma^2)$ distribution, for some unknown μ and σ^2, then we expect, roughly,

$$F(x_{(k)}) \approx \frac{k}{n+1} \tag{1.4}$$

where F is the cdf of $N(\mu, \sigma^2)$. Now if $Y \sim N(\mu, \sigma^2)$ then

$$F(y) = P(Y \leqslant y) = P\left(\frac{Y-\mu}{\sigma} \leqslant \frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

where Φ is the cdf of $N(0,1)$. So (1.4) is

$$\Phi\left(\frac{x_{(k)}-\mu}{\sigma}\right) \approx \frac{k}{n+1}.$$

Hence

$$x_{(k)} \approx \sigma\Phi^{-1}\left(\frac{k}{n+1}\right) + \mu.$$

So we can plot $x_{(k)}$ against $\Phi^{-1}\left(\frac{k}{n+1}\right)$, $k = 1, \ldots, n$, and see if the points lie on an approximate straight line (with gradient σ, intercept μ).

SLIDES. Normal Q-Q plot slides go here.

Exponential Q-Q plot

The exponential distrbution with mean μ has cdf $F(x) = 1 - e^{-x/\mu}$, $x > 0$. If data $x_1, \ldots, x_n$ have this distribution (with μ unknown) then we expect, roughly,

$$F(x_{(k)}) \approx \frac{k}{n+1}$$

with F as above. So

$$1 - e^{-x_{(k)}/\mu} \approx \frac{k}{n+1}$$

hence

$$x_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right).$$

So we can plot $x_{(k)}$ against $-\log(1 - \frac{k}{n+1})$ and see if the points lie on an approximate straight line (with gradient μ, intercept 0).

Pareto Q-Q plot

The Pareto distribution has cdf

$$F(x) = \begin{cases} 0 & \text{if } x < \alpha \\ 1 - \left(\frac{\alpha}{x}\right)^\theta & \text{if } x \geqslant \alpha \end{cases}$$

with parameters $\alpha, \theta > 0$. If data $x_1, \ldots, x_n$ have this distribution (with α, θ unknown) then we expect, roughly,

$$F(x_{(k)}) \approx \frac{k}{n+1}$$

with F as above. So

$$1 - \left(\frac{\alpha}{x_{(k)}}\right)^\theta \approx \frac{k}{n+1}$$

hence

$$\log x_{(k)} \approx \log \alpha - \frac{1}{\theta} \log\left(1 - \frac{k}{n+1}\right).$$

So we can plot $\log x_{(k)}$ against $-\log(1 - \frac{k}{n+1})$ and see if the points lie on an approximate straight line (with gradient $1/\theta$, intercept $\log \alpha$).

SLIDES. Danish fire data slides go here.

1.5 Multivariate normal distribution

We don't need to know much about the multiarite normal distribution – an intutive picture as in the bivariate example below is enough to start with. See also Part A Probability. The reason for including it here is that the asymptotic distribution of the MLE $\widehat{\theta}(\mathbf{X})$ in Section 1.7 is multivariate normal if θ is a vector (and is univariate normal if θ is a scalar).

The univariate normal distribution has two parameters, μ and σ^2. In the multivariate case, μ and σ^2 are replaced by a vector $\boldsymbol{\mu}$ and a matrix Σ.

First let $\mathbf{Z} = (Z_1, \ldots, Z_p)$ where $Z_1, \ldots, Z_p \overset{\text{iid}}{\sim} N(0,1)$. Then the pdf of $\mathbf{Z}$ is

$$\begin{aligned} f(\mathbf{z}) &= \prod_{j=1}^{p} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2} z_j^2 \right) \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^{p} z_j^2 \right) \\ &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right), \quad \mathbf{z} \in \mathbb{R}^p. \end{aligned}$$

In this case we will write $\mathbf{Z} \sim N(\mathbf{0}, I)$ where it is understood that $\mathbf{0}$ is a p-vector of zeroes and I is the $p \times p$ identity matrix.

Now let $\boldsymbol{\mu}$ be a p-vector and Σ a $p \times p$ symmetric, positive definite matrix, and let $|\Sigma|$ denote the determinant of Σ. We say that $\mathbf{X} = (X_1, \ldots, X_p)$ has a multivariate normal (MVN) distribution with *mean vector* $\boldsymbol{\mu}$ and *covariance matrix* Σ, written $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if its pdf is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

Observe that this pdf reduces to the $N(\mathbf{0}, I)$ pdf above when we substitute $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I$.

If $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, then

- $E(X_j) = \mu_j$
- $\text{var}(X_j) = \Sigma_{jj}$ and $\text{cov}(X_j, X_k) = \Sigma_{jk}$
- if $\mathbf{a}$ is any non-random p-vector, then $\mathbf{a}^T \mathbf{X} \sim N(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \Sigma \mathbf{a})$.

We simply state these properties without proof.

Taking $\mathbf{a} = (0, \ldots, 1, \ldots, 0)$, with the 1 being in the jth place, the third result gives us that the marginal distribution of X_j is $X_j \sim N(\mu_j, \Sigma_{jj})$.

Example 1.10 (Bivariate normal distribution)**.** Suppose $p = 2$. Let $-1 < \rho < 1$ and

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

On substituting this $\boldsymbol{\mu}$ and Σ into the MVN pdf above, we find that the pdf of (X_1, X_2) is

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} (x_1^2 - 2\rho x_1 x_2 + x_2^2) \right), \quad x_1, x_2 \in \mathbb{R}.$$

Here, the marginal distribution of X_1 is $N(0,1)$, and similarly $X_2 \sim N(0,1)$. The quantity ρ is the correlation between X_1 and X_2:

$$\text{corr}(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\,\text{var}(X_2)}} = \rho.$$

Also, X_1 and X_2 are independent if and only if $\rho = 0$.

SLIDES. Bivariate normal slide goes here.

1.6 Information

Definition. In a model with scalar parameter θ and log-likelihood $\ell(\theta)$, the *observed information* $J(\theta)$ is defined by

$$J(\theta) = -\frac{d^2\ell}{d\theta^2}.$$

When $\theta = (\theta_1, \ldots, \theta_p)$ the *observed information matrix* is a $p \times p$ matrix $J(\theta)$ whose (j,k) element is

$$J(\theta)_{jk} = -\frac{\partial^2\ell}{\partial\theta_j \partial\theta_k}.$$

This matrix is symmetric.

Example 1.11. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$. Then we have

$$\begin{aligned}
\text{likelihood} \quad & L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum_i x_i}}{\prod_i x_i!} \\
\text{log-likelihood} \quad & \ell(\theta) = -n\theta + \sum_i x_i \log\theta - \log(\prod_i x_i!) \\
\text{observed information} \quad & J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = \frac{\sum_i x_i}{\theta^2}.
\end{aligned}$$

Suppose we expand $\ell(\theta)$ in a Taylor series about $\widehat{\theta}$:

$$\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + \frac{1}{2}(\theta - \widehat{\theta})^2\ell''(\widehat{\theta}).$$

Assuming $\ell'(\widehat{\theta}) = 0$ we have

$$\ell(\theta) \approx \ell(\widehat{\theta}) - \frac{1}{2}(\theta - \widehat{\theta})^2 J(\widehat{\theta}). \tag{1.5}$$

The larger $J(\widehat{\theta})$ is, the more concentrated $\ell(\theta)$ is about $\widehat{\theta}$ and the more information we have about θ. Note that $J(\theta)$ is a function of θ and in the quadratic approximation (1.5), J is evaluated at $\theta = \widehat{\theta}$.

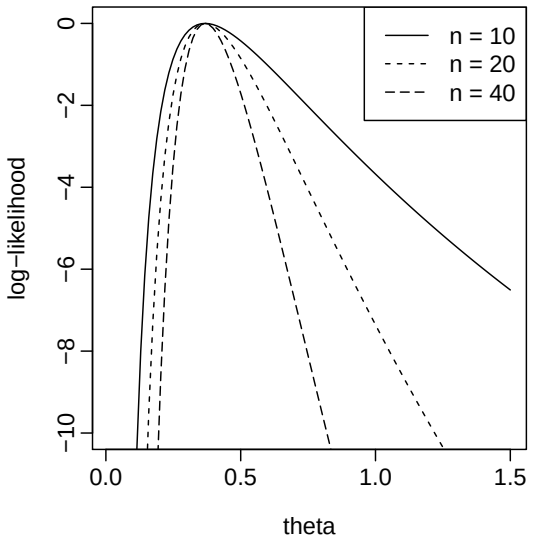

Figure 1.1. (Following Davison p102): the log-likelihood for an exponential with $\theta = \overline{x} = e^{-1}$; the curvature increases with n.

Before conducting an experiment we have no data so we cannot evaluate $J(\theta)$. But we can find its expected value.

Definition. In a model with scalar parameter θ, the *expected* or *Fisher information* is defined by

$$I(\theta) = E\bigg(- \frac{d^2\ell}{d\theta^2}\bigg).$$

When $\theta = (\theta_1, \ldots, \theta_p)$ the *expected* or *Fisher information matrix* is a $p \times p$ matrix $I(\theta)$ whose (j,k) element is

$$I(\theta)_{jk} = E\bigg(- \frac{\partial^2 \ell(\theta)}{\partial\theta_j \partial\theta_k}\bigg).$$

This matrix is symmetric.

Note:

(i) When calculating $I(\theta)$ we treat ℓ as $\ell(\theta; \mathbf{X})$ and take expectations over $\mathbf{X}$ (see the example below), e.g. in the scalar case

$$I(\theta) = E\bigg(- \frac{d^2\ell(\theta;\mathbf{X})}{d\theta^2}\bigg)$$

where the expectation is over $\mathbf{X}$.

(ii) If $X_1, \ldots, X_n$ are iid from $f(x;\theta)$ then $I(\theta) = n \times i(\theta)$ where $i(\theta)$ is the expected information in a sample of size 1. That is, in the scalar case,

$$\ell(\theta;\mathbf{X}) = \log\left(\prod_{j=1}^n f(X_j;\theta)\right) = \sum_{j=1}^n \log f(X_j;\theta)$$

and so

$$I(\theta) = \sum_{j=1}^n E\bigg(- \frac{d^2 \log f(X_j;\theta)}{d\theta^2}\bigg) = n \times i(\theta)$$

where

$$i(\theta) = E\bigg(- \frac{d^2 \log f(X_1;\theta)}{d\theta^2}\bigg).$$

Example 1.12. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ exponential with pdf $f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}$, $x \geqslant 0$. Note $E(X_i) = \theta$.

We have

$$L(\theta) = \prod_i \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum x_i/\theta}$$

$$\ell(\theta) = -n \log\theta - \frac{\sum x_i}{\theta}$$

$$J(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2} = -\frac{n}{\theta^2} + \frac{2\sum x_i}{\theta^3}.$$

To find $I(\theta)$ we treat $J(\theta)$ as a function of $\mathbf{X}$ (rather than $\mathbf{x}$), i.e. treat it as $J(\theta; \mathbf{X})$, and then take expectations. So

$$\begin{aligned} I(\theta) &= E\bigg(-\frac{n}{\theta^2} + \frac{2\sum X_i}{\theta^3} \bigg) \\ &= -\frac{n}{\theta^2} + \frac{2}{\theta^3} \sum_{i=1}^{n} E(X_i) \\ &= -\frac{n}{\theta^2} + \frac{2}{\theta^3} n\theta \quad \text{since } E(X_i) = \theta \\ &= \frac{n}{\theta^2}. \end{aligned}$$

Example 1.13. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. So we have a vector of parameters $\theta = (\mu, \sigma^2)$.

$$\begin{aligned} L(\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\bigg[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \bigg] \\ \ell(\mu, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \end{aligned}$$

Differentiating,

$$\begin{aligned} \{J(\theta)\}_{11} &= -\frac{\partial^2 \ell}{\partial \mu^2} = \frac{n}{\sigma^2} \\ \{J(\theta)\}_{22} &= -\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_i (x_i - \mu)^2 \\ \{J(\theta)\}_{12} &= -\frac{\partial^2 \ell}{\partial \mu \, \partial (\sigma^2)} = \frac{1}{\sigma^4} \sum_i (x_i - \mu). \end{aligned}$$

So

$$\begin{aligned} \{I(\theta)\}_{11} &= E\bigg(-\frac{\partial^2 \ell}{\partial \mu^2} \bigg) = \frac{n}{\sigma^2} \\ \{I(\theta)\}_{22} &= E\bigg(-\frac{\partial^2 \ell}{\partial (\sigma^2)^2} \bigg) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_i E[(X_i - \mu)^2] \\ &= -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} n \operatorname{var}(X_i) = -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} n\sigma^2 = \frac{n}{2\sigma^4} \\ \{I(\theta)\}_{12} &= E\bigg(-\frac{\partial^2 \ell}{\partial \mu \, \partial (\sigma^2)} \bigg) = \frac{1}{\sigma^4} \sum_i E(X_i - \mu) = 0 \end{aligned}$$

and so

$$I(\theta) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}.$$

1.7 Properties of MLEs

MLEs are intuitively appealing and have good properties. For example, subject to certain regularity conditions,

- we'll see shortly that

$$\widehat{\theta} \overset{\text{D}}{\approx} N(\theta, I(\theta)^{-1}) \tag{1.6}$$

  where this is an asymptotic distribution, i.e. a good approx for large n
- the asymptotic distribution (1.6) is centered at θ, i.e. $\widehat{\theta}$ is "asymptotically unbiased"
- $\widehat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$, i.e. we have convergence in probability to the true parameter value θ (we won't prove this)
- the asymptotic variance in (1.6) is as small as possible (for an unbiased estimator, see the SB2a course in Part B).

Invariance property of MLEs

MLEs also have an invariance property.

Example 1.14. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$. We may want to estimate $\psi = P(X_1 = 0)$
$= e^{-\theta}$. What is the MLE of ψ?

More generally suppose we would like to estimate $\psi = g(\theta)$ where g is a 1–1 function. What is the MLE of ψ?

In terms of ψ the likelihood $L^*(\psi)$ is given by

$$\begin{aligned} L^*(\psi) &= \prod_{i=1}^{n} f(x_i; g^{-1}(\psi)) \\ &= L(g^{-1}(\psi)). \end{aligned}$$

So

$$\begin{aligned} \sup_{\psi} L^*(\psi) &= \sup_{\psi} L(g^{-1}(\psi)) \\ &= \sup_{\theta} L(\theta) \end{aligned}$$

and the maximum of $L^*(\psi)$ is attained at the ψ such that $g^{-1}(\psi) = \widehat{\theta}$, i.e. at $\psi = g(\widehat{\theta})$. That is: $\widehat{\psi} = g(\widehat{\theta})$.

This is known as the *invariance property* of MLEs. It holds for all g, not just for 1–1 functions.

Example 1.15 (answer)**.** We previously found that $\widehat{\theta} = \overline{x}$. So the invariance property tells us immediately that $\widehat{\psi} = e^{-\widehat{\theta}} = e^{-\overline{x}}$.

Iterative calculation of $\widehat{\theta}$

Often, but not always, $\widehat{\theta}$ satisfies the *likelihood equation*

$$\frac{d\ell}{d\theta}(\widehat{\theta}) = 0. \tag{1.7}$$

We often have to solve (1.7) numerically. One way is using Newton–Raphson.

Suppose $\theta^{(0)}$ is an initial guess for $\widehat{\theta}$. Then

$$0 = \frac{d\ell}{d\theta}(\widehat{\theta}) \approx \frac{d\ell}{d\theta}(\theta^{(0)}) + (\widehat{\theta} - \theta^{(0)})\frac{d^2\ell}{d\theta^2}(\theta^{(0)}).$$

Rearranging,

$$\widehat{\theta} \approx \theta^{(0)} + \frac{U(\theta^{(0)})}{J(\theta^{(0)})}$$

where $U(\theta) = \frac{d\ell}{d\theta}$ is called the *score function*.

So we can start at $\theta^{(0)}$ and iterate to find $\widehat{\theta}$ using

$$\theta^{(n+1)} = \theta^{(n)} + J(\theta^{(n)})^{-1}U(\theta^{(n)}), \quad n \geqslant 0. \tag{1.8}$$

An alternative is to replace $J(\theta^{(n)})$ by $I(\theta^{(n)})$ and this is known as "Fisher scoring".

The above is for a scalar parameter θ. It extends straightforwardly when θ is the vector $\theta = (\theta_1, \ldots, \theta_p)$: the iterative formula is still (1.8), though in this case $J(\theta)$ is the Fisher information *matrix* defined earlier and $U(\theta)$ is the score *vector* defined by $U(\theta) = (\frac{d\ell}{d\theta_1}, \ldots, \frac{d\ell}{d\theta_p})^T$.

Asymptotic normality of $\widehat{\theta}$

First let θ be a scalar and consider the MLE $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$, a RV. Subject to regularity conditions,

$$\{I(\theta)\}^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, 1)$$

as sample size $n \to \infty$. Note that on the LHS, both $I(\theta)$ and $\widehat{\theta}$ depend on n. Here $\xrightarrow{D}$ means "converges in distribution".

So for large n we have

$$\widehat{\theta} \overset{\text{D}}{\approx} N(\theta, I(\theta)^{-1}). \tag{1.9}$$

The approximation (1.9) also holds when θ is a vector, we just need to remember that we have a MVN with mean vector θ and covariance matrix $I(\theta)^{-1}$.

In our sketch proof of asymptotic normality we use Slutsky's theorem (we state this theorem without proof). The notation $\xrightarrow{P}$ denotes convergence in probability.

Theorem 1.16 (Slutsky's theorem)**.** *Suppose $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$ as $n \to \infty$, where c is constant. Then (i) $X_n + Y_n \xrightarrow{D} X + c$, (ii) $X_nY_n \xrightarrow{D} cX$, (iii) $X_n/Y_n \xrightarrow{D} X/c$ if $c \neq 0$.*

Sketch proof of asymptotic normality (θ scalar). Assume $\widehat{\theta}$ solves

$$\frac{d\ell}{d\theta}(\widehat{\theta}) = 0.$$

Then

$$\begin{aligned} 0 = \frac{d\ell}{d\theta}(\widehat{\theta}) &\approx \frac{d\ell(\theta)}{d\theta} + (\widehat{\theta} - \theta)\frac{d^2\ell(\theta)}{d\theta^2} \\ &= U(\theta) - (\widehat{\theta} - \theta)J(\theta). \end{aligned}$$

So

$$\{I(\theta)\}^{1/2}(\widehat{\theta}-\theta) \approx \{I(\theta)\}^{1/2}\frac{U(\theta)}{J(\theta)} = \frac{U(\theta)/\{I(\theta)\}^{1/2}}{J(\theta)/I(\theta)}. \tag{1.10}$$

First consider the numerator in (1.10): $U(\theta) = \frac{d\ell}{d\theta}$ and $\ell(\theta) = \sum_{j=1}^{n} \log f(X_j;\theta)$, so

$$U(\theta) = \sum_{j=1}^{n} U_j(\theta)$$

where $U_j(\theta) = \frac{d}{d\theta}\log f(X_j;\theta)$ are iid for $j = 1,\ldots,n$.

Now $1 = \int_{-\infty}^{\infty} f(x;\theta)\,dx$. This integral, and the ones below, are all over the interval $(-\infty,\infty)$. So differentiating once wrt θ, and then again,

$$0 = \int \frac{df}{d\theta}\,dx = \int \left(\frac{d}{d\theta}\log f\right) f\,dx \tag{1.11}$$

$$0 = \int \left(\frac{d^2}{d\theta^2}\log f\right) f\,dx + \int \left(\frac{d}{d\theta}\log f\right)^2 f\,dx. \tag{1.12}$$

From (1.11), $0 = E(U_j)$.

From (1.12), $0 = -i(\theta) + E(U_j^2)$.

Hence

$$E(U) = 0$$
$$\mathrm{var}(U) = \sum_{j=1}^{n} \mathrm{var}(U_j) = ni(\theta) = I(\theta).$$

So, applying the CLT to the sum $U = \sum_{j=1}^{n} U_j$,

$$\frac{U(\theta)}{\{I(\theta)\}^{1/2}} = \frac{\sum_{j=1}^{n} U_j}{\{\mathrm{var}(\sum_1^n U_j)\}^{1/2}} \xrightarrow{D} N(0,1) \quad \text{as } n\to\infty. \tag{1.13}$$

Next consider the denominator in (1.10): let $Y_j = \frac{d^2}{d\theta^2}\log f(X_j;\theta)$ and $\mu_Y = E(Y_j)$. Then

$$\frac{J(\theta)}{I(\theta)} = \frac{\sum_1^n Y_j}{n\mu_Y} = \frac{\overline{Y}}{\mu_Y}.$$

By the weak law of large numbers (Prelims/Part A Probability), $\overline{Y}$ converges in probability to μ_Y as $n\to\infty$, written $\overline{Y} \xrightarrow{P} \mu_Y$. Hence

$$\frac{J(\theta)}{I(\theta)} \xrightarrow{P} 1 \quad \text{as } n\to\infty. \tag{1.14}$$

Putting (1.10), (1.13) and (1.14) together using Slutsky's theorem (part (iii)), we have

$$\{I(\theta)\}^{1/2}(\widehat{\theta}-\theta) \xrightarrow{D} N(0,1) \quad \text{as } n\to\infty. \qquad \square$$

The regularity conditions required for the proof include:

- the true value of θ is in the interior of the parameter space Θ
- the MLE is given by the solution of the likelihood equation
- we can differentiate sufficiently often wrt θ
- we can interchange differentiation wrt θ and integration over x.

This means that cases where the set $\{x : f(x;\theta) > 0\}$ depends on θ are excluded. E.g. the result does not apply to the uniform $U(0,\theta)$ distribution since the range $0 < x < \theta$ on which $f > 0$ depends on θ.

2 Confidence Intervals

Introduction

We start with a recap, some of it brief, from Prelims.

Any estimate, maybe a maximum likelihood estimate $\widehat{\theta} = \widehat{\theta}(\mathbf{x})$, is a *point estimate*, i.e. just a number. We would like to assess how accurate/precise such an estimate is.

Definition. Let $0 < \alpha < 1$. The random interval $(a(\mathbf{X}), b(\mathbf{X}))$ is called a $100(1-\alpha)\%$ *confidence interval (CI) for* θ if

$$P(a(\mathbf{X}) < \theta < b(\mathbf{X})) = 1 - \alpha.$$

Note: $a(\mathbf{X})$ and $b(\mathbf{X})$ are not allowed to depend on θ.

We also call $(a(\mathbf{X}), b(\mathbf{X}))$ a CI with *confidence level* $1-\alpha$.

The random interval $(a(\mathbf{X}), b(\mathbf{X}))$ is an *interval estimator*. The corresponding *interval estimate* is $(a(\mathbf{x}), b(\mathbf{x}))$, which is a numerical interval – it is just the numerical interval for a particular set of data.

In words: the interval $(a(\mathbf{X}), b(\mathbf{X}))$ traps θ with probability $1-\alpha$.

Warning: the interval $(a(\mathbf{X}), b(\mathbf{X}))$ is *random* and θ is *fixed*.

Most commonly people use 95% confidence intervals, i.e. $\alpha = 0.05$.

Interpretation: if we repeat an experiment many times, and construct a CI each time, then (approx) 95% of our intervals will contain the true value θ (i.e. we imagine "repeated sampling").

Notation: for $\alpha \in (0,1)$ let z_α be such that $P(N(0,1) > z_\alpha) = \alpha$, i.e. $1 - \Phi(z_\alpha) = \alpha$ and so $z_\alpha = \Phi^{-1}(1-\alpha)$.

Example 2.1. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma_0^2)$, where μ is unknown and σ_0^2 is known. Then

$$\left(\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $1-\alpha$ CI for μ. Write this interval as $\left(\overline{X} \pm z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$.

Why is this a $1-\alpha$ CI? First, recall that if $X_1, \ldots, X_n$ are independent, $X_i \sim N(\mu_i, \sigma_i^2)$, and $a_1, \ldots, a_n$ are constants, then

$$\sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i \mu_i,\ \sum_{i=1}^n a_i^2 \sigma_i^2). \tag{2.1}$$

In our example $X_i \sim N(\mu, \sigma_0^2)$, and using (2.1) we obtain $\overline{X} \sim N(\mu, \frac{\sigma_0^2}{n})$. Standardising,

$$\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0,1).$$

Hence

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

and after rearranging the inequalities we get

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha.$$

Hence our CI is $\left(\overline{X} \pm z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$. This is a *central* (*equal tail*) CI for μ.

One-sided confidence limits

With the setup of the example above, we have

$$P\left(\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} > -z_\alpha\right) = 1 - \alpha$$

so

$$P\left(\mu < \overline{X} + z_\alpha\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

and so $\left(-\infty, \overline{X} + z_\alpha\frac{\sigma_0}{\sqrt{n}}\right)$ is a "one-sided" CI, and $\overline{X} + z_\alpha\frac{\sigma_0}{\sqrt{n}}$ is called an *upper* $1-\alpha$ *confidence limit for* μ.

Similarly

$$P\left(\mu > \overline{X} - z_\alpha\frac{\sigma_0}{\sqrt{n}}\right) = 1 - \alpha$$

and $\overline{X} - z_\alpha\frac{\sigma_0}{\sqrt{n}}$ is a *lower* $1-\alpha$ *confidence limit for* μ.

2.1 CIs using CLT

There were plenty of examples of finding CIs using the CLT in Prelims, e.g. opinion polls.

SLIDES. Opinion poll slides go here.

2.2 CIs using asymptotic distribution of MLE

For large n, we have $\sqrt{I(\theta)}(\widehat{\theta} - \theta) \overset{\text{D}}{\approx} N(0,1)$. Hence

$$P\left(-z_{\alpha/2} < \sqrt{I(\theta)}(\widehat{\theta} - \theta) < z_{\alpha/2}\right) \approx 1 - \alpha. \tag{2.2}$$

Rearranging,

$$P\left(\widehat{\theta} - \frac{z_{\alpha/2}}{\sqrt{I(\theta)}} < \theta < \widehat{\theta} + \frac{z_{\alpha/2}}{\sqrt{I(\theta)}}\right) \approx 1 - \alpha.$$

In general $I(\theta)$ depends on θ, so $\widehat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I(\theta)}}$ are not suitable for $a(\mathbf{X})$ and $b(\mathbf{X})$.

Following the procedure developed in Prelims, we estimate $I(\theta)$ by $I(\widehat{\theta})$ [or by $J(\widehat{\theta})$] and so obtain an approximate CI of

$$\left(\widehat{\theta} \pm \frac{z_{\alpha/2}}{I(\widehat{\theta})^{1/2}}\right) \tag{2.3}$$

[or alternatively $\left(\widehat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{J(\widehat{\theta})}}\right)$].

[Why does replacing $I(\theta)$ by $I(\widehat{\theta})$ work?

First, we are assuming $\widehat{\theta} \xrightarrow{P} \theta$ and that $I(\theta)$ is continuous, hence $\left(\frac{I(\widehat{\theta})}{I(\theta)}\right)^{1/2} \xrightarrow{P} 1$. (Results of this type, but maybe not this exact one, are part of Part A Probability.) So we have

$$I(\widehat{\theta})^{1/2}(\widehat{\theta}-\theta) = \left(\frac{I(\widehat{\theta})}{I(\theta)}\right)^{1/2} \times I(\theta)^{1/2}(\widehat{\theta}-\theta)$$

where in the product on the RHS the first term is converging to 1 in probability and the second term is converging to $N(0,1)$ in distribution. Hence by Slutsky's Theorem part (ii) the LHS converges in distribution to $1 \times N(0,1)$, i.e.

$$I(\widehat{\theta})^{1/2}(\widehat{\theta}-\theta) \xrightarrow{D} N(0,1). \tag{2.4}$$

Result (2.4) tells us that (2.2) holds with $I(\theta)$ replaced by $I(\widehat{\theta})$. Then the same rearrangement as that following (2.2) leads to the CI (2.3).]

Example 2.2. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. Then $\widehat{\theta} = \overline{X}$ and $I(\theta) = \frac{n}{\theta(1-\theta)}$ and the interval (2.3) is $\left(\widehat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right)$.

Suppose $n = 30$, $\sum x_i = 5$. Then the above formula gives a 99% interval of $(-0.008, 0.342)$, i.e. the interval contains negative values even though we know $\theta > 0$!

We can avoid negative values by reparametrising the problem as follows. Let $\psi = g(\theta) = \log\frac{\theta}{1-\theta}$, so ψ is the "log odds". Since $\theta \in (0,1)$ we have $\psi \in (-\infty, \infty)$, so using a normal approx can't produce impossible ψ values. Now $\widehat{\theta} \overset{\text{D}}{\approx} N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$ and the delta method gives

$$\begin{aligned}
\widehat{\psi} &\overset{\text{D}}{\approx} N\left(\psi, \frac{\theta(1-\theta)}{n} g'(\theta)^2\right) \\
&\sim N\left(\psi, \frac{1}{n\theta(1-\theta)}\right) \\
&\sim N\left(\psi, \frac{(1+e^{\psi})^2}{ne^{\psi}}\right).
\end{aligned} \tag{2.5}$$

We can use (2.5) to find an approx $1-\alpha$ CI for ψ, say (ψ_1, ψ_2), i.e. $P(\psi_1 < \psi < \psi_2) \approx 1-\alpha$. Then, since $\theta = \frac{e^{\psi}}{1+e^{\psi}}$,

$$P\left(\frac{e^{\psi_1}}{1+e^{\psi_1}} < \theta < \frac{e^{\psi_2}}{1+e^{\psi_2}}\right) = P(\psi_1 < \psi < \psi_2) \approx 1-\alpha.$$

This CI for θ definitely won't contain negative values.

2.3 Distributions related to $N(0,1)$

Definition. Let $Z_1, \ldots, Z_r \overset{\text{iid}}{\sim} N(0,1)$. We say that $Y = Z_1^2 + \cdots + Z_r^2$ has the *chi-squared distribution with r degrees of freedom.* Write $Y \sim \chi^2_r$.

It is not hard to show that a χ^2_r is the same distribution as the Gamma$(\frac{r}{2}, \frac{1}{2})$ distribution, with pdf

$$f(y) = \frac{1}{2^{r/2}\Gamma(r/2)} y^{r/2-1} e^{-y/2}, \quad y > 0.$$

We won't need this pdf – what is important is that a chi-squared distribution is a sum of squared iid $N(0,1)$'s.

If $Y \sim \chi^2_r$, then $E(Y) = r$ and $\text{var}(Y) = 2r$.

If $Y_1 \sim \chi^2_r$ and $Y_2 \sim \chi^2_s$ are independent, then $Y_1 + Y_2 \sim \chi^2_{r+s}$.

SLIDES. Plot of χ^2-pdfs goes here.

Example 2.3. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Then $\frac{X_i}{\sigma} \sim N(0,1)$ and

$$\frac{\sum_{i=1}^n X_i^2}{\sigma^2} \sim \chi^2_n.$$

Hence

$$P\left(c_1 < \frac{\sum_{i=1}^n X_i^2}{\sigma^2} < c_2\right) = 1 - \alpha$$

where c_1, c_2 are such that $P(\chi^2_n < c_1) = P(\chi^2_n > c_2) = \frac{\alpha}{2}$. So

$$P\left(\frac{\sum_{i=1}^n X_i^2}{c_2} < \sigma^2 < \frac{\sum_{i=1}^n X_i^2}{c_1}\right) = 1 - \alpha$$

and we've found a $1 - \alpha$ CI for σ^2 (an exact interval).

Definition. Let $Z \sim N(0,1)$ and $Y \sim \chi^2_r$ be independent. We say that

$$T = \frac{Z}{\sqrt{Y/r}}$$

has a *(Student) t-distribution with r degrees of freedom*. Write $T \sim t_r$.

If $T \sim t_r$, then the pdf of T is

$$f(t) \propto \frac{1}{\left(1 + \frac{t^2}{r}\right)^{(r+1)/2}}, \quad -\infty < t < \infty.$$

As with the χ^2 distribution, we won't need this pdf – what is important is the definition of t_r in terms of χ^2_r and $N(0,1)$.

As $r \to \infty$, we have $t_r \xrightarrow{D} N(0,1)$.

SLIDES. Plot of t-pdfs goes here.

2.4 Independence of $\overline{X}$ and S^2 for normal samples

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$.

Consider $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean, and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$, the sample variance.

Theorem 2.4. *$\overline{X}$ and S^2 are independent and their marginal distributions are given by*

(i) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

(ii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Proof. Let $Z_i = (X_i - \mu)/\sigma$, $i = 1, \ldots, n$. Then $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} N(0,1)$, so have joint pdf

$$f(\mathbf{z}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = (2\pi)^{-n/2} e^{-\sum_i z_i^2/2}, \quad \mathbf{z} \in \mathbb{R}^n. \tag{2.6}$$

We now consider a transformation of variables from $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ to $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ (Part A Probability).

Let $\mathbf{z} = (z_1, \ldots, z_n)^T$, $\mathbf{y} = (y_1, \ldots, y_n)^T$, and let $\mathbf{y} = A\mathbf{z}$ where A is an orthogonal $n \times n$ matrix whose first row is $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$.

[A orthogonal: $A^T A = I$, where I is the $n \times n$ identity matrix.]

Since $\mathbf{z} = A^T\mathbf{y}$, we have $\partial z_i/\partial y_j = a_{ji}$ and hence the Jacobian is $J = J(y_1, \ldots, y_n) = \det(A^T)$, hence $|J| = 1$. [Since $\partial z_i/\partial y_j = a_{ji}$, the Jacobian matrix of partial derivatives is A^T, and so the Jacobian that we need is the determinant of this, i.e. $J = \det(A^T)$. Then, since $A^T A = I$, taking determinants gives $\det(A)^2 = 1$, and hence $|J| = 1$.]

We have

$$\sum_{i=1}^n y_i^2 = \mathbf{y}^T\mathbf{y} = \mathbf{z}^T A^T A\mathbf{z} = \mathbf{z}^T\mathbf{z} = \sum_{i=1}^n z_i^2. \tag{2.7}$$

Hence the pdf of $\mathbf{Y}$ is

$$\begin{aligned} g(\mathbf{y}) &= f(\mathbf{z}(\mathbf{y})) \cdot |J| \\ &= (2\pi)^{-n/2} e^{-\sum_i y_i^2/2} \cdot 1 \quad \text{using (2.6), (2.7) and } |J| = 1. \end{aligned}$$

Hence $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(0,1)$.

Now

$$Y_1 = (\text{first row of } A) \times \mathbf{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i = \sqrt{n}\,\overline{Z}$$

and then

$$\sum_{i=1}^n (Z_i - \overline{Z})^2 = \sum_{i=1}^n Z_i^2 - n\overline{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2.$$

So we have shown that

- $Y_1, \ldots, Y_n$ are independent
- $\overline{Z}$ is a function of Y_1 only

- $\sum_{i=1}^n (Z_i - \overline{Z})^2$ is a function of $Y_2, \ldots, Y_n$ only

and hence $\overline{Z}$ and $\sum_{i=1}^n (Z_i - \overline{Z})^2$ are independent.

Therefore $\overline{X}$ and S^2 are independent since $\overline{X} = \sigma\overline{Z} + \mu$ and $S^2 = \frac{\sigma^2}{n-1}\sum_{i=1}^n (Z_i - \overline{Z})^2$.

Finally,

(i) [know this from Prelims] $Y_1 \sim N(0,1)$, so $\overline{X} = \sigma\overline{Z} + \mu = \frac{\sigma}{\sqrt{n}} Y_1 + \mu \sim N(\mu, \frac{\sigma^2}{n})$.

(ii) $\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n (Z_i - \overline{Z})^2 = \sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$. □

So for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ we have, independently,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{2.8}$$

Using $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ as the $N(0,1)$ and $\frac{(n-1)S^2}{\sigma^2}$ as the χ^2 in the definition of a t-distribution, we obtain the important result that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{2.9}$$

(since the unknown σ in the numerator and denominator cancels).

Observe that estimating σ by S takes us from the $N(0,1)$ distribution in (2.8) to the t_{n-1} distribution in (2.9).

The quantity T defined by $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ is called a *pivotal quantity* or *pivot*. In general, a *pivot* is a function of $\mathbf{X}$ and the parameter θ whose distribution does not depend on θ. In the case of T, we have $\theta = (\mu, \sigma^2)$ and the distribution of T is t_{n-1}. In the example below we use T to find a CI for μ when σ^2 is unknown.

We can get other exact CIs in similar ways. From part (ii) of the theorem, $\frac{(n-1)S^2}{\sigma^2}$ is also a pivot (with a χ^2_{n-1} distribution). We can use it to find CI for σ^2 (see Problem Sheet 2).

Example 2.5. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$. Since $\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ we have

$$P\left(-t_{n-1}(\tfrac{\alpha}{2}) < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{n-1}(\tfrac{\alpha}{2}) \right) = 1 - \alpha$$

where $t_{n-1}(\frac{\alpha}{2})$ is the constant such that $P\big(t_{n-1} > t_{n-1}(\frac{\alpha}{2})\big) = \frac{\alpha}{2}$.

Rearranging the inequalities,

$$P\left(\overline{X} - t_{n-1}(\tfrac{\alpha}{2})\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1}(\tfrac{\alpha}{2})\frac{S}{\sqrt{n}} \right) = 1 - \alpha.$$

Hence $\big(\overline{X} \pm t_{n-1}(\frac{\alpha}{2})\frac{S}{\sqrt{n}}\big)$ is a $1-\alpha$ CI for μ.

SLIDES. Sleep data slides go here.

3 Hypothesis Testing

3.1 Introductory example: t-test (Sleep data)

Consider the number of hours of sleep gained, given a low dose of the drug, by the 10 patients:

$$0.7,\ -1.6,\ -0.2,\ -1.2,\ -0.1,\ 3.4,\ 3.7,\ 0.8,\ 0.0,\ 2.0.$$

Do the data support the conclusion that (a low dose of) the drug makes people sleep more, or not?

- We will start from the default position that the drug has no effect,
- and we will only reject this default position if the data contain "sufficient evidence" for us to reject it.

So we would like to consider

(i) the "null hypothesis" that the drug has no effect, and

(ii) the "alternative hypothesis" that the drug makes people sleep more.

We will denote the "null hypothesis" by H_0, and the "alternative hypothesis" by H_1.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown. (Recall from the sleep data slides, end of Section 2.4, that a normality assumption for the sleep data looked reasonable).

We interpret H_0 and H_1 as follows:

- H_0 says that "$\mu = \mu_0$ (and σ^2 is unknown)"
- H_1 says that "$\mu > \mu_0$ (and σ^2 is unknown)"

where $\mu_0 = 0$ for the sleep data example, but μ_0 might be non-zero in other examples.

Let

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

The idea is that a small/moderate value of t_{obs} is consistent with H_0 (here "small" includes negative values of t_{obs}). Whereas a very large value of t_{obs} is not consistent with H_0 and points us towards H_1 – since $\overline{x}$, and hence t_{obs}, will tend to be larger under H_1 as $\mu > \mu_0$ under H_1.

For the sleep data, $t_{\text{obs}} = 1.326$. [$\overline{x} = 0.75$, $\mu_0 = 0$, $s^2 = 3.2$, $n = 10$.] Is this t_{obs} large?

Let

$$t(\mathbf{X}) = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

If H_0 is true then $t(\mathbf{X}) \sim t_{n-1}$. So if H_0 is true then the probability of observing a value of $t(\mathbf{X})$ of t_{obs} or more is

$$\begin{aligned} p &= P(t(\mathbf{X}) \geqslant t_{\text{obs}}) \\ &= P(t_9 \geqslant 1.326) \\ &= 0.109. \end{aligned}$$

[The value of $P(t_9 \geqslant 1.326)$ can be obtained using R by typing
`1 - pt(1.326, 9)`
i.e. $1 - F_9(1.326)$, where F_9 is the cdf of a t_9 distribution. Alternatively, from statistical tables, $P(t_9 \leqslant 1.383) = 0.9$ and so $P(t_9 \geqslant 1.326)$ is just a little more than 0.1. Knowing that p is a bit more than 0.1 is accurate enough for us here.]

This value of p is called the *p-value* or *significance level*.

The value $p = 0.109$ is not particularly small. Assuming H_0 is true, we'd observe a value of $t(\mathbf{X})$ of at least 1.326 over 10% of the time, which is not a particularly unlikely occurrence. So we do not have much evidence to reject H_0, so we'll retain H_0, our conclusion is that the data are consistent with H_0 being true.

We are really examining whether the data are consistent with H_0, or not. So usually we speak in terms of "rejecting H_0" or "not rejecting H_0", or of the data being "consistent with H_0" or "not consistent with H_0" (rather than "accepting H_0" or "accepting H_1").

Wasserman (2005) puts it this way: "Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that [they are] guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0."

The other half of the sleep data is the number of hours of sleep gained, by the same 10 patients, following a normal dose of the drug:

$$1.9,\ 0.8,\ 1.1,\ 0.1,\ -0.1,\ 4.4,\ 5.5,\ 1.6,\ 4.6,\ 3.4.$$

Is there evidence that a normal dose of the drug makes people sleep more (than not taking a drug at all), or not?

Consider the same assumptions about $X_1, \ldots, X_n$ and the same H_0 and H_1.

This time we have

$$t_{\text{obs}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = 3.68.$$

[$\overline{x} = 2.33$, $\mu_0 = 0$, $s^2 = 4.0$, $n = 10$.]

If H_0 is true, then the probability of observing a value of $t(\mathbf{X})$ of 3.68 or more is

$$\begin{aligned} p &= P(t(\mathbf{X}) \geqslant 3.68) \\ &= P(t_9 \geqslant 3.68) \\ &= 0.0025. \end{aligned}$$

This value of p is very small. Assuming H_0 is true, we'd observe a value of $t(\mathbf{X})$ of at least 3.68 only 0.25% of the time (i.e. a very rare event). We can conclude that there is strong evidence to reject H_0 in favour of the alternative hypothesis H_1.

How small is small for a p-value? We might say something like:

$p < 0.01$	very strong evidence against H_0
$0.01 < p < 0.05$	strong evidence against H_0
$0.05 < p < 0.1$	weak evidence against H_0
$0.1 < p$	litte or no evidence against H_0

[This table from Wasserman (2005).]

One-sided and two-sided alternative hypotheses

The alternative hypothesis $H_1 : \mu > \mu_0$ is a *one-sided* alternative. The larger t_{obs} is, the more evidence we have for rejecting H_0.

Consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$. This alternative H_1 is also one-sided, and the p-value would be $p = P(t_{n-1} \leqslant t_{\text{obs}})$.

A different type of alternative hypothesis is $H_1 : \mu \neq \mu_0$. This is a *two-sided* alternative. If t_{obs} was very large, i.e. very positive, then that would provide evidence to reject H_0. Similarly if t_{obs} was very small, i.e. very negative, then that would also provide evidence to reject H_0. Let

$$t_0 = |t_{\text{obs}}| = \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right|.$$

The p-value for a test of $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu \neq \mu_0$ is the probability, under H_0, that $t(\mathbf{X})$ takes a value at least as extreme as t_{obs}, i.e. the p-value is

$$\begin{aligned} p &= P(|t(\mathbf{X})| \geqslant t_0) \\ &= P(t(\mathbf{X}) \geqslant t_0) + P(t(\mathbf{X}) \leqslant -t_0) \\ &= 2P(t(\mathbf{X}) \geqslant t_0). \end{aligned}$$

Note: this p-value, and the other p-values above, are all calculated under the assumption that H_0 is true. In future we will write things like $p = P(t(\mathbf{X}) \geqslant t_{\text{obs}} \,|\, H_0)$ or $p = P(|t(\mathbf{X})| \geqslant t_0 \,|\, H_0)$ to indicate this.

3.2 Tests for normally distributed samples

Example 3.1 (z-test)**.** Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ is unknown and where $\sigma^2 = \sigma_0^2$ is known.

Suppose we wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. Then we can use the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma_0/\sqrt{n}}.$$

If H_0 is true then $Z \sim N(0, 1)$.

Let

$$z_{\text{obs}} = \frac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}}.$$

A large value of z_{obs} casts doubt on the validity of H_0 and indicates a departure from H_0 in the direction of H_1. So the p-value for testing H_0 against H_1 is

$$\begin{aligned} p &= P(Z \geqslant z_{\text{obs}} \,|\, H_0) \\ &= P(N(0,1) \geqslant z_{\text{obs}}) \\ &= 1 - \Phi(z_{\text{obs}}). \end{aligned}$$

The z-test of $H_0 : \mu = \mu_0$ against $H_1' : \mu < \mu_0$ is similar but this time a small, i.e. very negative, value of z_{obs} casts doubt on H_0 (in the direction of H_1'). So the p-value is

$$\begin{aligned} p' &= P(Z \leqslant z_{\text{obs}} \,|\, H_0) \\ &= P(N(0,1) \leqslant z_{\text{obs}}) \\ &= \Phi(z_{\text{obs}}). \end{aligned}$$

Finally, consider testing $H_0 : \mu = \mu_0$ against $H_1'' : \mu \neq \mu_0$. Let $z_0 = |z_{\text{obs}}|$. A large value of z_0 indicates a departure from H_0 (in the direction of H_1''), so the p-value is

$$\begin{aligned} p'' &= P(|Z| \geqslant z_0 \,|\, H_0) \\ &= P(N(0,1) \geqslant z_0) + P(N(0,1) \leqslant -z_0) \\ &= 2(1 - \Phi(z_{\text{obs}})). \end{aligned}$$

Example 3.2 (t-test)**.** This example is really a repeat of Section 3.1. But it is included to show the similarities with the previous example (z-test). The setup is as for the z-test except that here σ^2 is unknown, the test statistic T below replaces Z, and the cdf of a t_{n-1} distribution replaces Φ.

Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and assume both μ and σ^2 are unknown. Consider testing $H_0 : \mu = \mu_0$ (and σ^2 unknown) against three possible alternatives:

(i) $H_1 : \mu > \mu_0$ (and σ^2 unknown)

(ii) $H_1' : \mu < \mu_0$ (and σ^2 unknown)

(iii) $H_1'' : \mu \neq \mu_0$ (and σ^2 unknown).

We can use the test statistic

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}.$$

If H_0 is true then $T \sim t_{n-1}$.

Let $t_{\text{obs}} = t(\mathbf{x}) = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$ and $t_0 = |t_{\text{obs}}|$. Then, as in Section 3.1,

(i) for the test of H_0 against H_1, the p-value is $P(t_{n-1} \geqslant t_{\text{obs}})$

(ii) for the test of H_0 against H_1', the p-value is $P(t_{n-1} \leqslant t_{\text{obs}})$

(iii) for the test of H_0 against H_1'', the p-value is $2P(t_{n-1} \geqslant t_0)$.

SLIDES. Slides on normal tests go here.

3.3 Hypothesis testing and confidence intervals

SLIDES. Slide on hypothesis testing and confidence intervals goes here.

The maize data example suggests a connection between hypothesis testing and confidence intervals: the connection appears to be that the p-value of a test of $H_0 : \mu = \mu_0$ being less than α is equivalent to the corresponding $100(1 - \alpha)\%$ confidence interval not containing μ_0.

We will illustrate the connection with a proof of this in one particular case.

Example 3.3. Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown.

We have already seen that:

(i) a $100(1 - \alpha)\%$ confidence interval for μ is given by

$$\left(\overline{x} \pm \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2) \right) \tag{3.1}$$

(ii) for the t-test of $\mu = \mu_0$ against $\mu \neq \mu_0$, the p-value is $p = P(|t_{n-1}| \geqslant t_0)$, where $t_0 = |\frac{\overline{x}-\mu_0}{s/\sqrt{n}}|$.

So

$$\begin{aligned} p < \alpha &\iff t_0 > t_{n-1}(\alpha/2) \\ &\iff \frac{\overline{x}-\mu_0}{s/\sqrt{n}} > t_{n-1}(\alpha/2) \quad \text{or} \quad \frac{\overline{x}-\mu_0}{s/\sqrt{n}} < -t_{n-1}(\alpha/2) \\ &\iff \mu_0 < \overline{x} - \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2) \quad \text{or} \quad \mu_0 > \overline{x} + \frac{s}{\sqrt{n}} t_{n-1}(\alpha/2). \end{aligned}$$

That is, $p < \alpha$ if and only if the CI (3.1) does not contain μ_0.

3.4 Hypothesis testing general setup

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$ where $\theta \in \Theta$ is a scalar or vector parameter.

Suppose we are interested in testing

- the *null hypothesis* $H_0 : \theta \in \Theta_0$
- against the *alternative hypothesis* $H_1 : \theta \in \Theta_1$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and possibly but not necessarily $\Theta_0 \cup \Theta_1 = \Theta$.

Suppose we can construct a *test statistic* $t(\mathbf{X})$ such that large values of $t(\mathbf{X})$ cast doubt on the validity of H_0 and indicate a departure from H_0 in the direction of H_1. Let $t_{\text{obs}} = t(\mathbf{x})$, the value of $t(\mathbf{X})$ actually observed. Then the *p-value* or *significance level* is $p = P(t(\mathbf{X}) \geqslant t_{\text{obs}} \,|\, H_0)$.

Note: p is calculated under the assumption that H_0 is true. We write $P(\ldots \,|\, H_0)$ to indicate this.

A small value of p corresponds to a value of t_{obs} unlikely to arise under H_0 and is an indicator that H_0 and the data $\mathbf{x}$ are inconsistent.

Warning: The p-value is NOT the probability that H_0 is true. Rather: assuming H_0 is true, it is the probability of $t(\mathbf{X})$ taking a value at least as extreme as the value t_{obs} that we actually observed.

A hypothesis which completely determines f is called *simple*, e.g. $\theta = \theta_0$. Otherwise a hypothesis is called *composite*, e.g. $\theta > \theta_0$ or $\theta \neq \theta_0$. Here "completely determines" means a hypothesis corresponds to a single function f, not a family of such functions. So e.g. saying that something is an "exponential distribution" does *not* completely determine f, it only determines the family to which f belongs, there are infinitely many members of that family, one for each parameter value $\theta \in (0, \infty)$.

Example 3.4. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 both unknown, so $\theta = (\mu, \sigma^2)$. Then $H_0 : \mu = \mu_0$ is a composite hypothesis because it corresponds to $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0,\ \sigma^2 > 0\}$ and this set contains more than one value of θ.

In a case like this σ^2 is called a *nuisance parameter*.

Suppose we want to make a definite decision: i.e. either reject H_0, or don't reject H_0. Then we can define our test in terms of a *critical region* $C \subset \mathbb{R}^n$ such that:

- if $\mathbf{x} \in C$ then we reject H_0
- if $\mathbf{x} \notin C$ then we don't reject H_0.

Errors in hypothesis testing

There are two possible types of error:

- type I error: rejecting H_0 when H_0 is true
- type II error: not rejecting H_0 when H_0 is false.

	don't reject H_0	reject H_0
H_0 true	✓	type I error
H_0 false	type II error	✓

Here, ✓ means "correct decision made".

First, consider testing the simple $H_0 : \theta = \theta_0$ against the simple $H_1 : \theta = \theta_1$.

The *type I error probability* α, also called the *size* of the test, is defined by

$$\begin{aligned} \alpha &= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\ &= P(\mathbf{X} \in C \,|\, \theta_0). \end{aligned}$$

The *type II error probability* β is defined by

$$\begin{aligned} \beta &= P(\text{don't reject } H_0 \,|\, H_1 \text{ true}) \\ &= P(\mathbf{X} \notin C \,|\, \theta_1) \end{aligned}$$

and $1 - \beta = P(\text{reject } H_0 \,|\, H_1 \text{ true})$ is called the *power* of the test.

Note: power $= 1 - \beta = P(\mathbf{X} \in C \,|\, H_1)$ which is the probability of correctly detecting that H_0 is false.

If H_0 is composite, $H_0 : \theta \in \Theta_0$ say, then we define the size of the test by

$$\alpha = \sup_{\theta \in \Theta_0} P(\mathbf{X} \in C \,|\, \theta).$$

If H_1 is composite then we have to define the power as a function of θ: the *power function* $w(\theta)$ is defined by

$$\begin{aligned} w(\theta) &= P(\text{reject } H_0 \,|\, \theta \text{ is the true value}) \\ &= P(\mathbf{X} \in C \,|\, \theta). \end{aligned}$$

Ideally, we'd like $w(\theta)$ to be near 1 for H_1-values of θ (i.e. for $\theta \in \Theta_1$) and to be near 0 for H_0-values of θ (i.e. for $\theta \in \Theta_0$).

Warning: A large p-value is not strong evidence in favour of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.

3.5 The Neyman–Pearson lemma

Consider testing a simple null hypothesis against a simple alternative:

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta = \theta_1. \tag{$*$}$$

Suppose we choose a small type I error probability α (e.g. $\alpha = 0.05$). Then among tests of this size we could aim to minimise the type II error probability β, i.e. maximise the power $1 - \beta$. Note that if we do this, as in the Neyman–Pearson lemma below, then H_0 and H_1 are treated asymmetrically.

Theorem 3.5 (Neyman–Pearson lemma)**.** *Let $L(\theta; \mathbf{x})$ be the likelihood. Define the critical region C by*

$$C = \left\{ \mathbf{x} : \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})} \leqslant k \right\}$$

and suppose the constants k and α are such that $P(\mathbf{X} \in C \,|\, H_0) = \alpha$. Then among all tests of $(*)$ of size $\leqslant \alpha$, the test with critical region C has maximum power.

Equivalently: the test with critical region C minimises the probability of a type II error.

Proof. [Proof below for continuous RVs, for discrete RVs replace $\int$ by $\sum$.]

Consider any test of size $\leqslant \alpha$, with a critical region A say. Then we have

$$P(\mathbf{X} \in A \,|\, H_0) \leqslant \alpha. \tag{3.2}$$

The critical region C is one possible A. Define

$$\phi_A(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A \\ 0 & \text{otherwise} \end{cases}$$

and let C and k be as in the statement of the theorem. Then

$$0 \leqslant \big\{ \phi_C(\mathbf{x}) - \phi_A(\mathbf{x}) \big\} \big[L(\theta_1; \mathbf{x}) - \tfrac{1}{k} L(\theta_0; \mathbf{x}) \big]$$

since $\{\dots\}$ and $[\dots]$ are both $\geqslant 0$ if $\mathbf{x} \in C$, and both $\leqslant 0$ if $\mathbf{x} \notin C$.

[That is, if $\mathbf{x} \in C$, then $\phi_C(\mathbf{x}) - \phi_A(\mathbf{x}) = 1 - \phi_A(\mathbf{x}) \geqslant 0$, and $L(\theta_1; \mathbf{x}) - \frac{1}{k} L(\theta_0; \mathbf{x}) \geqslant 0$ from the definition of C. Similarly, if $\mathbf{x} \notin C$, then $\phi_C(\mathbf{x}) - \phi_A(\mathbf{x}) = -\phi_A(\mathbf{x}) \leqslant 0$, and $L(\theta_1; \mathbf{x}) - \frac{1}{k} L(\theta_0; \mathbf{x}) < 0$ from the definition of C.]

So

$$\begin{aligned} 0 &\leqslant \int_{\mathbb{R}^n} \big\{ \phi_C(\mathbf{x}) - \phi_A(\mathbf{x}) \big\} \big[L(\theta_1; \mathbf{x}) - \tfrac{1}{k} L(\theta_0; \mathbf{x}) \big] \, d\mathbf{x} \\ &= P(\mathbf{X} \in C \,|\, H_1) - P(\mathbf{X} \in A \,|\, H_1) \\ &\quad - \tfrac{1}{k} \big[P(\mathbf{X} \in C \,|\, H_0) - P(\mathbf{X} \in A \,|\, H_0) \big]. \end{aligned} \tag{3.3}$$

Now $P(\mathbf{X} \in C \,|\, H_0) = \alpha$, so $[\dots]$ in (3.3) is $\geqslant 0$ by (3.2). Hence

$$0 \leqslant P(\mathbf{X} \in C \,|\, H_1) - P(\mathbf{X} \in A \,|\, H_1).$$

Thus $P(\mathbf{X} \in C \,|\, H_1) \geqslant P(\mathbf{X} \in A \,|\, H_1)$, i.e. the power of the test is maximised by using critical region C. □

The test given by the NP lemma is the most powerful test of $(*)$. Its critical region C is called the most powerful critical region or best critical region.

Example 3.6. Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma_0^2)$ where σ_0^2 is known. Find the most powerful test of size α of $H_0 : \mu = 0$ against $H_1 : \mu = \mu_1$, where $\mu_1 > 0$.

Note: H_1 is the hypothesis that μ takes one particular value – the value μ_1. Also, we are assuming that $\sigma^2 = \sigma_0^2$ is known. So H_1 is a simple hypothesis and the NP lemma applies.

For a general value of μ, the likelihood is

$$L(\mu; \mathbf{x}) = (2\pi\sigma_0^2)^{-n/2} \exp\left[-\frac{1}{2\sigma_0^2} \sum (x_i - \mu)^2 \right].$$

Step 1: by the NP lemma, the most powerful test is of the form

$$\text{reject } H_0 \iff \frac{L(0; \mathbf{x})}{L(\mu_1; \mathbf{x})} \leqslant k_1$$

where k_1 is a constant (i.e. does not depend on $\mathbf{x}$). Now

$$\begin{aligned}
\frac{L(0; \mathbf{x})}{L(\mu_1; \mathbf{x})} \leqslant k_1 &\iff \exp\left[-\frac{1}{2\sigma_0^2} \sum x_i^2 \right] \exp\left[\frac{1}{2\sigma_0^2} \sum (x_i - \mu_1)^2 \right] \leqslant k_1 \\
&\iff \exp\left[\frac{1}{2\sigma_0^2} \left(-\sum x_i^2 + \sum x_i^2 - 2\mu_1 \sum x_i + n\mu_1^2 \right) \right] \leqslant k_1 \\
&\iff \frac{1}{2\sigma_0^2}(-2\mu_1 n\overline{x} + n\mu_1^2) \leqslant k_2 \\
&\iff -\mu_1 \overline{x} \leqslant k_3 \\
&\iff \overline{x} \geqslant c
\end{aligned}$$

where k_1, k_2, k_3, c are constants that don't depend on $\mathbf{x}$ – all that matters is that they don't depend on $\mathbf{x}$, they can depend on $n, \sigma_0^2, \ldots$. So the form of the critical region is $\{\mathbf{x} : \overline{x} \geqslant c\}$. [Note: if the alternative hypothesis was $H_1 : \mu = \mu_1$, where $\mu_1 < 0$, then the final line of our iff calculation would give a critical region of the form $\{\mathbf{x} : \overline{x} \leqslant c\}$.]

Step 2: we now choose c so that the test has size α:

$$\begin{aligned}
\alpha &= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\
&= P(\overline{X} \geqslant c \,|\, H_0).
\end{aligned}$$

If H_0 is true then $\overline{X} \sim N(0, \sigma_0^2/n)$. So

$$\begin{aligned}
\alpha &= P(\overline{X} \geqslant c \,|\, H_0) \\
&= P\left(\frac{\overline{X}}{\sigma_0/\sqrt{n}} \geqslant \frac{c}{\sigma_0/\sqrt{n}} \,\middle|\, H_0 \right) \\
&= P\left(N(0,1) \geqslant \frac{c}{\sigma_0/\sqrt{n}} \right)
\end{aligned}$$

and hence $\frac{c}{\sigma_0/\sqrt{n}} = z_\alpha$. So the required value of c is $c = z_\alpha \sigma_0/\sqrt{n}$ and the most powerful test has critical region

$$C = \left\{ \mathbf{x} : \overline{x} \geqslant \frac{z_\alpha \sigma_0}{\sqrt{n}} \right\}.$$

E.g. the most powerful test of size 0.05 rejects H_0 if and only if $\overline{x} \geqslant 1.64\sigma_0/\sqrt{n}$.

We now calculate the power function of this test:

$$\begin{aligned} w(\mu) &= P(\text{reject } H_0 \,|\, \mu \text{ is the true value}) \\ &= P\left(\overline{X} \geqslant \frac{z_\alpha \sigma_0}{\sqrt{n}} \,\middle|\, \mu \right) \\ &= P\left(\frac{\overline{X} - \mu}{\sigma_0/\sqrt{n}} \geqslant z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}} \,\middle|\, \mu \right) \end{aligned}$$

and if μ is the true parameter value then $\frac{\overline{X}-\mu}{\sigma_0/\sqrt{n}} \sim N(0,1)$, so

$$\begin{aligned} w(\mu) &= P\left(N(0,1) \geqslant z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}} \right) \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}} \right). \end{aligned}$$

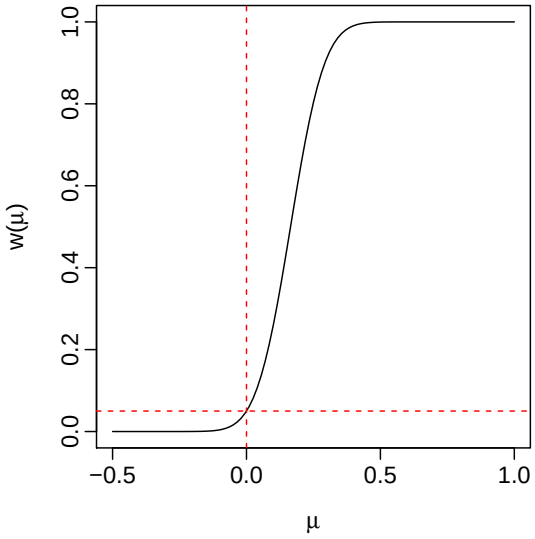

Figure 3.1. The power function $w(\mu)$ when $\alpha = 0.05$, $\sigma_0 = 1$, $n = 100$.

3.6 Uniformly most powerful tests

Consider testing the simple $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \in \Theta_1$.

Let $\theta_1 \in \Theta_1$.

When testing the simple null $\theta = \theta_0$ against the simple alternative $\theta = \theta_1$, the critical region from the NP lemma may be the same for all $\theta_1 \in \Theta_1$. If this holds then C is said to be *uniformly most powerful* (UMP) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$.

We can often find UMP tests of simple null hypotheses against one-sided alternatives, but not when testing against two-sided alternatives.

Example 3.7. In the example in Section 3.5 we have the same $C = \{\mathbf{x} : \overline{x} \geqslant z_\alpha \sigma_0/\sqrt{n}\}$ for each $\mu_1 > 0$. Hence this C gives a UMP test of $H_0 : \mu = 0$ against $H_1 : \mu > 0$.

SLIDES. Slides on insect traps example go here.

Example 3.8. In this example we will: (i) construct another UMP test, (ii) show that not all sizes α are possible, (iii) do a sample size calculation.

(i) Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$. Consider testing $H_0 : \lambda = 1$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 < 1$.

The likelihood is

$$L(\lambda; \mathbf{x}) = \prod \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}.$$

From the NP lemma, we should

$$\begin{aligned}
\text{reject } H_0 &\iff \frac{L(1;\mathbf{x})}{L(\lambda_1;\mathbf{x})} \leqslant k_1 \\
&\iff \frac{e^{-n}}{e^{-n\lambda_1}\lambda_1^{\sum x_i}} \leqslant k_1 \\
&\iff \left(\frac{1}{\lambda_1}\right)^{\sum x_i} \leqslant k_2 \\
&\iff \sum x_i \leqslant k \quad \text{since } 1/\lambda_1 > 1
\end{aligned}$$

where k, k_1, k_2 are constants that don't depend on $\mathbf{x}$.

With critical region $C = \{\mathbf{x} : \sum x_i \leqslant k\}$, the size of the test is given by

$$\begin{aligned}
\alpha &= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\
&= P\left(\sum X_i \leqslant k \,\middle|\, \lambda = 1\right).
\end{aligned}$$

This critical region C does not depend on which value of $\lambda_1 < 1$ we are considering, so it gives a UMP test of $H_0 : \lambda = 1$ against $H_1 : \lambda < 1$.

(ii) If H_0 is true, then $\sum X_i \sim \text{Poisson}(n)$. Wlog k can be an integer, and then we have

$$\begin{aligned}
\alpha &= P\left(\sum X_i \leqslant k \,\middle|\, \lambda = 1\right) \\
&= P(\text{Poisson}(n) \leqslant k) \\
&= \sum_{j=0}^{k} \frac{e^{-n}n^j}{j!}
\end{aligned}$$

So if $n = 5$ the possible values of α are $\alpha = 0.0067, 0.04, 0.12, 0.26, \ldots$, if $k = 0, 1, 2, 3, \ldots$. That is, not all sizes α are possible – this occurs because our test statistic $\sum X_i$ is a discrete RV. This discreteness issue does not affect the p-value: if we let $t_{\text{obs}} = \sum x_i$, then the p-value of data $\mathbf{x}$ is

$$p = P\left(\sum X_i \leqslant t_{\text{obs}} \,\middle|\, H_0\right) = P(\text{Poisson}(n) \leqslant t_{\text{obs}}) = \sum_{j=0}^{t_{\text{obs}}} \frac{e^{-n}n^j}{j!}.$$

(iii) Suppose that, before collecting any data, we want to determine a suitable sample size. Suppose we want $\alpha = 0.01$ and that we also want to ensure a power of at least 0.95 at $\lambda = \frac{1}{2}$. How large should n be?

As above, our test is of the form: reject $H_0 \iff \sum x_i \leqslant k$. We want k such that

$$\begin{aligned} 0.01 &= P(\text{reject } H_0 \,|\, H_0 \text{ true}) \\ &= P\left(\sum X_i \leqslant k \,\middle|\, H_0\right) \\ &= P\left(\frac{\sum X_i - n}{\sqrt{n}} \leqslant \frac{k-n}{\sqrt{n}} \,\middle|\, H_0\right) \end{aligned}$$

and, by the CLT (i.e. for large n), if H_0 is true then $\frac{\sum X_i - n}{\sqrt{n}} \overset{\text{D}}{\approx} N(0,1)$, so

$$\begin{aligned} 0.01 &\approx P\left(N(0,1) \leqslant \frac{k-n}{\sqrt{n}}\right) \\ &= \Phi\left(\frac{k-n}{\sqrt{n}}\right). \end{aligned}$$

Now $\Phi(-2.326) = 0.01$, hence $\frac{k-n}{\sqrt{n}} \approx -2.326$, so $k \approx n - 2.326\sqrt{n}$.

The power requirement is $w(\frac{1}{2}) \geqslant 0.95$, so

$$\begin{aligned} 0.95 &\leqslant w(\tfrac{1}{2}) \\ &= P(\text{reject } H_0 \,|\, \lambda = \tfrac{1}{2}) \\ &= P\left(\sum X_i \leqslant n - 2.326\sqrt{n} \,\middle|\, \lambda = \tfrac{1}{2}\right) \\ &= P\left(\frac{\sum X_i - n/2}{\sqrt{n/2}} \leqslant \frac{n/2 - 2.326\sqrt{n}}{\sqrt{n/2}} \,\middle|\, \lambda = \tfrac{1}{2}\right) \end{aligned}$$

and, by the CLT (i.e. for large n), $\frac{\sum X_i - n/2}{\sqrt{n/2}} \overset{\text{D}}{\approx} N(0,1)$ if $\lambda = \frac{1}{2}$, so

$$0.95 \leqslant \Phi\left(\frac{n/2 - 2.326\sqrt{n}}{\sqrt{n/2}}\right).$$

Now $\Phi(1.645) = 0.95$, so we require

$$\frac{n/2 - 2.326\sqrt{n}}{\sqrt{n/2}} \geqslant 1.645$$

which gives

$$\sqrt{n} \geqslant 2\left(2.326 + \frac{1.645}{\sqrt{2}}\right) = 2\left(\Phi^{-1}(0.99) + \frac{\Phi^{-1}(0.95)}{\sqrt{2}}\right)$$

i.e. $\sqrt{n} \geqslant 6.98$, so $n \geqslant 48.7$. So the recommended sample size would be $n = 49$.

3.7 Likelihood ratio tests

We now consider testing

- the null hypothesis $H_0 : \theta \in \Theta_0$
- against the *general alternative* $H_1 : \theta \in \Theta$.

So now H_0 is a special case of H_1: we say that the statistical model under H_0 is "nested within" H_1 (i.e. $\Theta_0 \subset \Theta$). We want to see (i.e. test) if simplifying to the H_0-model is reasonable.

The *likelihood ratio* $\lambda(\mathbf{x})$ is defined by

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} L(\theta;\mathbf{x})}{\sup_{\theta\in\Theta} L(\theta;\mathbf{x})}. \tag{3.4}$$

A *(generalised) likelihood ratio test* (LRT) of H_0 against H_1 has critical region of the form $C = \{\mathbf{x} : \lambda(\mathbf{x}) \leqslant k\}$, where k is a constant.

For a test of size α we must choose k so that

$$\sup_{\theta\in\Theta_0} P(\lambda(\mathbf{X}) \leqslant k \,|\, \theta) = \alpha.$$

So in principle we must look at the distribution of $\lambda(\mathbf{X})$, or use an approximation, to determine k. In practice this means either (i) we simplify the inequality $\lambda(\mathbf{X}) \leqslant k$ to an inequality in terms of a function of $\mathbf{X}$ whose distribution we know exactly (see the next example), or (ii) we use the approximate χ^2-distribution of $-2\log\lambda(\mathbf{X})$ (see further below).

Example 3.9. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Consider testing

- $H_0 : \mu = \mu_0$, with any $\sigma^2 > 0$
- against $H_1 : \mu \in (-\infty, \infty)$, with any $\sigma^2 > 0$.

Here the likelihood is

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right].$$

For the numerator of (3.4) we maximise L over σ^2 with $\mu = \mu_0$ fixed. The maximum is at $\sigma^2 = \widehat{\sigma}_0^2 = \frac{1}{n}\sum(x_i - \mu_0)^2$.

For the denominator of (3.4) we maximise L over μ and σ^2. The maximum is at $\mu = \widehat{\mu} = \overline{x}$, $\sigma^2 = \widehat{\sigma}^2 = \frac{1}{n}\sum(x_i - \overline{x})^2$.

Subsituting these values into the likelihood we obtain

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{L(\mu_0, \widehat{\sigma}_0^2)}{L(\widehat{\mu}, \widehat{\sigma}^2)} \\
&= \frac{\left[\frac{2\pi}{n}\sum(x_i - \mu_0)^2\right]^{-n/2} e^{-n/2}}{\left[\frac{2\pi}{n}\sum(x_i - \overline{x})^2\right]^{-n/2} e^{-n/2}} \\
&= \left[\frac{\sum(x_i - \mu_0)^2}{\sum(x_i - \overline{x})^2} \right]^{-n/2}.
\end{aligned}$$

Now note that $\sum(x_i - \mu_0)^2 = \sum(x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2$. (To see this write $\sum(x_i - \mu_0)^2 = \sum\{(x_i - \overline{x}) + (\overline{x} - \mu_0)\}^2$, expand the RHS and then simplify.) Then substituting into the expression for $\lambda(\mathbf{x})$ gives

$$\lambda(\mathbf{x}) = \left[1 + \frac{n(\overline{x} - \mu_0)^2}{\sum(x_i - \overline{x})^2} \right]^{-n/2}.$$

So the test is:

$$\begin{aligned} \text{reject } H_0 &\iff \lambda(\mathbf{x}) \leqslant k \\ &\iff \left| \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \right| \geqslant k_1. \end{aligned}$$

This is the familiar t-test, so we know that we should take $k_1 = t_{n-1}(\alpha/2)$ for a test of size α (i.e. we know the exact distribution of a function of $\lambda(\mathbf{x})$).

Likelihood ratio statistic

The statistic $\Lambda(\mathbf{X})$ defined by

$$\Lambda(\mathbf{X}) = -2 \log \lambda(\mathbf{X})$$

is called the *likelihood ratio statistic*. In terms of Λ, the critical region of the LRT becomes $\{\mathbf{x} : \Lambda(\mathbf{x}) \geqslant c\}$, for some constant c.

If H_0 is true then, under the regularity conditions discussed at the end of Section 1.7 (and which we assume from now on), as $n \to \infty$ we have

$$\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2_p \tag{3.5}$$

where $p = \dim \Theta - \dim \Theta_0$. Here $\dim \Theta$ is the dimension of the whole parameter space, which we can think of as the number of independent parameters in Θ. Similarly, $\dim \Theta_0$ is the dimension of Θ_0.

For a test with approximate size α, we reject H_0 if and only if $\Lambda(\mathbf{x}) \geqslant c$ where c is such that $P(\chi^2_p \geqslant c) = \alpha$. [Using notation similar to what we've used before, we might write $c = \chi^2_p(\alpha)$.] The size is approximately α because the distribution is approximately χ^2_p, assuming that we have a large sample.

Why is (3.5) true? (Sketch)

Consider the simplest case: $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta$, where $\dim \Theta = 1$, e.g. $\Theta = (0, \infty)$ or $\Theta = (-\infty, \infty)$. Note: in this case $\dim \Theta_0 = 0$.

So $p = \dim \Theta - \dim \Theta_0 = 1$, so we are looking for $\Lambda(\mathbf{X})$ to converge to a χ^2_1-distribution. We have

$$\begin{aligned} \Lambda(\mathbf{X}) &= -2 \log \left(\frac{L(\theta_0)}{L(\widehat{\theta})} \right) \\ &= 2[\ell(\widehat{\theta}) - \ell(\theta_0)]. \end{aligned}$$

Assuming that $\widehat{\theta}$ satisfies $\ell'(\widehat{\theta}) = 0$,

$$\begin{aligned} \ell(\theta_0) &\approx \ell(\widehat{\theta}) + (\widehat{\theta} - \theta_0)\ell'(\widehat{\theta}) + \frac{1}{2}(\widehat{\theta} - \theta_0)^2 \ell''(\widehat{\theta}) \\ &= \ell(\widehat{\theta}) - \frac{1}{2}(\widehat{\theta} - \theta_0)^2 J(\widehat{\theta}). \end{aligned}$$

Hence

$$\begin{aligned} \Lambda(\mathbf{X}) &= 2[\ell(\widehat{\theta}) - \ell(\theta_0)] \\ &\approx (\widehat{\theta} - \theta_0)^2 I(\theta_0) \frac{J(\widehat{\theta})}{I(\theta_0)}. \end{aligned}$$

If H_0 is true then $(\widehat{\theta} - \theta_0)\sqrt{I(\theta_0)} \overset{\text{D}}{\approx} N(0,1)$ and $J(\widehat{\theta})/I(\theta_0) \approx 1$, so

$$\Lambda(\mathbf{X}) \overset{D}{\to} [N(0,1)]^2 \times 1 \sim \chi^2_1.$$

It's convenient to use Θ for something to do with H_0 below. So let's rewrite our definition of Λ so that it avoids using Θ at all. From now on we write Λ as

$$\Lambda = -2\log\lambda = -2\log\left(\frac{\sup_{H_0} L}{\sup_{H_1} L}\right) \tag{3.6}$$

and, assuming n is large, when H_0 is true we have $\Lambda \overset{\text{D}}{\approx} \chi^2_p$ where $p = \dim H_1 - \dim H_0$.

Goodness of fit tests

SLIDES. First Hardy–Weinberg equilibrium slide goes here.

Suppose we have n independent observations, where each observation falls into one of the categories $1, \ldots, k$. Let n_i be the number of observations in category i, so $\sum_i n_i = n$ (this sum, and other sums over i below, are over $i = 1, \ldots, k$).

Let π_i be the probability that a single observation falls into category i, where $\sum_i \pi_i = 1$. Let $\pi = (\pi_1, \ldots, \pi_k)$.

The likelihood is given by

$$L(\pi) = \frac{n!}{n_1! \cdots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k}.$$

This is a *multinomial distribution*. The log-likelihood is

$$\ell(\pi) = \sum n_i \log \pi_i + \text{constant}.$$

With no restrictions on π other than $\sum_i \pi_i = 1$, the dimension of this general model is $k - 1$: i.e. there are $k - 1$ independent parameters that we can vary, and once say $\pi_1, \ldots, \pi_{k-1}$ are known, then π_k is determined by $\pi_k = 1 - \sum_{i=1}^{k-1} \pi_i$.

Suppose we want to test the fit of the more restrictive model where category i has probability $\pi_i = \pi_i(\theta)$, where $\theta \in \Theta$. That is, we wish to test

- the null hypothesis $H_0 : \pi_i = \pi_i(\theta)$, where $\theta \in \Theta$
- against the general alternative H_1 : the π_i are unrestricted except for $\sum_i \pi_i = 1$.

Suppose that $\dim \Theta = q$ (where $q < k-1$). Then the parameter space for H_1 has dimension $\dim H_1 = k-1$, the restricted parameter space for H_0 has dimension $\dim H_0 = \Theta = q$, and so when H_0 is true the approximate distribution of the likelihood ratio statistic Λ is χ^2_p where $p = (k-1) - q$. [In the Hardy–Weinberg equilibrium example, we have $q = 1$.]

(i) For the numerator in (3.6) we can maximise the log-likelihood $\sum n_i \log \pi_i(\theta)$ over $\theta \in \Theta$ to obtain the MLE $\widehat{\theta}$.

(ii) Let $g(\pi) = \sum \pi_i - 1$. For the denominator in (3.6) we need to maximise the log-likelihood $f(\pi) = \sum n_i \log \pi_i$ subject to the constraint $g(\pi) = 0$.

Using a Lagrange multiplier λ, we want

$$\frac{\partial f}{\partial \pi_i} - \lambda \frac{\partial g}{\partial \pi_i} = 0 \quad \text{for } i = 1, \ldots, k.$$

That is $\frac{n_i}{\pi_i} - \lambda = 0$, so $\pi_i = n_i/\lambda$, for $i = 1, \ldots, k$.

Now $1 = \sum \pi_i = \sum n_i/\lambda = n/\lambda$, hence $\lambda = n$, and so

$$\widehat{\pi}_i = \frac{n_i}{n} \quad \text{for } i = 1, \ldots, k.$$

So

$$\begin{aligned}\Lambda &= -2\log\left(\frac{\sup_{H_0} L}{\sup_{H_1} L}\right)\\ &= -2\log\left(\frac{L(\pi(\widehat{\theta}))}{L(\widehat{\pi})}\right)\\ &= 2[\ell(\widehat{\pi}) - \ell(\pi(\widehat{\theta}))]\\ &= 2[\sum n_i \log \widehat{\pi}_i - \sum n_i \log \pi_i(\widehat{\theta})]\\ &= 2\sum n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right)\end{aligned}$$

We compare Λ to a χ^2_p, where $p = k - 1 - q$, since this is the approximate distribution of Λ under H_0.

SLIDES. Hardy–Weinberg equilibrium slides go here.

SLIDES. Slides on insect counts example go here.

Pearson's chi-squared statistic

Write

$$\Lambda = 2\sum O_i \log\left(\frac{O_i}{E_i}\right)$$

where $O_i = n_i$ is the *observed* count in category i, and $E_i = n\pi_i(\widehat{\theta})$ is the *expected* count in category i under H_0.

For x near a, we have

$$x\log\left(\frac{x}{a}\right) \approx (x-a) + \frac{(x-a)^2}{2a}.$$

Hence

$$\begin{aligned}\Lambda &\approx 2\sum\left[(O_i - E_i) + \frac{(O_i - E_i)^2}{2E_i}\right]\\ &= \sum \frac{(O_i - E_i)^2}{E_i}\end{aligned}$$

since $\sum O_i = \sum E_i = n$. The statistic

$$P = \sum \frac{(O_i - E_i)^2}{E_i}$$

is called *Pearson's chi-squared statistic* and this also has an approximate χ^2_p-distribution under H_0, where $p = k - 1 - q$.

Two-way contingency tables

SLIDES. First hair and eye colour slide goes here.

Suppose each of n independent individuals is classified according to two sets of categories. Suppose the first set of categories corresponds to the rows of a table (e.g. hair colour), and the second set of categories corresponds to the columns of a table (e.g. eye colour). Suppose

there are r rows and c columns. Then there are rc cells in the table and each individual falls into precisely one cell (e.g. corresponding to their hair and eye colour).

	column (eye colour)				
row (hair colour)	1	2	$\cdots$	c	row sum
1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
column sum	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Let $n_{i+} = \sum_{j=1}^{c} n_{ij}$ be the number of individuals in the ith row of the table. Let $n_{+j} = \sum_{i=1}^{r} n_{ij}$ be the number of individuals in the jth column of the table.

Let π_{ij} denote the probability that an individual falls into cell (i, j) of the table (i.e. that the individual's hair colour is i and eye colour is j). The likelihood is

$$\begin{aligned} L(\pi) &= \frac{n!}{n_{11}!\, n_{12}! \cdots n_{rc}!} \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \cdots \pi_{rc}^{n_{rc}} \\ &= n! \prod_i \prod_j \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!} \end{aligned}$$

where the products are over all i and all j. The log-likelihood is

$$\ell(\pi) = \sum_i \sum_j n_{ij} \log \pi_{ij} + \text{constant}$$

where the sums are over all i and all j.

Suppose we would like to test the null hypothesis H_0 that the row into which an individual falls is independent of the column into which that same individual falls (i.e. test if hair and eye colour are independent). That is, we would like to test

- the null hypothesis $H_0 : \pi_{ij} = \alpha_i \beta_j$ for all i, j, where $\sum \alpha_i = 1$ and $\sum \beta_j = 1$
- against the general alternative H_1 : the π_{ij} are unrestricted except for $\sum_i \sum_j \pi_{ij} = 1$.

We can find Λ as follows.

(i) To find $\sup_{H_0}$: we need to maximise $\sum_i \sum_j n_{ij} \log(\alpha_i \beta_j)$ subject to the two constraints $\sum \alpha_i = 1$ and $\sum \beta_j = 1$. We can do this using two Lagrange multiplies [and this is part of a question on Sheet 3]. We find

$$\widehat{\alpha}_i = \frac{n_{i+}}{n}, \quad \widehat{\beta}_j = \frac{n_{+j}}{n}.$$

(ii) We have found $\sup_{H_1}$ already, in the goodness of fit section, with only slightly different notation. [Exercise: check this.] We have $\widehat{\pi}_{ij} = n_{ij}/n$.

So

$$\begin{aligned}
\Lambda &= -2\log\left(\frac{\sup_{H_0} L}{\sup_{H_1} L}\right) \\
&= 2\Big(\sup_{H_1} \ell - \sup_{H_0} \ell\Big) \\
&= 2[\sum_i \sum_j n_{ij} \log \widehat{\pi}_{ij} - \sum_i \sum_j n_{ij} \log(\widehat{\alpha}_i \widehat{\beta}_j)] \\
&= 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij} n}{n_{i.} n_{.j}}\right) \\
&= 2\sum_i \sum_j O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right)
\end{aligned}$$

where $O_{ij} = n_{ij}$ is the *observed* number of individuals in cell (i, j), and $E_{ij} = n\widehat{\alpha}_i\widehat{\beta}_j$ is the *expected* number of individuals in cell (i, j) under H_0. As before, $\Lambda \approx P$ where P is Pearson's chi-squared statistic

$$P = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

To calculate the degrees of freedom for Λ, and P:

(i) $\dim H_1 = rc - 1$ (exactly as for the goodness of fit tests)

(ii) under H_0 we have $r - 1$ degrees of freedom due to the paramaters α_i (i.e. there are r parameters α_i and as before we lose one degree of freedom since $\sum \alpha_i = 1$), and a further $c-1$ degrees of freedom due to the paramaters β_j, so $\dim H_0 = (r-1)+(c-1)$.

Hence, under H_0, both Λ and P have an approximate χ^2_p-distribution, where $p = \dim H_1 - \dim H_0 = (r-1)(c-1)$.

SLIDES. Hair and eye colour slides go here.

4 Bayesian Inference

So far we have followed the frequentist (or classical) approach to statistics. That is, we have treated unknown parameters as a fixed constants, and we have imagined repeated sampling from our model in order to evaluate properties of estimators, interpret confidence intervals, calculate p-values, etc.

We now take a different approach: in Bayesian inference, *unknown parameters* are treated as *random variables*.

SLIDES. Introductory slides go here.

4.1 Introduction

Suppose that, as usual, we have a probability model $f(\mathbf{x} \mid \theta)$ for data $\mathbf{x}$. Previously we wrote $f(\mathbf{x}; \theta)$. Now we write $f(\mathbf{x} \mid \theta)$ to emphasise that we have a model for data $\mathbf{x}$ *given*, i.e. *conditional on*, the value of θ.

Suppose also that, before observing $\mathbf{x}$, we summarise our beliefs about θ in a *prior density* $\pi(\theta)$. [Or in a *prior mass function* $\pi(\theta)$ if θ is discrete.] This means that we are now treating θ as a RV.

Once we have observed data $\mathbf{x}$, our updated beliefs about θ are contained in the conditional density of θ given $\mathbf{x}$, which is called the *posterior density* (of θ given $\mathbf{x}$), written $\pi(\theta \mid \mathbf{x})$.

Theorem 4.1 (Bayes' Theorem)**.**

(i) If events $B_1, B_2, \ldots$ partition the sample space, then for any event A we have

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A)}.$$

(ii) For continuous RVs Y and Z, the conditional density $f_{Z \mid Y}(z \mid y)$ satisfies

$$f_{Z \mid Y}(z \mid y) = \frac{f_{Y \mid Z}(y \mid z) f_Z(z)}{f_Y(y)}.$$

Proof.

(i) Proved last year.

(ii) To make the notation simpler, we omit subscripts on pdfs. By definition of conditional density,

$$f(z \mid y) = \frac{f(y, z)}{f(y)} \tag{4.1}$$

and also

$$f(y \mid z) = \frac{f(y, z)}{f(z)}. \tag{4.2}$$

From (4.2) we have $f(y, z) = f(y \mid z) f(z)$, and substituting this expression for $f(y, z)$ into (4.1) gives the result. □

To find the marginal density of Y, we integrate the joint pdf $f(y, z)$ over all z, i.e.

$$\begin{aligned} f(y) &= \int_{-\infty}^{\infty} f(y, z)\, dz \\ &= \int_{-\infty}^{\infty} f(y \,|\, z) f(z)\, dz. \end{aligned} \tag{4.3}$$

So, in the case of continuous RVs, by Bayes' Theorem (with $\mathbf{x}$ and θ in place of Y and Z) the posterior density is

$$\pi(\theta \,|\, \mathbf{x}) = \frac{f(\mathbf{x} \,|\, \theta)\pi(\theta)}{f(\mathbf{x})} \tag{4.4}$$

where, as in (4.3), the denominator $f(\mathbf{x})$ can be written

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

and where this integral is an integral over all θ.

We treat $\pi(\theta \,|\, \mathbf{x})$ as a function of θ, with data $\mathbf{x}$ fixed. Since $\mathbf{x}$ is fixed, the denominator $f(\mathbf{x})$ in (4.4) is just a constant. Hence

$$\begin{gathered} \pi(\theta \,|\, \mathbf{x}) \propto f(\mathbf{x} \,|\, \theta) \times \pi(\theta) \\ \text{posterior} \propto \text{likelihood} \times \text{prior}. \end{gathered} \tag{4.5}$$

Example 4.2. Conditionally on θ, suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta)$. That is, $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$, i.e. $f(x \,|\, \theta) = \theta^x (1 - \theta)^{1-x}$ for $x = 0, 1$. So

$$\begin{aligned} f(\mathbf{x} \,|\, \theta) &= \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} \\ &= \theta^r (1 - \theta)^{n - r} \end{aligned}$$

where $r = \sum x_i$.

A natural prior here is a Beta(a, b) pdf:

$$\pi(\theta) = \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} \quad \text{for } 0 < \theta < 1$$

where $B(a, b)$ is the beta function. Since the pdf $\pi(\theta)$ integrates to 1, the normalising constant $B(a, b)$ is given by

$$B(a, b) = \int_0^1 \theta^{a-1} (1 - \theta)^{b-1}\, d\theta.$$

We will use (without proof) the following expression for $B(a, b)$ in terms of the gamma function:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

where $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\, du$. Remember: $\Gamma(a+1) = a\Gamma(a)$ for $a > 0$, and $\Gamma(n) = (n-1)!$ when n is a positive integer. The values a and b satisfy $a > 0$ and $b > 0$ and are assumed known – their values reflect our beliefs about θ before observing any data.

Using (4.5),

$$\begin{aligned}\pi(\theta \,|\, \mathbf{x}) &\propto \theta^r(1-\theta)^{n-r} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{r+a-1}(1-\theta)^{n-r+b-1}. \end{aligned} \tag{4.6}$$

In reaching (4.6), in addition to dropping $f(\mathbf{x})$ which is not a function of θ, we have also dropped the constant $1/B(a,b)$. The RHS of (4.6) depends on θ exactly as for a Beta$(r+a, n-r+b)$ density. That is, the constant normalising (4.6) is $B(r+a, n-r+b)$ and we have

$$\pi(\theta \,|\, \mathbf{x}) = \frac{1}{B(r+a, n-r+b)} \theta^{r+a-1}(1-\theta)^{n-r+b-1} \quad \text{for } 0 < \theta < 1.$$

So acquring data $\mathbf{x}$ has the effect of updating (a,b) to $(r+a, n-r+b)$.

It is important to note that in the above Bernoulli-Beta example, as in most/all(?) other examples we will meet, there is *no need* to do any integration to find $\pi(\theta \,|\, \mathbf{x})$. We find $\pi(\theta \,|\, \mathbf{x})$ by comparing (4.6) with a Beta$(r+a, n-r+b)$ pdf.

Example 4.3. Conditional on θ, suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(\theta)$. Suppose the prior for θ is a Gamma(α, β) pdf:

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for } \theta > 0$$

where $\alpha > 0$ and $\beta > 0$ are assumed known.

Using posterior $\propto$ likelihood $\times$ prior, we have

$$\begin{aligned}\pi(\theta \,|\, \mathbf{x}) &\propto \left(\prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} \right) \times \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{r+\alpha-1} e^{-(n+\beta)\theta} \quad \text{for } \theta > 0 \end{aligned} \tag{4.7}$$

where $r = \sum x_i$. The $\beta^\alpha/\Gamma(\alpha)$ term, and the the $x_i!$ terms, have all been omitted: we are interested in $\pi(\theta \,|\, \mathbf{x})$ as a function of θ, and these omitted terms are constant with respect to θ, so omitting them simply adjusts the constant of proportionality.

The dependence on θ in (4.7) is as for a Gamma pdf, so $\pi(\theta \,|\, \mathbf{x})$ is the pdf of a Gamma$(r+\alpha, n+\beta)$.

Again: *no need* to do any integration to get the normalising constant in (4.7).

4.2 Inference

SLIDES. Slides on MRSA example go here.

All information about the parameter θ is contained in the posterior density, i.e. contained in $\pi(\theta \,|\, \mathbf{x})$.

Posterior summaries

Sometimes summaries of $\pi(\theta \,|\, \mathbf{x})$ are useful, e.g.:

- the posterior mode (the value of θ at which $\pi(\theta \,|\, \mathbf{x})$ is maximised)
- the posterior mean $E(\theta \,|\, \mathbf{x})$ (this expectation is over θ, and $\mathbf{x}$ is fixed)

- the posterior median (the value m such that $\int_{-\infty}^{m} \pi(\theta \,|\, \mathbf{x})\, d\theta = \frac{1}{2}$)
- the posterior variance $\text{var}(\theta \,|\, \mathbf{x})$
- other quantiles of $\pi(\theta \,|\, \mathbf{x})$ (i.e. in addition to the median).

Example 4.4. Conditional on θ, suppose $X \sim \text{Binomial}(n, \theta)$. We can write this as

$$X \,|\, \theta \sim \text{Binomial}(n, \theta).$$

We read "$X \,|\, \theta$" as "X given θ". Suppose the prior for θ is $U(0, 1)$.

The likelihood

$$f(x \,|\, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

is proportional to the Bernoulli likelihood in Section 4.1, and the prior is a $\text{Beta}(1, 1)$. Hence, by the first example in Section 4.1, $\pi(\theta \,|\, x)$ is a $\text{Beta}(x + 1, n - x + 1)$ pdf, i.e.

$$\theta \,|\, x \sim \text{Beta}(x + 1, n - x + 1).$$

We calculate the posterior mean:

$$\begin{aligned}
E(\theta \,|\, x) &= \int_0^1 \theta\, \pi(\theta \,|\, x)\, d\theta \\
&= \frac{1}{B(x+1, n-x+1)} \int_0^1 \theta^{x+1} (1-\theta)^{n-x}\, d\theta \\
&= \frac{1}{B(x+1, n-x+1)} B(x+2, n-x+1) \\
&= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)} \\
&= \frac{\Gamma(x+2)}{\Gamma(x+1)} \frac{\Gamma(n+2)}{\Gamma(n+3)} \\
&= (x+1) \frac{1}{n+2}.
\end{aligned}$$

So the posterior mean is $E(\theta \,|\, x) = \frac{x+1}{n+2}$. So even when all trials are successes (i.e. when $x = n$) this point estimate of θ is $\frac{n+1}{n+2}$, so is less than 1 (which seems sensible, especially if n is small).

The posterior mode is x/n, the same as the MLE. For large n, i.e. when the likelihood contribution dominates that from the prior, the posterior mean will be close to the MLE/posterior mode.

Interval estimation

The Bayesian analogue of a confidence interval is a *credible interval* (or *posterior interval*).

Definition. A $100(1-\alpha)\%$ *credible set* for θ is a subset C of Θ such that

$$\int_C \pi(\theta \,|\, \mathbf{x})\, d\theta = 1 - \alpha. \tag{4.8}$$

Note: $\int_C \pi(\theta \mid \mathbf{x})\, d\theta = P(\theta \in C \mid \mathbf{x})$, so (4.8) says $P(\theta \in C \mid \mathbf{x}) = 1 - \alpha$.

A credible *interval* is when the set C is an interval, say $C = (\theta_1, \theta_2)$. If $P(\theta \leqslant \theta_1 \mid \mathbf{x}) = P(\theta \geqslant \theta_2 \mid \mathbf{x}) = \alpha/2$, then the interval (θ_1, θ_2) is called *equal tailed*.

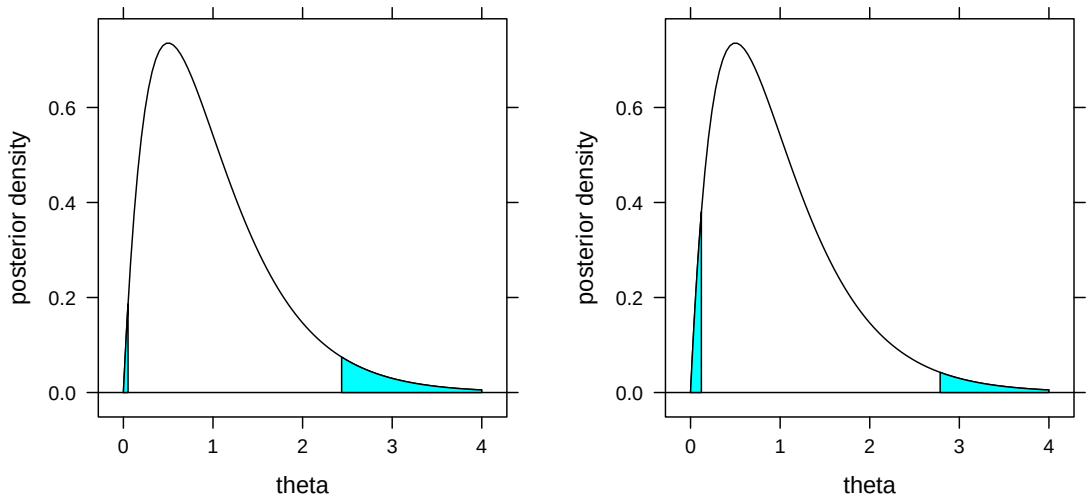

Figure 4.1. Two 95% credible intervals (the pdfs are Gamma$(2, 2)$). Left: the credible interval is (0.05, 2.44), the tail areas (shaded in blue) are 0.005 in lower tail, 0.045 in upper tail. Right: the interval (0.12, 2.79) is an equal tailed credible interval, both lower and upper tail areas are 0.025.

In words, (4.8) says:

> the probability that θ lies in C, given the observed data $\mathbf{x}$, is $1 - \alpha$.

This straightforward probability statement is what we want to be able to say about an interval estimate – this is a strength of the Bayesian approach.

The above statement is *not* true of a confidence interval. The interpretation of a frequentist confidence interval C' is different, and more tricky – we say something like:

> if we could recalculate C' for a large number of datasets collected in the same way as $\mathbf{x}$, then about $100(1 - \alpha)\%$ of these sets would contain the true value of θ.

Definition. We call C a *highest posterior density (HPD)* credible set if $\pi(\theta \mid \mathbf{x}) \geqslant \pi(\theta' \mid \mathbf{x})$ for all $\theta \in C$ and $\theta' \notin C$.

In words: for an HPD interval, the posterior density at any point $\theta \in C$ is at least as high as the posterior density at any point $\theta' \notin C$.

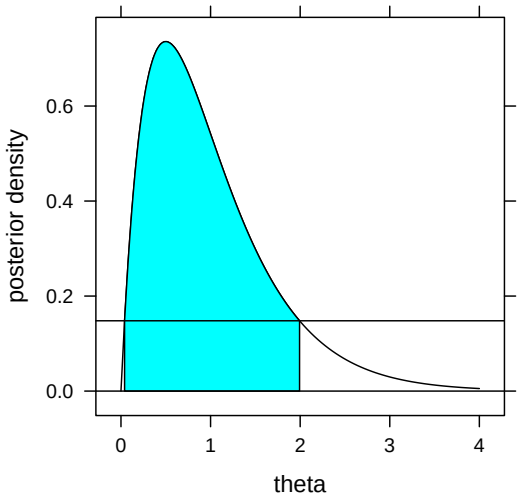

Figure 4.2. A 90% HPD interval (the pdf is Gamma(2, 2)). The HPD interval is (0.04, 2), the shaded area is 0.9. The density at any $\theta \in (0.04, 2)$ is higher than at any $\theta' \notin (0.04, 2)$.

An HPD interval has minimal width among all $100(1-\alpha)\%$ credible intervals. On the other hand, advantages of equal tailed intervals are that they have a direct interpretation in terms of $\alpha/2$ and $1-\alpha/2$ quantiles, they are usually easier to calculate, and if we change the parametrisation of a distribution from θ to $\phi = \phi(\theta)$ then the interval transforms from (θ_1, θ_2) to $(\phi(\theta_1), \phi(\theta_2))$ (this does not hold for HPD intervals in general).

Multi-parameter models

The parameter θ may be a vector. If so, everything above remains true provided that we understand each integral over θ to be a multiple integral over all components of θ. [Usually we have continuous parameters, but if any parameter is discrete then, for that parameter, integrals are replaced by summations.]

E.g. $\theta = (\psi, \lambda)$ say, in which case the prior is a *bivariate* density $\pi(\psi, \lambda)$, as is the posterior $\pi(\psi, \lambda \,|\, \mathbf{x})$. All information about ψ is contained in the marginal posterior density

$$\pi(\psi \,|\, \mathbf{x}) = \int \pi(\psi, \lambda \,|\, \mathbf{x})\, d\lambda.$$

That is, as usual, to find a marginal distribution we integrate over the other components of the density (i.e. integrate over λ here).

Prediction

Let X_{n+1} represent a future observation and let $\mathbf{x} = (x_1, \ldots, x_n)$ denote the observed data. Assume, conditional on θ, that X_{n+1} has density $f(x_{n+1} \,|\, \theta)$ independent of $X_1, \ldots, X_n$.

The density of X_{n+1} given $\mathbf{x}$, called the *posterior predictive density*, is a conditional density. We write it as $f(x_{n+1} \,|\, \mathbf{x})$. Here $\mathbf{x} = (x_1, \ldots, x_n)$ as usual. We have

$$\begin{aligned} f(x_{n+1} \,|\, \mathbf{x}) &= \int f(x_{n+1}, \theta \,|\, \mathbf{x})\, d\theta \\ &= \int f(x_{n+1} \,|\, \theta, \mathbf{x}) \pi(\theta \,|\, \mathbf{x})\, d\theta. \end{aligned}$$

For the first equality above: the density for z is found by integrating that for (z, θ) over all θ. For the second: $P(A \cap B \,|\, C) = P(A \,|\, B \cap C)P(B \,|\, C)$, or in terms of conditional densites $f(u, v|w) = f(u|v, w)f(v|w)$.

Now $f(x_{n+1} \,|\, \theta, \mathbf{x}) = f(x_{n+1} \,|\, \theta)$ by the independence. Hence

$$f(x_{n+1} \,|\, \mathbf{x}) = \int f(x_{n+1} \,|\, \theta)\pi(\theta \,|\, \mathbf{x})\, d\theta.$$

So, given $\mathbf{x}$, the predictive density is found by combining the density for x_{n+1} under the model (i.e. $f(x_{n+1} \,|\, \theta)$) with the posterior density.

If X_{n+1} is discrete, then of course $f(x_{n+1} \,|\, \mathbf{x})$ is a pmf (not a pdf).

For an example: see Sheet 4.

4.3 Prior information

How do we choose a prior $\pi(\theta)$?

(i) We use a prior to represent our beliefs about θ before collecting data: e.g. MRSA example, we might ask a scientific expert who might anticipate θ around 10, say with $\theta \in (5, 17)$ with probability 0.95.

If we approach this by asking several different experts for their beliefs, each expert's opinion might lead to a different prior, so we would want to repeat our analysis with each different prior.

(ii) We might have little prior knowledge, so we might want a prior that expresses "prior ignorance". E.g. if a probability is unknown, we might consider the prior $\theta \sim U(0, 1)$.

(But even uniform priors, apparently expressing "ignorance", can lead to problems when there are a large number of parameters.)

(iii) In the Bernoulli/Beta and Poisson/Gamma examples (Section 4.1), the posterior was of the same form as the prior, i.e. Beta-Beta and Gamma-Gamma. This occurred because the likelihood and prior had the same functional form – in such situations the prior and likelihood are said to be *conjugate*. Conjugate priors are convenient for doing calculations by hand.

There are also other possibilities, and note that (iii) can overlap with (i) and (ii). In some complex situations it might be hard to write down a representative prior distrbution.

Example 4.5. Conditional on θ, suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ where σ^2 is known. Suppose the prior is $\theta \sim N(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are known.

Then

$$\begin{aligned}\pi(\theta \,|\, \mathbf{x}) &\propto f(\mathbf{x} \,|\, \theta)\pi(\theta) \\ &\propto \exp\left[-\frac{1}{2}\sum \frac{(x_i - \theta)^2}{\sigma^2}\right] \exp\left[-\frac{1}{2}\frac{(\theta - \mu_0)^2}{\sigma_0^2}\right]\end{aligned}$$

where as usual we have ignored constants that don't depend on θ.

Now complete the square:

$$\frac{(\theta-\mu_0)^2}{\sigma_0^2}+\sum\frac{(x_i-\theta)^2}{\sigma^2}=\theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)-2\theta\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{x}}{\sigma^2}\right)+\text{constant}$$
$$=\frac{1}{\sigma_1^2}(\theta-\mu_1)^2+\text{constant}$$

where

$$\mu_1=\frac{\frac{1}{\sigma_0^2}\mu_0+\frac{n}{\sigma^2}\overline{x}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}} \tag{4.9}$$

$$\frac{1}{\sigma_1^2}=\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}. \tag{4.10}$$

Hence

$$\pi(\theta\,|\,\mathbf{x})\propto\exp\left(-\frac{1}{2\sigma_1^2}(\theta-\mu_1)^2\right)$$

and so $\pi(\theta\,|\,\mathbf{x})$ is a $N(\mu_1,\sigma_1^2)$ pdf. That is, $\theta\,|\,\mathbf{x}\sim N(\mu_1,\sigma_1^2)$.

(4.9) says that the posterior mean μ_1 is a weighted average of the prior mean μ_0 and the sample mean $\overline{x}$ (with weights $\frac{1}{\sigma_0^2}$ and $\frac{n}{\sigma^2}$).

(4.10) says "posterior precision = prior precision + data precision" where the *precision* of a RV is defined as being 1/variance.

This is another example of conjugacy: prior, likelihood and posterior are all normal.

Improper priors

If $\sigma_0^2\to\infty$ in the previous example, then $\pi(\theta\,|\,\mathbf{x})$ is approx $N(\overline{x},\sigma^2/n)$, i.e. the likelihood contribution dominates the prior contribution as $\sigma_0^2\to\infty$. This corresponds to a prior $\pi(\theta)\propto c$, a constant, i.e. a "uniform prior". But this $\pi(\theta)$ is not a probability distribution since $\theta\in(-\infty,\infty)$ and we can't have $\int_{-\infty}^{\infty}c\,d\theta$ equalling 1.

Definition. A prior $\pi(\theta)$ is called *proper* if $\int\pi(\theta)\,d\theta=1$, and is called *improper* if the integral can't be normalised to equal 1.

An improper prior can lead to a proper posterior which we can use for inference (e.g. uniform prior in in the normal-normal example) . But we can't use an improper posterior for meaningful inference.

Prior ignorance

If no reliable information is available, we might want a prior which has minimal effect on our inference. E.g. if $\Theta=\{\theta_1,\ldots,\theta_m\}$ then $\pi(\theta_i)=1/m$ for $i=1,\ldots,m$ does not favour any one value of θ over any other and in this sense is "non-informative" for θ.

Example 4.6. If $\Theta=(0,1)$ we might think that $\pi(\theta)=1$ for $0<\theta<1$, i.e. $\theta\sim U(0,1)$, represents prior ignorance. However, if we are ignorant about θ, then we are also ignorant

about $\phi = \log\big(\theta/(1-\theta)\big)$. Here $\phi \in \mathbb{R}$ is called the *log-odds*. The pdf of ϕ is

$$\begin{aligned} p(\phi) &= \pi\big(\theta(\phi)\big) \times \frac{d\theta}{d\phi} \\ &= 1 \times \frac{d}{d\phi}\left(\frac{e^\phi}{1+e^\phi}\right) \\ &= \frac{e^\phi}{(1+e^\phi)^2} \quad \text{for } -\infty < \phi < \infty. \end{aligned}$$

[Sketch $p(\phi)$.] The pdf $p(\phi)$ has a maximum at $\phi = 0$ and further $P(-3 < \phi < 3) \approx 0.9$. This does not seem to correspond to ignorance about ϕ, rather this prior is saying that the most likely values of ϕ are close to 0. That is, the prior that was apparently expressing "ignorance" about θ actually expresses some *knowledge* about ϕ.

Jeffreys priors

Suppose θ is a scalar parameter.

The problem with the $\phi = \log\big(\theta/(1-\theta)\big)$ example above is that the representation of "ignorance" changes if we change the parametrisation from θ to ϕ. A solution to this issue is the *Jeffreys prior* defined by

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

where as usual $I(\theta)$ is the expected (Fisher) information.

Recall that if $X_1, \ldots, X_n$ are all from $f(x \,|\, \theta)$ then $I(\theta) = ni(\theta)$ where $i(\theta)$ is the expected Fisher information in a sample of size 1,

$$i(\theta) = -E\left(\frac{d^2}{d\theta^2} \log f(X_1 \,|\, \theta)\right)$$

where the expectation is over X_1 with θ held fixed. Then the Jeffreys prior is

$$\pi(\theta) \propto i(\theta)^{1/2}$$

(the $n^{1/2}$ factor difference between $I(\theta)^{1/2}$ and $i(\theta)^{1/2}$ can be absorbed into the constant of proportionality).

Sometimes Jeffreys rule leads to an improper prior.

Example 4.7. Consider a single Bernoulli trial with success probability θ. We have

$$\begin{aligned} f(x \,|\, \theta) &= \theta^x (1-\theta)^{1-x} \quad \text{for } x = 0, 1 \\ \ell(\theta) &= x \log\theta + (1-x)\log(1-\theta) \\ -\frac{d^2\ell}{d\theta^2} &= \frac{x}{\theta^2} + \frac{1-x}{(1-\theta)^2}. \end{aligned}$$

Hence

$$\begin{aligned} i(\theta) &= E\left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2}\right] \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \quad \text{since } E(X) = \theta \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned}$$

So the Jeffreys prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ for $0 < \theta < 1$. This is a Beta$(\frac{1}{2}, \frac{1}{2})$.

Jeffreys priors can be extended to a vector parameter θ by taking

$$\pi(\theta) \propto |I(\theta)|^{1/2} \tag{4.11}$$

where the RHS means the square root of the determinant of the information matrix. However a simpler and more common approach for vector θ is to find the Jeffreys prior for each component of θ separately, and then to take the product of these (i.e. assume prior independence) to get the whole prior. This can lead to a different prior to (4.11)

Example 4.8. Suppose $\theta \in \mathbb{R}$ and $f(x \,|\, \theta) = g(x - \theta)$ for some function g, e.g. g could be the pdf of a $N(0, 1)$, or the pdf of a t_1. Then θ is called a *location parameter* and Jeffreys rule leads to $\pi(\theta) \propto 1$ for $\theta \in \mathbb{R}$.

Example 4.9. Suppose $\sigma > 0$ and $f(x \,|\, \sigma) = \frac{1}{\sigma} g(x/\sigma)$ for some function g, e.g. σ could be the standard deviation of a normal, or (the reciprocal of) the β parameter of a Gamma. Then σ is called a *scale parameter* and Jeffreys rule leads to $\pi(\sigma) \propto 1/\sigma$ for $\sigma > 0$.

Example 4.10. Combine the previous two examples: if $f(x \,|\, \theta, \sigma) = \frac{1}{\sigma} g(\frac{x-\theta}{\sigma})$ then we have a *location-scale* family of distributions. The approach of finding the Jeffreys prior for each component parameter separately and assuming prior independence of θ and σ simply uses the product of the above two priors:

$$\pi(\theta, \sigma) \propto 1 \times \frac{1}{\sigma} \quad \text{for } \theta \in \mathbb{R},\ \sigma > 0.$$

(Using the square root of the determinant of $I(\theta)$ leads to a different prior.)

We finish this section by showing that Jeffreys prior does not depend on the parametrisation used – we do the case of a scalar parameter θ. So consider a one-to-one transformation of the parameter, from θ to ϕ: let $\phi = h(\theta)$, with inverse $\theta = g(\phi)$ (i.e. $g = h^{-1}$). By transformation of variables, if θ has pdf $\pi(\theta)$ then ϕ has pdf

$$p(\phi) = \pi\big(g(\phi)\big)\, |g'(\phi)|. \tag{4.12}$$

We want to show that (i) and (ii) give the same prior for ϕ, where:

(i) we determine $\pi(\theta)$ using Jeffreys rule for θ, then transform it to give a prior $p(\phi)$ for ϕ

(ii) we determine $p(\phi)$ using Jeffreys rule for ϕ directly.

For (i): we have $\pi(\theta) \propto [i(\theta)]^{1/2}$, and transforming this to ϕ using (4.12) gives

$$p(\phi) \propto \big[i\big(g(\phi)\big)\big]^{1/2}\, |g'(\phi)| = [i(\theta)]^{1/2}\, |g'(\phi)|. \tag{4.13}$$

For (ii): we need to find the relationship between $i(\phi)$ and $i(\theta)$. Let $\ell(\theta) = \log f(X_1 \,|\, \theta)$ and recall that in Section 1.7 we saw that

$$i(\theta) = E\left[\left(\frac{d\ell}{d\theta}\right)^2\right]. \tag{4.14}$$

We have

$$\frac{d\ell}{d\phi} = \frac{d\ell}{d\theta}\frac{d\theta}{d\phi}.$$

So squaring both sides of this equation, taking expectations, and using (4.14) gives

$$i(\phi) = i(\theta)[g'(\phi)]^2$$

as $\frac{d\theta}{d\phi} = g'(\phi)$. Hence Jeffreys rule for ϕ gives

$$\pi(\phi) \propto [i(\phi)]^{1/2} = [i(\theta)]^{1/2} \, |g'(\phi)|$$

in agreement with (4.13), as required.

4.4 Hypothesis testing and Bayes factors

Suppose we want to compare two hypotheses H_0 and H_1, exactly one of which is true. The Bayesian approach attaches prior probabilities $P(H_0), P(H_1)$ to H_0, H_1, where $P(H_0) + P(H_1) = 1$. The *prior odds* of H_0 relative to H_1 is

$$\text{prior odds} = \frac{P(H_0)}{P(H_1)} = \frac{P(H_0)}{1 - P(H_0)}.$$

[The odds of any event A is $P(A)/(1 - P(A))$.]

We can compute the posterior probabilities $P(H_i \,|\, \mathbf{x})$ for $i = 0, 1$ and compare them. By Bayes' Theorem,

$$P(H_i \,|\, \mathbf{x}) = \frac{P(\mathbf{x} \,|\, H_i) P(H_i)}{P(\mathbf{x})} \quad \text{for } i = 0, 1 \tag{4.15}$$

where

$$P(\mathbf{x}) = P(\mathbf{x} \,|\, H_0) P(H_0) + P(\mathbf{x} \,|\, H_1) P(H_1).$$

Note: here $P(H_i \,|\, \mathbf{x})$ is the probability of H_i conditioned on the data, whereas p-values in Section 3 *can't* be interpreted in this way.

The *posterior odds* of H_0 relative to H_1 is

$$\text{posterior odds} = \frac{P(H_0 \,|\, \mathbf{x})}{P(H_1 \,|\, \mathbf{x})}.$$

Using (4.15),

$$\frac{P(H_0 \,|\, \mathbf{x})}{P(H_1 \,|\, \mathbf{x})} = \frac{P(\mathbf{x} \,|\, H_0)}{P(\mathbf{x} \,|\, H_1)} \times \frac{P(H_0)}{P(H_1)}$$

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}$$

where the *Bayes factor* B_{01} of H_0 relative to H_1 is given by

$$B_{01} = \frac{P(\mathbf{x} \,|\, H_0)}{P(\mathbf{x} \,|\, H_1)}. \tag{4.16}$$

So the change from the prior odds to the posterior odds depends on the data only through the Bayes factor B_{01}. The Bayes factor tells us how the data shifts the strength of belief in H_0 relative to H_1. If our prior model has $P(H_0) = P(H_1)$ then, given the data, we have that H_0 is B_{01} times more likely than H_1.

General setup

We are assuming we have:

(i) prior probabilities $P(H_i)$, $i = 0, 1$, where $P(H_0) + P(H_1) = 1$

(ii) a prior for θ_i under H_i which we write as $\pi(\theta_i \,|\, H_i)$ for $\theta_i \in \Theta_i$, $i = 0, 1$, where Θ_i denotes the parameter space under H_i

(iii) a model for data $\mathbf{x}$ under H_i which we write as $f(\mathbf{x} \,|\, \theta_i, H_i)$.

The two priors $\pi(\theta_i \,|\, H_i)$, $i = 0, 1$, could be of different forms, as could the two models $f(\mathbf{x} \,|\, \theta_i, H_i)$, $i = 0, 1$. E.g. the prior under H_0 could be an exponential distribution (one parameter), the prior under H_1 could a lognormal distribution (which has two parameters).

Sometimes, as in examples below, (i) and (ii) might be combined: the prior density might be $\pi(\theta)$ for $\theta \in \Theta$ where

- $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$
- the prior probabilities are $P(H_i) = \int_{\Theta_i} \pi(\theta)\, d\theta$
- $\pi(\theta_i \,|\, H_i)$ is the conditional density of θ given H_i, i.e.

$$\pi(\theta_i \,|\, H_i) = \frac{\pi(\theta)}{\int_{\theta \in \Theta_i} \pi(\theta)\, d\theta}.$$

This simplification does not hold for simple hypotheses: for if $H_i : \theta = \theta_i$, i.e. $\Theta_i = \{\theta_i\}$, then $P(H_i)$ would be an integral over the set Θ_i, which is just a single point, and so $P(H_i)$ would be zero (which is not what we want).

We can write the numerator and denominator in (4.16) as follows. Conditioning on θ_i (i.e. using the partition theorem/law of total probability), we have

$$P(\mathbf{x} \,|\, H_i) = \int_{\Theta_i} f(\mathbf{x} \,|\, \theta_i, H_i) \pi(\theta_i \,|\, H_i)\, d\theta_i. \tag{4.17}$$

The quantity $P(\mathbf{x} \,|\, H_i)$ is called the *marginal likelihood* for H_i: it is the likelihood $f(\mathbf{x} \,|\, \theta_i, H_i)$ averaged over Θ_i, weighted according to the prior $\pi(\theta_i \,|\, H_i)$. So we see that the Bayes factor (4.16) is somewhat similar to the likelihood ratio of Section 3, but not the same: similar because it is a ratio of (marginal) likelihoods; but not the same because here we are averaging over θ in (4.17), whereas in Section 3 we maximised over H_0 and H_1 to find the likelihood ratio statistic.

1. Here we are treating H_0 and H_1 in the same way. There is not the asymmetry that there was in Section 3 where we treated the null hypothesis H_0 in a different way to the alternative hypothesis H_1.
2. The Bayes factor B_{10} of H_1 relative to H_0 is just $B_{10} = (B_{01})^{-1}$, since the ratios above are simply inverted.
3. Bayes factors can only be used with proper priors: from (4.16) and (4.17) we see that B_{01} depends on two constants of propotionality, one for each the of priors $\pi(\theta_i \,|\, H_i)$, $i = 0, 1$, so both of these constants must be known.

From now on suppose our model is $f(\mathbf{x} \mid \theta)$ under both H_0 and H_1.

If $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ are both simple then

$$B_{01} = \frac{f(\mathbf{x} \mid \theta_0)}{f(\mathbf{x} \mid \theta_1)}$$

since the prior $\pi(\theta_i \mid H_i)$ corresponds to $\theta = \theta_i$ with probability 1. So B_{01} is just the likelihood ratio in favour of H_0.

If $H_i : \theta \in \Theta_i$, $i = 0, 1$, are both composite then

$$B_{01} = \frac{\int_{\Theta_0} f(\mathbf{x} \mid \theta)\pi(\theta \mid H_0)\, d\theta}{\int_{\Theta_1} f(\mathbf{x} \mid \theta)\pi(\theta \mid H_1)\, d\theta}.$$

If $H_0 : \theta = \theta_0$ is simple and $H_1 : \theta \in \Theta_1$ is composite, then

$$B_{01} = \frac{f(\mathbf{x} \mid \theta_0)}{\int_{\Theta_1} f(\mathbf{x} \mid \theta)\pi(\theta \mid H_1)\, d\theta}.$$

By analogy with the likelihood ratio statistic, the quantity $2 \log B_{01}$ is often used to summarise the evidence for H_0 compared to H_1, with rough interpretation as below (table from Davison, 2003).

B_{01}	$2 \log B_{01}$	Evidence for H_0
< 1	< 0	Negative (i.e. evidence supports H_1)
1–3	0–2	Hardly worth a mention
3–20	2–6	Positive
20–150	6–10	Strong
> 150	> 10	Very strong

Example 4.11. ["IQ test"] Suppose $X \sim N(\theta, \sigma^2)$ where $\sigma^2 = 100$. So $f(x \mid \theta) = \frac{1}{\sqrt{200\pi}} e^{-\frac{1}{200}(x-\theta)^2}$.

Let $H_0 : \theta = 100$ ("average") and $H_1 : \theta = 130$.

Suppose we observe $x = 120$. Then

$$B_{01} = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)} = \frac{f(120 \mid 100)}{f(120 \mid 130)} = 0.223.$$

So $B_{10} = 1/0.223 = 4.48$, so positive evidence for H_1.

Suppose the prior probabilities are $P(H_0) = 0.95$, $P(H_1) = 0.05$. Using posterior odds = Bayes factor $\times$ prior odds,

$$\frac{p_0}{1 - p_0} = B_{01} \times \frac{0.95}{0.05}$$

where $p_0 = P(H_0 \mid x)$. Solving,

$$p_0 = \frac{19 B_{01}}{1 + 19 B_{01}} = 0.81.$$

This posterior probability of H_0 is substantially decreased from its prior value (corresponding to $B_{01} = 0.223$ being small) but is still high.

Example 4.12. ["Weight"] Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where $\sigma^2 = 9$.

Let $H_0 : \theta \leqslant 175$ ("true weight $\leqslant 175$ pounds") and $H_1 : \theta > 175$.

Assume prior $\theta \sim N(\mu_0, \sigma_0^2)$ where $\mu_0 = 170$, $\sigma_0^2 = 5^2$.

Prior probability $P(H_0) = P(N(\mu_0, \sigma_0^2) \leqslant 175) = \Phi\left(\frac{175-170}{5}\right) = \Phi(1) = 0.84$. So the prior odds is $\frac{P(H_0)}{P(H_1)} = \frac{\Phi(1)}{1-\Phi(1)} = 5.3$.

Suppose we observe $x_1, \ldots, x_n$ where $n = 10$, $\overline{x} = 176$. Then from the normal example in Section 4.3 the posterior is $N(\mu_1, \sigma_1^2)$ where

$$\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = 175.8, \qquad \sigma_1^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} = 0.869.$$

Posterior probability $P(H_0 \,|\, \mathbf{x}) = P(N(\mu_1, \sigma_1^2) \leqslant 175) = \Phi\left(\frac{175-175.8}{\sqrt{0.869}}\right) = 0.198$. So the posterior odds is $\frac{P(H_0 \,|\, \mathbf{x})}{1-P(H_0 \,|\, \mathbf{x})} = 0.24$.

So Bayes factor $B_{01} = \frac{\text{post odds}}{\text{prior odds}} = 0.0465$ and $B_{10} = (B_{01})^{-1} = 21.5$.

So the data provide strong evidence in favour of H_1, we conclude that $\theta \leqslant 175$ is unlikely.

Example 4.13. [From Carlin and Louis (2008).] Suppose we have two products: P_0, an old standard product; and P_1, newer and more expensive. Let θ be the probability that a customer prefers P_1. Let the prior $\pi(\theta)$ be Beta(a, b).

Assume that the number of customers X (out of n) that prefer P_1 is $X \sim$ Binomial(n, θ). Then from Section 4.1 we know that the posterior $\pi(\theta \,|\, x)$ is Beta$(x + a, n - x + b)$.

Let's say $\theta \geqslant 0.6$ means that P_1 is a substantial improvement over P_0, so take $H_0 : \theta \geqslant 0.6$ and $H_1 : \theta < 0.6$.

Suppose $a = b = 1$, i.e. prior $\theta \sim U(0, 1)$. Then

$$P(H_0) = \int_{0.6}^{1} \pi(\theta)\, d\theta = 0.4 \quad \text{and} \quad P(H_1) = \int_{0}^{0.6} \pi(\theta)\, d\theta = 0.6.$$

Suppose we have $x = 13$ "successes" from $n = 16$ customers. Then

$$P(H_0 \,|\, x) = \int_{0.6}^{1} \pi(\theta \,|\, x)\, d\theta = \int_{0.6}^{1} \frac{1}{B(14, 4)} \theta^{13}(1-\theta)^3\, d\theta = 0.964$$
$$P(H_1 \,|\, x) = \int_{0}^{0.6} \pi(\theta \,|\, x)\, d\theta = \int_{0}^{0.6} \frac{1}{B(14, 4)} \theta^{13}(1-\theta)^3\, d\theta = 0.046.$$

So

$$\text{prior odds} = 0.4/0.6 = 0.67$$
$$\text{posterior odds} = 0.964/0.046 = 20.96$$
$$\text{Bayes factor } B_{01} = \frac{\text{posterior odds}}{\text{prior odds}} = 31.1.$$

Conclusion: we interpret $B_{01} = 31.1$ as a strong evidence for H_0.

We can also calculate the Bayes factor using marginal likelihoods (see (4.16) and (4.17)). The quantity $\pi(\theta \mid H_i)$ is the prior for θ, conditional on H_i being true. So

$$\pi(\theta \mid H_0) = \begin{cases} 0 & \text{if } 0 < \theta < 0.6 \\ \pi(\theta)/P(H_0) & \text{if } 0.6 \leqslant \theta < 1 \end{cases}$$

$$= \begin{cases} 0 & \text{if } 0 < \theta < 0.6 \\ 1/P(H_0) & \text{if } 0.6 \leqslant \theta < 1 \end{cases}$$

and similarly

$$\pi(\theta \mid H_1) = \begin{cases} 1/P(H_1) & \text{if } 0 < \theta < 0.6 \\ 0 & \text{if } 0.6 \leqslant \theta < 1. \end{cases}$$

So

$$P(x \mid H_0) = \int_0^1 f(x \mid \theta)\pi(\theta \mid H_0)\, d\theta$$
$$= \int_{0.6}^1 \binom{16}{13}\theta^{13}(1-\theta)^3 \times \frac{1}{P(H_0)}\, d\theta$$

and similarly

$$P(x \mid H_1) = \int_0^{0.6} \binom{16}{13}\theta^{13}(1-\theta)^3 \times \frac{1}{P(H_1)}\, d\theta$$

and then the Bayes factor is $B_{01} = P(x \mid H_0)/P(x \mid H_1)$.

4.5 Asymptotic normality of posterior distribution

From (4.5) we have $\pi(\theta \mid \mathbf{x}) \propto L(\theta)\pi(\theta)$, where $L(\theta)$ is the likelihood. Let $\widetilde{\ell}(\theta) = \log \pi(\theta \mid \mathbf{x})$ be the log posterior density,

$$\widetilde{\ell}(\theta) = \text{constant} + \log \pi(\theta) + \ell(\theta)$$
$$= \text{constant} + \log \pi(\theta) + \sum_{i=1}^n \log f(x_i \mid \theta)$$

and there are n terms in the sum, so expect the likelihood contribution to dominate $\widetilde{\ell}(\theta)$ for large n.

Let $\widetilde{\theta}$ be the posterior mode, assume $\widetilde{\ell}'(\widetilde{\theta}) = 0$, and assume $\widetilde{\theta}$ lies in the interior of the parameter space Θ. Then

$$\widetilde{\ell}(\theta) \approx \widetilde{\ell}(\widetilde{\theta}) + (\theta - \widetilde{\theta})\widetilde{\ell}'(\widetilde{\theta}) + \frac{1}{2}(\theta - \widetilde{\theta})^2\widetilde{\ell}''(\widetilde{\theta})$$
$$= \text{constant} - \frac{1}{2}(\theta - \widetilde{\theta})^2\widetilde{J}(\widetilde{\theta}) \tag{4.18}$$

where $\widetilde{J}(\theta) = -\widetilde{\ell}''(\theta)$. Note: in (4.18) $\widetilde{\ell}(\widetilde{\theta})$ is just a constant since it does not depend on θ.

So

$$\pi(\theta \mid \mathbf{x}) = \exp(\widetilde{\ell}(\theta))$$
$$\propto \exp(-\tfrac{1}{2}(\theta - \widetilde{\theta})^2\widetilde{J}(\widetilde{\theta}))$$

is our approximation which, as it's a function of θ, is of the form of a normal pdf with mean $\widetilde{\theta}$ and variance $\widetilde{J}(\widetilde{\theta})^{-1}$. That is, we have

$$\theta \,|\, \mathbf{x} \approx N(\widetilde{\theta}, \widetilde{J}(\widetilde{\theta})^{-1}) \tag{4.19}$$

In large samples, the likelihood contribution to $\pi(\theta \,|\, \mathbf{x})$ is much larger than the prior contribution, resulting in $\widetilde{\theta}$ and $\widetilde{J}(\widetilde{\theta})$ being essentially the same as the MLE $\widehat{\theta}$ and observed information $J(\widehat{\theta})$. Hence we also have

$$\theta \,|\, \mathbf{x} \approx N(\widehat{\theta}, J(\widehat{\theta})^{-1}). \tag{4.20}$$

[We can also obtain (4.20) via a Taylor expansion about $\widehat{\theta}$.]

The corresponding frequentist result looks similar: in Section 1.7 we saw

$$\widehat{\theta} \approx N(\theta, I(\theta)^{-1}) \quad \text{and} \quad \widehat{\theta} \approx N(\theta, J(\theta)^{-1}). \tag{4.21}$$

However, note that in (4.19) and (4.20) the parameter θ is a RV, and $\widetilde{\theta} = \widetilde{\theta}(\mathbf{x})$ and $\widehat{\theta} = \widehat{\theta}(\mathbf{x})$ are constants. In contrast, in (4.21) the quantity $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$ is a RV, and θ is treated as a constant.

Using the asymptotic results:

(i) the frequentist approximation $\widehat{\theta} \approx N(\theta, J(\theta)^{-1})$ leads to a 95% confidence interval of $\left(\widehat{\theta} \pm 1.96 J(\widehat{\theta})^{-1/2}\right)$

(ii) the Bayesian approximation $\theta \,|\, \mathbf{x} \approx N(\widehat{\theta}, J(\widehat{\theta})^{-1})$ leads to a 95% confidence interval of $\left(\widehat{\theta} \pm 1.96 J(\widehat{\theta})^{-1/2}\right)$.

The set of values of θ in (i) and (ii) are the same, but their interpretations – as frequentist in (i), and Bayesian in (ii) – are different.

SLIDES. Normal approximation slides go here.