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.

That is, we assume we have a parametric family.

Example $X \sim \text{Poisson}(\lambda)$.

Then $\theta = \lambda \in \Theta = (0, \infty)$.

Example $X \sim N(\mu, \sigma^2)$

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

Suppose we have data $\underline{x} = (x_1, \dots, x_n)$, numerical values. We regard these as observed values of i.i.d. random variables $X_1, \dots, X_n$ with the same distribution as X, so $\underset{\sim}{X} = (X_1, \dots, X_n)$ is a random sample.

Having observed $\underset{\sim}{X} = \underline{x}$, what can we infer/say about θ?

E.g. we might wish to:

- make a point estimate of θ
- construct an interval estimate for θ
- test a hypothesis about θ, e.g. test whether $\theta = 0$.

Approximately:

first two thirds of the course on the frequentist approach to questions like these

Last third will look at the Bayesian approach.

Notation

Since the distribution of X depends on θ, we write the probability mass function (p.m.f.) / probability denisity function (p.d.f.) of X as $f(x;\theta)$.

If X discrete: we have $f(x;\theta) = P(X=x)$, the p.m.f.

X continuous: $f(x;\theta)$ is the p.d.f.

We write $f(\underset{\sim}{x};\theta)$ for the joint pmf / pdf of $\underset{\sim}{X} = (X_1, \dots, X_n)$.

Assuming the X_i are independent,

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

Example $X_i \sim \text{Poisson}(\theta)$.

Then $f(x;\theta) = \dfrac{e^{-\theta}\,\theta^{x}}{x!}$, $\quad x = 0, 1, 2, \ldots$

So $f(\underset{\sim}{x};\theta) = \prod_{i=1}^{n} \dfrac{e^{-\theta}\,\theta^{x_i}}{x_i!} = \dfrac{e^{-n\theta}\,\theta^{\sum x_i}}{\prod x_i!}$.

Estimators

An **estimator** is any function $t(\underline{X})$ we might use to estimate θ.

Note: the function t is not allowed to depend on θ.

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

The estimator $T = t(\underline{X})$ is **unbiased** for θ if

$$E(T) = \theta \quad \text{for all } \theta.$$

Likelihood

The likelihood for θ, based on $\underset{\sim}{x}$, is $L(\theta; \underset{\sim}{x}) = f(\underset{\sim}{x}; \theta)$

where L is regarded as a function of θ, for a fixed $\underset{\sim}{x}$.

We often write $L(\theta)$ for $L(\theta; \underset{\sim}{x})$.

The log-likelihood is $\ell(\theta) = \log L(\theta)$

or sometimes $\ell(\theta; \underset{\sim}{x})$

or sometimes $\ell(\theta; \underset{\sim}{X})$.

Maximum likelihood

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

The maximum likelihood estimator is $\hat{\theta}(\underset{\sim}{X})$.

1.2 Delta method

Suppose $X_1, \ldots, X_n$ are iid with $E(X_i) = \mu$, $\mathrm{var}(X_i) = \sigma^2$.

By Central Limit Theorem (CLT),

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \qquad \text{for lage } n.$$

We would often like to know the asymptotic (i.e. lage n) distribution of $g(\bar{X})$ for some function g.

E.g. $\hat{\theta} = 1/\bar{X}$ and we want lage sample dist. of $\hat{\theta}$.

Taylor expansion:

$$g(\bar{X}) = g(\mu) + (\bar{X}-\mu)g'(\mu) + \dots$$

Approximate: $g(\bar{X}) \approx g(\mu) + (\bar{X}-\mu)g'(\mu)$ ①

Take expectations in ①: $E[g(\bar{X})] \approx g(\mu) + g'(\mu)\underbrace{E[\bar{X}-\mu]}_{0}$

$$= g(\mu) \text{ since } E(\bar{X}) = \mu$$

variance in ①: $\text{var}[g(\bar{X})] \approx \text{var}[g'(\mu)(\bar{X}-\mu)]$

$$= g'(\mu)^2 \, \text{var}(\bar{X})$$

$$= g'(\mu)^2 \frac{\sigma^2}{n} \qquad \text{since} \quad \text{var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Also from ①, $g(\bar{X})$ is approx normal since $\bar{X}$ is approx normal. Hence

$$g(\bar{X}) \overset{D}{\approx} N\left(g(\mu),\ g'(\mu)^2 \frac{\sigma^2}{n} \right)$$

asymptotic distribution

asymp. mean

asymp. variance

This is the <u>delta method</u>.

Example $X_1, \dots, X_n$ iid exponential with parameter or rate λ.

So pdf $f(x;\lambda) = \lambda e^{-\lambda x}, \quad x > 0$

and $\mu = E(X_i) = \dfrac{1}{\lambda}, \qquad \sigma^2 = \mathrm{var}(X_i) = \dfrac{1}{\lambda^2}.$

Let $g(\bar{X}) = \log \bar{X}$. With $g(u) = \log u$,

asymptotic mean $\quad g(\mu) = \log \mu = -\log \lambda$

asymptotic variance $\quad g'(\mu)^2 \dfrac{\sigma^2}{n} = \dfrac{1}{\mu^2} \cdot \dfrac{\sigma^2}{n} = \lambda^2 \cdot \dfrac{1}{n\lambda^2} = \dfrac{1}{n}$

Hence $g(\bar{X}) = \log \bar{X} \approx N\left(-\log \lambda, \frac{1}{n}\right).$

1.3 Order statistics

The order statistics of $x_1, \ldots, x_n$ are their values in increasing order, denoted $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

The sample median m is

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

The

lower quartile has $1/4$ of the sample less than it

upper quartile has $3/4$ ---------------

(defined in terms of $x_{(\lfloor n/4 \rfloor)}$ etc)

inter-quartile range IQR = upper quartile − lower quartile

The random variable versions of these are defined similarly

For random variables X_i,

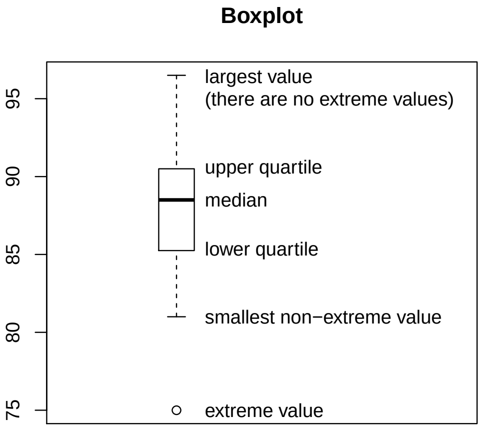
order statistics $\quad X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

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

and so on.

Boxplots

A boxplot, or box-and-whisker plot, is a convenient way of summarising data, particularly when the data is made up of several groups.



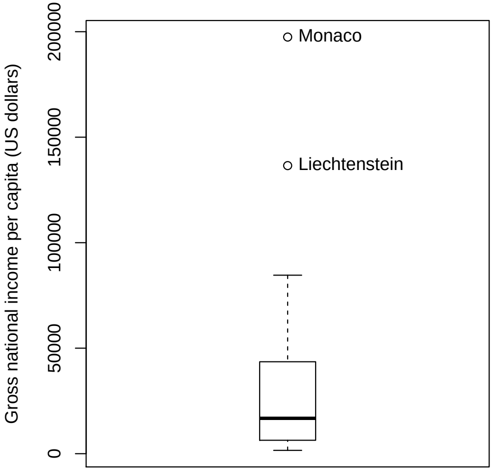
The box extends from one quartile to the other, and the central line in the box is the median.

The whiskers are drawn from the box to the most extreme observations that are no more than $1.5 \times \mathrm{IQR}$ from the box. (Alternatively $r \times \mathrm{IQR}$ can be used for other values of r.)

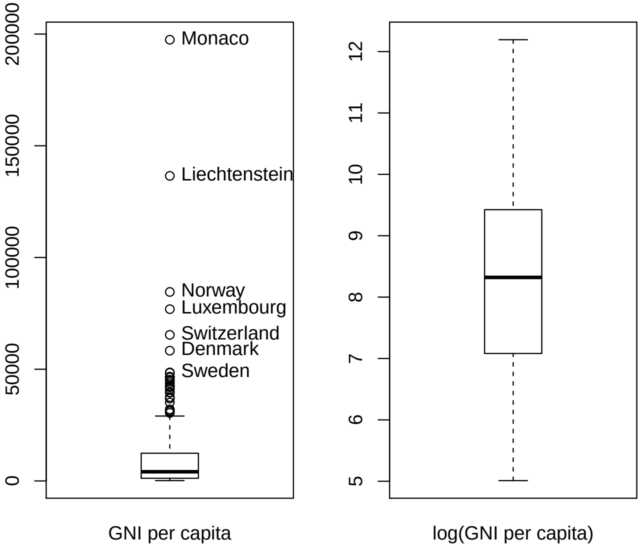
Observations which are more extreme than this are shown separately.

Gross national income per capita for 50 “sovereign states in Europe.”

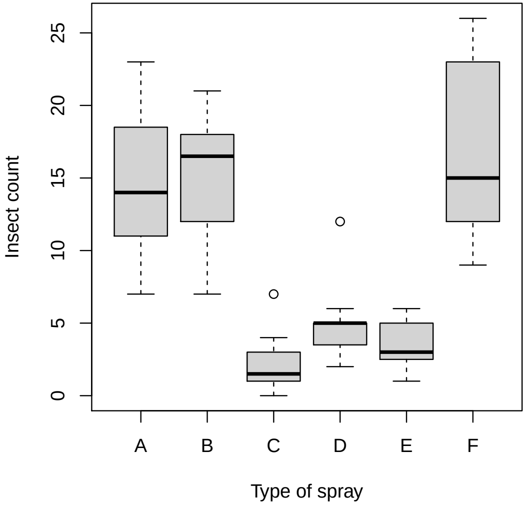
http://en.wikipedia.org/wiki/List_of_sovereign_states_in_Europe_by_GNI_(nominal)_per_capita



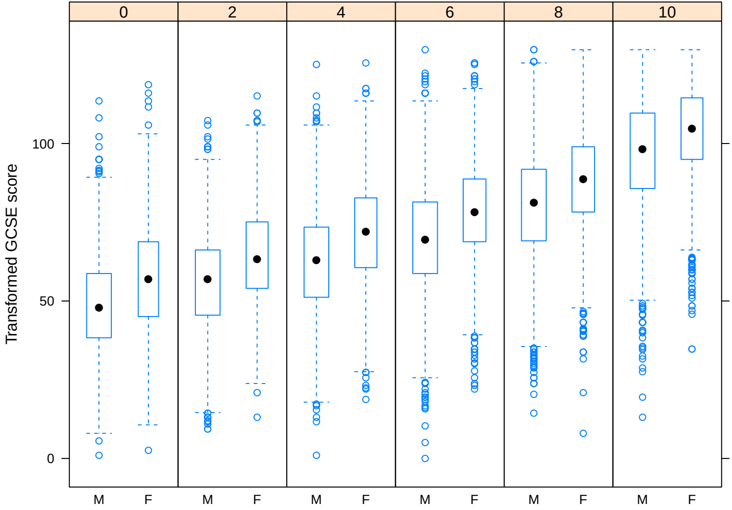
Now for 182 countries worldwide (including Europe).



Parallel boxplots are often useful to show the differences between subgroups of the data. Below: `InsectSprays` data from R.



Comparative boxplots of transformed GCSE scores by A-level chemistry exam score ($0 = $ worst, $2, 4, 6, 8, 10 = $ best) and gender.



Distribution of $X_{(r)}$

Assume the X_i are iid from a continuous distribution with cdf F, pdf f.

So $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ with probability 1.

What is the distribution of $X_{(r)}$?

$\underline{r=1}$ The cdf of $X_{(1)}$ is

$$F_{(1)}(x) = P(X_{(1)} \leq x)$$

$$= 1 - P(X_{(1)} > x)$$

$$= 1 - P(X_1 > x, \ldots, X_n > x)$$

$$= 1 - P(X_1 > x) \ldots P(X_n > x) \quad \text{since } X_i \text{ indep}$$

$$= 1 - [1 - F(x)]^n$$

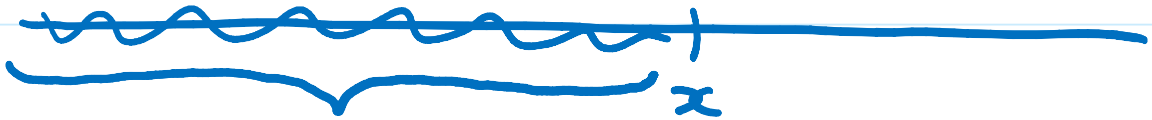
So pdf $f_{(1)}(x) = F_{(1)}'(x) = n[1 - F(x)]^{n-1} \cdot f(x)$

Theorem 1.1 The pdf of $X_{(r)}$ is

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

Proof By induction. We did the case $r=1$ above.
So assume true at r.

For any r:



the number of $X_i \leqslant x$ is Binomial$(n, F(x))$.

So for any r the cdf of $X_{(r)}$ is

$$F_{(r)}(x) = P(X_{(r)} \leq x)$$

$$= \sum_{j=r}^{n} \binom{n}{j} F(x)^j \left[1-F(x)\right]^{n-j}$$

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

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

Differentiating,

$$f_{(r+1)}(x) = f_{(r)}(x)$$

$$- \binom{n}{r} F(x)^{r-1} [1-F(x)]^{n-r-1} [r - nF(x)] f(x)$$

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

using ind. hypothesis

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

So result follows by induction. $\square$

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

$x \qquad x+\delta x$

prob of X_i in this interval

$= F(x)$ $\qquad \approx f(x)\delta x$ $\qquad \approx 1-F(x)$

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

$r-1$ of the X_i in $(-\infty, x)$

1 — — — — $[x, x+\delta x)$

$n-r$ — — — — $[x+\delta x, \infty)$

Approximately, 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 $f_{(r)}(x)$

(i.e. divide by δx and let $\delta x \to 0$).

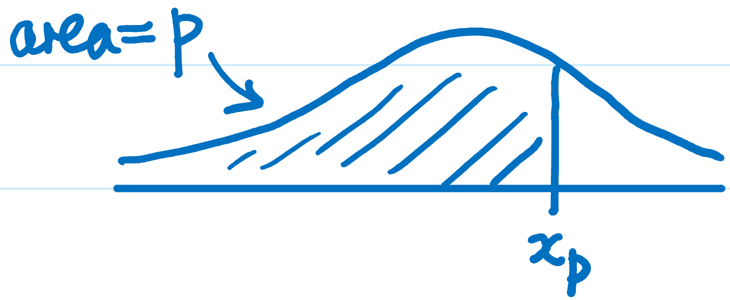
1.4 Q-Q plots

"quantile-quantile plot"

A Q-Q plot can be used to assess if it is reasonable to assume a set of data comes from a certain distribution.

The p^{th} quantile is the value x_p such that

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



Lemma 1.2 Suppose X a continuous random variable taking values in (a,b) with a strictly increasing cdf $F(x)$ for $x \in (a,b)$.

Let $Y = F(X)$. Then $Y \sim U(0,1)$.

$F(X)$ is sometimes called the <u>probability integral transform</u> of X.

We can write the result as $\quad F(X) \sim U$

or, applying F^{-1}, $\quad X \sim F^{-1}(U)$.

Lemma 1.3 If $U_{(1)}, \dots, U_{(n)}$ 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) $\operatorname{var}[U_{(r)}] = \dfrac{r}{(n+1)(n+2)}\left(1 - \dfrac{r}{n+1}\right)$

Note: $\operatorname{var}[U_{(r)}] = \dfrac{1}{n+2} p_r(1-p_r)$ where $p_r = \dfrac{r}{n+1}$

$$\leq \frac{1}{n+2} \cdot \frac{1}{4} = O\left(\frac{1}{n}\right).$$

Question: is it reasonable to assume data $x_1, \ldots, x_n$ are a random sample from F?

By Lemma 1.2 we can generate a random sample $X_1, \ldots, X_n$ from F by first taking $U_1, \ldots, U_n \overset{iid}{\sim} U(0,1)$ and then setting $X_k = F^{-1}(U_k)$.

The order statistics are $X_{(k)} = F^{-1}(U_{(k)})$. ①

If F is a reasonable distribution to assume, then we expect $x_{(k)}$ to be fairly close to $E[X_{(k)}]$.

Now

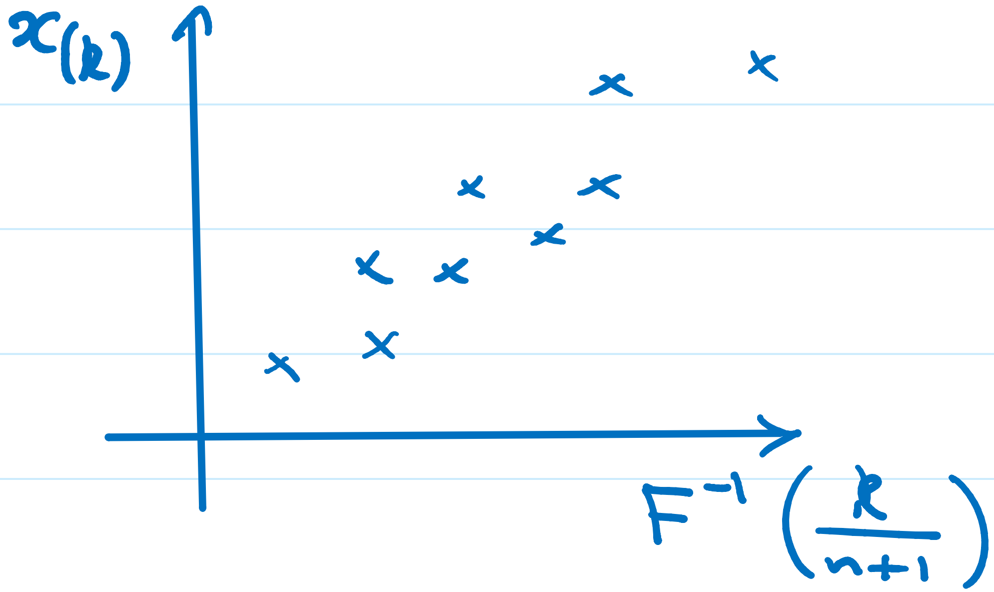
$$E[X_{(k)}] = E[F^{-1}(U_{(k)})] \quad \text{from } ①$$

$$\approx F^{-1}\left(E[U_{(k)}]\right) \quad \text{(eg delta method)}$$

$$= F^{-1}\left(\frac{k}{n+1}\right) \quad \text{by Lemma 1.3.}$$

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

In a **Q-Q plot** we plot the values of $x_{(k)}$ against $F^{-1}\left(\frac{k}{n+1}\right)$ for $k=1,\dots,n$



A Q-Q plot is a plot of observed values $x_{(k)}$ against the corresponding approx expectations $F^{-1}\left(\frac{k}{n+1}\right)$.

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

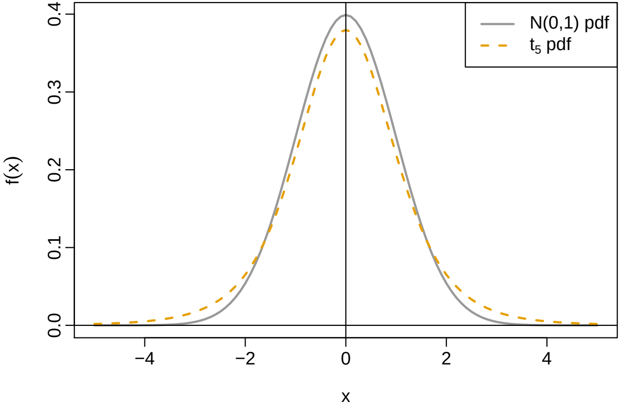
Of course we need to specify a candidate cdf F.

Comparing $N(0, 1)$ and t distributions

A t-distribution with r degrees of freedom has pdf

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

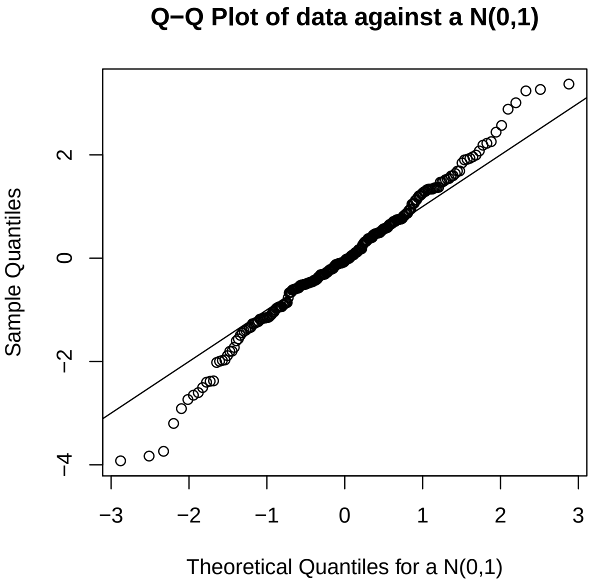
[More on t-distributions later.] Consider $r = 5$.



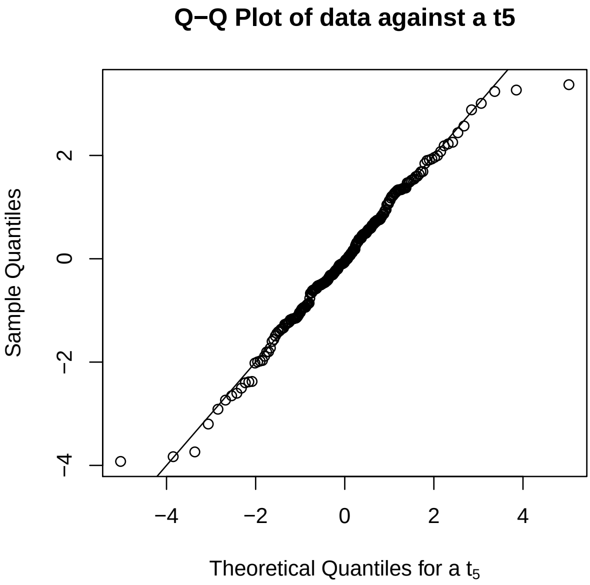
Suppose we simulate data $(x_1, \ldots, x_{250})$ from a t_5 distribution.

Using Q-Q plots we can consider the questions:

- is it reasonable to assume $(x_1, \ldots, x_{250})$ is from a $N(0,1)$?
- is it reasonable to assume $(x_1, \ldots, x_{250})$ is from a t_5?



A $N(0,1)$ assumption is not good – as expected.



A t_5 assumption is ok – as expected.

In practice F usually depends on an unknown parameter θ, so F and F^{-1} are unknown.

How do we handle this?

Normal Q-Q plot

If data $\underset{\sim}{x}$ are from a $N(\mu, \sigma^2)$ distribution, for some unknown μ, σ^2; then we have

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

where F is the cdf for $N(\mu, \sigma^2)$.

If $Y \sim N(\mu, \sigma^2)$ then

$$P(Y \leq y) = P\Big(\underbrace{\frac{Y-\mu}{\sigma}}_{N(0,1)} \leq \frac{y-\mu}{\sigma}\Big)$$

$$= \Phi\left(\frac{y-\mu}{\sigma}\right) \quad \text{where } \Phi \text{ is } N(0,1) \text{ cdf.}$$

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

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

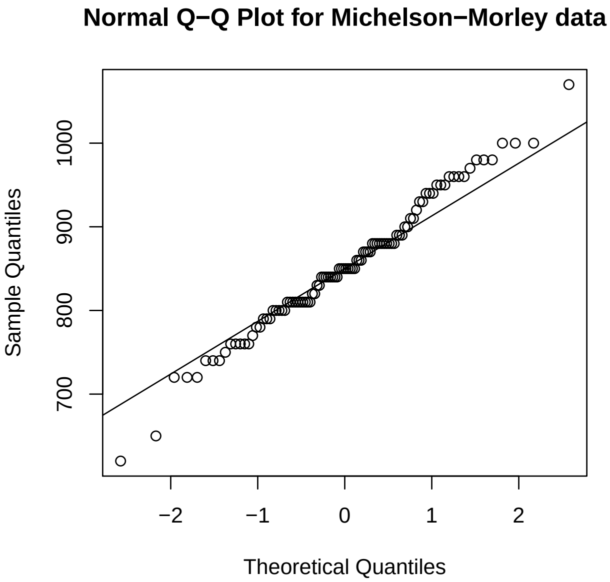
So we can plot $x_{(k)}$ against $\Phi^{-1}\left(\frac{k}{n+1}\right)$

for $k = 1 \ldots n$ and see if the points lie on an approx. straight line

(with gradient σ, intercept μ).

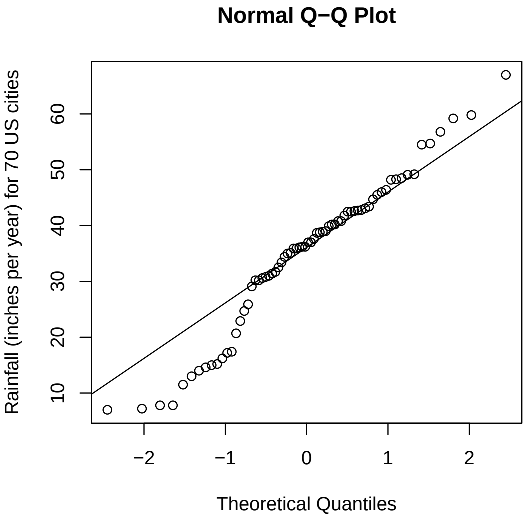
Normal Q-Q plots

20 observations from each experiment. Is a $N(\mu, \sigma^2)$ distribution plausible for these 100 observations?



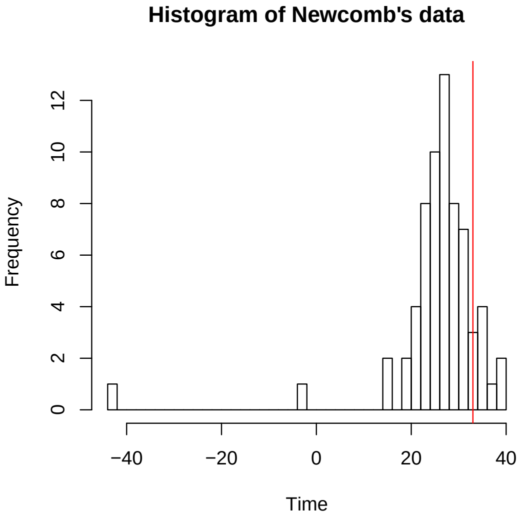
From the plot a normal distribution seems reasonable.

Below: `precip` data from R – average precipitation for 70 US cities.

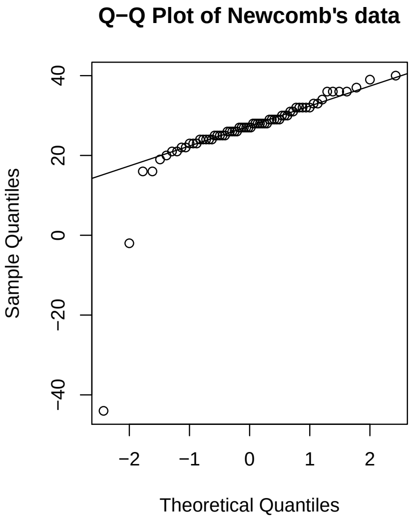


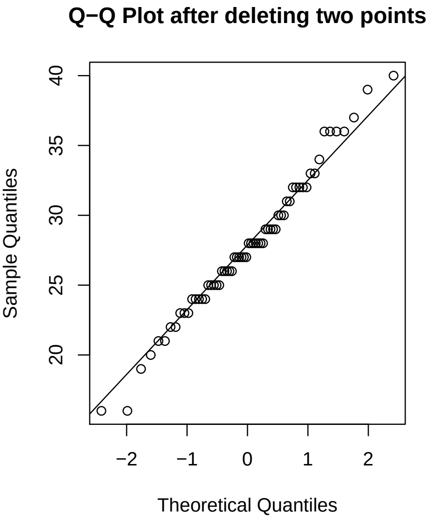
A normal assumption doesn't look good – problems in the lower tail.

Below: Newcomb's (1882) speed of light data – measurements are the time (in deviations from 24800 nanoseconds) to travel about 7400m. The currently accepted time (on this scale) is 33.



This time the problems are different – two (very small) outlying observations. If these are removed, a normal assumption looks ok.





Exponential Q-Q plot

The exponential distribution with mean μ has cdf $F(x) = 1 - e^{-x/\mu}$, $x > 0$.

If data $\underset{\sim}{x}$ have this distribution (μ unknown) then

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

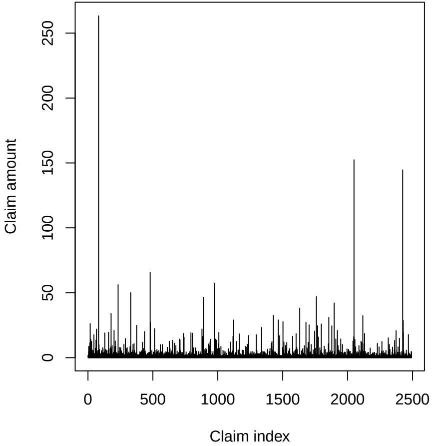
Hence $x_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right)$.

So plot $x_{(k)}$ against $-\log\left(1-\frac{k}{n+1}\right)$

and see if points lie on approx straight line (gradient μ, intercept 0).

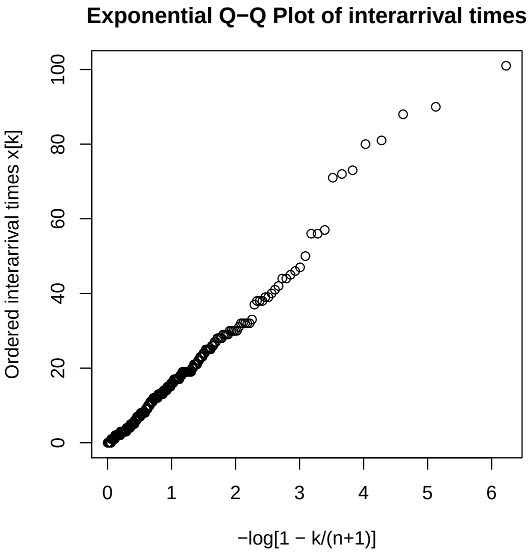
Example: Danish fire data (Davison, 2003)

Data on the times, and amounts, of major insurance claims due to fire in Denmark 1980–90.

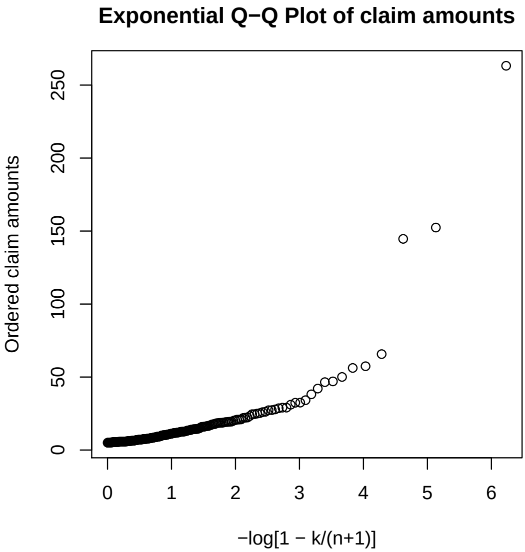


Following Davison, let's consider the 254 largest claim amounts, and the interarrival times between these claims.

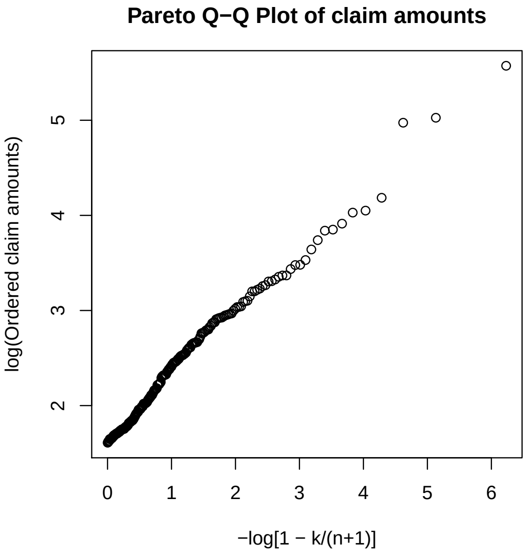
Is it reasonable to assume exponential interarrival times? See below – inter-arrivals look fairly close to exponential.



Is it reasonable to assume exponential claim amounts? See below – an exponential assumption is not reasonable.



Is it reasonable to assume Pareto claim amounts? See below – the Pareto fits fairly well.



1.5 Multivariate normal distribution

See lecture notes for some reminders about the multivariate normal distribution (Prelims Stats; Part A Prob).

1.6 Information

Definition In a model with scalar parameter θ and log-likelihood $\ell(\theta)$, the **observed information** $J(\theta)$ is defined by $J(\theta) = -\dfrac{d^2\ell}{d\theta^2}$.

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

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

Example $X_1, \dots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} \right) = -n\theta + \sum x_i \log\theta$$

$$- \log\left(\prod x_i!\right)$$

observed information:

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

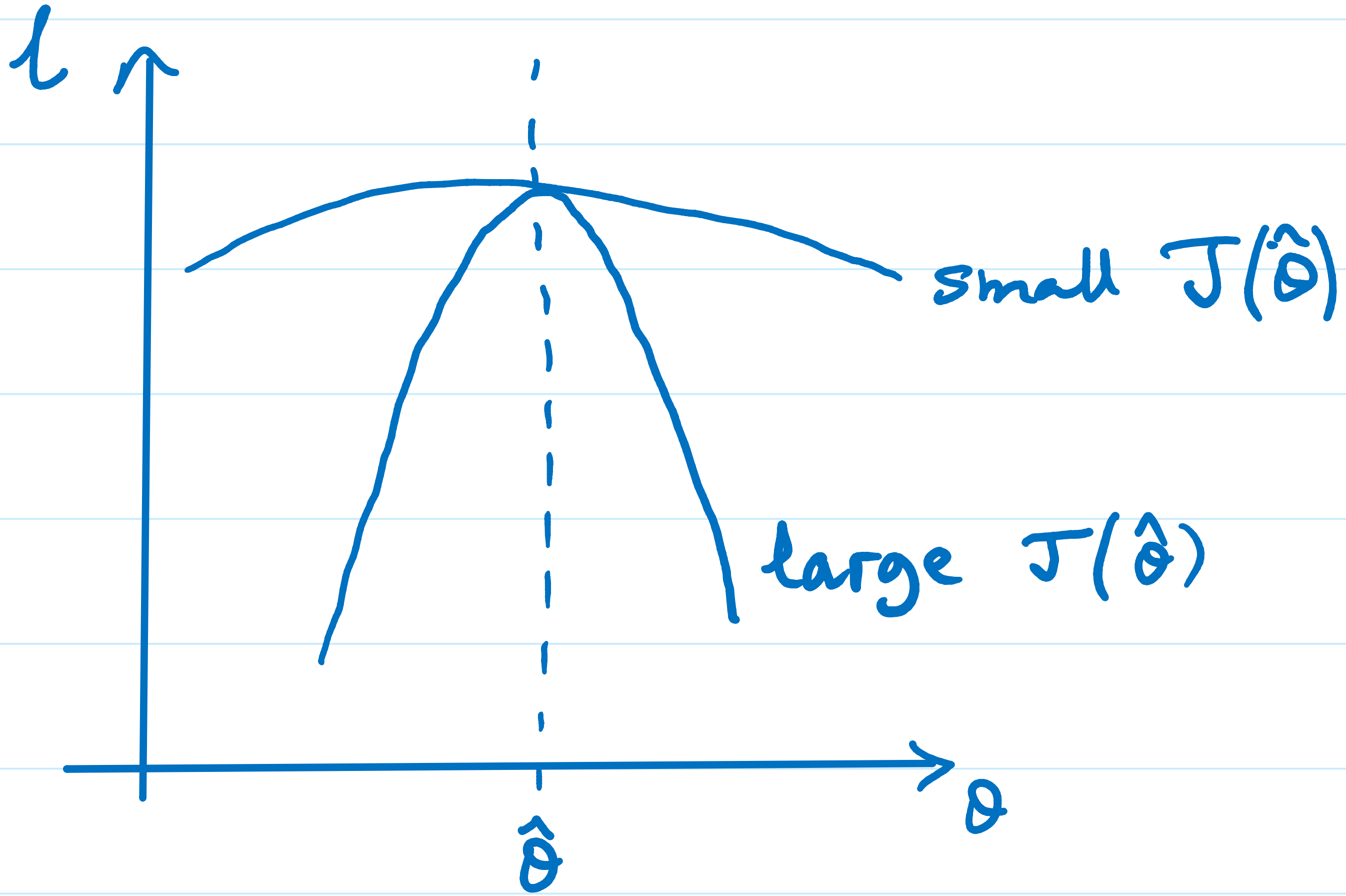
Expanding $\ell(\theta)$ as a Taylor series about $\hat{\theta}$:

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

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

$$\ell(\theta) \approx \ell(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 J(\hat{\theta})$$

a quadratic approx to $\ell(\theta)$



The larger $J(\hat{\theta})$ is, the more concentrated $\ell(\theta)$ is about $\hat{\theta}$ and the "more information" we have about θ.

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

$$I(\theta) = E\left[-\frac{d^2 l}{d\theta^2} \right].$$

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

$$I(\theta)_{jk} = E\left[- \frac{\partial^2 l}{\partial \theta_j \partial \theta_k} \right].$$

Note:

(i) when calculating $I(\theta)$ we treat log-lik ℓ as $\ell(\theta; \underset{\sim}{X})$ and take expectations over $\underset{\sim}{X}$.

(ii) if $X_1, \dots, X_n$ are iid then $I(\theta) = n . i(\theta)$ where $i(\theta)$ is the expected information in a sample of size 1.

So (i) is saying $I(\theta) = E\left[- \dfrac{d^2 \ell(\theta; \underset{\sim}{X})}{d\theta^2} \right]$.

<u>Example</u> $X_1, \dots, X_n \overset{iid}{\sim}$ exponential with pdf

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta}, \qquad x > 0.$$

Note $E(X_i) = \theta$.

$$\ell(\theta) = \log\left(\prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta}\right) = -n\log\theta - \frac{\sum x_i}{\theta}$$

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

$$I(\theta) = E\left[\frac{-n}{\theta^2} + \frac{2\sum x_i}{\theta^3}\right]$$

$$= \frac{-n}{\theta^2} + \frac{2}{\theta^3}\sum E(x_i)$$

$$= \frac{-n}{\theta^2} + \frac{2}{\theta^3} \cdot n\theta \qquad \text{since } E(x_i) = \theta$$

$$= \frac{n}{\theta^2}.$$

1.7 Properties of MLEs

Invariance property

Example $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$.

What is the MLE of $\psi = P(X_1 = 0) = e^{-\theta}$?

More generally, suppose we want to estimate ψ, where $\psi = g(\theta)$ and g is a 1-1 function.

For max likelihood estimation of ψ we maximise $f(\underline{x};\, g^{-1}(\psi))$ with respect to ψ.

As the maximum value of f is $f(\underline{x};\, \hat{\theta})$

the maximising value of ψ satisfies $g^{-1}(\psi) = \hat{\theta}$

i.e. $\psi = g(\hat{\theta})$

That is, the MLE of ψ is $\underline{\hat{\psi} = g(\hat{\theta})}$.

invariance property of MLEs

Example continued $(\psi = e^{-\theta})$

We know $\hat{\theta} = \bar{x}$.

The invariance property tells us $\hat{\psi} = e^{-\hat{\theta}}$

$$= e^{-\bar{x}}.$$

Iterative calculation of $\hat{\theta}$

Often $\hat{\theta}$ satisfies the likelihood equation $l'(\hat{\theta}) = 0$.
We often have to solve this equation numerically, e.g using Newton-Raphson.

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

$$0 = l'(\hat{\theta}) \approx \underbrace{l'(\theta^{(0)})}_{U} + (\hat{\theta} - \theta^{(0)}) \underbrace{l''(\theta^{(0)})}_{-J}$$

Rearranging: $\hat{\theta} \approx \theta^{(0)} + \dfrac{U(\theta^{(0)})}{J(\theta^{(0)})}$

where $U(\theta) = \dfrac{dl}{d\theta}$ is called the score function.

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

$$\theta^{(n+1)} = \theta^{(n)} + \frac{U(\theta^{(n)})}{J(\theta^{(n)})}, \quad n \geq 0$$

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

Asymptotic normality of $\hat{\theta}$

Let θ be a scalar and consider the MLE $\hat{\theta}(\underset{\sim}{X})$, which is a random variable.

Subject to regularity conditions, as $n \to \infty$,

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

So for large n we have the asymptotic distribution:

$$\hat{\theta} \approx N(\theta, I(\theta)^{-1}).$$

The above asymptotic distribution also holds when θ is a vector, when it denotes a multivariate normal.

Slutsky's Theorem Suppose $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$ as $n \to \infty$, where c is a constant.

Then (i) $X_n + Y_n \xrightarrow{D} X + c$

(ii) $X_n Y_n \xrightarrow{D} cX$

(iii) $\dfrac{X_n}{Y_n} \xrightarrow{D} \dfrac{X}{c}$ if $c \neq 0$.

Sketch proof of asymptotic normality, θ scalar

Assume $\hat{\theta}$ solves $\ell'(\hat{\theta}) = 0$.

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

$$= U(\theta) - (\hat{\theta} - \theta) J(\theta).$$

Hence $\hat{\theta} - \theta \approx \dfrac{U(\theta)}{J(\theta)}$.

So $I(\theta)^{1/2}(\hat{\theta} - \theta) \approx I(\theta)^{1/2} \cdot \dfrac{U(\theta)}{J(\theta)}$

$$= \frac{U(\theta) / I(\theta)^{1/2}}{J(\theta) / I(\theta)} = \frac{\text{TOP}}{\text{BOTTOM}} \quad (1).$$

For TOP:

$$U(\theta) = \frac{d}{d\theta} \log\left(\prod_{j=1}^{n} f(X_j; \theta) \right) = \sum_{j=1}^{n} U_j$$

where $U_j = \frac{d}{d\theta} \log f(X_j; \theta)$.

The U_j are iid. We'll apply the CLT.

Now $1 = \int f(x; \theta)\, dx \quad (*)$ 1-dim integral

Note: $\frac{df}{d\theta} = \left(\frac{d}{d\theta} \log f \right) . f$

Diff (*) with respect to θ:

$$0 = \int \frac{df}{d\theta}\, dx = \int \underbrace{\left(\frac{d}{d\theta}\log f\right)}_{U_j} \cdot f \, dx \qquad (a)$$

$$\text{Diff again:} \quad 0 = \int \left(\frac{d^2}{d\theta^2}\log f\right) f \, dx + \int \underbrace{\left(\frac{d}{d\theta}\log f\right)^2}_{U_j^2} f \, dx \qquad (b)$$

From (a): $\quad 0 = E(U_j)$

(b): $\quad 0 = -i(\theta) + E(U_j^2)$.

So $E(U) = \sum E(U_j) = 0$.

And $\mathrm{var}(U) = \sum \mathrm{var}(U_j)$ since U_j indep.

$$= n \cdot i(\theta)$$

$$= I(\theta)$$

Hence $\mathrm{TOP} = \dfrac{U(\theta)}{I(\theta)^{1/2}} = \dfrac{\sum U_j}{\sqrt{\mathrm{var}(\sum U_j)}}$

$$\xrightarrow{D} N(0,1) \quad \text{by CLT. (2)}$$

For BOTTOM:

Let $Y_j = \frac{d^2}{d\theta^2} \log f(X_j ; \theta)$ and $\mu_Y = E(Y_j)$.

Then $\text{BOTTOM} = \frac{J(\theta)}{I(\theta)} = \frac{\sum Y_j}{n \mu_Y} = \frac{\bar{Y}}{\mu_Y}$

$\xrightarrow{P} 1$ using WLLN (3)

Combining (1), (2), (3) and Slutsky (iii) gives

$$I(\theta)^{1/2} \cdot (\hat{\theta} - \theta) \xrightarrow{D} N(0,1). \qquad \square$$

The regularity conditions for the proof include:

- true value of θ is in interior of Θ
- MLE is given by solution of likelihood eq.
- can diff sufficiently often w.r.t. θ
- can interchange diff and integration suff. often

This means cases where the set $\{x : f(x;\theta) > 0\}$ depends on θ are excluded.

E.g. $U(0, \theta)$ is excluded.

2. Confidence Intervals

Let $\alpha \in (0,1)$.

A 1-α confidence interval is an interval $C=(a,b)$, where $a=a(\underset{\sim}{X})$ and $b=b(\underset{\sim}{X})$ such that

$$P(\theta \in C) = 1-\alpha.$$

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

In words: (a,b) traps θ with probability $1-\alpha$

Warning: C is random and θ is fixed

Most common is $\alpha = 0.05$, i.e. 95% $\underbrace{\text{C.I}}_{\text{confidence interval}}$

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

Example $X_1, \dots, X_n \stackrel{iid}{\sim} N(\theta, 1)$

and $\left(\bar{X} \pm \dfrac{1.96}{\sqrt{n}}\right)$ is a 95% C.I. for θ.

We'll usually want a central (equal tail) interval as above.

[One-sided intervals of the form (a, ∞) or $(-\infty, b)$ are possible.]

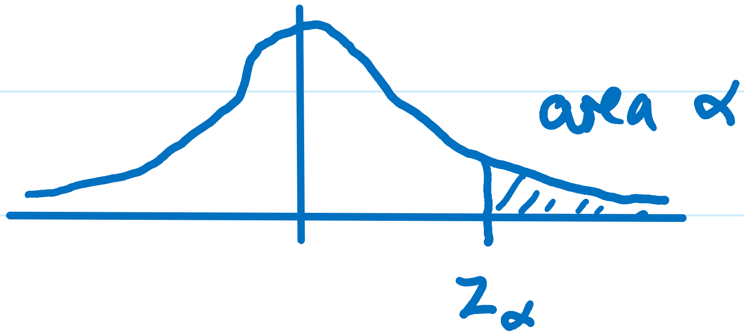
2.1 CIs using CLT

Plenty of examples in Prelims, and similar to the next section.

2.2 CIs using asymptotic distribution of MLE

We know $I(\theta)^{1/2} \cdot (\hat{\theta} - \theta) \xrightarrow{D} N(0,1)$

Let z_α be such that $P(N(0,1) > z_\alpha) = \alpha$



Then $1-\alpha \approx P\left(-z_{\frac{\alpha}{2}} < I(\theta)^{1/2} \cdot (\hat{\theta} - \theta) < z_{\alpha/2}\right)$ (1)

$$= P\left(\hat{\theta} - \frac{z_{\alpha/2}}{\sqrt{I(\theta)}} < \theta < \hat{\theta} + \frac{z_{\alpha/2}}{\sqrt{I(\theta)}}\right)$$

In general $I(\theta)$ depends on θ so (as in Prelims) replace $I(\theta)$ by $I(\hat{\theta})$ to get approx $1-\alpha$ C.I. of

$$\left(\hat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I(\hat{\theta})}} \right). \qquad (2)$$

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

We are assuming $\hat{\theta} \xrightarrow{P} \theta$ and that $I(\theta)$ is continuous, hence $\left(\frac{I(\hat{\theta})}{I(\theta)}\right)^{1/2} \xrightarrow{P} 1$.

So $$I(\hat{\theta})^{1/2} \cdot (\hat{\theta} - \theta) = \underbrace{\left(\frac{I(\hat{\theta})}{I(\theta)}\right)^{1/2}}_{\xrightarrow{P} 1} \cdot \underbrace{I(\theta)^{1/2}(\hat{\theta} - \theta)}_{\xrightarrow{D} N(0,1)}.$$

Hence by Slutsky (ii), $I(\hat{\theta})^{1/2} \cdot (\hat{\theta} - \theta) \xrightarrow{D} N(0,1)$.

So 1 holds with $I(\theta)^{1/2}$ replaced by $I(\hat{\theta})^{1/2}$ and then the same rearrangement as above leads to C.I. 2.

Example $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(θ). Then $\hat{\theta} = \bar{X}$ and

$$I(\theta) = \frac{n}{\theta(1-\theta)}$$ and interval 2 is

$$\left(\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \right).$$

If $n = 30$, $\sum_1^n x_i = 5$, then this gives a 99% interval of: $(-0.008, 0.342)$.

But we know $\theta > 0$!

We can avoid negative values by reparametrising as follows.

Let $\psi = g(\theta) = \log \frac{\theta}{1-\theta}$ "log odds" (of success)

Since $\theta \in (0,1)$ we have $\psi \in (-\infty, \infty)$ so using a normal distribution can't produce impossible ψ values

Note $\hat{\theta} \approx N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$ and delta method gives $\hat{\psi} \approx N\left(\psi, \frac{\theta(1-\theta)}{n} g'(\theta)^2\right)$. (3)

Use ③ to find approx $1-\alpha$ C.I. for ψ, say (ψ_1, ψ_2).

$$1-\alpha \approx P(\psi_1 < \psi < \psi_2)$$

$$= P\left(\frac{e^{\psi_1}}{1+e^{\psi_1}} < \theta < \frac{e^{\psi_2}}{1+e^{\psi_2}}\right) \quad \text{since } \theta = \frac{e^{\psi}}{1+e^{\psi}}.$$

This $1-\alpha$ C.I. for θ definitely won't contain negative values.

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

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

In fact $\chi_r^2 \sim \text{Gamma}\left(\frac{r}{2}, \frac{1}{2}\right)$.

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

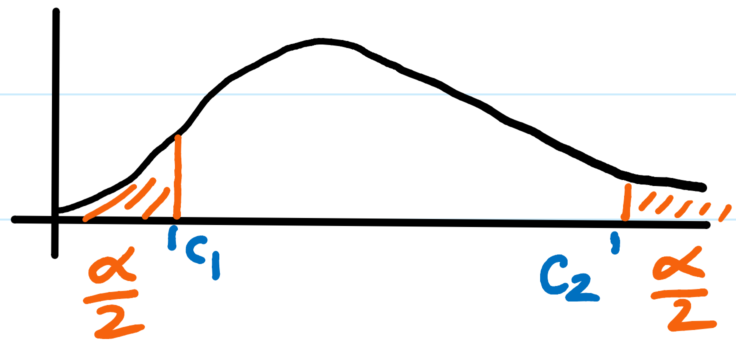
If $Y_1 \sim \chi_r^2$ and $Y_2 \sim \chi_s^2$ are independent, then $Y_1 + Y_2 \sim \chi_{r+s}^2$.

<u>Example</u> $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$.

Then $\frac{X_i}{\sigma} \sim N(0,1)$, hence $\frac{\sum X_i^2}{\sigma^2} \sim \chi^2_n$.

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

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



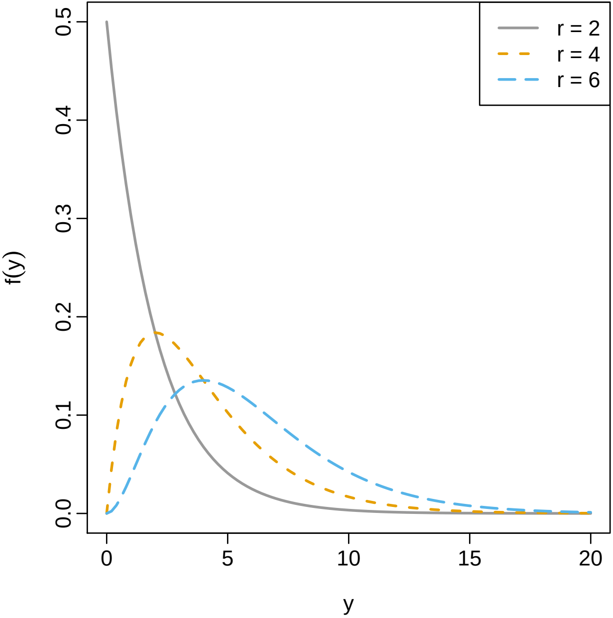
So $P\left(\frac{\sum X_i^2}{c_2} < \sigma^2 < \frac{\sum X_i^2}{c_1}\right) = 1 - \alpha$

and we've found a $1-\alpha$ CI for σ^2.

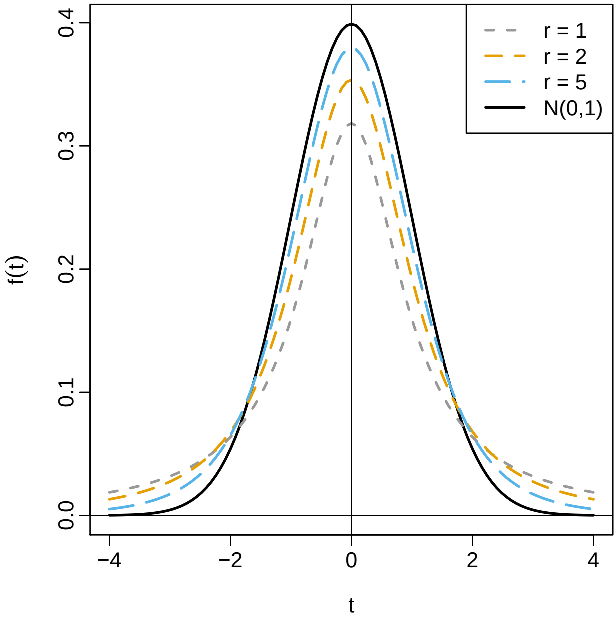
<u>Definition</u> Let $Z \sim N(0,1)$ and $Y \sim \chi^2_r$ be independent. We say that $T = \dfrac{Z}{\sqrt{Y/r}}$ has a (Student) t-distribution with r degrees of freedom. Write $T \sim t_r$.

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

Chi-squared pdfs



t distribution pdfs



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

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

Consider $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ sample mean

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ sample variance

<u>Theorem 2.1</u> $\bar{X}$ and S^2 are independent and their marginal distributions are

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

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

Proof Let $Z_i = \dfrac{X_i - \mu}{\sigma}$. Then $Z_1 \dots Z_n \overset{iid}{\sim} N(0,1)$

and so have joint pdf

$$f(\underset{\sim}{z}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = (2\pi)^{-n/2} e^{-\frac{1}{2}\sum z_i^2} \qquad ①$$

Now consider a transformation from $\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ to $\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$.

Let $\underline{Y} = A \underline{Z}$ where A is an orthogonal $n \times n$ matrix with first row $\left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$.

Orthogonal: $A^T A = I$, so $(\det A)^2 = 1$.

If $\underset{\sim}{y} = A\underset{\sim}{z}$ then $\underline{z} = A^T \underline{y}$ so $z_i = \sum_k a_{ki} y_k$

and $\dfrac{\partial z_i}{\partial y_j} = a_{ji}$.

Hence the Jacobian $J = J(y_1, \ldots, y_n) = \det(A^T)$, so $|J| = 1$ (2)

Also $\sum_1^n y_i^2 = \underset{\sim}{y}^T \underset{\sim}{y} = \underset{\sim}{z}^T A^T A \underline{z} = \underline{z}^T \underline{z} = \sum_1^n z_i^2$ (3)

Hence the pdf of $\underset{\sim}{Y}$ is $g(\underline{y}) = f(\underset{\sim}{z}(\underset{\sim}{y})) \cdot |J|$

$$= (2\pi)^{-n/2} e^{-\frac{1}{2}\sum y_i^2} \cdot 1$$

Hence $Y_1, \ldots Y_n \overset{iid}{\sim} N(0,1)$. using (1), (2), (3).

Now $Y_1 = (\text{first row of } A) \cdot \underset{\sim}{Z} = \frac{1}{\sqrt{n}} \sum_1^n Z_i = \sqrt{n}\, \bar{Z}$

and $$\sum_1^n (Z_i - \bar{Z})^2 = \sum_1^n Z_i^2 - 2\bar{Z}\sum Z_i + n\bar{Z}^2$$

$$= \sum_1^n Z_i^2 - n\bar{Z}^2$$

$$= \sum_1^n Y_i^2 - Y_1^2$$

$$= \sum_2^n Y_i^2$$

So $\bar{Z}$ is a function of Y_1 only

$\sum_1^n (Z_i - \bar{Z})^2$ - - - - - - $Y_2, \dots, Y_n$ only

and the Y_i are indep, hence $\bar{Z}$ and $\sum_1^n (Z_i - \bar{Z})^2$ are indep.

Then $\bar{X}$ and S^2 are indep because $\bar{X} = \sigma \bar{Z} + \mu$

and $S^2 = \dfrac{\sigma^2}{n-1} \sum_1^n (Z_i - \bar{Z})^2$.

Finally:

(i) $Y_1 \sim N(0,1)$ so $\bar{X} = \sigma \bar{Z} + \mu = \frac{\sigma}{\sqrt{n}} Y_1 + \mu \sim N(\mu, \frac{\sigma^2}{n})$

(of course!)

(ii) $\frac{(n-1)S^2}{\sigma^2} = \sum_1^n (Z_i - \bar{Z})^2 = \sum_2^n Y_i^2 \sim \chi^2_{n-1}$. $\square$

So we have $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

and these two random variables are <u>independent</u>

From the definition of a t_{n-1} distribution this gives

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

(the σ in numerator & denominator cancels).

The quantity $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ is called a pivotal quantity or pivot, meaning that it is a function of $\underset{\sim}{X}$ and $\theta = (\mu, \sigma^2)$ whose distribution does not depend on θ.

Similarly $\dfrac{(n-1)S^2}{\sigma^2}$ is another pivot.

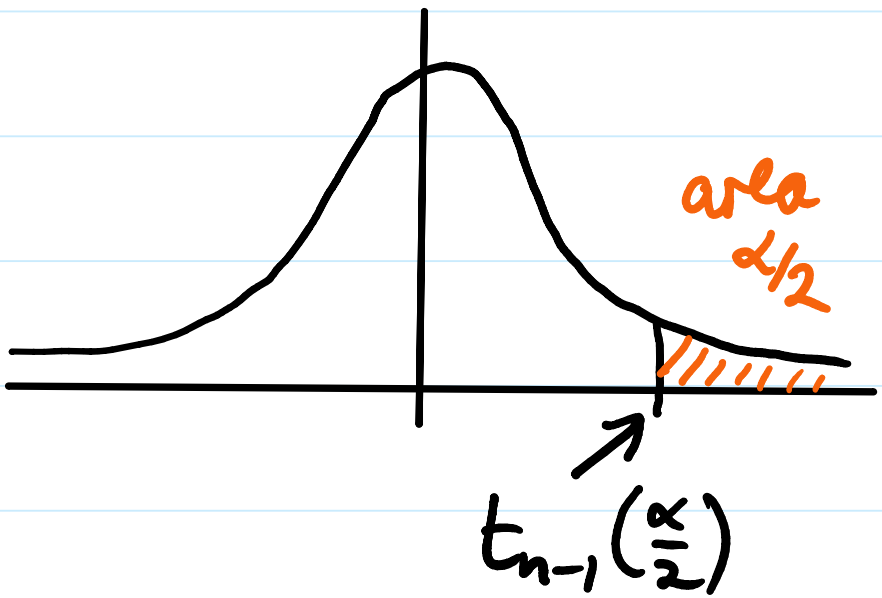
Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

Let's find a C.I. for μ.

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

where $t_{n-1}\left(\frac{\alpha}{2}\right)$ is such that

$$P\left(t_{n-1} > t_{n-1}\left(\tfrac{\alpha}{2}\right)\right) = \frac{\alpha}{2}.$$



Hence

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

and we have a $1-\alpha$ C.I. for μ.

When $\sigma = \sigma_0$ is known, the corresponding interval from Prelims is

$$\left(\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma_0}{\sqrt{n}}\right).$$

Student's Sleep data

"Student" = W.S. Gosset

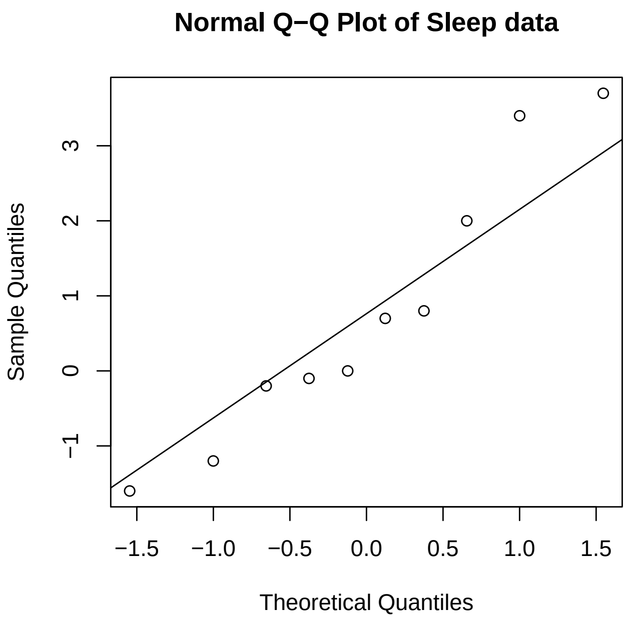
Below is half of Student's sleep data (1908):

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

The data give the number of hours of sleep gained, by 10 patients, following a low dose of a drug.

[The other half of the data give the sleep gained following a normal dose of the drug.]

A point estimate of the sleep gained is $\overline{x} = 0.75$ hours.


Normal Q−Q Plot of Sleep data
Sample Quantiles
−1
0
1
2
3
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

Treating the sample as iid $N(\mu, \sigma^2)$, with μ and σ^2 unknown, a 95% CI for μ is

$$\left(\overline{x} \pm t_{n-1}(\tfrac{\alpha}{2})\frac{s}{\sqrt{n}}\right) = (-0.53, 2.03)$$

using $\overline{x} = 0.75$, $s^2 = 3.2$, $n = 10$, $\alpha = 0.05$, $t_9(0.025) = 2.262$.

The value of $t_9(0.025)$ comes from statistical tables, or from R.

Here, it would be *incorrect* to use a $N(0,1)$ distribution instead of a t_9.

E.g. Suppose we "assume" $\sigma^2 = s^2 = 3.2$ (the sample variance) and calculate the interval

$$\left(\overline{x} \pm 1.96\sqrt{\frac{3.2}{10}}\right) = (-0.36, 1.86).$$

The interval $(-0.53, 2.03)$ obtained using the t_9 distribution is wider than the interval $(-0.36, 1.86)$.

The interval from the t_9 distribution is the correct one here. Since σ^2 is unknown, we need to estimate it (our estimate is s^2). Since we are estimating σ^2, there is more uncertainty than if σ^2 were known, and the t_9 distribution correctly takes this uncertainty into account.

Sleep data (low dose)

Number of hours of sleep gained, by 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.

Sleep data (normal dose)

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?

3. Hypothesis Testing

3.1 Introductory example

We denote the null hypothesis by H_0 and the alternative hypothesis by H_1.

Suppose $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown.

We'll consider

$H_0: \mu = \mu_0$ (and σ^2 is unknown) "drug has no effect"

$H_1: \mu > \mu_0$ (and σ^2 unknown) "drug makes people sleep more" (on average)

$\mu_0 = 0$ for sleep data, but sometimes $\mu_0 \neq 0$.

Let $t_{obs} = t(\underline{x}) = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ and let $t(\underline{X}) = \dfrac{\bar{X} - \mu_0}{S/\sqrt{n}}$.

The idea is:

- a small/moderate value of t_{obs} is consistent with H_0

  including negative

- whereas a very large value of t_{obs} is not consistent with H_0 and instead suggests H_1.

For sleep data (low dose) $t_{obs} = 1.326$.

Is this t_{obs} large?

If H_0 is true then $t(\underset{\sim}{X}) \sim t_{n-1}$ and the probability of observing a value $\geq t_{obs}$ is

$$p = P(t(\underset{\sim}{X}) \geq t_{obs}) = P(t_9 \geq 1.326) = 0.109.$$

This p is called the <u>p-value</u> or <u>significance level</u>.

$p = 0.109$ is not particularly small, we'd observe $t(\underset{\sim}{x}) \geqslant 1.326$ over 10% of the time (if H_0 true).

So we don't have much evidence to reject H_0, so we'll retain H_0 and say the data are consistent with H_0 being true.

What we have done here: look to see if data seem inconsistent with H_0

- if so, reject H_0
- if not, keep / retain H_0.

Usually we say "don't reject H_0" or "reject H_0"
(retain H_0)

or say data are "consistent with H_0" or "not consistent with H_0"

(rather than "accept H_0" or "accept H_1")

2nd example (sleep data, full dose)

$t_{obs} = 3.68$

This time p-value $= P(t(\underline{X}) \geq 3.68) = P(t_9 \geq 3.68)$

$= 0.0025.$

This is very small, we'd see $t(\underline{X}) \geq 3.68$ only 0.25% of the time if H_0 true, very rare.

We conclude that there is very strong evidence to reject H_0 in favour of 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 - - - - - - -
$0.05 < p < 0.1$	weak - - - - - - - - - -
$0.1 < p$	little or no - - - - - - - - - .

$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ μ, σ both unknown

null $H_0: \mu = \mu_0$, alternative $H_1: \mu > \mu_0$

$$T = T(\underline{X}) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

when H_0 true

$\underbrace{\text{The further } \bar{x} \text{ above } \mu_0,}_{\text{the further } t(\underline{x}) \text{ above zero}}$ the more evidence to reject H_0.

One-sided and two-sided alternative hypotheses

$H_1: \mu > \mu_0$ is a one-sided alternative. The larger t_{obs} the more evidence to reject H_0.

Similarly $H_1: \mu < \mu_0$ is also one-sided. The p-value would be $p = P(t_{n-1} \leq t_{obs})$.

A different type of alternative is $H_1: \mu \neq \mu_0$. This is a two-sided alternative.

$H_1: \mu \neq \mu_0$ two-sided

For this H_1:

if t_{obs} is very lage (i.e. very positive) then we have evidence to reject H_0

AND also

if t_{obs} is very small (i.e. very negative) then we have evidence to reject H_0

Let $t_0 = |t_{obs}| = \left| \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \right|$

The p-value is the probability $t(\underset{\sim}{X})$ takes a value "at least as extreme" as t_{obs}:

$$p = P(|t(\underset{\sim}{X})| \geq t_0)$$

$$= P(t(\underset{\sim}{X}) \geq t_0) + P(t(\underset{\sim}{X}) \leq -t_0)$$

$$= 2P(t(\underset{\sim}{X}) \geq t_0).$$

This p-value, and all others, are calculated under the assumption that H₀ is true. From now on we write $p = P(t(\underset{\sim}{X}) \geq t_{obs} \mid H_0)$ or $p = P(|t(\underset{\sim}{X})| \geq t_0 \mid H_0)$ to indicate this.

3.2 Tests for normally distributed samples

Example (z-test) $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ unknown and $\sigma^2 = \sigma_0^2$ known.

To test $H_0: \mu = \mu_0$ against $H_1: \mu > \mu_0$ we use the test statistic $Z = \dfrac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$.

Let $z_{obs} = \dfrac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}}$.

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

p-value $p = P(Z \geq z_{obs} \mid H_0) = P(N(0,1) \geq z_{obs})$

$= 1 - \Phi(z_{obs})$

For H_0 versus $H_1': \mu < \mu_0$

p-value $p' = P(Z \leq z_{obs} \mid H_0) = \Phi(z_{obs})$

For H_0 versus $H_1'': \mu \neq \mu_0$

p-value $p'' = P(|Z| \geq z_0 \mid H_0) = 2(1 - \Phi(z_0))$

where $z_0 = |z_{obs}|$.

Example (t-test) $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with μ and σ^2 unknown.

There are similar expressions for p-values like p, p', p'' above, after replacing:

- Z by $$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$
- Φ by cdf of t_{n-1}.

t-test (one sample)

[Example from Dalgaard (2008).] Data on the daily energy intake (in kJ) of 11 women:

$$5260,\ 5470,\ 5640,\ 6180,\ 6390,\ 6515,$$
$$6805,\ 7515,\ 7515,\ 8230,\ 8770.$$

Do these values deviate from a recommended value of 7725 kJ?

We consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where $\mu_0 = 7725$, and we make the standard assumptions for a t-test.

We have $t_{\text{obs}} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = -2.821$.

The p-value is $p = 2P(t_{10} \geqslant |t_{\text{obs}}|) = 0.018$. So we conclude that there is good evidence to reject the null hypothesis that the mean intake is 7725 kJ.

Testing $H_0 : \mu = 7725$ against $H_1^- : \mu < 7725$,

the p-value is $p^- = P(t_{10} \leqslant t_{\text{obs}}) = 0.009$.

Conclusion: there is good evidence to reject H_0 in favour of H_1^-.

Testing $H_0 : \mu = 7725$ against $H_1^+ : \mu > 7725$,

the p-value is $p^+ = P(t_{10} \geqslant t_{\text{obs}}) = 0.991$.

Conclusion: there is no evidence to reject H_0 in favour of H_1^+.

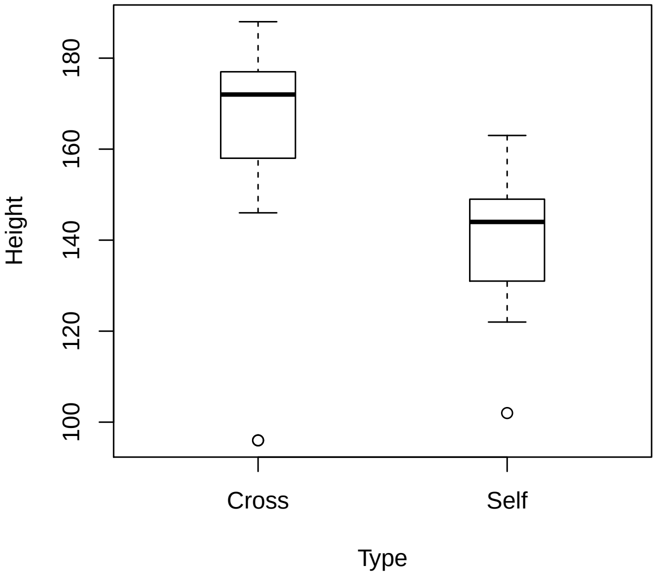
t-test (two sample)

Darwin's *Zea mays* data – heights of young maize plants.

Height (eights of an inch)			
Crossed		Self-fertilized	
188	146	139	132
96	173	163	144
168	186	160	130
176	168	160	144
153	177	147	102
172	184	149	124
177	96	149	144
163		122	

Are the heights of the two types of plant the same?

[In fact, the plants were in pairs – one cross- and one self-fertilized in each pair – we ignore this pairing for now. We'll see how to deal with pairing later.]



180
160
140
120
100
Height
Cross
Self
Type

Assume we have two independent samples $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu_X, \sigma^2)$, and $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu_Y, \sigma^2)$, where σ^2 is unknown.

Suppose we would like to test $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X \neq \mu_Y$.

Let

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}$$

where $S^2 = \frac{1}{m+n-2}[\sum(X_i - \overline{X})^2 + \sum(Y_i - \overline{Y})^2]$.

Assuming H_0 is true, we have $T \sim t_{m+n-2}$.

For the maize data, the observed value of T is

$$t = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}} = 2.437.$$

The alternative hypothesis $(\mu_X \neq \mu_Y)$ is two-sided, so the p-value of this test is

$$p = 2P(t_{28} \geqslant 2.437) = 0.021.$$

Conclusion: there is good evidence to reject the null hypothesis $\mu_X = \mu_Y$.

t-test (paired)

Suppose we have pairs of RVs (X_i, Y_i), $i = 1 \ldots, n$. Let $D_i = X_i - Y_i$.

Suppose $D_1, \ldots, D_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, with σ^2 unknown, and that we want to test a hypothesis about μ. We can use the test statistic

$$\frac{\overline{D} - \mu_0}{S_D/\sqrt{n}}$$

which has a t_{n-1} distribution under $H_0 : \mu = \mu_0$. (Here, S_D^2 is the sample variance of the D_i.)

The kind of situation where a paired test is used is when there are two measurements on the same "experimental unit", e.g. in the sleep data, low and normal doses were given to the same 10 patients.

Two sample t and paired t

Is the amount of sleep gained with a low dose the same as the amount gained with a high dose?

low (X)	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
normal (Y)	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
difference (D)	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

- Two sample t-test of $H_0 : \mu_X = \mu_Y$ against $H_1 : \mu_X \neq \mu_Y$: the p-value is 0.079.
- Paired t-test (of $\mu_0 = 0$), based on the differences D_i: the p-value is 0.0028.

The paired test uses the information that the observations are paired: i.e. we have one low and one high dose observation per patient. The two sample test ignores this information. Prefer the paired test here.

Could consider one-sided alternatives here.

Hypothesis testing and confidence intervals

For the maize data:

- the 95% (equal tail) confidence interval for $\mu_X - \mu_Y$ is $(3.34, 38.53)$ (see Sheet 2, Question 5)
- when testing $\mu_x = \mu_Y$ against $\mu_x \neq \mu_Y$, the p-value is 0.021.

So, observe that

(i) the p-value less than 0.05

(ii) the 95% confidence interval does not contain 0 (= the value of $\mu_X - \mu_Y$ under H_0).

(i) and (ii) both being true is not a coincidence – there is a connection between hypothesis tests and confidence intervals.

3.3 Hypothesis testing and confidence intervals

Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

(i) A $1-\alpha$ C.I. for μ is

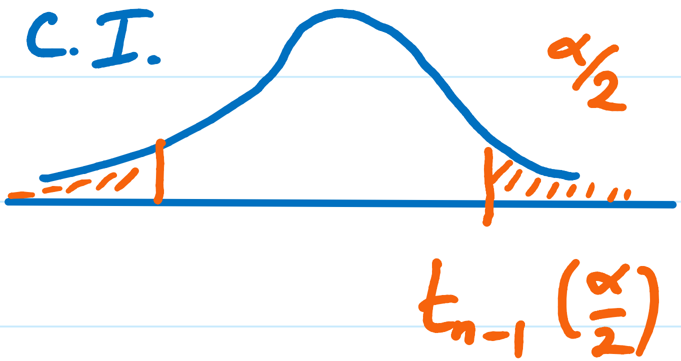
$$\left(\bar{x} \pm t_{n-1}\left(\tfrac{\alpha}{2}\right) \cdot \frac{s}{\sqrt{n}}\right) \qquad (1)$$

(ii) For t-test of $\mu = \mu_0$ against $\mu \neq \mu_0$,

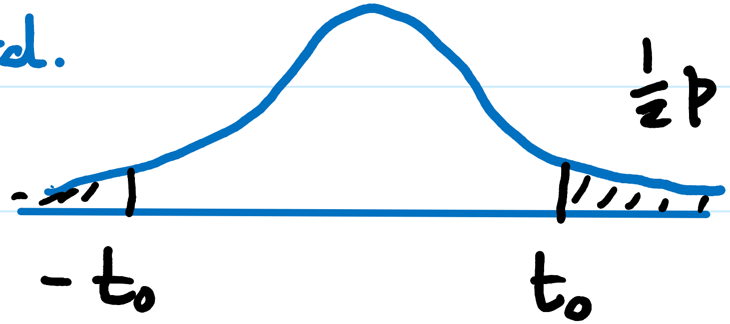
p-value is $p = P(|t_{n-1}| \geq t_0)$

where $t_0 = |t(\underline{x})| = \left|\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}\right|$.

For C.I.



p-val.



So $p < \alpha \iff t_0 > t_{n-1}\left(\frac{\alpha}{2}\right)$

$$\iff t(\underline{x}) > t_{n-1}\left(\tfrac{\alpha}{2}\right) \quad \text{or} \quad t(\underline{x}) < -t_{n-1}\left(\tfrac{\alpha}{2}\right)$$

$$\iff \mu_0 < \bar{x} - t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{s}{\sqrt{n}} \quad \text{or}$$

$$\mu_0 > \bar{x} + t_{n-1}\left(\tfrac{\alpha}{2}\right)\frac{s}{\sqrt{n}}$$

That is: $p < \alpha \iff$ C.I. ① does not contain μ_0.

3.4 Hypothesis testing general setup

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

Consider 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** $t(\underset{\sim}{X})$ such that large values of $t(\underset{\sim}{X})$ indicate a departure from H_0 in the direction of H_1.

Let $t_{obs} = t(\underset{\sim}{x})$, the value of $t(\underset{\sim}{X})$ observed.

Then the **p-value** or **significance level** is

$$p = P\left(t(\underset{\sim}{X}) \geq t_{obs} \mid H_0\right).$$

A small p is an indicator that H_0 and the data 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 $t(\underline{X})$ 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 $\theta \neq \theta_0$.

Example $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

$H_0 : \mu = \mu_0$ is composite because it corresponds to $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \ \sigma^2 > 0\}$ ← this set contains more than one point

Here σ^2 is called a nuisance parameter.

Suppose we want to make a definite decision:

either reject H_0

or don't reject H_0.

Then we can define a test in terms of a critical region $C \subset \mathbb{R}^n$:

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

Errors in hypothesis testing

There are two possible types of error:

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

✓ = correct decision

Consider simple $H_0: \theta = \theta_0$ versus simple $H_1: \theta = \theta_1$.

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

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$
$$= P(\underset{\sim}{X} \in C \mid \theta_0)$$

The **type II error probability** β is defined by

$$\beta = P(\text{don't reject } H_0 \mid H_1 \text{ true})$$
$$= P(\underset{\sim}{X} \notin C \mid \theta_1)$$

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

Note: power $= 1-\beta = P(\underset{\sim}{X} \in C \mid \theta_1)$

= probability of correctly detecting that H_0 is false.

If H_0 is composite, $H_0 : \theta \in \Theta_0$ say, then the size is defined by

$$\alpha = \sup_{\theta \in \Theta_0} P(\underset{\sim}{X} \in C \mid \theta)$$

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

$$w(\theta) = P(\text{reject } H_0 \mid \theta \text{ is the true value})$$

$$= P(\underset{\sim}{X} \in C \mid \theta)$$

Ideally we'd like

$w(\theta)$ to be near 1 for H_1-values of θ

- - - - - - - - - 0 for H_0-values of θ.

3.5 The Neyman–Pearson Lemma

Consider testing simple $H_0: \theta = \theta_0$ against simple $H_1: \theta = \theta_1$. (*)

Suppose we choose a small type I error probability α (e.g. $\alpha = 0.05$). Then, among all tests of this size we could aim to:

$$\begin{cases} \text{minimise the type II error probability } \beta \\ \text{i.e. maximise the power } 1-\beta \end{cases}$$

This approach treats H_0 and H_1 asymmetrically.

Theorem 3.1 (N-P Lemma) Let $L(\theta; \underset{\sim}{x})$ be the likelihood. Define the critical region C by

$$C = \left\{ \underset{\sim}{x} : \frac{L(\theta_0; \underset{\sim}{x})}{L(\theta_1; \underset{\sim}{x})} \leq k \right\}$$

and suppose constants k and α are such that

$P(\underset{\sim}{X} \in C \mid H_0) = \alpha$. ← "C has size α"

Then among all tests of $(*)$ of size $\leq \alpha$, the test with critical region C has maximum power.

<u>Proof</u> (for cts random variables – for discrete replace $\int$ by Σ)

Consider any test of size $\leq \alpha$, with critical region A say.

Then $P(\underline{X} \in A \mid H_0) \leq \alpha$ ①.

(C is one possibility for A).

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

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

Then $$0 \leq \left\{ \phi_C(\underline{x}) - \phi_A(\underline{x}) \right\} . \left[L(\theta_1; \underline{x}) - \frac{1}{k} L(\theta_0; \underline{x}) \right]$$

since $\{\ldots\}$ and $[\ldots]$ are both ≥ 0 if $\underline{x} \in C$
and both ≤ 0 if $\underline{x} \notin C$

So $$0 \le \int_{\mathbb{R}^n} \{\phi_C(\underline{x}) - \phi_A(\underline{x})\}\left[L(\theta_1;\underline{x}) - \frac{1}{k}L(\theta_0;\underline{x})\right] d\underline{x}$$

$$= P(\underline{X}\in C \mid H_1) - P(\underline{X}\in A \mid H_1) - \frac{1}{k}\Big[\underbrace{P(\underline{X}\in C \mid H_0)}_{\alpha} - \underbrace{P(\underline{X}\in A \mid H_0)}_{\le \alpha \text{ by } ①}\Big]$$

$$\le P(\underline{X}\in C \mid H_1) - P(\underline{X}\in A \mid H_1).$$

That is, $$\underbrace{P(\underline{X}\in C \mid H_1)}_{\text{power of region } C} \ge \underbrace{P(\underline{X}\in A \mid H_1)}_{\text{power of region } A} \qquad ②$$

And ② says the power is maximised by using region C. □

__Example__ $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$, σ_0^2 known.

Find most powerful test of $H_0: \mu = 0$ against $H_1: \mu = \mu_1$, where $\mu_1 > 0$.

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

__Step 1__ H_0, H_1 both simple, so N-P applies and most powerful test is of the form

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

k_1 a constant, i.e. doesn't depend on $\underline{x}$.

$$\Longleftrightarrow \exp\left[-\frac{1}{2\sigma_0^2}\sum x_i^2\right] \exp\left[\frac{1}{2\sigma_0^2}\sum (x_i-\mu_1)^2\right] \leq k_1$$

$$\Longleftrightarrow \exp\left[\frac{1}{2\sigma_0^2}\left(-\sum x_i^2 + \sum x_i^2 - 2\mu_1 \sum x_i + n\mu_1^2\right)\right] \leq k_1$$

$$\Longleftrightarrow \frac{1}{2\sigma_0^2}\left(-2\mu_1 n\bar{x} + n\mu_1^2\right) \leq k_2 \qquad (k_2 = \log k_1)$$

$$\Longleftrightarrow -\mu_1 \bar{x} \leq k_3$$

$$\Longleftrightarrow \bar{x} \geq c$$

where k_1, k_2, k_3, c are constants that don't depend on x (they can depend on $n, \sigma_0^2, \ldots$).

Step 2 Choose c so that the test has size α.

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

$$= P(\bar{X} \geq c \mid H_0) \qquad \text{and under } H_0,\ \bar{X} \sim N\left(0, \frac{\sigma_0^2}{n}\right) \quad ③$$

$$= P\left(\frac{\bar{X}}{\sigma_0/\sqrt{n}} \geq \frac{c}{\sigma_0/\sqrt{n}} \,\middle|\, H_0\right)$$

$$= P\left(N(0,1) \geq \frac{c}{\sigma_0/\sqrt{n}}\right) \qquad \text{by } ③$$

Hence $\dfrac{c}{\sigma_0/\sqrt{n}} = z_\alpha$. So most powerful critical region is $C = \left\{ \underset{\sim}{x} : \bar{x} \geq z_\alpha \dfrac{\sigma_0}{\sqrt{n}} \right\}$.

Let's also calculate the power function of this test.

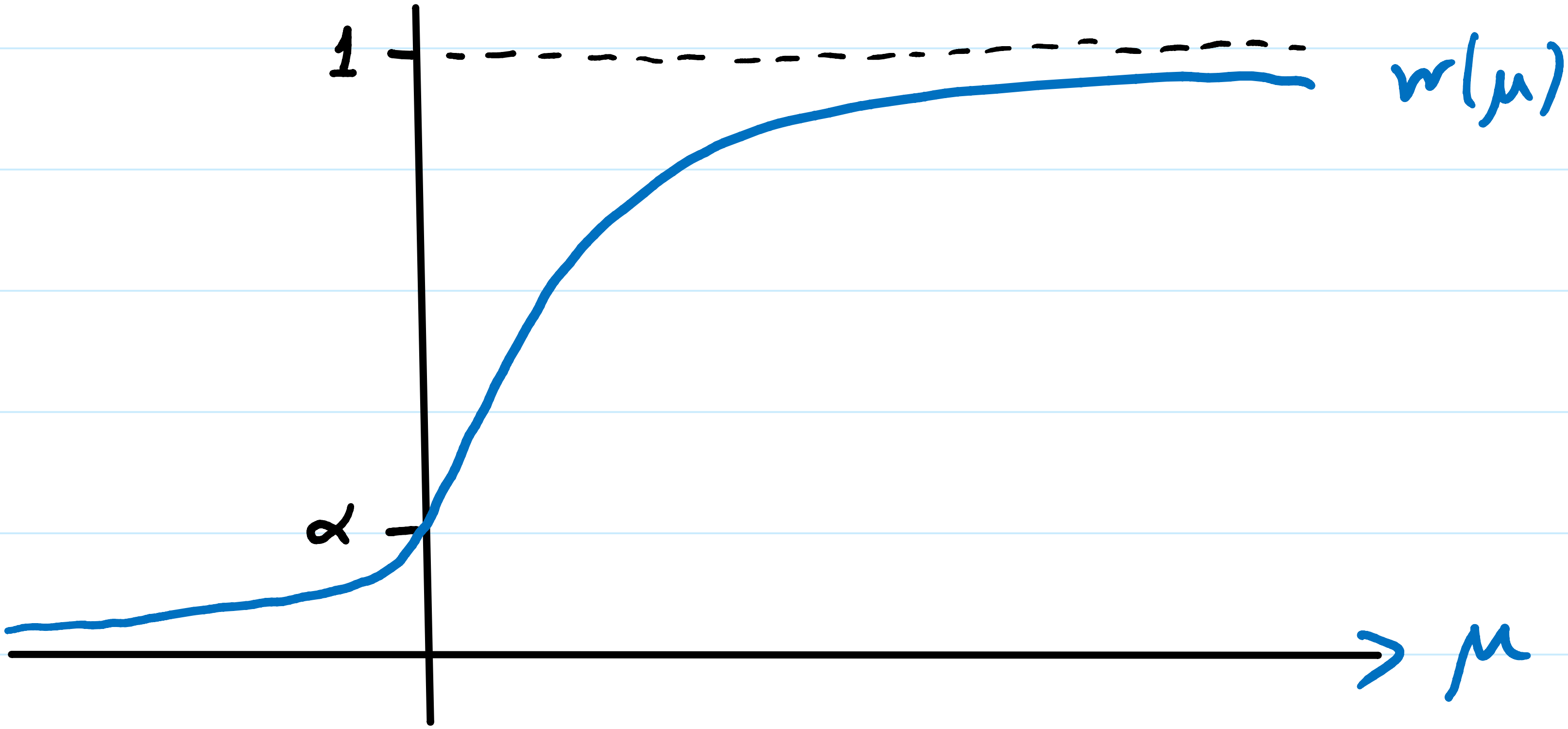
$$w(\mu) = P(\text{reject } H_0 \mid \mu \text{ is the true value})$$

$$= P\left(\bar{X} \geq z_\alpha \frac{\sigma_0}{\sqrt{n}} \,\middle|\, \mu\right) \qquad \text{if } \mu \text{ is true value, } \bar{X} \sim N\left(\mu, \frac{\sigma_0^2}{n}\right) \text{ (4)}$$

$$= P\left(\frac{\bar{X}-\mu}{\sigma_0/\sqrt{n}} \geq z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}} \,\middle|\, \mu\right)$$

$$= P\left(N(0,1) \geq z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}}\right) \qquad \text{by (4)}$$

$$= 1 - \Phi\left(z_\alpha - \frac{\mu}{\sigma_0/\sqrt{n}}\right)$$


1
α
π(μ)
μ

__Last example__: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma_0^2)$, σ_0^2 known.

We were testing $H_0: \mu = 0$ against $H_1: \mu = \mu_1$,

where μ_1 was a single value satisfying $\mu_1 > 0$.

Critical region was $\bar{x} \geqslant c$, or $\sum x_i \geqslant k$ (where $k = nc$)

Equation linking k and α was $\alpha = P(\sum X_i \geqslant k \mid H_0)$.

$\sum X_i$ was normal, so any value of α possible by choosing k appropriately.

If e.g. the $X_i \sim$ Poisson, then not all values of α possible as $P(\sum X_i \geqslant k \mid H_0)$ will decrease in jumps as k increases.

3.6 Uniformly most powerful tests

Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_1$.

When testing simple $\theta = \theta_0$ against simple $\theta = \theta_1 \in \Theta_1$, the critical region from N-P lemma may be the same for each $\theta_1 \in \Theta_1$. Then C is said to be uniformly most powerful (UMP) for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$.

Previous example: $N(\mu, \sigma_0^2)$, σ_0^2 known.

The critical region C we found for $\mu = 0$ versus $\mu = \mu_1$ was the same for all $\mu_1 > 0$.

Hence our C is UMP for testing $\mu = 0$ against $\mu > 0$.

$$C = \left\{ \underset{\sim}{x} : \ \bar{x} \geqslant z_\alpha \frac{\sigma_0}{\sqrt{n}} \right\}$$

Insect traps

33 insect traps were set out across sand dunes and the numbers of insects caught in a fixed time were counted (Gilchrist, 1984). The number of traps containing various numbers of the taxa *Staphylinoidea* were as follows.

Count	0	1	2	3	4	5	6	$\geqslant 7$
Frequency	10	9	5	5	1	2	1	0

Suppose $X_1, \ldots, X_{33} \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$.

Consider testing $H_0 : \lambda = 1$ against $H_1 : \lambda = \lambda_1$, where $\lambda_1 > 1$.

The NP lemma leads to a test of the form

$$\text{reject } H_0 \iff \sum x_i \geqslant c.$$

If the test has size α, then $\alpha = P(\sum X_i \geqslant c \mid H_0)$.

Under H_0, we have $\sum X_i \sim \text{Poisson}(33)$ exactly. However, instead of using this we can use a normal approximation:

$$\alpha = P\left(\frac{\sum X_i - 33}{\sqrt{33}} \geqslant \frac{c - 33}{\sqrt{33}} \,\middle|\, H_0 \right)$$

and, by the CLT, if H_0 is true then $\frac{\sum X_i - 33}{\sqrt{33}} \overset{\text{D}}{\approx} N(0,1)$, so

$$\alpha \approx 1 - \Phi\left(\frac{c - 33}{\sqrt{33}} \right).$$

Hence $\frac{c-33}{\sqrt{33}} \approx z_\alpha$, so $c \approx 33 + z_\alpha \sqrt{33}$.

So we have a critical region

$$C = \{\mathrm{x} : \sum x_i \geqslant 33 + z_\alpha \sqrt{33}\}.$$

Note that C does not depend on which value of $\lambda_1 > 1$ we are considering, so we actually have a UMP test of $\lambda = 1$ against $\lambda > 1$.

If $\alpha = 0.01$ then $c \approx 47$; if $\alpha = 0.001$ then $c \approx 51$.

The observed value of $\sum x_i$ is 54.

So in both cases the observed value of 54 is $\geqslant c$, so in both cases we'd reject H_0.

An alternative way of thinking about this is to calculate the p-value:

$$\begin{aligned} p &= P(\text{we observe a value at least as extreme as } 54 \mid H_0) \\ &= P\Big(\sum X_i \geqslant 54 \mid H_0\Big) \\ &\approx 0.0005 \end{aligned}$$

which is very strong evidence for rejecting H_0.

Note that a test of size α rejects H_0 if and only if $\alpha \geqslant p$. That is, the p-value is the smallest value of α for which H_0 would be rejected. (This is true generally, not just in this particular example.)

In practice, no-one tells us a value of α, we have to judge the situation for ourselves. Our conclusion here is that there is very strong evidence for rejecting H_0.

3.6 Likelihood ratio tests

Now consider testing $H_0 : \theta \in \Theta_0$ against the <u>general alternative</u> $H_1 : \theta \in \Theta$ (where $\Theta_0 \subset \Theta$).

So now H_0 is a special case of H_1.

H_0 is "<u>nested within</u>" H_1.

We test to see if simplifying to the H_0-model is reasonable.

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

$$\lambda(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta; \underline{x})}{\sup_{\theta \in \Theta} L(\theta; \underline{x})} = \frac{\text{TOP}}{\text{BOTTOM}} \quad (1)$$

A (generalised) likelihood ratio test (LRT) has critical region of the form

$$C = \{ \underline{x} : \lambda(\underline{x}) \leqslant k \}.$$

Sometimes we can calculate the distribution of a function of $\lambda(\underline{X})$,

more often we will approximate the distribution of a function of $\lambda(\underline{X})$.

Example $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, μ, σ^2 unknown.

Let $H_0: \mu = \mu_0$ (and any $\sigma^2 > 0$)

$H_1: \mu \in (-\infty, \infty)$ (and any $\sigma^2 > 0$).

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

For TOP of ①: max L over σ^2 with $\mu = \mu_0$ fixed.

Max is at $\sigma^2 = \hat{\sigma}_0^2 = \frac{1}{n} \sum (x_i - \mu_0)^2$.

For BOTTOM of ①: max L over μ and σ^2.

Max is at $\mu = \hat{\mu} = \bar{x}$, $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$.

Substitute into ① to get

$$\lambda(\underline{x}) = \frac{L(\mu_0, \hat{\sigma}_0^2)}{L(\hat{\mu}, \hat{\sigma}^2)} \leftarrow (2\pi\hat{\sigma}^2)^{-n/2} \exp\left[-\frac{1}{2\hat{\sigma}^2}\sum(x_i-\hat{\mu})^2\right]$$

$$-n/2$$

$$= \frac{\left[\frac{2\pi}{n}\sum(x_i-\mu_0)^2\right]^{-n/2} e^{-n/2}}{\left[\frac{2\pi}{n}\sum(x_i-\bar{x})^2\right]^{-n/2} e^{-n/2}}$$

Now note $\sum (x_i - \mu_0)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$.

Substitute into $\lambda(x)$ to find

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

So LRT is reject $H_0 \iff \lambda(\underline{x}) \leq k$

$$\iff \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq k_1 .$$

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

Likelihood ratio statistic

$\Lambda(\underset{\sim}{X}) = -2\log\lambda(\underset{\sim}{X})$ is called the **likelihood ratio statistic**.

The critical region $\{\underset{\sim}{x} : \lambda(\underset{\sim}{x}) \leq k\}$ becomes $\{\underset{\sim}{x} : \Lambda(\underset{\sim}{x}) \geq c\}$.

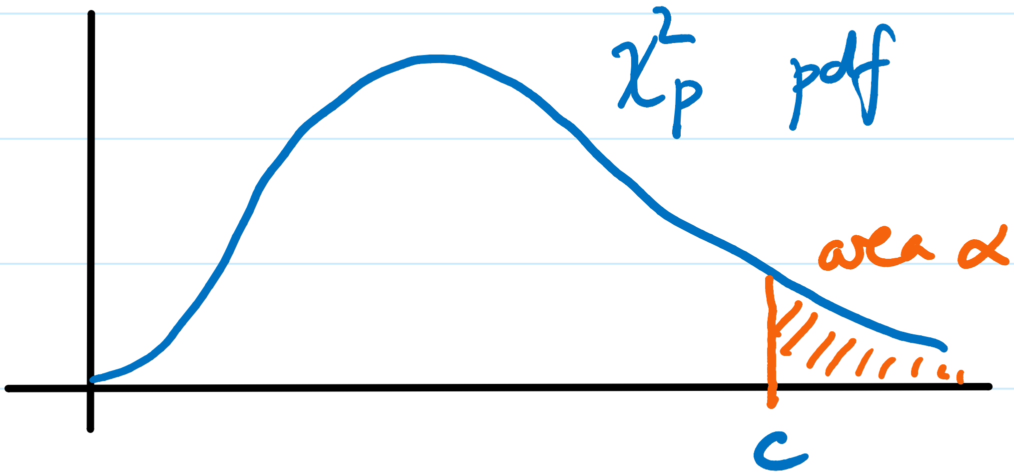
If H_0 is true then, under regularity conditions, as $n \to \infty$, we have $\Lambda(\underset{\sim}{X}) \xrightarrow{D} \chi^2_p$ ②

where $p = \dim H_1 - \dim H_0$.

$\dim H_1$ = # independent parameters in Θ

$\dim H_0$ = - - - - - - - - - - Θ_0.

Since $\Lambda(\underline{X}) \approx \chi^2_p$ for large n, under H_0, we get an approx test of size α by choosing c such that $P(\chi^2_p > c) = \alpha$.



Why is ② true?

Sketch proof for scalar θ, so $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta$ with $\dim \Theta = 1$.

So here $p = \dim \Theta - \dim \Theta_0 = 1 - 0 = 1$.

Taylor expansion: $\ell(\theta_0) \approx \ell(\hat{\theta}) + (\hat{\theta} - \theta_0)\ell'(\hat{\theta})$

$$+ \tfrac{1}{2}(\hat{\theta} - \theta_0)^2 \ell''(\hat{\theta})$$

$$= \ell(\hat{\theta}) - \tfrac{1}{2}(\hat{\theta} - \theta_0)^2 J(\hat{\theta}) \quad ③$$

assuming $\ell'(\hat{\theta}) = 0$.

So $\Lambda(\underset{\sim}{X}) = -2\log\left(\frac{L(\theta_0)}{L(\hat{\theta})}\right)$

$$= 2\left[l(\hat{\theta}) - l(\theta_0)\right]$$

$$\approx \underbrace{(\hat{\theta}-\theta_0)^2 I(\theta_0)}_{\approx [N(0,1)]^2} \cdot \underbrace{\frac{J(\hat{\theta})}{I(\theta_0)}}_{\approx 1}$$ using 3

under H_0, for large n

$$\approx \chi^2_1.$$

We now write the LR statistic as

$$\Lambda = -2 \log \lambda = -2 \log \left(\frac{\sup_{H_0} L}{\sup_{H_1} L} \right) \quad (1)$$

Goodness of fit tests

Hardy–Weinberg equilibrium

In a sample from the Chinese population of Hong Kong, blood types occurred with the following frequencies (Rice, 1995):

	Blood type			
	M	MN	N	Total
Frequency	342	500	187	1029

If gene frequencies are in Hardy–Weinberg equilibrium, then the probability of an individual having blood type M, MN, or N should be

$$\begin{aligned} P(M) &= (1-\theta)^2 \\ P(MN) &= 2\theta(1-\theta) \\ P(N) &= \theta^2. \end{aligned}$$

Consider n independent observations, each in one of categories $1, \dots, k$.

Let n_i = # observations in category i (frequency),

so $\sum_{i=1}^{k} n_i = n$

π_i = probability of an observation being in category i, so $\sum_{i=1}^{k} \pi_i = 1$.

Let $\pi = (\pi_1, \dots, \pi_k)$

Likelihood $L(\pi) = \frac{n!}{n_1! \ldots n_k!} \pi_1^{n_1} \cdots \pi_k^{n_k}$ multinomial distribution

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

Consider $H_0: \pi_i = \pi_i(\theta)$ for $i = 1, \ldots, k$, where $\theta \in \Theta$

(e.g. $\pi_1 = (1-\theta)^2$, $\pi_2 = 2\theta(1-\theta)$, $\pi_3 = \theta^2$, $\theta \in (0,1)$)

versus $H_1: \pi_i$ unrestricted except for $\sum \pi_i = 1$.

Then $\dim H_1 = k-1$,

and suppose $\dim H_0 = q < k-1$.

$$\Lambda = -2\log\left(\frac{\sup_{H_0} L}{\sup_{H_1} L}\right) \qquad (1)$$

The degrees of freedom for Λ are:

$$p = \dim H_1 - \dim H_0 = (k-1) - q.$$

(i) For TOP in (1): maximise over θ to get MLE $\theta = \hat{\theta}$

(ii) For BOTTOM in (1): maximise $f(\pi) = \sum n_i \log \pi_i$

subject to the constraint $g(\pi) = \sum \pi_i - 1 = 0.$

With Lagrange multiplier λ, we need

$$\frac{\partial f}{\partial \pi_i} = \lambda \frac{\partial g}{\partial \pi_i} \qquad i = 1 \ldots k$$

i.e. $$\frac{n_i}{\pi_i} = \lambda . 1$$

So $\pi_i = \dfrac{n_i}{\lambda}$ and then $1 = \sum \pi_i = \dfrac{\sum n_i}{\lambda} = \dfrac{n}{\lambda}$

and so $\lambda = n$.

So the MLEs under H_1 are $\hat{\pi}_i = \dfrac{n_i}{n}$.

So $\Lambda = -2\log\left(\dfrac{L(\pi(\hat\theta))}{L(\hat\pi)}\right)$

$$= 2\left[l(\hat\pi) - l(\pi(\hat\theta))\right]$$

$$= 2\left[\sum n_i \log\hat\pi_i - \sum n_i \log \pi_i(\hat\theta)\right]$$

$$= 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\,\pi_i(\hat\theta)}\right) \qquad \text{since } \hat\pi_i = \frac{n_i}{n}.$$

Compare this Λ to a χ^2_p where $p = k-1-q$ to carry out the test.

Pearson's chi-squared statistic

$$\Lambda = 2 \sum_{i=1}^{k} o_i \log\left(\frac{o_i}{e_i}\right)$$

where $o_i = n_i$ observed

$e_i = n.\pi_i(\hat{\theta})$ expected under H_0

Using $x \log \frac{x}{a} \approx x - a + \frac{(x-a)^2}{2a}$ gives

$$\Lambda \approx 2 \sum \left[o_i - e_i + \frac{(o_i - e_i)^2}{2e_i} \right]$$

$$= \sum \frac{(o_i - e_i)^2}{e_i} = P$$ Pearson's χ^2 statistic

Hardy–Weinberg equilibrium

In a sample from the Chinese population of Hong Kong, blood types occurred with the following frequencies (Rice, 1995):

	Blood type			
	M	MN	N	Total
Frequency	342	500	187	1029

If gene frequencies are in Hardy–Weinberg equilibrium, then the probability of an individual having blood type M, MN, or N should be

$$\begin{aligned} P(M) &= (1-\theta)^2 \\ P(MN) &= 2\theta(1-\theta) \\ P(N) &= \theta^2. \end{aligned}$$

The observed frequencies are $(n_1, n_2, n_3) = (342, 500, 187)$, with total $n = n_1 + n_2 + n_3 = 1029$.

The likelihood is

$$L(\theta) \propto [(1-\theta)^2]^{n_1} \times [\theta(1-\theta)]^{n_2} \times [\theta^2]^{n_3}$$

so the log-likelihood is

$$\ell(\theta) = (2n_1 + n_2)\log(1-\theta) + (n_2 + 2n_3)\log\theta + \text{constant}$$

from which we obtain

$$\widehat{\theta} = \frac{n_2 + 2n_3}{2n} = 0.425.$$

So $\pi_1(\widehat{\theta}) = (1-\widehat{\theta})^2$, $\pi_2(\widehat{\theta}) = 2\widehat{\theta}(1-\widehat{\theta})$, $\pi_3(\widehat{\theta}) = \widehat{\theta}^2$ and

$$\Lambda = 2\sum_i n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 0.032.$$

We compare Λ to a χ_p^2 where $p = \dim\Theta - \dim\Theta_0 = (3-1)-1 = 1$.

The value $\Lambda = 0.032$ is much less than $E(\chi_1^2) = 1$. The p-value is $P(\chi_1^2 \geqslant 0.032) = 0.86$, so there is no reason to doubt the Hardy–Weinberg model.

Pearson's chi-squared statistic leads to the same conclusion

$$P = \sum \frac{[n_i - n\pi_i(\widehat{\theta})]^2}{n\pi_i(\widehat{\theta})} = 0.0319.$$

Insect counts (Bliss and Fisher, 1953)

[Example from Rice (1995).] From each of 6 apple trees in an orchard that had been sprayed, 25 leaves were selected. On each of the leaves, the number of adult female red mites was counted.

Number per leaf	0	1	2	3	4	5	6	7	8+
Observed frequency	70	38	17	10	9	3	2	1	0

Does a Poisson(θ) model fit these data?

As usual for a Poisson, $\widehat{\theta} = \overline{x} = 1.147$, and

$$\pi_i(\widehat{\theta}) = \widehat{\theta}^i e^{-\widehat{\theta}}/i!, \quad i = 0, 1, \ldots, 7$$

$$\pi_8(\widehat{\theta}) = 1 - \sum_{i=0}^{7} \pi_i(\widehat{\theta}).$$

The expected frequency in cell i is $n\pi_i(\widehat{\theta})$.

Some expected frequencies are very small:

# per leaf	0	1	2	3	4	5	6	7	8+
Observed	70	38	17	10	9	3	2	1	0
Expected	47.7	54.6	31.3	12.0	3.4	0.8	0.2	0.02	0.004

The χ^2 approximation for the distribution of Λ applies when there are large counts.

The usual rule-of-thumb is that the χ^2 approximation is good when the expected frequency in each cell is at least 5.

To ensure this, we should pool some cells before calculating Λ or P.

After pooling cells $\geqslant 3$:

# per leaf	0	1	2	$\geqslant 3$
Observed	70	38	17	25
Expected	47.7	54.6	31.3	16.4

Then $\Lambda = 2\sum O_i \log\left(\frac{O_i}{E_i}\right) = 26.60$, and $P = \sum(O_i - E_i)^2/E_i = 26.65$.

These are to be compared with a χ^2 with $(4-1)-1 = 2$ degrees of freedom.

The p-value is $p = P(\chi^2_2 \geqslant 26.6) \approx 10^{-6}$, so there is clear evidence that a Poisson model is not suitable.

Two-way contingency tables

Hair and Eye Colour

The hair and eye colour of 592 statistics students at the University of Delaware were recorded (Snee, 1974) – dataset `HairEyeColor` in R.

	Eye colour			
Hair colour	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

Are hair colour and eye colour independent?

Cross-classify n individuals according to two sets of categories

		(eye colour)					
		1	2	$\cdots$	c	row sum
(hair colour)	1	n_{11}	$\cdots$	$\cdots$	n_{1c}	n_{1+}
	$\vdots$	$\vdots$			$\vdots$	$\vdots$
	r	n_{r1}	$\cdots$	$\cdots$	n_{rc}	n_{r+}
col sum		n_{+1}	$\cdots$	$\cdots$	n_{+c}	$\boxed{n_{++} = n}$

Let n_{ij} = frequency of (i,j) $\qquad \sum_i \sum_j n_{ij} = n$

π_{ij} = probability an individual falls into cell (i,j) $\qquad \sum_i \sum_j \pi_{ij} = 1$

Likelihood $L(\pi) = n! \prod_{i=1}^{r} \prod_{j=1}^{c} \dfrac{\pi_{ij}^{n_{ij}}}{n_{ij}!}$

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

Consider:

H_0: the two classifications are independent

(e.g. hair colour and eye colour are independent)

i.e. $\pi_{ij} = \alpha_i \beta_j$ where $\sum_i \alpha_i = 1$ and $\sum_j \beta_j = 1$

H_1: π_{ij} unrestricted except for $\sum_i \sum_j \pi_{ij} = 1$.

(i) Max under H_0 (Sheet 3): $\hat{\alpha}_i = \frac{n_{i+}}{n}$, $\hat{\beta}_j = \frac{n_{+j}}{n}$

(ii) Max under H_1 (done already): $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$.

We find $$\Lambda = 2 \sum_{i,j} n_{ij} \log\left(\frac{n_{ij} \cdot n}{n_{i+} \cdot n_{+j}}\right)$$

$$\approx \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where $o_{ij} = n_{ij}$ observed

$e_{ij} = n \hat{\alpha}_i \hat{\beta}_j$ expected # in (i,j) under H_0

Degrees of freedom of this Λ

$\dim H_1 = rc - 1$ probabilities $\pi_{11}, \dots, \pi_{rc}$

with $\sum_{i,j} \pi_{ij} = 1$.

$\dim H_0 = (r-1) + (c-1)$

$r-1$ for $\alpha_1 \dots \alpha_r$ with $\sum \alpha_i = 1$

$c-1$ for $\beta_1 \dots \beta_c$ with $\sum \beta_j = 1$

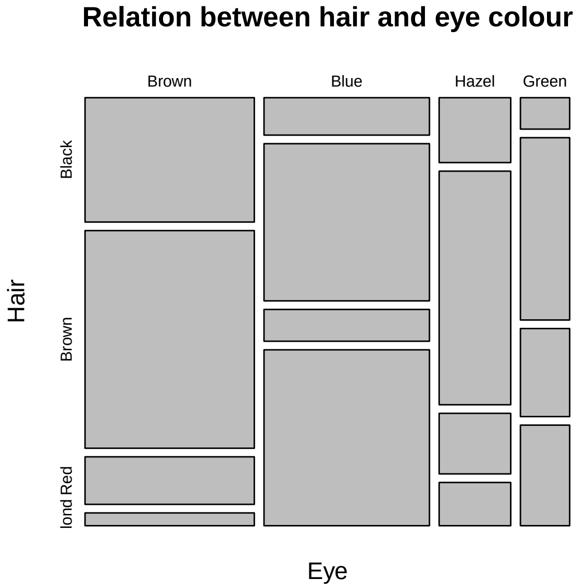
So $p = \dim H_1 - \dim H_0 = (r-1)(c-1)$

Hair and Eye Colour

The hair and eye colour of 592 statistics students at the University of Delaware were recorded (Snee, 1974) – dataset `HairEyeColor` in R.

	Eye colour			
Hair colour	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

Are hair colour and eye colour independent?


Relation between hair and eye colour
Brown
Blue
Hazel
Green
Black
Brown
lond Red
Hair
Eye

$$\Lambda = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\log\left(\frac{n_{ij}n}{n_{i+}n_{+j}}\right) = 146.4$$

$$\dim H_1 = 16 - 1 = 15$$

$$\dim H_0 = (4-1) + (4-1) = 6$$

Hence we compare Λ to a χ^2_p where $p = 15 - 6 = 9$.

The p-value is $P(\chi^2_9 \geqslant 146.4) \approx 0$.

So there is overwhelming evidence of an association between hair colour and eye colour (i.e. overwhelming evidence that they are not independent).

[Pearson's chi-squared statistic is $P = 138.3$.]

4. Bayesian Inference

Bayesian Inference

So far we have followed the frequentist approach:

- 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*.

In subjective Bayesian inference, probability is a measure of the strength of belief.

Before any data are available, there is uncertainty about the parameter θ. Suppose uncertainty about θ is expressed as a "prior" pdf (of pmf) for θ.

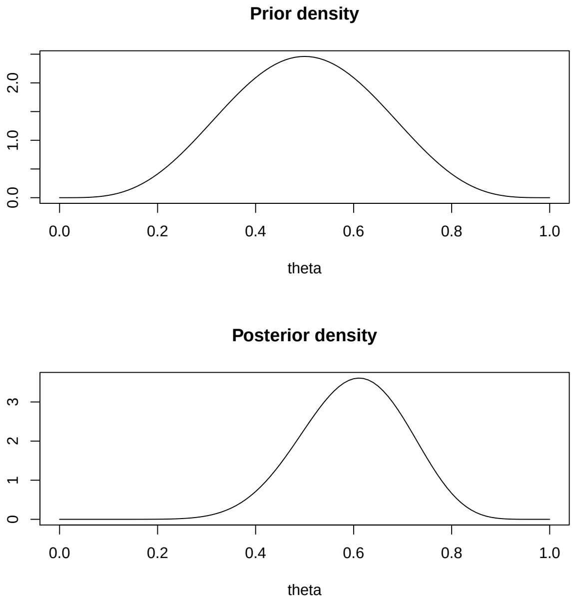
Then, once data are available, we can use Bayes' theorem to combine our prior beliefs with the data to obtain an updated "posterior" assessment of our beliefs about θ.

Example

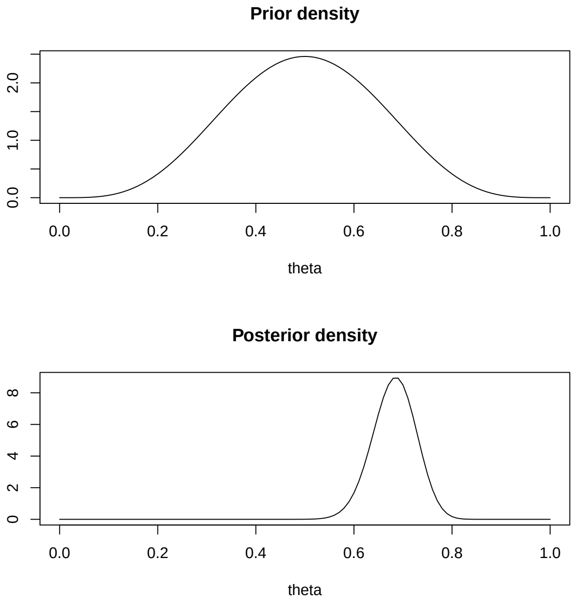
Suppose we have a coin which we think might be a bit biased.

Let θ be the probability of getting a head when we flip it.

Prior: Beta(5, 5). Data: 7 heads from 10 flips.



Prior: Beta(5, 5). Data: 70 heads from 100 flips.



4.1 Introduction

Suppose that, as usual, we have a probability model $f(\underset{\sim}{x} \mid \theta)$ for data $\underset{\sim}{x}$. $\Leftarrow$ likelihood

In this section we write $f(\underset{\sim}{x} \mid \theta)$ (rather than $f(\underset{\sim}{x}; \theta)$) to indicate that $\underline{x}$ is <u>conditional on θ</u>, we have a <u>conditional distribution/density</u>.

Suppose also, before observing $\underset{\sim}{x}$, we summarise our beliefs about θ in a <u>prior density</u> $\pi(\theta)$.

That is, we treat θ as a random variable.

Once we have observed $\underset{\sim}{x}$, our updated beliefs about θ are contained in the conditional density of θ given $\underset{\sim}{x}$, which is called the <u>posterior density</u> $\pi(\theta \mid \underset{\sim}{x})$.

Theorem (Bayes' theorem – continuous version)

For continuous random variables Y and Z, the conditional density $f(z|y)$ of Z given Y satisfies

$$f(z|y) = \frac{f(y|z)\,f(z)}{f(y)} \qquad (*).$$

Proof By definition of conditional density,

$$f(z|y) = \frac{f(y,z)}{f(y)} \quad ① \qquad \text{and} \qquad f(y|z) = \frac{f(y,z)}{f(z)} \quad ②.$$

From ② $f(y,z) = f(y|z)f(z)$ and substituting into ① gives $(*)$. $\square$

Note: marginal pdf of Y is

$$f(y) = \int_{-\infty}^{\infty} f(y,z)\,dz = \int_{-\infty}^{\infty} f(y|z)f(z)\,dz \qquad (**).$$

(similar expression for $f(z)$).

With $\underline{x}$ and θ in place of y and z we have

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{f(\underline{x})} \qquad \leftarrow \text{like } (*)$$

where $f(\underline{x}) = \int_{\text{all }\theta} f(\underline{x}|\theta)\pi(\theta)\,d\theta \qquad \leftarrow \text{like} (**).$

As usual for conditional densities, we treat $\pi(\theta|\underline{x})$ as a function of θ, with data $\underline{x}$ fixed.

Since $\underline{x}$ is fixed, $f(\underline{x})$ is just a constant, and so

$$\boxed{\begin{array}{c} \pi(\theta|\underline{x}) \propto f(\underline{x}|\theta) \times \pi(\theta) \\ \\ \text{posterior} \propto \text{likelihood} \times \text{prior} \end{array}}$$

Example Conditionally on θ, suppose $X_1 \ldots X_n \overset{iid}{\sim}$ Bernoulli(θ).

$$P(X_i = 1 \mid \theta) = \theta, \qquad P(X_i = 0 \mid \theta) = 1 - \theta$$

i.e. $f(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \qquad x_i = 0, 1.$

So likelihood $f(\underline{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$

$$= \theta^r (1-\theta)^{n-r} \quad \text{where } r = \sum_1^n x_i$$

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

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

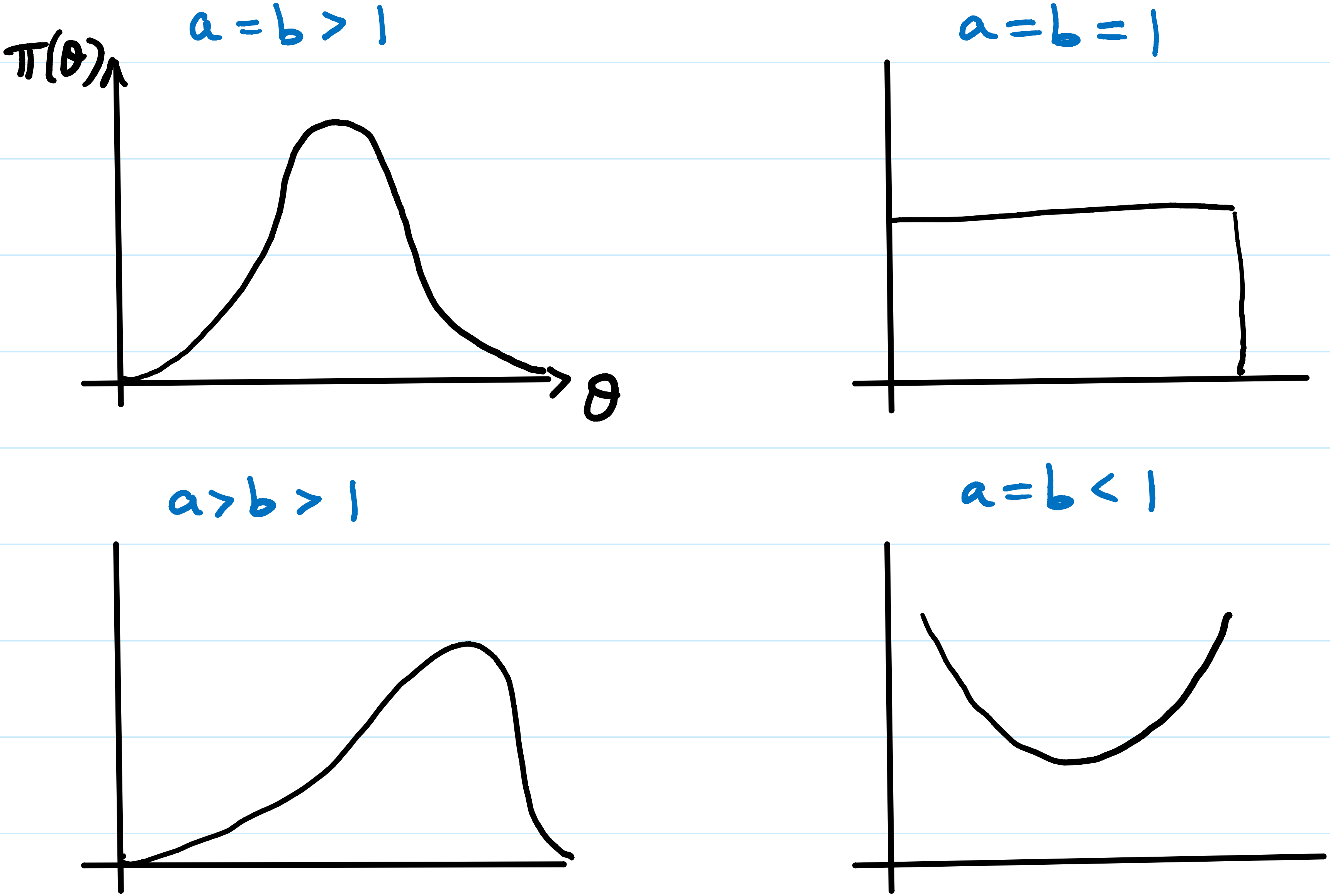
Here $B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}\,d\theta$ beta function

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

and $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u}\,du$

$\Gamma(a+1) = a\,\Gamma(a)$ for $a > 0$

$\Gamma(n) = (n-1)!$ for n positive integer.



a = b > 1
π(θ)
θ
a = b = 1
a > b > 1
a = b < 1

We are assuming a, b known, and $a > 0$, $b > 0$.

chosen to reflect our prior beliefs

Now posterior $\propto$ likelihood $\times$ prior, so

$$\pi(\theta \mid \underset{\sim}{x}) \propto \theta^{r}(1-\theta)^{n-r} \times \theta^{a-1}(1-\theta)^{b-1}$$

$$= \theta^{r+a-1}(1-\theta)^{n-r+b-1} \qquad (3)$$

The RHS of (3) depends on θ exactly as for a Beta$(r+a, n-r+b)$ density.

Hence the constant of proportionality in ③ must be

$$\frac{1}{B(r+a,\, n-r+b)},$$

and the posterior distribution is a Beta $(r+a,\, n-r+b)$.

So pdf $\pi(\theta\,|\,\underset{\sim}{x}) = \dfrac{1}{B(r+a,\, n-r+b)}\,\theta^{r+a-1}(1-\theta)^{n-r+b-1}$,

$0 < \theta < 1.$

Note: no need to do any integration.

Example Conditional on θ, suppose $X_1 \ldots X_n \overset{iid}{\sim}$ Poisson (θ).

Suppose prior for θ is a Gamma(α, β) pdf:

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0$$

where $\alpha > 0$, $\beta > 0$ known

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

$$\pi(\theta | \underline{x}) \propto \left(\prod_{i=1}^{n} \frac{e^{-\theta} \theta^{x_i}}{x_i!} \right) \times \theta^{\alpha-1} e^{-\beta\theta}$$

$$\propto \theta^{r+\alpha-1} e^{-(n+\beta)\theta} \qquad \text{where } r = \sum x_i .$$

So the posterior distribution is a Gamma,

$\pi(\theta \mid \underline{x})$ is a $\text{Gamma}(r+\alpha,\, n+\beta)$ pdf

[because $\pi(\theta \mid \underline{x})$ depends on θ as for a

$\text{Gamma}(r+\alpha,\, n+\beta)$].

Example (MRSA)

[Example from `www.scholarpedia.org`.]

Let θ denote the number of MRSA infections per 10,000 bed-days in a hospital.

Suppose we observe $y = 20$ infections in 40,000 bed-days, i.e. in $10{,}000N$ bed-days where $N = 4$.

- A simple estimate of θ is $y/N = 5$ infections per 10,000 bed-days.
- The MLE of θ is also $\widehat{\theta} = 5$ if we assume that y is an observation from a Poisson distribution with mean θN, so

$$f(y \mid \theta) = (\theta N)^y e^{-\theta N} / y!\,.$$

However, other evidence about θ may exist.

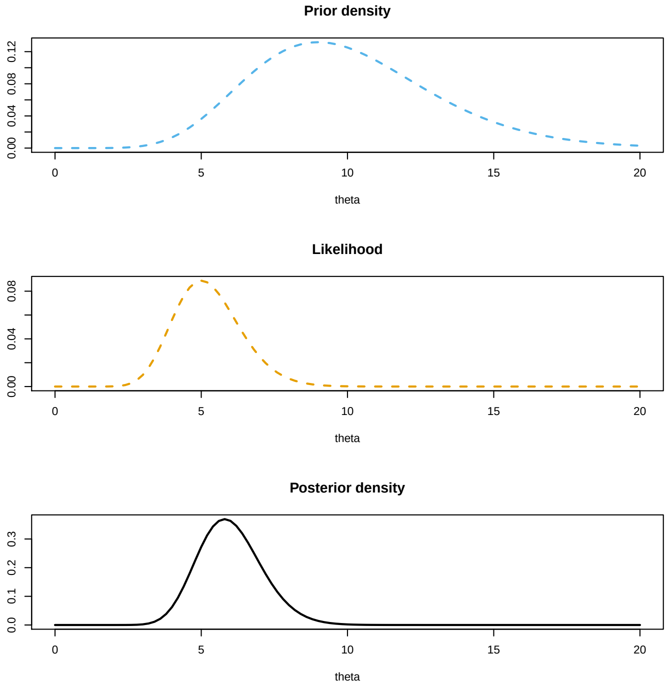
Suppose this other information, on its own, suggests plausible values of θ of about 10 per 10,000, with 95% of the support for θ lying between 5 and 17.

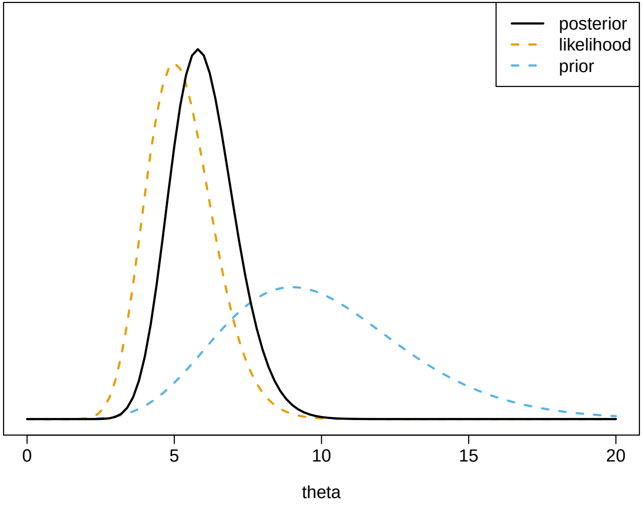
We can use a prior distribution to describe this. A Gamma pdf is convenient here:

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

Taking $\alpha = 10$, $\beta = 1$ gives approximately the properties above.

- The posterior combines the evidence from the data (i.e. the likelihood) and the other (i.e. prior) evidence. We can think of the posterior as a compromise between the likelihood and the prior.
- Calculated on board in lectures: the posterior is another Gamma.


Prior density
0.00
0.04
0.08
0.12
0
5
10
15
20
theta
Likelihood
0.00
0.04
0.08
0
5
10
15
20
theta
Posterior density
0.0
0.1
0.2
0.3
0
5
10
15
20
theta



posterior
likelihood
prior
0
5
10
15
20
theta

4.2 Inference

All information about θ is contained in the posterior density $\pi(\theta|\underline{x})$.

Posterior summaries

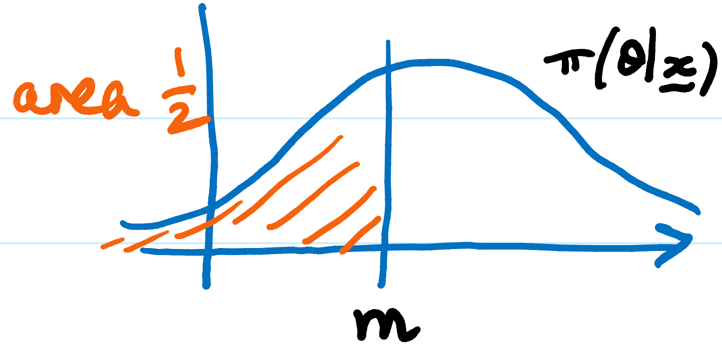
Sometimes summaries of $\pi(\theta|\underline{x})$ are useful, e.g.

(i) the posterior mode (value of θ at which $\pi(\theta|\underline{x})$ is max)

(ii) the posterior mean $E(\theta|\underline{x})$

expectation over θ
($\underline{x}$ is fixed)

(iii) posterior median, m such that $\int_{-\infty}^{m} \pi(\theta | \underline{x})\, d\theta = \frac{1}{2}$



(iv) $\text{var}(\theta | \underline{x})$

(v) other quantiles of $\pi(\theta | \underline{x})$.

Example Conditional on θ, suppose $X \sim \text{Binomial}(n, \theta)$.

We write this as: $X \mid \theta \sim \text{Binomial}(n, \theta)$.

Prior $\theta \sim U(0,1)$.

posterior $\propto$ likelihood $\times$ prior

$$\pi(\theta \mid \underset{\sim}{x}) \propto \binom{n}{x} \theta^{x} (1-\theta)^{n-x} \times 1$$

$$\propto \theta^{x} (1-\theta)^{n-x}$$

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

Posterior mean

$$E(\theta|\underline{x}) = \int_0^1 \theta \, \pi(\theta|\underline{x}) \, d\theta$$

$$= \frac{1}{B(x+1, n-x+1)} \int_0^1 \theta^{x+1} (1-\theta)^{n-x} \, d\theta$$

$$= \frac{1}{B(x+1, n-x+1)} \cdot B(x+2, n-x+1)$$

$$= \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)}$$

$$= \frac{x+1}{n+2}$$ using $\Gamma(a+1) = a\Gamma(a)$ twice

So even when all trials are successes ($x=n$), this point estimate is $\frac{n+1}{n+2} < 1$ (seems sensible especially if n small).

Posterior mode is $\frac{x}{n}$ (same as MLE).

For large n, i.e. when the likelihood contribution dominates that from the prior, posterior mean and mode will be close.

Interval estimation

Frequentist $\longrightarrow$ confidence interval

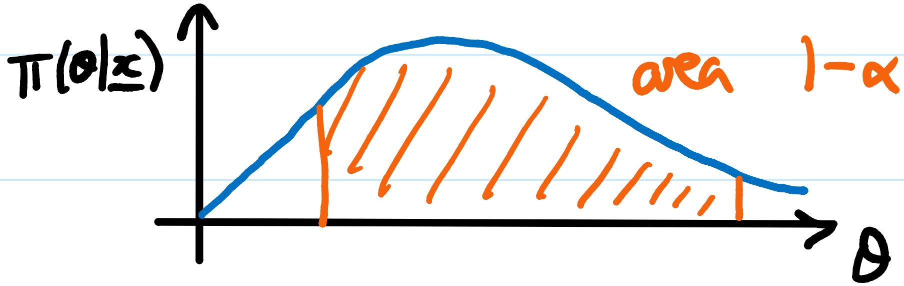
Bayesian $\longrightarrow$ credible interval

Let Θ be the parameter space.

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

$$\int_C \pi(\theta \mid \underline{x})\, d\theta = 1-\alpha.$$

Note this is just saying $P(\theta \in C \mid \underset{\sim}{x}) = 1-\alpha$

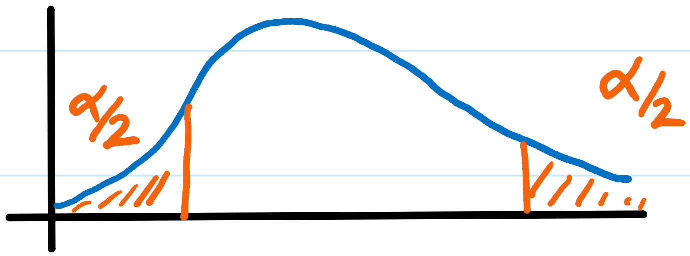


A credible interval is when set C is an interval,

$C = (\theta_1, \theta_2)$ say.

The interval (θ_1, θ_2) is called equal-tailed if

$P(\theta \le \theta_1 \mid \underset{\sim}{x}) = P(\theta \ge \theta_2 \mid \underset{\sim}{x})$



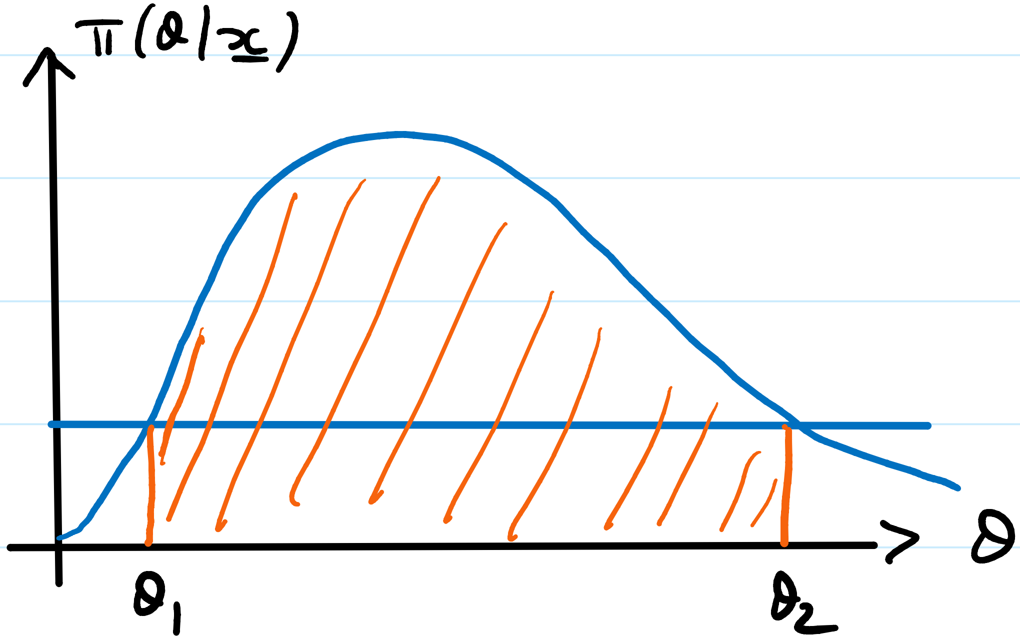
In words: "the probability that θ lies in C, given the observed data $\underline{x}$, is $1-\alpha$"

Very simple!

This is <u>not</u> true of a confidence interval.

Definition We call C a highest posterior density (HPD) credible set if $\pi(\theta|\underset{\sim}{x}) \geq \pi(\theta'|\underset{\sim}{x})$ for all $\theta \in C$ and all $\theta' \notin C$.

E.g. (θ_1, θ_2) here:



An HPD interval has minimal width among all $1-\alpha$ credible intervals.

Multi-parameter models

θ may be a vector. If so, everything above still applies, all integrals over θ mean multiple integrals over all components of θ.

e.g. $\theta = (\psi, \lambda)$, so posterior $\pi(\psi, \lambda \mid \underline{x})$.

All info about ψ is contained in the marginal posterior for ψ, which is $\pi(\psi \mid \underline{x}) = \int \pi(\psi, \lambda \mid \underline{x})\, d\lambda$

integrate over all λ to find marginal distribution

Prediction

Let X_{n+1} represent a future observation.

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 $\underline{x}$, called the posterior predictive density, is a conditional density, found by the usual rules of probability:

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

integrate over all θ to find marginal density

$\underline{x} = (x_1, \ldots, x_n)$ here

$$= \int \underbrace{f(x_{n+1} \mid \theta, \underline{x})}_{f(x_{n+1} \mid \theta) \text{ by the independence above}} \pi(\theta \mid \underline{x})\, d\theta$$

$$f(u, v \mid w) = f(u \mid v, w) f(v \mid w)$$

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

4.3 Prior information

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

(i) If substantial prior knowledge exists, we could ask a subject-area expert.

(ii) If we have little prior knowledge we might want a prior that expresses "prior ignorance"

is this possible? maybe $\theta \sim U(0,1)$ for a prior probability value

(iii) We might want to choose a "conjugate" prior for ease of calculation (by hand)

	prior		lik		posterior
e.g.	Beta	+	Bernoulli	$\rightarrow$	Beta
	Gamma	+	Poisson	$\rightarrow$	Gamma
	$\vdots$				

Note (iii) can overlap with (i) and (ii).

<u>Example</u> Conditional on θ, let $X_1 \dots X_n$ be independent $N(\theta, \sigma^2)$ where σ^2 known.

Let prior be $\theta \sim N(\mu_0, \sigma_0^2)$ where μ_0, σ_0^2 known.

Then $\pi(\theta | \underline{x}) \propto f(\underline{x} | \theta) \, \pi(\theta)$

$$\propto \exp\left[-\frac{1}{2} \sum \frac{(x_i - \theta)^2}{\sigma^2} \right] \exp\left[-\frac{1}{2} \frac{(\theta - \mu_0)^2}{\sigma_0^2} \right]$$

Now complete the square:

$$\frac{(\theta-\mu_0)^2}{\sigma_0^2} + \sum \frac{(x_i-\theta)^2}{\sigma^2} = \theta^2\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right) - 2\theta\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\bar{x}}{\sigma^2}\right)$$

$$+ \text{constant}$$

$$= \frac{1}{\sigma_1^2}(\theta-\mu_1)^2 + \text{constant}$$

after completing the square

where

$$\mu_1 = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}} \qquad ①$$

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \qquad ②$$

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

$\nwarrow$ a $N(\mu_1, \sigma_1^2)$ pdf

So $\theta | \underline{x} \sim N(\mu_1, \sigma_1^2)$.

weight $\frac{1}{\sigma_0^2}$

① says: posterior mean μ_1 = weighted av. of prior mean μ_0
and sample mean $\bar{x}$

weight $\frac{n}{\sigma^2}$

The precision of a random variable is $\frac{1}{\text{variance}}$.

② says: posterior precision = prior precision + data precision.

Improper priors

If $\sigma_0^2 \to \infty$ above then $\pi(\theta | \underline{x})$ is approx $N\left(\bar{x}, \frac{\sigma^2}{n}\right)$.
i.e. the likelihood contribution dominates the prior contribution as $\sigma_0^2 \to \infty$.

This corresponds to prior $\pi(\theta) \propto c$, a constant, i.e. a "uniform prior".

But this π is not a probability distribution since $\theta \in (-\infty, \infty)$ and we can't have $\int_{-\infty}^{\infty} c \, d\theta = 1$.

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 (e.g. uniform prior $\pi(\theta) \propto c$ for $\theta \in \mathbb{R}$ above) and we can use the posterior for inference.

But we can't use an improper posterior for meaningful inference.

Prior ignorance

If no reliable prior information is available we might want a prior which has minimal effect on our inference.

E.g. if $\Theta = \{\theta_1, \dots, \theta_m\}$ then $\pi(\theta_i) = \frac{1}{m}$, $i = 1 \dots m$ does not favour any value of θ, is "non-informative".

But things are not so simple when θ is continuous.

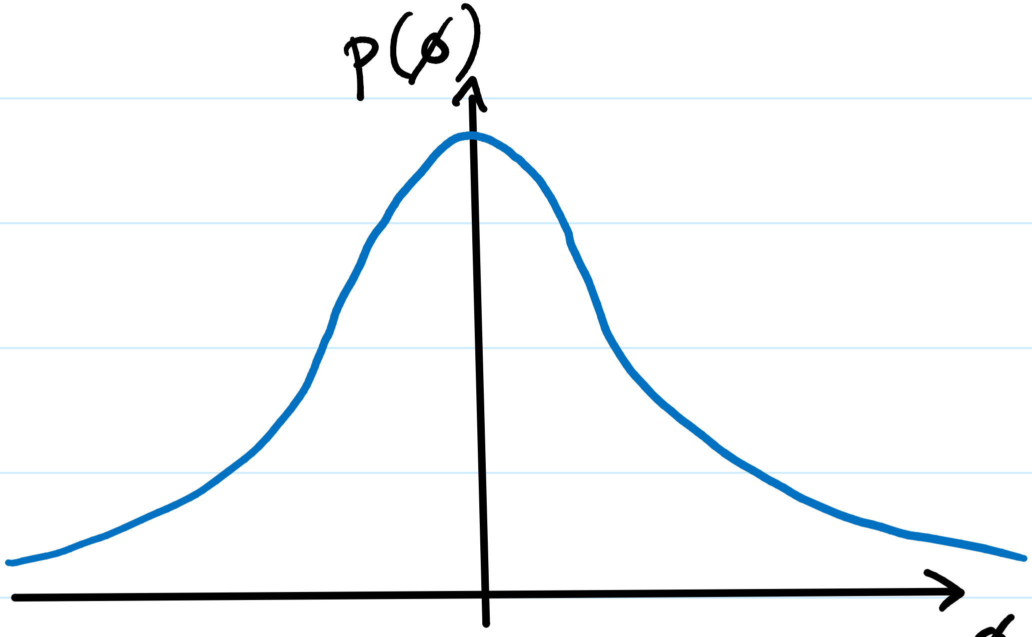
Example If $\Theta = (0,1)$ we might think $\theta \sim U(0,1)$ represents ignorance

However, if we are ignorant about θ then we are also ignorant about $\phi = \underbrace{\log\left(\frac{\theta}{1-\theta}\right)}_{\text{log odds}}$

θ has pdf $\pi(\theta) = 1, \quad 0 < \theta < 1.$

So ϕ has pdf $p(\phi) = \pi(\theta(\phi)) \dfrac{d\theta}{d\phi}$ $\qquad \theta = \dfrac{e^\phi}{1+e^\phi}$

$$= 1 \times \frac{e^\phi}{(1+e^\phi)^2}, \quad \phi \in \mathbb{R}.$$



this does not seem consistent with ignorance about ϕ.

Jeffreys priors

The problem with the ϕ-example above is that the representation of "ignorance" changes if we change parametrisation from θ to ϕ.

Suppose θ is a scalar.

A solution to the issue is the **Jeffreys prior** defined by $\pi(\theta) \propto I(\theta)^{1/2}$

square root of expected information

If $X_1 \ldots X_n$ are from $f(x|\theta)$, this is $\pi(\theta) \propto i(\theta)^{1/2}$.

In what sense is Jeffreys prior a "solution"?

Suppose $\phi = h(\theta)$.

Consider:

(i) Find $\pi(\theta)$ using Jeffreys rule, then transform this pdf to a pdf $p(\phi)$ for ϕ.

(ii) Determine prior for ϕ using $p(\phi) \propto I(\phi)^{1/2}$.

Then (i) and (ii) give the <u>same</u> prior for ϕ.

Example Suppose $X_1 \ldots X_n \sim \text{Bernoulli}(\theta)$.

Then $i(\theta) = \dfrac{1}{\theta(1-\theta)}$.

So Jeffrey prior is $\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, $0<\theta<1$.

This is a $\text{Beta}(\frac{1}{2}, \frac{1}{2})$.

Jeffreys priors:

- can be improper
- can be defined for vector θ by

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

BUT a simpler approach is more common: find the Jeffreys prior for each 1-dim. component of θ and take the product to get the whole prior (i.e. assume prior independence).

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

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

[Odds of event $A = P(A) / (1 - P(A))$.]

We can compute posterior probabilities $P(H_i | \underset{\sim}{x})$, $i=0,1$ and compare them.

By Bayes theorem,

$$P(H_i | \underset{\sim}{x}) = \frac{P(\underset{\sim}{x} | H_i) P(H_i)}{P(\underset{\sim}{x} | H_0) P(H_0) + P(\underset{\sim}{x} | H_1) P(H_1)} \quad i=0,1 \qquad (1)$$

Note: $P(H_i | \underset{\sim}{x})$ is the probability of H_i conditioned on data $\underset{\sim}{x}$, whereas p-values <u>can't</u> be interpreted this way.

The <u>posterior odds</u> of H_0 relative to H_1 is

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

Using ①,

$$\underbrace{\frac{P(H_0 \mid \underline{x})}{P(H_1 \mid \underline{x})}}_{\text{posterior odds}} = \underbrace{\frac{P(\underline{x} \mid H_0)}{P(\underline{x} \mid H_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(H_0)}{P(H_1)}}_{\text{prior odds}}$$

where the <u>Bayes factor</u> of H_0 relative to H_1 is

$$B_{01} = \frac{P(\underline{x} \mid H_0)}{P(\underline{x} \mid H_1)} \qquad ②$$

The change from prior odds to posterior odds depends on $\underline{x}$ only via the Bayes factor B_{01}.

B_{01} tells us how $\underline{x}$ shifts our strength of belief in H_0 relative to H_1.

General setup

We are assuming we have

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

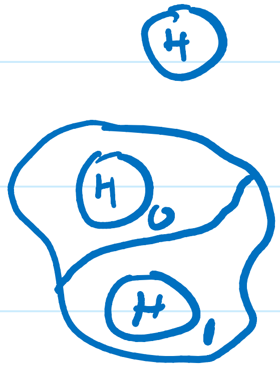
(ii) a prior distribution for θ_i under H_i,

i.e. $\pi(\theta_i \mid H_i)$ for $\theta_i \in \Theta_i$, $\quad i=0,1$.

(iii) a model under H_i for data $\underline{x}$ given by $f(\underline{x} \mid \theta_i, H_i)$

The two $\begin{cases} \text{priors in (ii) could be of different forms} \\ \text{models in (iii) could be of different forms.} \end{cases}$

Sometimes (see example later) (i) and (ii) might be combined. The prior might be $\pi(\theta)$ for $\theta \in \Theta$ where



- $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \phi$

- prior probabilities are $P(H_i) = \int_{\theta \in \Theta_i} \pi(\theta)\, d\theta$

- and $\pi(\theta_i \mid H_i)$ is the conditional density of θ given H_i,

i.e. $\pi(\theta_i \mid H_i) = \dfrac{\pi(\theta_i)}{\int_{\theta \in \Theta_i} \pi(\theta)\, d\theta}$.

Consider ②: conditioning on θ_i (law of total prob) we have

$$P(\underline{x} \mid H_i) = \int_{\Theta_i} f(\underline{x} \mid \theta_i, H_i)\, \pi(\theta_i \mid H_i)\, d\theta_i \qquad ③.$$

$P(\underline{x} \mid H_i)$ is called the marginal likelihood for H_i: it is the likelihood $f(\underline{x} \mid \theta_i, H_i)$ averaged over Θ_i, weighted according to the prior $\pi(\theta_i \mid H_i)$.

So here we average over θ. The Bayes factor is the ratio of two such averages:

$$B_{01} = \frac{P(\underline{x} \mid H_0)}{P(\underline{x} \mid H_1)}.$$

This is somewhat similar to the likelihood ratio of Sec. 3, except for LR we maximised over H_0, H_1 to find LR statistic Λ.

Note: 1. We are treating H_0, H_1 in the same way, whereas in Sec 3 we treated H_0, H_1 asymmetrically.

2. Bayes factor of H_1 relative to H_0 is just $B_{10} = B_{01}^{-1}$.

3. Bayes factors can only be used with proper priors: from (2), (3) B_{01} depends on two constants of proportionality (one for each $\pi(\theta_i \mid H_i)$) so these constants must be known.

Assume our model is $f(\underset{\sim}{x} \mid \theta)$.

If $H_i: \theta = \theta_i$, $i = 0, 1$, are both simple, then

$$B_{01} = \frac{f(\underset{\sim}{x} \mid \theta_0)}{f(\underset{\sim}{x} \mid \theta_1)} \quad \leftarrow \text{Lik ratio}$$

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

$$B_{01} = \frac{\int_{\Theta_0} f(\underset{\sim}{x} \mid \theta) \pi(\theta \mid H_0) \, d\theta}{\int_{\Theta_1} f(\underset{\sim}{x} \mid \theta) \pi(\theta \mid H_1) \, d\theta}.$$

Intepretation of Bayes factor:

B_{01}	Evidence for H_0
<1	negative (i.e. evidence supports H_1)
$1-3$	hardly worth a mention
$3-20$	positive
$20-150$	strong
>150	very strong

Example ("IQ") Suppose $X \sim N(\theta, \sigma^2)$ where $\sigma^2 = 100$.

So $f(x|\theta) = \dfrac{1}{\sqrt{200\pi}} e^{-\frac{1}{200}(x-\theta)^2}$.

Let $H_0: \theta = 100$, $H_1: \theta = 130$.

Suppose we observe $x = 120$.

Then $B_{01} = \dfrac{f(120|100)}{f(120|130)} = 0.223$.

$B_{10} = 1/0.223 = 4.48$, so positive evidence for H_1

Let prior probabilities be $P(H_0) = 0.95$, $P(H_1) = 0.05$.

Using post. odds = Bayes factor × prior odds,

$$\frac{p_0}{1-p_0} = B_{01} \times \frac{0.95}{0.05} \qquad \text{where } p_0 = P(H_0 \mid \underline{x})$$

Solving, $p_0 = \dfrac{19 B_{01}}{1 + 19 B_{01}} = 0.81$, so still a high

posterior probability of H_0.

<u>Example</u> ("Weight") $X_1, \dots, X_n \mid \theta \sim N(\theta, \sigma^2)$, $\sigma^2 = 3^2$

Let $H_0: \theta \leq 175$, $H_1: \theta > 175$

Prior: $\theta \sim N(\mu_0, \sigma_0^2)$, $\mu_0 = 170$, $\sigma_0^2 = 5^2$.

Prior prob: $P(H_0) = P(N(\mu_0, \sigma_0^2) \leq 175) = \Phi\left(\frac{175 - \mu_0}{\sigma_0}\right) = 0.84$

Prior odds: $\frac{P(H_0)}{P(H_1)} = \frac{0.84}{0.16} = 5.3.$

Observe $x_1, \dots, x_n$, $n = 10$, $\bar{x} = 176.$

Posterior $N(\mu_1, \sigma_1^2)$, $\mu_1 = \dots = 175.8$, $\sigma_1^2 = \dots = 0.869.$

Posterior prob: $P(H_0 | \underline{x}) = \Phi\left(\dfrac{175 - 175.8}{\sqrt{0.869}}\right) = 0.198.$

Post odds $= \dfrac{0.198}{0.802} = 0.24$

So Bayes factor $B_{01} = \dfrac{\text{post. odds}}{\text{prior odds}} = 0.0465.$

and $B_{10} = B_{01}^{-1} = 21.5$

Data provide strong evidence in favor of H_1

Example

[Example from Carlin and Louis (2008).]

Product P_0 – old, standard.

Product P_1 – newer, more expensive.

Assumptions:

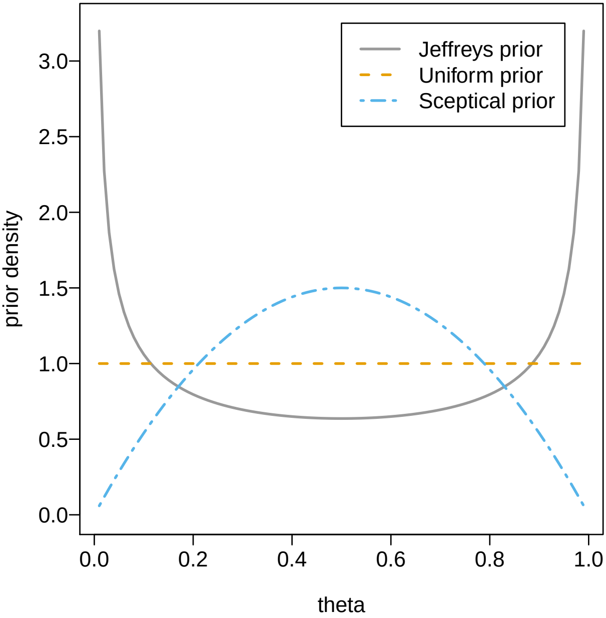
- the probability θ that a customer prefers P_1 has prior $\pi(\theta)$ which is Beta(a, b)
- the number of customers X (out of n) that prefer P_1 is $X \sim \text{Binomial}(n, \theta)$.

Let's say $\theta \geqslant 0.6$ means that P_1 is a substantial improvement over P_0. So take

$$H_0 : \theta \geqslant 0.6 \quad \text{and} \quad H_1 : \theta < 0.6.$$

We consider 3 possibile priors:

- Jeffreys' prior: $\theta \sim \text{Beta}(0.5, 0.5)$.
- Uniform prior: $\theta \sim \text{Beta}(1, 1)$.
- Sceptical prior: $\theta \sim \text{Beta}(2, 2)$, i.e. favours values of θ near $\frac{1}{2}$.



Jeffreys prior
Uniform prior
Sceptical prior
prior density
theta
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0

Prior odds $= P(H_0)/P(H_1)$ where

$$P(H_0) = \int_{0.6}^{1} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta$$

$$P(H_1) = \int_{0}^{0.6} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta.$$

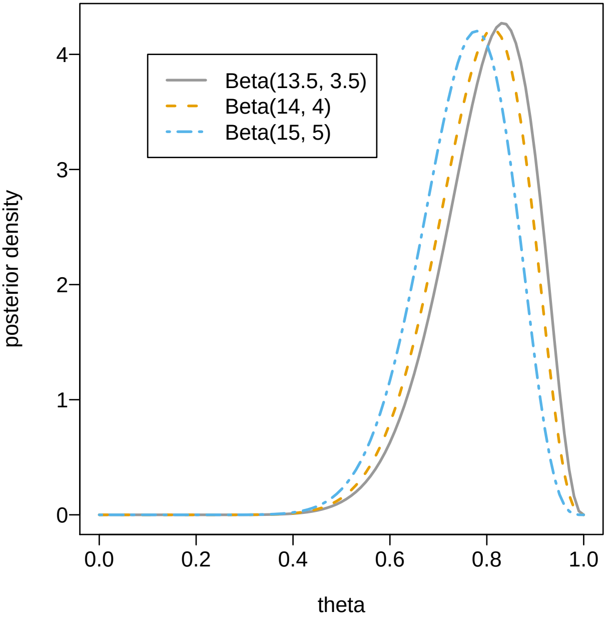
Suppose we have $x = 13$ "successes" from $n = 16$ customers.

Then (Section 4.1) the posterior $\pi(\theta \,|\, x)$ is $\mathrm{Beta}(x + a, n - x + b)$ with $x = 13$ and $n = 16$.

Posterior odds $= P(H_0 \,|\, x)/P(H_1 \,|\, x)$ where

$$P(H_0 \,|\, x) = \int_{0.6}^{1} \frac{1}{B(x + a, n - x + b)} \theta^{x+a-1} (1 - \theta)^{n-x+b-1} \, d\theta$$

$$P(H_1 \,|\, x) = \int_{0}^{0.6} \frac{1}{B(x + a, n - x + b)} \theta^{x+a-1} (1 - \theta)^{n-x+b-1} \, d\theta.$$


Beta(13.5, 3.5)
Beta(14, 4)
Beta(15, 5)
posterior density
theta
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

Prior	Prior odds	Posterior odds	Bayes factor
$\text{Beta}(0.5, 0.5)$	0.773	26.6	34.4
$\text{Beta}(1, 1)$	0.667	20.5	30.8
$\text{Beta}(2, 2)$	0.543	13.4	24.6

Conclusion: strong evidence for H_0.

4.5 Asymptotic normality of posterior distribution

We have $\pi(\theta|\underline{x}) \propto L(\theta)\,\pi(\theta)$

Let $\tilde{l}(\theta) = \log \pi(\theta|\underline{x})$

$$= \text{constant} + \underbrace{l(\theta)}_{\sum_{i=1}^{n} \log f(x_i|\theta),\ n \text{ terms},} + \underbrace{\log \pi(\theta)}_{\text{one term}}$$

expect likelihood contribution to dominate for large n

Let $\tilde{\theta}$ be the posterior mode, assume $\tilde{l}'(\tilde{\theta}) = 0$.

Then

$$\tilde{l}(\theta) \approx \tilde{l}(\tilde{\theta}) + (\tilde{\theta} - \theta)\underbrace{\tilde{l}'(\tilde{\theta})}_{=0} + \frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{l}''(\tilde{\theta})$$

$$= \tilde{l}(\tilde{\theta}) - \frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{J}(\tilde{\theta})$$

where $\tilde{J}(\theta) = -\tilde{l}''(\theta)$.

So $\pi(\theta | \underset{\sim}{x}) = \exp(\tilde{l}(\theta)) \propto \exp\left(-\frac{1}{2}(\theta - \tilde{\theta})^2 \tilde{J}(\tilde{\theta})\right)$

i.e. $\theta \,\Big|\, \underset{\sim}{x} \approx N\left(\tilde{\theta}, \tilde{J}(\tilde{\theta})^{-1}\right)$

$$\theta \mid \underset{\sim}{x} \approx N(\tilde{\theta}, \tilde{J}(\tilde{\theta})^{-1}) \qquad (1)$$

In large samples the likelihood contribution will dominate, resulting in $\tilde{\theta}$ and $\tilde{J}(\tilde{\theta})$ being close to the MLE $\hat{\theta}$ and observed information $J(\hat{\theta})$. Hence

$$\theta \mid \underset{\sim}{x} \approx N(\hat{\theta}, J(\hat{\theta})). \qquad (2)$$

(1), (2) look similar to the corresponding frequentist results, but note:

in (1), (2), θ is a random variable and $\tilde{\theta}(\underset{\sim}{x})$, $\hat{\theta}(\underset{\sim}{x})$ constants whereas in frequentist $\hat{\theta}(\underset{\sim}{X})$ is a random variable and θ constant.

Using the asymptotic results:

(i) frequentist $\hat{\theta} \approx N(\theta, J(\theta)^{-1})$ leads to 95% confidence interval of $\left(\hat{\theta} \pm 1.96\, J(\hat{\theta})^{-1/2}\right)$

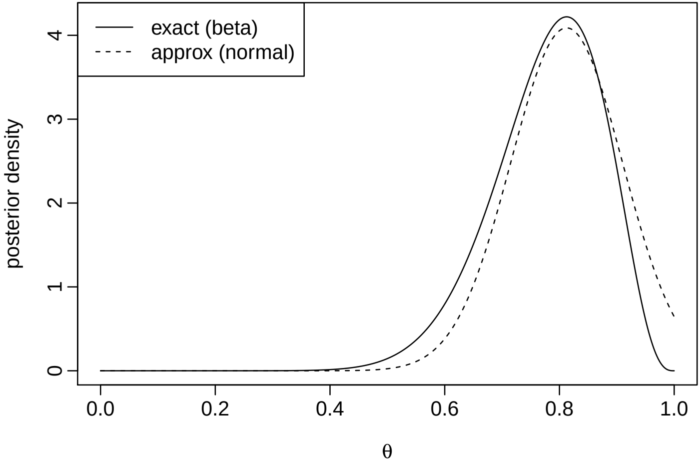
(ii) Bayesian (2) $\theta \mid \underset{\sim}{x} \approx N(\hat{\theta}, J(\hat{\theta})^{-1})$ leads to 95% credible interval of $\left(\hat{\theta} \pm 1.96\, J(\hat{\theta})^{-1/2}\right)$.

That is, the <u>same</u> interval of θ-values in both cases, but with <u>different</u> interpretations.

Normal approx to posterior (1)

Prior $\theta \sim U(0, 1)$.

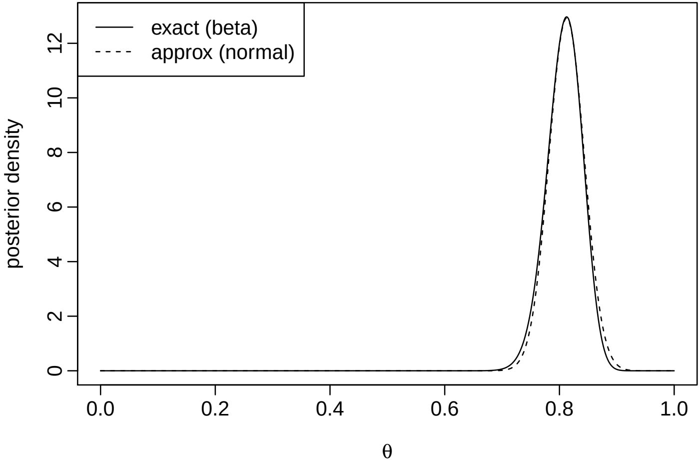
Bernoulli likelihood: $x = 13$ successes out of $n = 16$ trials.



Normal approx to posterior (2)

Prior $\theta \sim U(0,1)$.

Bernoulli likelihood: $x = 130$ successes out of $n = 160$ trials.



Part B courses

double unit, practicals, R

SB1 : applied, computational, regression models

SB2.1 : statistical inference, frequentist and Bayesian

SB2.2 : machine learning

SB3.1 : applied probability

SB3.2 : lifetime models