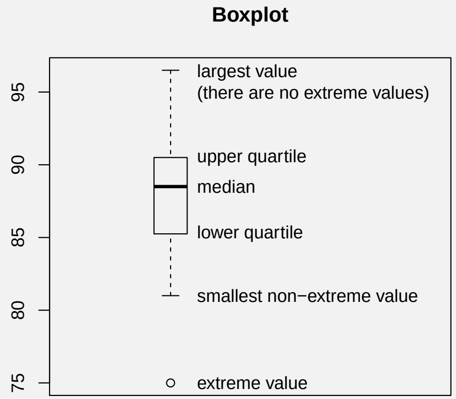
Boxplots

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

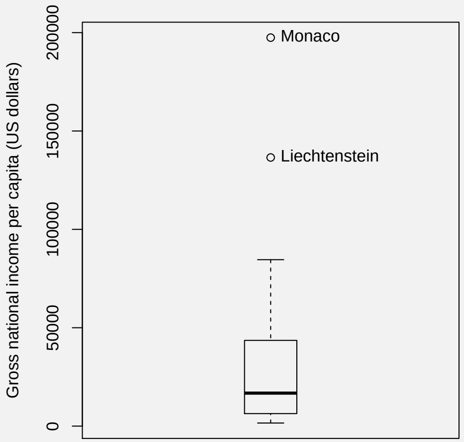


The box extends from one quartile to the other, and the central line in the box is the median.

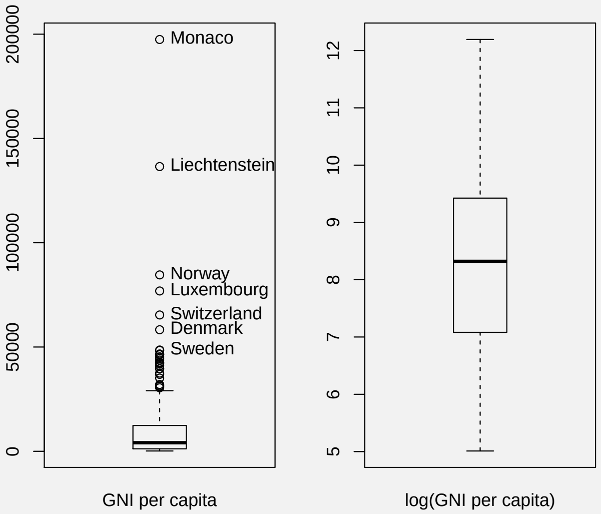
The whiskers are drawn from the box to the most extreme observations that are no more than 1.5×IQR from the box. (Alternatively $r\times$IQR can be used for other values of r.)

Observations which are more extreme than this are shown separately.

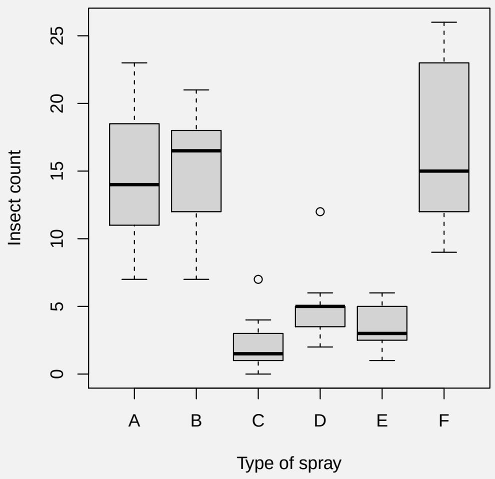
Gross national income per capita (World Bank data, 2009) for 50 “sovereign states in Europe.” `http://en.wikipedia.org/wiki/List_of_sovereign_states_in_Europe_by_GNI_(nominal)_per_capita`



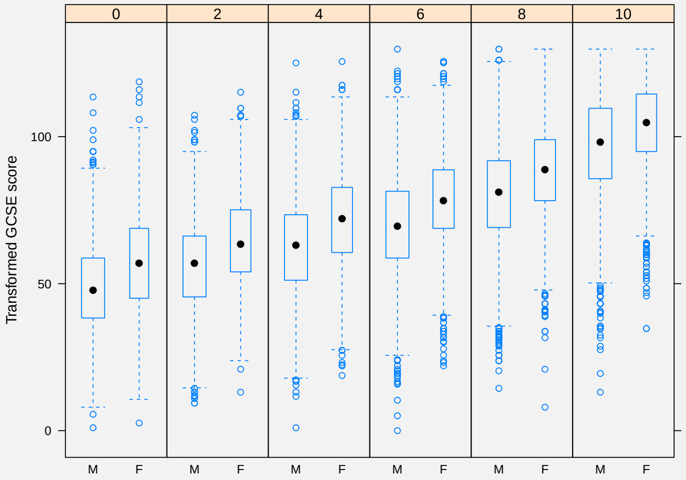
Now for 182 countries worldwide (including Europe).



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.

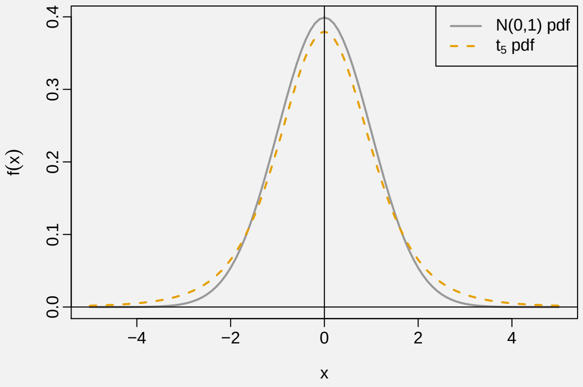


Comparing $N(0, 1)$ and t distributions

A t-distribution with r degrees of freedom has pdf

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

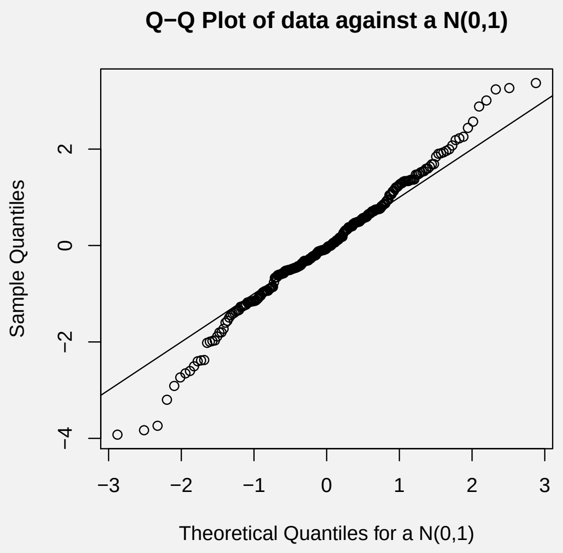
[More on t-distributions later.] Consider $r = 5$.



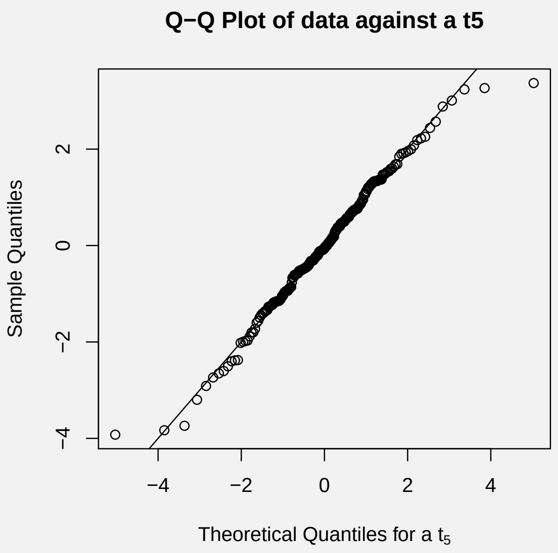
Suppose we simulate data $(x_1, \ldots, x_{250})$ from a t_5 distribution.

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?

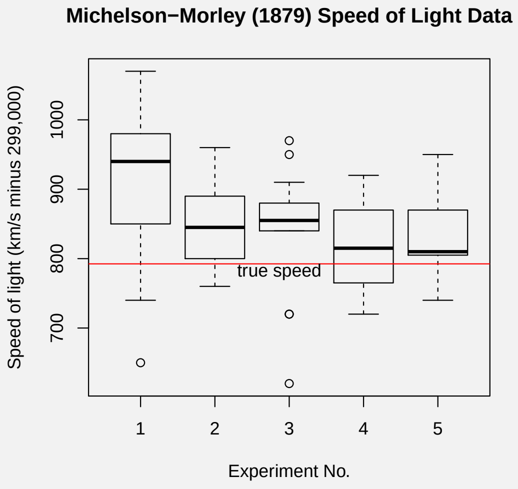


A $N(0,1)$ assumption is not good – as expected.

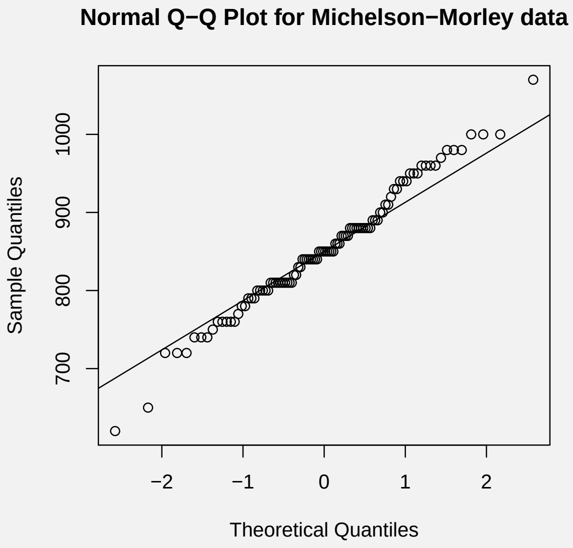


A t_5 assumption is ok – as expected.

Normal Q-Q plots

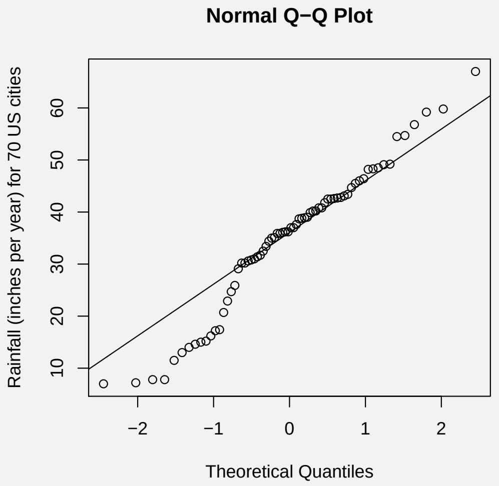


20 observations from each experiment. Is a $N(\mu, \sigma^2)$ distribution plausible for these 100 observations?



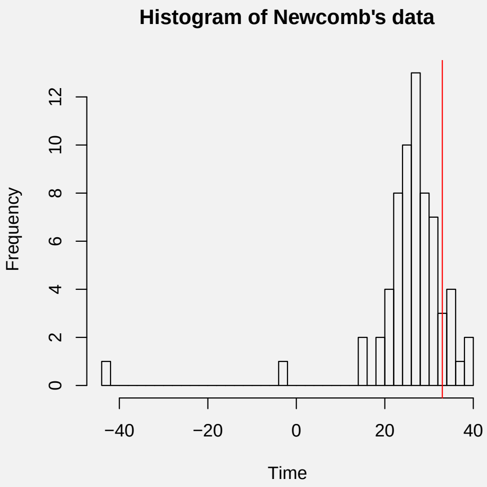
From the plot a normal distribution seems reasonable.

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

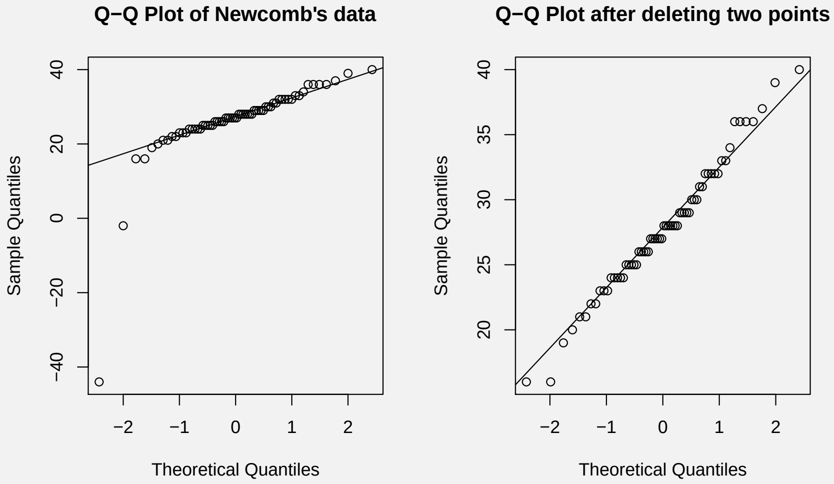


A normal assumption doesn't look good – problems in the lower tail.

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.

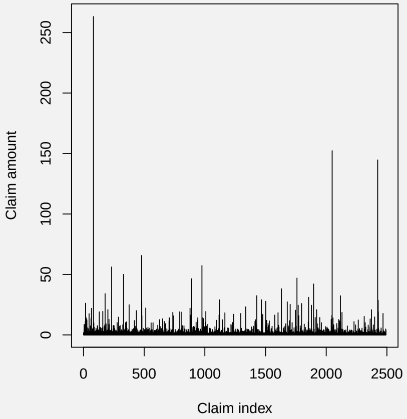


This time the problems are different – two (very small) outlying observations. If these are removed, a normal assumption looks ok.



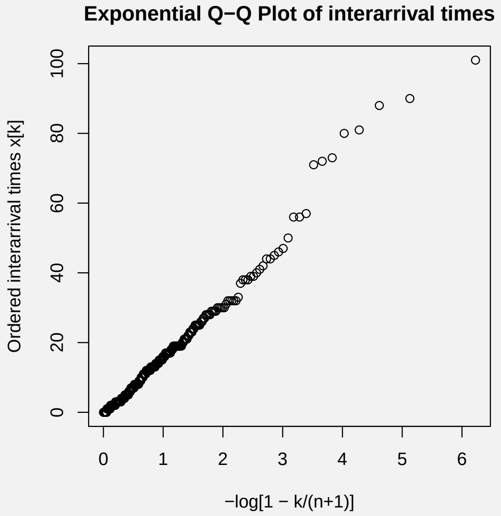
Example: Danish fire data (Davison, 2003)

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

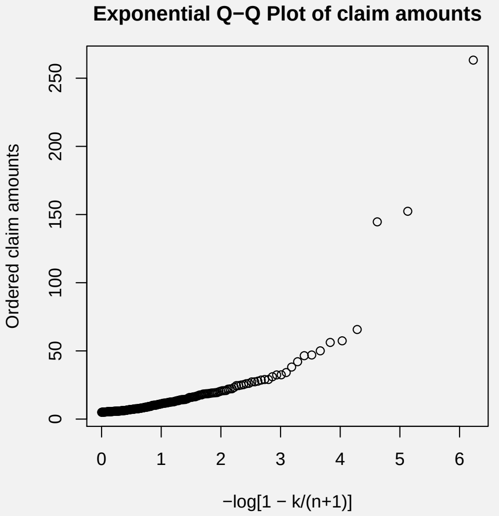


Following Davison, let's consider the 254 largest claim amounts, and the interarrival times between these claims.

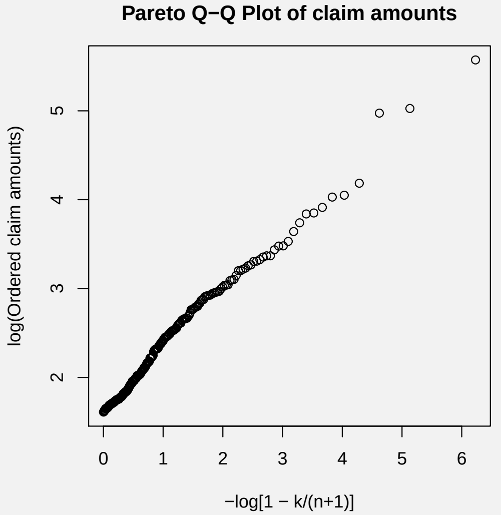
Is it reasonable to assume exponential interarrival times? See below – inter-arrivals look fairly close to exponential.



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.



Multivariate normal distribution

[Multivariate/bivariate normal: recap from Part A Probability.]

(1) Let $Z_1, \dots, Z_p \overset{\text{iid}}{\sim} N(0,1)$. Then joint pdf of $Z_1, \dots, Z_p$ is

$$f(\mathbf{z}) = \prod_{j=1}^{p} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2} z_j^2 \right)$$

$$= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^{p} z_j^2 \right)$$

$$= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right), \quad \mathbf{z} \in \mathbb{R}^p.$$

We write $\mathbf{Z} \sim N(\mathbf{0}, I)$ where $\mathbf{0}$ is a p-vector of zeroes and I is the $p \times p$ identity matrix.

(2) Now let $\boldsymbol{\mu}$ be a $p \times 1$-vector and let Σ be a $p \times p$ symmetric, positive definite matrix.

Then $\mathbf{X} = (X_1, \ldots, X_p)$ is said to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if its pdf is

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

We write $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$.

This pdf reduces to the $N(\mathbf{0}, I)$ pdf when $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I$.

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

- $\boldsymbol{\mu}$ is the *mean vector*: $E(X_j) = \mu_j$
- Σ is the *covariance matrix*: $\text{var}(X_j) = \Sigma_{jj}$ and $\text{cov}(X_j, X_k) = \Sigma_{jk}$
- the marginal distribution of X_j is $X_j \sim N(\mu_j, \Sigma_{jj})$.

Example (bivariate normal distribution)

Suppose $p = 2$. Let $-1 < \rho < 1$ and

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

Then the pdf of (X_1, X_2) is

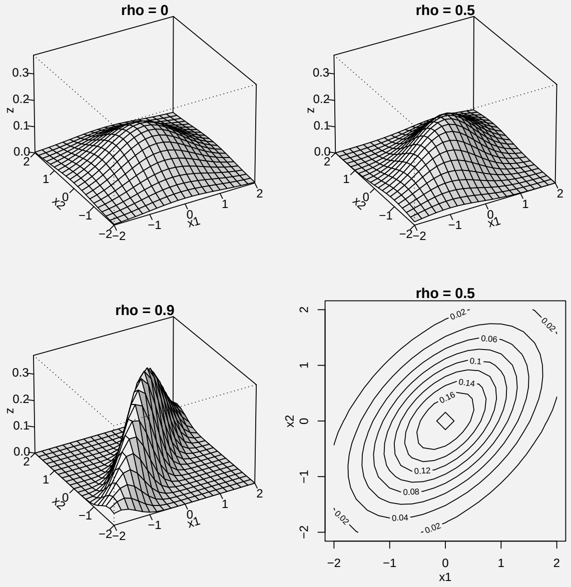
$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)} (x_1^2 - 2\rho x_1 x_2 + x_2^2) \right).$$

The marginal distributions are $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$.

The quantity ρ is the correlation between X_1 and X_2.

X_1 and X_2 are independent $\iff$ $\rho = 0$.

Below: pdf of (X_1, X_2) for $\rho = 0$, 0.5, 0.9, and contour-plot for $\rho = 0.5$.



Opinion polls (revision)

"Smoking should be banned completely"

In an ICM poll for the BBC (Feb 2006), 485 people agreed with this statement, and 501 disagreed.

Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th person agreed} \\ 0 & \text{otherwise} \end{cases}$$

and assume $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$, with $n = 986$. So $\widehat{p} = \overline{x}$.

By CLT,

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2} \right) \approx 1 - \alpha.$$

Estimating the variance $p(1-p)/n$ by $\widehat{p}(1-\widehat{p})/n$, we get an approximate $1 - \alpha$ CI for p of $(\overline{x} \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n})$.

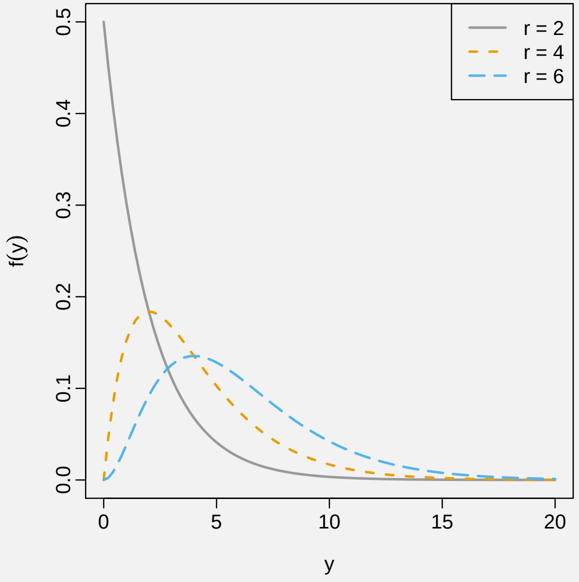
For the above data, and using $\alpha = 0.05$, an approximate 95% CI for p is (0.461, 0.523).

"Smoking should be banned in all public places but not completely"

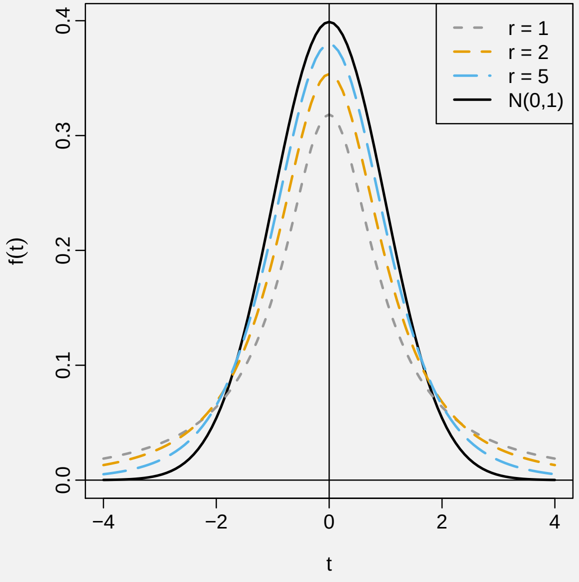
645 agreed, 341 disagreed

This time the approximate 95% CI for p is (0.624, 0.684).

Chi-squared pdfs



t distribution pdfs



Student’s Sleep data

“Student” = W.S. Gosset

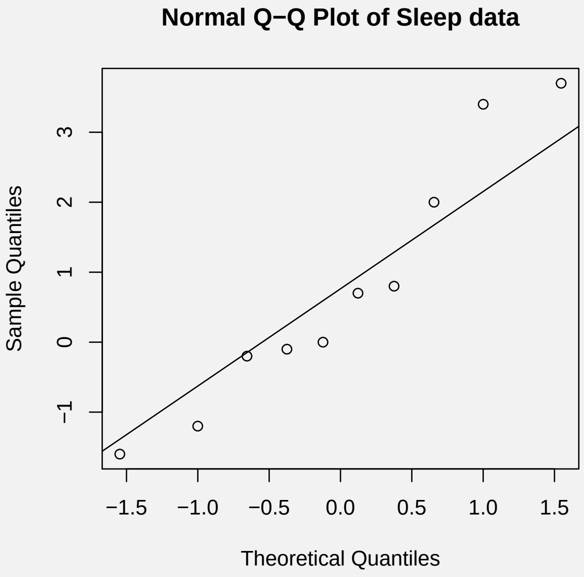
Below is half of Student’s sleep data (1908):

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

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.



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

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

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?

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

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^+.

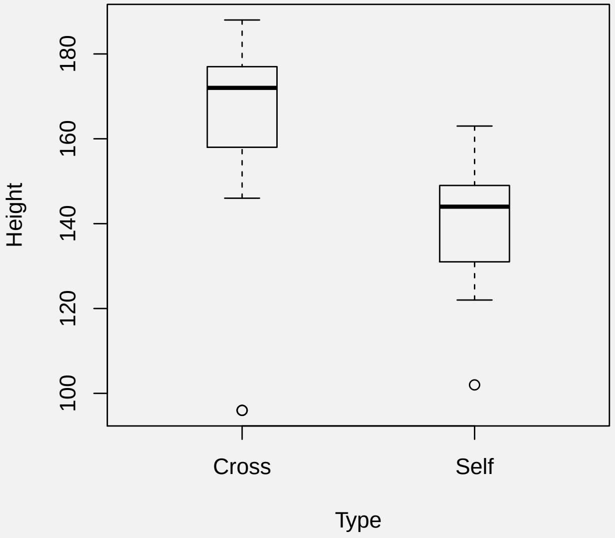
t-test (two sample)

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.]



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), $i = 1 \dots, n$. Let $D_i = X_i - Y_i$.

Suppose $D_1, \dots, D_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$, 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.

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

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

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

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:

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

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.

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^2_p where $p = \dim\Theta - \dim\Theta_0 = (3-1) - 1 = 1$.

The value $\Lambda = 0.032$ is much less than $E(\chi^2_1) = 1$. The p-value is $P(\chi^2_1 \geqslant 0.032) = 0.86$, 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

$$\begin{aligned} \pi_i(\widehat{\theta}) &= \widehat{\theta}^i e^{-\widehat{\theta}}/i!, \quad i = 0, 1, \dots, 7 \\ \pi_8(\widehat{\theta}) &= 1 - \sum_{i=0}^{7} \pi_i(\widehat{\theta}). \end{aligned}$$

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.

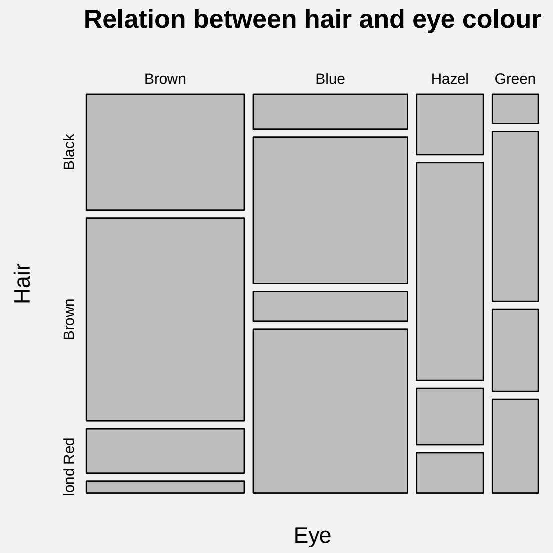
The p-value is $p = P(\chi^2_2 \geqslant 26.6) \approx 10^{-6}$, so there is clear evidence that a Poisson model is not suitable.

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?



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

$$\dim(\Theta) = 16 - 1 = 15$$

$$\dim(\Theta_0) = (4-1) + (4-1) = 6$$

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

The p-value is $P(\chi_9^2 \geqslant 146.4) \approx 0$.

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 $P = 138.3$.]

Bayesian Inference

So far we have followed the classical/frequentist approach to statistics:

- 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.

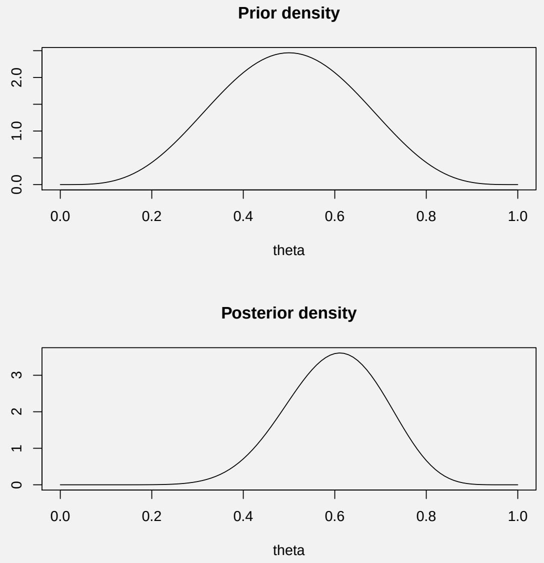
Then, once data are available, we can use Bayes' theorem to combine our prior beliefs with the data to obtain an updated “posterior” assessment of our beliefs about θ.

Example

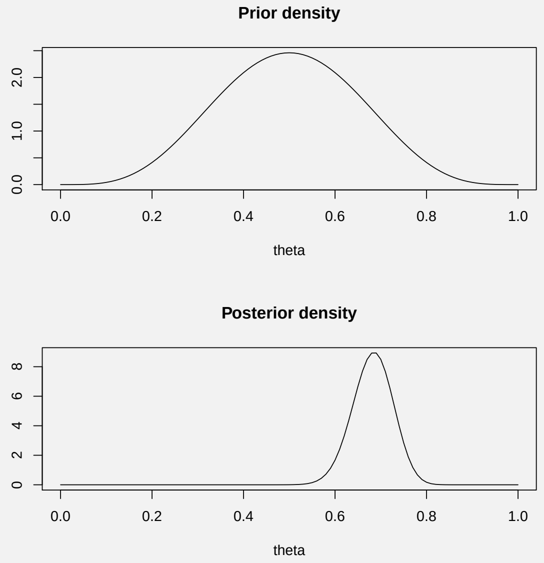
Suppose we have a coin which we think might be a bit biased.

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.



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.

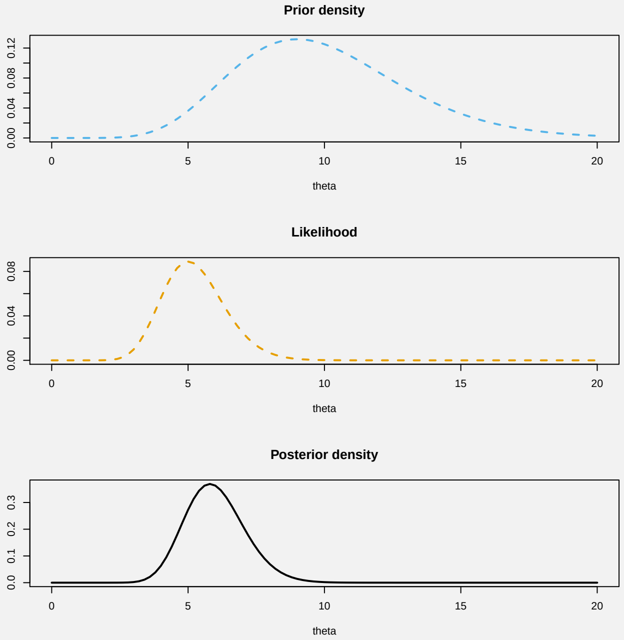
Suppose this other information, on its own, suggests plausible values of θ of about 10 per 10,000, with 95% of the support for θ lying between 5 and 17.

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.

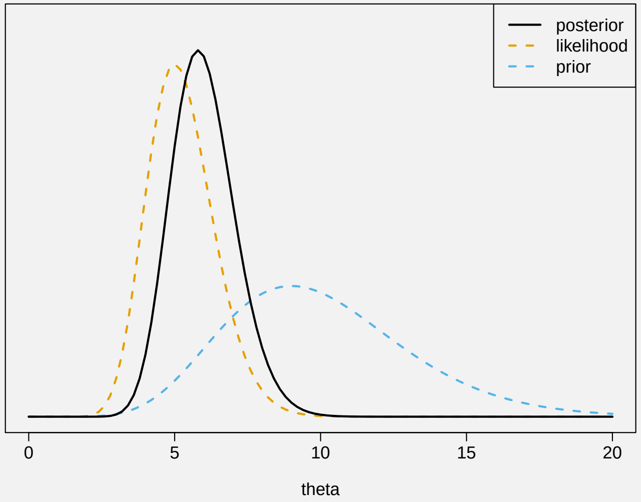
- The posterior combines the evidence from the data (i.e. the likelihood) and the other (i.e. prior) evidence. We can think of the posterior as a compromise between the likelihood and the prior.
- Calculated on board in lectures: the posterior is another Gamma.



Prior density

Likelihood

Posterior density



Prior information

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

(i) E.g. MRSA example, we might ask a scientific expert who might anticipate θ around 10, say with $\theta \in (5, 17)$ with probability 0.95.

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

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

Example

[Example from Carlin and Louis (2008).]

Product P_0 – old, standard.

Product P_1 – newer, more expensive.

Assumptions:

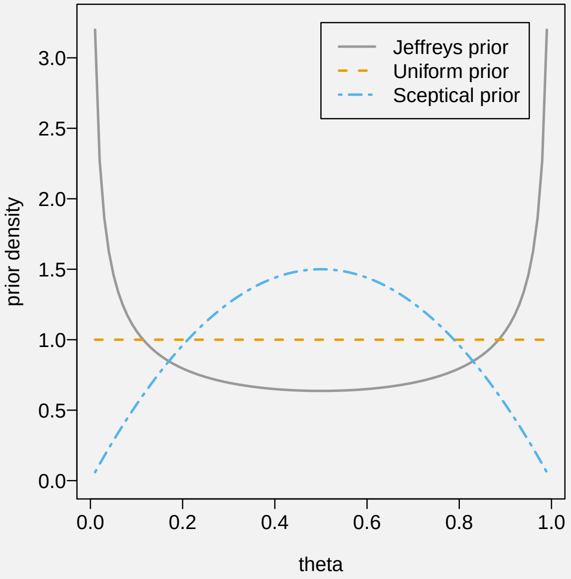
- the probability θ that a customer prefers P_1 has prior $\pi(\theta)$ which is $\text{Beta}(a, b)$
- the number of customers X (out of n) that prefer P_1 is $X \sim \text{Binomial}(n, \theta)$.

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 $\frac{1}{2}$.



Prior odds $= P(H_0)/P(H_1)$ where

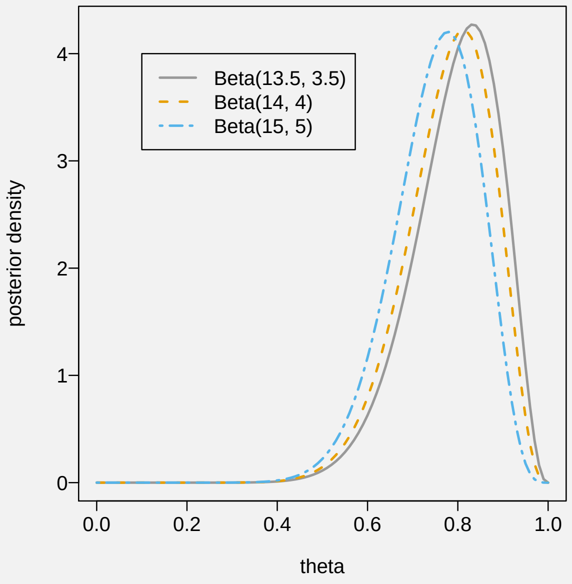
$$P(H_0) = \int_{0.6}^{1} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta$$
$$P(H_1) = \int_{0}^{0.6} \frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1}\, d\theta.$$

Suppose we have $x = 13$ "successes" from $n = 16$ customers.

Then (Section 4.1) the posterior $\pi(\theta \mid x)$ is $\text{Beta}(x+a, n-x+b)$ with $x = 13$ and $n = 16$.

Posterior odds $= P(H_0 \mid x)/P(H_1 \mid x)$ where

$$P(H_0 \mid x) = \int_{0.6}^{1} \frac{1}{B(x+a, n-x+b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1}\, d\theta$$
$$P(H_1 \mid x) = \int_{0}^{0.6} \frac{1}{B(x+a, n-x+b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1}\, d\theta.$$



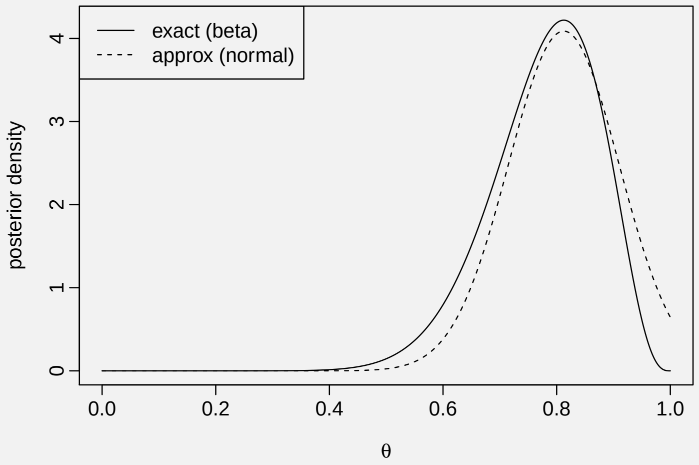
Prior	Prior odds	Posterior odds	Bayes factor
Beta$(0.5, 0.5)$	0.773	26.6	34.4
Beta$(1, 1)$	0.667	20.5	30.8
Beta$(2, 2)$	0.543	13.4	24.6

Conclusion: strong evidence for H_0.

Normal approx to posterior (1)

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

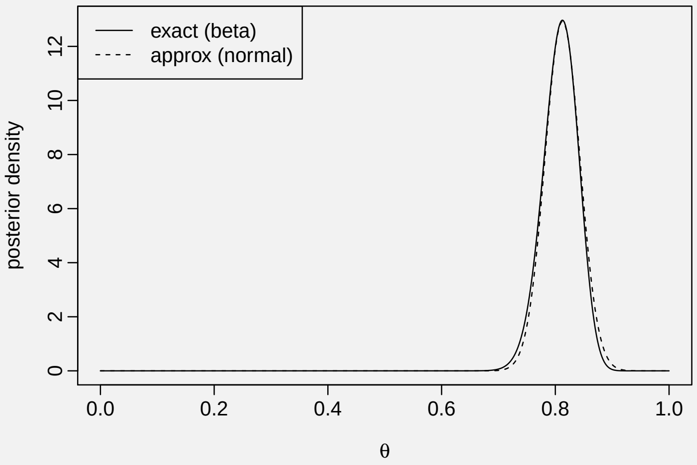
Bernoulli likelihood: $x = 13$ successes out of $n = 16$ trials.



Normal approx to posterior (2)

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

Bernoulli likelihood: $x = 130$ successes out of $n = 160$ trials.


exact (beta)
approx (normal)
posterior density
θ