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 $\underline{x} = (x_1, \ldots, x_n)$, numerical values. We regard these as observed values of i.i.d. random variables $X_1, \ldots, X_n$ with the same distribution as X, so $\underset{\sim}{X} = (X_1, \ldots, X_n)$ 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 $f(\underset{\sim}{x};\theta) = \displaystyle\prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\Sigma x_i}}{\Pi\, x_i!}$.

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 $X_1, \dots, X_n$ are iid with $E(X_i) = \mu$, $var(X_i) = \sigma^2$.

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

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

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

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

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

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

$$\underbrace{g(\bar{X}) \overset{D}{\approx}}_{\text{asymptotic distribution}} N\Big(\underbrace{g(\mu)}_{\text{asymp. mean}},\ \underbrace{g'(\mu)^2 \frac{\sigma^2}{n}}_{\text{asymp. variance}} \Big)$$

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

Example $X_1, \dots, X_n$ iid exponential with

parameter or rate λ.

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

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

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

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

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

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

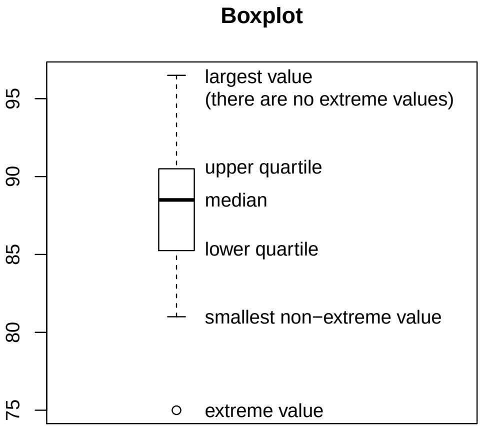
median

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

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.



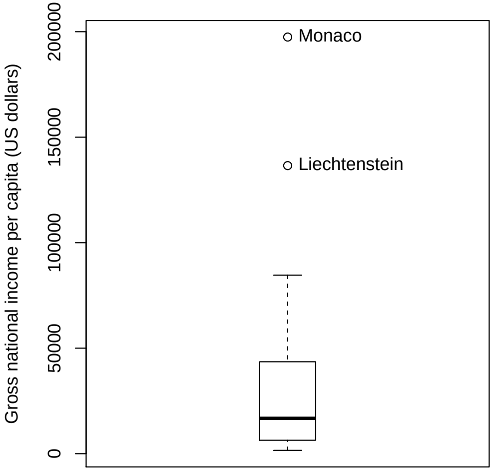
The box extends from one quartile to the other, and the central line in the box is the median.

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.)

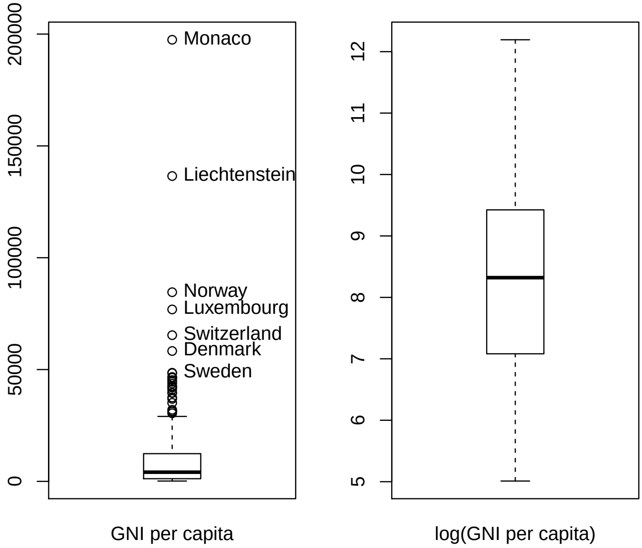
Observations which are more extreme than this are shown separately.

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

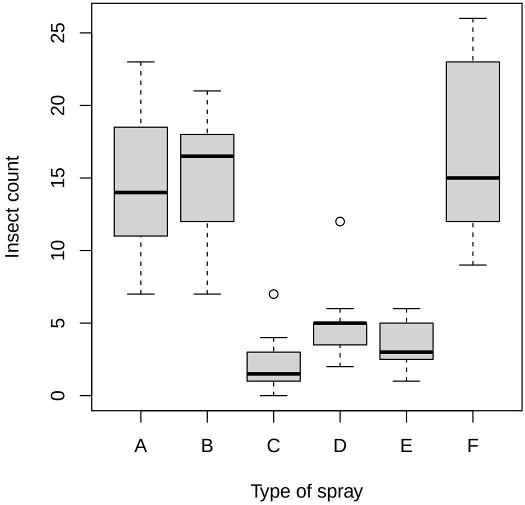
http://en.wikipedia.org/wiki/List_of_sovereign_states_in_Europe_by_GNI_(nominal)_per_capita



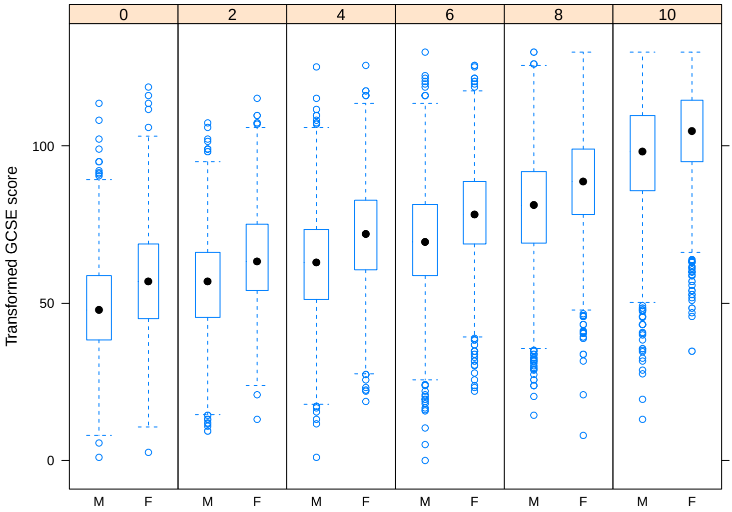
Now for 182 countries worldwide (including Europe).



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 ($0 = \text{worst}, 2, 4, 6, 8, 10 = \text{best}$) 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

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

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

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

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

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

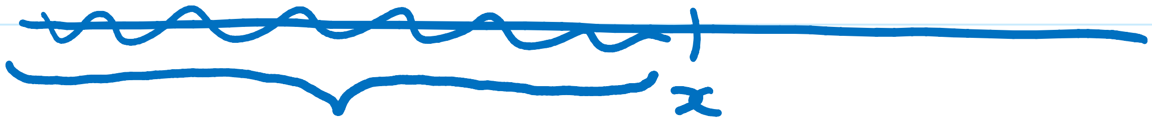
So pdf $f_{(1)}(x) = F_{(1)}'(x) = n[1 - F(x)]^{n-1} \cdot f(x)$

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,

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

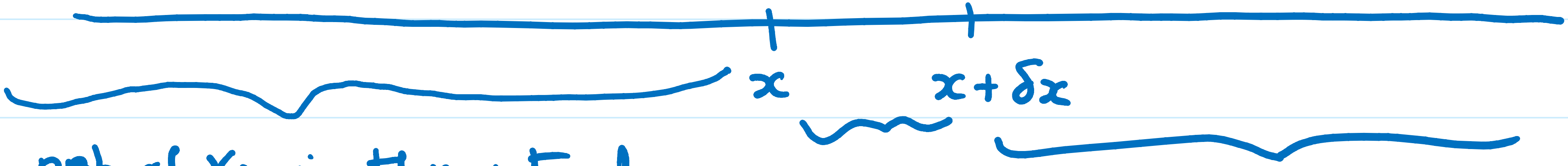
$$= \binom{n}{r} F(x)^{r} [1-F(x)]^{n-r-1} (n-r) f(x)$$

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 $f_{(r)}$



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

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

Omitting the δx gives $f_{(r)}(x)$

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

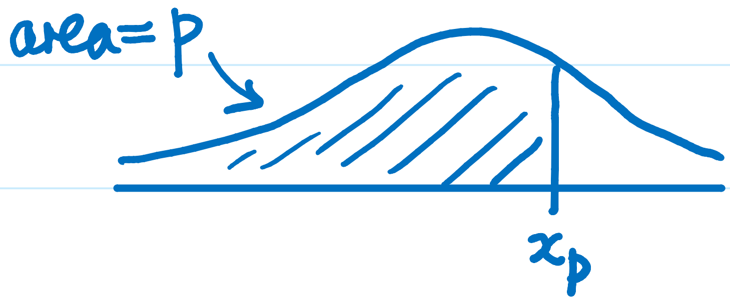
1.4 Q-Q plots

"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 **probability integral transform** of X.

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

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

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) $\text{var}[U_{(r)}] = \dfrac{r}{(n+1)(n+2)}\left(1 - \dfrac{r}{n+1}\right)$

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

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

<u>Question</u>: is it reasonable to assume data $x_1, \dots, x_n$ are a random sample from F ?

By Lemma 1.2 we can generate a random sample $X_1, \dots, X_n$ from F by first taking $U_1, \dots, U_n \overset{iid}{\sim} U(0,1)$ 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

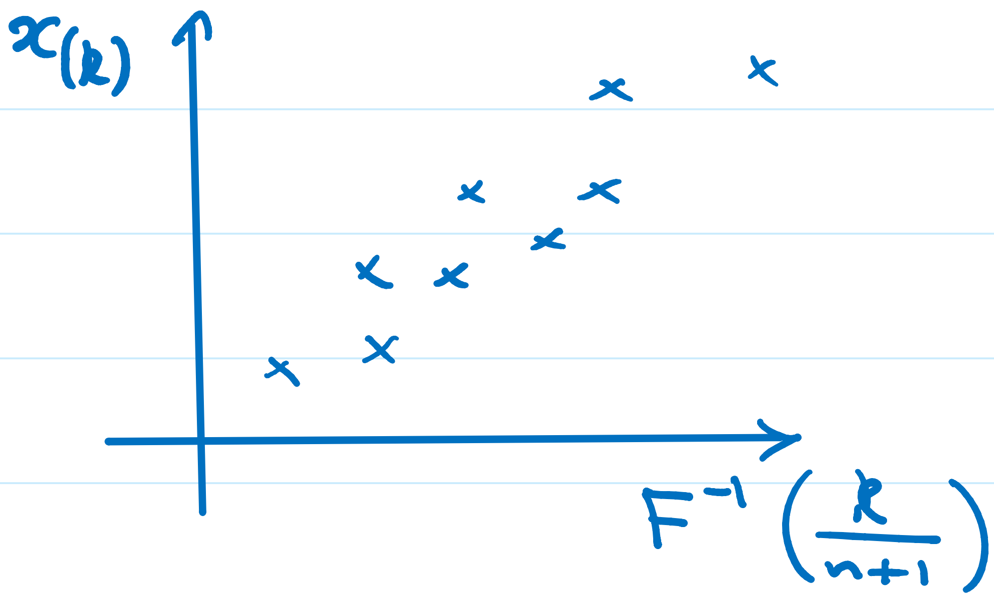
$$E[X_{(k)}] = E[F^{-1}(U_{(k)})] \quad \text{from } ①$$

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

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

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 $k=1,\ldots,n$



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.

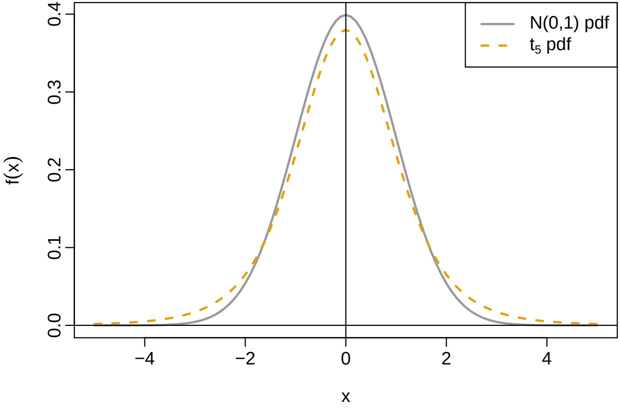
Of course we need to specify a candidate cdf F.

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

A t-distribution with r degrees of freedom has pdf

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

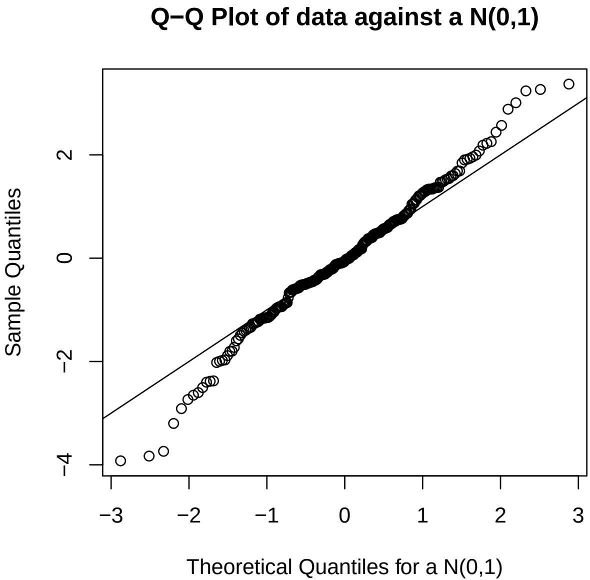
[More on t-distributions later.] Consider $r = 5$.



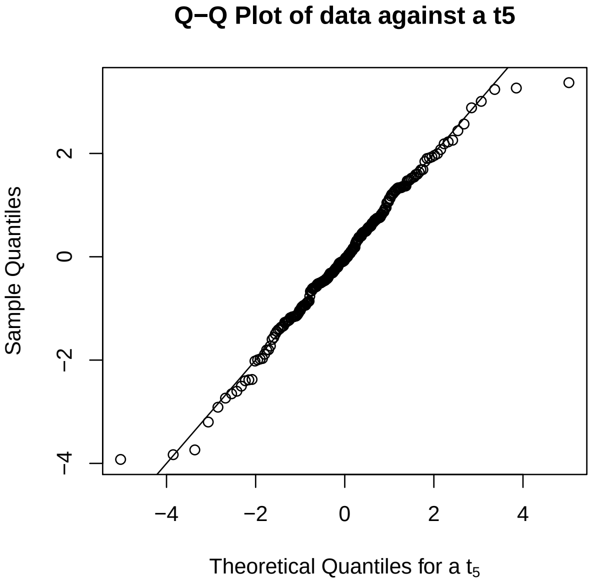
Suppose we simulate data $(x_1, \ldots, x_{250})$ from a t_5 distribution.

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 $\underline{x}$ 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

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

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

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

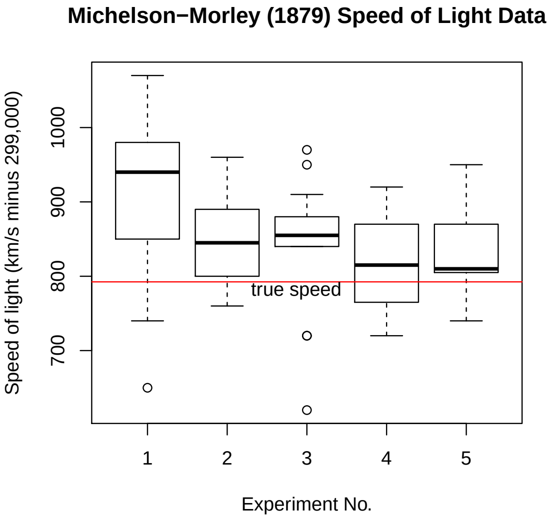
Hence $x_{(k)} \approx \sigma \Phi^{-1}\left(\frac{k}{n+1}\right) + \mu$.

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

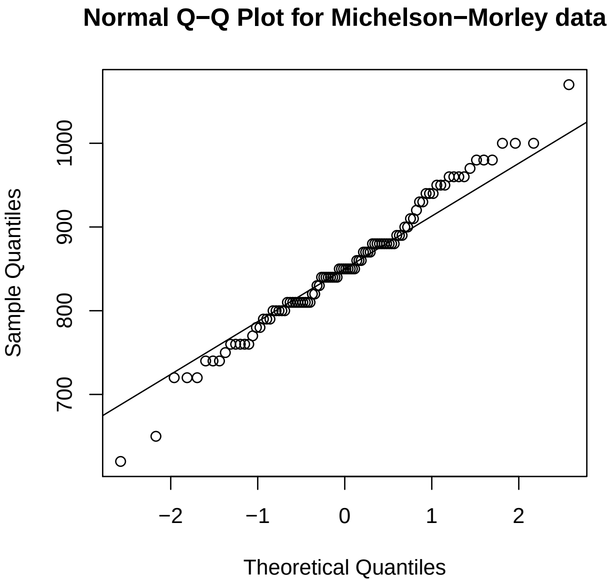
for $k = 1 \dots n$ and see if the points lie on an approx. straight line

(with gradient σ, intercept μ).

Normal Q-Q plots

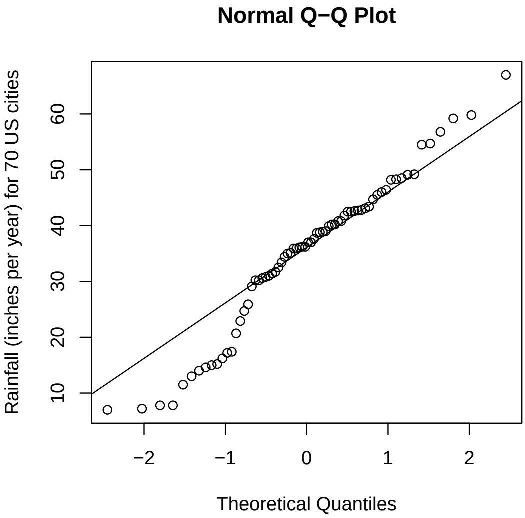


20 observations from each experiment. Is a $N(\mu, \sigma^2)$ distribution plausible for these 100 observations?



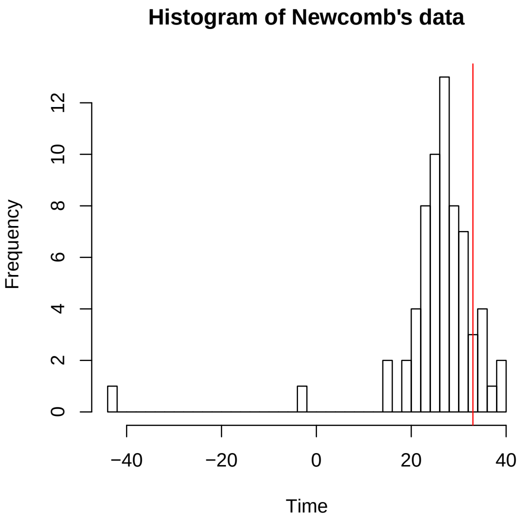
From the plot a normal distribution seems reasonable.

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

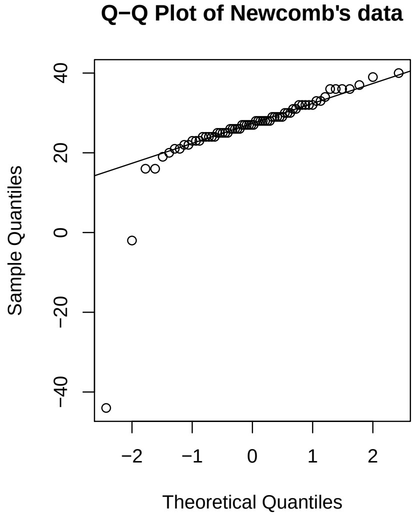


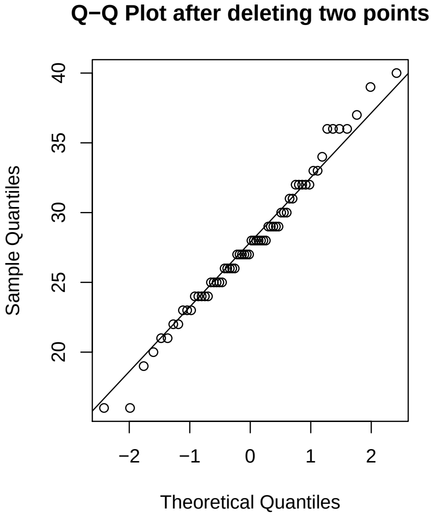
A normal assumption doesn't look good – problems in the lower tail.

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

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

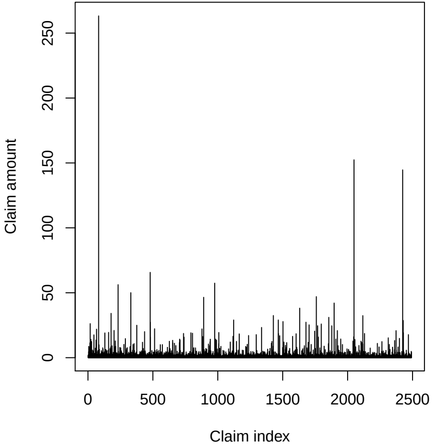
Hence $x_{(k)} \approx -\mu \log\left(1 - \frac{k}{n+1}\right)$.

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).

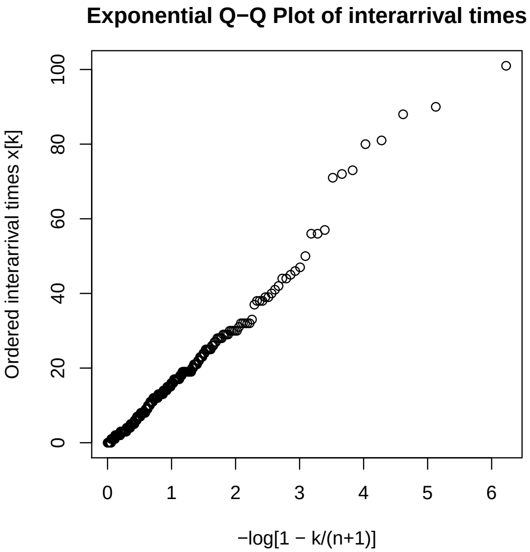
Example: Danish fire data (Davison, 2003)

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

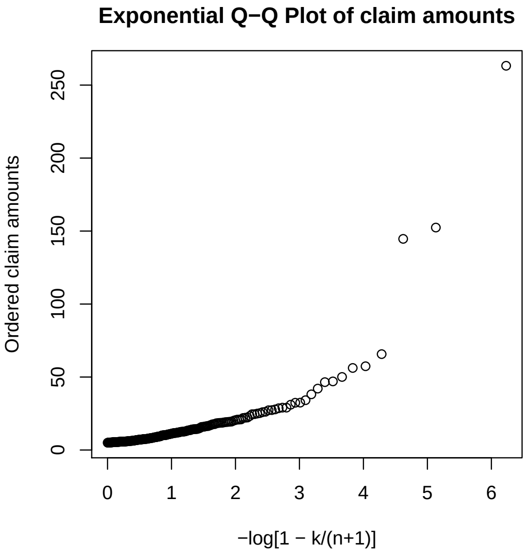


Following Davison, let's consider the 254 largest claim amounts, and the interarrival times between these claims.

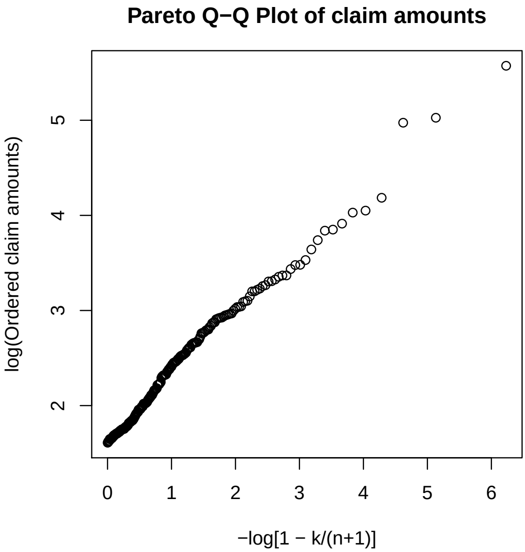
Is it reasonable to assume exponential interarrival times? See below – inter-arrivals look fairly close to exponential.



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 $l(\theta)$, the **observed information** $J(\theta)$ is defined by $J(\theta) = -\frac{d^2 l}{d\theta^2}$.

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

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

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:

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

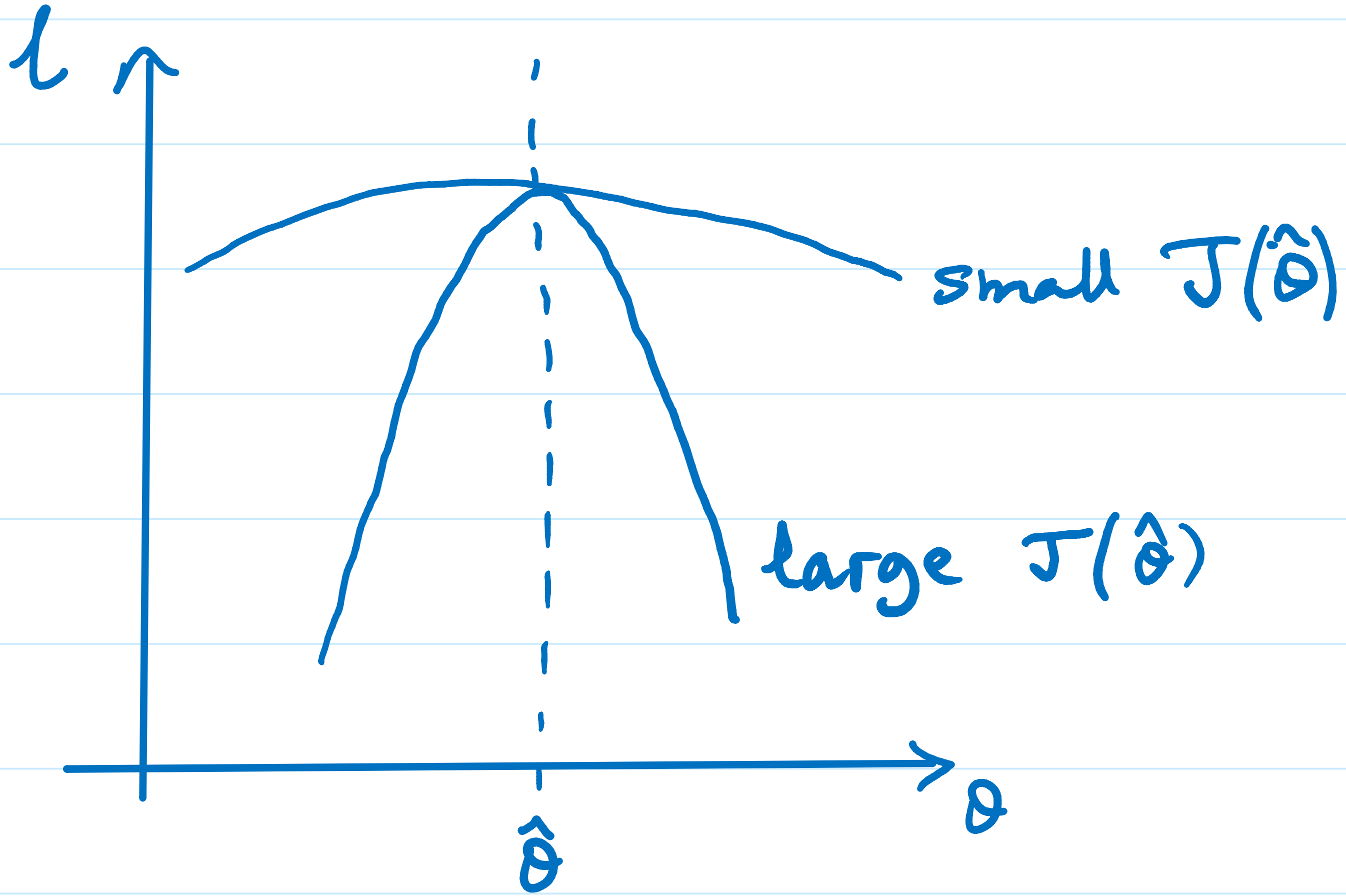
Expanding $\ell(\theta)$ as a Taylor series about $\hat{\theta}$:

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

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 $l(\theta; \underset{\sim}{X})$ and take expectations over $\underset{\sim}{X}$.

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

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

<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 $f(\underset{\sim}{x}\,;\, \underline{g^{-1}(\psi)})$ with respect to ψ.

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

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

where $U(\theta) = \frac{dl}{d\theta}$ 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 $$\hat{\theta} - \theta \approx \frac{U(\theta)}{J(\theta)}.$$

So $$I(\theta)^{1/2}(\hat{\theta} - \theta) \approx I(\theta)^{1/2} \cdot \frac{U(\theta)}{J(\theta)}$$
$$= \frac{U(\theta)/I(\theta)^{1/2}}{J(\theta)/I(\theta)} = \frac{\text{TOP}}{\text{BOTTOM}} \qquad (1).$$

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

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

$$= I(\theta)$$

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

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

For BOTTOM:

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

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

$\xrightarrow{P} 1$ 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.