Methods of Mathematical Statistics

Notes by Tim Brown and Davide Ferrari

Module 6: Point Estimation

Contents

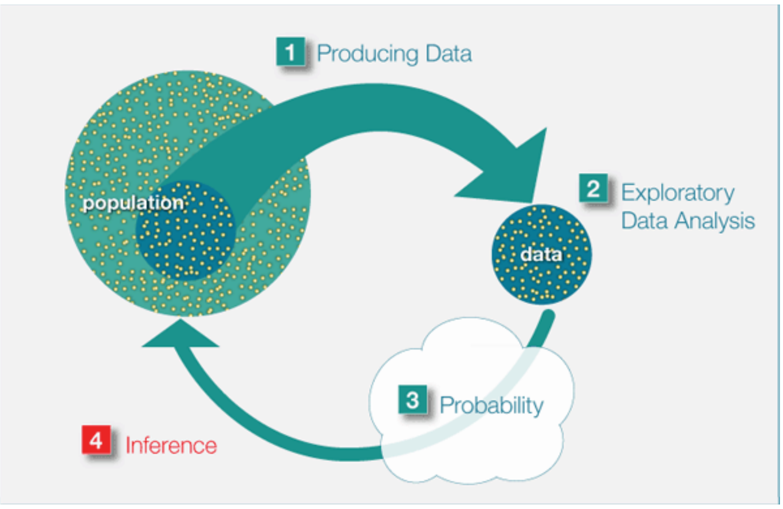

Figure 1: From 3, Probability to 4, Inference via 2, Exploratory Data Analysis

1 The Big Picture Again

Introduction

The data is assumed to be (for the moment) numbers

$$x_1, \ldots, x_n.$$

The model for the data is a *random sample*, that is a sequence of independent and identically distributed random variables

$$X_1, X_2, \ldots, X_n.$$

This model is equivalent to random selection from a hypothetical infinite population .

The goal is to use the data to learn about the distribution of the random variables from the data.

Introduction Ctd

A *Statistic* $T = \phi(X_1, \ldots, X_n)$ is a function of the sample and its realisation is denoted by $t = \phi(x_1, \ldots, x_n)$.

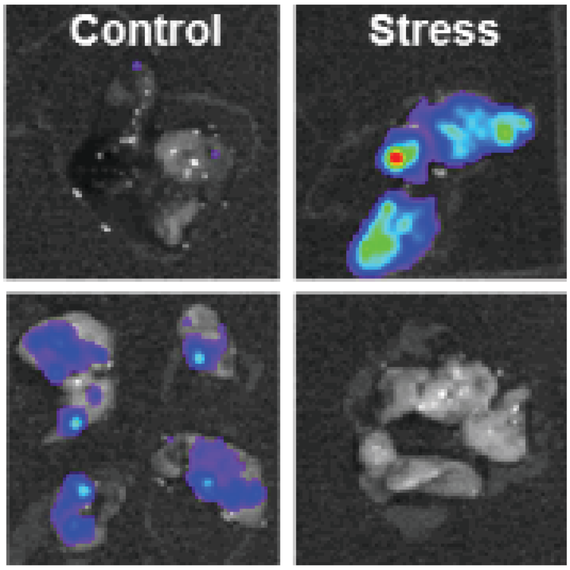

Figure 2: Differences in cancer progression

A statistic has two purposes:

- Describe/summarise the sample (Stage 2 in Figure 1)
- Estimate the true (unknown) distribution generating the sample (Stage 4 in Figure 1)

2 Descriptive Statistics, Explorary Data Analysis and Order Statistics — 6.1, 6.2 & 6.3

2.1 Example — Stress and Cancer

Stress and cancer

An experiment divided 10 mice randomly into control and stress groups, with the stress group receiving chronic stress .

The biologists measured

- Vascular endothelial growth factor C (VEGFC) — a protein involved in lymphangiogenesis — low or high levels of this are observed in many diseases
- Prostaglandin-endoperoxide synthase 2 (COX2) — a protein involved in inflammatory processes related to cancer.

Figure 2 shows cancer differences between the groups.

Data: Stress and cancer

From the widely reported study: "C. P. Le et al. Chronic stress in mice remodels lymph vasculature to promote tumour cell dissemination Nature Communications, 7, 2016".

Mouse	Group	VEGFC	COX2
1	Control	0.96718	14.05901
2	Control	0.51940	6.92926
3	Control	0.73276	0.02799
4	Control	0.96008	6.16924
5	Control	1.25964	7.32697
6	Stress	4.05745	6.45443
7	Stress	2.41335	12.95572
8	Stress	1.52595	13.26786
9	Stress	6.07073	55.03024
10	Stress	5.07592	29.92790

2.2 Summaries — 6.1, 6.2

Descriptive statistics

We begin by looking at the the important/useful statistics

Moment statistics – they describe central tendency, spread, skewness, kurtosis in the data

Frequency statistics – pdf, pmf, cdf from the data

Order statistics – quantiles

Looking at the data:

```
(x <- round(VEGFC, digit = 2))

##  [1] 0.97 0.52 0.73 0.96 1.26 4.06 2.41 1.53 6.07
## [10] 5.08

(y <- sort(x))

##  [1] 0.52 0.73 0.96 0.97 1.26 1.53 2.41 4.06 5.08
## [10] 6.07
```

Some basic summaries:

```
summary(x)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5200  0.9625  1.3950  2.3590  3.6475  6.0700

var(x)

## [1] 3.98761

IQR(x)

## [1] 2.685
```

$$
\begin{aligned}
&\text{Sample mean} = \overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{23.59}{10} = 2.3590\\
&\text{Sample median} = \text{middle observation} = \hat{\pi}_{0.5} = \frac{1.26 + 1.53}{2} = 1.3950\\
&\text{Sample variance} = s^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 = 3.98761\\
&\text{Sample st dev} = s = \sqrt{3.98761} = 1.9969\\
&\text{1st quartile} = \hat{\pi}_{0.25} = y_{3.25} = 0.9625\\
&\text{3st quartile} = \hat{\pi}_{0.75} = y_{7.75} = 3.6480\\
&\text{interquartile range} = \hat{\pi}_{0.75} - \hat{\pi}_{0.25} = 2.685
\end{aligned}
$$

Definition: $\hat{\pi}_p$ is an average of two consecutive ordered values y with weights determined by p and the choice of a method - see R documentation on quantiles for the details Note: $\hat{\pi}_{0.25}$ and $\hat{\pi}_{0.75}$ contain about 50% of the sample between them

2.3 Order Statistics — 6.3

Order statistics

Arrange the sample $x_1, \ldots, x_n$ in order of increasing magnitude and define

$$y_1 \leq y_2 \leq \ldots \leq y_n$$

(y_1 and y_n are the sample minimum and maximum, respectively). Then y_k is the called the *kth order statistic.*

For the above sample $y_1 = 0.52$, $y_2 = 0.73, \cdots, y_{10} = 6.07$. So what is $y_{3.25}$? We define $y_{3.25}$ to be 0.25 of the way from y_3 to y_4. Thus,

$$y_{3.25} = y_3 + 0.25(y_4 - y_3) = 0.96 + (0.97 - 0.96) \times 0.25 = 0.9625$$

Check that $y_{7.75} = 3.6480$.

2.4 Boxplots — 6.2

Boxplot for VEGFC

Convenient way of graphically depicting realisations corresponding to one or multiple groups. Main box: $\hat{\pi}_{0.25}, \hat{\pi}_{0.5}, \hat{\pi}_{0.75}$ Whiskers: $\hat{\pi}_{0.25} - k \times IQR, \hat{\pi}_{0.75} + k \times IQR$ (typically $k = 1.5$)

Figure 3 shows all mice, whilst 4 separates the groups.

2.5 Empirical cdf, pmf — histograms and smoothed histograms

Empirical cdf

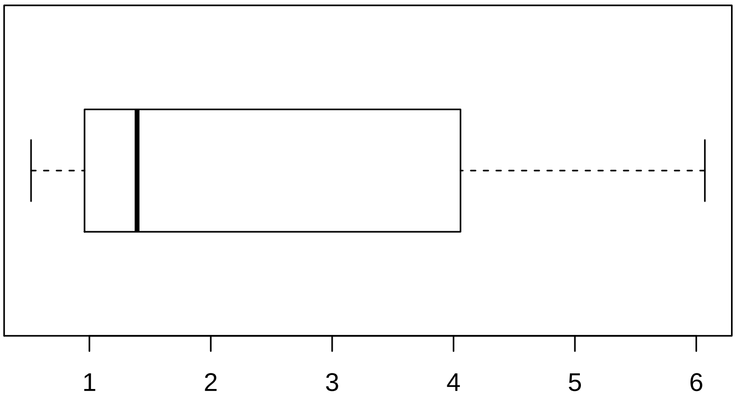

Figure 3: Boxplot for VEGFC for all mice

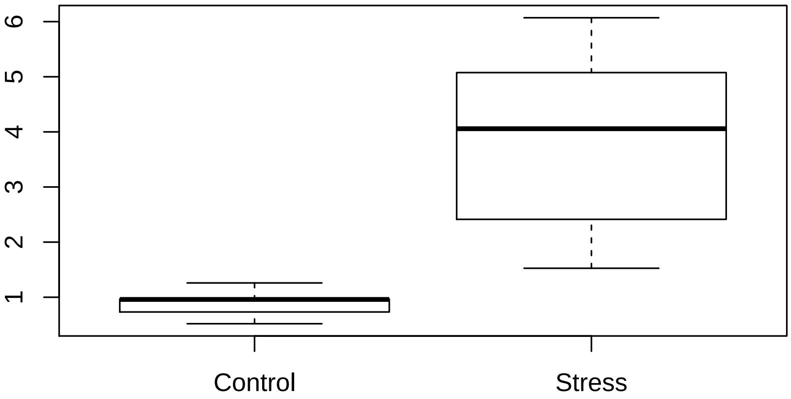

Figure 4: Boxplots for VEGFC in Groups

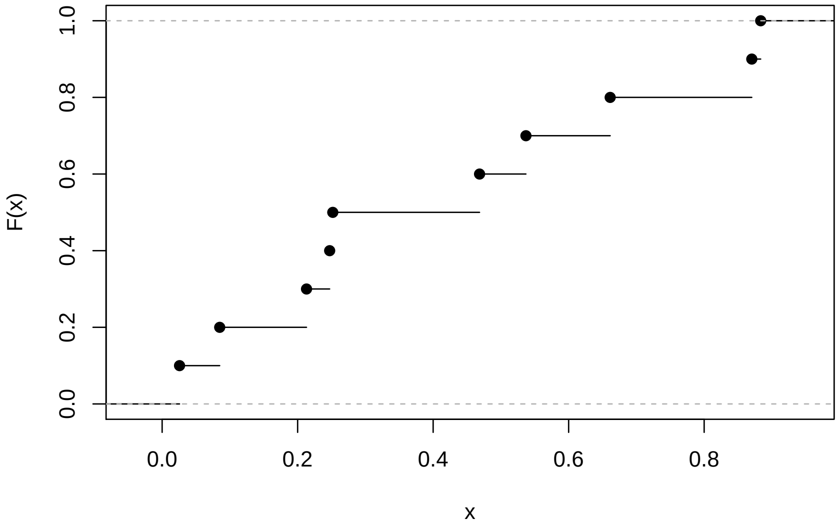

Figure 5: Empirical Distribution Function for 10 Uniform RV's on [0,1]

The sample cdf (empirical cdf) is defined as

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n},$$

where $I(\cdot)$ is the indicator function ($I(x_i \leq x) = 1$ if $x_i \leq x$ and $I(x_i \leq x) = 0$ if $x_i > x$). For example, in our data

$$\hat{F}(2) = \frac{\sum_{i=1}^{10} I(x_i \leq 2)}{10} = \frac{1}{10} \times 6.$$

Figure 5 shows an emperical distribution function for 10 random numbers uniformly distributed on [0,1].

The empirical cdf is a discrete cdf since it only increases at jumps. However, it will approximate the cdf of a continuous variable if the sample size is large. Figure 6shows cdfs based on $n = 50$ and $n = 200$ observations sampled from a standard normal distribution $N(0, 1)$.

TIf the underlying variable is discrete we use the pmf corresponding to the sample cdf $\hat{F}$

$$\hat{p}(x) = \frac{\sum_{i=1}^{n} I(x_i = x)}{n}$$

Figure 7 shows $\hat{p}(x)$ of size $n = 15$ from $Pois(5)$ (left) and the true pmf $p(x)$ of $Pois(5)$ below it.

Histograms and smoothed pdfs

However, if the underlying variable is continuous we would prefer to obtain an approximation of the pdf. There are several approaches that can be used:

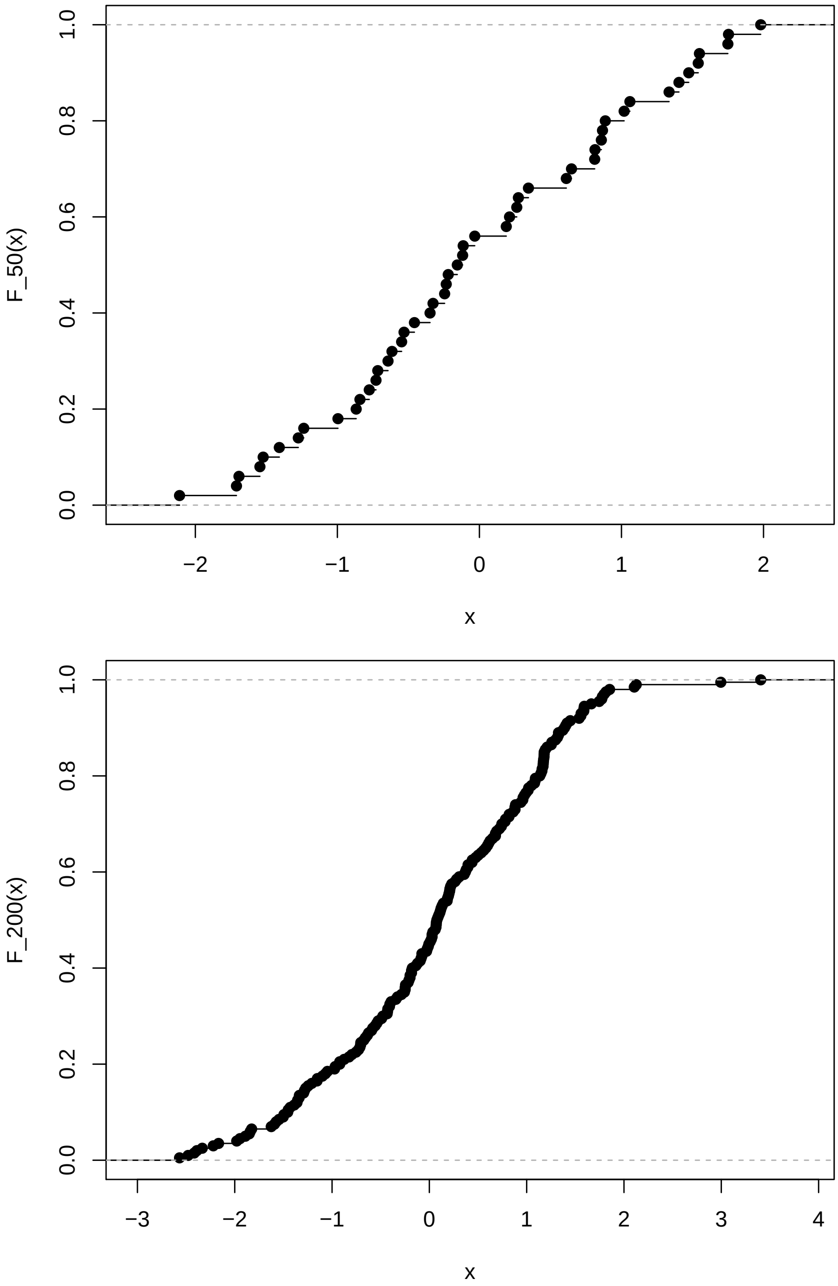

Figure 6: EDF's for 50 and 200 standard normal rv's

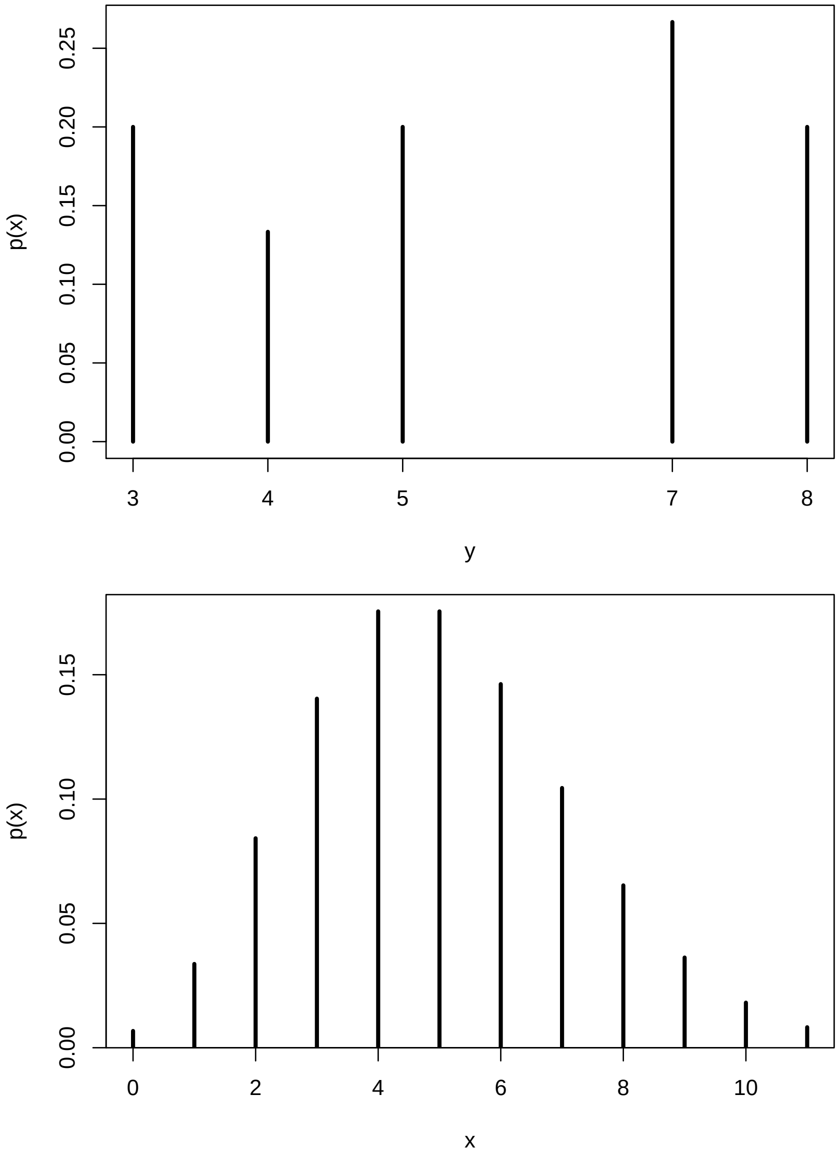

Figure 7: Empirical PMF for 15 random Poisson(5) rv's and PMF for Poisson(5)

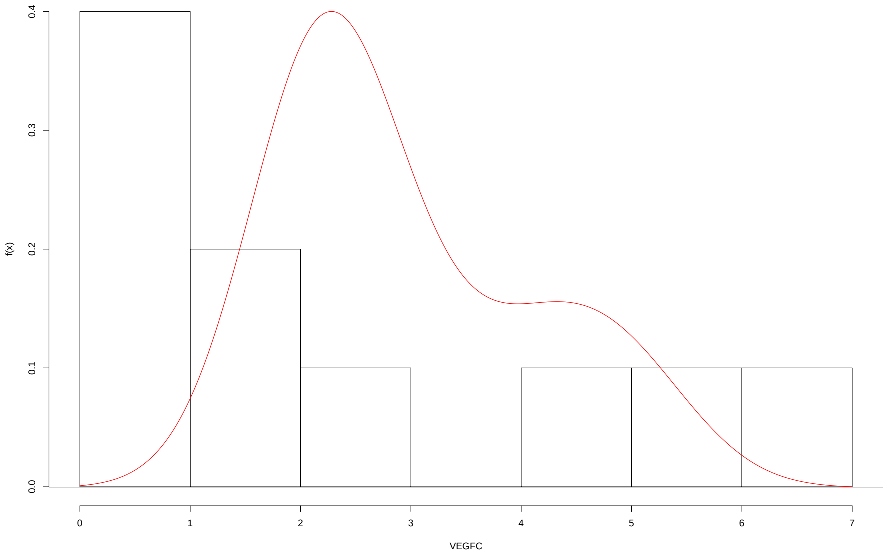

Figure 8: Histogram with smoothed pdf for the VEFGC data

1. Binned histogram, $\hat{f}_h$ (h is the bin length). First divide the entire range of values into a series of small intervals (bins) and then count how many values fall into each interval. A rectangle in the interval (a, b) $(b - a = h)$ has height
$$\hat{f}_h(x) = \frac{\sum_{i=1}^n I(a \leq x_i \leq b)/n}{h}$$

2. Smoothed pdf, $\hat{f}_h$ (h is the so-called bandwidth parameter)
$$\hat{f}_h(x) = \frac{1}{hn}\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right),$$

where $K(\cdot)$ is the kernel (a non-negative function that integrates to 1 and with mean zero) and h is a smoothing parameter

Figure 8 shows a histogram for the VEGFC data overlaid with a smoothed density plot.

Consider $n = 100$ observations from the Weibull distribution with pdf

$$f(x) = \frac{1}{2}xe^{-(x/2)^2}, x > 0.$$

The histogram, true density (solid black curve and smoothed pdf (red dashed curve) are shown in Figure 9.

Sample cdf and Quantile Quantile (QQ) plots

Graphical method for comparing the similarly of two probability distributions by plotting their quantiles (percentiles) against each other. The points in the plot are
$$\left\{y_k, F^{-1}\left(\frac{k}{n+1}\right)\right\}, \quad k = 1, \ldots, n.$$

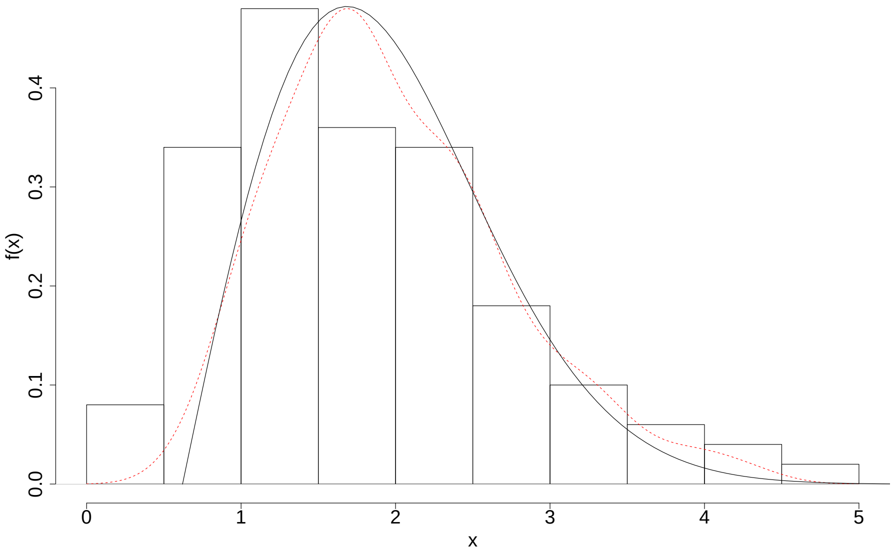

Figure 9: Weibull simulated data histogram, smoothed (red) and true (black) densities.

One axis shows the sample order statistics (that generate quantiles and percentiles)

$$y_k = \hat{\pi}_p, \text{ where } p = k/(n+1) \text{ one definition}$$

for $k = 1, \dots, n$. The other axis shows corresponding quantiles for a theoretical distribution:

$$F^{-1}\left(\frac{k}{n+1}\right)$$

Example: VEGFC

Is the sample from an exponential distribution with cdf $F(x) = 1 - e^{-\lambda x}$. Quantiles can be computed by inverting F:

$$F^{-1}(p) = -\ln(1-p)/\lambda \;\; \text{(e.g., set } \lambda = 0.5)$$

Sample quantiles

$$y_1 = 0.97, y_2 = 0.52, \dots, y_{10} = 5.08$$

The theoretical quantiles are obtained as

$$1/(10+1) = 0.09, 2/(10+1) = 0.18, \dots, 10/(10+1) = 0.91$$

$$F^{-1}(0.09) = 0.19, F^{-1}(0.18) = 0.40, \dots, F^{-1}(0.91) = 4.80$$

Figures 10 and 11 show the VEGFC data and exponential quantiles and density.

The right tail of the sample in Figures 10, 11 do not quite match the theoretical model, since the tail of the sample distribution looks heavier.

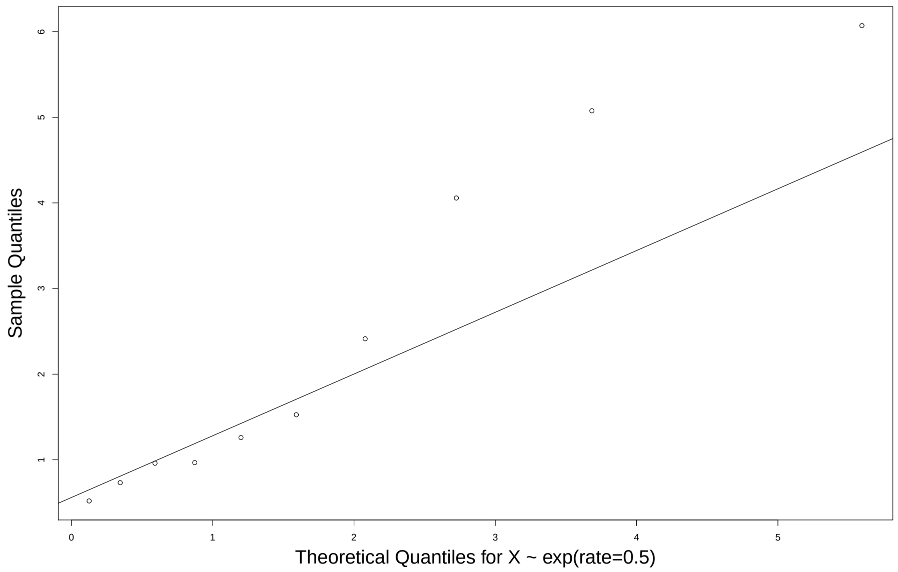

Figure 10: QQplot for VEGFC vs. Theoretical Quantiles for Exponential(0.5)

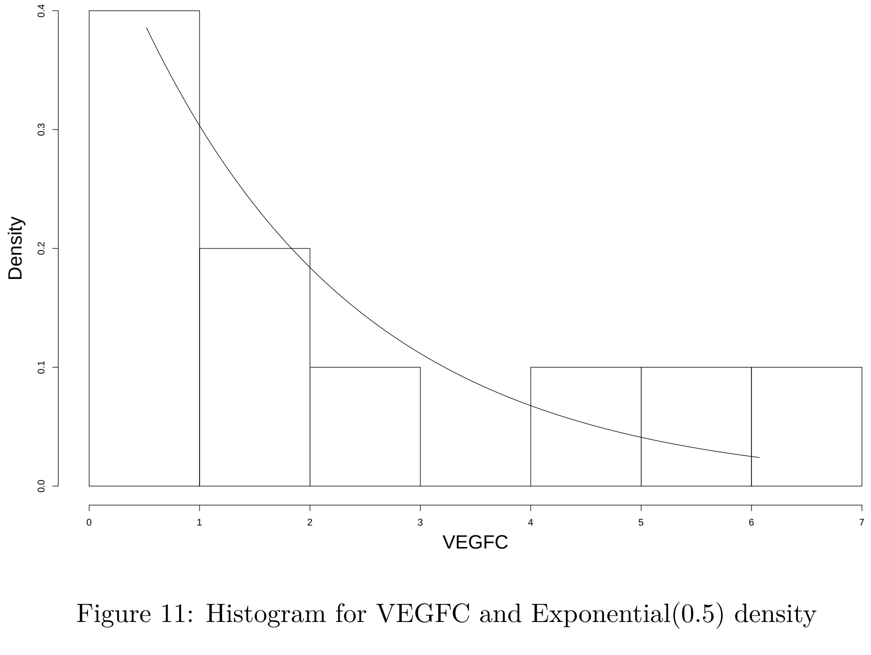

Figure 11: Histogram for VEGFC and Exponential(0.5) density

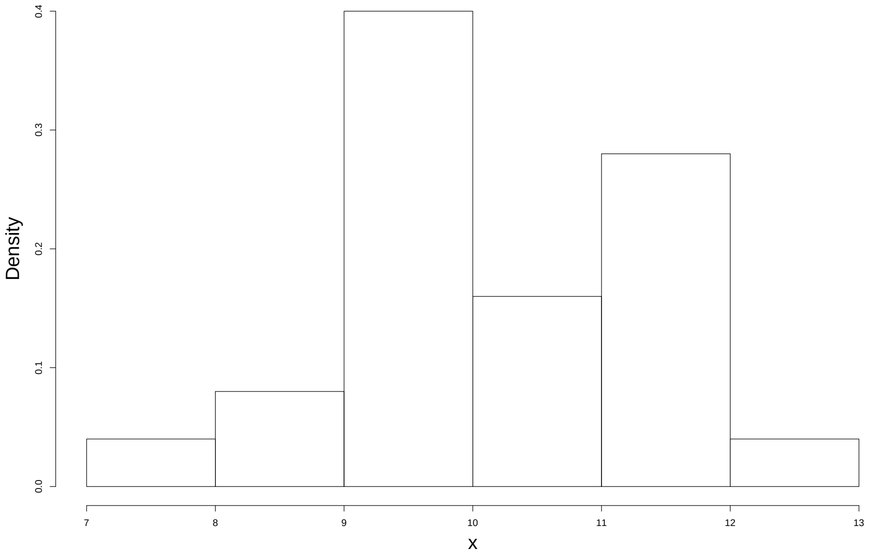

Figure 12: Histogram for 25 N(10,2) random numbers

Normal QQ plots

If $X \sim N(\mu, \sigma^2)$, then $X = \mu + \sigma Z$, where $Z \sim N(0,1)$. Therefore, if the normal model is correct

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

where $\Phi(z) = P(Z \leq z)$ is the standard normal cdf. So, if we plot the points

$$\left(y_k, \Phi^{-1}\left(\frac{k}{n+1}\right)\right), \quad k = 1, \ldots, n$$

the result should be a straight line with intercept μ and slope σ. The values $\Phi^{-1}(k/(n+1))$ are called normal scores.

Consider 25 observations from $X \sim N(10,2)$. The histogram in Figure 12 is not very helpful, whereas the qqplot in Figure 13 shows a good linear fit. Note that the departures at the extreme are typical.

3 Maximum Likelihood and Method of Moments — 6.4

3.1 Point estimation

Distributions of statistics

Consider sampling from $X \sim Exp(\lambda = 1/5)$. Typically λ is unknown but its value might be important practically. Consider various samples of size $n = 100$. The following R command produces 250 samples each of size 100 :

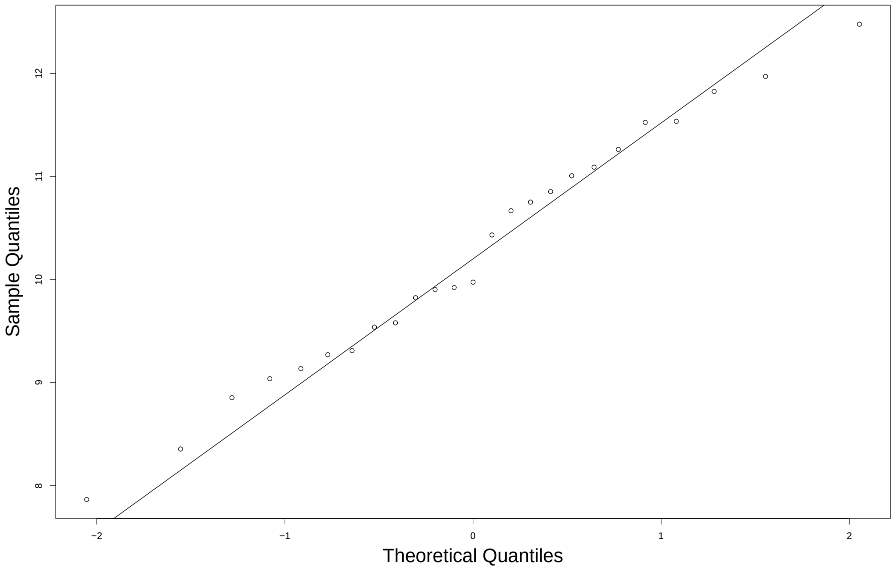

Figure 13: Qqplot for 25 N(10,2) random numbers

```
samples <- lapply(1:250, function(x) rexp(100, rate = 0.2))
```

The command `lapply` produces a list of length the same as the first argument.

It applies the function in the second argument to each member of the list in the first argument.

In this case, the function does not depend on its argument, producing 100 random exponential numbers of mean 5 each time it is called.

Distributions of statistics

Some information on the distribution of various statistcs from the samples can be obtained by using the R function `summary` in `lapply`, with `digits = 3` printing to 3 significant digits. The 2nd command prints out the summaries for the first four samples:

```
summaries <- lapply(samples, function(x) summary(x,
    digits = 3))
rbind(summaries[[1]], summaries[[2]], summaries[[3]],
    summaries[[4]])

##         Min. 1st Qu. Median Mean 3rd Qu. Max.
## [1,] 0.0344    1.60   3.60 5.09    7.88 19.8
## [2,] 0.1960    1.69   3.66 4.52    6.48 20.0
## [3,] 0.0364    1.89   4.12 5.01    6.33 28.6
## [4,] 0.0478    1.52   3.16 4.70    6.86 26.3
```

Distributions of statistics

```r
# par sets up graphical parameters mfrow gives the
# configuration of graphs 1 row and 2 columns here
# cex = 2.5 makes the text 2.5 times larger than
# normal
par(mfrow = c(1, 2), cex = 2.5)
# sapply operates on the list samples using the
# function mean and produces a vector rather than
# a list
x.bar <- sapply(samples, mean)
s.2 <- sapply(samples, var)
hist(x.bar, breaks = seq(from = 3.25, to = 7.25, by = 0.5))
abline(v = 5, lty = 2, col = "red")
hist(s.2, breaks = seq(from = 7.5, to = 77.5, by = 5))
abline(v = 25, lty = 2, col = "red")
```

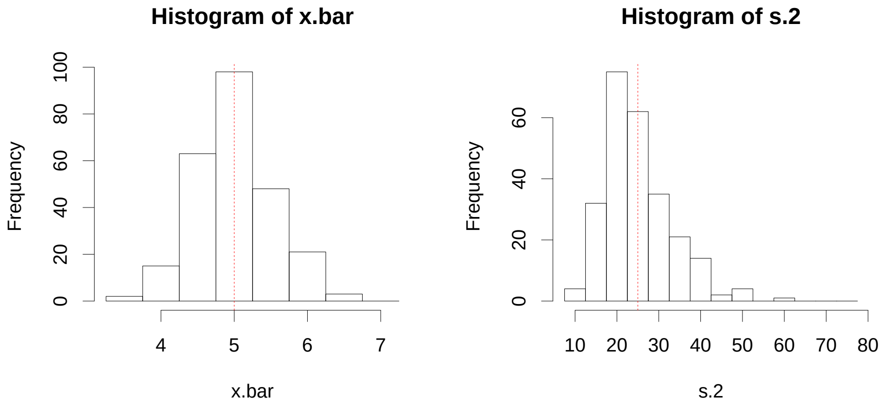

Figure 14: Vertical red, dashed lines are true $E(X) = 5$ and $Var(X) = 5^2$.

Recall that the statistic $T = \phi(X_1, \ldots, X_n)$ is a random variable and therefore has its own distribution.

The list of `summaries` contains 250 realisations of statistics $t = \phi(x_1, \ldots, x_n)$ where ϕ ranges over the min, max, mean and quartile statistics .

To use T for estimation, its distribution, including its mean and variance, is helpful.

For example, we know $E(\overline{X}) = \mu$ & $E(S^2) \approx \sigma^2$.

The following R commands compute the 250 realisations of $\bar{X}, S^2$ and show the histograms in Figure 14.

R commands for histograms of $\bar{X}, S^2$

The Law of Large Numbers tell us that sample mean $(\overline{X})$ estimates the population mean $(E(X) = 1/\lambda)$ and the sample standard deviation (S) estimates the population standard deviation ($SD(X) = \sqrt{Var(X)} = 1/\lambda$) for large sample sizes.

Suppose we are interested in estimating the parameter $\theta = 1/\lambda$. Should we use $\overline{X}$ or S? Which one is the better estimator?

Consider $B = 250$ samples of size $n = 100$ and compute

$$\overline{x}_1, \dots, \overline{x}_{250},$$

$$s_1, \dots, s_{250}$$

The chances for one estimator being closer to the true value $\theta = 1/\lambda = 5$ can be found out from simulations of repeated samples.

```
summary(x.bar)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.709   4.667   4.972   5.008   5.305   6.370

sd(x.bar)

## [1] 0.5316551

s <- sqrt(s.2)
summary(s)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.166   4.490   4.803   4.950   5.382   7.905
```

In this case, judging from the distributions of the estimators above $\overline{X}$ is superior to S.

Point Estimation

Sample mean $\bar{x}$ is an estimate of the *population* or *distribution* mean μ.

Sample standard deviation s is an estimate of the *distribution* standard deviation σ.

Sample histogram is an estimate of the pdf (or pmf) of the *distribution*.

Sample mean and standard deviation close to the *distribution* parameter most or all of the time?

Other estimates?

Point Estimation

Suppose the functional form of the pmf or pdf is known.

But this depends on a parameter $\theta \in \Theta$, the parameter space.

Example: exponential density

$$f(x; \theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0, \quad \theta \in \{\theta : 0 < \theta < \infty\}$$

Often wish to know the most "likely" pdf within this family based on a random sample $X_1, \dots, X_n$.

Recall a random sample is a collection of independent observations on the same distribution. That is, $X_1, \dots, X_n$ are independently and identically distributed (*iid*).

Point Estimation

Several procedures exist to find estimators.

An *estimator* is a random variable $u(X_1, \dots, X_n)$ that is a function u of the model rv's $X_1, X_2, \dots, X_n$.

An *estimator* has a pdf or pmf as a random variable.

An *estimate* is the observed value of the estimator $u(x_1, \dots, x_n)$, where $x_1, \dots, x_n$ is the observed data.

An *estimate* is a number (or might be a vector if there are several parameters being estimated).

How to choose amongst estimators?

Prob'ly statements about $\hat{\theta} = u(X_1, \dots, X_n)$ allow us to evaluate the estimator's accuracy in repeated samples — this the *frequentist* method of choosing between estimators.

EG:

1. Is $E(\hat{\theta}) = \theta$?
2. Is $Var(\hat{\theta})$ small?
3. Or is $P(\mid \hat{\theta} - \theta \mid < \epsilon)$ large? where ϵ is a desired closeness of our estimate to the population parameter

Start with a view — expressed through a pdf or pmf — about likely values of $\theta \in \Theta$. How does the data alter the assessment of the likely or expected value of θ ? - the *Bayesian* method of choosing between estimators?

Maximum likelihood estimation is a *frequentist* method for choosing an estimator — choose θ to maximise pdf or pmf.

Maximum likelihood estimation?

Maximum likelihood estimation is a *frequentist* method for choosing an estimator.

In probability, the focus was on calculating probabilities, expectations etc for random variables.

For discrete rv's the pmf *is* a probability but for continuous rv's the pdf *integrates* to give probabilities.

Either way, for independent rv's, the product of the pmf's or pdf's for our data gives is the "likelihood" of being at, or near, the data.

So one general possibility is to choose θ to *maximise* this likelihood.

Called maximum likelihood estimation.

In statistics, this likelihood focuses on the joint pmf or pdf as a function of the *parameter* values first.

3.2 Bernoulli rvs

Example: Maximum Likelihood for Bernoulli rv's

Suppose our population only has zero's and one's and that our model is a Bernoulli pmf for a randomly chosen member of the population (that is, the same probability of being one for all members of the population).

Then the probability mass function is

$$f(x;p) = p^x(1-p)^{1-x}, \quad x = 0,1 \quad 0 \le p \le 1$$

Observed values are $x_1, \cdots, x_n$ of $X_1, \cdots, X_n$ independent and identically distributed Bernoulli rv's

Using independence, the probability that the random sample was exactly the x's observed in the order that they were observed is

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

Example: Maximum Likelihood for Bernoulli rv's

Regard the sample $x_1, \cdots, x_n$ as known (we have in fact observed this) and regard this probability as a function of p.

Called the *likelihood* of p:

$$L(p) = L(p; x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i;p) = p^{\sum x_i}(1-p)^{n-\sum x_i}.$$

May then find the value of p that maximizes this likelihood, that is:

Find the parameter value p that maximizes the probability of what was observed.

Solution: Maximum Likelihood for Bernoulli rv's

Often helps to find the value of θ that maximizes the log of the likelihood rather than the likelihood.

Makes no difference because log of non-negative numbers is a one to one function whose inverse is the exponential, so any value θ that maximises the log-likelihood also maximises the likelihood and vice-versa.

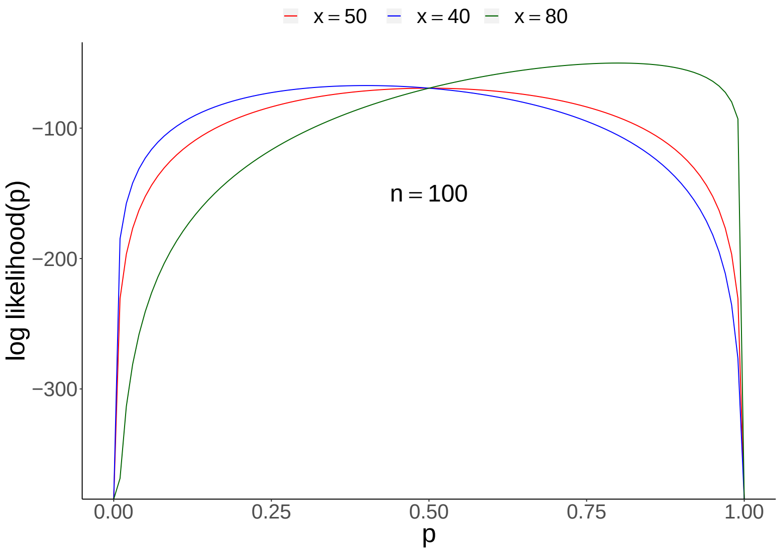

Figure 15: Log Likelihoods for Bernoulli trials for parameter p

Putting $x = \sum_{i=1}^{n} x_i$ so that x is the number of 1's in the sample,

$$\ln L(p) = x \ln p + (n - x) \ln(1 - p).$$

Find maximum of this log-likelihood with respect to p by differentiating and equating to zero

$$\frac{\partial \ln L(p)}{\partial p} = x\frac{1}{p} + (n - x)\frac{-1}{1 - p} = 0. \tag{1}$$

Finding the Bernoulli MLE - Odds are Useful

Define for $0 \leq p < 1$ the odds, o, corresponding to p by

$$o = \frac{p}{1 - p}. \tag{2}$$

Solving for p in terms of o gives

$$p = \frac{o}{1 + o}. \tag{3}$$

Odds are often talked about as "on" or "against" meaning that the ratio is for p or $1 - p$

With odds o against a win in a game of chance (or horse race or ...), the fair payout for paying \$1 to play the game is \$$(1 + o)$ because the expected profit is then $-1 * \frac{o}{1+o} + o * \frac{1}{1+o} = 0$.

Bernoulli MLE $\hat{p} = \bar{X}$.

Solve equation (1) by rewriting it as

$$\frac{p}{1-p} = \frac{x}{n-x}$$

So we can solve for p using equation (3) giving

$$p = \frac{\frac{x}{n-x}}{1 + \frac{x}{n-x}} = \bar{x}$$

Differentiating the log likelihood a second time gives

$$\frac{\partial^2 \ln L(p)}{\partial p^2} = -x\frac{1}{p^2} - (n-x)\frac{1}{(1-p)^2}$$

Is negative so the solution is a maximum.

3.3 General Description of MLE

Point Estimation: Maximum Likelihood

General procedure for random samples $X_1, \cdots, X_n$.

Likelihood function with m parameters $\theta_1, \cdots, \theta_m$ and data $x_1, \cdots, x_n$ is

$$L(\theta_1, \cdots, \theta_m) = \prod_{i=1}^{n} f(x_i; \theta_1, \cdots, \theta_m).$$

MLEs - *Maximum likelihood Estimators or Estimates* - $\hat{\theta}_1, \cdots, \hat{\theta}_m$ are values that maximize $L(\theta_1, \cdots, \theta_m)$.

MLEs are the same *function* of the data - the *estimates* - or random variables - *estimators.*

Often (but not always) useful to take logs and then differentiate and equate derivatives to zero to find MLE's.

Computation usually uses the data values rather than the random variables - if estimator, write in the rv's in capital letters

.

3.4 MLE for Exponential

Example — MLE for Exponential

Suppose $X_1, \cdots, X_n$ i.i.d.

$$f(x; \theta) = \frac{1}{\theta}e^{-x/\theta}, \quad x > 0, \quad \theta \in \{\theta : 0 < \theta < \infty\}$$

$$L(\theta) = \frac{1}{\theta^n} \exp\left(\frac{-\sum_{i=1}^n x_i}{\theta} \right)$$

$$\ln L(\theta) = -n \ln(\theta) - \frac{1}{\theta} \sum_{i=1}^n x_i$$

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} = 0$$

Yields $\hat{\theta} = \bar{X}$.

Solution — MLE for Exponential

This can be seen using the following inequality for all $x > 0$

$$f(x) \equiv \ln(x) - x \leq -1 = \ln(1) - 1 = f(1) \tag{4}$$

so that the value x which minimises $f(x)$ is $x = 1$

Written

$$\arg\max_x \left(f(x) \equiv \ln(x) - x \right) = 1 \tag{5}$$

Proof of (4) uses $f'(x) = \frac{1}{x} - 1, f''(x) = -\frac{1}{x^2} < 0$, so the only solution $x = 1$ to the equation $f'(x) = 0$ is a maximum

Rewrite $\ln L(\theta) = n \ln(\frac{1}{\theta}) - \frac{\sum_{i=1}^n x_i}{\theta}$ as $\ln L(\theta) = n\left(\ln(\frac{\bar{x}}{\theta}) - \frac{\bar{x}}{\theta} - \ln(\bar{x})\right) = f\left(\frac{\bar{x}}{\theta}\right) - \ln(\bar{x})$ and equation (5) shows that $\arg\max_\theta \left(\ln L(\theta)\right)$ satisfies $\frac{\bar{x}}{\theta} = 1$

3.5 MLE for Normal

Example — MLE for Normal

Suppose there is a random sample of size n from a population which is normally distributed with mean μ and variance σ^2. What are the maximum likelihood estimators of μ and σ^2 ?

Solution — MLE for Normal

Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$

Then the likelihood as a function of the parameters μ, σ^2 with data $x_1, \cdots, x_n$ is

$$\ln L(\mu, \sigma^2) = \frac{n}{2} \left(-\ln(2\pi) + \ln\left(\frac{1}{\sigma^2}\right) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{n\sigma^2} \right)$$

Rewrite using the Analysis of Variance identity, equation (4), in Module 5 as

$$\ln L(\mu, \sigma^2) = \frac{n}{2} \left(-\ln(2\pi) + \ln\left(\frac{1}{\sigma^2}\right) - \frac{v + (\bar{x} - \mu)^2}{\sigma^2} \right) \tag{6}$$

where

$$v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Solution — MLE for Normal Ctd

Hence, for any values of μ and σ^2, from (6) and the fact that the minimum value of $(\bar{x}-\mu)^2 = 0$ if $\bar{x} = \mu$,

$$\begin{aligned} \ln L(\mu,\sigma^2) &\leq \frac{n}{2}\left(-\ln(2\pi) + \ln\left(\frac{v}{\sigma^2}\right) - \ln(v) - \frac{v}{\sigma^2}\right) \\ &= \frac{n}{2}\left(f\left(\frac{v}{\sigma^2}\right) - \ln(2\pi) - \ln(v)\right) \\ &= \frac{n}{2}\left(-1 - \ln(2\pi) - \ln(v)\right) \qquad (7) \end{aligned}$$

where $f(x) = \ln(x) - x$ with maximum -1 for $x = 1$ (see MLE for exponential above.)

So

$$\arg\max_{(\mu,\sigma^2)} \left(\ln L(\mu,\sigma^2)\right) = (\bar{x}, v)$$

since the second two terms on the right of equation (7) depend only on data and the right side of (7) is $\ln L(\bar{x}, v)$ from equation (6).

3.6 Unbiased Estimators

Comment — MLE for Normal

ML Estimators for normal mean and variance are the corresponding rv's $\bar{X}$ and V where the data values $x_1, \cdots, x_n$ are replaced by the rv's $X_1, ..., X_n$.

In Module 5 it was shown that for iid $N(\mu,\sigma^2)$ rv's $X_1, \cdots, X_n$ the sample variance has the desirable property that

$$E(S^2) = \sigma^2. \qquad (8)$$

This property that the expected value of the estimator (in this case, S^2) is the parameter (in this case, σ^2) is called *unbiased.*

MLEs are often *biased*, as in this case, since $E(V) = E(\frac{n-1}{n}S^2) = \frac{(n-1)\sigma^2}{n} \neq \sigma^2$.

For small samples, the *unbiased* sample variance, S^2should be used over the biased MLE V because the two estimators differ only by a constant ($S^2 = \frac{n}{n-1}V$).

3.7 Method of Moments

Method of Moments

Often the maximum likelihood estimators can't be worked out by formulas and it is necessary to find maximum likelihood estimates by using a computer to maximise the likelihood function for each data set

Textbook p. 270 shows that this is the case for Gamma rv's where both the index and the scale parameter are to be estimated

A simple alternative to find estimators is to use enough moments so that t equating sample moments to population moments produces enough equations to find estimators for all the parameters

Called *Method of Moments*

Example — Method of Moments

Suppose $X_1, \cdots, X_n$ are iid Gamma (α, θ) so they have density

$$f(x; \alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} x^{\alpha-1} \exp\left(\frac{-x}{\theta}\right)$$

Find the method of moments estimators of α and θ.

Solution — Method of Moments

From Module 3, p.14, equation (22) the mean, μ, and variance, σ^2, for the Gamma distribution are:

$$\mu = \alpha\theta; \quad \sigma^2 = \alpha\theta^2$$

Method of moments says to equate these to the corresponding sample random variables $\bar{X}$ and V giving the estimating equations:

$$\alpha\theta = \bar{X}; \quad \alpha\theta^2 = V$$

Dividing the second equation by the first gives

$$\theta = \frac{V}{\bar{X}}$$

Solution — Method of Moments Ctd

Substituting this in the first equation gives

$$\alpha = \frac{\bar{X}}{\theta} = \frac{\bar{X}^2}{V}$$

Could use S^2 instead of V knowing that S^2 is unbiased (as is $\bar{X}$)

4 Simple Regression — 6.5

4.1 Setup

Simple Regression — Setup

Response Y such as Prostaglandin-endoperoxide synthase 2 (COX2), the protein involved in inflammatory processes related to cancer

Predictor x such as Vascular endothelial growth factor C (VEGFC) - a protein involved in lymphangiogenesis - low or high levels of this are observed in many diseases

Expect that Y depends on x.

Means $E(Y \mid x) = \mu(x)$ for some function μ of x.

Simplest linear model is $E(Y \mid x) = \alpha_1 + \beta x$.

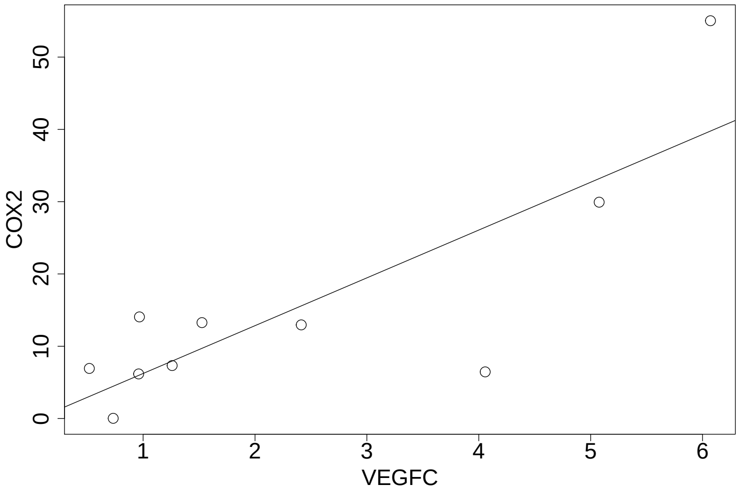

Figure 16: COX2 vs VEGFC with MLE Estimated Line

Observe independent pairs $(x_1, y_1), \cdots, (x_n, y_n)$.

So

$$Y_i = \alpha_1 + \beta x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ is a random error.

That is, given x, $Y \sim N(\alpha_1 + \beta x, \sigma^2)$

Simple Regression — Where to

Want to estimate the slope (β), the intercept (α), the variance of the errors (σ^2) and their standard errors, that is the standard deviations of the estimators.

In Module 7, confidence intervals for the estimates will be calculated.

Often the question is whether there is a relationship between the potential predictor x and the response y. Module 8 studies this, by testing the hypothesis that $\beta = 0$.

Often predictioons about future observations are needed. Module 7 considers this.

The first step is to establish the maximum likelihood estimates and their properties.

4.2 Maximum Likelihood Estimation

Simple Regression

Write $\alpha_1 = \alpha - \beta\bar{x}$ (*makes calculations easier*) so

$$\begin{aligned} Y_i &= \alpha - \beta\bar{x} + \beta x_i + \epsilon_i \\ &= \alpha + \beta(x_i - \bar{x}) + \epsilon_i \end{aligned}$$

Then $Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$

As the Y_i's are independent, the likelihood is

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

Simple Regression via Derivatives

MLEs Maximise the likelihood or alternatively minimise the negative:

$$-\ln L(\alpha, \beta, \sigma^2) = \frac{n}{2}\ln(2\pi\sigma^2) + \frac{\sum_{i=1}^{n}[Y_i - \alpha - \beta(x_i - \bar{x})]^2}{2\sigma^2}$$

Hence choose α and β to minimize the sum of squares

$$H(\alpha, \beta) = \sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \bar{x})]^2$$

Or solve

$$\begin{aligned} 0 &= \frac{\partial H(\alpha, \beta)}{\partial \alpha} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \bar{x})](-1) \\ 0 &= \frac{\partial H(\alpha, \beta)}{\partial \beta} = 2\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \bar{x})](-(x_i - \bar{x})) \end{aligned}$$

Simple Regression via Derivatives

Some algebra yields the *least square estimators* which are also the *MLEs*

$$\hat{\alpha} = \bar{Y}, \quad \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{9}$$

(*Matrix representations make this easier as will be seen in MAST90104.*)

To estimate σ^2, solve

$$0 = \frac{\partial H(\alpha, \beta, \sigma^2)}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{\sum_{i=1}^{n}[y_i - \alpha - \beta(x_i - \bar{x})]^2}{2(\sigma^2)^2}$$

which yields

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2.$$

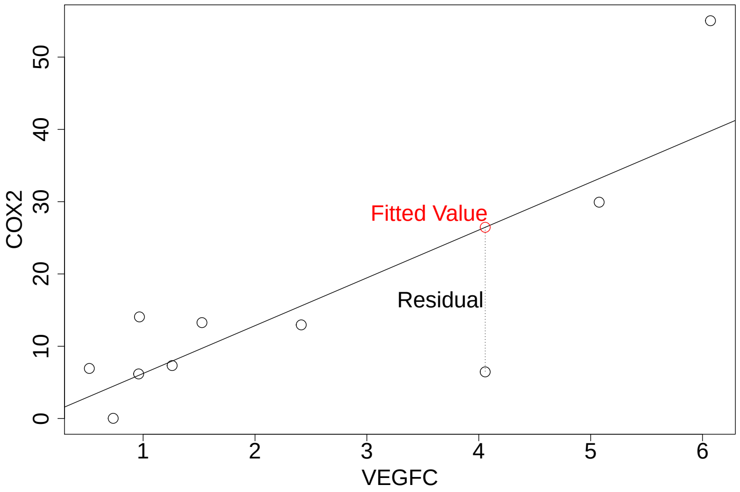

Figure 17: COX2 vs, VEGFC with Residual and Fitted illustrated

4.3 Establishing the derivative roots are MLEs.

Simpler Way — MLEs Simple Regression

Use same approach as MLEs for Normal Distribution with constant population mean and variance.

Key is a new Analysis of Variance identity:

$$\sum_{i=1}^{n}(Y_i-(\alpha+\beta(x_i-\bar{x}))^2=\sum_{i=1}^{n}\left(R_i^2+((\hat{\alpha}-\alpha)+(\hat{\beta}-\beta)(x_i-\bar{x}))^2\right) \quad (10)$$

where - see the next slide for illustration on the Stress and Cancer data -

1. the *fitted value* is $\hat{Y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$ - the rv representing the point on the (least squares) line
2. the *residual* is $R_i = Y_i - \hat{Y}_i$ - the vertical distance between the y-rv and the fitted value

Key Properties of Residuals and Fitted Values

Residuals add to zero, because the definitions of $\hat{\alpha}$ and $\hat{\beta}$ show

$$\begin{aligned}\sum_{i=1}^{n} R_i &= \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})) \\ &= \sum_{i=1}^{n}(Y_i - \bar{Y}) - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= 0 \end{aligned} \quad (11)$$

since the sum of deviations around the average is zero for both the x- and Y-values.
Hence for any line $l(x) = \alpha + \beta(x - \bar{x})$, the residuals are uncorrelated with the x-values on the line, that is

$$\sum_{i=1}^{n} R_i l(x_i) = 0. \tag{12}$$

Proof of equation (12)
Because the residuals add to zero - equation (11) -

$$\begin{aligned}
\sum_{i=1}^{n} R_i l(x_i) &= \alpha \sum_{i=1}^{n} R_i + \beta \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})) \\
&= \beta \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})) \\
&= \beta \left(\sum_{i=1}^{n} Y_i (x_i - \bar{x}) - \hat{\alpha} \sum_{i=1}^{n} (x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \\
&= \beta \left(\sum_{i=1}^{n} Y_i (x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) = 0
\end{aligned}$$

using the definition of $\hat{\beta}$ in (9).

Proof of Analysis of Variance Identity (10)
Using equation (12) twice in the last step (once with the fitted line)

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - (\alpha + \beta(x_i - \bar{x}))^2 &= \sum_{i=1}^{n} \left(R_i + \hat{Y}_i - l(x_i) \right)^2 \\
&= \sum_{i=1}^{n} \left(R_i^2 + 2R_i(\hat{Y}_i - l(x_i)) + (\hat{Y}_i - l(x_i))^2 \right) \\
&= \sum_{i=1}^{n} \left(R_i^2 + ((\hat{\alpha} - \alpha) - (\hat{\beta} - \beta)(x_i - \bar{x}))^2 \right) \\
&+ 2 \sum_{i=1}^{n} R_i (\hat{Y}_i - l(x_i))) \\
&= \sum_{i=1}^{n} \left(R_i^2 + ((\hat{\alpha} - \alpha) - (\hat{\beta} - \beta)(x_i - \bar{x}))^2 \right).
\end{aligned}$$

MLE's for Regression via Analysis of Variance
So the same argument used for finding the MLE for normal mean and variance applies for regression. The likelihood as a function of the parameters

μ, σ^2 with rv's $Y_1, \cdots, Y_n$ is

$$\ln L(\alpha, \beta, \sigma^2) = \frac{n}{2}\left(\ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{\sum_{i=1}^n (Y_i - (\alpha + \beta(x_i - \bar{x}))^2}{n\sigma^2}\right).$$

Rewriting the log likelihood using the Analysis of Variance identity (10):

$$\ln L(\alpha, \beta, \sigma^2) = \frac{-n}{2}\left(\ln(2\pi\sigma^2) - \frac{nV + \sum_{i=1}^n ((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)(x_i - \bar{x}))^2}{n\sigma^2}\right) \quad (13)$$

where

$$V = \frac{\sum_{i=1}^n (Y_i - (\hat{\alpha} + \hat{\beta}(x_i - \bar{x})))^2}{n}.$$

MLE's for Regression via AoV Ctd

Hence, for any values of α, β and σ^2, from (13) and the fact that the minimum value of $\sum_{i=1}^n ((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)(x_i - \bar{x}))^2 = 0$ if, and only if, $\alpha = \hat{\alpha}, \quad \beta = \hat{\beta}$,

$$\begin{aligned} \ln L(\mu, \sigma^2) &\leq \frac{n}{2}\left(-\ln(2\pi) + \ln\left(\frac{V}{\sigma^2}\right) - \ln(V) - \frac{V}{\sigma^2}\right) \\ &= \frac{n}{2}\left(f\left(\frac{V}{\sigma^2}\right) - \ln(2\pi) - \ln(V)\right) \quad (14) \end{aligned}$$

where $f(x) = \ln(x) - x$ has maximum -1 at $x = 1$.

So

$$\arg \max_{(\alpha, \beta, \sigma^2)} \left(\ln L(\alpha, \beta, \sigma^2)\right) = (\hat{\alpha}, \hat{\beta}, V)$$

since the maximum of f is achieved at 1, the second two terms on the right of equation (14) depend only on data and the right side of (14) is $\ln L(\hat{\alpha}, \hat{\beta}, V)$ from equation (13)

4.4 Properties of MLE's Regression

Distribution of $\hat{\alpha}$

RV's $Y_1, \cdots, Y_n$ are independent normally distributed random variables.

$\hat{\alpha} = \bar{Y}$ is a linear combination of these so, from Module 5 (p. 3), $\hat{\alpha}$ has a normal distribution with

Unbiased $\hat{\alpha}$:

$$E(\hat{\alpha}) = \frac{1}{n}\sum_{i=1}^n E(Y_i) = \frac{1}{n}\sum_{i=1}^n [\alpha + \beta(x_i - \bar{x})] = \alpha$$

Variance goes down with n^{-1} because as a sample mean $\bar{Y}$

$$Var(\hat{\alpha}) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^n Var(Y_i) = \frac{\sigma^2}{n}$$

Expectation of $\hat{\beta}$

Similarly $\hat{\beta}$ is a linear combination of the Y_i so has a normal distribution

Recall $E(Y_i) = \alpha + \beta(x_i - \bar{x})$ and $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Then

$$\begin{aligned} E(\hat{\beta}) &= \frac{\sum_{i=1}^n (x_i - \bar{x}) E(Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + (x_i - \bar{x})\beta)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})\alpha}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \beta}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta \end{aligned}$$

so $\hat{\beta}$ is unbiased because the sum of the deviations of the x's around their mean is 0.

Variance of $\hat{\beta}$

Recall for independent random variables, from Module 5 p.3,

$$Var(\sum_{i=1}^n a_i Y_i) = \sum_{i=1}^n a_i^2 Var(Y_i) \quad \text{so,}$$

$$\begin{aligned} Var(\hat{\beta}) &= Var\left(\sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} Y_i \right) \\ &= \sum_{i=1}^n \left[\frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} \right]^2 Var(Y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

and thus the variance of $\hat{\beta}$ depends on the pattern of the x's but diminishes like n^{-1} for reasonably spaced x's.

Analysis of Variance Extended To Obtain Distributions

Arguing as before that the cross-product term is zero because the deviations from the mean of the x's sum to 0, the Analysis of Variance identity can be improved by splitting the terms for $\hat{\alpha}$ and $\hat{\beta}$ into separate sums of squares:

$$\begin{aligned} &\sum_{i=1}^n [Y_i - \alpha - \beta(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2 \\ &+ n(\hat{\alpha} - \alpha)^2 + \sum_{i=1}^n (x_i - \bar{x})^2 (\hat{\beta} - \beta)^2. \qquad (15) \end{aligned}$$

Distribution of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$

Similarly to the Key Facts for Sample Mean and Variance (Module 5 p. 4 and 5) $Y_i \sim N(\alpha + \beta(x_i - \bar{x}), \sigma^2)$, $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$, $\hat{\beta} \sim N(\beta, \sigma^2/\sum_{i=1}^n (x_i - \bar{x})^2)$

So

$$\frac{[Y_i - \alpha - \beta(x_i - \bar{x})]^2}{\sigma^2}, \frac{(\hat{\alpha} - \alpha)^2}{\sigma^2/n}, \frac{(\hat{\beta} - \beta)^2}{\sigma^2/\sum(x_i - \bar{x})^2}$$

have χ^2 distributions with 1 degree of freedom each. Advanced results show these terms are independent.

Hence $\sum_{i=1}^n [Y_i - \alpha - \beta(x_i - \bar{x})]^2/\sigma^2$ has a χ_n^2 distribution.

Simple Regression

Using the moment generating function argument as for the Sample Mean and Variance the estimates $\hat{\alpha}$ and $\hat{\beta}$ when centred at their true value and normalised by their standard errors have t distributions but now with $n-2$ degrees of freedom

Again use the unbiased estimator of the variance, S^2:

$$\text{If } S^2 = \frac{\sum_{i=1}^n [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2}{n-2}, \text{ then } \frac{(n-2)S^2}{\sigma^2} \sim \chi_{n-2}^2 \qquad (16)$$

So

$$T_\alpha = \frac{\sqrt{n}(\hat{\alpha} - \alpha)}{S} \sim t_{n-2}, \quad T_\beta = \frac{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}(\hat{\beta} - \beta)}{S} \sim t_{n-2} \qquad (17)$$

4.5 Example — Stress and Cancer Data

Example — Simple Regression, Stress and Cancer.

Data Stress and Cancer example - find the maximum likelihood estimates of intercept and slope.

Hand computation possible because only 10 mice.

Intercept $\hat{\alpha} = \bar{y} = -15.22(2dp)$.

Slope $\hat{\beta} = 6.61(2dp)$.

Standard Deviation $\hat{\sigma}^2 = 95.36 = 9.77^2$.

Rare to do computations by hand!

Computers do the calculations very quickly

Form of the estimator is important in deriving its properties.

R Commands for Regression

```r
# Centre at Mean of VEFGC
VEGFC1 <- VEGFC - mean(VEGFC)
# Regress COX2 on centred VEFGC
regStress <- lm(COX2 ~ VEGFC1)
# Summary of Regression Output
summary(regStress)
```

```
##
## Call:
## lm(formula = COX2 ~ VEGFC1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.9988  -3.1019  -0.2101   3.7967  15.2614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   15.215      3.088   4.927  0.00115
## VEGFC1         6.614      1.630   4.056  0.00365
##
## (Intercept) **
## VEGFC1      **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.765 on 8 degrees of freedom
## Multiple R-squared:  0.6729,Adjusted R-squared:  0.632
## F-statistic: 16.46 on 1 and 8 DF,  p-value: 0.003651
```

Plot commands

```r
# Plot of data points
plot(VEGFC1, COX2, cex = 2, cex.axis = 2, cex.lab = 2,
    xlab = "Centred VEGFC", ylab = "COX2")
# Add the regression line
abline(regStress)
# Add the fitted value
points(VEGFC1[6], regStress$fitted.values[6], col = "red",
    cex = 2)
# Add the line to illustrate the residual
segments(x0 = VEGFC1[6], y0 = regStress$fitted.values[6],
    x1 = VEGFC1[6], y1 = COX2[6], lty = "dotted")
# Add the text for 'Residual'
text(x = VEGFC1[6] - 0.4, y = (regStress$fitted.values[6] +
    COX2[6])/2, "Residual", cex = 2)
# Add the text for 'Fitted Value'
text(x = VEGFC1[6] - 0.5, y = regStress$fitted.values[6] +
    2, "Fitted Value", cex = 2, col = "red")
```

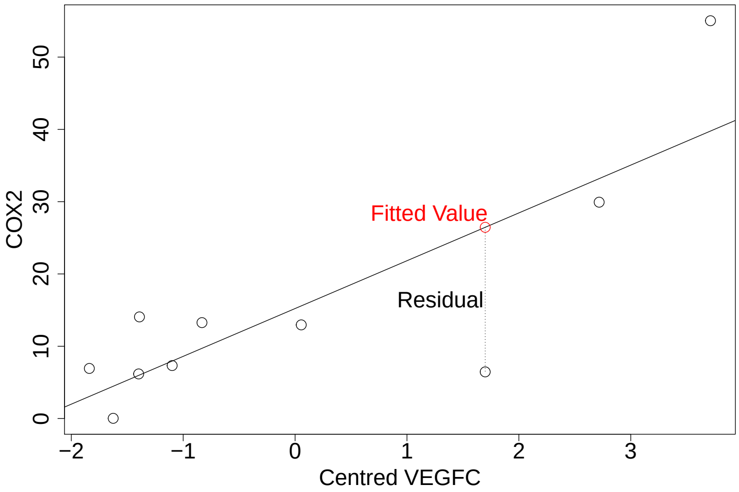

Figure 18: COX2 vs. Centred VEGFC with Residual and Fitted illustrated

Use Residuals to Check Model

An examination of residuals can check model assumptions, specifically:

Normal Distribution - via qqnorm.

Means vary linearly with x - plot residuals vs. x and look for systematic pattern in residuals.

Variance is constant over x - plot residuals vs. x and look for different standard deviations of residuals.

Plots more effective for larger data sets but the next slide illustrates the qqplot for the mice data with the R command:

```
# QQPlot of Normal Quantiles vs. Quantiles of Residuals
qqnorm(y=regStress$residuals,xlab="Theoretical Quantiles", ylab="Residual Quantiles",
cex=2,cex.axis=2,cex.lab=2,main="")
# Add line through 25th and 75th quantiles
qqline(regStress$residuals, col=2)
```

5 Asymptotic Distributions of Maximum Likelihood Estimators — 6.6

5.1 Fisher Information Definition

Fisher Information Definition

Suppose $f(x;\theta)$ is a pdf or pmf in x with parameter θ.

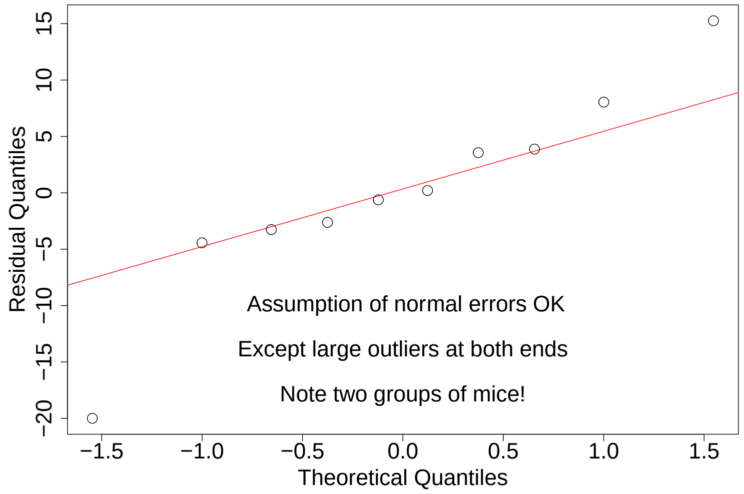

Figure 19: QQPlot for Residuals for Linear Regression of COX2 on VEGFC

Fisher *Information*, $I(\theta)$, is defined by:

$$I(\theta) \equiv Var(\frac{\delta}{\delta\theta}\ln f(X;\theta)) = \int_{-\infty}^{\infty}\frac{\delta}{\delta\theta}(\ln f(x;\theta))^2 f(x;\theta)\,dx. \qquad (18)$$

under regularity conditions permitting interchance of integrals and derivatives (which are usually satisfied - except if the range of X depends on θ).

Under these regularity conditions

$$E(\frac{\delta}{\delta\theta}\ln f(X;\theta)) = 0, \quad I(\theta) = E(-\frac{\delta^2}{\delta\theta^2}\ln f(X;\theta)) \qquad (19)$$

5.2 Example — Fisher Information for Exponential

Example - Fisher Information for Exponential

Calculate the Fisher Information for the exponential distribution with scale parameter $\theta > 0$ using both first and second derivatives of the log likelihood.

Solution — Fisher Information for Exponential

PDF

$$f(x;\theta) = \begin{cases} 0, & x < 0 \\ \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), & x \geq 0 \end{cases}$$

So , for $x \geq 0$, (see the determination of the MLE for exponential)

$$\frac{\delta}{\delta\theta}\ln f(x;\theta) = -\frac{1}{\theta} + \frac{x}{\theta^2} \quad \text{and} \quad \frac{\delta^2}{\delta\theta^2}\ln f(x;\theta) = \frac{1}{\theta^2} - \frac{2x}{\theta^3}$$

Hence using the fact that the mean of an exponential random variable is its scale parameter (see p. 7 of Module 3)

$$\begin{aligned}E(\frac{\delta}{\delta\theta}\ln f(X;\theta)) &= E\left(-\frac{1}{\theta}+\frac{X}{\theta^2}\right)\\ &= 0\end{aligned}$$

Solution — Fisher Information for Exponential Ctd

And using the fact that the variance of an exponential random variable is its scale parameter squared (see p. 7 of Module 3)

$$\begin{aligned}Var(\frac{\delta}{\delta\theta}\ln f(X;\theta)) = Var\left(-\frac{1}{\theta}+\frac{X}{\theta^2}\right) &= Var\left(\frac{X}{\theta^2}\right)\\ &= \frac{\theta^2}{\theta^4} = \frac{1}{\theta^2}\end{aligned}$$

Whereas

$$E\left(-\frac{\delta^2}{\delta\theta^2}\ln f(X;\theta)\right) = E\left(-\frac{1}{\theta^2}+\frac{2X}{\theta^3}\right) = -\frac{1}{\theta^2}+\frac{2\theta}{\theta^3} = \frac{1}{\theta^2}$$

Solution — Fisher Information for Exponential Ctd 2

Hence , as expected, using both methods the Fisher information for the exponential distribution

$$I(\theta) = \frac{1}{\theta^2} \tag{20}$$

5.3 MLE Asymptotically Unbiased and Normally Distributed, Variance Fisher Information

Distribution of MLE for large sample size

Suppose $\hat{\theta}$ is the MLE for n independent and identically distributed obervations *on an arbitrary pdf or pmf* that satisfies the regularity conditions needed, including interchanging derivatives and expectations as well as the requirement that the range of the rv's does not depend on θ.

Then, for large sample size, the distribution of $\hat{\theta}$ is approximately normal

Mean is the true parameter value θ

Variance is $\frac{1}{nI(\theta)}$

So

$$\sqrt{nI(\hat{\theta})}(\hat{\theta}-\theta) \approx N(0,1)$$

5.4 Example — Asymptotic Range for MLE

Example - Asymptotic Range for MLE

Find a 95% range of values in which the difference between the maximum likelihood estimator and the true parameter value is to lie if the underlying pdf for $\theta > 0$ is

$$f(x;\theta) = \begin{cases} 0, & x < 0 \\ \theta x^{\theta-1}, & x \geq 0 \\ 0, & x > 1 \end{cases}$$

and the sample size is large

Solution — Asymptotic Range for MLE

Derivatives of log density:

$$\frac{\delta}{\delta\theta} \ln f(x;\theta) = \frac{\delta}{\delta\theta} (\ln(\theta) + (\theta - 1)\ln(x)) = \frac{1}{\theta} + \ln(x)$$
$$\frac{\delta^2}{\delta\theta^2} \ln f(x;\theta) = -\frac{1}{\theta^2}$$

Because the second derivative is a constant, the expectation of the rv is that constant:

$$I(\theta) = E\left(-\frac{\delta^2}{\delta\theta^2} \ln f(X;\theta)\right) = \frac{1}{\theta^2}$$

Log likelihood for data $x_1, \cdots, x_n$ is

$$\sum_{i=1}^{n} \ln f(x_i;\theta) = \sum_{i=1}^{n} (\ln(\theta) + (\theta - 1)\ln(x_i))$$

Solution — Asymptotic Range for MLE Ctd

Derivative of log-likelihood equated to zero is

$$\sum_{i=1}^{n} \frac{1}{\theta} + \ln(x_i) = 0$$

So, noting that the second derivative is uniformly negative, the solution for the MLE is $\hat{\theta} = \left(\overline{-\ln X}\right)^{-1} = \frac{n}{\sum_{i=1}^{n} - \ln(X_i)}$

And $\hat{\theta} - \theta$ has an approximate normal distribution for large sample size with mean 0 and variance $1/(nI(\theta)) = \frac{\theta^2}{n} \approx \frac{\hat{\theta}^2}{n}$.

Hence

$$P\left(\frac{-2\hat{\theta}}{\sqrt{n}} < \hat{\theta} - \theta < \frac{2\hat{\theta}}{\sqrt{n}}\right) \approx 0.95$$

that is, the required interval is $(\frac{-2\hat{\theta}}{\sqrt{n}}, \frac{2\hat{\theta}}{\sqrt{n}})$

5.5 How Good is MLE? - Cramer-Rao Lower Bound

Cramer-Rao Lower Bound

With suitable regularity conditions (again the range of the rv must not depend on the parameter), there is a lower bound - called the *Cramer-Rao lower bound* - on the variance of an unbiased estimator

Cramer-Rao lower bound *is* the asymptotic variance for the MLE. So if an estimator, $Y = u(X_1, \ldots, X_n)$, is unbiased, then

$$Var(Y) \geq \frac{1}{nI(\theta)}$$

If an unbiased estimator has this variance, then it is best in the sense that it has minimum variance compared with other unbiased estimators

MLE's are also (close to) optimal for large sample size in that they are approximately unbiased and have the Cramer-Rao lower bound as their approximate variance

5.6 Example — Exponential

Example - Sample Mean for Exponential Samples

Show that the sample mean, the MLE, is the best unbiased estimator for the population mean for the exponential case for all sample sizes

Solution - Sample Mean for Exponential Samples

From equation (20) the Fisher information is the same as the reciprocal of the population variance θ^2

Hence

$$Var(\bar{X}) = \frac{\theta^2}{n} = \frac{1}{nI(\theta)}$$

which is the *Cramer-Rao* lower bound for the variance of *any* unbiased estimator

6 Bayesian Estimation — 6.8

6.1 Prior and Posterior Distributions

Bayesian Introduction

Bayesians start with their knowledge of the likely values of the parameter Θ now regared as a random variable with a pmf or pdf called the *prior* distribution (Latin for "before" collecting data).

Frequentists regard the parameter as an unknown number and use methods like Maximum Likelihood Estimation, Method of Moments or Best Unbiased Estimates to estimate the parameter.

In complicated, realistic problems frequentists often need to introduce smoothing to produce sensible estimates - for example, estimating a probability density function - and choosing the smoothing parameter is often the same as choosing a prior.

Both groups want to estimate an interval of parameter values (see Module 7) rather than just one number - easier for Bayesians.

Bayesian Introduction Ctd

Bayesians use the same *probability models* for their data but regard them as for the data *conditional* on the value of the parameter.

Given the data, they compute the *posterior* (Latin for "after" collecting data) distribution of Θ given the data $x_1, \cdots, x_n$ using Bayes Theorem.

6.2 Example - Binomial Probability

Example — Binomial Probability

Suppose coins are known to be fair or to have a bias in which heads comes up with probability 0.6, and suppose that 70% of coins are fair. If 80 heads are observed in 100 tosses of the coin, what now is the probability that the coin is fair?

Odds Version of Bayes Theorem

Suppose that F is the event that the coin is fair and D is the event that 80 heads are observed in 100 tosses of the coin.

Recall that if p is the probability of an event and o are the odds:

$$o = \frac{p}{1-p}, \quad p = \frac{o}{1+o}$$

.

Recall further that for any events A and B, $P(A \cap B) = P(A|B)P(B)$, by the definition of conditional probability (assuming $P(B) > 0$).

Bayes Theorem becomes the observation that the *posterior odds* of F is the *prior odds* of F mutiplied by the *likelihood ratio.*

Proof of Odds Version of Bayes Theorem

$$\begin{aligned}
\text{posterior odds} &= \frac{P(F|D)}{P(F^c|D)} \\
&= \frac{P(F \cap D)}{P(F^c \cap D)} \\
&= \frac{P(D|F)P(F)}{P(D|F^c)P(F^c)} \\
&= \text{likelihood ratio} \times \text{prior odds}
\end{aligned}$$

Example — Binomial Probability Restated

Suppose coins are known to be fair or to have a bias in which heads comes up with probability 0.6, and suppose that 70% of coins are fair. If 80 heads are observed in 100 tosses of the coin, what now is the probability that the coin is fair?

Solution - Binomial Probability

Applying the odds version of Bayes Theorem with the relevant Binomial probabilities gives

$$\text{poterior odds of a fair coin} = \frac{\binom{100}{80}(0.5)^{100}}{\binom{100}{80}(0.6)^{80}(0.4)^{20}} \times \frac{70}{30}.$$

R commands to work this out are:

```
o <- (0.5/0.6)^80 * (0.5/0.4)^20 * 0.7/0.3
o/(1 + 0)

## [1] 9.368657e-05
```

So the data changes our probability that our coin is fair to 0.000094 - small!

Comment — Odds Version of Bayes Theorem

Odds version of Bayes Theorem makes it clear why the probability of a false positive must be small for a rare disease test to be accurate - denominator must be tiny for the likelihood ratio to be large enough to compensate for the small possibility that a random person has the disease.

Fair coin example is the opposite - data made the numerator probability very small relative to the denominator probability.

Odds version works best when only two parameter values

Small sample size makes the prior odds important because the likelihood ratio is often not as extreme.

Choosing prior probabilities then becomes important.

6.3 Choosing Prior Probabilities

Choosing Prior Probabilities

Three techniques:

- Non-informative priors give equal weight or density to all parameter values - tricky when the parameter space is the whole line because this makes the total probability infinity.
- Prior beliefs to construct the prior - criticised as not objective by Frequentists because different people may get different results from the same data (something that also happens when Frequentists or Bayesians use different models for the data given the parameters - lies, damn lies ...).

Conjugate prior - a prior distribution that makes the posterior distribution a member of the same family - Bayes Theorem then updates the parameters from the prior to the posterior using the data (see Normal example) .

6.4 Bayesian Estimation

How to estimate?

Posterior distribution contains all the information - but what property of the posterior should be used to give a single estimate of Θ?

If it is desired to minimise $E_{\text{posterior}}(b - \Theta)^2$, then in Module 2 p.18, it was shown that we should choose b as the *mean* of the posterior distribution.

That is, with expected squared loss as the criterion, the Bayes estimate of a parameter is the *mean* of the *posterior distribution*.

With expected absolute value loss as the criterion, the Bayes estimate becomes the *median* of the *posterior distribution* - same as *mean* for *symmetric* distributions.

6.5 Bayes Theorem With PDF's

Bayes Theorem With PDF's

Suppose the vector (Θ, X) has a joint pdf where X is the data and Θ is the unknown parameter.

Let $f(x|\theta)$ be the conditional density of the data, X, given $\Theta = \theta$: same as $f(x;\theta)$ in the likelihood theory.

Let $h(\theta)$ be the prior density of Θ.

Joint density, as in Module 4, of (Θ, X) is $h(\theta)f(x|\theta)$.

Marginal density of X (often called the *predictive* density) is $k(x) = \int_{-\infty}^{\infty} f(x|\theta)h(\theta)\, d\theta$.

Bayes theorem says that the *posterior* density, $k(\theta|x)$, of Θ given $X = x$ is the joint density divided by the marginal:

$$k(\theta|x) = \frac{h(\theta)f(x|\theta)}{k(x)} = \frac{h(\theta)f(x|\theta)}{\int_{-\infty}^{\infty} f(x|\theta)h(\theta)\, d\theta}. \qquad (21)$$

Computing the posterior density

Argument for Bayes Theorem is the same if the data is not one number but a sample of numbers.

Posterior density from Bayes Theorem may often be recognised as a PDF j (say) times constants and/or functions of the data. Then this density j *is* the posterior density.

To see why this is true, suppose for some probability density function j

$$h(\theta)f(x|\theta) = l(x)j(\theta), \quad \text{written:} \quad h(\theta)f(x|\theta) \propto j(\theta) \qquad (22)$$

for some function l .

Computing the posterior density ctd

Bayes theorem then gives

$$k(\theta|x) = \frac{l(x)j(\theta)}{k(x)} \tag{23}$$

But the right side of (23) must integrate to one over θ that is, for any x, since j is a pdf,

$$\int_{-\infty}^{\infty} \frac{l(x)j(\theta)}{k(x)}\, d\theta = \frac{l(x)}{k(x)} \int_{-\infty}^{\infty} j(\theta)\, d\theta = \frac{l(x)}{k(x)} = 1$$

Which gives the posterior distribution as the pdf j,

$$h(\theta)f(x|\theta) \propto j(\theta) \implies k(\theta|x) = j(\theta). \tag{24}$$

6.6 Example — Normal

Example - Normal Data, Normal Prior

Previous measurements of VEGFC suggest that the mean value for a population of mice is normally distributed centred at 0.9 and with variance 0.25. The 5 VEFGC measurements in the stress group were 4.05745, 2.41335, 1.52595, 6.07073 and 5.07592. Given the population mean for VEGFC for stressed mice, the VEGFC measurements for the stress group are assumed to be independent and normally distributed with variance 0.49. What is the Bayes estimate for the population mean for mice who are stressed?

Solution — Normal Data, Normal Prior

Let Θ denote the unknown population mean for VEGFC and $x_1 = 4.05745, \cdots, x_5 = 5.07592$ be the data values.

Prior pdf is $h(\theta) = \frac{1}{\sqrt{2\times\pi 0.25}} \exp\left(-\frac{(\theta-0.9)^2}{2\times 0.25}\right)$.

PDF of data $x_1, \cdots, x_5$ given $\Theta = \theta$ is $f(x_1, \cdots, x_5|\theta) = \frac{1}{\sqrt{2\pi 0.49}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i-\theta)^2}{2\times 0.49}\right)$.

Analysis of Variance identity in Module 5, p.4, gives

$$f(x_1, \cdots, x_5|\theta) = \frac{1}{\sqrt{2\pi 0.49}} \exp\left(-\frac{\sum_{i=1}^{5}(x_i-\bar{x})^2 + 5(\bar{x}-\theta)^2}{2\times 0.49}\right)$$

.

Multiplying the prior with the density for the data given θ gives

$$h(\theta)f(x_1, \cdots, x_5|\theta) \propto \exp\left(-\left(\frac{(\theta-0.9)^2}{2\times 0.25} + \frac{5(\theta-\bar{x})^2}{2\times 0.49}\right)\right). \tag{25}$$

Solution — Normal Data, Normal Prior Ctd

Could this be a normal density? To see whether this is the case, *define* the inverses of the variances, $\tau_{\text{prior}}, \tau_{\text{data}}$ - called *precisions* - as well as their sum τ:

$$\tau_{\text{prior}} = \frac{1}{0.25} = 4, \quad \tau_{\text{data}} = \frac{5}{0.49} = 10.20, \quad \tau = \tau_{\text{prior}} + \tau_{\text{data}} = 14.20.$$

And *probabilities* associated with them:

$$p_{\text{prior}} = \frac{\tau_{\text{prior}}}{\tau} = 0.2816, \quad p_{\text{data}} = \frac{\tau_{\text{data}}}{\tau} = 0.7184.$$

Also define a *a weighted mean*, $\mu_{\text{posterior}}$, of the prior mean and the mean from the data and a *variance*, $\sigma^2_{\text{posterior}}$ (noting $\bar{x} = 3.829$) :

$$\mu_{\text{posterior}} = 0.9 p_{\text{prior}} + \bar{x} p_{\text{data}} = 3.004, \quad \sigma^2_{\text{posterior}} = \frac{1}{\tau} = 0.0704.$$

Solution — Normal Data, Normal Prior Ctd 2

Expanding the squares in the exponents in (25) and doing some algebra, after noting that all except terms involving θ are numbers, leads to

$$h(\theta) f(x_1, \cdots, x_5|\theta) \propto \exp\left(-\frac{\tau(\theta^2 - 2\mu_{\text{posterior}}\theta)}{2}\right), \tag{26}$$

Which on completing the square gives

$$h(\theta) f(x_1, \cdots, x_5|\theta) \propto \exp\left(-\frac{(\theta - \mu_{\text{posterior}})^2}{2\sigma^2_{\text{posterior}}}\right), \tag{27}$$

And this is in turn is proportional to a $N(\mu_{\text{posterior}} = 3.004, \sigma^2_{\text{posterior}} = 0.0704)$ pdf .

Solution — Normal Data, Normal Prior Ctd 3

Hence by (24) this normal density *is* the posterior pdf .

Bayes estimate (mean or median) is $\mu_{\text{posterior}} = 3.004$.

Variance of posterior distribution is $\sigma^2_{\text{posterior}} = 0.0704$.

Note that the Bayes estimate is a weighted average of the prior belief about the mean of Θ with the sample mean from the data - the weights are proportional to the prior variance and the ratio of the sample size to the data variance.

Makes sense to "shrink" from the sample mean to take into account prior beliefs.

Possible to make probability statements, on the basis of the posterior distribution, about our confidence in the estimate of the unknown mean of VEGFC for mice that are subject to stress.

Normal Data, Normal Prior

Argument works for any normally distributed data and normal prior to get the posterior distribution as normal, but with suitable changes to the definition of $\mu_{\text{posterior}}, \sigma^2_{\text{posterior}}$ (to take account of the specific values of the prior mean, the prior variance, the variance of the data and the sample mean).

If the prior came from the sampling distribution of previous observations, and the variance for the prior were the sampling variance divided by the sample size, the formula for the posterior mean and variance shows that the posterior distribution will be the sampling distribution for the combined sample.