Lecture Notes for Stat261

by

Mary Lesperance

Edits 3 January 2024

Department of Mathematics and Statistics
University of Victoria, Victoria, B.C.

Contents

Chapter 1

Background material

1.1 Distribution Summary

1. **Binomial**(n, p); $\quad f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n$

   where $\binom{n}{x} = \frac{n!}{(x!)(n-x)!}$.

   Consider n independent repetitions of an experiment each of which has only two possible outcomes, say (S, F) where

   $$P\{S\} = p \text{ is constant, i.e. the same for each experiment}$$

   Let $X = \#S$'s in n trials

   Then $X \sim$ Binomial (n, p).

   Example: Let X be the number of heads in n tosses of a fair coin.

   Note: We often use the Binomial Distribution when we **Sample with Replacement**.

   $E(X) = np \quad , \quad Var(X) = np(1-p)$

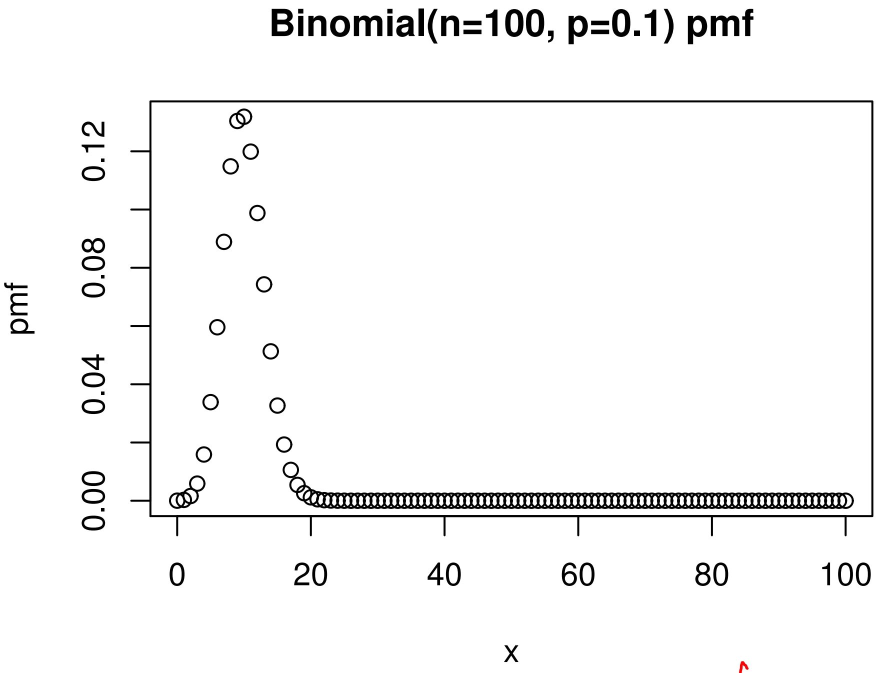


Figure 1.1: Binomial probability mass function (pmf), n=100, p=0.1

2. **Multinomial**$(n, p_1, \ldots, p_k)$; $\quad f(x_1, \ldots, x_k) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$

   where $\binom{n}{x_1 \ldots x_k} = \frac{n!}{(x_1!)(x_2)! \cdots (x_k)!}$ and

   $x_i = 0, 1, \ldots, n,$ such that $x_1 + \cdots + x_k = n$ and $p_1 + \cdots + p_k = 1.$

   Consider n independent repetitions of an experiment for which each outcome can be classified in exactly one of k mutually exclusive ways, $A_1, A_2, \ldots, A_k$.

   Let

   $$p_i = P\{\text{an outcome of one trial is of class } A_i\}$$
   $$X_i = \# \text{ outcomes that are of class } i \text{ out of } n \text{ repetitions}$$

   Then $(X_1, X_2, \ldots, X_k) \sim$ Multinomial $(n, p_1, \ldots, p_k)$

   $E(X_i) = np_i \,, \quad Var(X_i) = np_i(1-p_i)$

   $Cov(X_i, X_j) = -np_ip_j \quad i \neq j$

Note: $\sum_{i=1}^{k} p_i = 1 \quad \sum_{i=1}^{k} X_i = n$

Example: Toss a fair die $n = 100$ times and let $(X_1, X_2, \ldots, X_6)$ be the observed frequencies of the numbers $1, 2, 3, 4, 5, 6$ from the tosses of the die. Since the die is fair, then $p_i = 1/6$ for $i = 1, \ldots, 6$.

3. **Negative Binomial**(r, p); $f(x) = \binom{x+r-1}{r-1} p^r (1-p)^x$, $x = 0, 1, ...$

   Consider independent repetitions of an experiment each of which, has exactly two possible outcomes, say (S, F).

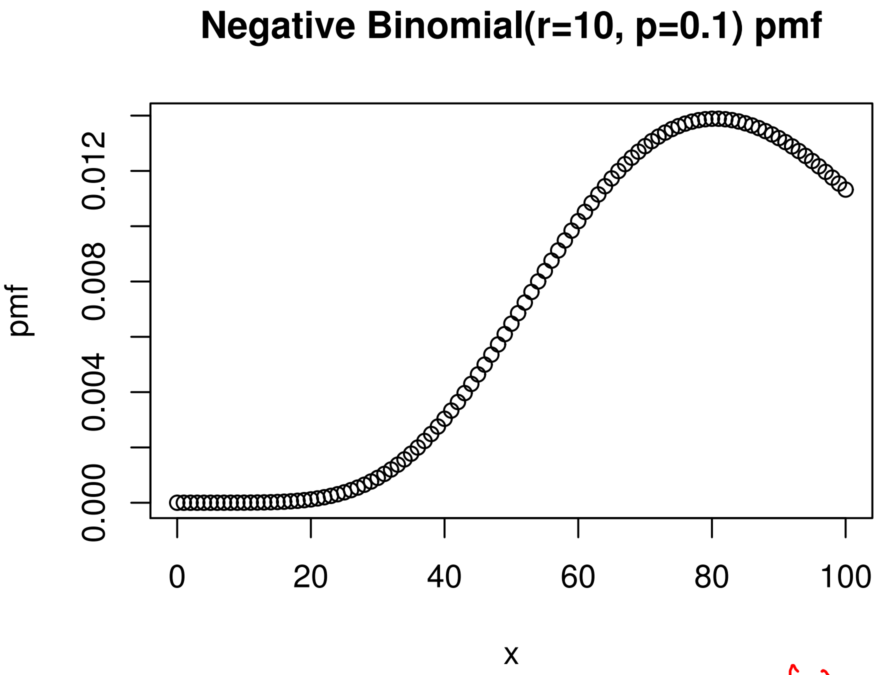
   Let $P(S) = p$ constant, i.e. the same for each experiment

   Let $X = \#$ F's before the r^{th} S

   Then $X \sim \text{NegBin}(r, p)$

   Example: Continue flipping a fair coin and stop when you observe the first head. $X =$ the number of tails before the first head has a Negative Binomial distribution with $r = 1$.

$$E(X) = \frac{r(1-p)}{p} \quad , \quad Var(X) = \frac{r(1-p)}{p^2}$$



(pmf)

Figure 1.2: Negative Binomial probability mass function, r=10, p=0.1

4. **Geometric**(p); is the same as Negative Binomial $(r = 1, p)$

5. **Hypergeometric(N, M, n)**; $f(x) = \binom{M}{x}\binom{N-M}{n-x}/\binom{N}{n}$ where $\max(0, n - N + M) \leq x \leq \min(n, M)$.

   Consider a finite population of size N. Let each object in the population be characterized as either a S or F, where there are $M \leq N$ S's in the population. Draw a random sample of size n from the population without replacement.

   Let $X = \#\ S$'s in the sample of size n.

   Then $X \sim \text{Hypergeometric}(N, M, n)$.

   Example: Suppose that a bin contains $N = 100$ balls, of which $M = 30$ are white and $N - M = 70$ are black. Choose a random sample of $n = 10$ balls from

$$E(X) = n\left(\underset{\overset{\uparrow}{p}}{\frac{M}{N}}\right), \quad Var(X) = \frac{N-n}{N-1}\, n\left(\underset{\overset{\uparrow}{p}}{\frac{M}{N}}\right)\left[1 - \left[\underset{\overset{\uparrow}{p}}{\frac{M}{N}}\right]\right]$$

the bin without replacement. X = the number of white balls in the sample has a Hypergeometric($N = 100, M = 30, n = 10$) distribution.

Example: A shipping container contains $N = 10,000$ iPhone 7's of which $M = 30$ are defective and the remainder are not defective. Choose a random sample of $n = 100$ iPhone 7's from the shipping container without replacement. Then X = the number of defectives in the sample has a Hypergeometric($N = 10,000, M = 30, n = 100$) distribution.

In this example, $n/N = 100/10,000 = 0.01 \leq 0.05$. Then X = the number of defectives in the sample is approximately distributed as Binomial($n = 100, p = 30/10,000 = 0.003$).

If $\frac{n}{N} \leq .05$, can use Binomial to approximate

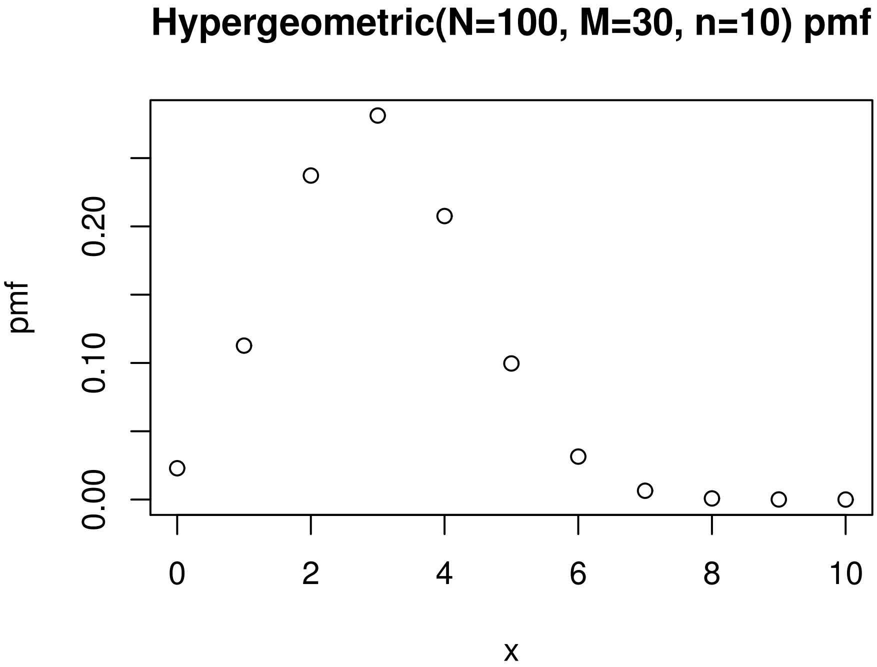


Figure 1.3: Hypergeometric probability mass function, N=100, M=30, n=10

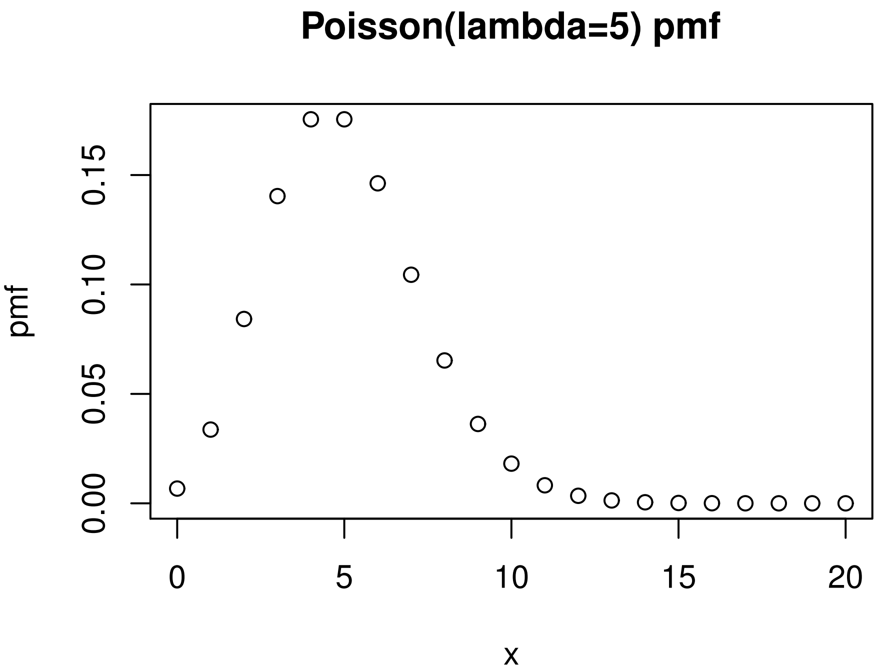
6. **Poisson** (λ); $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $x = 0, 1, \ldots$

Models the number of occurrences of random events in space and time, where the average rate, λ per unit time (or area, or volume) is constant.

$$\text{Let } X = \#\text{ events in } t \text{ units of time}$$
$$\text{Then } X \sim \text{Poisson } (\lambda t)$$

Example: Let X = the number of customers arriving at a bank in a given one hour time interval.

$E(X) = \lambda$, $Var(X) = \lambda$



(pmf)

Figure 1.4: Poisson probability mass function, $\lambda = 5$

7. **Exponential** (*mean* θ); $f(x) = \frac{1}{\theta}\, e^{-x/\theta} \quad x > 0 \quad \theta > 0$

Models lifetimes where there is no deterioration with age - or - waiting times between successive random events in a Poisson process. We also parameterize the exponential distribution using the rate parameter, $\lambda = 1/\theta$.

$E(X) = 0$, $Var(X) = \theta^2$

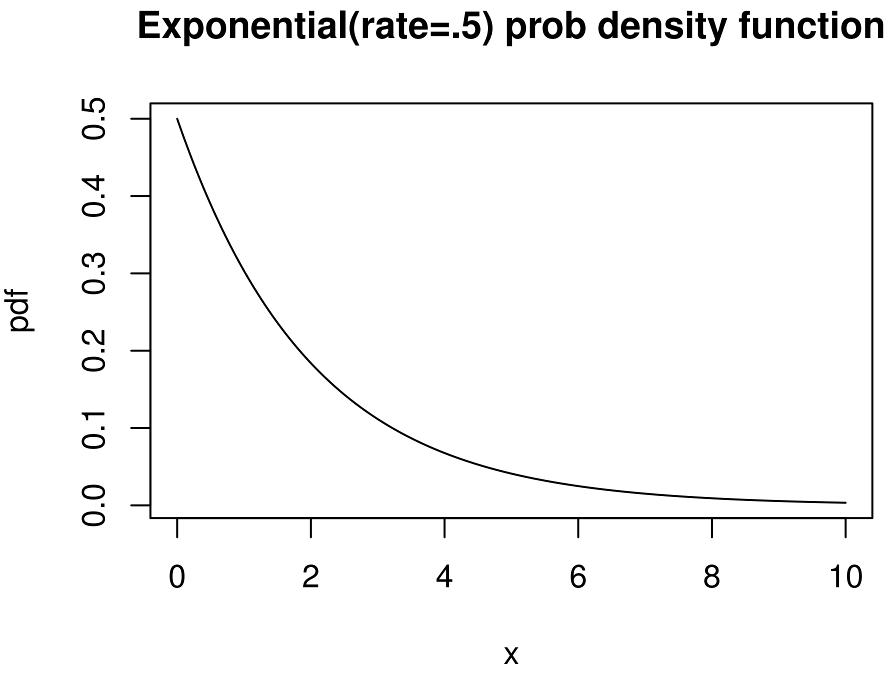


Figure 1.5: Exponential density function, $\overset{\text{mean}}{\theta} = 1/.5 = 2$

$\lambda = \text{rate} = .5$

8. **Gamma**(α, β); $\quad f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\, x^{\alpha-1}\, \exp(-x/\beta),\ x > 0,\ \alpha, \beta > 0.$

$E(X) = \alpha\beta$, $Var(X) = \alpha\beta^2$

flexible model used to model lifetimes.

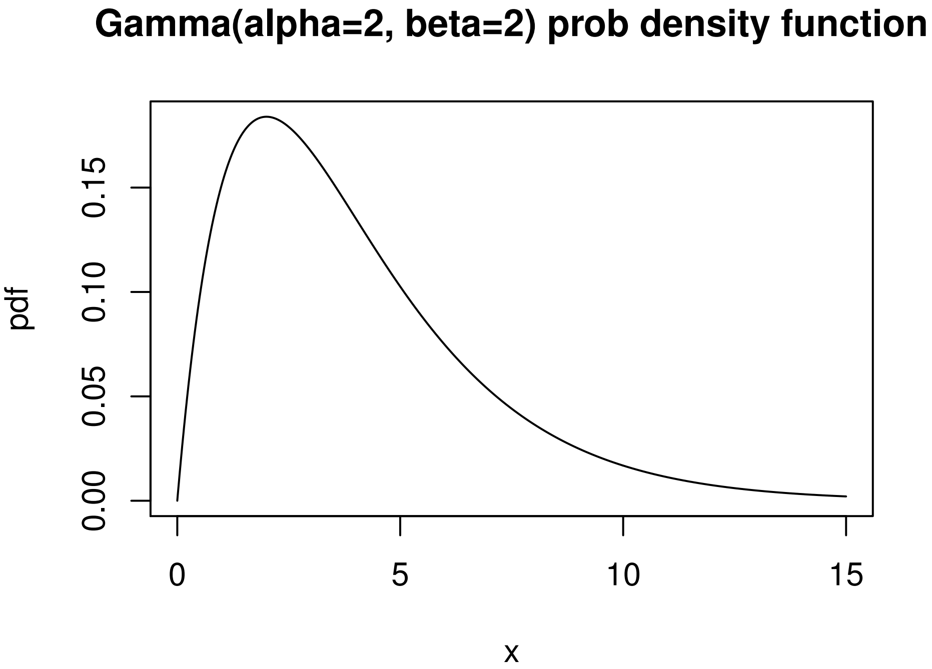


Figure 1.6: Gamma density function, $\alpha = 2, \beta = 2$

9. **Normal** (μ, σ^2); $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$, $x, \mu \in \Re$, $\sigma^2 > 0$.

   Many measurements are approximately normal.

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

$$\frac{X-\mu}{\sigma} \sim N(0,1).$$

$E(X) = \mu$, $Var(X) = \sigma^2$, sd = σ = standard deviation

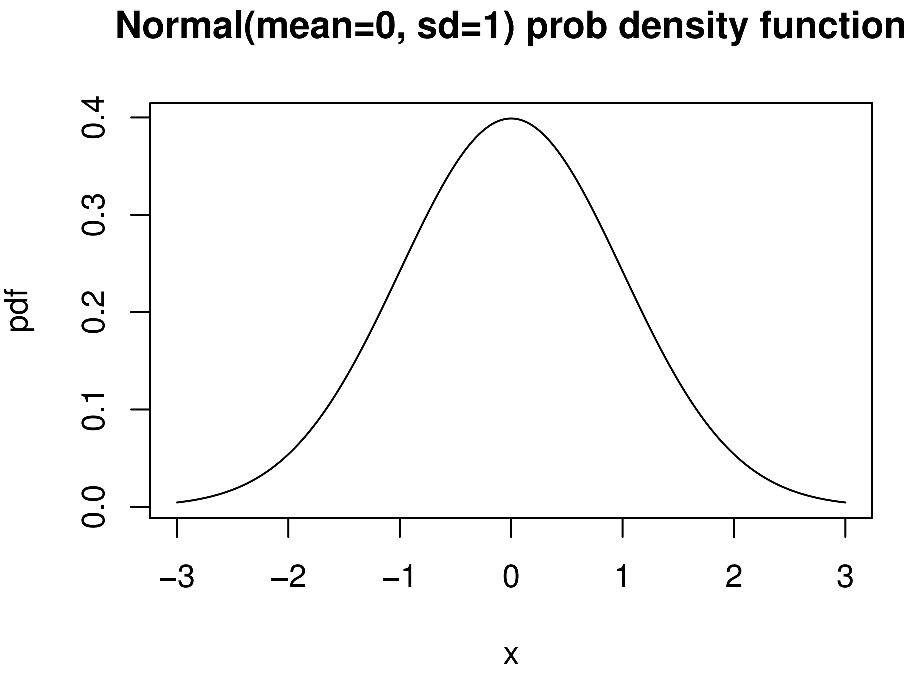


Figure 1.7: Normal density function, $\mu = 0, \sigma = 1$

Note: If $X_1, \ldots, X_n$ are independent with $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ and $a_1, \ldots, a_n$ are constants,

then $\sum_{i=1}^{n} a_i X_i \sim N\left(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2\right)$

Central Limit Theorem

Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of n independent random variables each with mean μ, variance σ^2. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \approx N(0,1) \quad \text{for large } n,$$

where $\approx$ means approximately distributed as.

1.1.1 R code for Distribution Figures

```
#Binomial
x <- 0:100
plot(x, dbinom(x,size=100,prob=.1), ylab='pmf', xlab='x')
title("Binomial(n=100, p=0.1) pmf")

#Negative Binomial
x <- 0:100
plot(x, dnbinom(x,size=10,prob=.1), ylab='pmf', xlab='x')
title("Negative Binomial(r=10, p=0.1) pmf")

#Hypergeometric
x <- 0:10
plot(x, dhyper(x,m=30,n=70,k=10), ylab='pmf', xlab='x')
title("Hypergeometric(N=100, M=30, n=10) pmf")

#Poisson
x <- 0:20
plot(x, dpois(x,lambda=5), ylab='pmf', xlab='x')
title("Poisson(lambda=5) prob mass function")

#Exponential
x <- seq(0,10,by=.01)
plot(x, dexp(x,rate=.5), ylab='pdf', xlab='x', type='l')
title("Exponential(rate=.5) prob density function")

#Gamma
x <- seq(0,15,by=.01)
plot(x, dgamma(x,shape=2,scale=2), ylab='pdf', xlab='x', type='l')
title("Gamma(alpha=2, beta=2) prob density function")

#Normal
x <- seq(-3,3,by=.01)
plot(x, dnorm(x,mean=0, sd=1), ylab='pdf', xlab='x', type='l')
title("Normal(mean=0, sd=1) prob density function")
```

1.2 Review Stat260

- A **random variable**, X is a quantity which is capable of taking various real values according to chance.

  Notation: X, Y, Z random variables

  x, y, z realized values of random variables

  $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ range of values

- **Discrete random variables**: have only finitely many or at most countably many possible values. For example,

  $$\begin{array}{lll} X \sim & \text{Binomial } (n, p) & \mathcal{X} = \{0, 1, 2, \ldots, n\} \\ X \sim & \text{Poisson } (\lambda) & \mathcal{X} = \{0, 1, 2, \ldots\} \end{array}$$

  The **Probability mass function, pmf,** of X is $f(x) = P(X = x)$.

- **Continuous random variables**: can take on any real value in an interval. For example,

  $$\begin{array}{ll} X \sim \text{Normal}\,(\mu, \sigma^2) & \mathcal{X} = \mathbb{R} \\ X \sim \text{Exponential}\,(\text{mean } \theta) & \mathcal{X} = (0, \infty) \end{array}$$

  The **Cumulative distribution function of X, cdf,** is $F(x) = P(X \leq x)$.

  The **Probability density function of X, pdf,** is $f(x) = \frac{d}{dx} F(x)$.

  Given $f(x)$ we can obtain $F(x)$,

  $$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx$$

- **Expectation of X:** $E(X)$, also called the (population) mean of X

  Discrete case: $$E(X) = \sum_{x \in \mathcal{X}} x \underbrace{f(x)}_{\uparrow \text{ pmf of } X}$$

  Continuous case: $$E(X) = \int_{-\infty}^{\infty} x \underbrace{f(x)}_{\uparrow \text{ pdf of } X} dx$$

  Recall: $E(aX + b) = aE(X) + b$ where a, b constants

- **Variance of X:** $$Var(X) = \sigma_X^2 = E\left\{(X - E(X))^2\right\} = E(X^2) - [E(X)]^2$$

  Discrete case: $$Var(X) = \sum_{x \in \mathcal{X}} (x - E(X))^2 f(x)$$

  Continuous: $$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\, dx$$

  Recall: (i) $\sqrt{Var(X)} = \sigma$ is called the **standard deviation (sd)** of X
  (ii) $Var(aX + b) = a^2 Var(X)$ a, b constants
  (iii) $Var(X + Y) = Var(X) + Var(Y) + 2\, Cov(X, Y)$
  (iv) $Cov(X, Y) = E(XY) - E(X)E(Y)$

- **Independent random variables**:

  Let X and Y be random variables with marginal pmf's (or pdf's) $f_1(x)$ and $f_2(y)$ respectively. Let $f(x, y)$ be the joint pmf (pdf) of X and Y. Then X and Y are **statistically independent** if and only if

  $$f(x, y) = f_1(x) f_2(y)$$ for all x and y

  Recall: If X, Y are independent, then

  $$Cov(X, Y) = 0 \text{ and } Var(X + Y) = Var(X) + Var(Y)$$

1.3 Notation

The following is a list of notation for these notes:

1. $\sim$: is distributed as
2. $\approx$: approximately distributed as
3. $L(\theta)$: the likelihood function as a function of θ
4. $\ell(\theta)$: log-likelihood as function of θ

Chapter 2

Likelihood methods

2.1 Introduction to Maximum Likelihood Estimation

Optional Text Reading: Section 9.1, pp. 3-8

Example 2.1.1. Canada Border Services processes hundreds of thousands of small parcels entering the country by mail. In their goal to combat the opioid crisis, they are interested in discerning the proportion of small parcels that contain ingredients that could be used to manufacturer illicit drugs, which we will call illegal parcels here. They do not have the resources to check ALL small parcels entering the country, and instead they perform a sample audit.

They randomly choose $n = 100$ small parcels and keep track of

$$X = \# \text{ illegal parcels out of } n = 100$$

$x = 0, 1, \ldots, 100 \,;\quad 0 < \theta < 1$

Let θ = probability that a randomly chosen parcel is illegal. The auditors are interested in estimating θ.

We begin by postulating a probability model to describe the sampling procedure.

We will assume that:

$$X \sim \text{Binomial}(n = 100, p = \theta) \quad x = 0, 1, ..., 100;\ 0 \leq \theta \leq 1$$

$$p(x; \theta) = \binom{100}{x} \theta^x (1 - \theta)^{100-x}$$

Question: What assumptions are required for the use of the Binomial distribution here?

The **Maximum Likelihood Estimate, MLE** of θ is the value of θ that <u>maximizes</u> $p(x; \theta)$ given the data x.

Is that a reasonable estimate of θ?

It is the parameter value that best explains the data in the sense that it maximizes the probability of the data assuming that the hypothesized probability model is true.

$$\text{MAXIMIZATION} \implies \text{CALCULUS}$$

Let's introduce some simplifications into the optimization problem:

- The factor $\binom{100}{x}$ will have no effect on the maximization of $p(x; \theta)$ over θ.

  To simplify the expression, we will omit multiplicative constants that do not involve θ.

  Definition: $\boxed{L(\theta) = cp(x; \theta)}$ is called the **Likelihood function**, where c is a positive constant that does not depend on θ. Therefore, we have:

  $$\max_{\theta}\ p\,(x; \theta) \iff \max_{\theta}\ L(\theta)$$

  In Example 2.1.1, $\quad c = 1/\binom{100}{x} \quad$ and $\qquad L(\theta) = \theta^x\,(1 - \theta)^{100-x}$

- Usually $L(\theta)$ is the product of terms in θ, however, it is generally easier to take derivatives of sums.

Definition: $\boxed{\ell(\theta) = \ln L(\theta)}$ is called the **Log-likelihood function**.

Note: $\ln(y)$ is a monotone increasing function of y, so we have the following:

$$\max_{\theta} \ell\left(\theta\right) \Leftrightarrow \max_{\theta} L(\theta) \Leftrightarrow \max_{\theta} p\left(x;\theta\right)$$

Returning to the auditing Example 2.1.1:

$$\begin{aligned} \ell\left(\theta\right) &= \ln L(\theta) = x \ln\theta + (100-x)\ln\left(1-\theta\right) \\ \ell'\left(\theta\right) &= \frac{x}{\theta} - \frac{100-x}{1-\theta} \qquad 0<\theta<1 \end{aligned}$$

At the maximum point, $\hat{\theta}$

$$\ell'(\hat{\theta}) = \ 0 = \ \frac{x}{\hat{\theta}} - \frac{100-x}{1-\hat{\theta}} \Longrightarrow \hat{\theta} = \frac{x}{100}$$

To ensure that $\hat{\theta} = \frac{x}{100}$ is a maximum, we check that the second derivative $\ell''(\hat{\theta}) < 0$.

$$\ell''\left(\theta\right) = -\frac{x}{\theta^2} - \frac{100-x}{\left(1-\theta\right)^2} < 0 \ \text{ for all } \ 0<\theta<1.$$

At the boundary values of 0 and 1, $L(0) = L(1) = 0 < L(\theta), 0 < \theta < 1)$. Therefore $\hat{\theta}$ is a maximum, the MLE.

Question: What is the MLE of θ when $x = 0$ or $x = n$?

Example: If $x = 7$ then $\hat{\theta} = .07$ An estimated 7% of small parcels were illegal, i.e. contained illicit ingredients.

We define two more quantities:

- The **Score Function**, $S(\theta)$ is defined as

$$\boxed{S(\theta) = \ell'(\theta) = \frac{d\ell(\theta)}{d\theta}}$$

- The **Information Function**, $I(\theta)$ is defined as

$$\boxed{I(\theta) = -\ell''(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}}$$

At the MLE, $\hat{\theta}$, the proportion of illegal parcels, $S(\hat{\theta}) = 0$ and $I(\hat{\theta}) > 0$.

Returning to the auditing Example 2.1.1 Sampling Method II:

The auditors check through the parcels in a random fashion until they find $r = 7$ parcels with questionable ingredients. They note

$$X_2 = \#\text{legal parcels until the 7'th illegal parcel is observed}$$

$$\begin{aligned}
X_2 &\sim \text{ Negative Binomial}(r = 7, p = \theta) \\
p(x;\theta) &= \binom{x+r-1}{r-1}\theta^r(1-\theta)^x \\
L(\theta) &= \theta^7(1-\theta)^x \quad \text{if } \ c = \frac{1}{\binom{x+r-1}{r-1}} \\
\ell(\theta) &= 7\ln\theta + x\ln(1-\theta) \\
\ell'(\theta) &= \frac{7}{\theta} - \frac{x}{1-\theta} \Rightarrow \hat{\theta} = \frac{7}{x+7} \\
\ell''(\theta) &= -\frac{7}{\theta^2} - \frac{x}{(1-\theta)^2} < 0 \qquad \text{for } 0 < \theta < 1.
\end{aligned}$$

If $x_2 = 93$, then $\hat{\theta} = .07$.

2.2 Likelihoods Based on Frequency Tables

Optional Text Reading: Section 9.1, pp. 8-10

Example 2.2.1. Two hundred specimens of a new high-impact plastic produced using a new 3D printing machine are tested by repeatedly striking them with a hammer until they fracture. Let Y_i be the random number of hits that are required to fracture the $i'th$ specimen so that the observed number of hits satisfies $y_i = 1, 2, 3, ...$ The following table summarizes the results in a Frequency Table. For example, 112 specimens fractured on the first hit and 30 specimens required 4 or more hits to fracture.

# hits required to fracture	1	2	3	≥ 4	Total
# specimens	112	36	22	30	200

Suppose that a specimen has a constant probability, θ, of *surviving a hit*, independently of previous hits received. Find the MLE of θ based on the Frequency Table results for 200 independent specimens. Compare estimated expected frequencies with the observed frequencies.

Note that we have $\begin{cases} \text{- n independent repetitions of an experiment} \\ \text{- each outcome must fall in exactly one category} \end{cases}$

Often data from n *independent* repetitions of an experiment are summarized in a frequency table,

Type of event or category	A_1	A_2	$\ldots$	A_k	Total
Observed frequency	X_1	X_2	$\ldots$	X_k	n

where each outcome of one of the n experiments must fall in *exactly* one category, $A_1, \ldots, A_k$, a partition of the sample space.

- Let $X_i = \#$ of times A_i occurs in n repetitions $[\sum_{i=1}^{k} X_i = n]$
- $p_i = P\{\text{an outcome of one trial is of type } A_i\}$ $[\sum_{i=1}^{k} p_i = 1]$
- $E(X_i) = np_i$

We can add a row in the table corresponding to expected cell frequencies.

Type of event or class	A_1	A_2	$\dots$	A_k	Total
Observed frequency	X_1	X_2	$\dots$	X_k	n
Expected frequency	np_1	np_2	$\dots$	np_k	n

The p_i's may be determined from a probability model that depends on an unknown parameter, θ, so that $p_i = p_i(\theta)$.

The distribution of the frequencies in the table is Multinomial$(n, p_1(\theta), \dots, p_k(\theta))$ and the probability of observing a particular set of frequencies $(x_1, ..., x_k)$ is:

$$P(x_1, \dots, x_k; \theta) = \binom{n}{x_1 x_2 \dots x_k} p_1(\theta)^{x_1} p_2(\theta)^{x_2} \dots p_k(\theta)^{x_k}.$$

The likelihood function is therefore,

$$L(\theta) = p_1(\theta)^{x_1} p_2(\theta)^{x_2} \dots p_k(\theta)^{x_k},$$

and the MLE, $\hat{\theta}$, is the value of θ that maximizes $L(\theta)$.

Using $\hat{\theta}$ we can compute $p_i(\hat{\theta}) = \hat{p}_i$ and $n\hat{p}_i = np_i(\hat{\theta})$, the *estimated expected frequencies*. The estimated expected frequencies are compared with the observed frequencies to give us an indication of how well the probability model fits.

Returning to Example 2.2.1:

$$\begin{aligned} &\text{Let } Y = \#\text{ hits required to fracture a random specimen}\\ p_1(\theta) &= P(Y = 1) = 1 - \theta\\ p_2(\theta) &= P(Y = 2) = \theta(1 - \theta)\\ p_3(\theta) &= P(Y = 3) = \theta^2(1 - \theta)\\ p_4(\theta) &= 1 - p_1 - p_2 - p_3 = 1 - (1 - \theta) - \theta(1 - \theta) - \theta^2(1 - \theta)\\ &= \theta^3 \end{aligned}$$

You may recognize that $Y - 1 \sim$ Negative Binomial$(r = 1, p = \theta)$ which is also known as the Geometric(θ) distribution. We can now write down the distribution of our data and obtain the Likelihood function.

$$P(x_1, x_2, x_3, x_4;\ \theta) = \binom{200}{112, 36, 22, 30} p_1(\theta)^{112}\ p_2(\theta)^{36}\ p_3(\theta)^{22}\ p_4(\theta)^{30}$$

$$\begin{aligned} L(\theta) &= p_1(\theta)^{112}\ p_2(\theta)^{36}\ p_3(\theta)^{22}\ p_4(\theta)^{30} \\ &= [1-\theta]^{112}\ [\theta(1-\theta)]^{36}\ [\theta^2(1-\theta)]^{22}\ [\theta^3]^{30} \\ &= [1-\theta]^{112+36+22}\ [\theta]^{36+2\cdot 22+3\cdot 30} \\ &= [1-\theta]^{170}\ \theta^{170} \end{aligned}$$

$$\ell(\theta) = 170\ \ln(1-\theta) + 170\ \ln\theta$$

$$\ell'(\theta) = S(\theta) = -\frac{170}{1-\theta} + \frac{170}{\theta}$$

$$\ell'(\hat\theta) = 0 \Longrightarrow \hat\theta = \frac{170}{340} = \frac{1}{2}$$

$$\ell''(\theta) = -\frac{170}{(1-\theta)^2} - \frac{170}{\theta^2} < 0 \ \text{ for } \ 0 < \theta < 1$$

Checking the boundary points, $0, 1$, $L(0) = L(1) = 0$ but $L(\theta) > 0$ for $\theta \neq 0, 1$, therefore $\hat\theta = \frac{1}{2}$ is a maximum, that is, it is the MLE of θ.

Substituting in the MLE for θ into the expressions for the $p's$, we obtain,

$$\begin{aligned}
\hat{p}_1 &= p_1(\hat{\theta}) = 1 - \hat{\theta} = \frac{1}{2} \\
\hat{p}_2 &= p_2(\hat{\theta}) = \hat{\theta}(1 - \hat{\theta}) = \frac{1}{4} \\
\hat{p}_3 &= p_3(\hat{\theta}) = \hat{\theta}^2(1 - \hat{\theta}) = \frac{1}{8} \\
\hat{p}_4 &= \hat{\theta}^3 = \frac{1}{8}
\end{aligned}$$

Using these estimates, we obtain estimated expected frequencies $n\hat{p}_i$ under the model:

# hits required to fracture	1	2	3	≥ 4	Total
Observed frequency	112	36	22	30	200
Estimated expected frequency	$200\hat{p}_1$ $= 100$	$200\hat{p}_2$ $= 50$	$200\hat{p}_3$ $= 25$	$200\hat{p}_4$ $= 25$	200

The estimated expected frequencies display poor agreement with the observed frequencies. We expect some variation between the observed and estimated expected frequencies. Does the poor agreement here suggest that something is wrong with the assumed probability model? We need to be able to quantify the differences between observed and estimated expected frequencies and decide if these are due to chance variation only or to an inappropriate model. It may be that the assumed model is incorrect, for example, the assumption of a constant probability of surviving a blow independently of previous blows may not be realistic.

2.3 Unusual example

There are some examples for which we cannot use Calculus to compute the maximum likelihood estimate. Here is one such example.

Example 2.3.1. The 'enemy' has an unknown number, N, drones, which have been numbered $1, 2, \dots N$. Spies have reported sighting 8 drones with numbers 137, 24, 86, 33, 92, 129, 17, 111. Assume that sightings are independent and that each of the drones has probability $\frac{1}{N}$ of being observed at each sighting. Find $\hat{N}$.

$$P(137, 24, 86, 33, 92, 129, 17, 111; N) = \begin{cases} \frac{1}{N^8} & \text{if } N \geq \max\{137, 24, 86, \ldots, 111\} \\ 0 & \text{otherwise} \end{cases}$$

As N decreases, $P(137, 24, 86, 33, 92, 129, 17, 111;\ N)$ increases provided that $N \geq 137$. Therefore, to maximize the probability of the observed data assuming this model, we need to make N as small as possible subject to $N \geq \max\{137, 24, 86, \ldots, 111\}$. Therefore, $\hat{N} = 137$ is the MLE of N. This is an example where we do NOT use Calculus to solve for the MLE.

2.4 Combining Independent Events

Optional Text Reading: Section 9.2

Example 2.4.1. Suppose that for Example 2.1.1 we observed the number of illegal parcels on each of two days, so that we observe

$$\begin{aligned} X_1 &= \#\text{ illegal parcels out of } n = 100 \text{ on day 1, and} \\ X_2 &= \#\text{ illegal parcels out of } n = 100 \text{ on day 2.} \end{aligned}$$

Assuming that the numbers of illegal parcels for day 1 is independent of the number of illegal parcels for day 2, we can write the JOINT probability mass function (pmf) for X_1 and X_2 as:

$$p(x_1, x_2; \theta) = \binom{100}{x_1} \theta^{x_1} (1-\theta)^{100-x_1} \binom{100}{x_2} \theta^{x_2} (1-\theta)^{100-x_2}.$$

The Likelihood function for θ now uses both data values and becomes,

$$\begin{aligned} L(\theta) &= \theta^{x_1} (1-\theta)^{100-x_1} \times \theta^{x_2} (1-\theta)^{100-x_2} \\ &= \theta^{(x_1+x_2)} \ (1-\theta)^{(200-x_1-x_1)} \end{aligned}$$

and the Log-likelihood function is,

$$\ell(\theta) = \ln L(\theta) = (x_1 + x_2) \ln \theta + (200 - x_1 - x_2) \ln(1-\theta).$$

We can proceed as above and obtain the Maximum Likelihood Estimate of θ. As an exercise, show that the MLE is $\hat{\theta} = (x_1 + x_2)/200$.

2.5 Relative Likelihood

Optional Text Reading: Section 9.3

In Example 2.1.1, we estimated θ, the probability that a randomly chosen parcel is illegal in a Binomial experiment, $X \sim \text{Binomial}(n = 100, \theta)$. We computed the MLE as $\hat{\theta} = x/100 = .07$ when $x = 7$. We know this to be the **most plausible** value of θ in the sense that it maximizes the probability of the observed data, assuming the Binomial model. Ultimately, we want to obtain a set of plausible values for θ which incorporate the variability in the data as described by the model.

Questions:

(1) What about $\theta = .06$? Is this a reasonable or plausible value for θ given the data we have?

(2) How can we produce a set of θ-values that are plausible given the data?

The relative plausibilities of other θ-values may be examined by comparing them with $\hat{\theta}$, the MLE.

Definition: The **Relative Likelihood function** (RLF) of θ is

$$\boxed{R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}.}$$

Since $L(\theta) = cp(x;\ \theta)$ where c does not depend on θ, then

$$R(\theta) = \frac{cL(\theta)}{cL(\hat{\theta})} = \frac{p(x;\ \theta)}{p(x;\ \hat{\theta})}.$$

Since $0 \le L(\theta) \le L(\hat{\theta})$ for all θ, then $0 \le R(\theta) \le 1$. □

Definition: The **Log Relative Likelihood function** of θ is

$$\boxed{r(\theta) = \ln R(\theta) = \ln L(\theta) - \ln L(\hat{\theta}) = \ell(\theta) - \ell(\hat{\theta}).}$$

Since $0 \leq R(\theta) \leq 1$, then $-\infty \leq r(\theta) \leq 0$. □

Note:

$$R(\theta_1) = \frac{cL(\theta_1)}{cL(\hat{\theta})} = \frac{\text{Probability of data when } \theta = \theta_1}{\text{max Probability of data for any value } \theta}.$$

- If $R(\theta_1) = 0.1$ then the data are 10 times more probable when $\theta = \hat{\theta}$ than when $\theta = \theta_1$, under the hypothesized model.
- If $R(\theta_2) = 0.5$, then the data are 2 times more probable when $\theta = \hat{\theta}$ than when $\theta = \theta_2$, under the hypothesized model.
- θ_2 is a more plausible parameter value than θ_1.
- $R(\theta)$ gives us a way of assessing and generating plausible values of θ given the data and the hypothesized model.
- For example, $\{\theta | R(\theta) \geq 0.5\}$ is a set of θ values that give the data at least 50% of the maximum possible probability under the hypothesized model.

Definition: A 100 p% **Likelihood interval** (LI) for θ is the set of θ values such that,

$$\boxed{R(\theta) \geq p \text{ or equivalently } \ln R(\theta) = r(\theta) \geq \ln p.}$$

Likelihood Interval Guidelines:

θ-values inside 10% LIs are referred to as plausible
50% LIs are referred to as very plausible

Question: Is a 10% LI contained in a 50% LI, or is a 50% LI contained in a 10% LI?

Likelihood intervals are similar, in practice, to Confidence Intervals, and we will see that they are mathematically related when the data are normally distributed. As an example, in the one-sample normal case when σ is known, the 14.7% Likelihood interval for the unknown mean, μ, corresponds to a 95% confidence interval.

Relative Likelihood is also used for Hypothesis Testing/Tests of Significance which we will see in Chapter 4.

Any report of results of an experiment should include $\hat{\theta}$ as well as an interval estimate such as a likelihood interval or a confidence interval.

Returning to Example 2.1.1, we construct a 100 p% LI for θ. We want to find all θ values such that $R(\theta) \geq p$, where $R(\theta) = L(\theta)/L(\hat{\theta})$ and

$$L(\theta) = \theta^x (1-\theta)^{n-x}.$$

Here $x = 7,\ \hat{\theta} = \frac{7}{100}$ and $L(\hat{\theta}) = (\frac{7}{100})^7 (\frac{93}{100})^{93}$.

To compute a 100 p% Likelihood interval, we want all θ such that

$$R(\theta) = \frac{\theta^7 (1-\theta)^{93}}{(\frac{7}{100})^7 (\frac{93}{100})^{93}} \geq p.$$

Equivalently, we can find the values θ such that $r(\theta) = \ln R(\theta) \geq \ln(p)$.

Here we find the roots θ of $r(\theta) - \ln(p) = 0$, where θ is in the interval $(0, 1)$ using the R function `uniroot()`. To use `uniroot()`, we need to supply the function that we wish to solve and starting values that bracket the roots. To determine starting values, we graph $r(\theta) - \ln(p)$ versus θ and overlay a horizontal line at zero. We give an example using $p = 0.1$ for a 10% Likelihood Interval.

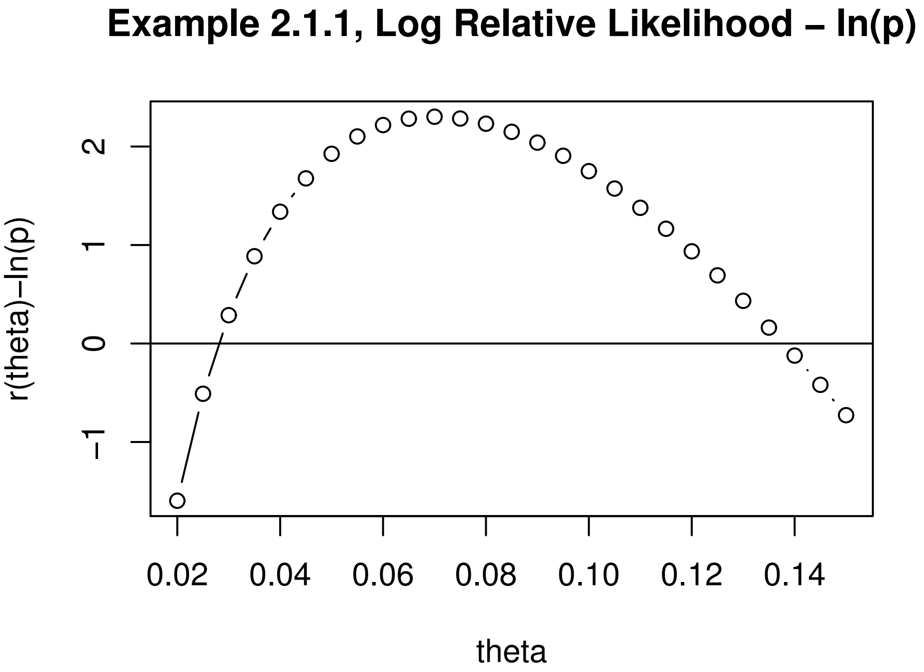


Figure 2.1: 10% Likelihood interval construction. The log relative likelihood minus ln(0.1) is plotted versus θ. A horizontal line at zero is overdrawn to assist with starting value determination.

From the graph, we see that there are two roots. The lower root lies within the interval [0.02, 0.04] and the upper root lies within the interval [0.1, 0.15]. These are the starting values that we supply to `uniroot()` in the code in the next section. The roots that R returned are in the $root slot below. The 10% Likelihood interval for θ is thus: (0.028, 0.138) and the MLE is $\hat{\theta} = 0.07$. Note that this interval is NOT symmetric about the value 0.07 with the right endpoint further from 0.07 than the left endpoint. This is displayed in the asymmetry of the plot above. [Aside: the 50% Likelihood interval for θ is (0.044, 0.10).]

```
> lower <- uniroot(logR.m.lnp, c(.02, .04), thetahat$maximum, p)
> lower
$root
[1] 0.02804908
$f.root
[1] 0.004036655
$iter
[1] 4
$init.it
[1] NA
$estim.prec
[1] 6.103516e-05

> upper <- uniroot(logR.m.lnp, c(.1, .15), thetahat$maximum, p)
> upper
$root
[1] 0.1378683
$f.root
[1] -1.041713e-05
$iter
[1] 4
$init.it
[1] NA
$estim.prec
[1] 6.103516e-05
```

Figure 2.2: R Output: Likelihood interval computations for Example 2.1.1

In Example 2.2.1, we worked through an example involving specimens of a new high impact plastic which were tested repeatedly by striking them with a hammer until they fractured. In that example,

$$\theta = P\{\text{specimen survives a hit independently of } \text{~~blows~~ hits received}\}$$

and the data are given again below.

# hits required to fracture	1	2	3	≥ 4	Total
# specimens	112	36	22	30	200

We computed the MLE, $\hat{\theta} = 0.5$

We know that $\hat{\theta} = 0.5$ is the **most plausible** value of θ in the sense that it maximizes $P(x_1, x_2, x_3, x_4;\ \theta)$, the probability of the observed data given θ.

We construct a 100 p% LI for θ. We want to find all θ values such that $R(\theta) \geq p$, where $R(\theta) = L(\theta)/L(\hat{\theta})$ and

$$\begin{aligned} L(\theta) &= p_1^{112}\ p_2^{36}\ p_3^{22}\ p_4^{30} \\ &= [1-\theta]^{170}\theta^{170}, \end{aligned}$$

$\hat{\theta} = \frac{1}{2}$ and so $\quad L(\hat{\theta}) = (\frac{1}{2})^{340}$.

To compute a 100 p% Likelihood interval, we want all θ such that

$$\begin{aligned} R(\theta) &= \frac{(1-\theta)^{170}\ \ \theta^{170}}{\left(\frac{1}{2}\right)^{340}} \geq p,\ \text{ or} \\ R(\theta) &= \left[4\,(1-\theta)\,\theta\right]^{170} \geq p. \end{aligned}$$

To solve this problem, we find the roots θ of $R(\theta) - p = 0$, for admissible values of θ in the interval $(0, 1)$. Below, we tabulate values for $R(\theta)$ and $r(\theta)$ for various values of θ to help us discern starting values for numerical root finding software.

		θ	$R(\theta)$			$r(\theta)$		
		.3	1.34×10^{-13}			-29.64		
		.4	.00968			-6.94		
		.45	.1811	$\longleftarrow$	10%	-1.71		
		.46	.3357			-1.09		
		.47	.5417	$\longleftarrow$	50%	$-.61$		
interval surrounding $\hat{\theta}$	$\longrightarrow$	.50	1			0		
		.53	.5417	$\longleftarrow$	50%	$-.61$		
		.54	.3357			-1.09		
		.55	.1811	$\longleftarrow$	10%	-1.71		
		.60	.00968			-6.94		
10%		.44	.0849			$\Longrightarrow$	$[.442, .558]$	10% LI
		.442	.10					
50%		.469	.5196			$\Longrightarrow$	$[.468, .532]$	50% LI
		.468	.4977					

Equivalently, we could have used $r(\theta)$ to compute the Likelihood Intervals. For a 100 p% Likelihood Interval we want all θ such that, $r(\theta) \geq \ln p$. To compute the endpoints of the interval, we solve the lower and upper roots of the equation in θ, $r(\theta) - \ln p = 0$ using the R code below:

$$\begin{array}{ll} \text{50\% LI} & r(\theta) = \ln 0.5 = -0.69 \\ \text{10\% LI} & r(\theta) = \ln 0.1 = -2.30, \end{array}$$

$$\begin{aligned} \text{where } r(\theta) &= \ell(\theta) - \ell(\hat{\theta}) \\ \ell(\theta) &= 170 \ln(1-\theta) + 170 \ln \theta \\ \ell(\hat{\theta}) &= 340 \ln 0.5 = -235.67. \end{aligned}$$

Alternatively (or additionally) a graph of the log relative likelihood can aid in choosing starting values for a root finding technique. Below is the graph of the log relative likelihood minus ln(.1) for Example 2.2.1.

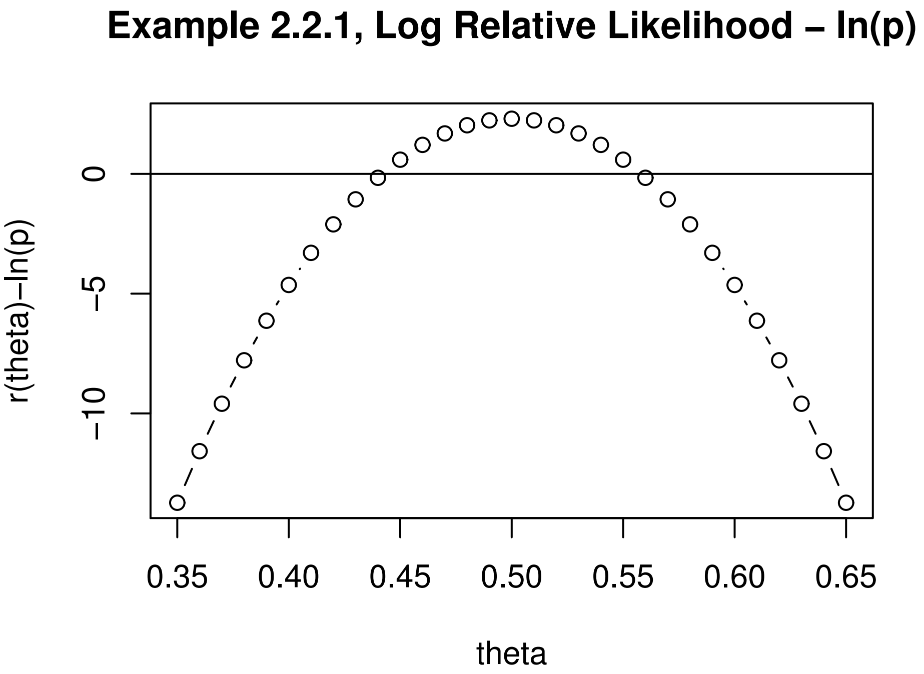


Figure 2.3: 10% Likelihood interval construction for example 2.2.1. The log relative likelihood - ln(.1) is plotted versus theta. A horizontal line at zero is overdrawn to assist with starting value determination.

2.5.1 R code for Example 2.1.1

```
# Example 2.1.1
# Log-likelihood function
ell <- function(theta){
  7*log(theta) + 93*log(1-theta)
}
theta <- seq(.02, .15,by=.005)
plot(theta,ell(theta),ylab='log likelihood',xlab='theta')
title('Example 2.1.1, Log-Likelihood')
```

```
#MLE of theta
thetahat <- optimize(ell, c(.05,.09), maximum=TRUE)
thetahat

#Log relative likelihood function
logR <- function(theta, thetahat){
  ell(theta) - ell(thetahat)
}
logR(theta,thetahat$maximum)
p <- .1  #10% likelihood interval
logR.m.lnp <- function(theta, thetahat, p) {logR(theta,thetahat)-log(p)}

plot(theta,logR.m.lnp(theta,thetahat$maximum, p), ylab='r(theta)-ln(p)',
     xlab='theta', type='b')
abline(h=0)
title('Example 2.1.1, Log Relative Likelihood - ln(p)')


#Likelihod intervals
lower <- uniroot(logR.m.lnp, c(.02, .04), thetahat$maximum, p)
lower

upper <- uniroot(logR.m.lnp, c(.1, .15), thetahat$maximum, p)
upper
```

2.5.2 R code for Example 2.2.1

```
# Example 2.2.1
# Log-likelihood function and plot
ell <- function(theta){
  170*log(theta) + 170*log(1-theta)
}
```

```
theta <- seq(.35,.65,by=.01)
plot(theta,ell(theta),ylab='log-likelihood',xlab='theta')
title('Example 2.2.1, Log-likelihood')


#MLE of theta - looks for maximum in interval (.4, .6)
thetahat <- optimize(ell, c(.4,.6), maximum=TRUE)
thetahat

#Log relative likelihood function and plot
logR <- function(theta, thetahat){
  ell(theta) - ell(thetahat)
}
logR(theta,thetahat$maximum)
p <- .1  #10% likelihood interval
logR.m.lnp <- function(theta, thetahat, p) {logR(theta,thetahat)-log(p)}

plot(theta,logR.m.lnp(theta,thetahat$maximum,p), ylab='r(theta)-ln(p)',
     xlab='theta', type='b')
title('Example 2.2.1, Log Relative Likelihood - ln(p)')
abline(h=0)
#The plot helps us to determine starting values for a root finding
#  technique used to solve for the Likelihood interval

#Likelihod intervals
#log relative likelihod minus ln(p)
#Find a root in the interval (.4, .5)

lower <- uniroot(logR.m.lnp, c(.4, .5), thetahat$maximum, p)
lower

#Find a root in the interval (.5, .6)
upper <- uniroot(logR.m.lnp, c(.5, .6), thetahat$maximum, p)
upper
```

2.6 Likelihood for Continuous Models

Optional Text Reading: Section 9.4

Suppose that have a sample $x_1, x_2, \ldots, x_n$ of independent observations on a continuous random variable X which has pdf $f(x;\theta)$ where θ is an unknown parameter.

Example 2.6.1. Times between successive failures X, of a computer system are thought to be independent and identically distributed (iid) random variables with an exponential distribution having mean θ.

Let $X_1, X_2 \ldots, X_n$ represent a random sample of observed times between successive failures so that

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

In the case of continuous measurements, the pdf evaluated at x does not represent the probability of observing x, however, we construct the likelihood using the pdf. In this case, the joint pdf of the independent sample is written as the product of the marginal pdf's and,

$$L(\theta) = \underset{\substack{\uparrow \\ \text{product symbol}}}{\prod_{i=1}^{n}} f(x_i;\theta).$$

$$= f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta)$$

Substituting in the example,

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\
&= \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \theta^{-n} e^{-\Sigma x_i/\theta} \\
\ell(\theta) &= -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i \\
S(\theta) = \ell'(\theta) &= -\frac{n}{\theta} + \frac{\Sigma x_i}{\theta^2} \Longrightarrow \hat{\theta} = \frac{\Sigma x_i}{n} = \bar{x}, \quad \text{sample mean!} \\
\ell''(\theta) &= \frac{n}{\theta^2} - \frac{2\Sigma x_i}{\theta^3} = \frac{n}{\theta^2} - \frac{2n\bar{x}}{\theta^3} \\
\ell''(\hat{\theta}) &= \frac{n}{\bar{x}^2} - \frac{2n\bar{x}}{\bar{x}^3} = \frac{n}{\bar{x}^2} - \frac{2n}{\bar{x}^2} = -\frac{n}{\bar{x}^2} \\
I(\hat{\theta}) &= \frac{n}{\bar{x}^2} > 0 \Longrightarrow \hat{\theta} = \bar{x} \quad \text{ is the MLE}
\end{aligned}$$

Suppose that we observed the following times between failures (to the nearest day):

70 11 66 5 20 4 35 40 29 8

1. What is an estimate of the expected time between failures? ($\sum x_i = 288$)
2. What values are plausible given the data? (10% and 50% LI's)
3. The computer manufacturer claims that the mean time between failures is 100 days. Comment.

Solution:

1. $X \sim \exp(\theta)$

   $\theta =$ Expected time between failures The estimated expected time between failures is, $\widehat{\theta} = \frac{\Sigma x_i}{n} = \frac{288}{10} = 28.8$.

2. Using $r(\theta)$, we obtain likelihood intervals for θ as follows:

$$\begin{aligned} r(\theta) &= \ln R(\theta) = \ln \frac{L(\theta)}{L(\hat{\theta})} \\ &= \ell(\theta) - \ell(\hat{\theta}) \\ \ell(\hat{\theta}) &= -n \ln \hat{\theta} - \frac{1}{\hat{\theta}} \Sigma x_i = -10 \ln 28.8 - 10 \\ &= -43.60 \\ \text{Plot } r(\theta) - \ln(p) &= -10 \ln \theta - \frac{288}{\theta} + 43.60 - \ln(p) \quad \text{versus} \quad \theta. \end{aligned}$$

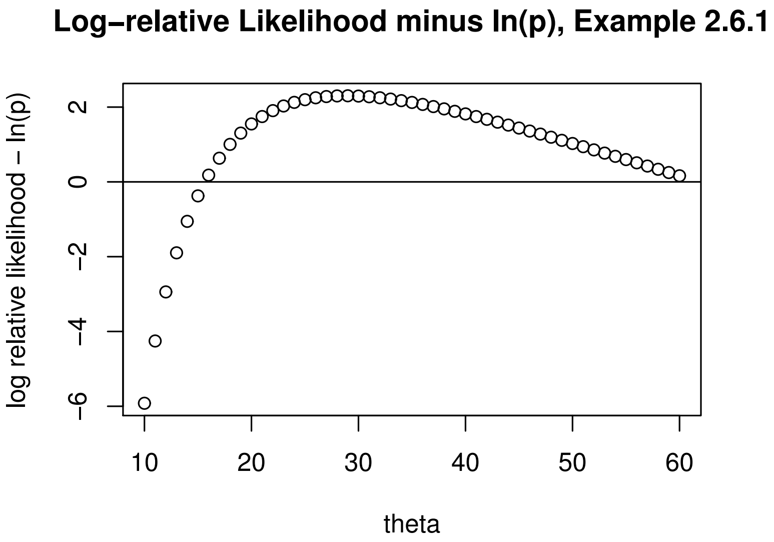


Figure 2.4: 10% Likelihood interval. Log relative likelihood minus ln(0.1) plotted versus θ. Horizontal line is at zero.

The likelihood intervals are:

$$\begin{aligned} 50\% \text{ LI} &: 20.28 \leq \theta \leq 42.83 \\ 10\% \text{ LI} &: 15.65 \leq \theta \leq 61.88 \end{aligned}$$

The data do not support the claim that the mean time between failures is 100 days. 100 days is not a plausible value for θ.

2.6.1 R code for Example 2.6.1

```
# Example 2.6.1

x <- c(70 , 11 , 66 , 5 , 20 , 4 , 35 , 40 , 29 , 8)
ell <- function(theta,x){
  n <- length(x)
  return(-n*log(theta) - sum(x)/theta)
}
theta <- seq(10,60,by=1)
plot(theta,ell(theta,x),ylab='log likelihood',xlab='theta')
title('Log Likelihood, Example 2.6.1')

#MLE
thetahat <- optimize(ell, c(20,40), maximum=TRUE, x=x)
thetahat
ell(thetahat$maximum,x)

#Log relative likelihood function
logR <- function(theta, thetahat,x){
  ell(theta,x) - ell(thetahat,x)
}
logR(theta,thetahat$maximum,x)
p <- .1  #10% likelihood interval
#log relative likelihood minus ln(p)
logR.m.lnp <- function(theta, thetahat, x, p) {logR(theta,thetahat,x)-log(p)}

plot(theta,logR.m.lnp(theta,thetahat$maximum,x,p),
     ylab='log relative likelihood - ln(p)',xlab='theta')
abline(h=0)    #add a horizontal line at zero
title('Log-relative Likelihood minus ln(p), Example 2.6.1')


#Likelihod intervals
lower <- uniroot(logR.m.lnp, c(10,20), thetahat$maximum, x, p)
lower
upper <- uniroot(logR.m.lnp, c(50,70), thetahat$maximum, x, p)
upper
```

```
p <- .5  #50% likelihood interval
lower <- uniroot(logR.m.lnp, c(10,25), thetahat$maximum, x, p)
lower
upper <- uniroot(logR.m.lnp, c(40,60), thetahat$maximum, x, p)
upper
```

2.7 Invariance

Optional Reading: Section 9.6

In the above Example 2.6.1, we might also be interested in estimating the probability that the time between failures is greater than 100 days, i.e.

$$\begin{aligned}\beta = P(X > 100) &= \int_{100}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx \\ &= e^{-100/\theta}\end{aligned}$$

How can we find the MLE of β?

Solution: We can reparameterize the Log-likelihood in terms of $\theta(\beta) = \frac{-100}{\ln \beta}$. Note that β is an INCREASING function of θ. [Exercise: Show that the derivative of β with respect to θ is positive for all $\beta \in (0,1)$.]

The Log-likelihood and Score function for β are:

$$\begin{aligned}\ell(\beta) &= -n \ln\left(\frac{-100}{\ln \beta}\right) + \frac{\ln \beta}{100}\left(\sum_{i=1}^{n} x_i\right), \\ \ell'(\beta) &= \frac{-n}{\left(\frac{-100}{\ln \beta}\right)} \frac{100}{(\ln \beta)^2} \frac{1}{\beta} + \frac{1}{100\beta}\left(\sum_{i=1}^{n} x_i\right).\end{aligned}$$

Setting the derivative to zero and solving [as an exercise] yields,

$$\hat{\beta} = e^{-100/\hat{\theta}} = 0.031.$$

The estimated probability that the time between failures is greater than 100 days is 0.031, very small.

We see that maximum Likelihood Estimates have some very nice properties - MLE's are invariant under one-to-one parametric transformations. In addition, a 10% Likelihood interval for β is $(e^{-100/\theta_1}, e^{-100/\theta_2}) = (e^{-100/15.65}, e^{-100/61.88}) = (0.0017, 0.199)$, where θ_1 and θ_2 are the endpoints of the 10% Likelihood interval for θ.

Example 2.7.1. Family income, X is measured on a scale such that $X = 1$ corresponds to a subsistence level income. The pdf of the income distribution is assumed to be the Pareto distribution

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

where $\theta > 0$. Data for a random sample of $n = 10$ families living in Toronto is: $1.02, 1.41, 1.75, 2.31, 3.42, 4.31, 9.21, 17.4, 38.6, 392.8$.

(a) Find the MLE of θ.

(b) Obtain an estimate of the median family income, β.

Figure 2.5 graphs the density of the Pareto distribution for various values of θ. Typically, there are many individuals with smaller incomes and only a few individuals with large incomes, and this is the behaviour displayed by the Pareto densities.

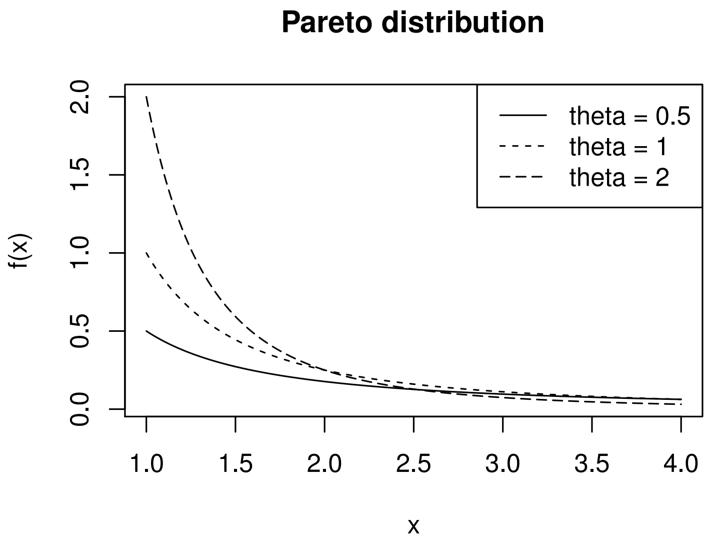


Figure 2.5: The Pareto density, for values of θ

Solution:

(a) Find the MLE of θ.

$$f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \theta x_i^{-(\theta+1)}$$

$$L(\theta) = \theta^n \prod_{i=1}^{n} x_i^{-(\theta+1)}$$

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

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i \Longrightarrow \hat{\theta} = \frac{n}{\Sigma_{i=1}^{n} \ln x_i} = \frac{1}{1.92} = 0.52208$$

$$\ell''(\theta) = \frac{n}{\theta^2} < 0 \text{ for all } \theta.$$

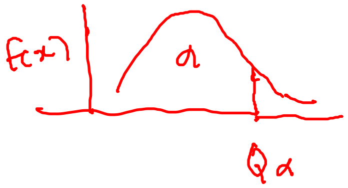
A 10% Likelihood interval for θ is (0.24, 0.96)

(b) Find an estimate of the median family income, β.

Definition: The $100\alpha^{\textbf{th}}$ **percentile** of continuous X is the variate value Q_α such that

$$\boxed{P\left(X \leq Q_\alpha\right) = F\left(Q_\alpha\right) = \alpha}$$

where $0 < x < 1$ and $F(x)$ is the cdf of X. The **Median** is the 50th percentile. □



Returning to Example 2.7.1,

$$\begin{aligned} 0.5 = P\left(X \leq \beta\right) &= \int_1^\beta \theta x^{-(\theta+1)dx} \\ &= -x^{-\theta}\Big|_1^\beta = 1 - \beta^{-\theta} \\ 0.5 = \beta^{-\theta} \quad &\Longrightarrow \quad \beta = 0.5^{-1/\theta} = 2^{1/\theta} \end{aligned}$$

Note that β is a $1-1$ function of θ!

Substituting $\theta = \ln 2 / \ln \beta$ into $L(\theta)$,

$$\begin{aligned} L\left(\theta\right) &= L\left(\frac{\ln 2}{\ln \beta}\right) \\ &= \left(\frac{\ln 2}{\ln \beta}\right)^n \prod x_i^{-(\ln 2/\ln \beta + 1)} \\ &= L_*\left(\beta\right) \end{aligned}$$

We can find $\hat{\beta}$ by maximizing $L_*\left(\beta\right)$. As an exercise, show that the maximizer of $L_*\left(\beta\right)$ is $\hat{\beta} = 2^{1/\hat{\theta}}$

MLE's are invariant under one-to-one parametric transformations. Also, $R\left(\theta\right) = R_*\left(\beta\right)$ where $\theta = \ln 2/\ln \beta$. Relative plausibilities do not depend upon the parametrization.

Definition: Invariance Property. Let $\theta = g\left(\beta\right)$ be a one-to-one transformation of β. Let $\hat{\theta}$ be the MLE of θ and $\theta_1 \leq \theta \leq \theta_2$ be a 100p% likelihood interval for θ.

The MLE of β is $\hat{\beta} = g^{-1}(\hat{\theta})$ and a 100p% Likelihood Interval for β is:

$$\begin{array}{ll} g^{-1}(\theta_1) \leq \beta \leq g^{-1}(\theta_2) & \text{if } g \text{ is monotone increasing} \\ g^{-1}(\theta_2) \leq \beta \leq g^{-1}(\theta_1) & \text{if } g \text{ is monotone decreasing} \quad \square \end{array}.$$

Returning to Example 2.7.1, the MLE of β is $\hat{\beta} = 2^{1/\hat{\theta}} = 2^{\sum \ln x_i / n} = 3.78$

Given that a 10% LI for θ is

$$0.24 \leq \theta \leq 0.96,$$

a 10% LI for β is (β monotone decreasing)

$$\begin{array}{c} 2^{1/0.96} \leq \beta \leq 2^{1/0.24} \\ 2.06 \leq \beta \leq 17.96, \end{array}$$

a set of plausible values for median income (relative to subsistence level) based on the Toronto sample data. □

Some Comments:

- What is the effect of increasing sample size, n on Likelihood intervals? Generally this produces a more sharply peaked likelihood which results in narrower Likelihood intervals. Likelihood intervals for θ will be more precisely estimated.
- Can we combine data from independent experiments or studies? Suppose we are given a random sample of family incomes (relative to subsistence level) for families living in London, England where it is assumed that the pdf of the income distribution is

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

When would it be appropriate to pool this data with the Toronto data and produce a common estimate of θ?

We can estimate θ for each sample, $\hat{\theta}_{\text{Toronto}}, \hat{\theta}_{\text{London}}$.

Plot the two log-relative likelihoods $r_{\text{Toronto}}(\theta)$, $r_{\text{London}}(\theta)$, on the same graph and look for values of θ that are plausible for both locations. If there are some, combine the two sets of data to produce a common estimate of θ.

2.7.1 R code for Example 2.7.1

```
#Example 2.7.1 Family Income Pareto distribution with data
# Pareto density plot
dpareto <- function(x,theta){
  theta*x^(-theta-1)
}
xx<-seq(1,4,by=.01)
pareto.dat<-cbind(dpareto(xx,.5), dpareto(xx,1), dpareto(xx,2))
matplot(xx, pareto.dat, type='l',col=1, lty=c(1,2,5),
        main='Pareto distribution',
        ylab='f(x)',xlab='x')
legend("topright",c(paste('theta',c(.5, 1, 2))),lty=c(1,2,5),col=1)
#We have an algebraic expression for the MLE, thetahat
x<- c(1.02, 1.41, 1.75, 2.31, 3.42, 4.31, 9.21, 17.4, 38.6, 392.8)
n <- length(x)
n
thetahat <- n/sum(log(x))
thetahat
ell <- function(theta,x){
  n <- length(x)
  ellres <- vector('numeric',length(theta))
  for(i in (1:length(theta)))
  { ellres[i] <- n*log(theta[i]) - sum((theta[i]+1)*log(x))
  }
  return(ellres)
}
### use this one!!! No loops
ell2 <- function(theta,x){
  n <- length(x)
  ellres <- n*log(theta) - (theta+1)*sum(log(x))
  return(ellres)
}

#Graph the Log Relative likelihood function
theta <- seq(.1,2,by=.01)
logR <- function(theta, thetahat,x){
  ell(theta,x) - ell(thetahat,x)
}
```

```
p <- .1 #10% likelihood interval
#log relative likelihood minus ln(p)
logR.m.lnp <- function(theta, thetahat, x, p) {logR(theta,thetahat,x)-log(p)}

plot(theta,logR.m.lnp(theta,thetahat,x,p),
    ylab=’log relative likelihood - ln(p)’,xlab=’theta’)
abline(h=0) #add a horizontal line at zero
title(’Log-relative Likelihood minus ln(p), Example 2.7.1’)

#Likelihod intervals
lower <- uniroot(logR.m.lnp, c(.2,.5), thetahat, x, p)
lower
upper <- uniroot(logR.m.lnp, c(.6,1.2), thetahat, x, p)
upper
#MLE and Likelihood intervals for beta, the median
2^(1/thetahat)
2^(1/c(lower$root,upper$root))
```

Chapter 3

Two Parameter Likelihoods

3.1 Maximum Likelihood Estimation

Optional Reading: Section 10.1

Example 3.1.1. Suppose $x_1, x_2, \ldots, x_n$ are independent observations on a random variable, $X \sim N(\mu, \sigma^2)$, where both μ, σ^2 are unknown. Find the joint MLE $(\hat{\mu}, \hat{\sigma}^2)$.

Solution:

$$L\left(\mu, \sigma^2\right) = c\prod_{i=1}^{n} f\left(x_i; \mu, \sigma^2\right)$$

where $f(x; \mu, \sigma^2)$ is the pdf of X.

$$\begin{aligned}
f\left(x; \mu, \sigma^2\right) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right\} \\
L\left(\mu, \sigma^2\right) &= c\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i-\mu\right)^2\right\} \\
&= \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\} \\
\ell\left(\mu, \sigma^2\right) &= -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2
\end{aligned}$$

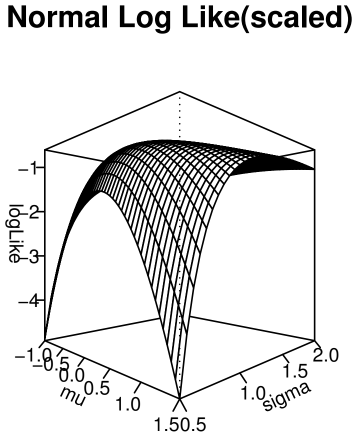


Figure 3.1: Log Likelihood for a random sample of size 100 from Normal(0,1)

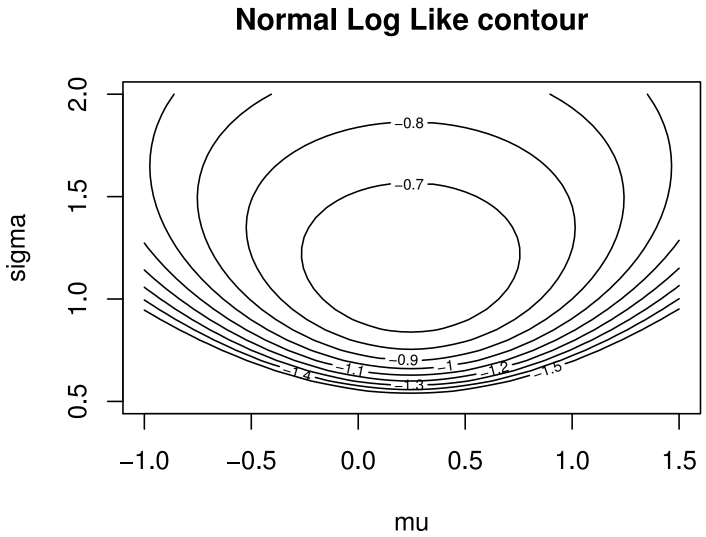


Figure 3.2: Log Likelihood contour plot for a random sample of size 100 from Normal(0,1)

Figures 3.1 and 3.2 display the Log-likelihood as a function of μ and σ for a sample of size 100 data values generated from the Normal(0,1) distribution. The figures show that the Log-likelihood is maximized somewhere near $\mu = 0.25$ and $\sigma = 1.1$.

Find the values $(\hat{\mu}, \hat{\sigma}^2)$ that maximize $\ell\left(\mu, \sigma^2\right)$. To do so, we take derivatives of $\ell(\mu, \sigma^2)$ with respect to μ and σ.

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \tag{3.1}$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.2}$$

At the joint maximizer, $(\hat{\mu}, \hat{\sigma}^2)$, both (3.1) and (3.2) are 0.

$(3.1) = 0 \Longrightarrow \sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \Longrightarrow \hat{\mu} = \sum_{i=1}^{n} x_i/n = \bar{x}$.

Substituting (3.1) into (3.2):

$(3.2) = 0 \Longrightarrow n/\hat{\sigma} = \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \Longrightarrow n\hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$
$\Longrightarrow \hat{\sigma}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$

Checking for a Maximum

The second derivatives, $\partial^2\ell/\partial\mu^2$ and $\partial^2\ell/\partial\sigma^2$ only give information in the direction of the μ, σ axes. We need criteria that tests for a maximum in all directions radiating from $\hat{\mu}, \hat{\sigma}$. We provide the criteria here, and a proof is given in the optional text pages 90 and 91.

As in the one parameter case, we define the **Observed Information matrix**, $I(\mu, \sigma^2)$,

$$\begin{aligned} I\left(\mu, \sigma^2\right) &= \begin{bmatrix} -\partial^2\ell/\partial\mu^2 & -\partial^2\ell/\partial\mu\partial\sigma \\ -\partial^2\ell/\partial\sigma\partial\mu & -\partial^2\ell/\partial\sigma^2 \end{bmatrix} \\ &= \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}. \end{aligned}$$

Note that $\partial^2\ell/\partial\sigma\partial\mu = \partial^2\ell/\partial\mu\partial\sigma$, and so $I_{21} = I_{12}$. At a relative maximum, $\left(\hat{\mu}, \hat{\sigma}^2\right), I\left(\hat{\mu}, \hat{\sigma}^2\right)$ must satisfy:

$$\begin{cases} \widehat{I}_{11} = I_{11}\left(\hat{\mu}, \hat{\sigma}^2\right) > 0 \\ \widehat{I}_{22} > 0 \\ \widehat{I}_{11}\widehat{I}_{22} - \widehat{I}_{21}^2 > 0 \end{cases}$$

In the example,

$$\begin{aligned} I_{11} &= -\frac{\partial^2\ell}{\partial\mu^2} = \frac{n}{\sigma^2} \\ I_{12} &= -\frac{\partial^2\ell}{\partial\mu\partial\sigma} = \frac{2\sum\left(x_i-\mu\right)}{\sigma^3} \\ I_{22} &= -\frac{n}{\sigma^2} + \frac{3}{\sigma^4}\sum\left(x_i-\mu\right)^2 \end{aligned}$$

Substituting in $\hat{\mu}$ and $\hat{\sigma}^2$,

$$\begin{aligned} \widehat{I}_{11} &= \frac{n^2}{\sum\left(x_i-\bar{x}\right)^2} > 0 \qquad \widehat{I}_{12} = 0 \\ \widehat{I}_{22} &= -\frac{n}{\hat{\sigma}^2} + \frac{3n\hat{\sigma}^2}{\hat{\sigma}^4} = \frac{2n}{\hat{\sigma}^2} > 0 \\ \widehat{I}_{11}\widehat{I}_{22} - \widehat{I}_{21}^2 &> 0 \Longrightarrow \left(\hat{\mu}, \hat{\sigma}^2\right) \text{ is the joint MLE} \end{aligned}$$

Example 2.2.1 revisited: Specimens of a new high impact plastic are tested by repeatedly striking them with a hammer until they fracture.

$$Y = \#\text{ blows required to fracture a specimen}$$

We assumed that:

$$P(Y=y)=\theta^{y-1}(1-\theta) \qquad x=1,2,3,\ldots$$

where $\theta = P$(surviving a hit independently of previous hits). Y has a geometric distribution. The assumption does not seem tenable and we found that the model did not yield estimated expected frequencies that were close to the observed frequencies.

It is suggested that, while the geometric distribution applies to most specimens, a fraction $1-\lambda, 0<\lambda<1$, of them will have flaws and always fracture on the first hit.

Compute estimates of λ, θ and compare observed and estimated expected frequencies under the model.

# hits required	1	2	3	≥ 4	Total
# specimens	112	36	22	30	200
	x_1	x_2	x_3	x_4	n

Solution: We need to construct the probability of the observed frequencies as a function of λ and θ.

Recall that we have 200 repetitions of an experiment where each outcome falls in one of the above categories. We modelled the probability of the observed frequencies using a Multinomial distribution.

Let $x_i =$ the number of specimens in category i; $\displaystyle\sum_{i=1}^{4} x_i = n = 200$

$$P(x_1,x_2,x_3,x_4)=\binom{200}{x_1\ x_2\ x_3\ x_4} p_1^{x_1}\ p_2^{x_2}\ p_3^{x_3}\ p_4^{x_4}$$

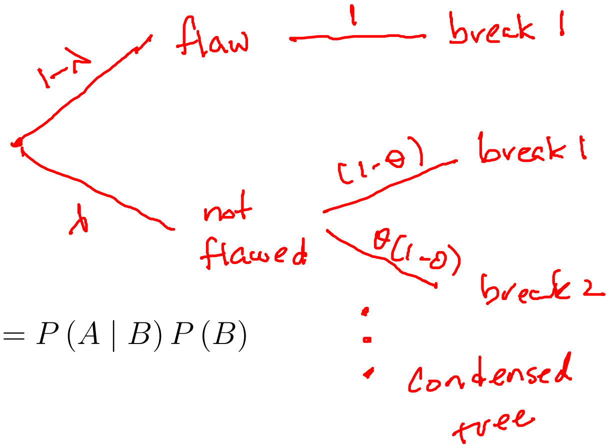
where $p_i = P\,(\text{a specimen falls in category } i)\,.$

$p_i = P(i \text{ hits required to fracture}) = P\,(Y=i) \qquad i=1,2,3$

$p_4 = P(\geq 4 \text{ hits required to fracture}) = P\,(Y\geq 4)$

We need to obtain expressions for $p_1, \dots, p_4$ in terms of λ and θ. It may be helpful to draw a tree diagram here. Try it as an exercise.

$$\begin{aligned}
&P(\text{item is flawed}) = 1 - \lambda, \\
&P(Y = 1 \mid \text{flawed}) = 1, \\
&P(\text{item is not flawed}) = \lambda, \text{ and} \\
&P(Y = y \mid \text{ not flawed}) = \theta^{y-1}(1 - \theta).
\end{aligned}$$



Recall: Let A and B be 2 events, then $P(A \text{ and } B) = P(A \mid B)\, P(B)$

$$\begin{aligned}
p_1 &= P(Y = 1) \\
&= P(Y = 1 \text{ and flawed}) + P(Y = 1 \text{ and not flawed}) \\
&= 1 - \lambda + \lambda (1 - \theta) = 1 - \lambda\theta
\end{aligned}$$

$$\begin{aligned}
p_2 &= P(Y = 2) \\
&= P(Y = 2 \text{ and flawed}) + P(Y = 2 \text{ and not flawed}) \\
&= 0 + \lambda\theta (1 - \theta)
\end{aligned}$$

$$p_3 = P(Y = 3) = \lambda\theta^2 (1 - \theta)$$

$$\begin{aligned}
p_4 &= 1 - p_1 - p_2 - p_3 \\
&= 1 - (1 - \lambda\theta) - \lambda\theta (1 - \theta) - \lambda\theta^2 (1 - \theta) = \lambda\theta^3
\end{aligned}$$

We can express the likelihood and log-likelihood in terms of θ and λ:

$$\begin{aligned}
L(\theta, \lambda) &= p_1^{x_1}\ p_2^{x_2}\ p_3^{x_3}\ p_4^{x_4} \\
&= [1 - \lambda\theta]^{112} [\lambda\theta (1 - \theta)]^{36} \left[\lambda\theta^2 (1 - \theta)\right]^{22} \left[\lambda\theta^3\right]^{30} \\
&= (1 - \lambda\theta)^{112} \lambda^{88} \theta^{170} (1 - \theta)^{58} \\
\ell(\theta, \lambda) &= 112 \ln (1 - \lambda\theta) + 88 \ln \lambda + 170 \ln \theta + 58 \ln (1 - \theta)
\end{aligned}$$

To compute the MLE's of θ and λ, we need to take derivatives, set them to zero and solve for $\hat{\theta}$ and $\hat{\lambda}$.

$$\frac{\partial \ell}{\partial \theta} = -\frac{112\lambda}{1-\lambda\theta} + \frac{170}{\theta} - \frac{58}{1-\theta} \tag{3.3}$$

$$\frac{\partial \ell}{\partial \lambda} = -\frac{112\theta}{1-\lambda\theta} + \frac{88}{\lambda} \tag{3.4}$$

$$(3.4) = 0 \Longrightarrow 112\hat{\lambda}\hat{\theta} = 88\left(1-\hat{\lambda}\hat{\theta}\right) \Longrightarrow \frac{112}{88}\hat{\lambda}\hat{\theta} = \left(1-\hat{\lambda}\hat{\theta}\right) \Longrightarrow \frac{200\hat{\lambda}\hat{\theta}}{88} = 1,$$

and

$$\hat{\lambda} = \frac{88}{200}\frac{1}{\hat{\theta}} \quad .$$

Substituting $\left(1-\hat{\lambda}\hat{\theta}\right) = \frac{112}{88}\hat{\lambda}\hat{\theta}$ into (3.3),

$$0 = -\frac{112\hat{\lambda}}{\frac{112}{88}\hat{\lambda}\hat{\theta}} + \frac{170}{\hat{\theta}} - \frac{58}{1-\hat{\theta}},$$

$$\Longrightarrow 0 = \frac{82}{\hat{\theta}} - \frac{58}{1-\hat{\theta}}.$$

$$\hat{\theta} = \frac{82}{140} = \frac{41}{70} = 0.5857$$

$$\text{and} \qquad \hat{\lambda} = \frac{88 \times 70}{200 \times 41} = \frac{154}{205} = 0.7512.$$

Substituting $\hat{\theta}$ and $\hat{\lambda}$ into the expressions for the $p_i's$, yields estimated probabilities:

$$\hat{p}_1 = 1 - \hat{\lambda}\hat{\theta} = 0.56$$
$$\hat{p}_2 = \hat{\lambda}\hat{\theta}\left(1-\hat{\theta}\right) = 0.18$$
$$\hat{p}_3 = \hat{\lambda}\hat{\theta}^2\left(1-\hat{\theta}\right) = 0.11$$
$$\hat{p}_4 = \hat{\lambda}\hat{\theta}^3 = 0.15.$$

# hits required	1	2	3	≥ 4	Total
observed frequency	112	36	22	30	200
$n\hat{p}_i$ = estimated frequency	112.00	36.46	21.35	30.19	200

The estimated and observed frequencies are very close! The new model fits the data very well.

We should check that we have attained a maximum using the second derivatives evaluated at the maximum. The information matrix entries are given below. As an exercise, check that they satisfy the criteria for a maximum.

$$I_{11} = -\frac{\partial^2 \ell}{\partial \theta^2} = \frac{112\lambda^2}{(1-\lambda\theta)^2} + \frac{170}{\theta^2} + \frac{58}{(1-\theta)^2} > 0$$

$$I_{22} = -\frac{\partial^2 \ell}{\partial \lambda^2} = \frac{112\theta^2}{(1-\lambda\theta)^2} + \frac{88}{\lambda^2} > 0$$

$$I_{21} = I_{12} = -\frac{\partial^2 \ell}{\partial \theta \partial \lambda} = \frac{112}{1-\lambda\theta} + \frac{112\theta\lambda}{(1-\lambda\theta)^2}$$

3.2 The Chi-Square Distribution

A continuous variate, X with pdf

$$f(x;\nu) = c_\nu \; x^{\nu/2-1}e^{-x/2} \text{ for } x > 0 \qquad \text{and } c_\nu \text{ a positive constant}$$

is said to have a $\chi^2_{(\nu)}$ distribution where ν is called the **degrees of freedom**.

It can be shown that $E(X) = \nu$,
$VAR(X) = 2\nu$.

We will be interested in the cdf of X,

$$F(x;\nu) = P\left(\chi^2_{(\nu)} \leq x\right) = \int_0^x f(x;\nu)\,dx.$$

Left tail areas are tabulated in the optional textbook on page 351, and a chi-square table is available on Brightspace. We will also use R to compute these.

Example: $\mathrm{P}\left(\chi^2_{(4)} \leq 7.779\right) = 0.9.$
Check that you can get this answer from the chi-square table provided on Brightspace or on page 351 of the optional text or using R (code below).

The chi-square density is graphed for various values of ν, the degrees of freedom.

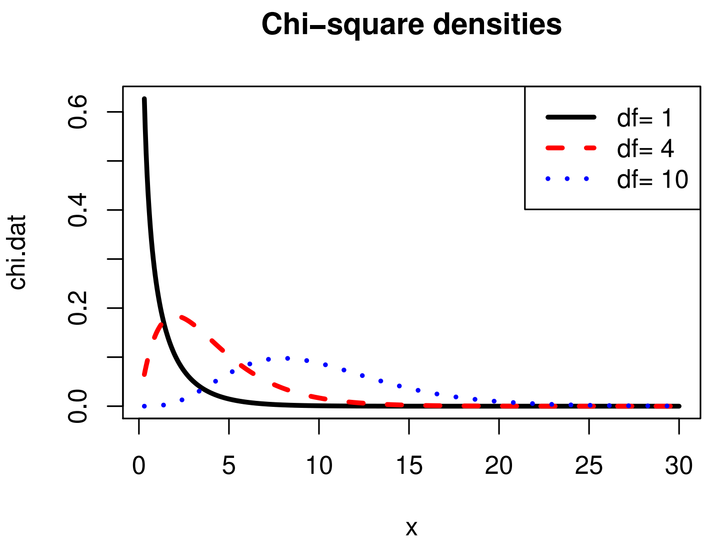


Figure 3.3: Chi-square densities

Properties of the Chi-square distribution:

1. Let $X_1, X_2, \ldots, X_n$ be independent random variates with $X_i \sim \chi^2_{(\nu_i)}$. Then $X_1 + X_2 + \cdots + X_n \sim \chi^2_{(\nu_1+\nu_2+\cdots+\nu_n)}$
2. Let $Z \sim N(0,1)$. Then $Z^2 \sim \chi^2_{(1)}$.
3. Let $Z_1, Z_2, \ldots Z_n$ be independent $N(0,1)$ random variables. Then $Z_1^2, Z_2^2, \ldots, Z_n^2$ are independent $\chi^2_{(1)}$ random variables and $Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi^2_{(n)}$

3.2.1 R code for Chi-square distribution

```
pchisq(7.779,4)  # find probability chi-square(4) < 7.779
qchisq(.9, 4)    # find .90 quantile of chi-square(4)


x <- seq(0.3,30,by=.05)
chi.dat <- cbind(dchisq(x,1), dchisq(x,4), dchisq(x,10))  #chi-square densities
matplot(x,chi.dat,type='l',col=1:3, lty=1:3, main='Chi-square densities')
legend("topright",c(paste('df=',c(1,4,10))),lty=1:3,col=1:3)
```

Chapter 4

Tests of Significance

4.1 Introduction to Tests of Significance

Optional reading: Chapter 12.1

We will consider a formal method for testing hypotheses about model parameters. To illustrate the ideas, I will consider a simple example which tests the following claim.

I claim that I have ESP

To test this claim, we will perform an experiment using a deck of cards. After shuffling the cards, a volunteer chooses a card and I must divine the colour of the suit. This is repeated 25 times and the number of correct responses is recorded.

Let X be the number of correct responses out of the 25 independent trials.

Notes:

1. Even if I do NOT have ESP, some correct responses will occur by chance.
2. If I do have ESP, I should be able to achieve more correct responses than would be expected by chance alone.

We define two hypotheses:

1. H_1: I have ESP. This is the claim or research hypothesis and is called the **alternative hypothesis**. (It is also denoted as H_A.)
2. H_0: I do not have ESP. This is usually the complement of H_1 and is called the **null hypothesis**.

We proceed as if we were performing a proof by contradiction, assuming that the null hypothesis is true until we obtain enough evidence against it.

To determine whether there is evidence against the null and in favour of the hypothesis of ESP, we compare the results obtained from the experiment with that which would be **expected under the null hypothesis** that I do NOT have ESP.

Large values of X observed, x_{obs}, will be interpreted as evidence against the null hypothesis and in favour of the alternative hypothesis.

Under H_0, the hypothesis that I do NOT have ESP, my responses are guesses, and each response has a .50 chance of being correct. Therefore Under H_0,
$X \sim$ Binomial$(n = 25, p = 0.5)$, so that I am expected to get $E(X) = np = 12.5$ correct, on average if the null hypothesis is true.

Note that the null and alternative hypotheses can be written in terms of p, the Binomial probability of success as:

$$H_1 : p > 0.5 \quad H_0 : p = 0.5.$$

Suppose that I get $x_{obs} = 18$ correct. Does this provide evidence against H_0?

To answer that question, Statisticians compute a **p-value** also called a **Significance level (SL)**. It is defined as the probability of observing a result as extreme or more in the direction of the alternative hypothesis, computed assuming that the null hypothesis is true.

Returning to the example, suppose that I get x_{obs} =18 correct responses. The p-value= $P(X \geq 18)$ is computed using the distribution of X under the null hypothesis, that is, Binomial$(n = 25, p = 0.5)$. Using R, we compute:

$$\text{p-value} = P(X \geq 18) = 1 - P(X \leq 17) = 0.02164263$$

using the code,

```
        1-pbinom(17, size=25, p=.5).
```

You may recall the Normal approximation to the Binomial,

$$X \approx N\left(\mu = np = \frac{25}{2},\ \sigma^2 = np\,(1-p) = \frac{25}{4}\right),$$

is appropriate when $np \geq 5$ and $n(1-p) \geq 5$. The p-value can be computed using this approximation as:

$$\text{p-value} = P(X \geq x_{obs}) = P\left(\frac{X-\mu}{\sigma} \geq \frac{x_{obs}-12.5}{\sqrt{\frac{25}{4}}}\right) \simeq P\left(Z \geq \frac{x_{obs}-12.5}{\sqrt{\frac{25}{4}}}\right),$$

where $Z \sim N(0,1)$.

The p-value is approximately,

$$P(X \geq 18) \simeq P(Z \geq 2.2) = 1 - 0.98610 = 0.0139,$$

where the normal probability is obtained from the $N(0,1)$ cdf table on Brightspace or on page 349 of the optional text. The approximation can be improved using a continuity correction to the normal approximation,

$$\begin{aligned} P(X \geq 18) &= 1 - P(X \leq 17) \simeq 1 - P\left(Z \leq \frac{17.5-12.5}{\sqrt{\frac{25}{4}}}\right) \\ &= 1 - P(Z \leq 2) = 1 - 0.97725 = 0.02275. \end{aligned}$$

Note: You should review how to use the $N(0,1)$ cdf table since we will resort to tables for the Quizzes and Exam.

How do we interpret the p-value in practice?

- Large $p-values$ suggest that results as extreme as x_{obs} would occur fairly often if H_0 were true and we have no evidence that H_0 is false. There is no inconsistency with H_0, **BUT** this does not prove that H_0 is true! It only indicates a lack of evidence against H_0.

- Small $p-values$ suggest that if H_0 were true, results as extreme as x_{obs} would occur very rarely. The data are then deemed to be inconsistent with H_0. Thus we say that we have evidence against H_0.

We say that:

$$\begin{aligned} p-value &> .10 && \text{- no evidence against } H_0 \\ .05 < p-value &\leq .10 && \text{- marginal evidence against } H_0 \\ .01 < p-value &\leq .05 && \text{- evidence against } H_0 \\ p-value &\leq .01 && \text{- strong evidence against } H_0 \end{aligned}$$

In our ESP example, we conclude that we have evidence against H_0, (p-value $= 0.02$). The data are consistent with the hypothesis that I have ESP. Of course, results like these could have occurred by chance, and this does NOT prove that I have ESP.

Ingredients for Tests of Significance

1. Test statistic, D - provides a ranking of all possible outcomes of an experiment according to how closely they agree with H_0, the null hypothesis.

   Small values of $D \Longrightarrow$ close agreement with H_0

   Large values of $D \Longrightarrow$ poor agreement with H_0

   (In the ESP example, $D = X =$ the number of correct responses.)

2. We need a measuring device to determine how far away from H_0 is the observed test statistic. We use the $p-value = P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right)$, the probability of a random D greater than or equal to the value observed, d_{obs}, computed assuming that H_0 is true.

   This is the probability of observing such poor agreement between the null hypothesis and the data if H_0 were true.

   If the $p-value$ is small, then such poor agreement would almost never occur when H_0 is true.

With data, we cannot prove or disprove a null hypothesis. All that we can do is to say whether our data is consistent or not consistent with the null hypothesis.

The p-value is a probability computed using our data and assuming that the null hypothesis is true.

Example 4.1.1. Blind Taste Test: Twenty-five individuals were given two similar glasses, one of Pepsi, one of Coke and each was asked to identify the one that was Coke. 60% (15) correctly identified Coke. Is this consistent with

$$H_0: \text{ there is no detectable difference between Pepsi and Coke.}$$

Solution: We shall initally proceed under assumption that H_0 is true, that there is no detectable differences between Pepsi and Coke and see if the data provides evidence against the null hypothesis. We need a probability model for the data under the assumption that H_0 is true. Let

$$X = \# \text{ individuals out of 25 who correctly identified Coke.}$$

If H_0 is true, then the responses would be guesses, with a .50 chance of being correct. Therefore under H_0, $X \sim Binomial\left(n = 25, p = \frac{1}{2}\right)$ and we would expect to observe a value of X near $E(X) = np = 12.5$ if H_0 is true. In this example, very small numbers of correct responses as well as large numbers of correct responses would suggest that there is a detectable difference between Pepsi and Coke. For example, if zero out of 20 responses were correct, that would suggest that the two drinks were detectably different, but not correctly identified.

We therefore define our statistic to be large when the number of observed correct, x_{obs}, is much smaller or much larger than 12.5. Let

$$D \equiv |X - 12.5|$$

be our test statistic which ranks possible values of X according to how close they are to H_0.

- If D is close to zero, then the data are in agreement with H_0.
- if D is large (close to 12), then the data are NOT in agreement with H_0.

In our example, 15 correctly identified Coke, so that,

$$d_{obs} = |x_{\text{obs}} - 12.5| = |15 - 12.5| = 2.5$$

$$\begin{aligned} p-value &= P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) \\ &= P\left(D = 2.5 \text{ or } 3.5 \text{ or } \ldots 12.5 \mid H_0 \text{ true}\right) \\ &= 1 - P\left(D = .5 \text{ or } 1.5 \mid H_0 \text{ true}\right) \end{aligned}$$

$$\begin{aligned} D = .5 &\Longrightarrow X = 13 \text{ or } 12 \\ D = 1.5 &\Longrightarrow X = 14 \text{ or } 11 \end{aligned}$$

$$\begin{aligned} p-value &= 1 - P\left(X = 11, 12, 13 \text{ or } 14 \mid H_0 \text{ true}\right) \\ &= 1 - \sum_{x=11}^{14} \binom{25}{x} \left(\frac{1}{2}\right)^{25} \\ &= 0.4243562 \quad \text{(using R)} \\ &= \ \text{EXACT } p-value \end{aligned}$$

R Code: `1- sum(dbinom(11:14, size=25, p=.5))`

If H_0 were true, results as extreme $X = 15$ would occur fairly often and we have no evidence against the null hypothesis. The data are consistent with the null hypothesis that there is no detectable difference between Pepsi and Coke (p-value = .42).

Example 4.1.1 continued: Suppose that 60% (150) of 250 individuals correctly identified Coke. Is there evidence against the null hypothesis?

Under $H_0 : X \sim Binomial\left(n = 250, p = \frac{1}{2}\right)$ and $E(X) = 125$.

$$\begin{aligned} D = |X - 125| \quad &\text{and} \quad d_{\text{obs}} = |150 - 125| = 25 \\ p-value = P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) &= 0.001883301 \end{aligned}$$

There is very strong evidence against the null hypothesis of no detectable difference between Coke and Pepsi (p-value = 0.002)!!

R Code: `1- sum(dbinom(101:149, size=250, p=.5))`

Since $n = 250$ is large, we can use a normal approximation to the Binomial to obtain,

$$\begin{aligned} p - value &= P\left(|X - 125| \geq 25\right) \\ &= P\left(\frac{|X - 125|}{\sqrt{250p(1-p)}} \geq \frac{25}{\sqrt{250p(1-p))}}\right) \\ &\simeq P\left(|Z| \geq \frac{25}{\sqrt{\frac{250}{4}}}\right) \\ &= P\left(|Z| \geq 3.16\right) = 0.001565402 \end{aligned}$$

R Code: `2*pnorm(-25/sqrt(250/4))`

The large SAMPLE SIZE yields a more precise estimate of the probability of correctly identifying Coke versus Pepsi!

4.2 Likelihood Ratio Tests for Simple Null Hypotheses

Optional reading: Chapter 12.2

We have looked at two test statistics to test hypotheses about the Binomial parameter p, $D = X$ and $D = |X - np|$. In general, the test statistic will depend upon the hypothesis being tested and it may be difficult to "come up" with a test statistic. The Likelihood Ratio Statistic (LRS) is a good statistic and it has intuitive appeal. We consider the LRS for simple null hypotheses. In many applications, the hypothesis to be tested can be formulated as a hypothesis concerning the values of unknown parameters in a probability model.

Definition: A **simple hypothesis** specifies numerical values for all of the unknown parameters in the model.

4.2.1 One Parameter Case

Consider a probability model with one unknown parameter θ. We wish to test $H_0 : \theta = \theta_0$ where θ_0 is a particular numerical value.

For example, $H_0 : p = \frac{1}{2}$ in the Binomial model.

With simple hypothesis tests, we are asking if a particular parameter value θ_0 is plausible given the data we have. We have already seen a function that tells us about the relative plausibilities of parameter values, $R(\theta)$ or $r(\theta)$.

Definition: The **Likelihood Ratio Statistic (LRS)** for testing $H_0 : \theta = \theta_0$ is

$$D \equiv -2r\left(\theta_0\right) = 2[\ell(\hat{\theta}) - \ell\left(\theta_0\right)],$$

where $\hat{\theta}$ is the MLE of θ. Since $\ell(\hat{\theta}) \geq \ell\left(\theta_0\right)$ for all values of θ_0, then $D \geq 0$.

$$\begin{aligned} D \text{ small} &\implies \text{outcome is such that } \theta_0 \text{ is a plausible parameter value} \\ D \text{ large} &\implies \text{outcome is such that } \theta_0 \text{ is not a plausible parameter value} \end{aligned}$$

Thus, D ranks possible outcomes of the experiment according to how well they agree with $H_0 : \theta = \theta_0$. Let d_{obs} be the observed numeric value of the Likelihood Ratio Statistic, then the $p - value$ is calculated as:

$$\begin{aligned} p - value = \mathrm{SL} &= P\left(D \geq d_{\mathrm{obs}} \mid H_0 \text{ true}\right) \\ &= P\left(D \geq d_{\mathrm{obs}} \mid \theta = \theta_0\right) \end{aligned}$$

Under the assumption that $H : \theta = \theta_0$ is true,

$$D \approx \chi^2_{(1)}$$

in most cases of **one-parameter simple** hypotheses, therefore, you can use the chi-squared table to obtain p-values as,

$$p - value \simeq P(\chi^2_{(1)} \geq d_{obs}).$$

Notation Notes:

- $\approx$ means approximately distributed as
- $\simeq$ means approximately equal to

Example 4.2.1. The measurement errors associated with a set of scales are independent normal with known $\sigma = 1.3$ grams. Ten ($n = 10$) weightings of an unknown mass μ give the following results in grams:

$$\begin{array}{ccccc} 227.1 & 226.8 & 224.8 & 228.2 & 225.6 \\ 229.7 & 228.4 & 228.8 & 225.9 & 229.6 \end{array}$$

Is the data consistent with $H_0 : \mu = \mu_0 = 226$. Derive the LRS for testing $\mu = 226$.

Let x_i represent the $i'th$ observed weighting, then

$$\begin{aligned} D &= 2\left[\ell\left(\hat{\mu}\right) - \ell\left(\mu_0\right)\right] \\ L\left(\mu\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right] \\ &= \left(\frac{1}{\sqrt{\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right] \end{aligned}$$

Since σ is assumed known, and equal to 1.3 grams, the term, $\left(\frac{1}{\sqrt{\sigma^2}}\right)^n$ is considered a constant and can be disregarded in the construction of the likelihood for μ.

$$\begin{aligned} \ell\left(\mu\right) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \quad \text{and} \quad \hat{\mu} = \frac{\sum x_i}{n} = \bar{x} = 227.49. \\ D &= -2r\left(\mu_0\right) = 2\left[\ell\left(\hat{\mu}\right) - \ell\left(\mu_0\right)\right], \quad \text{where } \mu_0 = 226 \\ &= 2\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu_0\right)^2\right] \end{aligned}$$

Consider the second term in the above expression:

$$\sum_{i=1}^{n}(x_i-\mu_0)^2=\sum_{i=1}^{n}[(x_i-\bar{x})+(\bar{x}-\mu_0)]^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-\mu_0)^2\,.$$

Why is the cross-product term zero?

Therefore,

$$\begin{aligned} D &= \frac{1}{\sigma^2}\left[-\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-\mu_0)^2\right] \\ &= \frac{n\,(\bar{x}-\mu_0)^2}{\sigma^2}=\frac{(\bar{x}-\mu_0)^2}{\sigma^2/n}. \end{aligned}$$

We want to find the distribution of D assuming that $H_0:\mu=\mu_0=226$ is true.

$$\begin{aligned} &\text{If } X\sim N\left(\mu_0,\sigma^2\right),\ \text{then}\quad \bar{X}\sim N\left(\mu_0,\frac{\sigma^2}{n}\right) \\ &\implies Z=\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\sim N\left(0,1\right)\implies D=Z^2\sim\chi^2_{(1)} \end{aligned}$$

For this example, the likelihood ratio test statistic is equivalent to the Z test statistic that you learned in your first course in statistics!

In the normal case, the LRS has an EXACT $\chi^2_{(1)}$ distribution!

$$\begin{aligned} d_{\text{obs}} &= \frac{10}{(1.3)^2}\left(227.49-226\right)^2=13.14 \\ p-value &= P\left(D\geq d_{\text{obs}}\mid\mu_0=226\right) \\ &= P\left(\chi^2_{(1)}\geq 13.14\right)=0.00029<.005 \end{aligned}$$

There is very strong evidence against $H_0:\mu=226\ \ (p-value<0.005)$. In our report, we should include our estimate of the mean, μ, together with an interval estimate which provides information about the margin of error in estimating the mean. We write in our report: **"The estimated mean weight is 227.49 grams and the data are not consistent with the hypothesis that the mean weight**

is 226 grams ($p - value < 0.005$, 10% Likelihood interval estimate 226.61-228.37 grams).”

I computed the 10% likelihood interval using the R code provided below. We will see later that likelihood intervals are related to confidence intervals, which are more commonly quoted in practice. In this example, the interval is very, very narrow, and the lower endpoint is close to 226 grams. Although the data suggest that the mean weight is **statistically** significantly different from 226, the investigator may find no **practical** difference between the observed data and the hypothesis that the mean is 226 grams.

4.2.2 R code for Example 4.2.1

```
#Compute the LRS and p-value
d<-10*(227.49-226)^2/(1.3^2)
1-pchisq(d,1)

#Define functions and data required to compute a likelihood interval.
sigma<-1.3
y<-c(227.1, 226.8, 224.8, 228.2, 225.6, 229.7, 228.4, 228.8, 225.9, 229.6)
mean(y)

#Log-likelihood function
ell <- function(mu,y,sigma){
  res<-vector("numeric",length(mu))
  for (i in 1:length(mu)){
    res[i]<--sum((y-mu[i])^2)/2/sigma^2
  }
  return(res)
}

#MLE
muhat <- optimize(ell, c(225,230), maximum=TRUE, y=y, sigma=sigma)
muhat

#Log relative likelihood function; plot for values of mu to help
#  determine starting values for computing a 10% likelihood interval
logR <- function(mu, muhat,y,sigma){
  ell(mu,y,sigma) - ell(muhat,y,sigma)
}

mu <- seq(225,230,by=.1)
plot(mu,logR(mu,muhat$maximum,y,sigma), ylab='log relative likelihood',xlab='mu')
abline(h=log(.1))    #add a horizontal line at ln(p)
title('Log-relative Likelihood, Example 4.2.1')

#Likelihod intervals
p <- .1  #10% likelihood interval
#log relative likelihood minus ln(p)
logR.m.lnp <- function(mu, muhat, y,sigma, p) {logR(mu,muhat,y,sigma)-log(p)}

lower <- uniroot(logR.m.lnp, c(226,227), muhat$maximum, y,sigma, p)
lower
upper <- uniroot(logR.m.lnp, c(228,229), muhat$maximum, y,sigma, p)
upper
```

4.2.3 LR Statistic for 2 or More Parameters

Suppose that we have a probability model that depends upon a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$. We wish to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$, for $\boldsymbol{\theta_0}$ a vector of numbers. Then H_0 is a simple hypothesis because it specifies a numerical value for each parameter in the model.

The **Likelihood ratio statistic** for testing the SIMPLE hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ is

$$D \equiv -2r\left(\boldsymbol{\theta_0}\right) = 2\left[\ell(\hat{\boldsymbol{\theta}}) - \ell\left(\boldsymbol{\theta_0}\right)\right],$$

where $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_p\right)'$ is the joint MLE.

Under the assumption that $\boldsymbol{\theta} = \boldsymbol{\theta_0}$, $D \approx \chi^2_{(k)}$ (in most cases), where k is the number of functionally independent unknown $\boldsymbol{\theta}$ parameters in the model.

Example 4.2.2. Multinomial distribution: Consider frequencies, $(X_1, X_2, \ldots, X_5) \sim$ Multinomial $(n, p_1, p_2, \ldots, p_5)$. Here $k = 4$ functionally independent parameters since

$\sum_{i=1}^{5} p_i = 1$. If four parameters are known, then the fifth is fully determined.

Example 4.2.3. Heart disease: In a long-term study of heart disease in a large group of men, it was noted that 63 men who had no previous record of heart problems died suddenly of heart attacks. The following table shows the number of such deaths recorded on each day of the week.

Day	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
# deaths	22	7	6	13	5	4	6	63
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	n

Test the null hypothesis that deaths are equally likely to occur on any day of the week.

Solution:

$$(X_1, \ldots, X_7) \sim \text{Multinomial } \ (n = 63, p_1, \ldots, p_7)$$

The null hypothesis is that deaths are equally likely to occur on any day of the week, and so, $H_0 : p_1 = p_2 = \cdots = p_7 = \frac{1}{7}$, a simple hypothesis.

The likelihood ratio statistic for testing H_0 is,

$$\begin{aligned} D &= -2r\left(p_1 = 1/7,\ p_2 = 1/7, \ldots,\ p_7 = 1/7\right) \\ &= 2\left[\ell\left(\hat{p}_1, \ldots, \hat{p}_7\right) - \ell\left(p_1 = 1/7, \ldots,\ p_7 = 1/7\right)\right] \end{aligned}$$

We need the joint MLE's for the $p's$!

$$\begin{aligned} L\left(p_1, \ldots, p_7\right) &= p_1^{x_1}\ p_2^{x_2}\ \cdots p_7^{x_7} = p_1^{22}\ p_2^{7}\ \cdots\ p_7^{6} \text{ where } \sum_{i=1}^{7} p_i = 1 \\ \ell\left(p_1,\ \ldots\ , p_7\right) &= 22\ln p_1 + 7\ln p_2 + \cdots + 6\ln p_7 \\ &= \sum_{i=1}^{7} x_i \ln p_i. \end{aligned}$$

We need to maximize $\ell\left(p_1,\ldots,p_7\right)$ subject to the constraint $\sum\limits_{i=1}^{7} p_i = 1$. To do that we can use Lagrange multipliers (if you have taken Math 200) or we can simply substitute the constraint into the log-likelihood and maximize over the parameters.

Method 1: Lagrange multipliers

The objective function, $g = \sum\limits_{i=1}^{7} x_i \ln p_i + \gamma\left(\sum_{i=1}^{7} p_i - 1\right)$ and the partial derivatives with respect to the $p's$ and γ are,

$$\begin{aligned} \frac{\partial g}{\partial p_i} &= \frac{x_i}{p_i} + \gamma \quad \text{for } i = 1,\ldots,7 & (4.1)\\ \frac{\partial g}{\partial \gamma} &= \sum p_i - 1 & (4.2) \end{aligned}$$

To maximize, set the partial derivatives to 0 and solve.

$$\begin{array}{lll} (4.1) = 0 & \Longrightarrow \hat{p}_i = -x_i/\hat{\gamma} & \quad i = 1,\ldots,7 \\ (4.2) = 0 & \Longrightarrow \sum_{i=1}^{7} \hat{p}_i = 1 & \Longrightarrow -\sum_{i=1}^{7} x_i/\hat{\gamma} = 1 \\ & & \Longrightarrow -\sum\limits_{i=1}^{7} x_i = \hat{\gamma} = -63 = -n \end{array}$$

Substituting the $\hat{\gamma}$ into the solution for (4.1), $\hat{p}_i = x_i/n = x_i/63, \quad i = 1,\ldots,7$, which is the sample proportion that fall in category i.

Method 2: Substitute the constraint into the log-likelihood

We have that,

$$\ell\left(p_1,\ \ldots\ ,p_7\right) = \sum_{i=1}^{7} x_i \ln p_i.$$

where $\sum_{i=1}^{7} p_i = 1$ and $\sum_{i=1}^{7} x_i = n$. Let $p_7 = 1 - p_1 - p_2 - p_3 - p_4 - p_5 - p_6$, and substitute this into the log-likelihood. The log-likelihood becomes,

$$\ell\left(p_1,\ \ldots\ ,p_7\right) = \sum_{i=1}^{6} x_i \ln p_i + x_7 \ln(1 - p_1 - p_2 - p_3 - p_4 - p_5 - p_6).$$

We take derivatives with respect to the $p's$, set the expressions equal to zero and solve for the MLE's of the $p's$.

$$\frac{\partial \ell}{\partial p_i} = \frac{x_i}{\hat{p}_i} - \frac{x_7}{\hat{p}_7} = 0 \quad \text{for } i = 1, \ldots, 6.$$

$$\text{Therefore, } \hat{p}_i = x_i \, \frac{\hat{p}_7}{x_7}. \tag{4.3}$$

Summing the six equations (4.3), results in,

$$\begin{aligned} \sum_{i=1}^{6} \hat{p}_i &= \frac{\hat{p}_7}{x_7} \sum_{i=1}^{6} x_i \\ 1 - \hat{p}_7 &= \frac{\hat{p}_7}{x_7}(n - x_7) \\ (1 - \hat{p}_7)x_7 &= \hat{p}_7(n - x_7) \\ \hat{p}_7 &= x_7/n, \quad \text{and substituting into (4.3),} \\ \hat{p}_i &= x_i/n. \end{aligned}$$

□

Now that we have the MLE's of the $p's$, we can compute our Likelihood Ratio Statistic:

$$\begin{aligned} d_{\text{obs}} &= 2\left[\ell\left(\hat{p}_1, \ldots, \hat{p}_7\right) - \ell\left(1/7, \ldots, 1/7\right)\right] \\ &= 2\left[\sum_{i=1}^{7} x_i \ln \frac{x_i}{n} - \sum_{i=1}^{7} x_i \ln \frac{1}{7}\right] \\ &= 23.27 \\ p - value &\simeq P\left(\chi^2_{(k)} \geq d_{\text{obs}} \mid H_0 \text{ true}\right), \quad \text{here } k = 6 \\ &= P\left(\chi^2_{(6)} \geq 23.27 \mid p_1 = p_2 = \cdots = p_7 = \frac{1}{7}\right) \\ &\simeq 0.0007 \end{aligned}$$

Here $k = 6$ functionally independent parameters since $\sum_{i=1}^{7} p_i = 1$. If six parameters are known, then the seventh is fully determined.

The p-value is very small (< 0.01) and we say that we have very strong evidence against the hypothesis that deaths are equally likely to occur on any day of the week. The estimated expected frequencies under the null hypothesis that $p_i = 1/7 = 0.1428571$ are all $np_i = 9$. From the table of observed frequencies, we note that there are many more heart attacks on Monday than would be expected under the null hypothesis.

4.2.4 R code for Example 4.2.3

```
#Heart disease example
freq <- c(22 , 7 , 6 , 13 , 5 , 4 , 6)
sum(freq)
ell <- function(p, freq){
  # Multinomial log-likelihood
  # freq = frequencies;  p = probabilities
  sum(freq * log(p))
}

logR <- function(p, phat, freq){
  #Log relative likelihood function
  ell(p, freq) - ell(phat, freq)
}

dobs <- -2 * logR(rep(1/7, 7), freq/sum(freq),freq)   #LRS observed
dobs
1-pchisq(dobs, 6)  #pvalue
```

4.3 Likelihood Ratio Tests for Composite Hypotheses

Optional reading: Chapter 12.3

We may have hypothesized, a priori, that heart attacks have different daily probabilities of occurrence, for example, they may be more likely to occur on Monday's, and formed our null hypothesis as:

$$H_0 : p_2 = p_3 = \cdots = p_7 = p \text{ and } \quad p_1 \text{ unspecified.}$$

Our **BASIC MODEL** for the multinomial frequencies is still $(x_1, x_2, \ldots, x_7) \sim$ Multinomial$(n = 63, p_1, \ldots, p_7)$.

The null hypothesis does not specify numerical values for every parameter in the model, since p is unknown, therefore it is NOT a simple hypothesis! H_0 is an example of a **composite hypothesis**. Note that if we knew p, we could obtain p_1 by subtraction since the $p's$ must sum to one so that $p_1 = 1 - 6p$.

Definition: A **Composite hypothesis** reduces the number of unknown parameters in the model, but not to zero.

To test the new, composite hypothesis, we need to find the MLE of the $p's$ assuming that H_0 is true. We substitute the hypothesized values into the log-likelihood and maximize over p.

$$\begin{aligned}\ell\left(p_1, p_2 = p, p_3 = p, p_4 = p, p_5 = p, p_6 = p, p_7 = p\right) &= x_1 \ln p_1 + \sum_{i=2}^{7} (x_i \ln p) \\ &= x_1 \ln p_1 + (\ln p) \sum_{i=2}^{7} x_i\end{aligned}$$

Instead of using a Lagrange multiplier, we substitute the constraint, $p_1 = 1 - 6p$, into the log likelihood as follows,

$$\begin{aligned}\ell_H(p_1, p) = \ell\left(p_1, p_2 = p, \ldots, p_7 = p\right) &= x_1 \ln\left[1 - 6p\right] + (\ln p) \sum_{i=2}^{7} x_i \\ &= 22 \ln\left(1 - 6p\right) + 41 \ln p.\end{aligned}$$

Taking the derivative with respect to p yields,

$$\frac{\partial \ell_H}{\partial p} = \frac{22(-6)}{1-6p} + \frac{41}{p}.$$

The MLE of p under H_0 is $\tilde{p} = \frac{41}{378} = 0.1085$ and $\tilde{p}_1 = 0.349$.

$\ell_H\left(\tilde{p}_1, \tilde{p}\right)$ is the largest value that $\ell\left(p_1, \ldots, p_7\right)$ can attain under H_0.

The **Likelihood Ratio Statistic** for testing the **composite** hypothesis H_0 is

$$D = 2\left[\underset{\substack{\uparrow \\ \text{max possible for this} \\ \text{prob model}}}{\ell\left(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_7\right)} - \underset{\substack{\uparrow \\ \text{max possible under } H_0}}{\ell\left(\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_7\right)}\right]$$

where $\tilde{p}_1 = 0.349$ and $\tilde{p}_i = \tilde{p} = .1085$ for $i = 2, ..., 7$.

D ranks possible outcomes according to how well they agree with H_0.

If D is small, then the maximum of the log likelihood is nearly as large assuming H_0 to be true as under the basic model.

Under H_0 : the asymptotic distribution of the LRS is approximately

$$D \approx \chi^2_{(k-q)}$$

where

$k = \#$ functionally independent unknown parameters in the basic model
$q = \#$ functionally independent unknown parameters in hypothesized model.

In our example, $k = 6$ and $q = 1$.

$$\begin{aligned}
d_{obs} &= 2\left[\ell\left(\hat{p}_1, \ldots, \hat{p}_7\right) - \ell\left(\tilde{p}_1, \tilde{p}, \ldots, \tilde{p}\right)\right] \\
&= 2\left[\sum_{i=1}^{7} x_i \ln \frac{x_i}{n} - x_1 \ln \tilde{p}_1 - \sum_{i=2}^{7} x_i \ln \tilde{p}\right] \\
&= 2\left[(-110.96) - (-114.22)\right] = 6.529754 \\
p - value &= P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) \\
&\simeq P\left(\chi^2_{(5)} \geq 6.529754\right) = 0.2580262
\end{aligned}$$

The p-value is large (> 0.1), and we have no evidence against the null hypothesis. We conclude that heart attacks were more likely to occur on Monday's for this sample, and equally likely to occur on the other days of the week. The estimated expected frequencies under H_0, are shown in the bottom row of the following table:

Day	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	Total
# deaths	22	7	6	13	5	4	6	63
Estimated expected freq under H_0	22=63$\tilde{p}_1$	6.8=63$\tilde{p}$	6.8	6.8	6.8	6.8	6.8	$n = 63$

4.3.1 R Code Example 4.2.3 continued

```
#hypothesize that p_2 = p_3 ... p_7 =p
p_tilde <- 41/378
p_tilde
p_1 <- 1-6*p_tilde
```

```
p_1
dobs <- -2*logR(c(p_1,rep(p_tilde,6)),freq/sum(freq),freq)
dobs

1-pchisq(dobs,5)  #pvalue
```

4.3.2 Summary of Likelihood Ratio testing

The following summarizes the material in Sections 4.2 and 4.3.

1. First we assume that the data arise from a BASIC probability model with k functionally independent unknown parameters. Compute the MLE's of these k unknown parameters, $\hat{\boldsymbol{\theta}}$.

2. Then we write the null hypothesis in terms of model parameters with q functionally independent unknown parameters. Note that for a simple hypothesis, $q = 0$. Compute the MLE's of the q unknown parameters under the null hypothesis, $\tilde{\boldsymbol{\theta}}$, when $q \neq 0$.

3. We test H_0 - is the data consistent with the null hypothesis model?

   (a) The likelihood ratio test statistic is computed. Small values of the statistic indicate good agreement between the data and the null hypothesis.

   (b) The p-value is the probability of obtaining results as extreme as those observed assuming that the null hypothesis is true. Small p-values indicate that the results are unlikely to occur if the null hypothesis is true. Large p-values suggest that the data are consistent with the null hypothesis.

Likelihood Ratio Statistic: $D = -2r\left(\tilde{\boldsymbol{\theta}}\right) = 2\ln\left[\frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})}\right] = 2[\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})]$.

$\tilde{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$ under the hypothesized model.

If H_0 is true, then $D \approx \chi^2_{(k-q)}$, for the cases that we will consider.

4.4 Tests for Binomial Probabilities

Optional reading: Chapter 12.4

Example 4.4.1. Should Pot Be Legalized? One hundred people, randomly selected from each of four provinces, were asked whether or not they think that pot should be legalized. The frequencies responding yes and no are tabulated below.

Province	B.C.	Alberta	Sask.	Ontario	Total
Yes	23	19	27	10	79
No	77	81	73	90	321
Totals	100	100	100	100	400

1. Test the hypothesis that the probability of a Yes response is the same in all four provinces.
2. Test the hypothesis that the three western provinces have respondents who are equally likely to say Yes, whereas Ontario responds differently.

1. Test the hypothesis that the probability of a Yes response is the same in all four provinces.

Step 1: BASIC model

Let $Y_i = \#$ Yes for province i out of $n_i = 100$.

Then $Y_i \sim$ Binomial (n_i, p_i) independent $i = 1, 2, 3, 4$, where $p_i = P(\text{Yes for province i})$.

The probability mass function for the data is,

$$f\left(y_1, y_2, y_3, y_4;\ p_1, p_2, p_3, p_4\right) = \prod_{i=1}^{4} \binom{n_i}{y_i} p_i^{y_i} \left(1 - p_i\right)^{n_i - y_i}$$

There are $k = 4$ functionally independent parameters.

$$L\left(p_1, p_2, p_3, p_4\right) = \prod_{i=1}^{4} p_i^{y_i} \left(1 - p_i\right)^{n_i - y_i}$$

$$\ell\left(p_1, p_2, p_3, p_4\right) = \sum_{i=1}^{4} \left[y_i \ln p_i + \left(n_i - y_i\right) \ln\left(1 - p_i\right)\right]$$

Taking the derivative with respect to p_i, setting equal to zero and solving, we obtain,

$$\frac{\partial \ell}{\partial p_i} = \frac{y_i}{p_i} - \frac{100 - y_i}{1 - p_i}, \quad \text{and} \;\; \hat{p}_i = \frac{y_i}{n_i}, \;\; i = 1, 2, 3, 4.$$

Step 2: Hypothesized model

$$H_0 : p_1 = p_2 = p_3 = p_4 = p \qquad \text{unspecified}$$

This is a composite hypothesis because p is unknown and must be estimated. There is $q = 1$ functionally independent unknown parameter. Substituting the null hypothesis into the log-likelihood, we obtain,

$$\begin{aligned} \ell_H(p) = \ell\left(p_1 = p, p_2 = p, p_3 = p, p_4 = p\right) &= \sum_{i=1}^{4} \left[y_i \ln p + (n_i - y_i) \ln\left(1 - p\right)\right] \\ &= (\sum_{i=1}^{4} y_i) \ln p + (400 - \sum_{i=1}^{4} y_i) \ln(1 - p) \end{aligned}$$

Taking derivatives, setting to zero and solving, we obtain,

$$\frac{\partial \ell_H}{\partial p} = \frac{\sum_{i=1}^{4} y_i}{p} - \frac{(400 - \sum_{i=1}^{4} y_i)}{1 - p}, \quad \text{and} \;\; \tilde{p} = \frac{\sum_{i=1}^{4} y_i}{400} = 0.1975. \quad = \frac{79}{400}$$

Step 3: Test the Hypothesis

$$\begin{aligned} D &= 2\left[\ell\left(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4\right) - \ell\left(\tilde{p}, \tilde{p}, \tilde{p}, \tilde{p}\right)\right] \\ &= 2\left\{\sum_{i=1}^{4} \left[y_i \ln \frac{y_i}{n_i \tilde{p}_i} + (n_i - y_i) \ln \frac{(n_i - y_i)}{n_i(1 - \tilde{p}_i)}\right]\right\}, \end{aligned}$$

where $\tilde{p}_i = \tilde{p} = 0.1975, i = 1, 2, 3, 4.$

Substituting in values for $\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4$, and $\tilde{p}$, we obtain $d_{obs} = 10.76$. The p-value for the test is computed as,

$$\begin{aligned} p - value &\simeq P(\chi^2_{(k-q)} \geq d_{obs}) \\ &= P(\chi^2_{(3)} \geq 10.76) \simeq 0.013. \end{aligned}$$

There is evidence against the hypothesis that the four provinces have the same probability of responding Yes. We would write, "the data are not consistent with the hypothesis that respondents from the four provinces are equally likely to support the legalization of pot (p-value = 0.013)." The estimated expected frequencies under the null hypothesis are given in the table below. Note that Ontario has fewer respondents in favour of legalizing pot than would be expected under the hypothesis that those sampled from the four provinces were equally likely to support legalization.

Province	B.C.	Alberta	Sask.	Ontario	Total
Yes	23	19	27	10	79
(Est Expected Yes)	(19.75)	(19.75)	(19.75)	(19.75)	79
No	77	81	73	90	321
Totals	100	100	100	100	400

Note: The form of the likelihood ratio statistic is:

$$D = 2\left[\sum_{all\ \ cells} ObsFreq\ \ln\left(\frac{ObsFreq}{ExpectedFreq}\right)\right],$$

where $ObsFreq$ is the observed frequency and $ExpectedFreq$ is the estimated expected frequency under the null hypothesis. This form for the likelihood ratio statistic will come up again.

2. Test the hypothesis that the three western provinces have respondents who are equally likely to say Yes, whereas Ontario responds differently.

Step 1: BASIC model
The BASIC model stays the same as for question 1.

Step 2: Hypothesized model

$$H_0' : p_1 = p_2 = p_3 = p_W,\ p_4 \qquad \text{unspecified}$$

This is a composite hypothesis because p_W and p_4 are unknown and must be estimated. There are $q = 2$ functionally independent unknown parameters. Substituting the null hypothesis into the log-likelihood, we obtain,

$$\begin{aligned}
\ell_{H'}(p_W, p_4) &= \ell\left(p_1 = p_W, p_2 = p_W, p_3 = p_W, p_4\right) \\
&= \sum_{i=1}^{3} \left[y_i \ln p_W + (n_i - y_i) \ln\left(1 - p_W\right)\right] + y_4 \ln p_4 + (100 - y_4) \ln(1 - p_4) \\
&= (\sum_{i=1}^{3} y_i) \ln p_W + (300 - \sum_{i=1}^{3} y_i) \ln(1 - p_W) + y_4 \ln p_4 + (100 - y_4) \ln(1 - p_4)
\end{aligned}$$

Taking derivatives, setting to zero and solving, we obtain,

$$\begin{aligned}
\frac{\partial \ell_{H'}}{\partial p_W} &= \frac{\sum_{i=1}^{3} y_i}{p_W} - \frac{(300 - \sum_{i=1}^{3} y_i)}{1 - p_W}, \quad \text{and } \tilde{p}_W = \frac{\sum_{i=1}^{3} y_i}{300} = 0.23. \\
\frac{\partial \ell_{H'}}{\partial p_4} &= \frac{y_4}{p_4} - \frac{(100 - y_4)}{1 - p_4}, \quad \text{and } \tilde{p}_4 = \frac{y_4}{100} = 0.10.
\end{aligned}$$

The estimated expected frequencies for the four provinces under the null hypothesis, H_0', are 23, 23, 23 and 10 respectively.

Step 3: Test the Hypothesis

$$\begin{aligned}
D &= 2\left[\ell\left(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4\right) - \ell\left(\tilde{p}_W, \tilde{p}_W, \tilde{p}_W, \tilde{p}_4\right)\right] \\
&= 2\left\{\sum_{i=1}^{4}\left[y_i \ln \frac{y_i}{n_i \tilde{p}_i} + (n_i - y_i) \ln \frac{(n_i - y_i)}{n_i(1 - \tilde{p}_i)}\right]\right\}.
\end{aligned}$$

Substituting in values for $\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4$, and $\tilde{p}_W$ and $\tilde{p}_4$, we obtain $d_{obs} = 1.814$. The p-value for the test is computed as,

$$\begin{aligned}
p - value &\simeq P(\chi^2_{(k-q)} \geq d_{obs}) \\
&= P(\chi^2_{(2)} \geq 1.814) \simeq 0.404.
\end{aligned}$$

We have no evidence against the hypothesis that the Western provinces are equally likely to support legalization of pot.

4.4.1 R Code for Example 4.4.1:

```
y<-c(23, 19, 27, 10)
n<-rep(100,4)
ell<-function(y,n,p){
  sum(y*log(p) + (n-y)*log(1-p))
}

LRS<-function(p,p0,y,n){
  2*(ell(y,n,p)-ell(y,n,p0))
}

phat<-y/n  #MLE under BASIC model
p0<-sum(y)/sum(n)  #MLE under (a) H0
p0
D<-LRS(phat,p0)    #observed LRS
D
1-pchisq(D,4-1)    #p-value

pW<-sum(y[1:3])/sum(n[1:3])
p1<-c(pW,pW,pW,y[4]/n[4])   #MLE under (b) H0
p1
D1<-LRS(phat,p1,y,n)            #observed LRS
D1
1-pchisq(D1,4-2)        #p-value
```

4.5 Tests for Multinomial Probabilities, Goodness of fit test

Optional reading: Chapter 12.5

ASIDE: Note the difference between independent Binomials and Multinomial by noting which marginal totals are fixed.

Example 4.5.1. (Example 2.2.1 revisited). (Goodness of Fit test).

200 specimens of a new high impact plastic are tested by repeatedly striking them with a hammer until they fracture. The data are as follows:

# hits required	1	2	3	≥ 4	Total
# specimens	112	36	22	30	200
frequencies	x_1	x_2	x_3	x_4	$n = 200$

Let

$$Y = \#\text{ hits required to fracture a specimen}$$

We assumed that:

$$f(y) = P(Y = y) = \theta^{y-1}(1-\theta) \qquad x = 1, 2, 3, \ldots$$

where $\theta = P(\text{surviving a hit independently of previous hits})$. Y has a geometric distribution. The assumption did not seem tenable and we found that the model did not yield estimated expected frequencies that were close to the observed frequencies. Now, we can formally test the Goodness of fit of this model using a Likelihood ratio test.

Step 1: BASIC model

$(X_1, X_2, X_3, X_4) \sim Multinomial\,(n = 200, p_1, p_2, p_3, p_4)$.

There are $k = 3$ functionally independent unknown parameters.

$$\begin{aligned}
L(p_1, p_2, p_3, p_4) &= p_1^{x_1}\ p_2^{x_2}\ p_3^{x_3}\ p_4^{x_4} \\
\ell(p_1, p_2, p_3, p_4) &= \sum_{i=1}^{4} x_i \ln p_i \\
\hat{p}_i &= \frac{x_i}{n} = \frac{x_i}{200} \quad \text{(already shown)}
\end{aligned}$$

Step 2: Hypothesized model

$$H_0 : \text{ Geometric } \left\{p_1 = 1-\theta,\ p_2 = \theta(1-\theta),\ p_3 = \theta^2(1-\theta),\ p_4 = \theta^3\right\}.$$

There is only $q = 1$ unknown parameter under the hypothesized model.

We computed the MLE for θ under H_0, $\tilde{\theta} = \frac{1}{2}$ so that

$$\tilde{p}_1 = \frac{1}{2},\ \tilde{p}_2 = \frac{1}{4},\ \tilde{p}_3 = \frac{1}{8},\ \tilde{p}_4 = \frac{1}{8}$$

3. Test the Hypothesis

$$\begin{aligned} D &= -2r\left(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4\right) \\ &= 2\left[\ell\left(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4\right) - \ell\left(\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \tilde{p}_4\right)\right] \\ &= 2\left[\sum_{i=1}^{4} x_i \ln\frac{x_i}{n} - \sum_{i=1}^{4} x_i \ln\tilde{p}_i\right] \\ &= 2\left[\sum_{i=1}^{4} x_i \ln\frac{x_i}{n\tilde{p}_i}\right] \end{aligned}$$

D has a form that we saw earlier,

$$\boxed{D = 2\left[\sum_{all\ \ cells} ObsFreq\ \ \ln\left(\frac{ObsFreq}{ExpectedFreq}\right)\right].}$$

The estimated expected frequencies are given in the table below:

# blows required	1	2	3	≥ 4	Total
# specimens	112	36	22	30	200
frequencies	x_1	x_2	x_3	x_4	$n = 200$
(Est exp freq)	(100)	(50)	(25)	(25)	(200)

Substituting into the formula, we obtain $d_{obs} = 7.048$.

Under $H_0 : D \approx \chi^2_{(k-q)}, \quad k - q = 3 - 1 = 2.$

$$
\begin{aligned}
p - value &= P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) \\
&\simeq P\left(\chi^2_{(2)} \geq 7.048\right) = 0.02948 \\
.025 < p &- value < .05
\end{aligned}
$$

There is evidence against the geometric model.

Exercise: Test the fit of the "extended" geometric model - where we assumed that a proportion λ were defective.

4.5.1 R Code for Example 4.5.1:

```
# Goodness of fit Plastic specimens

freq <- c(112, 36, 22, 30)
sum(freq)
ell <- function(p,freq){
  # Multinomial log-likelihood
  # freq = frequencies;  p = probabilities
  sum(freq*log(p))
}


LRS <- function(p0, phat,freq){
  #Log relative likelihood function
  2*(ell(phat,freq) - ell(p0,freq))
}

ptilde <- c(.5, .25, .125, .125)
dobs <- LRS(ptilde,freq/sum(freq),freq)   #LRS observed
dobs
1-pchisq(dobs,2)  #pvalue
```

4.6 Multinomial Probabilities - Tests for Independence in Contingency Tables

Optional reading: 12.6

Contingency tables are also called two-way tables or cross-tabulations.

Example 4.6.1. It was noted that married undergraduates seemed to do better academically than single students. The following observations were made on the examination results of 1500 engineering students. Students were asked to check a box on the examination booklet indicating if they were married or single.

	Fail	Pass	Total
Married	14	143	157
Single	283	1060	1343
Total	297	1203	1500

Are these observations consistent with the hypothesis of a common failure rate for single and married students? Use a Likelihood Ratio test to answer the question.

Step 1: BASIC Model

First, which totals are known before examination is written? Only 1500!

Assume that an engineering student falls in exactly one of the 4 categories independently of the other students and with a constant probability given in the table below:

	Fail	Pass	Total
Married	p_{11}	p_{12}	
Single	p_{21}	p_{22}	
Total			1

Because only the total number of students, 1500, is fixed in advance of the examination, then the data arise from a single Multinomial distribution and $\sum_{i,j=1}^{2} p_{ij} = 1$.

Let's introduce some general notation and label the frequencies as follows:

	Fail	Pass	Total
Married	x_{11}	x_{12}	r_1
Single	x_{21}	x_{22}	r_2
Total	c_1	c_2	$n = 1500$

We have that $(X_{11}, X_{12}, X_{21}, X_{22}) \sim$ Multinomial $(n = 1500, p_{11}, p_{12}, p_{21}, p_{22})$ and the number of functionally independent parameters is $k = 3$.

Then the MLE for p_{ij} under BASIC model is, $\hat{p}_{ij} = \frac{x_{ij}}{n}$.

Step 2: Hypothesized model

The hypothesis is that the failure probability is the same for married and single students. We write that in terms of the model as follows:

$$H_0 : P(\text{fail} \mid \text{married}) = P(\text{fail} \mid \text{single}) = P(\text{fail}) = \alpha \quad \text{unspecified.}$$

If H_0 is true, then

$$P\left(\text{pass} \mid \text{married}\right) = 1 - \alpha = P\left(\text{pass} \mid \text{single}\right) = P\left(\text{pass}\right)$$

In other words, the null hypothesis states that pass/fail on the examination is independent of marital status!

Let

$$\begin{aligned} \beta &= P\left(\text{married}\right) \quad \text{then,} \\ p_{11} &= P\left(\text{fail \& married}\right) = \alpha\beta \\ p_{12} &= P\left(\text{pass \& married}\right) = \left(1 - \alpha\right)\beta \\ p_{21} &= P\left(\text{fail \& single}\right) = \alpha\left(1 - \beta\right) \\ p_{22} &= P\left(\text{pass \& single}\right) = \left(1 - \alpha\right)\left(1 - \beta\right) \end{aligned}$$

Here $q = 2$, as we have two parameters that require estimation.

We will use the Likelihood ratio statistic for testing H_0, therefore we need to compute the MLE's of the p_{ij}'s under the hypothesized model.

$$\begin{aligned} L(\alpha,\beta) &= [\alpha\beta]^{14}\,[(1-\alpha)\,\beta]^{143}\,[\alpha\,(1-\beta)]^{283}\,[(1-\alpha)\,(1-\beta)]^{1060} \\ &= \alpha^{297}\,(1-\alpha)^{1203}\,\beta^{157}\,(1-\beta)^{1343} \end{aligned}$$

The Likelihood looks the same as that for two independent binomials!

$$\begin{aligned} \ell(\alpha,\beta) &= 297\ln\alpha + 1203\ln(1-\alpha) + 157\ln\beta + 1343\ln(1-\beta) \\ \tilde{\alpha} &= \frac{297}{1500} = \frac{c_1}{n} = \text{proportion who fail} \\ \tilde{\beta} &= \frac{157}{1500} = \frac{r_1}{n} = \text{ proportion married} \end{aligned}$$

We use these to compute the $\widetilde{p}_{ij}$'s and the estimated expected frequencies under H_0 in each of the cells in the table as follows:

$$\begin{aligned} e_{ij} &= \text{estimated expected freq. in } (i,j)^{th} \text{ cell under hypothesized model} \\ e_{11} &= n\widetilde{p}_{11} = n\widetilde{\alpha}\widetilde{\beta} = \frac{r_1c_1}{n} = \frac{(297)(157)}{1500} \\ e_{12} &= n\widetilde{p}_{12} = n(1-\widetilde{\alpha})\widetilde{\beta} = \frac{r_1c_2}{n} \\ e_{21} &= n\widetilde{p}_{21} = n\widetilde{\alpha}(1-\widetilde{\beta}) = \frac{r_2c_1}{n} \\ e_{22} &= n\widetilde{p}_{22} = n(1-\widetilde{\alpha})(1-\widetilde{\beta}) = \frac{r_2c_2}{n} \end{aligned}$$

The estimated expected frequencies under the independence hypothesis are included in the original data table in parentheses.

Observed frequencies (estimated expected frequencies under H_0)

	Fail	(Fail)	Pass	(Pass)	Total
Married	14	(31.086)	143	(125.914)	157
Single	283	(265.914)	1060	(1077.086)	1343
Total	297		1203		1500

Step 3: Test the hypothesis

From the last section, we have that the form of the Likelihood ratio statistic for the multinomial model is:

$$D = 2\left[\sum_{all\ cells} ObsFreq\ \ln\left(\frac{ObsFreq}{ExpectedFreq}\right)\right].$$

Rewriting that using the notation of this section,

$$\begin{aligned}
D &= 2\left[\sum_{\text{all cells}} x_{ij}\ \ln\frac{x_{ij}}{e_{ij}}\right]\\
d_{\text{obs}} &= 2\,(7.70) = 15.40\\
\text{p-value} &= P\left(D \geq d_{\text{obs}} \mid H \text{ true}\right)\\
&\simeq P\left(\chi^2_{(k-q)} \geq 15.40\right) \quad k=3, \quad q=2\\
&= P\left(\chi^2_{(1)} \geq 15.40\right) = .00009 < .001
\end{aligned}$$

We have very strong evidence against the hypothesis that there is a common failure rate for single and married students.

The data suggest that there is an association between marital status and whether students pass or fail the examination.

What is the nature of the association?

$$\text{Proportion of married who failed} = \frac{14}{147} = .089$$

157

$$\text{Proportion of single who failed} = \frac{283}{1343} = .211$$

The data suggests that married students are less likely to fail the exam. Does that mean that marriage causes better outcomes on examinations?

4.6.1 R Code for Example 4.6.1:

```
# Test of Independence

freq <- matrix(c(14, 143, 283, 1060), nrow=2, byrow=TRUE)
freq
sum(freq)
ell <- function(p,freq){
  # Multinomial log-likelihood
  # freq = frequencies;  p = probabilities
  sum(freq*log(p))
}

LRS <- function(p0, phat,freq){
  #Log relative likelihood function
  2*(ell(phat,freq) - ell(p0,freq)
}

rsum <- rowSums(freq)
csum <- colSums(freq)
rsum
csum

eij <- outer(rsum, csum)/sum(freq)  # exp freq
eij                                 # eij= r_i * c_j / n

# estimated probs under H0 are eij/sum(freq)

dobs <- LRS(c(eij)/sum(freq),c(freq)/sum(freq),c(freq))   #LRS observed
dobs
1-pchisq(dobs,1)  #pvalue, df=(#rows - 1)(#cols - 1)
```

4.7 Cause and Effect , Accuracy of χ^2 approximation

Optional reading: Sections 12.7

The statement "A and B are associated" means that A and B tend to occur together.

This does not mean that A causes B! There are 3 possible cause-effect relationships that could produce the association:

(i) A causes B
(ii) B causes A
(iii) some other factor C causes both A and B

We cannot claim that A causes B until we have ruled out (ii) and (iii).

4.7.1 Accuracy of the χ^2 approximation

There are situations where the χ^2 approximation to the distribution of the Likelihood ratio statistic for multinomial/binomial data is inaccurate.

The χ^2 approximation should not be trusted if there are categories for which $e_i \simeq 0$ but $x_i \geq 1$.

Rule of thumb: e_i should be ≥ 5.

Remedy: If there are several categories for which $e_i < 5$, pool adjacent categories to increase the corresponding $e_i's$.

4.8 The General Contingency Table

Consider n independent repetitions of an experiment and classify each outcome in two ways according to which of events, A_i or B_j, $i = 1, ..., a$, $j = 1, ..., b$ occur, where

(i) $A_1, \ldots, A_a$ is a partition of the A sample space
(ii) $B_1, \ldots, B_b$ is a partition of the B sample space,

so that each outcome belongs to exactly one of the $A_i's$, and each outcome belongs to exactly one of the $B_j's$.

The data is tabulated using the following notation:

Observed frequencies

	B_1	B_2	$\ldots$	B_b	Total
A_1	x_{11}	x_{12}		x_{1b}	r_1
A_2	x_{21}	x_{22}			r_2
$\vdots$					
A_a	x_{a1}	x_{a2}		x_{ab}	r_a
Total	c_1	c_2		c_b	n

Let $p_{ij} = P\left\{\text{an outcome falls in class } A_iB_j\right\}$, then $\sum_{i,j} p_{ij} = 1$

Under the assumption of independent repetitions, the BASIC model is

$$(X_{11}, X_{12}, \ldots, X_{ab}) \sim \text{Multinomial}\left(n, p_{11}, \ldots, p_{ab}\right).$$

There are $k = ab - 1$ functionally independent unknown parameters in the BASIC model.

$$P\left(x_{11}, x_{12}, \ldots, x_{ab}\right) = \binom{n}{x_{11}, \ldots, x_{ab}} p_{11}^{x_{11}}, \ldots, p_{ab}^{x_{ab}}$$

$$\ell\left(\boldsymbol{p}\right) = \sum_{i=1}^{a} \sum_{j=1}^{b} x_{ij} \ln p_{ij}, \text{ where } \boldsymbol{p} = (p_{11}, p_{12}, \ldots, p_{ab})'.$$

The Likelihood ratio statistic for testing hypotheses about $\boldsymbol{p}$ is:

$$\begin{aligned} D &= 2\left[\ell(\hat{\boldsymbol{p}}) - \ell(\tilde{\boldsymbol{p}})\right] \quad \text{where } \hat{p}_{ij} = \tfrac{x_{ij}}{n} \\ &= 2\left[\sum_i \sum_j x_{ij} \ln \tfrac{x_{ij}}{n\tilde{p}_{ij}}\right] \\ &= 2\left[\sum_i \sum_j x_{ij} \ln \tfrac{x_{ij}}{e_{ij}}\right] \quad (e_{ij} = n\tilde{p}_{ij}), \end{aligned}$$

where e_{ij} are the estimated expected frequencies under the hypothesized model, H_0.

$D \approx \chi^2_{(ab-1-q)}$ where q is the number of functionally independent unknown parameters in the hypothesized model.

The null Hypothesis of Independence can be written as:

$$\begin{aligned} & H_0 : P\left(A_i B_j\right) = P\left(A_i\right) P\left(B_j\right) && \text{for all } i, j \\ \text{or} \quad & H_0 : P\left(B_j \mid A_i\right) = P\left(B_j\right) && \text{for all } i, j \end{aligned}$$

Under H_0, the unknown parameters are:

$$\begin{aligned} & \alpha_i = P\left(A_i\right) \quad i = 1, \ldots, a \quad \textstyle\sum_{i=1}^{a} \alpha_i = 1 \\ & \beta_j = P\left(B_j\right) \quad j = 1, \ldots, b \quad \textstyle\sum_{j=1}^{b} \beta_j = 1 \\ \Longrightarrow \quad & q = (a-1) + (b-1) \end{aligned}$$

It can be shown [Exercise!] that $\tilde{\alpha}_i = \frac{r_i}{n}$ $\tilde{\beta}_j = \frac{c_j}{n}$, so that,

$$e_{ij} = n\tilde{p}_{ij} = n\tilde{\alpha}_i\tilde{\beta}_j = \frac{r_i c_j}{n}.$$

The Likelihood ratio test statistic for testing H_0 has an approximate χ^2 distribution with degrees of freedom computed as:

$$k - q = ab - 1 - [(a-1) + (b-1)] = (a-1)(b-1).$$

Example 4.8.1. The following data on heights of 210 married couples were presented by Yule in 1900.

		Wife		
Husband	Tall	Medium	Short	Total
Tall	18 (15.48)	28 (32.19)	19 (17.33)	65
Med	20 (23.57)	51 (49.03)	28 (26.4)	99
Short	12 (10.95)	25 (22.78)	9 (12.27)	46
Total	50	104	56	210

Test the hypothesis that heights of husbands and wives are independent.

Solution: The estimated expected frequencies, $e_{ij} = \frac{r_i c_j}{n}$ under the hypothesis of independence are given in parentheses in the table.

The Likelihood ratio test for testing H_0 is:

$$D = 2\left[\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}\ln\frac{x_{ij}}{e_{ij}}\right]$$

$$\begin{aligned}
p - value &= P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) \\
&\simeq P\left(\chi^2_{(k-q)} \geq 3.13\right) \qquad k - q = (a-1)(b-1) = 4 \\
&= P\left(\chi^2_{(4)} \geq 3.13\right) \simeq 0.54
\end{aligned}$$

There is no evidence against the hypothesis of independence. The data suggest that the heights of husbands and wives are not associated.

4.8.1 R Code for Example 4.8.1:

```
# Test of Independence - Yule example

freq <- matrix(c(18, 28, 19, 20, 51, 28, 12, 25, 9), nrow=3, byrow=TRUE)
freq
sum(freq)
```

```
ell <- function(p,freq){
  # Multinomial log-likelihood
  # freq = frequencies;  p = probabilities
  sum(freq*log(p))
}

LRS <- function(p0, phat,freq){
  #Log relative likelihood function
  2*(ell(phat,freq) - ell(p0,freq))
}

rsum <- rowSums(freq)
csum <- colSums(freq)
rsum
csum

eij <- outer(rsum, csum)/sum(freq)  # exp freq
eij                                 # eij= r_i * c_j / n

# estimated probs under H0 are eij/sum(freq)

dobs <- LRS(c(eij)/sum(freq),c(freq)/sum(freq),c(freq))   #LRS observed
dobs
1-pchisq(dobs,4)  #pvalue, df=(#rows - 1)(#cols - 1)
```

4.8.2 Pearson's χ^2 Goodness of Fit Statistic

Pearson's Goodness of Fit Statistic may be used with multinomial or binomial data.

$$\text{G.O.F.} = \sum_{\text{all cells}} \frac{(x_j - e_j)^2}{e_j} \quad \text{where } e_j = \text{estimated expected under } H_0$$

$$\text{G.O.F.} \approx \chi^2_{(k-q)}$$

When the e_j's are large, G.O.F. will be very nearly equal to the Likelihood ratio statistic.

For the Yule heights data in Example 4.8.1,

$$\text{G.O.F.} = 3.02, \quad \text{which yields } p - value \simeq 0.55,$$

and there is no evidence against H_0.

4.8.3 R Code for Pearson's GOF test, Example 4.8.1:

```
# Pearson Goodness-of-fit Statistic
#  use the freq and eij from the code in the previous section

GOF <- sum((freq-eij)^2/eij)
1-pchisq(GOF,4)
```

Chapter 5

Confidence Intervals

Optional reading: Section 11.4

A confidence interval is interpreted as a range of "reasonable" values for a parameter given the data. Earlier we considered the use of the relative likelihood function in determining which values of an unknown parameter θ are plausible in the light of the data. Values within the 10% Likelihood Interval are considered plausible because they give the data at least 10% of the maximum probability which is possible under the model.

Here we consider another way, based on tests of significance, of constructing a "reasonable" interval of values for an unknown parameters. We show how these intervals are related to Likelihood Intervals.

Definition: $[A, B]$ is a $100(1-\alpha)\%$ **confidence interval** for θ if $CP(\theta_0) = P(A \leq \theta_0 \leq B \mid \theta = \theta_0) = 1-\alpha$, (the coverage probability) for all parameter values θ_0. A 95% confidence interval would include the true parameter value θ_0 in 95% of repetitions of the experiment with θ fixed.

5.1 Invert a Test to Derive a Confidence Interval

Example 5.1.1. The Small Business Association has taken a sample of 50 small businesses to investigate the effect that the recession has had on profit levels. Test the government's claim that the recession has had no effect on profits, based on an observed sample average decrease in profits of \$1,000. Assume that measurements are independent and normally distributed with $\sigma = \$600$.

Solution:

Let X_i = change in profit in dollars for business i.

Step 1: BASIC model

Assume $X_i \sim N\left(\mu, \sigma^2 = 600^2\right), i = 1, .., 50$, and therefore $k = 1$.

$$L(\mu) = \prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$

$$\ell(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \quad \text{and we know} \quad \hat{\mu} = \bar{x}.$$

Step 2: Hypothesized Model $H_0 : \mu_0 = 0$, i.e. that there is no change in mean profits. Here $q = 0$, since there are no unknown parameters to estimate.

Step 3: Test the hypothesis

$$\begin{aligned} LRS = D &= -2r(\mu_0) = 2\left[\ell(\hat{\mu}) - \ell(\mu_0)\right] \\ &= \frac{n\left(\bar{X} - \mu_0\right)^2}{\sigma^2}. \end{aligned}$$

We obtained this result in the last chapter.

$$\text{Here } n = 50, \quad \sigma^2 = 600^2, \text{ and } \quad \bar{x} = -1000$$

$$\begin{aligned} d_{\text{obs}} &= \frac{50\left(-1000 - 0\right)^2}{600^2} = 8\cancel{5} \quad 138.889 \\ \text{p-value} &= P\left(D \geq d_{\text{obs}} \mid H_0 \text{ true}\right) \\ &= P\left(\chi^2_{(1)} \geq \underset{138,889}{\cancel{85}}\right) < .001 \end{aligned}$$

We have very strong evidence against H_0 of no effect due to the recession.

Now, we consider, for which parameter values, μ_0, does a likelihood ratio test of $H_0 : \mu = \mu_0$ yield a

$$p - value(\mu_0) \geq 0.05?$$

What values of μ_0 are reasonably consistent with the data?

$$p - value(\mu_0) = P\left(\chi^2_{(1)} \geq d_{\text{obs}} \mid H : \mu = \mu_0\right) \geq .05$$

where

$$d_{\text{obs}}(\mu_0) = \frac{n\left(\bar{x} - \mu_0\right)^2}{\sigma^2}.$$

From tables or R, $P\left(\chi^2_{(1)} \geq 3.841\right) = .05$

therefore $p - value(\mu_0) \geq .05$ if and only if $d_{\text{obs}}(\mu_0) \leq 3.841$. We must solve for the values μ_0 such that

$$\begin{aligned}
\frac{n\left(\bar{x} - \mu_0\right)^2}{\sigma^2} &\leq 3.841. \\
\left(\bar{x} - \mu_0\right)^2 &\leq 3.841\frac{\sigma^2}{n} \\
-1.96\frac{\sigma}{\sqrt{n}} &\leq \bar{x} - \mu_0 \leq 1.96\frac{\sigma}{\sqrt{n}} \\
\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} &\leq \mu_0 \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Here $\bar{x} = -1000, \quad \sigma = 600, \quad n = 50$, so the observed confidence interval is $[-1166, -834]$. Values of μ_0 within this interval are consistent with the data, and are reasonable estimates of μ. Since the confidence interval lies below zero, the government's claim that the recession has had no effect on profits is not plausible given the data.

One property of the Interval:

For a random $\bar{X}$, the interval $\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right]$ is a random interval. If we repeated the experiment, we would obtain a different $\bar{x}$ and a different interval estimate of μ. Let μ_T be the true unknown value of μ. We can compute the fraction of times that the random interval would include the true value μ_T in a large number of repetitions of the experiment.

$$\begin{aligned} & P\left\{\bar{X}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu_T \leq \bar{X}+1.96\frac{\sigma}{\sqrt{n}} \mid \mu=\mu_T\right\} \\ = & \ P\left\{-1.96 \leq \left(\frac{\bar{X}-\mu_T}{\frac{\sigma}{\sqrt{n}}}\right) \leq 1.96 \mid \mu=\mu_T\right\} = .95 \\ & \frac{\bar{X}-\mu_T}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \quad \text{if} \quad \mu=\mu_T \end{aligned}$$

With probability .95 the interval

$$\left[\bar{X}-1.96\frac{\sigma}{\sqrt{n}}, \bar{X}+1.96\frac{\sigma}{\sqrt{n}}\right]$$

contains μ_T, the true value.

This interval is called a **95% Confidence Interval** for μ.

We are confident that in 95 times out of 100 trials of the experiment the above interval will contain μ_T, the true value.

Definition: *A 95% Confidence Interval, CI, is an interval for the unknown parameter μ_T, such that in a large number of repetitions of the experiment, CI covers μ_T 95 times out of 100.*

One way to construct a 95% CI is to solve for the set of parameter values μ_0 such that for a test of $H : \mu = \mu_0$

$$\text{p-value}\,(\mu_0) \geq 0.05.$$

which we do in the next section.

5.2 Approximate Confidence Intervals

We consider a Basic model that has one, unknown parameter, θ, and let

$D = -2r(\theta_0)$, be the Likelihood Ratio Statistic for testing $H_0 : \theta = \theta_0$.

We know that $D \approx \chi^2_{(1)}$, so that $p - value \simeq P\left[\chi^2_{(1)} \geq d_{\text{obs}}(\theta_0)\right]$,

where $d_{\text{obs}}(\theta_0)$ is the observed value of D when $\theta = \theta_0$.

Since $P\left(\chi^2_{(1)} \geq 3.841\right) = .05$, then $p - value(\theta_0) \geq .05 \iff d_{\text{obs}}(\theta_0) \leq 3.841$.

Thus, an approximate 95% Confidence interval for θ is the set of θ_0 values such that,

$$\begin{aligned} d(\theta_0) &\leq 3.841 \\ \iff -2r(\theta_0) &\leq 3.841 \\ \iff r(\theta_0) &\geq -1.92 \\ \iff e^{r(\theta_0)} &= R(\theta_0) \geq e^{-1.92} = .147 \end{aligned}$$

Thus, an approximate 95% Confidence Interval for θ is just a 14.7% Likelihood Interval! Below is a table of common confidence interval levels and their corresponding likelihood intervals.

CI %	Corresponding LI %
90	25.8
95	14.7
96.8	10
99	3.6

Analogy: Pitching Horseshoes

Constructing a 95% confidence interval is like pitching horseshoes. In each case, there is a fixed target either the population μ or the stake. We are trying to bracket the target with some chancy device, either the random interval or the horseshoe.

There are several important ways in which confidence intervals differ. Customarily, only one Confidence Interval is constructed, for a target μ that is not visible. Consequently, the statistician does not know directly whether his/(her) Confidence Interval includes the true value; he/(she) must rely on indirect statistical theory for assurance in the long run, 95% of the Confidence Intervals similarly constructed would include the true value.

Example 5.2.1. (Example 4.4.1 revisited) In a random sample of 100 people from B.C., it was found that 23 out of 100 favour legalization of pot. Construct an approximate 95% Confidence Interval for the proportion of B.C. voters who support legalization of pot.

$$\begin{aligned} \theta &= \text{ proportion who support legalization of pot} \\ X &= \text{Number out of 100 sampled who support legalization of pot} \end{aligned}$$

$$X \sim Binomial(n = 100, \theta)$$

We want the set of all values θ_0, such that in a test $H_0 : \theta = \theta_0$

$$\text{p-value}\,(\theta_0) \geq .05$$

Method 1: Use Likelihood Ratio Statistic Using the methods of this section, we find a 14.7% Likelihood interval for θ. This works out to be $[.155, .319]$ using the R-code which is included at the end of this chapter.

Method 2: Use $D = |X - n\theta|$

Here we use $D = |X - n\theta|$, so that

$$p - value(\theta_0) = P\left\{|X - n\theta_0| \geq |x_{\text{obs}} - n\theta_0| \mid H : \theta = \theta_0\right\}.$$

For $n\theta, n(1-\theta) \geq 5$, $X \approx N\left(n\theta, n\theta\left(1-\theta\right)\right)$.

$$\begin{aligned} p - value(\theta_0) &= P\left\{\frac{|X - n\theta_0|}{\sqrt{n\theta_0\left(1-\theta_0\right)}} \geq \frac{|x_{\text{obs}} - n\theta_0|}{\sqrt{n\theta_0\left(1-\theta_0\right)}} \mid H_0 : \theta = \theta_0\right\} \\ &\simeq P\left\{|Z| \geq \frac{|x_{\text{obs}} - n\theta_0|}{\sqrt{n\theta_0\left(1-\theta_0\right)}}\right\} \end{aligned}$$

$$\text{where} \quad Z \sim N(0,1)$$

$$\text{p-value}(\theta_0) \geq .05 \iff \frac{|x_{\text{obs}} - n\theta_0|}{\sqrt{n\theta_0(1-\theta_0)}} \leq 1.96$$

(i) We can solve the quadratic for θ_0. [Exercise!]

(ii) Or we can approximate $VAR(X) = n\theta_0(1-\theta_0)$ with $n\hat{\theta}(1-\hat{\theta})$ where $\hat{\theta} = \frac{x_{\text{obs}}}{n} = .23$

Using method (ii) - Solve for θ_0.

$$|x_{\text{obs}} - n\theta_0| \leq 1.96\sqrt{n\hat{\theta}(1-\hat{\theta})}$$

$$\frac{x_{\text{obs}}}{n} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta_0 \leq \frac{x_{\text{obs}}}{n} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

An approximate 95% Confidence Interval for θ is:

$$\left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right]$$
$$= [.148, .312]$$

This gives us an indication of the precision of our estimate and the margin of error for our estimate is 0.082. Intervals constructed this way have the property that 19 times out of 20, such intervals will cover the true value.

This is the basis for quotes in the media such as,

> “is considered accurate to within 8.2 percentage points, 19 times out of 20”.

5.2.1 R Code for Example 5.2.1

```
# Binomial Example 95% CI using LI
# Log-likelihood function
ell <- function(theta){
  23*log(theta) + 77*log(1-theta)
}
theta <- seq(.1,.45,by=.01)
plot(theta,ell(theta),ylab='log likelihood',xlab='theta',
     type='l')
title('Example Binomial CI, Log-Likelihood')


#MLE of theta
thetahat <- optimize(ell, c(.1,.6), maximum=TRUE)
thetahat

#Log relative likelihood function
logR <- function(theta, thetahat){
  ell(theta) - ell(thetahat)
}
p <- .147  #14.7% likelihood interval
logR.m.lnp <- function(theta, thetahat, p) {logR(theta,thetahat)-log(p)}

plot(theta,logR.m.lnp(theta,thetahat$maximum,p), ylab='log relative likelihood',
     xlab='theta',type='l')
abline(h=0)   #add a horizontal line at zero
title('Example Binomial CI, Log Relative Likelihood')


#Likelihod intervals
lower <- uniroot(logR.m.lnp, c(.1, .23), thetahat$maximum, p)
lower

upper <- uniroot(logR.m.lnp, c(.23, .6), thetahat$maximum, p)
upper
```

5.3 Another Approximate Confidence Interval

Optional reading: 11.3

We learned in Section 4.2.1 that the Likelihood Ratio Statistic has an approximate $\chi^2_{(1)}$ distribution in the one parameter case assuming the null hypothesis to be true.

A related, and extremely useful result, is that in many cases, the MLE has an approximate normal distribution for large sample size, n. Again, consider a Basic model that has one, unknown parameter, θ, and let $\hat{\theta}$ be the MLE of θ where the 'true' value of $\theta = \theta_0$. Then for many probability models,

$$\left(\hat{\theta} - \theta_0\right) \sqrt{I(\hat{\theta})} \approx N(0,1),$$

Where $I(\hat{\theta})$ is the Information function defined in Section 2.1 evaluated at $\hat{\theta}$, $I(\theta) = -\ell''(\theta) = -\frac{d^2\ell(\theta)}{d\theta^2}$. This result can be rewritten as,

$$\hat{\theta} \approx N\left(\theta_0, I(\hat{\theta})^{-1}\right),$$

so that $\hat{\theta}$ is approximately normally distributed with the true value as its mean, and with asymptotic variance $1/I(\hat{\theta})$. The result generalizes to the multi-parameter case.

Using this result, we obtain an approximate $100(1-\alpha)\%$ Confidence Interval as,

$$\boxed{\left(\hat{\theta} - \frac{z_{1-\alpha/2}}{\sqrt{I(\hat{\theta})}},\ \hat{\theta} + \frac{z_{1-\alpha/2}}{\sqrt{I(\hat{\theta})}}\right)}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $N(0,1)$ distribution. For example, for $\alpha = .05$, $z_{1-\alpha/2} = 1.96$.

This result is used very frequently in applied statistics!!

Example 5.3.1. (Example 5.2.1 revisited) In a random sample of 100 people from B.C., it was found that 23 out of 100 favour legalization of pot. Construct an approximate 95% Confidence Interval for the proportion of B.C. voters who support legalization of pot using the normal approximation for the MLE.

$$\begin{aligned} \theta &= \text{ proportion who support legalization of pot} \\ X &= \text{Number out of 100 sampled who support legalization of pot} \end{aligned}$$

$$X \sim \mathit{Binomial}(n = 100, \theta)$$

The Log-likelihood, Score and Information function are respectively:

$$\begin{aligned} \ell(\theta) &= x \ln \theta + (n - x) \ln(1 - \theta) \\ \ell'(\theta) &= \frac{x}{\theta} - \frac{n - x}{1 - \theta} \\ \ell''(\theta) &= -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2} \\ I(\theta) &= \frac{x}{\theta^2} + \frac{n - x}{(1 - \theta)^2}, \end{aligned}$$

and after some simplification,

$$I(\hat{\theta})^{-1} = \frac{\hat{\theta}(1 - \hat{\theta})}{n}.$$

This yields an approximate Confidence Interval for θ, the proportion of B.C. voters who support legalization of pot as,

$$\begin{aligned} &\left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}, \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}}\right] \\ &= [.148, .312]\,, \end{aligned}$$

which is the same as what we obtained in the previous section using Method 2!

Chapter 6

Normal Theory

Optional reading: Sections 13.1, 13.2

The normal distribution plays a large role in modelling and the statistical analysis of continuous measurements. Many types of measurements have distributions which are approximately normal - the Central Limit Theorem helps to explain this. In the next sections, we will concentrate solely on models for normal measurements: Maximum Likelihood Estimation, tests of hypotheses and confidence intervals for normal measurements taken under varying conditions.

Before we do, recall:

(1) Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim N(\mu_i, \sigma_i^2)$.

Let $a_1, \ldots, a_n$ be constants. Then

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

(2) Let $Z_1, \ldots, Z_n$ be independent $N(0,1)$ random variables, then

$$Z_1^2 \sim \chi^2_{(1)} \text{ and } \sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}.$$

6.1 Basic Assumptions

We assume that measurements are independent and normally distributed with constant variance, $Y_i \sim N(\mu_i, \sigma^2)$.

Under this assumption, we assume that the effect of changing conditions is to alter μ. We can write the model in terms of independent error variables, $\varepsilon_1, \ldots, \varepsilon_n$ where

$$\varepsilon_i = Y_i - \mu_i \sim N\left(0, \sigma^2\right)$$

Then

$$Y_i = \mu_i + \varepsilon_i \text{ where } \varepsilon_i \sim N(0, \sigma^2) \text{ independent.}$$

A consequence is that,

$$P\left\{-3\sigma \leq \varepsilon_i \leq 3\sigma\right\} = .9973 \approx 1.$$

The smaller σ~~,~~ is, the smaller we expect ε_i to be. σ measures the amount of random variability (noise) that one would expect in repeated measurements taken under the same conditions.

Assumptions Concerning $\mu_1, \mu_2, \ldots, \mu_n$ We will express the n mean parameters as functions of q parameters, where $q < n$.

(1) **One-Sample Model**: n measurements taken under the same conditions

e.g. blood pressure measurements, Y_i for a group of patients all receiving the same drug

Assume: $\mu_1 = \cdots = \mu_n = \alpha$ unknown

There is $q = 1$ unknown mean parameter, assuming σ^2 is known.

(2) **Two-Sample Model**: 2 groups of sample measurements

e.g. salaries for co-op students: $2nd$ year, $3rd$ year

Assume: $\mu_{2nd} = \alpha$
$\mu_{3rd} = \alpha + \beta$

There are $q = 2$ unknown mean parameters, assuming σ^2 is known.

(3) **Straight Line Model**: n measurements taken under varying conditions

e.g. salaries for recently graduated students, Y_i, depend upon the number of co-op work terms, x_i they performed.

Here, the x_i are known constants,
Y_i vary,

Assume $Y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \alpha_i + \beta x_i$ and α, β are unknown parameters.

There are $q = 2$ unknown mean parameters, assuming σ^2 is known.

6.2 One Sample Model

Optional reading: Section 13.3

Example 6.2.1. Monthly salaries for Engineering and Math co-op students were collected. A sample of 10 salaries is given below (in $):

5050	4184	1787	2167	2650
5499	3163	3016	3120	4333

Compute a 95% Confidence Interval for the mean monthly salary, assuming that $Y_i =$ monthly salaries for person $i = 1, ..., n$ are $N(\mu, \sigma^2)$ and independent.

To assess the assumption of normality, we could consider a histogram as in Figure 6.1. Unfortunately, histograms are not very informative for small samples. A normal QQ (quantile-quantile) plot is given in Figure 6.2. If the data are approximately normally distributed, then this graph should resemble a straight line. In this case, it does(!) and we have no evidence against the normal assumption. We will learn about normal quantile-quantile (QQ) plots in Section 6.4.5.

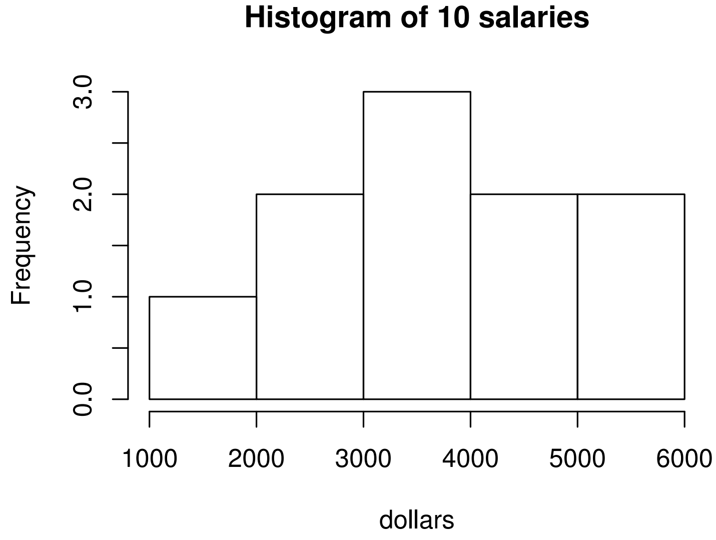


Figure 6.1: Histogram of 10 Salaries

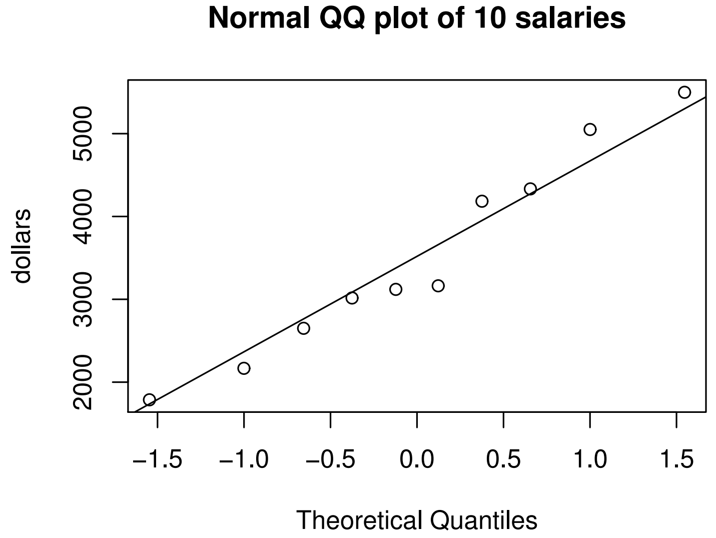


Figure 6.2: QQ plot of 10 Salaries

6.2.1 Confidence Intervals for μ

Confidence Interval for μ when σ^2 known

In the case that σ^2 is known, we saw from the chapter on confidence intervals that a 95% CI for μ is: $\left[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$. This is the set of μ_0 such that $p - value(\mu_0) \geq 0.05$ in a Likelihood ratio test of $H_0 : \mu = \mu_0$. The value $1.96 = z_{1-\alpha/2}$, where $\alpha = 0.05$.

Confidence Interval for μ when σ^2 unknown

One can show [EXERCISE!] that the Likelihood ratio test for testing $H_0 : \mu = \mu_0$ when σ^2 is unknown is

$$D = n \ln \left[1 + \frac{1}{n-1} T^2\right]$$

where

$$T = \frac{\widehat{\mu} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{\bar{y} - \mu_0}{\frac{s}{\sqrt{n}}}$$

and

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

We can compute a 95% Confidence interval for μ, by finding all μ_0 such that $p - value(\mu_0) \geq 0.05$. Instead note that D is a one-to-one increasing transformation of T^2 which I will call $g(D) = T^2$. $= (n-1)\left[e^{D/n} - 1\right]$

$$\begin{aligned} p - value(\mu_0) &= P\left\{D \geq d_{\text{obs}}(\mu_0) \mid H_0 : \mu = \mu_0\right\} \\ &= P\left\{g(D) \geq g(d_{\text{obs}}(\mu_0)) \mid H_0 : \mu = \mu_0\right\} \\ &= P\left\{T^2 \geq t^2_{\text{obs}}(\mu_0) \mid H_0 : \mu = \mu_0\right\}. \end{aligned}$$

We have that,

$$p - value(\mu_0) \geq 0.05 \iff t^2_{\text{obs}}\left(\mu_0\right) \leq a^2$$

where a^2 is chosen so that

$$\begin{aligned}&P\left\{T^2 \leq a^2 \mid H_0 : \mu = \mu_0\right\} = 0.95\\&= P\left\{-a \leq T \leq a \mid H_0 : \mu = \mu_0\right\}\\&= P\left\{-a \leq \frac{\overline{Y} - \mu_0}{\frac{s}{\sqrt{n}}} \leq a \mid H_0 : \mu = \mu_0\right\}.\end{aligned}$$

The exact distribution of T is known. Rewriting T as,

$$T = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \div \sqrt{\frac{s^2}{\sigma^2}},$$

we examine the two pieces in the expression.

(i) $Z = \frac{\overline{Y}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N\left(0,1\right)$ when $\mu = \mu_0$

(ii) $V = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ independent of Z. (Proved in Stat450)

$E\left[V\right] = n - 1$. (See Chapter 3)

$E\left[s^2\right] = E\left[\frac{\sigma^2}{n-1}V\right] = \frac{\sigma^2}{n-1}E\left[V\right] = \sigma^2$

Thus, s^2 is an unbiased estimate of σ^2.

Putting the pieces together,

$$\begin{aligned}T &= \frac{Z}{\sqrt{\frac{V}{(n-1)}}}\\&= \frac{N(0,1)}{\sqrt{\frac{\chi^2_{(n-1)}}{n-1}}} \sim t_{(n-1)},\end{aligned}$$

the Student's t distribution with $n-1$ degrees of freedom. Percentiles of the Student's t distribution are tabulated in Table B3. If the degrees of freedom are very large, > 60, then the t distribution approaches N(0,1). In general, for any $Z \sim N(0,1)$ independent of $V \sim \chi^2_{(\nu)}$,

$$\frac{Z}{\sqrt{\frac{V}{\nu}}} \sim t_{(\nu)}.$$

Returning to our example, we want to find a such that,

$$P\left\{-a \leq \frac{\bar{Y}-\mu_0}{\frac{s}{\sqrt{n}}} \leq a \mid H_0 : \mu = \mu_0\right\} = .95$$

$$n = 10, \text{ so} \quad P\left\{-a \leq t_{(9)} \leq a\right\} = 0.95.$$

Using the following R code, we obtain $a = 2.262157$.

R Code: `qt(.975,9)`

A 95% CI for μ is the set of all μ_0 such that

$$-2.262 \leq \frac{\overline{Y}-\mu_0}{\frac{s}{\sqrt{n}}} \leq 2.262.$$

Isolating μ_0, we obtain

$$\left[\bar{Y} - 2.262\frac{s}{\sqrt{n}}, \bar{Y} + 2.262\frac{s}{\sqrt{n}}\right].$$

Here, $\bar{y} = \$3496.9$, $n = 10$ and $s = \$1224.116$. A 95% Confidence interval for μ is $[\$2,621.22,\ \$4,372.58]$

A 95% confidence interval for the entire dataset using $n = 1151$, $\bar{y} = \$3445.382$, $s = \$920.52$, 97.5 percentile of $t_{(1150)}$ equal to 1.96 is:

$$[\$3,392.15,\ \$3,498.62],$$

a much narrower interval because n is larger.

Summary: 95% Confidence Interval for μ.

(a) when σ^2 known $\left[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]$

where $1.96 = 97.5^{th}$ percentile of $N(0,1)$

(b) when σ^2 unknown

$$\left[\bar{Y} - t^{.975}_{(n-1)}\frac{s}{\sqrt{n}},\ \bar{Y} + t^{.975}_{(n-1)}\frac{s}{\sqrt{n}}\right]$$

$t^{.975}_{(n-1)} = 97.5^{th}$ percentile of $t_{(n-1)}$

6.2.2 Hypothesis tests for μ

Optional reading: Section 13.3

Example 6.2.2. Returning to Example 6.2.1, let Y_i = monthly salary for person $i = 1, ..., n$, and $Y_i \sim N(\mu, \sigma^2)$ independent. Test the hypothesis that the mean salary is $\$3,000$ per month.

Solution: The Likelihood ratio test for testing $H_0 : \mu = \$3,000$ is

$$D = n \ln\left[1 + \frac{1}{n-1}\,T^2\right], \qquad \text{where } T = \frac{\bar{Y} - 3000}{\frac{s}{\sqrt{n}}}$$

$$\begin{aligned} p - value &= P\left\{D \geq d_{\text{obs}} \mid H_0 : \mu = 3000\right\} \quad \text{since } D \text{ monotone function of } T^2 \\ &= P\left\{T^2 \geq t^2_{\text{obs}} \mid H_0 : \mu = 3000\right\} \\ &= P\left\{|T| \geq |t_{\text{obs}}| \ \mid H_0 : \mu = 3000\right\} \end{aligned}$$

But $T \sim t_{(n-1=9)}$ when $\mu = 3000$, therefore

$$\begin{aligned} p - value &= P\left\{\left|t_{(9)}\right| \geq |t_{\text{obs}}|\right\} \\ |t_{\text{obs}}| &= \frac{|\bar{y} - 3000|}{\frac{s}{\sqrt{n}}} = 1.2836 \\ p - value &= P\left\{\left|t_{(9)}\right| \geq 1.2836\right\} = 0.2313, \end{aligned}$$

using R code: `2*(1-pt(1.2836,9))`. There is no evidence against the $H_0 : \mu = \$3,000$.

R code and output for the test:

```
> t.test(y,mu=3000)

One Sample t-test

data:  y
t = 1.2836, df = 9, p-value = 0.2313
alternative hypothesis: true mean is not equal to 3000
95 percent confidence interval:
 2621.22 4372.58
sample estimates:
mean of x
   3496.9
```

For the general normal linear model, we will use a t statistic for inference about mean parameters when σ^2 is unknown.

6.2.3 Inferences for σ^2

Example 6.2.3. Monthly salaries for Co-op students in Work Term 1 were collected over one year. The file 'SalaryWT1.csv' contains the salaries. A histogram and summary statistics are given below. GR.UG stands for Graduate and Undergraduate.

```
> summary(ywt1)
            Term     WTNum     GR.UG      SalMonth
 2015 - Fall   :111   W-1:352   GR: 50   Min.   :1406
 2015 - Spring:151              UG:302   1st Qu.:2717
 2015 - Summer: 90                       Median :3003
                                         Mean   :3149
                                         3rd Qu.:3526
                                         Max.   :7259
```

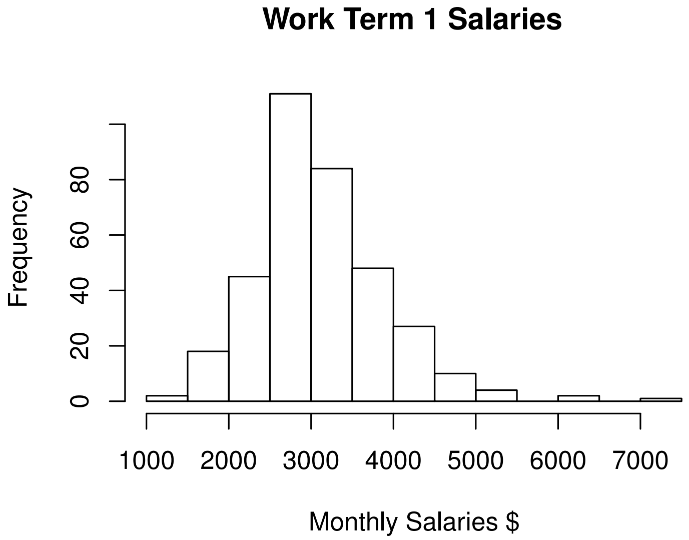


Figure 6.3: Histogram of Work Term 1 Salaries

Hypothesis Tests for σ^2

Optional reading: Section 13.3

Let $Y_i,\ i = 1, ..., n$ be the monthly salary for the i'th co-op student. We assume that $Y_i \sim N(\mu, \sigma^2)$, independent. For the Work Term 1 monthly salaries, we test the hypothesis that $H_0 : \sigma^2 = \sigma_0^2 = 750^2$ using a Likelihood Ratio test.

$$D = 2\,[\underset{\substack{\uparrow \\ \text{joint MLE}}}{\ell\,(\hat{\mu}, \hat{\sigma})} - \underset{\substack{\uparrow \\ \text{max under } H_0}}{\ell\,(\tilde{\mu}, \sigma_0^2)}]$$

Step 1: BASIC model

$$\ell\left(\mu, \sigma^2\right) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

In our example,

$$k = 2, \quad \hat{\mu} = \bar{y} = \$3149 \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum (y_i - \bar{y})^2}{n} = \frac{n-1}{n}s^2 = 778.9781^2.$$

Step 2: Hypothesized Model, $H_0 : \sigma^2 = 750^2$

We need to compute $\tilde{\mu}$, assuming that $\sigma^2 = 750^2$. Here $q = 1$.

As an exercise, show that maximizing $\ell\left(\mu, \sigma^2 = 750^2\right)$ over μ leads to $\tilde{\mu} = \bar{y}$.

Step 3: Test the hypothesis:

Substituting into the expression for the Likelihood ratio statistic,

$$\begin{aligned}
D &= 2\left[\ell\left(\hat{\mu}, \hat{\sigma}\right) - \ell\left(\tilde{\mu}, \sigma^2 = \sigma_0^2 = 750^2\right)\right].\\
D &= 2\left[-\frac{n}{2}\ln\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 + \frac{n}{2}\ln\left(\sigma_0^2\right) + \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right]\\
&= n\ln\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right) - n + n\frac{\hat{\sigma}^2}{\sigma_0^2} \qquad (6.1)
\end{aligned}$$

$$\hat{\sigma}^2 = 778.9781^2, \quad \sigma_0^2 = 750^2 \quad \text{and} \quad d_{\text{obs}} = 1.038.$$

Under $H_0 : \sigma^2 = 750^2$, $D \approx \chi^2_{(k-q)}$

$k = 2, \quad q = 1$ and $\quad D \approx \chi^2_{(1)}$

$$\begin{aligned}
\text{p-value} &= P\left(D \geq d_{\text{obs}} \mid H_0 : \sigma^2 = 750^2\right)\\
&\simeq P\left(\chi^2_{(1)} \geq 1.038\right) = 0.3083.
\end{aligned}$$

We have no evidence against $H_0 : \sigma^2 = 750^2$.

Confidence intervals for σ^2

Here we consider another way to construct Confidence Intervals, using **Pivotal Quantities**.

Recall the definition of a confidence interval.

Definition: *A* $100p\%$ *confidence interval* $[A, B]$ *for the unknown parameter* θ_T, *is an interval such that in a large number of repetitions of the experiment*, $[A, B]$ *covers* θ_T $100p$ *times out of* 100. $[0 \leq p \leq 1]$ i.e.

$$P\left\{A \leq \theta_T \leq B\right\} = p$$

We can use a pivotal quantity to construct a CI.

Definition: A **Pivotal Quantity,** Q**,** is a function of the data, and a monotone function of the unknown parameter, such that the distribution of Q does not depend upon θ_T.

Below are some examples of pivotal quantities.

(a) $Z = \frac{\bar{Y}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N\left(0,1\right)$ is a pivotal quantity for μ when σ^2 known

(b) $T = \frac{\bar{Y}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{(n-1)}$ is a pivotal quantity for μ when σ^2 unknown

(c) $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ is a pivotal quantity for σ^2

Here we construct a 95% confidence interval for σ^2 using a pivotal quantity.

Using χ^2 tables or R we find a, b such that

$$P\left\{a \leq \frac{(n-1)s^2}{\sigma^2} \leq b\right\} = .95$$

The convention is to choose a and b so that two tails have equal area, see Figure 6.4 below.

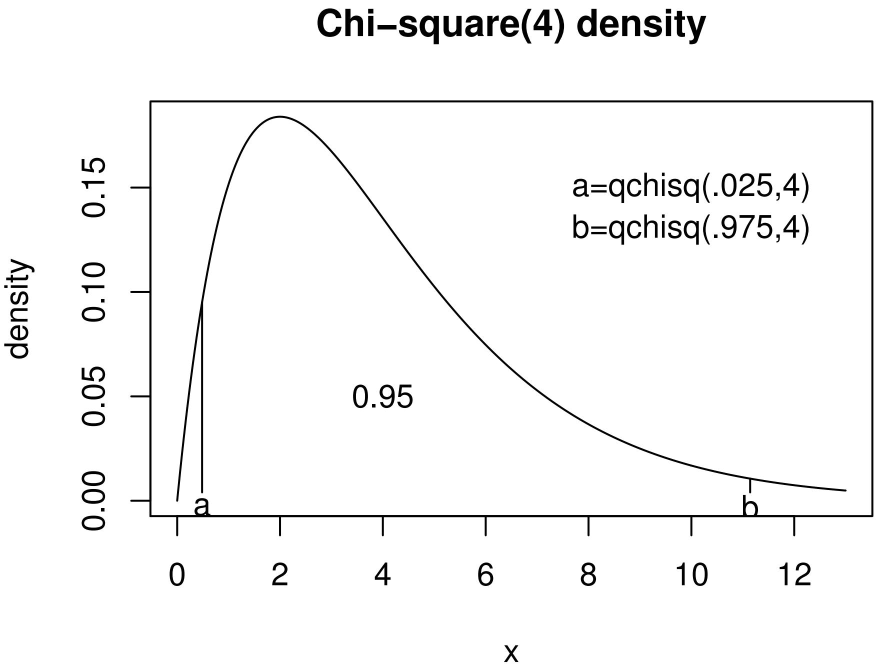


Figure 6.4: Chi-square on 4 degrees of freedom density

Then by a series of monotone transformations, we isolate σ^2.

$$P\left\{a \leq \frac{(n-1)\,s^2}{\sigma^2} \leq b\right\} = .95$$
$$P\left\{\frac{(n-1)\,s^2}{a} \geq \sigma^2 \geq \frac{(n-1)\,s^2}{b}\right\} = .95$$

The interval $\left[\frac{(n-1)s^2}{b}, \frac{(n-1)s^2}{a}\right]$ satisfies the definition of a 95% Confidence Interval for σ^2.

For the Work Term 1 data,

$$\begin{aligned} s &= 780.087, & n &= 352 \\ a &= 300.9897, & b &= 404.7974, \end{aligned}$$

and a 95% CI for σ^2 is $[726^2,\ 842^2]$. There is a great deal of variability in the data.

R code, Inferences for σ^2

```
LRS.sig<-function(y,sigma02){#LRS test for H_0: sigma^2 = sigma0^2
  n<-length(y)
  sigma2hat<-var(y)*(n-1)/n
  LRS<-n*(log(sigma02/sigma2hat) + (sigma2hat/sigma02) - 1)
  return(LRS)
}
n<-dim(ywt1)[1]
sd(ywt1$SalMonth)   #sd for Work Term 1 Co-op salary data
sqrt(var(ywt1$SalMonth)*(n-1)/n) #MLE of sigma

D<- LRS.sig(ywt1$SalMonth, 750^2)  #Likelihood Ratio test
D
1-pchisq(D,1)


#99% Confidence Interval for the Variance
sqrt((n-1)*var(ywt1$SalMonth)/qchisq(c(.995,.005),n-1))
qchisq(c(.005,.995),n-1)

#95% Confidence Interval for the Variance
sqrt((n-1)*var(ywt1$SalMonth)/qchisq(c(.975,.025),n-1))
qchisq(c(.025,.975),n-1)
```

6.3 The Two Sample Model

Optional reading: Section 13.4

Example 6.3.1. Monthly salaries for Co-op students in Work Term 1 and 2 were collected over one year. The file 'SalaryWT12.csv' contains the salaries. ~~A histogram~~ Boxplots and summary statistics by work term number are given below. Recall that the box on a boxplot encases the middle 50 percent of the data, i.e. from the 25'th to 75'th percentile. The solid line in the middle of the box indicates the median and outliers are plotted as small circles in both tails. The boxplots of the salaries for the two work terms suggest that these salaries have distributions which are similar.

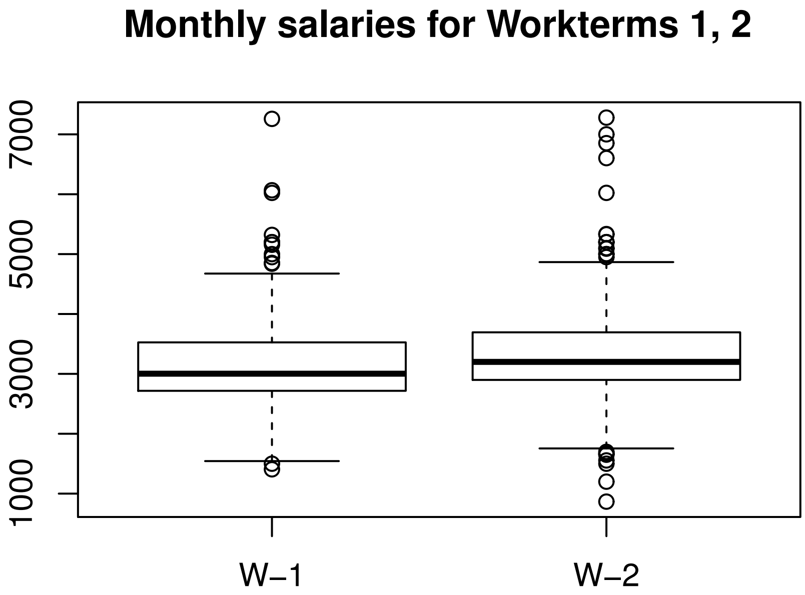


Figure 6.5: Boxplots of Work Term 1 and 2 Salaries

```
> by(ywt12$SalMonth,ywt12$WTNum,summary)  #summary statistics
ywt12$WTNum: W-1
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1406    2717    3003    3149    3526    7259
---------------------------------------------------------------------------
ywt12$WTNum: W-2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  866.6  2899.0  3200.0  3354.0  3691.0  7279.0

> by(ywt12$SalMonth,ywt12$WTNum,sd) #standard deviations
ywt12$WTNum: W-1
[1] 780.087
---------------------------------------------------------------------------
```

```
ywt12$WTNum: W-2
[1] 847.1914
```

6.3.1 Inferences for the differences between two means

Is this data consistent with the hypothesis that salaries for work term one and two are the same?

Two sample model, Variances assumed EQUAL and KNOWN

It is very unusual in practice to assume that the variances of the two groups are known. This case provides a 'baby' step en route to the case where variances are not assumed to be known.

Let

$$\begin{aligned} Y_{1i} &= \text{monthly salary for work term } 1, i = 1, ..., n_1 \\ Y_{2j} &= \text{monthly salary for work term } 2, j = 1, ..., n_2 \end{aligned}$$

We assume that,

$$\begin{aligned} Y_{1i} &\sim N(\mu_1, \sigma^2), \quad \text{independent}, \\ Y_{2j} &\sim N(\mu_2, \sigma^2), \quad \text{independent}, \\ &\text{and that } \sigma^2 = 818^2. \end{aligned}$$

We test the hypothesis, $H_0 : \mu_1 = \mu_2$, or equivalently, $H_0 : \mu_1 - \mu_2 = 0$.

Step 1: BASIC model

$$L(\mu_1, \mu_2) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_{1i} - \mu_1)^2\right] \times \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_2}(y_{2j} - \mu_2)^2\right]$$

As an exercise, show that $\hat{\mu}_1 = \bar{y}_1 = \sum\limits_{i=1}^{n_1} y_{1i}/n_1 = \$3,149$ and $\hat{\mu}_2 = \bar{y}_2 = \sum\limits_{j=1}^{n_2} y_{2j}/n_2 = \$3,354$.

Step 2: Hypothesized Model

Assuming $H_0 : \mu_1 = \mu_2 = \mu$, unknown, we need to estimate μ.

$$L\left(\mu_1 = \mu, \mu_2 = \mu\right) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_{1i} - \mu\right)^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}\left(y_{2j} - \mu\right)^2\right]$$

As an exercise, show that the MLE of μ is

$$\tilde{\mu} = \frac{\sum_{i=1}^{n_1} y_{1i} + \sum_{j=1}^{n_2} y_{2j}}{n_1 + n_2} = \bar{y}.$$

Step 3: Test the hypothesis

$$\begin{aligned} D &= 2\left[\ell\left(\hat{\mu}_1, \hat{\mu}_2\right) - \ell\left(\mu_1 = \tilde{\mu}, \mu_2 = \tilde{\mu}\right)\right] \\ &= 2\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_{1i} - \bar{y}_1\right)^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}_2\right)^2\right. \\ &\quad \left. + \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_{1i} - \bar{y}\right)^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}\left(y_{2j} - \bar{y}\right)^2\right] \end{aligned}$$

This can be simplified using,

$$\begin{aligned} \sum_{i=1}^{n_1}\left(y_{1i} - \bar{y}\right)^2 &= \sum_{i=1}^{n_1}\left[\left(y_{1i} - \bar{y}_1\right) + \left(\bar{y}_1 - \bar{y}\right)\right]^2 \\ &= \sum_{i=1}^{n_1}\left(y_{i1} - \bar{y}_1\right)^2 + n_1\left(\bar{y}_1 - \bar{y}\right)^2, \end{aligned}$$

since

$$\sum_{i=1}^{n_1}\left(y_{i1} - \bar{y}_1\right)\left(\bar{y}_1 - \bar{y}\right) = 0.$$

Simplifying for the work term 2 data in the same way yields,

$$D = \frac{1}{\sigma^2}\left[n_1\left(\bar{y}_1 - \bar{y}\right)^2 + n_2\left(\bar{y}_2 - \bar{y}\right)^2\right].$$

Noting that $\bar{y}$ is a function of both $\bar{y}_1$ and $\bar{y}_2$,

$$\begin{aligned} \bar{y} &= \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} \\ \text{and so } \bar{y}_1 - \bar{y} &= \frac{n_2\bar{y}_1 - n_2\bar{y}_2}{n_1 + n_2}. \end{aligned}$$

$$\text{Similarly, } \bar{y}_2 - \bar{y} = \frac{n_1\bar{y}_2 - n_1\bar{y}_1}{n_1 + n_2}.$$

Substituting into the expression for D yields,

$$\begin{aligned} D &= \frac{1}{\sigma^2}\left[\frac{n_1 n_2^2}{\left(n_1+n_2\right)^2}\left(\bar{y}_1 - \bar{y}_2\right)^2 + \frac{n_2 n_1^2}{\left(n_1+n_2\right)^2}\left(\bar{y}_1 - \bar{y}_2\right)^2\right] \\ &= \frac{1}{\sigma^2}\left(\bar{y}_1 - \bar{y}_2\right)^2 \frac{n_1 n_2}{n_1 + n_2} \\ &= \frac{1}{\sigma^2}\frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \end{aligned}$$

The $p - value$ for the test is,

$$p - value = P\left(D \geq d_{\text{obs}} \mid H_0 : \mu_1 = \mu_2\right).$$

We know that under $H_0 : \mu_1 = \mu_2$, $D \approx \chi^2_{(k-q=1)}$.

We can obtain the exact distribution of D under H_0 here. Since $Var(\bar{Y}_1) = \sigma^2/n_1$, $Var(\bar{Y}_2) = \sigma^2/n_2$ and $\bar{Y}_1$ and $\bar{Y}_2$ are independent, $Var(\bar{Y}_1 - \bar{Y}_2) = \sigma^2(1/n_1 + 1/n_2)$. Therefore,

$$\frac{\left(\bar{Y}_1 - \bar{Y}_2\right)^2}{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sim \chi^2_{(1)} \text{ exactly, and}$$

$$\frac{\left(\bar{Y}_1 - \bar{Y}_2\right)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = Z \sim N(0,1)\text{exactly.}$$

The $p-value$ is

$$\begin{aligned} p-value &= P\left[\chi^2_{(1)} \geq \frac{\left(\bar{y}_1 - \bar{y}_2\right)^2}{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right] \\ &= P\left[|Z| \geq \frac{|\bar{y}_1 - \bar{y}_2|}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right] \end{aligned}$$

$$\text{where } Z \sim N\left(0,1\right).$$

Returning to our example:

$$\begin{array}{ll} \bar{y}_1 = 3148.532, & \bar{y}_2 = 3354.483, \quad \sigma^2 = 818^2 \\ n_1 = 352, & n_2 = 308 \end{array}$$

and the

$$p-value = P\left[|Z| \geq \sqrt{10.4129}\right] = .00125.$$

There is strong evidence against the hypothesis that salaries for work terms 1 and 2 are the same. The increase in mean monthly salary is significantly different from zero.

A 95% confidence interval for the difference in the means is:

$$\begin{aligned} &\bar{y}_1 - \bar{y}_2 \pm z_{.975}\ \sigma\ \sqrt{1/n_1 + 1/n_2} \\ &= -205.95 \pm 125.09 \\ &= (-331.04, -80.86) \end{aligned}$$

We report, “The estimated mean monthly increase in salary in work term 2 over work term 1 is $205.95 (95% CI $80.86 - $331.04).”

Two sample model, Variances assumed EQUAL and UNKNOWN

In this section, we assume that the variances for the two groups are equal but unknown.

Let

$$\begin{aligned} Y_{1i} &= \text{monthly salary for work term } 1, i = 1, ..., n_1 \\ Y_{2j} &= \text{monthly salary for work term } 2, j = 1, ..., n_2 \end{aligned}$$

We assume that,

$$\begin{gathered} Y_{1i} \sim N(\mu_1, \sigma^2), \quad \text{independent,} \\ Y_{2j} \sim N(\mu_2, \sigma^2), \quad \text{independent,} \\ \text{and that } \sigma^2 \text{ is unknown.} \end{gathered}$$

We test the hypothesis, $H_0 : \mu_1 = \mu_2$, or equivalently, $H_0 : \mu_1 - \mu_2 = 0$.

Step 1: BASIC model

$$L\left(\mu_1, \mu_2, \sigma^2\right) = \left(\frac{1}{\sigma^2}\right)^{\frac{n_1}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_{1i} - \mu_1\right)^2\right] \left(\frac{1}{\sigma^2}\right)^{\frac{n_2}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_2}\left(y_{2j} - \mu_2\right)^2\right]$$

As an exercise, show that $\hat{\mu}_1 = \bar{y}_1 = \sum\limits_{i=1}^{n_1} y_{1i}/n_1 = \$3,149$, $\hat{\mu}_2 = \bar{y}_2 = \sum\limits_{j=1}^{n_2} y_{2j}/n_2 = \$3,354$, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2}.$$

Step 2: Hypothesized Model

Assuming $H_0 : \mu_1 = \mu_2 = \mu$, unknown, we need to estimate μ and the common σ^2.

$$L\left(\mu_1 = \mu, \mu_2 = \mu, \sigma^2\right) = \left(\frac{1}{\sigma^2}\right)^{\frac{n_1+n_2}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}\left(y_{1i} - \mu\right)^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}\left(y_{2j} - \mu\right)^2\right]$$

As an exercise, show that the MLE of μ is

$$\tilde{\mu} = \frac{\sum_{i=1}^{n_1} y_{1i} + \sum_{j=1}^{n_2} y_{2j}}{n_1 + n_2} = \bar{y},$$

and that

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n_1}(y_{1i} - \bar{y})^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y})^2}{n_1 + n_2}.$$

Step 3: Test the hypothesis

$$D = 2\left[\ell\left(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2\right) - \ell\left(\mu_1 = \tilde{\mu}, \mu_2 = \tilde{\mu}, \tilde{\sigma}^2\right)\right]$$

is a monotone function of T^2, where

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{pooled}\ \sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}.$$

and

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

$$D = (n_1 + n_2) \ln\left[1 + \frac{1}{n_1 + n_2 - 2} T^2\right]$$

[See the Appendix for a proof of this result.]

Returning to our example, $t_{obs} = -3.2504$, and the p-value is,

$$P(|t_{(n_1+n_2-2)}| \geq |t_{obs}|) = 0.001211.$$

A 95% confidence interval for the differences of the means is:

$$\bar{y}_1 - \bar{y}_2 \pm t^{.975}_{(n_1+n_2-2)}\ s_{pooled}\ \sqrt{1/n_1 + 1/n_2}$$
$$= (-330.37, -80.54).$$

The R output for this problem is below and code is in the following section:

```
> t.test(SalMonth~WTNum, data=ywt12,var.equal=TRUE)
Two Sample t-test

data:  SalMonth by WTNum
t = -3.2504, df = 658, p-value = 0.001211
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -330.36720  -81.53596
sample estimates:
mean in group W-1 mean in group W-2
         3148.532          3354.483
```

Two sample model, Variances assumed UNEQUAL and UNKNOWN

For the two sample model with unequal and unknown variances, we will use the Satterthwaite approximation to the degrees of freedom of the t-test that you learned in your first course in Statistics.

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \; t_{(\nu)},$$

where ν is the Satterthwaite approximation to the degrees of freedom.

The R output using this approximation appears below and code is in the following subsection.

```
> t.test(SalMonth~WTNum, data=ywt12,var.equal=FALSE)

Welch Two Sample t-test

data:  SalMonth by WTNum
t = -3.2326, df = 628.79, p-value = 0.001291
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -331.06376  -80.83941
sample estimates:
mean in group W-1 mean in group W-2
         3148.532          3354.483
```

The p-value for the test of the equality of the two means is 0.001291, which is very similar to the results of the test which assumes that the variances are equal.

R code for Inferences for Two sample model

```
###Inferences about the differences in the means, $\mu_1 - \mu_2$###

###Method 1:  Assume variances equal and known, Test and Confidence Interval
for Difference###

ywt12.means<-by(ywt12$SalMonth,ywt12$WTNum,mean)
ywt12.num<-by(ywt12$SalMonth,ywt12$WTNum,length)
ywt12.means
ywt12.num
zobs<-(ywt12.means[1]-ywt12.means[2])/818/sqrt(sum(1/ywt12.num))
zobs
pvalue<-2*pnorm(-abs(zobs))
pvalue

zobs^2
1-pchisq(zobs^2,1)

#95% confidence interval for the difference
(ywt12.means[1]-ywt12.means[2]) + qnorm(.975)*c(-1,1)*818*sqrt(sum(1/ywt12.num))
#difference in means
(ywt12.means[1]-ywt12.means[2])
#margin of error
qnorm(.975)*c(-1,1)*818*sqrt(sum(1/ywt12.num))

###Method 2:  Assume variances equal and unknown###
#this uses pooled estimate of variance for test

t.test(SalMonth~WTNum, data=ywt12,var.equal=TRUE)

###Method 3:  Assume variance are not equal###

t.test(SalMonth~WTNum, data=ywt12,var.equal=FALSE)
```

6.3.2 Testing Equality of Variances

On the last assignment, you will test the hypothesis of equal variances in the two sample model using a Likelihood Ratio test.

6.3.3 Appendix: Derive the 2-sample σ unknown but equal t-test

Derive the two-sample t-test.

$$\begin{array}{ll} Y_{1i} \sim N\left(\mu_1, \sigma^2\right) & i = 1, \ldots, n_1 \\ Y_{2j} \sim N\left(\mu_2, \sigma^2\right) & j = 1, \ldots, n_2 \end{array}$$

The likelihood ratio statistic for testing the equality of the two means is:

$$D = 2\left[\ell\left(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2\right) - \ell\left(\mu_1 = \tilde{\mu}, \mu_2 = \tilde{\mu}, \tilde{\sigma}^2\right)\right].$$

The Log-likelihood under the Basic model is:

$$\ell(\mu_1, \mu_2) = -\frac{n_1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum\left(y_{1i} - \mu_1\right)^2 - \frac{n_2}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum\left(y_{2j} - \mu_2\right)^2.$$

The MLE's for μ_1, μ_2 and σ^2 under the Basic model are solutions of the following equations:

$$\frac{\partial\ell(\mu_1, \mu_2)}{\partial\mu_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(y_{1i} - \mu_1) = 0 \tag{6.2}$$

$$\frac{\partial\ell(\mu_1, \mu_2)}{\partial\mu_2} = \frac{1}{\sigma^2}\sum_{j=1}^{n_2}(y_{2j} - \mu_2) = 0 \tag{6.3}$$

$$\frac{\partial\ell(\mu_1, \mu_2)}{\partial\sigma} = -\frac{(n_1 + n_2)}{\sigma} + \frac{1}{\sigma^3}\left[\sum_{i=1}^{n_1}\left(y_{1i} - \mu_1\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j} - \mu_2\right)^2\right] = 0 \tag{6.4}$$

The MLE's under the Basic model are:

$$\begin{aligned} \hat{\mu}_1 &= \sum_{i=1}^{n_1} y_{1i}/n_1 \\ \hat{\mu}_2 &= \sum_{j=1}^{n_2} y_{2j}/n_2 \\ \hat{\sigma}^2 &= \frac{\sum(y_{1i} - \hat{\mu}_1)^2 + \sum(y_{2j} - \hat{\mu}_2)^2}{n_1 + n_2} \end{aligned}$$

The Log-likelihood under the Reduced (hypothesized) model is:

$$\ell_H\left(\mu_1 = \mu, \mu_2 = \mu, \sigma^2\right) = -\frac{(n_1 + n_2)}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n_1}\left(y_{1i} - \mu\right)^2 + \sum_{j=1}^{n_2}\left(y_{2j} - \mu\right)^2\right]$$

Taking derivatives and solving yields the MLE's under the Reduced model are:

$$\begin{aligned}\tilde{\mu} &= \frac{\sum_i y_{1i} + \sum_j y_{2j}}{n_1 + n_2} \qquad \text{and} \\ \tilde{\sigma}^2 &= \frac{\sum_i \left(y_{1i} - \tilde{\mu}\right)^2 + \sum_j \left(y_{2j} - \tilde{\mu}\right)^2}{n_1 + n_2}.\end{aligned}$$

Substituting the above into the Likelihood Ratio Statistic,

$$\begin{aligned}D &= -\left(n_1 + n_2\right) \ln \hat{\sigma}^2 + \left(n_1 + n_2\right) \ln \tilde{\sigma}^2 \\ &= \left(n_1 + n_2\right) \ln \frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\end{aligned}$$

$$(n_1 + n_2)\tilde{\sigma}^2 = \sum_i \left(y_{1i} - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2 + \sum_j \left(y_{2j} - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2$$

The first term can be written as follows:

$$\begin{aligned}\sum_i \left(y_{1i} - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2 &= \sum_i \left(y_{1i} - \bar{y}_1 + \bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2 \\ &= \sum_i \left(y_{1i} - \bar{y}_1\right)^2 + n_1 \left(\bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2,\end{aligned}$$

and the similarly the second term,

$$\sum_j \left(y_{2i} - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2 = \sum_j \left(y_{2i} - \bar{y}_2\right)^2 + n_2 \left(\bar{y}_2 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}\right)^2.$$

The contents of the second terms can be simplified as,

$$\begin{aligned}\bar{y}_1 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} &= \frac{n_2(\bar{y}_1 - \bar{y}_2)}{n_1 + n_2} \\ \bar{y}_2 - \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} &= \frac{n_1(\bar{y}_2 - \bar{y}_1)}{n_1 + n_2}.\end{aligned}$$

Substituting into part of the expression for D,

$$
\begin{aligned}
\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} &= 1 + \frac{n_1\left[\frac{n_2(\bar{y}_1-\bar{y}_2)}{n_1+n_2}\right]^2 + n_2\left[\frac{n_1(\bar{y}_2-\bar{y}_1)}{n_1+n_2}\right]^2}{\sum_i\left(y_{1i}-\bar{y}_1\right)^2 + \sum_j\left(y_{2j}-\bar{y}_2\right)^2} \\
&= 1 + \frac{\frac{n_1 n_2}{n_1+n_2}\left(\bar{y}_1-\bar{y}_2\right)^2}{\sum_i\left(y_{1i}-\bar{y}_1\right)^2 + \sum_j\left(y_{2j}-\bar{y}_2\right)^2} \\
&= 1 + \frac{\left(\bar{y}_1-\bar{y}_2\right)^2}{\boxed{\left[\frac{1}{n_1+n_2-2}\right]}\left[\frac{\sum_i(y_{1i}-\bar{y}_1)^2+\sum_j(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]} \\
&= 1 + \frac{1}{n_1+n_2-2}T^2 \quad \text{where } T = \frac{\bar{y}_1-\bar{y}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.
\end{aligned}
$$

The Likelihood ratio statistic is a monotone function of T^2,

$$D = (n_1+n_2)\ln\left[1+\frac{1}{n_1+n_2-2}T^2\right].$$

6.4 The Straight Line Model

Optional reading: Section 13.5, 13.6

In this section we analyze data that is in the form of ordered pairs

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

y_i - called the **response** or **dependent** variable

x_i - called the **explanatory** or **predictor** variable

We wish to determine the form and strength of the relationship between the response variable (y) and the explanatory variable (x). Typically there are two possible goals for the analysis:

(1) **Explanation:** What is the relationship between y and x.

(2) **Prediction:** Given x, can we predict y accurately.

Example 6.4.1. We are interested in the relationship between monthly salaries for co-op students as a function of the work term number.

x_i = work term number for student i,

y_i = monthly salary for student i.

The first step is to graph the data and determine an appropriate model to fit to the data. The data are graphed below:

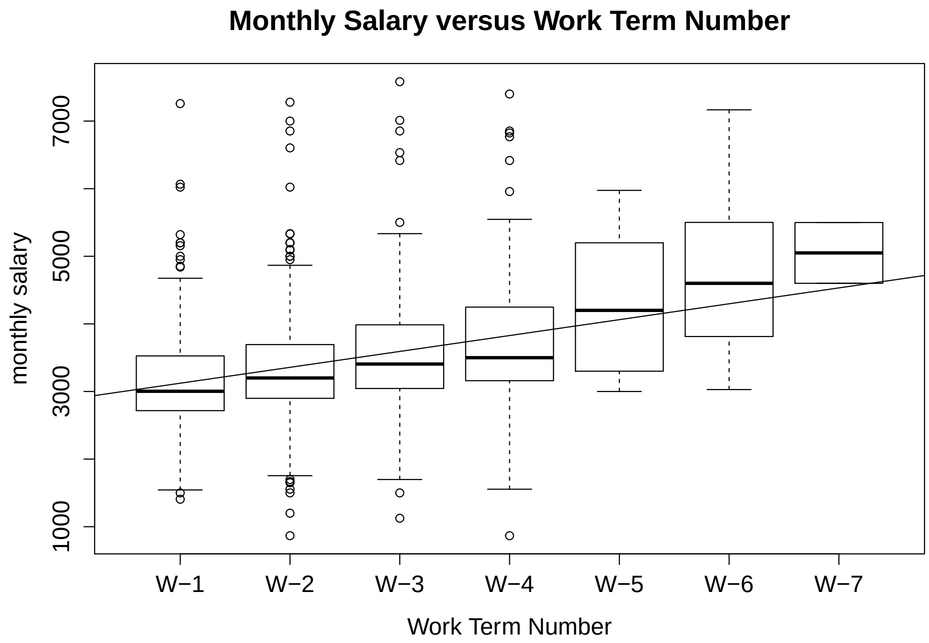


Figure 6.6: Boxplots of monthly salaries by work term number

From the graph, we note that,

(1) for a given number of work terms, the monthly salaries are subject to a large amount of variability:

(2) A linear relationship between monthly salary (Y) and work term number (X) seems appropriate.

In this course, we will primarily consider models where the $y's$ are linearly related to the $x's$.

We assume that $Y_i \sim N\left(\mu_i, \sigma^2\right)$ independent, where $\mu_i = E\left(Y_i\right) = \alpha + \beta x_i$. The mean monthly salary is linearly related to the work term number.

Another way to write this model is as:

$$Y_i = (\alpha + \beta x_i) + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$, independent.

6.4.1 Linear model parameter estimation

To estimate α, β and σ^2, we use Maximum likelihood estimation.

$$\begin{aligned}
L\left(\alpha, \beta, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - \mu_i\right)^2\right] \\
&= \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu_i\right)^2\right] \\
\ell\left(\alpha, \beta, \sigma^2\right) &= -n \ln \sigma \underbrace{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2}_{\hat{\alpha}, \hat{\beta} \text{ will maximize this}} \\
&\implies \hat{\alpha}, \hat{\beta} \text{ will minimize } \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i\right)^2 \\
&\hat{\alpha}, \hat{\beta} \text{ are often called LEAST SQUARES ESTIMATES}
\end{aligned}$$

We need to solve the system of equations,

$$\left.\frac{\partial \ell}{\partial \alpha}\right|_{\hat{\alpha}, \hat{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \tag{6.5}$$

$$\left.\frac{\partial \ell}{\partial \beta}\right|_{\hat{\alpha}, \hat{\beta}} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0 \tag{6.6}$$

Letting $\hat{\epsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$, the equations (6.5) and (6.6) can be written as:

$$\begin{aligned}
(6.5) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \hat{\epsilon}_i = 0 \\
(6.6) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \hat{\epsilon}_i x_i = 0
\end{aligned}$$

Solving (6.5) leads to

$$n\bar{y} - n\hat{\alpha} - \hat{\beta} n\bar{x} = 0 \implies \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Substituting for $\hat{\alpha}$ and solving (6.6) yields,

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i)x_i &= 0 \\ \Longrightarrow \sum(y_i - \bar{y})x_i - \hat{\beta}\sum(x_i - \bar{x})x_i &= 0 \\ \hat{\beta} = \frac{\sum(y_i - \bar{y})x_i}{\sum(x_i - \bar{x})x_i} &= \frac{S_{XY}}{S_{XX}}. \end{aligned}$$

Using algebraic manipulations, we derive some alternate formulae for S_{XY} and S_{XX} which will be useful later.

(i)

$$\begin{aligned} S_{XX} &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right)x_i = \sum_{i=1}^{n}x_i^2 - \bar{x}\sum_{i=1}^{n}x_i \\ S_{XX} &= \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \end{aligned}$$

(ii)

$$\begin{aligned} S_{XX} &= \sum_{i=1}^{n}\left(x_i - \bar{x}\right)x_i - \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\bar{x} \quad \text{since 2nd term=0} \\ S_{XX} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

(iii)

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{n}\left(y_i - \bar{y}\right)x_i = \sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i \\ S_{XY} &= \sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y} \end{aligned}$$

(iv)

$$\begin{aligned} S_{XY} &= \sum_{i=1}^{n} (y_i - \bar{y})\, x_i - \sum_{i=1}^{n} (y_i - \bar{y})\, \bar{x} \quad \text{since 2nd term=0} \\ S_{XY} &= \sum_{i=1}^{n} (y_i - \bar{y})\,(x_i - \bar{x}) \end{aligned}$$

(v)

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})\, y_i \quad \text{since} \sum_{i=1}^{n} (x_i - \bar{x})\, \bar{y} = 0$$

To estimate σ^2 we compute the derivative of the log likelihood function with respect to σ.

$$\begin{aligned} \frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \\ \left.\frac{\partial \ell}{\partial \sigma}\right|_{\hat{\alpha},\hat{\beta},\hat{\sigma}} &= 0 \implies n\hat{\sigma}^2 = \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\epsilon}^2 \end{aligned}$$

We can show that $E\left(\hat{\sigma}^2\right) \neq \sigma^2$ and it is therefore a biased estimate.

To estimate σ^2, we will use:

$$\begin{aligned} s^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2 \ \text{ where } \ \hat{\epsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \end{aligned}$$

Note that $\hat{\epsilon}_i$ is an estimate of $\epsilon_i = Y_i - (\alpha + \beta x_i)$ where we assumed that $\epsilon_i \sim N\left(0, \sigma^2\right)$ and independent. We call $\hat{\epsilon}_i$ a **Residual**.

Returning to Example 6.4.1, the fitted model from R is given below.

```
> Sal.lm<-lm(SalMonth~WTNumN, data=salarynz)
> summary(Sal.lm)

Call:
lm(formula = SalMonth ~ WTNumN, data = salarynz)

Residuals:
    Min      1Q  Median      3Q     Max
-2960.8  -522.6  -157.4   406.4  4136.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2887.40      56.26   51.33   <2e-16 ***
WTNumN        234.99      21.06   11.16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 874.7 on 1149 degrees of freedom
Multiple R-squared:  0.09779,Adjusted R-squared:  0.097
F-statistic: 124.5 on 1 and 1149 DF,  p-value: < 2.2e-16
```

Figure 6.7: R Output: Linear regression for salary data

- The estimated relationship between monthly salary and work term number is:

$$\text{Salary} = 2887.40 + 234.99 \times \text{Work Term number}.$$

- The estimate of σ is $s =$"Residual standard error" $= 874.7$ on 1149 degrees of freedom.
- We estimate that monthly salary increases by \$234.99 for each additional work term.
- The intercept estimate is the estimated monthly salary for zero work terms, but this is not meaningful here. Instead, we could quote the estimated monthly salary for work term 1, \$2887.40 + \$234.99.

6.4.2 Linear model Distribution theory

Distribution of $\hat{\beta}$

Recall: If $Y_1, \ldots, Y_n$ are independent with $Y_i \sim N(\mu_i, \sigma^2)$ then $\sum a_i Y_i \sim N\left(\sum a_i\mu_i, \sigma^2 \sum a_i^2\right)$, for constants $a_1, ..., a_n$.

We want to express $\hat{\beta} = \sum_{i=1}^{n} a_i Y_i$, as a linear combination of the Y_i's.

$$\begin{aligned}\hat{\beta} &= \frac{S_{XY}}{S_{XX}} \\ &= \frac{\sum (x_i - \bar{x})\, y_i}{S_{XX}} = \sum \frac{(x_i - \bar{x})\, y_i}{S_{XX}}\end{aligned}$$

$$\begin{aligned}\text{Let } a_i &= \frac{(x_i - \bar{x})}{S_{XX}} \\ \text{then } \hat{\beta} &\sim N\left(\sum a_i\mu_i,\ \sigma^2 \sum a_i^2\right) \\ E(\hat{\beta}) &= \sum a_i\mu_i = \sum \frac{(x_i - \bar{x})}{S_{XX}}\,(\alpha + \beta x_i) = \beta \\ VAR(\hat{\beta}) &= \sigma^2 \sum a_i^2 = \sigma^2 \ \sum \frac{(x_i - \bar{x})^2}{S_{XX}^2} = \frac{\sigma^2}{S_{XX}} \\ \hat{\beta} &\sim N\left(\beta, \frac{\sigma^2}{S_{XX}}\right)\end{aligned}$$

We can use the quantity

$$\frac{\hat{\beta} - \beta}{\sqrt{\frac{s^2}{S_{XX}}}} \sim t_{(n-2)} \tag{6.7}$$

for tests and confidence intervals for β. (It is a pivotal quantity for β as it is a function of the data and a function of the unknown parameter and its distribution is completely known.)

The quantity in the denominator of (6.7), $s/\sqrt{S_{XX}}$ is called the **Standard Error** of $\hat{\beta}$, s.e.($\hat{\beta}$), and it is listed on the output of Figure 6.7 under the column 'Std. Error'.

It is the square root of the estimated variance of $\hat{\beta}$. The standard error for $\hat{\beta}$ s.e.$(\hat{\beta})$, is 21.06.

Note that in (6.7), the degrees of freedom for the t distribution are the same as the denominator in the formula for s^2. This result holds generally. The estimate s is given in the output Figure 6.7 labelled as 'Residual standard error:' and the value here is 874.7 on 1149 degrees of freedom.

We next test the hypothesis that $\beta = 0$, and construct a 99% confidence interval for β.

$$p - value = P\left\{ \left|t_{(n-2)}\right| \geq \frac{\left|\hat{\beta} - 0\right|}{s\sqrt{\frac{1}{S_{XX}}}} \;\middle|\; H_0 : \beta = 0 \text{ true} \right\}$$

The R output in Figure 6.7, provides the observed value of the test statistic, (6.7), under the column 't value'. The observed value of our t-statistic for β is 11.16. The $p - value$ is given in the column 'Pr(> $|t|$)' and its value is listed as $< 2e - 16$. For very small or very large numbers, R uses exponential notation. $2e - 16$ means 2×10^{-16}.

$$p - value = P\left\{\left|t_{(1149)}\right| \geq 11.16\right\} < 2 \times 10^{-16}.$$

We have very strong evidence against $H_0 : \beta = 0$.

To compute a 99% CI for β, we use the general formulation,

$$\text{estimate } \pm \; t^{.995}_{(\nu)} \text{ s.e.(estimate)}.$$

Here we obtain the t-quantile from R as: `qt(.995, 1149)`, so our 99% confidence interval is:

$$234.99 \;\pm\; 2.58 \;\times 21.06 = \;[180.66, 289.32].$$

Distribution of $\hat{\alpha}$

We want to express $\hat{\alpha} = \displaystyle\sum_{i=1}^{n} a_i Y_i$, as a linear combination of the Y_i's.

$$\begin{aligned}
\hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} = \frac{1}{n}\sum y_i - \bar{x}\sum a_i y_i \\
&= \sum y_i \left[\underbrace{\frac{1}{n} - \bar{x}a_i}_{b_i}\right] \qquad a_i = \frac{x_i - \bar{x}}{S_{XX}} \\
&= \sum b_i y_i
\end{aligned}$$

Therefore, $\hat{\alpha} \sim N\left(\sum b_i\mu_i,\ \sigma^2 \sum b_i^2\right)$.

$$\begin{aligned}
E(\hat{\alpha}) &= \sum b_i\mu_i = \sum \left[\frac{1}{n} - \frac{\bar{x}\left(x_i - \bar{x}\right)}{S_{XX}}\right](\alpha + \beta x_i) \\
&= \alpha + \beta\bar{x} - \frac{\alpha\bar{x}}{S_{XX}}\sum\left(x_i - \bar{x}\right) - \frac{\beta\bar{x}}{S_{XX}}\sum\left(x_i - \bar{x}\right)x_i \\
&= \alpha \quad \text{since } S_{XX} = \sum\left(x_i - \bar{x}\right)x_i \quad \text{and} \quad \sum(x_i - \bar{x}) = 0
\end{aligned}$$

$$\begin{aligned}
VAR(\hat{\alpha}) &= \sum b_i^2\sigma^2 \\
\sum b_i^2 &= \sum\left[\frac{1}{n} - \frac{\bar{x}\left(x_i - \bar{x}\right)}{S_{XX}}\right]^2 \\
&= \sum\left[\frac{1}{n^2} - 2\frac{1}{n}\frac{\bar{x}\left(x_i - \bar{x}\right)}{S_{XX}} + \frac{\bar{x}^2\left(x_i - \bar{x}\right)^2}{S_{XX}^2}\right] \\
&= \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}^2}\sum\left(x_i - \bar{x}\right)^2 \\
&= \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}
\end{aligned}$$

$$\hat{\alpha} \sim N\left(\alpha, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]\right)$$

We can use the quantity

$$\frac{\hat{\alpha} - \alpha}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}}} \sim t_{(n-2)} \tag{6.8}$$

for tests and confidence intervals for α. (It is a pivotal quantity for α as it is a function of the data and a function of the unknown parameter and its distribution is completely known.)

The quantity in the denominator of (6.8), $s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}}$ is called the **Standard Error** of $\hat{\alpha}$, s.e.$(\hat{\alpha})$, and it is listed on the output of Figure 6.7 under the column 'Std. Error'. It is the square root of the estimated variance of $\hat{\alpha}$. The standard error for $\hat{\alpha}$ is 56.26.

Note that in (6.8), the degrees of freedom for the t distribution are the same as the denominator in the formula for s^2.

We next test the hypothesis that $\alpha = 0$, and construct a 99% confidence interval for α.

$$p-value = P\left\{ \left|t_{(n-2)}\right| \geq \frac{|\hat{\alpha}-0|}{s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{XX}}}} \,\middle|\, H_0 : \alpha = 0 \right\}$$

The R output in Figure 6.7, provides the observed value of the test statistic, (6.8), under the column 't value'. The observed value of our t-statistic for α is 51.33. The $p-value$ is given in the column 'Pr($>|t|$)' and its value is listed as $< 2e-16$. For very small or very large numbers, R uses exponential notation. $2e-16$ means 2×10^{-16}.

$$p-value = P\left\{\left|t_{(1149)}\right| \geq 11.16\right\} < 2\times10^{-16}.$$

We have very strong evidence against $H_0 : \alpha = 0$.

To compute a 99% CI for α, we use the general formulation,

$$\text{estimate} \ \pm \ t_{(\nu)}^{.995} \ \text{s.e.(estimate)}.$$

Here we obtain the t-quantile from R as: `qt(.995, 1149)`, so our 99% confidence interval is:

$$2887.40 \ \pm \ 2.58 \ \times 56.26 = \ [\$2742.25, \$3032.55].$$

Distribution of $\hat{\mu}_0 = \hat{\alpha} + \hat{\beta} x_0$

Given a particular value for x say x_0, the $E(Y) = \alpha + \beta x_0$ is estimated with $\hat{\alpha} + \hat{\beta} x_0 = \hat{\mu}_0$. We can obtain its distribution as follows.

$$\begin{aligned}
\hat{\mu}_0 &= \left[\bar{y} - \hat{\beta}\bar{x}\right] + \hat{\beta} x_0 \\
&= \bar{y} + \hat{\beta}\left(x_0 - \bar{x}\right) \\
&= \frac{1}{n}\sum y_i + \left(x_0 - \bar{x}\right) a_i y_i \qquad a_i = \frac{\left(x_i - \bar{x}\right)}{S_{XX}} \\
&= \sum \left[\underbrace{\frac{1}{n} + \left(x_0 - \bar{x}\right) a_i}_{c_i}\right] y_i \\
&\implies \hat{\mu}_0 \sim N\left(\sum c_i \mu_i, \sum c_i^2 \sigma^2\right)
\end{aligned}$$

We know that,

$$\begin{aligned}
E\left(\hat{\alpha}\right) = \alpha \qquad & E\left(\hat{\beta}\right) = \beta \\
& \implies E\left(\hat{\alpha} + \hat{\beta} x_0\right) = \alpha + \beta x_0 \\
\sum c_i^2 &= \sum \left[\frac{1}{n} + \left(x_0 - \bar{x}\right) a_i\right]^2, \qquad a_i = \frac{x_i - \bar{x}}{S_{XX}} \\
&= \sum \frac{1}{n^2} + \frac{2\left(x_0 - \bar{x}\right)}{n}\sum a_i + \left(x_0 - \bar{x}\right)^2 \sum a_i^2 \\
&= \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{XX}} \quad \text{since } \sum a_i = 0 \\
\hat{\alpha} + \hat{\beta} &\sim N\left(\alpha + \beta x_0, \left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}^2\right)}{S_{XX}}\right]\sigma^2\right)
\end{aligned}$$

We can construct confidence intervals and tests for $\mu_0 = \alpha + \beta x_0$ using

$$\frac{\hat{\alpha} + \hat{\beta} x_0 - (\alpha + \beta x_0)}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

A 99% CI for $\mu = \alpha + \beta x_0$

$$\left[\hat{\alpha} + \hat{\beta} x_0 \pm t_{(n-2)}^{.995} s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}\right]$$

Note that the Confidence interval is narrowest when $x_0 = \bar{x}$ and it increases as $|x_0 - \bar{x}|$ increases. Therefore, we can estimate $\alpha + \beta x_0$ most precisely when x_0 is close to $\bar{x}$, the mean of the x values used to fit the line.

6.4.3 R^2 and ANOVA

R^2

R^2 measures the proportion of the variation in Y explained by model. It is also called the Coefficient of Determination.

We obtain R^2 by decomposing the variation in Y into two parts, one part explained by the linear regression **(SSR)**, and one part unexplained (or error) by the regression, **(SSE)**.

If we did not have any $X's$, then we would estimate the mean of Y using $\bar{y}$ and the variance of Y using $s_Y^2 = \sum_{i=1}^n (y_i - \bar{y})^2/(n-1)$. The total variation in Y is the numerator of s_Y^2 and is called the **Total Sum of Squares,** SST, and is decomposed as follows.

$$\begin{aligned}
SST &= \sum (y_i - \bar{y})^2 \quad \text{adding and subtracting } \hat{y}_i \text{ within the brackets} \\
&= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2, \text{ where } \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i \\
&= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \\
&= \sum \hat{\epsilon}_i^2 + \sum (\hat{y}_i - \bar{y})^2 = SSE + SSR
\end{aligned}$$

The cross-product term is zero because of equations 6.5 and 6.6. SSE is the **sum of squares error** and SSR is the **sum of squares regression**.

Now we define R^2,

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Note that $0 \le R^2 \le 1$. If the regression line fits the data perfectly, then $y_i = \hat{y}_i$ and $\hat{\epsilon}_i = 0$ for all $i = 1, ..., n$. In that case, $SSE = 0$ and $R^2 = 1$.

If $\hat{y}_i = \bar{y}$ for all $i = 1, ..., n$, then $SSR = 0$ and $R^2 = 0$.

Returning to the R output for the co-op salary data, R^2 is called 'Multiple R-squared:' at the bottom of Figure 6.7, and is equal to 0.09779. Only about 10% of the variation in salaries is explained by the work term number.

R^2 has a deficiency in that it can be artificially inflated by adding more explanatory variables into the model. Adjusted R^2 incorporates a penalty for the number of explanatory variables in the model, and is the preferred measure of fit for linear regressions. Its formula is:

$$\text{Adjusted } R^2 = 1 - \frac{s^2}{s_Y^2}.$$

It is listed in Figure 6.7 as 'Adjusted R-squared'.

Anscombe's data

Anscombe's data provides a good illustration of issues with linear regression and R^2. See the file AnscombeR.pdf on Brightspace. The Anscombe dataset is built into R; simply type `anscombe` to see the dataset. The dataset consists of four pairs of x's and y's. We graph and fit linear models to each of the pairs.

```
> anscombe
   x1 x2 x3 x4    y1   y2    y3    y4
1  10 10 10  8  8.04 9.14  7.46  6.58
2   8  8  8  8  6.95 8.14  6.77  5.76
3  13 13 13  8  7.58 8.74 12.74  7.71
4   9  9  9  8  8.81 8.77  7.11  8.84
5  11 11 11  8  8.33 9.26  7.81  8.47
6  14 14 14  8  9.96 8.10  8.84  7.04
7   6  6  6  8  7.24 6.13  6.08  5.25
8   4  4  4 19  4.26 3.10  5.39 12.50
```

```
9   12 12 12  8 10.84 9.13  8.15  5.56
10   7  7  7  8  4.82 7.26  6.42  7.91
11   5  5  5  8  5.68 4.74  5.73  6.89
```

Below is the output from R for the fits of the linear models, $Y = \alpha + \beta X + \epsilon$, for each of the four pairs of x and y. Yes, they all have the SAME fit; SAME R^2, SAME coefficients estimates, SAME everything, so I only included one version. The graphs of the data pairs are quite different however.

```
> summary(ans.lm1)

Call:
lm(formula = y1 ~ x1, data = anscombe)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

Figure 6.8: R output, fit of linear model of Y on X

Figures 6.9 and 6.10 below show the scatterplots of the pairs of Anscombe's data together with the fitted linear model in the first columns. Plots of residuals versus fitted values from linear model fits are shown in the second columns. A linear model seems appropriate for the first pair, $(x1, y1)$. The second pair, $(x2, y2)$ require a quadratic model. The third pair has an outlier which raises the regression line. The fourth pair has an influential point which totally determines the line. Thus, although their R^2 values are all the same, we see that the linear model fits are all very different for the four pairs.

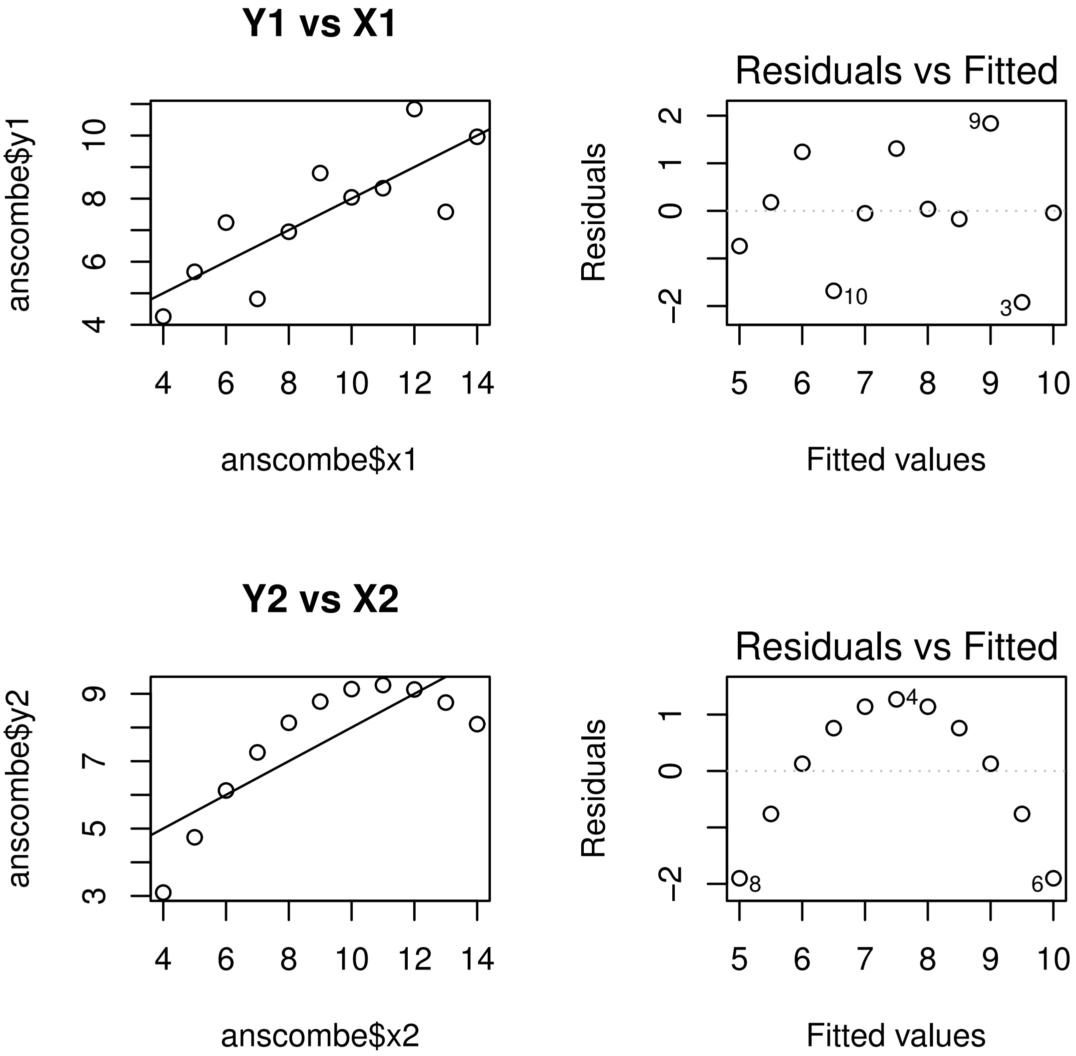


Figure 6.9: Anscombe pairs 1 and 2, Scatterplots; Residual plots

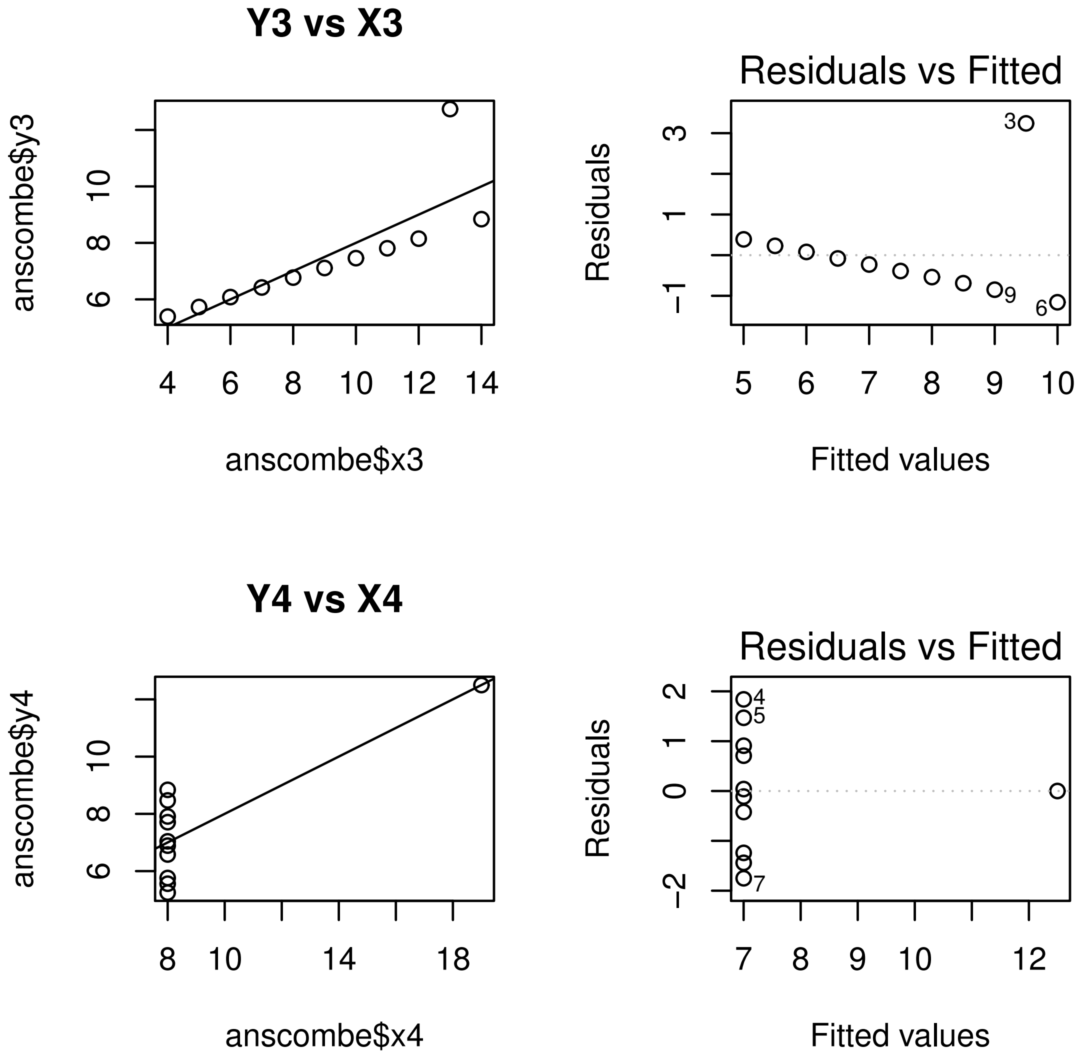


Figure 6.10: Anscombe pairs 3 and 4, Scatterplots; Residual plots

R Code for Anscombe analyses:

```
anscombe
plot(anscombe$x1, anscombe$y1, main=’Y1 vs X1’)
ans.lm1<-lm(y1˜x1, data=anscombe)
abline(ans.lm1)
```

```
plot(ans.lm1,which=1,add.smooth=FALSE)
summary(ans.lm1)

plot(anscombe$x2,  anscombe$y2, main='Y2 vs X2')
ans.lm2<-lm( y2~ x2, data=anscombe)
abline(ans.lm1)
plot(ans.lm2,which=1,add.smooth=FALSE)
summary(ans.lm2)

plot( anscombe$x3,  anscombe$y3, main='Y3 vs X3')
ans.lm3<-lm( y3~ x3,data=anscombe)
abline(ans.lm1)
plot(ans.lm3,which=1,add.smooth=FALSE)
summary(ans.lm3)

plot( anscombe$x4,  anscombe$y4, main='Y4 vs X4')
ans.lm4<-lm( y4~ x4,data=anscombe)
abline(ans.lm1)
plot(ans.lm4,which=1,add.smooth=FALSE)
summary(ans.lm4)
```

ANOVA

ANOVA stands for Analysis of Variance, and it is a tabulation of the sources of variation that we derived for R^2. It usually also includes test statistics, and usually, an $F-$test statistic with its $p-value$. For our simple linear regression, the ANOVA table has the following form.

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X variable	df_R	SSR	$\text{MSR} = \text{SSR}/\text{df}_R$	MSR/MSE	p-value
Residuals(Error)	df_E	SSE	$\text{MSE} = \text{SSE}/\text{df}_E$		

Table 6.1: Analysis of Variance Table

The quantities in the table are defined below:

- Df = degrees of freedom

- Sum Sq = sum of squares
- Mean Sq = mean square = Sum Sq/Df
- F value = value of F statistic for testing H_0 that all coefficients of X variable(s) are zero = MSR/MSE
- Pr(>F) = p-value for the test H_0 that all coefficients of the X variable(s) are zero; small p-values indicate that there is evidence against H_0
- MSE = $s^2 = \sum \hat{\epsilon}_i^2/\mathtt{df}_E$ is the estimate of σ^2.

The ANOVA table for the Co-op salary data appears below. Note that the p-value for the F-test is exactly the same as the p-value for the t test of the $H_0 : \beta = 0$ in Figure 6.7. That is because we have only one X variable in our model, namely WTNumN.

```
> anova(Sal.lm)
Analysis of Variance Table

Response: SalMonth
            Df    Sum Sq  Mean Sq F value    Pr(>F)
WTNumN       1  95290565 95290565  124.54 < 2.2e-16 ***
Residuals 1149 879171908   765163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

6.4.4 Checking Goodness of Fit

One method to check the fit of the model is to plot the data together with the fitted line. Are the data points scattered randomly about the fitted line? In Figure 6.9, the scatterplot of $(x1, y1)$ with the fitted line indicates a good model fit. The other Anscombe pairs do not fit the linear model well.

Another method is to plot the residuals. The model has the form,

$$Y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim N\left(0, \sigma^2\right) \text{ independent.}$$

To estimate ϵ_i, we use

$$\hat{\epsilon}_i = y_i - \left(\hat{\alpha} + \hat{\beta} x_i\right) = y_i - \hat{y}_i = \text{residual}_i.$$

If the model is correct, we would expect $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)$ to behave like a random sample from $N(0, \sigma^2)$.

A plot of residuals should be scattered about the centre line at zero, all within approximately $\pm 3s$. We plot the residuals versus the fitted values, $\hat{y}_i$, and look for:

1. constant variance,
2. patterns that suggest nonlinearity,
3. outliers,
4. influential points.

In the plots of residuals versus fitted values of Figures 6.9 and 6.10, the pairs 2, 3 and 4 indicate problems with the linear models.

We can also plot a histogram of the residuals to check for normality, or a Normal Q-Q plot which is explained in the next section.

6.4.5 Normal Q-Q plots

A Normal Q-Q plot of residuals is a graph of the **ordered** residuals from smallest to largest, versus the corresponding percentiles of the $N(0, 1)$ distribution. Suppose that $n = 10$ and we have 10 distinct ordered residuals, $\hat{\epsilon}_{(1)} < ... < \hat{\epsilon}_{(n)}$. The brackets are used to denote ordered values from (1) to (n).

- $\hat{\epsilon}_{(1)}$ is plotted versus the $100\left(\frac{1-.5}{10}\right)$ = 5th percentile of $N(0, 1)$ = -1.644854.
- $\hat{\epsilon}_{(2)}$ is plotted versus the $100\left(\frac{2-.5}{10}\right)$ = 15th percentile of $N(0, 1)$ = -1.036433
- ...
- $\hat{\epsilon}_{(10)}$ is plotted versus the $100\left(\frac{10-.5}{10}\right)$ = 95th percentile of $N(0, 1)$ = 1.644854

If the residuals are approximately normally distributed, then the graph should roughly look like a straight line. Figure 6.11 is a Normal Q-Q plot of a sample of size 200 generated in R from the $N(0, 1)$ distribution. The points fall roughly on a straight line. Figure 6.12 is a Normal Q-Q plot of a sample of size 200 generated in R from the $\chi^2_{(2)}$ distribution. The points do NOT fall on a straight line.

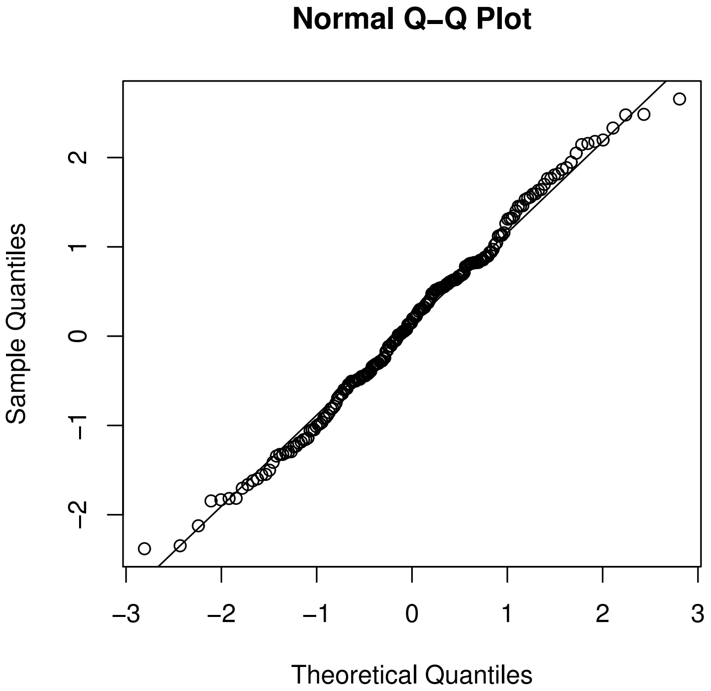


Figure 6.11: Normal QQ plot of Normal data

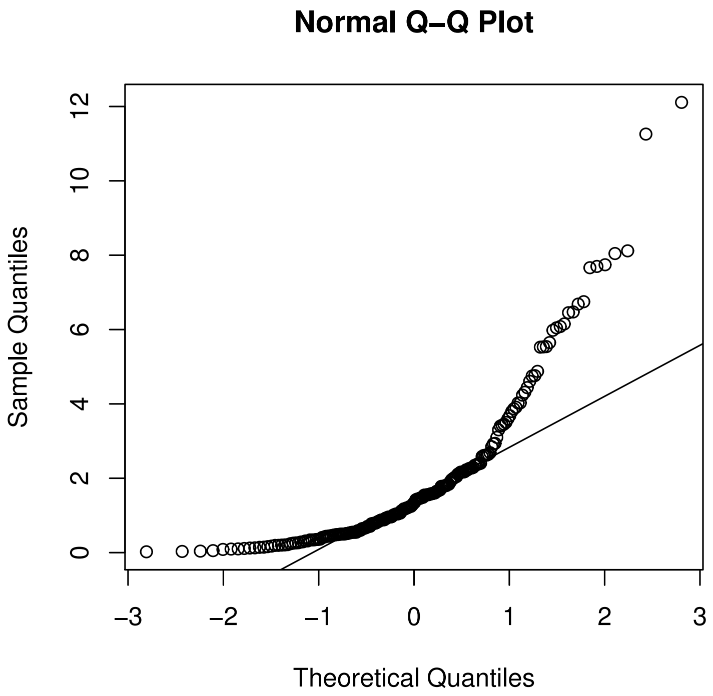


Figure 6.12: Normal QQ plot of Chi-square(2) data

R Code for Normal Q-Q plots:

```
set.seed(12345)
x1<-rnorm(200)
qqnorm(x1)
qqline(x1)   #overlays a line through the first and third quartiles

x3<-rchisq(200,df=2)
qqnorm(x3)
qqline(x3)
```

6.5 Analysis of Paired Measurements

Optional reading: Section 13.7

The analysis of paired measurements is an application of the one sample model.

Example 6.5.1. Twelve students in a statistics course recorded the scores listed below on their first and second tests in the course.

	Student											
	1	2	3	4	5	6	7	8	9	10	11	12
Test 1	64	28	90	30	97	20	100	67	54	44	100	71
Test 2	80	87	90	57	89	51	81	82	89	78	100	81

Test the hypothesis that there is no difference in the scores for the 2 tests.

Solution:

Note: The Test 1 and Test 2 pairs are not independent of each other. We would expect results from different individuals to be independent of one another however.

Let $X_i = i'th$ difference ($\text{Test}_1 - \text{Test}_2$), and assume $X_i \sim N(\mu, \sigma^2)$ independent, σ^2 is unknown. This is just a one sample model for the X_i's.

We will use the t statistic,

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

to test $H_0 : \mu = 0$.

$$\begin{aligned}
p - value &= P\left\{\left|t_{(n-1)}\right| \geq \frac{|\bar{x} - 0|}{s/\sqrt{n}} \mid H_0 : \mu = 0\right\} \\
\bar{x} &= -16.67 \\
s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} = 474.97 \\
t_{\text{obs}} &= \frac{|-16.67|}{\sqrt{(474.97)/12}} = 2.65 \\
p - value &= P\left\{\left|t_{(n-1)}\right| \geq 2.65\right\}
\end{aligned}$$

$$P\left\{t_{(11)} \geq 2.201\right\} = .025$$
$$P\left\{t_{(11)} \geq 2.718\right\} = .01$$
$$\Longrightarrow \; .02 < p - value \leq .05$$

There is evidence against the hypothesis that there is no difference in the scores for the 2 exams, with a mean difference of -16.67. (p-value = .02262)

A 95% Confidence Interval for the mean difference has the form,

$$\left[\bar{X} \pm t_{(n-1)} \frac{s}{\sqrt{n}}\right] = -16.67 \pm (2.201)\sqrt{\frac{474.97}{12}}$$
$$= [-30.5, -2.82]\,.$$

Note that the interval does not cover zero and the scores were significantly lower for Test 1 than Test 2.

Our concluding statement is that: Scores on Test 1 were significantly lower than those on Test 2 with a mean difference of 16.67 (s.e. 13.85).

R Code for Paired Measurements Example:

```
>T1<-c( 64, 28, 90, 30, 97, 20, 100, 67, 54, 44, 100, 71)
>T2<-c( 80, 87, 90, 57, 89, 51, 81, 82, 89, 78, 100, 81)
> var(T1-T2)
[1] 474.9697
> t.test(T1-T2)

One Sample t-test

data:  T1 - T2
t = -2.6491, df = 11, p-value = 0.02262
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -30.513786  -2.819547
sample estimates:
mean of x
-16.66667
```

Why pair?

Suppose instead we had randomly chosen a sample of Test 1 and Test 2 results, assuming,

$$\text{Test 1 results} \sim N\left(\mu_1, \sigma^2\right)$$

$$\text{Test 2 results} \sim N\left(\mu_2, \sigma^2\right)$$

and investigated $H_0 : \mu_1 - \mu_2 = 0$.

Suppose that by chance, the first group consisted of students ~~that were brighter~~ who studied more than the second group. Then any difference in the two test results may be due to the ~~intelligence difference~~ study time of the groups rather than to differences in the difficulty of the two tests. With pairing, differences in the two tests will not be obscured by the second sample of students being entirely different (independent) from the first.

Pairing is effective when there is considerable variation between subjects because it controls or reduces unwanted or extraneous variation.

Index