Design of Scientific Studies - Notes on Week 1

Nathan Taback

1 Why Design Scientific Studies?

Why should scientific studies be designed? Some reasons include avoiding bias, variance reduction, and system optimization.

1.1 Big Data and Designing Scientific Studies

1.1.1 What is Big data and why does it matter?

According to SAS (http://www.sas.com/en_id/insights/big-data/what-is-big-data.html)

> "… big data may be as important to business - and society - as the Internet has become. Why? More data lead to more accurate analyses."

The SAS webpage defines Big data (quoting Doug Laney) in terms of increasing volume (e.g., streaming social media) , velocity (e.g., sensors, smart metering), and variety (e.g., unstructured text documents, video, audio, financial transactions).

1.1.2 Is statistical sampling and randomization still relevant in the era of Big Data?

This question was asked by Professor Xiao-Li Meng asks (http://statistics.fas.harvard.edu/news/xiao-li-meng-chicago-statistician-year).

Meng then considers a the following: if we want to estimate a population mean which dataset would give us more accurate results: a 1% simple random sample or a dataset containing self-reports of the quantity of interest that covers 95% of the population?

The total error is captured by the mean square error (MSE). The mean square error of an estimator $\hat{\theta}$ of θ is,

$$MSE = E\left(\hat{\theta} - \theta\right) = Var\left(\hat{\theta}\right) + \left(E\left(\hat{\theta} - \theta\right)\right)^2.$$

In other words, $Total\,Error = Variance + Bias^2$. The term $E\left(\hat{\theta} - \theta\right)$ is called the bias of the estimator $\hat{\theta}$. If the bias is 0 then the estimator is called unbiased.

Suppose we have a finite population of measurements $\{x_1, \ldots, x_N\}$ of some quantity, say the total number of hours spent on the internet during a one-year period for every person in Canada. In 2015 the population of Canada (http://www.statcan.gc.ca/daily-quotidien/150617/dq150617c-eng.htm) is $N = 35,749,600$ or approximately 35.8 Million people. In order to estimate the mean number of hours spent on the Internet is it better to: take a simple random sample of 100 people and estimate the mean number of hours spent on the Internet; or use a large database (much larger than the random sample) that contains self-reports of hours spent on the Internet?

Suppose that $\bar{x}_a$ is the sample mean from the database and $\bar{x}_s$ is the mean from the random sample. Meng (http://www.stat.harvard.edu/Faculty_Content/meng/COPSS_50.pdf) shows that in order for $MSE\left(\bar{x}_a\right) < MSE\left(\bar{x}_s\right)$ it's sufficient that

$$f_a > \frac{n_s \rho_N^2}{1 + n_s \rho^2},$$

where f_a is the fraction of the population in the database, ρ is the correlation between the response being recorded and the recorded value, and n_s is the size of the random sample. For example, if $n_s = 100$ the database would need over 96% of the population if $\rho = 0.5$ to guarantee that $MSE(\bar{x}_a) < MSE(\bar{x}_s)$. In our example this would require a database with 34,319,616 Canadians.

This illustrates the main advantage of probabilistic sampling and the danger of putting faith in "Big Data" simply because it's Big!

2 Review of Statistical Theory

When an operation/experiment is repeated under nearly same conditions the fluctuation from one repetition to another is called noise or experimental variation or error. In statistics this refers to variation that is usually unavoidable. An experimental run has been performed when an apparatus has been setup and allowed to function under a specific set of experimental conditions. For example, machining a part under specific manufacturing conditions.

2.1 Data

Experimental data describes the outcome of the experimental run. For example 10 successive runs in a chemical experiment produce the following data:

```
set.seed(100)
# Generate a random sample of 10 observations from a N(60,10^2)
dat <- round(rnorm(10,mean = 60,sd = 10),1)
dat
```

```
##  [1] 55.0 61.3 59.2 68.9 61.2 63.2 54.2 67.1 51.7 56.4
```

2.2 Distributions

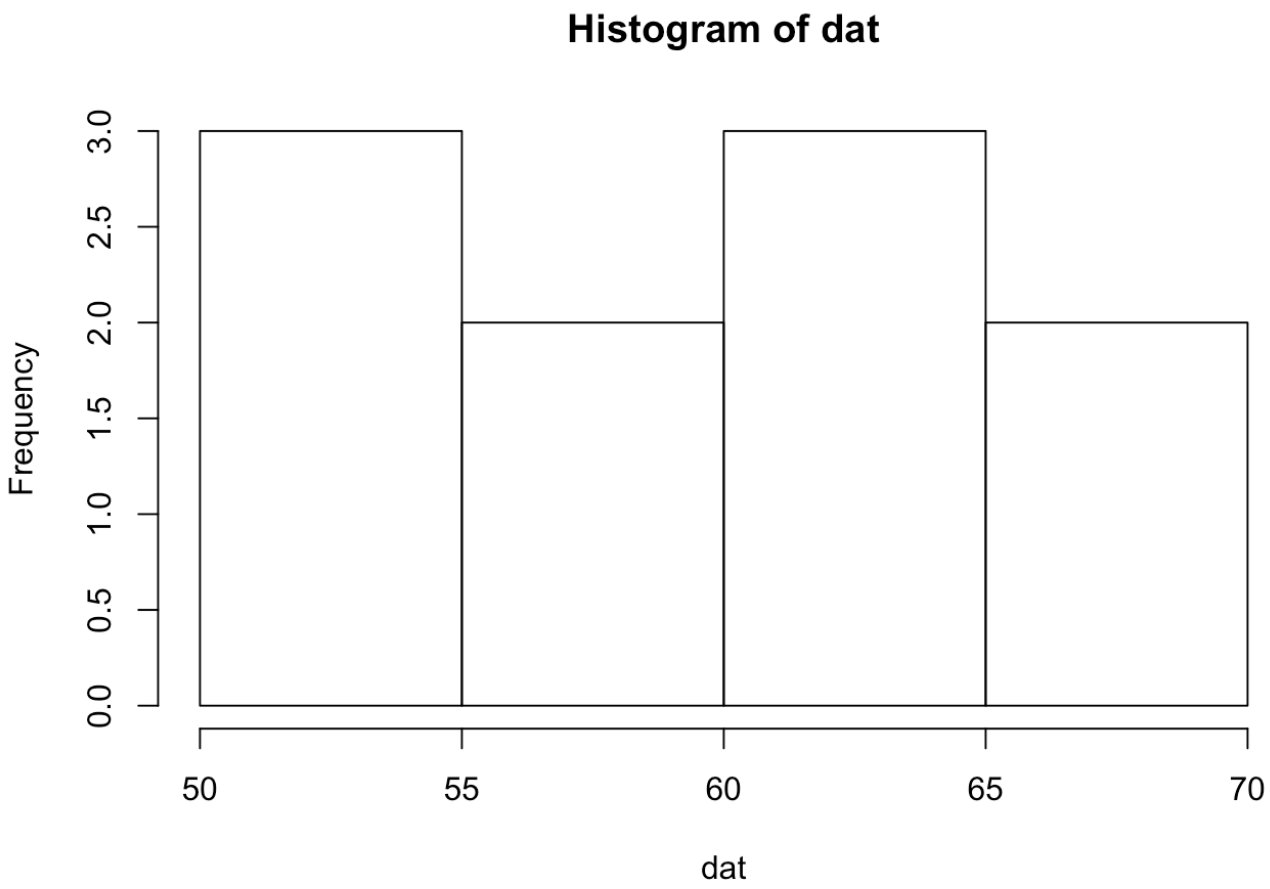
Distributions can be displayed graphically or numerically.

A histogram is a graphical summary of a data set.

```
summary(dat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   51.70   55.35   60.20   59.82   62.73   68.90
```

```
hist(dat)
```



The total aggregate of observations that might occur as a result of repeatedly performing a particular operation is called a population of observations. The observations that actually occur are some kind of sample from the population.

2.2.1 Continuous Distributions

A continuous random variable X is fully characterized by it's density function $f(x)$, where $f(x) \geq 0$, f is piecewise continuous, and $\int_{-\infty}^{\infty} f(x)dx = 1$.

The cumulative distribution function (CDF) of X is defined as:

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

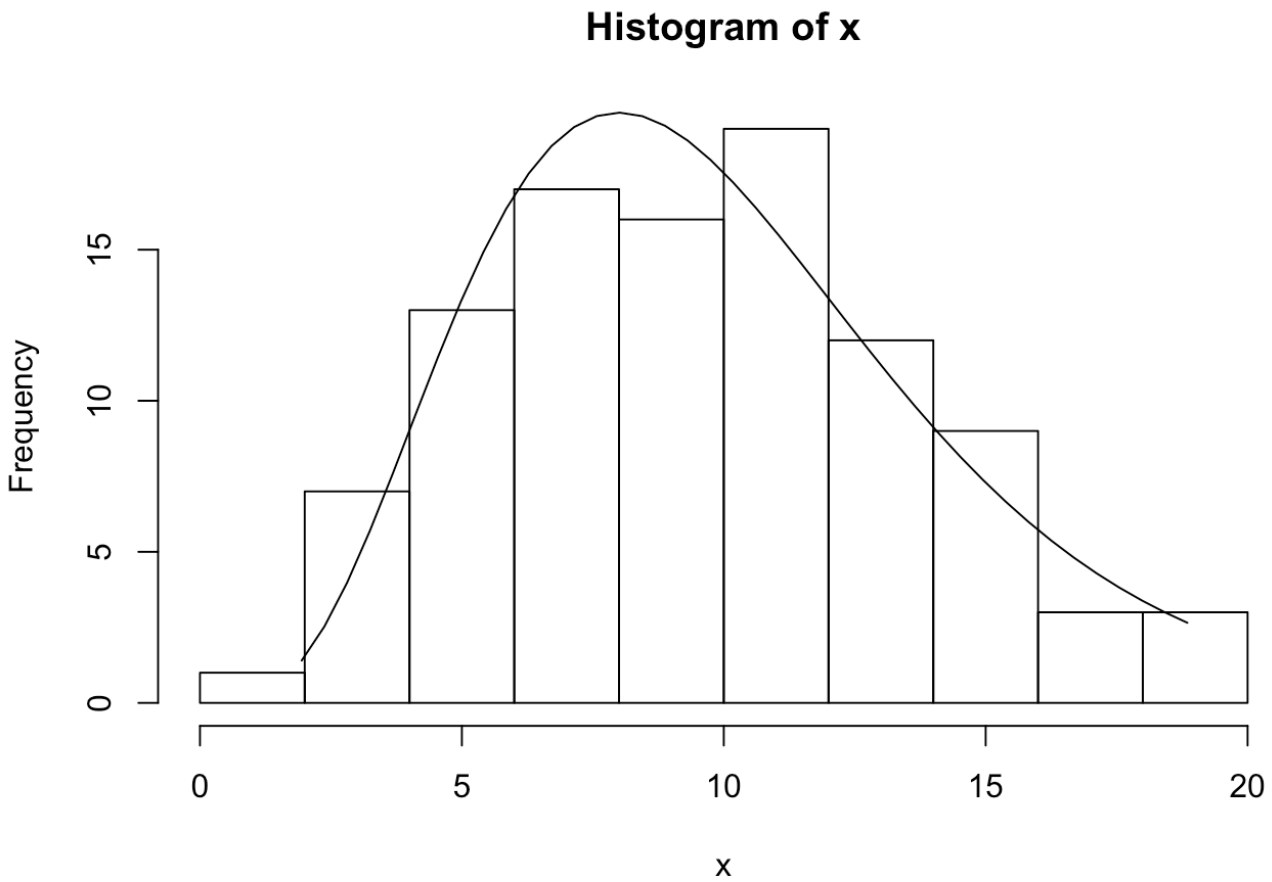
If f is continuous at x then $F'(x) = f(x)$ (fundamental theorem of calculus). The CDF can be used to calculate the probability that X falls in the interval (a, b). This is the area under the density curve which can also be expressed in terms of the CDF:

$$P\,(a < X < b) = \int_{a}^{b} f(x)dx = F(b) - F(a).$$

In R a list of all the common distributions can be obtained by the command `help("distributions")`.

The following R code draws a random sample of 100 observations from a Chi-square distribution on 10 degrees of freedom χ^2_{10}. The density function of the χ^2_{10} is superimposed over the histogram of the sample.

```
x<- rchisq(100,10) # draw a sample of 100 from chi-square 10
h <- hist(x) # create the histogram
# superimpoise chi-square density over histogram
xfit<-seq(min(x),max(x),length=40)
yfit <- dchisq(xfit,10) #chi-square density
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit,yfit)
```



2.3 Randomness

A random drawing is where each member of the population has an equal chance of being selected. The hypothesis of random sampling may not apply to real data. For example, cold days are usually followed by cold days. So daily temperature not directly representable by random drawings. In many cases we can't rely on the random sampling property although design can make this assumption relevant.

2.4 Parameters and Statistics

What is the difference between a parameter and a statistic? A parameter is a population quantity and a statistic is a quantity based on a sample drawn from the population.

Example: The population of all adult (18+ years old) males in Toronto, Canada. Suppose that there are N adult males. The quantity of interest, y, is age. A sample of size n is drawn from this population. The population mean is $\mu = \sum_{i=1}^{N} y_i/N$ and the sample mean is $\bar{y} = \sum_{i=1}^{n} y_i/n$.

2.5 Residuals and Degress of Freedom

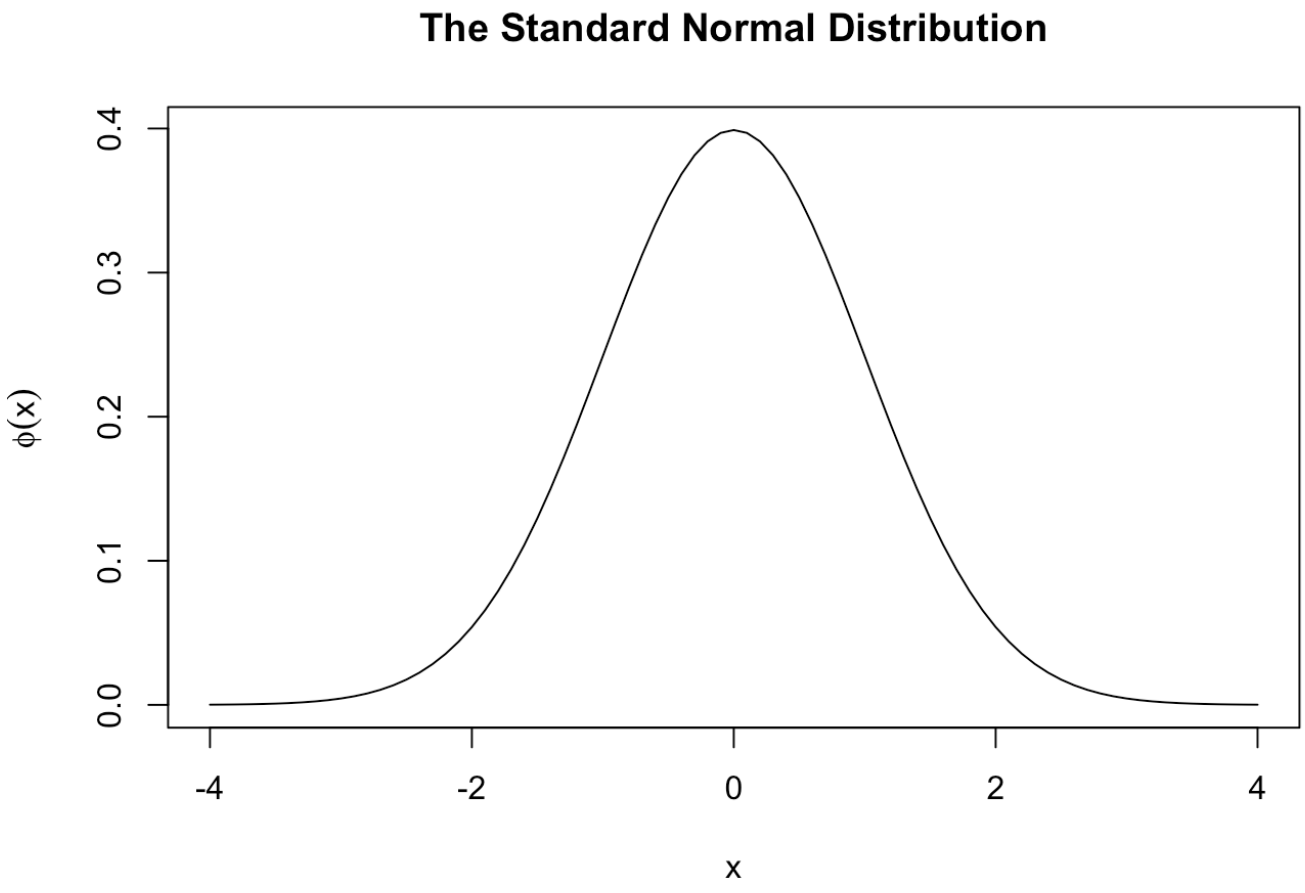
$y_i - \bar{y}$ is called a residual. Since $\sum(y_i - \bar{y}) = 0$ any $n - 1$ completely determine the the last observation. This is a constraint on the the residuals. So n residuals have $n - 1$ degrees of freedom since the last residual cannot be freely chosen.

2.6 The Normal Distribution

The density function of the normal distribution with mean μ and standard deviation σ is:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "The Standard Normal Distribution", yla
b=expression(paste(phi(x))))
```



A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by $X \sim N\left(\mu, \sigma^2\right)$.

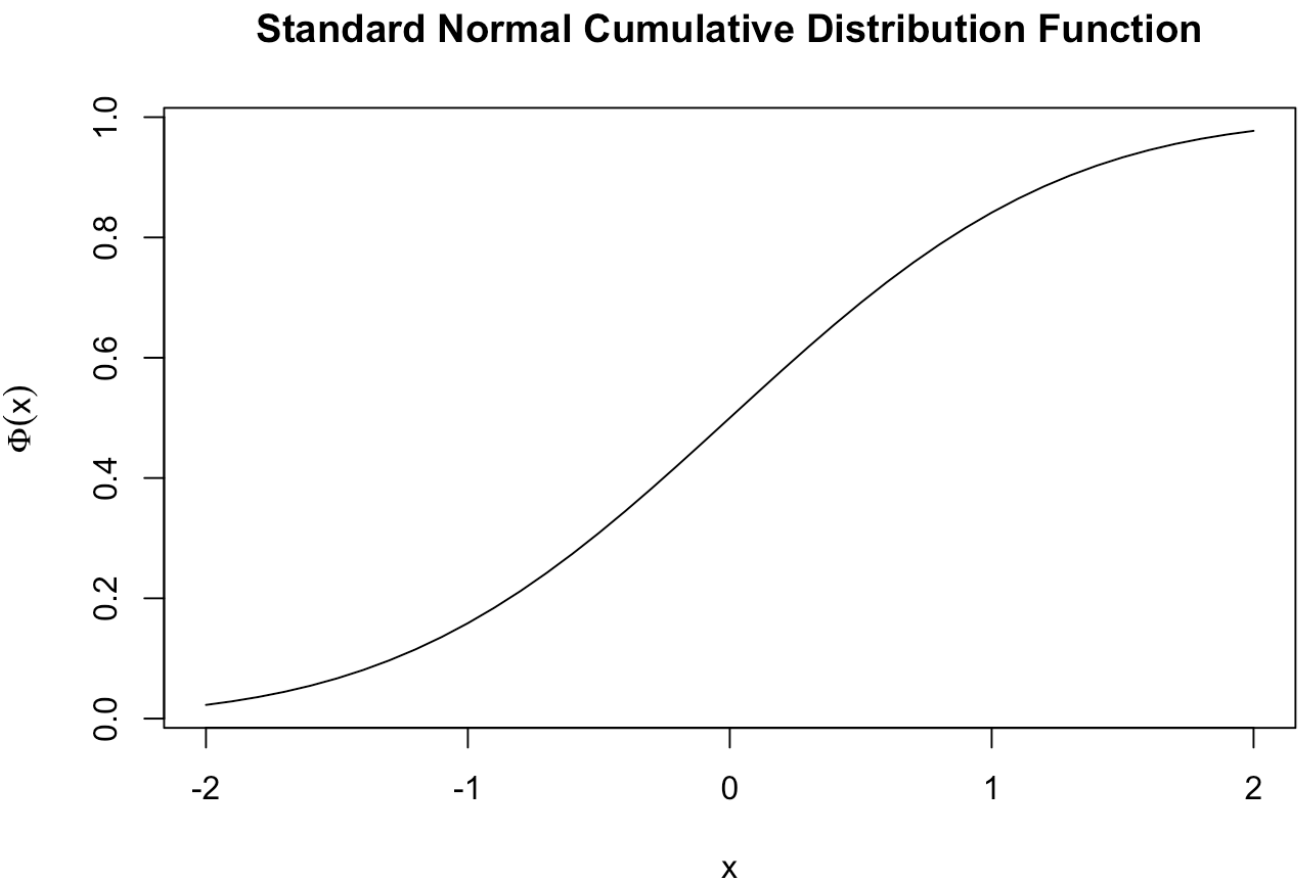
If $Y \sim N\left(\mu, \sigma^2\right)$ then $Z \sim N(0, 1)$, where $Z = \frac{Y-\mu}{\sigma}$.

The cumulative distribution function (CDF) of a $N(0, 1)$ distribution,

$$\Phi(x) = P(X < x) = \int_{-\infty}^{x} \phi(x)dx$$

is shown in the plot below using the R function for the normal CDF `pnorm()`.

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",xlab="x",ylab=expression(pas
te(Phi(x))),main = "Standard Normal Cumulative Distribution Function")
```



2.6.1 Exercises

1. Use R to calculate the $P(-1 < Z < 2)$, where $Z \sim N(0, 1)$. Answer:

```
pnorm(2)-pnorm(-1)
```

```
## [1] 0.8185946
```

2. Use R to calculate the $P(X > 5)$, where $X \sim N(6, 2)$. Answer:

```
1-pnorm(5,mean = 6,sd = sqrt(2))
```

```
## [1] 0.7602499
```

2.6.2 Normal Quantile Plots

Normal quantile plots are useful for assessing if data fits a normal distribution.

Suppose that $X_1, X_2, \ldots, X_n$ are a random sample from the uniform distribution on $[0, 1]$. The sample can be ordered from smallest to largest $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. It can be shown that

$$E\left(X_{(j)}\right) = \frac{j}{n+1}.$$

If the observations are uniform then the plot of the ordered observations $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ versus their expected values will be a straight line.

If we want to investigate if a sample $X_1, X_2, \ldots, X_n$ follows a certain distribution with CDF F_X then we can transform the sample to a uniform distribution on $[0, 1]$ by calculating $Y_i = F_X(X_i)$.

So, given a sample $X_1, X_2, \ldots, X_n$ plot

$$F(X_{(k)}) \quad \text{vs.} \quad \frac{k}{n+1}.$$

This is the same as

$$X_{(k)} \quad \text{vs.} \quad F^{-1}\left(\frac{k}{n+1}\right).$$

This means that the $k/(n+1)$ quantile is assigned to the k^{th} order statistic. But in some implementations sometimes the k^{th} quantile is assigned to $X_{(k)}$ is $(k-0.5)/n$ (see Rice, pg 352-355).

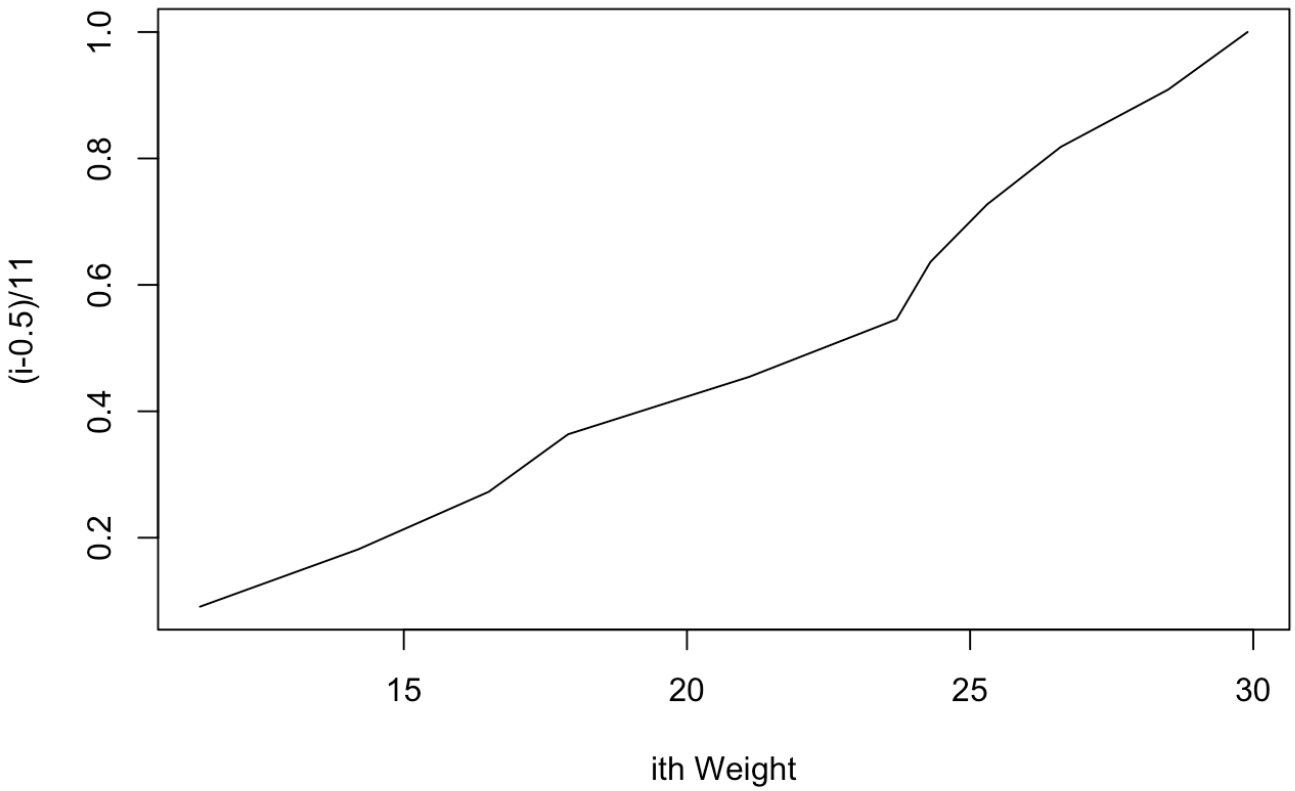
The following data from BHH are the weights from 11 tomato plants.

```
##  [1] 29.9 11.4 26.6 23.7 25.3 28.5 14.2 17.9 16.5 21.1 24.3
```

Do the weights follow a Normal distribution?

If the tomato weights are normally distributed then a plot of the ordered tomato weights, $y_{(1)} < y_{(2)} < \cdots < y_{(11)}$ versus the cumulative probabilities $p_i = (i-0.5)/N$, where N is the number of observations should be the same shape as the CDF of the Normal distribution.

```
plot(sort(tomato.data$pounds),1:11/length(tomato.data$pounds),type="l",xl
ab="ith Weight",ylab="(i-0.5)/11")
```



It's difficult to tell if the weights have the same shape as the Normal CDF. But, the curve can be stretched at the extreme ends so that it becomes a straight line. A method for stretching out the curve is developed below.

Assume that the weights, $y_i \sim N\left(\mu, \sigma^2\right)$.

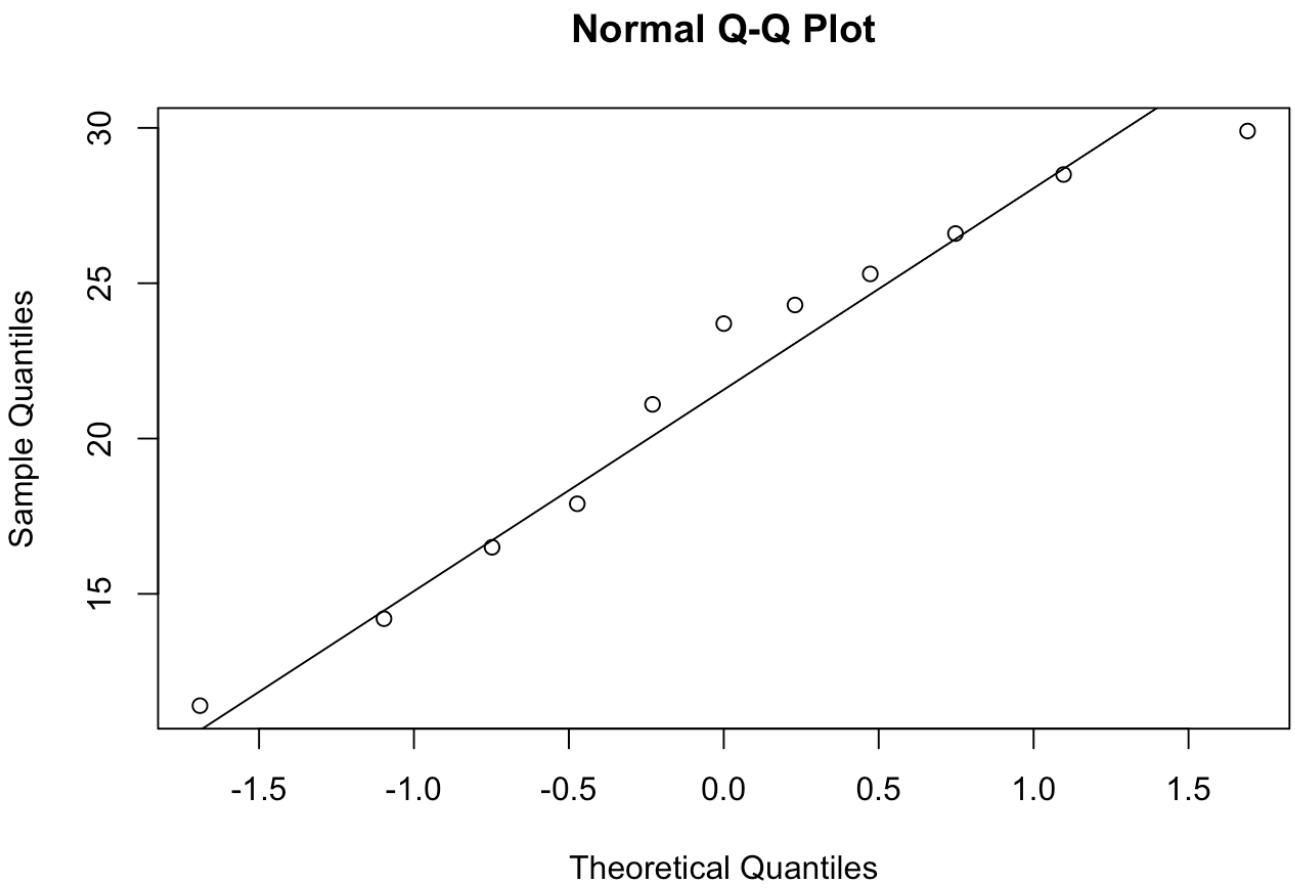
Then $\Phi(y_i)$ has a uniform distribution over $[0, 1]$. This means that the expected values of $\Phi(y_i), i = 1, \ldots N$ should be equally spaced over $[0, 1]$ or the N points $(p_i, \Phi(y_{(i)})$ should fall on a straight line. Applying the Φ^{-1} transform to the horizontal and vertical scales, the N points

$$\left(\Phi^{-1}(p_i), y_i\right), i = 1, \ldots, N,$$

should follow a straight line. These points form the **normal probability plot** of the tomato plant weights. (Wu and Hamada, pages 77-78)

A normal probability plot in R can be obtained using `qqnorm()` for the normal probability plot and `qqline()` to add the straight line.

```
qqnorm(tomato.data$pounds)
qqline(tomato.data$pounds)
```



In this case assuming that the data are generated from a normal distribution is a plausible assumption since most of the points are close the straight line.

2.6.3 Exercises

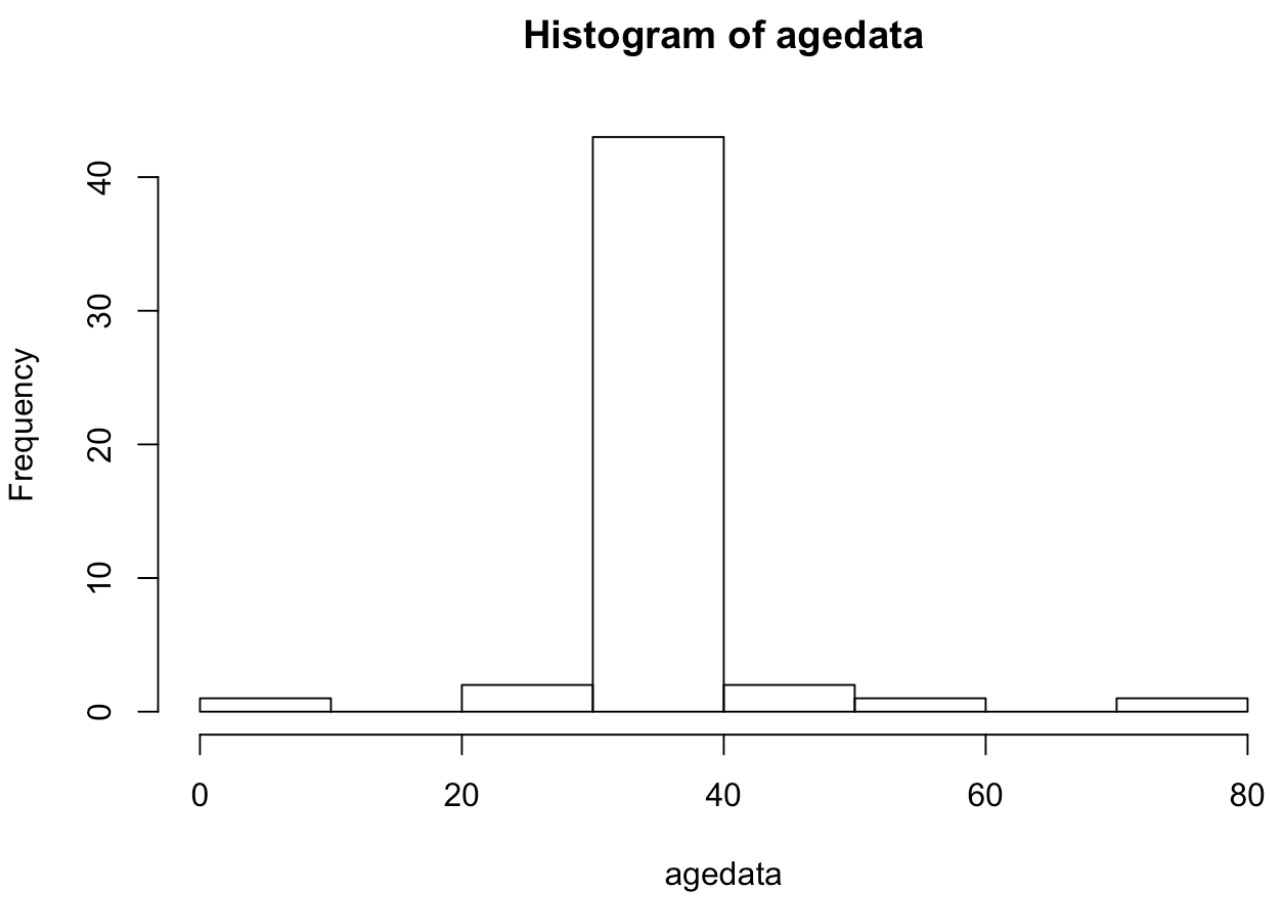
1. The following 50 dataset contains the ages of participants in a study of social media habits. Is it plausible to assume that the data are normally distributed?

```
##  [1] 34.2 35.1 35.5 36.5 47.0 34.7  9.8 36.1 40.6 34.8 38.3 34.7 37.3
33.8
## [15] 34.4 28.0 33.7 28.4 73.7 36.2 33.8 36.1 32.8 34.9 35.7 55.0 36.8
35.0
## [29] 33.7 33.6 37.9 35.9 38.6 34.8 34.3 39.3 33.1 33.4 30.9 36.2 35.2
36.1
## [43] 35.3 35.6 33.7 33.9 34.4 33.7 32.5 38.2
```

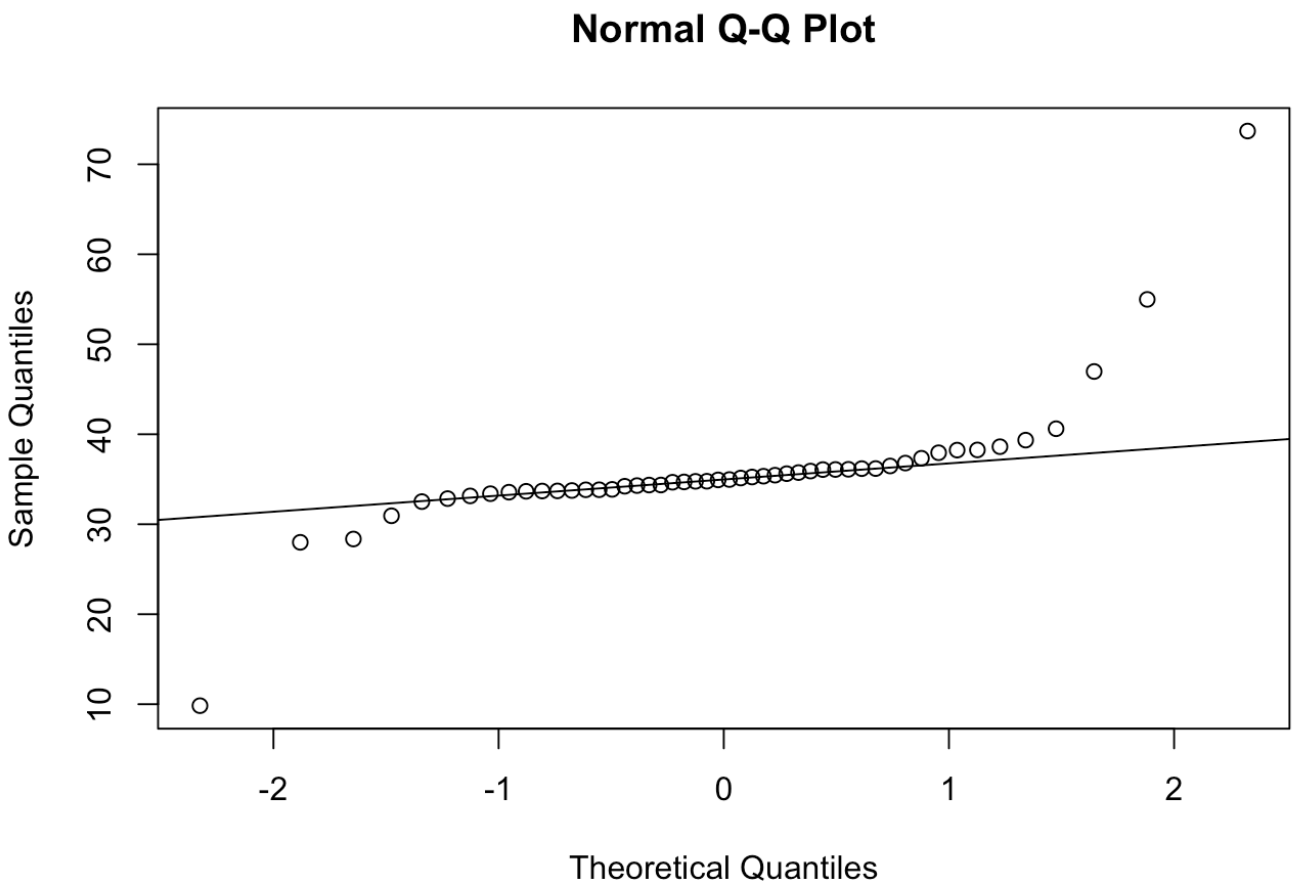
```
summary(round(agedata,1))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   33.72   34.95   35.86   36.20   73.70
```

```
hist(agedata)
```




```
qqnorm(agedata);qqline(agedata)
```



Answer: The histogram and the qqplot indicate the tails of the age distribution are heavier than the normal distribution. In the qqplot this is indicated by the marked deviations from the straight line for very small and large sample quantiles. Thus, the qqplot indicates that the data is not normally distributed.

2.7 Central Limit Theorem

The central limit theorem states that if $X_1, X_2, \ldots$ is an independent sequence of identically distributed random variables with mean $\mu = E(X_i)$ and variance $\sigma^2 = Var(X_i)$ then

$$\lim_{n\to\infty} P\left(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x\right) = \Phi(x),$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$ and $\Phi(x)$ is the standard normal CDF. This means that the distribution of $\bar{X}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Example: A fair coin is flipped 50 times. What is the distribution of the average number of heads?

Let $X_1, \ldots, X_{50}$ where $X_i = 1$, if the toss is a head and $X_i = 0$ if the toss is a tail. Since the coin is fair $P(X_i = 1) = 0.5, i = 1, \ldots, 50$. The average number of heads is $\sum_{i=1}^{50} X_i/50$. $E(X_i) = 0.5$ and $Var(X_i) = p(1-p) = 0.5(1-0.5) = 0.25$ so $\sum_{i=1}^{50} X_i/50 \overset{approx}{\sim} N(0.5, 0.25/\sqrt{50})$

2.8 Chi-Square Distribution

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables that have a $N(0, 1)$ distribution. The distribution of

$$\sum_{i=1}^{n} X_i^2,$$

has a chi-square distribution on n degrees of freedom or χ^2_n.

The mean of a χ^2_n is n with variance $2n$.

The chi-square distribution is a right-skewed distribution, but becomes normal as the degrees of freedom increases. In the plot below the χ^2_{50} density is very close to the $N(50, 100)$ density.

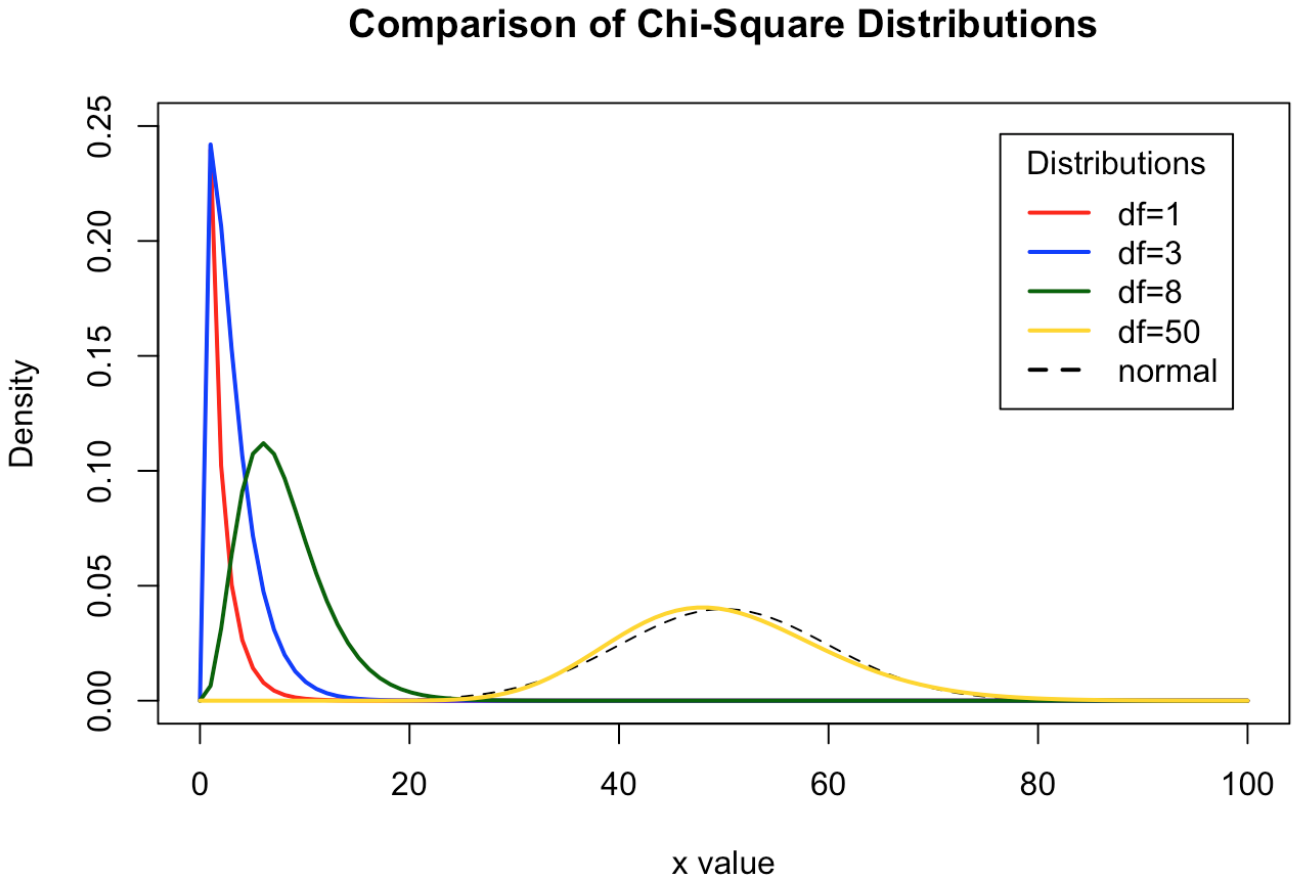
```
# Compare chi-square densities with normal density
x <- seq(0, 100, length=100)
hx <- dnorm(x,mean = 50,sd = sqrt(2*50)) #normal density

degf <- c(1, 3, 8, 50)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=50", "normal")

plot(x, hx, type="l", lty=2, xlab="x value",
     ylab="Density", main="Comparison of Chi-Square Distributions",yli
m=c(0,0.25))

for (i in 1:4){
  lines(x, dchisq(x,degf[i]), lwd=2, col=colors[i]) # dchisq is the chi-s
quare density
}

legend("topright", inset=.05, title="Distributions",
       labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```



Let $X_1, X_2, \ldots, X_n$ be independent with a $N(\mu, \sigma^2)$ distribution. The distribution of the sample variance $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$ has a χ^2_{n-1} distribution, namely,

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

2.8.1 Exercises

1. If $X_1, X_2, \ldots, X_50$ are a random sample from a $N(40, 25)$ then calculate $P(S^2 > 27)$.

Answer: We know that $\frac{49}{25} S^2 \sim \chi^2_{49}$. So

$$\begin{aligned} P\left(\frac{49}{25} S^2 > \frac{49}{25} \times 27\right) &= P\left(\chi^2_{49} > \frac{49}{25} \times 27\right) \\ &= P\left(\chi^2_{49} > 1.96 \times 27\right) \\ &= P\left(\chi^2_{49} > 52.92\right). \end{aligned}$$

The CDF of the χ^2_n in R is `pchisq()`.

Therefore,

```
1-pchisq(q = 52.92,df = 49)
```

```
## [1] 0.325325
```

So, $P(S^2 > 27) = 1 - P(S^2 < 27) = 0.325325$.

2.9 t Distribution

If $X \sim N(0,1)$ and $W \sim \chi^2_n$ then the distribution of $\frac{X}{\sqrt{W/n}}$ has a t distribution on n degrees of freedom or $\frac{X}{\sqrt{W/n}} \sim t_n$.

Let $X_1, X_2, \ldots$ is an independent sequence of identically distributed random variables that have a $N(0,1)$ distribution. The distribution of

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

where $S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)$. This follows since $\bar{X}$ and S^2 are independent.

The t distribution for small values of n has "heavier tails" compared to the normal. As the degrees of freedom increases the t-distribution is almost identical to the normal distribution.

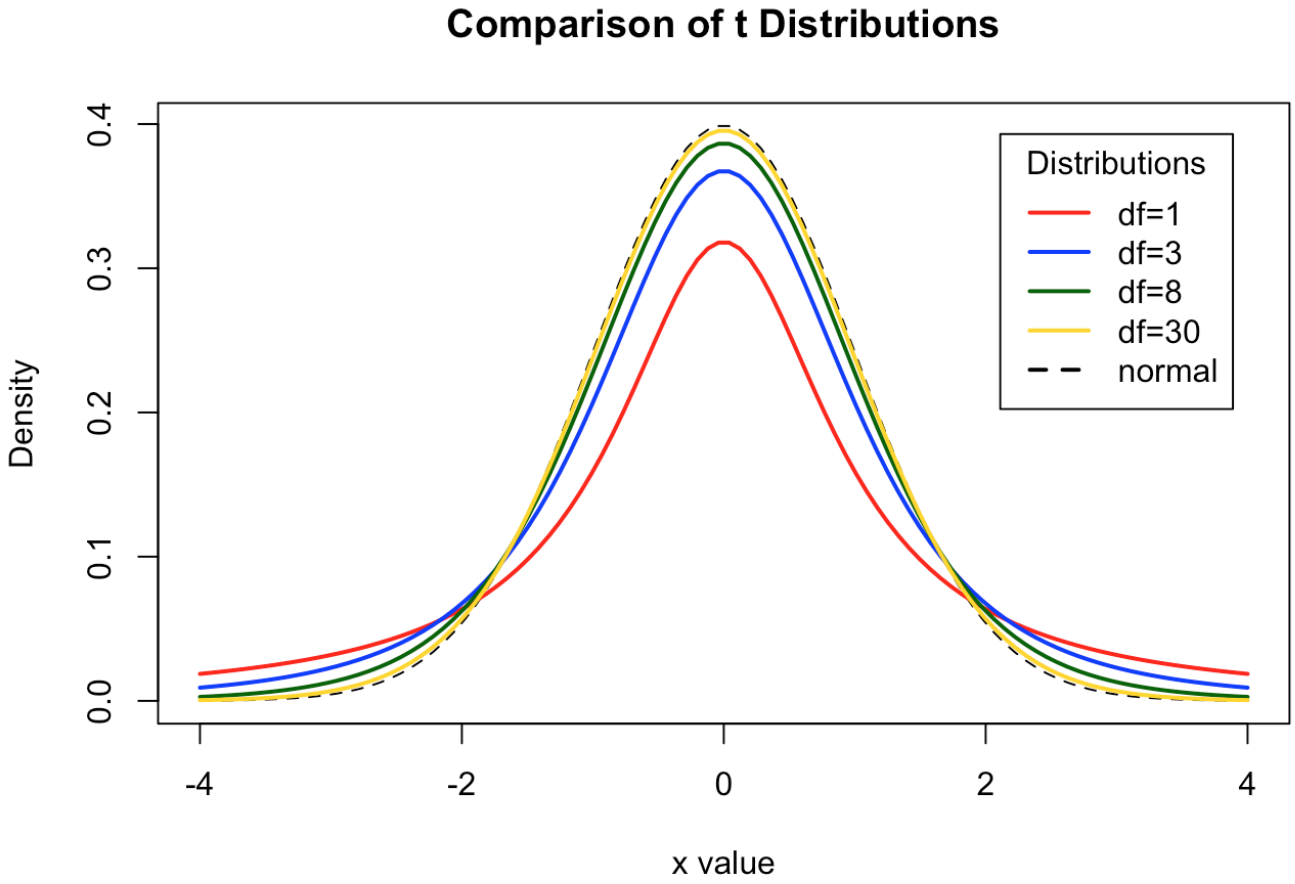
```
# Compare t densities with normal density
x <- seq(-4, 4, length=100)
hx <- dnorm(x) #normal density

degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")

plot(x, hx, type="l", lty=2, xlab="x value",
     ylab="Density", main="Comparison of t Distributions")

for (i in 1:4){
  lines(x, dt(x,degf[i]), lwd=2, col=colors[i]) # dt is the t density
}

legend("topright", inset=.05, title="Distributions",
       labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```



2.9.1 Exercises

1. Suppose that an experimenter obtained a random sample of body weights (kg) from 10 male subjects 66.9, 70.9, 65.8, 78, 71.6, 65.9, 72.4, 73.7, 72.9, 68.5. The distribution of weight in this population is known to be $N(70, \sigma^2)$. What is the probability that the average weight is between 68kg and 71kg?

Answer: The distribution of $\frac{\bar{X}-70}{\frac{S}{\sqrt{10-1}}}$ is t_9, where S is the sample standard deviation. So

$$P\left(68 < \bar{X} < 71\right) = P\left((68-70)/(5.6/\sqrt{9})) < (\bar{X}-70)/(5.6/\sqrt{9}) < (71-70)/(5.6/\sqrt{9})\right)$$

Use R to do the calculations. First put the data into a vector to calculate the standard deviation then use the t_9 CDF:

```
dat <- c(66.9, 70.9, 65.8, 78.0, 71.6, 65.9, 72.4, 73.7, 72.9, 68.5)
sd(dat) # The SD of the weights
```

```
## [1] 3.904186
```

```
a <- (68-70)/(sd(dat)/sqrt(9)); a
```

```
## [1] -1.536812
```

```
b <- (71-70)/(sd(dat)/sqrt(9)); b
```

```
## [1] 0.7684061
```

```
pt(b,df = 9)-pt(a,df = 9)
```

```
## [1] 0.689676
```

So,

$$\begin{aligned} P\left(68 < \bar{X} < 71\right) &= P\left(-1.536812 < t_9 < 0.7684061\right) \\ &= 0.689676 \end{aligned}$$

2.10 F Distribution

Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. The distribution of

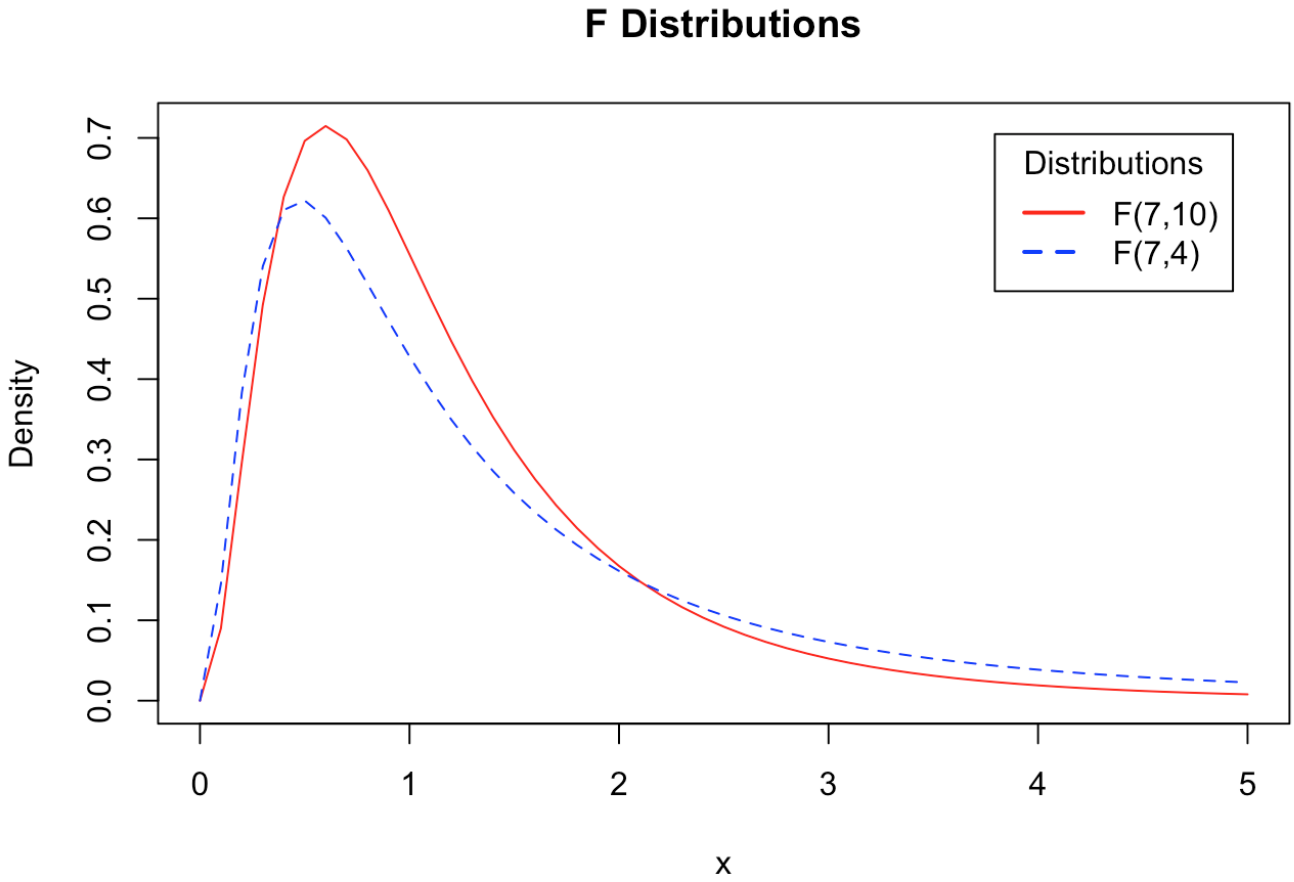
$$W = \frac{X/m}{Y/n} \sim F_{m,n},$$

where $F_{m,n}$ denotes the F distribution on m, n degrees of freedom. The F distribution is right skewed (see graph below). For $n > 2, E(W) = n/(n-2)$. It also follows that the square of a t_n random variable follows an $F_{1,n}$.

The F distribution is right skewed

```
# Compare t densities with normal density

colors <- c("red", "blue")
plot(seq(0,5,by=0.1), df(seq(0,5,by=0.1),7,10),type="l",col=colors[1],yla
b="Density",xlab="x",main = "F Distributions")
lines(seq(0,5,by=0.1), df(seq(0,5,by=0.1),7,4),type="l",lty=2,col=color
s[2])
labels <- c("F(7,10)","F(7,4)")
legend("topright", inset=.05, title="Distributions",
       labels, lwd=2, lty=c(1, 2), col=colors)
```



2.10.1 Exercises

1. Let $W \sim F_{7,10}$. Use R to calculate $P(3 < W < 4)$.

Answer: The CDF of the $F_{m,n}$ distribution in R is `pf()`.

```
pf(4,df1 = 7,df2 = 10)-pf(3,df1 = 7,df2 = 10)
```

```
## [1] 0.03258576
```

So, $P(3 < W < 4) = 0.0325858$.

2.11 Linear Regression

Lea (1965) discussed the relationship between mean annual temperature and mortality index for a type of breast cancer in women taken from regions in Europe (example from Wu and Hammada).

The data is shown below.

```
#Breast Cancer data
M <- c(102.5, 104.5, 100.4, 95.9, 87.0, 95.0, 88.6, 89.2, 78.9, 84.6, 8
1.7, 72.2, 65.1, 68.1, 67.3, 52.5)
T <- c(51.3, 49.9, 50.0,49.2, 48.5, 47.8, 47.3, 45.1, 46.3, 42.1, 44.2, 4
3.5, 42.3, 40.2, 31.8, 34.0)
```

A linear regression model of mortality versus temperature is obtained by estimating the intercept and slope in the equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$. The values of β_0, β_1 that minimize the sum of squares

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2,$$

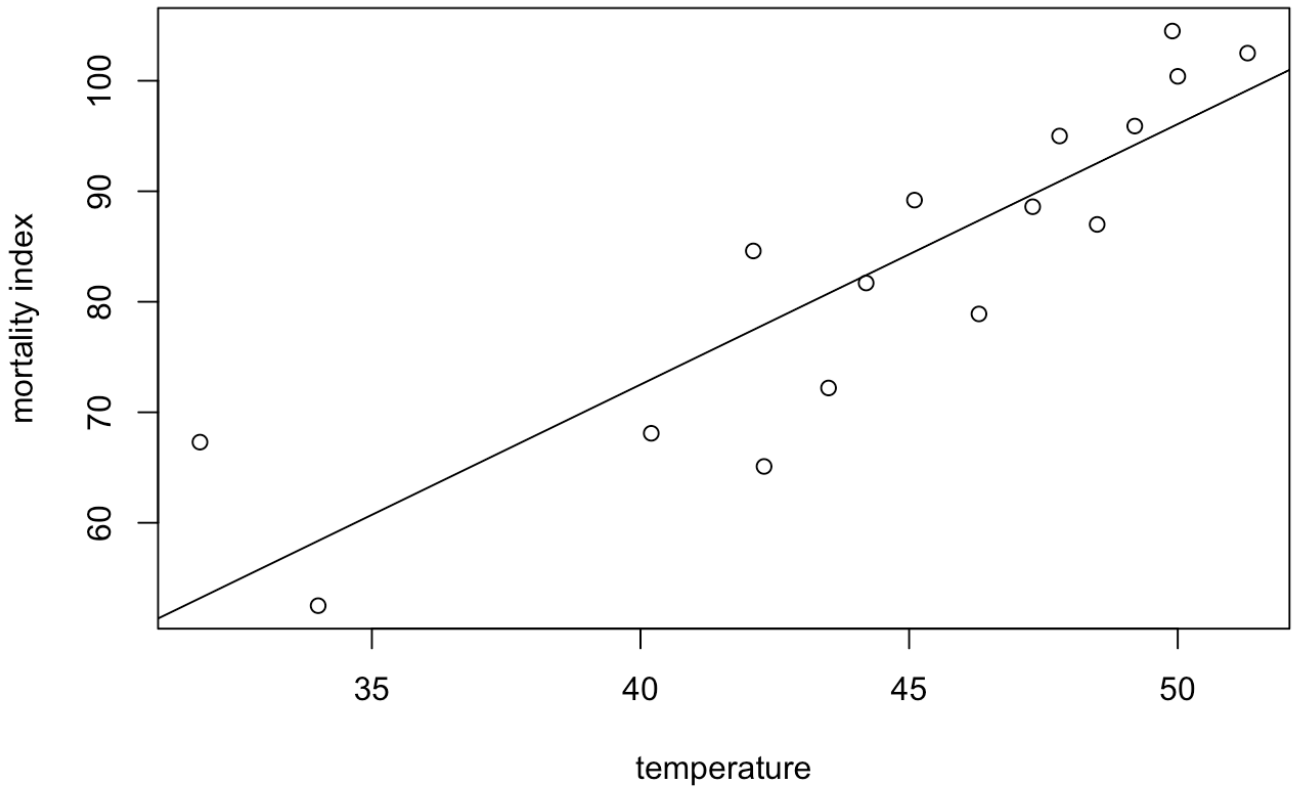
are called the least squares estimators. They are given by $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}, \hat{\beta_1} = r\frac{S_y}{S_x}$. r is the correlation between y and x, and S_x, S_y are the sample standard deviations of x and y respectively.

A scatter plot of the data shows a linear relationship between mortality and temperature. So a regression line is fit to the data.

```
plot(T,M,xlab="temperature",ylab="mortality index")
reg1 <- lm(M~T)
# Parameter estimates and ANOVA table
summary(reg1)
```

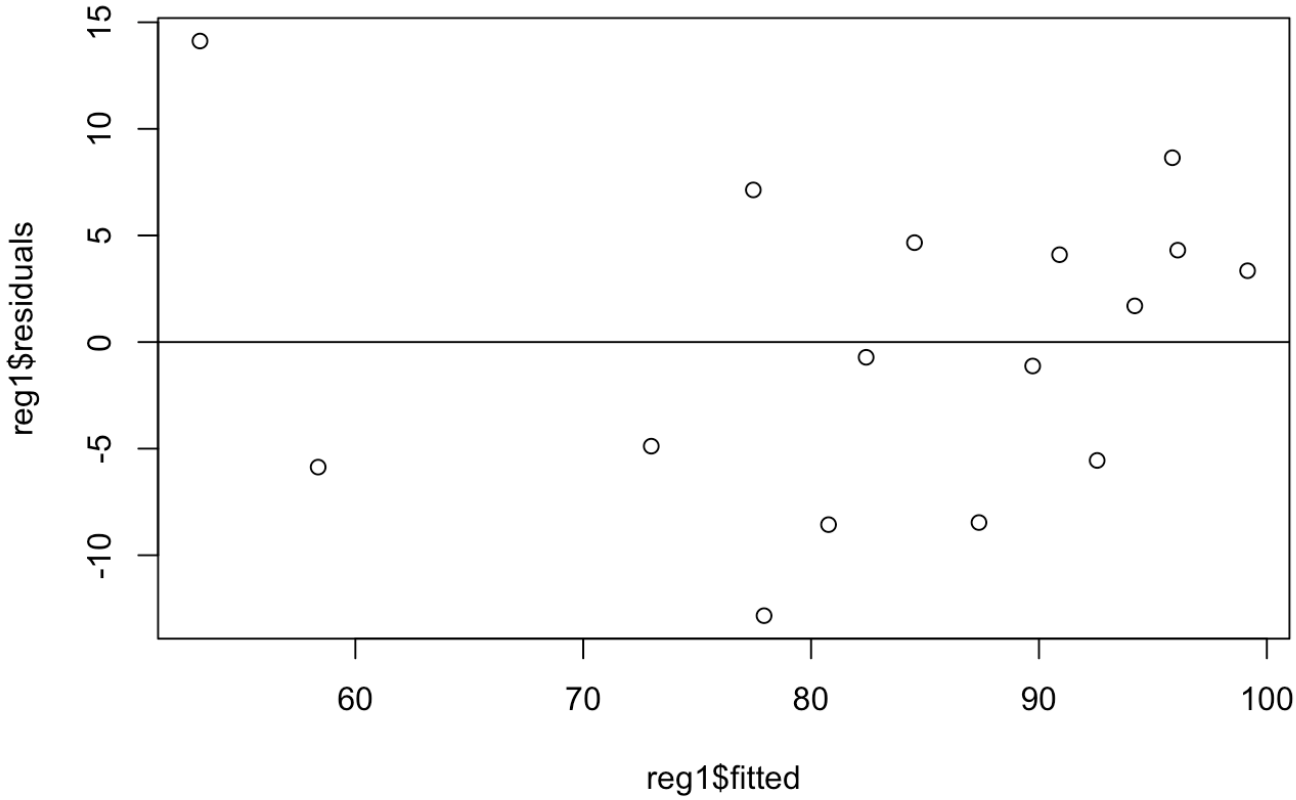
```
##
## Call:
## lm(formula = M ~ T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8358  -5.6319   0.4904   4.3981  14.1200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.7947    15.6719  -1.391    0.186
## T             2.3577     0.3489   6.758  9.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.545 on 14 degrees of freedom
## Multiple R-squared:  0.7654, Adjusted R-squared:  0.7486
## F-statistic: 45.67 on 1 and 14 DF,  p-value: 9.202e-06
```

```
# Add regression line to the plot
abline(reg1)
```

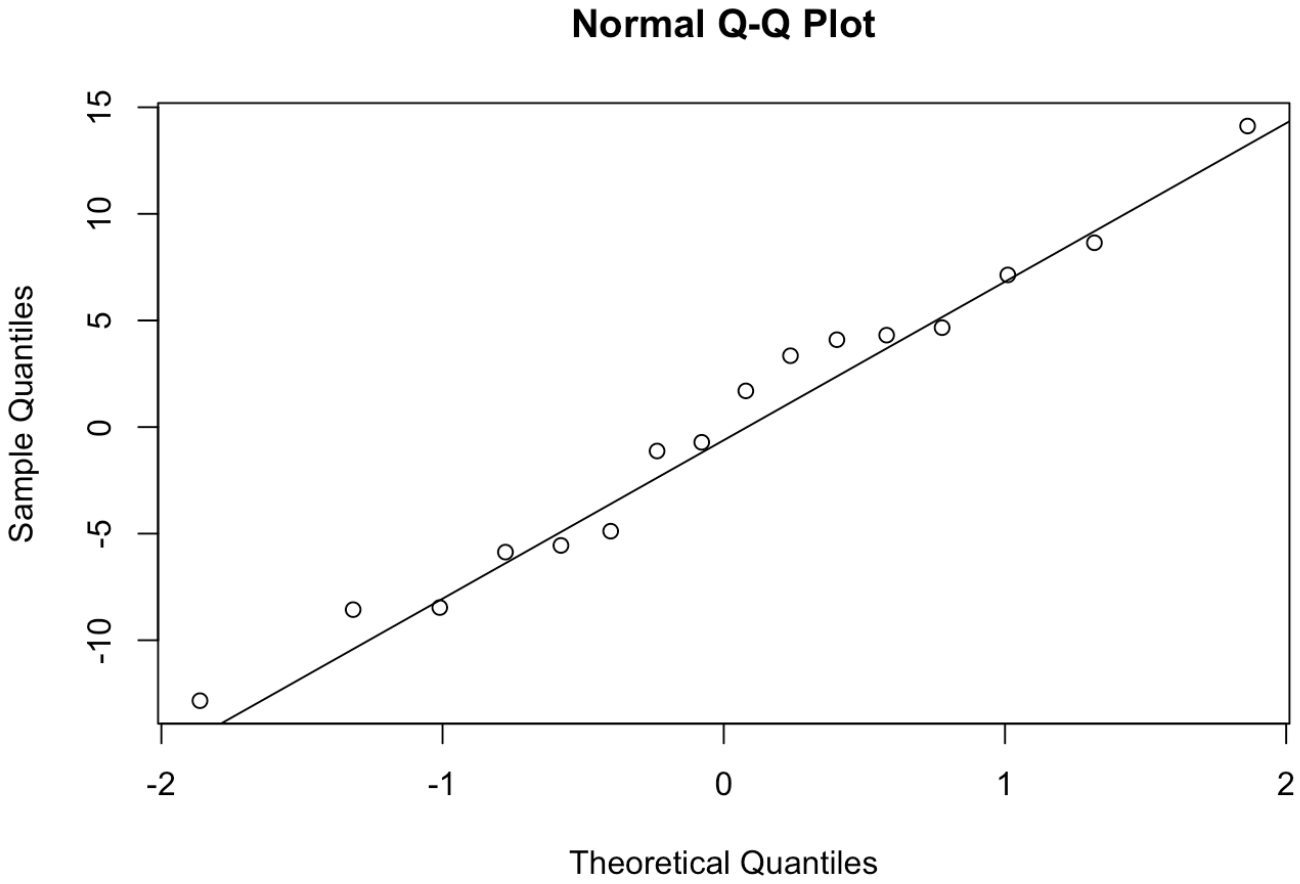


The two main assumptions of constant variance and normality of residuals should always be investigated.

```
#plot residuals vs. fitted
plot(reg1$fitted,reg1$residuals)
abline(h=0) # add horizontal line at 0
```




```
#check normality of residuals
qqnorm(reg1$residuals)
qqline(reg1$residuals)
```



If there is more than one independent variable then the above model is called a multiple linear regression model.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,\ i = 1, \ldots, n,$$

where $\epsilon_i \sim N(0, \sigma^2)$.

This can also be expressed in matrix notation as

$$y = X\beta + \epsilon,$$

where,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & a_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_k \end{pmatrix}$$

The least squares estimator is

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y.$$

The covariance matrix of $\hat{\beta}$ is $\left(X^T X\right)^{-1}\sigma^2$. An estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ is the predicted value of y_i.

2.11.1 Weighing Problem

Harold Hotelling in 1949 wrote a paper on how to obtain more accurate weighings through experimental design.

Suppose that we want to measure the mass of two apples A and B using an old-fashioned two-pan balance scale.

Which of the following two methods produces a more precise estimate of the weights of each apple?

Method 1

Weigh each apple separately.

Method 2

Obtain two weighings by

1. Weighing two apples in one pan.
2. Weighing one apple in one pan and the other apple in the other pan

Let w_1, w_2 be the weights of apples one and two. Each weighing has standard error σ. So the precision of the estimates from method 1 is σ.

If the objects are weighed together in one pan, resulting in measurement m_1, then in opposite pans, resulting in measurement m_2, we have two equations for the unknown weights w_1, w_2:

$$w_1 + w_2 = m_1$$
$$w_1 - w_2 = m_2.$$

This leads to $\hat{w}_1 = (m_1 + m_2)/2$ and $\hat{w}_2 = (m_1 - m_2)/2$. So, $Var\left(\hat{w}_1\right) = Var\left(\hat{w}_2\right) = \sigma^2/2$. The same precision with method 1 would require twice as many measurements.

The moral of the story is that the method used in the design of the experiment has an impact on the precision of the estimates obtained from the experiment.

This can also be viewed as a linear regression problem $y = X\beta + \epsilon$:

$$y = (m_1, m_2)', X = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \beta = (w_1, w_2)'.$$

The least-squares estimates can be found using R.

```
#step-by-step matrix mutiplication example for weighing problem

X <- matrix(c(1,1,1,-1),nrow=2,ncol=2) #define X matrix
Y <- t(X)%*%X # multiply X^T by X (X^T*X) NB: t(X) is the transpose of X
W <- solve(Y) # calculate the inverse
W%*%t(X) # calculate (X^T*X)^(-1)*X^T
```

```
##      [,1] [,2]
## [1,]  0.5  0.5
## [2,]  0.5 -0.5
```

2.12 Randomized Experiments and Observational Studies

According to Rosenbaum (Design of Observational Studies, pg. 6)

> "An observational study is an empiric investigation of effects caused by treatments when randomized experimentation is unethical or infeasible. The quality and strength of evidence provided by an observational study is determined largely by its design."

A technical definition of an observational study is given by Imbens and Rubin (2015)

> The process that determines which *experimental units* receive which *treatments* is called the *assignment mechanisim*. When the *assignment mechanism* is unknown then the design is called an *observational study*.

In randomized experiments (pg. 20, Imbens and Rubin, 2015):

> "… the assignment mechanism is under the control of the experimenter, and the probability of any assignment of treatments across the units in the experiment is entirely knowable before the experiment begins."

Randomized experiments are currently viewed as the most credible basis for determining cause and effect relationships. Health Canada (http://www.hc-sc.gc.ca/index-eng.php), the U.S. Food and Drug Administration (http://www.fda.gov), European Medicines Agency (http://www.ema.europa.eu/ema/), and other regulatory agencies all rely on randomized experiments in their approval processes for pharmaceutical treatments.

2.13 Principles of Experimental Design

2.13.1 Randomization

The primary objective in the design of experiments is the avoidance of bias or systematic error (Cox and Reid, 2005). One way to avoid bias is to use randomization. For example, if researchers want to compare a new medical treatment for a certain cancer to the current treatment then how should the researchers assign patients (experimental units) to the two medical treatments? One method (assignment mechanism) to assign patients to the two treatments is to have the researchers assign patients to treatments. Another method is flip a two-sided coin: if the result of the toss is heads then assign the patient to the new treatment; otherwise assign the patient to the standard treatment. What is the difference between these two methods of assigning patients to treatments? If the researcher decides in a haphazard manner which treatment a patient should receive then the assignment will be subject to the researchers' personal biases. For example, if the researcher does not *a-priori* believe that the new treatment will be better than the standard treatment then she might be reluctant to assign patients who she feels have a poor prognosis to the new treatment. If patients are assigned to the treatments based on the flip of a coin then it's possible to avoid the personal biases introduced by a researcher.

> "It is of the essence that randomization means the use of an objective physical device; it does not mean that allocation is vaguely haphazard or even that it is done in a way that looks effectively random to the investigator." (pg. 19, Cox and Reid, 2000)

2.13.2 Replication

An experiment should be replicated under the exact same experimental conditions so that variation among the replicates can be used to assess random errors that affect treatment comparisons. This type of *genuine replication* (Box, Hunter, and Hunter, 2005) is often not feasible since the experimental setup might be too time-consuming or expensive to replicate. This is a common problem in industrial experimentation.

2.13.3 Blocking

The technique of blocking can be used to control haphazard variation.

> "The central idea behind blocking is an entirely commonsense one of aiming to compare like with like. Using whatever prior knowledge is available about which baseline features of the units and other aspects of the experimental set-up are strongly associated with potential response, we group the units into blocks such that all the units in any one block are likely to give similar responses in the absence of treatment differences. Then, in the simplest case, by allocating one unit in each block to each treatment, treatments are compared on units within the same block." (Cox and Reid, 2005)

Experimental blocks can be created from responses from one subject having blood drawn every 2 hours over a 24 hour period; adjacent agricultural plots of land; or the two feet of the same subject in an exercise experiment.

3 Questions

1. A chemist has seven light objects to weigh on a balance pan scale. The standard deviation of each weighing is denoted by σ.

In a 1935 paper Frank Yates suggested an improved technique by weighing all seven objects together, and also weighing them in groups of three. The groups are chosen so that each object is weighed four times altogether, twice with any other object and twice without it.

Let $y_1, \ldots, y_8$ be the readings from the scale so that the equations for determining the unknown weights, $\beta_1, \ldots, \beta_7$, are

$$
\begin{aligned}
y_1 &= \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7 + \epsilon_1 \\
y_2 &= \beta_1 + \beta_2 + \beta_3 + \epsilon_2 \\
y_3 &= \beta_1 + \beta_4 + \beta_5 + \epsilon_3 \\
y_4 &= \beta_1 + \beta_6 + \beta_7 + \epsilon_4 \\
y_5 &= \beta_2 + \beta_4 + \beta_6 + \epsilon_5 \\
y_6 &= \beta_2 + \beta_5 + \beta_7 + \epsilon_6 \\
y_7 &= \beta_3 + \beta_4 + \beta_7 + \epsilon_7 \\
y_8 &= \beta_3 + \beta_5 + \beta_6 + \epsilon_8,
\end{aligned}
$$

where the $\epsilon_i, i = 1, \ldots, 8$ are independent errors.

In a 1949 paper Harold Hotelling suggested modifying Yates' procedure by placing in the other pan of the scale those of the objects not included in one of his weighings. In other words if the first three objects are to be weighed then the remaining four objects would be placed in the opposite pan.

a. Write Yates' procedure in matrix form $\mathbf{y} = X\beta + \epsilon$, where $\mathbf{y}' = (y_1, \ldots, y_8)$, $\beta' = (\beta_1, \ldots, \beta_7), \epsilon' = (\epsilon_1, \ldots, \epsilon_8)$, and X is an 8×7 matrix. Find the least squares estimate of β. (HINT: use the R code given above to carry out the matrix multiplication)
b. Write Hotellings procedure in matrix form $\mathbf{y} = X\beta + \epsilon$, where $\mathbf{y}' = (y_1, \ldots, y_8)$, $\beta' = (\beta_1, \ldots, \beta_7), \epsilon' = (\epsilon_1, \ldots, \epsilon_8)$, and X is an 8×7 matrix. Find the least squares estimate of β.
c. Find the variance of a weight using Yates' and Hotelling's procedures (you may use known results from regression analysis).
d. If the chemist wanted estimates of the weights with the highest precision then which procedure (Yates or Hotelling) would you recommend that the chemist use to weigh objects? Explain your reasoning.

2. What is the major difference between a randomized experiment and observational study?
3. Show that if $X \sim t_n$ then $X^2 \sim F_{1,n}$.

4 Solutions to Questions

1. a. Recall that for the linear model $y = X\beta + \epsilon$ the least squares estimate of β is $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$.

The following R code generates X and $\left(X^T X\right)^{-1} X^T$

```
X <- rbind( c(1,1,1,1,1,1,1),
            c(1,1,1,0,0,0,0),
            c(1,0,0,1,1,0,0),
            c(1,0,0,0,0,1,1),
            c(0,1,0,1,0,1,0),
            c(0,1,0,0,1,0,1),
            c(0,0,1,1,0,0,1),
            c(0,0,1,0,1,1,0))
X # print X
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    1    1    1    1    1    1
## [2,]    1    1    1    0    0    0    0
## [3,]    1    0    0    1    1    0    0
## [4,]    1    0    0    0    0    1    1
## [5,]    0    1    0    1    0    1    0
## [6,]    0    1    0    0    1    0    1
## [7,]    0    0    1    1    0    0    1
## [8,]    0    0    1    0    1    1    0
```

```
solve( t(X) %*% X ) %*% t(X) #calculate (X'X)^{-1}X'
```

```
##        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
## [1,] 0.0625  0.3125  0.3125  0.3125 -0.1875 -0.1875 -0.1875 -0.1875
## [2,] 0.0625  0.3125 -0.1875 -0.1875  0.3125  0.3125 -0.1875 -0.1875
## [3,] 0.0625  0.3125 -0.1875 -0.1875 -0.1875 -0.1875  0.3125  0.3125
## [4,] 0.0625 -0.1875  0.3125 -0.1875  0.3125 -0.1875  0.3125 -0.1875
## [5,] 0.0625 -0.1875  0.3125 -0.1875 -0.1875  0.3125 -0.1875  0.3125
## [6,] 0.0625 -0.1875 -0.1875  0.3125  0.3125 -0.1875 -0.1875  0.3125
## [7,] 0.0625 -0.1875 -0.1875  0.3125 -0.1875  0.3125  0.3125 -0.1875
```

$\hat{\beta}$ is obtained by multiplying the matrix $\left(X^T X\right)^{-1} X^T$ by the vector $y = (y_1, \ldots, y_8)'$. For example, $\hat{\beta}_1 = (1/16)(y_1 + 5y_2 + 5y_3 + 5y_4 - 3y_5 - 3y_6 - 3y_7 - 3y_8)$.

b. The X matrix for Hotelling's procedure is Yates' X with the zeros replaced by -1:

```
Xh <- ifelse(X == 0,-1,1)
Xh; solve( t(Xh) %*% Xh ) %*% t(Xh)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    1    1    1    1    1    1
## [2,]    1    1    1   -1   -1   -1   -1
## [3,]    1   -1   -1    1    1   -1   -1
## [4,]    1   -1   -1   -1   -1    1    1
## [5,]   -1    1   -1    1   -1    1   -1
## [6,]   -1    1   -1   -1    1   -1    1
## [7,]   -1   -1    1    1   -1   -1    1
## [8,]   -1   -1    1   -1    1    1   -1
```

```
##       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
## [1,] 0.125  0.125  0.125  0.125 -0.125 -0.125 -0.125 -0.125
## [2,] 0.125  0.125 -0.125 -0.125  0.125  0.125 -0.125 -0.125
## [3,] 0.125  0.125 -0.125 -0.125 -0.125 -0.125  0.125  0.125
## [4,] 0.125 -0.125  0.125 -0.125  0.125 -0.125  0.125 -0.125
## [5,] 0.125 -0.125  0.125 -0.125 -0.125  0.125 -0.125  0.125
## [6,] 0.125 -0.125 -0.125  0.125  0.125 -0.125 -0.125  0.125
## [7,] 0.125 -0.125 -0.125  0.125 -0.125  0.125  0.125 -0.125
```

To calculate $\hat{\beta}$ multiply the matrix $\left(X^TX\right)^{-1}X^T$ by the vector $y = (y_1, \ldots, y_8)'$. For example $\hat{\beta}_1 = (1/8)(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8)$.

c. The covariance matrix of $\hat{\beta}$ is $\left(X^TX\right)^{-1}\sigma^2$.

```
#Yates
solve( t(X) %*% X )
```

```
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,]  0.4375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
## [2,] -0.0625  0.4375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
## [3,] -0.0625 -0.0625  0.4375 -0.0625 -0.0625 -0.0625 -0.0625
## [4,] -0.0625 -0.0625 -0.0625  0.4375 -0.0625 -0.0625 -0.0625
## [5,] -0.0625 -0.0625 -0.0625 -0.0625  0.4375 -0.0625 -0.0625
## [6,] -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.4375 -0.0625
## [7,] -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.4375
```

```
#Hotelling
solve( t(Xh) %*% Xh )
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,] 0.125 0.000 0.000 0.000 0.000 0.000 0.000
## [2,] 0.000 0.125 0.000 0.000 0.000 0.000 0.000
## [3,] 0.000 0.000 0.125 0.000 0.000 0.000 0.000
## [4,] 0.000 0.000 0.000 0.125 0.000 0.000 0.000
## [5,] 0.000 0.000 0.000 0.000 0.125 0.000 0.000
## [6,] 0.000 0.000 0.000 0.000 0.000 0.125 0.000
## [7,] 0.000 0.000 0.000 0.000 0.000 0.000 0.125
```

The variance of an estimated weight is the diagonal element of the matrix multiplied by σ^2.

d. Pick the procedure with the lowest variance- Hotelling.

2. The major difference between a randomized experiment and an observational study is the assignment mechanism. The functional form of the assignment mechanism in an observational study is unknown, but in a randomized study it is known.

3. If $X \sim t_n$ then $X = Y/(\sqrt{W/n})$, where $Y \sim N(0,1)$ and $W \sim \chi^2_n$. So $X^2 = (Y/(\sqrt{W/n}))^2 = Y^2/(W/n)$, and $Y^2 \sim \chi^2_1$. Therefore, $X^2 \sim F_{1,n}$.