STA305/1004 - Review of Statistical Theory

September 10, 2019

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 5 observations
# from a N(60,10^2)
dat <- round(rnorm(5,mean = 60,sd = 10),1)
dat
```

```
## [1] 55.0 61.3 59.2 68.9 61.2
```

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.
##   55.00   59.20   61.20   61.12   61.30   68.90
```

Smallest Value

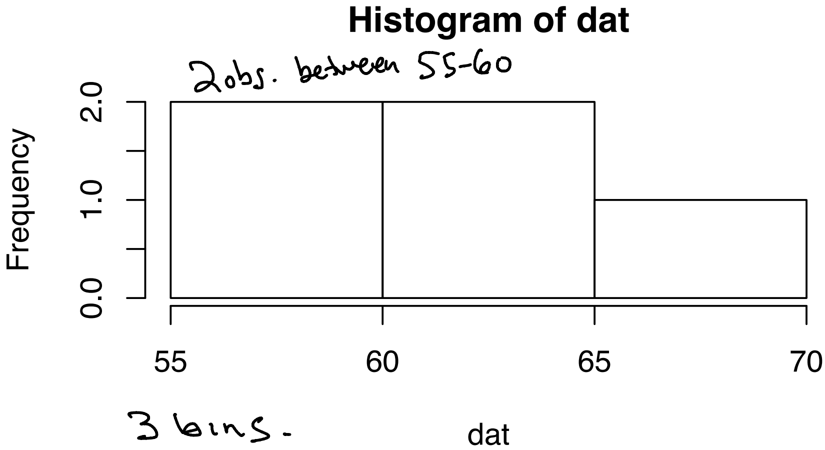
25th percentile.

50th percentile

75th percentile

Distributions

```r
hist(dat)
```



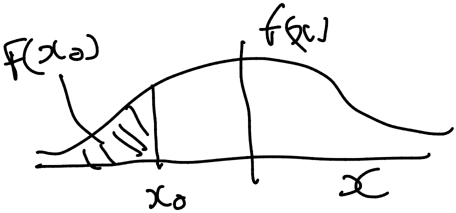
Distributions

- 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 a **sample** from the population.

Continuous Distributions

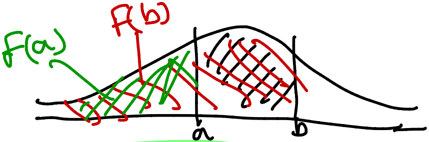
- A continuous random variable X is fully characterized by it's density function $f(x)$.
- $f(x) \geq 0$, f is piecewise continuous, and $\int_{-\infty}^{\infty} f(x)dx = 1$.
- The cumulative distribution function (CDF) of X is defined as:

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



$1 - F(x_0)$

Continuous Distributions



- 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:

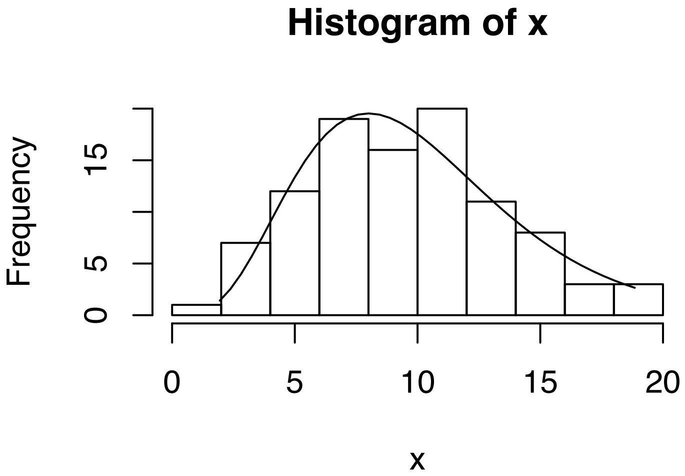
$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a).$$

- In R a list of all the common distributions can be obtained by the command `help("distributions")`.
- For example, the normal density and CDF are given by `dnorm()` and `pnorm()`.

cdf.

density.

100 observations (using `rchisq()`) 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.



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.

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 and the quantity of interest, y, is age.
- A sample of size n is drawn from this population.
- The population mean is $\mu = \sum_{i=1}^{N} y_i / N$. — fixed
- The sample mean is $\bar{y} = \sum_{i=1}^{n} y_i / n$.

Residuals and Degress of Freedom

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y} = \sum y_i - n\bar{y}$$
$$= n \cdot \bar{y} - n\bar{y} = 0$$

$y_i - \bar{y}$ 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.

The Normal Distribution

Sd σ σ^2 = variance.

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

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

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

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

$$\parallel$$

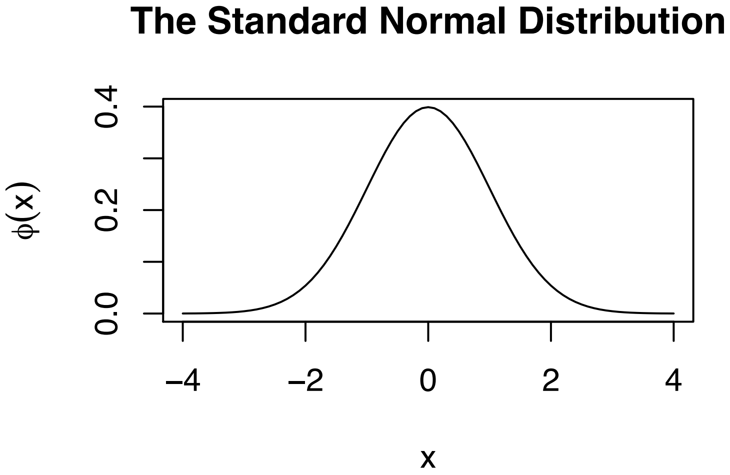
$$\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2} x^2\right)$$

$\mu = 0$ $\sigma^2 = 1$.

The Normal Distribution

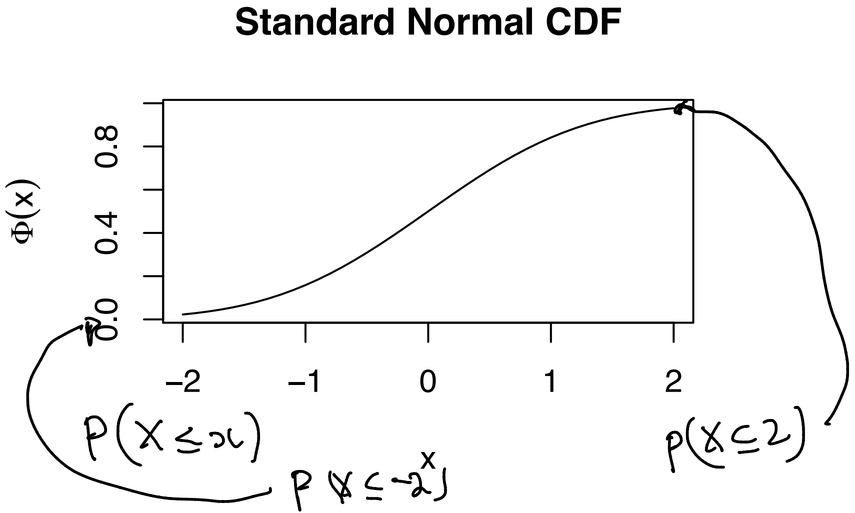
normal density function.

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



The Normal Distribution

```
plot(x <- seq(-2,2,by=0.1),pnorm(x),type="l",
     xlab="x",ylab=expression(paste(Phi(x))),
     main = "Standard Normal CDF")
```



The Normal Distribution

A random variable X that follows a normal distribution with mean μ and variance σ^2 will be denoted by

mean

$$X \sim N\left(\mu, \sigma^2\right).$$

variance.

If $Y \sim N\left(\mu, \sigma^2\right)$ then

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

where

$$Z = \frac{Y - \mu}{\sigma}.$$

location

$$\Rightarrow \quad Y = \mu + \sigma Z, \quad Z \sim N(0,1)$$

scale.

$$Y \sim N(\mu, \sigma^2)$$

The Normal Distribution

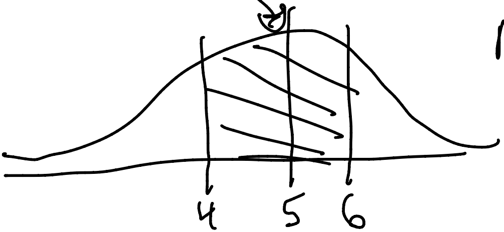
mean variance

$X \sim N(5, 3)$. Use R to find $P(4 < X < 6)$.

```
pnorm(6,mean = 5,sd = sqrt(3))-pnorm(4,mean = 5,sd = sqrt(3))
```

```
## [1] 0.4362971
```

Cumulative dist. function



$N(5,3)$

Normal Quantile Plots

The following data 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
```

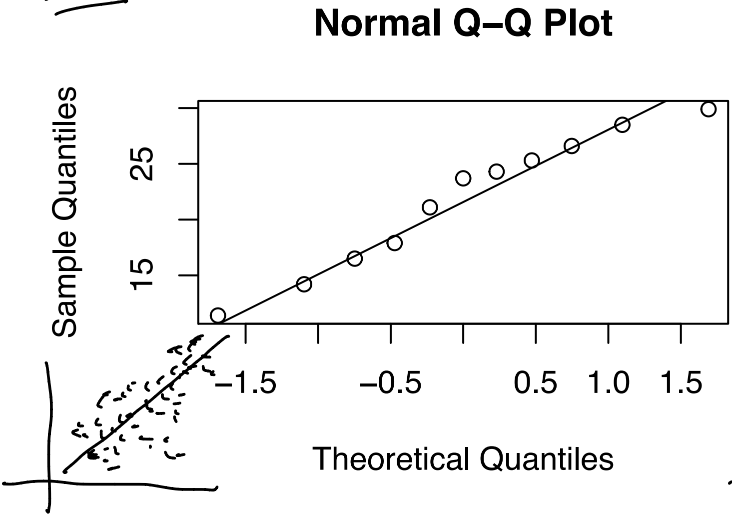
Do the weights follow a Normal distribution?

Normal Quantile Plots

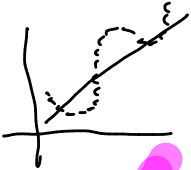
these plots indicate a systematic deviation from straight line.

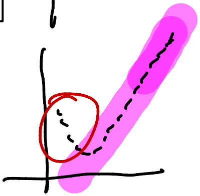
A normal quantile plot in R can be obtained using qqnorm() for the normal probability plot and qqline() to add the straight line.

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



Systematic deviation indicates non-normality.





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 $N\left(\mu, \frac{\sigma^2}{n}\right)$.

Central Limit Theorem

$P(X_i=1)=0.5$

$X_i = \begin{cases} 1 & \text{if } H \\ 0 & \text{if } T \end{cases}$

$\frac{\sum_{i=1}^{50} X_i}{50}$

$Var(X_i) = p(1-p)$

☑ Binomial
☐ Bernoulli
☐ t
☐ Normal

$Bin(n=50, p=0.5)$

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

$$E\left(\frac{\sum_{i=1}^{50} X_i}{50}\right) = \frac{1}{50}\sum_{i=1}^{50} E(X_i) = \frac{1}{50} \times 50 \times .5$$

$$= \frac{25}{50} = \frac{1}{2}.$$

$$Var\left(\frac{\sum_{i=1}^{50} X_i}{50}\right) = \frac{1}{50^2} Var(\sum X_i)$$

$$= \frac{1}{50^2} \sum Var(X_i) = \frac{1}{50^2}\, 50 \times .5 \times .5$$

$$\hat{p} = \frac{\sum x_i}{50} \sim N\left(0.5, \frac{0.5 \times .5}{50}\right)$$

$$\shortparallel$$

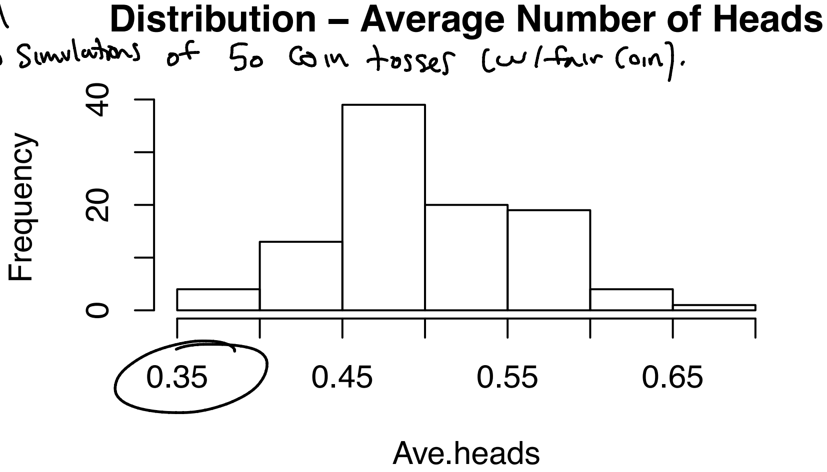
proportion of heads in 50 tosses of a fair coin.

Central Limit Theorem

of Simulations

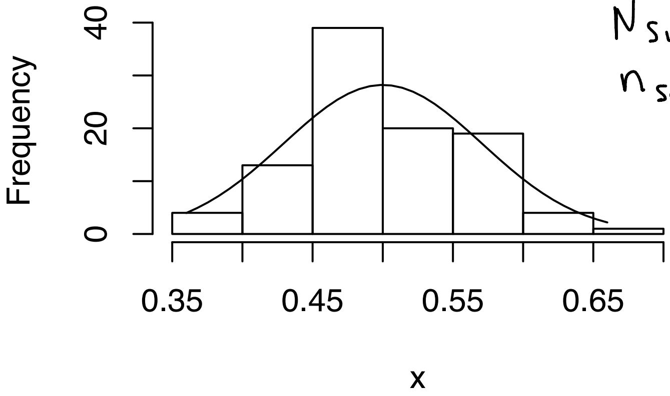
```
set.seed(100)
Total.heads <- rbinom(100,50,0.5); Ave.heads <- Total.heads/50;
hist(Ave.heads, main =  "Distribution - Average Number of Heads")
```

100 Simulations of 50 Coin tosses (w/ fair Coin).



Central Limit Theorem

```
set.seed(100)
x<- rbinom(100,50,0.5)/50 # draw a sample of 100 from bin(50,.5)
h <- hist(x, main = "", ) # create the histogram
# superimpoise normal density over histogram
xfit<-seq(min(x),max(x),length=40)
yfit <- dnorm(xfit,mean = .5,sd = sqrt((.5*.5)/50))
yfit <- yfit*diff(h$mids[1:2])*length(x)
lines(xfit,yfit)
```



$N_{sim} = 100$

$n_{\text{sample size}} = 50$

Chi-Square Distribution

Let $X_1, X_2, ..., X_n$ 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_n^2.

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

Chi-Square Distribution

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

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

distribution of Sample Variance.

t Distribution

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

t Distribution

One-Sample t-test

$H_0: \mu = \mu_1$

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

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

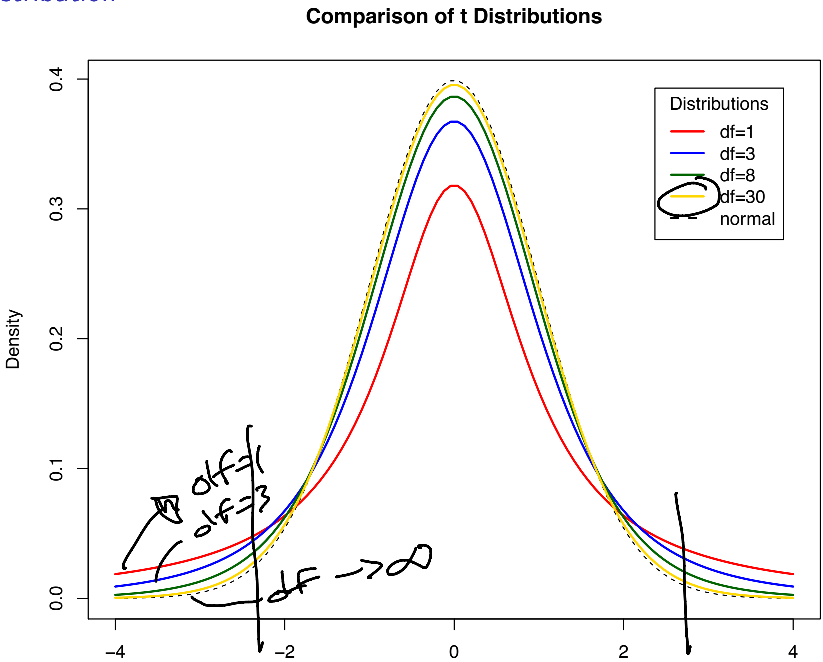
t test Statistic.

Sample sd.

where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$?

$t_{(n-1)}$

t Distribution



F Distribution

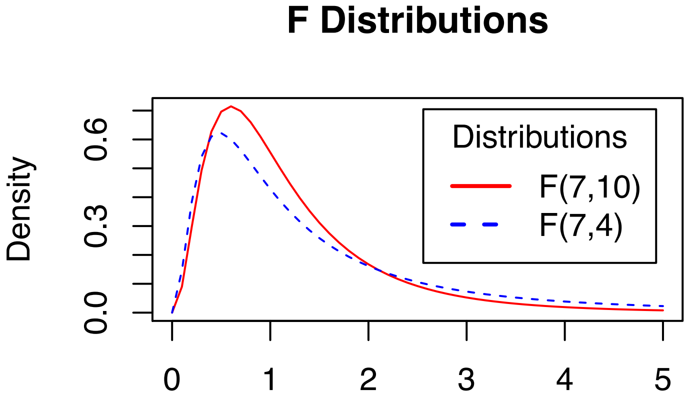
Let $X \sim \chi^2_m$ and $Y \sim \chi^2_n$ be independent. The distribution of

numerator df

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

denominator df

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 $F_{1,n}$.



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, 81.7, 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, 43.5, 42.3, 40.2, 31.8, 34.0)
```

Linear Regression

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

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

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

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

$$\frac{\partial L}{\partial \beta_0} = 0 \qquad \frac{\partial L}{\partial \beta_1} = 0$$

Solve for $\hat{\beta}_0$ $\hat{\beta}_1$

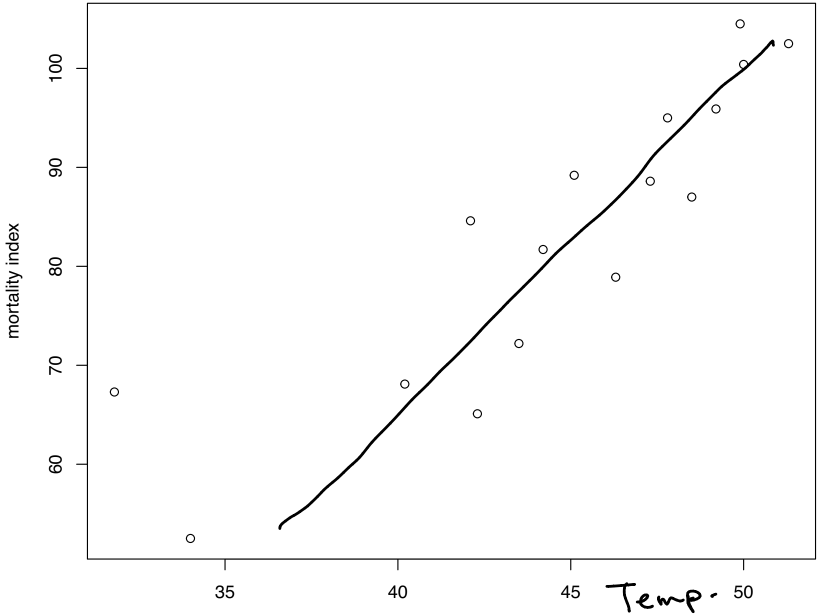
are called the least squares estimators. They are given by:

- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = r \frac{S_y}{S_x}$

r is the correlation between y and x, and S_x, S_y are the sample standard deviations of x and y respectively.

Linear Regression

```
plot(T,M,xlab="temperature",ylab="mortality index")
```



Linear Regression

Regression of T on M

```
reg1 <- lm(M~T)
summary(reg1) # Parameter estimates and ANOVA table
```

```
##
## 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
```

$y \sim x$

dependent ~ independent variables

$y \sim x_1 + x_2$

$\hat{\beta}_0$

$\hat{\beta}_1$

$H_0: \beta_0 = 0$

$t = -1.391$

$H_0: \beta_1 = 0$

% of variation in mortality explained by reg model

$$\left.\begin{aligned}\hat{\beta}_0 &= -21.7947 \\ \hat{\beta}_1 &= 2.3577\end{aligned}\right\}$$ estimates of intercept and slope.

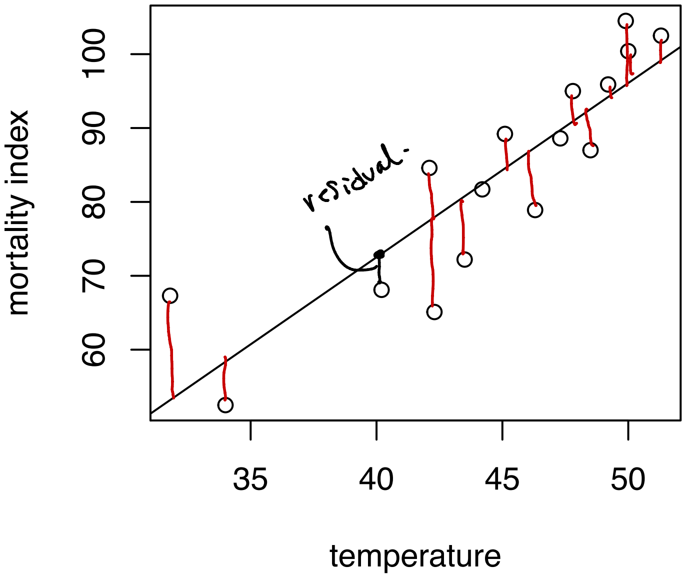
$R^2 = .7654$ $\therefore$ approx. 77% of the variation in mortality is explained by the regression model of mortality and temp.

$$\hat{y} = -21.7947 + 2.3577\,T$$

obtain fitted values by plugging in T values.

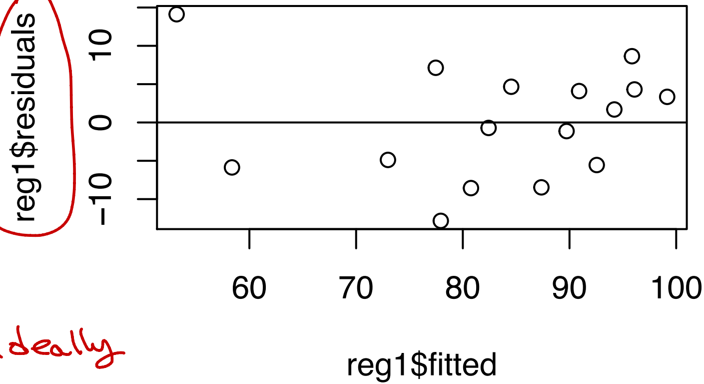
Linear Regression

```
plot(T,M,xlab="temperature",ylab="mortality index")
abline(reg1) # Add regression line to the plot
```



If there are no systematic patterns then little evidence of poor fit.

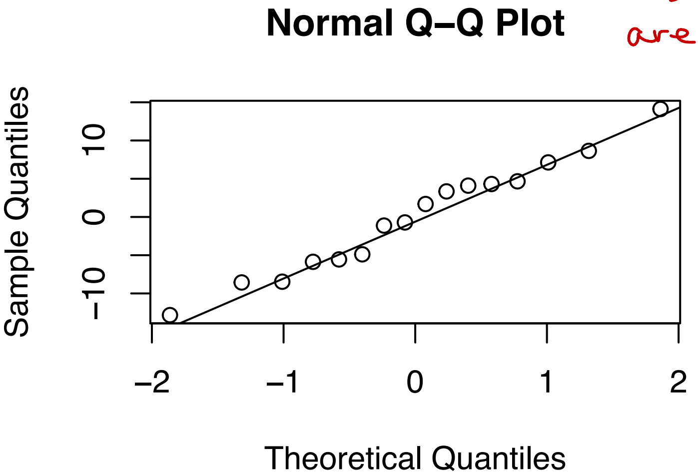
```
#plot residuals vs. fitted
plot(reg1$fitted,reg1$residuals);
abline(h=0) # add horizontal line at 0
```



Ideally have a random scatter of points.

p-values are valid provided residuals are normally distributed.

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



Linear Regression

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

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

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

This can also be expressed in matrix notation as

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

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ \vdots & & & \\ 1 & x_{n1} & & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The least squares estimator is

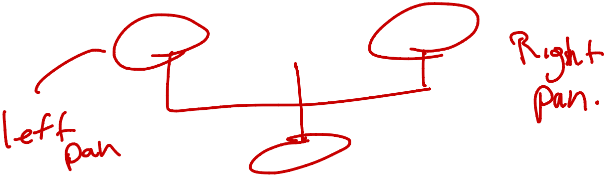
$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y.$$

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.

Weighing Problem



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

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

Weighing Problem

$Var(w_1) = \sigma^2$

This illustrates that experimental design can impact the precision of the estimates obtained.

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:

$m_1 + m_2 = 2w_1$

$$\left.\begin{aligned} w_1 + w_2 &= m_1 \\ w_1 - w_2 &= m_2 \end{aligned}\right\} \Rightarrow \hat{w}_1 = \left(\frac{m_1 + m_2}{2}\right)$$

$m_1 - m_2 = 2w_2$

$\Rightarrow \hat{w}_2 = \frac{m_1 - m_2}{2}$

$$Var(\hat{w}_1) = Var\left(\frac{m_1 + m_2}{2}\right)$$

$$= \frac{1}{4}(\sigma^2 + \sigma^2)$$

$$= \frac{2\sigma^2}{4} = \sigma^2/2 = Var(\hat{w}_2)$$

Weighing Problem

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

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

$$x_{i1} = \begin{cases} 1 & \text{if measurement is in Left pan} \\ -1 & \text{if measurement in Right pan.} \end{cases}$$

Weighing Problem

we could use lm( )

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 transpose of X
W <- solve(Y) # calculate the inverse
W %*% t(X) # calculate (X^T*X)^(-1)*X^T
```

transpose

%*% matrix multiplication.

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

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \hat{\omega}_1 \\ \hat{\omega}_2 \end{pmatrix}$$

```
W # print (X^T*X)^(-1) for SE
```

```
##      [,1] [,2]
## [1,]  0.5  0.0
## [2,]  0.0  0.5
```

$$\begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \sigma^2 = \begin{bmatrix} s.e(\hat{\omega}_1) \\ s.e(\hat{\omega}_2) \end{bmatrix}$$