STA305/1004 - Review of Statistical Theory

+ introduction to using R

January 13, 2016

HW #1 posted: due on Jan 26th @ 22:00

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
```

generate a random sample of n=5
N(μ=100, σ=10)

Distributions can be displayed graphically or numerically.

A histogram is a graphical summary of a data set.

```
summary(dat)
```

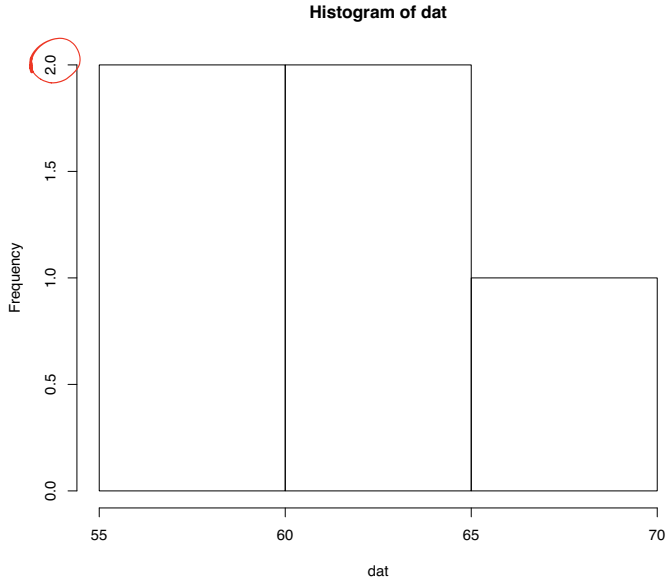
R Code

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   59.20   61.20   61.12   61.30   68.90
```

5 number summary + mean

Distributions

```
hist(dat)
```



2 + 2 + 1 = 5

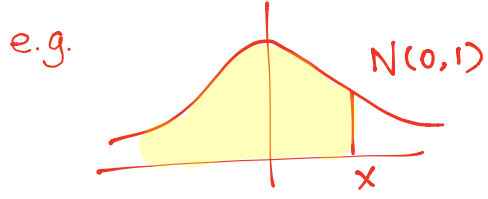
Distributions

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 some kind of 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:

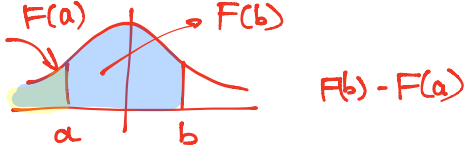
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx.$$



- 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:

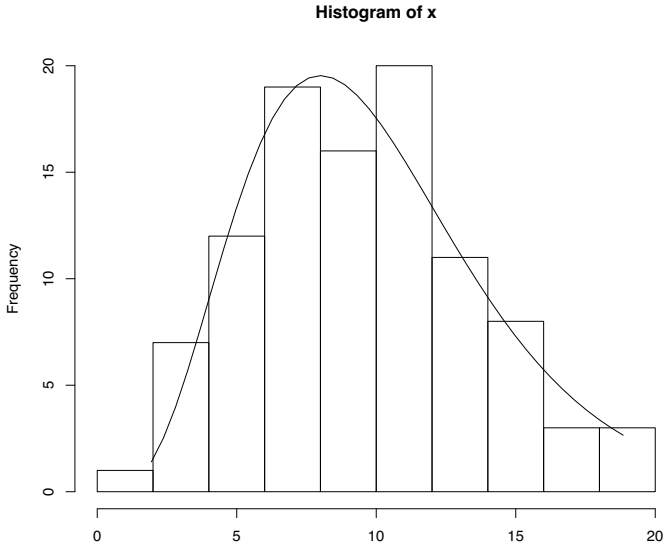
$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a).$$

- 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()`.



Continuous Distributions

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 $\mu = \sum_{i=1}^{N} y_i/N$. ← parameter
- The sample mean is $\bar{y} = \sum_{i=1}^{n} y_i/n$. ← statistic

Residuals and Degress of Freedom

n=10, 9 values then 10th is determined by

$y_i - \bar{y}$ 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

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

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

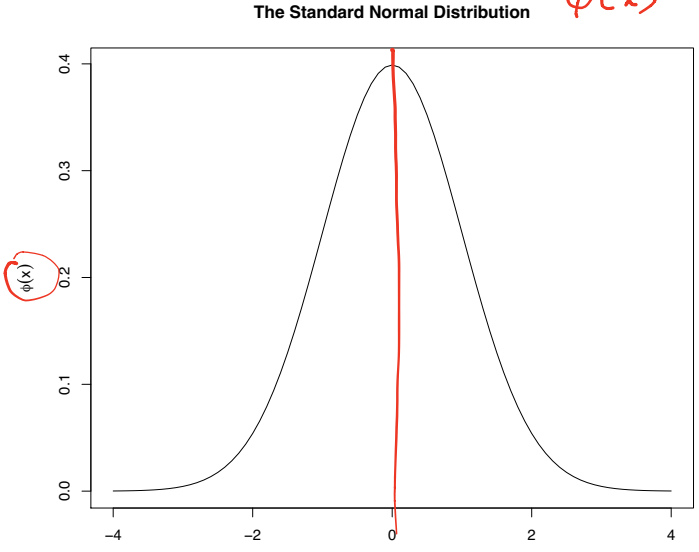
The cumulative distribution function (CDF) of a $N(0,1)$ distribution,

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

Not available in closed form
so use R

The Normal Distribution

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

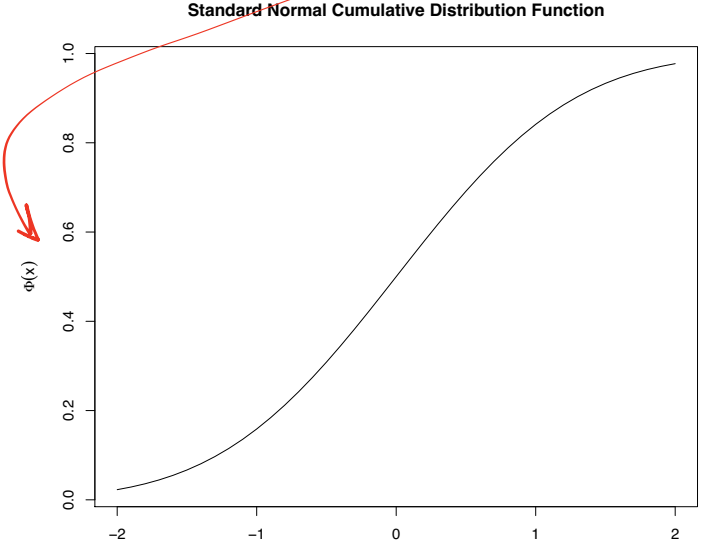


The Normal Distribution

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

CDF of N(0,1)

letter lower case "l"



The Normal Distribution

$Z = \frac{Y-\mu}{\sigma} \Rightarrow Y = \mu + \sigma Z$

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

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

$E(Z) = E\left(\frac{Y-\mu}{\sigma}\right) = \frac{E(Y)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$

$Var(Z) = Var\left(\frac{Y-\mu}{\sigma}\right) = \frac{1}{\sigma^2} Var(Y) = \sigma^2/\sigma^2 = 1$

↘ variance of Y-constant = variance of Y

The Normal Distribution

$$P\left(\frac{4-5}{\sqrt{3}} < \frac{X-5}{\sqrt{3}} < \frac{6-5}{\sqrt{3}}\right)$$

$X \sim N(5, 3)$. 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
```

$\Phi(6) - \Phi(4)$ ✗

CDF of $N(0,1)$

$\Phi\left(\frac{6-5}{\sqrt{3}}\right) - \Phi\left(\frac{4-5}{\sqrt{3}}\right)$ ✓

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

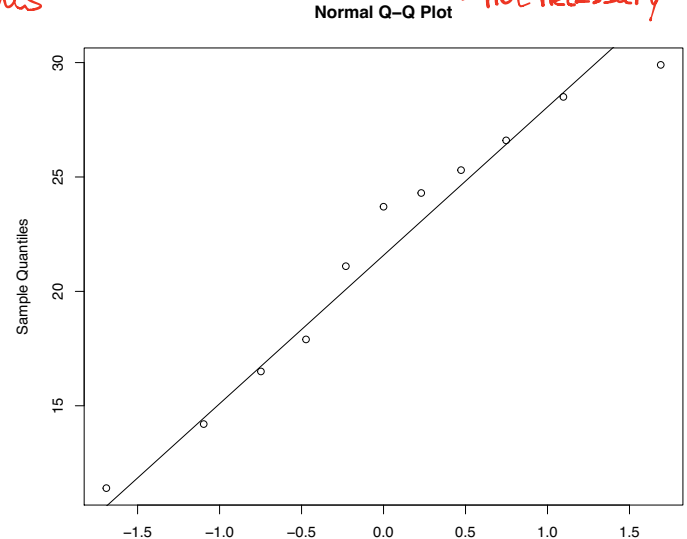
A normal quantile plot in R can be obtained using `qqnorm()` for the normal probability plot and `qqline()` to add the straight line.

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

add straight line

plot points

not necessary if run on separate lines



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

$$\bar{X} \overset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\text{Var}$$

Central Limit Theorem

$$\text{Total \# Hs} \sim \text{Bin}(50, 0.5)$$

$$\frac{\text{Total \# Hs}}{50} = \text{Ave \# Hs} = \frac{X_1 + X_2 + \cdots + X_{50}}{50} = \bar{X}$$

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

$E(X_i) = 0.5$

$\text{Var}(X_i) = 0.5 \times 0.5$

$X_i = \begin{cases} 1 & \text{Heads} \\ 0 & \text{Tails} \end{cases}$

$\bar{X} \overset{\text{approx}}{\sim} N\left(0.5, \frac{0.5 \times 0.5}{50}\right)$

$\bar{X} =$ prop of Hs

$P(X_i = 1) = 0.5$

X_i = ith coin toss, $i = 1, \cdots, 50$

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

$P(X_i = 1) = 0.5$

$X_1, X_2, \cdots, X_{50}$

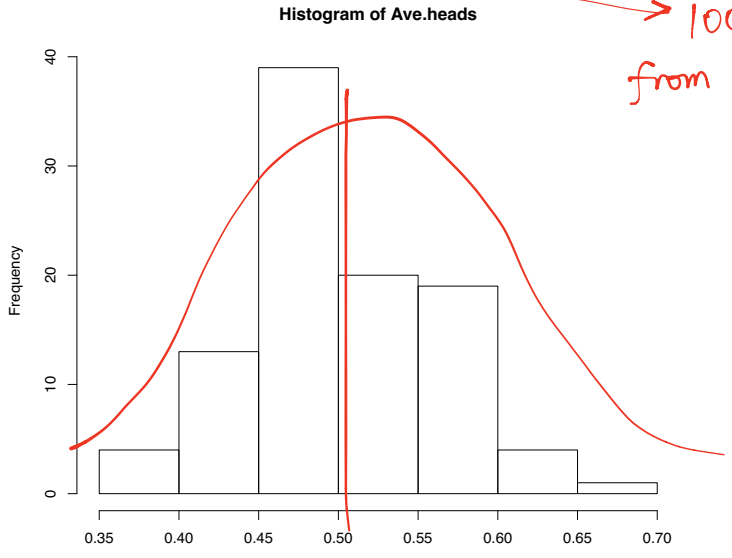
statistics of interest is $\underbrace{\frac{X_1 + X_2 + \cdots + X_{50}}{50}}_{\bar{X}}$

$X_1 + X_2 + \cdots + X_{50} \sim Bin(n = 50, p = 0.5)$

sample prop $\rightarrow \hat{p}$

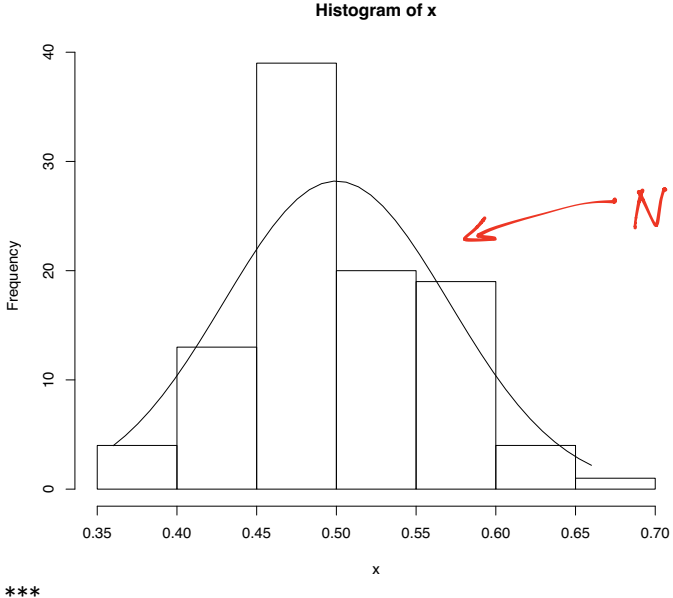
Central Limit Theorem

```
set.seed(100)
Total.heads <- rbinom(100,50,0.5); Ave.heads <- Total.heads/50;
hist(Ave.heads)
```



100 samples from Bin(50, 0.5)

Central Limit Theorem



$N(0.5, \frac{0.5 \times 0.5}{50})$

***

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^2_n.

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

Chi-Square Distribution

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

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

Why?

$$\sum\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-\mu)^2 \qquad \sum(X_i-\bar{X})=0$$

$$\frac{1}{\sigma^2}\sum(X_i-\bar{X})^2 + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \rightarrow (N(0,1))^2$$

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

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 $\frac{X}{\sqrt{W/n}} \sim t_n$.

t Distribution

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

Correct: $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n-1}}}$ mistake

t-test

$H_0: \mu = 0$

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

Why? ⟶ next page

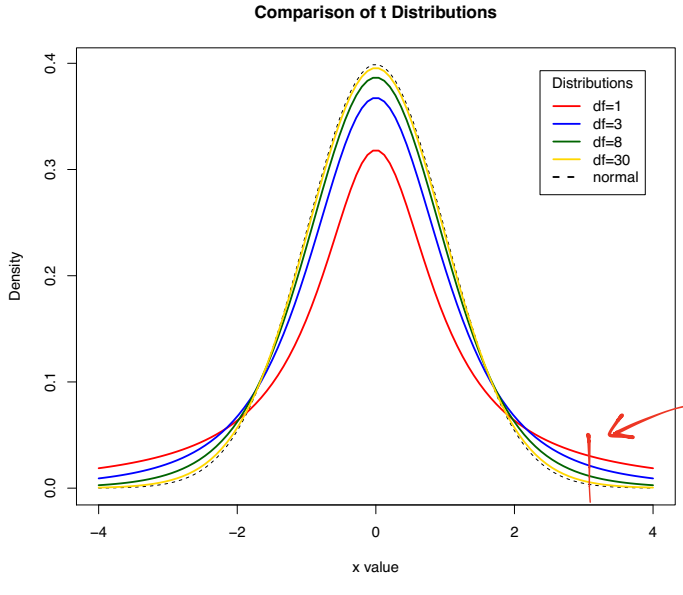
$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{s/\sqrt{n}/\sigma/\sqrt{n}}$$

$$= \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{s/\sigma} = \frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{s^2/\sigma^2}}$$

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ and } \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{(n-1)}}{n-1}$$

$$\therefore \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

t Distribution



Heavier tails than normal for df = small

F Distribution

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

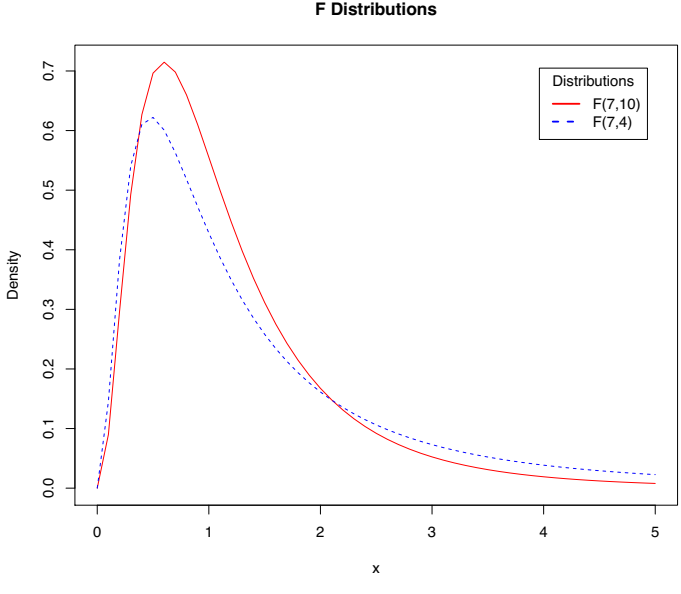
$$W = \frac{X/m}{Y/n} \sim F_{m,n},$$

numerator df

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 $F_{1,n}$.

F Distribution



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)
```

create a vector called M

$\hat{y}(34) - 52.5 = \text{residual}$ $\hat{y} = -21 + 2.35T$,

Linear Regression

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

$\hat{y}_{int}$ slope

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

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

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

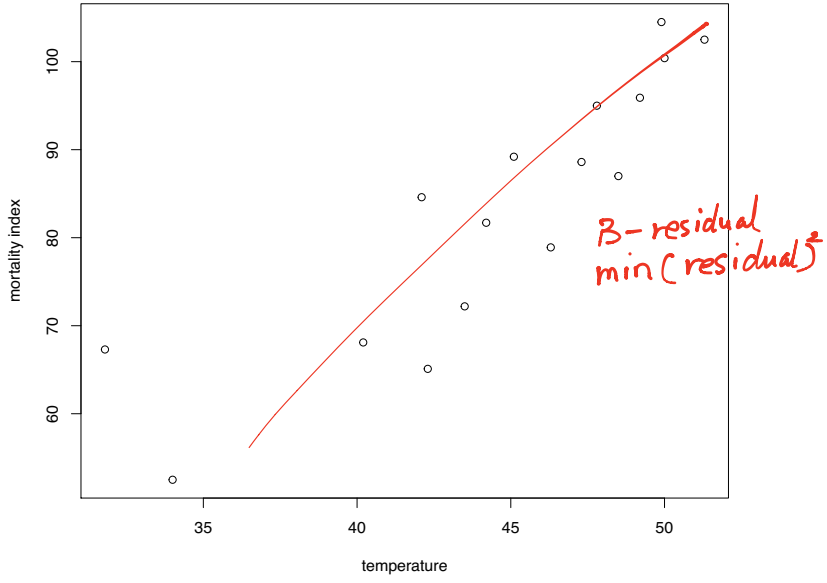
are called the least squares estimators. They are given by:

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

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

Linear Regression

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



***

Linear Regression

```
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
```

lm(·) $\hat{y}-y_i$ linear model

$\hat{\beta}_0$

$\hat{\beta}_1$

$H_0: \beta_0 = 0$

$H_0: \beta_1 = 0$

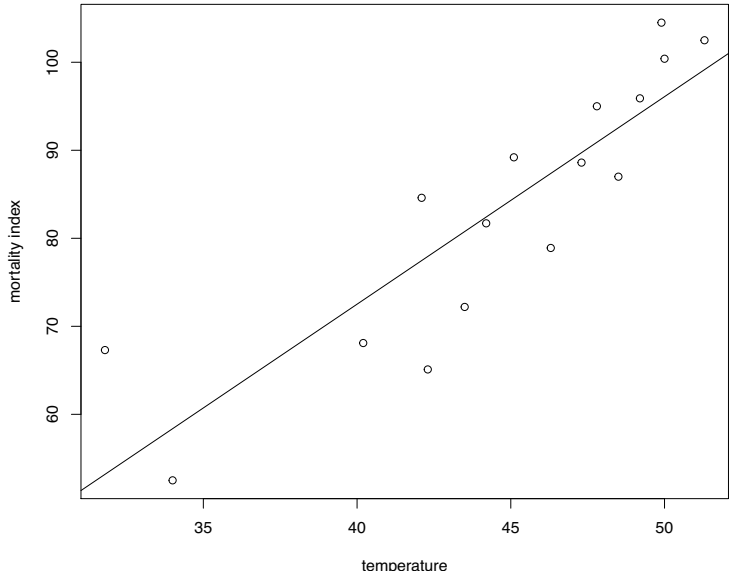
p-value

percent of variability in mortality explained by linear reg.

$\hat{y} = -21.79 + 2.35T$

Linear Regression

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

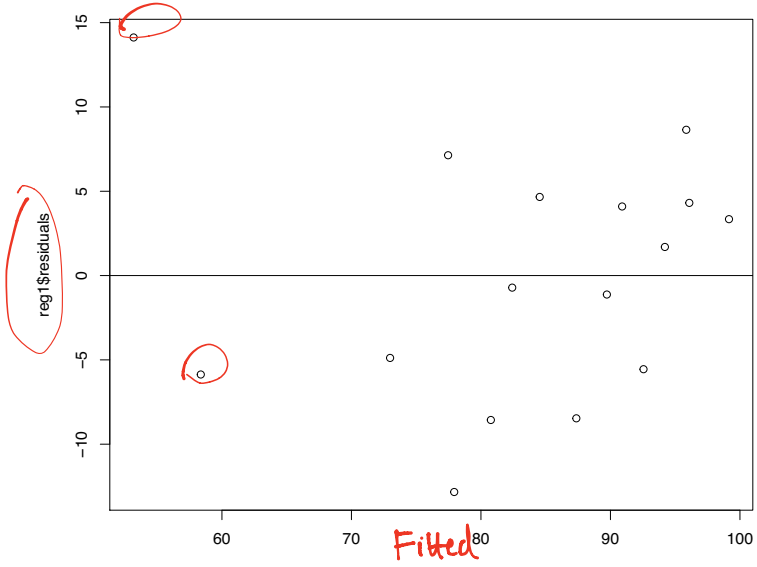


Linear Regression

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

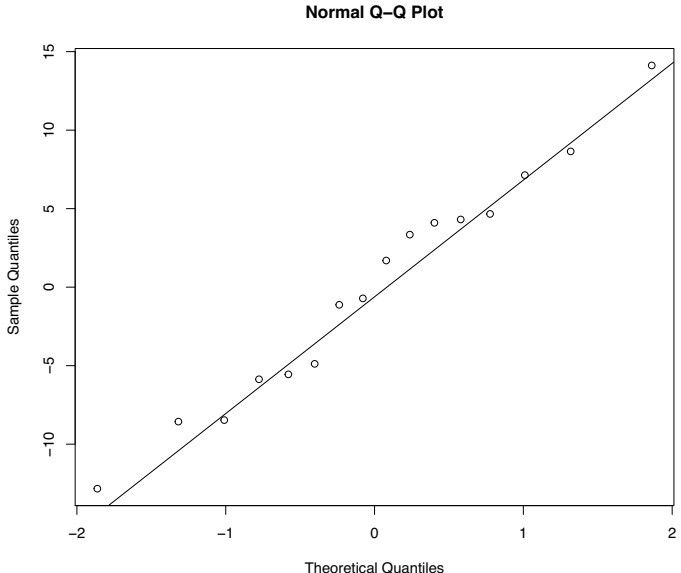
reg1 – regression object $

No pattern



Linear Regression

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



Assumption of normality seems plausible
⇒ p-values and CI for reg param are OK!

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

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

The least squares estimator is

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

use this to get est. of weights

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.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

lm(M ~ T + Diet + T:Diet)

multiple linear reg in R

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

measurement of the diff. in weight between apples

Weighing Problem

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 unknown weights

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

m_1, m_2 observed

Weighing Problem

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

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

$$y_i = w_1 x_{i1} + w_2 x_{i2}, \quad i = 1, 2$$

$$x_{in} = \begin{cases} 1 & \text{Left pan} \\ -1 & \text{Right pan} \\ 0 & \text{not included} \end{cases}$$

Weighing Problem

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
```

matrix multiplication

$$\hat{\beta} = \begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = (X^T X)^{-1} X^T y$$

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

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

$$\hat{\beta} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \left(\frac{m_1 + m_2}{2}, \frac{m_1 - m_2}{2} \right)$$

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

$$\mathrm{Var}(\hat{w}_1) = 0.5\,\sigma^2 = \mathrm{Var}(\hat{w}_2)$$

$$Cov(\hat{\beta}) = (X^T X)^{-1} \sigma^2$$

From R $\rightarrow$

$$= \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \sigma^2 = \begin{pmatrix} \frac{\sigma^2}{2} & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}$$

$$Var(\hat{\beta}_1) = Var(\hat{\beta}_2) = \frac{\sigma^2}{2}$$

END