

STA305/1004 - Review of Statistical Theory

Look for first assignment
by the end of this week.

January 11, 2017

Due: Jan. 27, 10:00 pm.
Friday

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.

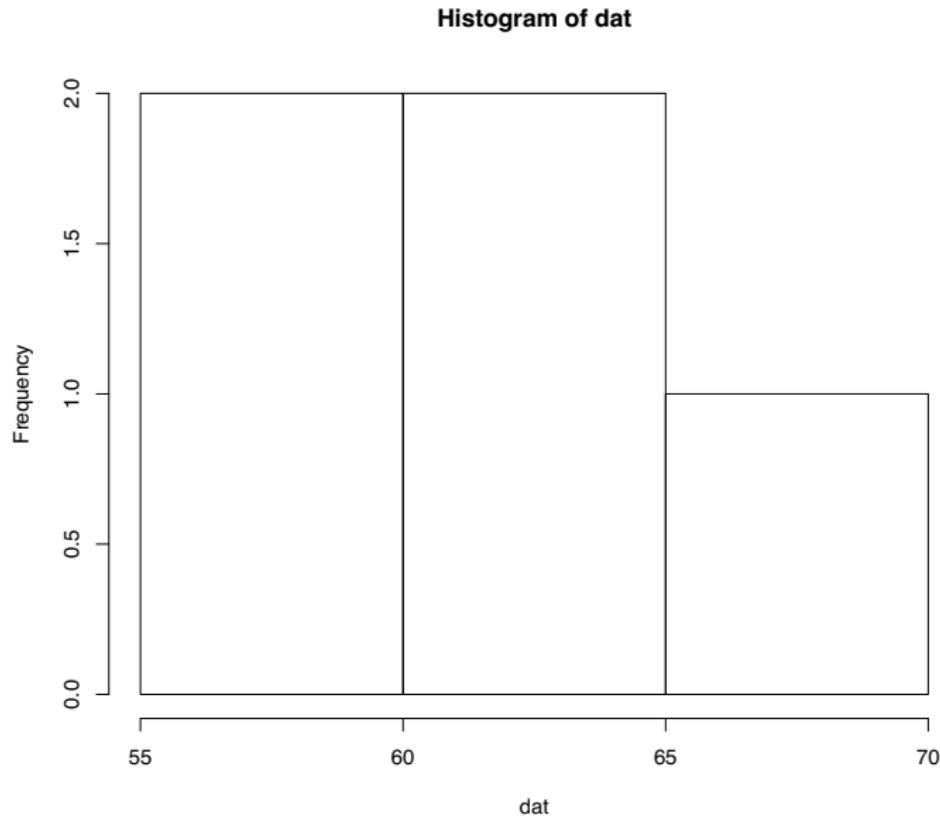
```
summary(dat)
```

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

5 Number
Summary.

Distributions

```
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.

Population \equiv All UoT Students

Sample \equiv Students in STA305

Continuous Distributions

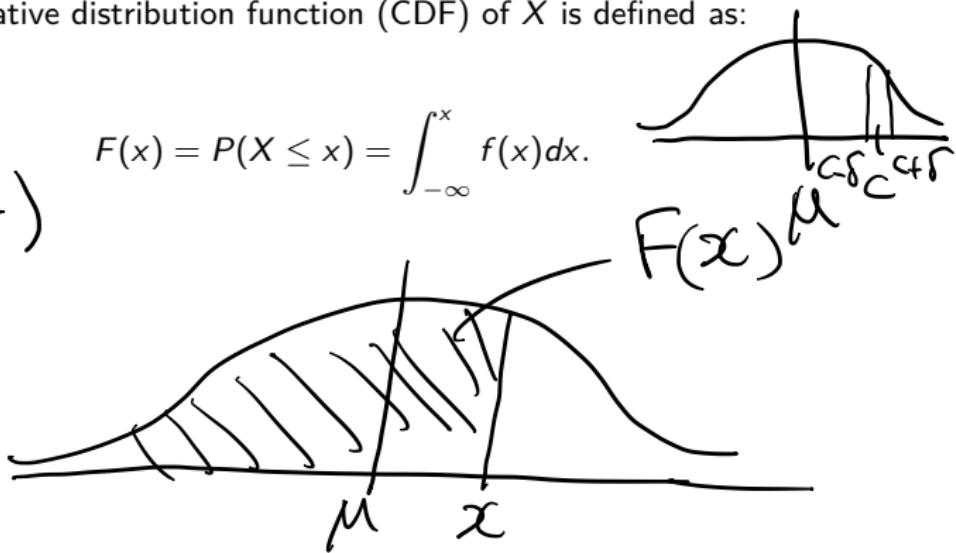
C IS a Constant

$$P(X=c) = 0$$

- ▶ A continuous random variable X is fully characterized by its 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:

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

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



Continuous Distributions

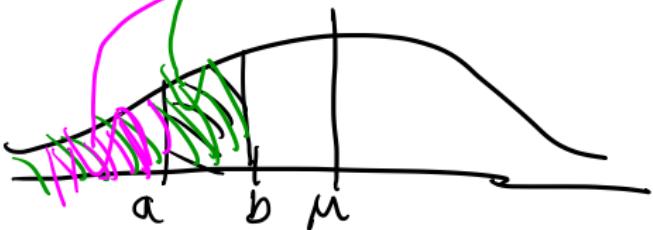
- ▶ If f is continuous at x then $F'(x) = f(x)$ (fundamental theorem of calculus).
- ▶ The CDF can be used to calculate the probability that X falls in the interval (a, b) . This is the area under the density curve which can also be expressed in terms of the CDF:

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

density

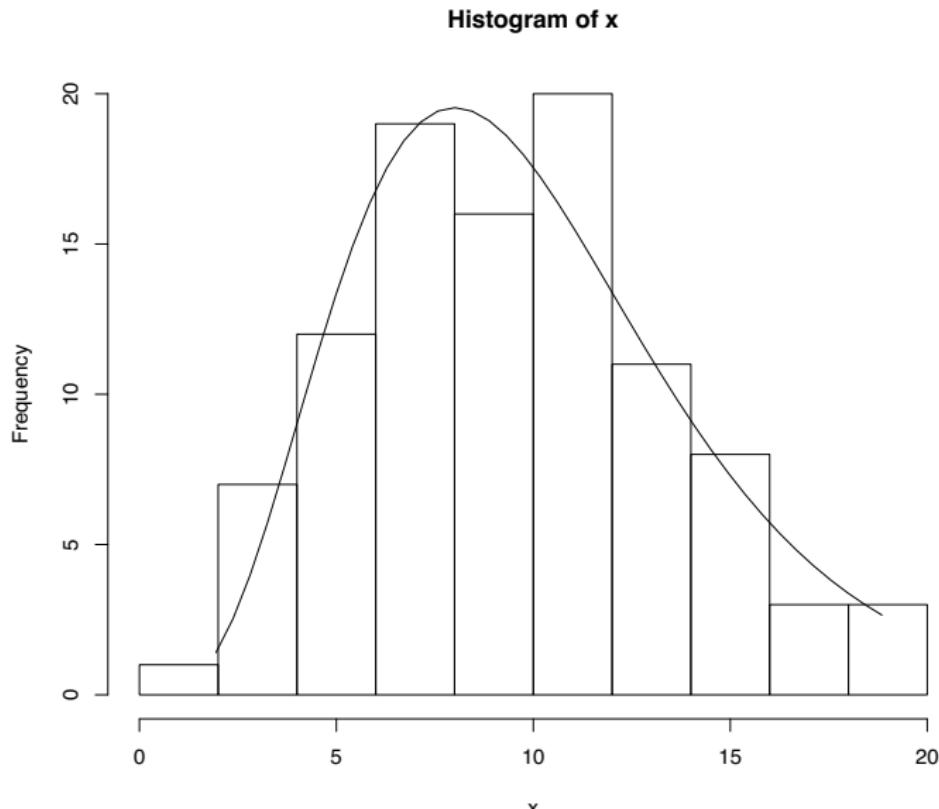
- ▶ In R a list of all the common distributions can be obtained by the command `help("distributions")`.
- ▶ For example, the normal density and CDF are given by `dnorm()` and `pnorm()`.

↓
CDF



Continuous Distributions

100 observations (using `rchisq()`) from a Chi-square distribution on 10 degrees of freedom χ_{10}^2 . The density function of the χ_{10}^2 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

$$N(\mu, \sigma^2)$$
$$x_1, x_2, \dots, x_{20}$$
$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20}$$

$$\mu$$

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

Residuals and Degrees of Freedom

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

$$\sum_{i=1}^n (y_i - \bar{y}) = \sum_{i=1}^n y_i - n\bar{y}$$

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

$y_i - \bar{y}$ is called a residual.

- ▶ Since $\sum(y_i - \bar{y}) = 0$ any $n - 1$ completely determine the last observation.
- ▶ This is a constraint on 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:

$$\text{Variance} = \sigma^2$$

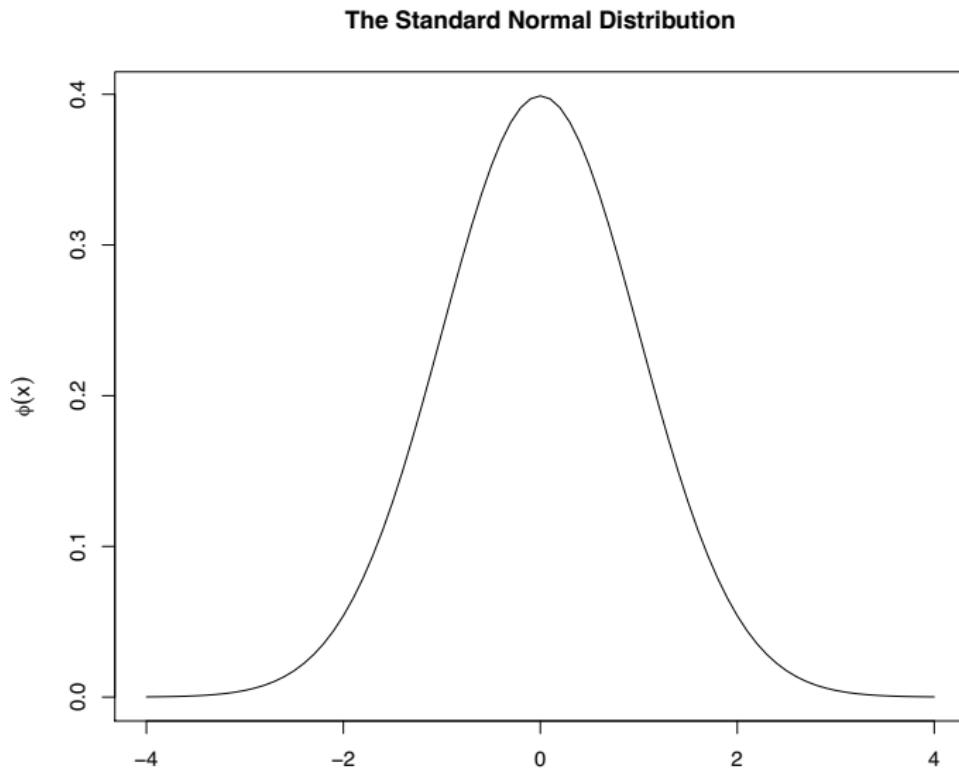
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

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

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

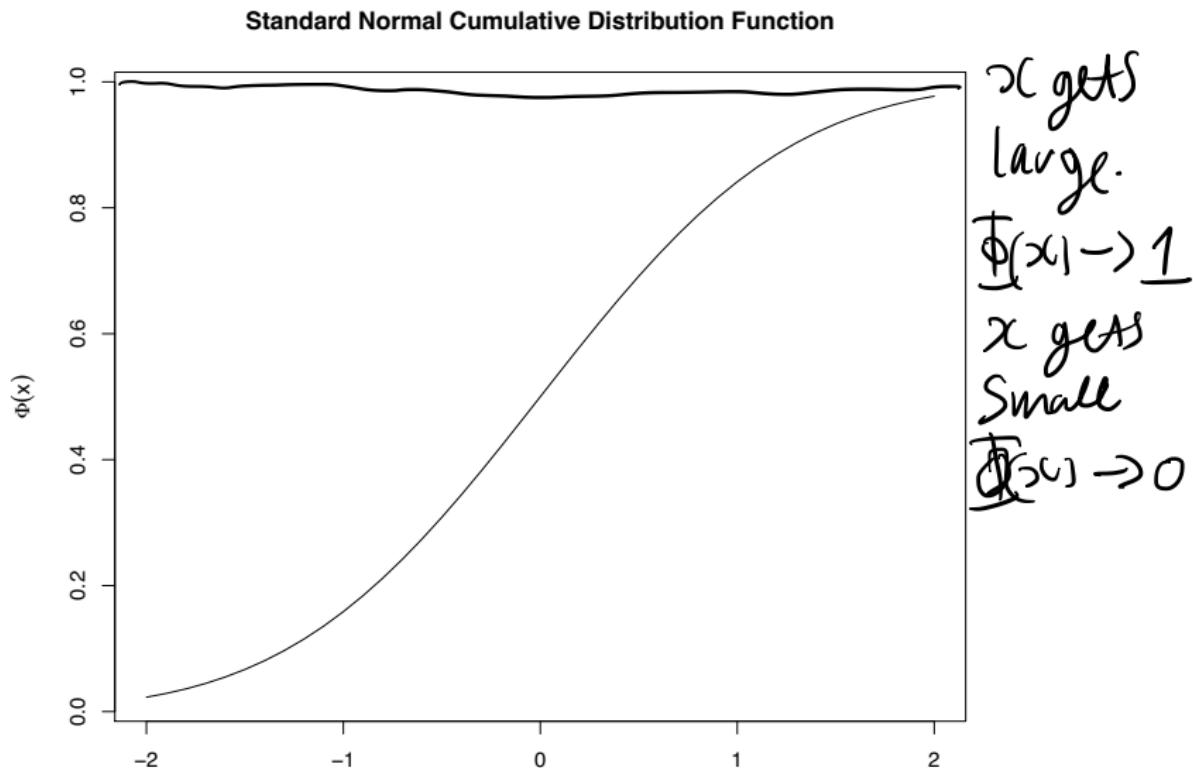
The Normal Distribution

```
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")
```



The Normal Distribution

$$Y \sim N(10, 2)$$

$$\text{Var}\left(\frac{Y-10}{\sqrt{2}}\right) = \frac{1}{2} \text{Var}(Y-10) = \frac{1}{2} \text{Var}(Y) = \frac{2}{2} = 1$$
$$\frac{Y-10}{\sqrt{2}} \sim N(0, 1)$$

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

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

$$E\left(\frac{Y-10}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}} E(Y-10)$$

$$= \frac{1}{\sqrt{2}} (E(Y) - 10)$$

$$= \frac{1}{\sqrt{2}} (10 - 10)$$

$$= 0$$

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

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

where

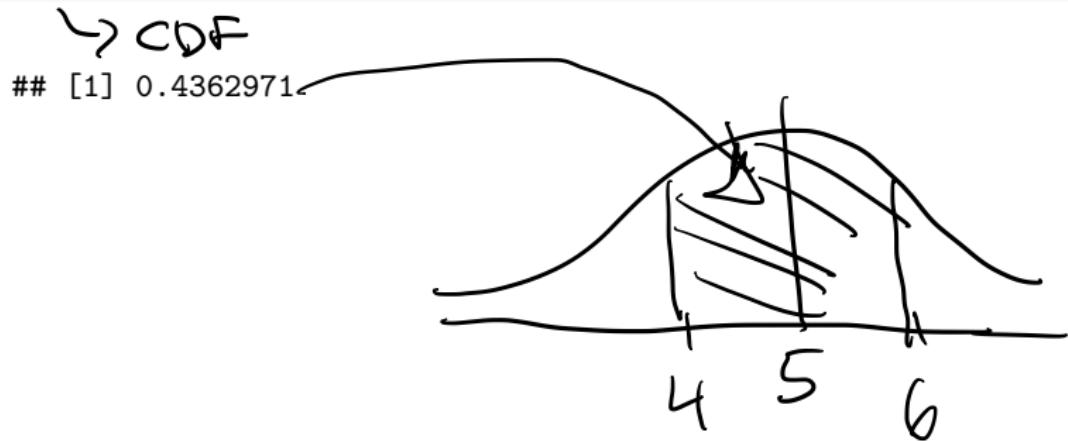
$$Z = \frac{Y - \mu}{\sigma}.$$

Any $N(\mu, \sigma^2)$ r.v. Y can be represented as $Y = \mu + \sigma Z$, where $Z \sim N(0, 1)$

The Normal Distribution

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

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



Normal Quantile Plots

The following data are the weights from 11 tomato plants.

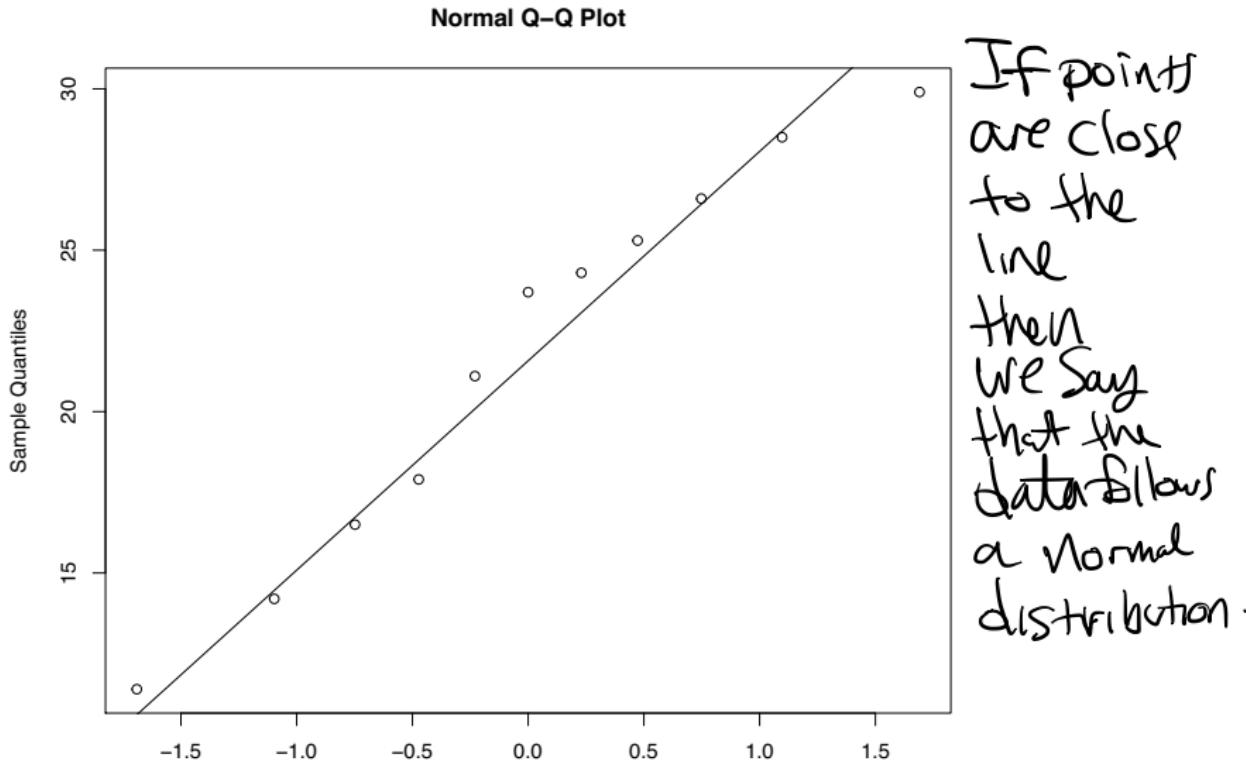
```
## [1] 29.9 11.4 26.6 23.7 25.3 28.5 14.2 17.9 16.5 21.1 24.3
```

Do the weights follow a Normal distribution?

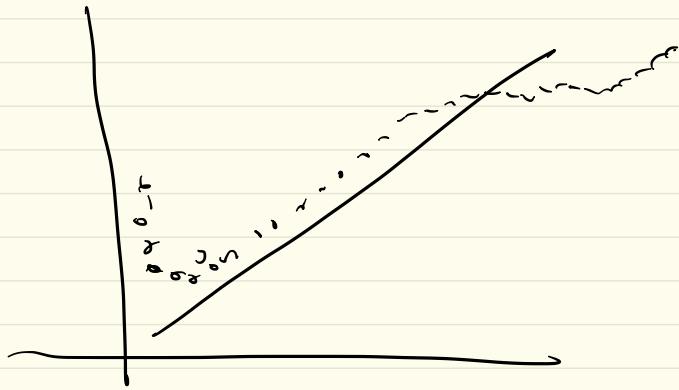
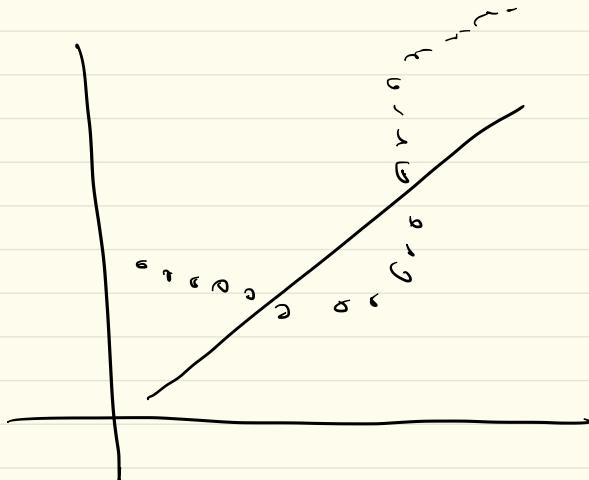
Normal Quantile Plots

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)
```



When does a normal quantile plot indicate non-normality?



If the points systematically deviate from the line then this indicates that the data does not follow a normal distribution.

Central Limit Theorem

Suppose that the density of X

We have a random sample.

Then the CLT States that



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

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

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

the distribution of \bar{X} is approx.

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

The distribution of the Sample Mean.

Assume that we collect ^{Samples of} random

Size $n=50$

Sample #

1

2

3

⋮

10,000

Sample mean

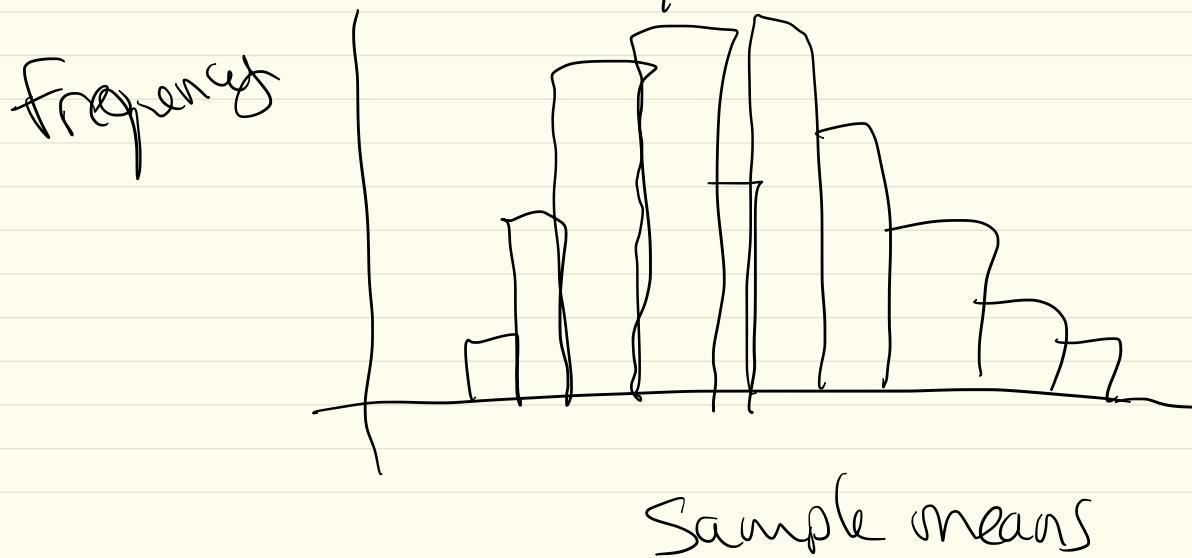
\bar{X}_1

\bar{X}_2

\bar{X}_3

$\bar{X}_{10,000}$

If we plot a histogram of
the 10,000 Sample means it
will look like a Normal
with mean μ and variance $\frac{\sigma^2}{n}$



Central Limit Theorem

$\text{Bin}(50, 0.5)$ is the distribution of the total # of Heads

$\frac{\#}{1}$	$\frac{\#}{2}$	$\frac{\#}{3}$...	$\frac{\#}{50}$
----------------	----------------	----------------	-----	-----------------

$\mu = \underline{50 \times 0.5}$

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

$$\frac{\text{Total } \# \text{ Hs}}{50} = \hat{P} = \sum_{i=1}^{50} X_i / 50 \sim N(25, \frac{0.5 \times 0.5}{50})$$

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

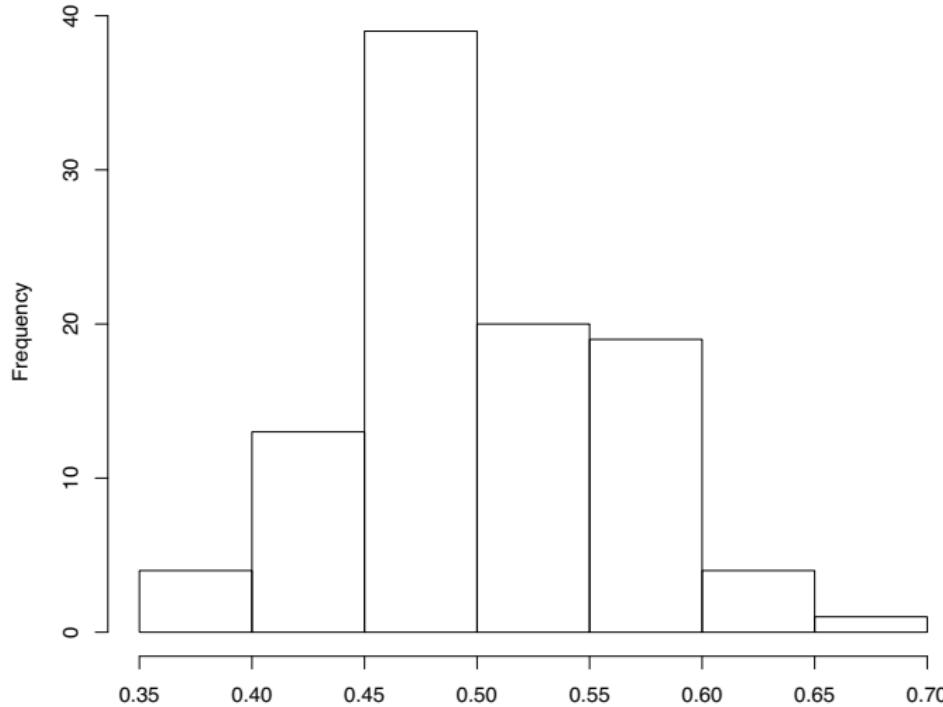
$$N(\mu, \sigma^2) \quad \sigma^2 = \frac{p(1-p)}{n} = \frac{0.5(1-0.5)}{50}$$

Central Limit Theorem

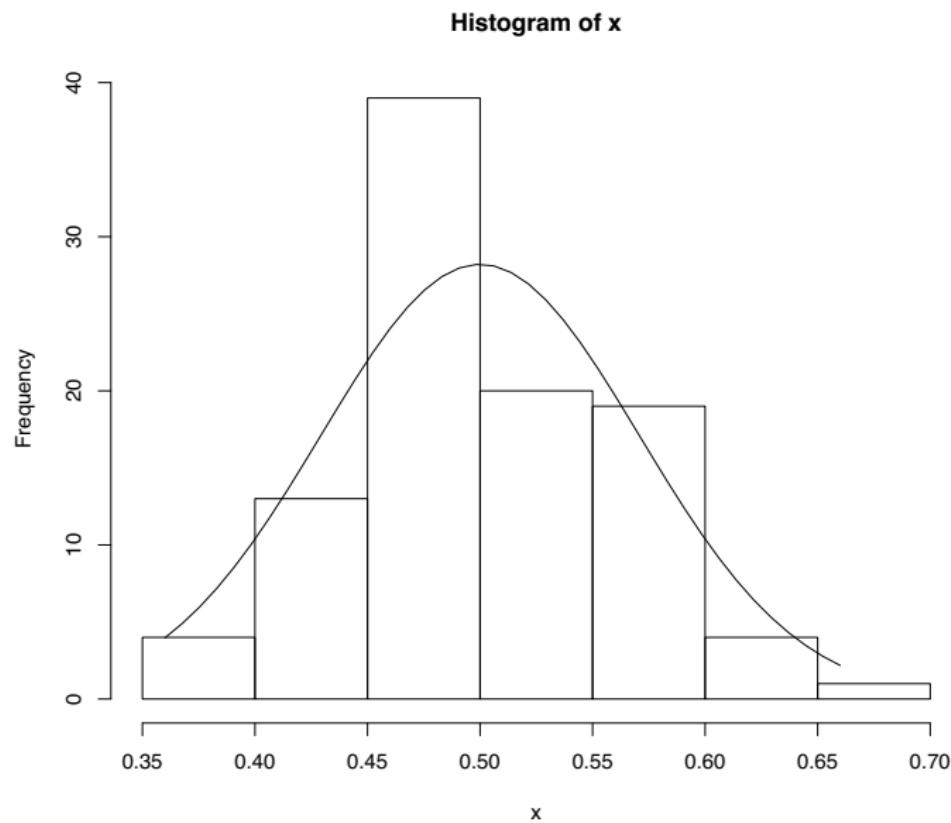
```
set.seed(100)
```

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

Histogram of Ave.heads



Central Limit Theorem



Chi-Square Distribution

Let X_1, X_2, \dots, 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

$$\frac{(n-1)}{\sigma^2} S^2 + (N(0,1))^2 = \chi^2_{(n)}$$

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

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

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \sim \chi^2_{(n)}$$

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

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

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

$$H_0: \mu = 0$$

$$H_a: \mu \neq 0$$

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

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

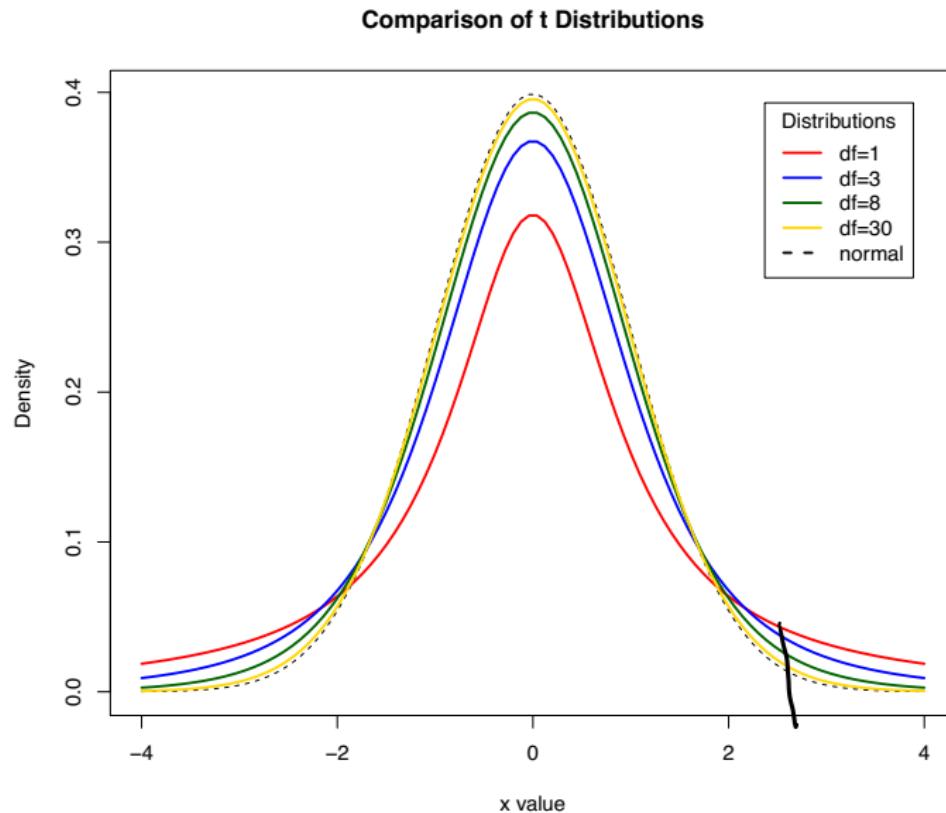
$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n-1}}}$

If we replace
 σ by s .

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

follows a
 $t(n-1)$

t Distribution



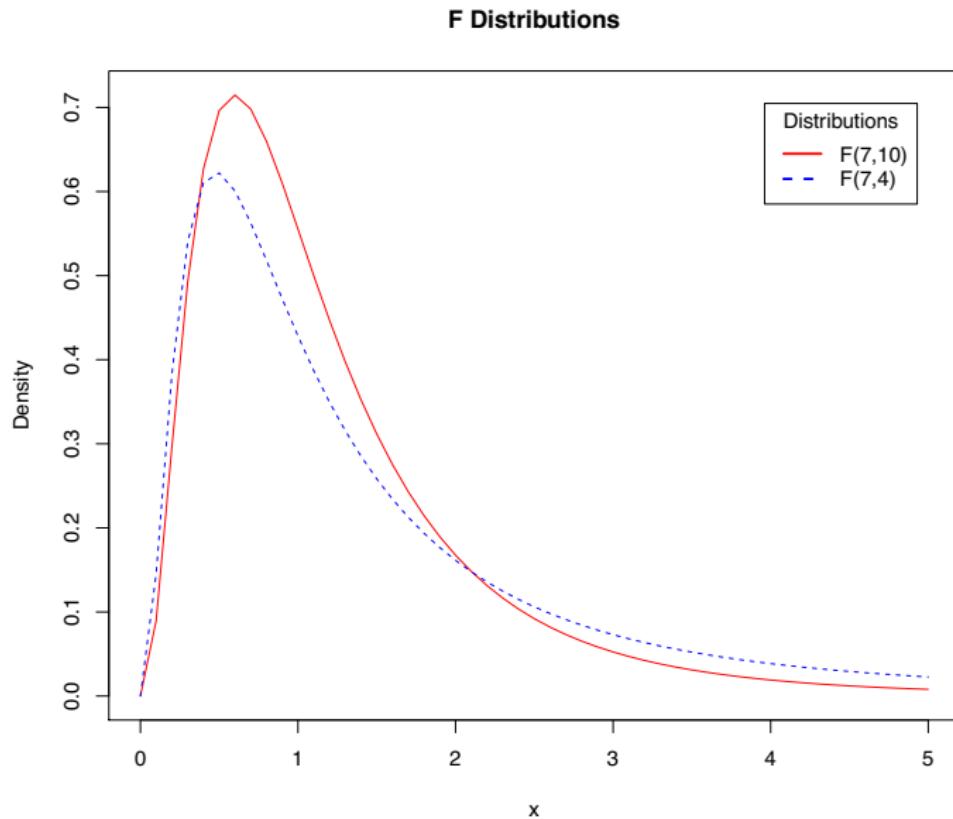
F Distribution

Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ 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 $F_{1,n}$.

F Distribution



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:

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

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

$$\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2, \quad \frac{\partial}{\partial \beta_0} = 0 \quad \frac{\partial}{\partial \beta_1} = 0$$

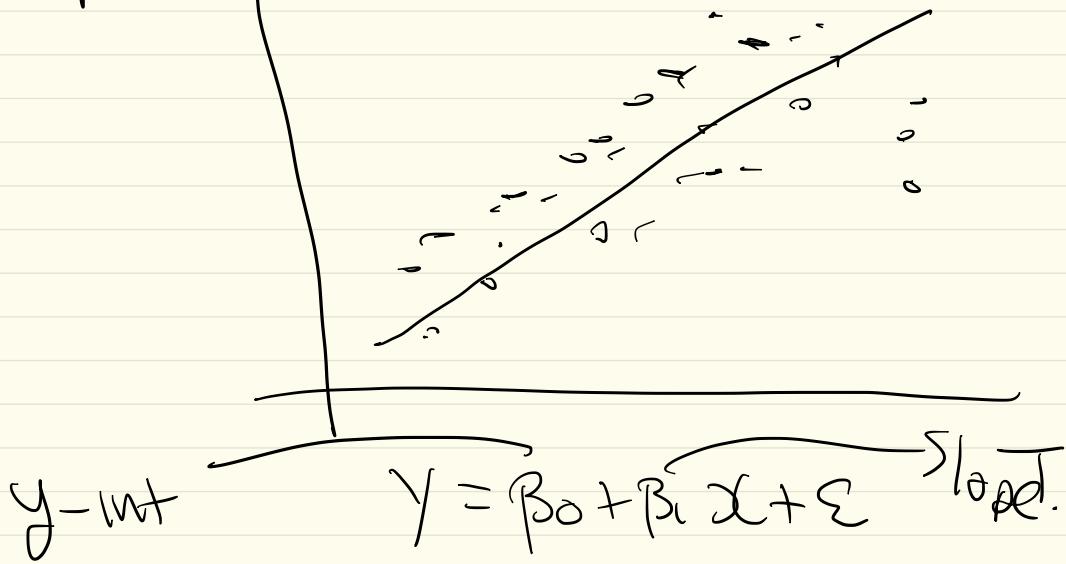
are called the least squares estimators. They are given by:

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

Correlation Coeff.

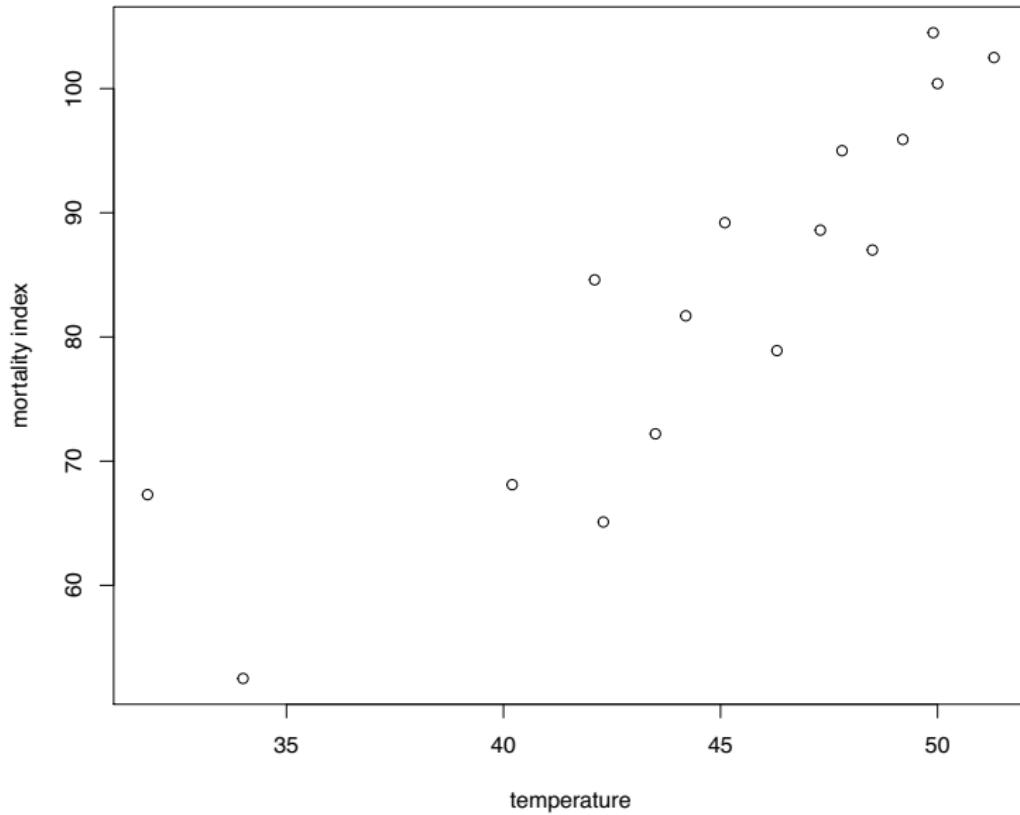
r is the correlation between y and x , and S_x, S_y are the sample standard deviations of x and y respectively.

\mathcal{N}



Linear Regression

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



Linear Regression

$$H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$$

```
reg1 <- lm(M~T)
```

```
summary(reg1) # Parameter estimates and ANOVA table
```

Strong evidence
against H_0

```
## 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
```

$$\frac{\hat{\beta}}{S.e(\hat{\beta})} = t$$

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \epsilon \\ \beta_1 &= 0 \quad Y = \beta_0 + \epsilon \end{aligned}$$

$$\hat{M}(T) = -21.79 + 2.36 T$$

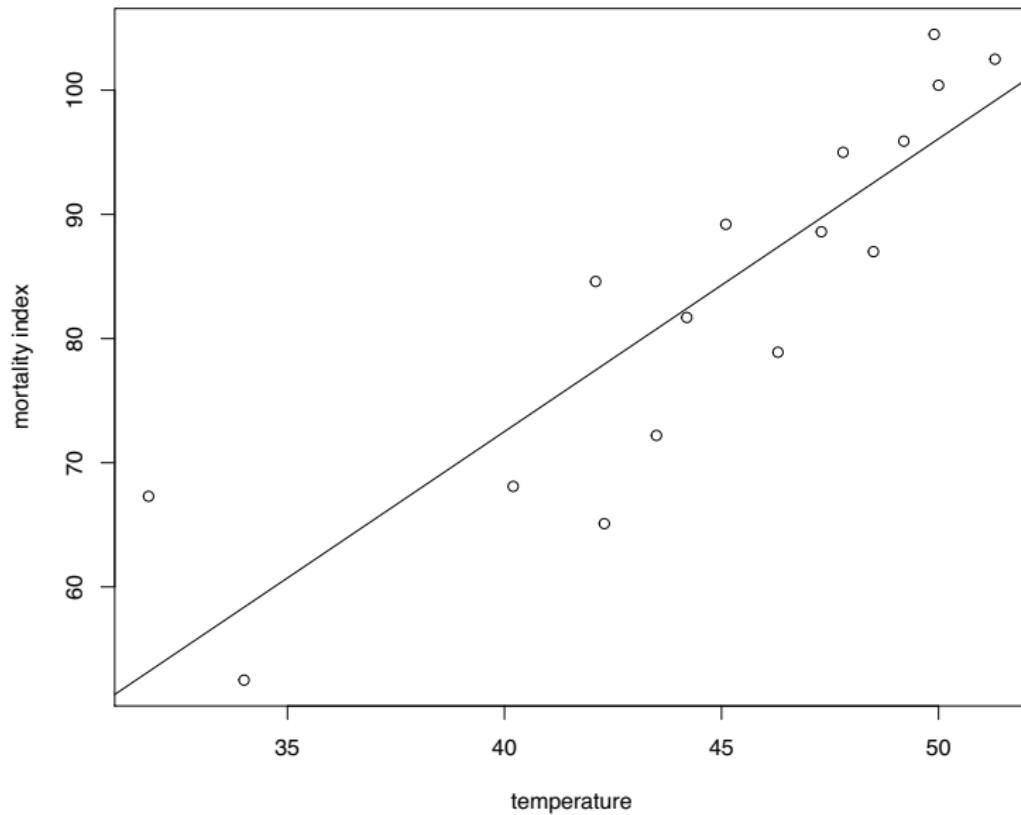
$$S.e(\hat{\beta}_1)$$

$$9.2 \times 10^{-6}$$

% of variation explained
by lin
Peg.

Linear Regression

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



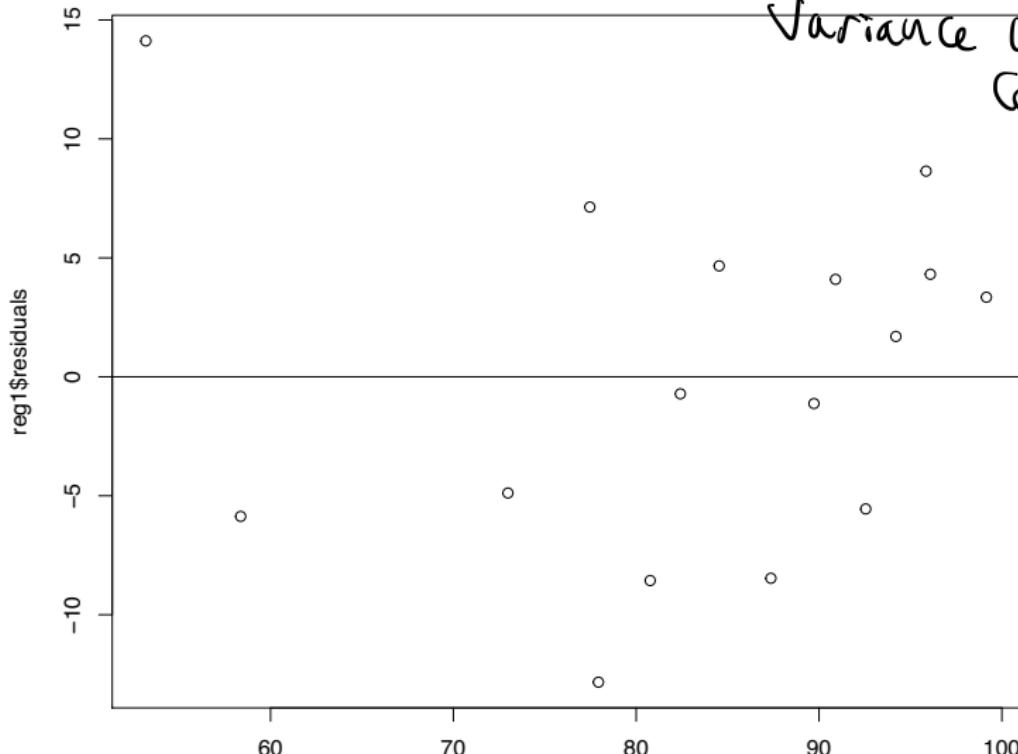
Linear Regression

#plot residuals vs. fitted

```
plot(reg1$fitted, reg1$residuals);
```

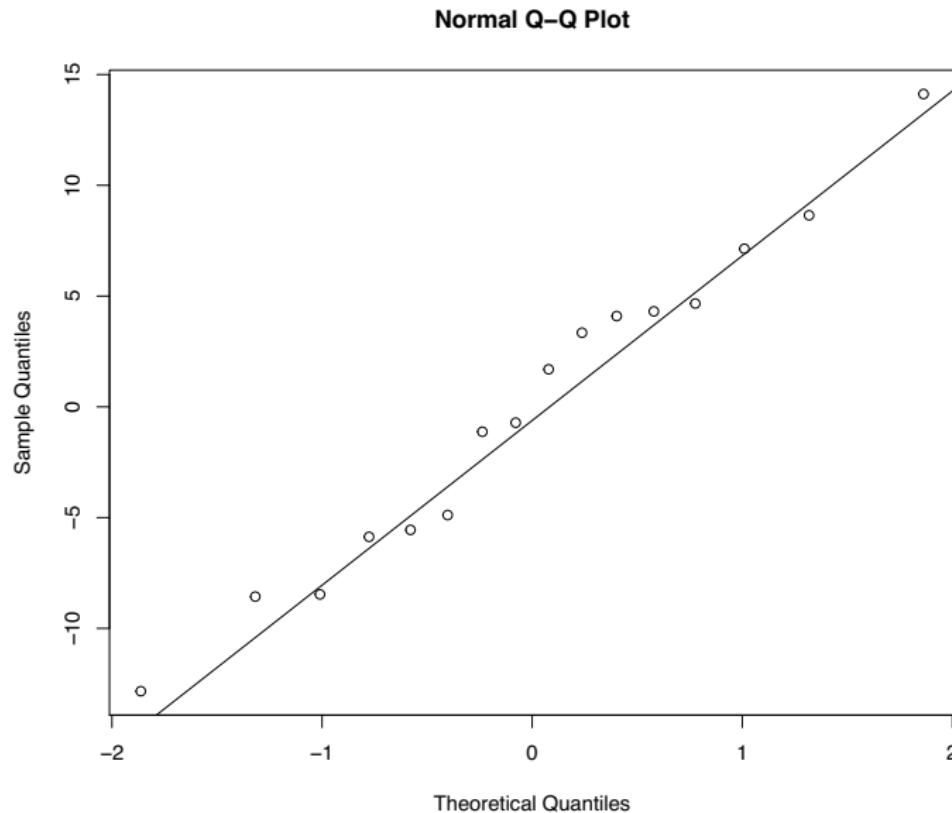
```
abline(h=0) # add horizontal line at 0
```

Random Scatter on
a residual plot
(Indicate) that
variance is
constant.



Linear Regression

```
#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.

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

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

This can also be expressed in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

The least squares estimator is

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

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is $(\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$. 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

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 :

$$\begin{cases} w_1 + w_2 = m_1 \\ w_1 - w_2 = m_2 \end{cases}$$

*→ Weigh together
→ Weigh ONE
in each pan.*

Weighing Problem

$$Y = w_1 x_{i1} + w_2 x_{i2}$$

$$x_{ik} = \begin{cases} 1 & \text{Left pan} \\ -1 & \text{Right pan} \end{cases}$$

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

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

$$\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix}$$

$$Y_i = w_1 x_{i1} + w_2 x_{i2}$$

Weighing Problem

The least-squares estimates can be found using R.

#step-by-step matrix multiplication 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
```

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

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \begin{pmatrix} \frac{m_1 + m_2}{2} \\ \frac{m_1 - m_2}{2} \end{pmatrix}$$

W # print $(X^T*X)^{-1}$ for SE

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

$$\begin{aligned} \text{Var}(\hat{m}_1) &= \sigma^2 / 2 \\ &= \text{Var}(\hat{m}_2) \end{aligned}$$