STA303/1004 - Intro to one-way ANOVA

January 9, 2018

Week 1 Topics

REVIEW

-Data summary: Five-number summary, Boxplots , tables

-Large-sample distribution theory: derived from Normal (CT, Z, χ^2, F)

-Statistical inference: confidence interval, hypothesis tests, errors, power

-Normality Test, Equal variance test

T-TESTS

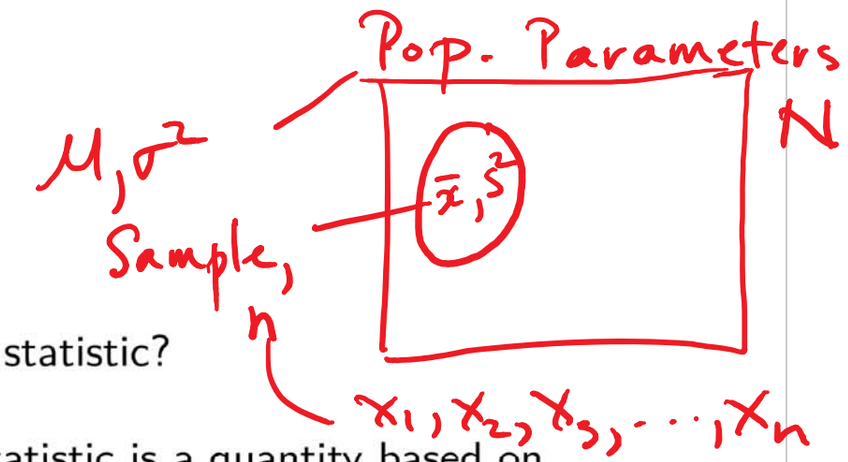
-One-sample t-test

-Paired t-test

-Two-sample t-test

-Non-parametric alternatives

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

The Normal Distribution

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

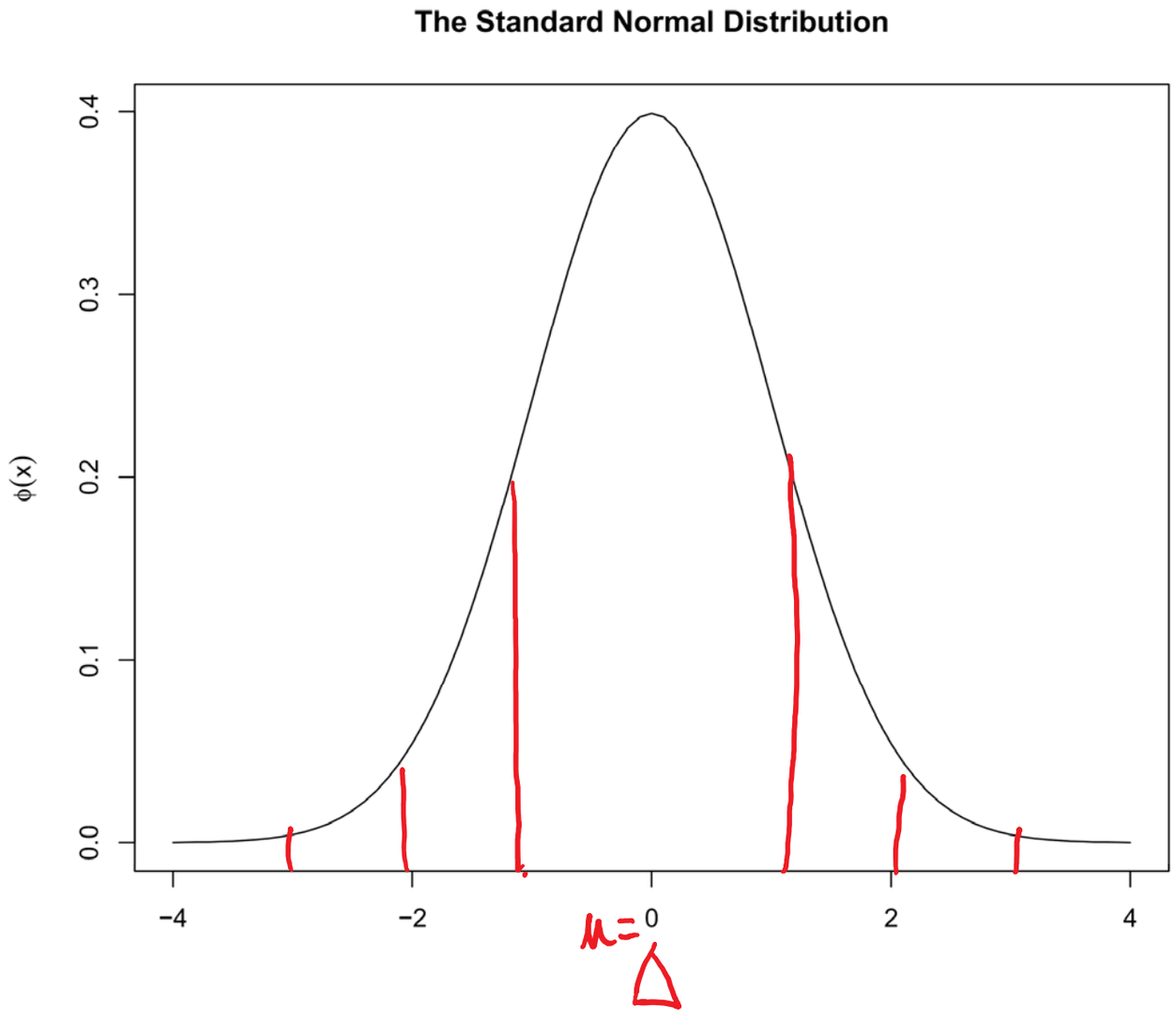
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

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

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

The Standard Normal Distribution

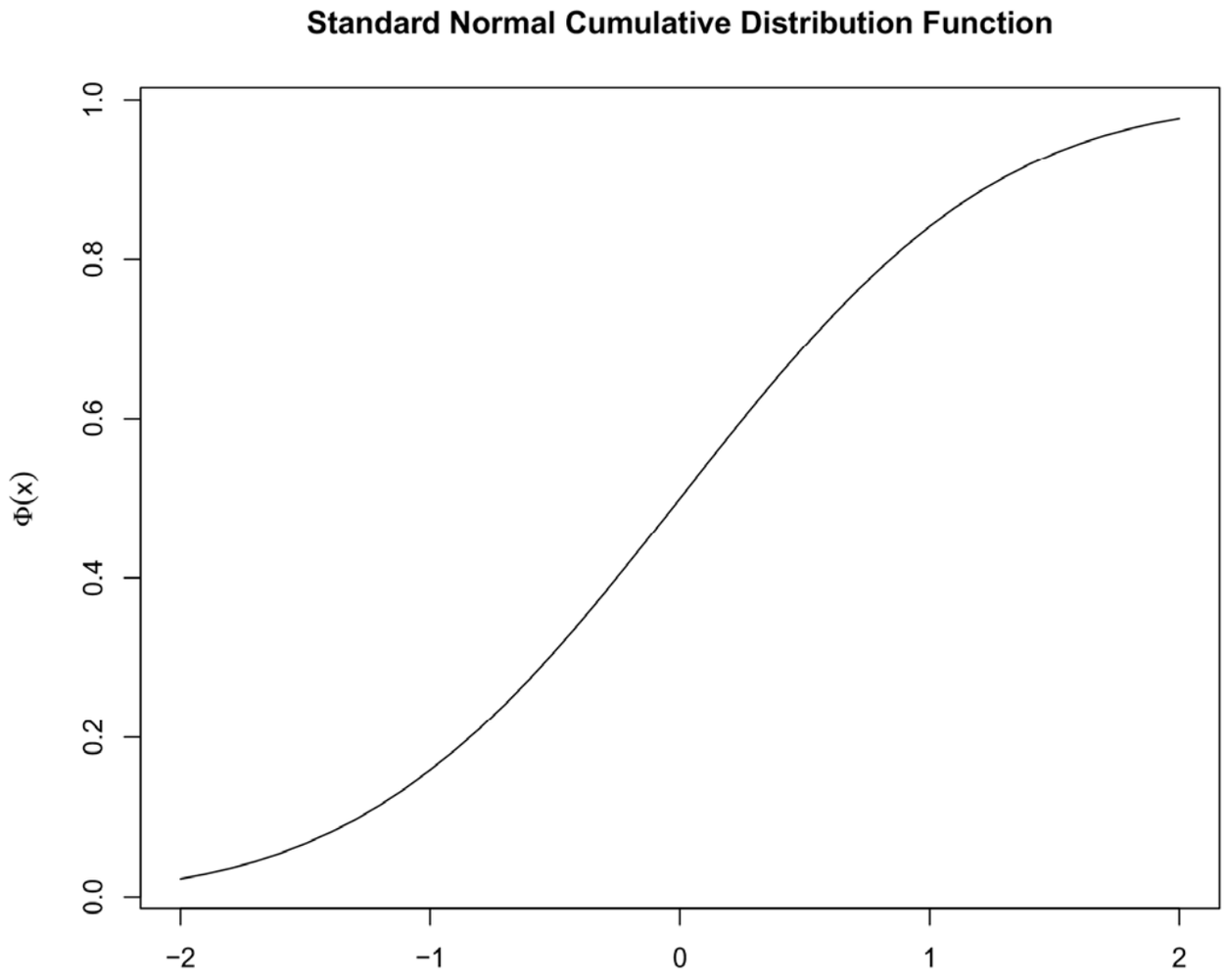
```
x <- seq(-4,4,by=0.1)
plot(x,dnorm(x),type="l",main = "The Standard Normal Distribution",
     ylab=expression(paste(phi(x))))
```



68-95-99.7% Rule

The Standard Normal CDF

```
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 and Standard Normal Distributions

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

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

where

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

The Normal Distribution

$X \sim N(0, 1)$. Use R to find $P(-2 < X < 2)$.

```
pnorm(2,mean = 0,sd = sqrt(1))-pnorm(-2,mean = 0,sd = sqrt(1))
```

```
## [1] 0.9544997
```

Normal Quantile-Quantile Plots

-used to visually assess Normality of a sample of measurements

-in R, use `qqnorm()` for the normal qq plot and `qqline()` to add the straight line.

Linear combination of IID Normal

If $X_i \sim N(\mu, \sigma_i^2)$ independently, then

$$V = a + \sum_1^n b_i X_i \sim N(a + \sum_1^n b_i \mu_i, \sum_1^n b_i^2 \sigma_i^2)$$

$\bar{X} = \frac{\sum X_i}{n}$

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

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

$Z^2 \sim \chi_1^2$

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 $2n$.

$H_0: \sigma^2 = \sigma_0^2 \quad (\sigma = \sigma_0)$

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

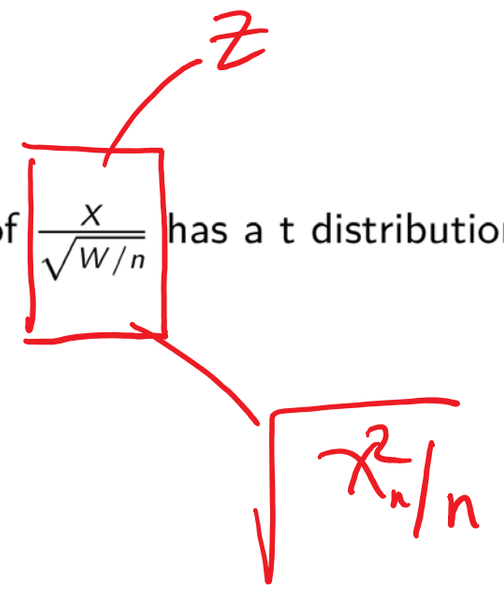
$$\frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1}$$

sample variance

t Distribution

$X \perp W$

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



$T_{df} \xrightarrow{D} Z$

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

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

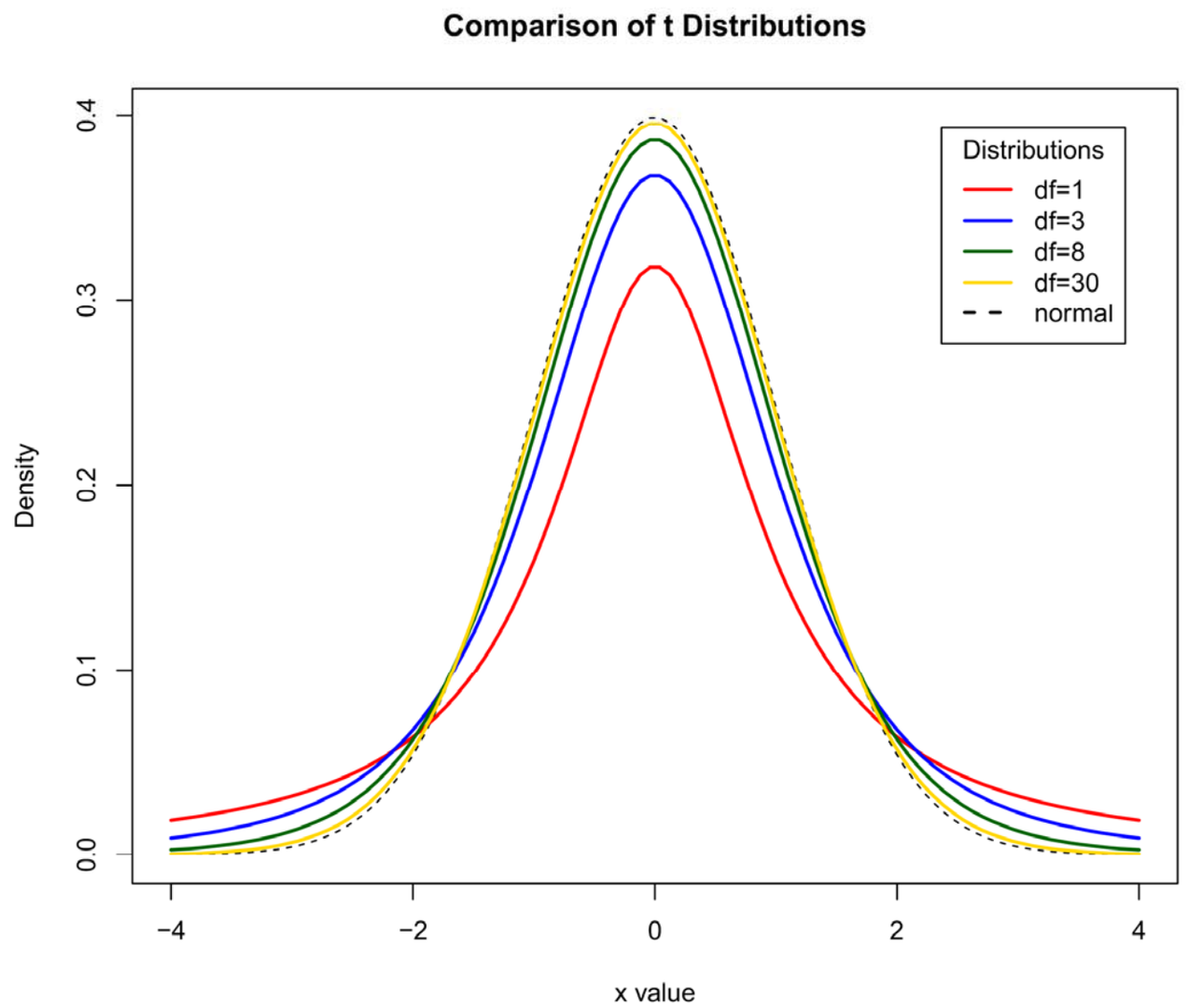
where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$?

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t_{n-1}$$

Z

$\sqrt{\chi^2_{n-1}/(n-1)}$

t Distribution



F Distribution

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

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

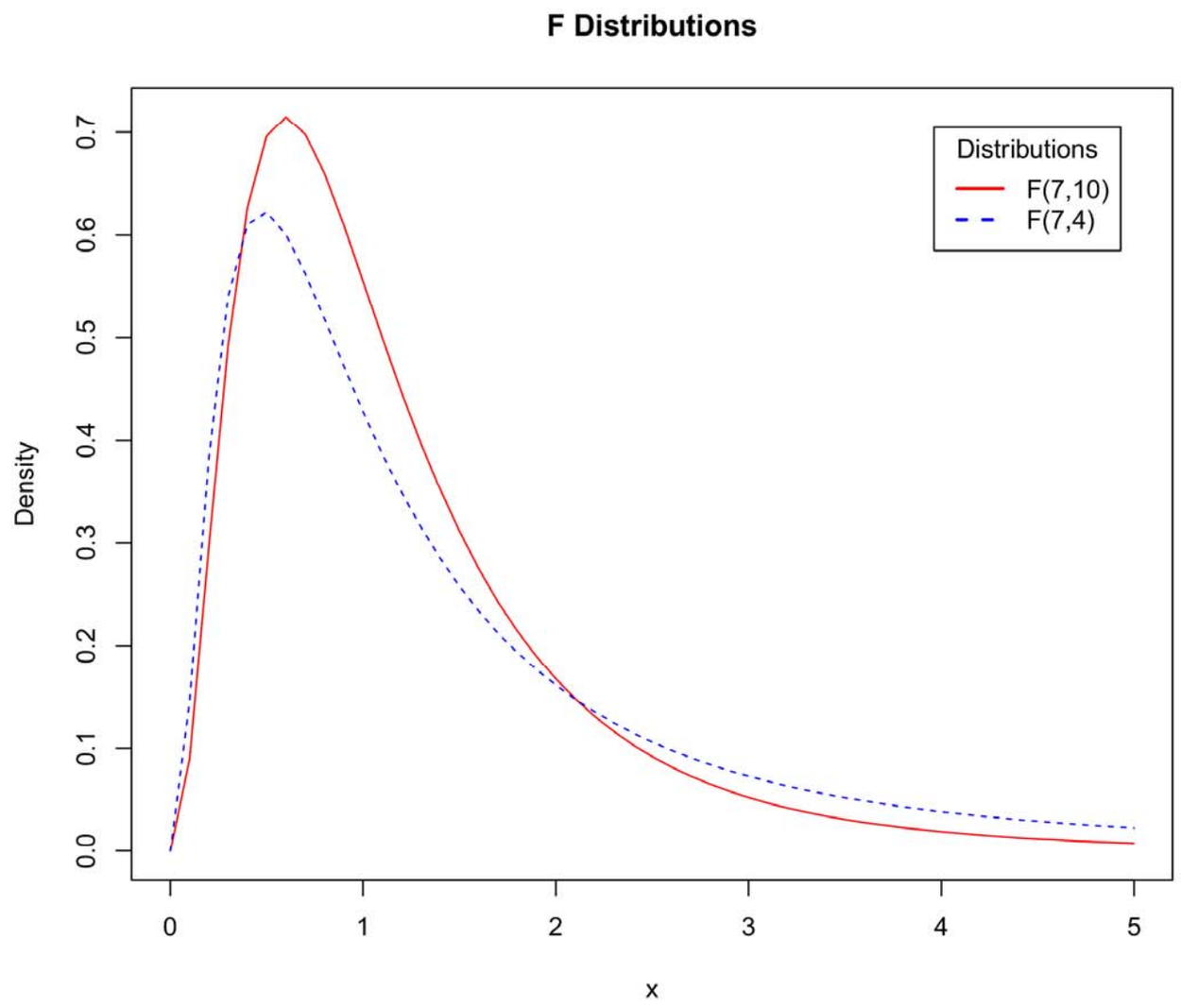
where $F_{m,n}$ denotes the F distribution on m, n degrees of freedom. The F distribution is right skewed (see graph below). For $n > 2, E(W) = n/(n-2)$. It also follows that the square of a t_n random variable follows an $F_{1,n}$.

$$T^2_{df} \overset{D}{=} F_{1,df}$$

χ^2 : $H_0: \sigma^2 = \sigma_0^2$

F : $H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$

F Distribution



The Sample Mean

If $X_1, \ldots, X_n \sim_{iid} N(\mu, \sigma^2)$ then

- $\bar{X} \sim N(\mu, \sigma^2/n)$
- $S^2 = \sum(X - \bar{X})^2/(n-1)$ and

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

- $\bar{X} \perp S^2$ and
-

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Simple Linear Regression

A simple linear regression model 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

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

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.

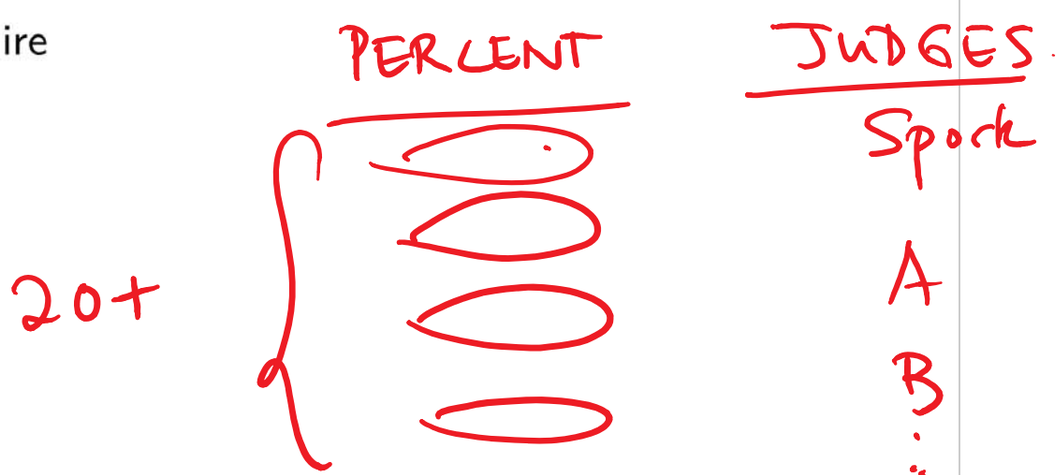
Case Study 1: The Spock Conspiracy Trial

- Boston, 1968
- Dr. Benjamin Spock (paediatrician and author) on trial for conspiring to violate the Selective Service Act.
- Accused of encouraging people to dodge military draft by his books that adviced on how mothers should raise children.
- Spock's jury had NO women.

Q: Is there evidence of gender bias in the jury selection for Spock's trial?

Case Study 1: Jury selection

- 300 names selected at random from city directory
- 35 to 200 jurors randomly selected (this group is called the venire)
- Then non-random selection or exclusion of jurors from the venire by both defence and prosecution
- For Spock's trial, only 1 woman in the venire but she was then dismissed by prosecution
- Defence argued that Spock's judge had history of women being underrepresented on his venires.
- Compared composition of recent venires of 6 other judges with that of Spock's judge
- Data: percent of women in each venire



Case Study 1: Two Key Questions

Spock's vs Others.

- Q1. Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges?
- Q2. Is there evidence that there are differences in women's representation in venires of the other 6 judges?

Others.

- Q: Conduct the relevant hypothesis test to answer Q1. Include the necessary assumptions, justifications and elements of a hypothesis test. What is your conclusion in plain English?

Case Study 1: The Spock Conspiracy Trial Data

The data is shown below.

```
#Juries data
juries<-read.csv(
  "/Users/Shivon/STA303_1002/LectureNotes/Lec1/juries.csv", header=T)
attach(juries)
#head(juries)
PERCENT
```

dim
length

```
##  [1]  6.4  8.7 13.3 13.6 15.0 15.2 17.7 18.6 23.1 16.8 30.8 33.6 40.
## [15] 27.0 28.9 32.0 32.7 35.5 45.6 21.0 23.4 27.5 27.5 30.5 31.9 32.
## [29] 33.8 24.3 29.7 17.7 19.7 21.5 27.9 34.8 40.2 16.5 20.7 23.5 26.
## [43] 29.5 29.8 31.9 36.2
```

```
JUDGE
```

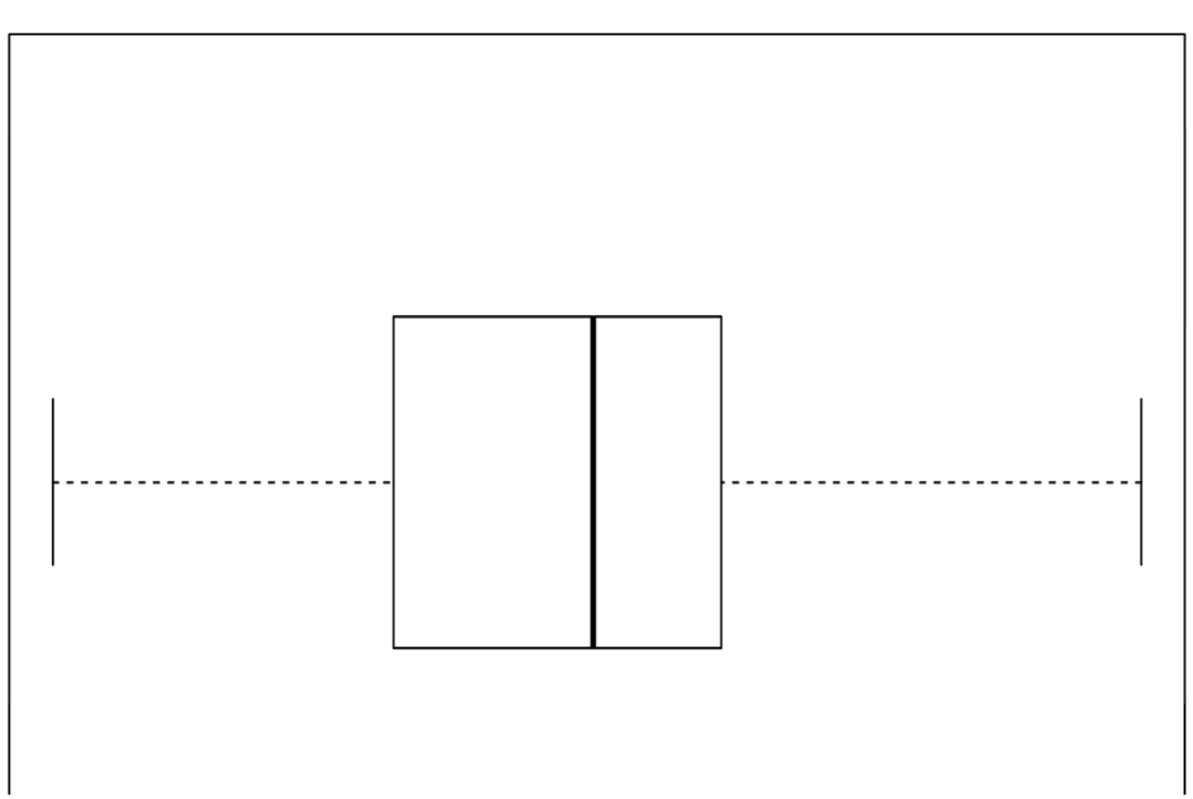
```
##  [1] SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS
## [11] A      A      A      A      B      B      B      B      B
## [21] C      C      C      C      C      C      C      C      C
## [31] D      E      E      E      E      E      E      F      F
## [41] F      F      F      F      F      F
## Levels: A B C D E F SPOCKS
```

Case Study 1: Data summary

```
summary(PERCENT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.40   19.95   27.50   26.58   32.38   48.90
```

```
boxplot(PERCENT, horizontal=T,main="Percent of women")
```



on all venues

5-num summary

$IQR = Q_3 - Q_1$

(robust measure of spread).

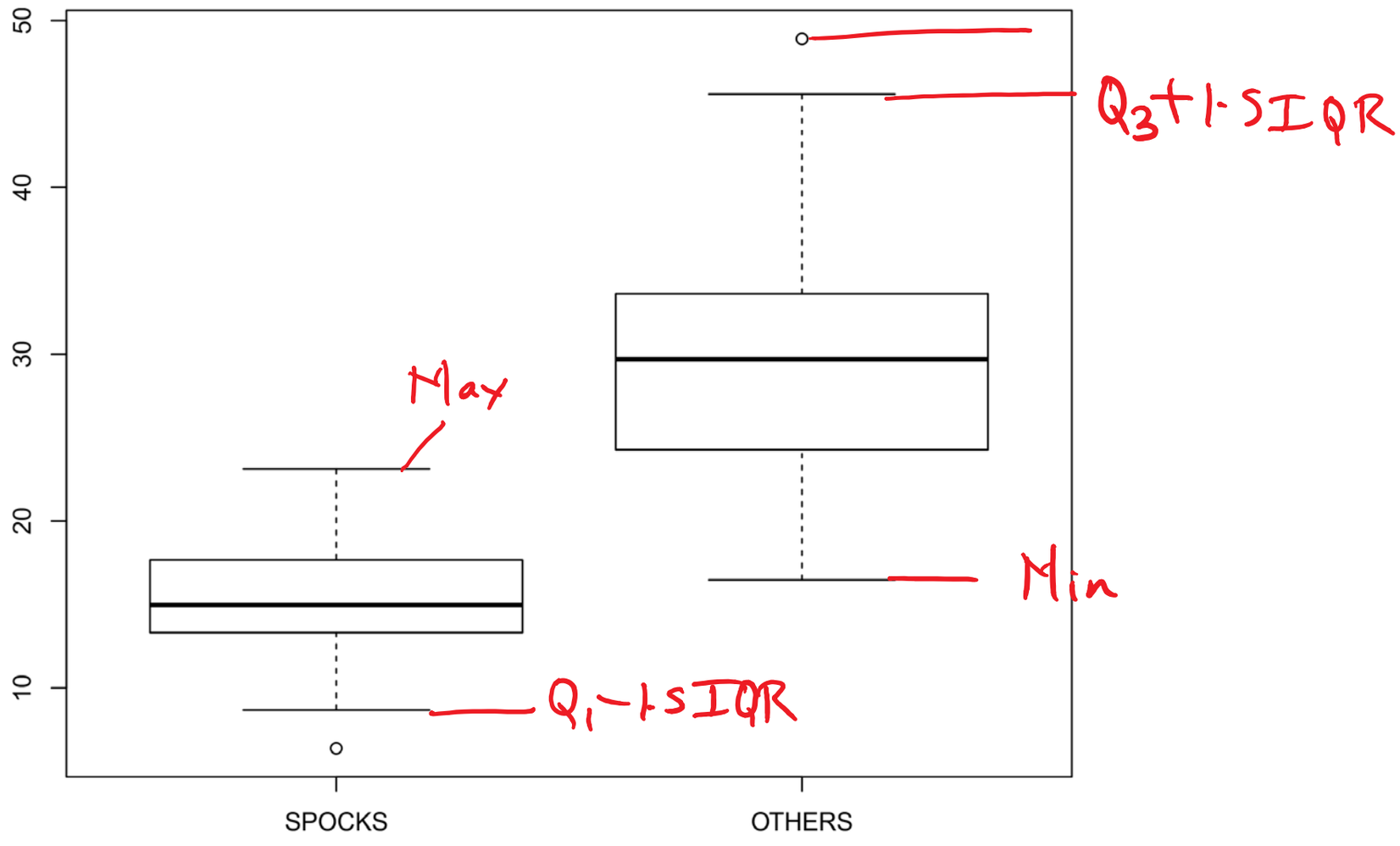
middle 50%

Outlier Rule:

1.5 IQR

Case Study 1: Two Sample t-tests

```
groupS<-PERCENT[JUDGE=="SPOCKS"]
groupNS<-PERCENT[JUDGE!="SPOCKS"]
boxplot(groupS, groupNS,xlab="JUDGE",names=c("SPOCKS","OTHERS"))
```



Case Study 1: One Sample t-test

(Test Assumptions) < 1. Random obs.
2. Approx. Normal pop.

```
#one sample t test
t.test(PERCENT, mu=50)
```

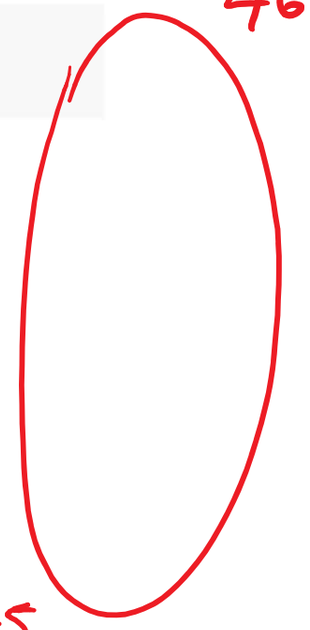
```
##
##  One Sample t-test
##
## data:  PERCENT
## t = -17.303, df = 45, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  23.85675 29.30847
## sample estimates:
## mean of x
##  26.58261
```

μ : true % of women on ventres

46 = n

$H_0: \mu = 50\%$

$H_a: \mu \neq 50\%$

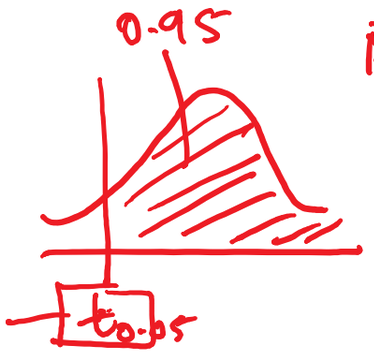


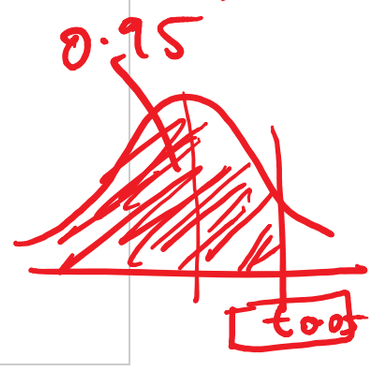
$\bar{x}, s, n$

$t = \frac{\bar{x} - 50}{s/\sqrt{n}} \sim t_{45}$

$P\text{-value} = 2P(T_{45} > |-17.303|)$

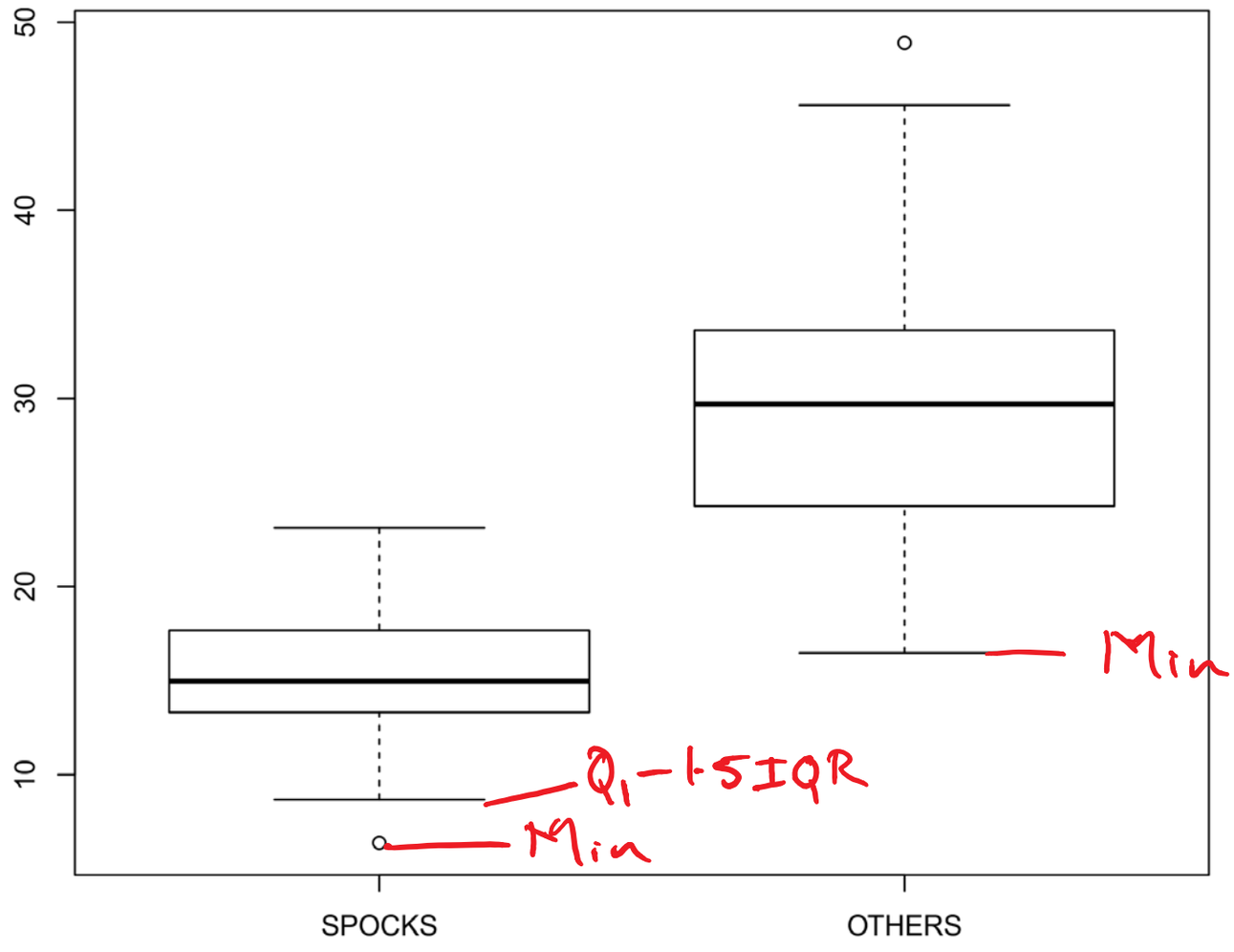
C.I: $\bar{x} \pm t_{45, 0.025} \frac{s}{\sqrt{n}}$

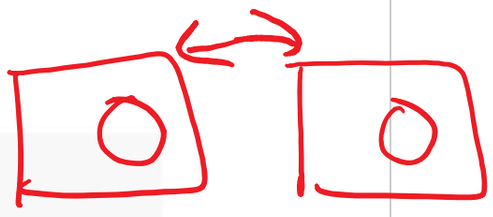




Case Study 1: Two Sample t-tests

```
groupS<-PERCENT[JUDGE=="SPOCKS"]
groupNS<-PERCENT[JUDGE!="SPOCKS"]
boxplot(groupS, groupNS,xlab="JUDGE",names=c("SPOCKS","OTHERS"))
```





1. Independent Samples

2. Samples are from approx. Normal populations.

3. $\sigma_1 = \sigma_2$

of group 1 — of group 2

Case Study 1: Checking equal variance assumption

```
var(groupS)
```

```
## [1] 25.38945
```

```
var(groupNS)
```

```
## [1] 55.21632
```

```
#Rule of Thumb
max(var(groupS),var(groupNS)) /min(var(groupS),var(groupNS))
```

> 4 ?

```
## [1] 2.174775
```

```
max(sd(groupS),sd(groupNS)) /min(sd(groupS),sd(groupNS))
```

> 2 ?

```
## [1] 1.474712
```

Case Study 1: Checking equal variance assumption

```
#F Test of Equal variances
var.test(groupS, groupNS)
```

$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1$ or $\sigma_1^2 = \sigma_2^2$

```
##
##  F test to compare two variances
##
## data:  groupS and groupNS
## F = 0.45982, num df = 8, denom df = 36, p-value = 0.2482
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1789822 1.7739665
## sample estimates:
## ratio of variances
##          0.4598178
```

→ Assume equal variances

Case Study 1: Two sample (unpooled) t-tests

```
#Welch-Satterthwaite (Unpooled)
t.test(groupS, groupNS,var.equal=F)
```

Assume $\sigma_1 \neq \sigma_2$

$H_0: \mu_S = \mu_0$ or $\mu_S - \mu_0 = 0$

```
##
##  Welch Two Sample t-test
##
## data:  groupS and groupNS
## t = -7.1597, df = 17.608, p-value = 1.303e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -19.23999 -10.49935
## sample estimates:
## mean of x mean of y
##  14.62222  29.49189
```

Case Study 1: Pooled t-test

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Assume $\sigma_1 = \sigma_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

```
#Pooled
t.test(groupS, groupNS,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  groupS and groupNS
## t = -5.6697, df = 44, p-value = 1.03e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.155294  -9.584045
## sample estimates:
## mean of x mean of y
##  14.62222  29.49189
```

Conc: There is evidence at 5% level that the % women on Spock's judge's venires is less than that of other judges. ($p = 0.00000103$)

Case Study 1: Paired t-test

```r
#Paired
t.test(groupS, groupNS,paired=TRUE)
```

```
## Error in complete.cases(x, y): not all arguments have the same lengt
```

Case Study 1: Pooled t-test (Left tailed)

```r
#Left-tailed Pooled
t.test(groupS,groupNS,alternative="less",var.equal=TRUE)
```

$H_0: \mu_S < \mu_D$

```
##
##  Two Sample t-test
##
## data:  groupS and groupNS
## t = -5.6697, df = 44, p-value = 5.148e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf -10.463
## sample estimates:
## mean of x mean of y
##  14.62222  29.49189
```

Case Study 1: Simple Linear Regression Approach

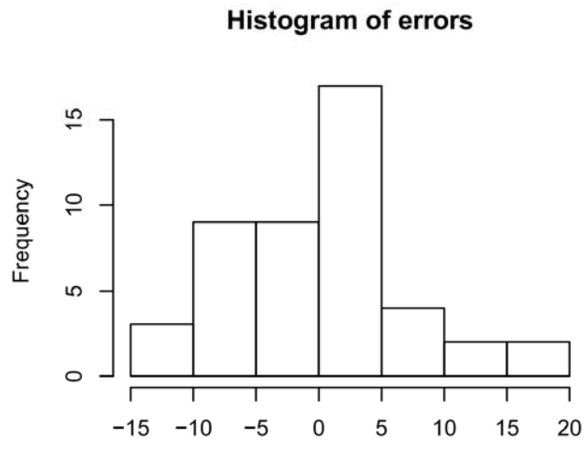
```
X=c(rep(1,length(groupS)), rep(0,length(groupNS)))
Y=PERCENT; model1<-lm(Y~X); summary(model1)
```

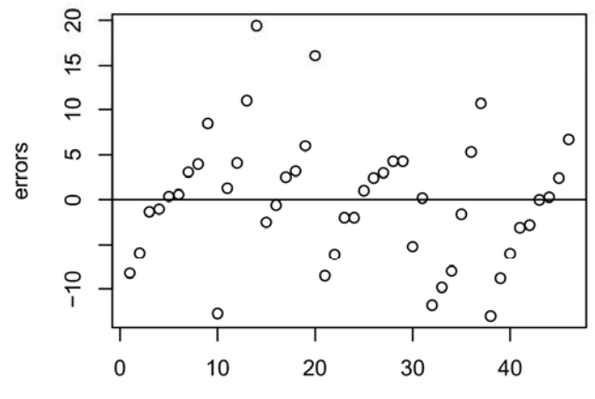
```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9919  -4.6669   0.2581   3.7854  19.4081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.492      1.160   25.42  < 2e-16 ***
## X            -14.870      2.623   -5.67 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.056 on 44 degrees of freedom
## Multiple R-squared:  0.4222, Adjusted R-squared:  0.409
## F-statistic: 32.15 on 1 and 44 DF,  p-value: 1.03e-06
```

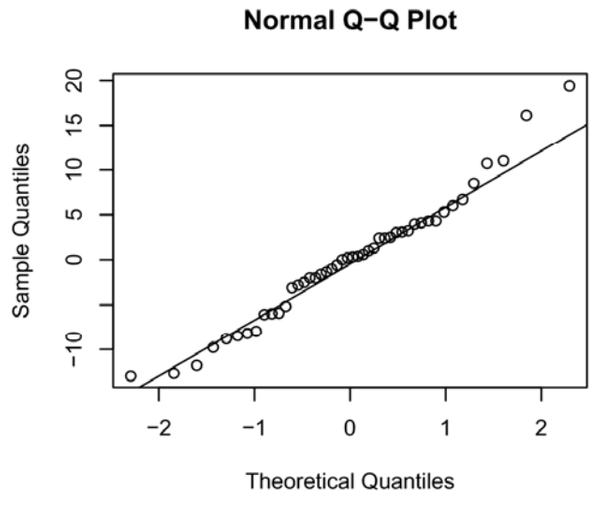
Case Study 1: Regression diagnostics

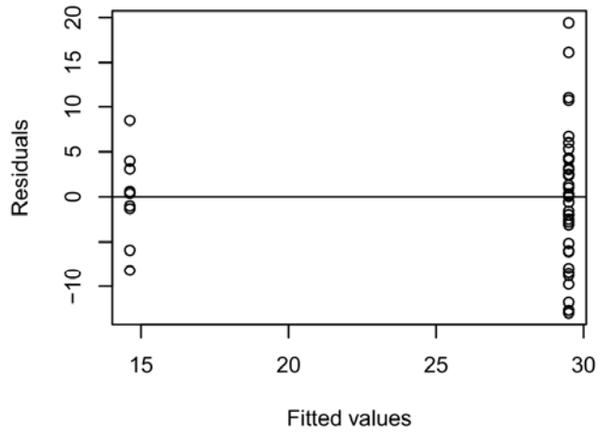
```
yhats=fitted(model1)
errors=residuals(model1)
# par(mfrow=c(2,2)) #partition plot window
# #plot (1,1)- histogram of residuals
# hist(errors, xlab="Residuals", breaks=5)
# #plot(1,2)- residuals vs index(time) with zero line
# # plot(errors)
# abline(0,0)
# #plot(2,1)-normal qq plot of residuals with qqline
# qqnorm(errors)
# qqline(errors)
# #plot(2,2)-residuals vs fitted values with zero line
# plot(yhats, errors, xlab="Fitted values", ylab="Residuals")
# abline(0,0)
```

Case Study 1: Regression diagnostics









Case Study 1: One-way ANOVA approach

```r
#ANOVA approach
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## X          1 1600.6 1600.62  32.145 1.03e-06 ***
## Residuals 44 2190.9   49.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tuesday, January 9, 2018 4:47 PM