STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 11, 2018

Week 1 Topics

REVIEW

-Data summary: Five-number summary, Boxplots

-Large-sample distribution theory: derived from Normal

-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:

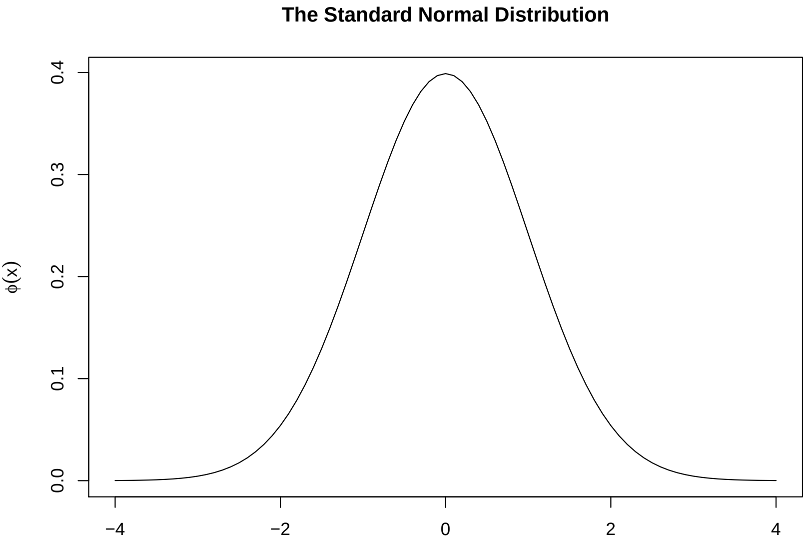
$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp \left(\frac{-1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right)$$

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

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

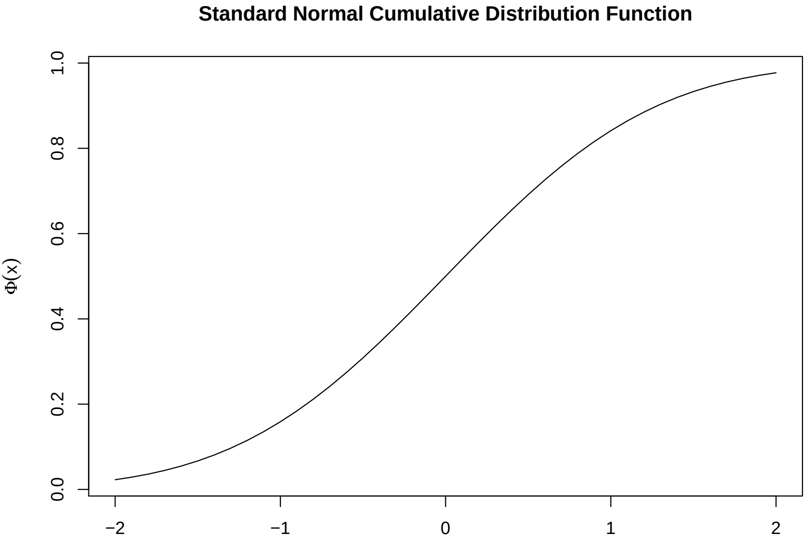
The Standard Normal Distribution

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



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 independent Normals

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

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

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

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

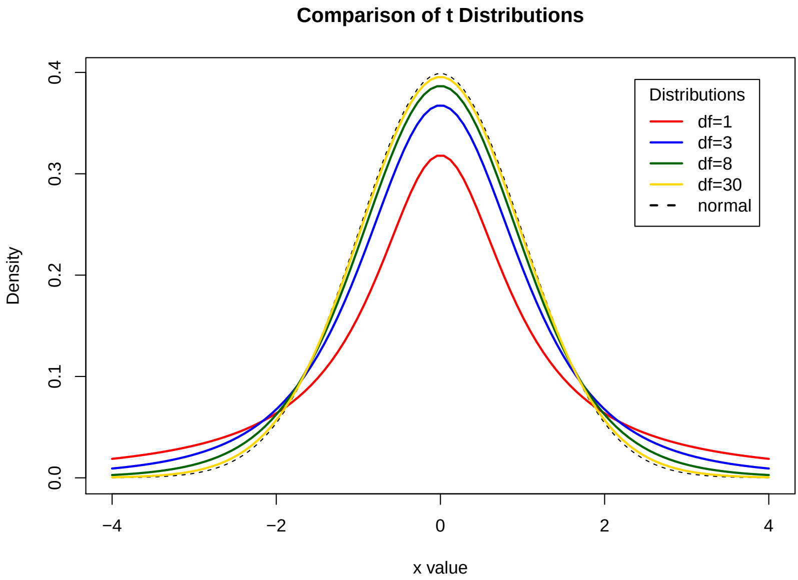
t Distribution

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

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

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

t Distribution



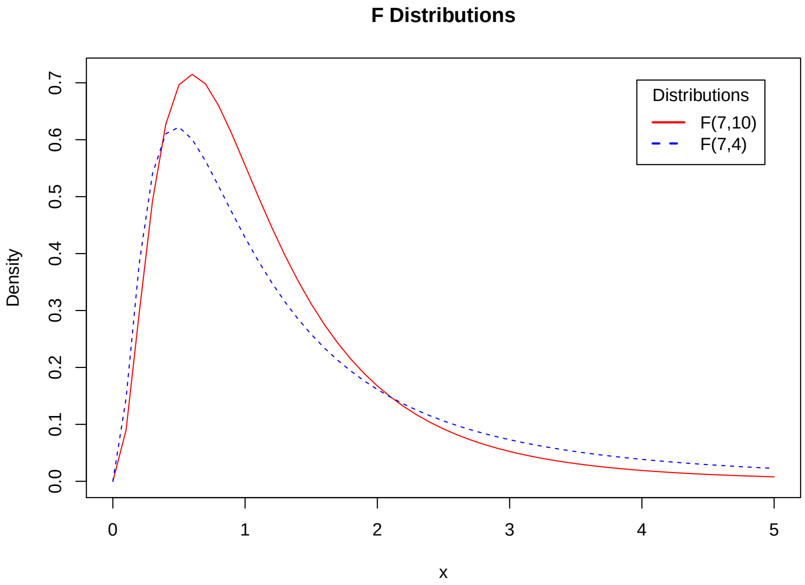
F Distribution

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 $F_{1,n}$.

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

- 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?
- 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.

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

```
##  [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
```

```r
JUDGE
```

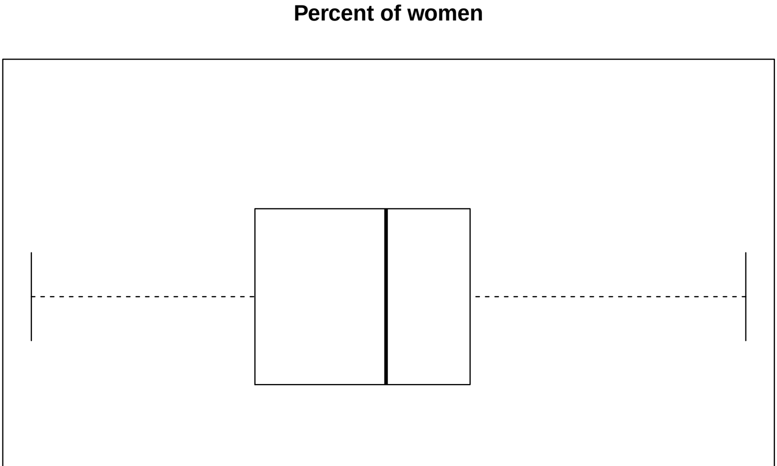
```
##  [1] SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS SPOCKS
## [11] A      A      A      A      B      B      B      B      B
## [21] C      C      C      C      C      C      C      C      C
## [31] D      E      E      E      E      E      E      F      F
## [41] F      F      F    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

Case Study 1: Data summary

```r
summary(PERCENT)
```

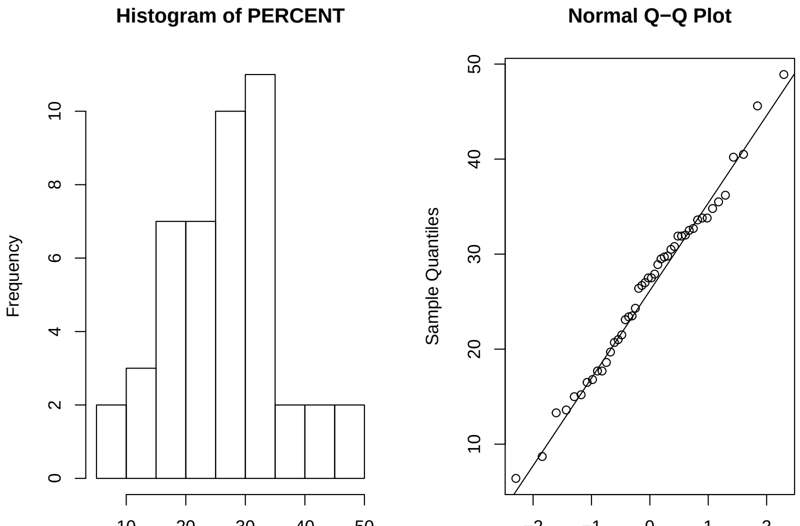
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.40   19.95   27.50   26.58   32.38   48.90 
```

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



Case Study 1: Check Normality

```r
par(mfrow=c(1,2))
hist(PERCENT)
qqnorm(PERCENT)
qqline(PERCENT)
```



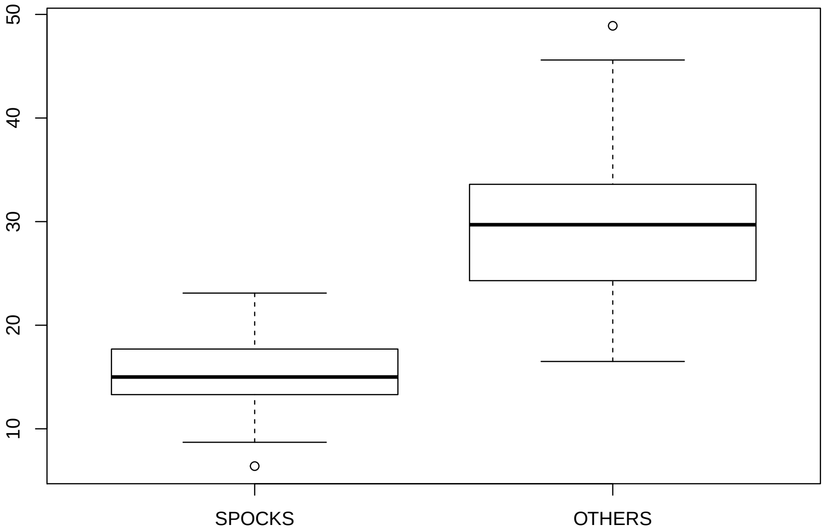
Case Study 1: Check Normality

```r
shapiro.test(PERCENT)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  PERCENT
## W = 0.98763, p-value = 0.9013
```

Case Study 1: Two Sample t-tests

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



Two-sample t-tests

- Purpose: To compare two population means
- Data: Two random samples $X_1, \dots, X_{n_x}$ and $Y_1, \dots, Y_{n_y}$ of sizes n_x and n_y from population 1 and population 2
- Null Hypothesis:

$$H_0 : \mu_x - \mu_y = D_0 \text{ (typically } D_0 = 0)$$

- Assumptions:
  - The two samples are iid from approximately Normal populations.
  - The two samples are independent of each other.
- Test statistic:

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{se(\bar{x} - \bar{y})}$$

Q: How do we estimate this standard error ("se")- standard deviation of $\bar{x} - \bar{y}$?

Case Study 1: Checking equal variance assumption

```r
var(groupS)
```

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

```r
var(groupNS)
```

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

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

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

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

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

Rule of thumb for checking equal variances

- Test:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad vs \quad H_a : \sigma_1^2 \neq \sigma_2^2$$

- Test statistic:

$$\frac{\text{larger sample variance}}{\text{smaller sample variance}} = \frac{S_{max}^2}{S_{min}^2}$$

- If test statistic is greater than 4, reject H_0

Variance Ratio F-test

- special case of Bartlett's test for homogeneity of variances (Bartlett, 1937)
- Null Hypothesis: $H_0 : \sigma_1^2 = \sigma_2^2$
- Underlying assumptions:
  - Random samples of sizes n_1 and n_2 are drawn from Normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively
  - Samples are independent
  - Samples are large (better when samples sizes are equal too)
- **Test statistic**:

$$F = \frac{S_1^2}{S_2^2} \sim_{H_0} F_{n_1-1, n_2-1}$$

- In R: `var.test()`
- For more than 2 variances:
  - `bartlett.test()`
  - Robust alternative: Levene's test (`levene.test()`)

Case Study 1: Checking equal variance assumption

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

```
##
##  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
```

Two-sample t-test (Satterthwaite approximation)

- Used when population variances cannot be assumed to be equal
- Test statistic: under H_0,

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \sim t_\nu$$

  where

$$\nu = \frac{(s_x^2/n_x + s_y^2/n_y)^2}{\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{(s_y^2/n_y)^2}{n_y - 1}}$$

- The *df* (degrees of freedom), ν is calculated by Satterthwaite approximation.
- ν may not be an integer so round down to the nearest integer

Pooled two-sample t-test

- Special case of two-sample t-test
- Assumes population variances are equal
- Pooled variance estimate

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$$

- Test statistic: under H_0

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{s_p^2(\frac{1}{n_x} + \frac{1}{n_y})}} \sim t_{n_x + n_y - 2}$$

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

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

```
##
##  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

```r
#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
```

Case Study 1: Paired t-test

```
#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)
```

```
##
##  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
```

Simple Linear Model Approach (Dummy variable)

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

$$X_i = \mathbb{1}_{A,i} = \begin{cases} 1 & \text{if } i\text{th observation is from "group A"} \\ 0 & i\text{th observation is NOT from "group A"} \end{cases}$$

Assumptions:

- The linear model is appropriate
- Gauss-Markov properties:
  - $E(\epsilon_i) = 0$
  - $\text{Var}(\epsilon_i) = \sigma^2$: Uncorrelated errors
- $\epsilon_i \sim$ Normal

Simple Linear Model: The Hypothesis Test

Test:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0$$

- The slope, β_1, captures the difference in means between groups
- Proof:
  - $E(Y|A) = E(Y|X == 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$
  - $E(Y|A^c) = E(Y|X == 0) = \beta_0 + \beta_1 \times 0 = \beta_0$
  - Hence, $\beta_1 = E(Y|A) - E(Y|A^c) = E(Y|X == 1) - E(Y|X == 0)$

Test statistic: Under the assumptions and H_0,

$$t = \frac{b_1}{se(b_1)} \sim t_{N-2=n_A+n_{others}-2}$$

Case Study 1: Simple Linear Regression Approach

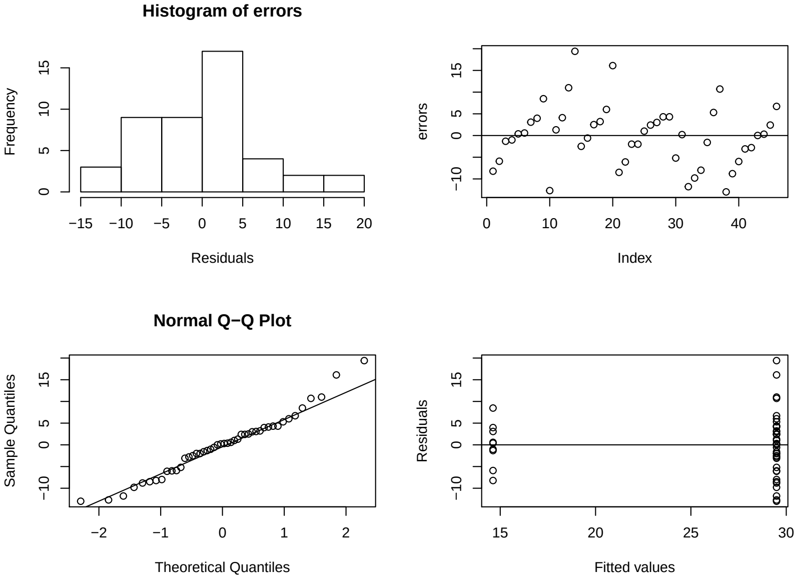
```r
X=c(rep(1,length(groupS)), rep(0,length(groupNS))) #X==1-Spock's judge,
Y=PERCENT; model1<-lm(Y~X); summary(model1)
```

```
## 
## 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

```r
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
```

Case Study 1: Partial results for (Q1)

Sample	SPOCK'S	OTHER
Mean	14.6222	29.4919
Standard deviation	5.0388	7.4308
Sample size	9	37

Hypothesis Test	Partial results
Equal variances assumed	Yes
t-test statistic	-5.67
df	44
P-value	≈ 0
Conclusion	Reject H_0

Notes:

- Equivalence: Pooled 2-sample t is a special case of One-way ANOVA
- Diagnostics: Gauss-Markov assumptions satisfied
- Caution: Unequal sample sizes

Robustness of t

- t-procedures are robust against assumptions of normality.
- In other words, t-procedures are often valid even when the assumption of normality is violated.
- They are not robust against strong skewness or outliers
- Can be used when sample size is small

- Non-parametric tests or "Distribution free" tests do not require that data follow any specific distribution.

Non-parametric alternatives

Gaussian	"Distribution free"
1-sample t	Sign test, Wilcoxon signed-rank test
2-sample t	Wilcoxon rank-sum test

In R: See `wilcox.test()`

R functions used

```
   summary()
   plot()
   boxplot()
t.test()
   pnorm()
   qqnorm()
   qqline()
   shapiro.test()
var.test()
   lm()
   fitted()
   residuals()
anova()
```