STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2019

Dr. Shivon Sue-Chee

UNIVERSITY OF TORONTO

January 10, 2019

Week 1 Topics

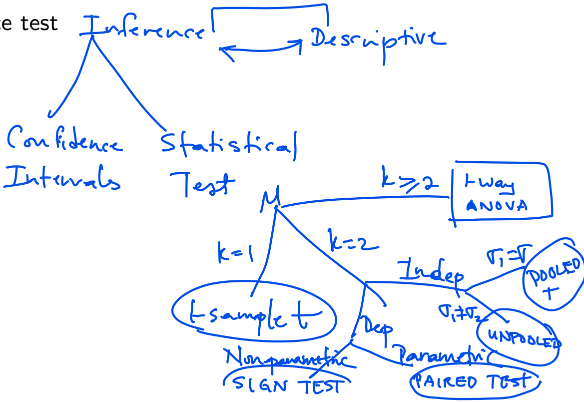
REVIEW

-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

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

numeric, cts

Case Study 1: Two Key Questions

① Spock's vs Others

② Amg others

- Q1. Is there evidence that women are underrepresented on Spock's judge's venires when compared to other judges? $k=2 \rightarrow$ 2-sample
- Q2. Is there evidence that there are differences in women's representation in venires of the other 6 judges? $k=6>2 \rightarrow$ 1-way ANOVA
- 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?

$k=7>2 \rightarrow$ 1-way ANOVA

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
```

$Q_1 - 1.5IQR$ 46 obs.

```
##  [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
```

43rd   46th

7 judges

$n_{spocks} = 9$

$n_{others} = 37$

```
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)
```

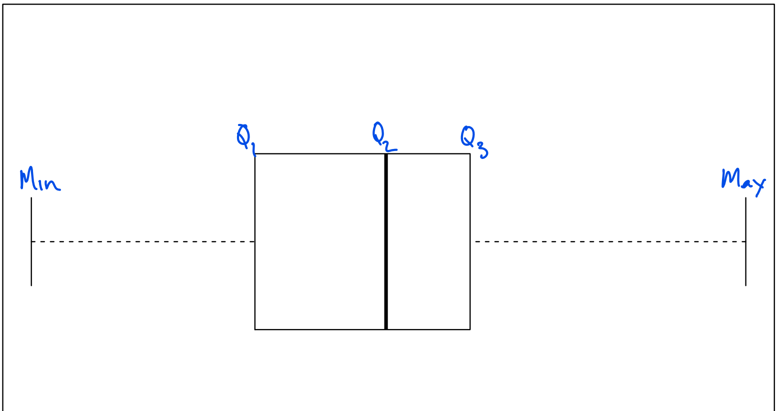
```
         1       2          3                 4         5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.40   19.95   27.50   26.58   32.38   48.90
```

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

Range = Max - Min

Percent of women

IQR = $Q_3 - Q_1$



Case Study 1: One Sample t-test

Assumptions: (1) Random sample
(2) Data is approximately Normal (or CLT applies)

Inf. → C.I, H.T

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

data $\uparrow$ μ_0 $\uparrow$

In R: ?t.test   help(t.test)

```
##
## 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
```

$\bar{x}$ (mean of x)

Claims (1) $H_0: \mu = \mu_0 = 50$

$H_a: \mu \neq 50$

(3) P-value ≈ 0

$p < 0.00001$

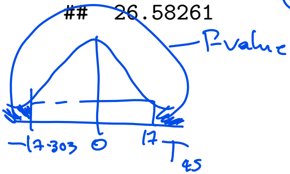
T.S. (2) $t = -17.303$

(4) At 5% level of significance, there is evidence that women are not fairly represented on the venires of the 7 judges.

$\bar{x} \pm t_{45,0.025} \dfrac{s}{\sqrt{n}}$

(2) $t = \dfrac{\bar{x} - \mu_0}{\sqrt{\dfrac{s^2}{n}}} = \dfrac{26.6 - 50}{s/\sqrt{46}} \sim T_{n-1}$

$46 - 1 = 45$



Case Study 1: Check Normality

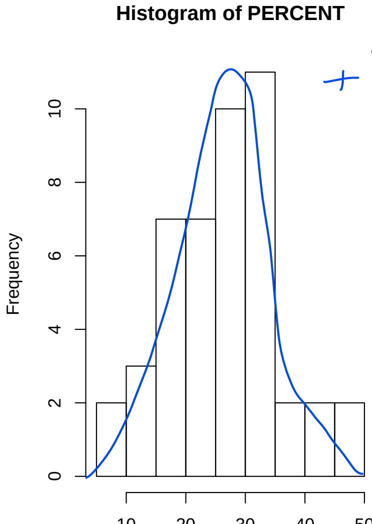
```
par(mfrow=c(1,2))
hist(PERCENT)
qqnorm(PERCENT)
qqline(PERCENT)
```

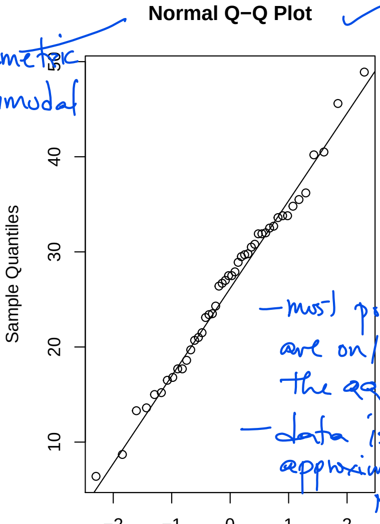
Pop is Normal — Data is Normal ✓

Pop. is NOT NORMAL or Unknown — large n — $\bar{X} \sim$ Normal by CLT ✓

small n ⊗

Symmetric + Unimodal





- most points are on/near the qq line

- data is approximately Normal

Case Study 1: Check Normality

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

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

1 H_0: Data is Normal

H_a: Data is NOT Normal

2 Test Statistic

3 large

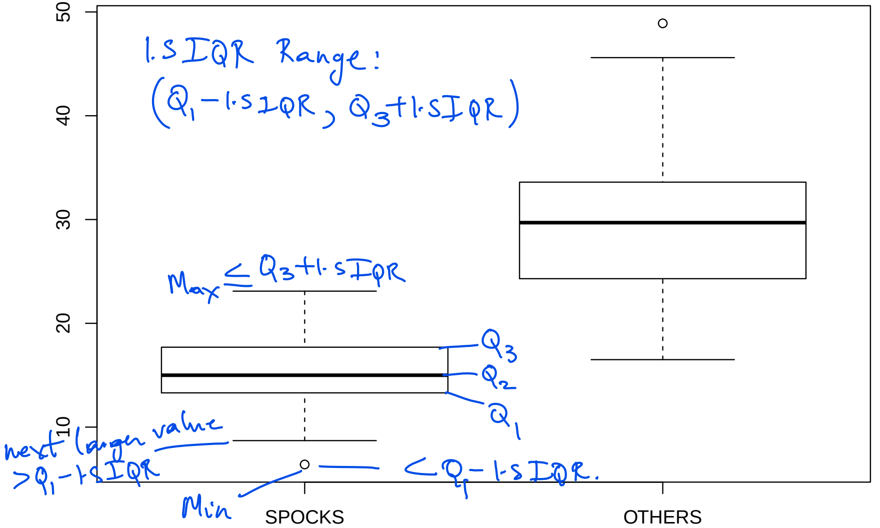
4 Evidence that data is Normal.

Case Study 1: Two Sample t-tests

$n_S = 9$

$n_J = 37$

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



Two-sample t-tests

μ_1, μ_2

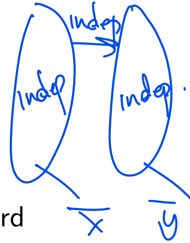
- Purpose: To compare two population means
- Data: Two random samples $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$ of sizes n_x and n_y from population 1 and population 2
- Null Hypothesis:

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

$\mathrm{Var}(ay) = a^2 \mathrm{Var}\, y$

- 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 $\bar{x} - \bar{y}$?

$$\mathrm{Var}(\bar{x} - \bar{y}) = \mathrm{Var}(\bar{x}) + \mathrm{Var}(\bar{y}) = \sigma_x^2 / n_x + \sigma_y^2 / n_y$$

$$\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}$$

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
```

$\frac{S^2_{max}}{S^2_{min}} < 4$

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

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

$\frac{S_{max}}{S_{min}} < 2$

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:

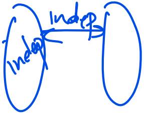
$$\frac{\text{larger sample variance}}{\text{smaller sample variance}} = \frac{S_{max}^2}{S_{min}^2}$$

- If test statistic is greater than 4, reject H_0

$$\frac{S_{max}}{S_{min}} > 2$$

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**:

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

$$F_{v,w} = \frac{\chi^2_v / v}{\chi^2_w / w}$$

$$\frac{(n_1-1) S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$$

$$\frac{(n_2-1) S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

- 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

$H_0: \sigma_1^2 = \sigma_2^2$

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

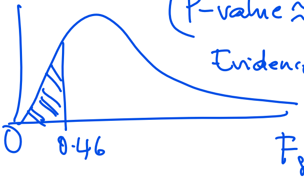
$n_1 = 9$

$n_2 = 37$

```
##
##  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
```

(P-value ≈ 0.25) $>$ ($\alpha = 0.10$)

Evidence that the pop. variances are the same.



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

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

$(n_x - 1) + (n_y - 1)$

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

Assume $\sigma_1 \neq \sigma_2$

```
#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
```

small

$\bar{x}_S$ $\bar{x}_O$

Evidence that the %'s of women is different for Spock's judge versus that of the other judges

Case Study 1: Pooled t-test

```
#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
```

$37+9-2=44$

small

– Same conclusion

$\bar{x}_S$ $\bar{x}_O$

Case Study 1: Paired t-test

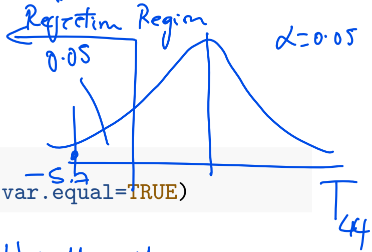
	Before	After
P_1	⊙	⊙
P_2		
⋮		
P_n		

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

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

- Equal sample sizes
- Dependent samples

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
```

$H_0: \mu_S = \mu_0$

$H_a: \mu_S < \mu_0$

Concl.

Evidence that the % of women on venires of Spock's judge is less than that of the other 6 judges.

Simple Linear Model Approach (Dummy variable)

Model:

response

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

1 predictor

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

check using Residual plots

$$\boxed{e_i \sim N(0, \sigma^2)}$$

Simple Linear Model: The Hypothesis Test

$\mu_A - \mu_{A^c}$

Test:

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

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

- 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, $\beta_1 = \underbrace{E(Y|A)}_{\mu_A} - \underbrace{E(Y|A^c)}_{\mu_{A^c}} = E(Y|X == 1) - E(Y|X == 0)$

μ_A μ_{A^c}

$X = \begin{cases} 1 \\ 0 \end{cases}$

$N = 46$

Test statistic: Under the assumptions and H_0,

$$t = \frac{b_1 - 0}{se(b_1)} \sim t_{N-2=n_A+n_{others}-2}$$

x_i	y_i
1	6.9
1	8.7
⋮	⋮
0	⋮

Case Study 1: Simple Linear Regression Approach

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

summary(lm(y~x))

$n_s = 9$

$n_0 = 37$

$N = 9+37 = 46$

```
##
## 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
```

$\hat{\beta}$s

$\hat{\beta}_0 = b_0$

b_1

$H_a: \beta_1 \neq 0$

$-5.67 = \frac{-14.87}{2.623}$

0.00000103

Same Conclusion

Case Study 1: Regression diagnostics

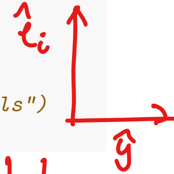
from linear model — $\hat{y}_i = b_0 + b_1 x_i$

y_i — observed data response values

lm(y~x)

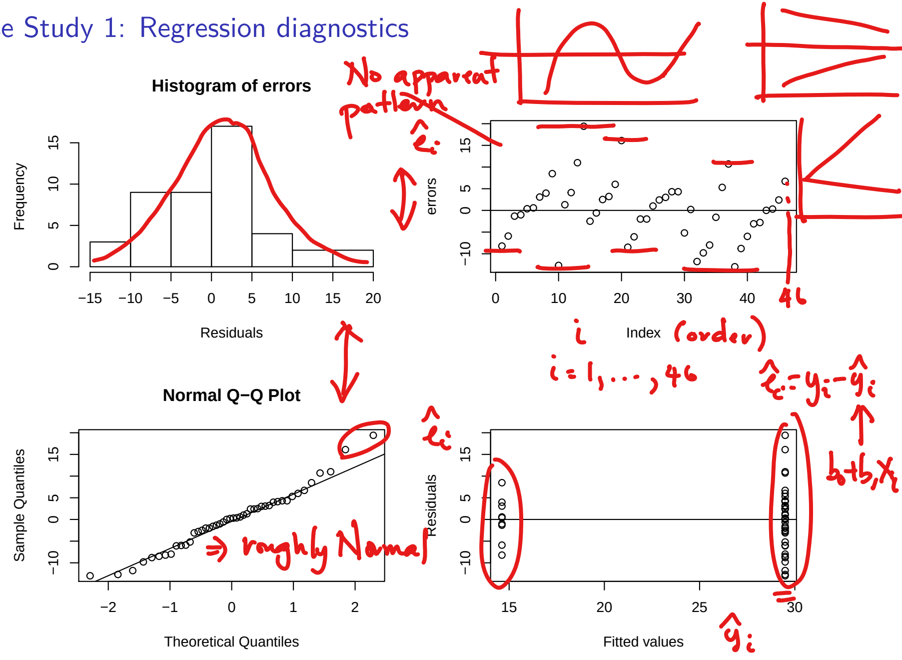
$\boxed{\hat{e}_i = y_i - \hat{y}_i}$

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



In R: (H-W) plot(model1) — 4 residual plot

Case Study 1: Regression diagnostics



Case Study 1: One-way ANOVA approach

$Z^2 \equiv \chi^2_1$

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

$t^2_{44} \equiv F_{1,44}$

```
## 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
```

0.00000103

Case Study 1: Partial results for (Q1)

	Sample	SPOCK'S	OTHER
$\bar{x}$	Mean	14.6222	29.4919
s	Standard deviation	5.0388	7.4308
n	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

$\rightarrow (-5.67)^2 = 32.145$

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

permutation

In R: See `wilcox.test()`

R functions used

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

A #1

II UTH ends with '00':

last 4 digits

$\boxed{\begin{array}{l}\text{set.seed(1234)}\\ \text{sample (1:99,1)}\end{array}}$

$Q_1 - 1.5\,IQR$

$Q_3 + 1.5\,IQR$