Hypothesis Testing

Chaklam Silpasuwanchai

Asian Institute of Technology

chaklam@ait.asia

January 2, 2020

Overview

Sources

- Mackenzie, Chapter 6, **Hypothesis Testing**, Human Computer Interaction: An Empirical Research Perspective, 1st ed. (2013)
- Yatani, Advanced Topics in Human-Computer Interaction, `http://yatani.jp/teaching/doku.php?id=2016hci:start`

Statistical Procedures

- Statistical tests comes in two forms: **parametric** and **non-parametric test**
- Parametric tests operate on data based on assumptions of a normal distribution or the t-distribution.
- Non-parametric tests can operate on any data.
- Parametric tests are superior to non-parametric tests, but require assumptions of distributions.

Measurement Scale	Defining Relations	Examples of Appropriate Statistics	Appropriate Statistical Tests
Nominal	• Equivalence	• Mode • Frequency	• Non-parametric tests
Ordinal	• Equivalence • Order	• Median • Percentile	
Interval	• Equivalence • Order • Ratio of intervals	• Mean • Standard deviation	• Parametric tests • Non-parametric tests
Ratio	• Equivalence • Order • Ratio of intervals • Ratio of values	• Geometric mean • Coefficient of variation	

Figure: Source: Fg. 6.1 (Mackenzie): Measurement scales of data, properties of data, and appropriate statistical tests

Type of data

- Nominal: techniques, gender, occupation
- Ordinal: likert-scale question; each interval is not equal
- Interval: tmperature; no exact zero
- Ratio: height/weight, speed, accuracy, time

Analysis of Variance

- **ANOVA**, or F-test, is the main statistical test for factorial experiment
- The main motivation to use statistical test is - if we see a difference in mean, is that difference **occur by chance** or is **significant**?
- The way to achieve is to determine whether IV has a significant effect on IV is based on **variances**
- Some definition: **Null hypothesis** is an assumption of no difference - e.g., there is no difference in the mean time between keyboard and mouse.

Example: One-factor design with 2 levels

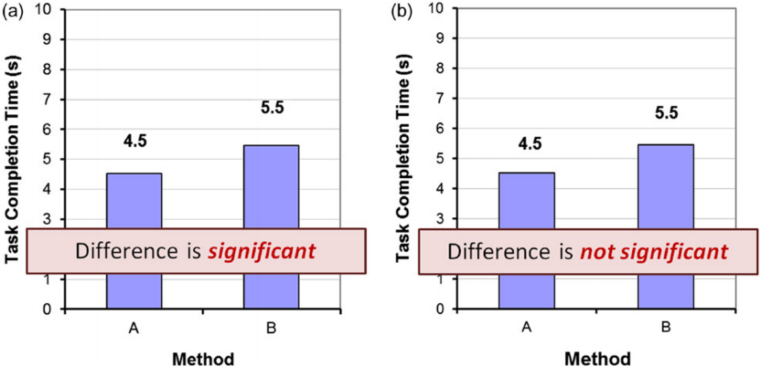


FIGURE 6.2

Difference in task completion time (in seconds) across two test conditions, Method A and Method B. Two hypothetical outcomes are shown: (a) The difference is statistically significant. (b) The difference is not statistically significant.

Figure: Source: Fg. 6.2 (Mackenzie)

Example: One-factor design with 2 levels

(a)

Participant	Method A	Method B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

(b)

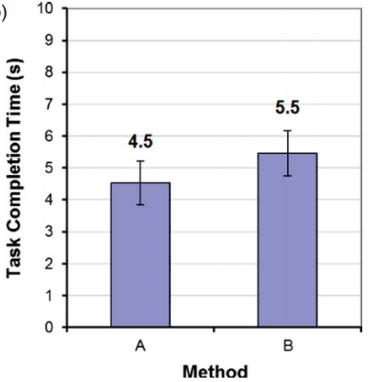


FIGURE 6.3

(a) Data for simulation in Figure 6.2a. (b) Bar chart with error bars showing ±1 standard deviation.

Figure: Source: Fg. 6.3 (Mackenzie)

Example: One-factor design with 2 levels

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

FIGURE 6.4

Analysis of variance table for data in Figure 6.3a.

Figure: Source: Fg. 6.4 (Mackenzie): P-value of 0.0121 means that there is less than 2% that the difference occurs by chance. By convention requires less than 0.05 to reject null hypothesis

> The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9}$ = 9.80, p < .05).

FIGURE 6.5

Example of how to report the results of an analysis of variance in a research paper.

Figure: Source: Fg. 6.5 (Mackenzie): F-value is calculated = between-group variances / within-group variances = 4.232 / .432

Example: One-factor design with 2 levels

Reporting format:

- Note that when we report the p value, p is cited as less than a more conservative threshold from the set .05, .01, .005, .001, .0005, .0001. Thus p is cited as $p < .05$ rather than $p = .0121$
- Must **strictly** follow the exact format for reporting - use parentheses, uppercase for F, lowercase for p, italics for F and p, space on both sides of equal sign, space after comma, space on both sides of the less than sign, degrees of freedom are subscript, three/four significant digits for F statistics, does not require 0 in front of p-value.

Example: One-factor design with 2 levels

(a)

Participant	Method A	Method B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

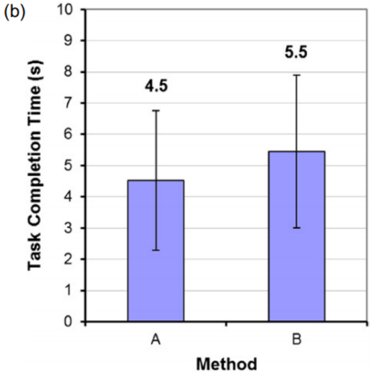


FIGURE 6.6

(a) Data for simulation in Figure 6.2b. (b) Bar chart with error bars showing ±1 standard deviation.

Figure: Source: Fg. 6.6 (Mackenzie)

Example: One-factor design with 2 levels

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

FIGURE 6.7

Analysis of variance for data in Figure 6.3b.

Figure: Source: Fg. 6.7 (Mackenzie). $F = 4.324/6.904 = .626$. Given p-value of .4491, there is around 45% that the difference occurs by chance.

> The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variances ($F_{1,9}$ = 0.626, ns).

FIGURE 6.8

Reporting a non-significant ANOVA result.

Figure: Source: Fg. 6.8 (Mackenzie). It means that we have not enough evidence to reject null hypothesis, but it **does not mean that null hypothesis is true either**.

Effect Size

Statistical considerations:

- A **t test** is the same as an analysis of variance if only two levels
- **Type 1** error: false positive (say difference when there isn't); **Type II** error: false negative (say no difference when there is)
- To fix Type I error, we check p value against our predefined significance level; this significance level is also called **alpha** (commonly .05 or less)
- To fix Type II error, we can report the **effect size** which measures how "strong" is the significance. SPSS reports **Partial Eta Squared** (η_p^2) - .02 means that the factor X by itself accounted for only 2% of the overall (effect + error) variance. Usually around > 0.06 is considered moderate, while > 0.14 is large.
- **Power analysis** can be used prior to the experiment, to determine the sample size needed. You need effect size, significance level(.05) and power (.7). Note that this effect size could be eta-square, Cohen's d, or omega-squared depending on the tools you use.

Effect Size

Statistical considerations:

- The most common metrics are **eta squared** and **partial eta squared**. The eta squared is a correlation ratio of SS-effect / SS-total, where SS-effect is sum of squares for the factor while SS-total is the total sum of square
- However, when you have many IVs, eta-squared becomes too small. To address this, we use partial eta-squared which is a ratio of SS-effect / (SS-effect+SS-error). Here SS-error means the sum of squre for the error term.
- For eta-squared, 0.01 is small, 0.06 is medium, and 0.14 is large
- For partial eta-squared, 0.01 (small), 0.09 (medium), and 0.25 (large)
- Normally, eta-squared is fine for HCI experiment with around 2 to 3 IVs

Example: One-factor design with 4 levels

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23

Figure: Source: Fg. 6.9a (Mackenzie)

Example: One-factor design with 4 levels

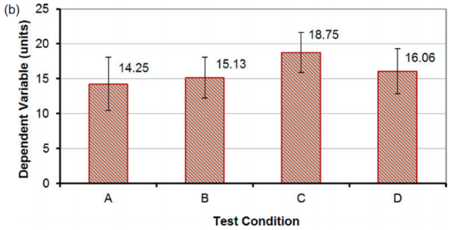


Figure: Source: Fg. 6.9b (Mackenzie)

ANOVA Table for Dependent Variable (units)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

Figure: Source: Fg. 6.9c (Mackenzie)

Example: One-factor design with 4 levels

To determine exactly which condition is different with which condition, a posthoc analysis is required - the most common method is either a Tukey's test or pairwise comparison with the Bonferroni correction

Scheffe for Dependent Variable (units)
Effect: Test Condition
Significance Level: 5 %

	Mean Diff.	Crit. Diff.	P-Value	
A, B	-.875	3.302	.9003	
A, C	-4.500	3.302	.0032	S
A, D	-1.813	3.302	.4822	
B, C	-3.625	3.302	.0256	S
B, D	-.938	3.302	.8806	
C, D	2.688	3.302	.1520	

Figure: Source: Fg. 6.11 (Mackenzie)

Example: Between-subjects designs

To check whether handedness has a effect on task completion time.

(a)

Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

(b)

Handedness	Task Completion Time (s)	
	Mean	*SD*
Left	22.0	1.93
Right	19.9	2.42

(c)

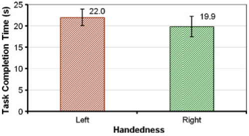


Figure: Source: Fg. 6.12 (Mackenzie)

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Pow er
Handedness	1	18.063	18.063	3.781	.0722	3.781	.429
Residual	14	66.875	4.777				

Figure: Source: Fg. 6.13 (Mackenzie)

Two-way analysis of variance

- Experiments with two IVs (factors) is called a **two-way design**
- Analysis of variance of two-way design tests **main effects** of each factor and **interaction effect**
- Interaction effect indicates an **relationship** between the IV which the relationship is shown in their effect on DV
- If experiment has two factors, three possibilities are possible: 2 factors are both within-subject, or both are between-subject, or one is within and another is between.

Interaction effects

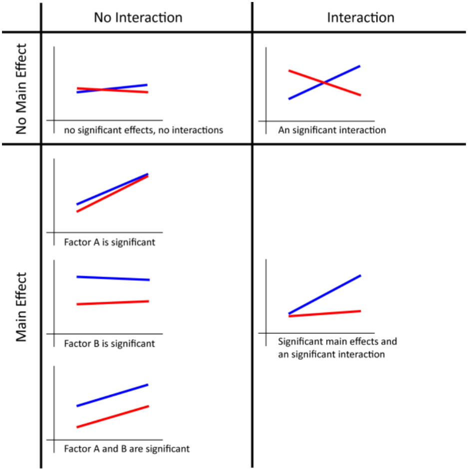


Figure: Source: Yatani's post-hoc tests

Example: 3 x 2 within-subjects design

Let's take both factors as within-subjects, the first factor is device with 3 levels - mouse, trackball, and stylus, and second factor is task with 2 levels - point-select and drag-select. We called this a 3 x 2 within-subjects design.

(a)

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

(b)

	Task 1	Task 2	*Mean*
Device 1	15.4	18.5	17.0
Device 2	15.8	15.3	15.6
Device 3	15.4	12.2	13.8
Mean	15.6	15.3	15.4

(c)

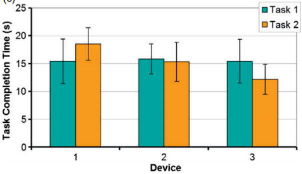


Figure: Source: Fg. 6.14 (Mackenzie)

Example: 3 x 2 within-subjects design

Three effects were observed - the main effect of device and task, and the interaction effect between device and task.

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	11	134.778	12.253				
Device	2	121.028	60.514	5.865	.0091	11.731	.831
Device * Subject	22	226.972	10.317				
Task	1	.889	.889	.076	.7875	.076	.057
Task * Subject	11	128.111	11.646				
Device * Task	2	121.028	60.514	5.435	.0121	10.869	.798
Device * Task * Subject	22	244.972	11.135				

Figure: Source: Fg. 6.15 (Mackenzie)

Example: 3 x 2 within-subjects design

Reporting:

The grand mean for task completion time was 15.4 seconds. Device 3 was the fastest at 13.8 seconds, while device 1 was the slowest at 17.0 seconds. The main effect of device on task completion time was statistically significant ($F_{2,22}$ = 5.865, p <.01). The task effect was modest, however. Task completion time was 15.6 seconds for task 1. Task 2 was slightly faster at 15.3 seconds; however, the difference was not statistically significant ($F_{1,11}$ = 0.076, ns). The results by device and task are shown in Figure x. There was a significant Device × Task interaction effect ($F_{2,22}$ = 5.435, p < .05), which was due solely to the difference between device 1 task 2 and device 3 task 2, as determined by a Scheffé post hoc analysis.

Figure: Source: Fg. 6.16 (Mackenzie)

Analysis of variance for counterbalancing testing

- Let's say we counterbalance a single factor with two levels (A and B). Group 1 (G1) does AB while Group 2 (G2) does BA.
- How can we confirm that the counterbalancing works? We can use analysis of variance to check
- We set a 2 x 2 mixed design with one within-subjects factor (method: A and B) and one between-subjects factor (groups: AB and BA)
- If the ANOVA shows that group effect is not significant, it means our counterbalancing works.
- If there is an interaction effect between method and group, it represents a phenomenon known as **asymmetric skills transfer**, meaning that it was different transitioning from A to B than from B to A

Chi-square test

- Is a **non-parametric tests** that is common for working with **counts/frequencies**
- The test compares the observed values, against the expected values, which are developed under the assumption that there is no difference among groups.
- Is non-parametric because it **does not work under any assumption of probability distribution**

Example: Counts of usage

Consider how males and females scroll, either with mouse wheel (MW), clicking and dragging the scrollbar (CD), or keyboard (KB),

(a)

Observed Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	28	15	13	56
Female	21	9	15	45
Total	49	24	28	101

(b)

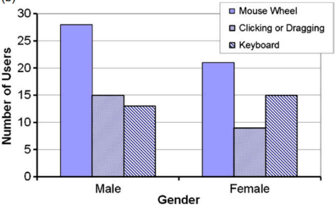


Figure: Source: Fg. 6.22 (Mackenzie)

Example: Counts of usage

To determine the effect, the chi-square test is used. The test statistics is written as χ^2 using the lowercase Greek letter *chi*. Each expected value is the row total multiplied by the column total, divided by the grand total. For example, Male-MW expected value is (56 x 49) /101 = 27.2. Each chi square is simply the square of (observed - expected) / expected. For example, Male-NW chi square is $(28.0 - 27.2)^2/27.2 = 0.025$

(a)

Expected Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	27.2	13.3	15.5	56.0
Female	21.8	10.7	12.5	45.0
Total	49.0	24.0	28.0	101

(b)

Chi Squares				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	0.025	0.215	0.411	0.651
Female	0.032	0.268	0.511	0.811
Total	0.057	0.483	0.922	**1.462**

Figure: Source: Fg. 6.23 (Mackenzie)

Example: Counts of usage

Similar to F-statistic, we need to determine the critical value. Given alpha of .05, and the degrees of freedom of (r-1)(c-1) = 2, we got the critical value as 5.99. Since the observed value (1.462) is not larger than 5.99, there are no significant differences. (*Note that if there is a difference, a similar posthoc test can be subsequently performed to determine the differences between methods*)

Significance Threshold (α)	Degrees of Freedom							
	1	2	3	4	5	6	7	8
.1	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36
.05	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51
.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09
.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13

Figure: Source: Fg. 6.24 (Mackenzie)

Non-parametric tests for ordinal data

- **Non-parametric tests** make no assumptions for probability distribution thus are applicable to a wider range of data than parametric tests
- Downsides of non-parametric tests are **loss of information** - while parametric tests works on interval or ratio data, non-parametric tests deal with ordinal data (ranks)
- Non-parametric tests ignore any property of the scale of data except ordinality
- For example, 49, 81, 82 are transformed to 1, 2, 3
- In HCI, non-parametric tests are often used for **questionnaires data** (e.g., using Likert scale) since they are **ordinal** data. Though non-parametric tests are limited to single factor analysis.

Non-parametric tests for ordinal data

Four most common non-parametric procedures that work based on the number of conditions and design (*Note that conditions here refer to* **levels** *of a single factor. Also note that since Kruskal-Wallis aand Friedman operate or 3 or more conditions, a statistically significant outcome is usually followed with a post hoc pairwise comparisons*)

Design	Conditions	
	2	3 or more
Between-subjects (independent samples)	Mann-Whitney U	Kruskal-Wallis
Within-subjects (correlated samples)	Wilcoxon Signed-Rank	Friedman

Figure: Source: Fg. 6.29 (Mackenzie)

Example: Mann-Whitney U

10 Mac users and 10 PC users are interviewed about their political views on a 10-point linear scale (1 = very left, 2 = very right). Turns out PC users are a little more "right-leaning"!

Mac Users	PC Users
2	4
3	6
2	5
4	4
9	8
2	3
5	4
3	2
4	4
3	5

Figure: Source: Fg. 6.30 (Mackenzie)

Example:Mann-Whitney U

- Here, the data are potentially interval-scale but the intervals between successive codes are not equal, as the difference between 1 and 2 and 3 and 4 may not be equal (unlike temperature). Since the data is at least ordinal, non-parametric tests are appropriate. Given 2 levels and between subject designs, **Mann-Whitney U** is suitable
- Here we found that $p = .1418$, thus we conclude that no differences were found.

(a)

Mann-Whitney U for Response
Grouping Variable: Category for Response

U	31.000
U Prime	69.000
Z-Value	-1.436
P-Value	.1509
Tied Z-Value	-1.469
Tied P-Value	.1418
# Ties	4

Figure: Source: Fg. 6.31 (Mackenzie)

Example: Wilcoxon Signed-Rank

10 users rated the design of two media players on a 10-point linear scale (1 = not cool, 10 = really cool). Which test should we use?

Mac Users	PC Users
2	4
3	6
2	5
4	4
9	8
2	3
5	4
3	2
4	4
3	5

Figure: Source: Fg. 6.32 (Mackenzie)

Example: Wilcoxon Signed-Rank

The Wilcoxon Signed-Rank test found that $p = .0242$, thus we conclude that no differences were found.

(a)

Wilcoxon Signed Rank Test for MPA, MPB

# 0 Differences	2
# Ties	2
Z-Value	-2.240
P-Value	.0251
Tied Z-Value	-2.254
Tied P-Value	.0242

Figure: Source: Fg. 6.33 (Mackenzie)

Example: Kruskal-Wallis

Is it significant?

A20-29	A30-39	A40-49
9	7	4
9	3	5
4	5	5
9	3	2
6	2	2
3	1	1
8	4	2
9	7	2

Figure: Source: Fg. 6-34 (Mackenzie).

(a)

Kruskal-Wallis Test for Acceptability
Grouping Variable: Category for Preference

DF	2
# Groups	3
# Ties	7
H	9.421
P-Value	.0090
H corrected for ties	9.605
Tied P-Value	.0082

Figure: Source: Fg. 6-35 (Mackenzie).

Example: Kruskal-Wallis

Since there are three conditions, we can further run post-hoc tests to find out the differences in pair. Here, we found the difference between group 1 and 3.

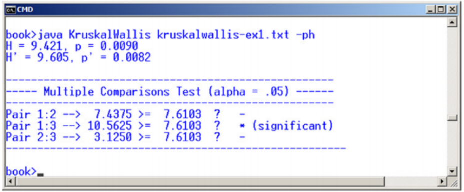


Figure: Source: Fg. 6.36 (Mackenzie)

Example: Friedman Test

So, what's the conclusion?

Participant	A	B	C	D
1	66	80	67	73
2	79	64	61	66
3	67	58	61	67
4	71	73	54	75
5	72	66	59	78
6	68	67	57	69
7	71	68	59	64
8	74	69	69	66

Friedman Test for 4 Variables

DF	3
# Groups	4
# Ties	2
Chi Square	8.475
P-Value	.0372
Chi Square corrected for ties	8.692
Tied P-Value	.0337

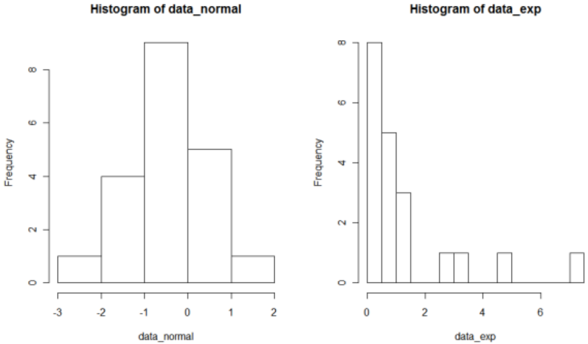
CMD

```
book>java Friedman friedman-ex1.txt -ph
H(3) = 8.475, p = 0.0372
H'(3) = 8.692, p' = 0.0337

---------------------------------------------------------------------
----------- Pairwise Comparisons (using Conover's F) -----------
---------------------------------------------------------------------
Pair 1:2 --> abs( 3.063 - 2.438) > 1.132 ?     -
Pair 1:3 --> abs( 3.063 - 1.438) > 1.132 ?     * (significant)
Pair 1:4 --> abs( 3.063 - 3.063) > 1.132 ?     -
Pair 2:3 --> abs( 2.438 - 1.438) > 1.132 ?     -
Pair 2:4 --> abs( 2.438 - 3.063) > 1.132 ?     -
Pair 3:4 --> abs( 1.438 - 3.063) > 1.132 ?     * (significant)
---------------------------------------------------------------------

book>
```

Figure: Source: Fg. 6-(37-39) (Mackenzie).

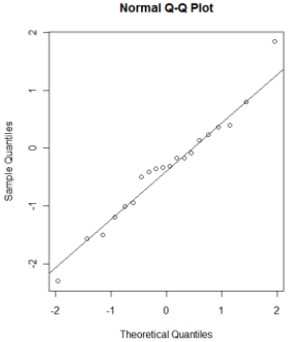
Normality check

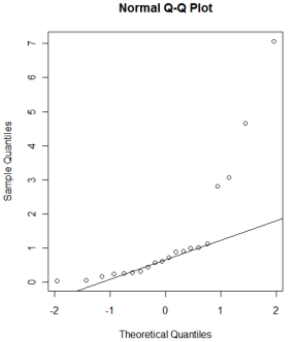
- To decide whether to use parametric vs. non-parametric tests largely depend on the assumption of normality and homogenity of variances. Normally, if your data is ratio, it is safe to use parametric tests. However, if unsure, one can also check for normality
- First easy way is to use **histogram** to check skewness



Normality check

- Another way is to use **Q-Q plot**. It plots your data against an ideal normal distribution. If the data points align approximately well with the line, it is normal.





Normality check

- Another more mathematical way is statistical test which measures the difference of observed data against a known normal distribution. Two common tests for normality is **Shapiro Wilk** and **Kolmogorov-Smirnov** test
- Shapiro-Wilk is more appropriate for small sample sizes (< 50)
- Both of this test can be easily done in SPSS
- For example, the null hypothesis of Shapiro-Wilk is that samples are taken from a normal distribution. Here, the p-value is larger than .05, thus is safe to say it's normal. The null hypothesis is same for Kolmogorov-Smirnov

Tests of Normality

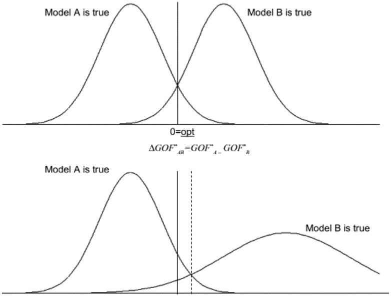
Course		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Time	Beginner	.177	10	.200*	.964	10	.827
	Intermediate	.166	10	.200*	.969	10	.882
	Advanced	.151	10	.200*	.965	10	.837

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

Homogeneity of variances

- The basic idea is that the variance in each group should be similar enough, if not, anova does not make sense
- Tests that can be use is **Levene's test** and **Bartlett's test** (p-value over 0.05 means that the variances are equal)
- In a repeated measures experiment, **Sphericity test** is used instead - p-value over .05 means that sphericity has not been violated.



Readings For Next Week

- Mackenzie, Chapter 3, **Interaction Elements**, Human Computer Interaction: An Empirical Research Perspective, 1st ed. (2013)

Questions