Experimentation in Software Engineering: Theory and Practice

Part - II Analyzing Your Data

Massimiliano Di Penta
University of Sannio, Italy
Giulio Antoniol
École Polytechnique de Montreal

Be Aware

- This is not a stat class
- We are not statisticians
- We do not do research in statistic
- We try to get the best we can out of our data

2

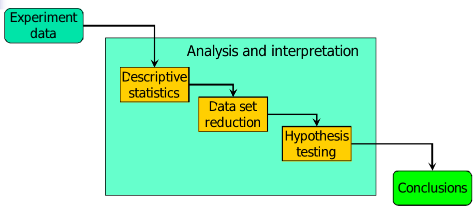
Analysis of results

Finally…

- Let's analyze the obtained experimental results!
- Overview of the data
  - Through descriptive statistics
- Removal of outliers
- Test of hypotheses related to the main factor
  - Paired, unpaired
- Analysis of co-factors
- Survey questionnaire analysis
- Interpretation and discussion of results

4

Analysis and interpretation



Convenient format for your data

- Comma separated value file where in each column you put
  - Subject ID
  - Lab #
  - System
  - Treatment
  - Ability / Experience / other co-factors
  - Values of the dependent variables being measured
  - Answers to survey questionnaire
- Very likely you might have a number of rows= # of subjects x # of labs

Data format: example

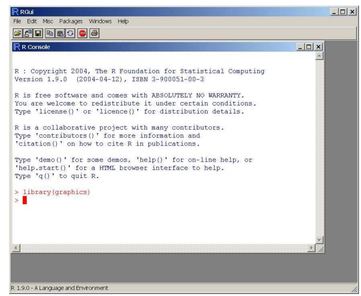
Exp	Subject	System	Method	Lab	Ability	Precision	Recall	FMeasure	Time
1	T1	WFMS	Conallen	1	h	0.59	0.6	0.55	114
1	T1	Claros	UML	2	h	0.79	0.78	0.77	85
1	T10	Claros	Conallen	1	l	0.64	0.54	0.58	92
1	T10	WFMS	UML	2	l	0.82	0.73	0.77	87
1	T11	WFMS	Conallen	1	l	0.79	0.65	0.7	134
1	T11	Claros	UML	2	l	0.76	0.8	0.74	115
1	T12	Claros	Conallen	1	h	0.78	0.92	0.82	119
1	T12	WFMS	UML	2	h	0.48	0.54	0.47	123
1	T13	WFMS	UML	1	l	0.38	0.2	0.25	116
1	T13	Claros	Conallen	2	l	0.67	0.57	0.61	104
1	T2	WFMS	Conallen	1	h	0.63	0.63	0.61	118

Some tool support

R

- Integrated suite of software facilities for data manipulation, calculation and graphical display.
  - http://www.r-project.org
  - Free implementation of of S (many similarities)
- Features:
  - Data handling and storage facility
  - Operators for calculations on arrays and matrices
  - A large collection of functions for data analysis
  - Graphical facilities
  - Simple and effective programming language
  - Fully expandible via packages

Getting Started



Packages - I

- All R functions and datasets are stored in packages.
- To load a particular package (e.g., the boot package)
  - > library(boot)
    - Packages are often inter-dependent, and loading one may cause others to be automatically loaded.

Packages - II

- Standard (base) packages
  - part of the R source code.
  - they contain the basic functions, and the standard statistical and graphical functions
- Contributed packages and CRAN
  - implement specialized statistical methods,
  - give access to data or hardware,
  - Some (the recommended packages) are distributed with every binary distribution of R.
  - Most are available for download from CRAN (http://cran.r-project.org/ and its mirrors)

The read.table() function

- Reads a data frame from a file or from the clipboard
- The first line of the file should have a name for each variable in the data frame
  - header=TRUE (otherwise header=FALSE)
- Each additional line of the file has its first item a row label and the values for each variable.
- If the file has one fewer item in its first line than in its second, this arrangement is presumed to be in force.

Read.table syntax

```
read.table(file, header = FALSE, sep = "", quote = "\"'", dec = ".",
        row.names, col.names, as.is = FALSE, na.strings = "NA",
        colClasses = NA, nrows = -1,
        skip = 0, check.names = TRUE, fill = !blank.lines.skip,
        strip.white = FALSE, blank.lines.skip = TRUE,
        comment.char = "#")
```

```
read.csv(file, header = TRUE, sep = ",", quote="\"", dec=".",
        fill = TRUE, ...)
```

- Examples:
  - read.csv("c:\table.csv") #reads a CSV file
  - read.table("clipboard",sep="\t",header=TRUE) #reads from the clipboard

Basics

- Assignment
  - a<-1+2
  - b<-c(1,3,3)
  - m<-mean(b)
- Outputs the value of a variable
  - a
    [3]
- Accessing a field of a data structure
  - t$System
- Subsetting
  - t$System[Method=="UML"]
  - t2<-subset(t,Method=="Conallen" & System=="Claros")
  - t3<-subset(t,select=-c(Method))

Data overview

Algoritmic Models

- We focused on statistical models inferred from/trained on past projects/activities
- Non algorithmic models may introduce undesired levels of subjectivity
- Non algorithmic models may be even more difficult to develop and validate

17

Statistic

- A statistic is an algebraic expression combining scores into a single number
- Statistics serve two functions:
  - they estimate parameters in population models
  - they describe the data.
- There are a large number of possible statistics

18

Definitions

Descriptive statistics: consists of methods for organizing and summarizing information.

Population: the collection of all individuals or items under consideration in a statistical study.

19

Definitions

Sample: that part of the population from wich information is collected.

Inferential statistics: consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

20

Descriptive statistics

- For each experiment collect descriptive statistics of the dependent variables
  - For each treatment of the main factor
- For nominal scale (categorical data):
  - Number of answers belonging to the various categories
  - Number of correct and wrong answers
- For ordinal scale:
  - Mean (if applicable), median, standard deviation, first and third quartile, min, max
  - Boxplots
- For ratio scale:
  - Also mean and standard deviation

R

- Median median(x)
- Mean mean(x)
- Quantiles quantile(x, prob)
  1st quartile is quantile(x, 0.25)
- Standard deviation sd(x)
- Descriptive Statistics
  summary(x)

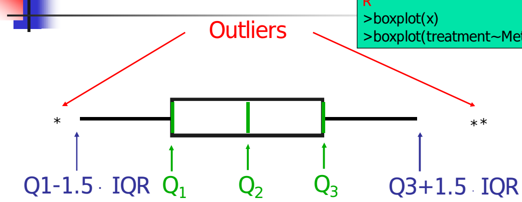
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	2	3	3	4	5

Descriptive Statistics: Conallen

Exp	UML				Conallen			
	N	mean	median	σ	N	mean	median	σ
All	64	0.64	0.67	0.15	62	0.67	0.70	0.14
Exp 1	13	0.64	0.72	0.17	13	0.63	0.62	0.08
Exp 2	28	0.58	0.57	0.15	27	0.67	0.73	0.16
Exp 3	15	0.71	0.74	0.12	14	0.67	0.69	0.16
Exp 4	8	0.72	0.70	0.13	8	0.73	0.74	0.13

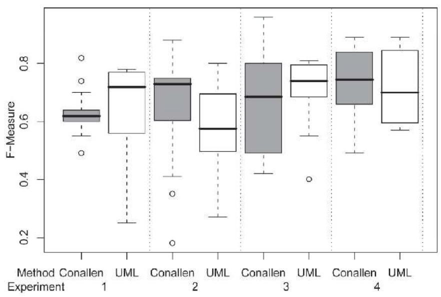
Box and Whisker Plot

R
>boxplot(x)
>boxplot(treatment~Method)



- The segments could also
  - Span between min and max
  - Span between two arbitrary quantiles sometimes $Q_{0.1}$ and $Q_{0.9}$

Boxplots: Conallen



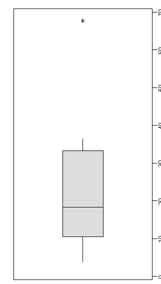
25

Effort in Months

Effort_month	Duration_month	Size_LOC
16,7	23	6050
22,6	15,5	8363
32,2	14	13334
3,9	9,2	5942
17,3	13,5	3315
67,7	24,5	38988
10,1	15,2	38614
19,3	14,7	12762
10,6	7,7	13510
36,4	15	26500

0%	25%	50%	75%	100%
3.900	10.480	18.300	39.030	67.700

IQR: 28.550

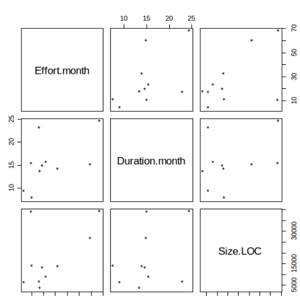


26

Scatter Plots

Effort_month	Duration_month	Size_LOC
16,7	23	6050
22,6	15,5	8363
32,2	14	13334
3,9	9,2	5942
17,3	13,5	3315
67,7	24,5	38988
10,1	15,2	38614
19,3	14,7	12762
10,6	7,7	13510
59,5	15	26500

- Assessing dependencies between variables
- Identify outliers (atypical values)
- Observe correlation



Sample Correlation Coefficient

- It is a measure of the closeness of relationship between two variables

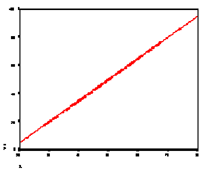
$$r = \frac{\sum x_i y_i}{\sqrt{\left(\sum x_i^2\right)\left(\sum y_i^2\right)}}$$

- Correlation coefficient range between –1.0 and 1.0

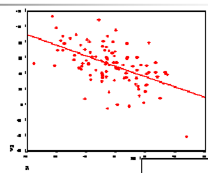
28

r

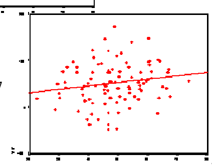
1.0



-0.54

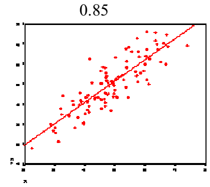


0.17

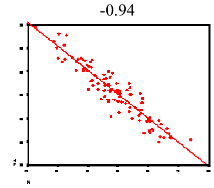


r

0.85

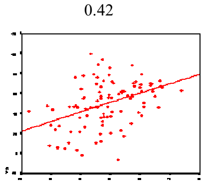


-0.94

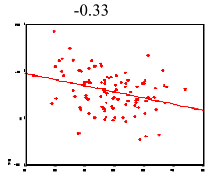


r

0.42

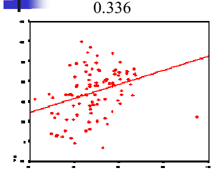


-0.33



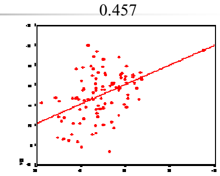
Outliers

0.336



An outlier that falls some distance away from the original regression line would decrease the size of the correlation coefficient

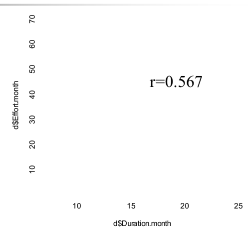
0.457

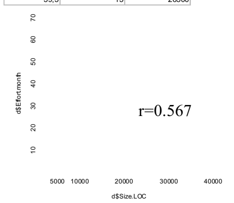


An outlier which falls near where the regression line would normally fall would necessarily increase the size of the correlation coefficient

Project effort

Effort_month	Duration_month	Size_LOC
16,7	23	6050
22,6	15,5	8363
32,2	14	13334
3,9	9,2	5942
17,3	13,5	3315
67,7	24,5	38988
10,1	15,2	38614
19,3	14,7	12762
10,6	7,7	13510
59,5	15	26500





Correlation and Causation

Much of the early evidence that cigarette smoking causes cancer was correlational.

It may be that people who smoke are more nervous and nervous people are more susceptible to cancer.

It may also be that smoking does indeed cause cancer.

The cigarette companies made the former argument, while some doctors made the latter.

Correlation and Causation

Suppose there exists a high correlation between the number of popsicles sold and the number of drowning deaths.

Does that mean that one should not eat popsicles before one swims?

Not necessarily.

Correlation and Causation

Both of the above variable are related to a common variable, the heat of the day.

The hotter the temperature, the more popsicles sold and also the more people swimming, thus the more drowning deaths.

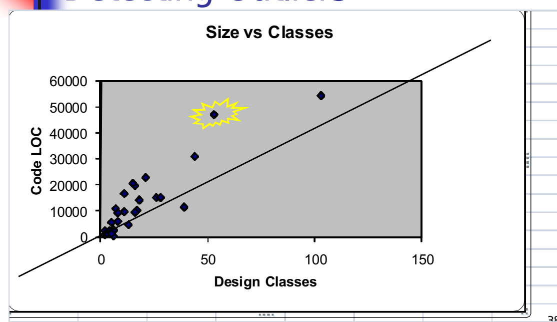
This is an example of correlation without causation.

Scatter Plot: R

- x<-c(5,15,9,15,7)
- y<-c(60,120,90,140,60)
- plot(x, y, xlim=c(0,20), ylim=c(0,200), xlabel="Registered vehicles", ylabel="Gasoline sales")

37

Detecting Outliers



38

Hypothesis testing

Data Level, Operations, and Statistical Methods

Data Level	Meaningful Operations	Statistical Methods
Nominal	Classifying and Counting	Nonparametric
Ordinal	All of the above plus Ranking	Nonparametric
Interval	All of the above plus Addition, Subtraction, Multiplication, and Division	Parametric
Ratio	All of the above	Parametric

40

Measurement Scales and Statistics

Scale	Relations	Statistics	Tests
Nominal	Equivalence	Mode, Frequency	Non parametric
Ordinal	Equivalence Comparison	Median, Percentile, Spearman r, Kendall τ, Kendall W	Non parametric
Interval	Equivalence Comparison Relation between intervals	Mean, Standard deviation Pearson product-moment correlation Multiple product-moment correlation	Non parametric
Ratio	Equivalence Comparisons Relation between intervals Ratio between pairs ov values	Geometric mean Coefficient of variation	Non parametric Parametric

41

Testing for normality

- In some cases you might be able to apply parametric tests
- Ensure you have enough data points per treatment (~30 at least)
- However this does not guarantee normality
- Look at boxplot/histogram
  - Bell-shaped histogram
- Empirical rules
- Use appropriate tests
  - Shapiro-Wilk
  - Anderson-Darling

42

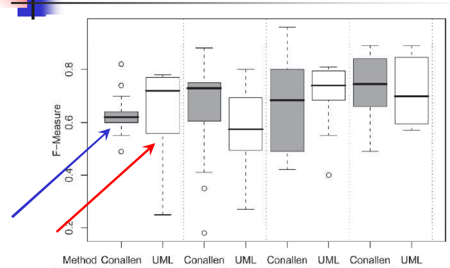
Empirical Rule

- Data are normally distributed (or approximately normal)

Distance from the Mean	Percentage of Values Falling Within Distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

43

Does our data follow a normal distribution?



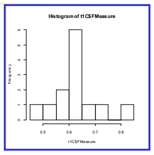
44

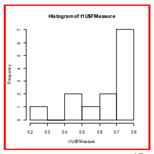
Testing for normality: example

```
>tU<-subset(t,Method=="UML" & Exp==1)
>tC<-subset(t,Method=="Conallen" & Exp==1)
>hist(tU$FMeasure)
>hist(tc$FMeasure)
>
>shapiro.test(tC$FMeasure)
 Shapiro-Wilk normality test

data:  t1C$FMeasure
W = 0.9401, p-value = 0.4585
>
>shapiro.test(tU$FMeasure)
 Shapiro-Wilk normality test
data:  tU$FMeasure
W = 0.8293, p-value = 0.01552
```

Remember the QQ Plot!





All data unpaired analysis

- We compare the mean / medians of independent populations
  - Values of the dependent variables obtained with different treatments
- Parametric test: unpaired t-test
- Non-parametric test: Mann Whitney U test (or Wilcoxon U test)
- Hypotheses:
  - Two-tailed:
    H_0: $\mu_2 = \mu_1$ H_a: $\mu_2 \neq \mu_1$
  - One-tailed (alternative: greater):
    H_0: $\mu_2 = \mu_1$ H_a: $\mu_2 > \mu_1$
  - One-tailed (alternative: less)
    H_0: $\mu_2 = \mu_1$ H_a: $\mu_2 < \mu_1$

Example

```
>t1<-subset(t,Exp==2)
>attach(t1)
>wilcox.test(FMeasure[Method=="Conallen"],
   FMeasure[Method=="UML"], paired=FALSE, alternative="greater")
 Wilcoxon rank sum test with continuity correction

data:  FMeasure[Method == "Conallen"] and FMeasure[Method == "UML"]
W = 512.5, p-value = 0.01199
alternative hypothesis: true location shift is greater than 0
```

For parametric statistics just replace wilcox.test with t.test

What if we have >2 treatments?

- Tests for multiple means
- Parametric: One-Way ANOVA
  - summary(aov(FMeasure~Method))
- Non-parametric: Kruskal-Wallis test
  - kruskal.test(FMeasure~Method)
- Hypotheses:
  - H_0: $\mu_3 = \mu_2 = \mu_1$
  - H_a: $\mu_2 \neq \mu_1 \vee \mu_2 \neq \mu_3 \vee \mu_3 \neq \mu_1$

All data paired analysis

- When each subject receives different treatments
- We would like to analyze the differences exhibited by subjects with different treatments

- $H_0: \mu_d = 0$ $H_a: \mu_d \neq 0$

- Available tests:
  - Parametric: paired t-test
  - Non-parametric: paired Wilcoxon test

49

Paired analysis: example

ID	F.Conallen	F.UML
T20	0.74	0.74
T21	0.74	0.51
T22	0.7	0.29
T24	0.88	0.62
T25	0.75	0.8
T26	0.66	0.39
T27	0.35	0.51
T28	0.62	0.59
T29	0.57	0.68
T30	0.73	0.43
T32	0.74	0.56

```
>wilcox.test(F.Conallen,F.UML, paired=TRUE,
    alternative="greater")

      Wilcoxon signed rank test with continuity
    correction

data:  F.Conallen and F.UML
V = 138, p-value = 0.04354
alternative hypothesis: true location shift is greater
    than 0
```

- Must have data in a paired format
  - or you can use a proper R script
- Need to remove subjects that took part to one lab only
- For parametric statistics just replace wilcox.test with t.test

50

Effect size measures

- With statistical tests we have shown that distributions of samples obtained with different treatments are significantly different
  - We can reject the null hypotheses
  - Ok fine.. but even if this is the case, difference could be quite small!
  - Who would care about a new method if it introduces a statistically significant improvement, but with a negligible practical effect?

51

Effect size measures

Cohen d effect size (independent samples)

- Indicates the magnitude of a main factor treatment effect on the dependent variables

$$d = \frac{\mu_2 - \mu_1}{\sigma} \qquad \sigma = \sqrt{(\sigma_1^2 + \sigma_2^2)/2}$$

pooled std. deviation

negligible	small	medium	large

0.2 0.5 0.8

52

Effect size measures

Cohen d effect size (dependent samples)

- To be used together with the paired t-test or Wilcoxon paired test

$$d = \frac{\mu_2 - \mu_1}{\sigma_D}$$

- σ_D is the standard deviation of the (paired) differences

negligible	small	medium	large

0.2 0.5 0.8

Non Parametric Effect Size

- Cliff, the delta index provides a useful representation of effect size
- Cliff's delta represents the degree of overlap between the two distributions of scores
- G1: 1 2 3 4 4 5
- G2: 1 1 2 2 2 3 3 3 4 5

Cliff's Calculation

	1	2	3	4	4	5	Mean
1	0	-1	-1	-1	-1	-1	-0.833
1	0	-1	-1	-1	-1	-1	-0.833
2	1	0	-1	-1	-1	-1	-0.5
2	1	0	-1	-1	-1	-1	-0.5
2	1	0	-1	-1	-1	-1	-0.5
3	1	1	0	-1	-1	-1	-0.167
3	1	1	0	-1	-1	-1	-0.167
3	1	1	0	-1	-1	-1	-0.167
4	1	1	1	0	0	-1	0.333
5	1	1	1	1	1	0	0.833
	0.8	0.3	-0.3	-0.7	-0.7	-0.9	-0.2501

negligible	small	medium	large

0.147 0.33 0.47

Conallen: summary of unpaired analysis

Exp	UML				Conallen				M-W	t-test	Effect
	N	mean	median	σ	N	mean	median	σ	p-value	p-value	size
All	64	0.64	0.67	0.15	62	0.67	0.70	0.14	0.19	0.13	0.20
Exp 1	13	0.64	0.72	0.17	13	0.63	0.62	0.08	0.82	0.82	−0.03
Exp 2	28	0.58	0.57	0.15	27	0.67	0.73	0.16	**0.01**	**0.01**	0.56
Exp 3	15	0.71	0.74	0.12	14	0.67	0.69	0.16	0.76	0.76	−0.29
Exp 4	8	0.72	0.70	0.13	8	0.73	0.74	0.13	0.36	0.36	0.12

Conallen: summary of paired analysis

		Difference		Wilcoxon	t-test	Effect
Exp	N	mean	median	*p*-value	*p*-value	size
All	51	0.02	0.00	0.27	0.19	0.12
Exp 1	13	−0.00	−0.11	0.61	0.53	−0.02
Exp 2	20	0.08	0.05	**0.04**	**0.03**	0.45
Exp 3	10	−0.06	−0.09	0.88	0.80	−0.27
Exp 4	8	0.02	0.02	0.31	0.34	0.15

Bonferroni correction

- As said, doing multiple t-tests introduce a higher error.
- I can still do t-tests and make the correction
- If I do N t-tests, I can reject the null hypotheses if the test p-values are such that:

$$p < p_{bonferroni} = \frac{\alpha}{N}$$

Bonferroni correction: example

- I have results from three treatments A, B, C
- I perform 3 t.tests (or Mann-Whitney tests)
  - t.test(A,B)
  - t.test(B,C)
  - t.test(A,C)
- I can reject the hypothesis if tests provide p-value <0.05/3=0.016

Alternative to Bonferroni

- Bonferroni correction is deemed to be too stringent
  - And criticized by many scientists
- There are less stringent alternatives…
  - Holm's correction
  - Benjamini and Hochberg correction

Holm's correction

- Rank your p-values from the smallest to the largest
- Given n the number of p-values (and thus of tests done)
- Multiply the first by n, the second by n-1, etc.
  - p-value significant if after multiplied is <0.05 (with significance 95%)

61

Holm's correction - Example

0.01	0.015	0.02	0.03

Bonferroni correction:

- 0.01 * 4 = 0.04
- 0.015 * 4 = 0.06
- 0.02 * 4 = 0.08
- 0.03 * 4 = 0.12

Holm's correction:

- 0.01 * 4 = 0.04
- 0.015 * 3 = 0.045
- 0.02 * 2 = 0.04
- 0.03 * 1 = 0.03

62

Benjamini and Hochberg

- Rank p-values
- The largest p-value remains as is
- The second largest value is multiplied by the number of genes divided by the rank
- The others are treated as the second

63

Benjamini and Hochberg - Example

0.01	0.015	0.02	0.03

Bonferroni correction:

- 0.01 * 4 = 0.04
- 0.015 * 4 = 0.06
- 0.02 * 4 = 0.08
- 0.03 * 4 = 0.12

BH correction:

- 0.01 * 4/4 = 0.01
- 0.015 * 4/3 = 0.02
- 0.02 * 4/2 = 0.04
- 0.03 * 1 = 0.03

64

R

p.adjust(p, method = p.adjust.methods,
  n = length(p))

- Method can be “bonferroni”, “holm”, “BH” and others
  - See help(p.adjust)
- Example:
  - p.adjust(c(0.01,0.015,0.02,0.03),method=“holm”)

65

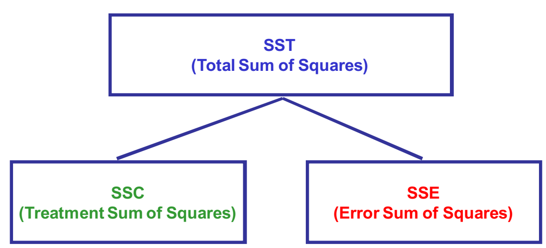
Analysis of co-factors

Analysis of co-factors

- We need to analyze whether
  - Co-factors directly influence the dependent variables
  - Co-factors interact with the main factor
  - Statistics to be used:
    - Two-Way ANOVA (effect of a single co-factor)
    - N-Way ANOVA (effect of more co-factors together)
    - Multiple pair-wise tests (with Bonferroni correction)
    - Interaction plots
    - Friedman test (non-parametric alternative to ANOVA)

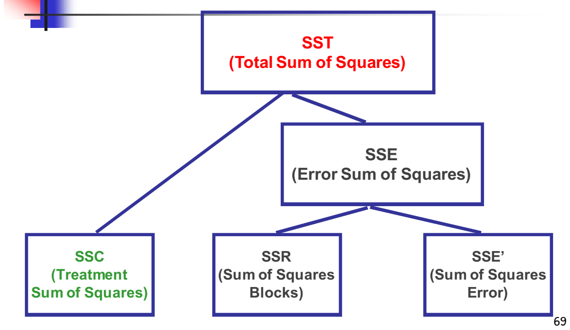
67

ANOVA: Partitioning Total Sum of Squares of Variation



68

Two-Way ANOVA

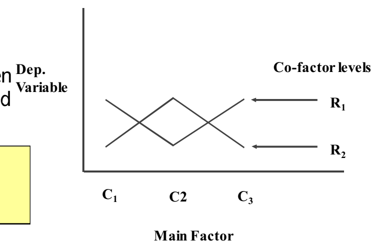


69

Interaction plots - I

- Graphically show the presence of interaction between the main factor and co-factors

R:
interaction.plot(MainFactor, CoFactor, DepVariable)



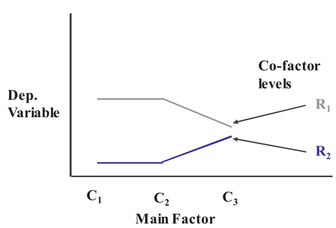
Lines cross → there is interaction

70

Interaction plots - II

- Graphically show the presence of interaction between the main factor and co-factors

R:
interaction.plot(MainFactor, CoFactor, DepVariable)



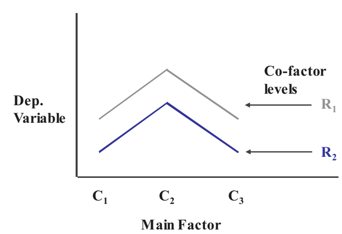
Lines converge → there is a mild interaction

71

Interaction plots - III

- Graphically show the presence of interaction between the main factor and co-factors

R:
interaction.plot(MainFactor, CoFactor, DepVariable)



Parallel lines → no interaction

72

ANOVA by Method & Ability

```
>summary(aov(FMeasure~Method*Ability))
                 Df  Sum Sq Mean Sq  F value  Pr(>F)
Method            1 0.01153 0.01153  0.6619 0.41832
Ability           1 0.02899 0.02899  1.6634 0.20086
Method:Ability    1 0.08462 0.08462  4.8555 0.03043 *
Residuals        80 1.39421 0.01743
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
   0.1 ' ' 1
```

- No direct effect of Method (overall data)
- No direct effect of Ability
- ...but significant interaction between method and Ability

73

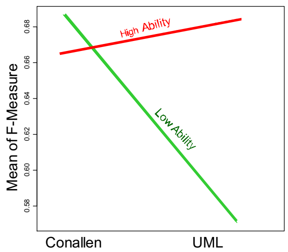
Generating LaTeX tables...

```
> library(xtable)
> xtable(summary(aov(FMeasure~Method*Ability)))

% latex table generated in R 2.7.0 by xtable 1.5-2 package
% Sat May 15 14:35:11 2010
\begin{table}[ht]
\begin{center}
\begin{tabular}{lrrrrr}
  \hline
 & Df & Sum Sq & Mean Sq & F value & Pr($>$F) \\
  \hline
Method         & 1 & 0.01 & 0.01 & 0.66 & 0.4183 \\
  Ability        & 1 & 0.03 & 0.03 & 1.66 & 0.2009 \\
  Method:Ability & 1 & 0.08 & 0.08 & 4.86 & 0.0304 \\
  Residuals      & 80 & 1.39 & 0.02 &  &  \\
   \hline
\end{tabular}
\end{center}
\end{table}
```

74

Interaction plots



- >interaction.plot(Method, Ability,FMeasure)
- Lines cross → there is interaction between two factors

75

ANOVA by Method & Experience

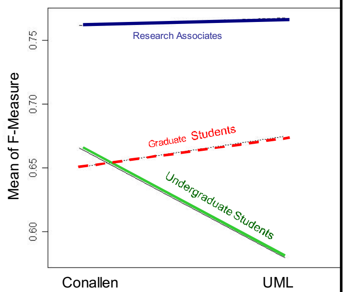
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Method	1	0.03	0.03	1.31	0.25
Experience	2	0.14	0.07	3.47	**0.034**
Method:Experience	2	0.08	0.04	2.05	0.13
Residuals	120	2.48	0.02		

- It looks like the Experience plays a significant role...
- Apparently, no significant interaction with Method

76

Interaction Plot: Method & Experience

- ...Interaction can be noticed if considering students only



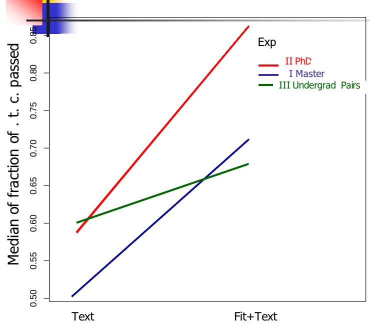
ANOVA By Method & Lab

- Needed to assess the presence of learning effect among labs
- No learning effect in this study
- No interaction with the Main Factor

Exp	Method p-value	Lab p-value	Method:Lab p-value
All	0.27	0.07	0.99
Exp 1	0.93	0.31	0.46
Exp 2	**0.04**	0.58	0.70
Exp 3	0.44	0.11	0.90
Exp 4	0.82	0.32	0.99

78

Fit Experiment: Interaction between Main factor and Experience



- Two way ANOVA:
  - the experience effect significant (p-value=0.039)
  - Interaction with main factor not significant (p-value=0.27)
- Subjects with different experience gained different benefits from the use of Fit tables.
- Slope (benefit) higher for highly experienced subjects

79

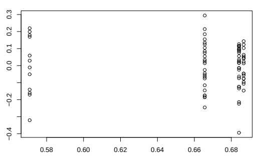
Analysis of Variance: Assumptions

- Observations are drawn from normally distributed populations
  - But ok.. ANOVA is pretty robust on that
  - Histogram or QQ plot ...
- Observations represent random samples from the populations
  - Samples should be independent
  - Design could mitigate this threat
  - Use repeated measures ANOVA where needed
- Variances of the populations are equal
  - Look at residuals

80

Looking at residuals…

Two-way ANOVA by Method & Ability



Two-way ANOVA by Method & Experience

Residuals

Fitted Model

```
>m<-lm(FMeasure ~ Method*Ability)
>plot(m$fitted,m$resid,xlab="Fitted Model",ylab="Residuals")

>m<-lm(FMeasure ~ Method*Experience)
>plot(m$fitted,m$resid,xlab="Fitted Model",ylab="Residuals")
```

81

Repeated measures ANOVA

Single Measures

↓

Two-sample t-test

↓

ANOVA
between-subject ANOVA

Treat 1	Treat 2	Treat 3	Control
Group 1	Group2	Group3	Group4

Assumptions

- Homogeneity of Variance
- Normality
- **Independence of observations**

Repeated Measures

↓

Paired-sample-t-test

↓

Repeated ANOVA
within-subject ANOVA

Treat 1	Treat 2	Treat 3	Control
Subj. 1 Subj. 2 Subj. 3 ….	Subj. 1 Subj. 2 Subj. 3 ….	Subj. 1 Subj. 2 Subj. 3 ….	Subj. 1 Subj. 2 Subj. 3 …..

Assumptions

- Homogeneity of Variance
- Homogeneity of Correlations
- Normality

82

Repeated ANOVA

One-way between-subject ANOVA

An individual score is specified by

$$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$$

μ = Grand mean

$\tau_j = \mu_j - \mu$ Treatment effect

$\varepsilon_{ij} = X_{ij} - \mu_j$ Residual error

One-way within-subject ANOVA

An individual score is specified by

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij}$$

μ = Grand mean

π_i = Subject effect

τ_j = Treatment effect (within-subject effect)

ε_{ij} = Residual error

83

Partitioning the Sum of Squares

Total Variation (SS_{total})

- Treatment effect (SS_{treat})
- Error (SS_{error})

Total Variation (SS_{total})

- Within subj. (SS_{within})
  - Treatment effect (SS_{treat})
  - Residual (SS_{res})
    - Subj. x Treat & Error
- Between subj ($SS_{between}$) Subject effects

84

Conallen example

- We should analyze the within effect among questions
  - At least
- Need to organize the table by question

Exp	Subject	Method	Question	Precision	Recall	Fmeasure
1	T1	Conallen	1	1	1	1
1	T1	UML	1	1	1	1
1	T1	Conallen	2	0	0	0
1	T1	UML	2	1	1	1
1	T1	Conallen	3	1	1	1
1	T1	UML	3	1	1	1
1	T1	Conallen	4	1	0.8	0.888889
1	T1	UML	4	1	1	1
1	T1	Conallen	5	0	0	0
1	T1	UML	5	1	1	1
1	T1	Conallen	6	1	0.666667	0.8
1	T1	UML	6	1	1	1
1	T1	Conallen	7	0	0	0
1	T1	UML	7	1	0.666667	0.8
1	T1	Conallen	8	0.033333	1	0.064516
1	T1	UML	8	1	0.666667	0.8
1	T1	Conallen	9	1	1	1
1	T1	UML	9	0	0	0
1	T1	Conallen	10	1	0.714286	0.833333
1	T1	UML	10	0.5	1	0.666667
1	T1	Conallen	11	1	1	1
1	T1	UML	11	0	0	0

```
aov(FMeasure ~ Method * Question + Error(Subject/(Method * Question)))
```


Results

Claros

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between Subjects					
Method	1	0.16	0.16	0.66	0.42
Residuals	55	13.71	0.25		
Within Subjects					
Question	1	6.25	6.25	49.22	**< 0.01**
Method:Question	1	0.25	0.25	1.98	0.16
Residuals	625	79.36	0.13		

WfMS

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between Subjects					
Method	1	0.22	0.22	0.84	0.36
Residuals	57	15.24	0.27		
Within Subjects					
Question	1	0.62	0.62	3.59	0.06
Method:Question	1	0.08	0.08	0.46	0.50
Residuals	647	110.87	0.17		

Categorical data

Categorical data

- We would like to test whether the proportions of correct and incorrect answers are significantly different
- Hypothesis being tested: H_0: $p_1=p_2$
  - Where p_1 and p_2 are proportion of data from 2 distributions
  - E.g. proportions of correct answers provided with two treatments
- Tests:
  - χ^2
  - Fisher's exact test (better for small data samples)
  - Prop test

Odds Ratio

- Effect size measure for categorical data
- Odds: number of times an event occurs / number of times the event does not occur

$$Odd = p/(1-p)$$

  - Used in medicine research, but also in sportive bets
  - Italy has 1:11 Odds to win the world cup, Brasil 1:5
- Odds Ratio: odds of the experimental group divided by the odds of the control group

$$OR = \frac{p/(1-p)}{q/(1-q)}$$

89

With a contingency table…

	T1	T2
pass	a	b
fail	c	d

	T1	T2
pass	10	8
fail	9	5

$$OR = \frac{a/c}{b/d} = \frac{10/9}{8/5} = 0.69$$

- Odds of passing test cases with T1 are 0.69 of those with T2
- … odds of passing test cases with T2 are 1.44 higher than with T1

90

Example: Fit to understand requirements

- Example
  - Use of (non executable) Fit Tables for comprehension of requirements [Ricca et al., IST 2009]
- Hypotheses:
  - H_{01}: the availability of Fit tables as a complement to requirements <u>does not</u> significantly affect the comprehension level
  - H_{02}: the availability of Fit tables as a complement to requirements <u>does not</u> significantly affect the comprehension effort

91

Testing the proportion of correct answers

- Main factor: availability of (non executable) Fit tables as complement to requirements
- Independent variable: # of correct answers provided to questions about requirements

> R1. *The library employee can insert, delete or update a member. For each member the following fields need to be specified: unique member ID, name, surname, address, date of birth and credit/debit. The member ID is automatically computed by summing day, month and year of his/her birth date and subtracting from the result the number of letters of name and surname. If the value obtained is not unique — i.e., it is an already existing ID — then the software subtracts 1 from it.*

- Simple design (single lab)

	R1	R2	R3	R4	R5	R6
Group 1	(+)	(-)	(+)	(-)	(+)	(-)
Group 2	(-)	(+)	(-)	(+)	(-)	(+)

92

Testing the proportion of correct answers

Correct answers

	Q1	Q2	Q3	Q4	Q5	Q6
Group A	6	3	7	1	3	0
Group B	0	2	2	6	2	3
P-value	0.0069	1	0.04	0.01	1	0.076

Overall correct answers

	Wrong	Correct
Fit (+)	18	27
Text (-)	37	8

```
>fisher.test(array(c(27,18,8,37),
     dim=c(2,2)))

     Fisher's Exact Test for Count
  Data

data:  array(c(27, 18, 8, 37), dim =
     c(2, 2))
p-value = 7.517e-05
alternative hypothesis: true odds ratio
     is not equal to 1
95 percent confidence interval:
  2.410347 20.947134
sample estimates:
odds ratio
  6.771132
```

- Fisher's test indicates a significant difference
- Odds ratio indicate that chances to answer correctly with Fit are about 7 times higher

93

Test for dependent samples

Exp	ID	Requireme	Treatment	Correct	Time
I	R1	1	Fit	Yes	7
I	R2	1	Fit	Yes	6
I	R3	1	Fit	No	9
I	R4	1	Fit	Yes	7
I	R5	1	Fit	Yes	8
I	R6	1	Fit	Yes	8
I	R7	1	Fit	Yes	8
I	R8	1	Fit	No	8
I	Y1	1	Text	No	11
I	Y3	1	Text	No	8

- There could be a dependency of the ordering effect of questions
- Cochran Q test: test for dependent samples on categorical data

```
>library(outliers)
>d2f<-subset(data,Exp=="II" & Treatment=="Fit")
>cochran.test(Correct~Requirement)
 Cochran test for outlying variance

data:  Correct ~ Requirement
C = 0.2523, df = 7.5, k = 6.0, p-value = 0.907
alternative hypothesis: Group 5 has outlying variance
sample estimates:
        1         2         3         4         5         6
0.0000000 0.2777778 0.1666667 0.1944444 0.3000000 0.2500000
```

94

Survey questionnaire analysis

How to analyze survey questionnaires

- Q1: I had enough time to perform the lab tasks (1-5)
- Q2: The objectives of the lab were perfectly clear to me (1-5)
- Q3: The questions were clear to me (1-5)

- Answers are expressed used a Likert scale thus
  - You can use statistical tests
- Answer >3 (at least weak agreement)
  - H_0: Q>3 (median of provided answers >3)
- Check if the experiment was more difficult for a particular treatment
  - H_0: $Q_{UML}=Q_{CONALLEN}$

96

Conallen: objective clarity

H_0:	$\tilde{Q} > 3$						$\tilde{Q}_{Conallen} = \tilde{Q}_{UML}$		
Exp	$\widetilde{Q1}$	p	$\widetilde{Q2}$	p	$\widetilde{Q3}$	p	Q1.p	Q2.p	Q3.p
All	2.00	**<0.01**	2.00	**<0.01**	2.00	**<0.01**	0.40	0.58	0.73
Exp 1	2.00	**<0.01**	2.00	**<0.01**	2.00	**<0.01**	0.65	0.59	1.00
Exp 2	2.00	**0.01**	2.00	**<0.01**	3.00	**0.02**	0.63	0.99	1.00
Exp 3	1.00	**<0.01**	1.00	**<0.01**	2.00	**<0.01**	0.94	0.88	0.60
Exp 4	2.00	**<0.01**	2.00	**0.03**	2.00	0.11	0.49	0.95	0.69

- Overall objectives clear
- No significant differences between treatments

Artifact Comprehension…

- Q4: I experienced no difficulty in reading the diagrams (1-5)
- Q5: I experienced no difficulty in reading the source code (1-5)
- Q8: I understood the meaning of Conallens' stereotypes (1-5)

H_0:	$\tilde{Q} \geq 3$						$\tilde{Q}_{Conallen} - \tilde{Q}_{UML}$			
Exp	$\widetilde{Q4}$	p	$\widetilde{Q5}$	p	$\widetilde{Q8}$	p	Q4 p	Q4 d	Q5 p	Q5 d
All	3.00	0.18	2.50	**<0.01**	2.00	**<0.01**	**<0.01**	−0.60	0.81	0.02
Exp 1	3.00	0.23	3.00	0.20	2.00	**0.01**	0.34	−0.44	0.98	−0.13
Exp 2	3.00	0.98	3.00	0.14	2.00	**0.04**	0.07	−0.48	0.49	0.21
Exp 3	2.00	**<0.01**	2.00	**<0.01**	2.00	**<0.01**	**0.02**	−0.89	0.60	−0.24
Exp 4	3.00	0.38	2.00	**<0.01**	1.50	**0.02**	0.07	−1.04	1.00	0.00

Time spent on various artifacts

- Q6: How much time (in percentage) did you spend looking at class diagrams?
- Q7: How much time (in percentage did you spend for source code browsing?

- We compute the Odds of looking at diagrams vs. looking at code for the two treatments…
- And then compute the OR
- Compare proportions using χ^2

Exp	Conallen	UML	OR	p-value
All	1.96	0.73	2.68	**<0.01**
Exp 1	2.06	0.61	3.39	**<0.01**
Exp 2	2.31	0.76	3.03	**<0.01**
Exp 3	1.51	0.85	1.77	**0.04**
Exp 4	1.17	0.57	2.03	**0.02**

Overall..

- No particular difficulty in the experimental tasks
- Diagrams more difficult to be understood than source code
  - But Conallen's diagrams easier to be understood
- When Conallen's diagrams are available, odds of looking at diagrams 2-3 times higher than for source code

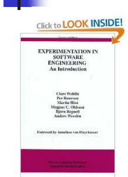
Summary of statistical tests

Summary of tests

Scale	Distrib. compared	Samples	Test
Ratio (parametric test)	Two	Indep.	t-test (unpaired)
		Dep.	t-test (paired)
	More than two	Indep.	ANOVA
		Dep.	Repeated mes. ANOVA
Ordinal (non-parametric)	Two	Indep.	Wilcoxon U test Mann-Whitney U test
		Dep.	Wilcoxon paired test
	More than two	Indep.	Kruskal-Wallis test Friedman test (blocked design)
Categorical	Two	Indep.	Fisher's exact test, χ^2 test
	Two or more	Indep.	χ^2 test
	Two or more	Dep.	Cochran Q test

References

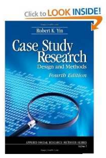
Suggested Books - I



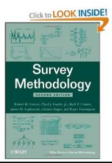
- **Experimentation in Software Engineering: An Introduction**
  Claes Wohlin, Per Runeson, Martin Höst , Springer, 1999
- **Basics of Software Engineering Experimentation**
  Natalia Juristo, Ana M. Moreno, Springer, 2010

104

Suggested Books - II

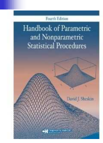


- Case Study Research: Design and Methods
  Robert K. Yin, Sage Publications, Inc; 4th edition (October 31, 2008)

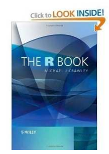


- Survey Methodology
  Robert M. Groves, Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, Roger Tourangeau, Wiley; 2 edition (July 14, 2009)

Suggested Books - III

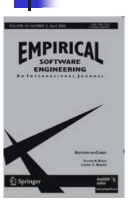


- Handbook of Parametric and Nonparametric Statistical Procedures: Fourth Edition
  David J. Sheskin, Chapman and Hall/CRC; 4th edition (January 19, 2007)



- The R Book
  Michael J. Crawley Wiley;
  1st edition (June 15, 2007)

Also…



- Empirical Software Engineering, (EMSE) Journal, edited by Springer
- International Conference on Empirical Software Engineering and Measurements (ESEM)
- And of course mainstream journals (TSE, TOSEM, IST…) and conferences (ICSE, ESEC / FSE, ASE…)

Thank you!