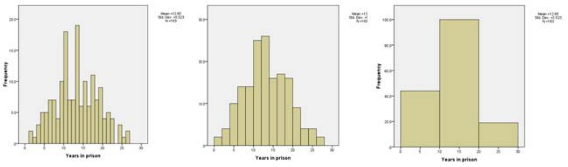
First and Second Steps in Statistics

Bernd Müller-Bierl
Universitair Ziekenhuis Brussel

TABLE OF CONTENTS

HISTOGRAMS AND BOXPLOTS

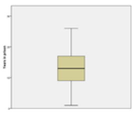
amount of time spent in prison for falsly convicted individuals



One year bin | Two year bin | Ten year bin

N = 163

Box and Whiskers Plot



5 number summary

	13	
9		17
1		26

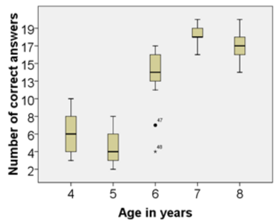
Units = years
IQR = 8
Range = 25

HISTOGRAMS AND BOXPLOTS

correct answers on a test by the participants' ages

Lower whisker = lower quartile (LQ) to LQ - 1.5 IQR
Higher whisker = higher quartile (HQ) to HQ + 1.5 IQR
Whiskers stop at extreme values

Outside points more than 1.5 IQR
Far outside points more than 3 IQR



THE MEAN AND THE STANDARD DEVIATION

some most important descriptive statistics

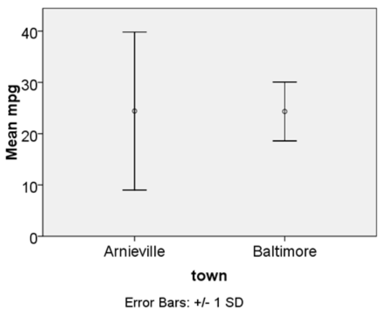
$$\bar{x} = \frac{1}{n}\sum_i x_i$$ Mean

$$\operatorname{var} x_i = \frac{\sum_i (\bar{x} - x_i)^2}{n-1}$$ Variance

$$\operatorname{sd} = \sqrt{\operatorname{var} x_i}$$ Standard deviation

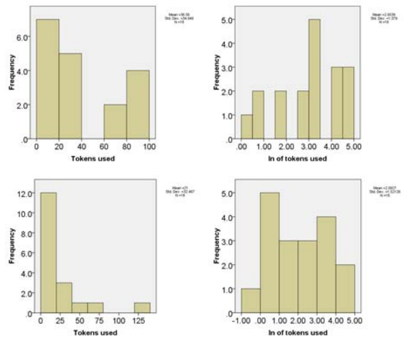
THE MEAN AND THE STANDARD DEVIATION

mean miles per gallon for two towns



THE MEAN AND THE STANDARD DEVIATION

amount of times Nim used his own name and Nim used the pronoun Me: positively skewed hystograms



PROPORTIONS AND BAR CHARTS

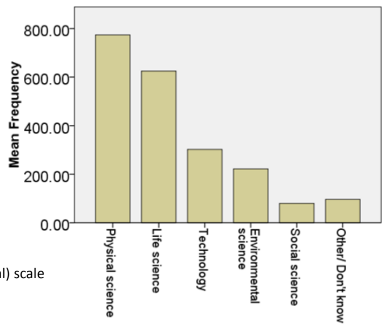
womans accused of being a witch

		Proportions
Single	51/241	= 0.21
Widowed	38/241	= 0.16
Divorced	4/241	= 0.02
Married	148/241	= 0.61

		Odds
Single	51/190	= 0.27
Widowed	38/203	= 0.19
Divorced	4/237	= 0.02
Married	148/93	= 1.59

PROPORTIONS AND BAR CHARTS

what comes to mind when "science" is mentioned: Bar chart



Categorical (nominal) scale

SAMPLING AND ALLOCATION

Simple Random Sample from a population of pizzas

All possible combinations (equal possible with SRS)	**All toppings equally likely**	**Vegetarian samples**	**One meat quota samples**
Mushrooms & Pepper	Mushroom & Pepper	Mushroom and Pepper	Mushroom & Saussage
Mushrooms & Olives	Peppers & olives	Mushrooms & Olives	Pepper & saussage
Mushrooms & Saussags	Saussage & Mushrooms	Pepper & Olives	Olives & Saussage
Pepper & Olives			
Pepper & saussage			
Olives & saussage			

'random' means that each possible sample is equally likely

Alternatives:

Cluster samples **example: choose SRS of schools, then SRS of pupils**

Quota samples **example: choose a sample to be half women, half men**

Convenience sample **example: choose the first 20 people that sign up your study**

SRS: Simple Random Sample

SAMPLING AND ALLOCATION

Random Allocation comparing viewers' reactions to English and Australian soap operas

Sampling **Allocating** **Measurement**

Polpulation → Sample → Group – Engl. Soaps → Despair rating

Polpulation → Sample → Group – Austr. Soaps → Despair rating

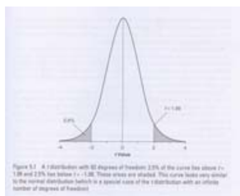
SRS: Simple Random Sample

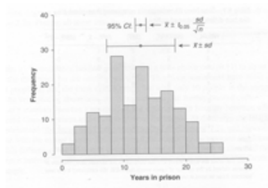
RA: Random Allocation

INFERENCE AND CONFIDENCE INTERVALS

amount of time spent in prison for falsly convicted individuals

The equation for the 95% interval is $CI_{95\%} = \bar{x} \pm t_{0.05} \frac{sd}{\sqrt{n}}$ $df = n - 1$





INFERENCE AND CONFIDENCE INTERVALS

do people like fresh or instant coffee more?

Within subject studies

The basic equation for the within-subject 95% confidence interval is

$$CI_{95\%} = \overline{x_1 - x_2} \pm t_{0.05} \frac{sd_{diff}}{\sqrt{n}} \qquad df = n-1$$

Table 5.1 *Data from 10 participants comparing how much they like two different types of coffee*

	$FRESH_i$	$INSTANT_i$	$DIFF_i$	$DIFF_i - \bar{x}$	$(DIFF_i - \bar{x})^2$
	5	3	2	1	1
	4	3	1	0	0
	6	5	1	0	0
	3	4	−1	−2	4
	4	4	0	−1	1
	5	3	2	1	1
	6	3	3	2	4
	3	3	0	−1	1
	5	3	2	1	1
	4	4	0	−1	1
Sum	45	35	10	0	14
Mean	4.5	3.5	1.0	0	1.56*

* This 1.56 is not the actual mean. It is the variance of the variable $DIFF_i$, the sum of squares divided by the number of cases minus one (14/9). We calculated it this way because it can be used in later calculations.

$$CI_{95\%} = 1.0 \pm 2.26 \frac{1.25}{\sqrt{10}} = 1.0 \pm 0.89 \qquad df = 9$$

more people like freshly brewed coffee

INFERENCE AND CONFIDENCE INTERVALS

significant teachear makes them learning willingly

Between subject studies

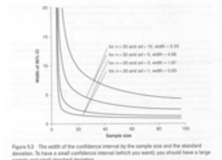


Figure 5.3 The width of the confidence interval by the sample size and the standard deviation. To have a small confidence interval (which you want), you should have a large sample and small standard deviation

makes me learn willingly 1 (not true) – 5 (true)

	boys	girls
Number	129	166
Mean	3.26	3.51
Standard Deviation	1.05	0.99
Confidence Intevals		

$$95\%CI = 3.26 \pm 1.98 \frac{1.05}{\sqrt{129}} = 3.26 \pm 0.18$$

$$95\%CI = 3.51 \pm 1.98 \frac{0.99}{\sqrt{166}} = 3.51 \pm 0.15$$

INFERENCE AND CONFIDENCE INTERVALS

when the sample sizes are different

Between subject studies

Defining the pooled variance as

$$pooled\ var = \frac{(n_1 - 1)var_1 + (n_2 - 1)var_2}{(n_1 - 1) + (n_2 - 1)}$$

the basic equation for the between-subject 95% confidence interval is

$$CI_{95\%} = \bar{x}_1 - \bar{x}_2 \pm t_{0.05} \sqrt{pooled\ var \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \qquad df = n_1 + n_2 - 2$$

$$CI_{95\%} = 3.51 - 3.26 \pm 1.98 \sqrt{1.03 \left(\frac{1}{166} + \frac{1}{129}\right)} = 0.25 \pm 0.24$$

mean for girls is higher

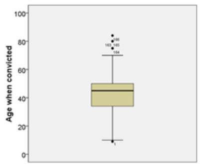
$$CI_{99\%} = 3.51 - 3.26 \pm 2.63 \sqrt{1.03 \left(\frac{1}{166} + \frac{1}{129}\right)} = 0.25 \pm 0.31$$

we cannot say that the mean for girls is higher

INFERENCE AND CONFIDENCE INTERVALS

Robust estimate of the standard error

Confidence intervals for medians



N = 166

	45	
34		50
9		84

Units = years
IQR = 16
Range = 75

Standard error of median estimate

$$k = \frac{n+1}{2} - z_{0.01} \sqrt{\frac{n}{4}}$$

$$se = \frac{X_{n-k+1} - X_k}{2 \times z_{0.01}} \qquad \text{Standard error } (X_i \text{ sorted})$$

$$95\%CI = median \pm z_{0.05} \times se = 45 \pm 3.80$$

Wilcox 2005 "robust estimates"

INFERENCE AND CONFIDENCE INTERVALS

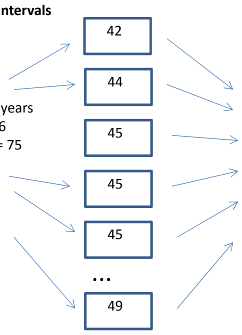
nonparametric evaluation of the median by bootstrapping

Bootstrapping confidence intervals

N = 166

	45	
34		50
9		84

Units = years
IQR = 16
Range = 75



Median $\pm 95\%CI$

Sampling from the data = sampling from the population

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

test of confidence intervalls within subjects

Null Hypotheses Significance Testing: Within subject t-test

H0: mean difference of population values equals zero

Significance testing is closely related to confidence interval construction. For the coffee example:

$$t = \frac{\overline{DIFF_i}}{se} = \frac{\overline{DIFF_i}}{sd/\sqrt{n}} = \frac{1.0}{1.25/\sqrt{10}} = 2.53 \qquad df = 9$$

$t_{0.05} = 2.26$ significant
$t_{0.01} = 3.25$ H0 not rejected

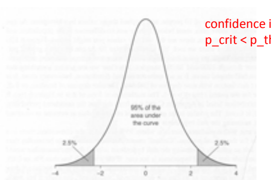


Figure 6.2 The t distribution with nine degrees of freedom

confidence interval contains zero if $t_{0.032} = 2.53$
p_crit < p_thresh: SIGNIFICANCE

reject H0 at the p=1% level
do not reject H0 at the p=5% level

p	
< 0.10	borderline evidence against H0
< 0.05	reasonable strong evidence against H0
< 0.025	strong evidence against H0
< 0.01	very strong evidence against H0

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

between subject tests need a pooled variance

Comparing two groups assuming equal variances: Between subject t-test

Let the standard deviations (or variances) be approximately same in two populations

With the pooled variance

$$pooled\ var = \frac{(n_1-1)\text{var}_1 + (n_2-1)\text{var}_2}{(n_1-1)+(n_2-1)}$$

the t-statistics becomes

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{pooled\ var(1/n_1 + 1/n_2)}} \qquad df = n_1 + n_2 - 2$$

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

acupuncture versus waiting

Comparing two groups not assuming equal variances

	acupunture	waiting	
Mean	-11..7	-6.1	$10.9^2/7.3^2 > 2$ SD's are different
Standard Deviation	7.3	10.9	
Total Number	12	11	

For acupuncture: $95\%CI = -11.7 \pm 2.20\frac{7.3}{\sqrt{12}} = -11.7 \pm 4.6$ $\quad t_{0.05} \quad df = 11$

For waiting: $95\%CI = -6.1 \pm 2.23\frac{10.9}{\sqrt{11}} = -6.1 \pm 7.3$ $\quad t_{0.05} \quad df = 10$

The difference between the means is given by: $95\%CI = (\bar{x}_1 - \bar{x}_2) \pm t_{0.05}\sqrt{\frac{\text{var}1}{n_1} + \frac{\text{var}2}{n_2}}$

take $t_{0.05}$ $df = 10$ this yields $95\%CI = -5.6 \pm 2.23(3.9) = -5.6 \pm 8.70$

Null hypothesee not rejected

t-statistics $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\text{var}1/n_1 + \text{var}2/n_2}} = \frac{-5.6}{3.9} = -1.44$ H0 not rejected

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

Wilcoxon and Mann-Whitney-Wilcoxon test as distribution free alternatives

The Wilcoxon signed rank test: An alternative to the within-subject t-test

```
  3   7   3   9  14   5   8  10  22   2    data
2.5   5 2.5   7   9   4   6   8  10   1    rank
tied    tied
```

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

retrievong times for happy memory and a sad memory

Table 6.1 *Hypothetical data comparing reaction times, in seconds, for retrieving a happy memory and a sad memory*

Participant	$HAPPY_i$	SAD_i	$DIFF_i$	$RANK_i$	$T+_i$	$T-_i$
1	3.9	10.4	−6.5	16	–	16
2	10.0	7.1	2.9	6.5	6.5	–
3	7.1	12.1	−5.0	14	–	14
4	9.7	6.3	3.4	8.5	8.5	–
5	10.0	8.8	1.2	4	4	–
6	9.9	5.1	4.8	13	13	–
7	6.1	14.9	−8.8	21	–	21
8	18.9	6.6	12.3	22	22	–
9	5.9	8.8	−2.9	6.5	–	6.5
10	2.8	11.0	−8.2	19.5	–	19.5
11	3.8	20.4	−16.6	23	–	23
12	5.7	11.4	−5.7	15	–	15
13	8.5	7.7	0.8	3	3	–
14	5.4	9.0	−3.6	10	–	10
15	6.2	14.4	−8.2	19.5	–	19.5
16	6.0	6.4	−0.4	2	–	2
17	6.4	14.5	−8.1	18	–	18
18	9.2	9.1	0.1	1	1	–
19	2.3	9.5	−7.2	17	–	17
20	7.0	30.6	−23.6	24	–	24
21	10.1	14.4	−4.3	12	–	12
22	6.1	9.8	−3.7	11	–	11
23	7.4	7.4	0.0	–	–	–
24	8.7	5.3	3.4	8.5	8.5	–
25	5.2	8.1	−2.9	5	–	5
					$\sum T+_i = 66.5$	$\sum T-_i = 233.5$

Definition of T

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

robust within subject test

Wilcoxon and Mann-Whitney-Wilcoxon test as distribution free alternatives

The Wilcoxon signed rank test: An alternative to the within-subject t-test

It is then $\quad z = \frac{T - n(n-1)/4}{\sqrt{n(n+1)(2n+1)/24}}$

Inserting T=66.5 and n=24 (since one participant is excluded) yields

$$z = \frac{66.5 - 24(23)/4}{\sqrt{24(25)(49)/24}} = -2.39$$

$z_{0.05} = 1.96$ significant

$z_{0.01} = 2.58$ H0 not rejected

That means the data is significant at p=0.05 level, but not at the p =0.01 level

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

number of correct plays out of 30

The Mann-Whitney-Wilcoxon test: An alternative to the between-subject t-test

Table 6.2 *The number of correct plays out of 30; data based on Helsen and Starkes (1999). The ranks are done for the entire sample, from 1 to 30*

Novices		Experts	
Number correct	Rank	Number correct	Rank
8	1.0	22	5.5
9	2.0	23	9.0
17	3.0	23	9.0
20	4.0	24	14.0
22	5.5	25	17.0
23	9.0	26	19.0
23	9.0	26	19.0
23	9.0	27	23.0
24	14.0	27	23.0
24	14.0	27	23.0
24	14.0	28	27.0
24	14.0	28	27.0
26	19.0	29	29.5
27	23.0	29	29.5
27	23.0		
28	27.0		
Sum of ranks = 190.5		Sum of ranks = 274.5.	

HYPOTHESIS TESTING: t TESTS AND ALTERNATIVES

robust between subject test

The Mann-Whitney-Wilcoxon test: An alternative to the between-subject t-test

The test statistic for the MWW, called the Mann-Whitney U, is the smaller of

$$\left(n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_1\right) \quad \text{and} \quad \left(n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_2\right)$$

For these data, these values are

$$\left((16)(14) + \frac{16(16+1)}{2} - 190.5\right) = 169.5$$
$$\left((16)(14) + \frac{14(14+1)}{2} - 274.5\right) = 54.5$$

So that $U = 54.5$. The corresponding z-value (z-statistic) is given by

$$z = \frac{n_1 n_2 / 2 - U}{\sqrt{(n_1 n_2 / 12)(n_1 + n_2 + 1)}}$$

$$z = \frac{(16)(14)/2 - 54.5}{\sqrt{((16)(14)/12)(16+14+1)}} = 2.39 \qquad z_{0.05} = 1.96$$ reject H0 at 5% level

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

ANOVA – Analysis of variance

Within subject design — comparing the values of several variables for one group

Between subject design — comparing the values of one variable for several groups

Here we look at:

- one-way between subjects ANOVA
- repeated measures ANOVA

The corresponding nonparametric tests are the
- Kruskal-Wallis test
- Friedman test

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

score: -5...5 for a task

ANOVA – Analysis of variance – one way between subjects

Table 7.1 *Data created to match very closely Festinger and Carlsmith's (1959) classic study of cognitive dissonance*

Control (mean = −0.45) x_{i1}	$(x_{i1} - \bar{x}_{\bullet 1})^2$	\$1 (mean = 1.35) x_{i2}	$(x_{i2} - \bar{x}_{\bullet 2})^2$	\$20 (mean = −0.05) x_{i3}	$(x_{i3} - \bar{x}_{\bullet 3})^2$
0	0.20	3	2.72	1	1.10
−3	6.50	1	0.12	2	4.20
3	11.90	1	0.12	3	9.30
2	6.00	3	2.72	0	0.00
−2	2.40	2	0.42	1	1.10
−1	0.30	3	2.72	3	9.30
2	6.00	3	2.72	0	0.00
3	11.90	2	0.42	−2	3.80
−3	6.50	2	0.42	2	4.20
−5	20.70	2	0.42	1	1.10
2	6.00	2	0.42	0	0.00
−3	6.50	2	0.42	0	0.00
3	11.90	−4	28.62	−1	0.90
0	0.20	4	7.02	−2	3.80
−2	2.40	0	1.82	−1	0.90
−2	2.40	−3	18.92	−4	15.60
−2	2.40	4	7.02	−3	8.70
−2	2.40	1	0.12	−1	0.90
−1	0.30	1	0.12	0	0.00
2	6.00	−2	11.22	0	0.00
SS = 112.95 var = 5.94 df = 19		SS = 88.55 var = 4.66 df = 19		SS = 64.95 var = 3.42 df = 19	

$$GM = \frac{1}{n}\sum_i x_i = 0.28$$

$$n = 60$$

$$\bar{x}_{\bullet 1} = -0.45$$

$$\bar{x}_{\bullet 2} = 1.35$$

$$\bar{x}_{\bullet 3} = -0.05$$

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

model equation for score – one way between subjects

$$X_{ij} = \bar{X}_{\bullet\bullet} + (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) + (X_{ij} - \bar{X}_{\bullet j})$$
$$X_{ij} = \mu + (\mu_j - \mu) + (X_{ij} - \mu_j)$$
$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{\bullet\bullet})^2}_{\text{SSTO TOTAL}} = \underbrace{n\sum_{j=1}^{p}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2}_{\text{SSBG TREATMENT}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{\bullet j})^2}_{\text{SSWG ERROR}}$$

$$(np-1) = (p-1) + p(n-1)$$

$$F = \frac{\text{MSBG}}{\text{MSWG}}$$

α_j = part of X_{ij} due to treatment

ε_{ij} = part of X_{ij} due to error

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

Sum of Squares

ANOVA – Analysis of variance – one way between subjects

SSTO = SSBG + SSWG
Total = Model + Error

$$\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = n\sum_{j=1}^{p}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet j})^2$$

$$F = \frac{\text{MSBG}}{\text{MSWG}} = \frac{\text{model}}{\text{error}}$$

total variance in the data = variance explained by the model + unexplained variance

model = treatment effect

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

Sum of Squares

ANOVA – Analysis of variance

Total variation $= \sum_i (x_i - GM)^2 = 302.18$

Total variation = within-group variation + between-groups variation

Within-group variation $= 112.95 + 88.55 + 64.95 = 266.45$

Between-groups variation $= 302.18 - 266.45 = 35.73 = \sum_{j=1}^{3} n_j (\bar{x}_{\bullet j} - GM)^2$

Within-group variation $= 266.45/302.18 = 88\%$

Between-groups variation $= 12\%$ — 12% of the variation is explained by the differences between the groups

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

ANOVA – Analysis of variance

mean sums of squares error

$$MSe = \frac{within-group\ variation}{df_e} = \frac{266.45}{57} = 4.67$$

$$MSb = \frac{between-group\ variation}{df_b} = \frac{35.73}{2} = 17.87$$

$$F(2,57) = \frac{MSb}{MSe} = \frac{17.87}{4.67} = 3.83 \qquad F_{0.05}(2,50) = 3.18$$ significant

the means are different

ANOVA table

	Sum of squares	df	Mean square	F	p	eta-sq
Between	35.733	2	17.867	3.822	0.028	0.12
Within	266.450	57	4.675			
Total	302.183	59				

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

company efficiencies and seasons

ANOVA – Analysis of variance – repeated measures ANOVA

Table 7.3 *Company efficiency and seasons*

Company	Autumn	Winter	Spring	Summer	$\bar{x}_{j\cdot}$ (*Mean*$_j$.)
1	30	24	35	28	29.25
2	34	31	52	47	41.00
3	30	45	41	42	39.50
4	51	58	66	52	56.75
5	67	55	77	69	67.00
6	35	56	58	61	52.50
$\bar{x}_{\cdot j}$ (*Mean*.$_j$)	41.17	44.83	54.83	49.83	
			Grand mean (GM) or 'a mean for all seasons' – sorry!		47.67

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

model equation for score in a randomized block design

$$X_{ij} = \bar{X}_{\bullet\bullet} + (\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet}) + (X_{i\bullet} - \bar{X}_{\bullet\bullet}) + (X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})$$

$$X_{ij} = \mu + (\mu_{\bullet j} - \mu) + (\mu_{i\bullet} - \mu) + (X_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu)$$

$$X_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}$$

$$\underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{\bullet\bullet})^2}_{\substack{\text{SSTO}\\ \text{TOTAL}}} = \underbrace{n\sum_{j=1}^{p}(\bar{X}_{\bullet j} - \bar{X}_{\bullet\bullet})^2}_{\substack{\text{SSM}\\ \text{TREATMENT}}} + \underbrace{p\sum_{i=1}^{n}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2}_{\substack{\text{SSBG}\\ \text{SSBlocks}}} + \underbrace{\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij} - \bar{X}_{i\bullet} - \bar{X}_{\bullet j} + \bar{X}_{\bullet\bullet})^2}_{\substack{\text{SSE}\\ \text{ERROR}}}$$

$$(np-1) = (p-1) + (n-1) + (n-1)(p-1)$$

$$\text{SSWG} = \sum_{i=1}^{n}(X_{ij} - \bar{X}_{i\bullet})^2 = \text{SSM} + \text{SSE}$$

$F = \frac{\text{MSBG}}{\text{MSE}}$ population mean for blocks (enterprises) equal

$F = \frac{\text{MSM}}{\text{MSE}}$ population mean for treatments (seasons) equal

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

company efficiencies and seasons

ANOVA – Analysis of variance – repeated measures ANOVA

SSTO = SSBG + SSWG

SSWG = SSM (model) + SSE (error)

$$SSTO = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2$$

$$SSWG = \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2 \qquad df_{WG} = n(p-1)$$

$$SSM = n\sum_{j=1}^{p}(\bar{x}_{\bullet j} - \bar{x}_{\bullet\bullet})^2 \qquad df_M = (p-1)$$

$$SSE = SSWG - SSM \qquad df_{err} = df_{WG} - df_M \qquad F = \frac{MS_{model}}{MS_{error}}$$

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

company efficiencies and seasons

ANOVA – Analysis of variance – repeated measures ANOVA

$$SS_{WITHIN} = \sum_{i=1}^{4}\sum_{j=1}^{6}(x_{ij} - \bar{x}_{i\bullet})^2 = (30-29.25)^2 + (24-29.25)^2 + \ldots + (61-52.5)^2 = 1309.5$$

$$SS_{WITHIN} = SS_e + SS_{TREATMENT}$$ or $$SS_{SUBJECT\times TREATMENT} = SS_e + SS_{MODEL}$$

$$SS_{TREATMENT} = n\sum(x_{\bullet j} - GM)^2 = 6\times\left[(41.17-47.67)^2 + \ldots + (49.83-47.67)^2\right] = 638.00$$

Variability between the treatment levels = between seasons

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

ANOVA – Analysis of variance

The final necessary sum of squares for the error is

$$SS_e = SS_{WITHIN} - SS_{TREATMENT} = 671.50$$

The degrees of freedom are

p-1 for treatments where k is the number of treatments (4-1=3)
(n-1)(p-1) for error where n is the sample size ((6-1)(4-1)= 15)
n(p-1) for within ((6)(4-1) =18)

The MS values are calculated in the same way as the between subject ANOVAs

$$MS_{TREATMENT} = SS_{TREATMENT} / df_{TREATMENT} = 638.00/3 = 212.67$$

$$MS_e = SS_e / df_e = 671.50/15 = 44.77$$

$$F = MS_{TREATMENT} / MS_e = 4.75$$

effect size $\eta_p^2 = SS_{TREATMENT} / SS_{WITHIN} = 0.487$

COMPARING MORE THAN TWO GROUPS OR MORE THAN TWO VARIABLES

ANOVA – Analysis of variance

```
ANOVA table
         Sum of squares   df      Mean square        F        p        eta-sq p
Treatment  638.00          3      212.67            4.75     0.016    0.487
Error       671.50        15       44.77
Within     1309.50        18
```

efficiencies are different to a 0.016 p-value for the different seasons

REGRESSION AND CORRELATION

actual vs. estimated car velocities

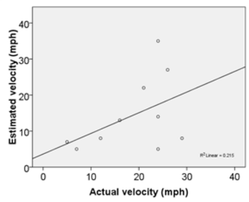
Regression line

Table 8.1 *Data and some preliminary calculations for estimating the velocity of cars. The final row is the sum of all the values in that column. Estimate$_i$ = E_i; Actual$_i$ = A_i; m(x) = mean(x); predicted value = pred$_i$; error (or residual) = e_i*

E_i	A_i	$E_i - m(E)$	$A_i - m(A)$	$(E_i - m(E))(A_i - m(A))$	$(E_i - m(E))^2$	$(A_i - (A))^2$	pred$_i$	e_i	e_i^2
8	12	−6.4	−6.8	43.5	41.0	46.2	10.5	−2.5	6.3
8	29	−6.4	10.2	−65.3	41.0	104.0	20.3	−12.3	150.1
35	24	20.6	5.2	107.1	424.4	27.0	17.4	17.6	310.4
13	16	−1.4	−2.8	3.9	2.0	7.8	12.8	0.2	0.0
5	7	−9.4	−11.8	110.9	88.4	139.2	7.6	−2.6	6.9
7	5	−7.4	−13.8	102.1	54.8	190.4	6.5	0.5	0.3
14	24	−0.4	5.2	−2.1	0.2	21.0	17.4	−3.4	11.4
22	21	7.6	2.2	16.7	57.8	4.8	15.7	6.3	40.2
27	26	12.6	7.2	90.7	158.8	51.8	18.5	8.5	71.8
5	24	−9.4	5.2	−48.9	88.4	27.0	17.4	−12.4	153.3
Sum=144	188	0	0	358.8	956.4	625.6	144	0	750.6

Note: Values printed have been rounded to the tenth place, but for calculations more decimals places are used.

$$e_i = E_i - pred_i$$



$$y_i = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

REGRESSION AND CORRELATION

Regression line

$SS_{total} = 956.4$ $\quad \mathrm{var} = SS_{total}/(n-1) = 106.3$ $\quad n-1 = 9$

$SS_{error} = 750.6$

$$SS_{total} = SS_{error} + SS_{model}$$

$SS_{model} = 205.8$ $\quad 205.8/956.4 = 0.22 = 22\%$ can be accounted for the model

Model number of degrees of freedom = k =1 = number of variables used in the equation

$MSS_{model} = SS_{model}/1 = 205.8$

Number of degrees of freedom associated with the error term = n-k-1 = 8

$MSS_{error} = SS_{error}/8 = 93.8$

$F(1,8) = 205.8/93.8 = 2.19$

$t(8) = \sqrt{2.19} = 1.48$ $\qquad p = 0.18$

We should not reject the hypothesis that there is no relationship between actual and estimated velocities

REGRESSION AND CORRELATION

Pearson's correlation

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

$$r = \frac{358.8}{\sqrt{(625.6)(956.4)}} = 0.46$$

r = 0.1, 0.3, 0.5 'small', 'medium', 'large'

t-test if r differs significantly from 0:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Wright and London give also a 95% *CI* for r.

REGRESSION AND CORRELATION

Spearman's r_s

Perform Pearson's correlation on the ranks of the data

compare to:

Perform Pearson's correlation on the log-log data

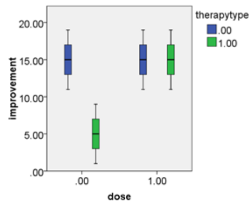
FACTORIAL ANOVAs AND MULTIPLE REGRESSION

2 x 2 design

Two-Way ANOVA

Table 9.1 *Data for a 2×2 two-way ANOVA showing the amount of improvement in depression depending on the type of psychotherapy and drug dose*

	Psychotherapy 1		Psychotherapy 2	
Drug dose:	Low	Medium	Low	Medium
Improvement	11	11	1	11
	13	13	3	13
	15	15	5	15
	17	17	7	17
	19	19	9	19
Mean (*M*)	15	15	5	15
Variance	10	10	10	10
SS_{within}	40	40	40	40



FACTORIAL ANOVAs AND MULTIPLE REGRESSION

model equation for score in a completely randomized factorial p x q design

$$X_{ijk} = \bar{X}_{\bullet\bullet\bullet} + (\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet}) + (X_{\bullet\bullet k} - \bar{X}_{\bullet\bullet\bullet}) + (X_{\bullet jk} - \bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet k} + \bar{X}_{\bullet\bullet\bullet}) + (X_{ijk} - \bar{X}_{\bullet jk})$$

$$X_{ijk} = \mu + (\mu_{j\bullet} - \mu) + (\mu_{\bullet k} - \mu) + (\mu_{jk} - \mu_{\bullet k} - \mu_{j\bullet} + \mu) + (X_{ijk} - \mu_{jk})$$

$$X_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

$$\underbrace{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}(X_{ijk} - \bar{X}_{\bullet\bullet\bullet})^2}_{\substack{\text{SSTO}\\\text{TOTAL}}} = \underbrace{nq\sum_{j=1}^{p}(\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet})^2}_{\substack{\text{SSA}\\\text{TREATMENT 1}}} + \underbrace{np\sum_{k=1}^{q}(\bar{X}_{\bullet\bullet k} - \bar{X}_{\bullet\bullet\bullet})^2}_{\substack{\text{SSB}\\\text{TREATMENT 2}}} +$$

$$+ \underbrace{n\sum_{j=1}^{p}\sum_{k=1}^{q}(\bar{X}_{\bullet jk} - \bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet k} + \bar{X}_{\bullet\bullet\bullet})^2}_{\substack{\text{SSAB}\\\text{INTERACTION}}} + \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}(X_{ijk} - \bar{X}_{\bullet jk})^2}_{\substack{SSWG\\ERROR}}$$

$$(npq-1) = (p-1) + (q-1) + (p-1)(q-1) + pq(n-1)$$

$F = \frac{\text{MSA}}{\text{MSE}}$ population mean for treatment 1 equal

$F = \frac{\text{MSB}}{\text{MSE}}$ population mean for treatment 2 equal

$$SS_{within,1} = (11-15)^2 + (13-15)^2 + (15-15)^2 + (17-15)^2 + (19-15)^2 = 40 \qquad pq(n-1) = 16$$

$$SS_{err} = SS_{within} = \sum SS_{within,i} = 160 \qquad GM = 12.5$$

$$SS_{PSYCH} = 10(15-12.5)^2 + 10(10-12.5)^2 = 125 \qquad p-1=1$$

$$SS_{DOSE} = 10(10-12.5)^2 + 10(15-12.5)^2 = 125 \qquad k-1=1$$

$$SS_{TOTAL} = \sum (x_i - GM)^2 = 535 \qquad npq-1=19$$

$$SS_{interaction} = 535 - 125 - 125 - 4 \times 40 = 125 \qquad (p-1)(q-1)=1$$

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

2 x 2 design

Two-Way ANOVA

Table 9.2 *An ANOVA table for the data in Table 9.1. The main effects and interaction are all statistically significant*

Effect	SS	df	MS	F	p	η_p^2
Psychotherapy	125	1	125	12.50	0.003	0.44
Dose level	125	1	125	12.50	0.003	0.44
Interaction	125	1	125	12.50	0.003	0.44
Within (error)	160	16	10			
Total	535	20 (including 1 for the constant)				

$F_{effect} = MS_{effect} / MS_{err}$

$\eta_p^2 = SS_{effect} / (SS_{effect} + SS_{err})$

F for entire ANOVA:

$MS_{model} = \sum SS_{effects} / \sum df_{effects} = 375/3 = 125$

$F(3,16) = MS_{model} / MS_{err} = 12.5 \qquad p = 0.001$

$\eta^2 = 0.70$

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

2 x 5 design

2 x 5 design

$$\frac{1}{11}\sum_{i=1}^{12}(X_{ijk} - X_{\bullet\bullet\bullet})^2$$

Table 9.3 *The variances for each of the 10 conditions, as well as the means for the two alcohol groups, the five task groups, and the whole sample*

	Task complexity					
	0	2	4	6	8	Row means
Placebo	27.38	21.38	19.50	20.21	19.74	0.068
Alcohol	13.25	40.35	20.54	19.77	28.17	0.943
Column means	1.79	−0.22	−1.69	−0.27	2.92	var₊₊ = 27.94 mean₊₊ = 0.506

$X_{\bullet j\bullet}$ (row means); $X_{\bullet\bullet k}$ (column means)

$$SS_{total} = var(n_{++} - 1) = 27.94 \cdot 119 = 3325.34$$

$$SS_{within} = \sum_{ij} var_{ij}(n_{ij} - 1) = 27.38 \cdot 11 + \ldots + 28.17 \cdot 11 = 2532.89$$

$$SS_{DRINK} = \sum_{i+} (mean_{i+} - mean_{++})^2 n_{i+}$$

$$= (0.068 - 0.506)^2 \cdot 60 + (0.943 - 0.506)^2 \cdot 60 = 23.03$$

$$SS_{TASK} = \sum_{+j} (mean_{+j} - mean_{++})^2 n_{+j}$$

$$= (1.79 - 0.506)^2 \cdot 24 + \cdots + (2.92 - 0.506)^2 \cdot 24 = 322.20$$

$$SS_{DRINK \times TASK} = SS_{total} - SS_{within} - SS_{DRINK} - SS_{TASK}$$

$$= 3325.34 - 2532.89 - 23.03 - 322.20 = 447.42$$

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

2 x 5 design

2 x 5 design

Table 9.5 *An ANOVA table with the values filled in*

Effect	SS	df	MS	F	p	η_p^2
DRINK	23.03	1	23.03	1.00	0.32	0.01
TASK	322.20	4	80.55	3.50	0.01	0.11
Interaction	447.22	4	111.80	4.86	0.001	0.15
Within (error)	2532.89	110	23.03			
Total	3325.34	119				

$F_{effect} = MS_{effect} / MS_{err}$

$\eta_p^2 = SS_{effect} / (SS_{effect} + SS_{err})$

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

Multiple Regression

Table 9.6 *Data for the multiple regression example. These are available on the book's web page*

No.	Kindness	Income	Charity	No.	Kindness	Income	Charity	No.	Kindness	Income	Charity
1	4	37	3.852	18	3	24	3.709	35	5	28	2.953
2	11	16	3.270	19	8	28	3.958	36	8	23	3.730
3	10	19	2.772	20	6	16	2.644	37	6	31	4.234
4	3	31	3.973	21	5	27	3.209	38	9	24	4.156
5	9	17	2.623	22	1	34	4.599	39	6	32	3.950
6	7	26	3.987	23	8	22	2.934	40	1	32	3.958
7	7	29	3.068	24	8	24	3.786	41	2	21	2.793
8	11	31	5.132	25	7	20	1.800	42	1	26	2.559
9	9	17	3.265	26	1	33	3.363	43	7	28	3.595
10	7	14	2.941	27	9	12	2.993	44	1	38	3.992
11	9	24	4.095	28	5	24	4.488	45	5	29	4.399
12	8	16	3.117	29	0	19	3.052	46	2	23	2.710
13	7	24	3.867	30	1	37	3.852	47	6	19	2.950
14	10	24	3.775	31	1	25	2.524	48	3	33	4.154
15	5	17	3.089	32	7	20	2.453	49	5	32	4.242
16	6	21	2.391	33	3	17	1.434	50	7	26	4.402
17	9	25	3.846	34	2	24	1.583				

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

Multiple Regression

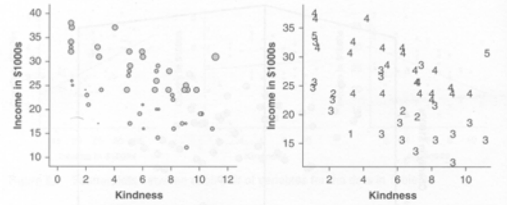


Figure 9.5 Two scatterplots of income with kindness. In the left panel the amount given to charity is proportional to the width of the circles. In the right panel the amount given is shown with its numeral

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

Multiple Regression

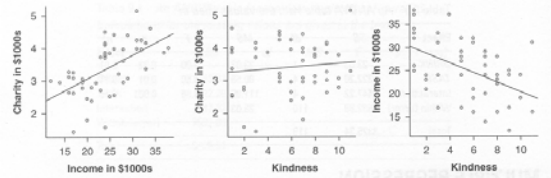


Figure 9.3 Scatterplots between each pair of variables for the data in Table 9.6

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

Multiple Regression

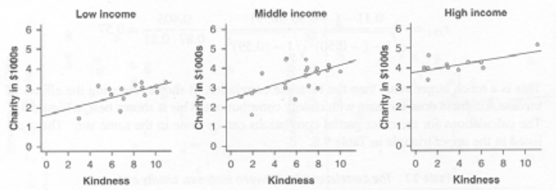


Figure 9.6 Scatterplots between charity contributions with kindness plotted separately for low, middle, and high income individuals

FACTORIAL ANOVAs AND MULTIPLE REGRESSION

Multiple Regression

partial correlation: $r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2}\sqrt{1 - r_{yz}^2}}$

Table 9.8 *The correlations are shown in the lower left-hand triangle, the partial correlations are shown in the upper right-hand triangle, and histograms showing the distributions are plotted along the diagonal*

	Kindness	Charity	Income
Kindness		0.57	−0.69
Charity	+0.11		0.74
Income	−0.50	+0.59	

CATEGORICAL DATA ANALYSIS

Effect size measures for 2 x 2 tables

Table 10.1 *A contingency table, sometimes called cross-tabs or cross-tabulation, for the frequency of correct and incorrect choices for each of the four confederates broken down by race of the participant and whether the confederate was identified. The first number is the frequency. The numbers in parentheses are the percentages. Below these are the odds of a correct response. Data are from Wright et al. (2001: Table 1)*

	Black confederate		White confederate	
Sample	Blacks	Whites	Blacks	Whites
South Africa				
Number correct	17 (68%)	8 (68%)	15 (60%)	21 (84%)
Number incorrect	8 (32%)	17 (32%)	10 (40%)	4 (16%)
Odds of correct response	2.125	0.471	1.500	5.250
England				
Number correct	19 (95%)	24 (77%)	8 (35%)	14 (52%)
Number incorrect	1 (5%)	7 (23%)	15 (65%)	13 (48%)
Odds of correct response	19.000	3.419	5.333	1.077

Odds ratio for the white participants viewing a white confederate in South Africa = 5.25/1.50 = 3.5

CATEGORICAL DATA ANALYSIS

Effect size measures for 2 x 2 tables

Table 10.2 *Data for identifying the white confederate in the South African data from Wright et al. (2001). Also shown are the equations for three measures of effect size: the odds ratio, phi and Cohen's κ*

	Participants' race		Three measures of effect size	
	White	Black		
Correct	A 21	B 15	odds ratio	AD/BC
Incorrect	C 4	D 10	phi	$\frac{(AD - BC)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$
			Cohen's κ	$\frac{2(AD - BC)}{(A+B)(B+D)+(C+D)(A+C)}$

cf. Pearson's r

0-0.2 poor, 0.2-0.4 fair, 0.4-0.6 moderate, 0.6-0.8 substantial, >0.8 almost perfect

CATEGORICAL DATA ANALYSIS

Effect size measures for 2 x 2 tables

1. Take the ln of the observed OR. Here, ln (3.50) = 1.253.
2. Calculate the standard error on the log odds ratio:

$$se(\ln OR) = \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}} \quad \text{here} \quad \sqrt{\frac{1}{21}+\frac{1}{15}+\frac{1}{4}+\frac{1}{10}} = 0.681$$

3. Calculate the 95% confidence interval of ln *OR*:
   lower bound = ln *OR* – 1.96 *se*(ln *OR*) here 1.253 – 1.96(0.681) = –0.082
   upper bound = ln *OR* + 1.96 *se*(ln *OR*) here 1.253 + 1.96(0.681) = 2.588
4. Back-transform these into odds ratios by exponentiating them (with the EXP or e^x key on your calculator):

   $\exp(-0.082) = e^{-0.082} = 0.92$
   $\exp(2.588) = e^{2.588} = 13.30$

Thus, the 95% confidence interval goes from 0.92 (just below chance which is 1.00) to 13.30. Note that the observed value (3.50) is not halfway between these.

CATEGORICAL DATA ANALYSIS

Effect size measures for 2 x 2 tables

Null hypotheses: No association between the two variables.

$$\chi^2 = \frac{n(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} = 3.57 \quad \text{or} \quad \chi^2 = \sum SR_{ij}^2 = 3.56 \qquad p = 0.06$$

$$df = (2-1)(2-1) = 1$$

Table 10.4 *Observed data (O_{ij}) for the white confederate in South Africa (Wright et al., 2001), showing the calculations for expected values (E_{ij}) and the standardized residuals (SR_{ij}) for each cell. These can be used to calculate the χ^2 statistic for the entire contingency table*

	Black	White (*RT*)	Row total	
Correct	$O_{11} = 15$ $E_{11} = 18$ $SR_{11} = -0.71$	$O_{12} = 21$ $E_{12} = 18$ $SR_{12} = 0.71$	$RT_1 = 36$	$E_{ij} = \frac{RT_i CT_j}{n}$
Incorrect	$O_{21} = 10$ $E_{21} = 7$ $SR_{21} = 1.13$	$O_{22} = 4$ $E_{22} = 7$ $SR_{22} = -1.13$	$RT_2 = 14$	$SR_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$
Column total (CT_j)	$CT_1 = 25$	$CT_2 = 25$	$n = 50$