Analyzing Data

Instructor: Nguyen Ngoc Vu, Ph.D.

Preliminary Analysis

Descriptive statistics

Measures of central tendency

- Mean
- Median
- Mode

Mean

- Sum of the values divided by the number of cases

$$\bar{x} = \frac{\sum x_i}{n}$$

Calculating the mean for high temperatures

Date	High Temperature
2-Jan	59
3-Jan	60
4-Jan	43
5-Jan	42
6-Jan	35
7-Jan	32
8-Jan	32
9-Jan	46
10-Jan	41
11-Jan	52
Sum	442

- Add values

$$\sum x_i = 442$$

- Number of cases

$$n = 10$$

- Calculate mean

$$\bar{x} = \frac{\sum x_i}{n} = \frac{442}{10} = 44.2$$

Median

- The median represents the middle of the ordered sample data
- When the sample size is odd, the median is the middle value
- When the sample size is even, the median is the midpoint/mean of the two middle values

Calculating the median for high temperatures

Date	High Temperature	
7-Jan	32	
8-Jan	32	
6-Jan	35	
10-Jan	41	
5-Jan	42	**<===Middle values**
4-Jan	43	**<===Middle values**
9-Jan	46	
11-Jan	52	
2-Jan	59	
3-Jan	60	

$$median = \frac{42+43}{2} = 42.5$$

Mode

- The mode is the value that occurs most frequently
- It is the least useful (and least used) of the three measures of central tendency
- The mode may help to correct false impressions if you know the mean and the median but don't actually see the data.
- A set of data can be bimodal, multimodal or with no mode.

e.g. ***101 99 1 1***

The mean is (101 + 99 + 1 + 1)/4 = 202/4 = 50.5 and the median = (99+1)/2 = 50. But the mode here is ***1. In this case, the mean and median values are misleading.***

Calculating the mode for high temperatures

Date	High Temperature	
2-Jan	59	
3-Jan	60	
4-Jan	43	
5-Jan	42	
6-Jan	35	
7-Jan	32	**<===Mode**
8-Jan	32	**<===Mode**
9-Jan	46	
10-Jan	41	
11-Jan	52	

$$mode = 32$$

Measures of central tendency and levels of measurement

- Mean assumes numerical values and requires interval or ratio data
- Median requires ordering of values and can be used with ratio, interval and ordinal data
- Mode only involves determination of most common value and can be used with ratio, interval, ordinal, and nominal data

Comparison of mean and median

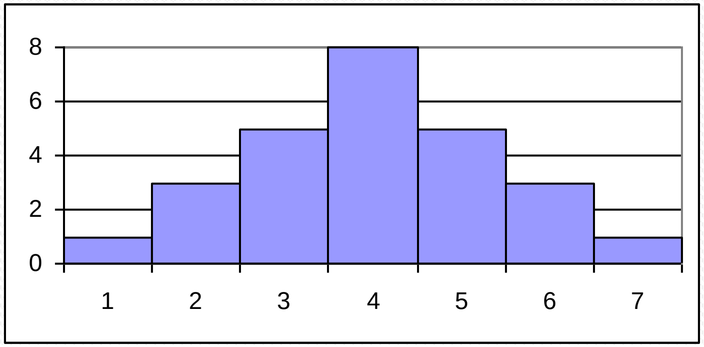
- Mean
  - Uses all of the data
  - Has desirable statistical properties
  - Affected by extreme high or low values (outliers)
  - May not best characterize skewed distributions
- Median
  - Not affected by outliers
  - May better characterize skewed distributions

The mean and median and the distribution of the data

- For symmetric distributions, the mean and the median are the same
- For skewed distributions, the mean lies in the direction of the skew (the longer tail) relative to the median

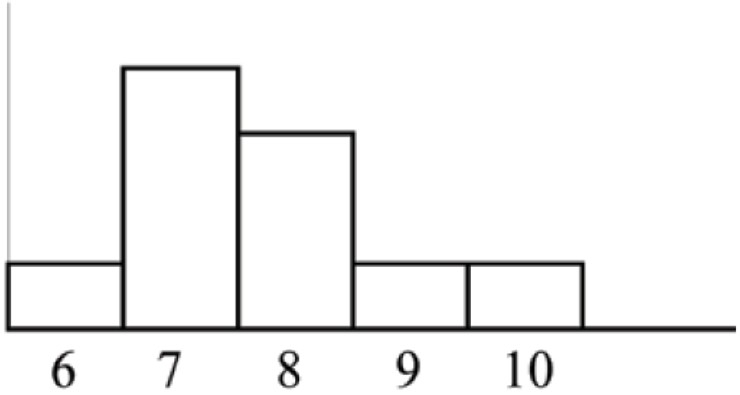
Distribution shapes

Symmetric: bell shaped



Positively skewed

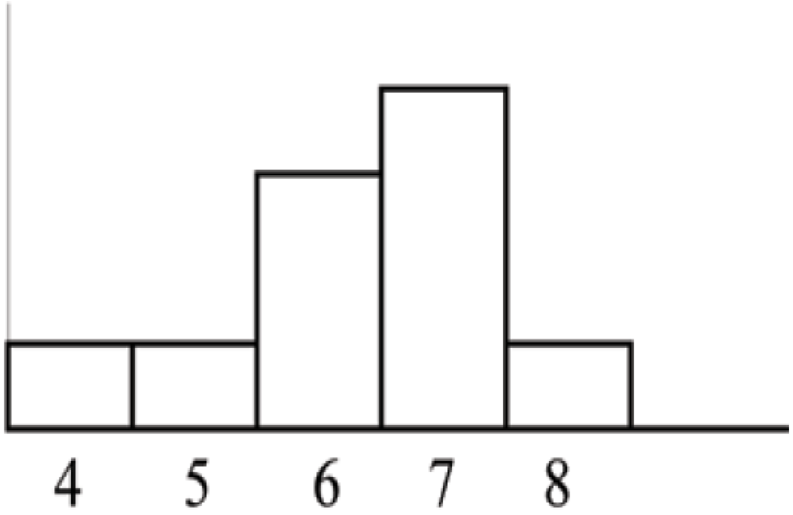
e.g.3 Distribution skewed to the right (Data set: 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8 ; 8 ; 8 ; 9 ; 10)



The mean is 7.7, the median is 7.5, and the mode is 7. *Notice that the mean is the largest statistic, while the mode is the smallest.* Again, the mean reflects the skewing the most. (Positively skewed)

Negatively skewed

e.g.2 Distribution skewed to the left (Data set: 4 ; 5 ; 6 ; 6 ; 6 ; 7 ; 7 ; 7 ; 7 ; 7 ; 7 ; 8)



The mean is 6.3, the median is 6.5, and the mode is 7. *Notice that the mean is less than the median and they are both less than the mode.* The mean and the median both reflect the skewing, but the mean more so. (Negatively skewed)

Measures of variation

- Range
- Variance and standard deviation
- Interquartile range

Range

- Range is the difference between the minimum and maximum values

Calculating the range for high temperatures

	High	
Date	**Temperature**	
7-Jan	32	**<===Lowest Value**
8-Jan	32	
6-Jan	35	
10-Jan	41	
5-Jan	42	
4-Jan	43	
9-Jan	46	
11-Jan	52	
2-Jan	59	
3-Jan	60	**<===Highest Value**

$$range = 60 - 32 = 28$$

Variance and standard deviation

- The variance s^2 is the sum of the squared deviations from the mean divided by the number of cases minus 1

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- The standard deviation s is the square root of the variance

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Calculating the variance and standard deviation for high temperatures

	High	Difference	Difference
Date	Temperature	X - mean	Squared
2-Jan	59	14.80	219.04
3-Jan	60	15.80	249.64
4-Jan	43	-1.20	1.44
5-Jan	42	-2.20	4.84
6-Jan	35	-9.20	84.64
7-Jan	32	-12.20	148.84
8-Jan	32	-12.20	148.84
9-Jan	46	1.80	3.24
10-Jan	41	-3.20	10.24
11-Jan	52	7.80	60.84
Sum	442		931.60
n	10		
Mean	44.2		

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{931.60}{10-1} = 103.51$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{103.51} = 10.2$$

Interpretation of standard deviation

- If distribution of data approximately bell shaped, then
  - About 68% of the data fall within one standard deviation of the mean
  - About 95% of the data fall within two standard deviations of the mean
  - Nearly all of the data fall within three standard deviations of the mean

Interquartile range

- Difference between upper (third) and lower (first) quartiles
- Quartiles divide data into four equal groups
    - Lower (first) quartile is 25th percentile
    - Middle (second) quartile is 50th percentile and is the median
    - Upper (third) quartile is 75th percentile

Calculating the interquartile range for high temperatures

Date	High Temperature		
7-Jan	32		
8-Jan	32		
6-Jan	35	**<===Bottom Half Middle Value = First Quartile = 35**	
10-Jan	41		
5-Jan	42	**<===Middle Value**	**Median = Second Quartile = 42.5**
4-Jan	43	**<===Middle Value**	
9-Jan	46		
11-Jan	52	**<===Top Half Middle Value = Third Quartile = 52**	
2-Jan	59		
3-Jan	60		

$$\textit{interquartile range} = 52 - 35 = 17$$

Interquartile range and outliers

- Value can be considered to be an outlier if it falls more than 1.5 times the interquartile range above the upper quartile or more than 1.5 times the range below the lower quarter
- Example for high temperatures
  - Interquartile range is 17
  - 1.5 times interquartile range is 25.5
  - Outliers would be values
    - Above 52 + 25.5 = 77.5 (none)
    - Below 25 – 25.5 = 9.5 (none)

Comparison of range, standard deviation, and interquartile range

- Sensitivity to extreme values
  - Range – extremely sensitive
  - Standard deviation – very sensitive
  - Interquartile range – not sensitive
- Standard deviation
  - Has desirable statistical properties
  - Suggests numbers of cases in different intervals for bell-shaped distributions

Statistical techniques to explore relationships among variables

Correlation

Correlational Research

- The purpose of correlational research is to discover relationships between two or more variables.
- *Relationship* means that an individuals status on one variable tends to reflect his or her status on the other.

Correlational Research

- Helps us understand related events, conditions, and behaviors.
  - Is there a relationship between educational levels of farmers and crop yields?
- To make predictions of how one variable might predict another
  - Can high school grades be used to predict college grades?

Pearson Product-Moment Correlation

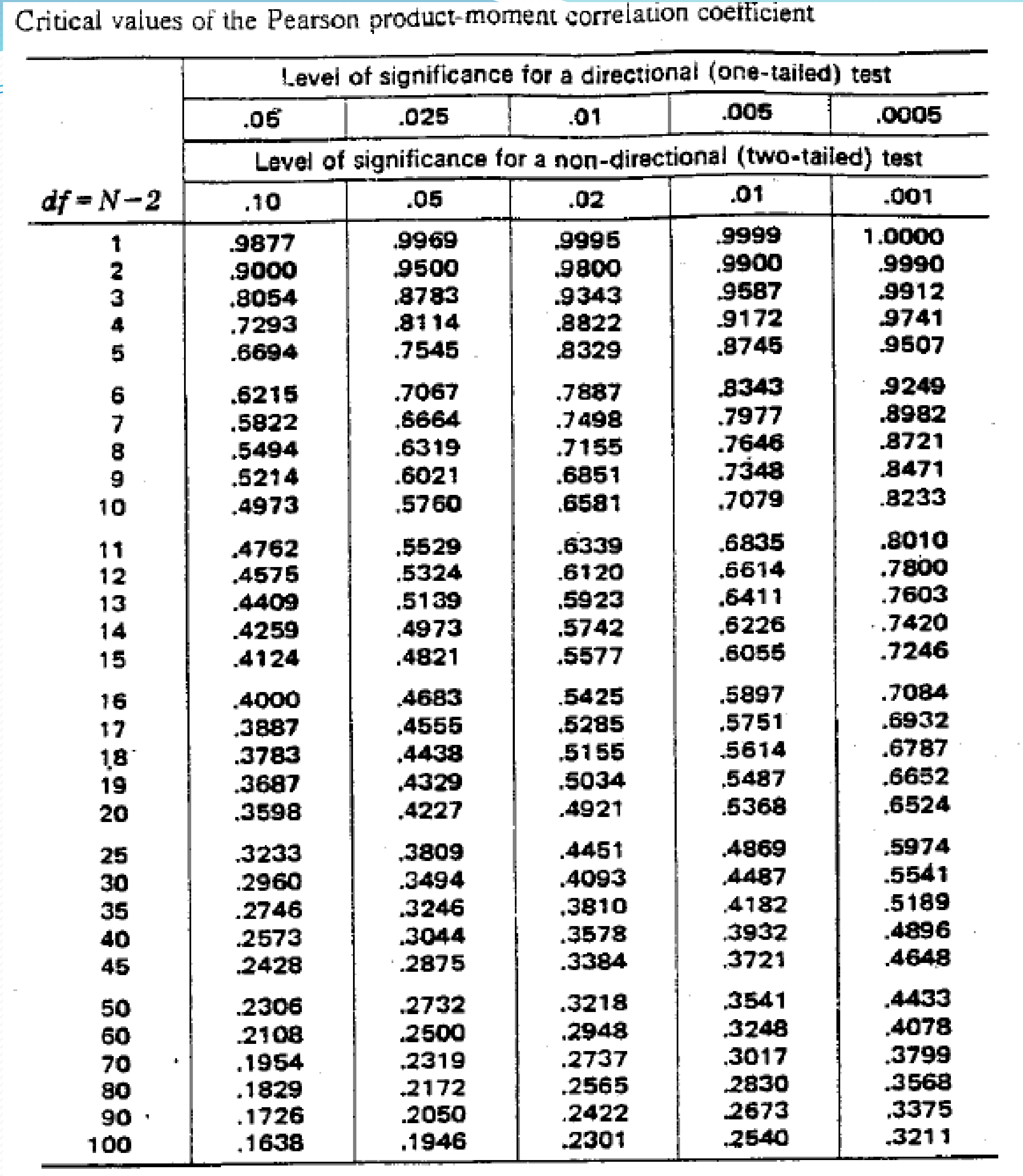
- Used when both the criterion and predictor variable contain continuous interval data such as test scores, years of experience, money, etc.

Examples of when to use the Pearson Correlation

Predictor Variable	Criterion Variable
Years of Experience in Extension	Job Satisfaction score
Family Income	End of Course (EOC) Test Scores
Distance from Krispy Kreme donut shop.	Weight

Formula

$$r_{Exp} = \frac{N_{Exp}(\sum X_E Y_E) - (\sum X_E)(\sum Y_E)}{\sqrt{\left[N_{Exp} \sum X_E{}^2 - (\sum X_E)^2\right]\left[N_{Exp} \sum Y_E{}^2 - (\sum Y_E)^2\right]}}$$

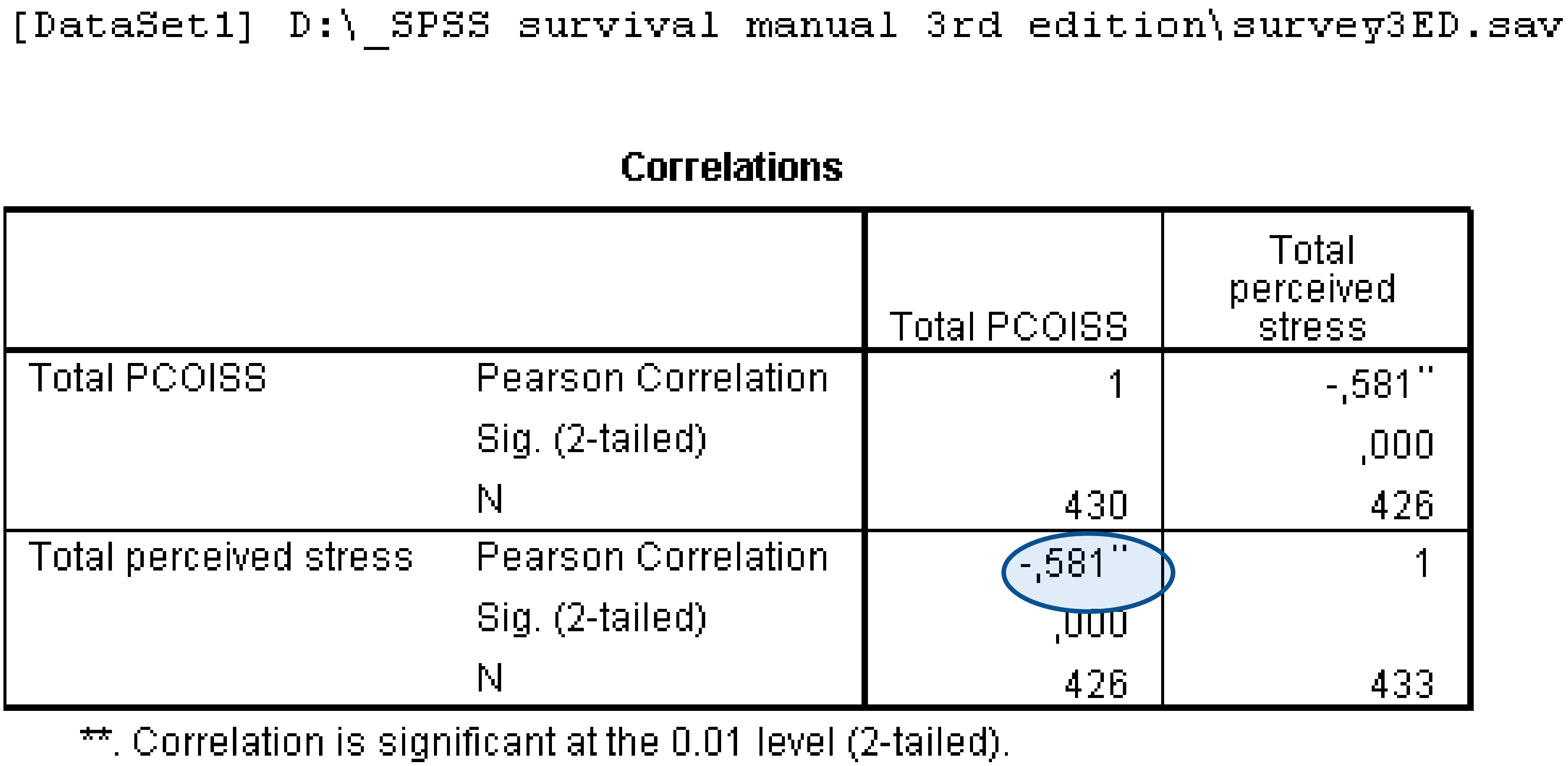


Critical values of the Pearson product-moment correlation coefficient

	Level of significance for a directional (one-tailed) test				
	.05	.025	.01	.005	.0005
	Level of significance for a non-directional (two-tailed) test				
df = *N* − 2	.10	.05	.02	.01	.001
1	.9877	.9969	.9995	.9999	1.0000
2	.9000	.9500	.9800	.9900	.9990
3	.8054	.8783	.9343	.9587	.9912
4	.7293	.8114	.8822	.9172	.9741
5	.6694	.7545	.8329	.8745	.9507
6	.6215	.7067	.7887	.8343	.9249
7	.5822	.6664	.7498	.7977	.8982
8	.5494	.6319	.7155	.7646	.8721
9	.5214	.6021	.6851	.7348	.8471
10	.4973	.5760	.6581	.7079	.8233
11	.4762	.5529	.6339	.6835	.8010
12	.4575	.5324	.6120	.6614	.7800
13	.4409	.5139	.5923	.6411	.7603
14	.4259	.4973	.5742	.6226	.7420
15	.4124	.4821	.5577	.6055	.7246
16	.4000	.4683	.5425	.5897	.7084
17	.3887	.4555	.5285	.5751	.6932
18	.3783	.4438	.5155	.5614	.6787
19	.3687	.4329	.5034	.5487	.6652
20	.3598	.4227	.4921	.5368	.6524
25	.3233	.3809	.4451	.4869	.5974
30	.2960	.3494	.4093	.4487	.5541
35	.2746	.3246	.3810	.4182	.5189
40	.2573	.3044	.3578	.3932	.4896
45	.2428	.2875	.3384	.3721	.4648
50	.2306	.2732	.3218	.3541	.4433
60	.2108	.2500	.2948	.3248	.4078
70	.1954	.2319	.2737	.3017	.3799
80	.1829	.2172	.2565	.2830	.3568
90	.1726	.2050	.2422	.2673	.3375
100	.1638	.1946	.2301	.2540	.3211

Practice

- Open survey3ED.sav
- Research question: Is there a relationship between the amount of control people have over their internal states and their levels of perceived stress?
- Variables: Total perceived stress & Total PCOISS (Perceived Control of Internal States Scale)



```
[DataSet1] D:\_SPSS survival manual 3rd edition\survey3ED.sav
```

Correlations

		Total PCOISS	Total perceived stress
Total PCOISS	Pearson Correlation	1	-,581**
	Sig. (2-tailed)		,000
	N	430	426
Total perceived stress	Pearson Correlation	-,581**	1
	Sig. (2-tailed)	,000	
	N	426	433

**. Correlation is significant at the 0.01 level (2-tailed).

Is there a relationship?

- First you must determine something called degrees of freedom (df). For a correlation study, the degrees of freedom is equal to 2 less than the number of subjects you had. If you collected data from 27 pairs, the degrees of freedom would be 25. Use the critical value table to find the intersection of alpha .05 (see the columns) and 25 degrees of freedom (see rows). The value found at the intersection (.381) is the minimum correlation coefficient r that you would need to confidently state 95 times out of a hundred that the relationship you found with your 27 subjects exists in the population from which they were drawn.

Report

- The relationship between perceived control of internal states (as measured by PCOISS) and perceived stress (as measured by the Perceived Stress scale) was investigated using Pearson product-moment correlation coefficient. There was a strong, negative correlation between two variables, r=-.58, n=426, p<.0005, with high levels of perceived control associated with lower levels of perceived stress.

Further practice

- Check the strength of correlation between Total perceived stress and Total life satisfaction (survey3ED.sav)
- Check the strength of correlation between Total self-esteem and Total life satisfaction (survey3ED.sav)
- Check the strength of correlation between Total social desirability and Total life satisfaction (survey3ED.sav)

Statistical techniques to compare groups

Chi square

T-tests

ANOVA

Chi-square test for goodness-of-fit

Chi-square test for goodness-of-fit

Determines if the observed frequencies are different from what we would expect to find.

- **Assumptions**
  -None of the expected values may be less than 1
  -No more than 20% of the expected values may be less than 5

Calculating the Chi-Square

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

f_e = expected frequencies

f_o = observed frequencies

Language	Number of Students
Chinese	23
Spanish	20
French	15
German	13
Japanese	29
Total	100

We expect equal choice for each foreign language

Q: Is the difference observed in the data collected significant?

Determining the Degrees of Freedom

$$df = (r - 1)(c - 1)$$

where

r = the number of rows

c = the number of columns

Chi-Square Table

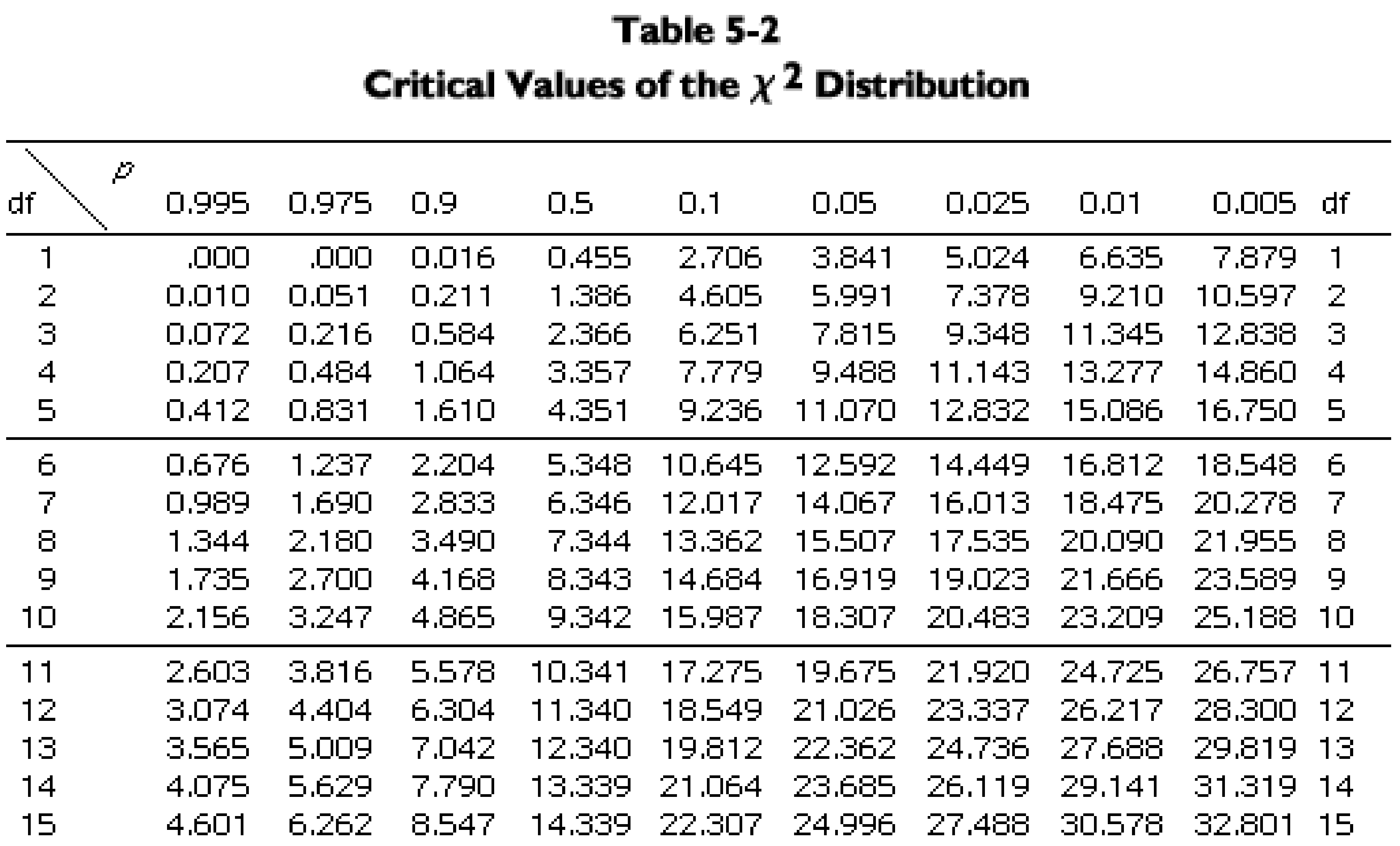
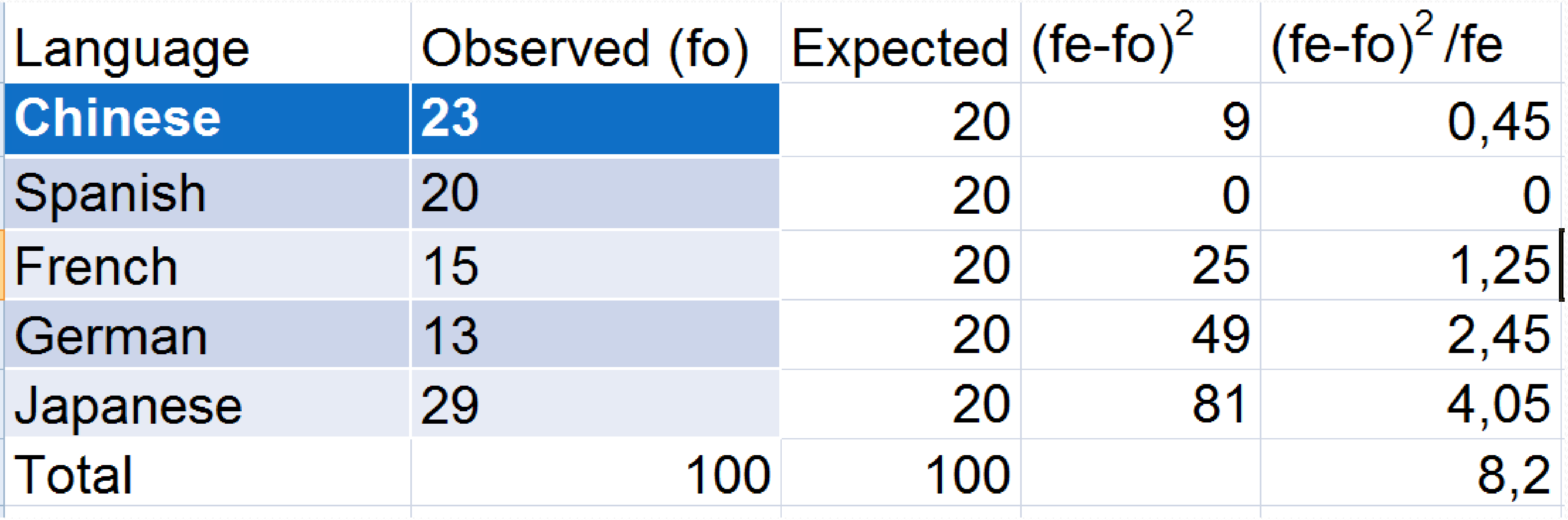


Table 5-2

Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

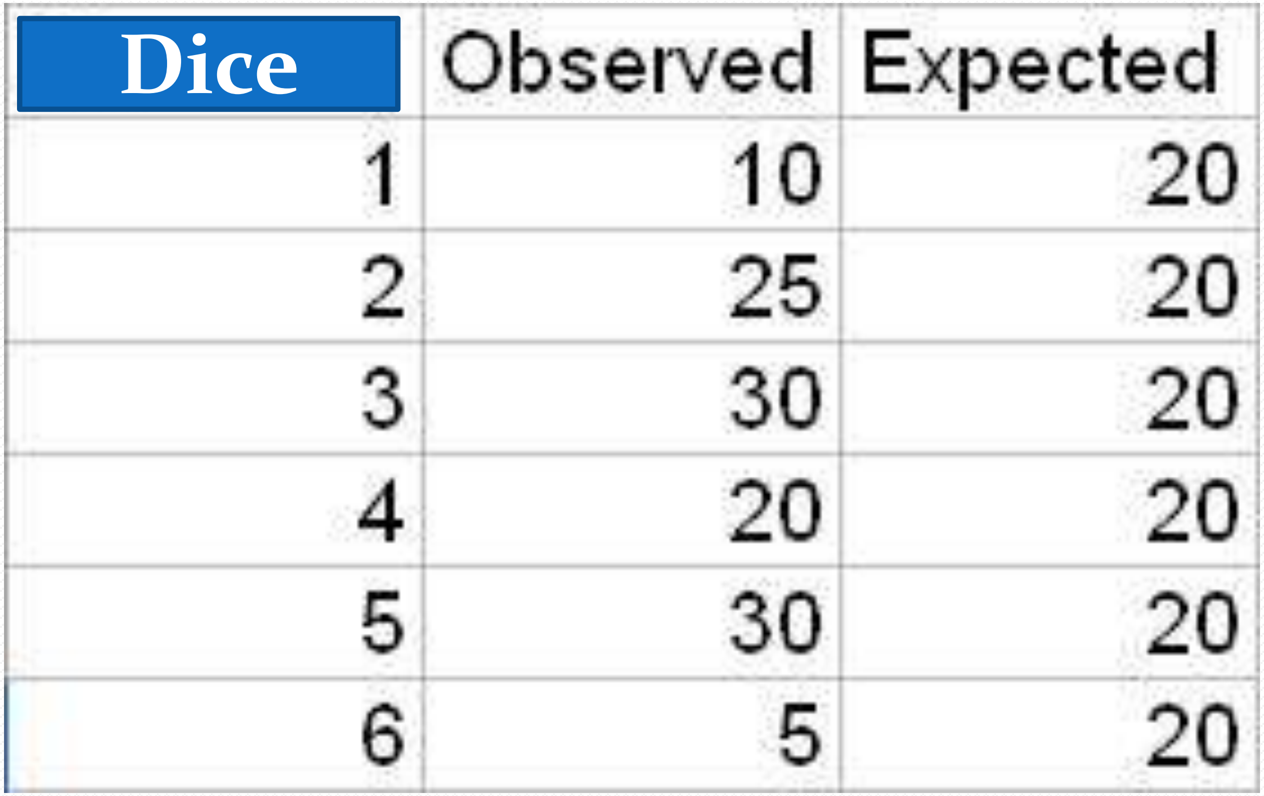
Calculating the Chi-Square



Language	Observed (fo)	Expected	$(fe\text{-}fo)^2$	$(fe\text{-}fo)^2$ /fe
Chinese	23	20	9	0,45
Spanish	20	20	0	0
French	15	20	25	1,25
German	13	20	49	2,45
Japanese	29	20	81	4,05
Total	100	100		8,2

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e} = 8,2 < 9,49$$

No significant difference between observed and expected values

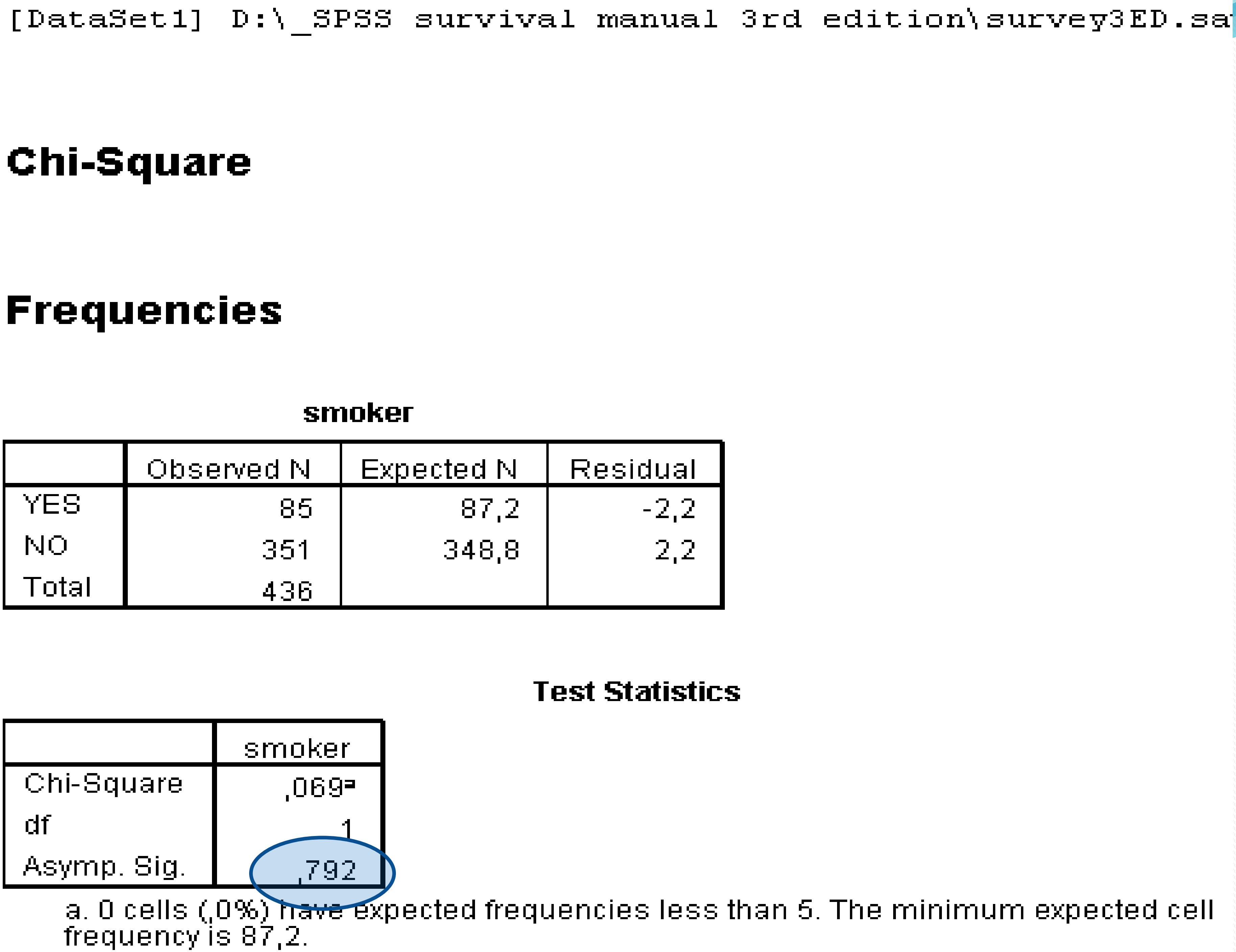


Dice	Observed	Expected
1	10	20
2	25	20
3	30	20
4	20	20
5	30	20
6	5	20

Task: Decide if the dice turning is fair.
P0,05 = 11,07 (df=5)

SPSS practice

- Open survey3ED.sav
- Research question: Is the number of smokers in the data file the same to that reported in literature from a previous nationwide study (20%)
- Variables: smoker (Y/N). Hypothesis: 20% smokers; 80% non smokers or .2/.8
- Analyze -> parametric tests -> Chi-square



```
[DataSet1] D:\_SPSS survival manual 3rd edition\survey3ED.sa
```

Chi-Square

Frequencies

smoker

	Observed N	Expected N	Residual
YES	85	87,2	-2,2
NO	351	348,8	2,2
Total	436		

Test Statistics

	smoker
Chi-Square	,069[a]
df	1
Asymp. Sig.	,792

a. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 87,2.

Report

- A chi-square goodness-of-fit test indicates there was no significant difference in the proportion of smokers identified in the current sample (19.5%) as compared with the value of 20% obtained in a previous nationwide study, Chi-square (1, n=436) = .07, p=.79

Chi-square for testing group independence

Chi-square for testing group independence

- The Chi Square Test of Independence tests the association between 2 nominal variables.
- **Assumptions:**
  -None of the expected values may be less than 1
  -No more than 20% of the expected values may be less than 5
- **Hypotheses:**
  Null: There is no association between the two variables.
  Alternate: There is an association between the two variables.

Calculating the Chi-Square

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$$

f_e = expected frequencies

f_o = observed frequencies

Calculating the Chi-Square

	Passed	Failed	Total
Experimental	73	12	85
Control	43	39	82
Total	116	51	167

Practice

	Like	Not Like	Total
Experimental	22	34	
Control	15	41	
Total			

Does the use of Facebook change students' interest in Writing classes?

Determining the Degrees of Freedom

$$\mathbf{df = (r - 1)(c - 1)}$$

where

r = the number of rows

c = the number of columns

Chi-Square Table

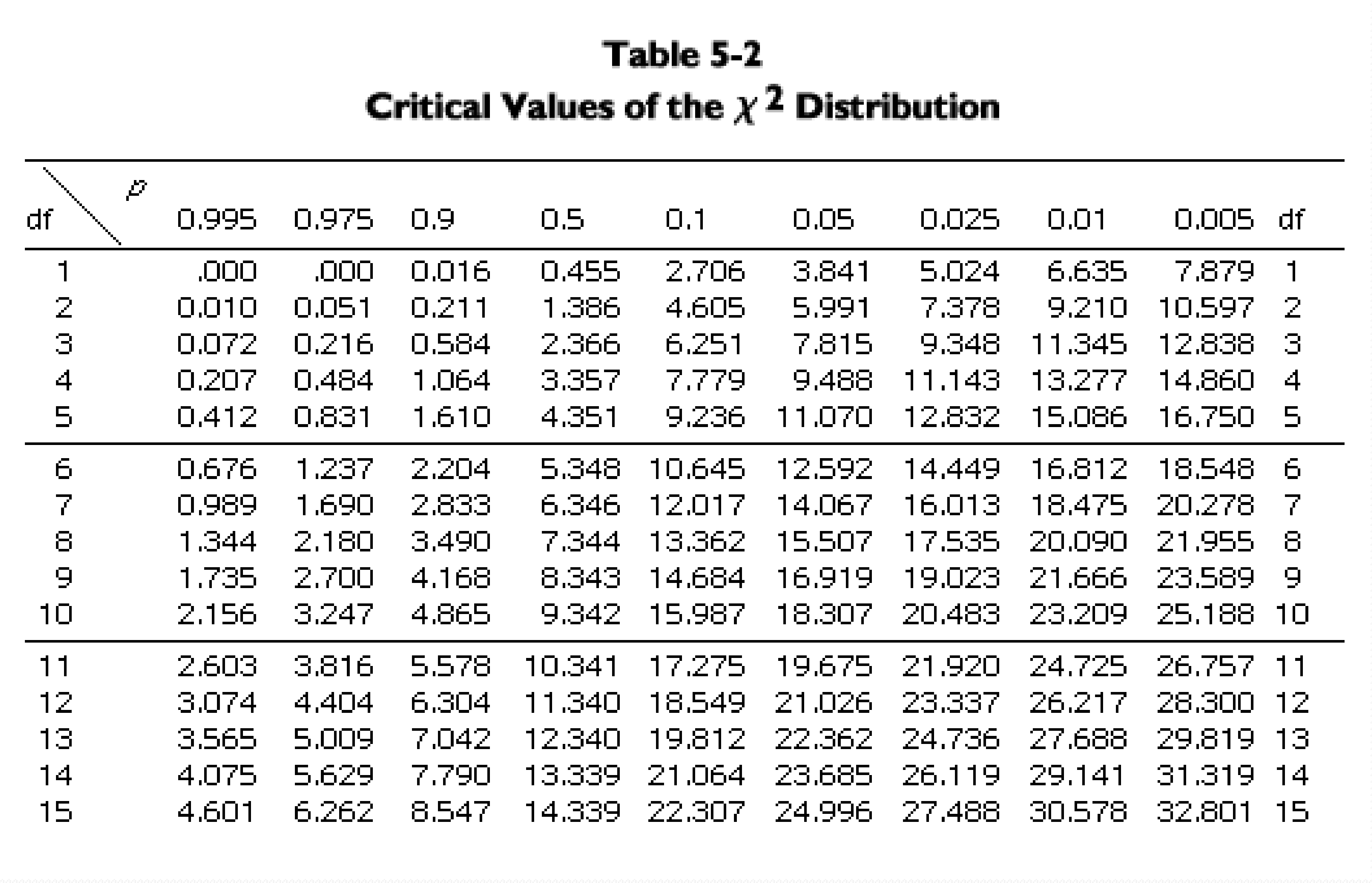


Table 5-2

Critical Values of the χ^2 Distribution

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	30.578	32.801	15

Calculating the Chi-Square

	Passed	Failed	Total
Experimental	73 (59.042)	12 (25.958)	85
Control	43 (56.958)	39 (25.042)	82
Total	116	51	167

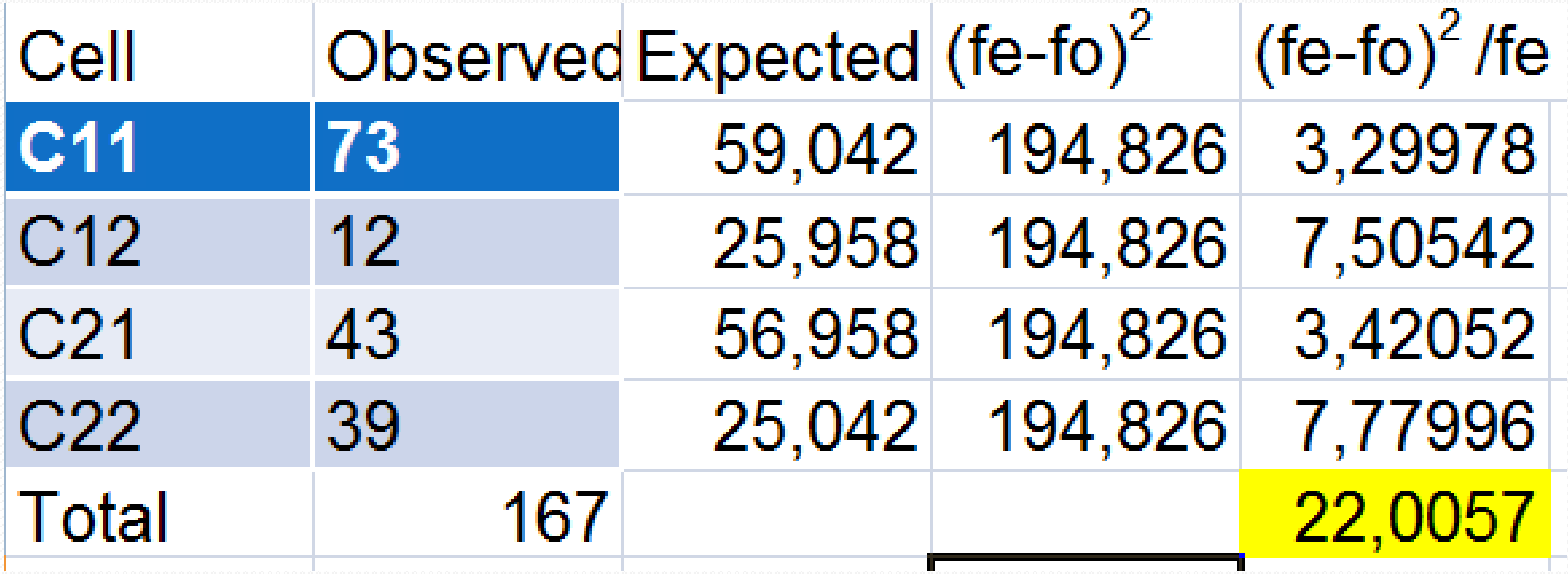
Expected value at cell ij

$$E_{ij} = \frac{T_i \text{ x } T_j}{N}$$

$$E_{11} = \frac{85 \text{ x } 116}{167} \qquad E_{12} = \frac{85 \text{ x } 51}{167}$$

$$E_{21} = \frac{82 \text{ x } 116}{167} \qquad E_{22} = \frac{82 \text{ x } 51}{167}$$

Calculating the Chi-Square



Cell	Observed	Expected	$(fe-fo)^2$	$(fe-fo)^2$ /fe
C11	73	59,042	194,826	3,29978
C12	12	25,958	194,826	7,50542
C21	43	56,958	194,826	3,42052
C22	39	25,042	194,826	7,77996
Total	167			22,0057

$$\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e} = 22{,}01 > 3{,}84$$

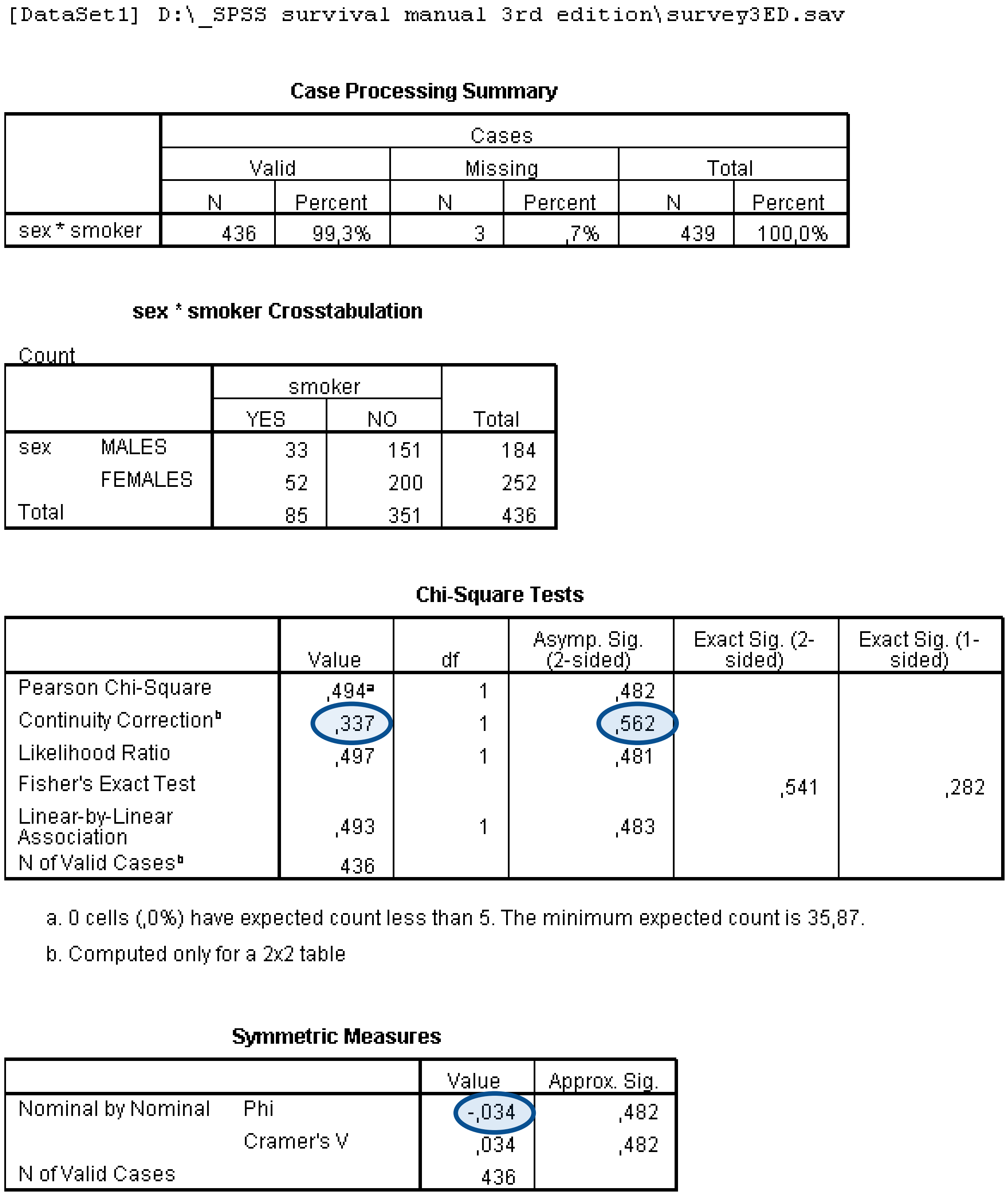
Significant difference between observed and expected values

Limitations of the Chi-Square Test

- The chi-square test does **not** give us much information about the ***strength*** of the relationship or its ***substantive significance*** in the population.
- The chi-square test is **sensitive** to ***sample size***. The size of the calculated chi-square is **directly proportional** to the size of the sample, independent of the strength of the relationship between the variables.
- The chi-square test is also **sensitive** to **small expected frequencies** in one or more of the cells in the table.

SPSS practice

- Open survey3ED.sav
- Research question: Are males more likely to smoke than females?
- Variables: sex (rows); smoker (col)
- Analyze -> Descriptive Stats -> Crosstabs
- Tick Chi-square & Phi and Crammer’s V



```
[DataSet1] D:\_SPSS survival manual 3rd edition\survey3ED.sav
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
sex * smoker	436	99,3%	3	,7%	439	100,0%

sex * smoker Crosstabulation

Count

		smoker		Total
		YES	NO	
sex	MALES	33	151	184
	FEMALES	52	200	252
Total		85	351	436

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	,494[a]	1	,482		
Continuity Correction[b]	,337	1	,562		
Likelihood Ratio	,497	1	,481		
Fisher's Exact Test				,541	,282
Linear-by-Linear Association	,493	1	,483		
N of Valid Cases[b]	436				

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 35,87.

b. Computed only for a 2x2 table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-,034	,482
	Cramer's V	,034	,482
N of Valid Cases		436	

Effect size

- 2x2 tables: phi coefficient
  - Cohen's (1998) criteria: .10: small; .30: medium; .50: large
- Larger tables: Crammer's V
  - R-1 or C-1 = 1 (2 categories): small=.10, medium=.30, large=.50
  - R-1 or C-1 = 2 (3 categories): small=.07, medium=.21, large=.35
  - R-1 or C-1 = 3 (4 categories): small=.06, medium=.17, large=.29

Report

- A chi-square test for independence (with Yates Continuity Correction) indicated no significant association between gender and smoking status, chi-square (1, n=436) = .34, p=.56, phi=-.03

Further practice

- staffsurvey3ED.sav: Use chi-square test for independence to compare the proportion of permanent versus casual staff (***employstatus***) who indicate they would recommend the organisation as a good place to work (***recommend***).
- sleep3ED.sav: Use a chi-square test for independence to compare the proportion of males and females (***sex***) who indicate they have a sleep problem (***problem)***

T-test

One sample T-test

- Compare the mean score of a sample to a known value. Usually, the known value is a population mean.
- **Assumption:**
  -The dependent variable is normally distributed. You can check for normal distribution with a Q-Q plot.

- The average sleep time is supposed to be 8 hours a day (m).
- We think college students sleep a different amount, maybe more - maybe less.
- We survey ten students to see how much they sleep.
- The data are as follows (each cell represents a student):

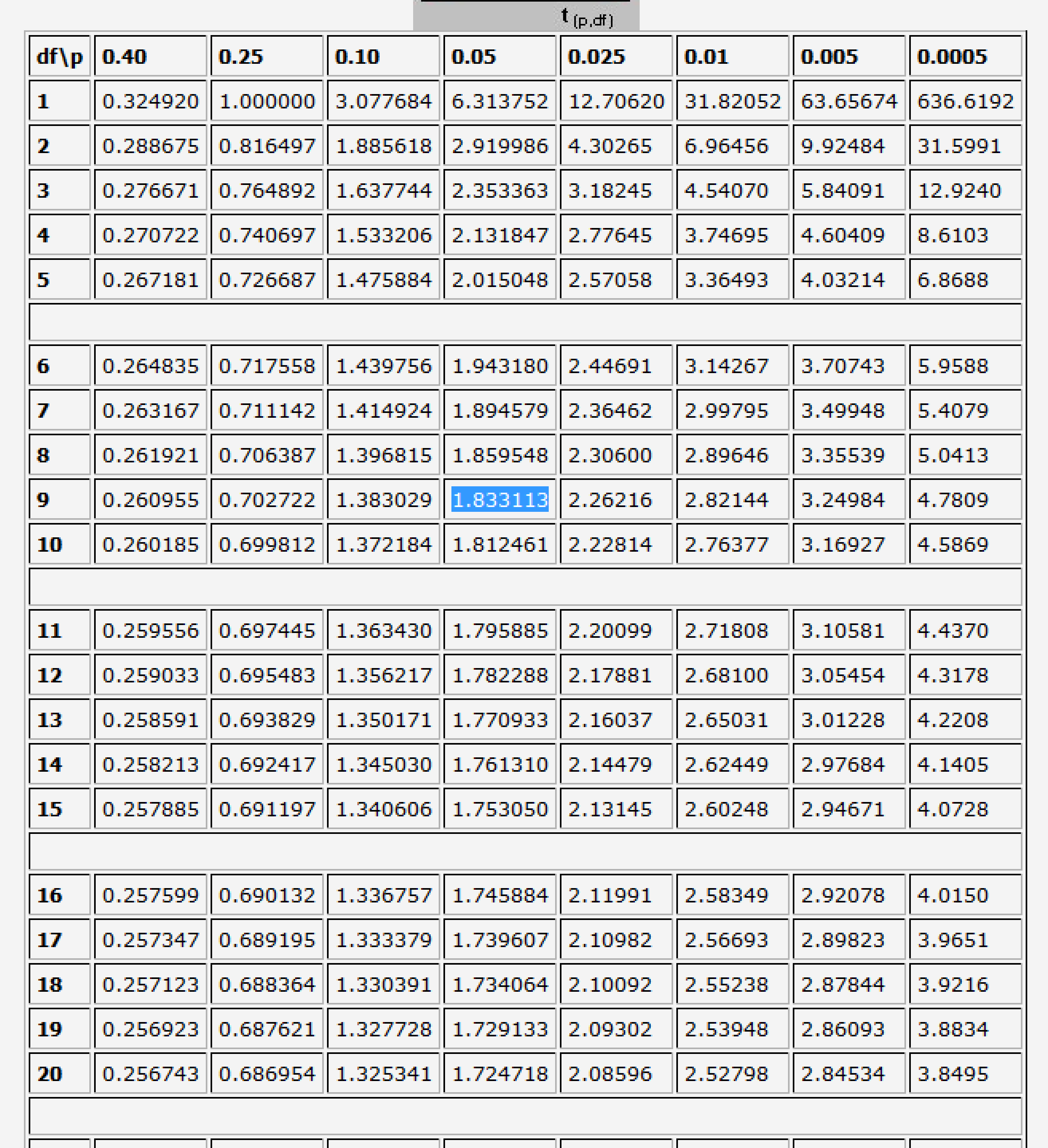
6	**5**	**4**	**3**	**7**
5	**5**	**5**	**6**	**6**

$$SD = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

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

Where,
SD = Standard deviation
$\bar{X}$ = Sample mean
n = number of observations in sample

Where,
t = one sample t-test value
μ = population mean



$t_{(p,df)}$

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495

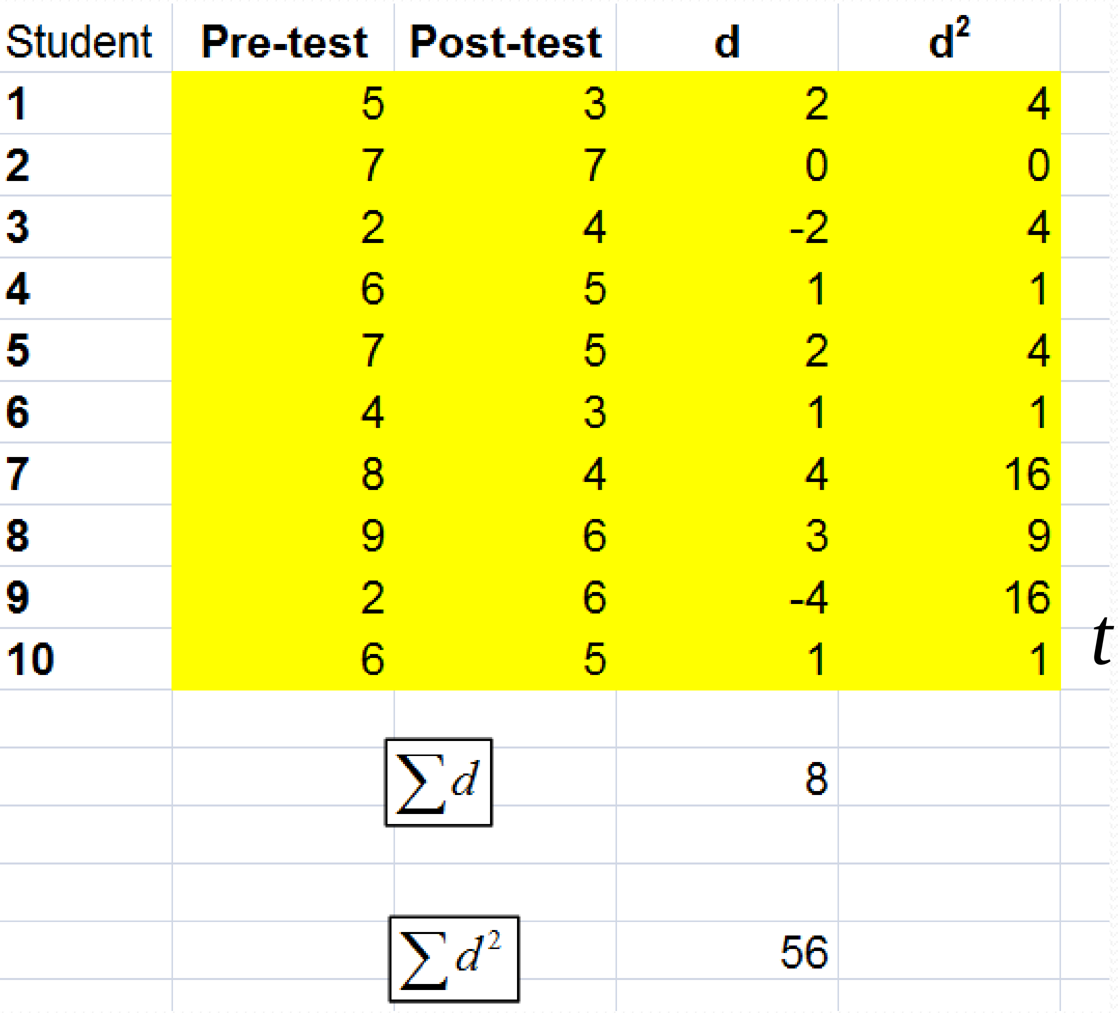
Paired sample T-test

Paired sample T-test

- Compare the means of two variables. It computes the difference between the two variables for each case, and tests to see if the average difference is significantly different from zero.
- **Assumption:**
  -Both variables should be normally distributed. You can check for normal distribution with a Q-Q plot.

Formula

$$t = \frac{\sum d}{\sqrt{\dfrac{n(\sum d^2) - (\sum d)^2}{n-1}}}$$



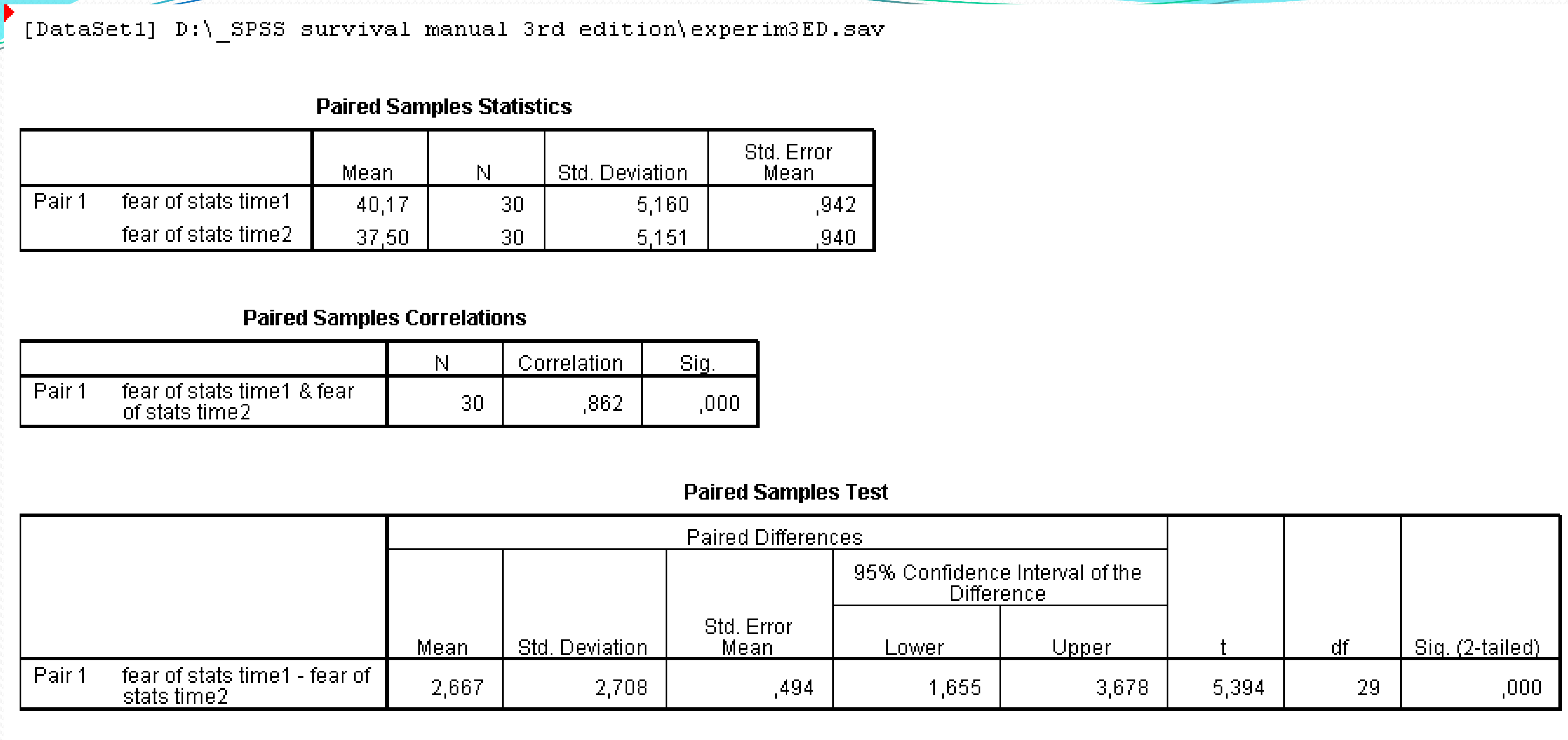
Student	Pre-test	Post-test	d	d^2
1	5	3	2	4
2	7	7	0	0
3	2	4	-2	4
4	6	5	1	1
5	7	5	2	4
6	4	3	1	1
7	8	4	4	16
8	9	6	3	9
9	2	6	-4	16
10	6	5	1	1
		$\boxed{\sum d}$	8	
		$\boxed{\sum d^2}$	56	

$$t = \frac{\sum d}{\sqrt{\dfrac{n(\sum d^2) - (\sum d)^2}{n-1}}}$$

$$t = \frac{8}{\sqrt{\dfrac{10(56) - (8)^2}{10-1}}} = 1{,}078$$

Practice

- Open experim3ED.sav
- Research question: Does the intervention have an impact on participants' fear of statistics score?
- Select:
  - Fost1: fear of stats time 1
  - Fost2: fear of stats time 2



```
[DataSet1] D:\_SPSS survival manual 3rd edition\experim3ED.sav
```

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	fear of stats time1	40,17	30	5,160	,942
	fear of stats time2	37,50	30	5,151	,940

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	fear of stats time1 & fear of stats time2	30	,862	,000

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	fear of stats time1 - fear of stats time2	2,667	2,708	,494	1,655	3,678	5,394	29	,000

Effect Size

(Cohen 1988, pp. 284-7)

Size	Eta squared (% of variance explained)
Small	.01
Medium	.06
Large	.14

$$Eta\ squared = \frac{t^2}{t^2 + (N-1)} = \frac{5.39^2}{5.39^2 + (30-1)} = 0.50$$

Report

- A paired-samples t-test was conducted to evaluate the impact of the intervention on students' scores on the Fear of Statistics Test (FOST). There was a statistically significant decrease in FOST scores from Time 1 (M=40.17, SD=5.16) to Time 2 (M=37.5, SD = 5.15), t(29) = 5.39, p<.0005 (two-tailed). The mean decrease in FOST scores was 2.27 with a 95% confidence interval ranging from 1.66 to 3.68. The eta squared statistic (.5) indicated a large effect size.

More practice

- Use the same experim3ED.sav file. Compare Fost1 & Fost3; Depression time 1 & time 3

Independent samples T-test

Independent samples T-test

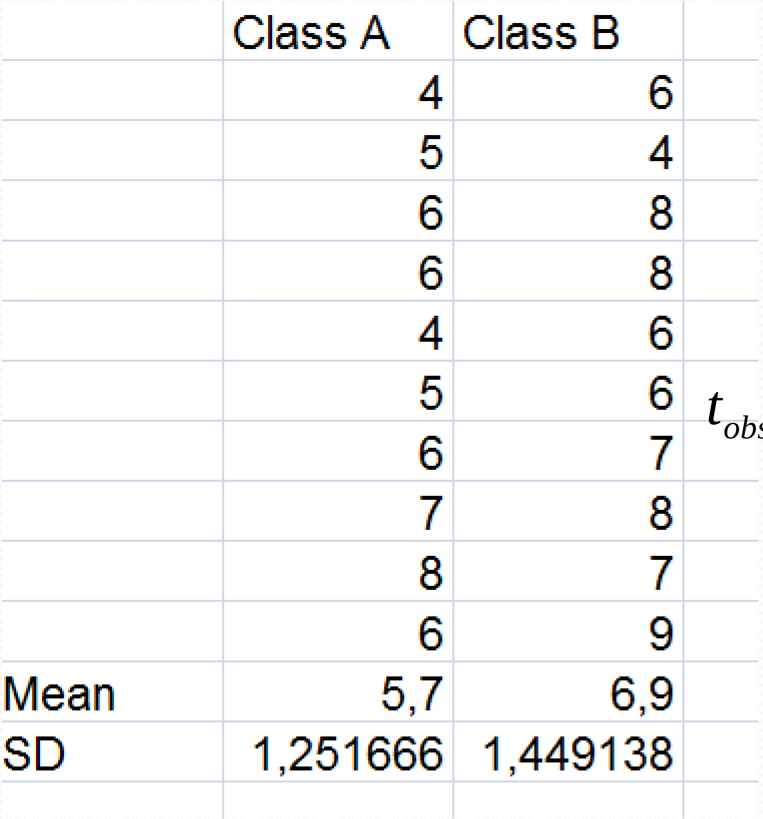
- Compare the mean scores of two groups on a given variable.
- **Assumptions:**
  -The dependent variable is normally distributed-> check for normal distribution with a Q-Q plot.
  -The two groups have approximately equal variance on the dependent variable -> check this by looking at the Levene's Test.
  -The two groups are independent of one another.

Formula

$$t_{obsExpCon} = \frac{M_{Exp} - M_{Con}}{\sqrt{(SD^2_{Exp} / N_{Exp}) + (SD^2_{Con} / N_{Con})}}$$

Degree of freedom

$$V = N_{Exp} + N_{Con} - 2$$



	Class A	Class B
	4	6
	5	4
	6	8
	6	8
	4	6
	5	6
	6	7
	7	8
	8	7
	6	9
Mean	5,7	6,9
SD	1,251666	1,449138

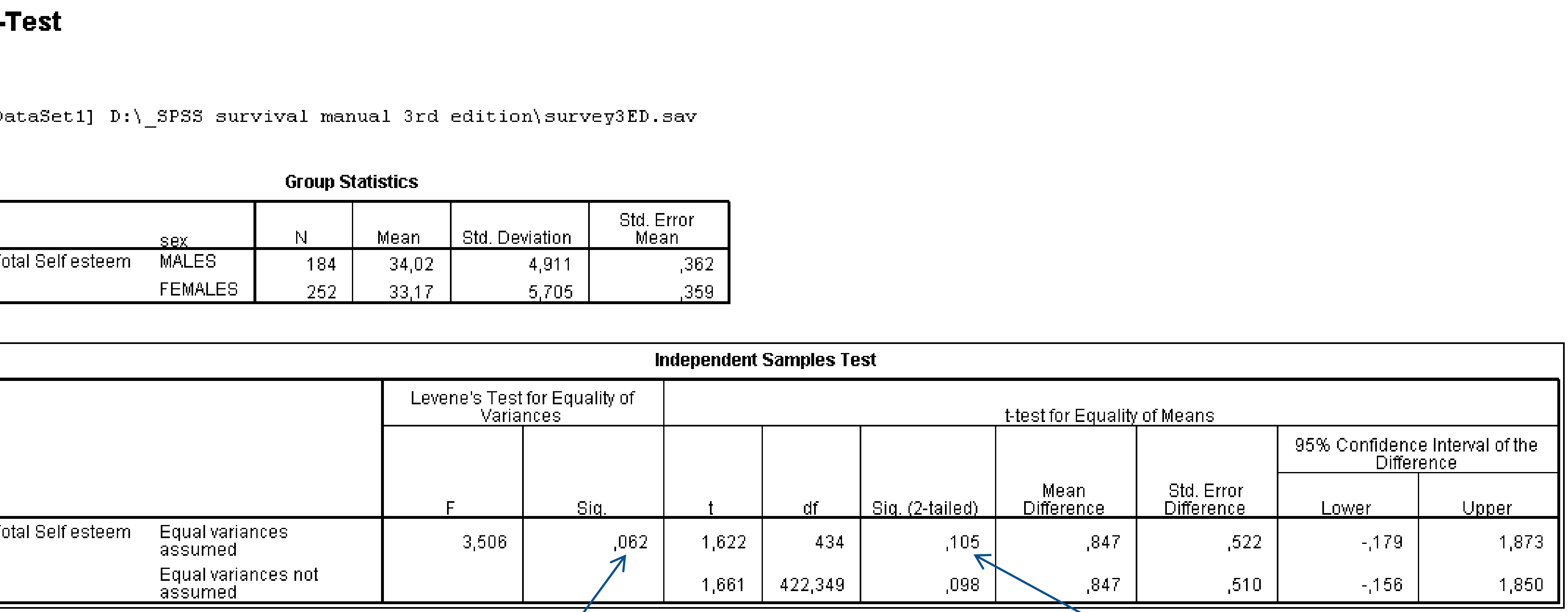
$$t_{obsExpCon} = \frac{M_{Exp} - M_{Con}}{\sqrt{(SD^2_{Exp} / N_{Exp}) + (SD^2_{Con} / N_{Con})}}$$

$$t_{obsExpCon} = \frac{6,9 - 5,7}{\sqrt{(1,449 * 1,449 / 10) + (1,251 * 1,251 / 10)}}$$

$$t_{obsExpCon} = 1,982293$$

Practice

- Open survey3ED.sav
- Research Question: Is there a significant difference in the mean self-esteem scores for males and females?
- To check Q-Q plot for each group: Data -> Select cases -> if …. =



-Test

```
ataSet1] D:\_SPSS survival manual 3rd edition\survey3ED.sav
```

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
otal Self esteem	MALES	184	34,02	4,911	,362
	FEMALES	252	33,17	5,705	,359

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
otal Self esteem	Equal variances assumed	3,506	,062	1,622	434	,105	,847	,522	-,179	1,873
	Equal variances not assumed			1,661	422,349	,098	,847	,510	-,156	1,850

If > .05 , variances of 2 groups are the same ->use the first line

If =< .05 , the difference is significant

Effect Size

Size	Eta squared (% of variance explained)	Cohen's d (standard deviation units)
Small	.01	.2
Medium	.06	.5
Large	.138	.8

$$Eta\ squared = \frac{t^2}{t^2 + (N1 + N2 - 2)} = \frac{1.62^2}{1.62^2 + (184 + 252 - 2)} = 0.006$$

Report

- An independent-samples t-test was conducted to compare the self-esteem scores for males and females. There was no significant difference in scores for males (M=34.02, SD=4.91) and females (M=33.17, SD = 5.71; t(434)=1.62, p=.11 (two-tailed).

- The magnitude of the differences in the means (mean difference = .85, 95 CI:-1.80 to 1.87) was very small (eta squared = .006/Cohen's d =

Practice

- staffsurvey3ED.sav: Compare the mean staff satisfaction scores (totsatis) for permanent and casual staff (employstatus). Is there a significant difference in mean satisfaction scores?
- sleep3ED.sav: compre the mean sleepiness ratings (Sleepiness and Associated Sensations Scale total score: totSAS) for males and females (sex). Is there a significant difference in mean sleepiness scores?

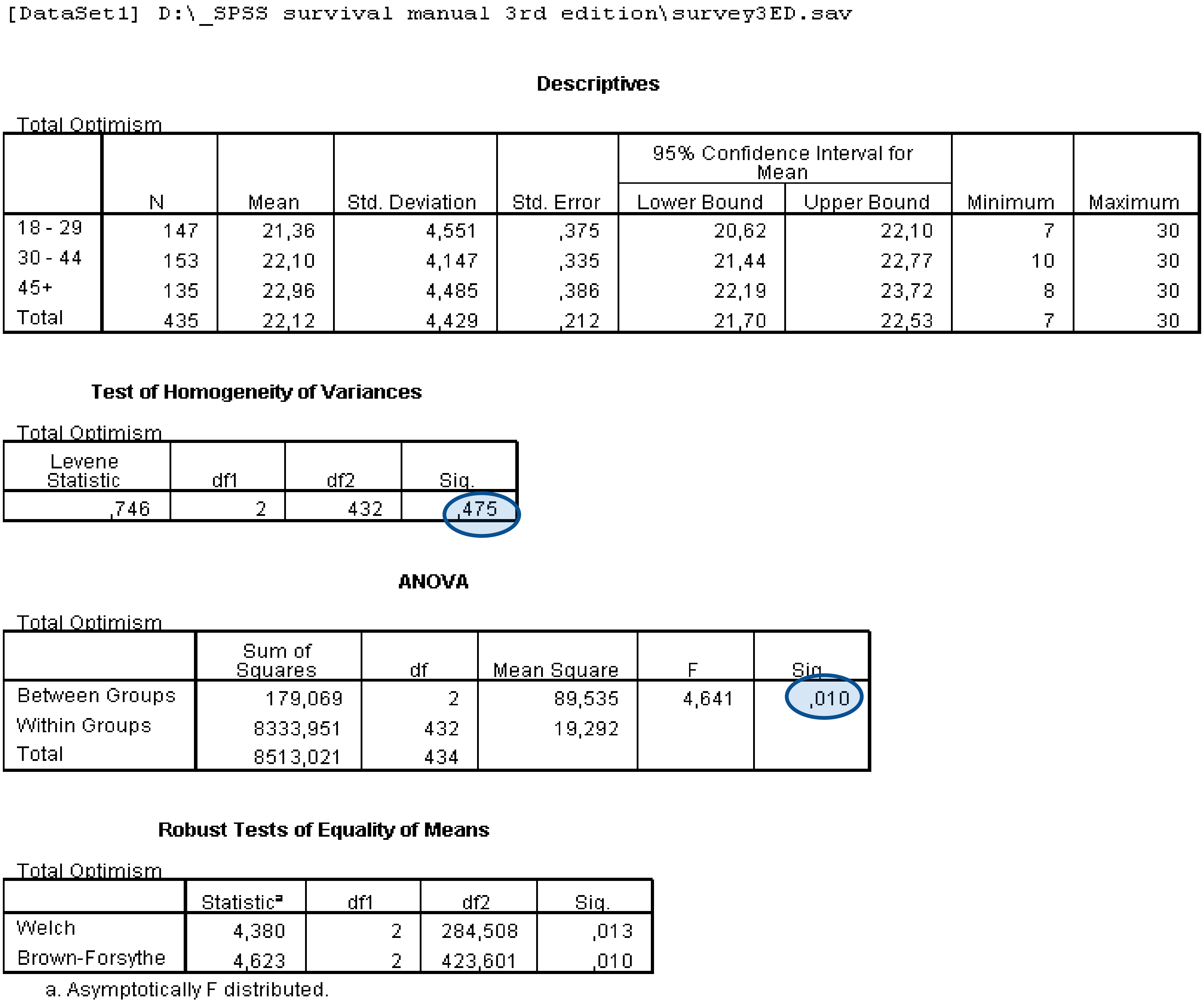
One-way ANOVA

One-way ANOVA

- Compare the mean of one or more groups based on one independent variable (or factor).
- **Assumptions:**
  -The dependent variable(s) is normally distributed-> check for normal distribution with a Q-Q plot.
  -The two groups have approximately equal variance on the dependent variable -> check this by looking at the Levene's Test.

Practice

- Open survey3ED.sav
- Research question: Is there a difference in optimism scores for young, middle-aged and old subjects?
- Dependent list: Total optimism variable
- Factor: Age 3 group (Agegp3)
- Options: Descriptive, Homegeneity of variance test, Brown-Forsythe, Welsh and Means Plot
- Click on Post Hoc



```
[DataSet1] D:\_SPSS survival manual 3rd edition\survey3ED.sav
```

Descriptives

Total Optimism

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
18 - 29	147	21,36	4,551	,375	20,62	22,10	7	30
30 - 44	153	22,10	4,147	,335	21,44	22,77	10	30
45+	135	22,96	4,485	,386	22,19	23,72	8	30
Total	435	22,12	4,429	,212	21,70	22,53	7	30

Test of Homogeneity of Variances

Total Optimism

Levene Statistic	df1	df2	Sig.
,746	2	432	,475

ANOVA

Total Optimism

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	179,069	2	89,535	4,641	,010
Within Groups	8333,951	432	19,292		
Total	8513,021	434			

Robust Tests of Equality of Means

Total Optimism

	Statistic[a]	df1	df2	Sig.
Welch	4,380	2	284,508	,013
Brown-Forsythe	4,623	2	423,601	,010

a. Asymptotically F distributed.

Calculating effect size

$$Eta\ squared = \frac{Sum\ of\ squares\ between - groups}{Total\ sum\ of\ squares} = \frac{197.07}{8513.02} = .02$$

Report

- A one-way between group analysis of variance was conducted to explore the impact of age on levels of optimism. Subjects were divided into three groups according to their age (Group 1: 29yrs or less; Group 2:30-44 yrs; Group 3:45 yrs and above). There was a statistically significant difference at the p<.05 level in optimism scores for the three age groups: F(2,432) = 4.6, p = .01.
- Despite reaching statistical significance, the actual difference in mean scores between the groups was quite small. The effect size, calculated using eta squared, was .02. Post-hoc comparisons using the Turkey HSD test indicated that the mean score for Group 1 (M=21.36, SD=4.55) was significantly different from Group 3 (M=22.96, SD = 4.49). Group 2 (M=22.10, SD=4.15) did not differ significantly from either Group 1 or 3.

Practice

- staffsurvey3ED.sav: Conduct one-way ANOVA with post-hoc tests (if approriate) to compare staff satisfaction scores (***totsatis***) across each of the length of service categories (use the ***servicegp3*** variable)
- sleep3ED.sav: Conduct one-way ANOVA with post-hoc tests (if approriate) to compare the mean sleepiness ratings (Sleepiness and Associated Sensations Scale total score: ***totSAS)*** for the three age groups defined by the variable ***agegp3*** (<=37, 38-50, 51+)