Parametric Significance Tests

Parametric Signifiance Tests

Experiment design	Two Conditions	Three(*) or more conditions
Between-group	*t* Test *(independent samples)*	One-way ANOVA
Within-group	*t* Test *(paired samples)*	Repeated Measures ANOVA

- Require normal distribution of the scores
- Variances should be nearly equal

Why?

(*) Can actually be applied also on two conditions

t Test

- Between-group design (independent samples)
  - Means are contributed from different groups
- Within-group design (paired samples)
  - Means are contributed from the same group

- Can be two-tailed or one-tailed

- Simplified version of ANalysis Of VAriance (ANOVA) with only two groups or conditions

t Test: Example of Two-tailed Test

- H_0: There is no significant difference in the task completion time between individuals who use the word-prediction software (WP group) and those who do not use the software (NP group).

Commonly accepted!

- Select reasonable significance level: $\alpha = 5\%$
- If t-test shows significance at $p < .05$, we can conclude that under $\alpha = 5\%$, i.e., in 95% of the time, the test result correctly applies to the entire population.

t Test : Example of Two-tailed Test

- Sample data from between-group experiment

Group	Participants	Task completion time	Coding
No prediction	Participant 1	245	0
No prediction	Participant 2	236	0
No prediction	Participant 3	321	0
No prediction	Participant 4	212	0
No prediction	Participant 5	267	0
No prediction	Participant 6	334	0
No prediction	Participant 7	287	0
No prediction	Participant 8	259	0
With prediction	Participant 1	246	1
With prediction	Participant 2	213	1
With prediction	Participant 3	265	1
With prediction	Participant 4	189	1
With prediction	Participant 5	201	1
With prediction	Participant 6	197	1
With prediction	Participant 7	289	1
With prediction	Participant 8	224	1

Testing for Normal Distribution

- Kolmogorov-Smirnov Test
- Shapiro-Wilk Test ($n \leq 50$)
  - D'Agostino for $n > 50$
  - Similar test Shapiro-Francia for $n > 50$

Shapiro-Wilk Test (Conover, pp. 450)

- Assumption: sample is a random sample
- Compute test denominator D for ‘No Prediction’

$$D = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{8} (x_i - 270.125)^2 = 12260.875$$

- Order ‘No Prediction’ samples

  X = {245, 236, 321, 212, 267, 334, 287, 259}

  from smallest to largest

  212 < 236 < 245 < 259 < 267 < 287 < 321< 334

- $x^{(i)}$ denotes the *i*th order statistic, e.g., $x^{(4)} = 259$

Shapiro-Wilk Test (Conover, pp. 450)

- Obtain coefficients $a_1, \ldots, a_k$ with $k = n/2 = 4$

  $a_1 = 0.6052$, $a_2 = 0.3164$, $a_3 = 0.1743$, $a_4 = 0.0561$

  (represent what the order statistics should look like if the population is normal)

- Compute test statistics

$$T = \frac{1}{D}\left[\sum_{i=1}^{k} a_i\left(x^{(n-i+1)} - x^{(i)}\right)\right]^2$$

$$= \frac{1}{D}\left[a_1\left(x^{(8)} - x^{(1)}\right) + a_2\left(x^{(7)} - x^{(2)}\right) + a_3\left(x^{(6)} - x^{(3)}\right) + a_4\left(x^{(5)} - x^{(4)}\right)\right]^2$$

$$= \ldots$$

Shapiro-Wilk Tables

TABLE A16 Coefficients for the Shapiro-Wilk Test[a]

i \ n	2	3	4	5	6	7	8	9	10
1	0.7071	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.0588	0.5739
2	—	0.0000	0.1667	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291
3	—	—	—	0.0000	0.0875	0.1401	0.1743	0.1976	0.2141
4	—	—	—	—	—	0.0000	0.0561	0.0947	0.1224
5	—	—	—	—	—	—	—	0.0000	0.0399

Shapiro-Wilk Test (Conover, pp. 450)

$$T = \frac{1}{D}\left[a_1\left(x^{(8)} - x^{(1)}\right) + a_2\left(x^{(7)} - x^{(2)}\right) + a_3\left(x^{(6)} - x^{(3)}\right) + a_4\left(x^{(5)} - x^{(4)}\right)\right]^2$$

$$= \frac{1}{D}\begin{bmatrix} a_1(334-212) + a_2(321-236) + \\ a_3(287-245) + a_4(267-259) \end{bmatrix}^2$$

$$= \frac{1}{D}\left[122\,a_1 + 85\,a_2 + 42\,a_3 + 8a_4\right]^2$$

$$= \frac{108.4978^2}{12260.875} \approx 0.96$$

- Sample: 212 < 236 < 245 < 259 < 267 < 287 < 321< 334
- Coefficients: $a_1 = 0.6052$, $a_2 = 0.3164$, $a_3 = 0.1743$, $a_4 = 0.0561$

Shapiro-Wilk Test (Conover, pp. 450)

- Look up quantile of Shapiro-Wilk test for $n = 8$ and at $\alpha = .05$, which is 0.818
- As $T \approx 0.96 > 0.818$, we accept H_0 saying that the sample is normally distributed

  (T close to 1.0, the sample behaves like normal sample, otherwise sample looks nonnormal)

- Repeat for 'With Prediction' sample: $T \approx 0.92$

Shapiro-Wilk Tables

TABLE A17 Quantiles of the Shapiro-Wilk Test Statistic[a]

n	0.01	0.02	0.05	0.10	0.50	0.90	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986

Testing for Equality of Variances

- Levene's test:

$$F_{Levene} = \frac{(N-t)\cdot\sum_{i=1}^{t} n_i(\overline{D}_i - \overline{D})^2}{(t-1)\cdot\sum_{i=1}^{t}\sum_{j=1}^{n_i}(D_{ij} - \overline{D}_i)^2}$$

t : number of treatments

n_i : number of observations from treatment i

$N = n_1 + n_2 + \ldots + n_t$ (overall size of combined samples)

y_{ij} : observation j from treatment i ($\mathrm{j} = 1,\ldots,\mathrm{n}_{\mathrm{i}}$ and $\mathrm{i} = 1,\ldots \mathrm{t}$)

$\bar{y}_i$: mean of sample data from treatment i

$D_{ij} = \left|\mathrm{y}_{\mathrm{ij}} - \bar{y}_i\right|$ (absolute deviation of observation j from treatment i mean)

$\overline{D}_i$: average of the n_{i} absolute deviations from treatment i

$\overline{D}$: average of all N absolute deviations

Testing for Equality of Variances

- Computing Levene's test score:

$$t = 2; \quad \mathrm{n}_1 = \mathrm{n}_2 = 8; \; \mathrm{N} = 16$$

$$F_{Levene} = \frac{(N-t)\cdot\sum_{i=1}^{t} n_i(\overline{D}_i - \overline{D})^2}{(t-1)\cdot\sum_{i=1}^{t}\sum_{j=1}^{n_i}(D_{ij} - \overline{D}_i)^2}$$

$$= \frac{14\cdot 8\cdot((\overline{D}_1 - \overline{D})^2 + (\overline{D}_2 - \overline{D})^2)}{1\cdot(\sum_{j=1}^{n_1}(D_{1j} - \overline{D}_1)^2 + \sum_{j=1}^{n_2}(D_{2j} - \overline{D}_2)^2)}$$

$$\overline{D}_1 = 32.90625$$

$$\overline{D}_2 = 29$$

$$\overline{D} = 30.953125$$

Testing for Equality of Variances

- Computing Levene's test score:

$$\ldots = \frac{854.4921875}{\sum_{j=1}^{n_1}(D_{1j} - \overline{D}_1)^2 + \sum_{j=1}^{n_2}(D_{2j} - \overline{D}_2)^2}$$

$$= \frac{854.4921875}{3598.3046875 + 2138} \approx 0.149$$

- Compare with critical value of F-distribution[1] for $\alpha = .05$: $F_{Levene} \approx 0.149 < 4.600$
- We retain H_0 and conclude that the variances are equal

[1] http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm

Critical Values of F-Distribution

5% significance level

$F_{.05}(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6
1	161.448	199.500	215.707	224.583	230.162	233.986
2	18.513	19.000	19.164	19.247	19.296	19.330
3	10.128	9.552	9.277	9.117	9.013	8.941
4	7.709	6.944	6.591	6.388	6.256	6.163
5	6.608	5.786	5.409	5.192	5.050	4.950
6	5.987	5.143	4.757	4.534	4.387	4.284
7	5.591	4.737	4.347	4.120	3.972	3.866
8	5.318	4.459	4.066	3.838	3.687	3.581
9	5.117	4.256	3.863	3.633	3.482	3.374
10	4.965	4.103	3.708	3.478	3.326	3.217
11	4.844	3.982	3.587	3.357	3.204	3.095
12	4.747	3.885	3.490	3.259	3.106	2.996
13	4.667	3.806	3.411	3.179	3.025	2.915
14	4.600	3.739	3.344	3.112	2.958	2.848
15	4.543	3.682	3.287	3.056	2.901	2.790
16	4.494	3.634	3.239	3.007	2.852	2.741
17	4.451	3.592	3.197	2.965	2.810	2.699

t Test: Example of Two-tailed Test

- Central question: does the IV affect the DV?
- Recall
  - IV: Typing support (No prediction vs. With prediction)
  - DV: Task completion time

Group	Participants	Task completion time	Coding
No prediction	Participant 1	245	0
No prediction	Participant 2	236	0
No prediction	Participant 3	321	0
No prediction	Participant 4	212	0
No prediction	Participant 5	267	0
No prediction	Participant 6	334	0
No prediction	Participant 7	287	0
No prediction	Participant 8	259	0
With prediction	Participant 1	246	1
With prediction	Participant 2	213	1
With prediction	Participant 3	265	1
With prediction	Participant 4	189	1
With prediction	Participant 5	201	1
With prediction	Participant 6	197	1
With prediction	Participant 7	289	1
With prediction	Participant 8	224	1

- No prediction:
  - M=270.125
- With prediction:
  - M=228

t Test: Example of Two-tailed Test

- If the IV did not affect the DV, then the *population* means for the two groups are equal
- However, because of sampling error, the two means are not exactly equal even if the H_0 is true **!**
- For the example: not known whether the difference of 270.125 and 228 can be explained by sampling error
- Thus, question is: how large is the sampling error in $\bar{X}_1 - \bar{X}_2$ given that the population means are equal? **!**
- Aim: precisely estimate how unusual the difference between means is relative to its standard error:

$$\frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

?

t Test: Example of Two-tailed Test

- Standard error of single distribution: $s \cdot \sqrt{\frac{1}{n}}$ with s is SD
- Standard error of the difference between two means randomly and independently sampled from the same population is: $\sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

  with σ is population SD and group sizes n_1 and n_2

- How to estimate the SD σ?
- Pooled variance ${s_p}^2$: compute variance for each group and combine it by weighting based on their df

$$ {s_p}^2 = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2} $$

t Test: Example of Two-tailed Test

- Answer: precisely estimate how unusual the difference between means is relative to its standard error:

$$(*) \quad \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1){s_1}^2 + (n_2 - 1){s_2}^2}{n_1 + n_2 - 2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

- It happens that $(*)$ has a t distribution with $n_1 + n_2 - 2$ degrees of freedom, under the assumptions that
  - Residual term has normal distribution
  - Variances of ${s_1}^2$ and ${s_1}^2$ are equal
- Basically, this is the independent samples t-test

!

t Test: Example of Two-tailed Test

- Determine degree of freedom:

$$
\begin{aligned}
d_f &= df_1 + df_2 \\
&= (n_1 - 1) + (n_2 - 1) \\
&= (8-1) + (8-1) = 14
\end{aligned}
$$

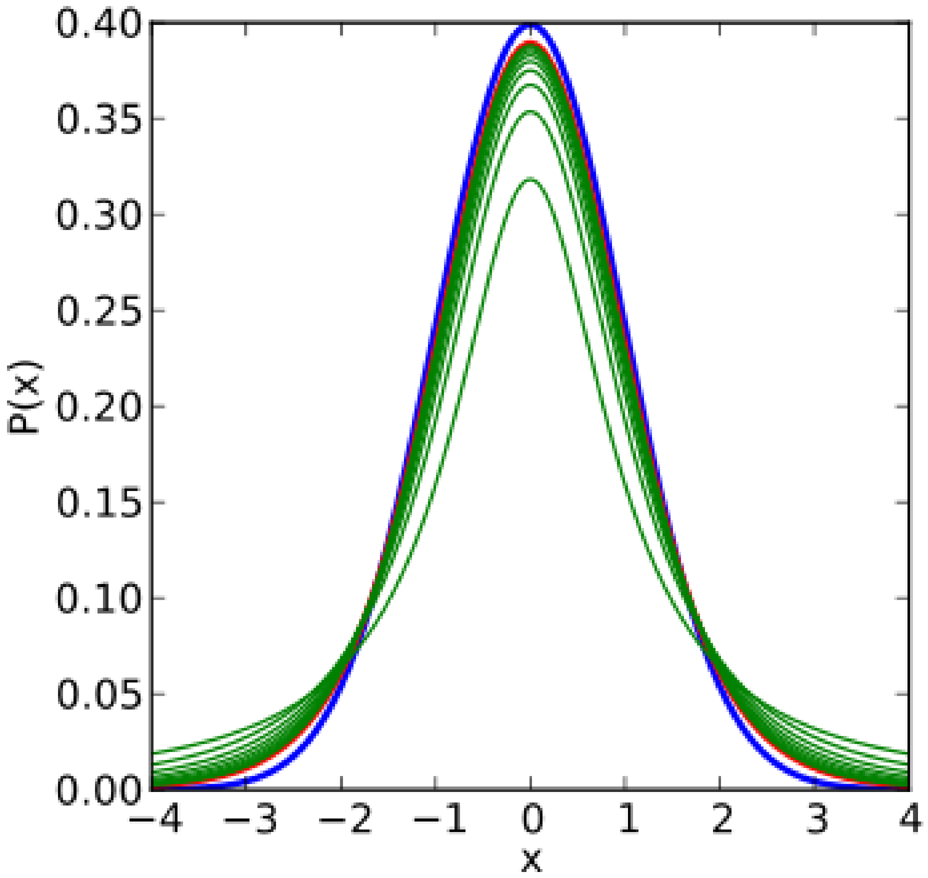
- Look up critical t-value (α=.05, two-tailed test, df=14): t-critical = 2.1448

t Table

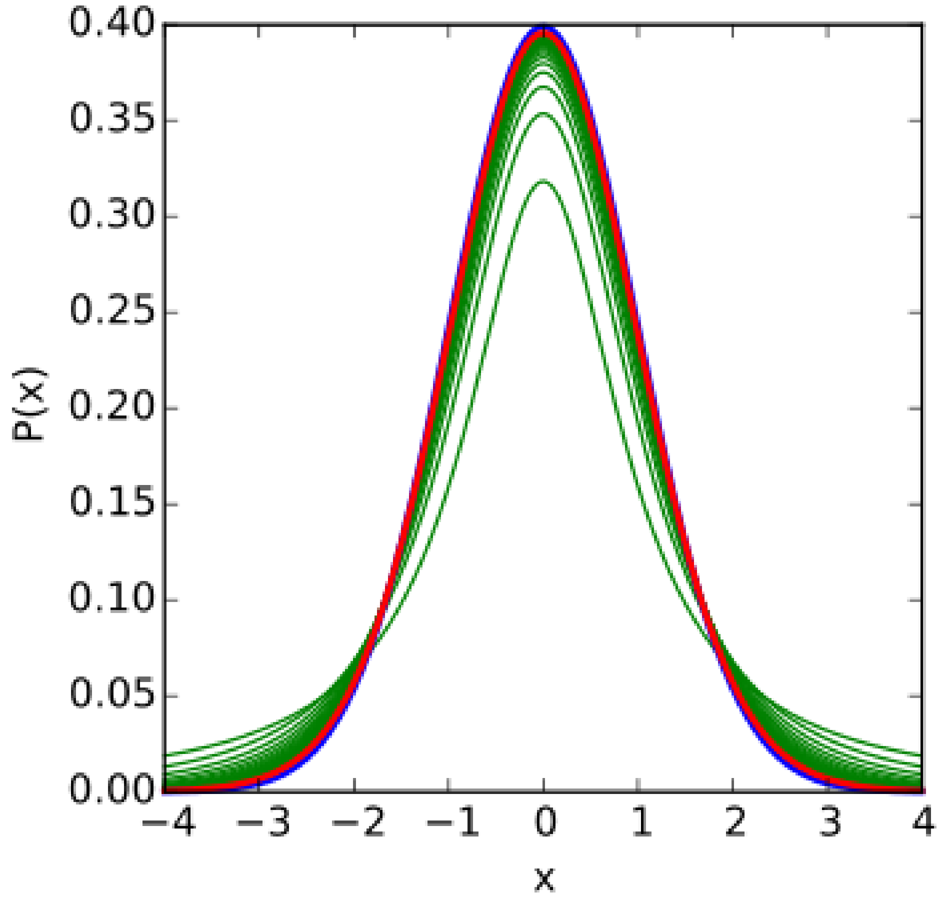
df			
1-tailed	0.1	0.05	0.025
2-tailed	0.2	0.1	0.05
1	3.0777	6.3138	12.7062
2	1.8856	2.9200	4.3027
	…		
13	1.3502	1.7709	2.1604
14	1.3450	1.7613	2.1448

- Source: http://statisticslectures.com/tables/ttable/

Let‘s look at the t distribution



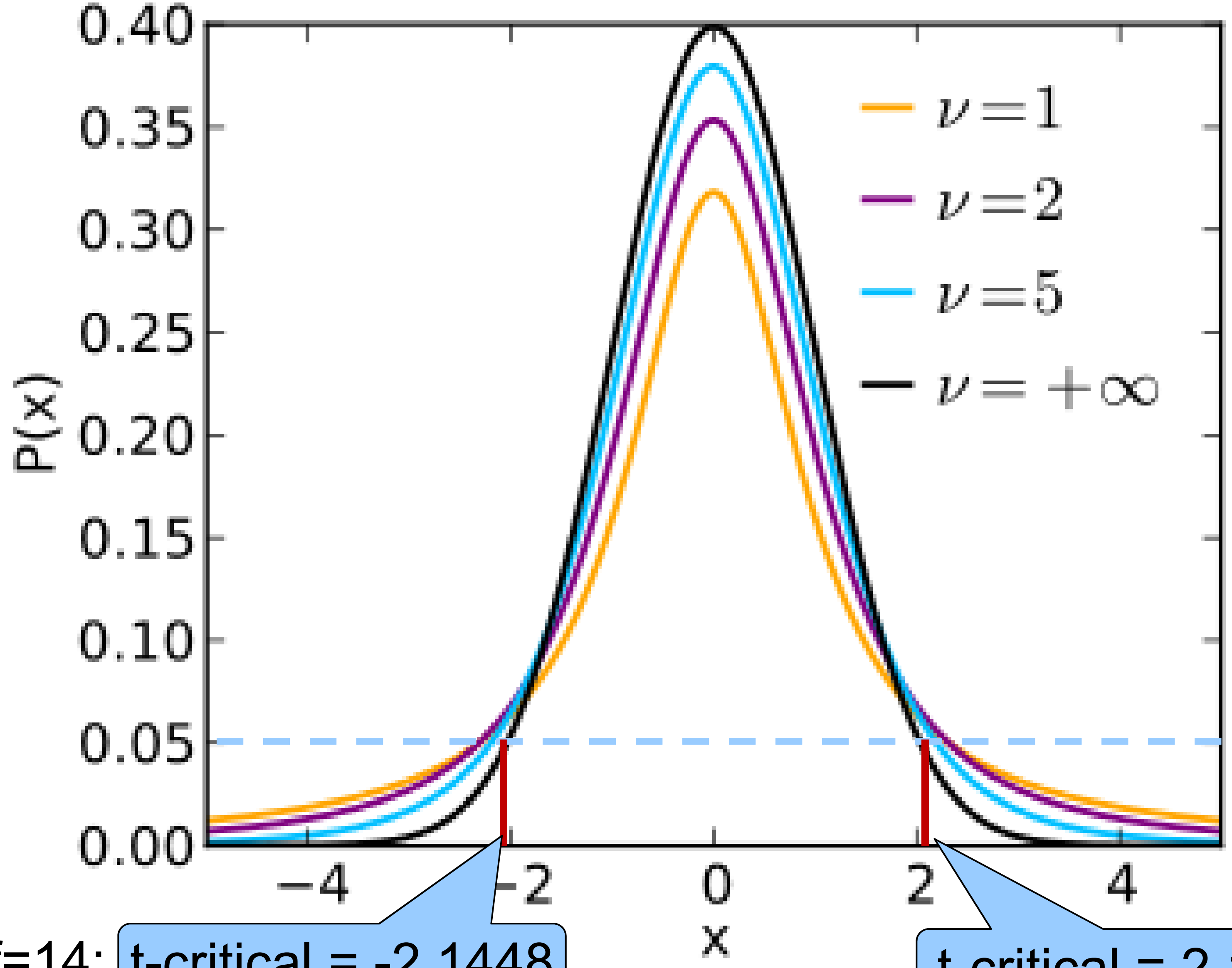
df =10



df =30

Source: Wikipedia

Let‘s look at the t distribution



For df=14: t-critical = -2.1448

t Test: Example of Two-tailed Test

- Calculate pooled variance

Estimate the population variance

$$s_p^2 = \frac{s_1^2 df_1 + s_2^2 df_2}{df_1 + df_2}$$

$$= \frac{s_1^2 (n_1 - 1) + s_2^2 (n_2 - 1)}{(n_1 - 1) + (n_2 - 1)}$$

$$= \frac{\sum_{i=1}^{n} (x_{1_i} - \bar{x}_1)^2 + \sum_{i=1}^{n} (x_{2_i} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} = \ldots$$

t Test: Example of Two-tailed Test

- Calculate pooled variance (continued)

$$s_p^2 = \ldots = \frac{\sum_{i=1}^{n}(x_{1_i} - \bar{x}_1)^2 + \sum_{i=1}^{n}(x_{2_i} - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)}$$

$$= \frac{12260.875 + 8866}{(8-1)+(8-1)} = 1509.0625$$

$$\bar{x}_1 = 270.125, \bar{x}_2 = 228$$

t Test: Example of Two-tailed Test

- Apply test statistics

In textbooks, you may see various versions

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2 n_1 + s_p^2 n_2}{n_1 n_2}}}$$

$$= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p \boxed{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}$$

Normalization factor for pooled variance

$$= \frac{270.125 - 228}{\sqrt{\frac{1509.0625}{8} + \frac{1509.0625}{8}}} \approx \frac{42.125}{19.423} \approx 2.169$$

t Test: Example of Two-tailed Test

- Returned t-value $t \approx 2.169 > 2.1448$
- Higher than t-value for specific df=14 at 95% confidence interval, which is 2.1448

  → Reject H_0

- Given the significance level $\alpha = 5\%$. There is a significant difference in the task completion time between ‘No prediction’ group and ‘With prediction’ group, $t\ (14)=2.169$, $p < .05$.

Please Note: p-value vs. α

- Given the significance level $\alpha = 5\%$. There is a significant difference in the task completion time between 'No prediction' group and 'With prediction' group, $t\,(14)=2.169$, $p < .05$.

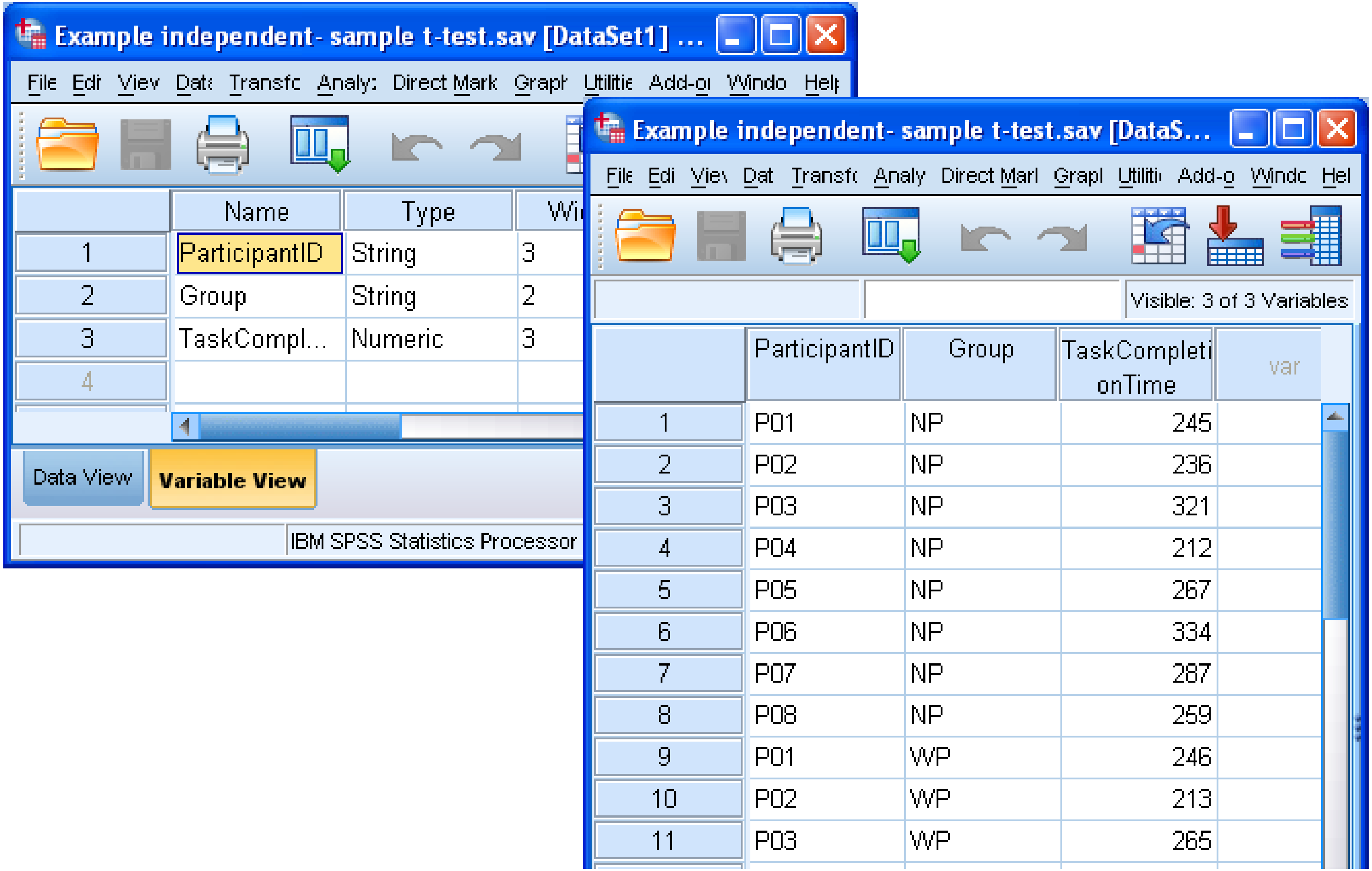
- p-value: probability of observing an effect given that H_0 is true (i.e., we actually should not observe that effect) **!**
- Significance level α: probability of rejecting H_0 given that it is actually true (Type I error) **!**

t Test: Example of Two-tailed Test

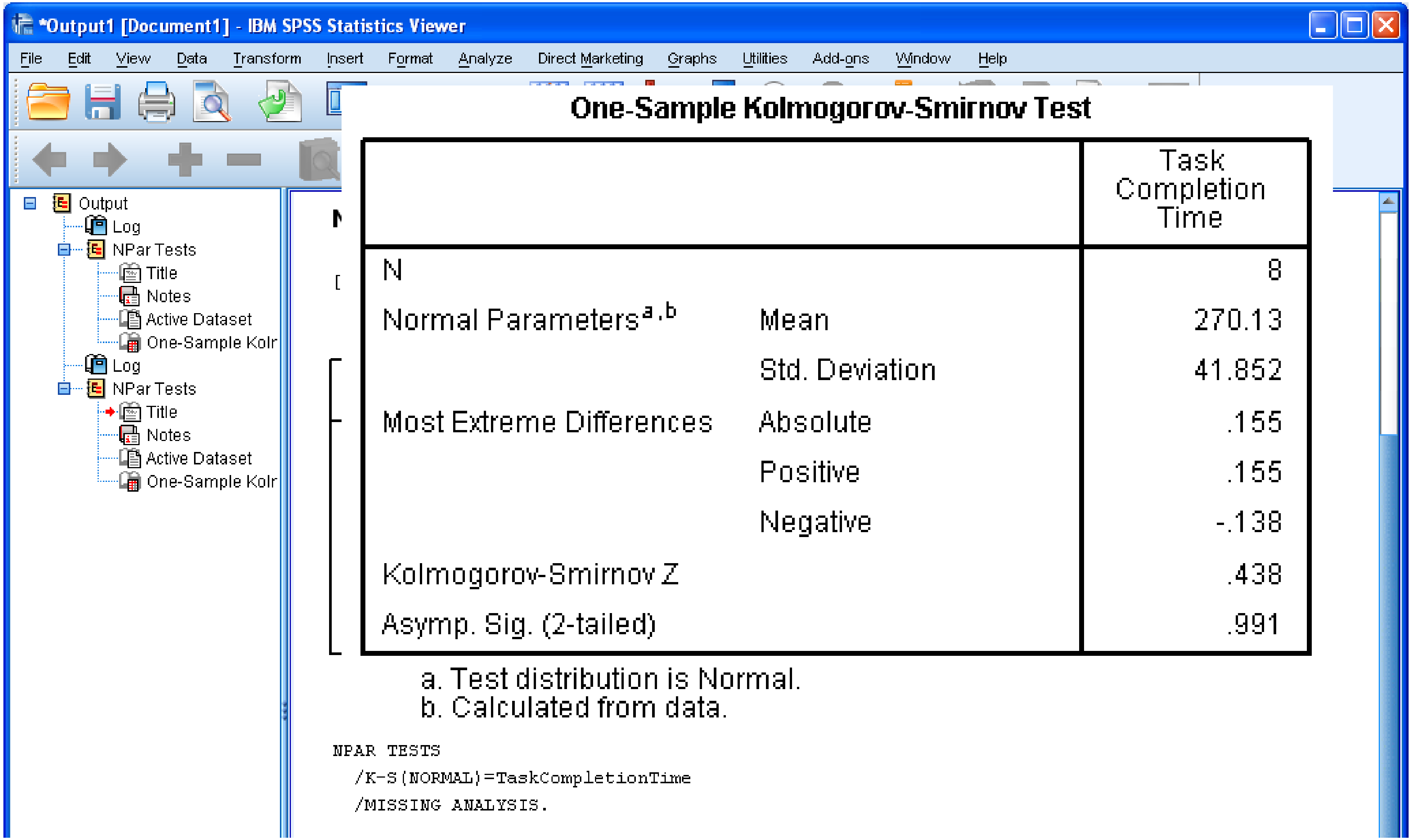
- Example computations with Gnumeric and SPSS 20 in Dropbox
  http://dl.dropbox.com/u/8830452/RMinHCI/DataAnalysis.zip

- Also: computations with sample data from within-group design

Example in SPSS



Example in SPSS (cont.)



One-Sample Kolmogorov-Smirnov Test

		Task Completion Time
N		8
Normal Parameters[a,b]	Mean	270.13
	Std. Deviation	41.852
Most Extreme Differences	Absolute	.155
	Positive	.155
	Negative	-.138
Kolmogorov-Smirnov Z		.438
Asymp. Sig. (2-tailed)		.991

a. Test distribution is Normal.
b. Calculated from data.

```
NPAR TESTS
  /K-S(NORMAL)=TaskCompletionTime
  /MISSING ANALYSIS.
```

Example in SPSS (cont.)

... ➜ **NPar Tests**

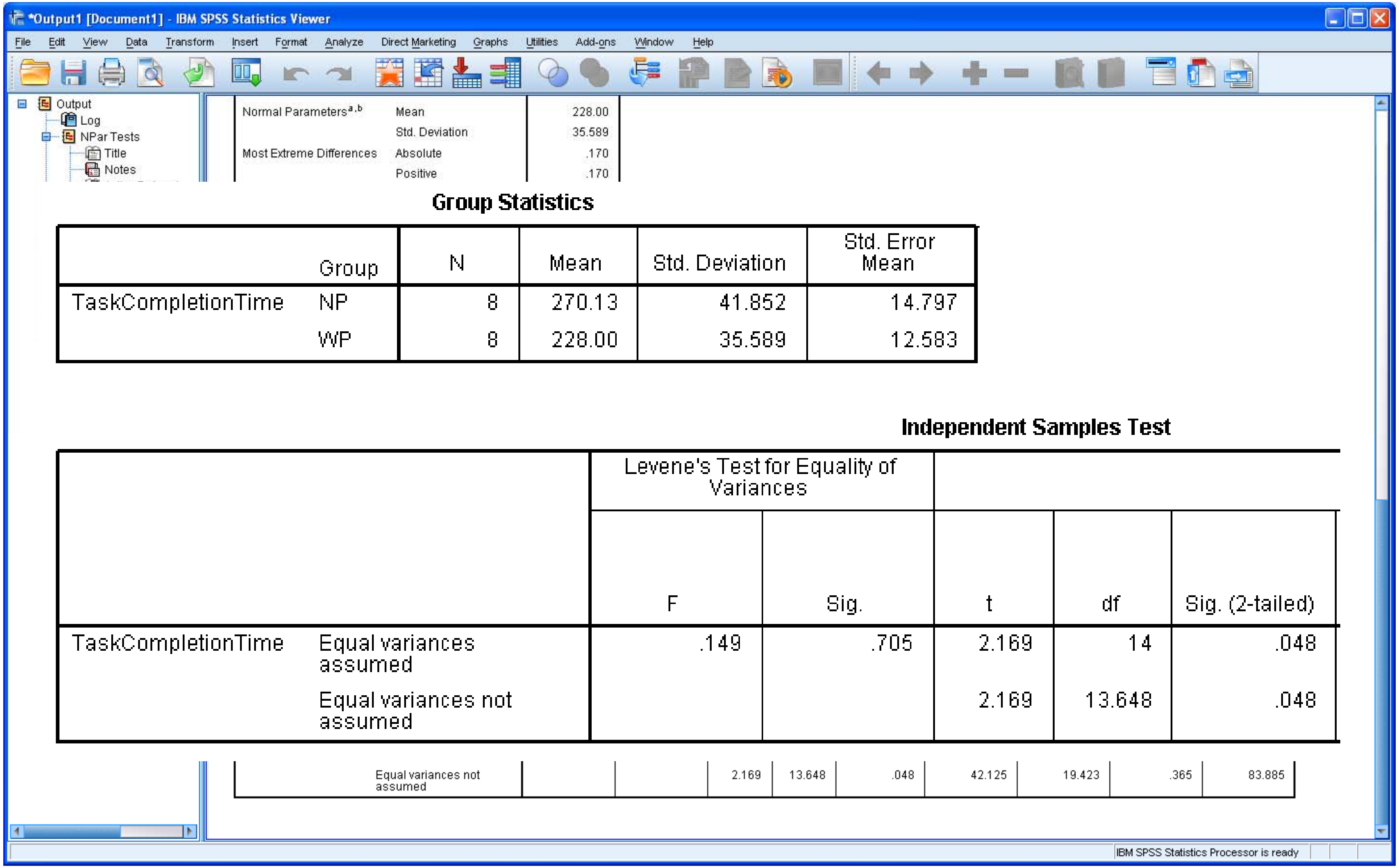
```
[DataSet1] X:\Dropbox\Public\RMinHCI\DataAnalysis\Example indepe
```

One-Sample Kolmogorov-Smirnov Test

		Task Completion Time
N		8
Normal Parameters[a,b]	Mean	228.00
	Std. Deviation	35.589
Most Extreme Differences	Absolute	.170
	Positive	.170
	Negative	-.137
Kolmogorov-Smirnov Z		.480
Asymp. Sig. (2-tailed)		.975

a. Test distribution is Normal.
b. Calculated from data.

Example in SPSS (cont.)



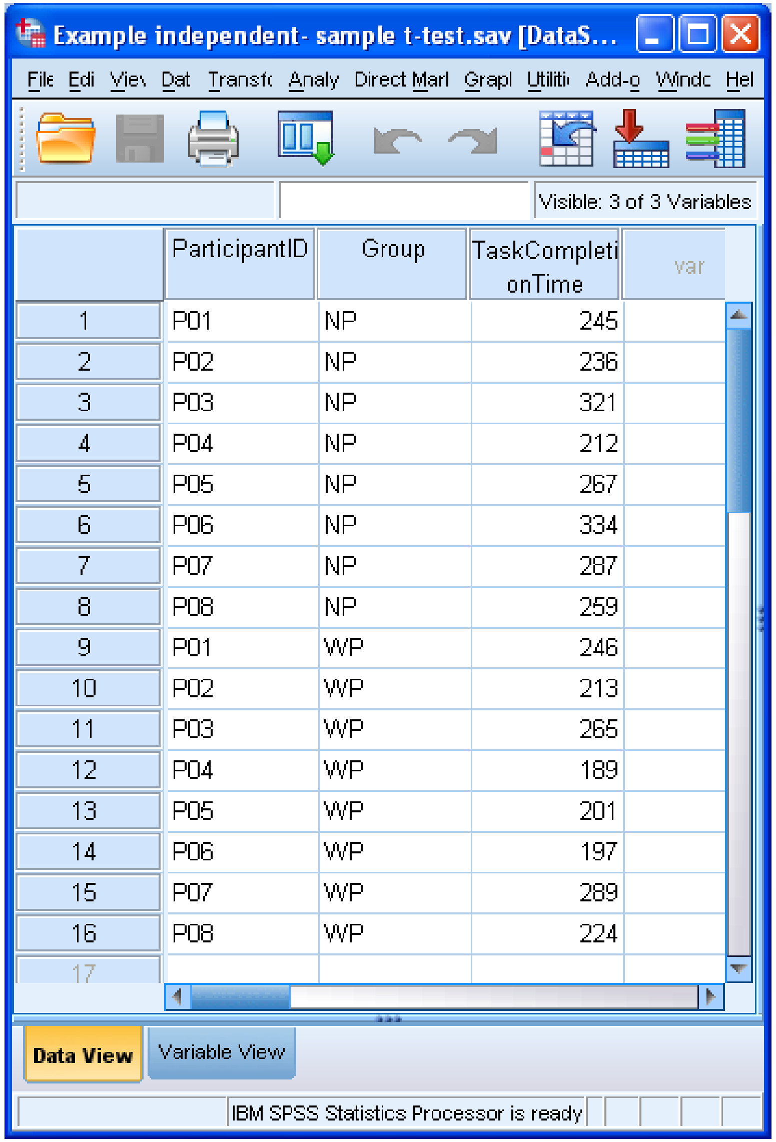
Group Statistics

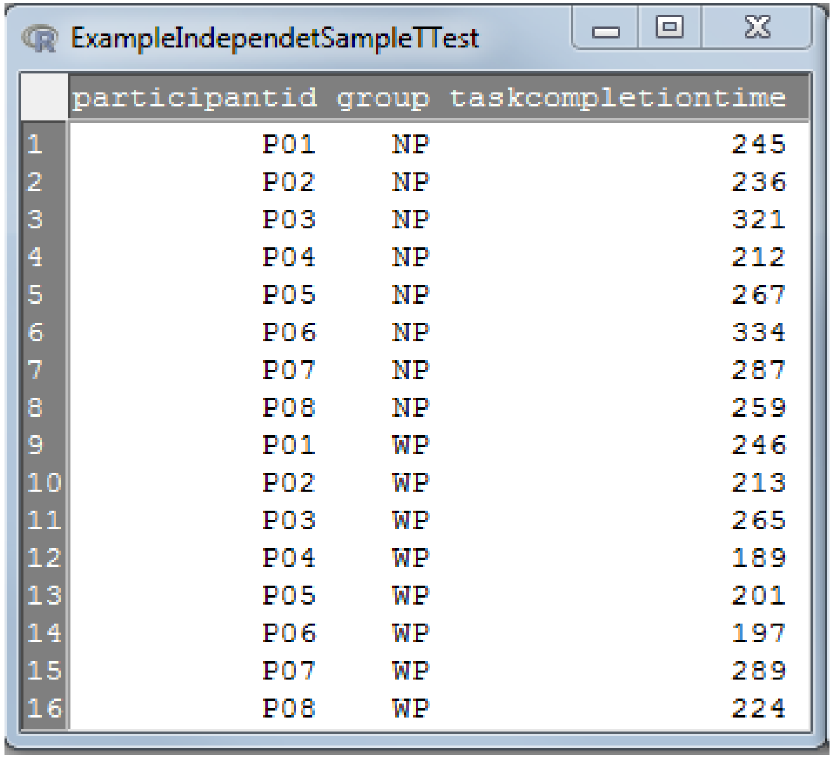
	Group	N	Mean	Std. Deviation	Std. Error Mean
TaskCompletionTime	NP	8	270.13	41.852	14.797
	WP	8	228.00	35.589	12.583

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
TaskCompletionTime	Equal variances assumed	.149	.705	2.169	14	.048
	Equal variances not assumed			2.169	13.648	.048

SPSS vs. R Commander (Rcmdr)





Organizing Data in Rcmdr

ExampleIndependetSampleTTest

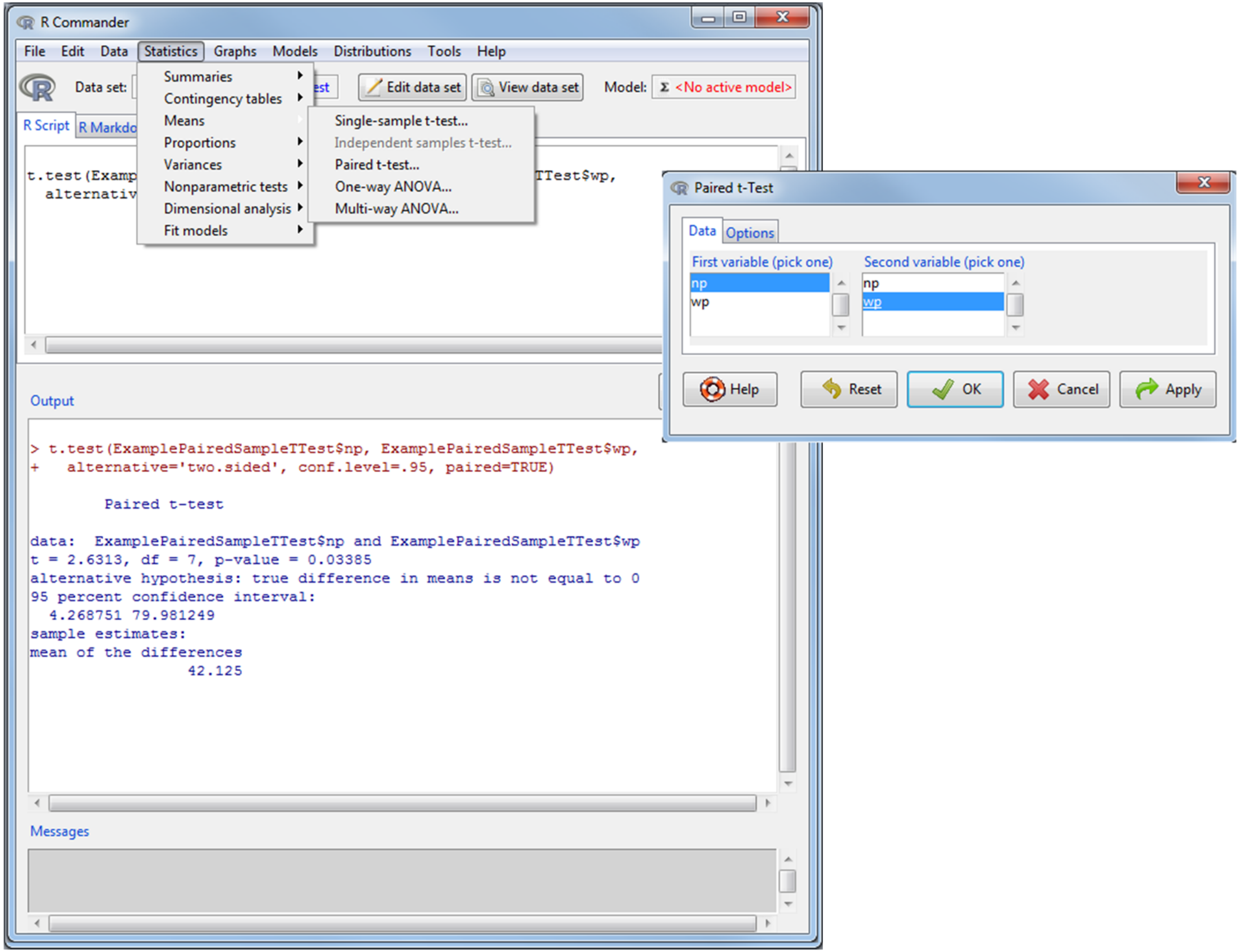
	participantid	group	taskcompletiontime
1	P01	NP	245
2	P02	NP	236
3	P03	NP	321
4	P04	NP	212
5	P05	NP	267
6	P06	NP	334
7	P07	NP	287
8	P08	NP	259
9	P01	WP	246
10	P02	WP	213
11	P03	WP	265
12	P04	WP	189
13	P05	WP	201
14	P06	WP	197
15	P07	WP	289
16	P08	WP	224

ExamplePair...

	participantid	np	wp
1	P01	245	246
2	P02	236	213
3	P03	321	265
4	P04	212	189
5	P05	267	201
6	P06	334	197
7	P07	287	289
8	P08	259	224
9		NA	NA
10		NA	NA
11		NA	NA
12		NA	NA
13		NA	NA
14		NA	NA
15		NA	NA
16		NA	NA

- Independent sample vs. paired sample

Example in R Commander (Rcmdr)



t Test: Example of One-tailed Test

- H_0: Individuals who use word-prediction software can type faster than those who do not use word-prediction software.

- Difference?
  - Direction is specified in hypothesis
  - Implying that word-prediction may improve typing speed or have no impact at all
  - One-tailed t test is appropriate

t Test: Example of One-tailed Test

	A	B	C
1		*Variable 1*	*Variable 2*
2	*Mean*	270.125	228
3	*Variance*	1751.55357142857	1266.57142857143
4	*Observations*	8	8
5	*Pooled Variance*	1509.0625	
6	*Hypothesized Mean Difference*	0	
7	*Observed Mean Difference*	42.125	
8	*df*	14	
9	*t Stat*	2.1687839769566	
10	*P (T<=t) one-tail*	0.0239062481778	
11	*t Critical one-tail*	1.76131013577595	
12	*P (T<=t) two-tail*	0.0478124963556	
13	*t Critical two-tail*	2.14478668791818	

ANalysis Of VARiance (ANOVA)

- One-way ANOVA
  - One IV with two or more conditions
  - Returns a value F (thus, also called F-test)
- Factorial ANOVA
  - Like one-way ANOVA, but two or more IVs involved
- Both are for between-group designs

- One-way/Factorial Repeated Measures ANOVA

One-way ANOVA: Example

Group	Participants	Task completion time	Coding
Standard	Participant 1	245	0
Standard	Participant 2	236	0
Standard	Participant 3	321	0
Standard	Participant 4	212	0
Standard	Participant 5	267	0
Standard	Participant 6	334	0
Standard	Participant 7	287	0
Standard	Participant 8	259	0
Prediction	Participant 1	246	1
Prediction	Participant 2	213	1
Prediction	Participant 3	265	1
Prediction	Participant 4	189	1
Prediction	Participant 5	201	1
Prediction	Participant 6	197	1
Prediction	Participant 7	289	1
Prediction	Participant 8	224	1
Speech-based dictation	Participant 1	178	2
Speech-based dictation	Participant 2	289	2
Speech-based dictation	Participant 3	222	2

Source	Sum of squares	df	Mean square	*F*	Significance
Between-group	7842.250	2	3921.125	2.174	0.139
Within-group	37880.375	21	1803.827		

- Three conditions
  - No prediction (standard)
  - With prediction
  - Speech-based dictations

One-way Repeated Measures ANOVA

	Standard	Prediction	Speech
Participant 1	245	246	178
Participant 2	236	213	289
Participant 3	321	265	222
Participant 4	212	189	189
Participant 5	267	201	245
Participant 6	334	197	311
Participant 7	287	289	267
Participant 8	259	224	197

Source	Sum of square	Df	Mean square	*F*	Significance
Entry method	7842.25	2	3921.125	2.925	0.087
Error	18767.083	14	1340.506		

One-way Repeated Measures ANOVA

- Significance does not mean that all means are actually different
- Conduct a pairwise t-tests
- Relative high risk of error due to multiple tests

- Reduce risk of error by using post-hoc test like Student-Newman-Keuls-Test
- Or apply Bonferroni correction, i.e., compare $\alpha_{new} = \alpha/m$, m being the number of conditions !

Split-plot ANOVA

- Involves between-group & within-group factors
- Example experiment design
  - Group 1 for task "Transcription"
  - Group 2 for task "Composition"
  - Both groups experience 3 conditions (K, P, S)

	Keyboard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composition	Group 2	Group 2	Group 2

- Could be considered two separate experiments with group 1 and group 2, respectively?

Split-plot ANOVA data layout

Task type	Participant number	Task type coding	Standard	Prediction	Speech
Transcription	Participant 1	0	245	246	178
Transcription	Participant 2	0	236	213	289
Transcription	Participant 3	0	321	265	222
Transcription	Participant 4	0	212	189	189
Transcription	Participant 5	0	267	201	245
Transcription	Participant 6	0	334	197	311
Transcription	Participant 7	0	287	289	267
Transcription	Participant 8	0	259	224	197
Composition	Participant 9	1	256	265	189
Composition	Participant 10	1	269	232	321
Composition	Participant 11	1	333	254	202
Composition	Participant 12	1	246	199	198
Composition	Participant 13	1	259	194	278
Composition	Participant 14	1	357	221	341
Composition	Participant 15	1	301	302	279
Composition	Participant 16	1	278	243	229

Split-plot ANOVA summary report

Source	Sum of square	df	Mean square	F	Significance
Task type	2745.187	1	2745.187	0.995	0.335
Error	38625.125	14	2758.937		

Table 4.18 Results of the split-plot test for the between-group variable.

Source	Sum of square	df	Mean square	F	Significance
Entry method	17564.625	2	8782.313	5.702	0.008
Entry method * task type	114.875	2	57.437	0.037	0.963
Error (entry method)	43126.5	28	1540.232		

Table 4.19 Results of the split-plot test for the within-group variable.

Summary: Parametric Methods

Commonly used significance tests for comparing means and their application context

Experiment design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

Correlation, Effect Sizes, and Confidence Intervals

Statistical Power: Lessons Learned?

- One-tailed tests are more powerful in rejecting H_0 than two-tailed tests as they do not require such strong differences in the conditions
- Increasing the number of participants allows to find the smallest effect having statistical significance (=rejecting H_0 at „any costs“)
- Thus, it is extremely important to report the effect size of the experiment result

!

- Reality check!!!

Types of Effect Sizes

- Correlation family
  - Effect size based on „variance explained“
  - Example: Pearson‘s r

- Difference family
  - Effect size based on differences between means
  - Example: Cohen‘s d

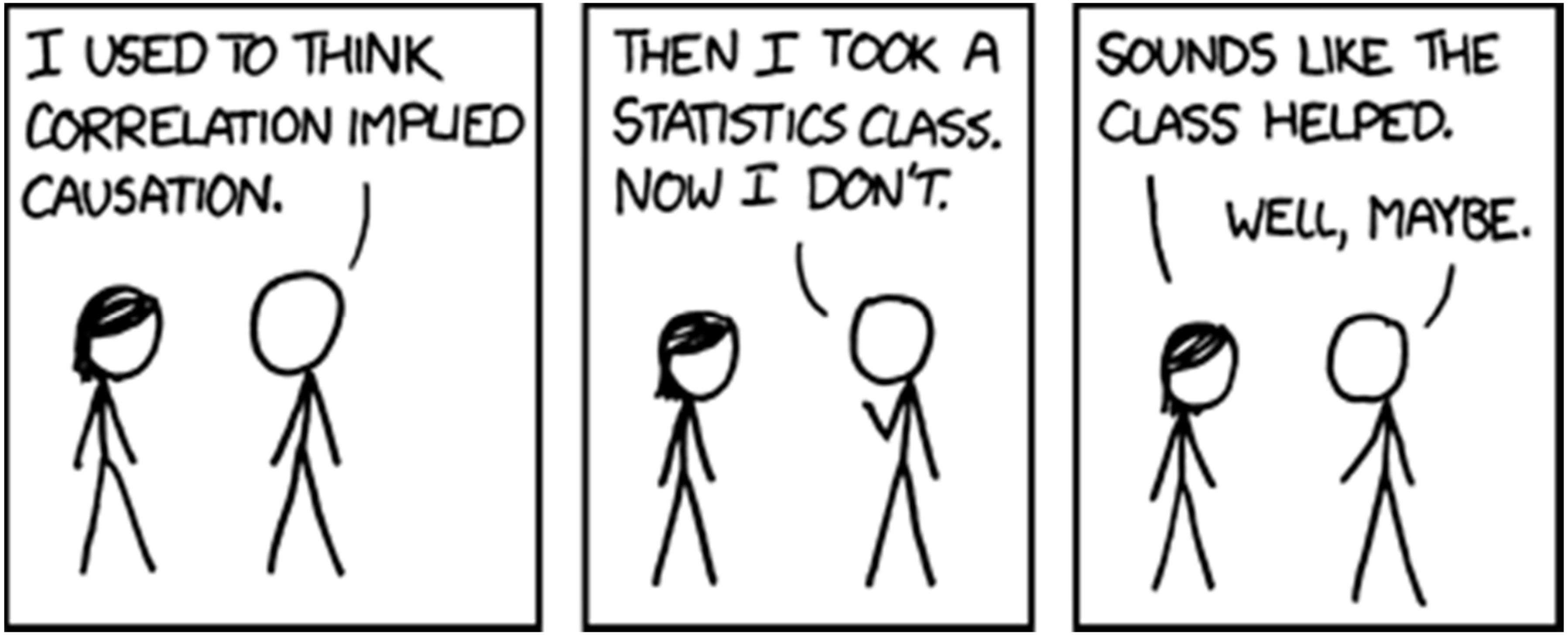
- And others, e.g., for categorical variables

Correlation Analysis

- Identify relationships between two factors
- Pearson‘s correlation coefficient (Pearson’s r)
  - Linear relationships
  - Perfect positive prediction when r = 1,
    no linear relationship when r = 0,
    perfect negative relationship when r = -1.
- Negative relationship equally good as positive
- Pearson‘s r as the most common effect size

- Caution: correlation does not imply causation !

Got it?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

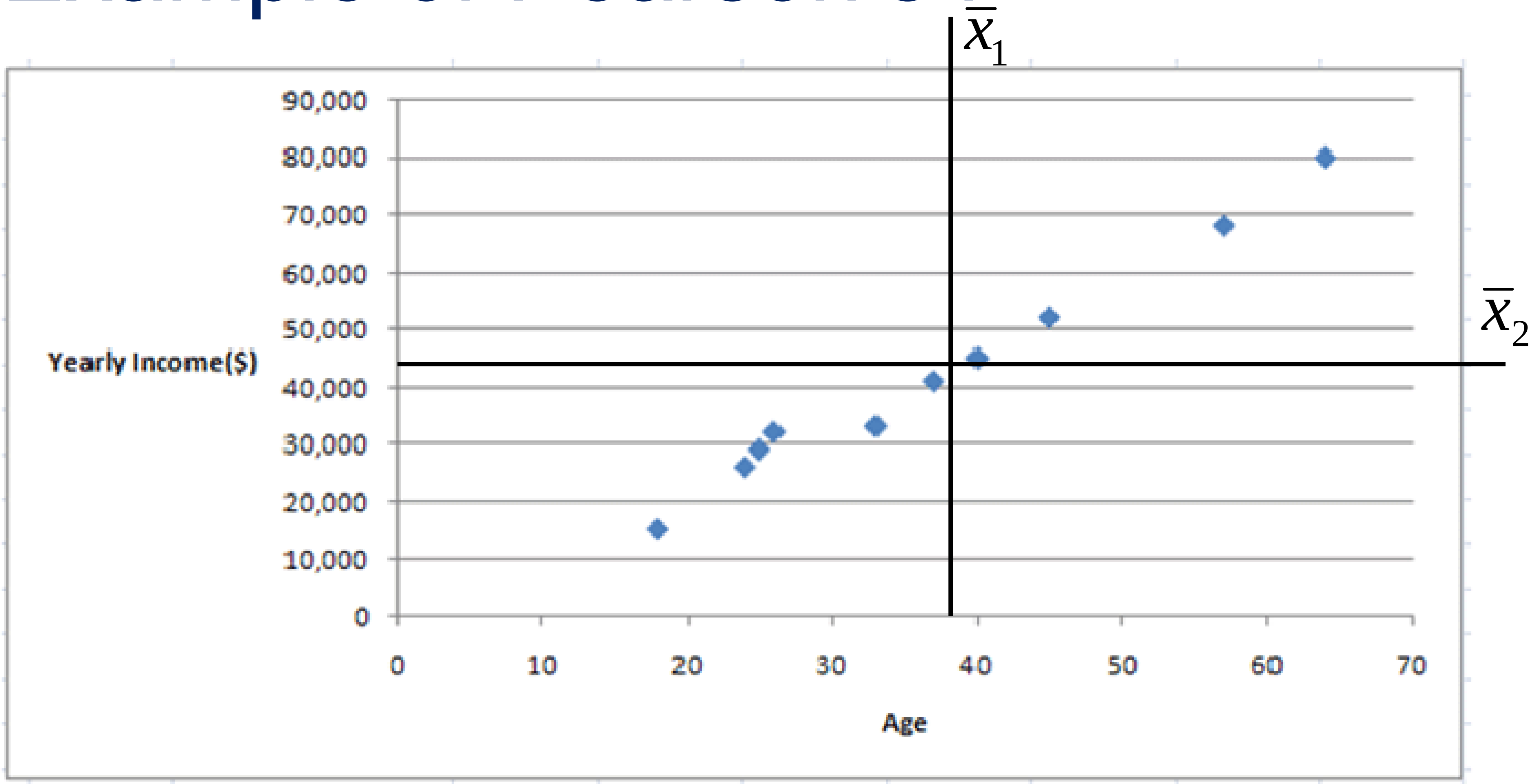
Computing Pearson's r

- Deviation score formula

$$r = \frac{\sum_{i=1}^{n}(x_{1_i} - \bar{x}_1)(x_{2_i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1_i} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{2_i} - \bar{x}_2)^2}}$$

- The i-th score on the j-th treatment: x_{j_i}

Example of Pearson's r



- Age and yearly income have a strong positive relationship ($r(8) = .99, p < .05$).

Source: http://www.statisticslectures.com/topics/pearsonr/

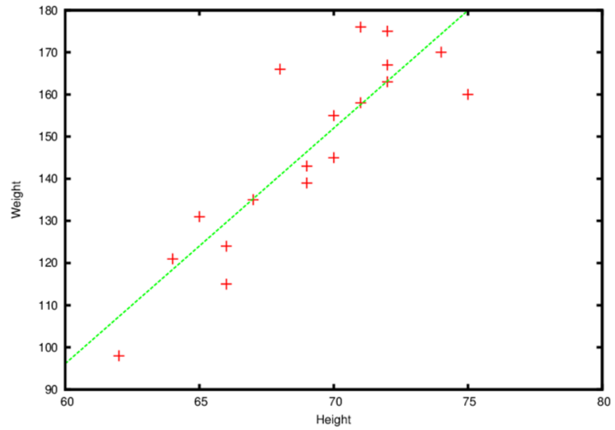
Effect Sizes in Pearson‘s r

- Allow to compare different experiments in an objective way
  - r = .10 (small effect): in this case the effect explains 1% of the total variance
  - r = .30 (medium effect): explains 9% of total variance
  - r = .50 (large effect): explains 25% of the variances

- Spearman’s r for non-linear relationships

Excursion DMML: Goodness of Fit

- Compare the regression sum of squares (SSM) with the total sum of squares (SST):



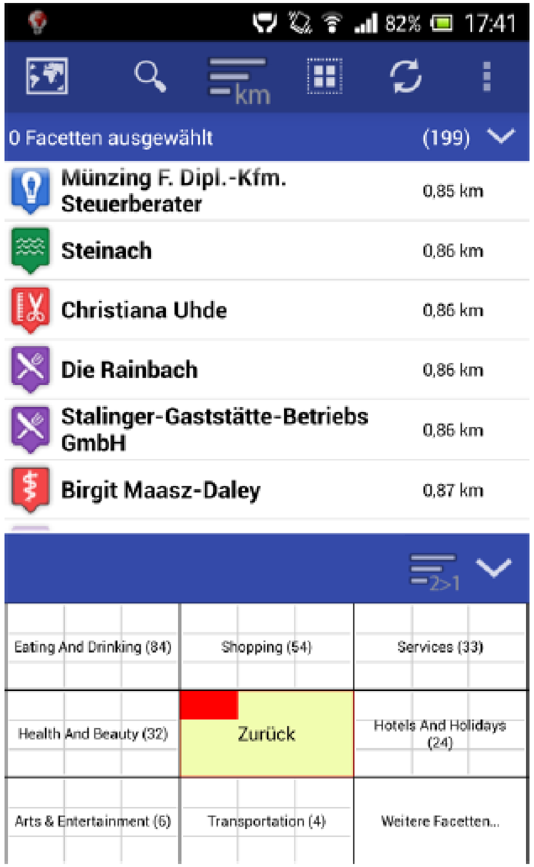
$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1\ldots m}(f(\boldsymbol{x_i}) - \bar{y})^2}{\sum_{i=1\ldots m}(y_i - \bar{y})^2}$$

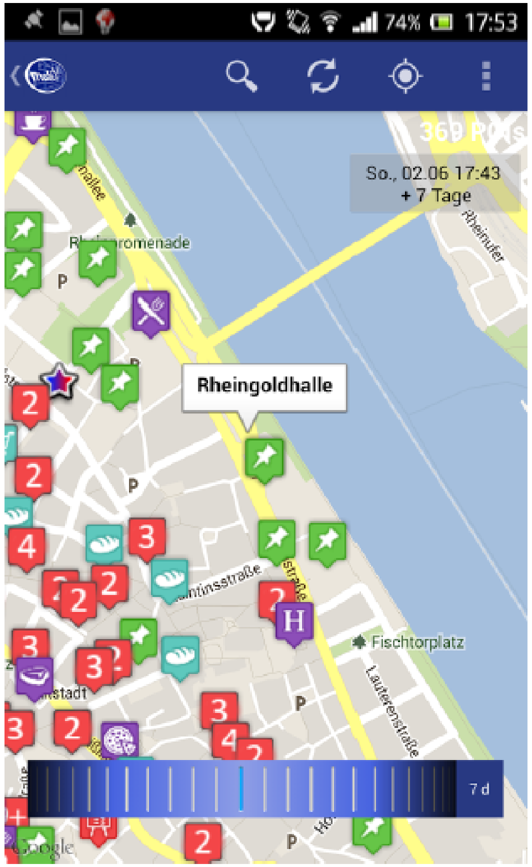
- Denotes how much variability can be explained with the regression model
- Note: R^2 increases when more explanatory variables are added to the model → use adjusted R^2 measure

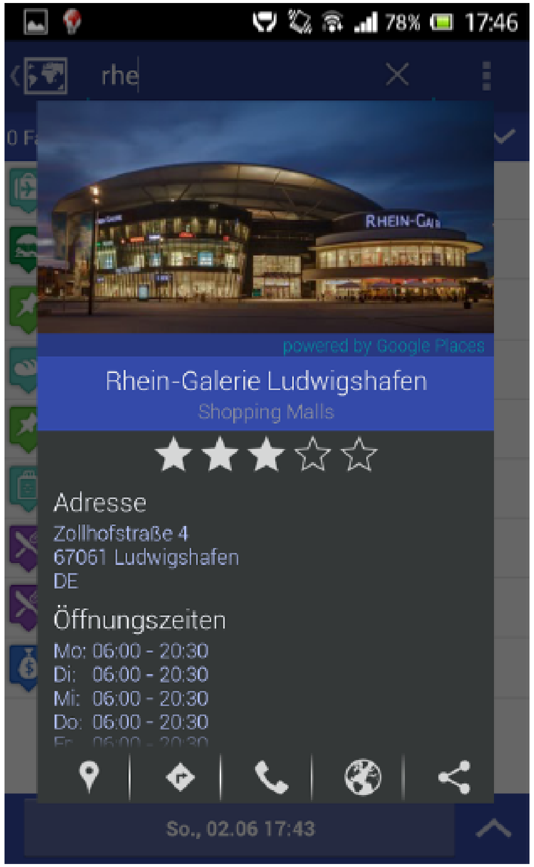
$$Adjusted R^2 = 1 - \left(\frac{m-1}{m-d}\right)(1 - R^2)$$ with

  - m number of objects and
  - $d + 1$ number of parameters in the model

Example: Entity Resolution in mobEx

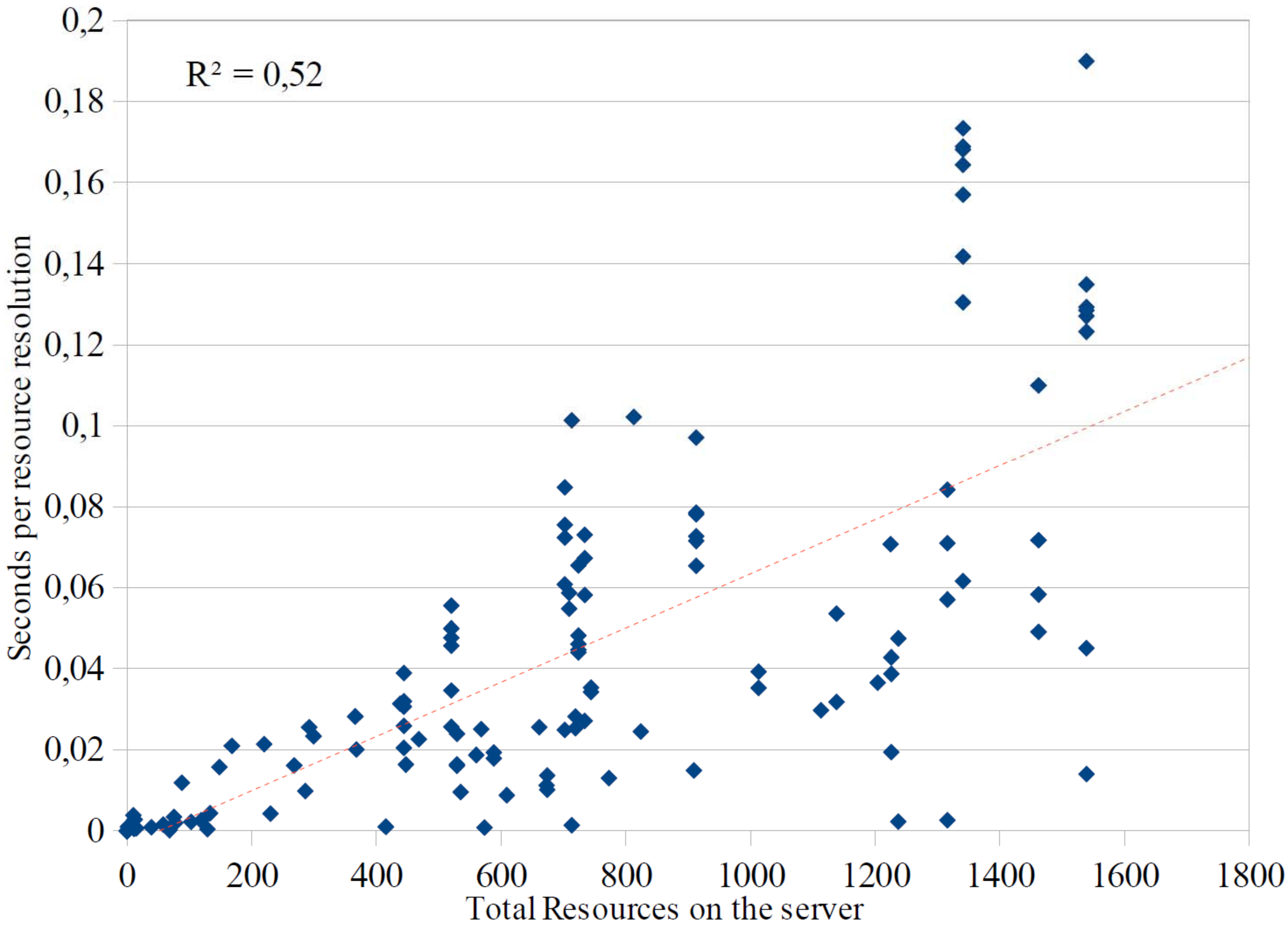






- Integration of nine data providers
- Query-time entity resolution
- Delivery of results once first merges are available

Example: Runtime Performance



Difference between Means: Cohen's d

- Most common measure of how much the treatment affects the dependent variable

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$ (population standardized mean difference)

- Can be estimated from the means' estimators and the pooled variance

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \quad \text{with} \quad t_{independent} = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Note: the pooled variance disappears

Results in: $d = t_{independent} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Difference between Means: Cohen's d

- Cohen's d for independent samples t test

$$d = t_{independent} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

  - Small effect: $d \geq .2$
  - Medium effect: $d \geq .5$
  - Large effect: $d \geq .8$

- Cohen's d for dependent samples t test

$$d = t_{dependent} \cdot \sqrt{\frac{2(1-r)}{n}}$$

$$t_{dependent} = \frac{\bar{X}_D}{s_D \cdot \sqrt{1/n}}$$

where n is the number of pairs, r is correlation between paired scores

Difference between Means: Cohen’s d

- Cohen’s d can range from 0 to positive infinity
- Most values of d vary from 0 to 1
- Cohen’s d is like a Z score in that its denominator is a standard deviation

Confidence Intervals (CI)

- Compute an upper and lower limit of an interval $[l, u]$
  - Such that the probability that the *fixed* parameter (e.g., population mean) is contained in $[l, u]$ is $1 - \alpha$
  - $1 - \alpha$ is the confidence level coverage
  - α is the Type I error rate

- Typically $\alpha = 0.05$, i.e., we compute 95% CIs (also written as: $CI_{.95}$)

- CIs can be computed for single means and differences between means (two distributions)

CI on Mean (Single Distribution)

- Compute the 95% CI $[l, u]$ given mean $\bar{x}$, standard deviation s and a constant c then

$$l = \bar{x} - c \cdot \frac{s}{\sqrt{n}} \text{ and } u = \bar{x} + c \cdot \frac{s}{\sqrt{n}}$$

- For large samples with $n \geq 30$ (*Central Limit Theorem*)
  - Use standard normal distribution: $c = z_{(1-\alpha/2)}$
  - Example for $\alpha = 0.05$: $c = z_{0.975} = 1.96$
- For smaller samples with $n < 30$
  - Use approximation provided by Student's t-distribution at degree of freedom $\nu = n - 1$:

$$c = z_{(1-\alpha/2;\, \nu)}$$

CI on Mean (Single Distribution)

- More formally, a confidence interval is given as
$$p[\theta_L(\boldsymbol{X}) \leq \theta \leq \theta_U(\boldsymbol{X})] = 1 - \alpha$$
  where
  - θ is some parameter of interest (mean, difference in mean, variance, etc.)
  - $\alpha = \alpha_L + \alpha_U$ (i.e., lower and upper can be different)
  - $\theta_L(\boldsymbol{X})$ and $\theta_U(\boldsymbol{X})$ the lower and upper random confidence limits of the observed data $\boldsymbol{X}$ (we call them random limits, since they are based on random data)

CI on Mean (Single Distribution)

- Example for μ over normally distributed data ($n \geq 30$):

$$p\left[z_{(\alpha/2)} \leq \frac{\bar{X}-\mu}{s_{\bar{X}}} \leq z_{(1-\alpha/2)}\right] = 1 - \alpha$$

$$\Leftrightarrow p\left[z_{(\alpha/2)} \cdot s_{\bar{X}} - \bar{X} \leq -\mu \leq z_{(1-\alpha/2)} \cdot s_{\bar{X}} - \bar{X}\right] = 1 - \alpha$$

$$\Leftrightarrow p\left[\bar{X} - z_{(1-\alpha/2)} \cdot s_{\bar{X}} \leq \mu \leq \bar{X} - z_{(\alpha/2)} \cdot s_{\bar{X}}\right] = 1 - \alpha$$

$$\Leftrightarrow p\left[\bar{X} - z_{(1-\alpha/2)} \cdot s_{\bar{X}} \leq \mu \leq \bar{X} + z_{(1-\alpha/2)} \cdot s_{\bar{X}}\right] = 1 - \alpha$$

$-z_{(\alpha/2)} = z_{(1-\alpha/2)}$ since the z-distribution is symmetric

$\bar{X}$: sample mean
μ : population mean (parameter we seek to estimate for)
$s_{\bar{X}}$: population standard deviation of the sampling distribution of the mean (standard error)
$\alpha/2 = \alpha_U = \alpha_L$: significance level of computing the CI

CI on the Difference between Means

- Analog to the single distribution case, the equation in the two group situation with $n_1, n_2 < 30$ is defined as:

$$p\left[t_{(\alpha_L;\,\nu)} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{(1-\alpha_U;\,\nu)}\right] = 1 - \alpha$$

$$\Leftrightarrow p\left[(\bar{X}_1 - \bar{X}_2) - t_{(1-\alpha_L;\,\nu)} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \right.$$

$$\left. \leq (\bar{X}_1 - \bar{X}_2) + t_{(1-\alpha_U;\,\nu)} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right] = 1 - \alpha$$

with $\nu = n_1 + n_2 - 2$ degrees of freedom

- Further transformation needed to provide CI over standardized mean difference

Confidence Intervals

- A 95% CI can detect with a probability of 95% an interval that contains the true value of the parameter θ (e.g., the mean). Thus, after computing the CI, it is a binary decision. Either the parameter value is included in the CI or not. It is not longer a matter of probability!

- Thus, common misunderstandings of CIs are:
  - A 95% CI does contain 95% of the sample data.
  - A 95% CI is a range of plausible values for the sample mean. → but plausible values for the CI parameters
  - A 95% CI contains with 95% probability the population mean / population parameter.
  - A particular 95% CI has a 95% probability that a sample mean is falling into the interval when one does repeat the experiment.

Continued on Part 3 …