University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Some words on statistical significance

Sebastian Padó

Jun 12, 2015

Topics

1. Data analysis vs. Evaluation vs. Significance
2. Significance testing: how to do it
    1. Traditional methods
    2. Simulation-based significance testing
3. Effect Sizes

- NB. Statistics is a huge and developing field
    - I could spend a semester talking about this topic
    - Please participate!

Data and Models (I)

- The „simple case“: Experimental work
  - Measurement of **variable of interest**
    - E.g., F0 formant
  - Experimental **manipulation**
    - E.g., gender
- Significance-related questions:
  - **Q1.** Does the variable **change significantly** with manipuation?
    - E.g., do women have a higher F0?

Data and Models (II)

- The more complex case: Computational work
  - Gold-standard values of a variable (gold labels) given
    - E.g. part of speech tags
  - Have some **model** make **predictions**
  - **Evaluation measure** compares predictions and gold labels
    - E.g. accuracy
- Significance-related questions:
  - **Q2**. Are the predictions **significant**?
  - **Q3**. Is Model A **significantly different** from Model B?
    - Q3 subsumes Q2: „significantly different from **chance**“
    - Q3 corresponds to a subset of Q1:
      comparing two models == binary manipulation

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Why Significance?

- Why do we need to care about significance if we do a proper evaluation?
- Evaluation gives us **numbers** but does not tell us whether they are **meaningful**
- Examples:
  - Q: Binary classification task. A model achieves 50% accuracy. Is this a reasonable model?
    - No. 50% accuracy is **chance level** performance.
  - Q: In Exp 1, a model gets 11/20 (55%) examples correct. In Exp2, it gets 1100/2000 (55%) examples correct. Which experiment is more significant?

Pretheoretic Intuitions

- Larger differences are more meaningful
- Results on larger datasets are more meaningful
- Lower variance makes differences more meaningful

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Hypothesis Testing

In practice, most significance testing is operationalized as **hypothesis testing**

- Formulate a **null hypothesis** about your data
  - No **relationship** between two measured phenomena
  - Any differences are **due to chance**
- We want to reject the null hypothesis (in favor of an **alternative hypothesis**)
  - Gather evidence from the data
  - Select **appropriate** statistical test

Errors

- Just because there appears to be a difference, there needn‘t be one, and vice versa
- **Type I error**: incorrect rejection of a true null hypothesis
  - **Type II error**: failure to reject a false null hypothesis
- Which one is more problematic?
  - General assumption: Type I (scientists are conservative)
- Hypothesis testing always relative a chosen level of type-I errors (**p-value, α**)
  - (1-p) is called the **significance level** (e.g. 0.95 == p=5%)
- NB. Choosing a p level does not **reduce** the number of errors you make – you just trade Type I against Type II

University of Stuttgart
Germany

Notation, Terminology

- Terminology: „significant at p=0.05“
- Notation: asterisks (p>0.05: -, p=0.05:*, p=0.01:**, p=0.001:***)

	Probabilistic Models			Similarity-based Models				
	B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Accuracy	0.50	0.62	0.75	0.50	0.68	0.56	0.68	0.70
Coverage	1.00	0.44	0.75	1.00	0.98	0.94	0.98	0.98
Backoff Accuracy	0.50	0.55	**0.69**	0.50	0.68	0.56	0.68	**0.70**

		Probabilistic Models			Similarity-based Models				
		B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Prob.	B_p								
	SOV_p	-							
	SO_p	*	-						
Similarity	B_s	-	-	*					
	SOV_Σ	*	-	-	*				
	SOV_Π	-	-	-	-	-			
	SO_Σ	*	-	-	*	-	-		
	SO_Π	**	$*^\dagger$	-	**	-	$*^\dagger$	-	

(from Zarcone et al. 2012)

Finding the right test

- Traditional hypothesis testing is „made-to-measure"
  - For each type of variable and setup: different tests
  - Two-step procedure:
    - Compute test statistic (with nice mathematical properties)
    - Translate test statistic into p-value
  - Today: type numbers into R and try to understand output
- Examples:
  - chi squared: are two sets of counts (proportions) from the same distribution?
  - t-test: are two numeric samples from the same distribution?
  - ANOVA: are >2 samples from the same distribution?
  - ...

Example

- Effect lengths in group A: 23, 14, 35, 23, 26, 30
- Effect lengths in group B: 15, 20, 28, 26
- Q: Are the lap times **significantly different**?
- Use independent two-sample t-test
  - Requires that samples are independent (unpaired)
  - Requires that values are normally distributed
    - Should NOT use the test when assumptions are not met!
  - Null hypothesis: Means of two samples are identical
    - Higher statistic = more support to reject the null hypothesis

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Scaling behavior of t

- If the difference between the sample means becomes larger...
  - ...t increases
- If the sample sizes become larger...
  - ...t increases
- If the variances become larger...
  - ...t decreases

- Meets our intuitions!

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

t-test table

t **Table** from http://www.sjsu.edu/faculty/gerstman/StatPrimer

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318

- df = degrees of freedom
  - I‘m not going to go into that
- For a two-sample setup with n total measurements, df=n-2

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

t values

Why are there no negative values of *t* in the table?

- Because it‘s just the **absolute difference** that matters

Wait – it‘s somewhat more subtle than that

  - „**Two-tailed**“ test: Alternative hypothesis: Means of sample 1 and sample 2 are significantly **different**
  - „**One-tailed**“ test: Alternative hypothesis: Mean of sample 1 is significantly **larger** (**smaller**) than mean of sample 2
- One-tailed test becomes significant more easily
  - But: is based on an additional assumption
  - Must check sign of difference manually!
- Recommendation: use more conservative two-tailed test

In R

```
> a = c(15, 20, 28, 26)
> b = c(23,14,35,23,26,30)
> t.test(a,b)
   Welch Two Sample t-test
data:  a and b
t = -0.7028, df = 7.448, p-value = 0.5036
alternative hypothesis: true difference in means is not
   equal to 0
95 percent confidence interval:
 -12.61180   6.77847
sample estimates:
mean of x mean of y
 22.25000  25.16667
```

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Another Example

From an earlier slide:

„In Exp 1, model gets 11/20 (55%) ex. correct. In Exp2, it gets 1100/2000 (55%) ex. correct. Significant?"

- Use Pearson's Chi Squared test
  - compares observed values O and expected values E
  - compute E as row means

	null hyp.		experiment	
correct	10	10.5	11	10.5
incorrect	10	9.5	9	9.5

$$\chi^2 = \sum \frac{(E-O)^2}{E}$$

  - here: X^2 = (10-10.5)^2/10.5 + ... = 0.1
  - In this application of Chi Squared: df = 1

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Chi Square Table

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

from http://faculty.southwest.tn.edu/jiwilliams/probability.htm

In R

```
> a=matrix(c(10,10,11,9),nrow=2)
> a
     [,1] [,2]
[1,]   10   11
[2,]   10    9
>chisq.test(a,correct=F)

  Pearson's Chi-squared test

data:  a
X-squared = 0.1003, df = 1, p-value = 0.7515
```

Another Example

From an earlier slide:

„In Exp 1, model gets 11/20 (55%) ex. correct. In Exp2, it gets 1100/2000 (55%) ex. correct. Significant?"

- Use Pearson's Chi Squared test

	null hyp.		experiment	
correct	1000	1050	1100	1050
incorrect	1000	950	900	950

$$\chi^2 = \sum \frac{(E-O)^2}{E}$$

- hier: X^2 = (1000-1050)^2/1050 + ... = 10.03

University of Stuttgart
Germany

IMS

Institut für Maschinelle Sprachverarbeitung

In R

```
> b = matrix(c(1000,1000,1100,900),nrow=2)
> b
     [,1] [,2]
[1,] 1000 1100
[2,] 1000  900
> chisq.test(b,correct=F)

  Pearson's Chi-squared test

data:  b
X-squared = 10.0251, df = 1, p-value = 0.001544
```

Scaling Behavior of χ^2

- Larger difference between E and O:
  - Nominator grows quadratically: Chi squared increases
- Corpus size increases:
  - Assuming E and O grow linearly
  - Nominator grows quadratically, denominator grows linearly: Chi squared increases

- Again, meets our expectations

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Further Aspects: Multiple Comparisons

- Setting p values becomes tricky when you need to make many comparisons
  - p=0.05 means we **expect** 1 out of 20 comparisons to come out as significant even though it is not
  - Comparison of n models requires $O(n^2)$ comparisons
- **Bonferroni correction** simply divides *p* by *m*, the number of comparisons
  - This way, the overall Type I error rate remains constant..
  - ..but individual effects are harder to find

	Probabilistic Models			Similarity-based Models				
	B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Accuracy	0.50	0.62	0.75	0.50	0.68	0.56	0.68	0.70
Coverage	1.00	0.44	0.75	1.00	0.98	0.94	0.98	0.98
Backoff Accuracy	0.50	0.55	**0.69**	0.50	0.68	0.56	0.68	**0.70**

		Probabilistic Models			Similarity-based Models				
		B_p	SOV_p	SO_p	B_s	SOV_Σ	SOV_Π	SO_Σ	SO_Π
Prob.	B_p								
	SOV_p	-							
	SO_p	*	-						
Similarity	B_s	-	-	*					
	SOV_Σ	*	-	-	*				
	SOV_Π	-	-	-	-	-			
	SO_Σ	*	-	-	*	-	-		
	SO_Π	**	$*^\dagger$	-	**	-	$*^\dagger$	-	

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Further Aspects: (Non-)parametrics

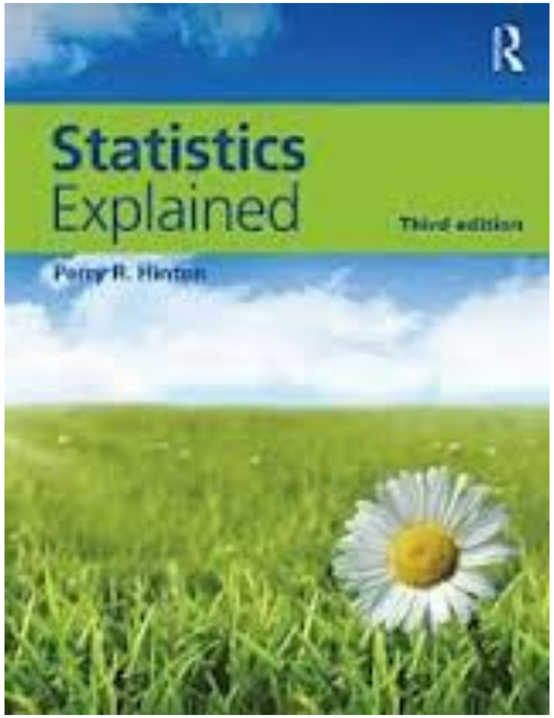
- The standard tests are **parametric**
  - Assume data follows some distribution (typically normal)
    - Wrong for many applications in language!
  - Can invalidate the outcome of significance tests
  - Always test for normality (e.g. Kolmogorov-Smirnov)
- Alternative approach: nonparametric tests
  - Avoid assuming a distribution, but are typically weaker
  - Many are based on **rank comparisons**
  - Rank-based analogue to t-test: Wilcoxon signed-rank test
    - Form all pairs of measurements from the two samples and count how often each of the samples is higher
  - Chi Square is actually nonparametric ;-)

A Good Book

Perry Hinton: Statistics Explained

Explains lots of statistics...

...and when to use them



Questions

Questions on traditional significance tests?

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

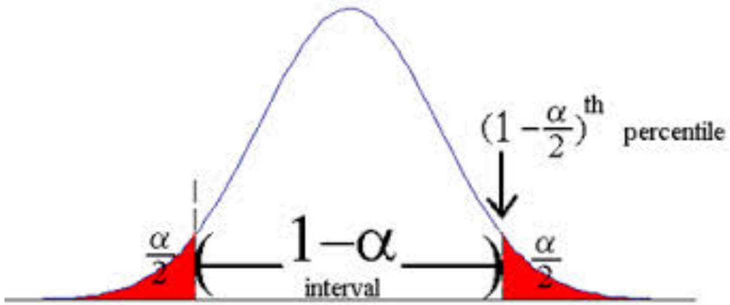
We Need Something Else

- Traditional significance tests are often unsuitable for Q3 (testing differences between computational models)
- Why?

- Traditional tests compare **means** or **counts**
- We want to use arbitrary evaluation metrics
  - F-Score: not a mean, nor a count
  - BLEU: not a mean, nor a count
  - ...
- Also, these metrics are almost certainly not normal...

Confidence Intervals

- This looks like a detour, but it will be relevant
- Think of your model evaluation as a **draw from a (e.g. normal) distribution**



https://onlinecourses.science.psu.edu/stat504/node/19

- If you knew the form of this distribution, you could compute **an interval** within which the true model quality is located **for a given p level/α**.
  - E.g. „The true accuarcy of the POS tagger is between 60% and 70%, at a 95% confidence level“
  - Easiest case: simply look at the percentiles

The relation to significance testing

- How does this relate to significance testing?
  - Compute confidence intervals not for the quality of one model, but for the **difference in quality between two models**
  - E.g. „The true difference in accuracy between the two POS taggers is between 3% and 5% at a 95% confidence level"
- Q: How does this relate to significance?
  - A: A difference is significant if the **confidence interval does not include 0** (cf. null hypothesis!)
- Q: What happens if you lower the p-level (higher threshold)?
  - A: The confidence interval gets „broader"

Simulation-Based Hypothesis Testing

- How to get from one number („the quality“) to a distribution we can compute a confidence interval from?
- **Resampling methods**: Create new, similar datasets from an existing dataset → simulation
  - Often used: **Bootstrap resampling**
    - Visualize evaluation as **set of bins** (e.g. sentences)
    - **Sampling from bins with replacement,**
  - More specifically: for a dataset of size m,
    - repeat for a large number n of times:
      - draw m results from sample with repl., compute statistic
    - treat n values as „sample from the quality distribution“

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

Boostrapping for Accuracy

- Example for simple case (one model instead of difference)
- POS tagger applied to 5 sentences (m=5)

Sent #	1	2	3	4	5
Correct	4	7	1	2	3
Total	5	8	3	5	7

- Overall accuracy: (4+7+...+3)/(5+8+...+7) = 60.7%
- Bootstrapping four values (n=3): draw randomly from [1..5]
  - 1: (1,2,1,1,4) => (4+7+4+4+2)/... = 0.75
  - 2: (1,4,3,3,2) => (4+2+1+1+7)/... = 0.625
  - 3: (4,4,3,2,1) => ... = 0.615

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

In R

- Library **boot** provides function **boot**
  - Takes a data frame
  - And a function that takes the frame and a vector of indices and computes the overall quality
  - And a number of samples to be taken
  - Returns an object that can compute confidence intervals
- Quality function for accuracy:

```
computeAcc <- function(data, indices) {
   sample <- data[indices,]
    acc <- sum(sample$corr)/sum(sample$total)
    acc }
```

In R

```
> library("boot")
> c = data.frame("corr" = c(4,7,1,2,3),
  "total"=c(5,8,3,5,7))
> sum(c$corr)/sum(c$total)
[1] 0.6071429
> b <- boot(c,computeAcc,100)
> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates
Intervals :
Level      Normal              Basic
95%    ( 0.3813,  0.8162 )   ( 0.3854,  0.8333 )
```

different ways of computing confidence intervals from distribution

Changing the confidence level

- Standard call assumes 95% confidence level

```
boot.ci(b) == boot.ci(b,conf=0.95)
...
Intervals :
Level      Normal              Basic
95%   ( 0.3813,  0.8162 )   ( 0.3854,  0.8333 )
```

- If I now do boot.ci(b,conf=0.99), how the numbers change as expected:

```
...
Intervals :
Level      Normal              Basic
99%   ( 0.3130,  0.8845 )   ( 0.3756,  0.8459 )
```

Example: Significant differences

- Two POS taggers applied to 5 sentences (m=5)

Sent #	1	2	3	4	5
Correct 1	4	7	1	2	3
Correct 2	3	8	2	2	2
Total	5	8	3	5	7

- Overall accuracy of both models; 60.7%
- Bootstrapping four values (n=3): draw randomly from [1..5]
  - 1: (1,2,1,1,4) => 0.75 - 0.68 = +0.07
  - 2: (1,4,3,3,2) => 0.63 - 0.71 = -0.07
  - 3: (4,4,3,2,1) => 0.62 – 0.62 = 0.0
- This time the sign **does** matter!

University of Stuttgart
Germany

R code for significant differences

Institut für Maschinelle Sprachverarbeitung

```
> c = data.frame("corr1" = c(4,7,1,2,3), "corr2" = c(3,8,2,2,2),
   "total"=c(5,8,3,5,7))
> computeAccDiff <- function(data, indices) {
+    sample <- data[indices,]
+    acc1 <- sum(sample$corr1)/sum(sample$total)
+    acc2 <- sum(sample$corr2)/sum(sample$total)
+    acc1-acc2 }
> b <- boot(c,computeAccDiff,100)
> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates
Intervals :
Level      Normal                Basic
95%   (-0.1486,  0.1515 )   (-0.1379,  0.1897 )
```

University of Stuttgart
Germany

Same code, very different models

```
> c = data.frame("corr1" = c(1,1,1,1,1), "corr2" = c(4,7,2,4,6),
  "total"=c(5,8,3,5,7))
> computeAccDiff <- function(data, indices) {
+    sample <- data[indices,]
+    acc1 <- sum(sample$corr1)/sum(sample$total)
+    acc2 <- sum(sample$corr2)/sum(sample$total)
+    acc1-acc2 }
> b <- boot(c,computeAccDiff,100)
> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 100 bootstrap replicates
Intervals :
Level      Normal                Basic
95%   (-0.7603, -0.5396 )   (-0.8120, -0.5560 )
```

Advanced Aspects

- What types of variables can bootstrapping be applied to?
  - Any, that‘s the beauty of it...
- How many bootstrap samples should I draw?
  - As many as you want. There are O(n!) different samples.
  - Definitely more than 1/p (Drawing the 100th percentile (p=0.01) from a sample of size 20 will not be precise..)
- What information do I have to keep for each bin?
  - You need to compute overall quality: „**sufficient statistics**“
- Advice on picking bins?
  - More bins: better randomization, also more processing effort
  - Bins should be **as independent as possible**
    - Sentences usually good level, unless at discourse level

Questions

(More) Questions on simulation-based significance testing?

University of Stuttgart
Germany

Hypothesis Testing: A Broader View

- Significance testing is a helpful extension to evaluation
- But it has its flaws, too – in particular for Q1 (data analysis)
  - What happens if you make the dataset bigger and bigger?
  - **Even the smallest effects become significant if they occur consistently**
  - Example: two cities, mean height = 1.60m/1.62m, sd=0.20m
    - Sample size 100: t=0.7, p=0.43
    - Sample size 1000: t=2.2, p=0.10
    - Sample size 10000: t=7.0, p=10^{-7}

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

- Is this reasonable?
  - **Yes**. Significance tests **whether some effect can be attributed to chance** or not.

Current discussions

- Journal of Basic and Applied Social Psychology has banned significance testing (as of January 2015)
  - http://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991
- „The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid [...]“
- „No [**inferential** statistics procedures are necessary], because the state of the art remains uncertain. However, BASP will require strong **descriptive** statistics, including **effect sizes**. We also encourage the presentation of **frequency or distributional data** when this is feasible. Finally, we encourage the use of **larger sample sizes** than is typical in much psychology research [...]“

Effect Sizes

- The New Hot Thing (?)
- Quantify the **strength of the relationship** between variables
  - E.g. what part of the **variability of the target variable** can the **experimental manipulation** explain?
- Exist in various instantiations: for correlations, for mean differences (two-sample, multiple=sample, ...)
  - Examples: Cohen‘s d, Hedges‘ g, eta squared, ...
  - Generally range between 0 (no explantory power) and 1 (explains everything)

Effect sizes vs. significances

- It is however not easy to interpret effect sizes (compare inter-annotator agreement)
  - Rule-of-thumbs exist for individual measures
    - E.g. eta squared: 0.02 small, 0.13 medium, 0.26 large
    - Large effect sizes can be uninteresting if the input variable is unrelated to the aims of the study
    - Small effect sizes can be interesting if output variable is very valuable (life expectancy)
- General suggestion: Always report significance and effect size together (Wilkinson 1999)
  - Significant effect but low size: as discussed on slide 37
  - Insignificant effect but large size: potentially big finding

Revisiting the Example from Slide 37

- Example: two cities, mean height = 1.60m/1.62m, sd=0.20m
- What is the effect size of the city variable?
  - In principle, a two-sample t-test setup
  - However, analyses of variance (AOV) in R directly provide eta squared (under the somewhat misleading name „R squared“)
    - Since AOV is a generalization of t-test, let‘s use just that..

```
> library(reshape2)
> d <- data.frame("id"=seq(1,1000),
"city1"=rnorm(1000,mean=1.6,sd=0.2),
"city2"=rnorm(1000,mean=1.62,sd=0.2))
> d <- melt(d,id.vars=c("id"))
> a <- aov(value~variable,data=d)
```

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

R Example (continued)

```
> summary.lm(a)
Residuals:
     Min       1Q   Median       3Q      Max
-0.66241 -0.13281  0.00043  0.13603  0.65272
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.599557   0.006269 255.173   <2e-16 ***
variablecity2 0.019200   0.008865   2.166   0.0304 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1982 on 1998 degrees of freedom
Multiple R-squared:  0.002342,   Adjusted R-squared:  0.001843
F-statistic: 4.691 on 1 and 1998 DF,  p-value: 0.03044
```

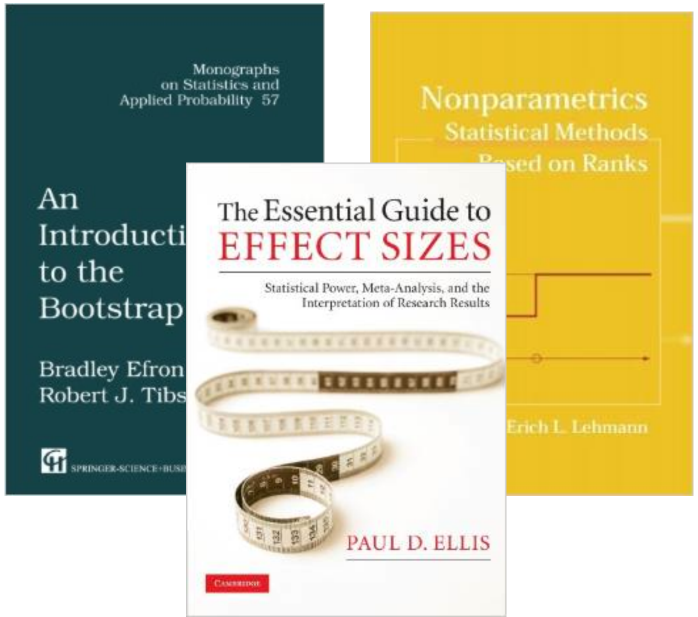
Significance of effect

Effect size

University of Stuttgart
Germany

Institut für Maschinelle Sprachverarbeitung

More Literature



- B. Efron and R. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall 1993.
- Erich Lehmann. Nonparametrics. Springer 2006.
- Paul Ellis. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. CUP 2010.
- Wilkinson, L., & APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 54, 594-604 (2010).
- A. Yeh. More accurate tests for the statistical significance of result differences. Proceedings of COLING 2000.