WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Frequency distribution, cross tabulation, elementary hypothesis testing

Class 3: Marketing Research

Service and Digital Marketing
October 16, 2017

Chapter objectives

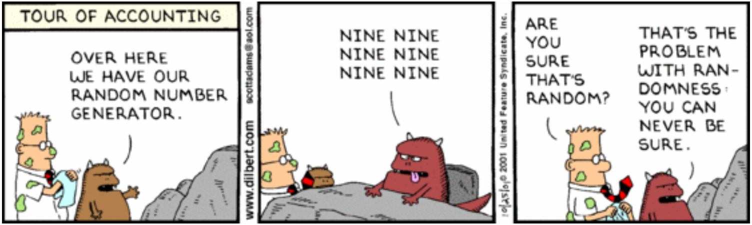
- Create descriptive statistics and graphs
- Calculate means and standard deviations of a distribution of observations
- Conduct χ^2 analyses and tests
- Understand how to use cross tables in practice and be able to interpret the results of different associated statistics

Cyanide and Happiness © Explosm.net

Hypothesis testing

Overview

1. Formulate H_0 & H_1
2. Choose the level of significance α (e.g. $\alpha = 0.05$)
3. Select an appropriate test, based on assumptions about the properties of the underlying data (and grouping variables)
4. Calculate a test statistic T
5. Reject H_0 if $p(T) \leq \alpha$ or do not reject H_0 if $p(T) > \alpha$



Step 1: Formulation of the hypothesis (1)

- A **null hypothesis (H_0)** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An **alternative hypothesis (H_1)** is one in which some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions (= **research question**).

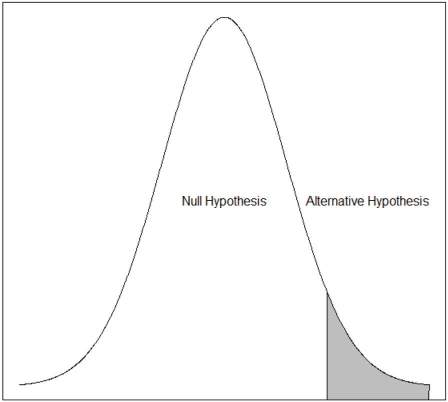
A null hypothesis may be rejected, but it can never be accepted based on a single test. The null hypothesis is formulated in such a way that its rejection leads to the acceptance of the desired conclusion.

Example:

$H_0 = \bar{x}_{female} = \bar{x}_{male}$
$H_1 = \bar{x}_{female} \neq \bar{x}_{male}$

Step 1: Formulation of the hypothesis (2)

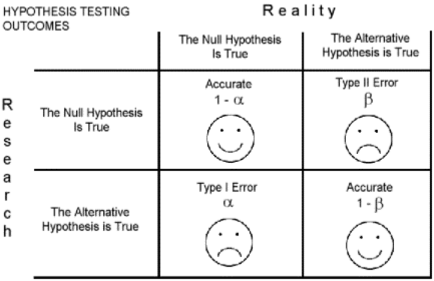
Normal Distribution: $\mu = 0$, $\sigma = 1$



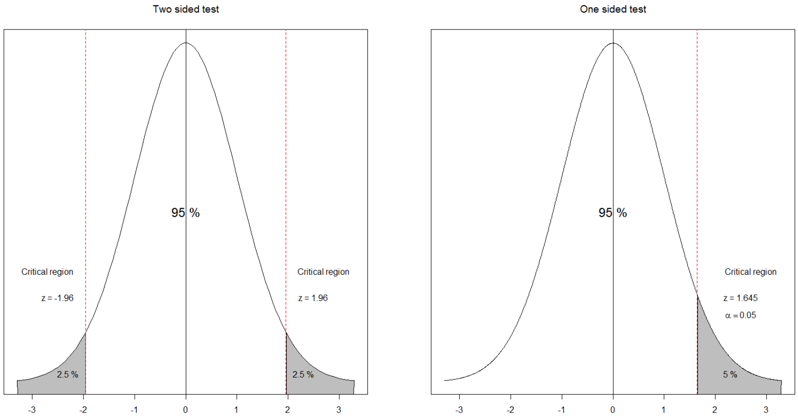
Alternative = the null hypothesis becomes implausible
but when?

Step 2: Choose significance level (1)

- **Type I error** (significance level, α): occurs when the sample results lead to the rejection of the null hypothesis when it is in fact true (common values: 0.05 or 0.01).
- **Type II error** (β): occurs when, based on the sample results, the null hypothesis is not rejected when it is in fact false.
- **Test power** ($1 - \beta$): the probability of rejecting the null hypothesis when it is false.



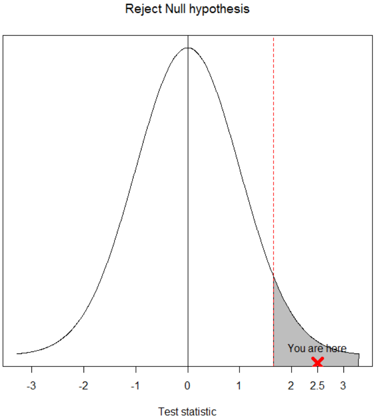
Step 2: Choose significance level (2)

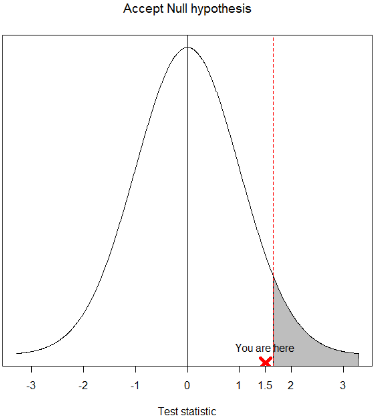


Step 3: Test selection

- select the appropriate test based on the measurement level of the variable being tested and test specific assumptions
  - measurement level of the variables (nominal, ordinal, interval/ratio)
  - **independent** samples (two or more group comparisons on different observation units, also called *unpaired*) or **related** samples (two or more measurement points on the same observation units, also called *paired*)
- the test statistic measures how close the sample is to the null hypothesis and follows a certain distribution (e.g. normal, t, χ^2,...)

Step 4: Accept/Reject





Overview of univariate tests

Measurement level	One sample	Two samples: Independent	Two samples: Dependent
Nominal	binom. test, χ^2, z-test	χ^2	McNemar
Ordinal	KS-test	U-test	Wilcoxon
Interval (or Ratio)	T-test, Z-test	T-test, Z-test	T-test

Measurement level		K-samples: Independent	K-samples: Dependent
Nominal		χ^2	Cochran
Ordinal		Kruskal	Friedman
Interval (or Ratio)		ANOVA	repeated ANOVA

Nominal variables & one sample: χ^2 test

- to compare a certain sample proportion of a **nominal** variable with an expected population proportion
- *Do the number of individuals or objects that fall in each category differ significantly from the number you would expect?*

Case 1: equal proportions

Case 2: χ^2 test of independence

Case 1: χ^2 test for equal proportions (1)

- Tests the null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.

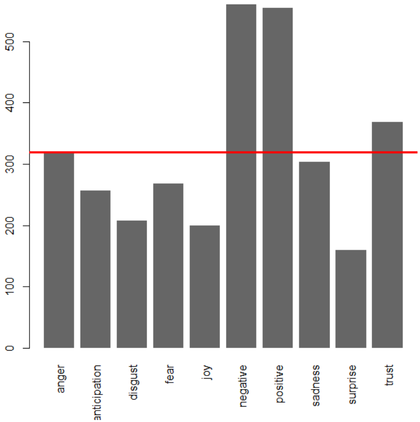
Example: Trump's twitter behavior (cont.)

4901 words obtained from tweets based on Trump's **Android** phone (during the presidential election campaign in 2016) had been categorized into 10 sentiments using the NRC Word-Emotion Association lexicon.

Research question: is the sentiment of words equally distributed? or is there any sentiment that is overrepresented?

H_0: all sentiments (= categories) are equal (uniform distributed)
H_1: at least one sentiment differs (is more/less frequent)

Case 1: χ^2 test for equal proportions (2)



- Under H_0 you would expect a proportion of $1/10$ for all categories $\rightarrow 1/10 \times N = 1/10 \times 3197 = 319.7$ observations per category
- We are testing the deviations of the observed values (o_i) from the expected values (e_i)

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

with $df = no.categories - 1$

$$\chi^2 = \frac{(321 - 319.7)^2 + (256 - 319.7)^2 + \ldots + (369 - 319.7)^2}{319.7} = \frac{175646.1}{319.7} = 549.4091$$

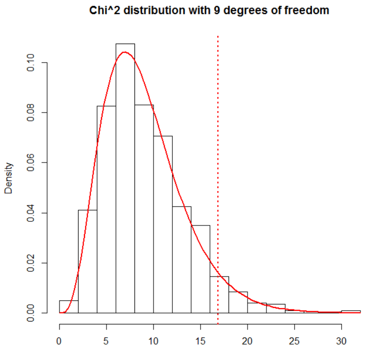
with $df = 10 - 1 = 9$

Case 1: χ^2 test for equal proportions (3)

```
# calculate absolute values (frequencies)
> observed <- table(dat)

# compare it to the expected value for all categories being equally likely
# i.e. the number of observations devided by the number of categories
> expected <- sum(table(dat))/10
> residuum <- observed - expected
> cbind(observed, expected, residuum)

             observed expected residuum
anger             321    319.7      1.3
anticipation      256    319.7    -63.7
disgust           207    319.7   -112.7
fear              268    319.7    -51.7
joy               199    319.7   -120.7
negative          560    319.7    240.3
positive          555    319.7    235.3
sadness           303    319.7    -16.7
surprise          159    319.7   -160.7
trust             369    319.7     49.3

# calculate the test stastistic
> sum(residuum^2)/(expected)
549.4091
```

Case 1: χ^2 test for equal proportions (4)

```
# plot chi-square distribution
> x <- rchisq(1000, 9)
> hist(x, prob=TRUE,
+ main="Chi^2 distribution with 9 degrees of freedom",
+ xlab="")
> curve(dchisq(x, df=9), col="red", lwd=2, add=TRUE )
# the corresponding 95%-quantile
# (= for a significance level of 5%) we would be here
> abline(v=qchisq(1-0.05, 9), col="red", lty="dotted",
+ lwd=2)
# our test statistic is here 549.4091

# calculate area under the chi-square distribution
# curve (= p-value)
> 1 - pchisq(549.4091, 9)
0
```



Case 1: χ^2 test for equal proportions (5)

```
# fast: use built-in function:
> chisq.test(table(dat))

        Chi-squared test for given probabilities

data:  table(dat)
X-squared = 549.4091, df = 9, p-value < 2.2e-16
```

Interpretation:

H_0 (i.e., the 10 sentiments are distributed uniformly) is **rejected** ($p < .001$).
Words with a positve/negative classified sentiment are more frequent!

Case 2: χ^2 test of independence (1)

- to examine the relationship between two independent, nominal variables
- **independence**: the occurrence of one does not affect the probability of the other (= no additional information about one variable can be extracted from the other). Examples: tossing a coin, measuring peoples height,...

two independent variables

	eye-color: blue	eye-color: brown
female	25	25
male	25	25

two dependent variables

	eye-color: blue	eye-color: brown
female	50	0
male	0	50

Note: The direction of the dependency is unclear!

Case 2: χ^2 test of independence (2)

Example: Trump's twitter behaviour (cont.)

Overall, we have 628 tweets from the iPhone, and 762 tweets from the Android. We can also see a difference involves sharing links or pictures in tweets

Research question: Is there an association between content (w/o picture or link) of the tweet and the device used (Samsung Galaxy vs. iPhone)?

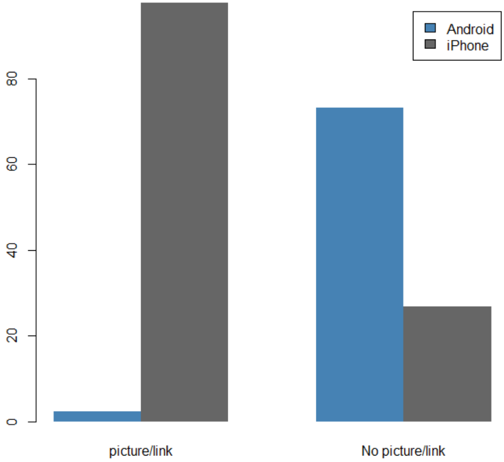
	Picture/link	No picture/link
Android	10	543
iPhone	423	199

H_0: the device used (rows) is independent of the content (columns).
H_1: the device used is not independent of the content.

Case 2: χ^2 test of independence (3)

```
> dat <- matrix(c(10,543,423,199), ncol=2, byrow=T)
> rownames(dat) <- c("Android","iPhone")
> colnames(dat) <- c("picture/link","No picture/link")
> dat
        picture/link No picture/link
Android           10             543
iPhone           423             199

# make a staked barplot
> barplot(dat, col=c("steelblue","grey40"))

# add information (row and column sums)
> addmargins(dat)
# to get conditional information: devide either by row or by column sums
# 1.) conditional on column information
> prop.table(dat, 2)*100

# make a staked barplot again (using conditional information)
> barplot(prop.table(dat, 2)*100, col=c("steelblue","grey40"))
# we have no space for a legend, so we force the bars to be side by side
> barplot(prop.table(dat, 2)*100, col=c("steelblue","grey40"), beside=TRUE,
+ legend=TRUE)
```

Case 2: χ^2 test of independence (4)



Obviously tweets from the iPhone are more likely to contain either a picture or a link.

Case 2: χ^2 test of independence (5)

	pic/link	No pic/link	total
Android	e_{11}	e_{12}	553
iPhone	e_{21}	e_{22}	622
total	433	742	1175

- $e_{11} = 553 \times 443/1175 = 208.4928$
- $e_{12} = 553 \times 742/1175 = 349.2136$
- $e_{21} = 443 \times 622/1175 = 234.5072$
- $e_{22} = 742 \times 622/1175 = 392.7864$

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

with $df = (no.rows - 1) \times (no.columns - 1)$

$$\begin{aligned} \chi^2 &= \frac{(10-208)^2}{208} + \frac{(423-349)^2}{349} + \frac{(543-234)^2}{234} + \frac{(199-392)^2}{392} \\ &= 548 \\ df &= (2-1) \times (2-1) = 1 \end{aligned}$$

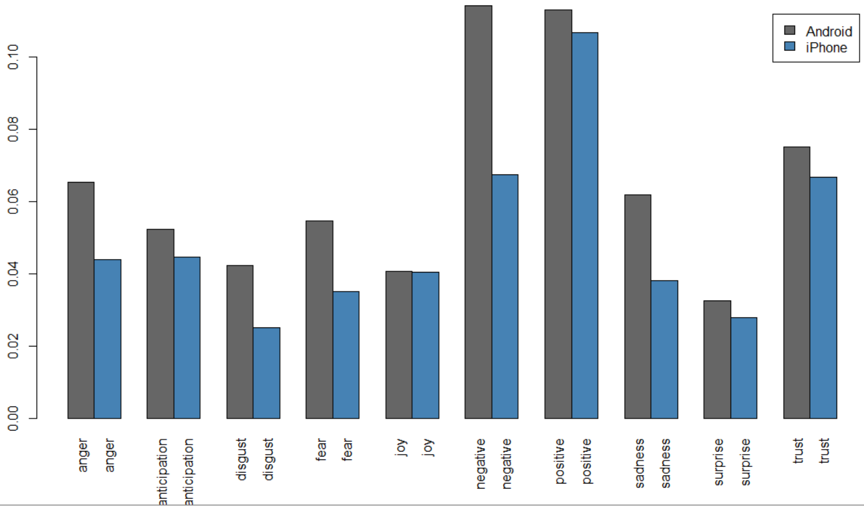
Case 2: χ^2 test of independence (6)

```
> chisq.test(dat)

        Pearson's Chi-squared test with Yates'
        continuity correction

data:  dat
X-squared = 548.41, df = 1, p-value < 2.2e-16
```

Interpretation:

H_0 is **rejected** ($p < .001$).
Tweets from the iPhone are significantly more likely to contain either a picture or a link.

Android vs. iPhone in sentiments



Assignments

See learning platform!

Submission deadline: 23. 10. at 08:00 am

(via the learning platform `www.learn.wu.ac.at`)

Oral presentation of solutions on Monday!

(Recap: random selection of four students to present their solution).

How to report statistical results

1. **Introduction:**
   - describe the managerial background (who should care and why?) and formulate the research question
   - formulate the hypothesis and describe the statistical problem
2. **Statistical method used:** description of the statistical method used and describe why it was used
3. **Report of results and findings:**
   - presentation of the data (tables or graphs)
   - analysis and interpretation of the statistical results
   - managerial implications (in light of the research question)

Degrees of freedom

Describes the number of values in the final calculation of a statistic that are free to vary. In general each parameter being estimated costs a degree of freedom.

For contingencies:
Suppose four numbers (a, b, c and d) that must add up to a total of m: you are free to choose the first three numbers at random, but the fourth must be chosen so that it makes the total equal to m.

a	10
b	12
c	8
d	(restricted)
m	40

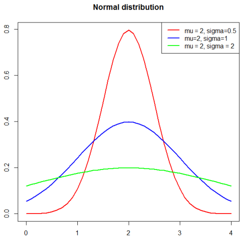
Degrees of freedom are often used to characterize various distributions. E.g. chi-square distribution, t-distribution, F distribution.

Most common test distributions

- Normal distribution:

  defined by μ and σ^2,
  $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$



Test Distributions (2)

- χ^2 distribution:

  defined for $df = k > 0$,
  and $x \geq 0$
  $X \sim \chi^2_k$

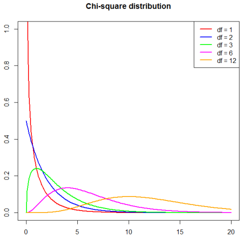
  $$f(x) = \frac{x^{(k/2-1)} exp^{-x/k}}{2^{k/2}\Gamma(k/2)}$$

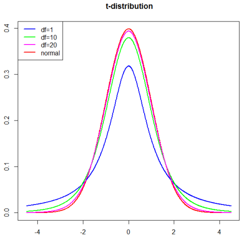
- t-distribution:

  defined for $df = k > 0$,
  $Z \sim t(k)$

  $$f(Z) = \frac{X}{\sqrt{Y/k}}$$

  with $X \sim N(0,1)$ and
  $Y \sim \chi^2_k$





- F-distribution:

  defined for $df_1 = m > 0$
  and $df_2 = n > 0$,
  $Z \sim F(m, n)$

  $$Z = \frac{X/m}{Y/n}$$

  with $X \sim \chi^2_m$ and $Y \sim \chi^2_n$