Research Methods:

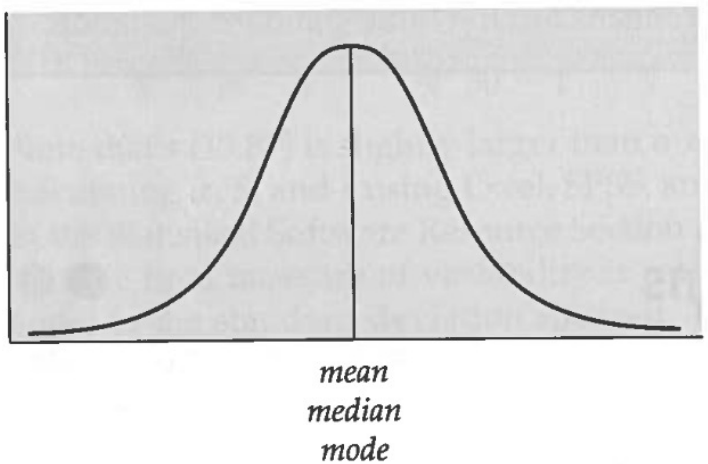
Nonparametric Procedures

Laurens Rook (Delft University of Technology)

Delft University of Technology

Statistics Lectures 1-5

- All the statistics lectures thus far (*z*-score, *z*-test, *t*-test, regression) assumed that the data were normally distributed



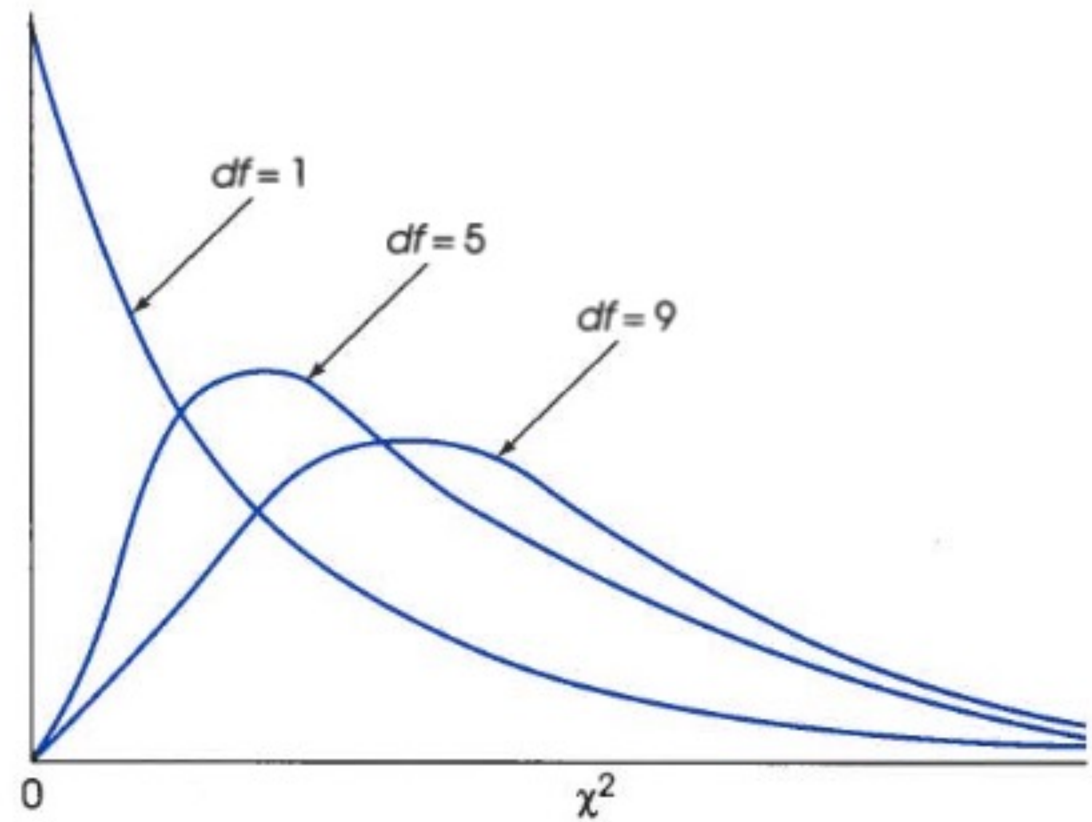
- We even discussed a test (Shapiro-Wilk) to find out if this is the case for a variable under investigation

Chi-square distributions

- What if we have nominal (categorical) data that do not follow a normal a distribution?

- Nonparametric statistical tests: distribution-free tests on proportions or relationships within a population for nominal / categorical data (described with the chi-square test statistic)

Examples



The shape of the chi-square distribution for different values of *df*. As the number of categories increases, the peak (mode) of the distribution has a larger chi-square value.

Nonparametric Procedures

1.

Nonparametric Procedures (Ch. 10, mod. 21)

- Procedures to analyze categorical data that are not normally distributed:

    - The chi-square goodness-of-fit test --> for one nominal variable / category
    - Chi-square test of independence --> for two nominal variables / categories

Chi-square in JASP

- The chi-square goodness-of-fit

  --> we discuss how to **manually** compute and interpret it

- The chi-square test of independence does exist in JASP

  --> we discuss how to **manually** compute or run it in **JASP** and how interpret it

1. Chi-square goodness-of-fit test

- What is it?
- When do you use it?
- How do the H_0 and H_a look like?
- How to calculate & interpret the test output?

What is it?

- The chi-square goodness-of-fit test = a nonparametric procedure to determine how well an observed frequency distribution fits an expected frequency distribution (--> you look at "counts")

  - observed frequencies (the frequency with which participants fall into a category)

  - expected frequencies (the frequency expected in a category if the sample data represent the population)

TU Delft

Assumptions

- The chi-square goodness-of-fit test is:
  - nonparametric (not a standard normal curve)
  - used for **one** nominal (categorical) variable
  - based on a random sample with independent observations
  - does not work for very small frequencies per cell (< 5)

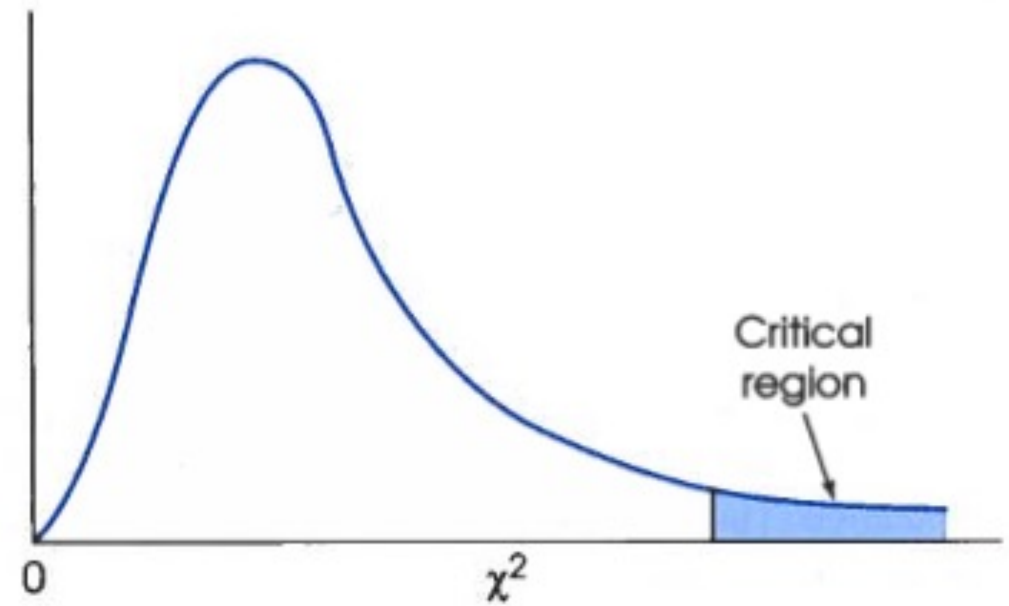
Formulation of the hypothesis

- We compare observed frequencies to expected frequencies:

- $\mathbf{H_0}$ = the observed frequencies do fit the expected frequencies for the population (we have <u>no</u> sign. difference)
- $\mathbf{H_a}$ = the observed frequencies do NOT fit the expected frequencies for the population (we have a sign. difference)

Region of rejection

Chi-square distributions are positively skewed. The critical region is placed in the extreme tail, which reflects large chi-square values.



Calculation of the chi-square

- **For one nominal variable**, we first determine the observed and expected frequencies in a given distribution:

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

$\boldsymbol{\Sigma}$ = the sum (of O and E per category)

O = the observed frequency

E = the expected frequency

Procedure (manual calculation) – I

1. Determine the *E* and *O* for the "positive" occurrence in the category
2. Determine the *E* and *O* for the "negative" occurrence in the category

   - --> build a contingency table with E and O per category (note that this is a 2 x 2 table)

3. Compute the χ^2_{obt} using the formula

Procedure (manual calculation) – II

- Next, we determine the region of rejection by looking up the critical value:

1. Compute the *df* (no. of categories - 1)
2. Look up the associated χ^2_{cv} in Table A6 (book, Appendix A):

If $\chi^2_{obt} \geq \chi^2_{cv}$ --> **reject $\mathbf{H_0}$** (there is significant difference); if $\chi^2_{obt} < \chi^2_{cv}$ --> **accept $\mathbf{H_0}$** (there is no significant difference);

Example

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

- A researcher believes that the percentage of people who exercise in California is greater than the national exercise rate. The national rate is 20%, the researcher gathers a random sample of 120 individuals from California and finds that 31 of them exercise regularly.

- What is the result of the χ^2 goodness-of-fit test, and what can be concluded from it?

TABLE A.6 Critical Values for the χ^2 Distribution

df	.10	.05	.025	.01	.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.992	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997

Extra slide on the degrees of freedom

- Do we have $df = 119$ or $df = 1$ in the example case?

- With parametric tests (that are based on variation and standard deviations) we now would have had sample size -1 = 119

- For this nonparametric test (which looks at the frequencies expected/observed), however, we look at the number of categories (for the 1 nominal variable) – 1 = 1

TU Delft

Interpretation of the chi-square

	Exercise	No Exercise
Observed	31	89
Expected	24	96

The column highlighted is often forgotten in the equation – but should be added for a correct outcome!

$$\chi^2_{obt} = \Sigma \frac{(O-E)^2}{E} = \frac{(31-24)^2}{24} + \frac{(89-96)^2}{96} = \frac{49}{24} + \frac{49}{96} = 2.04 + 0.51 = 2.55$$

degrees of freedom (number of categories $-$ 1) $=$ 2 – 1 $=$ 1; Table A6, Appendix A $-$ $->$ $\chi^2_{cv} = 3.841$

$\chi^2_{obt} < \chi^2_{cv}$ $-->$ H_0 accepted; $\chi^2_{obt} \geq \chi^2_{cv}$ $-->$ H_0 rejected;

We have: $2.55 < 3.841 -->$

H_0 accepted. The percentage of people exercising in California is not significantly greater than the national exercise rate.

2. Chi-square test of independence

- What is it?
- When do you use it?
- How do the H_0 and H_a look like?
- How to calculate & interpret the (JASP) test output?

- --> we test how well an observed frequency distribution **of two nominal variables** fits an expected pattern of frequencies

Assumptions

- The chi-square test of independence is:
  - nonparametric (not a standard normal curve)
  - used for **two** nominal (categorical) variables
  - based on a random sample with independent observations

Calculation of the chi-square test of independence

- **For two nominal variables**, we first determine the observed and expected frequencies in a given distribution:

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

$\boldsymbol{\Sigma}$ = the sum (of O and E per category)

O = the observed frequency

E = the expected frequency

See also Jackson book (Ed. 3), p. 358

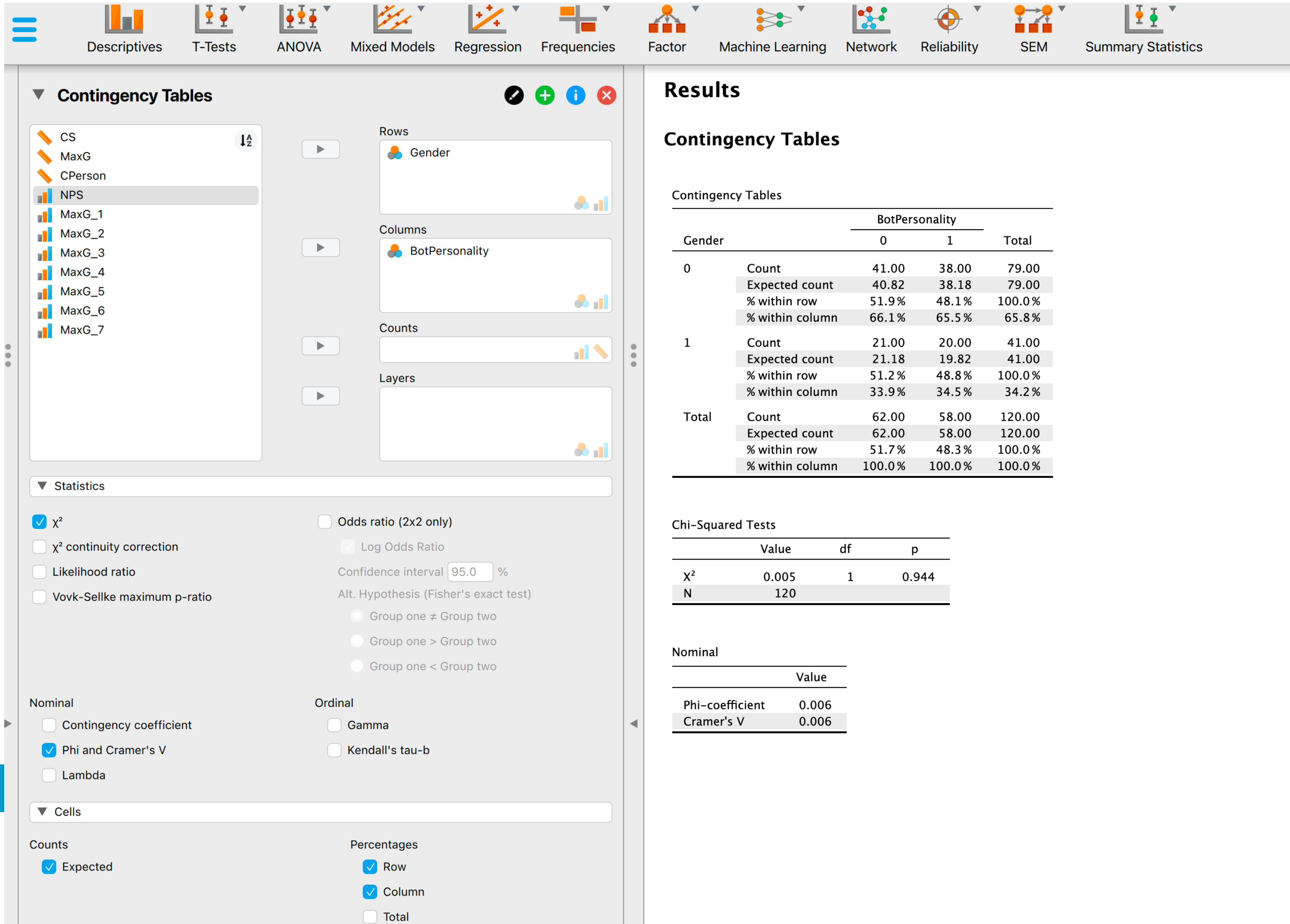
TUDelft

Formulation of the hypothesis

- For the two nominal variables, we compare observed frequencies to expected frequencies:

- $\mathbf{H_0}$ = we have no significant difference between the groups
- $\mathbf{H_a}$ = we have a significant difference between the groups

Example output in JASP



Contingency Tables

		BotPersonality		
Gender		0	1	Total
0	Count	41.00	38.00	79.00
	Expected count	40.82	38.18	79.00
	% within row	51.9 %	48.1 %	100.0 %
	% within column	66.1 %	65.5 %	65.8 %
1	Count	21.00	20.00	41.00
	Expected count	21.18	19.82	41.00
	% within row	51.2 %	48.8 %	100.0 %
	% within column	33.9 %	34.5 %	34.2 %
Total	Count	62.00	58.00	120.00
	Expected count	62.00	58.00	120.00
	% within row	51.7 %	48.3 %	100.0 %
	% within column	100.0 %	100.0 %	100.0 %

Chi-Squared Tests

	Value	df	p
Χ²	0.005	1	0.944
N	120		

Nominal

	Value
Phi-coefficient	0.006
Cramer's V	0.006

Output in Exam --> How to interpret?

Contingency Tables

Contingency Tables

Gender		BotPersonality 0	BotPersonality 1	Total
0	Count	41.00	38.00	79.00
	Expected count	40.82	38.18	79.00
	% within row	51.9%	48.1%	100.0%
	% within column	66.1%	65.5%	65.8%
1	Count	21.00	20.00	41.00
	Expected count	21.18	19.82	41.00
	% within row	51.2%	48.8%	100.0%
	% within column	33.9%	34.5%	34.2%
Total	Count	62.00	58.00	120.00
	Expected count	62.00	58.00	120.00
	% within row	51.7%	48.3%	100.0%
	% within column	100.0%	100.0%	100.0%

Chi-Squared Tests

	Value	df	p
X^2	0.005	1	0.944
N	120		

χ^2 $(1, N = 120) = .005$, $p = .944$ (*ns.*) --> we **accept H_0** that there is no difference between the groups (men and women equally interacted with a high or low personalized chatbot)

Effect size: Phi coefficient

- Phi coefficient = a test to determine the effect size (or strenght) for a chi-square test of a 2 x 2 contingency table

  --> i.e., to what extent the two nominal variables move together (reminiscent of correlation coefficients)

- Interpreting the magnitude (rules of thumb, book):

  - Phi = .10 --> (very) small effect;
  - Phi = .30 --> medium effect
  - Phi = .50 --> large effect

In JASP

Phi = .006 -->
no to (very) small effect

Contingency Tables

Contingency Tables

Gender		BotPersonality 0	BotPersonality 1	Total
0	Count	41.00	38.00	79.00
	Expected count	40.82	38.18	79.00
	% within row	51.9%	48.1%	100.0%
	% within column	66.1%	65.5%	65.8%
1	Count	21.00	20.00	41.00
	Expected count	21.18	19.82	41.00
	% within row	51.2%	48.8%	100.0%
	% within column	33.9%	34.5%	34.2%
Total	Count	62.00	58.00	120.00
	Expected count	62.00	58.00	120.00
	% within row	51.7%	48.3%	100.0%
	% within column	100.0%	100.0%	100.0%

Chi-Squared Tests

	Value	df	p
X^2	0.005	1	0.944
N	120		

Nominal

	Value
Phi-coefficient	0.006
Cramer's V	0.006

Cramer’s V = phi, but it is used when you have a contingency table bigger than the 2 x 2

A note

- Chi-squared tests have one / two categorical variables in the lecture & book examples

- Tests of independence may also be about variables with many (for instance: 5) categories

  - --> i.e., 5 different colors in an M&M bag
  - these variables also have dichotomous outcomes, as you'd assess "color presence in the M&M bag" (yes, no)

TU Delft

The binomial test

2.

The binomial test: A variation with z-score

- **The binomial test has the same assumptions as the chi-square goodness-of-fit test -->**
  - the measurement scale is dichotomous (yes/no, male/female, positive/negative, pass/fail, etc.)
  - test responses are independent
  - sample size is small, but representative of the population (but must be ≥ 10 to justify use of unit normal table for the z critical region)

- Instead of the χ^2 the z-score is used

How is this even possible?

- We are discussing nominal / categorical variables with dichotomous outcomes. How can we use the (parametric) z-score?

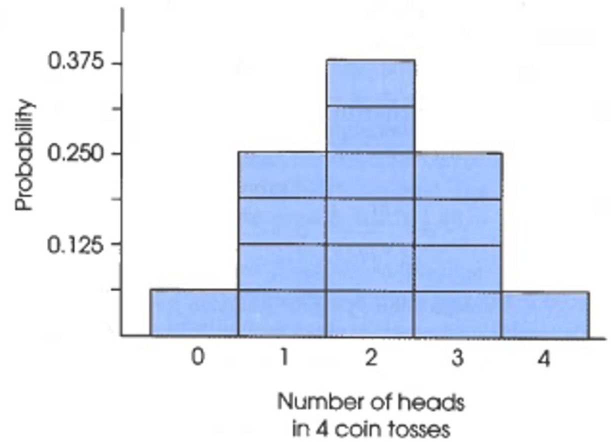
- For the binomial test, you have a sample of *X* individuals, for which you count how often they are classified in category *A* or *B* (--> yes/no, pass/fail, etc.)

- This gives you a distribution of probabilities for a binomial distribution (see next slide)

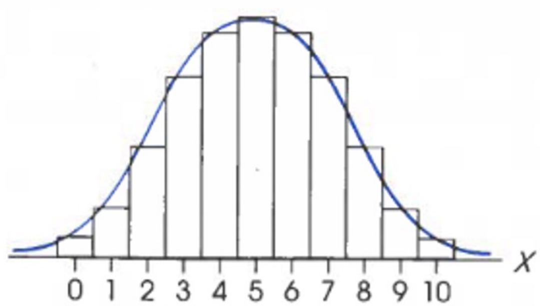
TU Delft

The binomial distribution

A binomial distribution for the number of heads obtained in four tosses of a balanced coin.



The relationship between the binomial distribution and the normal distribution. The binomial distribution is always a discrete histogram, and the normal distribution is a continuous, smooth curve. Each X value is represented by a bar in the histogram or a section of the normal distribution.



The first figure illustrates that a binomial distribution may already begin to mimic a normal distribution after 4 runs.

The second figure nicely visualizes that collecting larger samples of (nonparametric) categorical variables eventually creates a binomial distribution that resembles the normal distribution!

Implications of using the unit normal table

- With a larger (> 10) sample size, you approximate the standard normal curve with your categorical variable --> i.e., you can run the binomial test with the z-score (instead of the chi-squared test)

- It should be noted that you then no longer work with the (observed / expected) frequencies as you did with the chi-squared tests), but with *the **probability** that an outcome may occur* (as with the z-test & t-tests!)

- For a JASP output, this implies that you now can explore the proportion and associated p-value (instead of the frequencies / counts) to evaluate the H_0/H_a!

TU Delft

Example:

- In 2018, Windows had a 86% market share and Apple had a 14% in the UK. Was this also true in the classroom?

- **H_0** = the observed frequencies do fit the expected frequencies for the population (we have no sign. difference)
- **H_a** = the observed frequencies do NOT fit the expected frequencies for the population (we have a sign. difference)

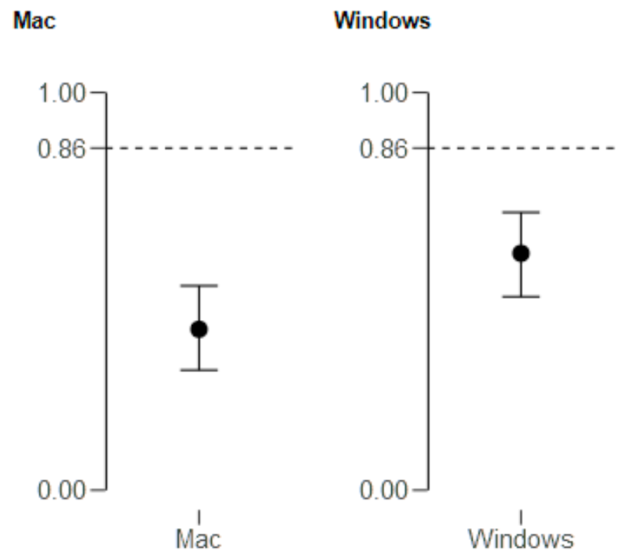
Example from: Goss-Sampson, 2020 (jasp.com)

Results: Testing against 86% (Windows)

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.86.



- For this binomial test, you test, whether the proportion of Windows laptops in the classroom differs from the 86% UK market share (your H_a) or not (your H_0)

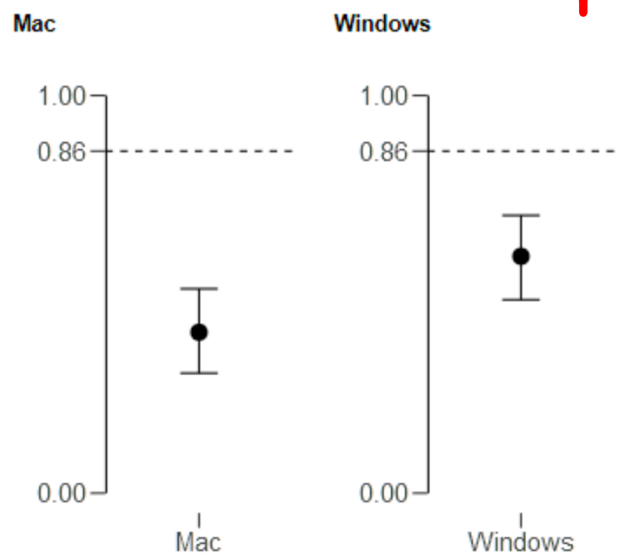
TU Delft

Results: Testing against 86% (Windows)

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.86.



- Students in the classroom used significantly fewer Windows laptops than expected for the UK market share (59.6%, p < .001)

NOTE from the red section that – as you now use the z-score – you look at the proportion & p-value to judge your H_0 and H_a!

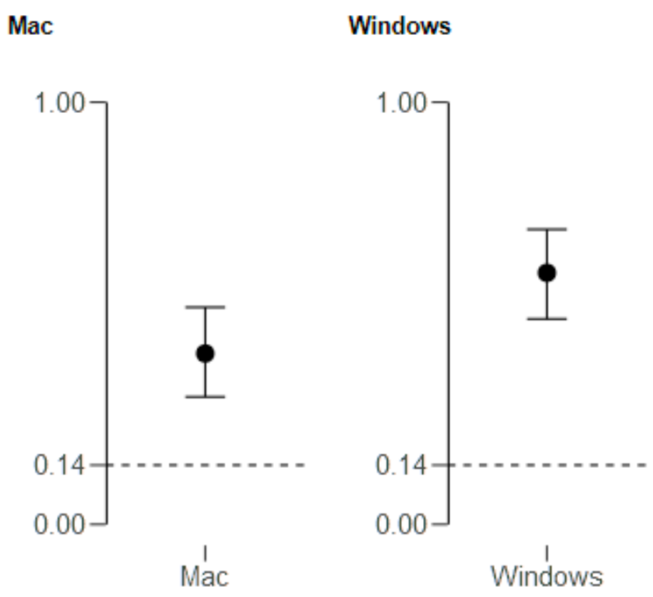
TUDelft

Results: Testing against 14% (Mac)

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.14.



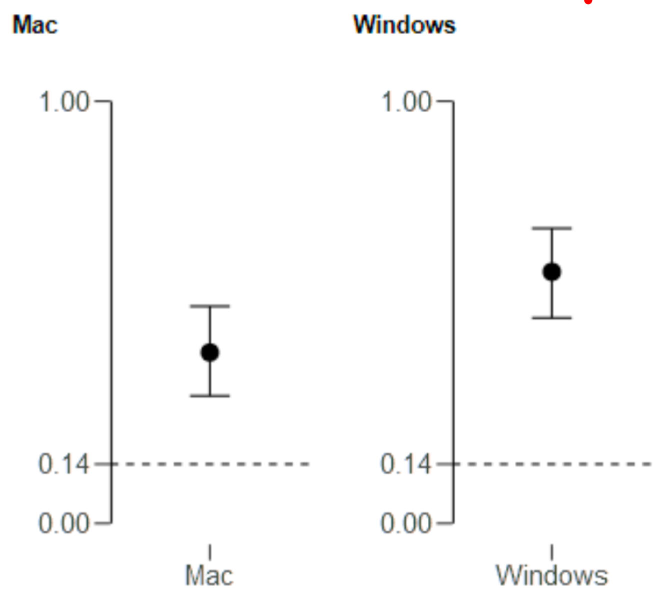
- For this binomial test, you test, whether the proportion of Macbooks in the classroom differs from the 14% UK market share (another H_a) or not (another H_0)

Results: Testing against 14% (Mac)

Binomial Test

	Level	Counts	Total	Proportion	p
Laptop	Mac	36	89	0.404	< .001
	Windows	53	89	0.596	< .001

Note. Proportions tested against value: 0.14.



- Students in the classroom used significantly more MacBooks than expected for the UK market share (40.4%, $p < .001$)

NOTE from the red section that – as you now use the z-score – you look at the proportion & p-value to judge your H_0 and H_a!

TU Delft

Binomial test: Conclusion

- With larger samples, you may obtain a binomial distribution for your categorical outcome(s) that is technically not a normal distribution – but resembles it

- This allows you to switch from chi-squared to z-score --> and test your hypotheses using probability, proportions, and associated p-values instead of frequencies /counts)

- As such, the binomial test behaves like t-test (which also approximates the normal distribution)

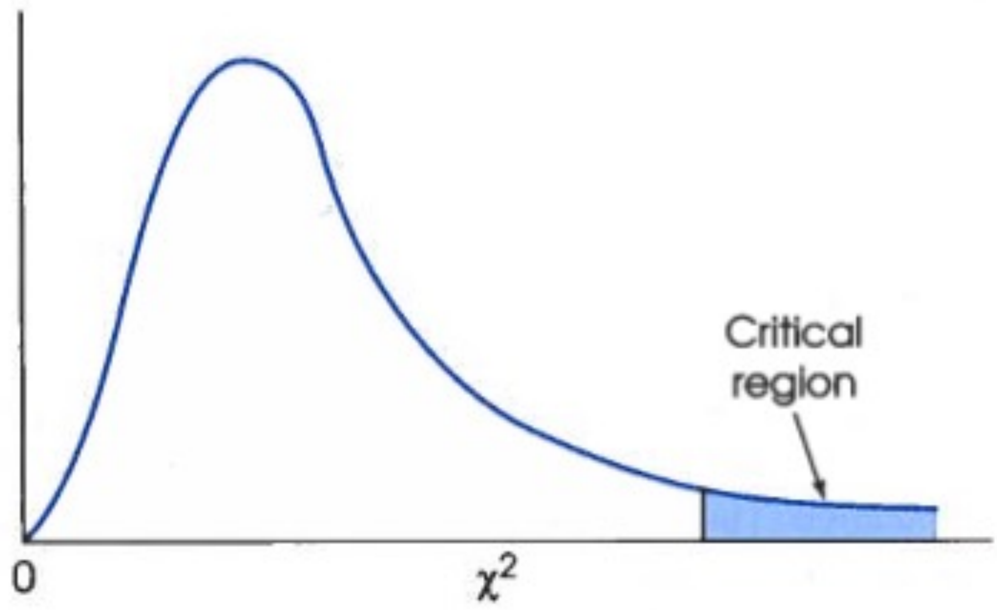
FYI: Multinomial test

- The multinomial test is an extension of the binomial test for **three or more categorical variables**
- Using the same H_0 and H_a formulations as before
- … but with a larger number of descriptives for the categorical variables

In the exam

- You may be asked to interpret the output of a binomial test
- The question to answer always is: Based on the output, do you accept or reject the H_0 for a specified categorical variable?
- Formulate the answer as shown in the Windows-Mac examples

That's all…



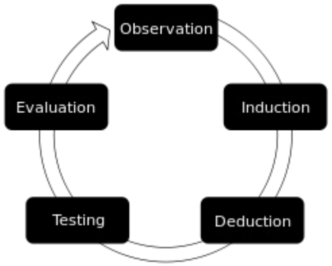
Common mistakes in exams

- Building only half of the contingency table (for O, but not for E)
    - which gives you an incorrect chi-squared outcome
- Mixing up the chi square versions at test

Summary of Statistics Modules

3.

The goals of science



Jackson (Ch. 1, mod. 1) describes 3 basic goals of scientific research:

- Description: via careful observation of behavior
- Prediction: via identification of factors impacting behavior
- Explanation: via causes and mechanisms that explain the when and why of behavior

Experiments

- A causal relation: A → B (with treatment & levels)
  - A_1 (low)
  - A_2 (high)
  - A_3 (none)

- In the laboratory, this brings high internal validity (but also low external validity, because the lab is an artificial environment)

- In the field, this brings higher external validity (real life situation) at the expense of internal validity

TU Delft

Descriptives

- Principles of descriptive statistics:
  - frequency distributions and graphs
  - measures of central tendency and variation
  - lessons learned (from prior exam rounds)
    - **Kurtosis** (how flat/peaked the normal distribution is)
    - **Skewness** (how right-peaked / left-peaked the distribution is ~ proxy for (non)normality)
      - Magnitude $\pm$ 0.00 - $\pm$2.00
      - SW test for normality is possible but cannot be trusted (in small and big data samples)

TU Delft

Probability and Hypothesis Testing

- **NHST** works with two competing hypotheses*:
- H_0 (the null hypothesis): the effect does not exist
- H_a (the alternative hypothesis): an effect (a difference between two groups) exists, and is significant
  - this H_a is your research hypothesis, the statement you wish to support, whereas H_0 is the one you seek to reject

* The examples here apply to two-tailed tests

Z-test (similar for t-test)

- We convert the z-score to the associated proportion (Table A1, Appendix A) to get the ***z*** **value obtained**

- We compare the ***z*** **critical value** (that marks the edge of the region of rejection in a sampling distribution) with the ***z*** **value obtained**

- **Critical value =** the edge of the region of rejection in a sampling distribution. Values equal or beyond it fall in the region of rejection for the H_0

Z-test and t-test

- Theoretical differences for z-score, z-test, t-test
- Theoretical assumptions underlying each application (when / for what type of measures)
- Theoretical concepts (Type I & II error, statistical power)

Correlation

- You learned, practically, to identify strong, moderate, and weak correlation coefficients, and their valence

    - Pearson's r
    - Spearman's rank-order correlation coefficient

- Understand, theoretically, how correlations allows for prediction, but not for making cause-and-effect claims

Linear & multiple regression

- They are advanced correlational techniques using the H_0 vs. H_a logic

- You test the steepness of slope *b* of your regression equation --> and interpret results from the coefficient output table in JASP

- The stronger the relationships between the variables in your data, the more accurate your prediction --> as indicated in the model output in JASP)

TU Delft

Nonparametric Procedures

- Procedures to analyze data that are not normally distributed:
- We discussed chi-square tests that apply to comparisons of one or two nominal (categorical) variables
    - The chi-square goodness-of-fit test --> for one nominal variable / category
    - Chi-square test of independence --> for two nominal variables / categories
- We looked at **the binomial test** as a way to test hypotheses for categorical variables in larger samples based on the z-score

These stats in exam form

- You could be asked to compute a simple z-score, chi-squared, z-test or t-test (with conversion table provided)
- You could be asked to interpret a JASP output:
  - of descriptives, *z*-test/*t*-test, correlation, regression analysis, chi-square or binomial test
  - be provided with a specification of H_0 and H_a
  - and be asked whether H_0 and H_a must be accepted / rejected based on this

Good luck!