VIII. Statistical Tests (Part I)

- ▶ Introduction to Statistical Testing
- ▶ Testing with the Binomial Distribution

Statistical Testing

- ▶ How to establish scientific knowledge based on empirical observation
- ▶ How to draw a **conclusion** from a **sample** that is valid for the **population**
- ▶ With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone.
- ▶ Hypothesis testing can be used to determine whether a statement (our assumption) about the value of a population parameter should or should not be rejected.

Introduction to Statistical Testing (2/6)

Statistical Testing

- ▶ We begin by making a tentative assumption about a population parameter - the **null hypothesis**.
- ▶ Next we define the **alternative hypothesis**, which is the opposite of what is stated in the null hypothesis.
- ▶ Then we define a value for α. Usually $\alpha = 5\%$ or $\alpha = 1\%$. This is our risk for a type I error (see later).

Introduction to Statistical Testing (3/6)

Statistical Testing

- ▶ We perform a **statistical test** that yields a **test statistic** which (under the null hypothesis) follows a certain **distribution**.
- ▶ Then we decide if we **accept** or **reject** our **null hypothesis**:
  - ▶ If we calculate the test statistic by hand we use tables that list the critical values for rejection of the null hypothesis. Since it is very common to use $\alpha = 5\%$ or $\alpha = 1\%$, we find these critical values in the table.
  - ▶ If we use software, we usually get the (exact) **p-value**. We reject the null hypothesis, if this p-value is less than α.
- ▶ There is usually a decision that follows the acception or the rejection of the null hypothesis.

Be careful!

- ▶ Whatever your result is, it never proves the hypothesis.
- ▶ Pay attention to the descriptive measures as well to avoid pitfalls (especially when using software).
- ▶ P-value misunderstandings are very common.

Interpreting a p-value

The **p-value** of a test is the probability of obtaining an effect at least as extreme as the one you observed in your sample data, assuming the truth of the null hypothesis.

To put this in other words: If the null hypothesis is true, the **p-value** is the probability of obtaining your sample data. It answers the question if your sample data is unusual if the null hypothesis is true. If it is too unusual, we reject the null hypothesis.

Type I and Type II error

	H0 true in population	**H0 wrong** in population
Test result: **accept H0**	Correct conclusion!	Type II Error
Test result: **reject H0**	Type I Error	Correct conclusion!

Type I or Type II error - which one is worse?

There is no general rule. It always depends on the context and on the type of problem you analyze!

We are already familiar with the binomial distribution, a discrete probability distribution that gives us the probability for exactly k successes if we run an experiment n times:

Binomial Distribution

$$P(k \text{ successes of n trials}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

for $k \in \{0, 1, \cdots, n\}$

The binomial test is used when an experiment has two possible outcomes and you have an idea about what the probability of success is. A binomial test is run to see if observed test results differ from what was expected.

Testing with the Binomial Distribution (2/5)

1. Pick a value for α (common values are 5% or 1%) **before** your test and formulate your hypotheses.
2. Determine if the **left** or the **right** tail of the binomial distribution is your rejection region for H_0.
3. In your experiment, you had exactly k successes out of n trials.
4. If your rejection region is on the **right** hand side of the binomial distribution, reject H_0 if $\sum_{i=k}^{n} P(X = i) \leq \alpha$
5. If your rejection region is on the **left** hand side of the binomial distribution, reject H_0 if $\sum_{i=0}^{k} P(X = i) \leq \alpha$
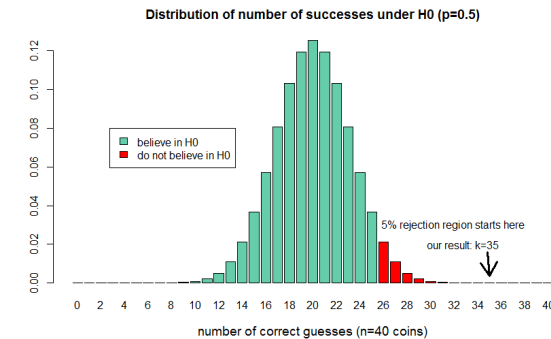6. Otherwise, do not reject H0.

Testing with the Binomial Distribution (3/5)

Example: Extrasensory Perception

In an experiment on extrasensory perception (ESP), an individual is asked to state the result of (head or tail) of tossing an unbiased coin. If the individual states 35 times out of 40 results correctly, determine whether the results are significant at a 5% level.

1. $\alpha = 5\%$. H_0 : P(success) $\approx$ 50%, H_1 : P(success) > 50%
2. If k is the number of **sucesses**, clearly the **right tail** of the distribution is of interest.

Distribution of number of successes under H0 (p=0.5)

- believe in H0
- do not believe in H0

5% rejection region starts here
our result: k=35

number of correct guesses (n=40 coins)

Testing with the Binomial Distribution (4/5)

Example: Extrasensory Perception (contd.)

3. We counted k=35 successes.
4. We calculate $P(X = 35) + P(X = 36) + \cdots P(X = 40)$. Using Excel, we get a **p-value** of **0.000000691306**
5. (We are not interested in the left hand side of the distribution. We do not want to reject H0 if our candidate is extremely untalented.)
6. Reject H_0 because 0.000000691306 < 0.05

Testing with the Binomial Distribution (5/5)

This is what inferential statistics is about.

Distribution of number of successes under H0 (p=0.5)

- believe in H0
- do not believe in H0

5% rejection region starts here
our result: k=35

number of correct guesses (n=40 coins)

If your engine revolutions are too high, there is going to be **something wrong** with your engine - it can get seriously damaged. *We should therefore change gears (or slow down a bit)*
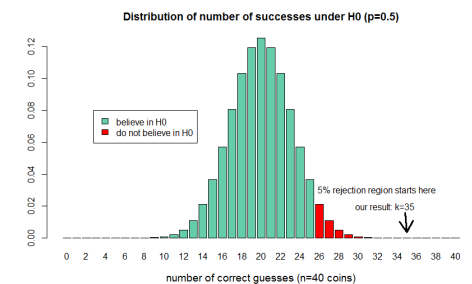
If your test statistic happens to be in the rejection area of the distribution, there is probably **something wrong** with your H_0. *We should therefore reject it.*

IX. Statistical Tests (Part II)

- χ^2 Tests for Frequencies
- χ^2 Test for Contingency Tables

χ^2 Test for Frequencies (1/6)

The results obtained in samples do not always agree exactly with the theoretical results expected according to the rules of probability. For example, although theoretical considerations lead us to expect 50 heads and 50 tails when we toss a fair coin 100 times, we are not surprised if we get 47 tails and 53 heads or similar results. Given n events (E_1, E_2, $\cdots$, E_n) we can summarise the results from an experiment as follows:

Event	E_1	E_2	$\cdots$	E_n
Observed Frequency	o_1	o_2	$\cdots$	o_n
Expected Frequency	e_1	e_2	$\cdots$	e_n

The expected frequencies can be calculated using the probabilities for each event and the sample size. We can now calculate a *measure of discrepancy* between the observed and the expected values. This measure is called a χ^2-statistic.

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

Note that always $\chi^2 \geq 0$ (and $\chi^2 = 0$ if and only if all observed frequencies agree **exactly** with the expected frequencies). The larger the value of χ^2, the more discrepancy between the observed and the expected frequencies.

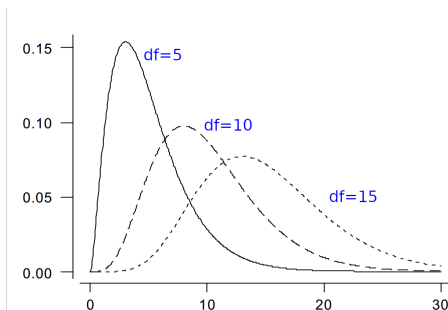
χ^2 Test for Frequencies (2/6)

Under H_0, the χ^2 statistic follows the χ^2 sampling distribution with n-1 degrees of freedom.

Mathematical Background: χ^2-distribution

- The area under the curve is equal to 1
- The curve starts at zero and indefinitely extends to the right (never touches the x-axis)
- The parameter **df** (the degrees of freedom) determines the shape of the curve. For our problem, df=n-1.

χ^2 Test for Frequencies (3/6)

The χ^2 Test Procedure

1. Determine α (e.g. 5% or 1%) and state the null and alternative hypothesis
2. H_0 states that there is **no** difference between the the distributions, H_1 states that there is a difference.
3. Calculate the empirical χ^2-value
4. df=n-1 (where n is the numer of categories)

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

5. Make your decision for or against H_0:

Critical value decision rule
(You need a table)

- if $\chi^2_{\text{empirical}} > \chi^2_{\text{critical}}$ **reject** H_0
- if $\chi^2_{\text{empirical}} \leq \chi^2_{\text{critical}}$ **do not reject** H_0

Decision rule based on software output

- if the p-value $< \alpha$ **reject** H_0
- if the p-value $\geq \alpha$ **do not reject** H_0

χ^2 Test for Frequencies (4/6)

EXCEL-Functions

=CHISQ.TEST(observed; expected)	performs a χ^2 Test, returns the p-value.
=CHISQ.INV(probability; df)	Returns the inverse of the left-tailed probability of the χ^2distribution.
=CHISQ.DIST(x; df; TRUE)	Returns the area under the curve from 0 to x.

χ^2 Test for Frequencies (5/6): Example

Example 1: Safety at Work
You collected data about industrial injuries over the past 10 years:

day of the week	Mon.	Tues	Wed.	Thurs.	Fri
Number of accidents	135	108	120	114	146

Test the hypothesis that the number of accidents does not depend on the day of the week (H_0) using a significance level of $\alpha = 5\%$.

χ^2 Test for Frequencies (6/6): Example (continued)

Example 1: Safety at Work (contd.)
At first we calculate the expected probabilities:

day of the week	Mon.	Tues	Wed.	Thurs.	Fri	SUM
Number of accidents	135	108	120	114	146	**623**
expected under H_0	124.6	124.6	124.6	124.6	124.6	**623**

Now we calculate the test statistic:

$$\chi^2 = \frac{(135 - 124.6)^2}{124.6} + \frac{(108 - 124.6)^2}{124.6} + \cdots + \frac{(146 - 124.6)^2}{124.6} \approx 7,83$$

In the table ($\alpha = 5\%$, df=4) we find the critical value 9,488. Since 9,448 > 7.83 we do not reject H_0. The number of accidents does not depend on the day of the week.

χ^2 Test for Contingency Tables (1/4)

The χ^2 Test Procedure for Tables

1. Determine α (e.g. 5% or 1%) and state the null and alternative hypothesis.
2. Calculate sums for the rows and columns and calculate the table with the expected values.
3. H_0 states that there is **no** difference between the the distributions, H_1 states that there is a difference.
4. Calculate the empirical χ^2-value using the table with the **observed** and the **expected frequncies**:

> If our tables (for the observed and the expected frequencies) have n rows and m columns we calculate the following test statistic:
>
> $$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

5. If your table has n rows and m columns, df $= (n - 1) \cdot (m - 1)$
6. Make your decision for or against H_0 using the χ^2-table.

Example: Sex and Political Views

Consider the following tabulated results depicting the results of a survey:

sex/party	Democrat	Independent	Republican
F	762	327	468
M	484	239	477

(Example from Agresti, A. (2007). An Introduction to Categorical Data Analysis, 2nd ed., New York: John Wiley & Sons, page 38.)

One might come up with the following question:

Is there a relationship between the variables **sex** and **political views**?

Solution: Sex and Political Views (contd.)

1. $\alpha = 5\%$

2. We calculate row and columnsums first:

sex/party	Democrat	Independent	Republican	sums
F	762	327	468	**1557**
M	484	239	477	**1200**
sums	**1246**	**566**	**945**	**N=2757**

and now we caluclate the expected frequencies using these sums:

sex/party	Democrat	Independent	Republican	sums
F	703.6714	319.6453	533.6834	**1557**
M	542.3286	246.3547	411.3166	**1200**
sums	**1246**	**566**	**945**	**N=2757**

Hint: For each cell, multiply the corresponding rowsum with the corresponding columnsum and divide by the sample size, e.g. $\frac{1246 \cdot 1557}{2757} = 703.6714$

Solution: sex and political views (continued)

3. H_0 states there is no relationship between sex and political views, H_1 states there is a relationship.

4. Now we calculate our measure of discrepancy:

$$\chi^2 = \frac{(762 - 703.6714)^2}{703.6714} + \frac{(327 - 319.6453)^2}{319.6453} + \cdots + \frac{(477 - 411.3166)^2}{411.3166} \approx 30.07$$

5. Our parameter $df = (2 - 1) \cdot (3 - 1) = 2$ since we have 2 rows and 3 columns.

6. The critical value (df=2, 5%) = 5.99. We reject H0, because 30.97 > 5.99.