Testing a hypothesis

Testing hypotheses

- This module investigates the use of data science and statistics to answer questions about the process that generated data we observe.
- Key concepts in this module:
  - Formulation of a **hypothesis**
  - Evaluating the evidence against the hypothesis: a **hypothesis test** based on a **test statistic**
  - This is compared to a **reference distribution** to determine how unusual the outcome is

The general testing of hypothesis framework

The scientific method

- Formulate a hypothesis
- Test it on the basis of data

What does it mean to test a hypothesis?

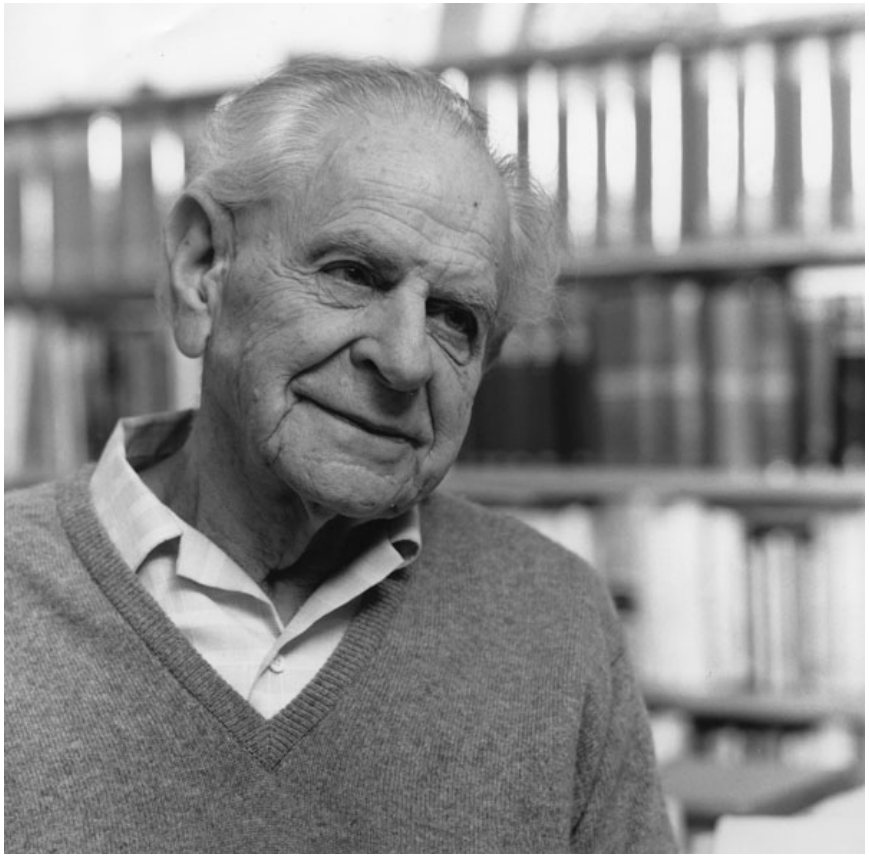
Can the data prove it true? No, but a test can prove the hypothesis to be false.

Hume and the problem of induction

David Hume (1711 – 1776)

- The problem of induction: no matter how many instances of white swans we might have observed, this does not justify the conclusion that all swans are white.
- Natural instinct, rather than reason, explains the human practice of making inductive inferences.
- How can we learn from data then?

Deductive method of testing



Karl Popper (1902-1994)

Predictions are deduced from the theory... and compared with the results of practical applications and experiments... they can be *falsified*

The Logic of Scientific Discovery, 1934

The scientific method in practice

- The hypothesis being tested is referred to as "null hypothesis," to connote that it is the "default," status quo understanding of how things work.
- The **complement** (opposite) of the null – i.e. the state of nature that would prevail if the null hypothesis were false – is referred to as the **alternative** hypothesis (though in some cases we specify a more restrictive set of alternatives).
- If our test statistic is very unlikely to be observed under the null, then this is seen as evidence against the null and we **reject the null**.
- When we reject the null we make a discovery: we find out that the current understanding of the world is not sufficient.

Distinguishing Coke and Pepsi

- It has been suggested that the distinction between Coke and Pepsi rests to a large amount on the visual cue of the container, so that based on taste alone the two a essentially indistinguishable.
- This is a hypothesis that can be tested!
- The first task is to design an experiment. How would we test this hypothesis?

Coke vs. Pepsi

- The null hypothesis ("nothing extraordinary is going on") is that the taster cannot distinguish Coke and Pepsi and is therefore just guessing.
- Suppose we have people taste 10 cups that are filled at random with Coke or Pepsi. They get 7 out of 10 right on average. What can we conclude from this experiment?
- This is the second task: evaluate your test result in the context of a *reference distribution*.
- The key point is: If the null hypothesis is true and the taster is just guessing, then it's possible to compute the chances of getting 7 or more correct. If those chances are small, then this is seen as evidence against the null hypothesis and we reject the null.

The logic behind testing hypotheses, a different example:

- A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.
- In this case "nothing extraordinaty is going on" means that the drug has no effect, so that's our null hypothesis H_0.
- Note that in this case the company would like to reject H_0!
- So the logic of testing is often indirect: One assumes that nothing extraordinary is happening and then hopes to reject this null hypothesis H_0.

Setting up a test statistic

- A **test statistic** measures how far away the data are from what we would expect if H_0 were true.
- The most common test statistic is the **z-statistic**

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

- 'Observed' is the outcome of an appropriate statistic for this situation. In the Coke-Pepsi example an appropriate statistic would be the number of correct answers, or the percentage of correct answers.
- 'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that* H_0 is true.
- In the Coke-Pepsi example, the expected number of correct answers under H_0 (guessing) is 5, the SE is computed to be 1.58, so

$$z = \frac{7-5}{1.58} = 1.27$$

p-values measure the evidence against H_0

- Large values of the test-statistic are evidence against the null.

- The strength of the evidence is measured by the **p-value** (or: **observed significance level**):
- The p-value are the chances of getting a value of z as extreme or more extreme than the observed z, assuming H_0 is true.
- Note that if H_0 is true, then z follows a normal curve by the CLT. This is our reference distribution (**null distribution**). Using this normal curve as null distribution we can find the p-value: 20.4 %.
- The smaller the p-value, the stronger the evidence against H_0. Typically the criterion for rejecting H_0 is a p-value smaller than 5%. Then the result is called **statistically significant**.

More about p-values

- Note that the p-value **does not** give the chances that H_0 is true, as H_0 is either true or not - there are no chances involved. Rather, it gives the chances of seeing a statistic as extreme, or more extreme, than the observed one, assuming H_0 is true.
- Rejecting for large values of $|z|$, i.e. if z is large or if z is small, is a **two-sided test**. But in the Coke-Pepsi example, we might only want to reject if z is large, i.e. there are many correct answers. This is a **one-sided test**. The p-value will be different! (10.2%)
- Whether to use one-sided or two-sided tests depends on the question at hand. In the Coke-Pepsi experiment, a small z (i.e. few correct answers) is actually evidence that the taster can distinguish Coke and Pepsi, but mixes up the two brands! So a two-sided test would be appropriate if you want to allow for this possibility.

More on testing

There are two ways that a test can result in a wrong decision:

H_0 is true, but is erroneously rejected: Type I error ("false positive")

H_0 is false, but we fail to reject it: Type II error

Rejecting H_0 if the p-value is smaller than 5% means

P(type I error)$\leq$ 5%.

Testing hypotheses - Recap

Remember the basics:

- The null hypothesis describes the situation where "nothing special is happening".
- It's not possible to prove that a hypothesis is correct, but it's possible to disprove it.
- Given data, we can assess whether the data are compatible with the null hypothesis. If not, we reject the null hypothesis.
- This is done via a test statistic (which summarizes the information in the data and measures how far away from the null situation the data are) and a reference distribution which describes the behavior of the test statistic in the null case.
- If the test statistic is far out in the reference distribution (namely: there is less than a 5% chance of getting such an extreme outcome), then we reject the null hypothesis.

Constructing the reference distribution

The null distribution describes how the test statistic behaves under the null hypothesis. Think about it as a histogram of many replicates of the test statistic obtained from many samples under the null hypothesis.

Sometimes one can actually find the null distribution this way: In the Coke-Pepsi experiment, 10 guesses can be easily simulated.

Often that's not possible. Then there are three typical ways of finding an (approximate) null distribution:

- The **CLT** may apply - approximate with Normal distribution.
- We may approximate the null distribution with the **bootstrap** if it's possible to include the null hypothesis into the bootstrap sampling (see homework).
- It **permutation test** may be applicable: The null hypothesis may justify scrambling certain aspects of the data. Each scramble will produce a reference sample.

Contingency tables

Suppose several categorical variables are recorded for each subject, e.g. marital status and sex. Then the data are often tabulated in a **contingency table**:

```
cross
```

```
  Div/Sep Married Never Married Widowed
F      94     157           111      78
M      61     359           122      18
```

Contingency tables

- A contingency table shows the number of observations for each combination of the possible values of these two categorical variables
- Adding up all of the numbers gives the number of observations

```
sum(cross)
```

```
[1] 1000
```

- The sums by row and column give the **marginal counts**.

```
  F   M
440 560
```

```
      Div/Sep       Married Never Married       Widowed
          155           516           233            96
```

Contingency tables

```
cross
```

	Div/Sep	Married	Never Married	Widowed
F	94	157	111	78
M	61	359	122	18

-**Note**: this is just a sample, not the entire population

-We are interested in whether marital status and sex are related.

- To come up with a summary that addresses this question, it is useful to start thinking about what we would do if we had the entire population (not just a sample of 1000 subjects)

A population model for contingency tables

We have two random variables: Marital Status, Sex

One variable can take on the values *Divorced/Separated*, *Married*, *Never Married*, *Widowed* and the other the values *F*,*M*

	Div/Sep	Married	Never Married	Widowed	
F	π_{11}	π_{12}	π_{13}	π_{14}	q_1
M	π_{21}	π_{22}	π_{23}	π_{24}	q_2
	p_1	p_2	p_3	p_4	1

- π_{11} is the **population (relative) frequency** of Female and Divorced/separated
- q_1 is the population frequency of Female
- p_1 is the population frequency of Divorced/separated

What does it mean for Marital Status to be independent of Sex?

A basic fact about probabilities

We say that the random variables X and Y are **independent** if

$$\pi_{jk} = P[X = x_j, Y = y_k] = P[X = x_j]P[Y = y_k] = p_j q_k$$

where $P[X = j, Y = k]$ is the probability that $X = x_j$ **and** $Y = y_k$.

An aside: What does independence mean?

	Div/Sep	Married	Never Married	Widowed	
F	π_{11}	π_{12}	π_{13}	π_{14}	q_1
M	π_{21}	π_{22}	π_{23}	π_{24}	q_2
	p_1	p_2	p_3	p_4	1

- The proportion of females that are Divorced/Separated is the same of the proportion of males that are Divorced/Separated
- It is the same of the overall population proportion of Divorced/Separated

This can be checked using the formula on the previous slide.

Back to our data

```
cross
```

```
   Div/Sep Married Never Married Widowed
F       94     157           111      78
M       61     359           122      18
```

- This table represents the outcomes of measuring `sex` and `marital.stat` for 1000 subjects
- We can think of each subject as an experiment with two random outcomes: sex and marital status of the subject.
- Are the two outcomes independent?

Relative frequencies

From our contingency table, we can compute the relative frequency of each possible outcome:

```
n.obs <- sum(cross)
pi.hat <- cross/sum(cross)
pi.hat
```

```
  Div/Sep Married Never Married Widowed
F   0.094   0.157         0.111   0.078
M   0.061   0.359         0.122   0.018
```

We can think of these as **estimates** $\hat{\pi}_{jk}$ of the probabilities $\pi_{jk} = P[X = x_j, Y = y_k]$, where X represents marital status and Y represents sex.

Marginal relative frequencies

We can also compute the **marginal** relative frequencies

```
      Div/Sep        Married Never Married        Widowed
        0.155          0.516          0.233          0.096

   F    M
0.44 0.56
```

- We can think of these as **estimates** $\hat{p}_j$, $\hat{q}_k$ of the **marginal probabilities** $p_j = P[X = x_j]$ and $q_k = P[Y = y_k]$
- These are just estimates because we only have a **sample** of all people, we don't have data on the entire human population
- If sex and marital status **were independent** then because

$$\pi_{jk} = p_j q_k$$

we would expect:

$$\hat{\pi}_{jk} \approx \hat{p}_j \hat{q}_k$$

Expected counts under independence

With this logic we can calculate the **expected counts** – the contingency table for sex and marital status that we would expect to observe under independence, assuming $\hat{p} = p$ and $\hat{q} = q$

```
independence_prob = matrix(0, nrow=2, ncol=4)
for (j in 1:2) {
    for (k in 1:4) {
        independence_prob[j,k] = q.hat[j] * p.hat[k]
    }
}
expected_counts = n.obs * independence_prob
round(expected_counts)
```

```
     [,1] [,2] [,3] [,4]
[1,]   68  227  103   42
[2,]   87  289  130   54
```

Evidence against independence

In R, one common measure of discrepency can be easily computed:

```
chisq.test(cross, simulate.p.value=TRUE)
```

```
    Pearson's Chi-squared test with simulated p-value (base
    replicates)

data:  cross
X-squared = 111.33, df = NA, p-value = 0.0004998
```

It uses Pearson's X^2 as a measure of discrepency

$$\chi^2(O, E) = \sum_i \sum_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}}$$

Evidence against independence

Alternatively one can compute a p-value based on theoretical approximations (lots of the work of early statisticians has been in figuring out what the p-values would be without having the power of computing.)

```
chisq.test(cross, simulate.p.value=FALSE)
```

```
    Pearson's Chi-squared test

data:  cross
X-squared = 111.33, df = 3, p-value < 2.2e-16
```