# *Experimental Methods in Computer Science*

**(*Metodologias Experimentais em Informática*)**

**Henrique Madeira**

**Master in Informatics Engineering**
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia da Universidade de Coimbra
2023/2024





1

## *Hypothesis Testing*

## *Test for a proportion*





2

## Inferences for proportion

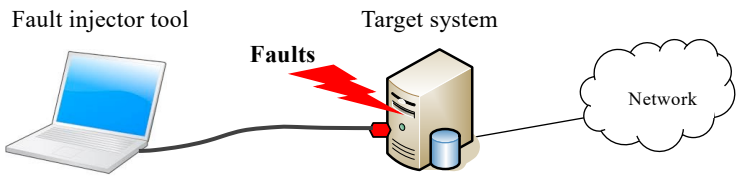
- Very often in computer/software experiments the dependent variable has only two possible outcomes. For example:
  - Error detected **or** error not detected
  - Vulnerability detected **or** vulnerability not detected
  - Silent data corruption **or** no silent data corruption (either the corruption was detected or there was no corruption at all)
  - System crashed **or** system did not crash
  - Robust behavior of web service **or** non robust behavior
  - Test case succeed **or** test case failed
  - Message arrived within specified timeframe **or** arrived outside the specified timeframe
  - Safety behavior **or** non safety behavior
  - Etc, etc, etc

Henrique Madeira, DEI-FCTUC, 2018-2023



3

## Example of a traditional fault injection scenario




- Injected 1000 faults and in 756 faults the system detected errors
- Injected 1000 faults and in 89 faults the system crashed
- Injected 1000 faults and in 56 faults the system produced an erroneous output without any warning (silent data corruption)
- …

The dependent variable is binary (two mutually exclusive outcomes). We can assume that a binomial distribution is a good approximation for these cases

Henrique Madeira, DEI-FCTUC, 2018-2023



4

## Binomial model

A binomial variable has the following properties:

- The variable is binary; it can take only one of two possible values.
- The variable is observed a known number of times (called **n**).
  - Each observation is often called a trial.
  - The number of times that the outcome of interest (e.g., error detection) is observed is **x**. It is often called the number of "**successes**" (in observing the outcome of interest).
- The probability that the outcome of interest occurs is the same for each trial.
- The trials are independent and the outcome of one trial does not affect the outcome of the any other trial.

Henrique Madeira, DEI-FCTUC, 2018-2023



5

## Binomial model

A binomial variable has the following properties:

- The variable is binary; it can take only one of two possible values.
- The variable is observed a known number of times (called **n**).
  - Each observation is often called a trial.
  - The number of times that the outcome of interest (e.g., error detection) is observed is **x**. It is often called the number of "**successes**" (in observing the outcome of interest).
- The probability that the outcome of interest occurs is the same for each trial.
- The trials are independent and the outcome of one trial does not affect the outcome of the any other trial.

These two bullets deserve some discussion.
Impact on the experiments to assure validity.

Henrique Madeira, DEI-FCTUC, 2018-2023



6

## Sampling distribution of the sample proportion

- Sample Proportion: $\hat{p} = \frac{x}{n}$ (**sample** = set of trials)
- $\hat{p}$ is the proportion of the sample with the outcome of interest. It is an estimate of the population proportion ***p***
- $\hat{p}$ varies from sample to sample in a random way
- For **large *n*** of samples the sampling distribution can be considered as a **normal distribution**. But the large number of samples should:
  - include a number of successes and non successes larger or equal to 10 (i.e., ***np* ≥ 10 and *n(1-p)* ≥ 10**);
  - be at least 20 times smaller than the population (i.e., population should be much larger)
- Consequently, we assume that the mean of the sampling distribution is approximately equal to the true population proportion ***p.***

Henrique Madeira, DEI-FCTUC, 2018-2023



7

## Sampling distribution of the sample proportion

- Sample Proportion: $\hat{p} = \frac{x}{n}$ (**sample** = set of trials)
- $\hat{p}$ is the proportion of the sample ... ate of the population proportion ***p***
- $\hat{p}$ varies from sample to sample in ... way
- For **large *n*** of samples the s... ng distribution ca... onsidered as a **normal distribution**. But the lar... number of samples s... d:
  - include a number of ...uccesses and non successes ...arger or equal to 10 (i.e., ***np* ≥ 10 and *n(1-p)* ≥ 10**);
  - be at least 20 times smaller than the population (i.e., population should be much larger)
- Consequently, we assume that the mean of the sampling distribution is approximately equal to the true population proportion ***p.***

Could be a problem in computer dependability and security experiments. Why?

Generally, not a problem in computer experiments. Why?

Henrique Madeira, DEI-FCTUC, 2018-2023



8

## Confidence Intervals (CI) for population proportion

Considering that for larger samples the sampling distribution of the sample proportion is approximately normal:

- The standard error (SE) of sample proportion is given by

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

- The confidence interval (CI) for population proportion is

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023



9

## Test statistic and effect size

Considering that for larger samples the sampling distribution of the sample proportion is approximately normal:

- Test statistic

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p\,)}{n}}}$$

Proportion observed in the sample

Proportion expected/ assumed for the population

Number of observations

- Effect size

$$d = \frac{|\,\hat{p} - p\,|}{p\,(1-p\,)}$$

Henrique Madeira, DEI-FCTUC, 2018-2023



10

## Test statistic and effect size: small sample

For a small sample from a normal population, use the **t statistic** instead of the z statistic. The degrees of freedom is n – 1

- Test statistic

$$t = \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p\,)}{n}}}$$

**But care should be taken with quite small samples!**

- Effect size

$$d = \frac{|\,\hat{p} - p\,|}{p\,(1-p\,)}$$





11

## Hypothesis test for a proportion: **Steps**

Follows the same basic steps of the other hypothesis testing:

1. State the hypothesis to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

| Large samples | Small samples |
|---|---|
| $z = \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p\,)}{n}}}$ | $t = \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p\,)}{n}}}$ |

Using the standard Z table or online calculators.

(or ***t*** table/calculators with df = n-1 for small samples)





12

## Example 5: Hypothesis test for a proportion

- You developed a new code vulnerability scanning tool to be used in the development of web applications. The CTO of your company has decided that the company can market the tool if it is able to detect at least 75% of the code vulnerabilities with a false positive rate below 15%.
- To test the tool, you used a benchmark composed by representative web applications that have been previously seeded with representative vulnerabilities. A total of 237 vulnerabilities have been injected in the application code. The tool detected 185 of those vulnerabilities, but also indicated 38 wrong vulnerabilities (false positives). The false positives have been confirmed by manual inspection of the code.
- **Can you report to the CTO with 95% confidence that your tool can detect more than 75% of the vulnerabilities and with less than 15% false positives?**





13

## Example 5: Hypothesis test for a proportion

**Steps:**

1. State the hypothesis or claim to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

As the problem has two independent claims (proportion of vulnerabilities detected and proportion of false positives) we do two separated hypothesis tests.





14

# Example 5: Hypothesis test for a proportion

## Step 1 - State the hypothesis be tested

We will test first the hypothesis related to the proportion of vulnerabilities detected

- **$H_0$: $p \leq 0.75$**
  The proportion of vulnerabilities detected is no better than 0.75 (0.75 is the target stated by the CTO)

- **$H_1$: $p > 0.75$**
  The proportion of vulnerabilities detected by the tool is higher than 0.75 (your tool is better than the limit stated by the CTO)





15

# Example 5: Hypothesis test for a proportion

## Step 2 - Compute the test statistic

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p\,)}{n}}} = \frac{0.7806 - 0.75}{\sqrt{\frac{0.75(1-0.75)}{237}}} = 1.088$$

**Z = 1.088**





16

# Example 5: Hypothesis test for a proportion
## Step 3 – Obtain p value

Looking at the Z table for $z = 1.088$ (one tailed) we find that ***p*** is between 0.1401 and 0.1379.

Using a calculator

***p*** = 0.1383 = 13.83%

| *A* *Z* | *B* *Area Between Mean and Z* | *C* *Area Beyond Z* |
|---|---|---|
| 0.71 | 0.2611 | 0.2389 |
| 0.72 | 0.2642 | 0.2358 |
| 0.73 | 0.2673 | 0.2327 |

Excerpt from the Z table

| | | |
|---|---|---|
| 1.07 | 0.3577 | 0.1423 |
| 1.08 | 0.3599 | 0.1401 |
| 1.09 | 0.3621 | 0.1379 |
| 1.10 | 0.3643 | 0.1357 |





17

# Example 5: Hypothesis test for a proportion
## Step 4 – Make a decision

As ***p*** = 13.83% we cannot reject the null hypothesis. It means that I cannot prove that my tool is better than 0.75 with 95% confidence.

**<u>As the tool fail in the first test, I do not need to test the false positives</u>**

**→ Conclusion: you should continue trying to improve the tool!!!**

Small ***p*** values provide evidence against the null hypothesis, as it means that the observed data are unlikely when the null hypothesis is true.

**Conventions:**

- $p \geq 0.10$ → the observed difference is **"not significant"**
- $0.05 \leq p < 0.10$ → the observed difference is **"marginally significant"**
- $0.01 \leq p < 0.05$ → the observed difference is **"significant"**
- $p < 0.01$ → the observed difference is **"highly significant"**





18

## Measure the confidence intervals: Proportion of vulnerabilities detected

**Proportion of vulnerabilities detected ~ coverage of the tool**

$$\hat{p} = \frac{x}{n} = \frac{185}{237} = 0.7806$$

Z = 1.96
from Z table for 95% confidence, two tailed

CI for a confidence level of 95%

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.7806 \pm 1.96 \times \sqrt{\frac{0.7806(1-0.7806)}{237}} = 0.7806 \pm 0.0527$$

**Proportion of vulnerabilities detected (95% confidence) = 0.7806 ± 0.0527**

Coverage of the tool between 72.79% and 83.33% with 95% confidence

Henrique Madeira, DEI-FCTUC, 2018-2023



19

## Measure the confidence intervals: Proportion of false positives

$$\hat{p} = \frac{x}{n} = \frac{38}{223} = 0.1704$$

For the false positives we should consider all the vulnerabilities detected: 185 (correct) + 38 (incorrectly indicated as vulnerabilities)

CI for a confidence level of 95%

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1704 \pm 1.96 \times \sqrt{\frac{0.1704(1-0.1704)}{223}} = 0.1704 \pm 0.0493$$

**Proportion of false positives (95% confidence) = 0.1704 ± 0.0493**

False positive detection of the tool is between 12.11% and 21.98% with 95% confidence

Henrique Madeira, DEI-FCTUC, 2018-2023



20

## Two-proportion z-test
### Hypothesis test for the difference between proportions

Hypothesis test to determine whether the difference between two proportions is significant.

For the previous example, this test could be useful if you want to compare the proportions of vulnerabilities detected (coverage) by two competing tools, T1 and T2.

We consider the measurements obtained with one tool are independent from the measurements obtained with the other (unrelated measurements → **independent samples**).

Follow the same steps as for hypothesis test for a proportion, but it requires a slightly different test statistic.





21

## Two-proportion z-test
### Possible sets of hypothesis under test

Let us assume we have two population proportions, P1 and P2

The following hypothesis can be tested about the difference between the two population proportions

| Null hypothesis | Alternate hypothesis | Number of tails |
|---|---|---|
| $P1 = P2$ | $P1 \neq P2$ | 2 |
| $P1 \geq P2$ | $P1 < P2$ | 1 |
| $P1 \leq P2$ | $P1 > P2$ | 1 |

Depending on the experiment and the goals of the researcher/engineers one or more of these hypothesis could be relevant to be tested.





22

## Test statistic

To compute the standard error (SE) we need a **pooled sample proportion**
(and assume it is similar to the entire population)

$$\hat{p} = \frac{p_1 \times n_1 + p_2 \times n_2}{n_1 + n_2}$$

p1 is the sample proportion from population 1,
p2 is the sample proportion from population 2,
n1 is the size of sample 1
n2 is the size of sample 2.

The standard error (SE) is:

$$SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

**The test statistic**

$$z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Henrique Madeira, DEI-FCTUC, 2018-2023



23

## Two-proportion z-test: Steps

Follows the same basic steps of the other hypothesis testing:

1. State the hypothesis to be tested
2. Compute the test statistic
3. Obtain p value
4. Make a decision

Using

$$z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Using the standard Z table or a calculator

Henrique Madeira, DEI-FCTUC, 2018-2023



24