p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Lecture 8: Hypothesis testing part II
Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 30, 2020

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Today

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Learning outcome

- Be able to explain the concept of p-hacking
- Be able to explain the following hypothesis tests
    - One-sample and two-sample z-test
    - One-sample and two-sample t-test
    - Paired t-test
    - Binomial test (exact, approximate)
    - McNemar's test (exact, approximate)

  For each of these tests, be able to describe the typical set up for the experiment, the general purpose of the test, data produced by the experiment, random variables, parameter of interest, null hypothesis, alternative hypothesis, test statistic, null distribution, the computation of p-value

- Be able to compare two classifiers using the paired t-test and McNemar's test for different scenarios

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Today

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Recall: one-tailed vs two-tailed tests

- *p*-value indicates how "surprising" the observation is
- "Surprising" observation usually means potential novelty
- In one of the examples from the previous lecture, we have shown that we reject the null hypothesis for the one-tailed test and we fail to reject the null hypothesis for the two-tailed test given the same significance level
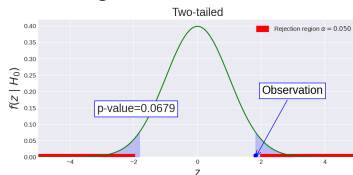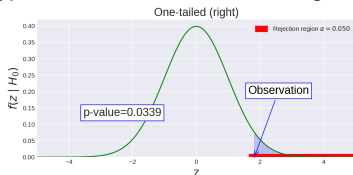
- In this example, if we use the two-tailed test, we will not claim that we have observed potential novelty with the experiment, whereas if we use the one-tailed test, we claim that we do observe potential novelty
- The conclusion we draw depends on which test we conduct

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Variation of the p-value

- *p*-value is computed from data
- Data is random - *p*-value is random
- With the same experiment set up, if we switch to a different sample, *p*-value will be different

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking

- Many factors can result in a different *p*-value
- *p*-hacking refers to situations where researchers are **trying multiple things until they get the desired result**
- This action can be a conscious decision, a subconscious decision or even an unconscious action
- *p*-hacking can be tricky to identify
- Suggestions to avoid *p*-hacking, e.g. one should always **report effect sizes and confidence intervals**
- Reference:
  - `https://www.nature.com/news/scientific-method-statistical-errors-1.14700`
  - Why Most Published Research Findings Are False?

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

CHALMERS | GÖTEBORGS UNIVERSITET

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true**

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true** - **try to prove yourself wrong**

p-hacking
Test statistics and hypothesis tests
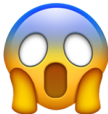Compare two classifiers
A/B testing
Summary

p-hacking (cont.)

What should I do!?

- Be honest and explicit about your assumptions
- Be "conservative"
- Be skeptical about your result - **don't let go of any doubt**!
- Assume the first success is always **too good to be true** - **try to prove yourself wrong** - be a proper scientist

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Today

CHALMERS | GÖTEBORGS UNIVERSITET

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer

- Recall that in this course, we only consider H_0 **with an equal sign in them**, i.e. the **null distribution is fully specified**; the description of H_0 is based on this assumption

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer

- Recall that in this course, we only consider H_0 **with an equal sign in them**, i.e. the **null distribution is fully specified**; the description of H_0 is based on this assumption
- For **symmetric null distributions**, e.g. **standard Gaussian distribution**, **student's t distribution**, **binomial distribution with** $p = 0.5$, etc, we only illustrate examples with the two-tailed alternative hypothesis H_A in this lecture without loss of generality; the one-tailed version can be easily derived

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer

- Recall that in this course, we only consider H_0 **with an equal sign in them**, i.e. the **null distribution is fully specified**; the description of H_0 is based on this assumption
- For **symmetric null distributions**, e.g. **standard Gaussian distribution**, **student's t distribution**, **binomial distribution with** $p = 0.5$, etc, we only illustrate examples with the two-tailed alternative hypothesis H_A in this lecture without loss of generality; the one-tailed version can be easily derived
- For the **exact binomial test with** $p \neq 0.5$, the null distribution is not symmetric; in this case, the computation of the two-tailed p-value is not uniquely defined; in this lecture, we will not go into details for these cases; we will only look at the one-tailed tests for asymmetric binomial null distributions

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer

- Recall that in this course, we only consider H_0 **with an equal sign in them**, i.e. the **null distribution is fully specified**; the description of H_0 is based on this assumption
- For **symmetric null distributions**, e.g. **standard Gaussian distribution**, **student's t distribution**, **binomial distribution with** $p = 0.5$, etc, we only illustrate examples with the two-tailed alternative hypothesis H_A in this lecture without loss of generality; the one-tailed version can be easily derived
- For the **exact binomial test with** $p \neq 0.5$, the null distribution is not symmetric; in this case, the computation of the two-tailed p-value is not uniquely defined; in this lecture, we will not go into details for these cases; we will only look at the one-tailed tests for asymmetric binomial null distributions
- For each hypothesis test, the purpose of the Python code snippet is to provide a better understanding of the calculation; in practice, there are alternative libraries and built-in functions for these tests that might result in a more compact implementation

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer

- Recall that in this course, we only consider H_0 **with an equal sign in them**, i.e. the **null distribution is fully specified**; the description of H_0 is based on this assumption
- For **symmetric null distributions**, e.g. **standard Gaussian distribution**, **student's t distribution**, **binomial distribution with** $p = 0.5$, etc, we only illustrate examples with the two-tailed alternative hypothesis H_A in this lecture without loss of generality; the one-tailed version can be easily derived
- For the **exact binomial test with** $p \neq 0.5$, the null distribution is not symmetric; in this case, the computation of the two-tailed p-value is not uniquely defined; in this lecture, we will not go into details for these cases; we will only look at the one-tailed tests for asymmetric binomial null distributions
- For each hypothesis test, the purpose of the Python code snippet is to provide a better understanding of the calculation; in practice, there are alternative libraries and built-in functions for these tests that might result in a more compact implementation
- We will come back to this lecture next week for some examples (to give you some time getting familiar with the concepts before the exercises)

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- Typical set up for the experiment
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- Typical set up for the experiment
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- **Typical set up for the experiment**
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test
- **Data**: symbolic description of the data produced by the experiment
- **Random variable** and **assumption** corresponding to the data

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- **Typical set up for the experiment**
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test
- **Data**: symbolic description of the data produced by the experiment
- **Random variable** and **assumption** corresponding to the data
- **Parameter of interest** and the **estimates**

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- **Typical set up for the experiment**
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test
- **Data**: symbolic description of the data produced by the experiment
- **Random variable** and **assumption** corresponding to the data
- **Parameter of interest** and the **estimates**
- **Hypotheses** H_0 and H_A

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- **Typical set up for the experiment**
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test
- **Data**: symbolic description of the data produced by the experiment
- **Random variable** and **assumption** corresponding to the data
- **Parameter of interest** and the **estimates**
- **Hypotheses** H_0 and H_A
- **Test statistic**
- **Null distribution**
  - PDF/PMF: description of the PDF/PMF
  - Python: code snippet of the PDF/PMF

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Disclaimer (cont.)

For each of the hypothesis tests we introduce, we present the following components:

- **Typical set up for the experiment**
  - **Test subjects**, e.g. the number of samples, the number of groups, etc
  - Description of the **experiment** and the **result**
  - Description of the **data type** produced in the result
- **Purpose**: the general purpose of the test
- **Data**: symbolic description of the data produced by the experiment
- **Random variable** and **assumption** corresponding to the data
- **Parameter of interest** and the **estimates**
- **Hypotheses** H_0 and H_A
- **Test statistic**
- **Null distribution**
  - PDF/PMF: description of the PDF/PMF
  - Python: code snippet of the PDF/PMF
- **p-value**
  - Definition: an expression of p-value in terms of a probability
  - Python: code snippet to illustrate the computation of the p-value (see page 10)

z-test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

One-sample z-test

- **Typical set up for the experiment**:
  - **One sample** of independent test subjects, e.g. a sample of patients, a sample of customers, etc
  - Run the same experiment on each subject and collect the outcomes, e.g. give a new drug to a sample of patients and measure the effect on each individual patient; test a new web design on a sample of customers and record the time they spend on the web page, etc
  - The result contains one i.i.d. sample with **continuous numerical values**
- **Purpose**: to test if the mean of the result differs from a predefined constant
- **Data**: $x_1, \cdots, x_N$, e.g. blood pressure after taking a new drug
- **Random variable** and **assumption**: $X_1, \cdots, X_N$
  - X_i i.i.d.
  - X_i Gaussian or large N (CLT)
  - X_i standard deviation σ known
- **Parameter of interest**: μ
- **Parameter estimate**: $\bar{x}$, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$
- **Hypotheses** H_0 and H_A: given c a constant

$$H_0: \quad \mu = c$$
$$H_A: \quad \mu \neq c$$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

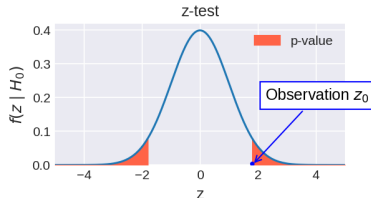# One-sample z-test (cont.)

- **Test statistic**:

$$z_0 = \frac{\bar{x} - c}{\sigma / \sqrt{N}}$$

- **Null distribution**: standard normal distribution
  - PDF: $f(z \mid H_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
  - Python: stats.norm.pdf(z, 0, 1)
- ***p*-value**
  - Definition: $p = 2 \min \left(P(Z \leq z_0 \mid H_0), P(Z \geq z_0 \mid H_0) \right)$
  - Python: $2 * \min \left(\text{stats.norm.cdf}(z_0, 0, 1), 1\text{-stats.norm.cdf}(z_0, 0, 1) \right)$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Two-sample z-test

- **Typical set up for the experiment**:
  - **Two samples** of independent test subjects, where the two samples $\mathcal{X}$ and $\mathcal{Y}$ letters with a calligraphic font are typically used to denote sets are independent from one another, e.g. two samples of independent patients, two samples of independent customers, etc
  - Run two sets of experiments A and B on the test subjects from the two samples $\mathcal{X}$ and $\mathcal{Y}$, respectively, and collect the outcomes, e.g. give different drugs to the two samples of patients and measure the effect on each individual patient; test two web designs on two samples of customers and record the time they spend on the web page, etc
  - The result contains two i.i.d. samples with **continuous numerical values**
- **Purpose**: to test if two alternative options have different effects by testing if the mean of the result from one sample differs from the mean of the other sample
- **Data**: $x_1, \cdots, x_{N_X}$ and $y_1, \cdots, y_{N_Y}$, e.g. blood pressure measured after taking two different drugs
- **Random variable** and **assumption**: $X_1, \cdots, X_{N_X}$, $Y_1, \cdots, Y_{N_Y}$
  - X_i and Y_j independent
  - X_i i.i.d.; Y_j i.i.d.
  - X_i Gaussian or large N_X; Y_j Gaussian or large N_Y
  - X_i and Y_j have known standard deviation σ_X and σ_Y, respectively
- **Parameter of interest**: μ_X, μ_Y
- **Parameter estimate**: $\bar{x}$, $\bar{y}$
- **Hypotheses** H_0 and H_A: given c a constant

$$H_0: \quad \mu_X - \mu_Y = c$$
$$H_A: \quad \mu_X - \mu_Y \neq c$$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Two-sample z-test (cont.)

- **Test statistic**:

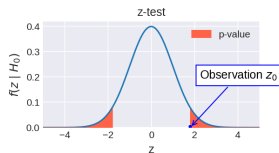$$z_0 = \frac{\bar{x} - \bar{y} - c}{\sqrt{\frac{\sigma_X^2}{N_X} + \frac{\sigma_Y^2}{N_Y}}}$$

  Hint: $\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma_X^2/N_X + \sigma_Y^2/N_Y\right)$
- **Null distribution**: standard normal distribution
  - PDF: $f(z \mid H_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
  - Python: stats.norm.pdf(z, 0, 1)
- ***p*-value**
  - Definition: $p = 2\min\left(P(Z \le z_0 \mid H_0), P(Z \ge z_0 \mid H_0)\right)$
  - Python: $2 * \min\left(\text{stats.norm.cdf}(z_0, 0, 1), 1\text{-stats.norm.cdf}(z_0, 0, 1)\right)$

One-sample t-test

One-sample t-test

- **Typical set up for the experiment** (same as the one-sample z-test):
  - One sample of independent test subjects, e.g. a sample of patients, a sample of customers, etc
  - Run the same experiment on each subject and collect the outcomes, e.g. give a new drug to a sample of patients and measure the effect on each individual patient; test a new web design on a sample of customers and record the time they spend on the web page, etc
  - The result contains one i.i.d. sample with **continuous numerical values**
- **Purpose**: to test if the mean of the result differs from a predefined constant
- **Data**: $x_1, \cdots, x_N$, e.g. blood pressure after taking a new drug
- **Random variable** and **assumption**: $X_1, \cdots, X_N$
  - X_i i.i.d.
  - X_i Gaussian or large N
  - X_i standard deviation σ **unknown**
- **Parameter of interest**: μ
- **Parameter estimate**: $\bar{x}$
- **Hypotheses** H_0 and H_A: given c a constant

$$H_0 : \quad \mu = c$$
$$H_A : \quad \mu \neq c$$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

One-sample t-test (cont.)

- **Test statistic**:
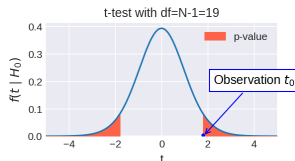
$$t_0 = \frac{\bar{x} - c}{s/\sqrt{N}}$$

where $s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ is the sample standard deviation

- **Null distribution**:
  - Student-t distribution with degree of freedom $df = N - 1$
  - Python: stats.t.pdf($t, df = N - 1$)

- **p-value**:
  - Definition: $p = 2\min\left(P(T \le t_0 \mid H_0), P(T \ge t_0 \mid H_0)\right)$
  - Python: $2 * \min\left(\text{stats.t.cdf}(t_0, df = N - 1), 1-\text{stats.t.cdf}(t_0, df = N - 1)\right)$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Two-sample t-test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Two-sample t-test

- **Typical set up for the experiment** (same as the two-sample z-test):
  - Two samples of independent test subjects, where the two samples $\mathcal{X}$ and $\mathcal{Y}$ are independent from one another, e.g. two samples of independent patients, two samples of independent customers, etc
  - Run two sets of experiments A and B on the test subjects from the two samples $\mathcal{X}$ and $\mathcal{Y}$, respectively, and collect the outcomes, e.g. give different drugs to the two samples of patients and measure the effect on each individual patient; test two web designs on two samples of customers and record the time they spend on the web page, etc
  - The result contains two i.i.d. samples with **continuous numerical values**
- **Purpose**: to test if two alternative options have different effects by testing if the mean of the result from one sample differs from the mean of the other sample
- **Data**: $x_1, \cdots, x_{N_X}$ and $y_1, \cdots, y_{N_Y}$, e.g. blood pressure measured after taking two different drugs
- **Random variable** and **assumption**: $X_1, \cdots, X_{N_X}, Y_1, \cdots, Y_{N_Y}$
  - X_i and Y_j independent
  - X_i i.i.d.; Y_j i.i.d.
  - X_i Gaussian or large N_X; Y_j Gaussian or large N_Y
  - X_i and Y_j have **unknown** standard deviation σ_X and σ_Y, respectively
- **Parameter of interest**: μ_X, μ_Y
- **Parameter estimate**: $\bar{x}$, $\bar{y}$
- **Hypotheses** H_0 and H_A: given c a constant

$$H_0: \quad \mu_X - \mu_Y = c$$
$$H_A: \quad \mu_X - \mu_Y \neq c$$

CHALMERS | GÖTEBORGS UNIVERSITET

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Two-sample t-test (cont.)

- **Test statistic**:

$$t_0 = \frac{\bar{x} - \bar{y} - c}{\sqrt{\frac{s_X^2}{N_X} + \frac{s_Y^2}{N_Y}}}$$
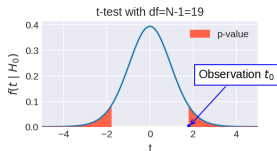
  with degree of freedom $df = \dfrac{(s_X^2/N_X + s_Y^2/N_Y)^2}{(\frac{s_X^2}{N_X})^2/(N_X - 1) + (\frac{s_Y^2}{N_Y})^2/(N_Y - 1)}$

- **Null distribution**:
  - Student-t distribution with degree of freedom df
  - Python: stats.t.pdf($t, df = df$)
- **p-value**:
  - Definition: $p = 2\min\left(P(T \leq t_0 \mid H_0), P(T \geq t_0 \mid H_0)\right)$
  - Python: $2 * \min\left(\text{stats.t.cdf}(t_0, \text{df}=df), 1\text{-stats.t.cdf}(t_0, \text{df}=df)\right)$

Paired t-test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Paired t-test

- **Typical set up for the experiment**:
  - One sample of independent test subjects, e.g. one sample of independent patients
  - Run two sets of experiments A and B on all subjects from the sample and collect the outcomes, e.g. measure the blood pressure of the patients **before** giving them a new drug (experiment A); measure the blood pressure of the patients **after** giving them the new drug (experiment B)
  - The result contains two samples with **continuous numerical values**
- **Purpose**: to test if two alternative options have different effects by testing if the mean of the difference between two results differs from a predefined constant
- **Data**: $x_1, \cdots, x_N, y_1, \cdots, y_N$
- **Random variable** and **assumption**: $X_1, \cdots, X_N, Y_1, \cdots, Y_N$
  - $X_i - Y_i$ i.i.d.
  - $X_i - Y_i \sim \mathcal{N}\left(\mu_{X-Y}, \sigma^2_{X-Y}\right)$ with unknown standard deviation
- **Parameter of interest**: μ_{X-Y}
- **Parameter estimate**: $m_{X-Y} = \frac{1}{N} \sum_{i=1}^{N} (x_i - y_i)$
- **Hypotheses** H_0 and H_A: given c a constant

$$H_0: \quad \mu_{X-Y} = c$$
$$H_A: \quad \mu_{X-Y} \neq c$$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Paired t-test

- **Test statistic**:

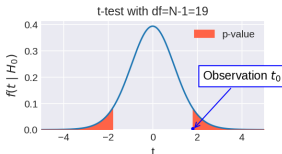$$t_0 = \frac{m_{X-Y} - c}{s_{X-Y}/\sqrt{N}}$$

  where $s_{X-Y} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(x_i - y_i - m_{X-Y}\right)^2}$

- **Null distribution**:
  - Student-t distribution with degree of freedom $N-1$
  - Python: stats.t.pdf($t, df = N - 1$)
- **p-value**:
  - Definition: $p = 2 \min \left(P(T \le t_0 \mid H_0), P(T \ge t_0 \mid H_0)\right)$
  - Python: $2 * \min \left(\text{stats.t.cdf}(t_0, df = N - 1),\ 1\text{-stats.t.cdf}(t_0, df = N - 1)\right)$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Exercise 1

- A company claims that a new drug E they have developed can increase the average sleeping hours of people with insomnia. Design three different hypothesis tests to test this statement.

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Exercise 2

- One of the tests you have designed is a two-sample test. After the experiments, you realized the test subjects being selected in the second group are parents or siblings of the first group. Would that be a problem? Can you still use the result somehow?

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Binomial test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
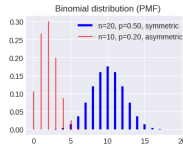Paired t-test
Binomial test
McNemar's test

Binomial distribution

- Discrete distribution
- Applies to discrete numerical data - the number of success from n independent Bernoulli trials with probability of success p
- PMF:
  - Equation

$$f_X(k \mid n, p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, \cdots, n, \ p \in [0,1]$$

  where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient (n choose k)
  - Shape
    - When $p = 0.5$, the PMF is symmetric
    - When $p \neq 0.5$, the PMF is asymmetric

- Parameters: p and n; n is typically known

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(exact) Binomial test

- **Typical set up for the experiment**:
  - One sample of independent test subjects, e.g. one sample of independent patients
  - Run the same experiment on all subjects from the sample and collect the outcomes, e.g. give a new drug to a sample of patients and measure how many patients are cured
  - The result contains one sample with **nominal categorical values with two categories**, which is then summarized into one **discrete numerical value** - the number of "success"
- **Purpose**: to test if the proportion of "success" differs from a predefined constant
- **Data**: N independent Bernoulli trials with k_0 "success", e.g. the number of cured patients within the sample of size N
- **Random variable** and **assumption**: $K \sim Binomial(N, p)$ with known N and unknown success rate p
- **Parameter of interest**: p
- **Parameter estimate**: $\hat{p} = \frac{k_0}{N}$
- **Null hypothesis**: given π a constant,

$$H_0 : p = \pi$$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(exact) Binomial test (cont.)

- **Test statistic**: k_0
- **Null distribution**:

$$P(X = k) = \binom{N}{k}\pi^k(1 - \pi)^{N-k}$$

  - Binomial distribution with parameters N and π
  - Python: stats.binom.pmf(k, N, π)
- As discussed in the disclaimer (cf. page 10), we only introduce the following scenarios:
  - One-tailed (left) binomial test with any $\pi \in (0, 1)$
  - One-tailed (right) binomial test with any $\pi \in (0, 1)$
  - Two-tailed binomial test with $\pi = 0.5$, where the null distribution is symmetric

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

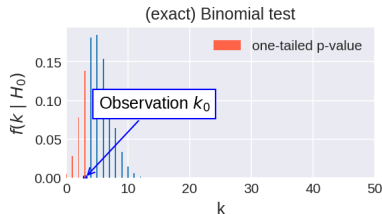# (exact) One-tailed (left) binomial test

- **Hypotheses** H_0 and H_A:

$$H_0 : \qquad p = \pi$$
$$H_A : \qquad p < \pi$$

- *p*-**value**:
  - Definition: $p = P(K \leq k_0 \mid H_0)$
  - Python: stats.binom.cdf(k_0, n=N, p=π)

(exact) Binomial test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(exact) One-tailed (right) binomial test
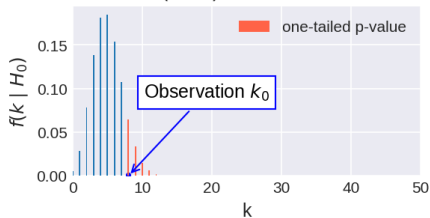
- **Hypotheses** H_0 and H_A:

$$H_0: \quad p = \pi$$
$$H_A: \quad p > \pi$$

- *p*-**value**:
  - Definition: $p = P(K \geq k_0 \mid H_0)$
  - Python: $1 - \text{stats.binom.cdf}(k_0, n = N, p = \pi) + \text{stats.binom.pmf}(k_0, n = N, p = \pi)$

(exact) Binomial test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(exact) Two-tailed binomial test

- **Hypotheses** H_0 and H_A:

$$H_0 : \quad p = 0.5$$
$$H_A : \quad p \neq 0.5$$

- *p*-**value**:
  - Definition: $p = 2\min\left(P(K \leq k_0 \mid H_0), P(K \geq k_0 \mid H_0)\right)$
  - Python:
    - $c = \text{stats.binom.cdf}(k_0, n = N, p = 0.5)$
    - $2 * \min\left(c, 1 - c + \text{stats.binom.pmf}(k_0, n = N, p = 0.5)\right)$

(exact) two-tailed Binomial test with $p = 0.5$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test
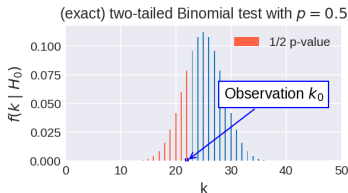
(large N) Binomial test

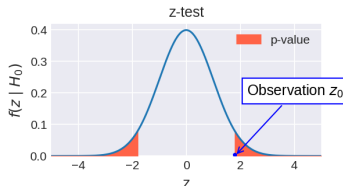Same set up as page 30, but with large N

- **Test statistic**:

$$z_0 = \frac{k_0 - N\pi}{\sqrt{N\pi(1-\pi)}}$$

- **Null distribution**: standard normal distribution
  - PDF: $f(z \mid H_0) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$
  - Python: stats.norm.pdf(z, 0, 1)
- **p-value**:
  - Definition: $p = 2\min\left(P(Z \leq z_0 \mid H_0), P(Z \geq z_0 \mid H_0)\right)$
  - Python: $2 * \min\left(\text{stats.norm.cdf}(z_0, 0, 1), 1\text{-stats.norm.cdf}(z_0, 0, 1)\right)$

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

McNemar's test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

McNemar's test

- **Typical set up for the experiment**:
  - **One sample** of independent test subjects, e.g. one sample of independent patients
  - Within the sample, there are two **groups**; each subject belongs to one and only one group, e.g. within the sample of patients, we have one group with high blood pressure and another group with normal blood pressure
  - Run two sets of experiments A and B on all test subjects from the sample and collect the outcomes, e.g. measure the blood pressure (hight or normal) of the patients **before** giving them a new drug (experiment A); measure the blood pressure (hight or normal) of the patients **after** giving them the new drug (experiment B)
  - The result contains one sample with **nomimal categorical values with two categories** measured from each test subject, e.g. high blood pressure and normal blood pressure
- **Purpose**: to test if an action have different effects on two different **groups**
- **Data**: N independent Bernoulli trials with outcomes $x_1, \cdots, x_N$ and $y_1, \cdots, y_N$ for the two experiments, respectively; $x_i, y_i \in \{0, 1\}$

	$x_i = 0$	$x_i = 1$	
$y_j = 0$	n_{00}	n_{10}	$n_{00} + n_{10}$
$y_j = 1$	n_{01}	n_{11}	$n_{01} + n_{11}$
	$n_{00} + n_{01}$	$n_{10} + n_{11}$	N

where n_{mn} is the count of $x_i = m$ and $y_j = n$
- **Random variable** and **assumption**: i.i.d. $X_i \sim Bernoulli(p_X)$ and i.i.d. $Y_i \sim Bernoulli(p_Y)$

CHALMERS | GÖTEBORGS UNIVERSITET

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

McNemar's test (cont.)

Example

- A company is trying to determine the effectiveness of a drug on lowering blood pressure
- The company tested the drug on a sample of 229 **independent** patients
- There are **two groups** within this sample: a high blood pressure group (105 patients) and a normal blood pressure group (124 patients); each patient belongs to one and only one of these two groups
- The blood pressure of each patient is measured **before** (to determine the group) and **after** (to determine the effect) taking the drug
- The data is summarized as follows:

	After (high blood pressure)	After (normal blood pressure)	
Before (high blood pressure)	90	15	105
Before (normal blood pressure)	22	102	124
	112	117	229

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(small discordance $n_{01} + n_{10}$) McNemar's test (cont.)

- **Parameter of interest**: **discordance**

$$p = \min\left(P(X_i = 0, Y_i = 1 \mid X_i \neq Y_i), P(X_i = 1, Y_i = 0 \mid X_i \neq Y_i)\right)$$

- **Parameter estimate**: $\hat{p} = \frac{\min(n_{01}, n_{10})}{n_{01} + n_{10}}$
- **Hypotheses** H_0 and H_A:
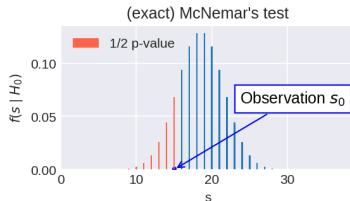
$$H_0: \quad p = 0.5$$
$$H_A: \quad p \neq 0.5$$

- H_0: the drug does not have any effect on blood pressure control
- H_1: the drug has effect on blood pressure control

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(small discordance $n_{01} + n_{10}$) McNemar's test (cont.)

- **Test statistic**: $s_0 = \min(n_{01}, n_{10})$
- **Null distribution**:
  - Binomial distribution with parameters $(n_{01} + n_{10}, 0.5)$
  - Python: stats.binom.pmf(s, $n_{01} + n_{10}$, 0.5)
- **p-value**:
  - Definition: $p = 2P(S \le s_0 \mid H_0)$
  - Python: $2 * $ stats.binom.cdf$(s_0, n_{01} + n_{10}, 0.5)$

(exact) McNemar's test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(large discordance $n_{01} + n_{10}$) McNemar's test

Same set up as page 37, but with large $n_{01} + n_{10}$, e.g. $n_{01} + n_{10} > 25$

- **Parameter of interest**: **discordance** (note: it is different from the previous definition (cf. page 39)) $p_{01} = P(X = 0, Y = 1)$ and $p_{10} = P(X = 1, Y = 0)$
- **Parameter estimate**: $\hat{p}_{01} = \frac{n_{01}}{N}$ and $\hat{p}_{10} = \frac{n_{10}}{N}$
- **Hypotheses** H_0 and H_A:

$$H_0 : \qquad p_{01} = p_{10}$$
$$H_A : \qquad p_{01} \neq p_{10}$$

- **Test statistic**:

$$s_0 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$$

Note: "-1" is called the continuity correction (https://en.wikipedia.org/wiki/Continuity_correction)

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
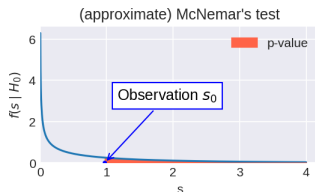Two-sample t-test
Paired t-test
Binomial test
McNemar's test

(large discordance $n_{01} + n_{10}$) McNemar's test (cont.)

- **Null distribution**:
  - Large $n_{01} + n_{10}$: s_0 follows a chi-squared distribution with $df = 1$
  - Python: stats.chi2.pdf(s, $df = 1$)
- **p-value**:
  - Definition: $P(S \geq s_0 \mid H_0)$
  - Python: 1-stats.chi2.cdf(s_0, $df = 1$)

(approximate) McNemar's test

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

z-test
One-sample t-test
Two-sample t-test
Paired t-test
Binomial test
McNemar's test

Exercise 3

- Run both the exact McNemar's test and the approximate McNemar's test on the data set provided on page 38.

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Today

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

K-fold cross validation

- **Classifiers**: A and B
- **Data**: evaluation metric; **continuous numerical data**, e.g. accuracies $p_1^A, \cdots, p_K^A$ and $p_1^B, \cdots, p_K^B$ on the K validation sets

	fold 1	fold 2	$\cdots$	fold K
classifier A	p_1^A	p_2^A	$\cdots$	p_K^A
classifier B	p_1^B	p_2^B	$\cdots$	p_K^B
$p_i^A - p_i^B$	$p_1^A - p_1^B$	$p_2^A - p_2^B$	$\cdots$	$p_K^A - p_K^B$

- **Random variable** and **assumption**: $P_1^A, \cdots, P_N^A, P_1^B, \cdots, P_N^B$
  - $P_i^A - P_i^B$ i.i.d.
  - $P_i^A - P_i^B \sim \mathcal{N}\left(\mu_{A-B}, \sigma_{A-B}^2\right)$ with unknown standard deviation
- **Parameter of interest**: μ_{A-B}
- **Parameter estimate**: $m_{A-B} = \frac{1}{K} \sum_{i=1}^{K} \left(p_i^A - p_i^B\right)$
- **Hypotheses** H_0 and H_A:

$$H_0: \quad \mu_{A-B} = 0$$
$$H_A: \quad \mu_{A-B} \neq 0$$

- **Test statistic**: $t = \frac{m_{A-B}}{s_{A-B}/\sqrt{K}}$, where $s_{A-B} = \sqrt{\frac{1}{K-1} \sum_{i=1}^{K} \left(p_i^A - p_i^B - m_{A-B}\right)^2}$
- **Hypothesis test**: paired t-test

CHALMERS | GÖTEBORGS UNIVERSITET

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Training-validation split and leave-one-out cross validation

- **Classifiers**: A and B
- **Data**: classifiers A and B tested on the validation data; the outcome x_i^A and x_i^B can be either correct (0) or incorrect (1); **nominal categorical data** with two categories correct or incorrect

	classifier A correct	classifier A incorrect
classifier B correct	n_{00}=count(A correct, B correct)	n_{10}=count(A incorrect, B correct)
classifier B incorrect	n_{01}=count(A correct, B incorrect)	n_{11}=count(A incorrect, B incorrect)

- **Random variable** and **assumption**: $X_i^A \sim Bernoulli(p_A), X_i^B \sim Bernoulli(p_B)$
- **Test statistic**:
  - Small discordance (e.g. $n_{01} + n_{10} < 25$): $s_0 = \min(n_{01}, n_{10})$
  - Large discordance: $s_0 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}$
- **Hypothesis test**: McNemar's test
- **Note**: why McNemar's test? - One consideration while developing hypothesis test is to have a reasonably low type I error

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Exercise 4

- You have a labeled data set $\mathcal{D} = \{(x_1, y_1), \cdots, (x_N, y_N)\}$
- You developed two classifiers using the 10-fold validation
- Construct a table of the resulting F1 scores for these two classifiers
- Design a hypothesis test to compare these two classifiers

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Exercise 5

- You have a labeled data set $\mathcal{D} = \{(x_1, y_1), \cdots, (x_N, y_N)\}$
- You developed two classifiers using the training-validation split
- Construct a table of the results for these two classifiers
- Design a hypothesis test to compare these two classifiers

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Today

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

What is A/B testing?

- A/B testing refers to the methodology for testing two variants of the same type of product
- It is widely used for improving marketing strategies
- Example:
    - **Web design**: a company would like to increase the number of their subscribers by testing a new design for the "subscribe" button on their web page
    - **Product pricing**: a company would like to introduce a new product to the market; they compare two potential prices to see which one maximizes their sales while maintaining a reasonable margin
- We will revisit this topic next week as part of the exercise

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Today

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis tests, comparison of two classifiers

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis tests, comparison of two classifiers

Next:

- Unsupervised learning, clustering, k-means, Gaussian Mixture Models

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
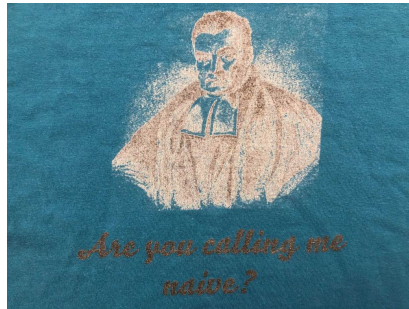A/B testing
Summary

Summary

So far:

- Data types and data containers
- Descriptive data analysis: descriptive statistics, visualization
- Probability distributions, events, random variables, PMF, PDF, parameters
- CDF, Q-Q plot, how to compare two distributions (data vs theoretical, data vs data)
- Modeling
- Parameter estimation: maximum likelihood estimation (MLE) and maximum a posteriori estimation (MAP)
- Classification, multinomial naive Bayes classifier, Gaussian naive Bayes classifier
- Central limit theorem, interval estimation
- Hypothesis tests, comparison of two classifiers

Next:

- Unsupervised learning, clustering, k-means, Gaussian Mixture Models

Before next lecture:

- Gaussian distribution
- The Bayes' rule

p-hacking
Test statistics and hypothesis tests
Compare two classifiers
A/B testing
Summary

I heard my name again!