National Ph.D. Program in *Artificial Intelligence for Society*

# Statistics for Machine Learning

Lesson 08 - Confidence intervals and Hypotheses testing.

Andrea Pugnana, Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
andrea.pugnana@di.unipi.it salvatore.ruggieri@unipi.it

# From point estimate to interval estimate

## Estimator and point estimate

A *statistics* is a function $h(X_1, \ldots, X_n)$ of r.v.'s.
An *estimator* of a parameter $\theta$ is a statistics $T_n = h(X_1, \ldots, X_n)$ intended to provide information about $\theta$.
A *point estimate* $t$ of $\theta$ is $t = h(x_1, \ldots, x_n)$ over realizations of $X_1, \ldots, X_n$.

- Sometimes, a *range* of plausible values $l < \theta < u$ is useful, as it provides uncertainty information
- Idea: *confidence interval* is an interval for which we can be confident the unknown parameter $\theta$ is in with a specified probability (called *confidence level*)

# Confidence intervals

CONFIDENCE INTERVALS. Suppose a dataset $x_1, \ldots, x_n$ is given, modeled as realization of random variables $X_1, \ldots, X_n$. Let $\theta$ be the parameter of interest, and $\gamma$ a number between 0 and 1. If there exist sample statistics $L_n = g(X_1, \ldots, X_n)$ and $U_n = h(X_1, \ldots, X_n)$ such that

$$\mathrm{P}(L_n < \theta < U_n) = \gamma$$

for every value of $\theta$, then

$$(l_n, u_n),$$

where $l_n = g(x_1, \ldots, x_n)$ and $u_n = h(x_1, \ldots, x_n)$, is called a $100\gamma\%$ *confidence interval for* $\theta$. The number $\gamma$ is called the *confidence level*.

- Sometimes, only have $P(L_n < \theta < U_n) \geq \gamma$ *[conservative* $100\gamma\%$ *confidence interval]*
  - E.g., the interval found using Chebyshev's inequality
- There is no way of knowing if $l_n < \theta < u_n$ (interval is correct or not)
- We only know that we have probability $\gamma$ of covering $\theta$
- Notation: $\gamma = 1 - \alpha$ where $\alpha$ is called the *significance level*
  - $100\gamma = 95\%$ *confidence level*, i.e. probability that interval includes the parameter
  - $\alpha = 0.05$ *significance level*, i.e. probability that interval does not include the parameter

**Seeing theory simulation**

# Confidence intervals for the mean: summary

- $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n \sim F$ with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
- Problem: what is a confidence interval for $\mu$?
  - Normal data $F = \mathcal{N}(\mu, \sigma^2)$
    - with known variance: $Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$
    - with unknown variance: $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$
  - General data (with unknown variance)
    - large sample, i.e., large $n$: $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$
    - bootstrap (next lesson)
  - Bernoulli data $F = Ber(\mu)$
    - confidence interval for proportions: $T = \frac{\bar{X}_n - \mu}{\sqrt{\bar{X}_n(1-\bar{X}_n)}/\sqrt{n}}$

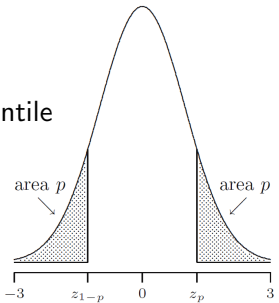
# Critical values

**Critical value**

The (right) *critical value* $z_p$ of $Z \sim \mathcal{N}(0,1)$ is the number with right tail probability $p$:

$$P(Z \geq z_p) = p$$

- The right tail is $P(Z \geq z_p) = 1 - P(Z \leq z_p) = 1 - \Phi(z_p)$
  - This is why Table B.1 of the textbook is given for $1 - \Phi()$
- $1 - \Phi(z_p) = p$ means $\Phi(z_p) = 1 - p$, i.e., $z_p$ is the $(1-p)$th quantile
- By symmetry, $P(Z \geq z_p) = P(Z \leq -z_p) = p$, and then
  $$z_{1-p} = -z_p$$
  - E.g., $z_{0.975} = -z_{0.025} = -1.96$ and $z_{0.025} = -z_{.975} = 1.96$



# CI for the mean: normal data with known variance

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$
- Estimator $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ and the scaled mean:

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{1}$$

- Confidence interval for $Z$:

$$P(c_l \leq Z \leq c_u) = \gamma \quad \text{or} \quad P(Z \leq c_l) + P(Z \geq c_u) = \alpha = 1 - \gamma$$

- Symmetric split:

$$P(Z \leq c_l) = P(Z \geq c_u) = \alpha/2$$

Hence $c_u = -c_l = z_{\alpha/2}$, and by (1):

$$P(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha = \gamma$$

$(\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for $\mu$

# One-sided confidence intervals

- One-sided confidence intervals (*greater-than*):

$$P(L_n < \theta) = \gamma$$

  Then $(l_n, \infty)$ is a $100\gamma\%$ or $100(1-\alpha)\%$ one-sided confidence interval
- $l_n$ is called the *lower confidence bound*
- Normal data with known variance:

$$P(\bar{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}} \leq \mu) = 1 - \alpha = \gamma$$

$(\bar{x}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$ is a $100\gamma\%$ or $100(1-\alpha)\%$ one-sided confidence interval for $\mu$

**See R script**

# CI for the mean: normal data with unknown variance

- Use the unbiased estimator of $\sigma^2$ and its estimate:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

  - and then $S_n^2/n$ is an unbiased estimator of $Var(\bar{X}_n) = \sigma^2/n$

- The following transformation is called the *studentized mean*: $T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$

DEFINITION. A continuous random variable has a *t-distribution with parameter* $m$, where $m \geq 1$ is an integer, if its probability density is given by

$$f(x) = k_m \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} \qquad \text{for } -\infty < x < \infty,$$

where $k_m = \Gamma\left(\frac{m+1}{2}\right) / \left(\Gamma\left(\frac{m}{2}\right)\sqrt{m\pi}\right)$. This distribution is denoted by $t(m)$ and is referred to as the $t$-distribution with $m$ *degrees of freedom.*

  - Student/Gosset t-distribution $X \sim t(m)$: **Some history on its discovery**
    - $E[X] = 0$ for $m \geq 2$, and $Var(X) = m/(m-2)$ for $m \geq 3$
    - For $m \to \infty$, $X \to \mathcal{N}(0,1)$

**See R script**

# CI for the mean: normal data with unknown variance

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

**Critical value**

The (right) *critical value* $t_{m,p}$ of $T \sim t(m)$ is the number with right tail probability $p$:

$$P(T \geq t_{m,p}) = p$$

- Same properties as $z_p$
- From the studentized mean:

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

to confidence interval:

$$P(\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}) = 1 - \alpha = \gamma$$

$(\bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}})$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for $\mu$

**See R script**

# CI for the mean: general data with unknown variance

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n$
- A variant of CLT states that for $n \to \infty$

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \to \mathcal{N}(0,1)$$

- For large $n$, we make the approximation: *[how large should n be?]*

$$T = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \approx \mathcal{N}(0,1)$$

and then

$$P(\bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}) \approx 1 - \alpha = \gamma$$

$(\bar{x}_n - z_{\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n + z_{\alpha/2}\frac{s_n}{\sqrt{n}})$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for $\mu$

**See R script**

# General form of Wald confidence intervals

$$\theta \in \hat{\theta} \pm z_{\alpha/2} se(\hat{\theta}) \qquad \text{or} \qquad \theta \in \hat{\theta} \pm t_{\alpha/2} se(\hat{\theta})$$

- They originate from the **Wald test statistics**:

$$T = \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

- Importance of standard error $se(\hat{\theta})$ of estimators!
- Limitation: asymptotic, symmetric intervals

# CI for proportions (e.g., classifier accuracy)

- Dataset $x_1, \ldots, x_n$ realization of random sample $X_1, \ldots, X_n \sim Ber(p)$
  - $x_i = \mathbb{1}_{y_\theta^+(w_i)=c_i}$ is 1 for correct classification, 0 for incorrect classification *[over a test set]*
  - $p$ is the (unknown) accuracy of classifier $y_\theta^+()$
- $B = \sum_{i=1}^n X_i \sim Bin(n, p)$ and $\bar{X}_n = B/n$
  - For large $n$, $Bin(n, p) \approx \mathcal{N}(np, np(1-p))$ for $0 \ll p \ll 1$ *[**De Moivre–Laplace**]*
    - and then $\bar{X}_n = B/n \approx \mathcal{N}(p, p(1-p)/n)$
    - $se(\bar{X}_n) = \sqrt{np(1-p)}/n \approx \sqrt{\bar{X}_n(1-\bar{X}_n)/n}$, because we don't known $p$
    - Consider $T = (\bar{X}_n - p)/se(\bar{X}_n) \approx \mathcal{N}(0,1)$ and then $P(-z_{\alpha/2} \leq T \leq z_{\alpha/2}) = \gamma$ implies:

$$P(\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \leq p \leq \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}) = 1 - \alpha = \gamma$$

$(\bar{x}_n - z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}}, \bar{x}_n + z_{\alpha/2}\sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}})$ is a $100\gamma\%$ or $100(1-\alpha)\%$ confidence interval for $p$

  - Drawbacks: symmetric, large sample, skewness, etc. *[see **Wilson score interval** and **Exact or Clopper–Pearson interval**]*

**See R script**

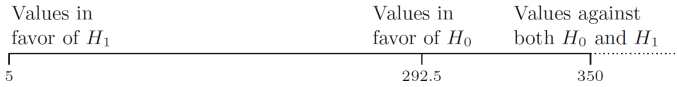
# Hypothesis testing

- Hypotheses testing consists of contrasting two conflicting hypotheses based on observed data
- Consider the German tank problem:
  - Military intelligence states that $N = 350$ tanks were produced *[H0 or null hypothesis]*
  - Alternative hypothesis: *[H1 or alternative hypothesis]*
    $N < 350$ (*one-tailed or one-sided test*), or $N \neq 350$ (*two-tailed or two-sided test*)
  - Observed serial tank id's: 61 19 56 24 16
- Statistical test: How likely is the observed data under the assumption that the null hypothesis holds?
  - If it is NOT (sufficiently) likely, we reject the null hypothesis in favor of H1
  - If it is (sufficiently) likely, we cannot reject the null hypothesis
- Why '*we cannot reject the null hypothesis*' and not instead '*we accept the null hypothesis*'?
  - Other hypotheses, e.g., $N = 349$ or $N = 351$, could also be not rejected
    and then, we cannot say which of $N = 349$ or $N = 350$ or $N = 351$ is actually true

# Test statistic

TEST STATISTIC. Suppose the dataset is modeled as the realization of random variables $X_1, X_2, \ldots, X_n$. A *test statistic* is any sample statistic $T = h(X_1, X_2, \ldots, X_n)$, whose numerical value is used to decide whether we reject $H_0$.

- In the German tank example: *[See Lesson 19]*
  - $H_0 : N = 350$
  - $H_1 : N < 350$
  - Observed serial tank id's: 61 19 56 24 16
- We use $T = \max\{X_1, X_2, X_3, X_4, X_5\}$
- If $H_0$ is true, i.e., $N = 350$, then $E[T] = \frac{5}{6}(N+1) = \frac{5}{6}351 = 292.5$



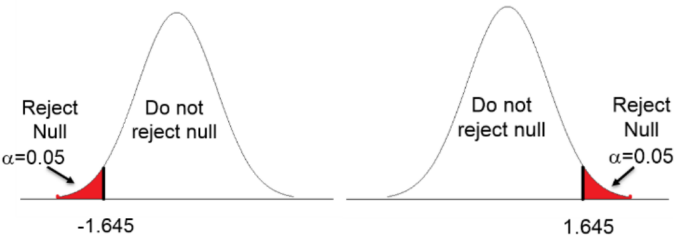

- If $H_0$ is true, we have:

$$P(T \leq 61) = P(\max\{X_1, X_2, X_3, X_4, X_5\} \leq 61) = \frac{61}{350} \cdot \frac{60}{349} \cdots \frac{57}{346} = 0.00014$$

very unlikely: either we are unfortunate, or $H_0$ can be rejected

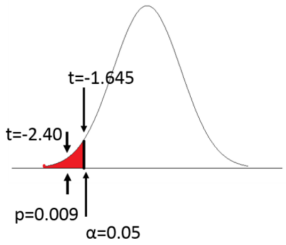
# Statistical test of hypothesis: one-tailed – critical region

- $H_0$: $\theta = v$ *[Null hypothesis]*
- $H_1$: $\theta < v$ (resp. $H_1$: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when $H_0$ is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- $c_l$ s.t. $P(T \leq c_l) = \alpha$ (resp. $c_u$ s.t. $P(T \geq c_u) = \alpha$) *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t \leq c_l$ (resp. $t \geq c_u$): $H_0$ is rejected *[Critical region]*
  - otherwise: $H_0$ cannot be rejected



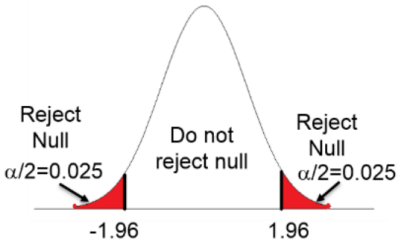
# Statistical test of hypothesis: one-tailed – p-value

- $H_0$: $\theta = v$ *[Null hypothesis]*
- $H_1$: $\theta < v$ (resp. $H_1$: $\theta > v$) *[Left-tailed/Right-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when $H_0$ is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- $p = P(T \leq t)$ (resp. $p = P(T \geq t)$) *[p-value]*
  - evidence against $H_0$ - the smaller the stronger evidence
- Output of the test at confidence level $100(1-\alpha)\%$ using $p$-values
  - $p \leq \alpha$: $H_0$ is rejected
  - otherwise: $H_0$ cannot be rejected

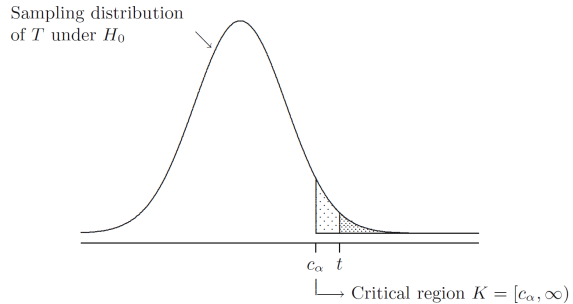


# Statistical test of hypothesis: two-tailed

- $H_0$: $\theta = v$ *[Null hypothesis]*
- $H_1$: $\theta \neq v$ *[Two-tailed test]*
- $100(1-\alpha)\%$, e.g., 95% or 99% or 99.9% *[Confidence level]*
  - i.e., $\alpha = 0.05$ or $\alpha = 0.01$ or $\alpha = 0.001$ *[Significance level]*
- $T = h(X_1, \ldots, X_n)$ test statistics when $H_0$ is true
- $x_1, \ldots, x_n$: observed dataset, and $t = h(x_1, \ldots, x_n)$ *[t-value]*
- $c_l$ s.t. $P(T \leq c_l) = \alpha/2$ and $c_u$ s.t. $P(T \geq c_u) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t \leq c_l$ or $t \geq c_u$: $H_0$ is rejected *[Critical region]*
  - otherwise: $H_0$ cannot be rejected



# Critical values and p-values




- *Critical region* $K$: the set of values that reject $H_0$ in favor of $H_1$ at significance level $\alpha$
- *Critical values*: values on the boundary of the critical region
- *p-value*: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that $H_0$ is true
- $t \in K$ iff $p$-value $\leq \alpha$

# Type I and Type II errors

| | | True state of nature | |
|---|---|---|---|
| | | $H_0$ is true | $H_1$ is true |
| **Our decision on the basis of the data** | Reject $H_0$ | *Type I error* | Correct decision |
| | Not reject $H_0$ | Correct decision | *Type II error* |

- Type I error is we falsely reject $H_0$: $P(\text{Reject } H_0 | H_0 \text{ is true})$ *[$\alpha$-risk, false positive rate]*
  - E.g., unjust speed-limit fine
  - we reject $H_0$ when $p < \alpha$, so this error occur with probability $100\alpha\%$
  - this error can be controlled by setting the significance level $\alpha$ to the largest acceptable value
    - how much is an *acceptable value*?
  - A possible solution is to solely report the *p*-value, which conveys the maximum amount of information and permits decision makers to choose their own level
- Type II error is we falsely do not reject $H_0$: $P(\text{Not Reject } H_0 | H_1 \text{ is true})$ *[$\beta$-risk, false negative rate]*
  - E.g.,lack of a true speed-limit sanction
  - $1 - \beta = P(\text{Reject } H_0 | H_1 \text{ is true})$ is called the *power* of the test

# Relation with confidence intervals

- $H_0$: $\mu = 120$ (null hypothesis)
- $H_1$: $\mu > 120$ (alternative hypothesis)
- $\alpha = 0.05$ (significance level)
- $c_u = 120 + z_{0.05}\frac{2}{\sqrt{3}} = 121.9$
- $H_0$ rejected when:

$$
\begin{aligned}
& t = \bar{x}_3 \geq c_u \\
\Leftrightarrow \quad & \bar{x}_3 \geq 120 + z_{0.05}\frac{2}{\sqrt{3}} \\
\Leftrightarrow \quad & 120 \leq \bar{x}_3 - z_{0.05}\frac{2}{\sqrt{3}} \\
\Leftrightarrow \quad & 120 \text{ is not in the } 95\% \text{ one-tailed c.i. for } \mu
\end{aligned}
$$

because $(\bar{x}_3 - z_{0.05}\frac{2}{\sqrt{3}}, \infty)$ is a one-tailed c.i. for $\mu$

# One sample tests for the mean: summary

- $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n \sim F$ with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$

  **Question:** how consistent is the dataset with the null hypothesis that $\mu = \mu_0$

  - expected level over the population given blood measurement levels over $n$ persons
  - expected accuracy over the distribution given results on $n$ test instances for a classifier

- $H_0 : \mu = \mu_0 \quad H_1 : \mu \neq \mu_0$ (or $H_1 : \mu > \mu_0$, or $H_1 : \mu < \mu_0$)
- We distinguish a few cases:
  - Normal data $F = \mathcal{N}(\mu, \sigma^2)$
    - with known variance: $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$ *[z-test]*
    - with unknown variance: $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ *[t-test]*
  - General data (with unknown variance)
    - large sample, i.e., large $n$, $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ *[t-test]*
    - symmetric distribution *[Wilcoxon test]*
    - bootstrap t-test
  - Bernoulli data $F = Ber(\mu)$
    - Test of proportions : $B^\star = \frac{\bar{X}_n - \mu_0}{\sqrt{\mu_0(1-\mu_0)}/\sqrt{n}}$ *[Binomial test]*

**See R script**

# Misues of $p$-values

Misinterpretations of p-values, [Greenland et al, 2016]

- ~~The p-value is the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.~~ A p-value indicates the degree of compatibility between a dataset and a particular hypothetical explanation
- ~~The 0.05 significance level is the one to be used:~~ No, it is merely a convention. There is no reason to consider results on opposite sides of any threshold as qualitatively different.
- ~~A large p-value is evidence in favor of the test hypothesis:~~ A p-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller p-values
- ~~If you reject the test hypothesis because $p \leq 0.05$, the chance you are in error is 5%:~~ No, the chance is either 100% or 0%. The 5% refers only to how often you would reject it, and therefore be in error.

# Two sample tests for the mean: summary

- $x_1, \ldots, x_n$ realizations of $X_1, \ldots, X_n \sim F_1$ with $E[X_i] = \mu_1$ and $Var(X_i) = \sigma_X^2$
- $y_1, \ldots, y_m$ realizations of $Y_1, \ldots, Y_m \sim F_2$ with $E[Y_i] = \mu_2$ and $Var(Y_i) = \sigma_Y^2$

**Question:** how consistent is the dataset with the null hypothesis that $\mu_1 = \mu_2$

  - blood measurements over $n$ persons for control and (medical) treatment groups of patients
  - accuracy over $n$ benchmark datasets for two classifiers

- $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$ Wald test statistics: $T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{Var(\bar{X}_n - \bar{Y}_m)}} = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$
- We distinguish a few cases:
  - $F_1, F_2$ are normal distributions
    - $\sigma_X^2$ and $\sigma_Y^2$ are known *[z-test]*
    - $\sigma_X^2$ and $\sigma_Y^2$ are unknown and $\sigma_X^2 = \sigma_Y^2$ *[t-test]*
    - $\sigma_X^2$ and $\sigma_Y^2$ are unknown and $\sigma_X^2 \neq \sigma_Y^2$ *[Welch test]*
  - $F_1, F_2$ are general distributions
    - Large sample *[t-test]*
    - $F_1(x - \Delta) = F_2(x)$ location-shift *[Wilcoxon test]*
    - Bootstrap two sample test
  - Bernoulli data *[test of proportions]*
  - Paired data *[paired t-test]*

# Two sample tests for proportions

- $X_1, \ldots, X_n \sim Ber(\mu_1)$ and $Y_1, \ldots, Y_m \sim Ber(\mu_2)$
- $H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$
- Large sample *[prop.test]*
  - $\bar{W}_{n+m} = (X_1 + \ldots + X_n + Y_1 + \ldots + Y_m)/(n+m)$ the overall average
  - Test statistics when $H_0$ is true

$$Z = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\bar{W}_{n+m}(1-\bar{W}_{n+m})}\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim \mathcal{N}(0,1)$$

  - $z$ value is $\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\bar{w}_{n+m}(1-\bar{w}_{n+m})}\sqrt{\frac{1}{n}+\frac{1}{m}}}$ and $p$-value $p = P(|Z| \geq |z|) = 2(1 - \Phi(|z|))$
- **Fisher exact test** (based on odds ratio) for small samples *[fisher.test]*

**See R script**

# Paired data

- Datasets $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are measurement for the same experimental unit
  - unit: a person before and after a (medical) treatment
  - unit: a dataset/fold used to train two different classifiers
- The theory is essentially based on taking differences $x_1 - y_1, \ldots, x_n - y_n$ and thus reducing the problem to that of a one-sample test.
- $H_0 : \mu_1 = \mu_2 \Rightarrow H_0 : \mu_1 - \mu_2 = 0$
- Advantage: better power / lower Type II risk of the test w.r.t. unpaired version
  - $P_{paired}(p \leq \alpha | H_1) \geq P_{unpaired}(p \leq \alpha | H_1)$

**See R script**

# Optional references

- On confidence intervals and statistical tests (with R code)


Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)
Nonparametric Statistical Methods.
3rd edition, *John Wiley & Sons, Inc.*


- On p-values


Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman (2016)
**Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations**.
European Journal of Epidemiology 31, pages 337–350