Hypothesis testing for two samples

Giorgio Corani - (IDSIA, SUPSI)

Bayesian Data Analysis and Probabilistic Programming

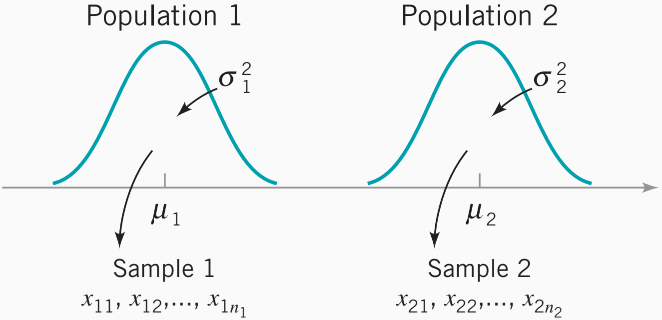
TODO

- in fondo: inserire analisi frequentista delle due proporzioni beauty and gender
- nel notebook aggiungere analisi bayesiana dell'esempio
- discutere che analisi frequentista equivale a prior uniforme?
- aggiungere anche posterior predictive check?

Credits

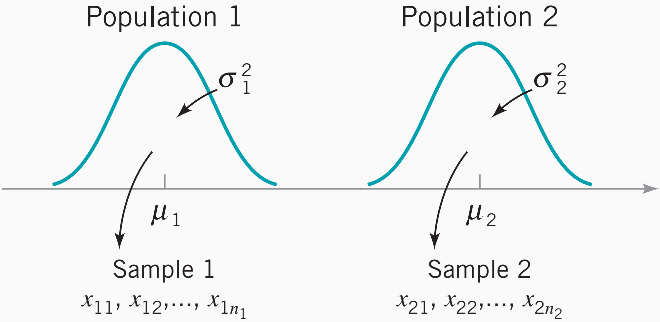
- The examples are mostly from D. P. Montgomery, *Introduction to Statistical Process Control*, 6th Edition, Wiley.

How to compare two populations



- The first population has mean μ_1 and variance σ_1^2.
- The second population has mean μ_2 and variance σ_2^2.

How to compare two populations



- The sample sizes are n_1 e n_2.
- We assume the samples of the populations to be *independent* from each other.

The assumption of equal variances

- We assume $\sigma_1^2 = \sigma_2^2$.
- This allows estimating σ^2 as a weighted average of s_1^2 e s_2^2. This is generally more accurate than estimating the two variances independently.

Comparing the mean of two populations

- The two-tailed test is:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 \\
H_1 &: \mu_1 \neq \mu_2
\end{aligned}
$$

Comparing the mean of two populations

We have:

- $\bar{x}_1$ e $\bar{x}_2$: empirical means of the two samples
- s_1^2 e s_2^2: empirical variances of the two samples.

Sampling distribution of $\bar{x}_1 - \bar{x}_2$

- It is the distribution of $\bar{x}_1 - \bar{x}_2$ if we extract many times two samples of size n_1 e n_2 from the two populations.

- Assuming
  - $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
  - n_1 and n_2 >15-20 (for the normality of $\bar{x}_1$ e $\bar{x}_2$):

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Test statistic, assuming σ to be known

- Given:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

under H_0 we have:

$$\frac{\bar{x}_1 - \bar{x}_2 - \overbrace{(\mu_1 - \mu_2)}^{\text{ipotizzato 0 in } H_0}}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

- Yet, σ is. unknown and we cannot use this statistic.

t-test

- The statistic of the t is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- which follows a t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.
- note that s_p replaces σ in the statistic

Pooled variance

- In order to estimate σ^2 we use a weighted average of s_1^2 and s_2^2:

$$s_P^2 = \frac{(n_1 - 1)}{n_1 + n_2 - 2} \cdot s_1^2 + \frac{(n_2 - 1)}{n_1 + n_2 - 2} \cdot s_2^2$$
$$s_P = \sqrt{s_P^2}$$

- s_P^2: *pooled* variance
  - the weight are in practice proportional to the sample sizes (actually, they are proportional to the degrees of freedom).
  - if $n_1 = n_2$, s_p^2 is the simple mean of s_1^2 and s_2^2.

The test statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

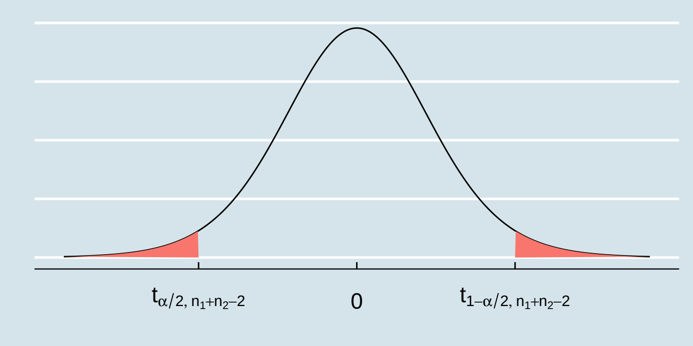
- $\bar{x}_1 - \bar{x}_2$ is the sample estimate of $\mu_1 - \mu_2$.
- $s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the *standard error* of $\bar{x}_1 - \bar{x}_2$, i.e., a measure of how the estimate $\bar{x}_1 - \bar{x}_2$ is spread around the actual value of $\mu_1 - \mu_2$

Rejection region

- The rejection region is a set of values of the test statistic for which the null hypothesis is rejected.

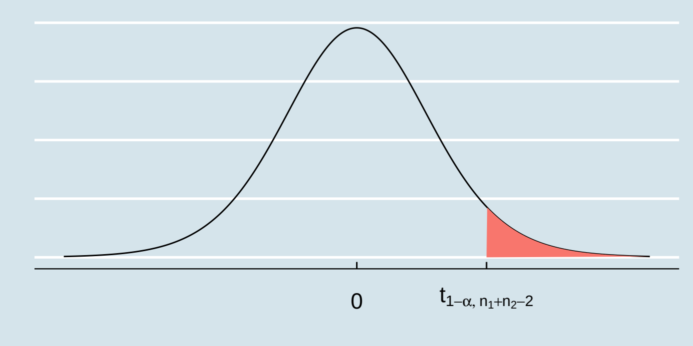
Two-tailed test

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$
  - Rejection region: $t_0 < t_{n_1+n_2-2,\alpha/2}$ e $t_0 > t_{n_1+n_2-2,1-\alpha/2}$
  - Each tails contains probability $\alpha/2$



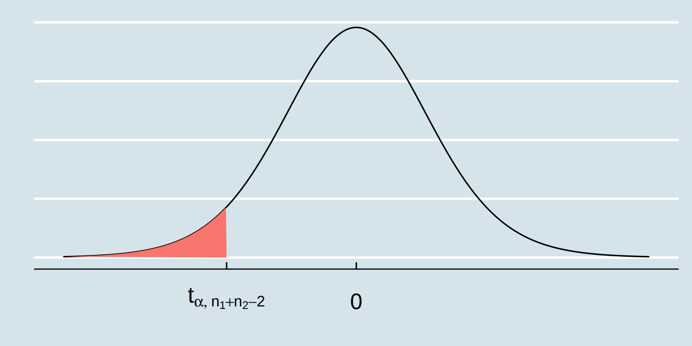
Right-tailed test

- $H_0 : \mu_1 \leq \mu_2$
- $H_1 : \mu_1 > \mu_2$
    - The most *positive* values of the statistic support H_1.
    - Rejection region: $t_0 > t_{n_1+n_2-2,\alpha}$ (contains probability α).



Left-tailed test

- $H_0 : \mu_1 \geq \mu_2$
- $H_1 : \mu_1 < \mu_2$
    - The most *negative* values of the statistic support H_1.
    - Rejection region: $t_0 < -t_{n_1+n_2-2,\alpha}$ (contains probability α)



Example: comparing mean yields of catalysts

- Two catalysts are being compared: catalyst 1 is currently in use, but catalyst 2 is acceptable.
- Catalyst 2 is cheaper: it should be adopted, providing it does not change the process yield.
- An experiment is run in the pilot plant and results are in the next slide. Is there any difference between the mean yields?
- The two-tailed test is:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Comparing mean yields of catalysts

- $n_1 = n_2 = 8$
- $\bar{x}_1$ = 92.25, s_1 =2.39
- $\bar{x}_2$ = 92.73, s_2 =2.98
- We adopt α=0.05.

Comparing mean yields of catalysts

- Since $n_1 = n_2$, s_p^2 is the average of s_1^2 and s_2^2:

$$s_p^2 = \frac{7}{14}s_1^2 + \frac{7}{14}s_2^2 = \frac{2.39^2 + 2.98^2}{2} = 7.3$$
$$s_p = \sqrt{s_p^2} = \sqrt{7.3} = 2.7$$

Statistic and critical values

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

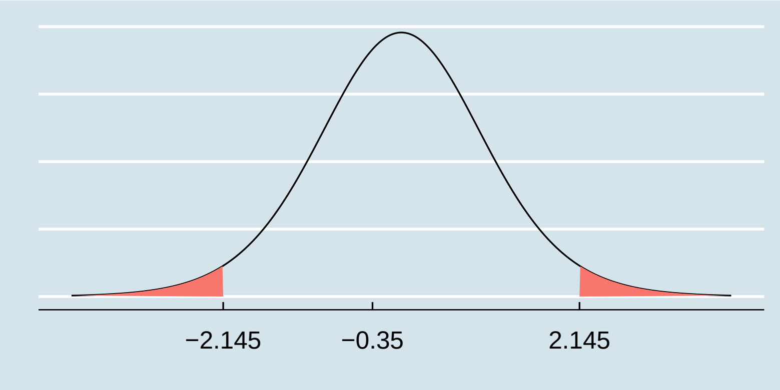
$$= \frac{92.25 - 92.73}{2.7\sqrt{\frac{1}{8} + \frac{1}{8}}}$$

$$= -0.35$$

- The critical values are $\pm t_{.975,14} = \pm 2.145$.

Decision

- The statistic is in *non-rejection* region: we do not have strong evidence that the mean yield of the two catalysts is different.



Confidence interval (CI) of $\mu_1 - \mu_2$

- The CI contains the plausible values of $\mu_1 - \mu_2$:

$$\bar{x}_1 - \bar{x}_2 \pm t_{1-\alpha/2, n_1+n_2-2} \cdot s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- $t_{1-\alpha/2, n_1+n_2-2}$: quantile $(1-\alpha/2)$ of the t distributon with $(n_1 + n_2 - 2)$ degrees of freedom; it is the critical value of the test.

- $s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ is the standard error of $(\bar{x}_1 - \bar{x}_2)$

CI vs hypothesis test

- If the hypothesis $\mu_1 = \mu_2$ is plausible given the data:
  - the two-tailed test does not reject H_0
  - the CI contains 0.
- If the hypothesis $\mu_1 = \mu_2$ is *not* plausible:
  - the two-tailed test *rejects* H_0
  - the CI does not contain 0.

Confidence interval (CI)

- The degrees of freedom are 8-1+8-1 = 14

$$\bar{x}_1 - \bar{x}_2 \pm t_{.975,14} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(92.25 - 92.73) \pm 2.145 \cdot 2.7 \sqrt{\frac{1}{8} + \frac{1}{8}} = (-3.38, 2.42)$$

- 0 is a plausible value for $\mu_1 - \mu_2$, as it is within the CI.
- Indeed the test does not refuse H_0.

Example of one-tailed test

- A study reports the weight of calcium in standard cement and cement doped with lead, after a stress test.
- Reduced levels of calcium imply low hydration in the cement, possibly allowing water to attack various the cement structure.
- The dopes cement is more expensive, and we want evidence (α=0.05) of its higher performance compared to standard cement.

One-tailed test

- The alternative hypothesis H_1 is what we try to demonstrate (*doped* cement has higher performance).

$$H_0 : \mu_{\text{standard}} \geq \mu_{\text{doped}}$$
$$H_1 : \mu_{\text{standard}} < \mu_{\text{doped}}$$

Data

$$n_{\text{standard}} = 10$$
$$\bar{x}_{\text{standard}} = 87.0$$
$$s_{\text{standard}} = 5.0$$

$$n_{\text{doped}} = 15$$
$$\bar{x}_{\text{doped}} = 90.0$$
$$s_{\text{doped}} = 4.0$$

Test statistic

$$s_P^2 = \frac{9 \cdot (5)^2 + 14 \cdot (4)^2}{10 + 15 - 2} = 19.52$$
$$s_P = \sqrt{19.52} = 4.4$$

The statistic is:

$$t_0 = \frac{\bar{x}_{\text{standard}} - \bar{x}_{\text{doped}}}{s_p \sqrt{\frac{1}{n_{\text{standard}}} + \frac{1}{n_{\text{doped}}}}}$$
$$= \frac{87 - 90}{4.4\sqrt{\frac{1}{10} + \frac{1}{15}}} = -1.67$$

Rejection region

- If $\bar{x}_{\text{doped}} > \bar{x}_{\text{standard}}$, the statistic is negative and thus in favor of H_1.
- The precise criterion is that we reject H_0 if $t_0 < t_{0.05,23} = -1.71$.
- The statistic (-1.67) is in rejection region: there is no strong evidence that *doped* cement performs better than standard cement.

Comparing two proportions

Hypothesis test for two proportions

- We want to check whether π_1 e π_2, the proportion of successes in two populations, are significantly different.
- We observe the *sample* proportion of successes, $p_1 = \frac{X_1}{n_1}$ and $p_2 = \frac{X_2}{n_2}$, while π_1 and π_2 cannot be observed.
- The term success and failure refer to the outcome being 1 or 0.
- If both samples contains at least 5 successes and failures, then p_1 and p_2 are approximately normally distributed. We use this approximation in order to define the sampling distribution of the statistic.

Comparing two proportions

- We have two samples of size n_1 e n_2, containing X_1 and X_2 successes.
- The two-tailed test is:

$$
\begin{aligned}
H_0 &: \pi_1 = \pi_2 \\
H_1 &: \pi_1 \neq \pi_2
\end{aligned}
$$

The test statistic

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

with :

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

- Under H_0, the statistic $Z \sim N(0,1)$

The test statistic

- If we extract many times two samples of size n_1 and n_2 from two populations with $\pi_1 = \pi_2$ and compute the statistic, it will be different every time, following approximately a $N(0, 1)$ distribution.
- It the test is one-tailed the statistics remains the same, but the rejection region changes.

Rejection regions

H_1	Rejection region	p-value
$\pi_1 \neq \pi_2$	$z < z_{\alpha/2}$ e $z > z_{1-\alpha/2}$	$2(1 - \Phi(\lvert z \rvert))$
$\pi_1 > \pi_2$	$z > z_{1-\alpha}$	$1 - \Phi(z)$
$\pi_1 < \pi_2$	$z < z_{\alpha}$	$\Phi(z)$

CI of $\pi_1 - \pi_2$

- If we extract many times two samples of size n_1 and n_2 from two populations with $\pi_1 = \pi_2$ and compute the CI, it will contain the actual value of $\pi_1 - \pi_2$ in $(1 - \alpha)$ of the experiments.
- The CI is:

$$p_1 - p_2 \pm z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- The CI and the two-tailed test approximate differently the standard error and they might sometimes draw inconsistent conclusions.

Esempio: valutare l'efficacia di un new

- Per valutare l'efficacia di un new si svolge un *randomized trial*.
- In modo casuale ad alcuni pazienti viene somministrato il new; ad altri il old.
- Alla fine del periodo di cura, è necessario analizzare se c'è una differenza statisticamente significativa fra i due gruppi.

Example: assess the effect of a process innovation

- From a traditional production process we have 262 boards, of which 154 without any defect.
- From a innovative production process we have 227 boards, of which 163 without any defect.
- Is the new process significantly more accurate than the previous one?

Comparing the two proportions

- Both samples contain more than 5 successes and 5 failures; the normal approximation is sound.
- The test is:

$$H_0 : \pi_{\text{new}} \leq \pi_{\text{old}}$$
$$H_1 : \pi_{\text{new}} > \pi_{\text{old}}$$

- and we use $\alpha = 0.01$.

Comparing the two proportions

- The rejection region contains positive values of $p_{\text{new}} - p_{\text{old}}$ and thus also of the statistic.
- Rejection region: $Z_0 > \Phi^{-1}(.99)$ = 2.33

Comparing the two proportions

$$
\begin{aligned}
p_{\text{new}} &= 163/227 = 0.72 \\
p_{\text{old}} &= 154/262 = 0.59 \\
\bar{p} &= (163 + 154)/(227 + 262) = 0.65 \\
Z &= \frac{p_{\text{new}} - p_{\text{old}}}{\sqrt{\bar{p} \cdot (1 - \bar{p}) \cdot 1/n}} = 3.01 > 2.33
\end{aligned}
$$

- The statistic is in rejection region.

Beauty and sex ratio

Chap. 9.4 of “Regression and other stories”, A. Gelman, J. Hill, A. Vehtari

- Book published from Cambridge University Press (2020).
- The book is also freely available online.

- Keep in mind these example, we will re-analyze the data in a later lecture using a Bayesian approach.

Beauty and sex ratio

- Some years ago a researcher analyzed data from a survey of 3000 Americans and observed a correlation between attractiveness of parents and the sex of their children.
- He considers 3000 couple of parents, among which 300 classified as highly attractive.
- The proportion of girls among the children of “standard” parents is 48% (X=1296, n=2700).
- The proportion of girls among the children of “highly attractive” parents is 56% (X=168, n=300).

Is the difference significant?

The test is:

$$
\begin{aligned}
H_0 &: \pi_{\text{attr}} \leq \pi_{\text{std}} \\
H_1 &: \pi_{\text{attr}} > \pi_{\text{std}}
\end{aligned}
$$

Is the difference significant?

$$
\begin{aligned}
p_{\text{attr}} &= 0.56 \\
p_{\text{std}} &= 0.48 \\
\bar{p} &= \frac{168 + 1296}{300 + 2700} = 0.488 \\
\text{sd err} &= \sqrt{\bar{p} \cdot (1 - \bar{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})} = 0.03 \\
Z &= \frac{p_{\text{new}} - p_{\text{old}}}{\text{sd err}} = 2.63 > 2.33,
\end{aligned}
$$

where 2.33 is the 99-th quantile (α=0.01).

- The difference in the proportion of girls between the two groups is statistically *significant*.