Statistical Methods for Data Science

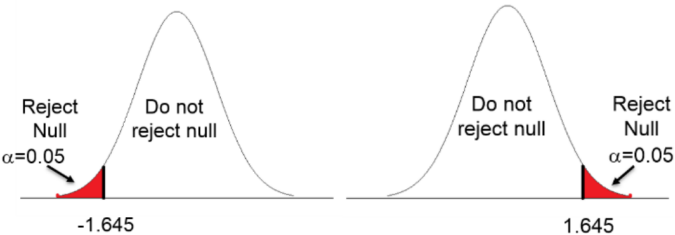
Lesson 21 - One-sample t-test and application to linear regression.

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

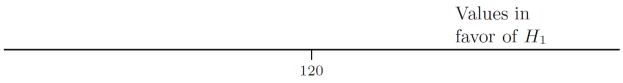
Statistical test of hypothesis: one-tailed

- 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. $P(T < c_l) = \alpha$ (resp. c_u s.t. $P(c_u < T) = \alpha$) *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $t < c_l$ (resp. $c_u < t$): H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected



Example: speed limit

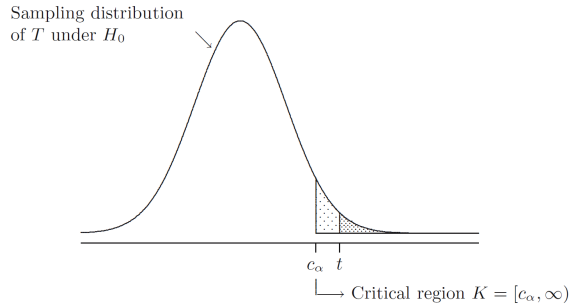
- Speed limit: 120 Km/h
- A device conduts 3 measurements: $X_1, X_2, X_3 \sim N(\mu, 4)$ (true speed + measur. error)
- Based on $T = \bar{X}_3 = (X_1 + X_2 + X_3)/n3 \sim N(\mu, 4/3)$:
  - if $T > c_u$ the drived is fined
  - otherwise it is not
- What should c_u be to unjustly fine only 5% of drivers? *[Type I error]*
- One-tailed statistical test
  - H_0: $\mu = 120$ (null hypothesis)
  - H_1: $\mu > 120$ (alternative hypothesis)
  - $\alpha = 0.05$ (significance level), or $100(1-\alpha)\% = 95\%$ (confidence level)
  - $T = \bar{X}_3$ (test statistics)
- Assuming H_0 is true, find t such that $P(T \geq c_u) = 0.05$



Example: speed limit

- $X_1, X_2, X_3 \sim N(\mu, 4)$ and then $T = \bar{X}_3 \sim N(\mu, 4/3)$
- $Z = \frac{T-120}{2/\sqrt{3}} \sim N(0,1)$
- $P(T \geq c_u) = P(\frac{T_3-120}{2/\sqrt{3}} \geq \frac{c_u-120}{2/\sqrt{3}}) = P(Z \geq \frac{c_u-120}{2/\sqrt{3}})$
- Right critical value: $P(Z \geq z_\alpha) = \alpha$
- Hence $\frac{c_u-120}{2/\sqrt{3}} = z_{0.05}$, i.e., $c_u = 120 + z_{0.05}\frac{2}{\sqrt{3}} = 121.9$
- In summary, for $\alpha = 0.05$ we should reject $H_0 : \mu = 120$ in favor of $H_1 : \mu > 120$ if the observed (average) speed t is $t \geq 121.9$

Critical values and p-values



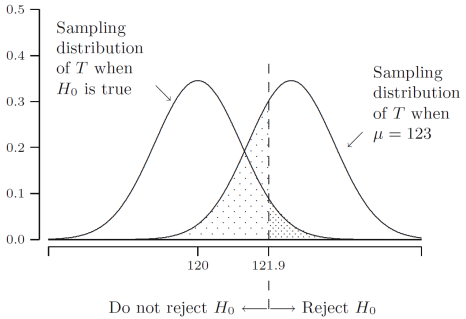
- *Critical region* K: the set of values that reject H_0 in favor of H_1 at significance level α
- *Critical values*: values on the boundary of the critical region
- *p-value*: the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that H_0 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: we falsely reject H_0 *[α-risk, false positive rate]*
  - E.g., unjust fine
  - Type I error is equal to α
- Type II error: we falsely do not reject H_0 *[β-risk, false negative rate]*
  - E.g., lack of a true fine
  - How large is type II error?

Type II error



- Type II error: probability of not being fined when $\mu > 120$ but $t < 121.9$
- Assume $\mu = 125$, hence $T = \bar{X}_3 \sim N(125, 4/3)$
  - Type II error is $P(T < 121.9 | \mu = 125) = P(\frac{T-125}{2/\sqrt{3}} < \frac{121.9-125}{2/\sqrt{3}}) = \Phi(-2.68) = 0.0036$
- Assume $\mu = 123$, hence $T = \bar{X}_3 \sim N(123, 4/3)$
  - Type II error is $P(T < 121.9 | \mu = 123) = P(\frac{T-123}{2/\sqrt{3}} < \frac{121.9-123}{2/\sqrt{3}}) = \Phi(-0.95) = 0.1711$
- Type II error can be arbitrarily close to $1 - \alpha$

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 with 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 μ

Statistical tests for the mean

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ (or $H_1 : \mu > \mu_0$, or $H_1 : \mu < \mu_0$)
- Normal data
  - 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

Normal data with known σ^2: z-test

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ *[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]*
- $Z = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}}$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and z value is $\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$
- $P(Z \leq -z_{\alpha/2}) = \alpha/2$ and $P(Z \geq z_{\alpha/2}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|z| \geq z_{\alpha/2}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

Normal data with unknown σ^2: t-test

- $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$ *[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]*
- $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}}$ test statistics when H_0 is true
- $x_1, \ldots, x_n$: observed dataset, and t value is $\frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$
- $P(T \leq -t_{\alpha/2,n-1}) = \alpha/2$ and $P(T \geq t_{\alpha/2,n-1}) = \alpha/2$ *[Critical values]*
- Output of the test at confidence level $100(1-\alpha)\%$ using critical values
  - $|t| \geq t_{\alpha/2,n-1}$: H_0 is rejected *[Critical region]*
  - otherwise: H_0 cannot be rejected

See R script

General data, large sample: t-test

- $T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \to N(0,1)$ for $n \to \infty$ *[Variant of CLT]*
- We can use z-test with $\sigma^2 = S_n^2$
- Or, since $t(n) \to N(0,1)$ for $n \to \infty$, we can use t-test directly!

See R script

General data, symmetric distribution: Wilcoxon test

- $X_1, \ldots, X_n \sim F$ with $f(\mu - x) = f(\mu + x)$
- $H_0 : \mu = 67$
- $H_1 : \mu \neq 67$
- $W = \min\{\sum rank^+, \sum rank^-\}$, with ranking w.r.t. $|x_i - \mu_0|$

x	71	79	40	70	82	72	60	76	69	75
$x - \mu_0$	4	12	-27	3	15	5	-7	9	2	8
$rank$	3	8	10	2	9	4	5	7	1	6
$rank^+$	3	8		2	9	4		7	1	6
$rank^-$			10				5			

- $w = \min\{40, 15\} = 15$
- Ignore cases where $|x_i - \mu_0| = 0$. If the values have tied, then consider the mean value.
- Normal approximation for $n > 50$
- Exact test for $n \leq 50$
- In general, a statistical test of the median!

See R script

General data: bootstrap test

`boot.ci` method in R confidence intervals:

- `type='stud'`: $(\bar{x}_n - q_{1-\alpha/2}\frac{s_n}{\sqrt{n}}, \bar{x}_n - q_{\alpha/2}\frac{s_n}{\sqrt{n}})$ with quantiles over the distribution of t^*

EMPIRICAL BOOTSTRAP SIMULATION FOR THE STUDENTIZED MEAN.
Given a dataset $x_1, x_2, \ldots, x_n$, determine its empirical distribution function F_n as an estimate of F. The expectation corresponding to F_n is $\mu^* = \bar{x}_n$.

1. Generate a bootstrap dataset $x_1^*, x_2^*, \ldots, x_n^*$ from F_n.
2. Compute the studentized mean for the bootstrap dataset:

$$t^* = \frac{\bar{x}_n^* - \bar{x}_n}{s_n^*/\sqrt{n}},$$

where $\bar{x}_n^*$ and s_n^* are the sample mean and sample standard deviation of $x_1^*, x_2^*, \ldots, x_n^*$.

Repeat steps 1 and 2 many times.

- $t_0 = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$ r number of repetitions
- one-sided p-value, i.e., $P(T \geq t_0)$, estimated as $|\{i = 1, \ldots, r \mid t_i^* \geq t_0\}|/r$
- two-sided p-value, i.e., $P(|T| \geq |t_0|)$, estimated as $|\{i = 1, \ldots, r \mid |t_i^*| \geq |t_0|\}|/r$

See R script

Hypothesis testing in linear regression

- Simple linear regression: $Y_i = \alpha + \beta x_i + U_i$ with $U_i \sim \mathcal{N}(0, \sigma^2)$
- We have $\hat{\beta} \sim \mathcal{N}(\beta, Var(\hat{\beta}))$ where $Var(\hat{\beta}) = \sigma^2/SXX$ is unknown *[prove it!]*
- The studentized statistics is $t(n-2)$-distributed: *[proof omitted]*

$$T = \frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} \sim t(n-2)$$

- $H_0 : \beta = 0 \qquad H_1 : \beta \neq 0$
- p-value is $p = P(|T| > |t|) = 2 \cdot P(T > \left|\frac{\hat{\beta}-0}{se(\hat{\beta})}\right|)$
- H_0 can be rejected in favor of H_1 at $\alpha = 0.05$, if $p < 0.05$, or, equivalently, if $|t| > t_{n-2,0.025}$.
- A similar approach applies to the intercept.

See R script