Recap of Hypothesis Testing (one sample)

Giorgio Corani - (IDSIA, SUPSI)

Bayesian Data Analysis and Probabilistic Programming

Cosa è un'ipotesi statistica

- A statistical hypothesis is an assumption about the parameter of a distribution.
- For instance, we can test whether the mean diameter μ of the production of manufactured cylinders equals the nominal value of 5 cm:

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

- H_0 è is the *null* hypothesis, H_1 is the alternative hypothesis.

Hypothesis test

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

- H_0 and H_1 contain all the possible values of μ.

Hypothesis test

- H_0 e H_1 refer to a parameter of a distribution, such as:
  - the mean μ of a normal
  - the proportion π of a binomial
- We make no test about sample mean and sample proportion, ($\bar{x}$, p, etc), which are measured on the sample and there is no uncertainty about them.
- The uncertainty is about the parameter of the distribution / population.

Strong and weak decision

- We try to prove the alternative hypothesis.
- Rejecting the null hypothesis is a *strong* decision: the sample provides strong evidence about the null hypothesis being false.
- Non-rejecting the null hypothesis is a *weak* decision
  - By non-rejecting H_0 we do not conclude that H_0 is true. We rather conclude that there is not enough evidence to reject it.
  - A frequentist test cannot reject H_0.

Type-1 error: rejecting H_0 when H_0 is true.

- α: probabilità di fare un errore di tipo I.

- Typical values for α=0.05 o α=0.01.
- α: *significance* of the test (e.g., 0.05, 0.01)
- $1 - \alpha$: *confidence* of the test (e.g., 0.95, 0.99)
- If the test rejects H_0, we say that a statistically significant effect has been observed with confidence $1 - \alpha$.

Type II error: not rejecting H_0 when it is false.

- This corresponds to letting free
- The type II error (β) is controlled by using a large enough sample. In order to decide the sample size, the *power* of the test has to be studied.

Esempio di test di uguaglianza (“test a due code”)

- La media della nostra produzione deve essere esattamente a 5cm.
- Questo test può scoprire sregolazioni della media verso l’alto o il basso:

$$H_0 : \mu = 5$$
$$H_1 : \mu \neq 5$$

- Se il test non rifiuta H_0 (decisione *debole*), la produzione continua regolarmente.
- Se il test rifiuta H_0 (decisione *forte*), concludiamo che la produzione si sia sregolata; è necessario intervenire sul processo.

Test direzionali (“a una coda”)

- In alcune applicazione ci interessano gli spostamenti del parametro in un’unica direzione (solo verso l’alto o il basso).

Example of one-tailed test (from Montgomery, chap 4)

- A system manager wants to know whether the mean response time of a computer network command exceeds 75 millisec. If that is the case, there is evidence that the network is not working at its nominal speed.

The test is:

$$H_0 : \mu \leq 75$$
$$H_1 : \mu > 75$$

- If we reject H_0, we conclude the network to be *significantly* slower than its nominal speed.

Example of one-tailed test

- If instead the provider wants to demonstrate that the network is even faster than its nominal value, he will use the test:

$$H_0 : \mu \geq 75$$
$$H_1 : \mu < 75$$

- If the test rejects H_0, we conclude the network to be *significatnly* faster than its nominal speed.

Equals sign in H_0 e H_1

In general:

- H_0 contains either $=$, or $\geq$, or $\leq$).
- H_1 contains either $\neq$, or $<$, or $>$).

Test d'ipotesi per la media

- Vogliamo testare l'ipotesi:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

- μ: media (ignota) della variabile nella popolazione
- μ_0: valore di riferimento

Procedura

- Prendiamo un campione dalla popolazione
- Calcoliamo una statistica (*statistica del test*).
- Non rifiutiamo H_0 se il valore osservato della statistica è plausibile sotto l'ipotesi nulla (decisione **debole**).
- Rifiutiamo H_0 se il valore osservato della statistica non è plausibile sotto l'ipotesi nulla (decisione **forte**).

La statistica del test (assumendo σ nota)

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\bar{x}$: media del campione (media *campionaria*)
- n: numero di misure nel campione (*dimensione del campione*).
- σ^2: varianza della popolazione, che assumiamo di conoscere. Più avanti rilasseremo questa assunzione.
- Se H_0 è vera, $Z_0 \sim N(0, 1)$.

Intuizione

- Teorema del limite centrale:

$$\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$$

  - la normalità di $\bar{x}$ è garantita purchè n sia sufficientemente ampio.

- Questa è la distribuzione che otterremmo misurando $\bar{x}$ in tanti diversi campioni di dimensione n estratti da una popolazione con media μ e varianza σ^2.

- Se l'ipotesi nulla e' vera:

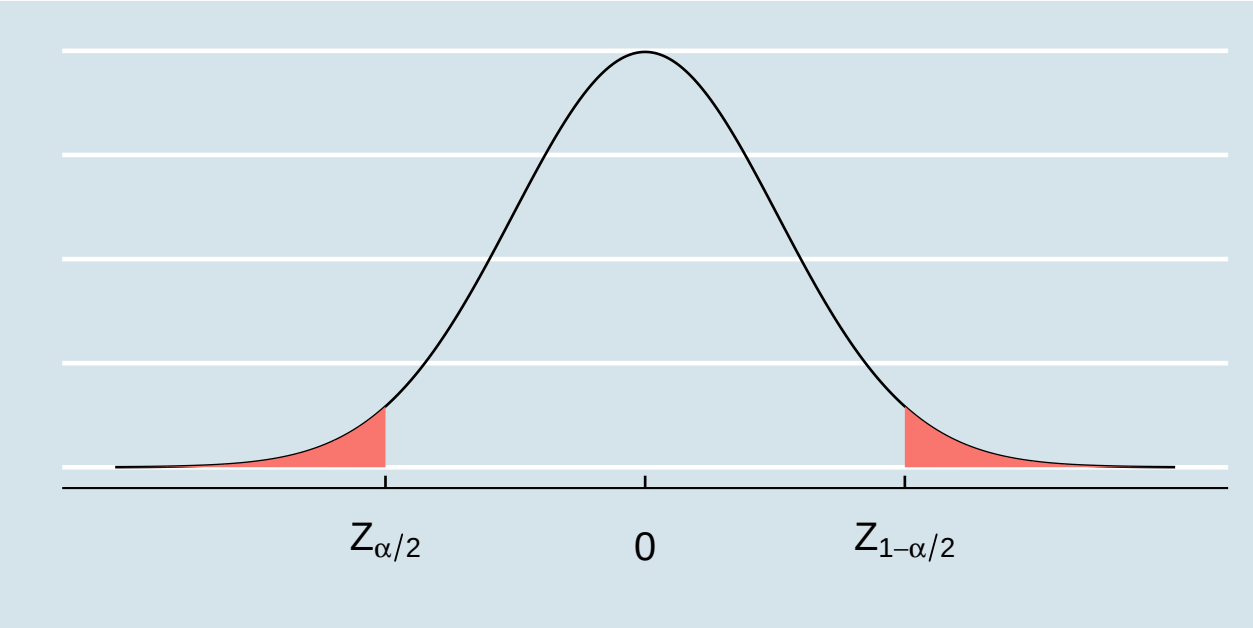
$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

.

- Per avere la normalità di Z_0 è necessario n >5-10, oppure che la popolazione sia normale.

$Z_{\alpha/2}$

- $Z_{\alpha/2}$ e $Z_{1-\alpha/2}$: percentile $\frac{\alpha}{2}$ ed $(1 - \frac{\alpha}{2})$ della normale standard.
- Per simmetria, $Z_{1-\alpha/2} = -Z_{\alpha/2}$ nel caso Gaussiano.
- Entrambe le code rosse contengono probabilità $\alpha/2$.

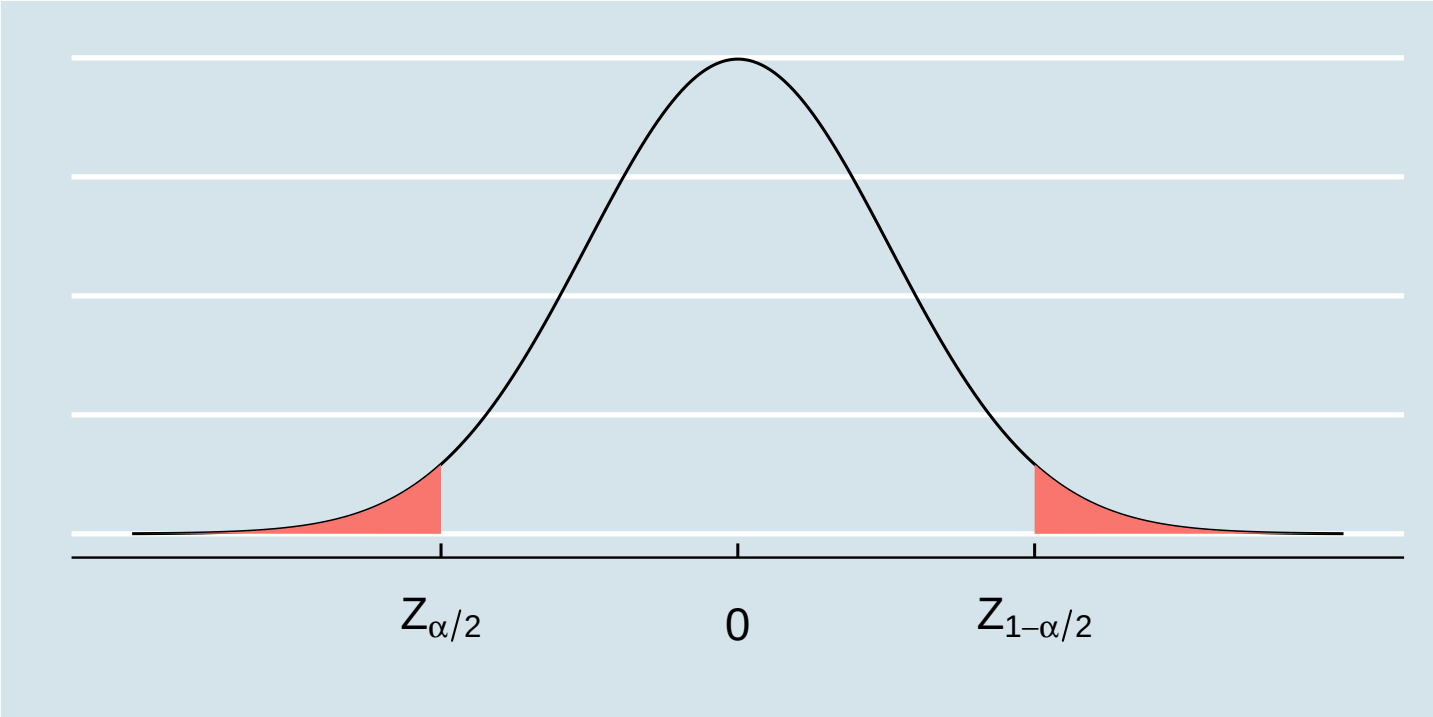


Regione di rifiuto e non-rifiuto

- Supponiamo di ripetere molte volte questo esperimento:
  - estrarre un campione di dimensione n dalla popolazione
  - calcolare Z_0
- Se l'ipotesi nulla è vera, una proporzione $(1 - a)$ dei valori di Z_0 cadrà tra $Z_{\alpha/2}$ e $Z_{1-\alpha/2}$.
- È poco plausibile che Z_0 cada fuori da questi limiti (succede solo con probabilità α).
- Se succede questo *rifiutiamo* l'ipotesi nulla.

Regione di rifiuto del test a due code

- Rifiutiamo H_0 se la statistica Z_0 cade in una delle due regioni rosse (*regioni di rifiuto*).



Regione di rifiuto ($\alpha = 0.05$)

$$\alpha = 0.05$$
$$\alpha/2 = 0.025$$
$$1 - \alpha/2 = 0.975$$
$$z_{1-\alpha/2} = \Phi^{-1}(0.975) = 1.96$$
$$z_{\alpha/2} = \Phi^{-1}(0.025) = -1.96$$

Le due regioni di rifiuto sono:

- $Z_0 > 1.96$
- $Z_0 < -1.96$

$Z_{\alpha/2}$ ($\alpha = 0.01$)

$$\alpha = 0.01$$
$$\alpha/2 = 0.005$$
$$1 - \alpha/2 = 0.995$$
$$z_{1-\alpha/2} = \Phi^{-1}(0.995) = 2.58$$
$$z_{\alpha/2} = \Phi^{-1}(0.005) = -2.58$$

Le due regioni di rifiuto sono:

- $Z_0 > 2.58$
- $Z_0 < -2.58$

Test a una coda

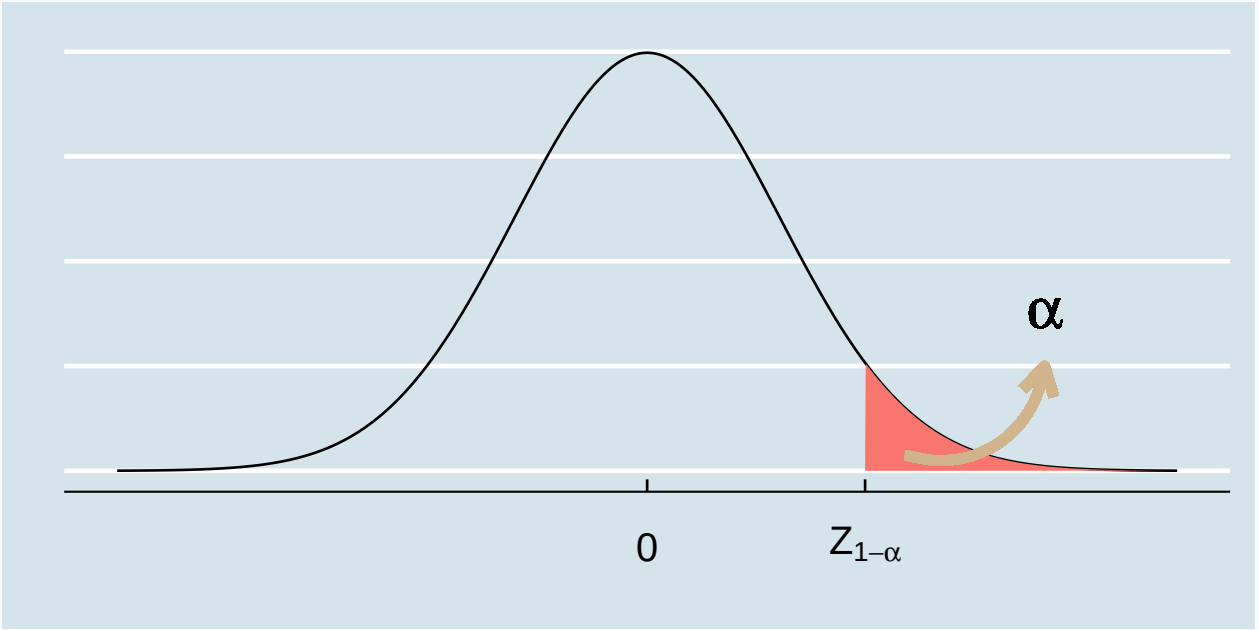
- H_1 rappresenta l'ipotesi che vogliamo dimostrare.
- Se vogliamo provare a dimostrare $\mu > \mu_0$ faremo il test:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

Regione di rifiuto (test con coda destra)

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

- Il rifiuto di H_0 richiede $\bar{x} > \mu_0$, quindi $Z_0 > 0$.
- Più precisamente la regione di rifiuto è $Z_0 > Z_{1-\alpha}$.

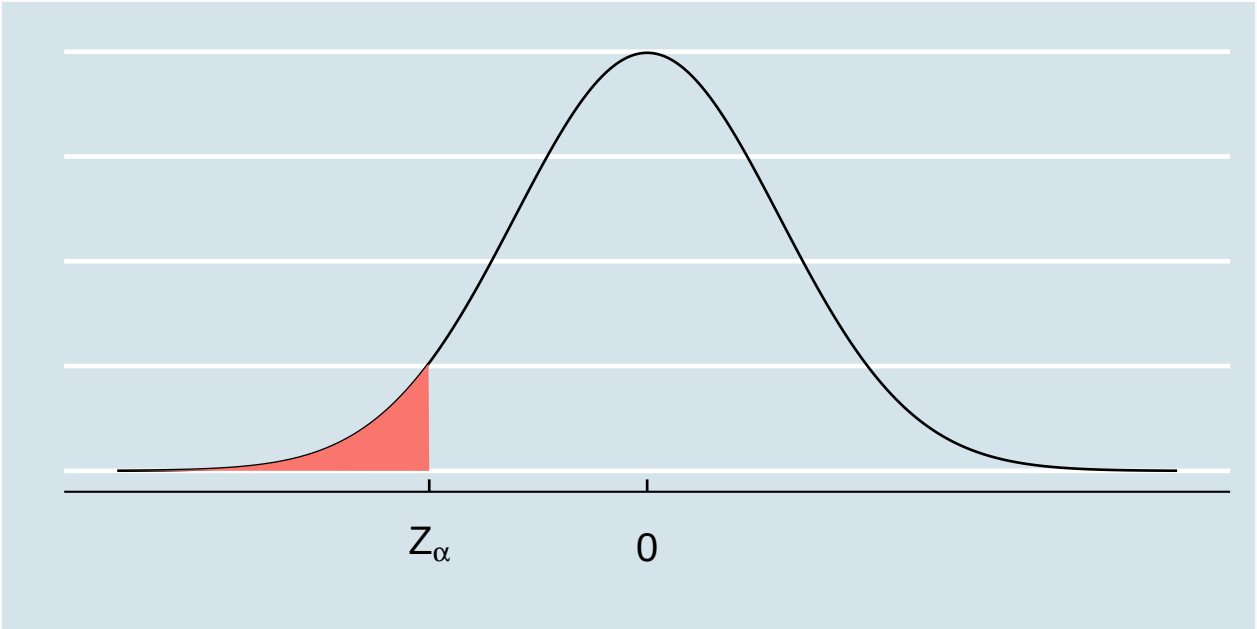


Regione di rifiuto (test con coda sinistra)

- Invece se vogliamo dimostrare $\mu < \mu_0$ faremo il test:

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- Rifiutiamo H_0 se $Z_0 < Z_\alpha$ (ricordiamo che $Z_\alpha = -Z_{1-\alpha}$).



$$\alpha = 0.05$$
$$z_\alpha = \Phi^{-1}(0.05) = -1.64$$
$$z_{1-\alpha} = \Phi^{-1}(0.95) = 1.64$$

$$\alpha = 0.01$$
$$z_\alpha = \Phi^{-1}(0.01) = -2.32$$
$$z_{1-\alpha} = \Phi^{-1}(0.99) = 2.32$$

Esempio

- Da specifica, il tempo di risposta di una rete di computer è di 75 ms.
- Il cliente sospetta che la rete sia più lenta del valore di specifica. Compie 25 misure, ottenendo un tempo di risposta medio $\bar{x} = 79.25$.
- La σ dei tempi di risposta, nota da precedenti analisi, è 8 ms.
- Svolgere il test con significatività $\alpha = 0.05$.

Soluzione

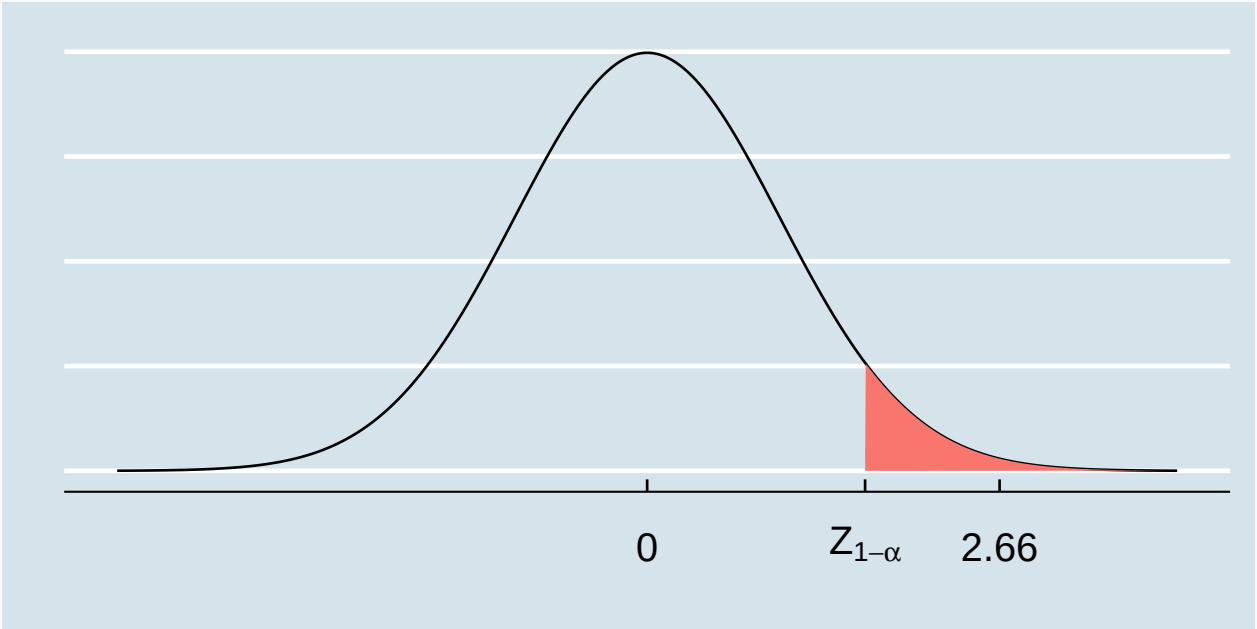
$$H_0 : \mu \leq 75$$
$$H_1 : \mu > 75$$

- La *regione di rifiuto* contiene tutti i valori superiori al *valore critico* $Z_{1-\alpha} = \Phi^{-1}(0.95) = 1.64$.

$$Z_0 = \frac{\bar{x} - 75}{8\sqrt{25}} = \frac{79.25 - 75}{8/\sqrt{25}} = 2.66$$

Esito del test

- Prendiamo una decisione forte, rifiutando H_0 e concludendo che il tempo di risposta è *significativamente* superiore a 75 millisecondi.



Esempio - viscosità dell'asfalto

- Il valore ottimale di viscosità media dell'asfalto è 3200 cps (centipoise) .
- Sulla base dell'esperienza è la viscosità può essere assunta normalmente distribuita, con $\sigma = 118$.
- Sono state svolte 15 misure, la cui media è $\bar{x} = 3210$.
- Testare la conformità della viscosità media della produzione assumendo $\alpha = 0.05$.

Soluzione

Il test è:

$$H_0 : \mu = 3200$$
$$H_1 : \mu \neq 3200$$

- I valori critici di questo test a due code sono $Z_{\alpha/2}$ e $Z_{1-\alpha/2}$, pari a ± 1.96.

Soluzione

- La statistica è:

$$Z_0 = \frac{\bar{x} - \mu_0}{8\sqrt{25}} = \frac{3210 - 3200}{118/\sqrt{15}} = 0.328$$

- Non rifiutiamo H_0 (decisione debole).

- Non abbiamo evidenza che la viscosità media della nostra produzione sia diversa dal valore di specifica.

Discussione

- La performance del processo potrebbe comunque essere insoddisfacente a causa di alta variabilità, per cui le singole unità potrebbero essere troppo lontane dalla media.

- Vedremo più avanti come valutare la performance del processo considerandone anche la variabilità (*analisi di capacità del processo*).

t-test

Test d'ipotesi con varianza ignota (t-test)

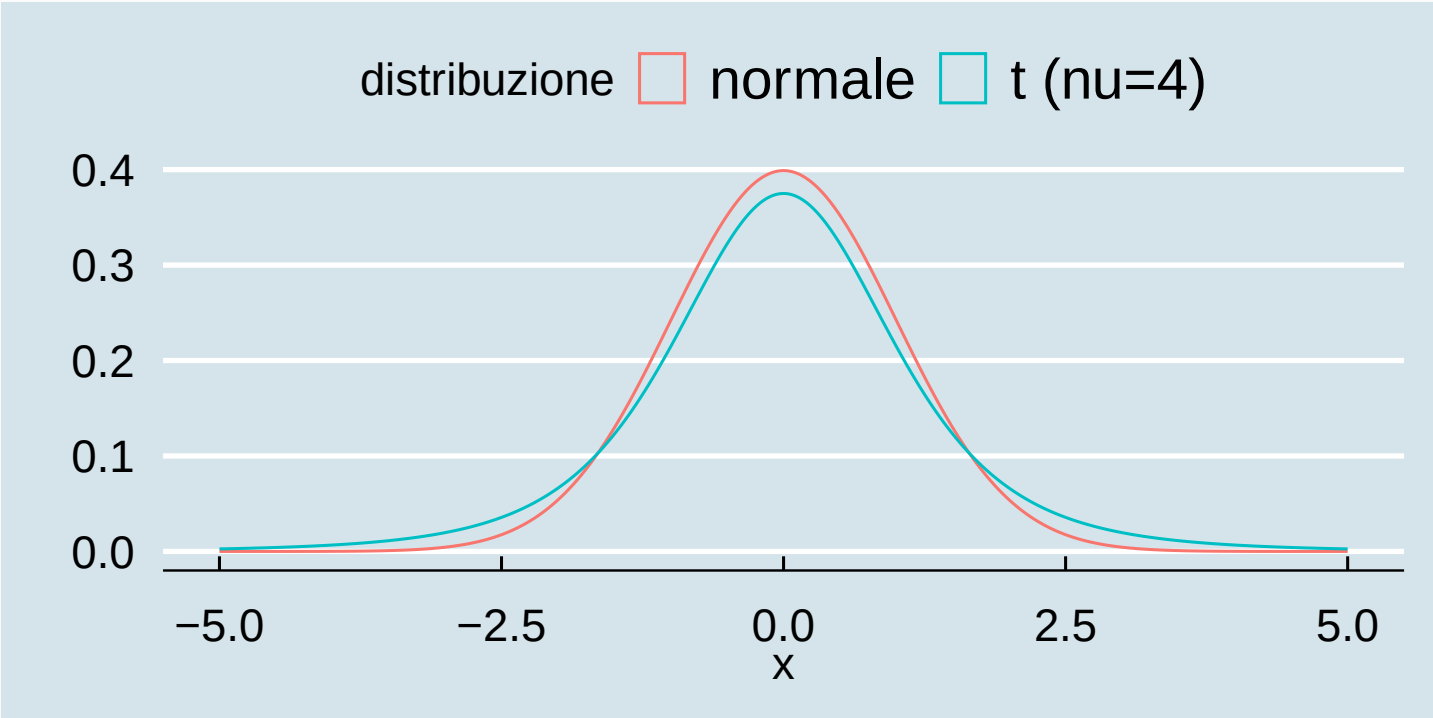
- In generale, la σ della popolazione è ignota; invece conosciamo la deviazione standard *campionaria*, s.
- Sostituendo s a σ, la statistica del test diventa:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Sotto H_0, la statistica segue una distribuzione t con $n - 1$ gradi di libertà (i gradi di libertà sono denotati da ν).

La distribuzione t (“t di Student”)

- Come la $N(0, 1)$, la distribuzione t è simmetrica e centrata in 0.
- La t ha densità non trascurabile anche oltre le 3 deviazioni standard, dove invece la densità normale è praticamente nulla.



La distribuzione t ("t di Student")

- Per ν <10, la t ha code più lunghe della $N(0,1)$; questo implica un aumento del 10-30% dei valori critici:

$$z_{.975} = 1.96$$
$$t_{\nu=4,.975} = 2.78$$
$$t_{\nu=10,.975} = 2.23$$
$$t_{\nu=30,.975} = 2.04$$

- Per ν >30 la t è praticamente equivalente ad una $N(0,1)$.

t - test

- Supponiamo di estrarre molte volte un campione di dimensione n dalla popolazione e di calcolare per ogni campione la statistica:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Il valore di t_0 calcolato per ogni campione sarà diverso e segue una distribuzione t con $n - 1$ gradi di libertà.

- I valori critici del test ($t_{n-1,\alpha/2}$, $t_{n-1,1-\alpha/2}$, $t_{n-1,1-\alpha}$ etc) si leggono in tabella (dipendono da n-1 e da α).

Esempio di t-test

- Un processo di produzione monitora la lunghezza di un componente, che da specifica è di 5.2 mm.
- Si vogliono identificare sregolazioni del processo verso l'alto e verso il basso.
- Il campione ha n=15 misure, con $\bar{x}$=4.66, s=1.52
- Calcolare test e CI usando α=0.05.

Soluzione

$$H_0 : \mu = 5.2$$
$$H_1 : \mu \neq 5.2$$

- Da tabella:
  - $t_{14,\alpha/2} = -2.145$
  - $t_{14,1-\alpha/2} = 2.145$
  - Nel caso normale (cioè se σ fosse noto), i valori critici sarebbero più vicini a 0: ± 1.96.

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{4.66 - 5.2}{1.52/\sqrt{15}} = -1.376$$

- Il test *non rifiuta* H_0: la statistica è all'interno dei valori critici.

Verifica tramite intervallo di confidenza

Il CI calcolato con lo stesso α è:

$$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} = 4.66 \pm 2.145\frac{1.52}{\sqrt{15}} = (3.82, 5.50)$$

- Il valore ipotizzato dall'ipotesi nulla, cioè 5.2, è all'interno del CI; è quindi un valore plausibile per la media della popolazione.
- Coerentemente, il t-test non rifiuta l'ipotesi nulla.

Esercizio

- Un'associazione di consumatori vuole dimostrare che le scatole di biscotti prodotte da una certa azienda siano più leggere del valore dichiarato (375 g.).
- L'associazione pesa 26 scatole, con il seguente risultato:
  - $\bar{x} = 368$ gr.
  - $s = 15$ gr.
- C'è evidenza che l'azienda stia vendendo scatole di biscotti più leggero del dichiarato?
  - Adottare $\alpha = 0.05$.

Test per la proporzione di una popolazione

- Consideriamo una popolazione in cui ogni elemento ha una caratteristica binaria:
  - conforme, difettoso
  - guarisce, non guarisce
  - preferisce sito A, preferisce sito B (AB test)
- Di solito l'esito d'interesse è quello più raro ed è detto *successo*.
- Vogliamo testare l'ipotesi che la proporzione π di successi nella popolazione sia uguale a π_0 (ipotesi nulla).

Test per la proporzione di una popolazione

- Misuriamo la proporzione campionaria $p = \frac{X}{n}$:
  - X: numero di successi nel campione
  - n: dimensione del campione
- Verifichiamo che il campione sia sufficientemente ampio:
  - $X > 5$ (numero di successi >5)
  - $n - X > 5$ (numero di insuccessi >5)

Test per la proporzione di una popolazione

Se il campione è sufficientemente ampio, possiamo approssimare con una $N(0,1)$ la distribuzione della statistica:

$$Z_0 = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

- $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$ è l'errore standard di p sotto H_0
- cioè la deviazione standard di p su infiniti campioni estratti da una popolazione binaria la cui proporzione di successi è π_0

Test per la proporzione di una popolazione

- Il test a due code è ad esempio:

$$H_0 : \pi = \pi_0$$
$$H_1 : \pi \neq \pi_0$$

- Valori critici e p-value si calcolano usando la $N(0, 1)$.

Esempio

- In un campione casuale di 899 persone che lavorano a casa ci sono 414 donne.
- È plausibile che la proporzione di donne nella popolazione di persone che lavorano da casa sia il 50%?
- Svolgere il test con $\alpha = 0.05$ e calcolare il p-value.

Soluzione

- Il campione contiene più di 5 successi ed insuccessi, quindi possiamo fare il test.
- Il test è a due code:

$$H_0 : \pi = 0.5$$
$$H_1 : \pi \neq 0.5$$

- I valori critici sono ± 1.96.

Soluzione

$$p = \frac{414}{899} = 0.46$$

$$\begin{aligned} Z_0 &= \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \\ &= \frac{.46 - .5}{\sqrt{.46(1-.46)/899}} \\ &= -2.41 \end{aligned}$$

- Rifiutiamo H_0
- p-value : $2 \cdot (1 - \Phi(|Z_0|)) = 2 \cdot 0.008 = 0.016 < \alpha$
- p-value decisamente più piccolo di α.

Esercizio: frazione di difettosi

- Prima di firmare un contratto, volete forte evidenza che la frazione di difettosi del fornitore sia <10%.
- Quindi analizzate un campione di n=250 pezzi trovandone 11 difettosi (p=11/250= 0.044).
- Svolgere il test d'ipotesi ($\alpha = 0.05$) per decidere se accettare la fornitura.

Soluzioni

p-value (rete di computer)

$$H_0 : \mu \geq 75$$
$$H_1 : \mu < 75$$

$$Z_0 = \frac{\bar{x} - 75}{8\sqrt{25}} = \frac{79.25 - 75}{8/\sqrt{25}} = 2.66$$

$$p\text{-value: } = 1 - \Phi(Z_0) = 1 - 0.996 = 0.0039$$

- Rifiutiamo H_0; il p-value è un ordine di grandezza inferiore ad α e la statistica è quindi molto oltre il valore critico.

p-value (viscosità asfalto)

$$H_0 : \mu = 3200$$
$$H_1 : \mu \neq 3200$$

$$Z_0 = \frac{\bar{x} - \mu_0}{8\sqrt{25}} = \frac{3210 - 3200}{118/\sqrt{15}} = 0.328$$

$$\begin{aligned} P\text{-value:} \quad &= 2 \cdot [1 - \Phi(Z_0)] = 2 \cdot [1 - \Phi(.328)] \\ &= 2 \cdot [1 - 0.628] = 0.743 \end{aligned}$$

- Non rifiutiamo H_0. Il p-value è molto superiore ad α: siamo ampiamente in regione di *non-rifiuto*.

t-test sulle scatole di biscotti

$$H_0 : \mu \geq 375$$
$$H_1 : \mu < 375$$

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{368 - 375}{15/\sqrt{26}} = -2.38$$

- Il valore critico del test a una coda è $t_{25,0.05} = -1.7$: rifiutiamo H_0.
- Non calcoliamo il CI, in quanto servirebbe il CI unilaterale.

Frazione di difettosi

- Il campione contiene più di 5 successi ed insuccessi; possiamo quindi procedere con il test:

$$H_0 : \pi \geq 0.1$$
$$H_1 : \pi < 0.1$$

Frazione di difettosi

$$Z_0 = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

$$Z_0 = \frac{.044 - .1}{\sqrt{\frac{.1 \cdot .9}{250}}} = -2.95$$

- Il valore critico è $Z_{alpha} = Z_{.05} = -1.64$.
- La statistica è oltre il valore critico e quindi rifiutiamo H_0.
- Il p-value è $\Phi(Z_0) = \Phi(-2.95) = 0.0001$, molto più piccolo di α.

Some problems with frequentist hypothesis test

- It can only reject (strong decision) or non-reject (weak decision) H_0; it cannot make a strong decision in favor of H_0. For instance, you cannot conclude with high confidence that two classifiers are practically equivalent.
- There are many other drawbacks of frequentist hypothesis testing, which we do not cover.
- A point hypothesis (two-tailed test) of the type $H_0 : \mu = 0$ is always wrong, in the sense that the parameter has never exactly the hypothesised value.
- We would like instead to have the posterior probability of H_0 and H_1 being correct, given the data