7. Test d'ipotesi

Un tipico problema che ci si può trovare ad affrontare è il seguente:

Faccio una certa ipotesi (che indico con H_0 e che chiamo **ipotesi nulla**). In base ai dati che ho a disposizione devo decidere se accettare o rifiutare la verità di questa ipotesi.

Si potranno verificare quattro situazioni alternative:

1. L'ipotesi è vera e l'accetto $\rightarrow$ bene
2. L'ipotesi è vera ma in base ai dati la rifiuto $\rightarrow$ in questo caso si dice che si commette **errore di prima specie**
3. L'ipotesi è falsa ma in base ai dati la accetto $\rightarrow$ in questo caso si dice che si commette **errore di seconda specie**
4. L'ipotesi è falsa e la rifiuto $\rightarrow$ bene

Per chiarirsi le idee vediamo prima un esempio.

Esempio 7.0.2. Ho una moneta. Voglio verificare se è bilanciata o meno. La lancio n volte. Pongo $X_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa,} \\ 0 & \text{se all'}i\text{-esimo lancio esce croce.} \end{cases}$, $i = 1, \ldots, n$. Ho un campione statistico bernoulliano di numerosità n e parametro $p \in [0, 1]$ incognito, dove p è la probabilità che esca testa in un singolo lancio.

L'ipotesi nulla che dobbiamo testare è

$$H_0) \qquad p = 0.5.$$

Facciamo dunque n lanci. Otteniamo k teste ed $n - k$ croci:

$$x_1,\ x_2,\ \ldots,\ x_n \qquad \text{dove} \qquad x_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa,} \\ 0 & \text{se all'}i\text{-esimo lancio esce croce.} \end{cases}$$

e dunque $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i = \dfrac{k}{n}$.

Stabilisco una distanza massima ε tra $\overline{x}$ e 0.5 entro la quale accettare l'ipotesi $p = 0.5$ e oltre la quale rifiutarla. Ovvero: accetto H_0 se $|\overline{x} - 0.5| < \varepsilon$ e la rifiuto se $|\overline{x} - 0.5| \geq \varepsilon$. cioè se $\left| \displaystyle\sum_{i=1}^{n} x_i - \dfrac{n}{2} \right| \geq n\varepsilon$. Quanto vale la probabilità di commettere errore di prima specie, ovvero di rifiutarla quando esse invece è vera?

Commetto errore di prima specie con probabilità

$$\alpha := \mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \frac{n}{2}\right| \geq n\varepsilon\right).$$

Poiché le v.a. X_i sono indipendenti e seguono tutte la distribuzione di Bernoulli di parametro p, la v.a. $Y := \sum_{i=1}^{n} X_i$ è una v.a. binomiale di parametri n e p. Se l'ipotesi H_0 è vera, allora $p = 0.5$ dunque : $Y \sim B(n, 0.5)$ e

$$\alpha := \mathbb{P}\left(\left|Y - \frac{n}{2}\right| \geq n\varepsilon\right) = \mathbb{P}\left(Y \geq \frac{n}{2} + n\varepsilon\right) + \mathbb{P}\left(Y \leq \frac{n}{2} - n\varepsilon\right)$$

Vediamo alcuni casi

- $n = 50$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}(Y \geq 25 + 5) + \mathbb{P}(Y \leq 25 - 5) = 1 - F_Y(29) + F_Y(20).$$

```
> 1 - pbinom(c(29), size=50, prob=0.5, lower.tail=TRUE)
+ pbinom(c(20), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.2026388
```

- $n = 100$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}(Y \geq 50 + 10) + \mathbb{P}(Y \leq 50 - 10) = 1 - F_Y(59) + F_Y(40).$$

```
> 1 - pbinom(c(59), size=100, prob=0.5, lower.tail=TRUE)
+ pbinom(c(40), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.05688793
```

- $n = 200$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}(Y \geq 100 + 20) + \mathbb{P}(Y \leq 100 - 20) = 1 - F_Y(119) + F_Y(80).$$

```
> 1 - pbinom(c(119), size=200, prob=0.5, lower.tail=TRUE)
+ pbinom(c(80), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.005685156
```

- $n = 300$, $\varepsilon = 0.1$

$$\alpha = \mathbb{P}(Y \geq 150 + 30) + \mathbb{P}(Y \leq 150 - 30) = 1 - F_Y(179) + F_Y(120).$$

```
> 1 - pbinom(c(179), size=300, prob=0.5, lower.tail=TRUE)
+ pbinom(c(120), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.00063422
```

- $n = 50$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 25 + 2.5\right) + \mathbb{P}\left(Y \leq 25 - 2.5\right) = 1 - F_Y(27) + F_Y(22).$$

```
> 1 - pbinom(c(27), size=50, prob=0.5, lower.tail=TRUE)
+ pbinom(c(22), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.4798877
```

- $n = 100$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 50 + 5\right) + \mathbb{P}\left(Y \leq 50 - 5\right) = 1 - F_Y(54) + F_Y(45).$$

```
> 1 - pbinom(c(54), size=100, prob=0.5, lower.tail=TRUE)
+ pbinom(c(45), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.3682016
```

- $n = 200$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 100 + 10\right) + \mathbb{P}\left(Y \leq 100 - 10\right) = 1 - F_Y(109) + F_Y(90).$$

```
> 1 - pbinom(c(109), size=200, prob=0.5, lower.tail=TRUE)
+ pbinom(c(90), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.178964
```

- $n = 300$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 150 + 15\right) + \mathbb{P}\left(Y \leq 150 - 15\right) = 1 - F_Y(164) + F_Y(135).$$

```
> 1 - pbinom(c(164), size=300, prob=0.5, lower.tail=TRUE)
+ pbinom(c(135), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.0939037
```

- $n = 400$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 200 + 20\right) + \mathbb{P}\left(Y \leq 200 - 20\right) = 1 - F_Y(219) + F_Y(180).$$

```
> 1 - pbinom(c(219), size=400, prob=0.5, lower.tail=TRUE)
+ pbinom(c(180), size=400, prob=0.5, lower.tail=TRUE)
[1] 0.05104022
```

- $n = 500$, $\varepsilon = 0.05$

$$\alpha = \mathbb{P}\left(Y \geq 250 + 25\right) + \mathbb{P}\left(Y \leq 250 - 25\right) = 1 - F_Y(264) + F_Y(225).$$

```
> 1 - pbinom(c(274), size=500, prob=0.5, lower.tail=TRUE)
+ pbinom(c(225), size=500, prob=0.5, lower.tail=TRUE)
[1] 0.02832616
```

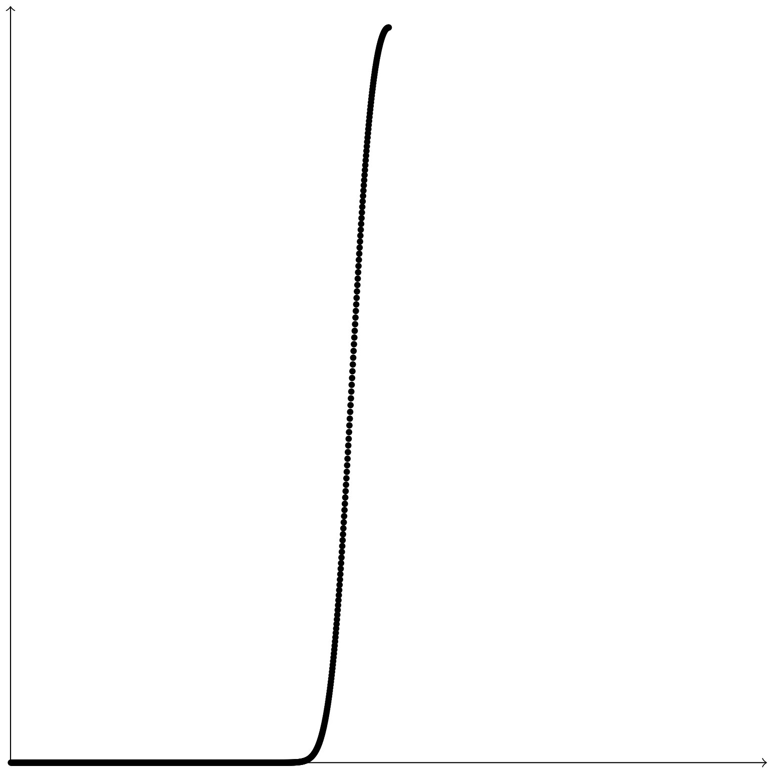

Figura 7.1: $\beta(p)$

Solitamente si vuole controllare (nel senso di tenere bassa, inferiore a 0.1 o a 0.05) la probabilità α di commettere errore di prima specie. Tale probabilità viene detta *livello di significatività* del test. Fissato il livello di significatività α, la numerosità n e la soglia di tolleranza ε andranno scelti di conseguenza come visto negli esempi precedenti.

Inoltre, fissato α, ci chiediamo quanto valga la probabilità di commettere errore di seconda specie, ovvero di accettare H_0 quand'essa invece è falsa.

Se H_0 è falsa, allora la probabilità di ottenere testa non è 0.5 ma assume un valore $p \neq 0.5$ (ignoto) e dunque $Y \sim B(n,p)$ e io accetto H_0 con probabilità

$$\beta(p) := \mathbb{P}_p\left(\left|Y - \frac{n}{2}\right| < n\varepsilon\right) = \mathbb{P}_p\left(Y < \frac{n}{2} + n\varepsilon\right) - \mathbb{P}_p\left(Y \leq \frac{n}{2} - n\varepsilon\right)$$

Si calcola $\beta(p)$ per vari valori di p. La funzione $\beta(p)$ è detta **curva operativa caratteristica (OC)** mentre $1 - \beta(p)$ cioè la probabilità di rifiutare H_0 quand'essa in effetti è falsa e il parametro incognito vale p, è detta **potenza del test**.

Esempio 7.0.3. Consideriamo la solita moneta e stavolta vogliamo vedere se è più probabile ottenere testa che ottenere croce. Vogliamo cioè testare l'ipotesi nulla

$$H_0) \qquad p \leq 0.5$$

Un test di questo tipo è detto *test unilaterale*.

Stabilisco una tolleranza massima ε entro la quale accettare l'ipotesi $p \leq 0.5$ e oltre la quale rifiutarla. Ovvero: accetto H_0 se $\overline{x} < 0.5 + \varepsilon$ e la rifiuto se $\overline{x} \geq 0.5 + \varepsilon$ cioè se $\sum_{i=1}^{n} x_i \geq \frac{n}{2} + n\varepsilon$. Quanto vale la probabilità di commettere errore di prima specie, ovvero di rifiutarla quando essa invece è vera?

Commetto errore di prima specie con probabilità

$$\alpha := \mathbb{P}\left(Y \geq \frac{n}{2} + n\varepsilon\right).$$

Se H_0 è vera, allora $Y \sim B(n,p)$ per qualche $p \leq 0.5$. Indico F_Y^p la sua funzione di ripartizione Vediamo alcuni casi

- $n = 50$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 25 + 5\right) = 1 - F_Y^p(29) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(29)\right\} = 1 - F_Y^{0.5}(29)$$

```
> 1 - pbinom(c(29), size=50, prob=0.5, lower.tail=TRUE)
```

- $n = 100$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 50 + 10\right) = 1 - F_Y^p(59) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(59)\right\} = 1 - F_Y^{0.5}(59).$$

```
> 1 - pbinom(c(59), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.02844397
```

- $n = 200$, $\varepsilon = 0.1$

$$\alpha = 1 - \mathbb{P}\left(Y < 100 + 20\right) = 1 - F_Y^p(119) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(119)\right\} = 1 - F_Y^{0.5}(119).$$

```
> 1 - pbinom(c(119), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.002842578
```

- $n = 50$, $\varepsilon = 0.05$

$$\alpha = 1 - \mathbb{P}\left(Y < 25 + 2.5\right) = 1 - F_Y^p(27) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(27)\right\} = 1 - F_Y^{0.5}(27)$$

```
> 1 - pbinom(c(27), size=50, prob=0.5, lower.tail=TRUE)
[1] 0.2399438
```

- $n = 100$, $\varepsilon = 0.05$

$$\alpha = 1 - \mathbb{P}\left(Y < 50 + 5\right) = 1 - F_Y^p(54) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(54)\right\} = 1 - F_Y^{0.5}(54)$$

```
> 1 - pbinom(c(55), size=100, prob=0.5, lower.tail=TRUE)
[1] 0.1841008
```

- $n = 200$, $\varepsilon = 0.05$

  $$\alpha = 1 - \mathbb{P}(Y < 100 + 10) = 1 - F_Y^p(109) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(109)\right\} = 1 - F_Y^{0.5}(109).$$

```
> 1 - pbinom(c(109), size=200, prob=0.5, lower.tail=TRUE)
[1] 0.08948202
```

- $n = 300$, $\varepsilon = 0.05$

  $$\alpha = 1 - \mathbb{P}(Y < 150 + 15) = 1 - F_Y^p(164) \leq \sup_{p \in [0,0.5)} \left\{1 - F_Y^p(164)\right\} = 1 - F_Y^{0.5}(145).$$

```
> 1 - pbinom(c(164), size=300, prob=0.5, lower.tail=TRUE)
[1] 0.04695185
```

In generale dunque un test d'ipotesi ha la seguente struttura:

1. Si ha un campione statistico $X_1, X_2, \ldots, X_n$, $X_i \sim \mathcal{D}(\theta)$, dove θ è un parametro reale.

2. Si formula un'ipotesi (che si chiama *ipotesi nulla* e si indica con H_0), solitamente nella forma
$$H_0) \qquad \theta \in \Theta_0$$
dove Θ_0 è un sottoinsieme di $\mathbb{R}$.

3. Si formula una regola di decisione per l'accettazione o il rifiuto di H_0. La regola di decisione è di questo tipo: si sceglie una statistica che fornisce una stima del parametro θ e un sottoinsieme $A \subset \mathbb{R}$, detto *regione di accettazione*. Dopodiché
   - se $Y \in A$, allora si accetta H_0;
   - se $Y \notin A$, allora si rifiuta H_0.

   L'insieme $A^c := \mathbb{R} \setminus A$ è detto *regione di rifiuto*.

Come già detto, è solitamente richiesto di *limitare* la probabilità di commettere errore di prima specie, cioè di limitare la probabilità di rifiutare l'ipotesi nulla quando essa è vera. Vediamo come questo sia possibile nel caso di campioni gaussiani.

7.1. Test d'ipotesi per la media di campioni gaussiani

7.1.1. Campione gaussiano di cui è nota la varianza

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu = \mu_0.$$

H_0 è vera se e solo se $\mathbb{E}\left[\overline{X}\right] = 0$ dunque accetto l'ipotesi nulla H_0 se la media campionaria si discosta da μ_0 per meno di un valore soglia ε ovvero se $|\overline{x} - \mu_0| < \varepsilon$ e la rifiuto altrimenti.

Il livello di di significatività (cioè la probabilità di commettere un errore di prima specie) è allora

$$\alpha = \mathbb{P}_{\mu_0}\left(\left|\overline{X} - \mu_0\right| \geq \varepsilon\right)$$

dove il pedice μ_0 indica che H_0 è vera, cioè che $\mu = \mu_0$.

Se H_0 è vera, $\overline{X} \sim \mathcal{N}\left(\mu_0, \dfrac{\sigma^2}{n}\right)$ e $Z := \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$. Dunque

$$\begin{aligned}
\alpha &= \mathbb{P}_{\mu_0}\left(\left|\overline{X} - \mu_0\right| \geq \varepsilon\right) = \mathbb{P}_{\mu_0}\left(\frac{\left|\overline{X} - \mu_0\right|}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(|Z| \geq \frac{\varepsilon\sqrt{n}}{\sigma}\right) \\
&= \mathbb{P}\left(Z \geq \frac{\varepsilon\sqrt{n}}{\sigma}\right) + \mathbb{P}\left(Z \leq \frac{-\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) + \Phi\left(\frac{-\varepsilon\sqrt{n}}{\sigma}\right) \\
&= 2\left(1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right)\right)
\end{aligned}$$

Se voglio fissare a priori α, deve essere allora $\Phi\left(\dfrac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \dfrac{\alpha}{2}$ cioè deve essere $\dfrac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\frac{\alpha}{2}}$ e dunque devo scegliere

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$|\overline{x} - \mu_0| < \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \leq \mu_0.$$

Accetto l'ipotesi nulla H_0 se la media campionaria è inferiore a $\mu_0+\varepsilon$ cioè se $\overline{x} < \mu_0+\varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} \geq \mu_0+\varepsilon\right).$$

dove il pedice indica che la media del campione è $\mu \leq \mu_0$.

Poiché $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ e $Z := \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, si ha

$$\begin{aligned}\mathbb{P}_{\mu\leq\mu_o}\left(\overline{X} \geq \mu_0+\varepsilon\right) &= \mathbb{P}_{\mu\leq\mu_0}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\mu_0-\mu+\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(Z \geq \frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right)\\ &= 1-\mathbb{P}\left(Z \leq \frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right) = 1-\Phi\left(\frac{(\mu_0-\mu+\varepsilon)\sqrt{n}}{\sigma}\right) \leq 1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right).\end{aligned}$$

Se voglio limitare superiormente $\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} > \mu_0+\varepsilon\right)$, cioè se voglio

$$\mathbb{P}_{\mu\leq\mu_0}\left(\overline{X} > \mu_0+\varepsilon\right) \leq \alpha \qquad \forall\mu \leq \mu_0$$

scelgo ε in modo da avere $1-\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = \alpha$ cioè $\frac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\alpha}$ e dunque scelgo

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$\overline{x} < \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ incognita e varianza σ^2 nota. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \geq \mu_0.$$

Accetto l'ipotesi nulla H_0 se la media campionaria è superiore a $\mu_0-\varepsilon$ cioè se $\overline{x} > \mu_0-\varepsilon$. La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu\geq\mu_0}\left(\overline{X} \leq \mu_0-\varepsilon\right).$$

Poiché $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, e $Z := \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$, si ha

$$\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right) = \mathbb{P}_{\mu \geq \mu_0}\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\mu_0 - \mu - \varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) = \mathbb{P}\left(Z \leq \frac{(\mu_0 - \mu - \varepsilon)\sqrt{n}}{\sigma}\right)$$
$$= \Phi\left(\frac{(\mu_0 - \mu - \varepsilon)\sqrt{n}}{\sigma}\right) \leq \Phi\left(\frac{-\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right).$$

Se voglio limitare superiormente $\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right)$ cioè se voglio

$$\mathbb{P}_{\mu \geq \mu_0}\left(\overline{X} \leq \mu_0 - \varepsilon\right) \leq \alpha \qquad \forall \mu \geq \mu_0$$

scelgo ε in modo da avere $\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \alpha$ cioè $\frac{\varepsilon\sqrt{n}}{\sigma} = z_{1-\alpha}$ e dunque scelgo

$$\varepsilon = \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ la loro media. Accetto H_0 se

$$\overline{x} > \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$$

e la rifiuto altrimenti.

7.1.2. Campione gaussiano di cui non è nota la varianza

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe ignote. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu = \mu_0.$$

Sappiamo che, se $\mu = \mu_0$, allora $T := \frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \sim t(n-1)$.

Inoltre H_0 è vera se e solo se $\mathbb{E}\left[\overline{X}\right] = \mu_0$ ovvero, per l'indipendenza di $\overline{X}$ e S^2, se e solo se $\mathbb{E}[T] = 0$. Dunque accetto l'ipotesi nulla H_0 se $|T| \leq \varepsilon$.

Il livello di di significatività (cioè la probabilità di commettere un errore di prima specie) è allora

$$\alpha = \mathbb{P}\left(|T| \geq \varepsilon\right).$$

Si ha quindi

$$\alpha = \mathbb{P}\left(|T| \geq \varepsilon\right) = \mathbb{P}\left(T \geq \varepsilon\right) + \mathbb{P}\left(T \leq -\varepsilon\right)$$
$$= 1 - F_T(\varepsilon) + F_T(-\varepsilon) = 2\left(1 - F_T(\varepsilon)\right)$$

Se voglio fissare a priori α, deve essere allora $F_T(\varepsilon) = 1 - \dfrac{\alpha}{2}$ dunque devo scegliere

$$\varepsilon = t_{n-1,1-\frac{\alpha}{2}}.$$

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t = \dfrac{(\overline{x} - \mu_0)\sqrt{n}}{\sqrt{s^2}}$ (dove $\overline{x}$ e s indicano la media e la deviazione campionaria del campione, rispettivamente). Accetto H_0 se

$$|t| \le t_{n-1,1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti, ovvero accetto H_0 se

$$\mu_0 - \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}} \le \overline{x} \le \mu_0 + \frac{t_{n-1,1-\frac{\alpha}{2}}\, s}{\sqrt{n}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe incognite. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \le \mu_0.$$

Diamo la seguente regola di accettazione: accettiamo H_0 se $\dfrac{(\overline{X} - \mu_0)\sqrt{n}}{S} \le \varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu \le \mu_0}\left(\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon\right).$$

Se H_0 è vera, allora $\mu \le \mu_0$ e dunque

$$\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} \le \frac{(\overline{X} - \mu)\sqrt{n}}{S} =: T \sim t(n-1).$$

Di conseguenza

$$\left\{\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon\right\} \subset \left\{\frac{(\overline{X} - \mu)\sqrt{n}}{S} > \varepsilon\right\}$$

Dunque, per ogni $\mu \le \mu_0$ si ha

$$\mathbb{P}_{\mu \le \mu_0}\left(\frac{(\overline{X} - \mu_0)\sqrt{n}}{S} > \varepsilon\right) \le \mathbb{P}_{\mu \le \mu_0}\left(\frac{(\overline{X} - \mu)\sqrt{n}}{S} > \varepsilon\right) = \mathbb{P}\left(T > \varepsilon\right) = 1 - F_T(\varepsilon)$$

Se vogliamo stabilire il livello di significatività α dovremmo scegliere ε in modo che

$$1 - F_T(\varepsilon) = \alpha$$

cioè $\varepsilon = t_{n-1,1-\alpha}$.

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t_0 = \dfrac{(\overline{x}-\mu_0)\sqrt{n}}{s}$. Accetto H_0 se

$$t_0 \leq t_{n-1,1-\alpha}$$

ovvero se

$$\overline{x} \leq \mu_0 + \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ e varianza σ^2 entrambe incognite. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \mu \geq \mu_0.$$

Diamo la seguente regola di accettazione: accettiamo H_0 se $\dfrac{(\overline{X}-\mu_0)\sqrt{n}}{S} \geq -\varepsilon$.

La probabilità di commettere un errore di prima specie è allora

$$\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon\right).$$

Se H_0 è vera, allora $\mu \geq \mu_0$ e dunque

$$\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} \geq \frac{(\overline{X}-\mu)\sqrt{n}}{S} =: T \sim t(n-1).$$

Di conseguenza

$$\left\{\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon\right\} \subset \left\{\frac{(\overline{X}-\mu)\sqrt{n}}{S} < -\varepsilon\right\}$$

Dunque

$$\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon\right) \leq \mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} < -\varepsilon\right) \qquad \forall\mu \geq \mu_0$$

e quindi, per ogni $\mu \geq \mu_0$ si ha

$$\begin{aligned}\mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu_0)\sqrt{n}}{S} < -\varepsilon\right) &= \mathbb{P}_{\mu\geq\mu_0}\left(\frac{(\overline{X}-\mu)\sqrt{n}}{S} < -\varepsilon\right)\\ &= \mathbb{P}\left(T < -\varepsilon\right) = F_T(-\varepsilon) = 1 - F_T(\varepsilon).\end{aligned}$$

Se vogliamo stabilire il livello di significatività α dovremmo scegliere ε in modo che

$$1 - F_T(\varepsilon) = \alpha$$

cioè $\varepsilon = t_{n-1,1-\alpha}$.

Presi i dati $x_1,\ x_2,\ \ldots,\ x_n$, sia dunque $t_0 = \dfrac{(\overline{x}-\mu_0)\sqrt{n}}{s}$. Accetto H_0 se

$$t_0 \geq -t_{n-1,1-\alpha}$$

e la rifiuto altrimenti, ovvero accetto H_0 se

$$\overline{x} \geq \mu_0 - \frac{t_{n-1,1-\alpha}\, s}{\sqrt{n}}$$

e la rifiuto altrimenti.

7.2. Test d'ipotesi per l'uguaglianza di medie di campioni gaussiani

Supponiamo di avere due campioni, entrambi gaussiani

$$\begin{aligned} &X\colon X_1,\ X_2,\ \ldots,\ X_n \qquad && X_i \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right), \\ &Y\colon Y_1,\ Y_2,\ \ldots,\ Y_k \qquad && Y_j \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right). \end{aligned}$$

Vogliamo testare l'ipotesi

$$H_0) \qquad \mu_X = \mu_Y$$

Osserviamo che $\mu_x = \mu_Y$ se e solo se $\mathbb{E}\left[\overline{X}-\overline{Y}\right] = 0$.

Per limitare la probabilità di commettere errore di prima specie, distinguiamo tre diversi casi

7.2.1. Primo caso: le varianze σ_X^2 e σ_Y^2 sono note

Considero la v.a. $W := \overline{X} - \overline{Y}$. Per le proprietà dei campioni gaussiani

$$W \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}\right).$$

Dunque H_0 è vera se e solo se $W \sim \mathcal{N}\left(0, \dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{k}\right)$. Dunque stabilisco il seguente criterio di accettazione:

$$\text{Accetto } H_0 \text{ se e solo se } |w| = |\overline{x} - \overline{y}| < \varepsilon.$$

La probabilità di commettere errore di prima specie vale allora

$$\alpha = \mathbb{P}_{\mu_X=\mu_Y}\left(|W| \geq \varepsilon\right) = \mathbb{P}_{\mu_X=\mu_Y}\left(\frac{|W|}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{k}}} \geq \frac{\varepsilon}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{k}}}\right)$$

D'altra parte, se H_0 è vera, allora $Z := \dfrac{W}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{k}}} \sim \mathcal{N}(0,1)$, e dunque dovremo scegliere $\dfrac{\varepsilon}{\sqrt{\dfrac{\sigma_X^2}{n} + \dfrac{\sigma_Y^2}{k}}} = z_{1-\frac{\alpha}{2}}$ ovvero

$$\varepsilon = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}}.$$

Dunque accettiamo l'ipotesi H_0 se

$$|\overline{x} - \overline{y}| < z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{k}}$$

e la rifiutiamo altrimenti.

Osservazione 7.2.1. Se $\sigma_X^2 = \sigma_Y^2 = \sigma_0^2$ e $k = n$, allora $\varepsilon = z_{1-\frac{\alpha}{2}} \sigma_0 \sqrt{\frac{2}{n}}$.

7.2.2. Secondo caso: le varianze σ_X^2 e σ_Y^2 sono ignote ma uguali

Consideriamo le due varianze campionarie

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{k-1} \sum_{j=1}^{k} (Y_j - \overline{Y})^2.$$

Sia σ^2 il comune valore di σ_X^2 e σ_Y^2. Sappiamo che

$$V_X := \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \quad V_Y := \frac{(k-1)S_Y^2}{\sigma^2} \sim \chi^2_{k-1}$$

dunque, per la Proprietà 5.3.2, $V_X + V_Y \sim \chi^2_{n-1+k-1} = \chi^2_{n+k-2}$

D'altra parte

$$V_X + V_Y = \frac{(n-1)S_X^2 + (k-1)S_Y^2}{\sigma^2} = \frac{n+k-2}{\sigma^2} \frac{(n-1)S_X^2 + (k-1)S_Y^2}{n+k-2}.$$

Consideriamo la statistica:

$$\overline{S}^2 := \frac{(n-1)S_X^2 + (k-1)S_Y^2}{n+k-2}.$$

Si ha

$$V_X + V_Y = \frac{(n+k-2)\overline{S}^2}{\sigma^2}. \tag{7.1}$$

Inoltre $\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma^2\left(\frac{1}{n} + \frac{1}{k}\right)\right)$, quindi

$$Z := \frac{\overline{X} - \overline{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{k}}} \sim \mathcal{N}(0,1) \qquad \text{se e solo se} \qquad \mu_X = \mu_Y.$$

Sia

$$T := \frac{Z\sqrt{n+k-2}}{\sqrt{V_X + V_Y}}.$$

Poiché i due campioni sono gaussiani e indipendenti le v.a. $\overline{X}$, S_X^2, $\overline{Y}$ e S_Y^2 sono indipendenti, quindi Z e $\overline{S}^2$ sono indipendenti, quindi $\mathbb{E}\left[\overline{X} - \overline{Y}\right] = 0$ se e solo se $\mathbb{E}[T] = 0$. Come criterio di accettazioneper l'ipotesi nulla H_0 scelgo dunque $|t| < \varepsilon$.

Inoltre, se $\mathbb{E}\left[\overline{X} - \overline{Y}\right] = 0$, allora per il Teorema 5.3.5

$$T \sim t(n+k-2).$$

Sostituendo l'espressione per Z e quella per $V_X + V_Y$ data nell' equazione (7.1) si ha

$$T = \frac{\overline{X}}{\overline{S}\sqrt{\frac{1}{n} + \frac{1}{k}}} \sim t(n+k-2)$$

Osserviamo anche che

$$\mathbb{E}\left[\overline{S}^2\right] = \frac{(n-1)\mathbb{E}\left[S_X^2\right] + (k-1)\mathbb{E}\left[S_Y^2\right]}{n+k-2} = \frac{(n-1)\sigma^2 + (k-1)\sigma^2}{n+k-2} = \sigma^2$$

e dunque possiamo usare $\overline{S}^2$ per stimare la varianza σ^2.

La probabilità di commettere errore di prima specie è allora

$$\alpha = \mathbb{P}\left(|T| \geq \varepsilon\right)$$

Fissato il livello di significatività α, devo dunque scegliere $\varepsilon = t_{n+k-2,1-\frac{\alpha}{2}}$.

Siano $x\colon x_1,\ x_2,\ \ldots,\ x_n$ e $y\colon y_1,\ y_2,\ \ldots,\ y_k$ i dati, $\overline{x}$ e $\overline{y}$ le rispettive medie, s_x^2 e s_y^2 le rispettive varianze e sia $\overline{s}^2 := \dfrac{(n-1)s_x^2 + (k-1)s_y^2}{n+k-2}$: accetto l'ipotesi nulla se

$$\frac{|\overline{x} - \overline{y}|}{\overline{s}\sqrt{\frac{1}{n} + \frac{1}{k}}} < t_{n+k-2,1-\frac{\alpha}{2}},$$

cioè se

$$|\overline{x} - \overline{y}| < t_{n+k-2,1-\frac{\alpha}{2}}\ \overline{s}\sqrt{\frac{1}{n} + \frac{1}{k}}$$

e la rifiuto altrimenti.

7.2.3. Terzo caso: le varianze σ_X^2 e σ_Y^2 sono ignote e diverse

Si può dimostrare che se n e k sono sufficientemente grandi e se H_0 è vera, allora la distribuzione della statistica

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\dfrac{S_X^2}{n}+\dfrac{S_Y^2}{k}}}$$

è *approssimativamente* gaussiana standard. Dunque, per ottenere approssimativamente un livello di significatività α si accetta l'ipotesi nulla $H_0) \qquad \mu_X = \mu_Y$ quando

$$\frac{|\overline{x}-\overline{y}|}{\sqrt{\dfrac{s_X^2}{n}+\dfrac{s_Y^2}{k}}} < z_{1-\frac{\alpha}{2}}$$

e la si rifiuta altrimenti.

Il problema di individuare un test che dia un livello di significatività α prescritto è ancora un problema aperto ed è noto come *problema di Behrens-Fisher*.

7.3. Test d'ipotesi per la varianza di campioni gaussiani

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 = \sigma_0^2.$$

H_0 è vera se e solo se $\mathbb{E}\left[S^2\right] = \sigma_0^2$ se e solo se $\mathbb{E}\left[\dfrac{S^2}{\sigma_0^2}\right] = 1$.

Sappiamo che la v.a. $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

Dunque accetto H_0 se $1-\varepsilon_1 < \dfrac{s^2}{\sigma_0^2} < 1+\varepsilon_2$, ε_1, ε_2 positivi, cioè se e solo se

$$(n-1)(1-\varepsilon_1) < \frac{(n-1)s^2}{\sigma^2} < (n-1)(1+\varepsilon_2).$$

Devo scegliere ε_1 e e_2 in modo da ottenere il livello di significatività α desiderato:

$$\begin{aligned}\alpha &= \mathbb{P}\left(\frac{S^2}{\sigma_0^2} > 1+\varepsilon_2\right) + \mathbb{P}\left(\frac{S^2}{\sigma_0^2} < 1-\varepsilon_1\right)\\ &= \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} > (n-1)(1+\varepsilon_2)\right) + \mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} < (n-1)(1-\varepsilon_1)\right).\end{aligned}$$

Una possibile scelta è allora

$$\mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} > (n-1)(1+\varepsilon_2)\right) = \frac{\alpha}{2} \qquad \text{cioè} \quad (n-1)(1+\varepsilon_2) = \chi^2_{n-1,1-\frac{\alpha}{2}}$$

$$\mathbb{P}\left(\frac{(n-1)S^2}{\sigma_0^2} < (n-1)(1-\varepsilon_1)\right) = \frac{\alpha}{2} \qquad \text{cioè} \quad (n-1)(1-\varepsilon_1) = \chi^2_{n-1,\frac{\alpha}{2}}.$$

Dunque accetto H_0 se

$$\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}$$

cioè se

$$\frac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}} < s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 \leq \sigma_0^2.$$

Accetto l'ipotesi nulla se $\dfrac{s^2}{\sigma_0^2} \leq 1+\varepsilon$.

Se la varianza è $\sigma^2 \leq \sigma_0^2$, allora $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ e la probabilità di commettere errore di prima specie è

$$\begin{aligned}
&\mathbb{P}_{\sigma^2\leq\sigma_0^2}\left(\frac{S^2}{\sigma_0^2} > 1+\varepsilon\right) = \mathbb{P}_{\sigma^2\leq\sigma_0^2}\left(\frac{(n-1)S^2}{\sigma^2} > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right)\\
&= \mathbb{P}\left(V > \frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right) = 1 - F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1+\varepsilon)\right)\\
&\leq 1 - F_V\left((n-1)(1+\varepsilon)\right).
\end{aligned}$$

Posso allora limitare superiormente con α la probabilità di commettere errore di prima specie imponendo

$$1 - F_V\left((n-1)(1+\varepsilon)\right) = \alpha$$

cioè scegliendo ε in modo che

$$(n-1)(1+\varepsilon) = \chi^2_{n-1,1-\alpha}.$$

Dunque accetto l'ipotesi nulla H_0 se

$$\frac{(n-1)s^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha}$$

cioè se

$$s^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha}$$

e la rifiuto altrimenti.

Test unilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione gaussiano di media μ (nota o incognita) e varianza σ^2 incognita. Vogliamo testare l'ipotesi nulla

$$H_0) \qquad \sigma^2 \geq \sigma_0^2.$$

Accetto l'ipotesi nulla se $\dfrac{s^2}{\sigma_0^2} \geq 1-\varepsilon$.

Se la varianza è $\sigma^2 \geq \sigma_0^2$, allora $V := \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ la probabilità di commettere errore di prima specie è allora

$$\begin{aligned}\mathbb{P}_{\sigma^2}\left(\frac{S^2}{\sigma_0^2} < 1-\varepsilon\right) &= \mathbb{P}_{\sigma^2}\left(\frac{(n-1)S^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon)\right) \\ &= F_V\left(\frac{\sigma_0^2}{\sigma^2}(n-1)(1-\varepsilon)\right) \leq F_V\left((n-1)(1-\varepsilon)\right).\end{aligned}$$

Posso allora limitare superiormente con α la probabilità di commettere errore di prima specie imponendo

$$F_V\left((n-1)(1-\varepsilon)\right) = \alpha$$

cioè scegliendo ε in modo che

$$(n-1)(1-\varepsilon) = \chi^2_{n-1,\alpha}.$$

Dunque accetto l'ipotesi nulla H_0 se

$$\frac{(n-1)s^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$$

cioè se

$$s^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha}$$

e la rifiuto altrimenti.

7.4. Test d'ipotesi per la media di campioni bernoulliani

Abbiamo già trattato questo argomento nell' esempio introduttivo.

Test bilaterale

Sia $X_1,\ X_2,\ \ldots,\ X_n$ è un campione di Bernoulli di parametro p incognito. Sappiamo che $\mathbb{E}\left[X_i\right] = p$ e che, vedi (5.1)

$$\mathbb{P}\left(\left|\overline{X}-p\right| > t\right) \leq \frac{1}{4n\,t^2} \qquad \forall t > 0.$$

dunque usiamo $\overline{x}$ come stima di p.

Testiamo l'ipotesi nulla

$$H_0) \qquad p = p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} - p_0 < \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p_0)$ di cui conosciamo la distribuzione. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\alpha = \mathbb{P}_{p=p_0}\left(\left|\overline{X} - p_0\right| < \varepsilon\right) &= \mathbb{P}_{p=p_0}\left(|Y - np_0| < n\varepsilon\right) \\ &= \mathbb{P}\left(Y < n(p_0 + \varepsilon)\right) - \mathbb{P}\left(Y \le n(p_0 - \varepsilon)\right).\end{aligned}$$

Test unilaterale

Testiamo l'ipotesi nulla

$$H_0) \qquad p \le p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} < p_0 + \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p)$ per qualche $p \le p_0$. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\mathbb{P}_{p\le p_0}\left(\overline{X} \ge p_0 + \varepsilon\right) &= \mathbb{P}_{p\le p_0}\left(Y \ge n(p_0 + \varepsilon)\right) \\ &\le \mathbb{P}_{p_0}\left(Y \ge n(p_0 + \varepsilon)\right).\end{aligned}$$

Per limitare superiormente il livello di sgnificatività α scelgo dunque ε in modo che $\mathbb{P}_{p_0}\left(Y \ge n(p_0 + \varepsilon)\right) \le \alpha$.

Test unilaterale

Testiamo l'ipotesi nulla

$$H_0) \qquad p \ge p_0.$$

Stabiliamo il criterio di accettazione: accetto H_0 se $\overline{x} > p_0 - \varepsilon$ e la rifiuto altrimenti.

Se H_0 è vera, allora la v.a. $Y := \sum_{i=1}^{n} X_i$ segue la distribuzione $B(n, p)$ per qualche $p \ge p_0$. La probabilità di commettere errore di prima specie è quindi

$$\begin{aligned}\mathbb{P}_{p\ge p_0}\left(\overline{X} \le p_0 - \varepsilon\right) &= \mathbb{P}_{p\ge p_0}\left(Y \le n(p_0 - \varepsilon)\right) \\ &\le \mathbb{P}_{p_0}\left(Y \le n(p_0 - \varepsilon)\right).\end{aligned}$$

Per limitare superiormente il livello di significatività α scelgo dunque ε in modo che $\mathbb{P}_{p_0}\left(Y \le n(p_0 - \varepsilon)\right) \le \alpha$.

7.5. Test del χ^2

Sia $X_1,\ X_2,\ \ldots,\ X_n$ un campione statistico. Supponiamo che le v.a. del campione siano discrete a valori $y_1,\ y_2,\ \ldots,\ y_k$. Consideriamo le densità di probabilità

$$p_j := \mathbb{P}\left(X_i = y_j\right), \qquad j = 1, \ldots, k.$$

Vogliamo testare l'ipotesi nulla

$$H_0 \qquad p_j = p_j^0 \qquad \forall j = 1, \ldots, k.$$

Per ogni $j = 1, \ldots, k$ considero le *frequenze campionarie*

$$N_j(\omega) = \#\left\{i \in \{1, \ldots, n\} \colon X_i(\omega) = y_k\right\} \qquad j = 1, \ldots, k$$

e le *frequenze campionarie relative*

$$F_j := \frac{N_j}{n}, \qquad j = 1, \ldots, k.$$

Sicuramente $N_j \sim B(n, p_j)$, quindi $\mathbb{E}\left[N_j\right] = np_j$ e $\mathbb{E}\left[F_j\right] = p_j$.

In particolare H_0 è vera se e solo se $\mathbb{E}\left[N_j\right] = np_j^0 \quad \forall j = 1, \ldots, k$, dunque un criterio di accettazione potrebbe essere quello di accettare H_0 se e solo se $\left|n_j - np_j^0\right| < \varepsilon \quad \forall j = 1, \ldots, k$.

Questo criterio però non ci permette di calcolare la probabilità di errore di prima specie. Vale però il seguente risultato:

Teorema 7.5.1 (di Pearson)**.** *Se $N_j \sim B(n, p_j)$, allora la funzione di ripartizione della v.a.*

$$T_n := \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

converge, per $n \to \infty$, alla funzione di ripartizione associata alla distribuzione χ^2_{k-1}.

Osservazione 7.5.1. Nelle applicazioni l'approssimazione è considerata accettabile se $np_j \geq 5 \quad \forall j = 1, \ldots, k$.

Formuliamo allora il seguente criterio di accettazione:

$$\text{accetto l'ipotesi nulla } H_0 \text{ se e solo se } t_n := \textstyle\sum_{j=1}^{k} \dfrac{(n_j - np_j^0)^2}{np_j^0} < \varepsilon.$$

La probabilità di commettere errore di prima specie è allora

$$\alpha := \mathbb{P}\left(T_n \geq \varepsilon\right) \simeq 1 - F_{\chi^2_{k-1}}(\varepsilon).$$

Scelgo dunque ε tale che $F_{\chi^2_{k-1}}(\varepsilon) = 1 - \alpha$, cioè $\varepsilon = \chi^2_{k-1,1-\alpha}$.

Osservazione 7.5.2. Il test si può applicare anche nel caso in cui $y_1,\ y_2,\ \ldots,\ y_k$ siano sostituiti da classi di modalità $I_1,\ I_2,\ \ldots,\ I_k$.

7.5.1. Test di normalità

Supponiamo di aver un campione $X_1,\ X_2,\ \dots,\ X_n$. Vogliamo testare l'ipotesi

$$H_0) \qquad \text{Il campione è normale}$$

Si può procedere nel seguente modo:

1. stimiamo μ e σ^2 rispettivamente con $\overline{x}$ e s^2;
2. standardizziamo i dati ponendo $z_i := \dfrac{x_i - \mu}{\sigma}$. Se il campione segue la distribuzione $\mathcal{N}\left(\mu, \sigma^2\right)$, allora $Z_i \sim \mathcal{N}\left(0, 1\right)$;
3. suddividiamo la retta reale in intervalli $I_1,\ I_2,\ \dots,\ I_k$ (simmetrici rispetto all'origine), ivi comprese due semirette simmetriche $[a, +\infty)$ e $(-\infty, -a]$;
4. contiamo $n_j := \#\left\{i \in \{1, \dots, n\} \colon z_i \in I_j\right\}$;
5. calcoliamo $p_j^0 := \mathbb{P}\left(Z_i \in I_j\right)$;
6. consideriamo la v.a. $T_n := \displaystyle\sum_{j=1}^{k} \frac{(N_j - np_j^0)^2}{np_j^0}$. Si può dimostrare che per $n \to \infty$ la funzione di ripartizione di T_n converge alla funzione di ripartizione associata alla distribuzione χ^2_{n-2-1}, dove il -2 è dovuto al fatto che abbiamo sostituito i due parametri μ e σ^2 con le loro stime provenienti dai dati $\overline{x}$ e s^2;
7. accettiamo l'ipotesi nulla se $t_n < \varepsilon$. Se imponiamo un livello di significatività α, sceglierermo allora $\varepsilon = \chi^2_{n-3,1-\alpha}$.

Bibliografia

[1] Luigi Barletti. Appunti del corso applicazioni di matematiche e statistica, a.a. 2007–08.

[2] Fabio Frascati. *Formulario di Statistica con R*. http://cran.r-project.org/doc/contrib/Frascati-FormularioStatisticaR.pdf, 2008.

[3] Antonia Morpoulou and Kyriaki Polikreti. Principal component analysis in monument conservation: Three application examples. *Journal of Cultural Heritage*, 10:73–81, 2009.

[4] John Verzani. *simpleR*. http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf, 2001.