Confidence interval and hypothesis testing

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

Confidence interval

Aim

Get a range of admissible values for our parameter.... Θ

read it

With 99% confidence Θ will be inside our estimated confidence interval...

Confidence?

- Consider that CI is random (Rely on in each sample) unlike Θ is fixed.
- We use $\hat{\Theta}$ for construct the CI but it belong to Θ
- Not is 95% of probability of Θ is in specific interval.
- Confidence means; if repeated the method (collect data and construct CI) for $\alpha = 0.05$ of 100 CI's, you expect of 95 of them capture parameter Θ.

Quantile

Remember the definition of z_α is

...

$$P(X < z_\alpha) = \alpha \tag{1}$$

Confidence level

$$1 - \alpha, \quad \alpha \in (0, 1) \tag{2}$$

Significance level of α.

Upper and lower bounds

Given α we are searching two values (under and above) of zero (remember that is Z) that:

...

- Z_α
- $Z_{1-\frac{\alpha}{2}}$
- The area between $-Z_{1-\frac{\alpha}{2}}$ and $Z_{1-\frac{\alpha}{2}}$ is equal to α

CI

...

$$P\left(-Z_{1-\frac{\alpha}{2}} \leq \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1-\alpha \tag{3}$$

The before intervals were constructe

...

$$P\left(\bar{x} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \tag{4}$$

Note that σ is population parameter, if n is large you could uses sample standard deviation S.

CI

General

$$\hat{\Theta} \pm \text{Margin of error} \tag{5}$$

Where $\hat{\Theta}$ is our best estimator.

There is a importaint point here, is that condifence interval rely on in the distrbution of $\hat{\Theta}$.

...

$$\hat{\Theta} \pm Z_{1-\frac{\alpha}{2}} S.E(\hat{\Theta}) \tag{6}$$

Precision - informative

...

If a interval is very wide then not is informative!

unknown σ

In this case we dont known the SE therefore is used a **Estimated Estandard Error (ESE)**

ESE for mean

$$\frac{S}{\sqrt{n}} \tag{7}$$

where S is the sample standard deviation.

CI with σ unknown

...

$$\left(\bar{x} - t_{(1-\frac{\alpha}{2},n-1)}\frac{S}{\sqrt{n}} \quad , \bar{x} + t_{(1-\frac{\alpha}{2},n-1)}\frac{S}{\sqrt{n}}\right) \tag{8}$$

where t_{n-1} comes from t distribution with n degrees of freedom.

Consider that when n is large t tend to Z.

```
from scipy.stats import t
from scipy.stats import norm
print(norm.ppf(0.95))
print(t.ppf(0.95, 25))
print(t.ppf(0.95, 100000))
```

mean difference pair data

Paired data

Two measurement of a same individual after a treatment.

$$\mu_d = \mu_{post} - \mu_{pre} \tag{9}$$

...

$$\left(\bar{x}_d - t_{(1-\frac{\alpha}{2}, n-1)} \frac{S_d}{\sqrt{n}} \quad , \bar{x}_d + t_{(1-\frac{\alpha}{2}, n-1)} \frac{S_d}{\sqrt{n}} \right) \tag{10}$$

where t_{n-1} comes from t distribution with n degrees of freedom, a where y is our variable of interest in dataset and therefore

$$\bar{x}_d = \frac{1}{n} \sum_{i=1}^{n} y_{post,i} - y_{pre,i} \tag{11}$$

No paired data

Two approaches

- Pooled $\sigma_A^2 = \sigma_B^2$
- Unpooled $\sigma_A^2 \neq \sigma_B^2$

Uses S_A and S_B as approximations to see what approach is better.

No paired data

Unpooled

SE Unpooled

$$SE = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} \tag{12}$$

Remember that in most practical applications we don't know σ then replace it with S.

...

used t distribution to estimate the area:

- Uses Welchs approximation (See this reference)
- or $min(n_A - 1, n_B - 1)$

No paired data

pooled

ESE pooled

$$\frac{\sqrt{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}}{n_A + n_B - 2}\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \tag{13}$$

Excersice

Construct a program to calcualte pooled and unpooled intervals.

laboratories CI

mean

...

- **CI(mean) simulation**
- **CI(mean) real data**

Laboratories

- First
- Central limit theorem
-

CI

⊖

...

given that limits are random the interval is random

...

Seeing-theory

Confidence interval

For what it is useful confidence interval?
Now assume that $\alpha \in [0, 1]$

$$P(\hat{\theta}_{low} < \theta < \hat{\theta}_{upper}) = 1 - \alpha. \tag{14}$$

Note that the interval is also random.

CI

Remember that CI is aming to find θ therefore is a mistake said that $(1-\alpha)*100$ times the parameter falls inside the interval (There is a common mistake). How find $\hat{\theta}_{low}, \hat{\theta}_{upper}$

A better approximation is that the probability of the interval contain the parameter is $1-\alpha$.

Interpretation of CI

it is uses the interval in n sampling evaluations then $(1-\alpha)$ times the interval contain θ.

CI(mean) with known σ^2

$X_n \sim N(\mu, \sigma^2)$, then

$$\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \tag{15}$$

$$P(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96) = 0.95 \tag{16}$$

before of some algebraic inequalities operations we have:

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \tag{17}$$

See simulation here: CI simulation (click)
Which is the pivotal quantity?

The value $Z_{\frac{\alpha}{2}}$ is whose that the area to the right of the point (in normal curve) is $\frac{\alpha}{2}$ is also the quantile of level $1 - \frac{\alpha}{2}$ the value that left to the left of the area $1 - \frac{\alpha}{2}$.

Proportion confidence interval P

$$\hat{P} = \frac{\sum x_i}{n} \tag{18}$$

where x_i is the number of successes, therefore using the central limit theorem we have that

$$\frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \sim N(0,1) \tag{19}$$

Interval

$$\left(\hat{P} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \quad , \quad \hat{P} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \tag{20}$$

Work

...

- Analize how chante the results for $\alpha = 0.01$

Considerations

It is important remember **Random sample**.
take in mind when you use $z_{f(\alpha)}$ that sample $n \geq 30$.

laboratories CI

Proportion

...

- **CI(proportion) simulation**
- **CI(proportion) real data**

Differece proportion

In two populations (A, B) that present a feature as ϕ determine if

...

$$P_A(\phi) - P_B(\phi) \tag{21}$$

The difference in the proportion of subjects or objects in A and B that present ϕ

...

$$\hat{P}_A - \hat{P}_B \pm Z_{1-\frac{\alpha}{2}} SE(\hat{P}_A - \hat{P}_B) \tag{22}$$

Where comes from the SE?

...

$$SE(\hat{P}_A - \hat{P}_B) = \sqrt{\frac{\hat{P}_A(1-\hat{P}_A)}{n_A} + \frac{\hat{P}_B(1-\hat{P}_B)}{n_B}} \qquad (23)$$

How interpret them?

if the interval is positive (L, U) we are going to said; *there are* 95% *of confidence that* P_A *is greater than* P_B *between* L *and* U*.*

Exercise

With the following Dataset determine the confidence interval for proportion difference among Males and Females whose score in any module is in Q_3.

Considerations

What happend if 0 is in the interval of differneces?, remeberm that here is neccesary random samples and that the samples is large this las requirement is

$$n_i \hat{P}_i \geq 10 \text{ for } i = A, B \tag{24}$$

$$n_i(1 - \hat{P}_i) \geq 10 \text{ for } i = A, B \tag{25}$$