LOGIC - HYPOTESIS TESTING

Null hypotesis, H_0 - NOTHING GOING ON

Alternative " , H_1 - SOMETHING EXTRA GOING ON

Ej.

$H_0 : P(T) = \frac{1}{2}$ (FAIR COIN) T = Tails

$H_1 : P(T) \neq$ (UNFAIR COIN)

TEST STATISTIC -

Measures how far away the data are from the expected if H_0 were true

The most common test statistic is the **z-statistic**.

$$z = \frac{\text{observed} - \text{expected}}{SE}$$

Ej. COIN TOS

EXPECTED : $10 \cdot \frac{1}{2} = 5$ $\quad SE = \sqrt{10} \cdot \sqrt{\frac{1}{2} \cdot \frac{1}{2}} = 1.58$

$$z = \frac{7 - 5}{1.58} = 1.27$$

P-Values as Measures of Evidence

Large values of $|z|$ are evidence against H_0

The strenght of the evidence is the

p-value (observed significance level)

Los valores de p por debajo del 5% se suelen considerar "estadísticamente significativos"

$$z = \frac{\text{observado} - \text{esperado}}{SE}$$

- If p-value is larger than 5% we will not reject the null hypotesis.

T-Test

- We use it when the sample is small.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \bar{x} \pm t_{n-1}\, SE$$

Statistical Significance vs Importance

TWO SAMPLE Z-TEST

Rating:

$p_1 \rightarrow n = 1000$ & 55%

$p_2 \rightarrow n = 1500$ & 58%

$H_0 \rightarrow p_1 = p_2$ (nothing invoked coming up)

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}}$$

$\hat{p}_1 = 55\%$ $\quad p_1 - p_2 = 0$

$\hat{p}_2 = 58\%$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\sqrt{(SE(p_1))^2 + (SE(p_2))^2}} \Rightarrow \frac{(0.58 - 0.55) - 0}{0.0202} = 1.48$$

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Minimos Tabla

p-value = 2 (7%)

14% > 5% so we cannot reject null hypothesis

Also confidence interval $(\hat{p}_2 - \hat{p}_1) \pm z\, SE(\hat{p}_2 - \hat{p}_1)$

In our case at 98% $z = 2$ $\quad [-1\%, 7\%]$

$0.55 \cdot 1000 = 550$
$0.58 \cdot 1500 = 870$
$\Big\} \rightarrow 1420$ out of 2500

pooled estimate $\frac{1420}{2500} = 56.8$

$$SE(\hat{p}_2 - \hat{p}_2) \sqrt{\frac{0.568 \cdot (1-0.568)}{1000} + \frac{0.568(1-0.568)}{1500}} = 0.020 \text{ ??}$$

which gives the same answer

TO COMPARE TWO MEANS

$$SE(\bar{x}_2 - \bar{x}_0) = \sqrt{(SE(\bar{x}_2))^2 + (SE(\bar{x}_2))^2}$$

$SE(\bar{x}_2) = \frac{\sigma_2}{\sqrt{n_2}}$ estimated $\frac{s_2}{\sqrt{n_2}}$ t-test

Paired Difference Test

paired t-test

$$t = \frac{\bar{d} - 0}{SE(\bar{d})}$$

$$SE(\bar{d}) = \frac{\sigma}{\sqrt{n}} \rightarrow s_d = 0.55$$

$$t = \frac{1.4 - 0}{0.55/\sqrt{15}} = 5.69$$

COMPUTER SIMULATIONS IN PLACE OF CALCULATIONS

Recall - confidence interval

$$\bar{x} \pm z\,SE(\bar{x})$$

If we instead of $\bar{x}$ are interested in other estimator: for example $\hat{\theta}$ for a parameter θ and the normal aproximation is not valid for that estimator $\hat{\theta}$

En estas situaciones las simulaciones pueden ser usados para estimar

LAW OF LARGE NUMBERS
TO APPROXIMATE QUANTITIES OF INTEREST

Monte Carlo Method

For the explanation we will use the average height of people living in USA

Sample $n = 100$

We are interested in a parameter θ of a population that we estimate with $\hat{\theta}$

Our statistic (estimator) $\hat{\theta}$ is the average so

$\hat{\theta}$ = average of sample $= \frac{1}{n} \sum_{i=1}^{n} x_i$

Monte Carlo Method or Simulation:

Approximation to a fixed quantity θ by the average of independent random variables with expected value θ. The larger the sample the smaller the SE of the Statistic

$$SE(\hat{\theta}) = \sqrt{E(\hat{\theta} - E(\hat{\theta}))^2}$$

E → variance

Example:

1. 1000 samples of 100 observations
2. Compute $\hat{\theta}$ for the samples $\hat{\theta}_1 \dots \hat{\theta}_{1000}$
3. Compute Standard Dev

$$s(\hat{\theta}_1 \dots \hat{\theta}_{1000}) = \sqrt{\frac{1}{1000-1} \sum_{i=1}^{1000} \left(\hat{\theta}_i - \text{average}(\hat{\theta}_i)\right)^2}$$

$\bar{\hat{\theta}} = \text{average}(\hat{\theta})$

Summary

- Random Sampling
- The more samples the less SE
- We can compute the SE with it

Plug-In / Bootstrap principle

The bootstrap principle uses the plug-in principle and the montecarlo method to aproximate quantities such as $SE(\hat{\theta})$

1. Draw sample $X_1 \dots X_n$ to compute $\hat{\theta}$
2. Repeat B times (Eg $B = 1000$) to get $\hat{\theta}_1 \dots \hat{\theta}_B$
3. If we have only 1 sample then the bootstrap simulate from the sample since we don't have access to the population.

   Basically it interacts with the sample like if the sample was the population

Non-parametic bootstrap

Sometimes we can known or supose characteristics of the data. for example we might known that it follows a normal distribution but not its SE or mean

BOOTSTRAP CONFIDENCE INTERVALS

$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

We can estimate the sampling distribution of $\hat{\theta}$ not just $SE(\hat{\theta})$

Making a histogram of the bootstrap copies

Also we can do $\hat{\theta} - \theta$

bootstrap pivotal interval

$$\left[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)},\ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\right]$$

Bootstraping in regression

We have data $(X_1, Y_1), \dots, (X_n, Y_n)$ from simple linear regresion model

$$Y_i = a + bX_i + e_i$$

From the data we can compute estimates $\hat{a}$, $\hat{b}$.

And Compute residuals:

$$\hat{e}_i = Y_i - \hat{a} - \hat{b} X_i$$

* Remember: residuals are the difference between the observed values and the values predicted by the model/regresion line

Residual = observed - predicted

Steps

1. Compute residuals $\hat{e}_i = Y_i - \hat{a} - \hat{b} X_i$
2. Resample from those residuals to get $e_1^* \dots e_n^*$
3. Compute the bootstrap responses $Y_i^* = \hat{a} + \hat{b} X_i + e_i^*$