

1. Comparisons of two proportions

[Bookmark this page](#)

Recitation problem statement

You are interested in comparing the proportions of people in their 20's that smoke in France and in the US. After you sample randomly and independently n people in their 20's in both countries, you observe that N_{US} sampled US Americans and N_F sampled French are smokers. Based on such an experiment, how would you test whether there is a significant difference between the proportions of smokers in both countries?

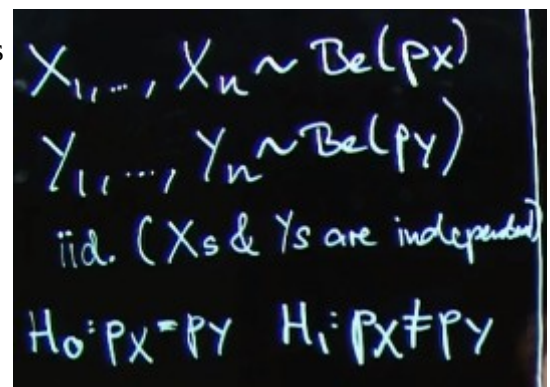
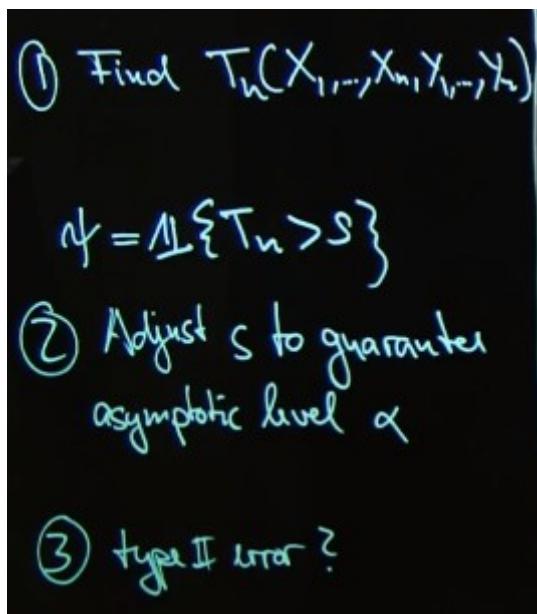
Note: In the following videos, we introduce a new term called "pivot". The formal definition of a pivotal quantity (or a pivot) is as follows. Let X_1, \dots, X_n be random samples and let T_n be a function of X and a parameter vector θ . That is, T_n is a function of X_1, \dots, X_n, θ . Let $g(T_n)$ be a random variable whose distribution is the same for all θ . Then, g is called a pivotal quantity or a pivot.

For example, let X be a random variable with mean μ and variance σ^2 . Let X_1, \dots, X_n be iid samples of X . Then,

$$g_n \triangleq \frac{\overline{X_n} - \mu}{\sigma}$$

is a pivot with $\theta = [\mu \ \sigma^2]^T$ being the parameter vector. The notion of a parameter vector here is not to be confused with the set of parameters that we use to define a statistical model.

Have a hypothesis test to check if the number of smokers is the same or different in the US and France



Comparison of two proportions

① $\hat{p}_X = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\mathbb{P}} p_X$, $\hat{p}_Y = \frac{1}{n} \sum_{i=1}^n Y_i \xrightarrow[n \rightarrow \infty]{\mathbb{P}} p_Y$

Consider $\hat{p}_X - \hat{p}_Y = g(\hat{p}_X, \hat{p}_Y)$, $g(x, y) = x - y$

CLT: $\sqrt{n} \left(\begin{pmatrix} \hat{p}_X \\ \hat{p}_Y \end{pmatrix} - \begin{pmatrix} p_X \\ p_Y \end{pmatrix} \right) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N} \left(0, \underset{\parallel}{\Sigma} \right)$

$= \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} X_i \\ Y_i \end{pmatrix}$

$\begin{pmatrix} p_X(1-p_X) & 0 \\ 0 & p_Y(1-p_Y) \end{pmatrix}$

Delta Method:

$\sqrt{n} (g(\hat{p}_X, \hat{p}_Y) - g(p_X, p_Y)) \rightarrow \mathcal{N} \left(0, \underbrace{\nabla g(p_X, p_Y)^T \Sigma \nabla g(p_X, p_Y)}_{\sigma_g^2} \right)$

$\nabla g(x, y) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow \sigma_g^2 = (1 \ -1) \begin{pmatrix} p_X(1-p_X) & 0 \\ 0 & p_Y(1-p_Y) \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$= p_X(1-p_X) + p_Y(1-p_Y)$

Want an expression that does not depend on any of the underlying parameters\\

Comparison of two proportions $\hat{p}_X = \frac{1}{n} \sum_{i=1}^n X_i$, $\hat{p}_Y = \frac{1}{n} \sum_{i=1}^n Y_i$

$\sqrt{n} (\hat{p}_X - \hat{p}_Y - (p_X - p_Y)) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, p_X(1-p_X) + p_Y(1-p_Y))$

$\Rightarrow \frac{\sqrt{n} (\hat{p}_X - \hat{p}_Y - p_X + p_Y)}{\sqrt{p_X(1-p_X) + p_Y(1-p_Y)}} \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1)$

Simplify by setting $p_X = p_Y$ for H_0 and using a plugin estimator since we do not currently know what p_X hat and p_Y hat are.

For H_0 , set $p_X = p_Y = p \in (0,1)$: $p_X(1-p_X) + p_Y(1-p_Y) = 2p \cdot (1-p)$

$$\hat{p} = \frac{1}{2}(\hat{p}_X + \hat{p}_Y) \xrightarrow[n \rightarrow \infty]{P} p$$

Slutsky's method: $\sqrt{n} \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{2\hat{p}(1-\hat{p})}} \xrightarrow[n \rightarrow \infty]{D} \mathcal{N}(0,1)$

This is the test statistic, T_n , that we want. The absolute value of it as it's a 2 sided test.

$$(2) T_n = \left| \sqrt{n} \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{2\hat{p}(1-\hat{p})}} \right| \xrightarrow[n \rightarrow \infty]{D} \mathcal{N}(0,1) \text{ under } H_0(p_X = p_Y)$$

Cutting off at $T_n > S$ to guarantee an asymptotic level of α

$$\begin{aligned} \mathbb{P}(T_n > s) &= \mathbb{P}\left(\left|\sqrt{n} \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{2\hat{p}(1-\hat{p})}}\right| > s\right) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(|Z| > s) \\ &= 2 \cdot (1 - \Phi(s)) \stackrel{!}{=} \alpha \end{aligned}$$

$\Rightarrow s = q_{\alpha/2}, 1 - \alpha/2$ quantile of $\mathcal{N}(0,1)$

$$(3) \hat{p}_X \xrightarrow[n \rightarrow \infty]{P} p_X, \hat{p}_Y \xrightarrow[n \rightarrow \infty]{P} p_Y, \hat{p} = \frac{1}{2}(\hat{p}_X + \hat{p}_Y) \xrightarrow[n \rightarrow \infty]{P} \frac{1}{2}(p_X + p_Y) =: \tilde{p}$$

$$T_n = \left| \sqrt{n} \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{2\hat{p}(1-\hat{p})}} \right| \xrightarrow[n \rightarrow \infty]{P} \begin{cases} p_X - p_Y \neq 0 \text{ under } H_1, \\ +\infty \Rightarrow \text{type II error} \xrightarrow[n \rightarrow \infty]{P} 0 \\ \rightarrow \sqrt{2\tilde{p}(1-\tilde{p})} \end{cases}$$

If the null hypothesis does not hold then the statistic test exceeds the limits with overwhelming probability (approaches infinity)

Remarks:

Different sample sizes

④ Different sample sizes : $X_1, \dots, X_{n_1}, Y_1, \dots, Y_{n_2}$
CLT: $Z_1, \dots, Z_n \stackrel{iid}{\sim} \mathcal{R}, \mathbb{E}[Z_i] = \mu, \text{Var}(Z_i) = \sigma^2$

$$\sqrt{n} \frac{\bar{Z}_n - \mu}{\sqrt{\sigma^2}} \xrightarrow[n \rightarrow \infty]{D} N(0,1); \quad \bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i \quad \left| \quad \begin{aligned} \mathbb{E}[\bar{Z}_n] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i] \\ &= \frac{n}{n} \mu = \mu \\ \text{Var}(\bar{Z}_n) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(Z_i) \\ &= \frac{n}{n^2} \cdot \text{Var}(Z_i) = \frac{1}{n} \sigma^2 \end{aligned} \right.$$

$$\sqrt{n} \frac{\bar{Z}_n - \mu}{\sqrt{\sigma^2}} \xrightarrow[n \rightarrow \infty]{D} N(0,1)$$

$\underbrace{\hspace{10em}}$
 $\mathbb{E}[-n-] = 0, \text{Var}(-n-) = 1$

