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 ?

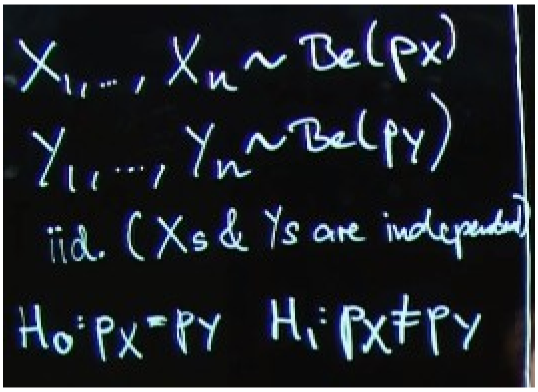
Note: In the following videos, we introduce a new term called "pivot". The formal definition of a pivotal quantity (or a pivot) is as follows. Let $X_1, \ldots, X_n$ be random samples and let T_n be a function of X and a parameter vector θ. That is, T_n is a function of $X_1, \ldots, X_n, \theta$. Let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

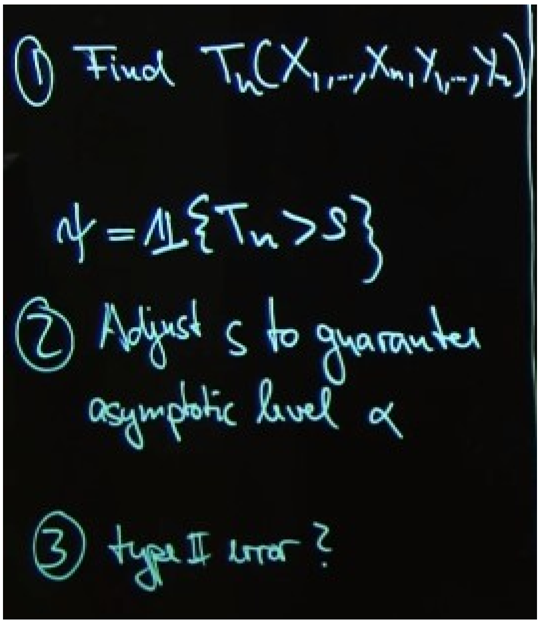
For example, let X be a random variable with mean μ and variance σ^2. Let $X_1, \ldots, X_n$ be iid samples of X. Then,

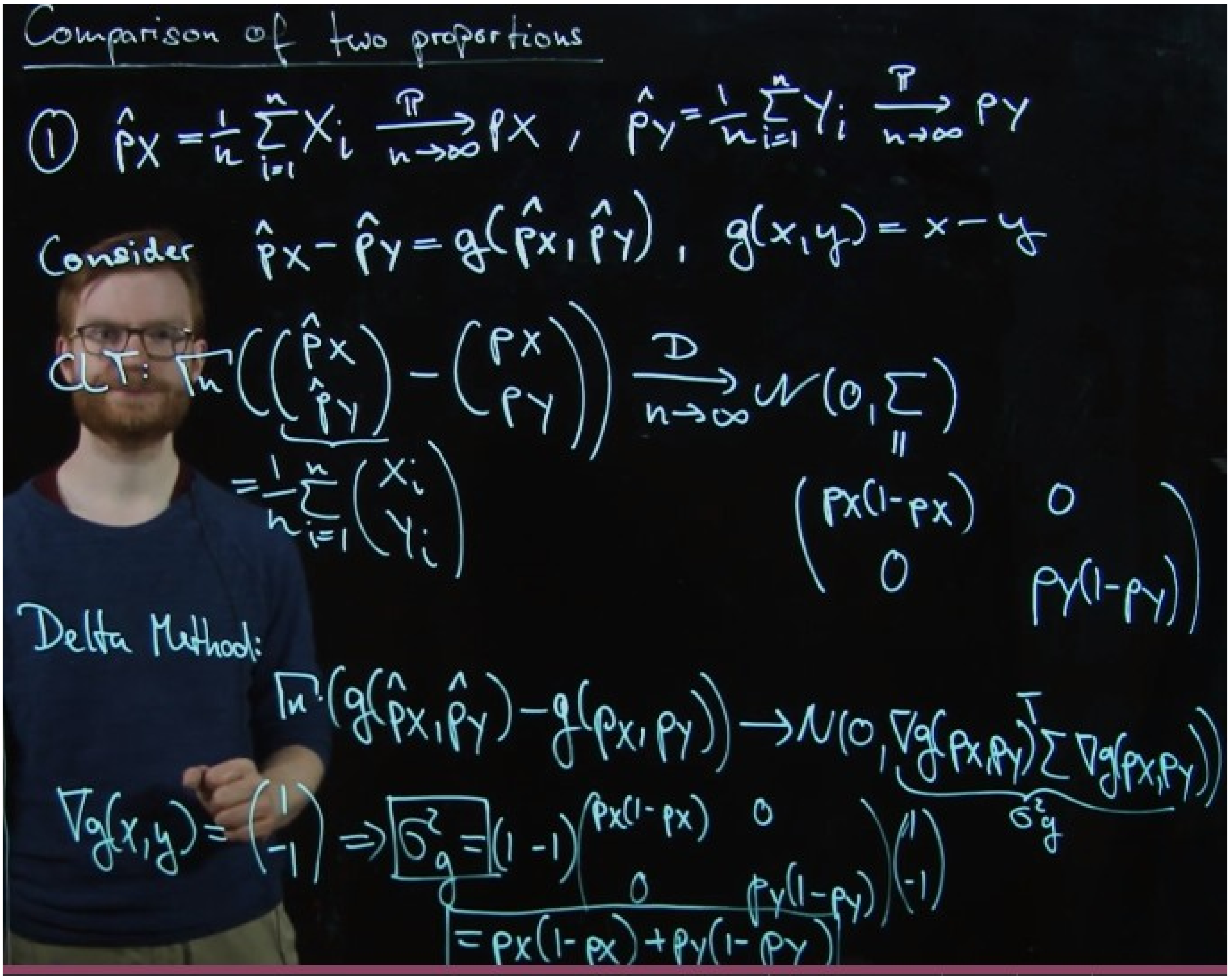
$$g_n \triangleq \frac{\overline{X_n} - \mu}{\sigma}$$

is a pivot with $\theta = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$ being the parameter vector. The notion of a parameter vector here is not to be confused with the set of paramaters that we use to define a statistical model.

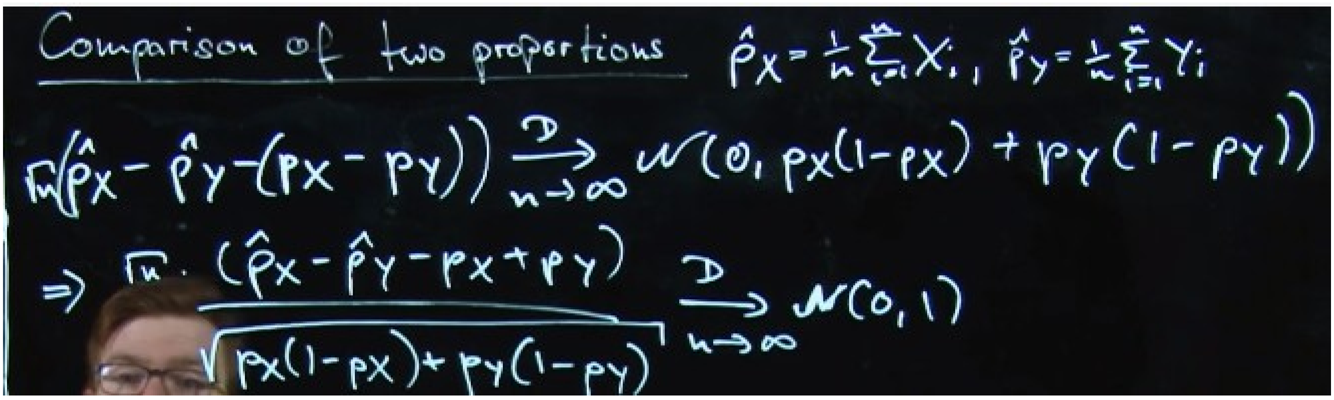
Have a hypothesis test to check if the number of smokers is the same or different in the US and France







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



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

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

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

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

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

Cutting off at Tn > S to guarantee an asymptotic level of alpha

$$\mathbb{P}(T_n > s) = \mathbb{P}\left(\left|\sqrt{n}\,\frac{\hat{p}_X - \hat{p}_Y}{\sqrt{2\hat{p}(1-\hat{p})}}\right| > s\right) \xrightarrow[n\to\infty]{} \mathbb{P}(|Z| > s), \quad Z \sim N(0,1)$$

$$= 2(1-\Phi(s)) \overset{!}{=} \alpha$$

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

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

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

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

Remarks:

Different sample sizes