# Estadística III

## Pruebas de bondad de ajuste

Alejandro López Hernández

FES Acatlán - UNAM

March 20, 2020

# Índice

La idea de las pruebas de bondad de ajuste es comparar la función de distribución de nuestros datos ($\hat{F}_n$) con una función de distribución dada ($F_0$). Nuestro objetivo será encontar estadístos que nos ayuden a aceptar o rechazar la siguiente prueba:

$$H_0 : F = F_0$$

Ejemplo: Supongamos que tenemos los datos

$$X = 0.254, 1.23, 4.566, 2.165, 1.23, 1.829, 5, 3.23$$

Una pregunta interesante es si los datos tiene una distribución *uniforme*, para probar la hipótesis, es necesario calcular $\hat{F}_n$ y con ella la prueba sería de la forma:

$$H_0 : F = F_0 \sim U(0, 5)$$

Algunas alternativas para medir las diferencias en las distribucciones son:

- $D_n^+ = \sup_{-\infty<t<\infty}(\hat{F}_n(t) - F_0(t))$
- $D_n^- = \sup_{-\infty<t<\infty}(F_0(t) - \hat{F}_n(t))$
- $D_n = \sup_{-\infty<t<\infty}|F_0(t) - \hat{F}_n(t)| = \max(D_n^+, D_n^-)$
- $V_n = D_n^+ + D_n^-$
- $C_n = \int(F_0(t) - \hat{F}_n(t))^2 dF_0(t)$
- $A_n = \int \frac{(\hat{F}_n(t)-F_0(t))^2}{F_0(t)(1-F_0(t))} dF_0(t)$
- $w_{n,k,g} = \int(F_0(t) - \hat{F}_n(t))^k g(F_0(t)) dF_0(t)$
- $w_{n,k,g} = \int(F_0(t) - \hat{F}_n(t))^k g(F_0(t)) dF_0(t)$

La prueba de Kolgomorov-Smirnov se define como

$$D_n = \sup_{-\infty<t<\infty} |F_0(t) - \hat{F}_n(t)|$$

Se define de esa manera debido al teorema de Gilvenko-Cantelli, que nos dice que $\sup_{-\infty<t<\infty} |F(t) - \hat{F}_n(t)| \to 0$ a.s, es decir que la máxima distancia entre la distribución empírica y la real tiende a 0. Por lo tanto si nuestra $F_0$ es la distribución real, se espera que $D_n$ sea pequeño.

Para calcular $D_n$ solo es necesario conocer las observaciones $X_1, X_2, ..., X_n$, si $X_{\{i\}}$ es el i-ésimo estadístico de orden, se puede probar que

$$D_n = \max_{1 \leq i \leq n} max(\frac{i}{n} - F_0(X_{\{i\}}), F_0(X_{\{i\}}) - \frac{i-1}{n})$$

La idea de este estadístico es medir el area que separa $\hat{F}_n$ de $F_0$, el estadístico se define como:

$$C_n = \int (F_0(t) - \hat{F}_n(t))^2 dF_0(t)$$

De forma analoga a el estadístico $D_n$, $C_n$ se puede escribir en función de los estadísticos de orden:

$$C_n = \frac{1}{12n} + \sum_{i=1}^{n} \left( F_0(X_{\{i\}}) - \frac{2i-1}{n} \right)$$

La idea es tambien medir el área en que separa $\hat{F}_n$ de $F_0$, sin embargo, el termino $F_0(t)(1 - F_0(t))$ busca tener una mayor ponderación en la región donde la distribución $F_0(t)$ tiene mayor incertidumbre. El estadistico se define como

$$A_n = \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t)$$

De forma analoga a los estadísticos anteriores $A_n$ se puede escribir como:

$$A_n = -n - \frac{1}{n}\left[\sum_{i=1}^{n}(2i-1)(\log F_0(X_{\{i\}}) + \log(1 - F_0(X_{\{n-i+1\}}))\right]$$