PROBABILIDAD Y ESTADÍSTICA

FCEN – UNCuyo

Licenciatura y Profesorado en Ciencias Básicas

M. Sc. Marcelo E. Alberto

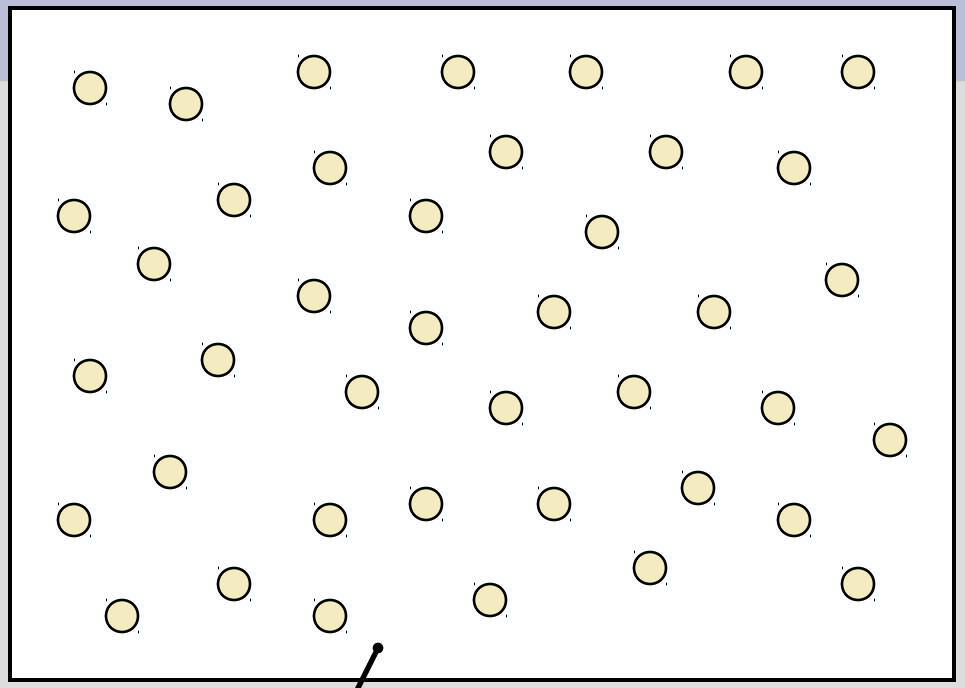
2018

Inferencia Estadística

Problemas de Inferencia

- Estimar el % de plantas de genotipo “A” que resulta infectada.
- Comparar efectividad de dos catalizadores químicos
- Evaluar existencia de asociación entre 2 variables
- Comparar si la relación entre Fotosíntesis Neta y Crecimiento es la misma en dos diferentes genotipos

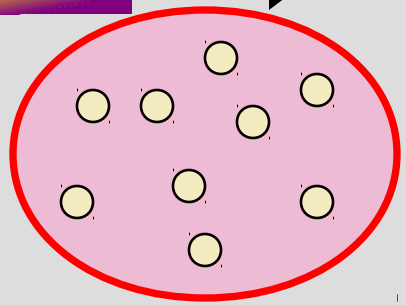
Población



PARÁMETROS

μ y σ

Muestra



ESTADÍSTICAS

$\bar{x}$ y s

Inferencia Estadística

Parámetro

Cantidad numérica que resume las características de una población.

Estadística

Cantidad numérica que resume las características de una muestra. Es cualquier función de los datos que no depende de parámetros desconocidos.

Inferencia Estadística

Muestra aleatoria de tamaño **n**

Conjunto de variables aleatorias con la misma función de probabilidad.

$$\{X_1, X_2, X_3, \cdots, X_n\}$$

$$X_i \sim P(\theta), \text{ con } 1 \leq i \leq n$$

Esperanza y Varianza de los elementos de la muestra

Dada una muestra aleatoria $\{X_1, \cdots, X_n\}$ resulta:

$$E[X_i] = \mu$$

$$V[X_i] = \sigma^2$$

$$\forall i = 1, \cdots, n$$

¿POR QUÉ?

Esperanza y Varianza de la muestra aleatoria

Dada una muestra aleatoria $\{X_1, \cdots, X_n\}$ resulta:

$$E[\bar{X}] = \mu$$

$$V[\bar{X}] = \frac{\sigma^2}{n}$$

¿POR QUÉ?

Propiedades Útiles para los ¿POR QUÉ?

Definiciones para la variable aleatoria W:

Promedio: $E[W]=\mu=\int_{\mathbb{R}} w f(w)\,dw$

Varianza: $V[W]=\sigma^2=\int_{\mathbb{R}} (w-\mu) f(w)\,dw$

$E[\alpha]=\alpha$

$V[\alpha]=0$

$E[\alpha W]=\alpha E[W]$

$V[\alpha W]=\alpha^2 V[W]$

$E[W\pm U]=E[W]\pm E[U]$

$V[W\pm U]=V[W]+V[U]$

siendo W y U variables aleatorias independientes y α constante

Teorema Central del Límite

Dada una muestra aleatoria $\{X_1, \cdots, X_n\}$ resulta: la variable aleatoria

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

tiene una distribución que se acerca a la distribución normal estándar conforme n tiende a infinito.

Teorema Central del Límite

Sea $f_n(z)$ la función densidad de la v.a. Z
para el tamaño muestral n
y sea $\phi(z)$ la función densidad de la distribución normal estándar.
El enunciado del Teorema Central del Límite puede expresarse así:

$$\lim_{n \to \infty} f_n(z) = \phi(z)$$

Nótese que el límite es una función

Teorema Central del Límite: consecuencias

Con la notación $Z \underset{n\to\infty}{\sim} N(0;1)$ entendemos que la v.a. Z tiene distribución **asintótica** normal estándar.

También diremos que la v.a. Z tiene distribución normal estándar **aproximada** para valores de n **suficientemente grandes** y lo anotaremos así:

$$Z \dot{\sim} N(0,1)$$

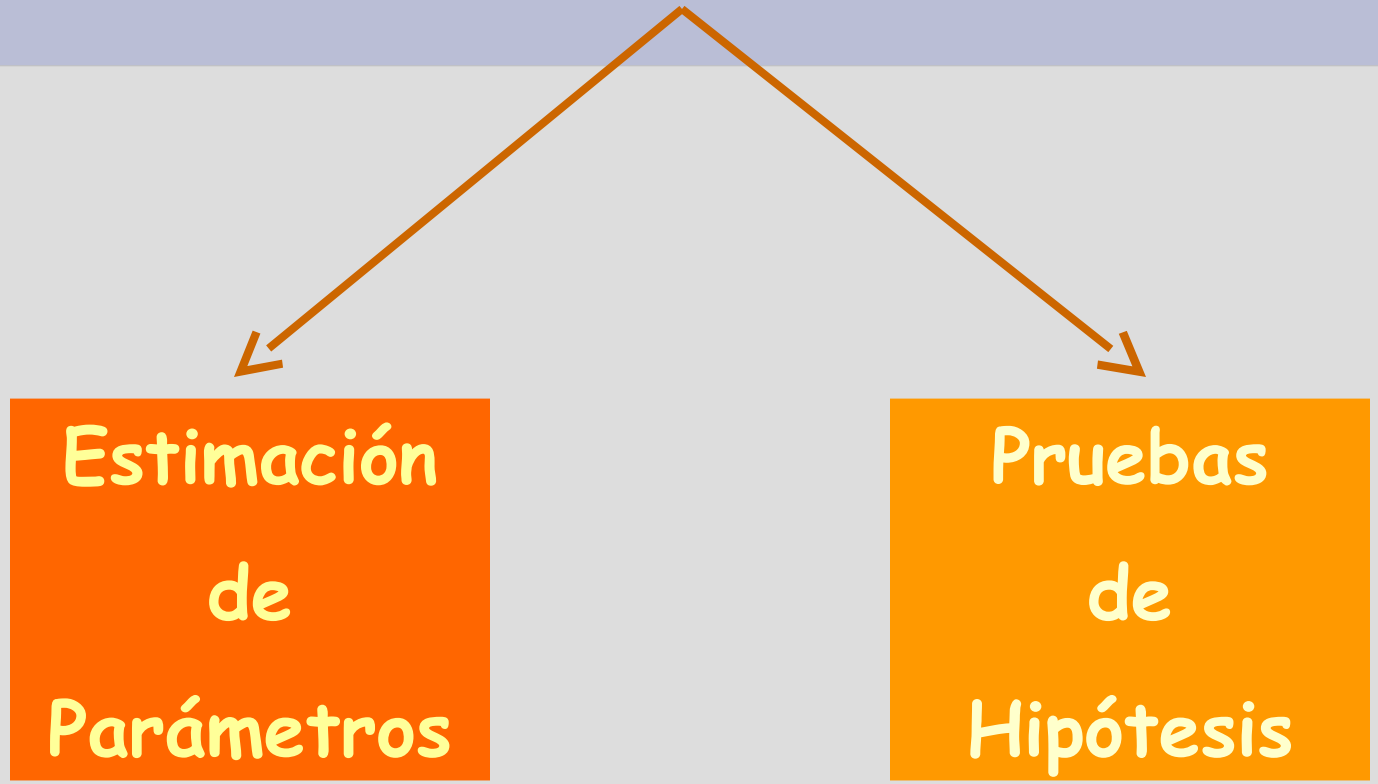
Teorema Central del Límite: consecuencias

Como consecuencia, tenemos que

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{X}$ tiene distribución **asintótica** normal con promedio μ y varianza $\frac{\sigma^2}{n}$.

Inferencia Estadística



Estimación de Parámetros

- Estimación Puntual
- Estimación por Intervalos de Confianza

Estimación Puntual

Se busca una estadística ***T*** que "estime" al parámetro θ. Un *estimador* con ciertas características:

- Insesgado:

$$E[T] = \theta$$

- Mínimo Error Cuadrático Medio:

$$ECM(T) = E\left[(T-\theta)^2\right]$$

Intervalo de Confianza

- Se busca construir un intervalo que contenga al parámetro estimado, con una probabilidad dada.
- Si **no conocemos** la distribución de probabilidades de X el teorema conocido como "Desigualdad de Chebychev" asegura que

$$P\left(\left|X-\mu\right| \leq k\sigma\right) \geq 1-\frac{1}{k^2}$$

$$\forall k > 1$$

Intervalo de Confianza

- Si $E[X]=\mu$, $V[X]=\sigma^2$ entonces

$$E[\overline{X}]=\mu \text{ , } V[\overline{X}]=\frac{\sigma^2}{n}$$

Según la desigualdad de Chebyshev

$$P\left(\overline{X}-k\frac{\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}+k\frac{\sigma}{\sqrt{n}}\right)\geq 1-\frac{1}{k^2}$$

$$\forall k>1$$

Intervalo de Confianza

Por tanto el intervalo

$$\left[\overline{X} - k \frac{\sigma}{\sqrt{n}} \quad ; \quad \overline{X} + k \frac{\sigma}{\sqrt{n}} \right]$$

estima μ con una confianza de

al menos $100\left(1 - \frac{1}{k^2}\right)\%$

para cualquier $k > 1$

Intervalo de Confianza

Otro Enfoque: se desea construir un intervalo que contenga al parámetro estimado, con una probabilidad *exacta.*

Es necesario poder asumir una distribución de probabilidades para X.

Intervalo de Confianza

Si es posible asumir $X \sim N(\mu;\sigma)$ entonces hay una probabilidad p de que el intervalo

$$\bar{x} - z_p \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_p \frac{\sigma}{\sqrt{n}}$$

contenga a la media poblacional.

El valor p se llama nivel de confianza y z_p es el percentilo $p\%$ de la distribución normal estándar, $N(0,1)$

Intervalo de Confianza

Si es posible asumir $X \sim N(\mu;\sigma)$ pero *se desconoce* el valor de σ entonces el intervalo se basa en la distribución de Student (Gosset)

$$\bar{x} - t_p(n-1)\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_p(n-1)\frac{s}{\sqrt{n}}$$

$t_p(n\text{-}1)$ es el percentilo p de la distribución Student con n-1 *grados de libertad.*

Intervalo de Confianza

Los anteriores intervalos de confianza para la media poblacional se construyen en base a *cantidades pivotales*:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1) \quad ; \qquad \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} \sim t(n-1)$$

σ Conocida

σ Desconocida

Ejercicio

Para los datos de agua de Lavalle calcular intervalos de confianza para las medias de pH, Ca, Mg y As mediante la desigualdad de Chebyshev y una cantidad pivotal correspondiente. Se desean niveles de confianza de 90% y 99%

Prueba de Hipótesis

Una *hipótesis* es una afirmación cuyo *valor de verdad* es desconocido. Se desea asignar valor de verdad a esta hipótesis (verdadera o falsa) según un *criterio estadístico*.

Ej.: $\mathcal{H}$: "el agua de pozo de Lavalle tiene pH 7"

$\mathcal{H}$: $\mu_{pH} = 7$

Prueba de Hipótesis

Una *prueba de hipótesis* o *test de hipótesis* o *dócima de hipótesis* es un proceso para asignar valor de verdad a una hipótesis en base a un criterio estadístico basado en una muestra aleatoria. Se plantea una pareja de hipótesis llamadas *hipótesis nula* e *hipótesis alternativa*. (cada una es la negación lógica de la otra):

Ej.: $\mathscr{H}_0$: $\mu_{pH} = 7$ vs $\mathscr{H}_1$: $\mu_{pH} \neq 7$

La prueba se realiza sobre $\mathscr{H}_0$

Prueba de Hipótesis

¿Qué podría suceder?

Según la información de la muestra	H_0 en la naturaleza	
✸	Verdadera	Falsa
Se presume verdadera	⚐	❄
Se presume falsa	❄	⚐

Prueba de Hipótesis

Según la información de la muestra ↓	H_0 Verdadera	H_0 Falsa
H_0 se presume Verdadera	✓	Error Tipo 2
H_0 se presume Falsa	Error Tipo 1	✓

Prueba de Hipótesis

	H_0 Verdadera	H_0 Falsa
Se acepta H_0	✓	Error 2
Se rechaza H_0	Error 1	✓

P(rechazar H_0 siendo verdadera) = α

P(aceptar H_0 siendo falsa) = β

Prueba de Hipótesis

α → Nivel de Significación

$1-\beta$ → Potencia

Prueba de Hipótesis

- Supuesto cierto modelo de distribución de probabilidad.
- En base a una muestra de tamaño *n.*
- Suponiendo que $\mathscr{H}_0$ es verdadera

Se evaluará la probabilidad p de que pueda obtenerse un promedio muestral igual o mayor que el obtenido.

Lo denominamos **valor** p.

Este valor es reportado en los resultados del análisis estadístico.

Prueba de Hipótesis

Criterio de Rechazo de $\mathscr{H}_0$

Si *valor-p* es un valor *arbitrariamente pequeño* se *rechazará* la hipótesis nula. El criterio es:

Rechazar $\mathscr{H}_0$ si y sólo si valor-p $<< \alpha$

Comparación de 2 muestras

Datos Categóricos	Datos Numéricos	
	Caso Paramétrico	Caso No Paramétrico
Proporción	***Media***	***Mediana***
Prueba basada en la distribución normal	Prueba basada en la estadística *t* de Student	Prueba de Mann–Witney Prueba de Wilcoxon

Comparación de k muestras (k>2)

Datos Categóricos	Datos Numéricos	
	Caso Paramétrico	Caso No Paramétrico
Proporción	***Media***	***Mediana***
Tablas de Contingencia Modelo Lineal Generalizado (MLG)*	Análisis de la Varianza de Fisher (ANOVA)*	Análisis de la Varianza no–paramétrico (Kruskal–Wallis)

Relación entre 2 variables

Datos Categóricos	Datos Numéricos
Modelos Lineales Generalizados (MLG)	Regresión Lineal Regresión no-lineal