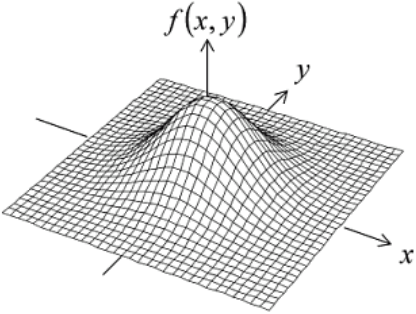
Análisis de Datos
2º de Biología
Introducción

Universidad Autónoma de Madrid

Departamento de Matemáticas

2019


$f(x, y)$
y
x

Introducción

- El curso: modelos lineales
  - Diseño de Experimentos
  - Regresión lineal simple
  - Regresión lineal múltilple
- Un breve repaso
  - Inferencia estadística
  - La distribución normal
  - Intervalos de confianza
  - Contrastes de hipótesis

Modelos lineales

Explican los valores de una variable aleatoria mediante una relación lineal entre los valores de otras variables que pueden influir en ella

Elementos del modelo básico:

$$
\begin{aligned}
\text{Variable a explicar} &= \text{constante común} \\
&+ \text{suma de efectos de las variables o factores} \\
&+ \text{errores o variaciones aleatorias}
\end{aligned}
$$

Modelos a estudiar

Variable explicada

Será una variable continua que, medida sobre una población homogénea, tendrá distribución normal

DISEÑO DE EXPERIMENTOS

Las variables explicativas (factores) son cualitativas

Tema 1: Análisis de la varianza unifactorial

Tema 2: Análisis de la varianza con varios factores

REGRESIÓN

Las variables explicativas son cuantitativas

Tema 3: Regresión lineal simple —una sola variable explicativa—

Tema 4: Regresión lineal múltiple —varias variables explicativas—

Tema 1. Análisis de la varianza unifactorial

Analiza y compara el comportamiento de una variable continua Y sobre distintos niveles (poblaciones — grupos — tratamientos) de un factor (variable explicativa)

Ejemplos

- Producción de un cultivo con distintos fertilizantes
  Variable explicada: Producción (¿Tm?) Factor: fertilizante
- Tamaño de una misma especie en hábitats distintos
  Variable explicada: tamaño (¿Kg?) Factor: hábitat
- Contaminación por un tóxico en puntos distintos
  Variable explicada: Contaminación (¿en ppm?) Factor: localización

Tema 2. Análisis de la varianza con varios factores

Analiza y compara el comportamiento de una variable continua Y en distintos niveles de varios factores (variables explicativas) y las posibles interacciones entre ellos.

Ejemplos

- Altura de una misma especie de conífera en distintos suelos y distintos climas (factores: suelo, clima)
- Producción de un cultivo de una misma variedad (de maíz, por ejemplo) en varias parcelas con varias orientaciones y con varios con distintos fertilizantes (factores: parcela, fertilizante)
- Efecto de tres medicamentos por edad del paciente y por sexo del paciente (factores: medicamento, edad —niño, joven, adulto— y sexo)

Tema 3. Regresión lineal simple

Analiza el comportamiento de una variable continua Y a través de los valores de otra variable continua X (variable explicativa)

Ejemplos

- Volumen de un huevo de *Cuculus canorus* en función de su longitud
- Peso de un caimán en función de su longitud
- Masa de madera en función de la altura del árbol (en una determinada especie de árbol)

Tema 4. Regresión lineal múltiple

Analiza el comportamiento de una variable continua Y a través de los valores de otras variables continuas $X_1 \ldots X_k$ (variables explicativas)

Ejemplos

- Crecimiento de un tipo de cultivo en función de las cantidades de distintos nutrientes en el agua de riego
- Volumen de madera (en una determinada especie leñosa) en función de la altura y del diámetro del ejemplar
- Volumen de grasa en el cuerpo humano en función del grosor de la piel a la altura del tríceps y en la región renal

Inferencia: elementos básicos

Variable aleatoria Continua o discreta. Función de densidad. Tablas.

Modelo Variables que intervienen. Relaciones y propiedades.

Muestra Observaciones a realizar. Procedimiento.

Datos Muestra realizada. Valores de las observaciones.

Estadístico Variable aleatoria que calcula (estima) un parámetro.

Conocimientos previos

Modelos de probabilidad Prueba de Bernoulli, Binomial, Poisson, Normal, t de Student, F de Snedecor, χ^2 (Ji cuadrado),...

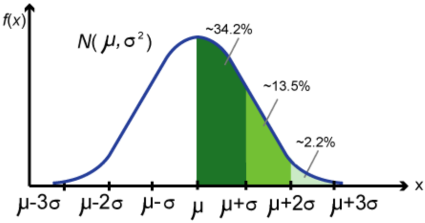
Estimación de parámetros Media muestral, varianza muestral, cuasivarianza.

Intervalos de confianza Para la media; para la varianza; para la diferencia de dos medias; para el cociente de dos varianzas.

Contrastes de hipótesis Hipótesis nula H_0 e hipótesis alternativa H_1; región $\mathcal{R}$ de rechazo de H_0.

La distribución normal

El modelo básico en nuestro estudio será el de la distribución normal: $N(\mu, \sigma)$. Media μ. Varianza: σ^2.



Función de densidad: $f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} \exp -\dfrac{1}{2}(\dfrac{x-\mu}{\sigma})^2$

$$X \sim N(\mu, \sigma) \implies Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Estimación de parámetros

Muestra aleatoria: $(X_1, \ldots, X_n)$

Cada X_i es el resultado que se obtendrá en la i-ésima observación de la variable X.

Muestra realizada (datos): $(x_1, \ldots, x_n)$

x_i es el resultado obtenido en la i-ésima observación X_i.

Estimación de parámetros

Estimador de la media μ

Media muestral: $\bar{X} = \dfrac{\sum X_i}{n}$; Estimación: $\bar{x} = \dfrac{\sum x_i}{n}$

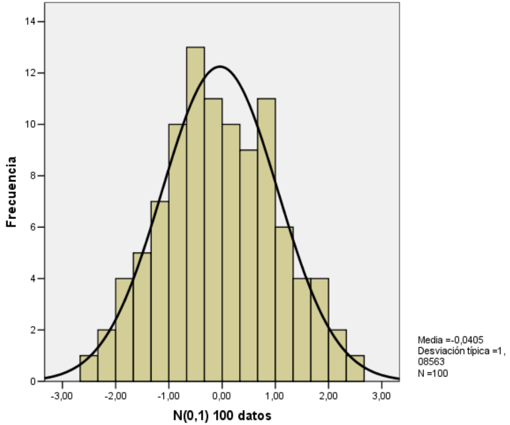
Estimadores de σ^2

Varianza muestral: $V_x = \dfrac{\sum (X_i - \bar{X})^2}{n}$; Estimación: v_x

Cuasivarianza: $S_x^2 = \dfrac{\sum (X_i - \bar{X})^2}{n-1}$; Estimación: s_x^2

Simulación

Histograma y curva Normal ajustada a 100 datos simulados con ordenador de una variable $N(0,1)$

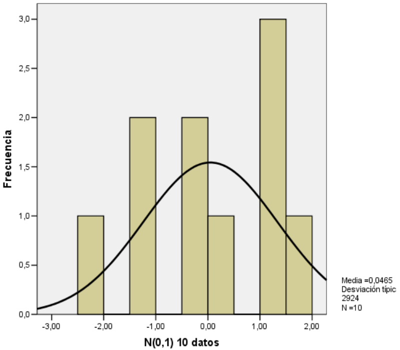


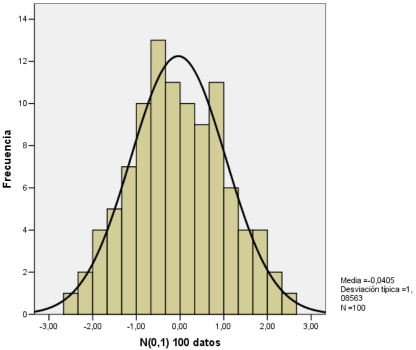
El efecto del azar y el tamaño de la muestra: simulaciones

con 10 datos	N(0,1)	N(2,1)	N(4,1)
Media	0,046	1,638	3,951
Mediana	0,005	1,850	3,885
Desviación típica	1,292	0,862	1,169
Mínimo	-2,184	-0,234	2,369
Máximo	1,733	2,656	5,583

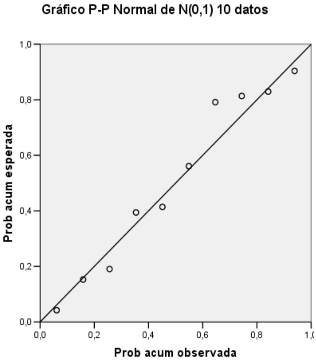
con 100 datos	N(0,1)	N(2,1)	N(4,1)
Media	-0,040	1,965	4,048
Mediana	-0,085	1,931	3,977
Desviación típica	1,086	1,006	1,062
Mínimo	-2,578	-0,474	1,933
Máximo	2,376	4,374	6,324

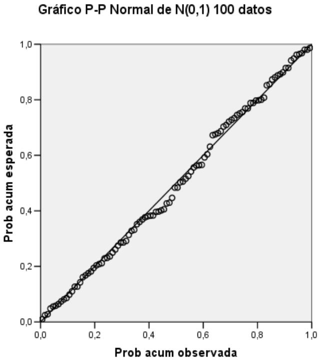
Histogramas





Diagramas pp





Intervalos de confianza

Dos poblaciones normales e independientes

$X \sim N(\mu_1, \sigma_1), \quad Y \sim N(\mu_2, \sigma_2)$ v. a. independientes

$(X_1, \dots, X_{n_1})$ m.a.s. de X; se calcula $\bar{x}$ y s_1^2

$(Y_1, \dots, Y_{n_2})$ m.a.s. de Y; se calcula $\bar{y}$ y s_2^2

Si las varianzas son iguales ($\sigma_1^2 = \sigma_2^2$) la varianza común se estima como

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Intervalo para comparar medias

Intervalo de confianza $1-\alpha$ para $\mu_1 - \mu_2$, con varianzas σ_1^2, σ_2^2 desconocidas pero iguales ($\sigma_1 = \sigma_2$)

$$\left((\bar{x} - \bar{y}) \pm t_{n_1+n_2-2;\frac{\alpha}{2}} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

En los temas 1 y 2 extenderemos estos intervalos al caso en el que tengamos más de dos poblaciones (normales e independientes).

Intervalo para comparar varianzas

Intervalo de confianza $1-\alpha$ para $\frac{\sigma_1^2}{\sigma_2^2}$

$$\left(\frac{1}{F_{n_1-1,n_2-1;\frac{\alpha}{2}}} \cdot \frac{s_1^2}{s_2^2}, \quad F_{n_2-1,n_1-1;\frac{\alpha}{2}} \cdot \frac{s_1^2}{s_2^2} \right)$$

Contrastes de hipótesis

Con la notación anterior,

Contraste t de igualdad de medias con varianzas iguales (pero desconocidas)

$$H_0 \equiv \mu_1 = \mu_2\,; \qquad \mathcal{R} = \left\{ |\bar{x} - \bar{y}| > t_{n_1+n_2-2;\frac{\alpha}{2}} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right\}$$

Previamente puede contrastarse la posiblidad de que las varianzas no sean iguales:

$$H_0 \equiv \sigma_1 = \sigma_2\,; \qquad \mathcal{R} = \left\{ \frac{s_1^2}{s_2^2} \notin \left[F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}, \quad F_{n_1-1,n_2-1;\frac{\alpha}{2}} \right] \right\}$$

NOTA: La región $\mathcal{R}$ es, en cada caso, la región de rechazo de la hipótesis nula H_0.