Clase 1

Probabilidades

Carrera de Especialización en Estadística,
Facultad de Ciencias Exactas y Naturales, UBA

2023

Sobre el cuestionario

¿Qué interpretás si te decimos que con probabilidad 0.8 una medición realizada por cierto equipo para determina un mesurando está entre 67 y 73?

Sobre el cuestionario

¿Qué interpretás si te decimos que con probabilidad 0.8 una medición realizada por cierto equipo para determina un mesurando está entre 67 y 73?

- no se
- ...si tomáramos infinitas mediciones, el promedio estaría entre 67 y 73.
- Significa que 8 de cada 10 casos la medición va a estar entre 67 y 73
- Interpreto que si el equipo hiciera infinitas mediciones, el 80% de esas mediciones estaría entre 67 y 73.

Sobre el cuestionario

¿Qué interpretás si te decimos que con probabilidad 0.8 una medición realizada por cierto equipo para determina un mesurando está entre 67 y 73?

Que si repetimos ese experimento m veces, la cantidad de veces que tenemos éxito (es decir, un mesurando entre 67 y 73) es n. El número n/m es la frecuencia relativa del evento. Cuando m es un número muy grande llamamos probabilidad del evento a ese cociente

Sobre el cuestionario

¿Qué interpretás si te decimos que con probabilidad 0.8 una medición realizada por cierto equipo para determina un mesurando está entre 67 y 73?

Interpreto que, a partir de un parámetro estaístico, el verdadero valor de la medición se encuentra entre 67 y 73 con una probabilidad de 0.8.

Unos ejercicios interesantes

Sea G_n una variable aleatoria con distribución normal, con esperanza $\mu = 70$ y varianza $\sigma^2 = 9/n$: $G_n \sim \mathcal{N}(70, 9/n)$.

1. Consideremos $n = 1$. Calcular la probabilidad de que G_1 diste de su media en más de dos unidades

$$\mathbb{P}\left(|G_1 - 70| > 2\right)$$

2. Hallar una expresión que dependa de n para la probabilidad de que G_n diste de su media en más de dos unidades.
3. Hallar n para que la probabilidad de que G_n diste de su media en más de dos unidades sea a lo sumo 0,01.
4. Determinar si fue necesario conocer el valor de μ para responder a las preguntas planteadas.

Concer μ

¿Fue necesario conocer el valor de μ para resolver el ejercicio?

$$G_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$\mathbb{P}\left(|G_n - 70| > \varepsilon\right) = \dots$$

Uno más

1. Sea $Z \sim \mathcal{N}(0, 1)$. Calcular

$$\mathbb{P}(-1.65 \leq Z \leq 1.65)$$

2. Sea $X \sim \mathcal{N}(\mu, \sigma^2)$. Calcular

$$\mathbb{P}(X - 1.65\sigma \leq \mu \leq X + 1.65\sigma)$$

3. Sea $Z \sim \mathcal{N}(0, 1)$. Hallar a de forma tal que

$$\mathbb{P}(-a \leq Z \leq a) = 0.99$$

¿Cómo seguimos?

¿Cómo seguimos?

- Variables Aleatorias

¿Cómo seguimos?

- Variables Aleatorias
- Vectores Aleatorios

Problema de Alturas

	A	B	C	D
1	altura	genero	contextura_madre	altura_madre
2	172.7	M	mediana	159.8
3	158.5	F	mediana	160.3
4	162.6	F	mediana	160.5
5	174.1	M	mediana	159.8
6	168.3	M	mediana	158.3
7	170.9	M	bajita	155.8
8	160.5	F	alta	163.5
9	169.3	M	mediana	160.5
10	172	M	mediana	159.7
11	173.5	M	bajita	155.8
12	171.5	M	mediana	157.7
13	158.8	F	mediana	160.9
14	159.5	F	alta	162.7
15	173.7	M	alta	163.6
16	160.7	F	alta	162.7
17	175.3	M	bajita	154.9
18	174.5	M	alta	163.2

- Variables
  - Y: Altura (cuantitativa)
  - X_1: Género
  - X_2: Contextura de la madre
  - X_3: Altura de la madre

Otro escenario: ¿Le damos un crédito?

Se quiere predecir si un cliente de un banco pagará un crédito.

Variables registradas:

- X_1: **balance** (saldo tarjeta)
- X_2: **income** (ingreso anual)
- X_3: **student** (si - no)

- $Y = \begin{cases} 1 & \text{(Yes)} \quad \text{default} \\ 0 & \text{(No)} \quad \text{c.c.} \end{cases}$

Paquete ISLR

```
> Default
   default student    balance    income
1       No      No  729.52650 44361.625
2       No     Yes  817.18041 12106.135
3       No      No 1073.54916 31767.139
4       No      No  529.25060 35704.494
5       No      No  785.65588 38463.496
6       No     Yes  919.58853  7491.559
7       No      No  825.51333 24905.227
8       No     Yes  808.66750 17600.451
9       No      No 1161.05785 37468.529
10      No      No    0.00000 29275.268
11      No     Yes    0.00000 21871.073
12      No     Yes 1220.58375 13268.562
13      No      No  237.04511 28251.695
14      No      No  606.74234 44994.556
15      No      No 1112.96840 23810.174
16      No      No  286.23256 45042.413
17      No      No    0.00000 50265.312
18      No     Yes  527.54018 17636.540
19      No      No  485.93686 61566.106
20      No      No 1095.07274 26464.631
21      No      No  228.95255 50500.182
22      No      No  954.26179 32457.509
23      No      No 1055.95660 51317.883
24      No      No  641.98439 30466.103
25      No      No  773.21172 34353.314
```

Vectores Aleatorios

El video que correspondiente a ese material se encuentra en este enlace

p_{XY}: Vectores aleatorios

Función de probabilidad puntual conjunta:

$$p_{XY}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x \cap Y = y)$$

X: precio kg. pan.
Y: precio docena facturas

X/Y	80	90	100	120	130	$p_X(\cdot)$
100	0.13	0.02	0.11	0.03	0.00	
120	0.06	0.01	0.04	0.02	0.00	
140	0.08	0.03	0.24	0.2	0.03	
$p_Y(\cdot)$						

Table: p_{XY} función de probabilidad puntual conjunta

$$\mathbb{P}(X = 100) = \ldots \qquad \mathbb{P}(Y = 130) = \ldots$$

Ojo al piojo

X/Y	0	1	$p_X(\cdot)$
0	0.07	0.23	0.30
1	0.53	0.17	0.70
$p_Y(\cdot)$	0.60	0.40	1.00

X/Y	0	1	$p_X(\cdot)$
0	0.01	0.29	0.30
1	0.59	0.11	0.70
$p_Y(\cdot)$	0.60	0.40	1.00

X/Y	0	1	$p_X(\cdot)$
0	0.18	0.12	0.30
1	0.42	0.28	0.70
$p_Y(\cdot)$	0.60	0.40	1.00

Vectores independientes

- $P(X \in A, Y \in B) = P(X \in A)P(X \in B)$

- $p_{X,Y}(x, y) = p_X(x)p_Y(y)$

X/Y	y_1	y_2	$p_X(\cdot)$
x_1	$p_X(x_1)p_Y(y_1)$	$p_X(x_1)p_Y(y_2)$	$p_X(x_1)$
x_2	$p_X(x_2)p_Y(y_1)$	$p_X(x_2)p_Y(y_2)$	$p_X(x_2)$
$p_Y(\cdot)$	$p_Y(y_1)$	$p_Y(y_2)$	

Esperanza de una función de un vector:

- $\mathbb{E}[h(X)] = \sum_x h(x) p_X(x)$
- $\mathbb{E}[h(X,Y)] = \sum_{x,y} h(x,y) p_{X,Y}(x,y)$
- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

Esperanza de una función de un vector:

- $\mathbb{E}[h(X,Y)] = \sum_{x,y} h(x,y) p_{X,Y}(x,y)$

- X, Y independientes, entonces $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$,

Esperanza de una función de un vector:

- $\mathbb{E}[h(X)] = \sum_x h(x)p_X(x)$
- $\mathbb{E}[h(X,Y)] = \sum_{x,y} h(x,y)p_{X,Y}(x,y)$
- $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$
- X, Y independientes, entonces $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$,
- $\mathbb{V}(X+Y) =$?

Propiedad

$X_1, \ldots, X_n$, variables aleatorias

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{E}\left(X_i\right)$$

$X_1, \ldots, X_n$ independientes, entonces $\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}\left(X_i\right)$

Suma de variables independientes: X, Y independientes

- $X \sim \mathcal{B}(n, p)$, $Y \sim \mathcal{B}(m, p)$, $X + Y \sim$

- $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\delta)$, $X + Y \sim$

Suma de variables independientes: X, Y independientes

- $X \sim \mathcal{B}(n, p)$, $Y \sim \mathcal{B}(m, p)$, entonces $X + Y \sim \mathcal{B}(n + m, p)$.

- $X \sim \mathcal{P}(\lambda)$, $Y \sim \mathcal{P}(\delta)$, entonces $X + Y \sim \mathcal{P}(\lambda + \delta)$.

Suma de muchas variables independientes

- $X_i \sim \mathcal{B}(n_i, p)$ independientes, entonces

$$\sum_{i=1}^{n} X_i \sim \mathcal{B}\left(\sum_{i=1}^{n} n_i, p\right)$$

- $X_i \sim \mathcal{P}(\lambda_i)$ independientes, entonces

$$\sum_{i=1}^{n} X_i \sim \mathcal{P}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Suma de Normales independientes

$X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ independientes, entonces

$$X + Y \sim$$

$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, independientes, entonces

$$\sum_{i=1}^{n} X_i \sim$$

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, independientes, con MISMA DISTRIBUCIóN, entonces

$$\sum_{i=1}^{n} X_i \sim$$

Suma de Normales independientes

$X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ independientes, entonces

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, independientes, entonces

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, independientes, con MISMA DISTRIBUCIóN, entonces

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$$

Suma de Normales - Ejemplo

La duración de una batería se distribuye normalmente con media de 900 días y desviación estándar 35 días.
Se ponen 10 de tales baterías en cierto equipo, de forma tal que en cuanto una batería deja de funcionar, se activa instantáneamente la siguiente. ¿Cuál es la probabilidad de que el equipo funcione al menos 9200 días?

Muestra: Ejemplo típico.

1. Realizamos un experimento muchas veces en idénticas condiciones, de manera independiente.
2. X_i: resultado de la i-ésima repetición.
3. X_i tienen todas la misma distribución porque se repite en idénticas condiciones. $X_i \sim F$
4. $X_1, \ldots, X_n$ son independientes (en sentido matemático), por que las repeticiones se realizan de manera independiente (en sentido de la vida).

Muestra: Ejemplo típico.

1. Realizamos un experimento muchas veces en idénticas condiciones, de manera independiente.
2. X_i: resultado de la i-ésima repetición.
3. X_i tienen todas la misma distribución porque se repite en idénticas condiciones. $X_i \sim F$
4. $X_1, \ldots, X_n$ son independientes (en sentido matemático), por que las repeticiones se realizan de manera independiente (en sentido de la vida).

   En tal caso, $X_i \sim F$ para todo i, y por consiguiente,

   - $P(X_i \leq t) = P(X_j \leq t) = P(X_1 \leq t)$
   - $E[X_i] = E[X_j] = E[X_1]$
   - $V(X_i) = V(X_j) = V(X_1)$.
   - mongo(X_i)=mongo(X_j)=mongo(X_1).

Muestra - variable i.i.d

Diremos que $X_1, \ldots, X_n$ son una muestra si son variables aleatorias independientes, idénticamente distribuídas.

$$X_1, \ldots, X_n \text{ , i.i.d.}$$

En tal caso, $X_i \sim F$ para todo i, y por consiguiente,

$$\text{mongo}(X_i) = \text{mongo}(X_j) = \text{mongo}(X_1).$$

Promedio de normales i.i.d

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, independientes, entonces

$$\sum_{i=1}^{n} X_i \sim$$

el promedio $\quad \overline{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i \sim$

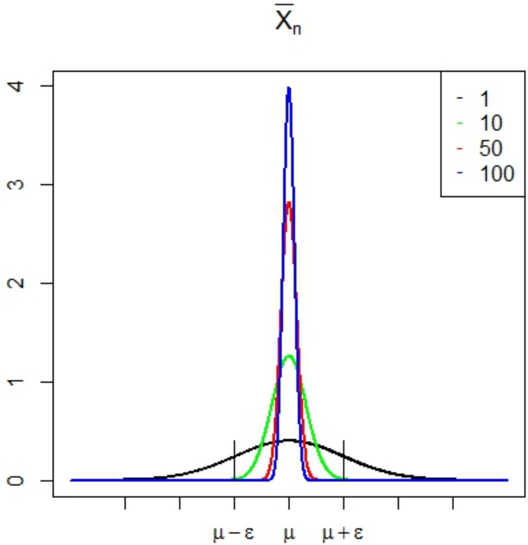
Promedio de normales i.i.d

$X_i \sim \mathcal{N}(\mu, \sigma^2)$, independientes, entonces

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(n\mu, n\sigma^2\right)$$

el promedio $\quad \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

La densidad del promedio de normales



LGN - Ejemplo normales - Parte 1:

- Sean $X_1, \ldots, X_n$ i.i.d., $X_i \sim \mathcal{N}(3, 4^2)$.

$$P\left(|\overline{X}_n - 3| > 0.01\right)$$

LGN - Ejemplo normales - Parte 2:

- Sean $X_1, \ldots, X_n$ i.i.d., $X_i \sim \mathcal{N}(\mu, 4^2)$. Encontrar n_0 de forma tal que $\forall n \geq n_0$

$$P\left(|\overline{X}_n - \mu| > 0.01\right) \leq 0.2$$

Ley de los grandes números (LGN)- Caso Normales

Sean $X_1, \ldots, X_n$ i.i.d., $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Para todo $\varepsilon > 0$, tenemos que

$$P\left(|\overline{X}_n - \mu| > \varepsilon\right) =$$

Ley de los grandes números (LGN)- Parte I

Sean $X_1, \ldots, X_n$ i.i.d., $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Para todo $\varepsilon > 0$, tenemos que

$$P\left(|\overline{X}_n - \mu| > \varepsilon\right) = 2\left(1 - \phi\left(\frac{\varepsilon * \sqrt{n}}{\sigma}\right)\right),$$

siendo $\phi(u) = P(Z \leq u)$ cuando $Z \sim \mathcal{N}(0,1)$. Luego

$$\lim_{n\to\infty} P\left(|\overline{X}_n - \mu| > \varepsilon\right) = 0$$

$$\lim_{n\to\infty} P\left(|\overline{X}_n - \mu| \leq \varepsilon\right) = 1$$

Convergencia en Probabilidad

Definition

Sean $(Y_n)_{n\geq 1}$, Y variables aleatorias. Diremos que $(Y_n)_{n\geq 1}$ converge a Y en probabilidad si para todo $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(|Y_n - Y| > \varepsilon\right) = 0$$

Notación: $Y_n \to Y$ en probabilidad

Promedio de Normales i.i.d. converge a la esperanza

Demostramos que si $(X_i)_{i\geq 1}$ i.i.d., $X_i \sim \mathcal{N}(\mu, \sigma^2)$, entonces

$$\overline{X}_n \to \mu \quad \text{en probabilidad.}$$

Ejemplo:

- Sean $X_1, \ldots, X_n$ i.i.d., $X_i \sim \mathcal{N}(\mu, 4^2)$. Encontrar n_0 de forma tal que $\forall n \geq n_0$

$$P\left(|\overline{X}_n - \mu| > 0.01\right) \leq 0.2$$

Ejercicio 2 - Práctica 3 - Modelo de mediciones

1. μ: magnitud que se desea determinar.
2. X_i: resultado de la i-ésima medición.
3. ε_i representa el error de la i-ésima medición.
4. La i- ésima medición se relaciona con el error y la magnitud de interés mediante el modelo

$$X_i = \mu + \varepsilon_i$$

5. Error (solo) aleatorio : $E(\varepsilon_i) = 0$.

Modelo de medición - Errores Normales

$$X_i = \mu + \varepsilon_i \ , \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

1. Asuma ahora que $\varepsilon_i \sim \mathcal{N}(0, 0.25)$
2. Obtenga la distribución de X_i.
3. Obtenga la distribución de $\overline{X}_n$, su esperanza y su varianza.
4. Calcule

$$\mathbb{P}(|\overline{X}_n - \mu| \leq 0.1) \ , \quad \text{para } n = 10 \text{ y para } n = 100$$

5. Obtenga una expresión para la probabilidad de que el promedio de las n mediciones diste de la verdadera magnitud μ en menos de 0.1 unidades en función de n. Estudie la monotonía y el límite cuando n tiende a infinito de esta probabilidad.
6. Determine cuán grande debe ser n para que $P\left(|\overline{X}_n - \mu| < 0.1\right) \geq 0.99$.

¿Qué vimos hoy?

- Repaso de la Normal
- Empezamos con vectores discretos
- Suma de variables aleatorias
- ¿Qué pasa con el promedio de normales?

Referencias All of Statistics - Wasserman

2.5 Bivariate Distributions

2.6 Marginal Distributions

2.7 Independent Random Variables

3.3 Variance and Covariance