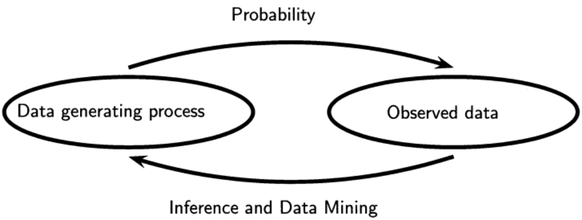
Probability

Felipe José Bravo Márquez

March 22, 2021

Probability and Statistics

- Probability is the language of uncertainty that is also the basis for statistical inference.
- The problem studied in probabilities is: given a data generating process, which are the properties of the outputs?
- The problem studied in statistical inference, data mining and machine learning is: given the outputs, what can we say about the process that generates the observed data?



[1] Figure taken from [Wasserman, 2013]

- A **random experiment** in the act of measuring a process whose output is uncertain.
- The set with all possible outputs of a random experiment is the **sample space** Ω.
- For example, $\Omega = \{1, 2, 3, 4, 5, 6\}$ is the sample space of the experiment of rolling of a die.
- An **event** $E \subseteq \Omega$ corresponds to a subset of those outputs.
- For example, $E = \{2, 4, 6\}$ is the event of observing an even number when rolling a die.

Probabilidades (II)

- Una probabilidad $\mathbb{P}$ es una función de valor real definida sobre Ω que satisface las siguientes propiedades:

Propiedades

1. Para cualquier evento $E \subseteq \Omega$, $0 \leq \mathbb{P}(E) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. Sean $E_1, E_2, \ldots, E_k \in \Omega$ conjuntos disjuntos

$$\mathbb{P}(\bigcup_{i=1}^{k} E_i) = \sum_{i}^{k} P(E_i)$$

- La probabilidad de un evento E, $\mathbb{P}(E)$ es la fracción de veces que se observaría el evento al repetir infinitamente el experimento.

Variable Aleatoria

- Una **variable aleatoria** es un mapeo

$$X : \Omega \to \mathbb{R}$$

  que asigna un valor real $X(e)$ a cualquier evento de Ω
- Ejemplo: Tiramos una moneda 10 veces. Sea $X(\omega)$ la cantidad de caras en la secuencia de resultados.
  - Si $w = CCSCCSCCSS$, entonces $X(\omega) = 6$

Ejemplo

- Tiramos una moneda 2 veces. Sea X la la cantidad de sellos obtenidos.
- La variable aleatoria y su distribución se resume como:

e	$\mathbb{P}(e)$	$X(e)$
CC	1/4	0
CS	1/4	1
SC	1/4	1
SS	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

Definiciones de V.A

- Sea X una V.A , se define **función de distribución acumulada** (CDF) o $F_X : \mathbb{R} \rightarrow [0, 1]$
$$F_X(x) = \mathbb{P}(X \leq x)$$

Variables Aleatorias Discretas

- Una V.A X es **discreta** si mapea las salidas a un conjunto contable.
- Se define la **función de probabilidad** o **función de masa de probabilidad** de una V.A X discreta como $f_X(x) = \mathbb{P}(X = x)$
- Entonces $f_X(x) \geq 0 \ \forall x \in \mathbb{R}$ y $\sum_i f_X(x_i) = 1$
- La CDF de X se relaciona con f_X de la siguiente manera:

$$F_X = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

Definiciones de V.A II

Variable Aleatoria continua

- Una V.A X es continua si:
- existe una función f_X tal que $f_X(x) \geq 0 \; \forall x$, $\int_{-\infty}^{\infty} f_X(x)dX = 1$

$$\int_{-\infty}^{\infty} f_X(x)dX = 1$$

- Para todo $a \geq b$:

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx$$

- La función f_X recibe el nombre de **función densidad de probabilidad** (PDF).
- La PDF se relaciona con la CDF como:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- Luego $f_X(x) = F'_X(x)$ en todos los puntos x donde F_X es diferenciable
- Para distribuciones continuas la probabilidad que X tomo un **valor particular** vale siempre **cero**.

Algunas Propiedades

1. $\mathbb{P}(x < X \leq y) = F(y) - F(x)$
2. $\mathbb{P}(X > x) = 1 - F(x)$
3. Si X es continua luego

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Cuantiles

- Sea X una V.A con CDF F. La CDF inversa o función cuantía se define como

$$F^{-1}(q) = inf\,\{x: \; F(x) > q\}$$

- Para $q \in [0, 1]$ si F es estrictamente creciente y continua, $F^{-1}(q)$ es el único valor real tal que $F(x) = q$
- Luego $F^{-1}(1/4)$ es l primer cuartil, $F^{-1}(1/2)$ la mediana (o segundo cuartil) y $F^{-1}(3/4)$ el tercer cuartil.

Algunas distribuciones

	Función de Probabilidad	Parámetros
Normal	$f_x = \frac{1}{\sqrt{2\pi\sigma}} \exp^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	μ, σ
Binomial	$f_x = \binom{n}{x} p^x (1-p)^{n-x}$	n, p
Poisson	$f_x = \frac{1}{x!}\lambda^x \exp^{-\lambda}$	λ
Exponencial	$f_x = \lambda \exp^{-\lambda x}$	λ
Gamma	$f_x = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp^{-\lambda x}$	λ, α
Chi-cuadrado	$f_x = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(\frac{k}{2}-1)} \exp^{-x/2}$	k

Distribución Normal

- X tiene una distribución Normal o Gaussiana de parámetros μ y σ, $X \sim N(\mu, \sigma^2)$ si
$$f_x = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$
- Donde $\mu \in \mathbb{R}$ es el "centro" o la **media** de la distribución y $\sigma > 0$ es la **desviación estándar**.
- Cuando $\mu = 0$ y $\sigma = 1$ tenemos una **Distribución Normal Estándar** denotada por Z.
- Denotamos por $\phi(z)$ a la PDF y por $\Phi(z)$ a la CDF de una Normal estándar.
- Los valores de $\Phi(z)$, $\mathbb{P}(Z \leq z)$ se encuentran tabulados.

Propiedades Útiles

1. Si $X \sim N(\mu, \sigma^2)$, luego $Z = (X - \mu)/\sigma \sim N(0, 1)$
2. Si $Z \sim N(0, 1)$, luego $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
3. Sean $X_i \sim N(\mu_i, \sigma_i^2)$,$i = 1, \ldots, n$ V.As independientes:
$$\sum_{i=1}^{n} X_i \sim N(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$$

Ejemplo Normal

- En R podemos acceder a las PDF, CDF, función cuantía y generación de números aleatorios de las distribuciones.
- Para una Normal son:

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Ejemplo

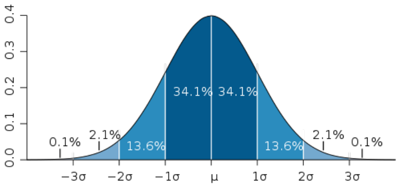
Sea $X \sim N(3, 5)$, encontrar $\mathbb{P}(X > 1)$
$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}(Z < \frac{1-3}{\sqrt{5}}) = 1 - \Phi(-0.8944) = 0.81$
En R:

```
 > 1-pnorm(q=(1-3)/sqrt(5))
[1] 0.8144533
```

O directamente:

```
> 1-pnorm(q=1,mean=3,sd=sqrt(5))
[1] 0.8144533
```

La regla 68-95-99.7 de una Normal



Sea X una V.A $\sim N(\mu, \sigma^2)$

- $\mathbb{P}(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.6827$
- $\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545$
- $\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$

En R para $X \sim N(0, 1)$:

```
> pnorm(1)-pnorm(-1)
[1] 0.6826895
> pnorm(2)-pnorm(-2)
[1] 0.9544997
> pnorm(3)-pnorm(-3)
[1] 0.9973002
```

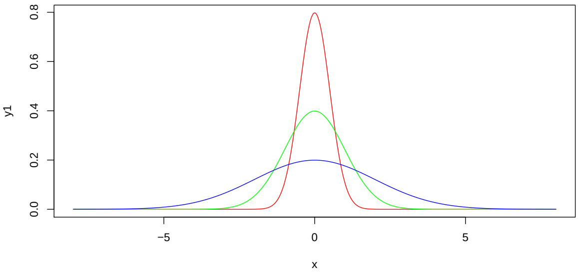
Simetría de la Normal

- La PDF de una normal es simétrica alrededor de μ
- Entonces $\phi(z) = \phi(-z)$
- $\Phi(z) = 1 - \Phi(-z)$

```
> dnorm(1)
[1] 0.2419707
> dnorm(-1)
[1] 0.2419707
> pnorm(0.95)
[1] 0.8289439
> 1-pnorm(-0.95)
[1] 0.8289439
```

Graficando la PDF de Normales con distinta varianza en R

```
x=seq(-8,8,length=400)
y1=dnorm(x,mean=0,sd=0.5)
y2=dnorm(x,mean=0,sd=1)
y3=dnorm(x,mean=0,sd=2)
plot(y1~x,type="l",col="red")
lines(y2~x,type="l",col="green")
lines(y3~x,type="l",col="blue")
```



Probabilidades Conjuntas y Condicionales

- La noción de función probabilidad (masa o densidad) se puede **extender** a más de una V.A
- Sean X Y dos V.A, $\mathbb{P}(X, Y)$ representa la **función de probabilidad conjunta**.
- Las variables son independientes entre sí, si

$$\mathbb{P}(X, Y) = \mathbb{P}(X) \times \mathbb{P}(Y)$$

- La **probabilidad condicional** para Y dado X se define como

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}$$

- Si X e Y son independientes $\mathbb{P}(Y|X) = \mathbb{P}(Y)$

Probabilidades Conjuntas y Condicionales (2)

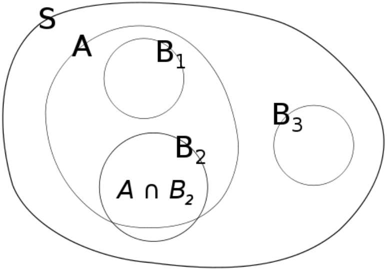


Figure: Fuente:
`en.wikipedia.org/wiki/Conditional_probability`

- Sea S el espacio muestral, A y B_n eventos.
- Las probabilidades son proporcionales al área.
- $\mathbb{P}(A) \sim 0.33$, $\mathbb{P}(A|B_1) = 1$
- $\mathbb{P}(A|B_2) \sim 0.85$ y $\mathbb{P}(A|B_3) = 0$

Teorema de Bayes y Probabilidades Totales

- La probabilidad condicional $\mathbb{P}(Y|X)$ y $\mathbb{P}(X|Y)$ pueden ser expresadas en función de la otra usando el **teorema de Bayes**

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

- Se entiende a $P(Y|X)$ como la fracción de veces que Y ocurre cuando se sabe que ocurre X.
- Luego sea $\{Y_1, Y_2, \ldots, Y_k\}$ un conjunto de salidas mutuamente excluyentes de una V.A X, el denominador del teorema de Bayes se puede expresar como:

$$\mathbb{P}(X) = \sum_{i=1}^{k} \mathbb{P}(X, Y_i) = \sum_{i=1}^{k} \mathbb{P}(X|Y_i)\mathbb{P}(Y_i)$$

Ejemplo

- Divido mis correos en tres categorías: A_1=“spam”, A_2=“baja prioridad”, A_3=“alta prioridad”
- Sabemos que $\mathbb{P}(A_1) = 0.7$, $\mathbb{P}(A_2) = 0.2$ y $\mathbb{P}(A_3) = 0.1$, claramente $0.7 + 0.2 + 0.1 = 1$
- Sea B el evento de que el correo contenga la palabra “gratis”.
- Sabemos que $\mathbb{P}(B|A_1) = 0.9$ $\mathbb{P}(B|A_2) = 0.01$ y $\mathbb{P}(B|A_3) = 0.01$ claramente $0.9 + 0.01 + 0.01 \neq 1$
- Cual es la probabilidad de que sea “spam” un correo que tiene la palabra “gratis”?
- Usando Bayes y Probabilidades totales:

$$\mathbb{P}(A_1|B) = \frac{0.9 \times 0.7}{(0.9 \times 0.7) + (0.01 \times 0.2) + (0.01 \times 0.1)} = 0.995$$

Esperanza

- Sea X una V.A, se define su **esperanza** o **momento de primer orden** como:

$$\mathbb{E}(X) = \begin{cases} \sum_x (x \times f(x)) & \text{Si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} (x \times f(x))dx & \text{Si } X \text{ es continua} \end{cases}$$

- Es el promedio ponderado de todos los posibles valores que puede tomar una variable aleatoria
- Para el caso de lanzar dos veces una moneda con X el número de caras:

$$\begin{aligned} \mathbb{E}(X) &= (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) \\ &= (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1 \end{aligned}$$

- Sean las variables aleatorias $X_1, X_2, \ldots, X_n$ y las constantes $a_1, a_2, \ldots, a_n$,

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i \mathbb{E}(X_i)$$

Varianza

- La varianza mide la "dispersión" de una distribución
- Sea X una V.A de media μ, se define la varianza de X denotada como σ^2, σ_X^2 o $\mathbb{V}(X)$ como:

$$\mathbb{V}(X) = \mathbb{E}(X-\mu)^2 = \begin{cases} \sum_{i=1}^{n} f_X(x_i)(x_i-\mu)^2 & \text{Si } X \text{ es discreta} \\ \int (x-\mu)^2 f_X(x)dx & \text{Si } X \text{ es continua} \end{cases}$$

- La **desviación estándar** σ se define como $\sqrt{\mathbb{V}(X)}$

Propiedades

- $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mu^2$
- Si a y b son constantes, luego $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$
- Si $X_1, \ldots, X_n$ son independientes y $a_1, \ldots, a_n$ son constantes, luego

$$\mathbb{V}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i)$$

Ley de los Grandes Números

Forma Débil

- Sean $X_1, X_2, \ldots X_n$ variables aleatorias IID de media μ y varianza σ^2
- El promedio $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$ converge en probabilidad a μ, $\overline{X_n} \xrightarrow{P} \mu$
- Esto es equivalente a decir que para todo $\epsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}(|\overline{X_n} - \mu| < \epsilon) = 1$$

- Entonces la distribución de $\overline{X_n}$ se concentra alrededor de μ cuando n crece.

Ejemplo

- Sea el experimento de lanzar una moneda donde la probabilidad de cara es p
- Para una V.A de distribución Bernoulli $E(X) = p$
- Sea $\overline{X_n}$ la fracción de caras después de n lanzamientos.
- La ley de los grandes números nos dice que $\overline{X_n}$ converge en probabilidad a p
- Esto no implica que $\overline{X_n}$ sea numéricamente igual a p
- Si n en grande la distribución de $\overline{X_n}$ estará concentrada alrededor de p.

Teorema Central del Límite

- Si bien la ley de los grandes números nos dice que $\overline{X_n}$ se acerca a μ
- Esto no es suficiente para afirmar algo sobre la distribución de $\overline{X_n}$

Teorema Central del Límite (CLT)

- Sean $X_1, \ldots, X_n$ variables aleatorias IID de media μ y varianza σ^2
- Sea $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$

$$Z_n \equiv \frac{\overline{X_n} - \mu}{\sqrt{\mathbb{V}(\overline{X_n})}} = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow Z$$

  donde $Z \sim N(0,1)$
- Esto es equivalente a:

$$\lim_{n \to \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Teorema Central del Límite (2)

- El teorema nos permite aproximar la distribución de $\overline{X_n}$ a una normal cuando n es grande.
- Aunque no sepamos la distribución de X_i, podemos aproximar la distribución de la media.

Notaciones alternativas que muestran que Z_n converge a una Normal

$$
\begin{aligned}
Z_n &\approx N(0,1) \\
\overline{X_n} &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X_n} - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X_n} - \mu) &\approx N(0, \sigma^2) \\
\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} &\approx N(0,1)
\end{aligned}
$$

Teorema Central del Límite (3)

- Supongamos que el número de errores de un programa computacional sigue una distribución de Poisson con parámetro $\lambda = 5$
- Si $X \sim \mathit{Poisson}(\lambda)$, $\mathbb{E}(X) = \lambda$ y $\mathbb{V}(X) = \lambda$.
- Si tenemos 125 programas independientes $X_1, \ldots, X_{125}$ nos gustaría aproximar $\mathbb{P}(\overline{X_n} < 5.5)$
- Usando el CLT tenemos que

$$\begin{aligned}
\mathbb{P}(\overline{X_n} < 5.5) &= \mathbb{P}\left(\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{5.5 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
&\approx \mathbb{P}\left(Z < \frac{5.5 - 5}{\frac{\sqrt{5}}{\sqrt{125}}}\right) = \mathbb{P}(Z < 2.5) = 0.9938
\end{aligned}$$

References I

Wasserman, L. (2013).
All of statistics: a concise course in statistical inference.
Springer Science & Business Media.