Probabilidades y Estadística para la Minería de Datos

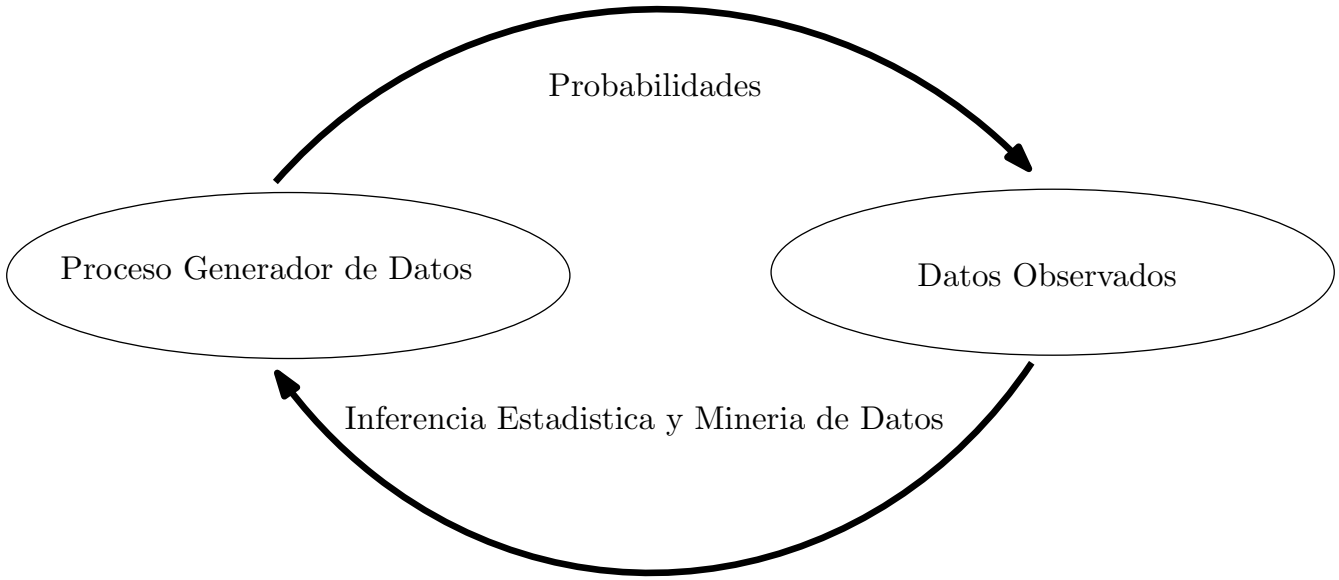
Felipe José Bravo Márquez

Universidad de Chile - Departamento de Ciencias de la Computación -Minería de Datos

4 de septiembre de 2012

Motivación

- Las probabilidades son el lenguaje de la incertidumbre que a la vez es la base de la inferencia estadística.
- El problema estudiado en probabilidades es: dado un proceso generador de datos, ¿cuáles son las propiedades de las salidas?
- El problema estudiado en inferencia estadística, minería de datos y machine learning es: dadas las salidas, ¿qué podemos decir del proceso que genera los datos observados?



Probabilidades

- Un **experimento aleatorio** en el acto de medir un proceso cuya salida es incierta
- El conjunto con todas las posibles salidas de un experimento aleatorio es el **espacio muestral** Ω
- Ej: $\Omega = \{1, 2, 3, 4, 5, 6\}$ es el espacio muestral del lanzamiento de un **dado**.
- Un **evento** $E \subseteq \Omega$ corresponde a un subconjunto de esas salidas
- Ej: $E = \{2, 4, 6\}$ es el evento de observar un número par al lanzar un dado

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

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

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 $\mathbb{P}(X > 1)$

$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}(Z < \frac{1-3}{\sqrt{5}}) = 1 - \Phi(-0{,}8944) = 0{,}81$

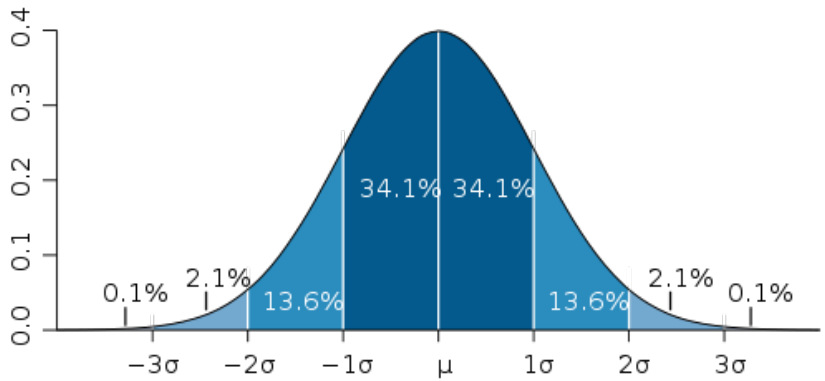
En R:

```
 > 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
```

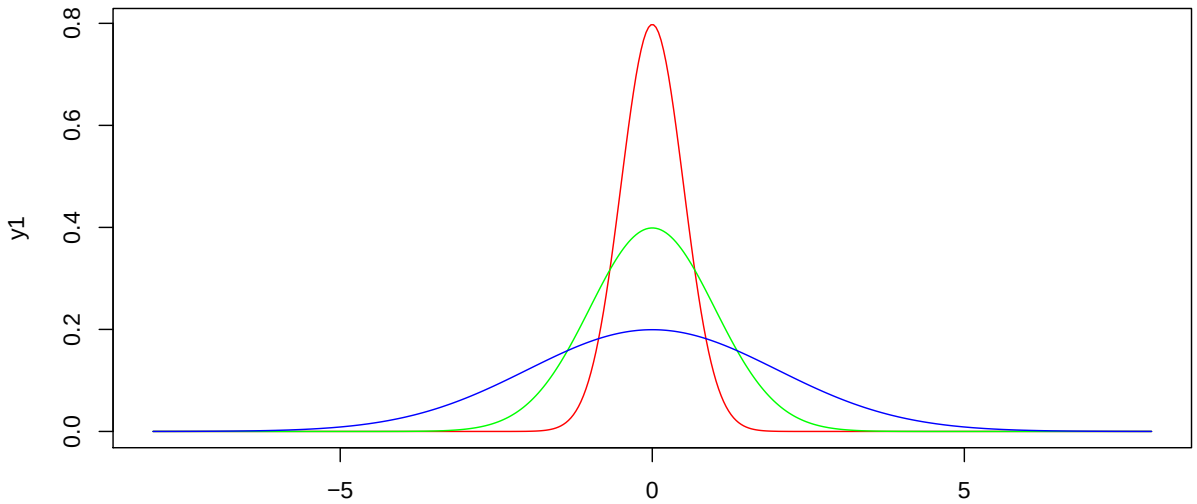
Simetría de la Normal

- 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)

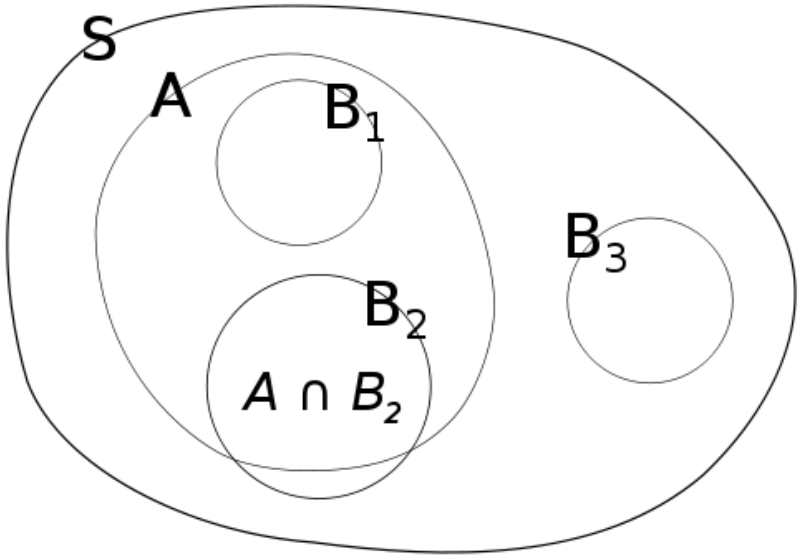


Figura: Fuente:
en.wikipedia.org/wiki/Conditional_probability

- 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 $X \sim Poisson(\lambda)$, $\mathbb{E}(X) = \lambda$ 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

$$\begin{aligned} \mathbb{P}(\overline{X_n} < 5{,}5) &= \mathbb{P}\left(\frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{5{,}5 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\ &\approx \mathbb{P}\left(Z < \frac{5{,}5 - 5}{\frac{\sqrt{5}}{\sqrt{125}}}\right) = \mathbb{P}(Z < 2{,}5) = 0{,}9938 \end{aligned}$$

Inferencia Estadística

- Para realizar conclusiones sobre una **población**, generalmente no es factible reunir todos los datos de ésta.
- Debemos realizar conclusiones razonables respecto a una población basado en la evidencia otorgada por **datos muestrales**.
- El proceso de realizar conclusiones sobre una población a partir de datos muestrales se conoce como **inferencia estadística**.

Inferencia Estadística (2)

- En inferencia estadística tratamos de **inferir** la distribución que genera los datos observados
- Ejemplo: Dado una muestra $X_1, \ldots, X_n \sim F$. ¿Cómo inferimos F?
- En algunos cosas sólo nos interesa inferir alguna propiedad de F como su **media**.
- Los modelos estadísticos que asumen que la distribución se puede modelar con un conjunto finito de parámetros $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ se llaman modelos **paramétricos**.
- Ejemplo: si asumimos que los datos vienen de una distribución normal $N(\mu, \sigma^2)$, μ y σ serían los parámetros del modelo.
- Un **estadístico** (muestral) es una medida cuantitativa calculada a partir de los datos.

Estimación Puntual

- La estimación puntual es el proceso de encontrar la **mejor aproximación** de una cantidad de interés a partir de una **muestra estadística**.
- La cantidad de interés puede ser: un parámetro en un modelo paramétrico, una CDF, una PDF, o una función de regresión.
- Por convención se denota a la estimación puntual del valor de interés θ como $\hat{\theta}$ o $\hat{\theta}_n$
- Es importante remarcar que mientras θ es un valor fijo desconocido, $\hat{\theta}$ depende de los datos y por ende es una variable aleatoria.

Estimación Puntual (2)

Definición Formal

- Sean $X_1, \ldots, X_n$ n observaciones IID de una distribución F
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es una función de $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

- El **sesgo** (bias) de un estimador se define como:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- Un estimador es insesgado si $\mathbb{E}(\hat{\theta}_n) = \theta$ o $\text{bias}(\hat{\theta}_n) = 0$

Estimación Puntual (3)

- La distribución de $\hat{\theta}_n$ se conoce como la **distribución muestral**
- La desviación estándar de $\hat{\theta}_n$ se conoce como **error estándar** *se*:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- El error estándar nos habla sobre la variabilidad del estimador entre todas las posibles muestras de un mismo tamaño.

Estimación Puntual (4)

- Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una población de media μ y varianza σ^2
- Se define la **media muestral** $\overline{X_n}$ o $\hat{\mu}$ como:

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Es un estimador insesgado:

$$\mathbb{E}(\overline{X_n}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n} \times \mathbb{E}(\sum_{i=1}^{n} X_i) = \frac{1}{n}(n \times \mu) = \mu$$

- Su error estándar sería $se(\overline{X_n}) = \sqrt{\mathbb{V}(\overline{X_n})}$ donde

$$\mathbb{V}(\overline{X_n}) = \mathbb{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathbb{V}(\sum_{i=1}^{n} X_i) = \frac{n}{n^2}\mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Entonces $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$

Ejemplos de Estimación Puntual (5)

- Por lo general no sabemos σ de la población.
- Cuando queremos estimar la varianza de una población a partir de una muestra hablamos de la **varianza muestral**:
- Existen dos estimadores comunes, una versión sesgada

$$s_n^2 = \frac{1}{n}\sum_i^n (X_i - \overline{X_n})^2$$

- Una versión sin sesgo

$$s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$$

- Cuando no sabemos la varianza de la población y queremos estimar la media, el error estándar es estimado:

$$\hat{se}(\overline{X_n}) = \frac{s}{\sqrt{n}}$$

Estimación Puntual (6)

- Sean $X_1, \ldots, X_n \sim$ Bernoulli(p) y sea $\hat{p}_n = \frac{1}{n}\sum_i X_i$
- Luego $\mathbb{E}(\hat{p}_n) = \frac{1}{n}\sum_i \mathbb{E}(X_i) = p$, entonces $\hat{p}_n$ es insesgado.
- El error estándar *se* sería

$$se = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$$

- El error estándar estimado $\hat{se}$:

$$\hat{se} = \sqrt{\hat{p}(1-\hat{p})/n}$$

Estimación Puntual (7)

- Se espera que un buen estimador sea insesgado y de mínima varianza.
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es **consistente** si converge al valor verdadero cuando el número de datos de la muestra tiende a infinito.
- La calidad de un estimador se puede medir usando el **error cuadrático medio** (MSE)

$$MSE = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2$$

Estimación Puntual (8)

- Si para un estimador $\hat{\theta}_n$, su *bias* $\to 0$ y su *se* $\to 0$ cuando $n \to \infty$, $\hat{\theta}_n$ es un estimador consistente de θ.
- Por ejemplo, para la media muestral $\mathbb{E}(\overline{X_n}) = \mu$ lo que implica que el *bias* $= 0$ y $se(\overline{X_n}) = \frac{\sigma}{\sqrt{n}}$ que tiende a cero cuando $n \to \infty$. Entonces $\overline{X_n}$ es un estimador consistente de la media.
- Para el caso del experimento Bernoulli se tiene que $\mathbb{E}(\hat{p}) = p \Rightarrow$ *bias* $= 0$ y $se = \sqrt{p(1-p)/n} \to 0$ cuando $n \to \infty$. Entonces $\hat{p}$ es un estimador consistente de p.

Intervalo de Confianza

- Sabemos que el valor de un estimador puntual **varía** entre una muestra y otra
- Es más razonable encontrar un **intervalo** donde sepamos que valor **real del parámetro** se encuentra dentro del intervalo con una cierta **probabilidad**.
- La forma general de un intervalo de confianza en las siguiente:

$$\text{Intervalo de Confianza} = \text{Estadístico Muestral} \pm \text{Margen de Error}$$

- Entre más ancho el intervalo mayor incertidumbre existe sobre el valor del parámetro.

Intervalo de Confianza (2)

Definición

- Un **intervalo de confianza** para un parámetro poblacional desconocido θ con un **nivel de confianza** $1-\alpha$, es un intervalo $C_n = (a, b)$ donde:

$$\mathbb{P}(\theta \in C_n) = 1 - \alpha$$

- Además $a = a(X_1, \ldots, X_n)$ y $b = b(X_1, \ldots, X_n)$ son funciones de los datos
- El valor α se conoce como el nivel de **significancia**, generalmente se toma como 0,05 lo que equivale a trabajar con un nivel de confianza de 95 %
- La significancia se puede interpretar como la probabilidad de equivocarnos.

Intervalo de Confianza (3)

Interpretación

- Existe mucha **confusión** de como interpretar un intervalo de confianza
- Una forma de interpretarlos es decir que si repetimos **un mismo experimento** muchas veces, el intervalo contendrá el valor del parámetro el $(1-\alpha)$ % de las veces.
- Esta interpretación es correcta, pero rara vez repetimos un mismo experimento varias veces.
- Una interpretación mejor: un día recolecto datos creo un intervalo de 95 % de confianza para un parámetro θ_1. Luego, en el día 2 hago lo mismo para un parámetro θ_2 y así reiteradamente *n* veces. El 95 % de mis intervalos **contendrá** los valores reales de los parámetros.

Intervalo de Confianza (4)

- Se tienen n observaciones independientes $X_1, \ldots, X_n$ IID de distribución $N(\mu, \sigma^2)$
- Supongamos que μ es **desconocido** pero σ^2 es **conocido**.
- Sabemos que $\overline{X_n}$ es un estimador insesgado de μ
- Por la ley de los grandes números sabemos que la distribución de $\overline{X_n}$ se concentra alrededor de μ cuando n es grande.
- Por el CLT sabemos que

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

  cuando n es grande
- Despejando, tenemos que $\mu = \overline{X_n} - \frac{\sigma}{\sqrt{n}} Z$

Intervalo de Confianza (5)

- Queremos encontrar un intervalo intervalo $C_n = (\mu_1, \mu_2)$ con un nivel de confianza $1 - \alpha$:

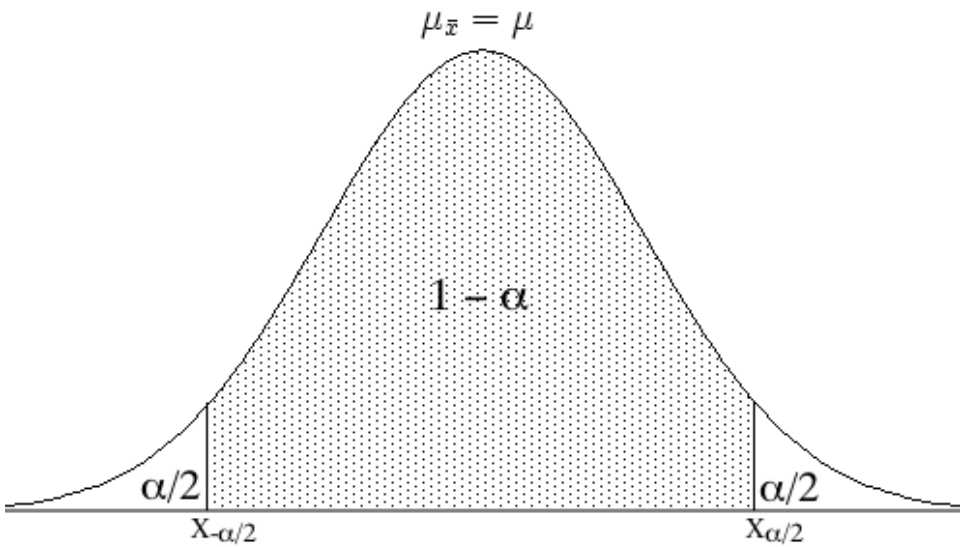
$$\mathbb{P}(\mu_1 \leq \mu \leq \mu_2) = 1 - \alpha$$

- Sea $z_a = \Phi^{-1}(1 - a)$, con $a \in [0, 1]$ donde Φ^{-1} es la función cuantía de una normal estandarizada
- Esto es equivalente a decir que z_a es el valor tal que $1 - \Phi(z_a) = \mathbb{P}(Z \geq z_a) = a$
- Por simetría de la normal $z_{\alpha/2} = -z_{(1-\alpha/2)}$

Intervalo de Confianza (6)

- Se tiene que

$$\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$



Intervalo de Confianza (7)

- El intervalo de confianza para μ es:

$$C_n = (\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$

- Entonces $z_{\alpha/2}$ nos dice cuantas veces tenemos que multiplicar el **error estándar** en el intervalo.
- Mientras menor sea α mayor será $z_{\alpha/2}$ y por ende más ancho será el intervalo.
- Demostración:

$$\begin{aligned}
\mathbb{P}(\mu \in C_n) &= \mathbb{P}(\overline{X_n} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X_n} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \\
&= \mathbb{P}(-z_{\alpha/2} < \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha/2}) \\
&= \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= 1 - \alpha
\end{aligned}$$

Intervalo de Confianza (8)

- Como $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ podemos usar la función cuantía de la normal para calcular intervalos de confianza en R

```
> alpha <- 0.05
> xbar <- 5
> sigma <- 2
> n <- 20
> se <-sigma/sqrt(n)
> error <- qnorm(1-alpha/2)*se
> left <- xbar-error
> right <- xbar+error
> left
[1] 4.123477
> right
[1] 5.876523
>
```

Distribución T

- En la práctica, si no conocemos μ es poco probable que conozcamos σ
- Si estimamos σ usando s, los intervalos de confianza se construyen usando la distribución **T-student**

Distribución T

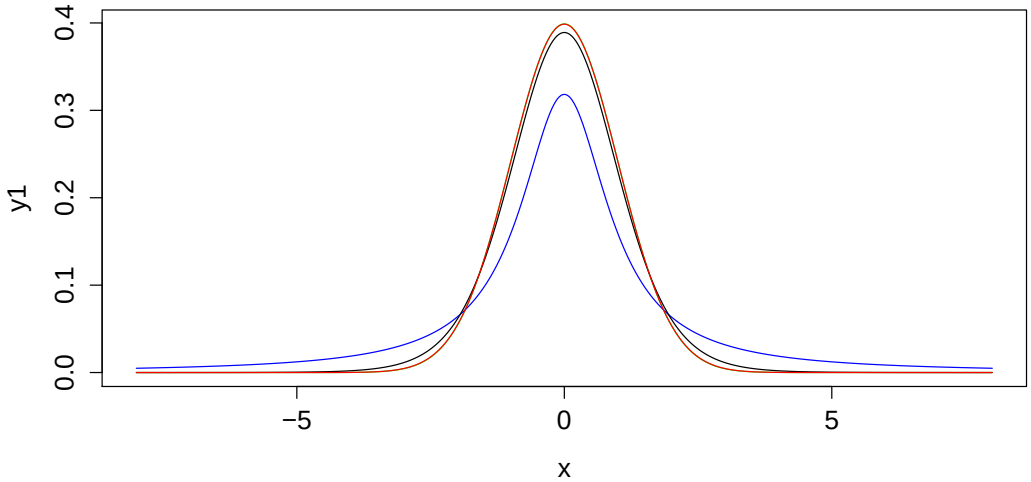
- Una V.A tiene distribución t con k grados de libertad cuando tiene la siguiente PDF:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})(1+\frac{t^2}{k})^{(k+1)/2}}$$

- Cuando $k = 1$ se le llama distribución de **Cauchy**
- Cuando $k \to \infty$ converge a una distribución normal estandarizada
- La distribución t tiene colas más anchas que la normal cuando tiene pocos grados de libertad

Distribución T (2)

```
x<-seq(-8,8,length=400)
y1<-dnorm(x)
y2<-dt(x=x,df=1)
y3<-dt(x=x,df=10)
y4<-dt(x=x,df=350)
plot(y1~x,type="l",col="green")
lines(y2~x,type="l",col="blue")
lines(y3~x,type="l",col="black")
lines(y4~x,type="l",col="red")
```



Intervalo de Confianza (9)

- Sea $s^2 = \frac{1}{n-1}\sum_i^n (X_i - \overline{X_n})^2$ tenemos:

$$T = \frac{\overline{X_n} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

- Sea $t_{n-1,a} = \mathbb{P}(T > a)$, equivalente a la función cuantía qt evaluada en $(1 - a)$
- El intervalo de confianza resultante es:

$$C_n = (\overline{X_n} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \overline{X_n} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}})$$

- Como las colas de la distribución t son más anchos cuando n es pequeño, los intervalos de confianza resultantes son más anchos

Intervalo de Confianza (10)

- Calculemos un intervalo de confianza para la media de `Petal.Length` de los datos del **Iris** con 95 % de confianza

```
>data(iris)
>alpha<-0.05
>n<-length(iris$Petal.Length)
>xbar<-mean(iris$Petal.Length)
>xbar
[1] 3.758
>s<-sd(iris$Petal.Length)
>se<-s/sqrt(n)
>error<-qt(p=1-alpha/2,df=n-1)*se
>left<-xbar-error
>left
[1] 3.473185
>right<-xbar+error
>right
[1] 4.042815
```

- Otra forma:

```
>test<-t.test(iris$Petal.Length,conf.level=0.95)
>test$conf.int
[1] 3.473185 4.042815
```

Test de Hipótesis

- Cuando queremos probar si alguna **propiedad** asumida sobre una población se contrasta con una muestra estadística usamos un **Test de Hipótesis**
- El test se compone de las siguientes hipótesis:
  - **Hipótesis Nula** H_0: Simboliza la situación actual. Lo que se ha considerado real hasta el presente.
  - **Hipótesis Alternativa** H_a: es el modelo alternativo que queremos considerar.
- La idea es encontrar suficiente **evidencia estadística** para rechazar H_0 y poder concluir H_a
- Si no tenemos suficiente evidencia estadística **fallamos en rechazar** H_0

Test de Hipótesis (2)

Metodología para Realizar un Test de Hipótesis

- Elegir una hipótesis nula H_0 y alternativa H_a
- Fijar un nivel de significancia α del test
- Calcular un estadístico T a partir de los datos
- El estadístico T es generalmente un valor estandarizado que podemos chequear en una tabla de distribución
- Definir un criterio de rechazo para la hipótesis nula. Generalmente es un valor crítico c.

Test de Hipótesis (3)

- Ejemplo: Se sabe que la cantidad de horas promedio de uso de Internet mensual en Chile país es de 30 horas
- Supongamos que queremos demostrar que el promedio es distinto a ese valor.
- Tendríamos que $H_0 : \mu = 30$ y $H_a : \mu \neq 30$
- Fijamos $\alpha = 0{,}05$ y recolectamos 100 observaciones
- Supongamos que obtenemos $\overline{X_n} = 28$ y $s = 10$
- Una forma de hacer el test es construir un intervalo de confianza para μ y ver si H_0 está en el intervalo.

```
> 28-qt(p=0.975,99)*10/sqrt(100)
[1] 26.01578
> 28+qt(p=0.975,99)*10/sqrt(100)
[1] 29.98422
```

- El intervalo sería la zona de aceptación de H_0 y todo lo que esté fuera de éste será mi región de rechazo.
- Como 30 está en la región de rechazo, rechazo mi hipótesis nula con un 5 % de confianza.

Test de Hipótesis (4)

- Otra forma de realizar el test es calcular el estadístico $T = \frac{\overline{X_n} - \mu_o}{\frac{s}{\sqrt{n}}}$
- En este caso sería

$$T = \frac{28 - 30}{\frac{10}{\sqrt{100}}} = -2$$

- Como $H_a : \mu \neq 30$, tenemos un test de dos lados, donde la región de aceptación es

$$t_{n-1,1-\alpha/2} < T < t_{n-1,\alpha/2}$$

```
 > qt(0.025,99)
[1] -1.984217
> qt(0.975,99)
[1] 1.984217
```

- Como T está en la región de rechazo, rechazamos la hipótesis nula.

Test de Hipótesis (5)

- Generalmente, además de saber si rechazamos o fallamos en rechazar una hipótesis nula queremos saber la evidencia que tenemos en contra de ella.
- Se define un **p-valor** como la probabilidad de obtener un resultado al menos tan extremo como el observado en los datos dado que la hipótesis nula es verdadera.
- Si el **p-valor** es menor que el nivel de significancia α, rechazamos H_0
- Ejemplo:

```
> data(iris)
> mu<-3 # La hipótesis nula
> alpha<-0.05
> n<-length(iris$Petal.Length)
> xbar<-mean(iris$Petal.Length)
> s<-sd(iris$Petal.Length)
> se<-s/sqrt(n)
> t<-(xbar-mu)/(s/sqrt(n))
> pvalue<-2*pt(-abs(t),df=n-1)
> pvalue
[1] 4.94568e-07 # es menor que 0.05 entonces rechazamos H0
```

Test de Hipótesis (6)

- La forma elegante de hacerlo en R:

```
> t.test(x=iris$Petal.Length,mu=3)

One Sample t-test

data:  iris$Petal.Length
t = 5.2589, df = 149, p-value = 4.946e-07
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 3.473185 4.042815
sample estimates:
mean of x
    3.758
```

Test de Hipótesis (7)

- Tenemos dos tipos de errores cuando realizamos un test de hipótesis
- Error tipo I: es cuando rechazamos la hipótesis nula cuando ésta es cierta.
- Este error es equivalente al nivel de significancia α
- Error tipo II: es cuando la hipótesis nula es falsa pero no tenemos evidencia estadística para rechazarla.
- Para mitigar los errores tipo I generalmente usamos valores de α más pequeños.
- Para mitigar los errores tipo II generalmente trabajamos con muestras más grandes.
- Existe un trade-off entre los errores tipo I y tipo II.

	Retener H_0	Rechazar H_0
H_0 es verdadera	✓	error tipo I
H_1 es verdadera	error tipo II	✓

Bilbiografía I


L. Wasserman *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.