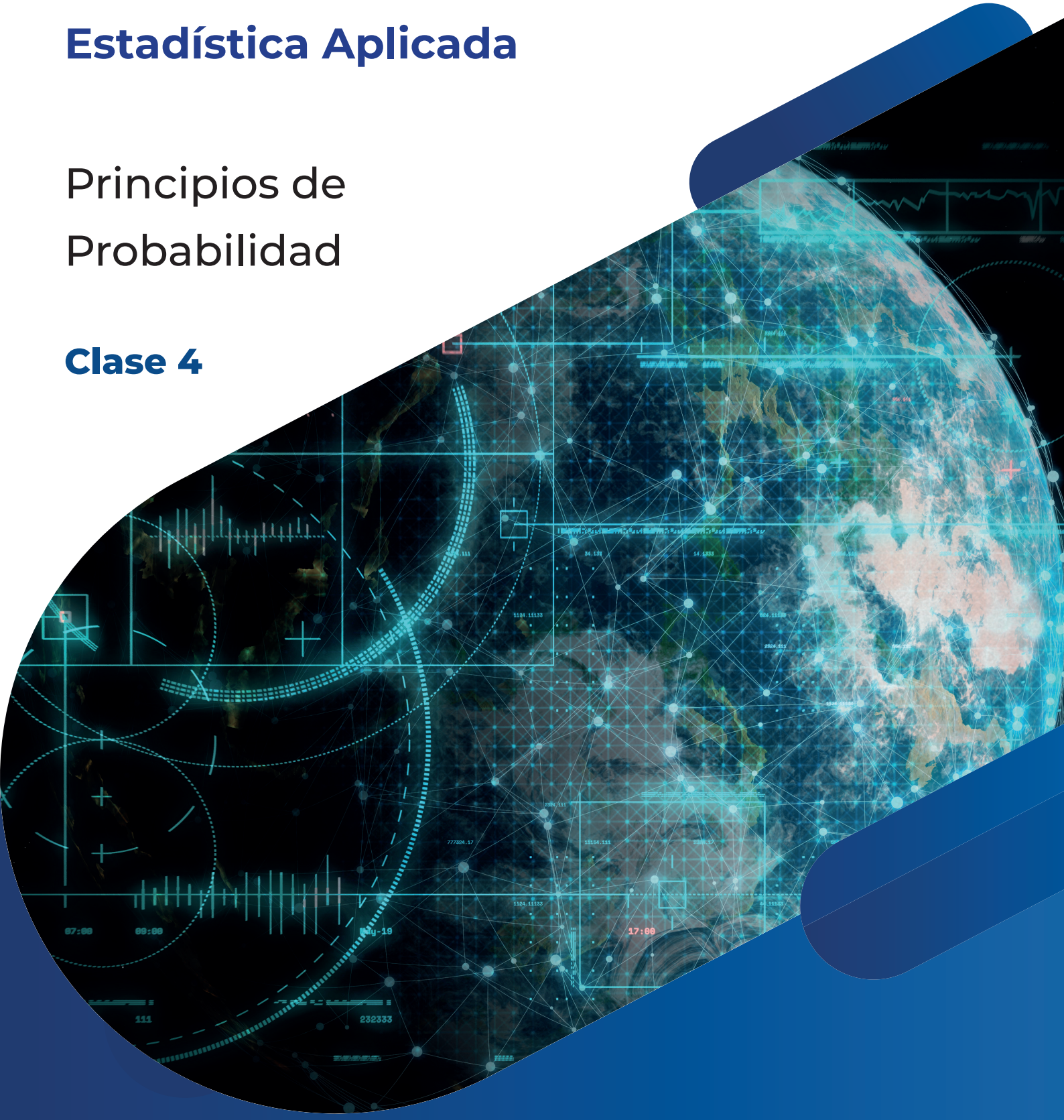
Estadística Aplicada

Principios de Probabilidad

Clase 4



Ingeniería en ciberseguridad

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, vamos a cubrir los conceptos de distribución. En estadística, cuando nos referimos a la palabra "distribución," en realidad queremos decir distribución de probabilidades. En el área de conocimiento de la estadística, existe una gran cantidad de distribuciones; algunas representan distribuciones que se presentan en problemas de la vida real, y otras son más bien "curiosidades" matemáticas que nos ayudan en la resolución de problemas.

En clases anteriores, habíamos discutido los dos tipos de variables numéricas, a saber: continuas y discretas, y en función de estos tipos, vamos a estudiar las distribuciones que aplican en cada caso. En esta clase nos centraremos en la distribución gaussiana y normal.

4) Medición de Probabilidad variables continuas

¿Qué es una distribución?

Cuando usamos el término "distribución" en estadística, usualmente queremos decir una distribución de probabilidad. Una distribución de probabilidad es una función que muestra los valores posibles para una variable y con qué frecuencia ocurren.

Cuando son datos numéricos, a menudo se ordenan de menor a mayor y se calcula la probabilidad de obtener cada uno de estos y por último los visualizaremos para examinar la forma, el centro y la cantidad de variabilidad en los datos.

Cuando son datos numéricos, a menudo se ordenan de menor a mayor y se calcula la probabilidad de obtener cada uno de estos; por último, los visualizaremos para examinar la forma, el centro y la cantidad de variabilidad en los datos.

Cuando examinamos un proceso generador de datos, podríamos decir que, en general, el proceso podría tener datos que van desde -∞ a +∞, pero en la práctica, la mayoría de los fenómenos, eventos y procesos, etcétera, que se examinan tienden a agruparse alrededor de un valor medio, y mientras más se alejan de esta media, la probabilidad de ocurrencia disminuye. Si graficamos la probabilidad de ocurrencia, veremos que sigue una forma de campana.

4.1) Distribución normal

En esta distribución de probabilidades, la variable es de tipo continua, por lo tanto, la variable puede tomar una cantidad infinita de valores. Para cada valor de la variable aleatoria, calcularemos su probabilidad y, si trazamos un diagrama en el cual en el eje X colocamos el valor de la variable y en el eje Y la probabilidad obtenida, veremos que van delineando la forma de una campana de Gauss; por eso, le llamamos una distribución gaussiana.

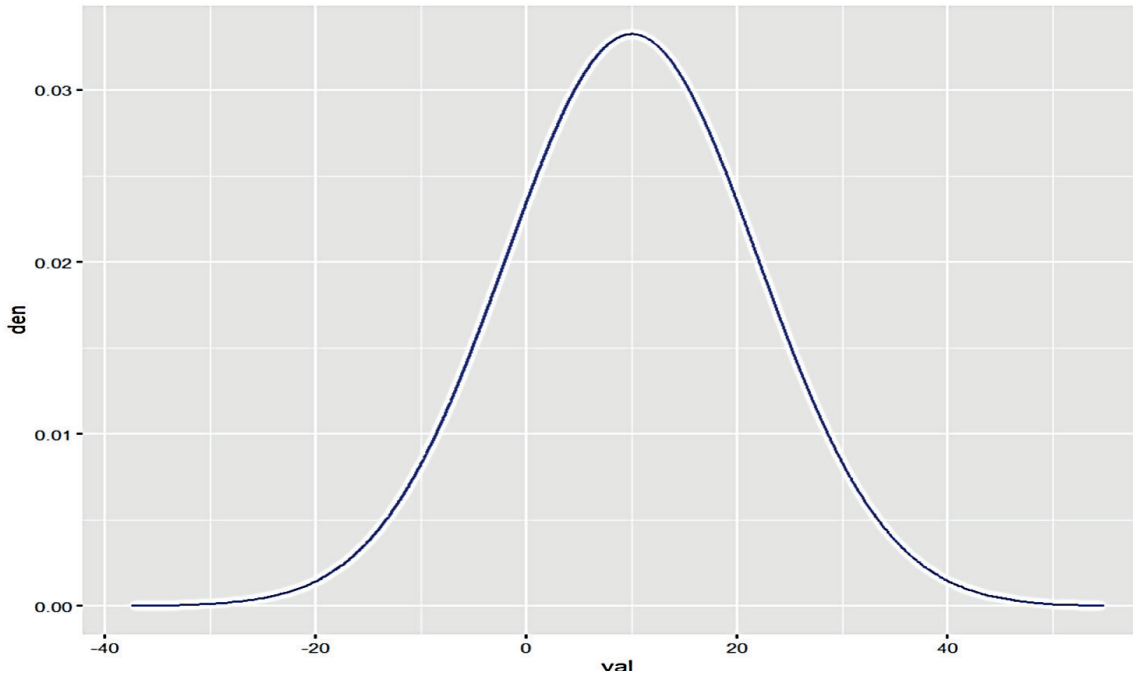
Esta distribución tiene las siguientes características:

- ✓ La curva es simétrica con respecto a su media. Hay más probabilidad de valores cercanos a la media, el 50% de los resultados caen debajo de la media, el otro 50% sobre la media.
- ✓ La curva decrece uniformemente en ambas direcciones a partir del valor central.
- ✓ Es asintótica, esto significa que la curva se acerca cada vez más al eje x, pero en realidad nunca llega a tocarlo. Esto es, los puntos extremos de la curva se extienden indefinidamente en ambas direcciones.
- ✓ La media, mediana y moda son iguales.

Figura 1

Distribución de una variable gaussiana

Creación de autor Alfonso Prado



Por supuesto, no todos los procesos que generan este tipo de distribución presentan exactamente esta curva. Note, por ejemplo, que esta curva está centrada en el valor 10 y, aunque no es fácil de visualizar, una desviación estándar de 12. Para entender este concepto, veamos la función matemática que genera esta curva.

Figura 2

Función CDF de la distribución gaussiana

$$\Phi_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\,du$$
$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du, \quad x \in \mathbb{R}.$$

Donde:

μ es la media, la mediana, la moda y el valor esperado

σ es la desviación estándar de la población

σ^2 es la varianza

Φ representa la función de densidad de probabilidad

Note que en la función figuran los valores de μ (media) y σ (desviación estándar), esto quiere decir que la distribución varía en función de estos argumentos y, en última instancia, del proceso generador de datos. Cuando la media de una distribución es cero y la desviación estándar es igual a uno, decimos que tenemos una distribución gaussiana normal o estándar.

4.2) Funciones de distribución normal

Funciones de probabilidad:

Las funciones de probabilidad son funciones estadísticas (es decir, no son funciones programáticas) que describen la distribución de probabilidad de una variable aleatoria.

Tabla 1

Funciones de probabilidad

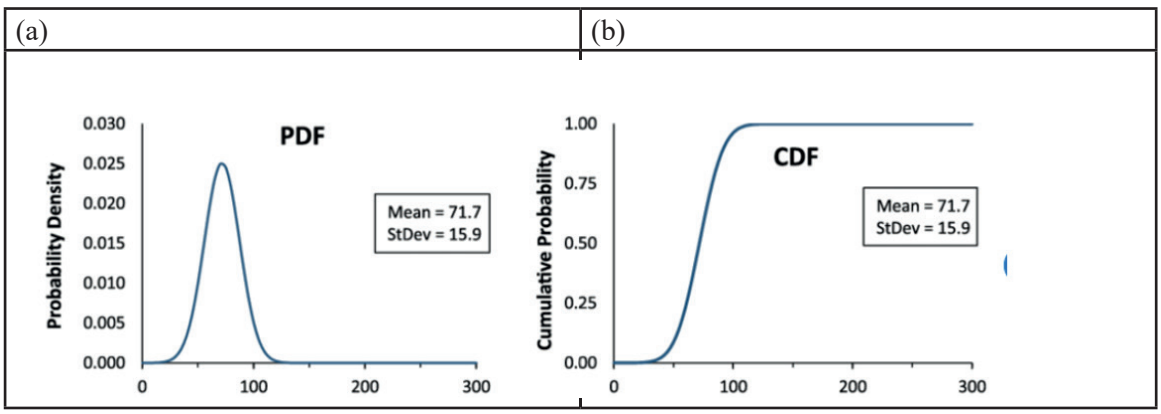
Variable aleatoria discreta X $\mathcal{E}$ { 1,2,3}	Variable aleatoria continua X $\mathcal{E}$ [1.10]
(PMF) La función de masa de probabilidad describe la probabilidad de valores individuales de la variable aleatoria y se describe con la fórmula: $p(x) = P(X = x)$	(PDF) La función de densidad de probabilidad describe la probabilidad de que una variable aleatoria se encuentre entre dos valores, ya que la probabilidad de un solo valor de variable continua prácticamente es 0 $F(x) = P(a \leq x \leq b)$
(CDF) La función de distribución acumulativa aplica tanto para variables discretas o continuas y se describe como: $F_X(x) = P(X \leq x)$ (CDF Inverso): Encontrar el valor debajo del cual obtenemos una distribución acumulativa dada .	

En distribución gaussiana las funciones PDF y CDF siguen la forma indicada en la siguiente figura

Figura 3

Función PDF (a) y CDF (b) de una distribución gaussiana

Creación de autor Alfonso Prado



Note que la función PDF nos indica que la máxima probabilidad se da en la media y que la suma de las probabilidades (o el área debajo de la curva) siempre será igual a 1 o 100%. Por otro lado, el CDF representa la probabilidad acumulada desde -∞ hasta un valor dado, por lo que en la medida que dicho valor dado incrementa, la probabilidad acumulada tiende a 1.

En base a lo anterior, podemos concluir que la probabilidad de un valor dado de la variable aleatoria va a depender de su distribución. Por ejemplo, la probabilidad de obtener el valor 100 en la figura 3a es relativamente baja; sin embargo, la probabilidad obtenida si la distribución estuviera centrada en 100 sería la más alta. Aquí entra el estadístico Z.

Estadístico Z

Eventualmente, necesitamos comparar dos distribuciones que tienen μ y σ diferentes para ver qué tan alejado está un valor x dado de la media. En estos casos, la comparación de las distribuciones se vuelve muy complicada, por lo que se estila usar el estadístico Z.

Z se denomina variable tipificada de X. Es una forma de ver a cualquier distribución en forma estandarizada en términos/múltiplos de su desviación estándar, en otras palabras, el valor Z es la cantidad de desviaciones estándar a la que está distanciada un valor x de su media.

Figura 4

Fórmula de cálculo del estadístico Z

$$Z = \frac{x - \mu}{\sigma}$$

Entonces, si tenemos una distribución gaussiana y tomando en cuenta el cálculo de las probabilidades de la figura 2, podemos asegurar que a Z= ±1 σ , tenemos una suma de probabilidades (o cobertura) del 68.26%, o, en otras palabras, dentro ±1 σ tenemos una probabilidad de encontrar el 68.26% de los valores.

A Z=±2σ tenemos una cobertura del área (o están incluidos) el 95.44% de las probabilidades.

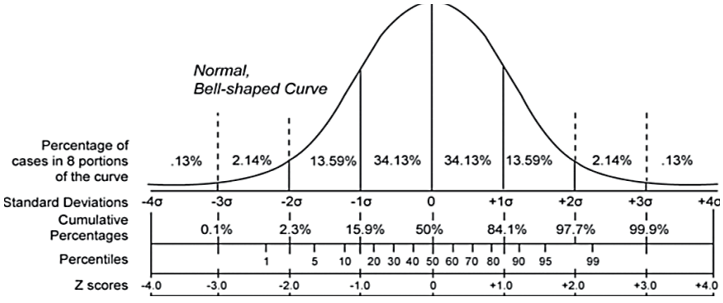
A Z=±3σ tenemos una cobertura del área (o están incluidos) el 99.72% de las probabilidades.

La figura 5 resume el valor Z, su equivalencia en sigmas y el porcentaje de cobertura de las probabilidades

Figura 5

Relación entre Z, σ y cobertura

Tomado de Webster (2017)



Sin embargo, de lo mencionado arriba, ningún problema de distribución se trata en términos de cobertura del 68.26% , 95.44% o 99.72% , más bien se utilizan los valores más estandarizados de Z que corresponden valores como se muestra en la tabla 3.

Tabla 3

Valor Z y cobertura de probabilidades

Cobertura	Valor Z
90%	±1.66
95%	±1.96
99%	±2.57

Funciones programáticas distribución gaussiana

Como habíamos indicado anteriormente, las funciones PDF y CDF son conceptos estadísticos, pero ahora veremos cuáles son las funciones programáticas que nos ayudan en el cálculo de probabilidades.

La siguiente tabla indica cómo calcular la función PDF, CDF y el inverso de CDF de una distribución gaussiana (valores por defecto mostrados).

Tabla 4

Funciones programáticas de la distribución

Creación de autor Alfonso Prado

Función de Distribución	Distrib. Normal	
PDF	dnorm()	dnorm(x, mean = 0, sd = 1, log = FALSE)
CDF	pnorm()	pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
Inverso CDF	qnorm()	qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

Función para obtener un set de datos de la distribución	Distrib. Normal	
	rnorm()	rnorm(n, mean = 0, sd = 1)

Función rnorm

Esta función sirve para generar un conjunto de datos que sigue una distribución dada por los argumentos mean (media) y sd (la desviación estándar). El argumento n representa la cantidad de elementos que se desea tenga la distribución.

Función dnorm

Esta función devuelve el valor de la función de densidad de probabilidad (PDF) de la distribución normal dada una determinada variable aleatoria x, una media poblacional μ y la desviación estándar poblacional σ.

Veamos un ejemplo.

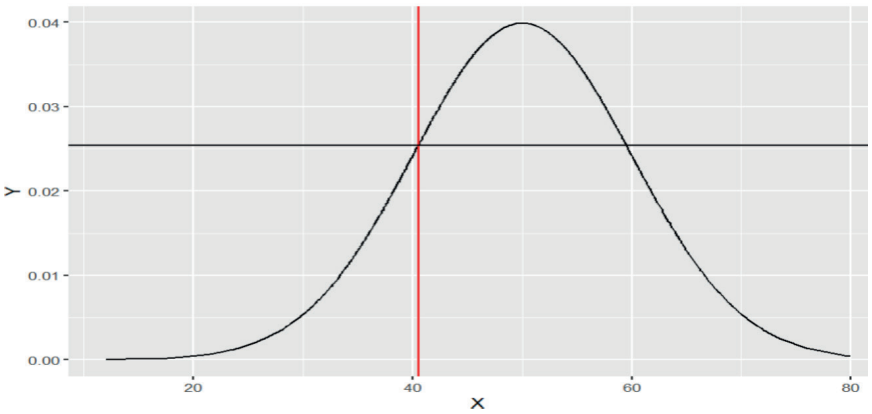
```
#Creamos una serie
serie4 <-rnorm(1000 , mean=50, sd=10)
#Obtenemos su densidad
df4 <- data.frame(X=serie4 , Y=dnorm(serie4, mean=50 ,sd=10))
#definimos un valor
x=40.5
#Obtenemos la proabilidad para x
prob_valor <- dnorm(x, mean=50, sd=10)
#Graficamos
g<- ggplot(data=df4 , aes(x=X, y=Y))+
  geom_line()+
  geom_vline(xintercept=x, color="red")+
geom_hline(yintercept=prob_valor)
g
```

Obtenemos la siguiente figura

Figura 6

Distribución y probabilidad puntual

Creación de autor Alfonso Prado



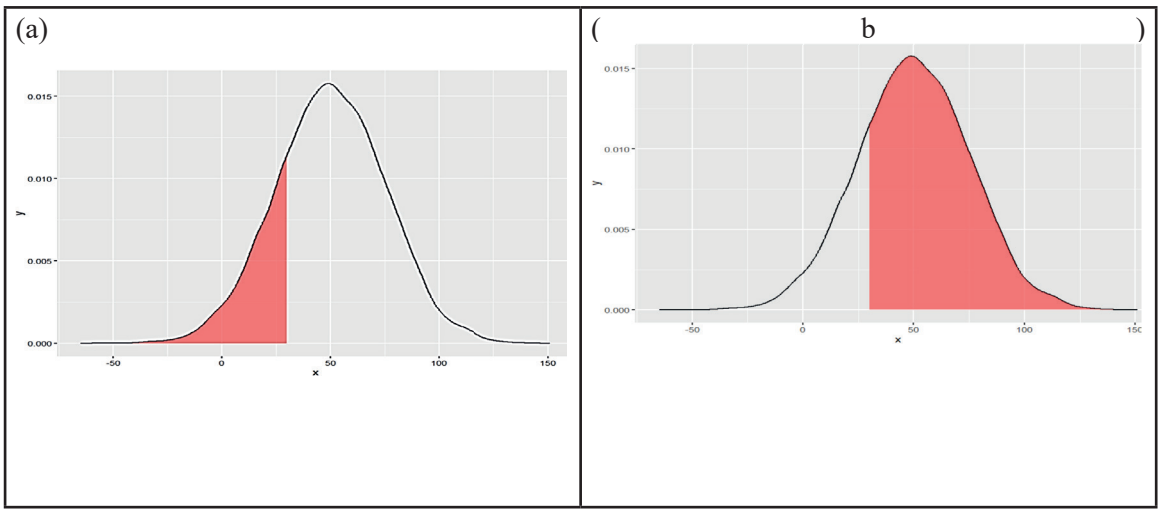
Función pnorm

Esta función calcula la probabilidad acumulada de una distribución dada por los argumentos de media y desviación mencionados anteriormente. Adicionalmente hay otro argumento lower.tail que es una variable lógica (TRUE/FALSE) y que indica el tipo de cálculo que se requiere. Por ejemplo, con lower.tail=TRUE (que es el valor por defecto) calculará la probabilidad acumulada desde -∞hasta el valor dado por el argumento q, pero con lower.tail=TRUE calculará la misma probabilidad desde valores mayores a q hasta el +∞. El siguiente gráfico explica el concepto. Es importante notar que mientras el parámetro lower.tail =TRUE calcula la probabilidad acumulada para un valor $x \leq q$ cuando usamos lower.tail=FALSE.

Figura 7

Probabilidad acumulada en rojo con lower.tail=TRUE (a) y lower.tail=FALSE (b)

Creación de autor Alfonso Prado



Función qnorm

Comúnmente se nos presenta el problema inverso al anterior, es decir, queremos obtener el valor debajo del cual se acumula una cierta probabilidad. Para esto, utilizamos la función qnorm. Sus argumentos son similares a los de pnorm, excepto que ahora el argumento q representa la probabilidad acumulada que se desea obtener y la función retorna el valor debajo o encima del cual se obtiene dicha probabilidad.

Figura 8

Código para validar el inverso de CDF

```
#Obtenemos cual es la probabilidad acumulada del valor -1, para una media 0
#y desviación 1
pnorm(-1, mean=0,  sd=1)
Resultado 0.1586
#Ahora preguntamos cual es el valor debajo del cual obtenemos el 15.86%
qnorm(.1586)
Resultado: -1
#En otras palabras si tengo una probabilidad acumulada del 15.86%
#cual es el valor de x que nos dará este CDF
```

Distorsiones a la curva de distribución

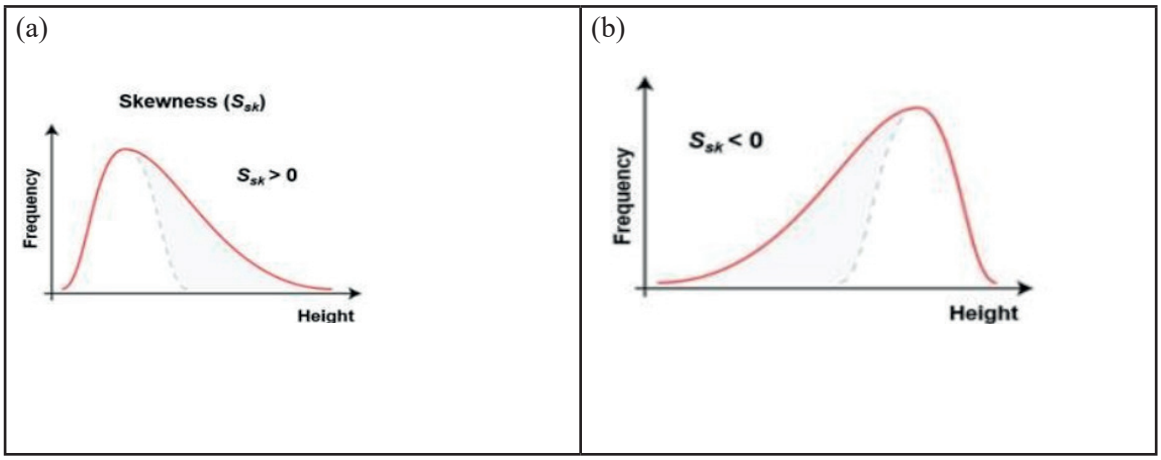
En la práctica la curva de densidad no es siempre tan perfecta, todo depende del proceso generador de la data. Hay dos fenómenos que se presentan:

Skewness: Es una medida de la simetría de la curva. Esta puede estar recostada sobre uno de sus lados. Podemos medirla con la función skewness(), que nos devuelve un indicador mayor que cero si la distribución tiene una cola derecha más amplia, y menor que cero si la cola izquierda es más amplia.

Figura 9

Skewness positivo (a) y negativo (b) de una distribución gaussiana

Creación de autor Alfonso Prado



COMO INTERPRETAR ESTOS VALORES

Asimetría baja (casi cero):

Los valores de asimetría cercanos a cero, generalmente dentro del rango de -0,5 a 0,5, se consideran indicadores de una asimetría baja o insignificante. En un conjunto de datos con una asimetría baja, la distribución es relativamente simétrica, con una distribución aproximadamente equilibrada de valores en ambos lados de la media.

Asimetría moderada (entre -0,5 y -1 o 0,5 y 1):

Los valores de asimetría entre -0,5 y -1 (asimetría negativa) o entre 0,5 y 1 (asimetría positiva) se consideran indicadores de una asimetría moderada. En este caso, la distribución está algo sesgada, pero la asimetría no es extrema. Hay una asimetría notable en la distribución, pero aún puede ser razonablemente interpretable.

Asimetría alta (superior a -1 o 1)

Los valores de asimetría que están significativamente por debajo de -1 (asimetría negativa fuerte) o significativamente por encima de 1 (asimetría positiva fuerte) indican un alto grado de asimetría. En los conjuntos de datos con alta asimetría, la distribución es altamente asimétrica, con una cola larga en un lado. La asimetría extrema puede dificultar el análisis de los datos y puede requerir una consideración especial en los análisis estadísticos.

Curtosis:

La curtosis es una medida de la asimetría de la curva, es decir, que tan pronunciadas son las colas en relación con una distribución normal. Podemos medir la curtosis mediante la función kurtosis().

COMO INTERPRETAR LOS VALORES

Curtosis alta (leptocúrtica):

Un valor alto de curtosis indica una distribución con un pico más pronunciado y agudo que una distribución normal. Implica que los datos están más concentrados alrededor de la media. Esto puede dar como resultado una distribución que parece más puntiaguda y menos dispersa.

Curtosis baja (platicúrtica):

Un valor bajo de curtosis indica una distribución con un pico más plano en comparación con una distribución normal. Sugiere que los datos están más dispersos. Esto puede dar como resultado una distribución que parece más plana y menos puntiaguda que una distribución normal.

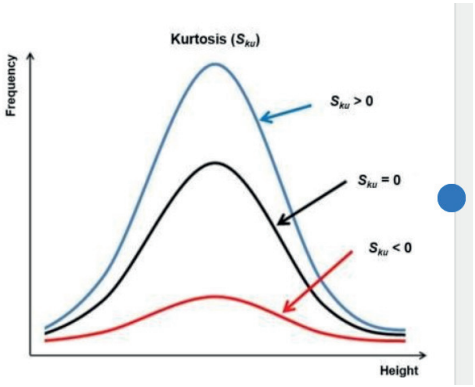
Mesocúrtica (curtosis cero):

Un valor de curtosis de 0 (o cercano a 0) sugiere una distribución que está muy cerca de una distribución normal (en forma de campana). Los datos tienen colas y un pico similares a los de una distribución normal. Esto a veces se denomina “mesocúrtica” e indica que los datos no presentan valores atípicos excesivos ni un comportamiento extremo en términos de sus colas.

Figura 10

Distorsión de la distribución por curtosis

Creación de autor Alfonso Prado



Graficación de distribución normal

Con frecuencia nos encontramos con situaciones en las que nos gustaría entender cómo se distribuye una variable en un conjunto de datos.

Existen algunas variantes de estos diagramas que se diferencian en la cantidad de información que presentan:

- ✓ Histogramas
- ✓ Polígonos de frecuencia
- ✓ Diagramas de densidad

HISTOGRAMAS Y POLIGONOS DE FRECUENCIA

Los histogramas son nada más que diagramas de barras que usan la función stat_count para variables discretas y stat_bin para variable continuas.

Geom_histogram

Muestra información sobre el uso de esta función

Enlace

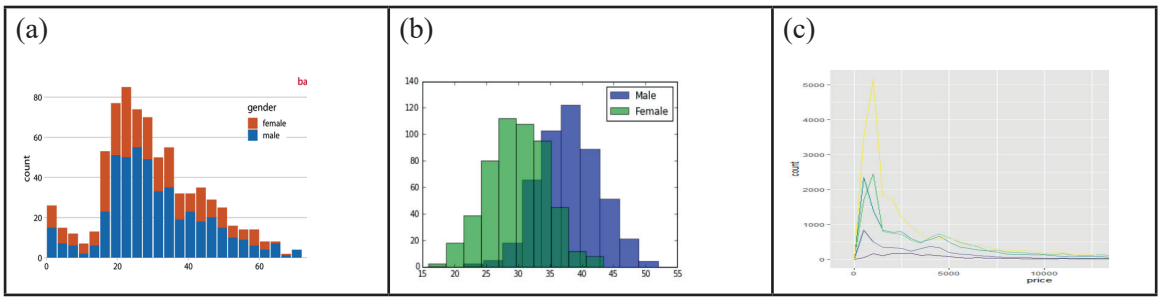
Debido a que los histogramas se generan al agrupar los datos, su apariencia visual exacta depende de la parametrización de la cantidad de columnas (bin) y del ancho del intervalo (binwidth). Estos argumentos tienen el mismo significado que el diagrama geom_bin_2d. Adicionalmente, pueden estar complementados por otros mapeos para obtener diversos histogramas para diferentes niveles de una variable categórica. Estos son conocidos como histogramas apilados. Funcionan bien para una pequeña cantidad de niveles, pero cuando estos niveles aumentan, el diagrama se vuelve difícil de leer. Una opción que puede ayudar en estos casos es usar transparencia (alpha), aunque puede ser confuso ya que aparecen nuevos colores no incluidos en la leyenda; en tal caso, la opción de polígonos de frecuencia podría ser más adecuada.

Los polígonos de frecuencia geom_freqpoly() igual muestran la frecuencia, pero con líneas que unen los puntos máximos de cada barra de conteo, por lo que son mucho menos “congestionados”

Figura11

Histogramas apilado (a) , con transparencia(b) y polígonos de frecuencia (c)

Creación de autor Alfonso Prado



Gráficos de densidades

La densidad representa el valor de la probabilidad para cada valor de una variable continua. Los diagramas de densidad muestran estos valores a través de una línea y son ampliamente usados para demostrar si un conjunto sigue una cierta distribución.

Función geom_density

Muestra el uso de esta función

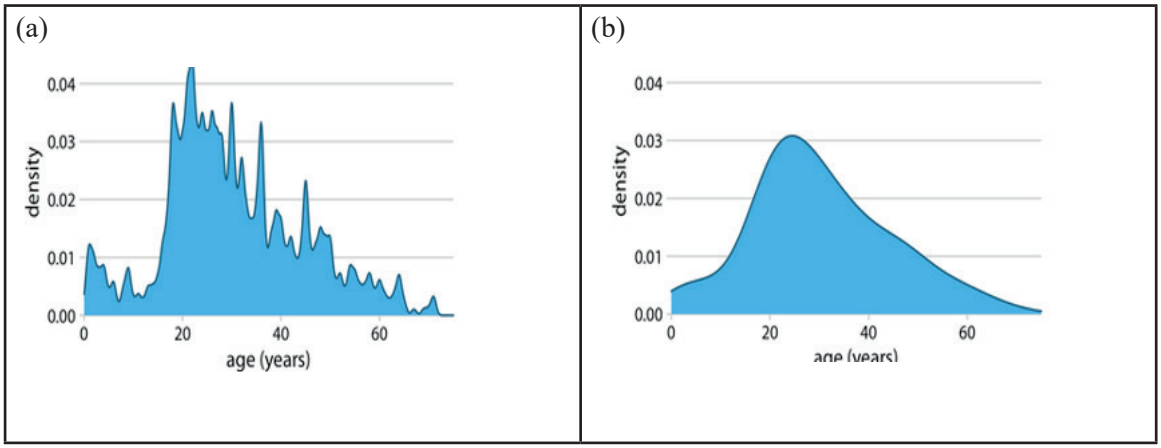
Enlace

Esta función tiene un parámetro de configuración conocido como “adjust” que impone un cierto nivel de suavizado. En general nos interesa ver la forma general de la curva mas no detalles, entonces “adjust” lo que hace es recortar los picos parciales que pueden aparecer. Si, por otro lado, nos interesa ver el detalle, bajaremos el valor de este parámetro.

Figura 12

Densidad de una variable (a) y efecto del parámetro adjust (b)

Creación de autor Alfonso Prado

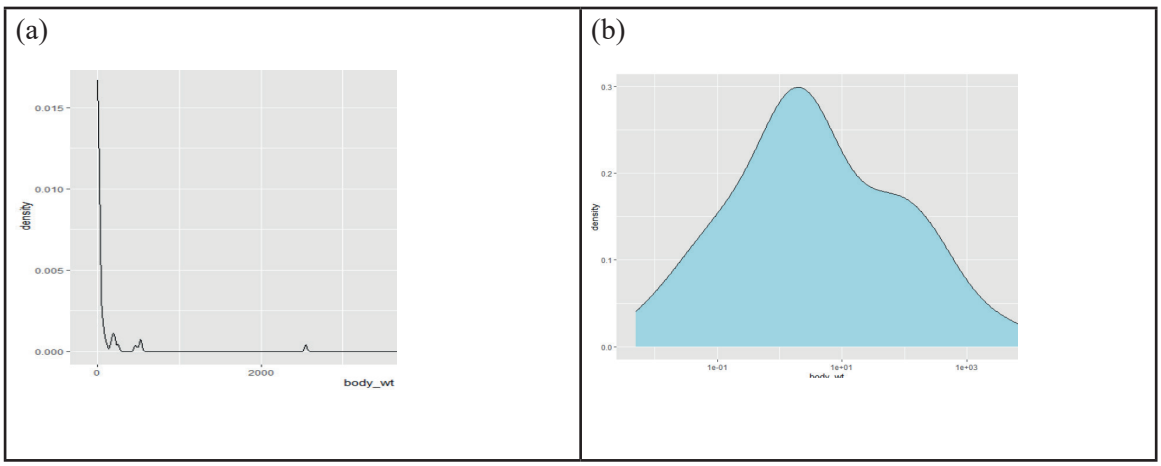


En casos en que los datos presentan alto skewness o curtosis el gráfico resultante podría ser poco informativo, en esos casos se puede utilizar escalas que permiten ver la forma de la curva, por ejemplo, scale_x_continuous(trans=”log10”)

Figura 13

Densidad poco informativa con skewness (a) con transformación log10 (b)

Creación de autor Alfonso Prado



También es posible que queramos analizar la densidad de una variable como se distribuye por otra variable categórica. Estos diagramas son conocidos como densidades apiladas

Gráficos para mostrar relación cuantil-cuantil (QQ)

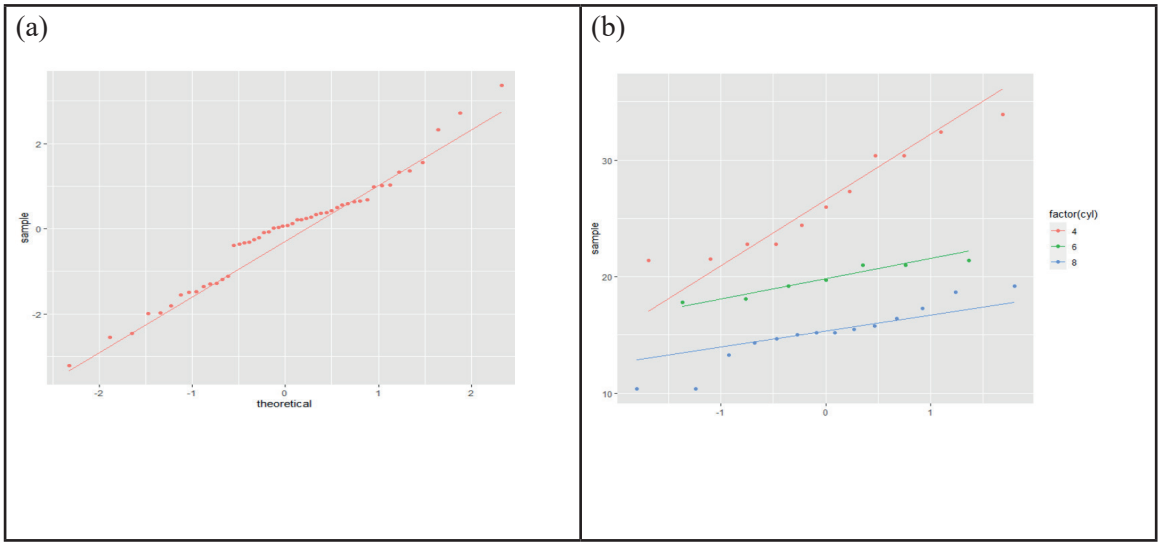
Hay ocasiones en las que no nos interesa ver exactamente la forma de la curva de densidad, sino únicamente comprobar si la misma se ajusta a una distribución normal o queremos comparar dos distribuciones para ver si son semejantes. En estos casos, los diagramas QQ son los apropiados.

Estos diagramas contienen dos elementos: por un lado, la distribución de la variable a analizar y una línea que indica cuál debería ser la distribución si esta fuera normal. La primera parte se obtiene mediante el geom_qq() y la segunda con el geom_qq_line().

Figura 14

Diagramas cuantil-cuantil de una distribución no normal (a) y distribución por varios factores (b)

Creación de autor Alfonso Prado



Definición de los términos citados en la Clase

DISTRIBUCIÓN NORMAL	Llamada también como distribución de Gauss, es la distribución de probabilidad más utilizada en estadística y teoría de probabilidad. Esto se debe a dos razones: - Su función de densidad es simétrica y con forma de campana lo que favorece su aplicación como modelo a gran número de variables. - Es además límite de otras distribuciones y aparece relacionada con resultados ligados a la teoría de las probabilidades gracias a sus propiedades matemáticas.
ASIMETRÍA	Es la falta de simetría entre los datos de una distribución. El concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a la izquierda y derecha de un valor central (media aritmética).

Referencias Citadas

Economipedia, Población estadística: Qué es, tipos y ejemplos, recuperado de https://economipedia.com/definiciones

Webster A., (2000), Estadística aplicada a los negocios, Irwin Professional Publishing

Wasserman, L. (2010). All of Statistics: A Concise Course in Statistical Inference. Springer Publishing Company.

Otzen T. Manterola c. (2017) ,Técnicas de Muestreo sobre una Población a Estudio, Int. J. Morphol. vol.35 no.1 Temuco, recuperado de http://dx.doi.org/10.4067/S0717-95022017000100037

Conceptos de estadística

Alvarez H., Conceptos de estadística, Universidad Santiago de Compostela, Recuperado de https://www.usc.gal/genp/docencia/ConceptosDeEstadistica.pdf

Profundización Clase 4.

Distribución Normal 1

Muestra la creación y visualización de una distribución normal

Enlace

Distribución normal 2

Muestra el uso de la función pnorm
Enlace

Distribución normal 3

Muestra el uso de la función qnorm
Enlace

PUCE Virtual

La excelencia no se improvisa

síguenos