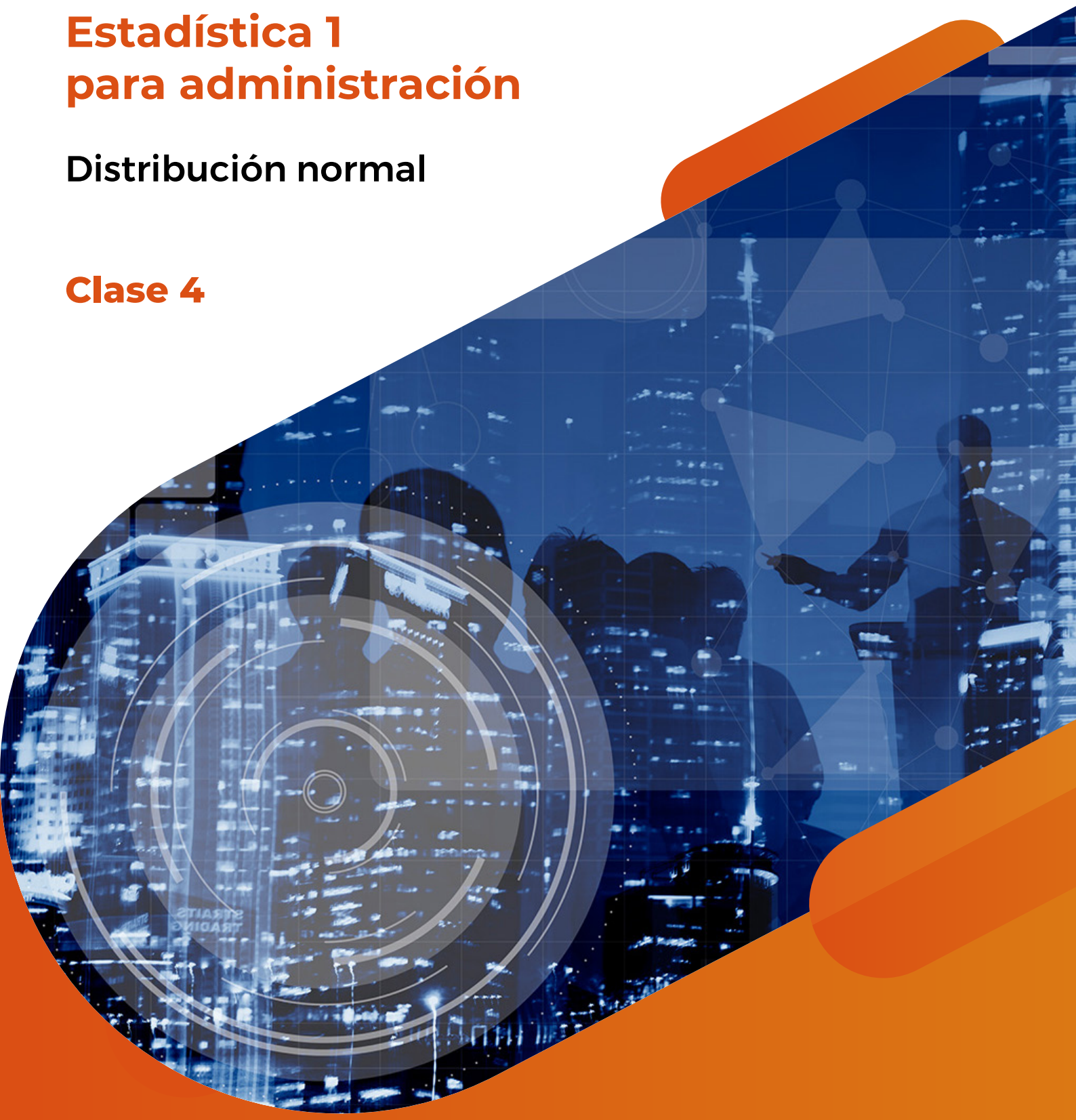
PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Distribución normal

Clase 4



La excelencia no se improvisa

INTRODUCCIÓN DE LA CLASE 4

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. Con este concepto en mente nuestra clase empieza con una presentación general de que es probabilidad y como calcularla en distintos escenarios como probabilidad condicional, probabilidad relativa a priori y posteriori.

Luego pasaremos a presentar el concepto de distribución de probabilidades en función de los diversos tipos de variables discutidos en la clases anteriores.

En el área de conocimiento de la estadística existen una gran cantidad de distribuciones, algunas representan distribuciones que se presentan en problemas de vida real, otras son más bien "curiosidades" matemáticas que nos ayudan en la resolución de problemas, de todas estas, las distribuciones gaussianas y Student T son las más comunes para variables continuas, así como la distribución Binomial y Poisson para variables discretas, que son en las que centraremos nuestro estudio.

RDA 1: Analizar información de un conjunto de datos, mediante tablas, gráficos y medidas descriptivas.

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 las probabilidades obtenidas, veremos que van delineando la forma de una campana de Gauss por eso le llamamos una distribución gaussiana.

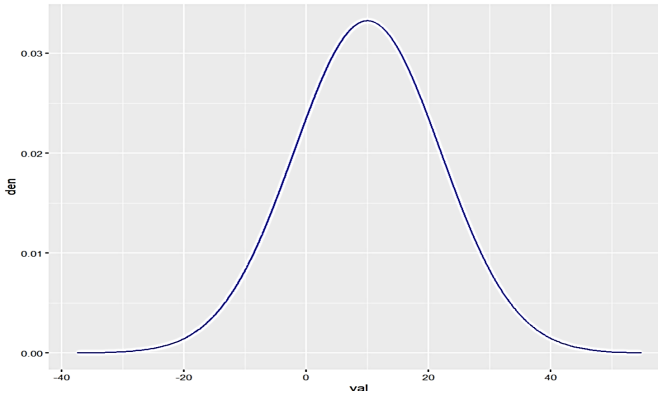
Esta distribución tiene las siguientes características

- ✓ 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 de genera esta curva.

Figura 2

Función CDF para la distribución gaussiana Prado A. (2025)

Adaptación de la fórmula tomada de Aula360 (s.f.) Apuntes Unidad 3 Distribución Normal

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

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.

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

Creación del autor Alfonso Prado

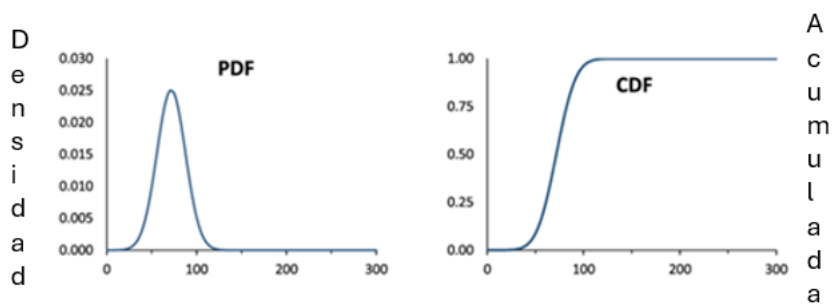
Variable aleatoria discreta $X \in \{1,2,3\}$	Variable aleatoria continua $X \in [1..10]$
(PMF) La función de masa de probabilidad describe la probabilidad de valores individuales de la variable aleatoria y se describe con la fórmula: $p(x) = P(X = x)$	(PDF) La función de densidad de probabilidad describe 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

(a)	(b)
	

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 bajo, sin embargo, la probabilidad obtenida si la distribución estuviera centrada en 100 sería la más7 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 complicado, por lo que se estila usar el estadístico Z.

El estadístico 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 Prado A. (2025)

Adaptación de la fórmula para la resolución de medias de Webster (2000) capitulo 4

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

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

A $Z = \pm 2\sigma$ tenemos una cobertura del área (o están incluidos) el 95.44% de las probabilidades.

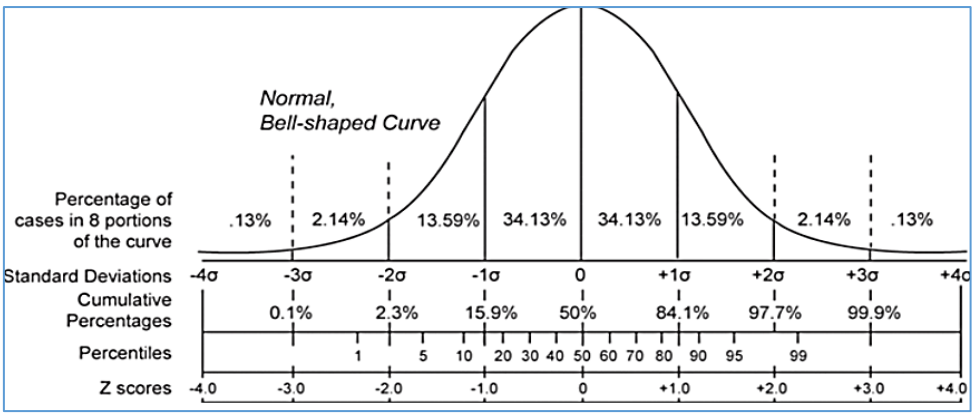
A $Z = \pm 3\sigma$ tenemos una cobertura del área (o están incluidos) el 99.72% de las probabilidades.

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 Prado A. (2025)

Adaptación de la fórmula para la resolución de medias de Webster (2000) capitulo 4



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 2.

Tabla 2

Valor Z y cobertura de probabilidades

Creación del autor Alfonso Prado

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

Funciones programáticas distribución gaussiana

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

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

Tabla 3

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 retorne un vector con la distribución dada.

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.

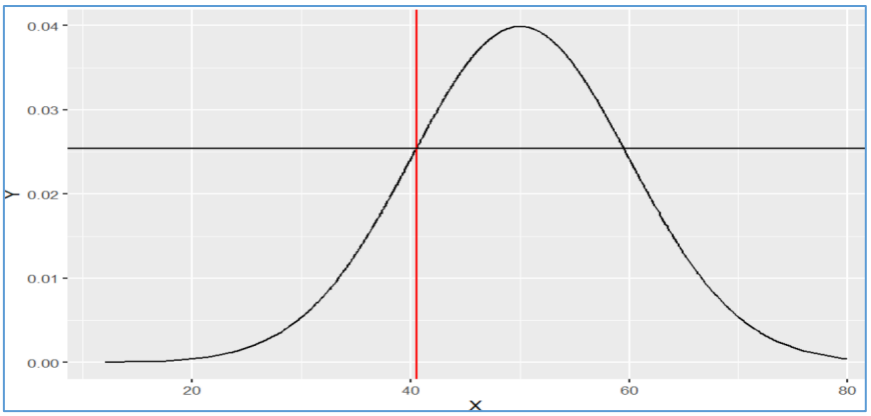
```
#Creamos una serie
serie4 <-rnorm(1000 , mean=50, sd=10)
#Obtenemos su densidad
df4 <- data.frame(X=serie4 , Y=dnorm(serie4, mean=50 ,sd=10))
#definimos un valor
x=40.5
#Obtenemos la proabilidad para x
prob_valor <- dnorm(x, mean=50, sd=10)
#Graficamos
g<- ggplot(data=df4 , aes(x=X, y=Y))+
  geom_line()+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



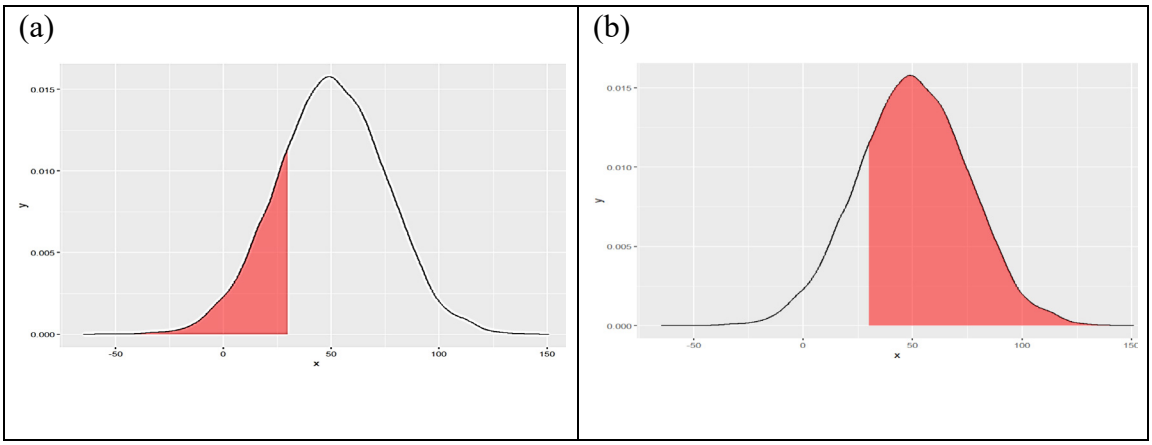
Función pnorm

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 $-\infty$ hasta el valor dado por el argumento q, pero con lower.tail=TRUE calculará la misma probabilidad desde valores mayores a q hasta el $+\infty$. 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 utilizaremos la función qnorm. Sus argumentos son similares al de pnorm excepto que ahora el argumento q representa la probabilidad acumulada que se desea obtener y la función retorna el valor debajo/encima del cual se obtiene dicha probabilidad. El siguiente pedazo de código explica el concepto.

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
```

Distribución normal

Funciones de la distribución

Medidas posición

En diversas ocasiones los investigadores pueden estar interesados en el valor que marca el 10% superior de una muestra, o por ejemplo debajo de qué valor se ubica el 35 % de los valores de una variable, y otra situación entre qué números limitan el 50% central de los datos de una distribución. Los valores que denotan estos cortes porcentuales son conocidos como cuantiles y dependiendo del número de cortes que se hagan sobre la muestra dichos cuantiles también se llamaran cuartiles (4 secciones), deciles (10 secciones) o percentiles (100 secciones). Todos estos cuantiles son explicados por una

misma fórmula, la de los percentiles, que se explica a continuación para los datos agrupados y no agrupados.

Algunos percentiles reciben nombres especiales debido a que corresponden con valores cabalísticos o de bastante recurrencia en el ámbito estadístico. Por ejemplo, los percentiles de 25, 50 y 75 son conocidos como cuartiles en orden a que dividen la muestra en cuatro partes iguales y los cortes para esta división son hechos en los datos que ocupan dichas posiciones porcentuales.

Veamos un ejemplo concreto

En el siguiente conjunto de datos se proporcionan los ingresos de pacientes a un hospital por causa de una determinada dolencia durante 50 días.
14,18, 14, 16, 18, 16, 17, 17, 17, 18, 10,19, 17, 16, 10,18, 15, 19, 16, 13, 17, 16, 14, 17, 16, 19, 17, 14, 17,16, 18, 18, 19, 11, 18, 17, 10, 18, 15, 17, 17, 16, 15, 10, 18, 19, 17, 15, 16, 15.
Calcular los percentiles P_{75} y P_5 usando la tabla de frecuencias.
En primer lugar debemos ordenar el dataset en forma ascendente

10,10,10,10, 11, 13, 14, 14, 14, 14, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 17, 17, 17, 17, 17, 17, 17, 18, 18, 18, 18, 18, 18, 18, 18, 18, 19, 19, 19, 19, 19
Como se busca el percentil k 75 y dado que n 50, se calcula a =75 (50)/100 37,5 aplicando la fórmula de los percentiles, se tiene:

$$\begin{aligned} P_{75} &= x_{[\lfloor 37{,}5 \rfloor]} + (37{,}5 - \lfloor 37{,}5 \rfloor)\left(x_{[\lfloor 37{,}5+1 \rfloor]} - x_{[\lfloor 37{,}5 \rfloor]}\right) \\ &= x_{[37]} + (37{,}5 - 37)\left(x_{[38]} - x_{[37]}\right) \\ &= 18 + (0{,}5)(18 - 18) \\ &= 18 \end{aligned}$$

A partir de lo anterior, se puede interpretar que el 75 % de los datos, es decir, el 75 % de los días del mes, se presentaron menos de 18 ingresos de pacientes al hospital.

Representación gráfica mediante diagrama de caja

Con frecuencia nos encontramos con situaciones en las que nos gustaría ver gráficamente cómo se distribuye una variable en particular en un conjunto de datos. Existen varias formas de diagramas distribuciones:

- ✓ Diagramas de caja

- ✓ Diagramas de violin

Diagramas de caja

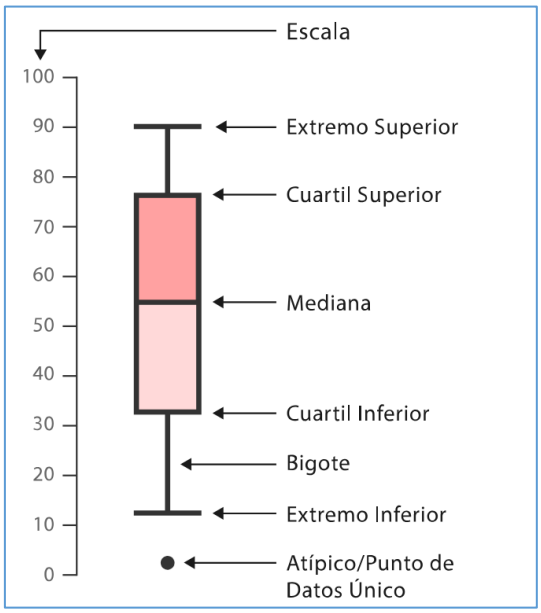
Un diagrama de caja o "Boxplot" es una representación gráfica que incluye el rango, los cuartiles y la media, haciendo posible visualizar cómo están distribuidos los datos y además permitiendo ver la dispersión de estos.

A pesar de estar registrados varios tipos de estilos, en este texto manejaremos una de las más comunes, por ejemplo, en algunas versiones se estila incluir líneas de valores máximo y mínimo conocidas como los "bigotes" (extremo superior e inferior en la figura) del diagrama de caja, así como la presencia o ausencia de datos atípicos

Figura 7

Diagrama general de caja

Fuente: https://datavizcatalogue.com/



Veamos un ejemplo

Del dataset midwest nos interesa saber cómo se distribuye la densidad poblacional de todos los estados y de cada uno de ellos por separado.

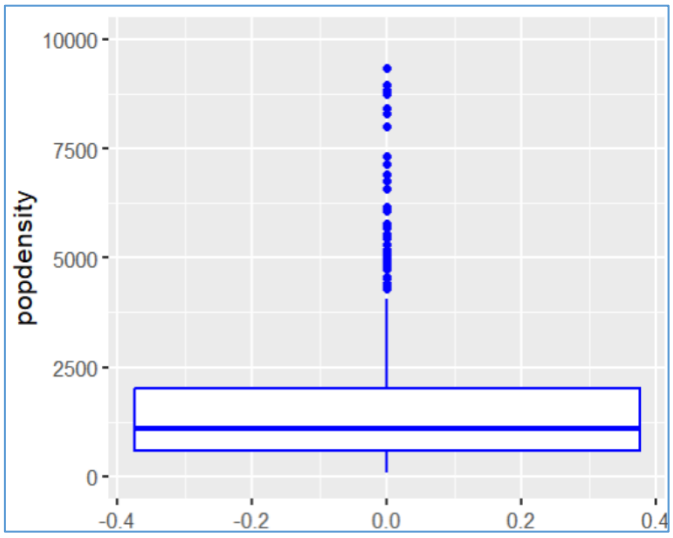
```
ggplot(data=midwest, aes(y=popdensity)) +
```

```
  geom_boxplot() +
  scale_fill_discrete()+ylim(0,10000)
```

Figura 8

Diagrama de caja densidad poblacional dataset midwest

Creación del autor Alfonso Prado



Vemos que la densidad está corrida hacia abajo, con múltiples valores atípicos. Investigando un poco más profundamente podemos ver los diagramas de caja por estado

```
ggplot(data=midwest, aes(y=popdensity, x=state , group=factor(state), fill=factor(state))) +
  geom_boxplot() +
  scale_fill_discrete()+ylim(0,10000)
```

Figura 9

Diagrama de caja densidad poblacional dataset midwest por estado

Creación del autor Alfonso Prado

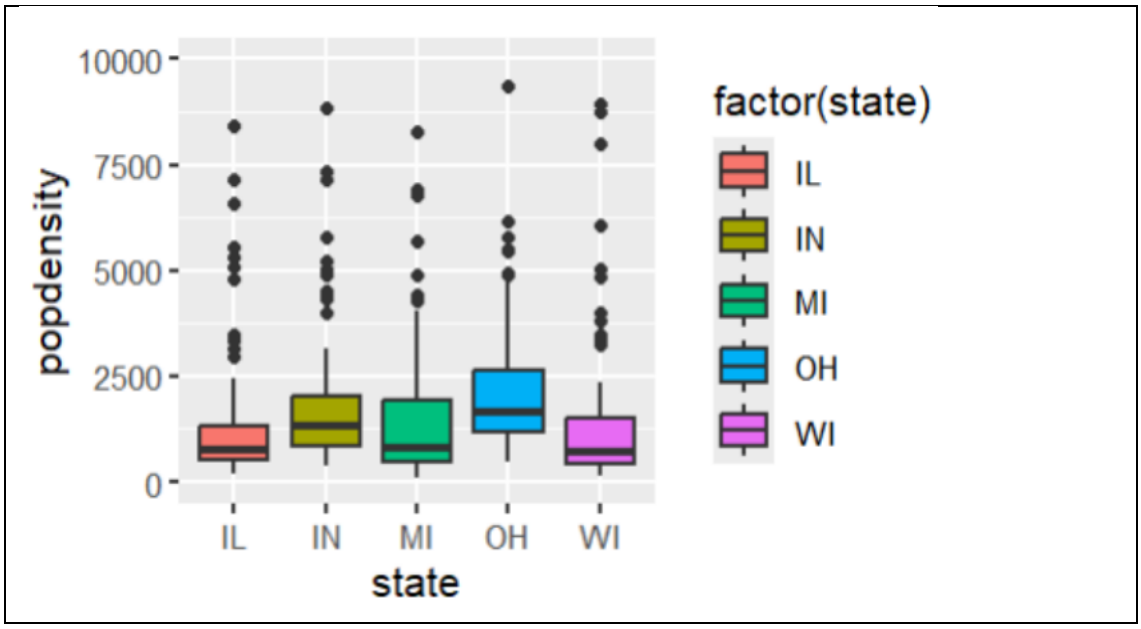


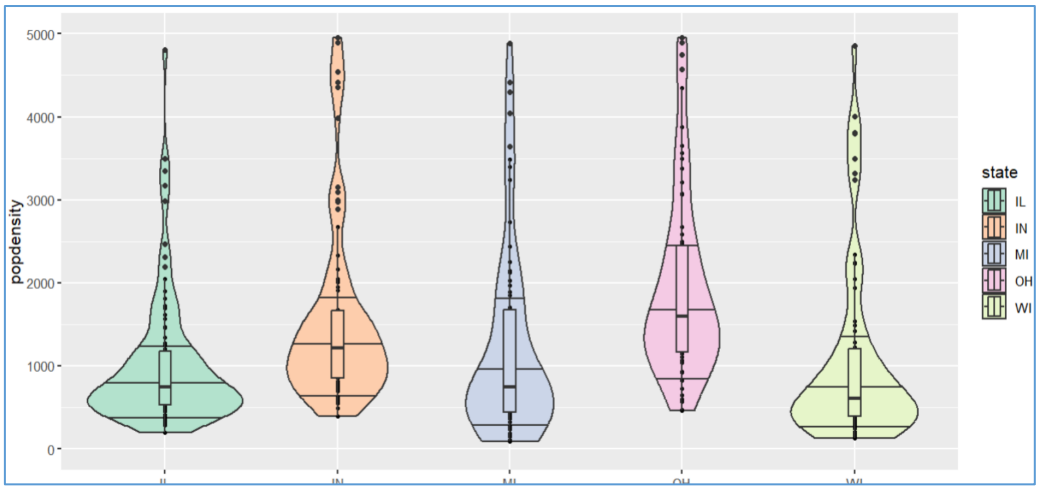
Diagrama de Violín

Los diagramas violín son igual que los boxplots pero incluyen una distribución al costado, es decir no solo interesa saber cual es el valor de 1Q, media , 3Q, etc. sino que dentro del Q1 queremos saber cuál es la distribución, tiene un sesgo (Asimetría y Curtosis)

Toda esta información está disponible en el diagrama de violín, Veamos un ejemplo de la densidad poblacional del dataset midwest por estado, podemos observar que ningún estado tiene una distribución gausiana perfecta, indicando que todos los estados tienen áreas donde la densidad poblacional es superior al resto del área. Además, podemos notar que el diagrama de caja ha sido configurado para mostrar los siguientes percentiles: draw_quantiles=c(.1,.5, .75).

Diagramas de caja y violín combinados
Presenta configuración

Figura 10
Diagrama de caja y violín del dataset midwest por estado
Creación del autor Alfonso Prado



REFERENCIAS CITADAS

Webster. A. , , (2017), Estadística aplicada a los negocios y economía, Irwin- McGraw-HILL

The Data Visualisation Catalogue, : https://datavizcatalogue.com/

Aula360, (s.f.), Apuntes Unidad 3 Distribución Normal, https://static.sumaysigue.uchile.cl

PUCE Virtual

La excelencia no se improvisa

síguenos