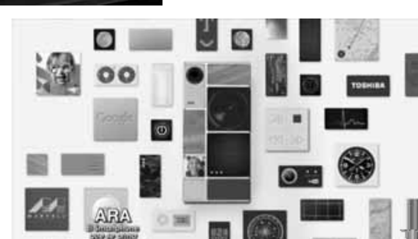
UNIDAD 1

INTRODUCCIÓN AL DISEÑO EXPERIMENTAL

TEMA 2: EXPERIMENTOS COMPARATIVOS SIMPLES: MEDIDAS DESCRIPTIVAS Y DISTRUBUCIONES DE PROBABILIDAD CONTINUA

Experimentos para comparar dos condiciones



Juan Pablo Sucre Reyes

Experimentos para comparar dos condiciones

• Incluye conceptos básicos de la estadística: *variables aleatorias, distribuciones de probabilidad, muestras aleatorias, distribuciones de muestreo y pruebas de hipótesis*.

Juan Pablo Sucre Reyes

1. Introducción

• Considera diferentes formulaciones = tratamientos = niveles del factor formulaciones.

• *Ejemplo: La fuerza de la tensión de adhesión del mortero de cemento portland es una de sus características clave. Se desea compara la fuerza de una formulación modificada (látex de polímeros) con la fuerza del mortero sin modificar. Se reúnen 10 observaciones de cada formulación.*

Tabla 2-1 Datos de la fuerza de la tensión de adhesión del experimento de la formulación del cemento portland

j	Mortero modificado y_{1j}	Mortero sin modificar y_{2j}
1	16.85	17.50
2	16.40	17.63
3	17.21	18.25
4	16.35	18.00
5	16.52	17.86
6	17.04	17.75
7	16.96	18.22
8	17.15	17.90
9	16.59	17.96
10	16.57	18.15

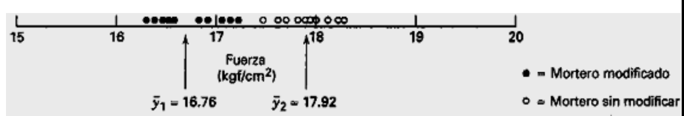


Figura 2-1 Diagrama de puntos de los datos de la fuerza de la tensión de adhesión de la tabl

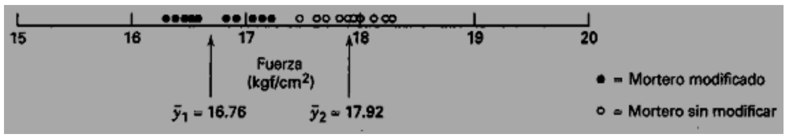
• *Conclusiones: Del examen visual (diagrama de puntos) se puede decir que la mezcla sin modificar es más fuerte, lo que se confirma con la comparación de las medias (diferencia no trivial). Pero esta diferencia no es suficiente para implicar que las formulaciones son diferentes (fluctuaciones del muestreo)*

• *Una prueba de hipótesis comparará las formulaciones en términos objetivos.*

Juan Pablo Sucre Reyes

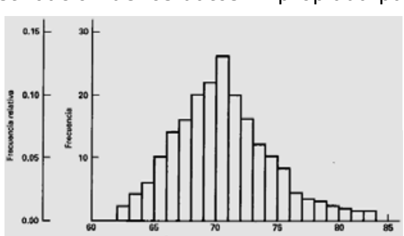
1.1 Conceptos estadísticos básicos

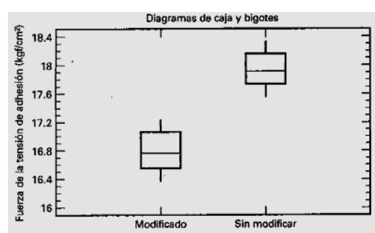
- *Corrida:* c/u observaciones del experimento, que difieren generando fluctuaciones o ruido en los resultados (error).
- *Error experimental (estadístico):* originado por la variación sin control (inevitable).
- *Variable aleatoria:* variable de respuesta que tiene error, pudiendo ser *discreta* (conjunto de valores posibles finito) o *continua* (conjunto de valores posibles intervalo)
- *Descripción gráfica de variabilidad:* métodos para analizar datos experimentales
- *a) Diagrama de puntos:* visualiza la localización o tendencia central de las observaciones y su dispersión. Apropiada para n ≤ 20 observaciones.



1.1 Conceptos estadísticos básicos

- *Descripción gráfica de variabilidad:* métodos para analizar datos experimentales
- *b) Histograma:* muestra la tendencia central, la dispersión y la forma general de la distribución de los datos. Apropiada para n > 20 observaciones.





- *c) Diagrama de caja:* muestra el mínimo, el máximo, los cuartiles inferior (percentil 25) y superior (percentil 75) y la mediana (percentil 50). La longitud de la caja expresa la variabilidad o dispersión.
- Estos métodos gráficos son útiles para resumir la información de una muestra de datos. Para describirlos con mayor detalle se usan *distribuciones de probabilidad*.

2. Resumen de datos cuantitativos

• *Distribución de frecuencia:* resumen tabular (datos) que muestra el N° (frecuencia) de elementos en cada una de varias clases (a definir) que no se superponen.

•*Ejemplo: Se anoto el tiempo [días] para completar las auditorías de final de año para una muestra de 20 clientes de Sanderson and Clifford (firma de contadores públicos). Los tres pasos necesarios para definir las clases son:*

Duración de la auditoría de fin de año (en días)

12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

•*1. Determine el N° de clases:* se forman especificando los rangos que se usarán para agrupar los datos (regla general: utilizar entre 5 y 20 clases).

• *N° de elementos de datos pequeño (n=20), se tendrá una distribución con 5 clases.*

•*2. Defina el ancho de cada clase:* Regla general: el mismo para todas las clases. Un N° grande de clases significa un ancho de clase menor, y viceversa. Así:

$$\text{Ancho de clase aproximado} = \frac{\text{valor de datos mayor} - \text{valor de datos menor}}{\text{número de clases}}$$

• *Dado que se decidió usar 5 clases, usando la ecuación se tiene (33-12)/5 = 4,2 que se redondeará para usar un ancho de clase = 5 días.*

2.1 Distribución de frecuencia

•*3. Determine los límites de clase:* Deben elegirse de modo que cada elemento de datos pertenezca a una y sólo una de las clases. El *límite de clase inferior* identifica el valor de datos menor asignado a la clase; y el *límite de clase superior* al valor mayor.

Distribución de frecuencia para los datos de duración de la auditoría

Duración de la auditoría (días)	Frecuencia
$\boxed{10–14}$	4
15–19	8
20–24	5
25–29	2
30–34	1
Total	20

• *Determinados el número, ancho y límites de clase se obtiene una distribución de frecuencia con el conteo del número de valores de datos que pertenecen a cada clase.*

• *Algunas conclusiones:*

• *Las duraciones de las auditorías que ocurren con más frecuencia están en la clase 15–19 días (8 de las 20).*

• *Sólo una auditoría requirió 30 o más días.*

•*4. Punto medio de clase:* valor medio entre los límites de clase inferior y superior.

• *En el caso de los datos de duración de la auditoría, los puntos medios de las cinco clases son 12, 17, 22, 27 y 32.*

2.2 Distribuciones de frecuencia relativa y frecuencia porcentural

• *Frecuencia relativa:* proporción de las observaciones que pertenecen a una clase. Si se tienen *n* observaciones:

$$\text{Frecuencia relativa de la clase} = \frac{\text{frecuencia de la clase}}{n}$$

• *Frecuencia porcentual:* de una clase es la frecuencia relativa multiplicada por 100.

Distribuciones de frecuencia relativa y frecuencia porcentual

Duración de la auditoría (días)		Frecuencia relativa	Frecuencia porcentual
10–14		0.20	20
15–19		0.40	40
20–24		0.25	25
25–29		0.10	10
30–34		0.05	5
	Total	1.00	100

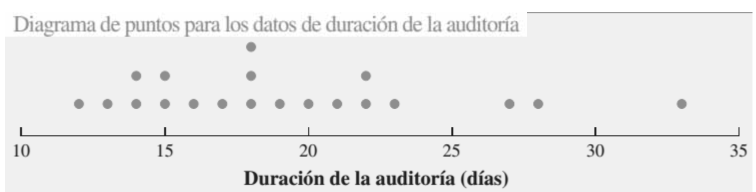
• *Algunas conclusiones: 0.40 de las auditorías, o 40%, requirió de 15 a 19 días, y sólo 0.05, o 5%, requirió 30 o más días.*

Juan Pablo Sucre Reyes

2.3 Diagrama de puntos

• El eje horizontal muestra el rango de los datos. Cada valor se representa por medio de un punto colocado sobre este eje.

• Muestran los detalles de los datos y son útiles para comparar la distribución de los datos de dos o más variables.

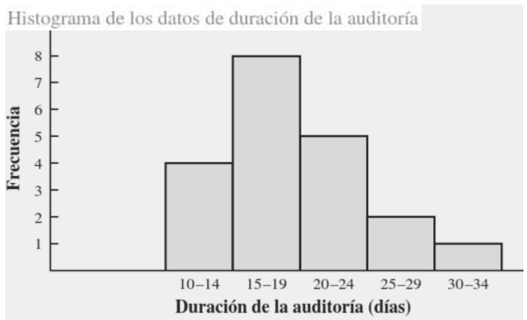


Juan Pablo Sucre Reyes

2.4 Histograma

• Presentación gráfica común de los datos cuantitativos, elaborada para datos previamente resumidos (*distribución de frecuencia, relativa o porcentual*).

• La variable de interés se coloca sobre el eje horizontal y la frecuencia, sobre el eje vertical con un rectángulo *(base determinada por los límites de clase sobre el eje horizontal, y altura = frecuencia, la frecuencia relativa o porcentual).*



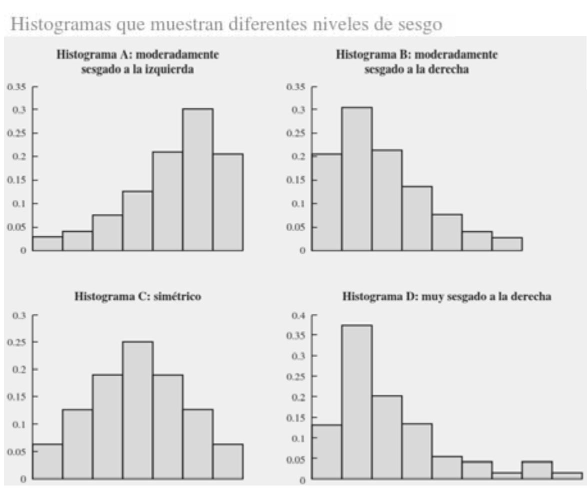
• Uno de los usos más importantes del histograma es proporcionar información acerca de la forma de una distribución.

Juan Pablo Sucre Reyes

2.4 Histograma

• Un histograma está sesgado a la izquierda si su cola se extiende más hacia esta dirección. Si está sesgado a la derecha si su cola se extiende más hacia esta dirección.

• En un histograma simétrico la cola izquierda imita la forma de la cola derecha (en aplicaciones nunca lo son).

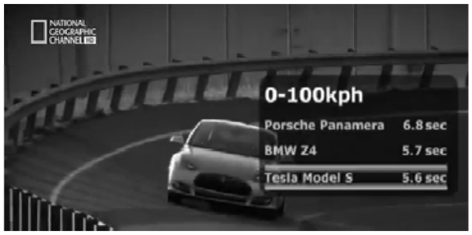


Juan Pablo Sucre Reyes

3. Medidas numéricas de posición, dispersión, forma y asociación

• Si se calculan para los datos de una muestra, se les llama *estadístico muestral*. Si se calculan para los datos de una población, se les llama *parámetros poblacionales*.

• En inferencia estadística, un estadístico muestral es un estimador puntual del parámetro poblacional correspondiente.



3.1 Medidas de posición o localización

• *Media:* medida de la ubicación central de los datos. Si los datos son para una muestra, la media se denota por x; si son para una población, se denota por μ.

MEDIA MUESTRAL

$$\bar{x} = \frac{\Sigma x_i}{n}$$

•*Ejemplo: En una oficina de colocación de empleos a nivel universitario envió un cuestionario a una muestra de 12 licenciados en administración de empresas recién egresados solicitando información sobre los sueldos mensuales iniciales:*

Sueldos mensuales iniciales para una muestra de 12 licenciados

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

•*Solución: Se tiene así:*

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + \cdots + x_{12}}{12} = \frac{3450 + 3550 + \cdots + 3480}{12} = \frac{42480}{12} = 3540$$

• El número de observaciones en una población se denota por N y el símbolo para la media poblacional es μ.

MEDIA POBLACIONAL

$$\mu = \frac{\Sigma x_i}{N}$$

3.1 Medidas de posición o localización

• *Mediana:* otra medida de ubicación central; es el valor de en medio cuando los datos están acomodados en orden ascendente (del valor menor al valor mayor).

MEDIANA

Ordene los datos de forma ascendente (del valor menor al valor mayor).

a) Para un número impar de observaciones, la mediana es el valor de en medio.
b) Para un número par de observaciones, la mediana es el promedio de los dos valores de en medio.

• *Ejemplo 1:* mediana de los tamaños de grupo para la muestra de cinco grupos de estudiantes universitarios:

32 42 46 46 54

• *Ejemplo 2:* mediana de los sueldos iniciales para los 12 licenciados en administración de empresas:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

• De donde:

$$\text{Mediana} = \frac{3490 + 3520}{2} = 3505$$

• *NOTA:* Siempre que un conjunto de datos contiene valores extremos, la mediana suele ser la medida preferida de posición central.

3.1 Medidas de posición o localización

•*Moda:* es es el valor que ocurre con mayor frecuencia.

• *Ejemplo 1:* moda de los tamaños de grupo en 5 grupos de estudiantes universitarios:

32 42 46 46 54

• *Ejemplo 2:* moda de los sueldos iniciales para los 12 licenciados administradores:

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

• Si los datos contienen exactamente dos modas, se dice que son bimodales. Si son más de dos, se dice que son multimodales.

• *Percentil:* muestra cómo se distribuyen los datos en el intervalo del valor menor al valor mayor.

El percentil *p*-ésimo es un valor tal que *por lo menos p* por ciento de las observaciones es menor o igual que este valor, y *por lo menos* (100 − *p*) por ciento de las observaciones es mayor o igual que este valor.

•

CÁLCULO DEL *p*-ÉSIMO PERCENTIL

Paso 1. Ordene los datos de modo ascendente (del valor menor al valor mayor).

Paso 2. Calcule un índice *i*

$$i = \left(\frac{p}{100}\right)n$$

donde *p* es el percentil de interés y *n* es el número de observaciones.

Paso 3. *a)* Si *i no es un entero, redondéelo.* El entero siguiente *mayor* que *i* denota la posición del *p*-ésimo percentil.

b) Si *i es un entero,* el *p*-ésimo percentil es el promedio de los valores en las posiciones *i* e *i* + 1.

3.1 Medidas de posición o localización

•*Ejemplo 1: percentil 85 para los datos de los sueldos iniciales mensuales de los 12 licenciados administradores:*

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925

$$i = \left(\frac{p}{100}\right)n = \left(\frac{85}{100}\right)12 = 10.2$$

• Como *i no es un entero, se redondea. La posición del percentil 85 es el siguiente entero mayor que 10,2, es decir, la posición 11 ó 3 730.*

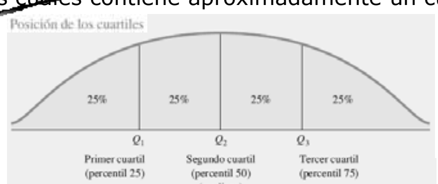
• *Ejemplo 2: percentil 50 para los datos de los sueldos iniciales:* $i = \left(\frac{50}{100}\right)12 = 6$

• *El percentil 50 es (3490 + 3520)/2 = 3505 (coincide con la mediana).*

Juan Pablo Sucre Reyes

3.1 Medidas de posición o localización

• *Cuartiles:* A menudo es recomendable dividir los datos en cuatro partes, cada una de las cuales contiene aproximadamente un cuarto, o 25% de las observaciones



Q_1 = primer cuartil, o percentil 25
Q_2 = segundo cuartil, o percentil 50 (también la mediana)
Q_3 = tercer cuartil, o percentil 75

•*Ejemplo 1: cuartiles para los datos de los sueldos iniciales mensuales de los 12 licenciados administradores:*

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925

• El cálculo de los cuartiles Q1 y Q3 requiere el uso de la regla para obtener los percentiles 25 y 75.

$$i = \left(\frac{p}{100}\right)n = \left(\frac{25}{100}\right)12 = 3$$

$$Q_1 = (3450 + 3480)/2 = 3465.$$

$$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)12 = 9$$

$$Q_3 = (3550 + 3650)/2 = 3600.$$

• Los cuartiles dividen los datos de los sueldos iniciales en cuatro partes, de las cuales cada una contiene 25% de las observaciones.

3310	3355	3450	3480	3480	3490	3520	3540	3550	3650	3730	3925
		$Q_1 = 3465$			$Q_2 = 3505$ (mediana)			$Q_3 = 3600$			

Juan Pablo Sucre Reyes

3.2 Medidas de variabilidad o dispersión

$$\text{Rango} = \text{valor mayor} - \text{valor menor}$$

- *Rango:* Medida más sencilla de variabilidad.
- Basada sólo en 2 de las observaciones (valores extremos influyen mucho en él).
- *Ejemplo 1: Rango de los sueldos iniciales para los 12 licenciados administradores:*

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

- *El rango = 3925 - 3310 = 615.*

- *Rango intercuartílico (RIC):* rango de la media de 50% de los datos. $\text{RIC} = Q_3 - Q_1$
- *Ejemplo 1: rango intercuartílico de los sueldos mensuales iniciales, los cuartiles son Q_3 = 3600 y Q_1= 3 465. Por tanto, RIC = 3600 – 3465 = 135.*

- *Varianza:* Utiliza todos los datos. Se basa en la diferencia entre el valor de cada observación (x_i) y la media (*X para una muestra; μ para una población*): desviación respecto de la media.

VARIANZA POBLACIONAL

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

VARIANZA MUESTRAL

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

- La varianza muestral s^2 es el estimador de la varianza poblacional σ^2.
- Útil para comparar la variabilidad de dos o más variables *(> valor ⇒ > variabilidad).*

Juan Pablo Sucre Reyes

3.2 Medidas de variabilidad o dispersión

- *Ejemplo 1: Varianza de los tamaños de grupo de la muestra de cinco grupos de estudiantes universitarios:*

Número de estudiantes en el grupo (x_i)	Tamaño de grupo medio ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media ($x_i - \bar{x})^2$
46	44	2	4
54	44	10	100
42	44	−2	4
46	44	2	4
32	44	−12	144
		0	256
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

La varianza es:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{256}{4} = 64$$

64 estudiantes ^{2}.

- *Ejemplo 2: varianza para los sueldos iniciales para los 12 licenciados administradores:*

Sueldo mensual (x_i)	Media muestral ($\bar{x}$)	Desviación respecto de la media ($x_i - \bar{x}$)	Desviación cuadrada respecto de la media ($x_i - \bar{x})^2$
3450	3540	−90	8100
3550	3540	10	100
3650	3540	110	12100
3480	3540	−60	3600
3355	3540	−185	34225
3310	3540	−230	52900
3490	3540	−50	2500
3730	3540	190	36100
3540	3540	0	0
3925	3540	385	148225
3520	3540	−20	400
3480	3540	−60	3600
		0	301850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

La varianza es:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{301\,850}{11} = 27\,440.91$$

27440,91 $us^2

Juan Pablo Sucre Reyes

3.2 Medidas de variabilidad o dispersión

• *Desviación estándar:* se deriva de la varianza:

$$\text{Desviación estándar muestral} = s = \sqrt{s^2}$$

$$\text{Desviación estándar poblacional} = \sigma = \sqrt{\sigma^2}$$

• La desviación estándar muestral s es el estimador de la poblacional σ.

• Más fácil de interpretar ya que se mide en las mismas unidades que los datos

•*Ejemplo 1: la desviación estándar de los tamaños de grupo de la muestra de cinco grupos de estudiantes universitarios:* $s = \sqrt{64} = 8.$ *estudiantes.*

•*Ejemplo 2: la desviación estándar de los sueldos mensuales iniciales:*

• $s = \sqrt{27\,440.91} = 165.65.$ *$us*

• *Coeficiente de variación :* medida relativa de la variabilidad; mide la desviación estándar con respecto a la media.

COEFICIENTE DE VARIACIÓN

$$\left(\frac{\text{desviación estándar}}{\text{media}} \times 100\right)\%$$

• La varianza muestral s^2 es el estimador de la varianza poblacional σ^2.

• Útil para comparar la variabilidad de dos o más variables *(> valor ⇒ > variabilidad).*

•*Ejemplo 1: CV de los tamaños de grupo:* $[(8/44) \times 100]\% = 18.2\%.$ *la desviación estándar muestral es 18,2% del valor de la media muestral.*

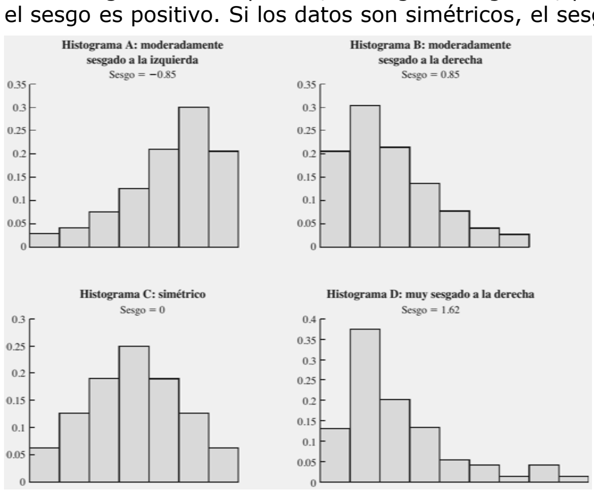
• *Ejemplo 2: CV de los sueldos mensuales iniciales:* $[(165.65/3\,540) \times 100]\% = 4.7\%.$

• *la desviación estándar muestral es sólo 4.7% del valor de la media muestral.*

Juan Pablo Sucre Reyes

3.3 Medidas de la forma de la distribución

• *Forma de la distribución:* La fórmula del sesgo es compleja (*fácil en software*). Para datos sesgados a la izquierda, el sesgo es negativo; para datos sesgados a la derecha, el sesgo es positivo. Si los datos son simétricos, el sesgo es cero.



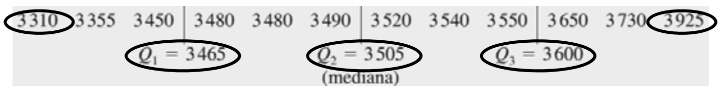
$$\text{Sesgo} = \frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

• En una distribución simétrica, la media y la mediana son iguales. Con datos sesgados positivamente, la media será mayor que la mediana; y con sesgo negativo, la media será menor que la mediana.

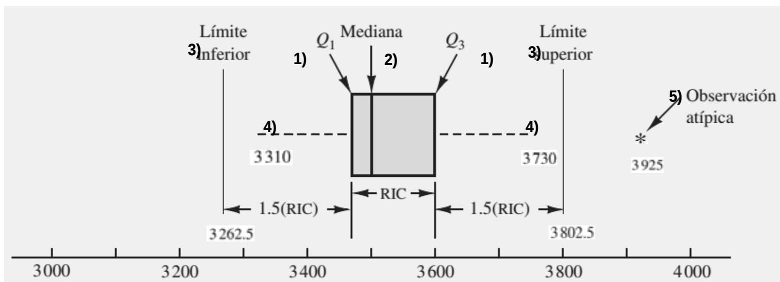
Juan Pablo Sucre Reyes

4. Análisis exploratorio de datos

- *Resumen de cinco números:* los datos se resumen con: *1.Valor menor 2. Primer cuartil (Q_1) 3.Mediana (Q_2) 4. Tercer cuartil (Q_3) 5.Valor mayor*
- Se colocan los datos en orden ascendente y se identifican los 5 números.
- *Ejemplo: sueldos mensuales para la muestra de 12 licenciados recién egresados:*



- *Diagrama de caja:* resumen gráfico basado en un resumen de 5 números. La clave es el cálculo de la mediana y los cuartiles Q_1 y Q_3. Se usa también el $RIC = Q_3 - Q_1$.
- *Ejemplo 1: diagrama de caja de los datos de los sueldos mensuales iniciales:*

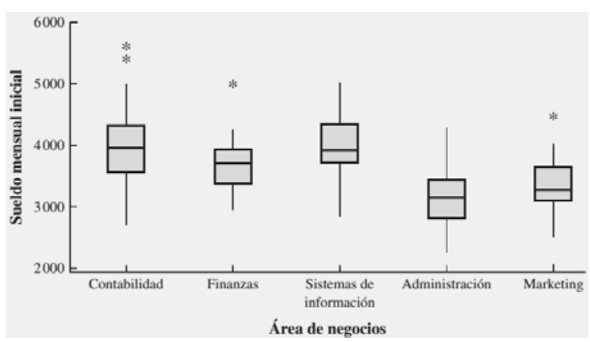


- Aunque los límites siempre se calculan, no siempre se trazan.

Juan Pablo Sucre Reyes

4. Análisis exploratorio de datos

- *Ejemplo 2: diagrama de caja de los sueldos mensuales iniciales de los licenciados en administración de empresas por área de especialización, se seleccionó una muestra de 111 licenciados recién graduados por campo de especialización y sueldo mensual inicial*



- *Conclusiones:*
- Sueldos más altos: contabilidad; sueldos más bajos: administración y marketing.
- Medianas: la de los sueldos de contabilidad y sistemas de información es similar y mayor. Le sigue finanzas, y administración y contabilidad tienen una mediana inferior.
- Observaciones atípicas: sueldos altos para contabilidad, finanzas y marketing.
- Variabilidad: Los sueldos de finanzas (menor variación), contabilidad (la mayor).

Juan Pablo Sucre Reyes

5. Medidas de asociación entre variables

• Medidas descriptivas de la relación entre dos variables: *covarianza y correlación*.

•*Covarianza:* medida descriptiva de la asociación lineal entre dos variables.

• Para una muestra de tamaño n con observaciones (x_1,y_1), (x_2,y_2),etc.; y para una población de tamaño N con observaciones (x_1,y_1), (x_2,y_2), etc. se tiene:

COVARIANZA MUESTRAL

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

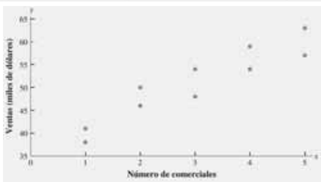
COVARIANZA POBLACIONAL

$$\sigma_{xy} = \frac{\Sigma(x_i - \mu_x)(y_i - \mu_y)}{N}$$

•*Ejemplo: El gerente de una tienda quiere determinar la relación entre el número de comerciales de televisión transmitidos el fin de semana y las ventas durante la semana siguiente. Los datos muestrales con las ventas (cientos de $us) se dan a continuación; además se muestra el diagrama de dispersión relacionado:*

•

Week	Number of Commercials x	Sales Volume ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46



La relación lineal entre x e y será medida por:

Cálculos de la covarianza muestral

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
2	50	−1	−1	1
5	57	2	6	12
1	41	−2	−10	20
3	54	0	3	0
4	54	1	3	3
1	38	−2	−13	26
5	63	2	12	24
3	48	0	−3	0
4	59	1	8	8
2	46	−1	−5	5
30	510	0	0	99

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{10-1} = 11$$

Juan Pablo Sucre Reyes

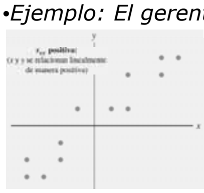
5.1 Interpretación de la covarianza

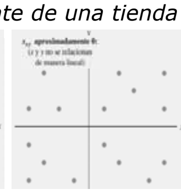
• Si el valor de S_{xy} es (+), los puntos con mayor influencia en S_{xy} deben estar en los cuadrantes I y III $\Rightarrow \exists$ asociación lineal positiva entre x e y *(si x ↑ ⇒ y ↑).*

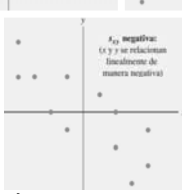
• Si el valor de S_{xy} es (-), los puntos con mayor influencia en S_{xy} deben estar en los cuadrantes II y IV $\Rightarrow \exists$ asociación lineal negativa entre x e y (*si x ↑ ⇒ y ↓).*

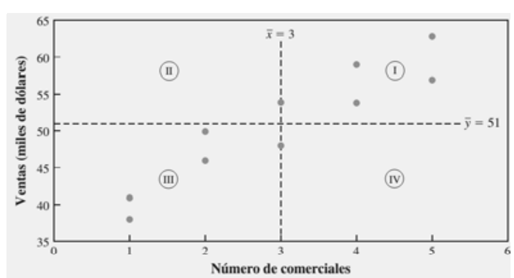
•Si los puntos están distribuidos de manera uniforme en los cuatro cuadrantes, el valor de S_{xy} será cercano a cero $\Rightarrow$ no $\exists$ una asociación lineal entre x e y.

•*Ejemplo: El gerente de una tienda quiere determinar la relación ….*









• *∃ relación lineal positiva,* $S_{xy} = 11$.

• *NOTA: un problema con la covarianza es que su valor depende de las unidades de medida para x e y ⇒ mejor usar correlación.*

Juan Pablo Sucre Reyes

5.2 Coeficiente de correlación

• Para los datos muestrales, es r_{xy} :

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO-MOMENTO DE PEARSON: DATOS MUESTRALES

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

donde
r_{xy} = coeficiente de correlación muestral
s_{xy} = covarianza muestral
s_x = desviación estándar muestral de x
s_y = desviación estándar muestral de y

•*Ejemplo 1: coeficiente de correlación muestral para la tienda de estéreos y equipos de sonido. Calculamos* S_x y S_y *con los datos de la tabla:*

• $s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49 \quad s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$ *y así:* $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = 0.93$

• Para los datos poblacionales, es ρ_{xy} :

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO-MOMENTO DE PEARSON: DATOS POBLACIONALES

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

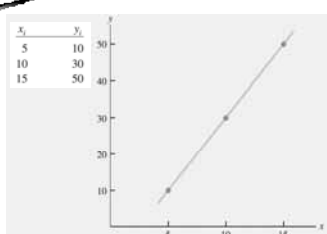
donde
ρ_{xy} = coeficiente de correlación poblacional
σ_{xy} = covarianza poblacional
σ_x = desviación estándar poblacional de x
σ_y = desviación estándar poblacional de y

• *NOTA: El coeficiente de correlación muestral* r_{xy} *proporciona una estimación del coeficiente de correlación poblacional* ρ_{xy} *.*

Juan Pablo Sucre Reyes

5.2.1 Interpretación del coeficiente de correlación

• *Ejemplo: El diagrama de dispersión representa una relación positiva perfecta entre x y y con base en los datos muestrales siguientes:*



Cálculos utilizados para obtener el coeficiente de correlación muestral

x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
5	10	−5	25	−20	400	100
10	30	0	0	0	0	0
15	50	5	25	20	400	100
30	90	0	50	0	800	200

$\bar{x} = 10 \quad \bar{y} = 30$

• *A partir de ello se calcula:*

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{200}{2} = 100 \quad s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{50}{2}} = 5 \quad s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{800}{2}} = 20$$

• *y de esta manera:* $r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$

• Un coeficiente de correlación muestral de +1 = relación lineal positiva perfecta entre x y y; de -1 = relación lineal negativa perfecta; de 0 = no existe relación lineal.

• Proporciona una medida de asociación lineal y no necesariamente de causalidad. Una correlación alta entre dos variables no significa que los cambios en una variable ocasionarán cambios en la otra.

Juan Pablo Sucre Reyes

5. Medidas de asociación entre variables

AUTOevaluación 46. A continuación se presentan cinco observaciones tomadas para dos variables.

x_i	6	11	15	21	27
y_i	6	9	6	17	12

a) Elabore un diagrama de dispersión con estos datos.
b) ¿Qué indica el diagrama de dispersión acerca de la relación entre *x* y *y*?
c) Calcule e interprete la covarianza muestral.
d) Determine e interprete el coeficiente de correlación muestral.

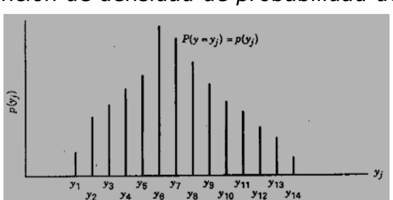
AUTOevaluación 51. Las temperaturas diarias altas (High) y bajas (Low) para 14 ciudades de todo el mundo se muestran en el siguiente cuadro (The Weather Channel, 22 de abril de 2009).

City	High	Low	City	High	Low
Athens	68	50	London	67	45
Beijing	70	49	Moscow	44	29
Berlin	65	44	Paris	69	44
Cairo	96	64	Rio de Janeiro	76	69
Dublin	57	46	Rome	69	51
Geneva	70	45	Tokyo	70	58
Hong Kong	80	73	Toronto	44	39

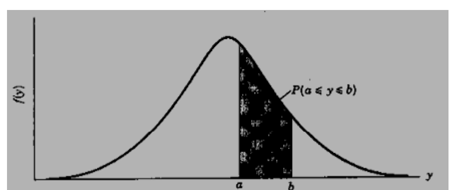
a) ¿Cuál es la media muestral de la temperatura alta?
b) ¿Cuál es la media muestral de la temperatura baja?
c) ¿Cuál es la correlación entre las temperaturas alta y baja? Comente.

6. Distribuciones de probabilidad

• Describen la estructura de la probabilidad de una variable aleatoria. *Cuando y es discreta, p(y) es la función de probabilidad de y. Cuando y es continua, f(y) es la función de densidad de probabilidad de y.*



a) Una distribución discreta



b) Una distribución continua

• *Resumen cuantitativo de las propiedades de las distribuciones de probabilidad:*

$$0 \le p(y_j) \le 1 \quad \text{todos los valores de } y_j$$
$$P(y = y_j) = p(y_j) \quad \text{todos los valores de } y_j$$
$$\sum_{\substack{\text{todos los} \\ \text{valores de } y_j}} p(y_j) = 1$$

$$0 \le f(y)$$
$$P(a \le y \le b) = \int_a^b f(y)\,dy$$
$$\int_{-\infty}^{\infty} f(y)\,dy = 1$$

6. Distribuciones de probabilidad

•*Media, varianza y valores esperados:* la media es una medida de su tendencia central o localización. Puede expresarse también en términos del valor esperado *E*.

$$\mu = \begin{cases} \int_{-\infty}^{\infty} y f(y)\,dy & y \text{ continua} \\ \sum_{\text{toda } y} y p(y) & y \text{ discreta} \end{cases} \qquad \mu = E(y) = \begin{cases} \int_{-\infty}^{\infty} y f(y)\,dy & y \text{ continua} \\ \sum_{\text{toda } y} y p(y) & y \text{ discreta} \end{cases}$$

• La varianza mide su variabilidad o dispersión. Puede expresarse en exclusivo en términos del valor esperado *E*. También conviene definir un operador de varianza *V*.

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy & y \text{ continua} \\ \sum_{\text{toda } y} (y-\mu)^2 p(y) & y \text{ discreta} \end{cases} \qquad \sigma^2 = E[(y-\mu)^2] \qquad V(y) \equiv E[(y-\mu)^2] = \sigma^2$$

• *Propiedades de los operadores E y V:* Sea y una variable aleatoria con media μ y varianza σ^2 y *c* es una constante, se tiene:

1. $E(c) = c$
2. $E(y) = \mu$
3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$

• Si hay dos variables aleatorias y_1 y y_2, se tiene:

7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$
8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2\ \mathrm{Cov}(y_1, y_2)$

donde $\mathrm{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)]$

9. $V(y_1 - y_2) = V(y_1) + V(y_2) - 2\ \mathrm{Cov}(y_1, y_2)$

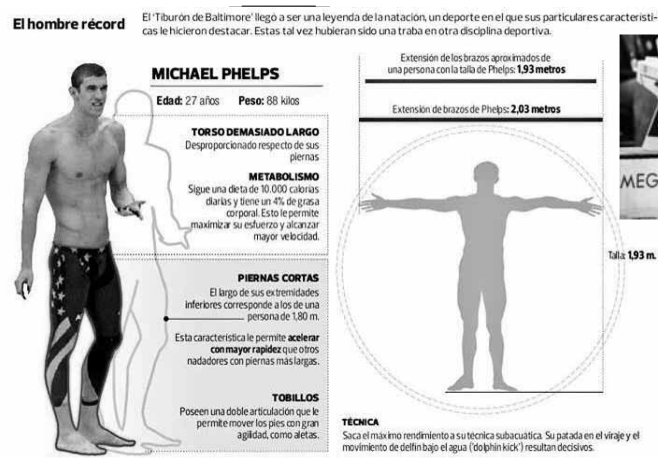
Si y_1 y y_2 son **independientes**, se tiene

10. $V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$
11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$

Sin embargo, observe que, en general,

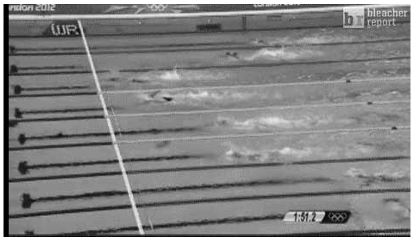
12. $E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

Variables aleatorias continuas = función de densidad de probabilidad



Variables aleatorias continuas = función de densidad de probabilidad

- $f(x)$ no proporciona las probabilidades directamente. El área bajo la gráfica $f(x)$ que corresponde a un intervalo dado representa la probabilidad de que la variable aleatoria continua x asuma un valor dentro de ese intervalo.
- Dado que el área bajo la gráfica $f(x)$ en cualquier punto en particular es cero, la probabilidad de cualquier valor particular de la variable aleatoria $x = 0$.
- 3 distribuciones de probabilidad continua: *uniforme, normal* (amplias aplicaciones y de uso extendido en inferencia estadística) y *exponencial.*



1. Distribución de probabilidad uniforme

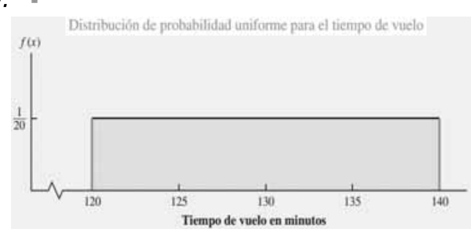
- Siempre que la probabilidad sea proporcional a la longitud del intervalo, la variable aleatoria está distribuida de manera uniforme.

FUNCIÓN DE DENSIDAD DE PROBABILIDAD UNIFORME

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{para } a \le x \le b \\ 0 & \text{en cualquier otro caso} \end{cases}$$

- *Ejemplo: Considere la variable aleatoria x que representa el tiempo de vuelo de un avión que viaja de Chicago a Nueva York. Suponga que este tiempo puede ser cualquier valor en el intervalo de 120 a 140 minutos. La probabilidad de que el tiempo de vuelo esté dentro de cualquier intervalo de 1 minuto es igual a la probabilidad de que esté dentro de cualquier otro intervalo de 1 minuto contenido dentro del intervalo mayor de 120 a 140 minutos. Así cada intervalo de 1 minuto es igualmente probable.*
- *Solución: para a=120 y b=140, se tiene:* -

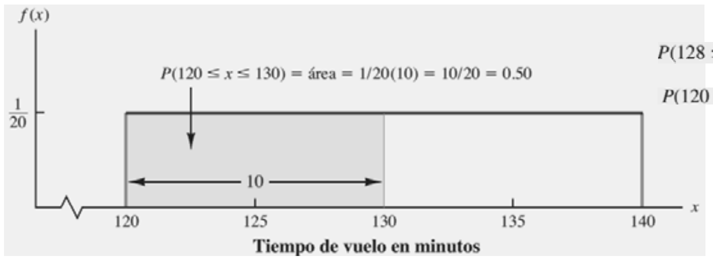
$$f(x) = \begin{cases} 1/20 & \text{para } 120 \le x \le 140 \\ 0 & \text{en cualquier otro caso} \end{cases}$$



- *¿Cuál es la probabilidad de que el tiempo de vuelo se encuentre entre 120 y 130 minutos? $P(120 \le x \le 130) = 0{,}50$.*

1.1 El área como medida de la probabilidad

- Identificada la función de densidad de probabilidad f(x), la probabilidad de que x tome un valor entre uno inferior x_1 y uno superior x_2 se obtiene al calcular el área bajo la gráfica f(x) en el intervalo de x_1 a x_2
- *Ejemplo: ¿Cuál es la probabilidad de que el tiempo de vuelo se encuentre entre 120 y 130 minutos?; ¿entre 128 y 136 minutos?, ¿entre 120 y 140 minutos?*



$P(128 \le x \le 136) = 8(1/20) = 0.40.$

$P(120 \le x \le 140) = 20(1/20) = 1$

- *Propiedades: 1) f(x)≥ 0 para todos los valores de x; 2) Área total bajo f(x) = 1.*
- *Valor esperado y varianza: para toda f(x)* $E(x) = \frac{a+b}{2}$ $\quad \mathrm{Var}(x) = \frac{(b-a)^2}{12}$

- *Ejemplo: Media y varianza para tiempos de vuelo de Chicago a Nueva York:*
- $E(x) = \frac{(120+140)}{2} = 130$ $\quad \mathrm{Var}(x) = \frac{(140-120)^2}{12} = 33.33$ *y además:* $\sigma = 5.77$ minutos

2. Distribución de probabilidad normal

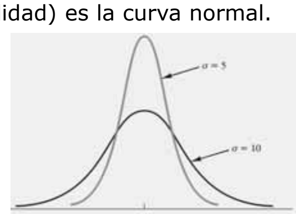
- La más importante para describir una variable aleatoria continua. Utilizado en una amplia variedad de aplicaciones (vida cotidiana) y sobretodo en inferencia estadística.
- *La curva normal:* La forma de la distribución normal es una curva con forma de campana y su función de densidad de probabilidad es:

FUNCIÓN DE DENSIDAD DE PROBABILIDAD NORMAL

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Donde: μ = media
σ = desviación estándar
π = 3.14159
e = 2.71828

- La distribución normal tiene 7 características principales:
- 1. La familia de distribuciones se diferencia por medio de dos parámetros: μ y σ.
- 2. El punto más alto de la curva está sobre la media, que coincide con la Me y la Mo.
- 3. La media de una distribución normal puede ser: negativa, cero o positiva.
- 4. Es simétrica: la forma de la curva a la izquierda de la μ es igual a la derecha. Sus extremos se extienden hacia el infinito (nunca tocan el eje x). No están sesgadas (0).
- 5. La σ determina qué tan plana y ancha (variabilidad) es la curva normal.



2.1 Propiedades de la distribución normal

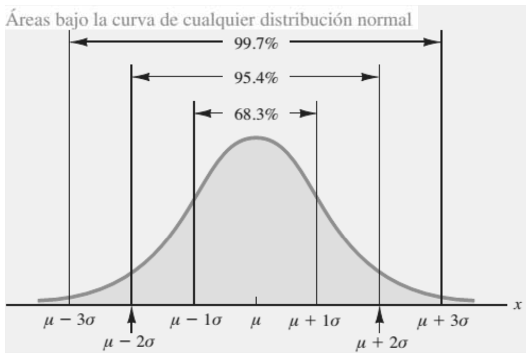
•6. Las probabilidades para la variable aleatoria normal = las áreas bajo la curva normal (*área total bajo la curva = 1, a la izquierda o derecha de μ = 0,50*)

•7. Los % de los valores (variable aleatoria normal) en intervalos de uso común son:

•*a) 68,3% de los valores se sitúan (+/-) a 1.σ de su μ.*

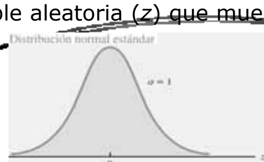
•*b) 95,4% de los valores se sitúan (+/-) a 2.σ de su μ.*

•*c) 99,7% de los valores se sitúan (+/-) a 3.σ de su μ.*



2.2 Distribución de probabilidad normal estándar

• Variable aleatoria (*z*) que muestra una distribución normal con una μ = 0 y σ =1.



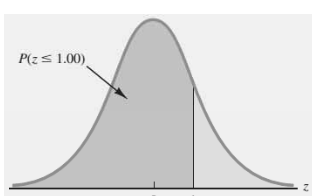
FUNCIÓN DE DENSIDAD NORMAL ESTÁNDAR

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

• Si la probabilidad de que una variable aleatoria normal esté dentro de cualquier intervalo específico = el área bajo la curva normal en ese intervalo; para la distribución normal estándar, las áreas bajo la curva normal ya se han estimado *(tablas).*

• Los tres tipos de probabilidades que se necesita calcular incluyen: 1) la probabilidad de que la variable aleatoria normal estándar *z* ≤ valor determinado

•*Ejemplo: Probabilidad de que z ≤ 1,00, esto es: P(z≤ 1,00).*



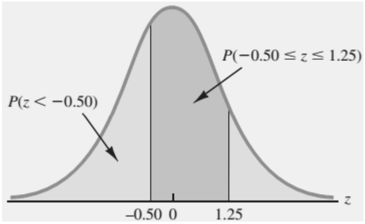
z	0.00	0.01	0.02
.			
.			
.			
0.9	0.8159	0.8186	0.8212
1.0	$\boxed{0.8413}$	0.8438	0.8461
1.1	0.8643	0.8665	0.8686
1.2	0.8849	0.8869	0.8888
.			
.			
.			

P(z ≤ 1.00)

2.2 Distribución de probabilidad normal estándar

•Los tres tipos de probabilidades que se necesita calcular incluyen: *2) la probabilidad de que z esté entre dos valores dados.*

•*Ejemplo1: Probabilidad de que z esté entre -0,50 y 1,25; es decir, P(-0,50 ≤ z≤ 1,25).*

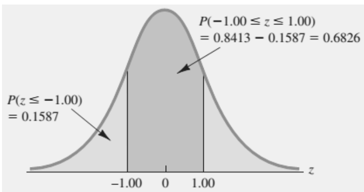


z	0.00	0.01	0.02	0.03	0.04	0.05
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944

z	0.00
−0.9	0.1841
−0.8	0.2119
−0.7	0.2420
−0.6	0.2743
−0.5	0.3085

• *3 pasos: 1° hallar el área bajo la curva a la izq. de z=1,25; 2° hallar el área bajo la curva a la izq. de z =-0,50. 3° restar el área del paso 2° del 1° = P(-0,50 ≤ z ≤ 1,25) = 0,8944 – 0,3085 = 0,5859.*

•*Ejemplo2: Probabilidad de que z esté dentro de 1.σ de μ; es decir, P(-1,00 ≤ z ≤ 1,00)*



z	0.00
−1.4	0.0808
−1.3	0.0968
−1.2	0.1151
−1.1	0.1357
−1.0	0.1587

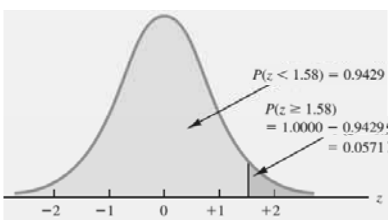
z	0.00
0.5	0.6915
0.6	0.7257
0.7	0.7580
0.8	0.7881
0.9	0.8159
1.0	0.8413

Juan Pablo Sucre Reyes

2.2 Distribución de probabilidad normal estándar

•Los tres tipos de probabilidades que se necesita calcular incluyen: *3) la probabilidad de que z ≥ valor determinado.*

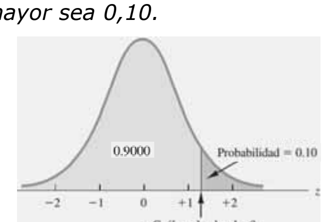
•*Ejemplo: Probabilidad de que z sea por lo menos igual a 1,58; es decir, P(z ≥ 1,58).*



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429

• *P(z ≥ 1,58) = 0,0571*

• *Ejemplo*: Determinar un valor de z tal que la probabilidad de obtener un valor de z mayor sea 0,10.*



z	0.06	0.07	0.08
1.0	0.8554	0.8577	0.8599
1.1	0.8770	0.8790	0.8810
1.2	0.8962	0.8980	$\boxed{\mathbf{0.8997}}$
1.3	0.9131	0.9147	0.9162
1.4	0.9279	0.9292	0.9306
.			
.			

Valor de probabilidad acumulada más cercano a 0.9000

Juan Pablo Sucre Reyes

2.3 Probabilidades para cualquier distribución de probabilidad normal

• Dada una distribución normal con cualquieras μ y σ, las diversas probabilidades se hallan *convirtiendo* primero a la distribución normal estándar. Luego se usa su tabla y los valores de z apropiados para obtener las probabilidades buscadas.

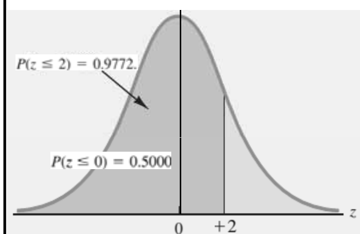
CONVERSIÓN A LA VARIABLE ALEATORIA NORMAL ESTÁNDAR

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

• *z = # de σ de la μ a las que está la variable aleatoria normal x.*

•*Ejemplo: Sea una distribución con μ=10 y σ=2. Halle la probabilidad de que la variable aleatoria x esté entre 10 y 14.*

•*Solución: En x=10, z=(10-10)/2= 0 y en x=14, z=(14 -10)/2=2. Así P(0 ≤ z ≤ 2)=?*



z	0.00
1.8	0.9641
1.9	0.9713
2.0	0.9772

P(0 ≤ z ≤ 2)= 0,9772 – 0,5000 =0,4772.

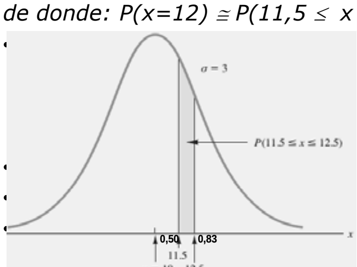
3. Aproximación normal de las probabilidades binomiales

• *RECUERDE:* experimento binomial = n ensayos independientes con dos resultados posibles: éxito *(p)* o *fracaso (1-p).* P(x) = probabilidad de x éxitos en n ensayos.

• Cuando n es grande, es difícil evaluar la función binomial. Si *np ≥ 5* y *n(1-p) ≥* 5, se usa la aproximación normal a la binomial, donde *μ=np* y *σ=√np(1-p)* (curva normal).

•*Ejemplo 1: Una empresa tiene una historia de cometer errores en 10% de sus facturas. Se tomó una muestra de 100 facturas y se quiere calcular la probabilidad de que 12 contengan errores.*

•*Solución: Al aplicar la aproximación normal: μ = 100.0,1= 10 y σ = 100.0,1.0,9= 3. de donde: P(x=12) ≅ P(11,5 ≤ x ≤ 12,5) ; 0,5 = factor de corrección de continuidad.*



Estandarizando valores x en z se tiene:

$$z = \frac{x-\mu}{\sigma} = \frac{12.5 - 10.0}{3} = 0.83 \quad \text{en } x = 12.5$$

$$z = \frac{x-\mu}{\sigma} = \frac{11.5 - 10.0}{3} = 0.50 \quad \text{en } x = 11.5$$

de donde por tablas:

P(11,5 ≤ x ≤ 12,5) = 0,7967 – 0,6915 = 0,1052

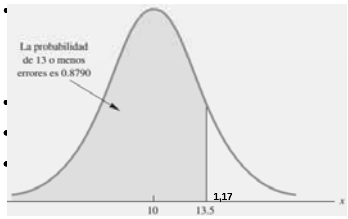
z	0.00	0.01	0.02	0.03
0.7	0.7580	0.7611	0.7642	0.7673
0.8	0.7881	0.7910	0.7939	0.7967

z	0.00
0.5	0.6915
0.6	0.7257

3. Aproximación normal de las probabilidades binomiales

•*Ejemplo 2: Una empresa tiene una historia de cometer errores en 10% de sus facturas. Se tomó una muestra de 100 facturas y se desea calcular la probabilidad de 13 o menos errores en la muestra de 100 facturas*

•*Solución: Al aplicar la aproximación normal: μ = 100.0,1= 10 y σ = 100.0,1.0,9= 3. de donde: P(x≤ 13) ≅ P(x ≤ 13,5) ; 0,5 = factor de corrección de continuidad.*



Estandarizando valor x = 13,5 en z se tiene:

$$z = \frac{13.5 - 10.0}{3.0} = 1.17$$

de donde por tabla:

P(x ≤ 13,5) = P (z ≤ 1,17) = 0,8790

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830

Juan Pablo Sucre Reyes

4. Distribución de probabilidad exponencial

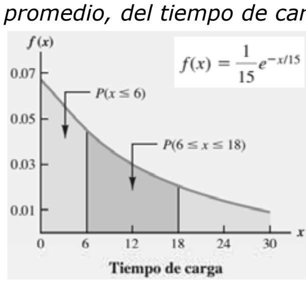
• Usada para variables aleatorias como: tiempo entre llegadas a un autolavado, tiempo para cargar un camión, distancia entre defectos en una carretera, etc. La f(x) es:

FUNCIÓN DE DENSIDAD DE PROBABILIDAD EXPONENCIAL

$$f(x) = \frac{1}{\mu} e^{-x/\mu} \quad \text{para } x \geq 0 \qquad \text{donde } \mu = \text{valor esperado o media}$$

• La media de la distribución y la desviación estándar de la distribución son iguales.

•*Ejemplo: Sea x = tiempo de carga para un camión (muelle Schips). Si la media, o promedio, del tiempo de carga es 15 minutos (μ = 15), se tiene f(x) y su gráfica:*



• *Determine la probabilidad de que cargar un camión tarde: a) 6 minutos o menos; b) 18 min. o menos; c) entre 6 y 18 minutos.*

•*Solución:* Calculo de área o probabilidad:

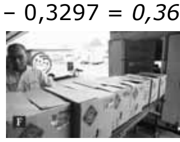
DISTRIBUCIÓN EXPONENCIAL: PROBABILIDADES ACUMULADAS

$$P(x \leq x_0) = 1 - e^{-x_0/\mu}$$

Así: $P(x \leq x_0) = 1 - e^{-x_0/15}$

• de donde: $P(x \leq 6) = 1 - e^{-6/15} = 0.3297$; $P(x \leq 18) = 1 - e^{-18/15} = 0.6988$ y

•P(6 ≤ x ≤ 18)= 0,6988 – 0,3297 = *0,3691*



Juan Pablo Sucre Reyes

4.1 Relación entre las distribuciones de Poisson y exponencial

• Recuerde que la función de probabilidad (discreta) de Poisson es:

$$f(x) = \frac{\mu^x e^{-\mu}}{x!} \quad \text{donde} \quad \begin{aligned} f(x) &= \text{probabilidad de } x \text{ ocurrencias en un intervalo} \\ \mu &= \text{valor esperado o número medio de ocurrencias en un intervalo} \\ e &= 2.71828 \end{aligned}$$

• La distribución de probabilidad exponencial continua está relacionada con la distribución de Poisson discreta. Si Poisson proporciona una descripción apropiada del número de ocurrencias por intervalo, la exponencial provee una descripción de la duración del intervalo entre ocurrencias.

•*Ejemplo: El número de autos que llegan a un autolavado en una hora se describe por medio de una distribución de Poisson con una media (μ) de 10 automóviles por hora.*

•*Solución:* la f(x) para la probabilidad de x llegadas por hora es: $f(x) = \frac{10^x e^{-10}}{x!}$

•el tiempo promedio entre la llegada de los vehículos es: $\frac{1 \text{ hora}}{10 \text{ automóviles}} = 0.1 \text{ hora/automóvil}$

•De donde la función de densidad de probabilidad exponencial que describe el tiempo entre las llegadas (*μ = 0,1 hora/automóvil*) es:

$$f(x) = \frac{1}{0.1} e^{-x/0.1} = 10e^{-10x}$$

Juan Pablo Sucre Reyes

GRACIAS POR SU ATENCIÓN…..

Juan Pablo Sucre Reyes