Capítulo 1

1. MARCO TEÓRICO

1.1 Histograma

La forma más común de representación gráfica de una distribución de frecuencias es el **histograma.** El histograma de una distribución de frecuencias se construye con rectángulos adyacentes, las alturas de los cuales representan las frecuencias de clase mientras que sus bases se extienden entre sucesivas fronteras de clase.

1.2 Medidas Descriptivas

Dado un conjunto de *n* medidas u observaciones, $x_1, x_2, \ldots x_n$, podemos describir su centro (medio o lugar central) de muchas maneras. Las más comunes son la **media aritmética** y la **mediana.**

La **media** o también conocida como **media aritmética** se define como:

Media $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

A veces es preferible usar la **mediana** como medida descriptiva del centro, o lugar, de un conjunto de datos. Esto es así, sobre todo, cuando se desea reducir los cálculos o eliminar el efecto de valores extremos (muy grandes o muy pequeños). La mediana de *n* observaciones, $x_1, x_2, \ldots x_n$, puede definirse como la mitad del conjunto de datos después de que las observaciones se han colocado en serie ordenada. Si las observaciones se organizan en serie ordenada y *n* es un número impar, la mediana es el valor de la observación con el número $\frac{n+1}{2}$; si *n* es un número par, la mediana se define como la media (promedio) de las observaciones con los números $\frac{n}{2}$ y $\frac{n+2}{2}$.

Ahora, con respecto a la **varianza** y la **desviación estándar** para una muestra representan medidas de dispersión alrededor de la media. Se calculan de manera parecida a aquellas para una población. La varianza s^2 es:

Varianza $$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

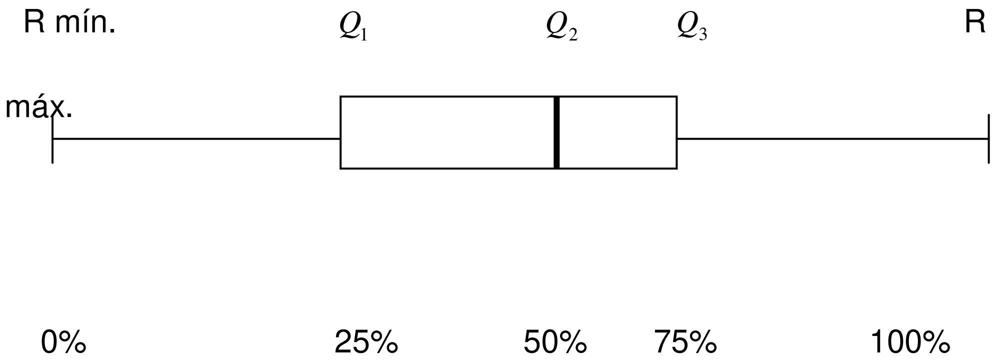
Desviación estándar $$s = \sqrt{s^2}$$

1.2.1 Medidas de posición de datos

Con respecto a otras medidas de dispersión, además de la varianza y la desviación estándar que son las más útiles en análisis estadístico, existen otras técnicas con las cuales puede medirse la dispersión de un conjunto de datos. Estas medidas adicionales de dispersión son los denominados **cuartiles**. Cada conjunto de datos tiene **tres cuartiles** que lo dividen en cuatro partes iguales. **El primer cuartil** es ese valor debajo del cual clasifica el 25% de las observaciones, y sobre el cual puede encontrarse el 75% restante. **El segundo cuartil** es justo la mitad. La mitad de las observaciones están por debajo y la mitad por encima; en este sentido, es lo mismo que la mediana. **El tercer cuartil** es el valor debajo del cual está el 75% de las observaciones y encima del cual puede encontrarse el 25% restante.

1.3 Diagrama de Caja

La información resumida contenida en los cuartiles pone relieve en la representación gráfica llamada **diagrama de caja**. La mitad central de los datos, que va del primero al tercer cuartiles, se representa con un rectángulo. La mediana se identifica con una barra dentro de esta caja. Una línea se extiende del tercer cuartil al máximo y otra del primer cuartil al mínimo.



Dibujo 1. Representación de Cuartiles. Libro Estadística aplicada a los negocios y la economía.

Los diagramas de caja son especialmente eficaces para la descripción gráfica de comparaciones entre conjuntos de observaciones. Son fáciles de entender y ejercen un poderoso impacto visual.

1.4 Coeficientes de Variación

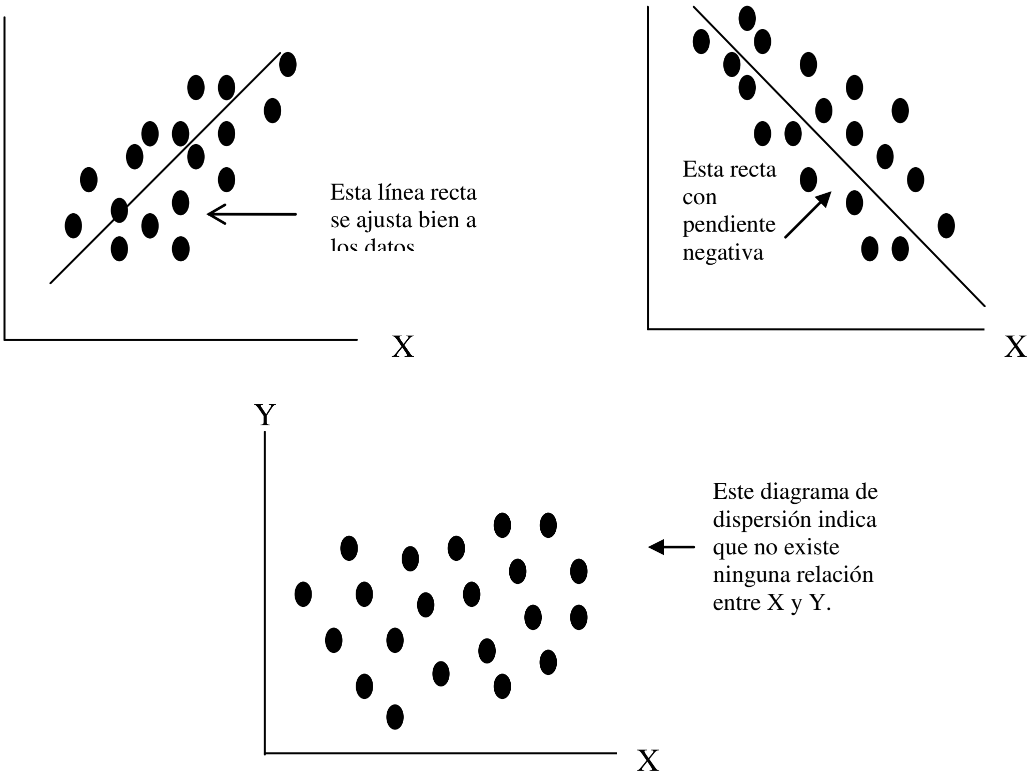
Como sabemos, un uso importante de la desviación estándar es servir como medida de dispersión. Sin embargo, se aplican ciertas limitaciones. Cuando se consideran 2 o más distribuciones que tienen medias significativamente diferentes, o que están medidas en unidades distintas, es peligroso sacar conclusiones respecto a la dispersión sólo con base en la desviación estándar.

Por tanto, con frecuencia debemos considerar el **Coeficiente de Variación** (CV), el cual sirve como medida relativa de dispersión. El coeficiente de variación determina el grado de dispersión de un conjunto de datos relativo a su medida. Se calcula la desviación estándar de una distribución por su media y multiplicando por 100.

Coeficiente de variación $$CV = \frac{s}{\overline{X}}(100)$$

1.5 Regresión Lineal

Es un modelo en el cual participan dos tipos de variables: La variable **X** que es la variable independiente (o variable explicativa), y la variable **Y** que es la dependiente (o variable de respuesta); éstas a su vez se relacionan y se presentan por medio de una línea recta.



Dibujo 2. Ejemplos de Diagrama de dispersión. Libro Estadística aplicada a los negocios y la economía.

1.5.1 Regresión Lineal Múltiple

En un modelo de regresión múltiple, **Y** es una función de dos o más variables independientes. Un modelo de regresión con *k* variables se puede expresar así:

$$Y = f(X_1, X_2, X_3, ..., X_k)$$

1.5.2 Autocorrelación

La **autocorrelación** ocurre cuando los términos de error no son independientes. Existe la autocorrelación positiva cuando los signos iguales se agrupan y existe una autocorrelación negativa cuando cada error es seguido de un error de signo opuesto. Los diagramas residuales nunca son tan obvios o tan fáciles de leer, afortunadamente existe una forma más confiable para detectar la autocorrelación en base en la prueba de **Durbin-Watson.**

1.5.2.1 Estadístico de Durbin-Watson

El estadístico de Durbin-Watson se calcula así:

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}$$

En donde e_i es el error en el periodo de tiempo t, y e_{i-1} es el error en el periodo anterior. La fórmula requiere que el término de error $(Y_i - \hat{Y}_i)$ se calcula para cada periodo y es difícil calcularlo manualmente. Este valor se utiliza para probar la hipótesis de que no existe correlación entre términos de error sucesivos, así:

$H_o : \rho_{e_i, e_{i-1}} = 0$ (No existe autocorrelación)

$$H_a : \rho_{e_i, e_{i-1}} \neq 0 \quad \text{(Existe autocorrelación)}$$

1.5.3 Análisis de Correlación

El coeficiente de correlación fue desarrollado por Carl Pearson a finales de siglo, y algunas veces se les llama el coeficiente de correlación producto-momento de Pearson. Representado con una r, el coeficiente de correlación puede asumir cualquier valor entre -1 y +1; es decir:

$$-1 \leq r \leq +1$$

Un valor de r= -1 indica una relación perfecta entre X y Y. Todas las observaciones quedan en una línea recta perfecta con una pendiente negativa. Por tanto, X y Y se moverán en direcciones opuestas, en cambio si r=1 la relación entre X y Y será positiva perfecta. Por el contrario, si se muestra muy poca o ninguna relación entre X y Y, r se aproxima a cero. En general, entre mayor sea el valor absoluto de r, más fuerte será la relación entre X y Y.

1.5.3.1 Coeficiente de Determinación

El coeficiente de determinación r^2 tiene significado sólo para las relaciones lineales. Dos variables pueden tener un r^2 de cero y sin embargo, están relacionadas en sentido curvilíneo.

1.5.4 Análisis de Varianza (Anova)

Dado el modelo de regresión, puede realizarse el **análisis de varianza** (ANOVA). El procedimiento del **ANOVA** prueba si alguna de las variables independientes tiene una relación con la variable dependiente. Si una variable independiente no está relacionada con la variable Y, su coeficiente debería ser cero. El procedimiento ANOVA prueba la hipótesis nula de que todos los valores β son cero contra la hipótesis alternativa de que por lo menos un β no es cero. Es decir:

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_A : \exists \beta_i \neq 0$$

Si no se rechaza la hipótesis nula, entonces no hay relación lineal entre Y y cualquiera de las variables independientes. Por otra parte, si la hipótesis nula se rechaza, por lo menos una variable independiente está relacionada linealmente con Y.

El proceso **ANOVA** establece una tabla y utiliza la prueba F, el formato general para una regresión múltiple es:

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrado Medio	Valor F
Entre muestras (tratamiento)	SCR	K	SCR/k	F=CMR/CME
Dentro de las muestras (error)	SCE	n-k-1	SCE/n-k-1	
Variación total	SCT	n-1		

Tabla I. ANOVA. Libro Probabilidad y Estadística para Ingenieros.