Finanzas Corporativas Aplicadas

Clase 2

César Núñez Cuevas
cnunezc@fen.uchile.cl

Pre Procesamiento de Datos

- Se refiere al proceso en el que se explora la información con el fin de generar un set de datos que permita ser trabajado correctamente.
- Missing Values o Data Inconsistente
- Data Ruidosa
- Conversión (Categorica)

Herramientas

- Medidas de Posición
- Medidas de Dispersión
- Prácticas Comunes

Promedio

- Se puede obtener al calcular el total de los valores por el tamaño del dataset.
- Fácil de calcular y entender.
- Muy sensible en muestras pequeñas.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Mediana

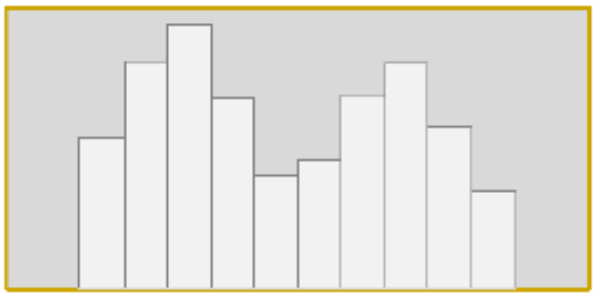
- Valor que divide la mitad una muestra.
- Es útil en estadística dado a que tiende a tener una variabilidad acotada.
- Hay formas distintas de calcularla dependiendo de la cantidad de datos (par o impar)

Impar
23
33
34
36
38
40
41
41
44

Par
12
30
31
37
38
40
41
41
44
45

Moda

- El valor que más se repite dentro del dataset.
- Tiene la función de poder analizar la forma de la distribución.
- Es la medida que entrega menos información.



Medidas de Dispersión

- Buscan representar cómo la data se desvía desde las medidas de dispersión.
- Sirven como indicadores de calidad de la información y cómo se mejora la variabilidad.
- Las medidas son: Rango, Varianza y Desviación Estándar.

Rango

- La diferencia entre el valor más alto y el bajo.
- Es la medida más simple en términos de dispersión.
- Puede ser engañosa en caso de valores extremos.

$$Rango(x) = X_{max} - X_{min}$$

Varianza

- Medida de variación sobre la media (promedio)
- Permite aproximar cuánto los puntos del set de datos están dispersos de la media.
- Se encuentra en unidades al cuadrado, por ende, no tiene interpretabilidad por si sola.
- NO negativa

$$\sigma^2 = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$$

Desviación Standard

- Distancia promedio de los puntos desde su promedio.
- Una baja desviación indica homogeneidad, una alta desviación indica heterogeneidad.
- Se encuentra en la misma unidad de medida que la variable. Ejemplo: Porcentajes.

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}}$$

Outlier

- Datos que es significativamente mayor o menor que los otros dentro del set de datos
- En caso de muestras pequeñas puede afectar de forma importante los cálculos. Por ejemplo: 2020
- Ojo con sesgos o diferentes muestras.

Outlier - Ejemplo

- Hay un outlier claro
- Hay efecto en el promedio y la desviación?

X	Hijos
1	0
2	3
3	1
4	2
5	10
Promedio	3.2
Des. Std.	3.5

Outlier - Ejemplo

- Generalmente se saca el outlier (en muestras grandes es el 1 % superior e inferior)
- Regularización de los promedios y dispersión.

X	Hijos
1	0
2	3
3	1
4	2
Promedio	1.5
Des. Std.	1.1

Intervalos de Confianza

- Permiten poder analizar la información aproximando un intervalo dónde el valor poblacional verdadero se encuentre.
- Utilizado para poder predecir estimar rangos de precios y detección de outliers.

$$IC(\mu) = \mu \pm t_\alpha \sigma$$

- α representa el nivel de "confianza" que se quiere asumir. Si es igual a 2,5 % se está asumiendo un nivel de confianza de un 95 %. Dado eso el valor de t_α será de 1,96
- En general, si hay un valor que se encuentre 3 desviaciones estándar fuera, puede ser considerado un outlier. (2,33 representa el 99 %)

Otras Métricas

- Covarianza: Permite capturar la relación (sinergía) entre dos variables.

$$\sigma_{xy} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \mu_i)(x_{ji} - \mu_k)$$

Otras Métricas

- Correlación: Permite estimar la dependencia de 2 variables (precios). Si es mayor a cero hay relación positica, si es menor a cero es negatica. Si es cero no hay relación o se puede decir que hay independencia.

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

$$\sigma_{xy} = \sigma_{xy} \cdot \sigma_x \cdot \sigma_y$$

Actividad

- Descargar información financiera de alguna acción. (Bolsa Santiago, Investing y Yahoo Finance)
- Calcular Métricas presentadas previamente en Excel.
- Interpretar resultados (brevemente).

Incomplete Data

- Qué sucede si hay información que no está?
- El proceso se compone de:
  - Inspección
  - Eliminación
  - Identificación
  - Reemplazo: Promedio o Interpolación

Incomplete Data

- Interpolación: Permite "conectar" 2 puntos dado que hay un valor vacío. (sólo series de tiempo). Puede ser con promedio o ecuación de la recta.

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1)$$

$$y - y_1 = m(x - x_1)$$

- En caso de valores vacíos para otras variables, usar otras técnicas. Promedio o eliminación son las más utilizadas.

Prácticas Comunes

- Los valores vacíos, Excel los asume como 0. Ver operaciones.
- Revisar siempre el formato de las fechas.
- Revisar siempre el formato númerico; marcador de decimales versus marcador de miles.
- Revisar cuál es el marcador para los parámetros de las funciones.
- Desbalance de variables. (70-30)

Contenidos

	Fecha	Entregable	Detalle
Clase 0	18-04		Introducción al Curso
Clase 1	25-04		Manejo de Excel - Pre Procesamiento de Datos
Clase 2	02-05		Pivot Tables - Data Visualization
Clase 3	09-05	Propuesta Gráficos	Introducción a Power BI
Clase 4	16-05	Presentación Profesor	Visualizaciones en Power BI
Clase 5	23-05	Presentación Dashboard	Bases de Datos Relacionadas en Power BI