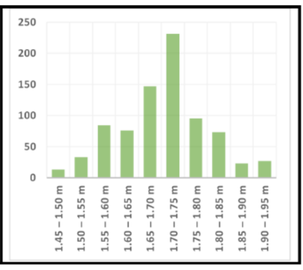
Histograma

El **histograma** es la representación visual de la distribución de los datos a lo largo del intervalo continuo de la variable que se está analizando. El histograma se compone por un conjunto de barras que representan la frecuencia en que aparece un dato dentro diferentes subintervalos de la variable analizada. Los histogramas nos permiten encontrar de manera visual estimaciones acerca de la concentración de los datos, además nos permite observar si existen vacíos en la distribución o valores inusuales.

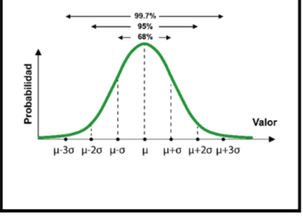
Rango de altura (m)	Cantidad de Alumnos
1.45 – 1.50 m	13
1.50 – 1.55 m	33
1.55 – 1.60 m	84
1.60 – 1.65 m	76
1.65 – 1.70 m	147
1.70 – 1.75 m	231
1.75 – 1.80 m	94
1.80 – 1.85 m	73
1.85 – 1.90 m	23
1.90 – 1.95 m	26



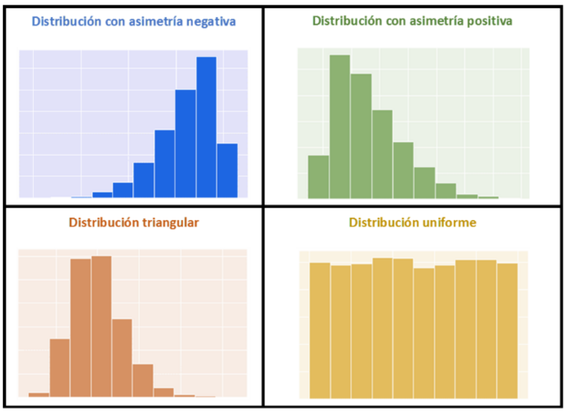
En nuestro primer ejemplo de histogramas, representamos las alturas de ochocientos alumnos, nuestro histograma evidencia que la mayoría de los alumnos tiene una altura cercana a 1.70 metros y que muy pocos alumnos presentan estaturas menores a 1.50 metros o mayores a 1.85 metros. En este ejemplo nuestro histograma se asemeja mucho a una gráfica típica del área de estadística conocida como la distribución normal.

La **distribución normal** es una representación de los datos con una forma simétrica respecto a una medida de tendencia central (usualmente la media) que se encuentra muy comúnmente en la naturaleza. La distribución normal suele asociarse al siguiente par de reglas empíricas:

1. El 68% de los datos se encuentran a menos de una desviación estándar de la media.
2. El 95% de los datos se a menos de dos desviaciones estándar de la media.



La distribución normal, es capaz de representar muchos conjuntos de datos de diferentes áreas del conocimiento o de fenómenos de la naturaleza, pero no es la única distribución que existe. Los histogramas también son útiles para darnos una visión aproximada de la distribución que rige a nuestra población y nos permite construir nuestras propias reglas empíricas producto de la constante observación de los datos con los que trabajamos.



In [ ]:
```
import pandas as pd
```

Heart Disease UCI Cleveland
https://www.kaggle.com/ronitf/heart-disease-uci

In [ ]:
```
df = pd.read_csv("data/heart/heart.csv")
```

In [ ]:
```
df.head(7) # head por defecto trae 5 filas, pero se puede dar el parámetro entre paréntesis
```

In [ ]:
```
import matplotlib.pyplot as plt
# para que se vea en este archivo, cuaderno o libreta
%matplotlib inline
```

In [ ]:
```
# Gráfica de histograma con el dato de Colesterol
plt.hist(df['chol'])
plt.show()
```

In [ ]:
```
plt.hist(df['chol'],bins=20) # Repartir el eje X en 20 intervalos
plt.show()
```

In [ ]:
```
plt.hist(df['chol'],bins=5) # Ahora solo 5 secciones
plt.show()
```

In [ ]:
```
plt.hist(df['chol'],bins=20,histtype='step') # Histograma de paso (es la silueta)
plt.show()
```

In [ ]:
```
# Por defecto los histogramas son verticales, pero se puede cambiar la orientación
plt.hist(df['chol'],bins=20,orientation='horizontal')
plt.show()
```

In [ ]:
```
# Dos graficas en el mismo plano, esto es por estar en la misma celda
plt.hist(df['chol'],bins=20,alpha=0.5)
plt.hist(df['trestbps'],bins=10,alpha=0.5)
plt.show()
```

In [ ]:
```
## En el histograma anterior no se sabe que se esta graficando
# Se agregan etiquetas
plt.hist(df['chol'],bins=20,alpha=0.5,label='Colesterol')
plt.hist(df['trestbps'],bins=10,alpha=0.5,label='trestbps')
plt.legend()
```

Otra libreria para hacer histogramas.

Documentación en: https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

In [ ]:
```
# Esta libreria por defecto muestra la distribución, para el ejemplo se anula con el false
import seaborn as sns

# Anular la distribución
sns.histplot(df['chol'], kde=False)
```

In [ ]:
```
sns.histplot(df['chol'],kde=True)
```

In [ ]:
```
# Hacer 2 histogramas al mismo tiempo
sns.jointplot(x="thalach", y="chol", data=df) # Qué relacion tiene el ritmo cardiaco y el colesterol
```

Esta gráfica nos presenta en el eje 'Y' el colesterol y en el eje 'X'el máximo ritmo cardíaco.

Los puntos es donde más se relaciona una concentración entre ambas variables. En la sección de la izquierda y derecha son extremos no hay muchos datos porque no hay muchos pacientes que tengan una máxima de ritmo cardíaco cercano entre 80 y 100 tampoco hay muchos pacientes que tengan un colesterol que rebase los 400. Por ello en esa zona no hay datos, más hacia abajo y a la derecha ya empiezan a relacionarse.

Relacionando las variables ya podemos saber para qué colesteroles hay un máximo ritmo cardíaco y podemos empezar a relacionar, pero ver los datos con puntos, es un poco difícil de analizar.

In [ ]:
```
# Cambiar los puntos por hexágonos
sns.jointplot(x="thalach", y="chol",kind='hex', data=df)
```

Ahora ya se ve un poco más claro, por las regiones oscuras son las regiones donde más personas tienen un colesterol cercano a 220 y un ritmo cardíaco cercano a 150, esta región está muy cargada entonces muchas personas tienen esta misma coincidencia. Pocas personas andan en los extremos claros.

In [ ]:
```
# Agrega fomatos al gráfico
sns.set()
sns.jointplot(x="thalach", y="chol",kind='hex', data=df)
```

In [ ]:
```

```