Probabilidades y Estadística

Clase 2
Estadística

Nicolás Araya Caro

Universidad Diego Portales
Escuela de Informática y Telecomunicaciones

14 de septiembre de 2023

Table of Contents

Pequeño resumen

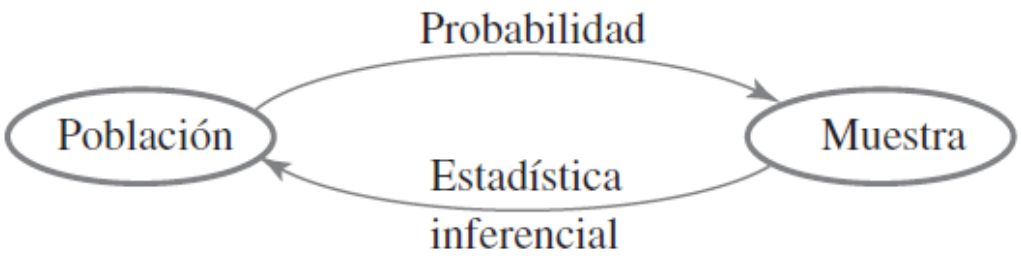


Figura 1.2 Relación entre probabilidad y estadística inferencial.

Recopilación de Datos

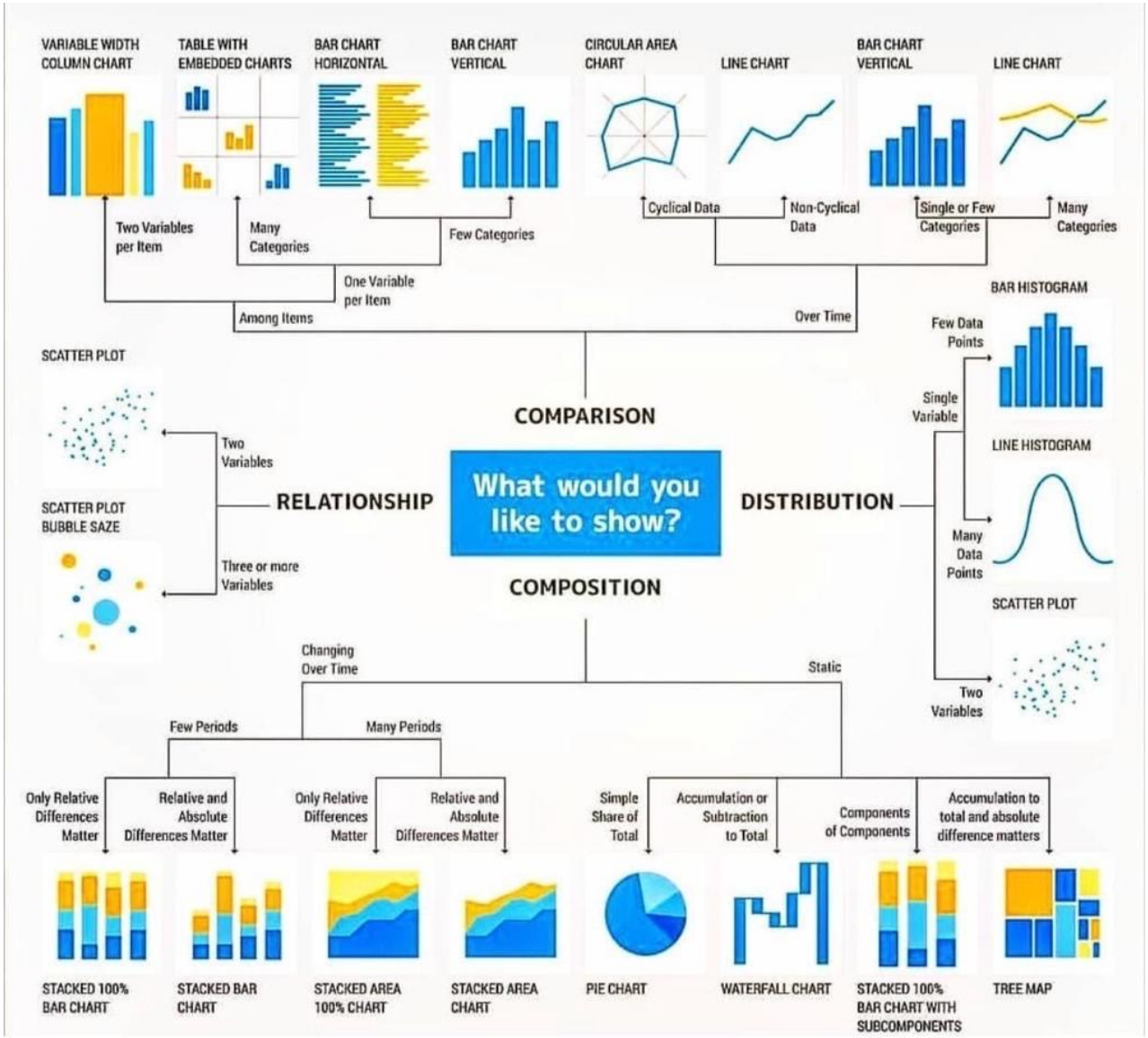
- La estadística se ocupa no sólo de la organización y análisis de datos una vez que han sido recopilados sino también con el desarrollo de técnicas de recopilación de datos. Si éstos no son apropiadamente recopilados, un investigador no puede ser capaz de responder las preguntas consideradas con un razonable grado de confianza.
- Cuando la recopilación de datos implica seleccionar individuos u objetos de un marco, el método más simple para garantizar una selección representativa es tomar una muestra aleatoria simple.
- En ocasiones se pueden utilizar métodos de muestreo alternativos para facilitar el proceso de selección, a fin de obtener información extra o para incrementar el grado de confianza en conclusiones (como el muestreo estratificado, sistemático etc).

Recordar: Variable

Una variable numérica es **discreta** si su conjunto de valores posibles es finito o se puede enumerar en una sucesión infinita (una en la cual existe un primer número, un segundo número, y así sucesivamente). Una variable numérica es **continua** si sus valores posibles abarcan un intervalo completo sobre la línea de números.

Table of Contents

Tipo de visualizaciones



Ejemplo Práctico

Desde aquí, usaremos el dataset diamonts.csv Para ejemplificar las visualizaciones. Este dataset consta de las siguientes variables:

1. Característica.
2. Corte.
3. Color.
4. Claridad.
5. Profundidad.
6. Tabla.
7. Precio.
8. X, Y, Z.

Table of Contents

Gráfica Tallos y Hojas

Considere un conjunto de datos numéricos $x_1, x_2, ..., x_n$ para el cual x_i se compone de por lo menos dos dígitos. Una forma rápida de obtener la representación visual informativa del conjunto de datos es construir una gráfica de tallos y hojas.
Pasos:

1. Seleccione uno o más de los primeros dígitos para los valores de tallo. Los segundos dígitos se convierten en hojas.
2. Enumere los posibles valores de tallos en una columna vertical.
3. Anote la hoja para cada observación junto al valor de tallo.
4. Indique las unidades para tallos y hojas en algún lugar de la gráfica

Conclusiones

Una gráfica de tallos y hojas da información sobre los siguientes aspectos de los datos:

- Identificar un valor típico o representativo.
- Grado de dispersión en torno al valor típico.
- Presencia de brechas en los datos.
- Grado de simetría en la distribución de los valores.
- Número y localización de crestas.
- Presencia de valores afuera de la gráfica.

Ejemplo

El consumo de alcohol por parte de estudiantes universitarios preocupa no sólo a la comunidad académica sino también, a causa de consecuencias potenciales de salud y seguridad, a la sociedad en su conjunto. El artículo "Health and Behavioral Consequences of Binge Drinking in College" presentó un amplio estudio sobre el consumo excesivo de alcohol en universidades a través de Estados Unidos. Un episodio de parranda se definió como cinco o más tragos en fila para varones y cuatro o más para mujeres. Se muestra una gráfica de tallo y hojas de 140 valores de:
$x =$ porcentaje de edades de los estudiantes de licenciatura bebedores.

ejemplo

0	4
1	1345678889
2	12234566667778899999
3	01122333445556666777778888999999
4	11122222334444556666667778888999
5	00111222233455666667777888899
6	01111244455666778

Tallo: dígitos de diez cifras
Hojas: dígitos de una cifra

Figura 1.4 Gráfica de tallo y hojas de porcentajes de bebedores en cada una de 140 universidades.

¿Qué Concluye?

Ejemplo Práctico

Aplicado al dataset diamonds.csv **(ver archivo .mlx)**:

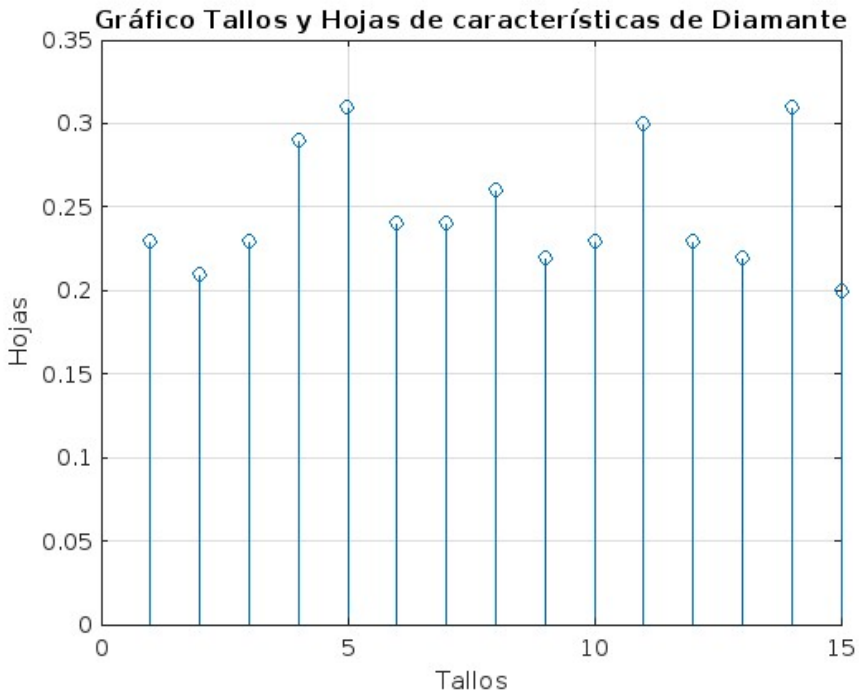


Figura: Ejemplo de diagrama Tallos y Hojas con matlab

¿Qué Concluye de este gráfico?

Table of Contents

Gráfica de Puntos

- Una gráfica de puntos es un resumen de datos numéricos cuando el conjunto de datos es razonablemente pequeño o existen pocos valores de datos distintos.
- Cada observación está representada por un punto sobre la ubicación correspondiente en una escala de medición horizontal.
- Cuando un valor ocurre más de una vez, existe un punto por cada ocurrencia y estos puntos se apilan verticalmente.
- Como con la gráfica de tallos y hojas, una gráfica de puntos da información sobre la localización, dispersión, extremos y brechas.

Ejemplo

La figura 1.6 muestra una gráfica de puntos para los datos de temperatura de diferentes sellos anulares. Un valor de temperatura representativo es uno que se encuentra entre la mitad de los 60 (°F) y existe poca dispersión en torno al centro. Los datos se alargan más en el extremo inferior que en el superior y la observación más pequeña, 31, apenas puede ser descrita como valor extremo.

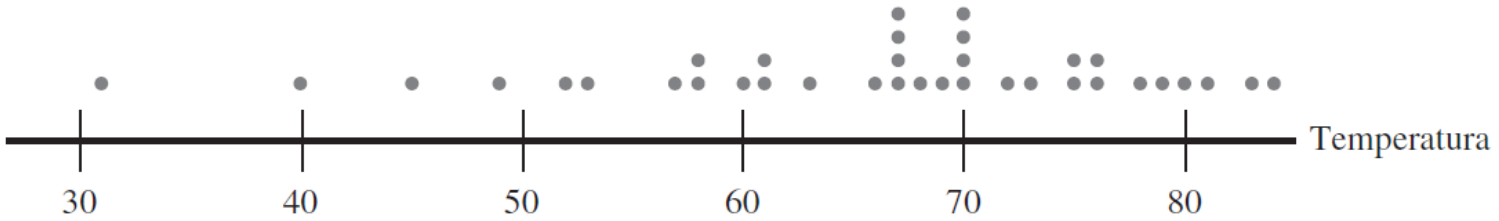


Figura 1.6 Gráfica de puntos de los datos de temperatura de los sellos anulares (°F).

¿Qué Concluye?

Table of Contents

Algunos datos numéricos se obtienen de estos casos:

1. Contando para determinar el valor de una variable (el número de citatorios de tráfico que una persona recibió durante el año pasado, el número de personas que solicitan empleo durante un periodo particular).
2. Tomando mediciones (peso de un individuo, tiempo de reacción a un estímulo particular).

La prescripción para trazar un histograma es en general diferente en estos dos casos.

Frecuencia

Una variable discreta **x** casi siempre resulta de contar, en cuyo caso posibles valores son 0, 1, 2, 3,... o algún subconjunto de estos enteros. De la toma de mediciones surgen variables continuas. Considérense datos compuestos de observaciones de una variable discreta x. La **frecuencia** de cualquier valor x particular es el **número de veces que ocurre un valor en el conjunto de datos**.

Tipos de Frecuencias

- **Frecuencia Absoluta (f_i):** número de veces que aparece un determinado valor en un estudio estadístico. (Sumatoria de la cantidad de ocurrencias de un determinado valor).
- **Frecuencia Relativa (n_i):** cociente entre la frecuencia absoluta de un determinado valor y el número total de datos (N). ($n_i = \frac{f_i}{N}$). La suma de las frecuencias relativas es igual a 1.
- **Frecuencia Acumulada (F_i):** suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado.
- **Frecuencia Relativa Acumulada (N_i):** cociente entre la frecuencia acumulada de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento.

"La distribución de frecuencias o tabla de frecuencias es una ordenación en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente".

Histograma

- **Construcción de un histograma para datos discretos:** En primer lugar, se determina la frecuencia y la frecuencia relativa de cada valor x. Luego se marcan los valores x posibles en una escala horizontal. Sobre cada valor, se traza un rectángulo cuya altura es la frecuencia relativa (o alternativamente, la frecuencia) de dicho valor.
- **Construcción de un histograma para datos continuos (anchos de clase iguales):** Se determina la frecuencia y la frecuencia relativa de cada clase (subdividir el eje de medición en un número adecuado de intervalos de clase o clases). Se marcan los límites de clase sobre un eje de medición horizontal. Sobre cada intervalo de clase, se traza un rectángulo cuya altura es la frecuencia relativa correspondiente (o frecuencia).

Nota: Del intervalo de clase se puede obtener la marca de clase, la cual consiste en la media de los limites del intervalo.

Formas de Histograma

1. **Unimodal**: presenta una sola cresta y luego declina.
2. **Bimodal**: tiene dos crestas diferentes (con más de dos crestas es multimodal).

Un histograma es **simétrico** si la mitad izquierda es una imagen de espejo de la mitad derecha. Un histograma Bimodal es positivamente **asimétrico** si la cola derecha o superior se alarga en comparación con la cola izquierda o inferior y negativamente asimétrico si el alargamiento es hacia la izquierda.

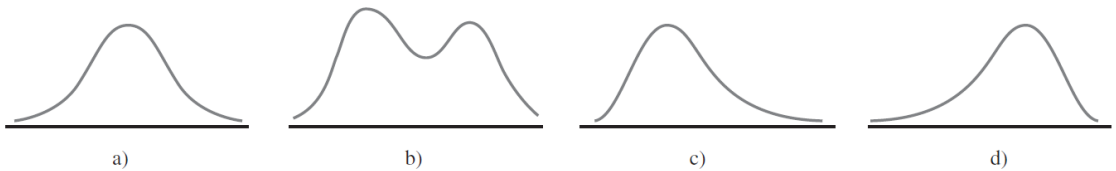


Figura 1.11 Histogramas alisados: a) unimodal simétrico; b) bimodal; c) positivamente asimétrico y d) negativamente asimétrico.

Ejemplo Discreto

¿Qué tan inusual es un juego de béisbol sin hit o de un hit en las ligas mayores y cuán frecuentemente un equipo pega más de 10, 15 o incluso 20 hits? La tabla 1.1 es una distribución de frecuencia del número de hits por equipo por juego de todos los juegos de nueve episodios que se jugaron entre 1989 y 1993.

Ejemplo

Tabla 1.1 Distribución de frecuencia de hits en juegos de nueve episodios

Hits/juego	Número de juegos	Frecuencia relativa	Hits/juego	Número de juegos	Frecuencia relativa
0	20	0.0010	14	569	0.0294
1	72	0.0037	15	393	0.0203
2	209	0.0108	16	253	0.0131
3	527	0.0272	17	171	0.0088
4	1048	0.0541	18	97	0.0050
5	1457	0.0752	19	53	0.0027
6	1988	0.1026	20	31	0.0016
7	2256	0.1164	21	19	0.0010
8	2403	0.1240	22	13	0.0007
9	2256	0.1164	23	5	0.0003
10	1967	0.1015	24	1	0.0001
11	1509	0.0779	25	0	0.0000
12	1230	0.0635	26	1	0.0001
13	834	0.0430	27	1	0.0001
				19 383	1.0005

Ejemplo

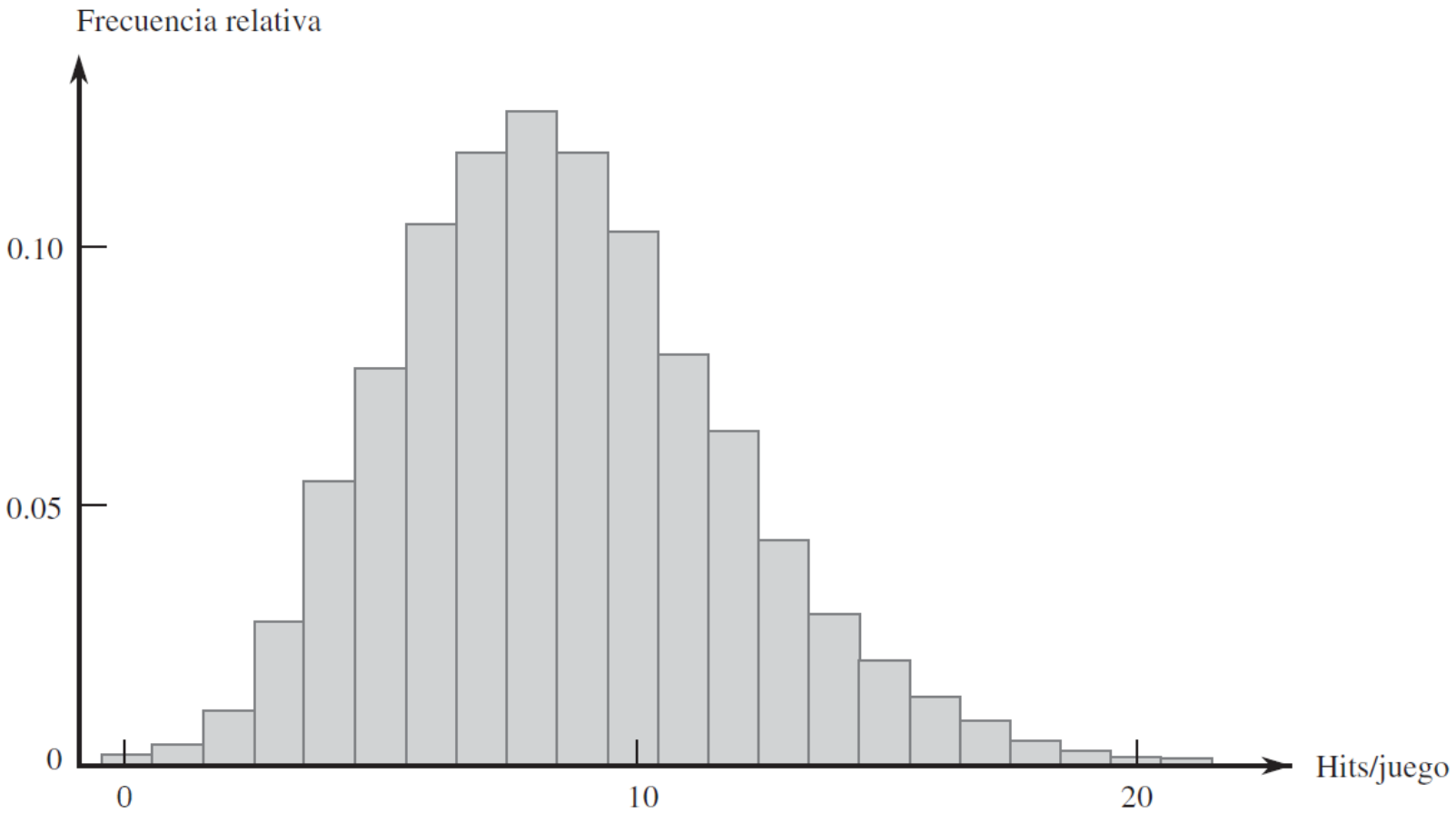


Figura 1.7 Histograma de número de hits por juego de nueve episodios.

Ejemplo Práctico continuo

Aplicado al dataset diamonds.csv **(ver archivo .mlx)**:

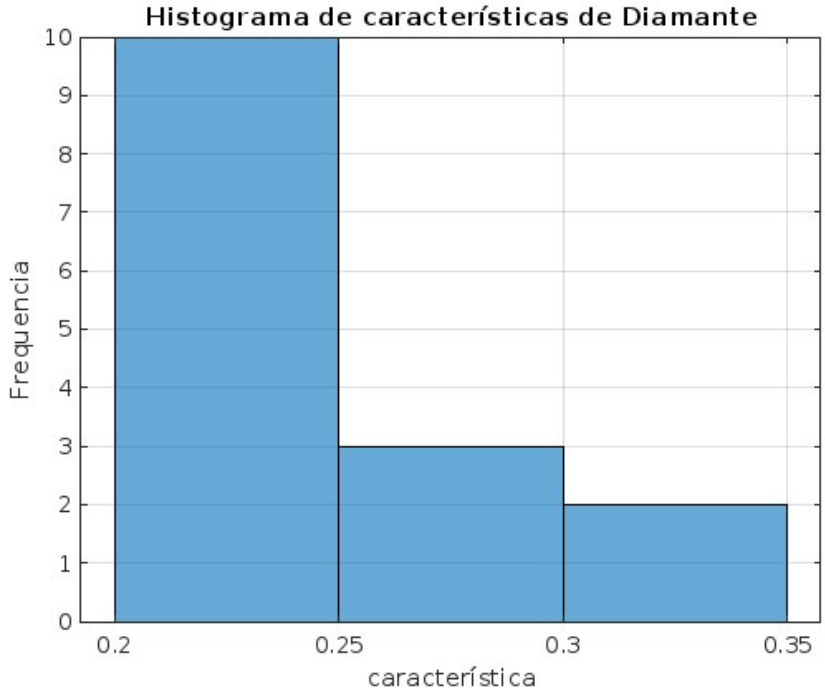


Figura: Ejemplo de Histograma con matlab

¿Qué Concluye de este gráfico?

Table of Contents

Table of Contents

Media

La **media muestral** $\bar{x}$ de las observaciones $x_1, x_2, ..., x_n$ está dada por:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

En cambio, la **media poblacional** μ (con N como el tamaño de la población) se define por:

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

Nota spoiler: más adelante veremos que es posible inferir en base a la media muestral $\bar{x}$, la media de la población μ

Mediana

La **mediana muestral** $\tilde{x}$ se obtiene ordenando primero las n observaciones de la más pequeña a la más grande (con cualesquiera valores repetidos incluidos de modo que cada observación muestral aparezca en la lista ordenada).

$$\tilde{x} = \begin{cases} \text{El valor medio único si } n \text{ es impar} & = \left(\frac{n+1}{2}\right)^{\text{n-ésimo}} \text{ valor ordenado} \\ \text{El promedio de los dos valores medios si } n \text{ es par} & = \text{promedio de } \left(\frac{n}{2}\right)^{\text{n-ésimo}} \text{ y } \left(\frac{n}{2}+1\right)^{\text{n-ésimo}} \text{ valores ordenados} \end{cases}$$

Nota: Existe una mediana poblacional $\tilde{\mu}$, pero no la veremos :)

Moda

La Moda es Aquel valor que se presente con mayor frecuencia (bueh).

Percentiles

La mediana divide el conjunto de datos en dos partes iguales. Para obtener medidas de ubicación más finas, se podrían dividir los datos en más de dos partes como por ejemplo:

- **Cuartiles:** Divide el conjunto en 4 partes iguales.
- **Deciles:** Divide el conjunto en 10 partes iguales.
- **Percentiles:** Divide el conjunto en 100 partes iguales.

Percentiles

¿Lo recuerdan?

Table of Contents

Estadigrafo

- Las medidas principales de variabilidad (dispersión) implican las desviaciones de la media, $(x_1 - \bar{x}), (x_2 - \bar{x}), ..., (x_n - \bar{x})$. Es decir, las desviaciones de la media se obtienen restando $\bar{x}$ de cada una de la n observaciones muestrales.
- Una desviación será positiva si la observación es más grande que la media (a la derecha de la media sobre el eje de medición) y negativa si la observación es más pequeña que la media. Si todas las desviaciones son pequeñas en magnitud, entonces todas las x_i se aproximan a la media y hay poca variabilidad.
- Alternativamente, si algunas de las desviaciones son grandes en magnitud, entonces algunas x_i quedan lejos de $\bar{x}$ lo que sugiere una mayor cantidad de variabilidad.

¿Cómo se puede evitar que las desviaciones negativas y positivas se neutralicen entre sí cuando se combinan? puede ser con valor absoluto, o elevar al cuadrado.

Varianza muestral

La varianza muestral, denotada por s^2 está dada por:

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

La **desviación estándar muestral**, denotada por s, es la raíz cuadrada (positiva) de la varianza:

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

Una alternativa para el numerador (S_{xx}) de s^2 es:

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Nota spoiler: más adelante veremos que es posible inferir en base a la varianza muestral s^2, la varianza de la población σ^2.

Varianza poblacional

La **varianza poblacional**, denotada por σ^2 está dada por:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{S_{xx}}{N}$$

La **desviación estándar poblacional**, denotada por σ, es la raíz cuadrada (positiva) de la varianza:

$$\sigma = \sqrt{\sigma^2}$$

Una alternativa para el numerador (S_{xx}) de σ^2 es:

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{N}$$

Rango

Rango: diferencia numérica entre el valor máximo y el valor mínimo (bueh).

Coef de variación

El coeficiente de variación, también denominado como coeficiente de variación de Pearson, es una medida estadística que nos informa acerca de la dispersión relativa de un conjunto de datos. A nivel muestral:

$$CV = \frac{s}{\bar{x}} \cdot 100$$

a nivel poblacional:

$$CV = \frac{\sigma}{\mu} \cdot 100$$

Gráfica de caja

Las gráficas de tallo y hojas e histogramas transmiten impresiones un tanto generales sobre un conjunto de datos, mientras que un resumen único tal como la media o la desviación estándar se enfoca en sólo un aspecto de los datos. En años recientes, se ha utilizado con éxito un resumen gráfico llamado gráfica de caja para describir varias de las características más prominentes de un conjunto de datos. Estas características incluyen:

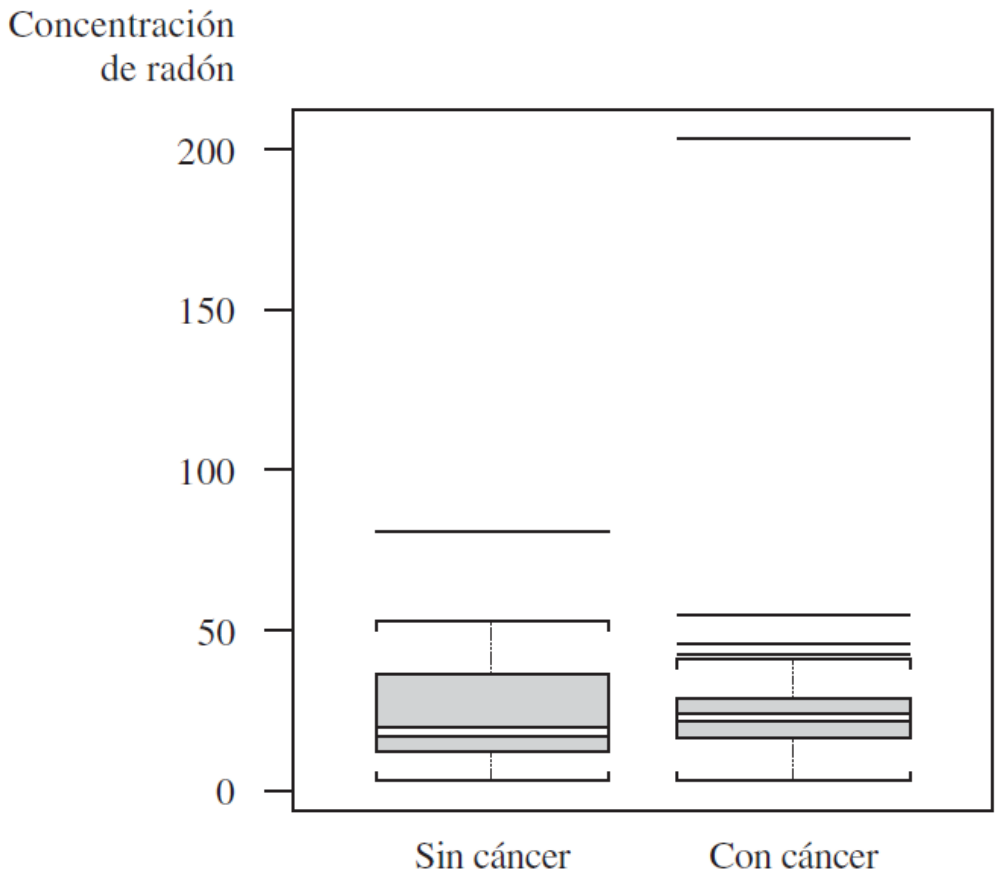
1. el centro
2. la dispersión
3. el grado y naturaleza de cualquier alejamiento de la simetría
4. la identificación de las observaciones "extremas o apartadas" inusualmente alejadas del cuerpo principal de los datos.

Gráfico de caja

Se ordenan las observaciones de la más pequeña a la más grande y se separa la mitad más pequeña de la más grande; se incluye la mediana $\tilde{x}$ en ambas mitades si n es impar. En tal caso el cuarto inferior es la mediana de la mitad más pequeña y el cuarto superior es la mediana de la mitad más grande. Una medida de dispersión que es resistente a los valores apartados es la dispersión de los cuartos fs, dada por fs = cuarto superior – cuarto inferior

Ejemplo

En años recientes, algunas evidencias sugieren que las altas concentraciones de radón bajo techo pueden estar ligadas al desarrollo de cánceres en niños, pero muchos profesionales de la salud aún no están convencidos. Un artículo reciente "Indoor Radon and Childhood Cancer" presentó los datos adjuntos sobre concentración de radón (Bq/m3) en dos muestras diferentes de casas. La primera consistió en casas en las cuales un niño diagnosticado con cáncer había estado residiendo. Las casas en la segunda muestra no incluían casos registrados de cáncer infantil.



Gráfica de caja de los datos del ejemplo 1.19, obtenida con S-Plus.

Ejemplo matlab

Aplicado al dataset diamonds.csv **(ver archivo .mlx)**:

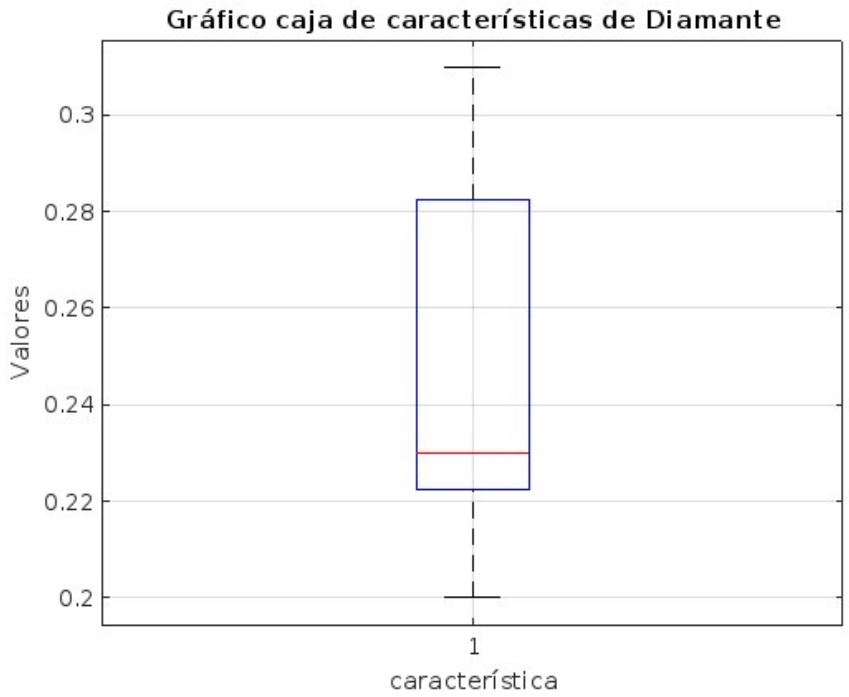


Figura: Ejemplo de Gráfico de caja con matlab

¿Qué Concluye de este gráfico?

Table of Contents

Ejercicio 1

Se desea determinar cual es la masa crítica antes de que los servidores colapsen. Al averiguar, ud. recopila datos de cuantas conexiones concurrentes entrantes hubo, en años pasados, antes de que los servidores se cayeran. Los datos que se recopilaron fueron [13, 19, 10, 19, 16, 20, 20, 15, 17, 16], en cientos de conexiones (por ejemplo, 19 implica 1900 conexiones). Calcule la media, moda y mediana de estos datos, junto a la varianza , desviación estándar y realice tabla de frecuencias.

Resolución

1. Ordenar los datos [10, 13, 15, 16, 16, 17, 19, 19, 20, 20]
2. calcular moda: en este caso hay 3 datos mas repetidos que son 16, 19 y 20 (trimodal).
3. mediana: hay 10 elementos, por lo que la mediana es el promedio entre los valores ubicados en las posiciones $x_5 = 16$ y $x_6 = 17$: $\tilde{x} = \frac{16+17}{2} = 16,5$
4. media aritmética: $\bar{x} = \frac{10+13+15+...+20+20}{10} = 16{,}5$
5. varianza: $s^2 = \frac{\sum_{i=1}^{10}(x_i - 16{,}5)^2}{9} = 9{,}45$
6. Desviación estándar: $s = \sqrt{s^2} = 3{,}074$

Tabla de Frecuencias

x_i	Absoluta f_i	Relativa n_i	Acumulada F_i	Relativa Acumulada N_i
10	1	0.1	1	0.1
13	1	0.1	2	0.2
15	1	0.1	3	0.3
16	2	0.2	5	0.5
17	1	0.1	6	0.6
19	2	0.2	8	0.8
20	2	0.2	10	1

Cuadro: Tabla de frecuencias

ejercicio 2

Un empresa de soporte técnico informático decide medir el desempeño de sus trabajadores en función de la cantidad de consultas resueltas en un mes. El resumen del ultimo mes entrego lo siguiente:

Cantidad de consultas resueltas	Frecuencia
$0 - 10$	2
$10 - 20$	9
$20 - 30$	12
$30 - 40$	3

La empresa, en su preocupación por sus trabajadores, espera que la cantidad de consultas resueltas sea la misma para cada uno, de tal manera de que no existan trabajadores estresados por la carga laboral. Para esto la empresa esta pensando en realizar cambios en la asignación de consultas, si y solo si, la variabilidad relativa de la cantidad de consultas resueltas por trabajador supera el 25 %. Demuestre si es que la empresa debe o no implementar el cambio mencionado.

Resolución

Para poder encontrar la variabilidad relativa para responde a la duda de la empresa debemos obtener la media y la desviación estándar. Para esto, extenderemos la tabla para encontrar los valores de las sumas de la siguiente forma:

Cantidad de consultas resueltas	**Frecuencia**	X_i	$X_i \cdot f_i$	$X_i^2 \cdot f_i$
$0 - 10$	2	5	10	50
$10 - 20$	9	15	135	2025
$20 - 30$	12	25	300	7500
$30 - 40$	3	35	105	3675

Resolución

con esto podemos obtener los estadisticos de la siguiente forma:

- Media: $\mu = \frac{\sum_{i=1}^{4} x_i \cdot f_i}{N} = \frac{550}{26} = 21{,}15$ *consultas*
- Desviación Estándar: $\sigma = \sqrt{\frac{\sum_{i=1}^{4} x_i^2 \cdot f_i}{N} - \mu^2} = \sqrt{62{,}13} = 7{,}88$
- $CV = \frac{7{,}88}{21{,}15} \cdot 100 = 37{,}26\,\%$

Finalmente podemos decir que la muestra de los 26 trabajadores respecto a la cantidad de consultas resueltas tiene una variabilidad relativa del 37.26 %, la cual supera el 25 % requerido por la empresa, por lo que la empresa si debería implementar el cambio en la asignación de las consultas.

ejercicio 3

El imperio ha sido muy meticuloso en sus censos. De cierto sistema planetario ha obtenido los siguientes datos para cantidad de habitantes de todos sus planetas (expresado en miles de millones de habitantes): {4, 11, 4, 3, 8, 8, 3, 4, 12, 9}
Estadígrafos de posición: Calcule moda, media y mediana de los datos entregados. Estadígrafos de dispersión: Calcular varianza , desviación estándar y coeficiente de variación de los datos entregados. Realice Histograma