Estadística Descriptiva

Oscar R. Diaz **19/02/2016** Capítulo 1

Introducción al Análisis Exploratorio de Datos.

Objetivos de aprendizaje:

Después de completar este capítulo usted debería ser capaz de:

A nivel de conocimiento:

1. Definir la estadística descriptiva y la estadística inferencial.
2. Definir una variable.
3. Definir qué es una medición.
4. Definir variables cualitativas y variables cuantitativas.
5. Definir variables discretas y continuas.
6. Identificar las escalas de medición nominal, ordinal, de escala y de razón.

A nivel de comprensión:

7. Organizar datos usando una distribución de frecuencias.
8. Seleccionar el gráfico adecuado para una serie dada de datos.

A nivel de aplicación:

9. Construir diagramas de barras y diagramas de barras comparativos.
10. Construir histogramas.

A nivel de análisis

11. Distinguir las principales características de una distribución de datos a partir de la distribución de frecuencias.
12. Distinguir las principales características de una distribución de datos a partir de un diagrama de barras.
13. Distinguir las principales características de una distribución de datos a partir de un histograma.
14. Usar diagramas de barras o histogramas para comparar dos o más series de datos.

Contenidos:

1. Introducción.
2. Estadística Descriptiva e Inferencial.
3. Variables y sus medidas.
4. Escalas de Medición.
5. Distribuciones de frecuencia y gráficos para variables cualitativas.
6. Distribuciones de frecuencia y gráficos para variables cuantitativas.

«A judicious man looks at statistics, not to get knowledge but to save himself from having ignorance foisted on him». (Thomas Carlyle)

1. Introducción.
¿Por qué estudiar Estadística?

En ingeniería es común trabajar con datos que provienen de mediciones. Estas mediciones se hacen generalmente bajo situaciones controladas de laboratorio con el propósito de tomar decisiones.

Un estudiante de ingeniería necesita un curso de probabilidad y estadística entre otras razones por:

1. Como profesional, deberá ser capaz de leer y entender estudios estadísticos realizados en su campo laboral. Para lograr este entendimiento, deberá conocer el vocabulario, símbolos, conceptos y procedimientos estadísticos usados en dichos estudios.
2. Como profesional también deberá ser capaz de dirigir investigaciones en su campo laboral, donde los procedimientos estadísticos como el diseño de experimentos, la recolección, la organización, el análisis y el resumen de datos le serán de mucha utilidad para hacer pronósticos confiables para el futuro. También deberá ser capaz de comunicar los resultados del estudio en sus propias palabras.
3. Por último, pero no menos importante, como profesional puede usar los conocimientos en probabilidad y estadística para ser un ciudadano con un pensamiento crítico que analice y tome posturas ante los problemas de la realidad nacional

2. Estadística Descriptiva e Inferencial

Nota histórica

El origen de la estadística descriptiva puede ser rastreado hasta los métodos de recolección de datos usados en los censos realizados por los babilonios y egipcios entre los años 4500 y 3000 A. C.

El emperador romano Augusto (27 A. C. – 17 D. C.) realizó censos de nacimientos y muertes de los ciudadanos romanos, así como de sus animales y cultivos anuales.

En ingeniería es común trabajar con datos que provienen de mediciones, que han sido tomadas bajo condiciones controladas, con el propósito de tomar alguna decisión. Por lo general la cantidad de datos es voluminosa y sin ningún significado a primera vista, por lo que se hace necesario procesar los datos a fin de que puedan proporcionar la información requerida por el usuario para la toma de decisiones.

El primer paso es introducir el concepto de orden y proceder a ordenar los datos. Luego de ordenar hay que resumir los datos de manera que puedan ser manipulados de una manera sencilla y ágil. Esto se logra por medio de métodos y técnicas estadísticas que generan diferentes tipos de resultados como gráficas, tablas o cifras que el tomador de decisiones usará para describir la situación de donde provienen los datos.

Según el tipo de estudio que se haga el razonamiento estadístico tiene dos grandes ramas:

- La estadística descriptiva
- La estadística inferencial

La estadística descriptiva se encarga de responder preguntas como: de los estudiantes de nuevo ingreso para el año 2000 de la Universidad de El Salvador ¿qué proporción eran hombres? ¿qué proporción eran mujeres? ¿cuántos de ellos se graduaron en el tiempo reglamentado? Después de graduados ¿cuántos obtuvieron trabajo en menos de un año? ¿cuántos de ellos estudiaron una maestría? Todas estas son preguntas para las cuales la Estadística Descriptiva nos proporciona herramientas para caracterizar, de manera conveniente, los datos en estudio de tal modo que tengan significado para el usuario y le permitan extraer información válida para la toma de decisiones.

El propósito de la Estadística Descriptiva es organizar y resumir datos de manera que resulte más fácil su comprensión.

La segunda rama de la estadística, la estadística inferencial, trata de hacer generalizaciones relativas a poblaciones — la totalidad de los elementos de interés— a partir de los resultados obtenidos por la observación de relativamente pocos elementos de la población (muestra) realizando estimaciones y pruebas de hipótesis, determinando relaciones entre variables y haciendo predicciones.

3. Variables y sus medidas.

En general, la estadística descriptiva trata de describirnos una situación. Esta situación se describe por medio de variables. En esta sección estudiaremos la naturaleza de las variables y los tipos de datos. Iniciamos definiendo el concepto de variable.

Una variable es una característica o atributo que puede asumir diferentes valores

Por ejemplo, 18 años puede ser la medición para la característica *edad* para una persona en particular. En una escala del 1 al 3, 2 podría significar el grado de satisfacción de ésta persona, y si hablamos del sexo de esta persona, arbitrariamente "1" indicaría que es una mujer.

El análisis estadístico no es posible sin números, y no puede haber números sin un proceso de **medición**.

Una medición es el proceso de asignar números a la característica que se desea estudiar

Las variables se pueden clasificar como cualitativas o cuantitativas. En las **variables cualitativas** (también conocidas como categóricas) la medición describe un elemento colocándolo en una categoría o grupo de acuerdo a alguna característica o atributo. Por ejemplo, si clasificamos personas de acuerdo al sexo (masculino o femenino) entonces, la variable *género* es cualitativa. Otros ejemplos de estas variables serían la preferencia religiosa, el estado civil, la afiliación política, la etnia y la localización geográfica.

Los valores de las variables cualitativas y cuantitativas difieren en el *tipo* más que en la *cantidad*. El sexo es un buen ejemplo. A pesar de que hombres y mujeres son claramente diferentes en la función reproductiva (una distinción cualitativa), esto no implica que un género sea "mayor que" o "menor que" el otro (una distinción cuantitativa)

En las **variables cuantitativas** las mediciones resultan en valores numéricos que podemos ordenar y realizar operaciones aritméticas con ellos. Por ejemplo, la variable *edad* es una variable numérica y las personas pueden ser ordenadas de acuerdo a su edad. Otros ejemplos de este tipo de variables son el peso de la persona, la altura y la temperatura corporal.

Las variables cuantitativas pueden ser clasificadas en dos grupos: *discretas* y *continuas.* las variables discretas pueden asumir valores como 0, 1, 2, 3 y se dice que son contables. Por ejemplo, el número de niños de una familia, el número de estudiantes en el salón de clase y el número de llamadas que recibe un operador cada día durante un mes.

Las variables discretas asumen valores que pueden ser contados

Las variables continuas pueden asumir un número infinito de valores en un intervalo entre dos valores específicos. Por ejemplo la temperatura, ya que la variable puede asumir un infinito número de valores entre dos temperaturas dadas.

Las variables continuas pueden asumir un infinito número de valores entre dos valores específicos.

4. Escalas de medición.

¿por qué son importantes las escalas de medición?

Para entender y usar apropiadamente las diferentes técnicas del análisis estadístico, es necesario identificar previamente la escala de medición correspondiente, ya que cada escala tiene sus propiedades matemáticas, que determinan el análisis estadístico apropiado en cada caso, de manera que los datos se puedan explorar convenientemente, organizarlos, resumirlos y presentarlos.

Además de clasificarse como cualitativas o cuantitativas, las variables pueden ser clasificadas por cómo son categorizadas, contadas o medidas. En 1946 S. S. Steven publicó el artículo «On the Theory of Scales of Measurement», en el cual introdujo un esquema muy elaborado para la clasificación de variables. Steven propuso que una variable puede ser clasificada en una de cuatro escalas: *nominal, ordinal, de intervalo y de razón*. Desde el punto de vista de las propiedades matemáticas y estadísticas, la escala de medición más rudimentaria es la nominal y la más completa la de razón.

Escala Nominal: En esta categoría la característica o variable de interés consiste en clases mutuamente excluyentes según determinada propiedad. Además, no existe un orden lógico particular para las distintas clasificaciones o categorías que permita, por ejemplo, ordenarlas; es decir que los números en esta escala sólo se usan como identificadores o nombres. Por ejemplo, si nuestro estudio incluye una variable *sexo,* codificamos femenino como 1 y masculino como 2. Pero los números 1 y 2 representan categorías de datos: son simples identificadores de una cualidad que se está midiendo y son completamente arbitrarios ya que puede usarse F o M o cualquier otra alternativa, para la codificación. A este nivel la operación matemática permitida es el *conteo.*

Escala Ordinal: Posee todas las características de la escala nominal, pero además los datos o mediciones en una escala ordinal pueden ser colocados en categorías que pueden ordenarse, de manera que reflejen diferentes grados o cantidades de la característica bajo estudio. Los números representan una cualidad que se está midiendo, y expresan si una observación tiene más de la cualidad medida que otra. Por ejemplo, un estudiante de inglés puede ser clasificado como «básico», «intermedio» o «avanzado» que codificamos con 1, 2 y 3 respectivamente. En este caso 3 indica que una persona está más avanzada que un 2 o que un 1. Sin embargo, note como en esta escala —por la falta de una unidad de medida común— no se puede distinguir las diferencias entre las categorías. ¿Es la diferencia entre «básico» e «intermedio» la misma que entre «intermedio» y «avanzado»? No se puede saber, hay un cierto orden, pero no una cantidad mensurable. Además del conteo, en esta categoría se pueden *ordenar* los datos.

Escala de Intervalo: Esta posee todas las características de la escala ordinal, con la propiedad adicional de que las mediciones son generalmente números y la diferencia entre un par de ellos da un resultado significativo debido a la existencia de una unidad de medida común y constante. Una limitante de esta escala es que

carece de un punto inicial o de referencia natural que indique la ausencia de atributo. Por ejemplo, en el caso de un termómetro cuyas lecturas son medidas en grados Celsius, el cero de esta escala (0°C) es arbitrariamente fijado al punto al cual el agua se congela (a nivel del mar). En contraste, la ausencia de calor (la temperatura a la cual la actividad molecular cesa) es aproximadamente -273 °C. Como consecuencia, no podemos decir que 0 °C indique la ausencia de calor. Esta falta de un cero natural impide establecer que un día con 30 °C sea tres veces más caliente que uno con 10 °C, pero si podemos decir que la distancia entre 25° y 30° es la misma que la existente entre 20° y 25°. En esta escala las diferencias y las sumas de datos tienen un significado numérico racional, pero no la multiplicación y división.

Escala de Razón: esta escala tiene todas las características de la escala de intervalo, pero además, tiene un punto cero natural que indica la ausencia del atributo. Esto tiene como consecuencia que además de las operaciones lógicas de ordenación y comparación, las diferencias y las sumas, la división y multiplicación de datos tenga un sentido numérico racional. Por ejemplo, el salario. Cero salario indica la ausencia de la característica medida (no hay dinero). Por consiguiente, si alguien gana mensualmente $ 1000 podemos decir que gana el doble de alguien que gana $ 500 en el mismo periodo.

La tabla 1 resume las principales características de las escalas.

Tabla 1.
Principales características y Propiedades de las Escalas de Medición

Escala de medición.	Propiedad del sistema numérico.	Operación matemática permitida.	Ejemplos
Nominal	Identidad	Conteo	Sexo
Ordinal	Magnitud	Ordenar	Nivel educativo
Intervalo	Distancia	Suma y resta	Temperaturas
Razón	Cero absoluto	Multiplicación y división.	Peso,

Nota: las propiedades y operaciones de una escala incluye las de la escala anterior.

5. Distribuciones de Frecuencia y Gráficos para Variables Cualitativas.

En esta sección veremos cómo las distribuciones de frecuencia y los gráficos de barra pueden ser usados para resumir datos categóricos. Usaremos éstos tipos de gráficos cuando el propósito sea mostrar *la distribución de los datos*, es decir *cómo se distribuyen los datos a través de las diferentes categorías que pueden observarse*, o comparar visualmente dos o más series de datos según alguna variable de interés.

Para construir la distribución de frecuencias y el gráfico de barras, siga los siguientes pasos:

1. Liste las categorías que componen las variables. Para evitar algún tipo de sesgo, lístelas alfabéticamente, o en orden descendente de frecuencia.
2. Registre la frecuencia asociada f con cada categoría y, si lo desea, su correspondiente frecuencia relativa. Incluya además el número total de datos, n, al final de la tabla.
3. Construya el gráfico con la información del paso 2.

Ejemplo 1: 25 estudiantes de ingeniería se sometieron a una prueba para determinar su grupo sanguíneo. Los resultados fueron los siguientes:

A	B	B	AB	O
O	O	B	AB	B
B	B	O	A	O
A	O	O	O	AB
AB	A	O	B	A

Construya una distribución de frecuencias y un gráfico de barras para los datos.

Solución:

Como los datos son categóricos, debemos usar clases discretas. De los datos identificamos que hay cuatro tipos sanguíneos: A, B, AB y O que pueden ser utilizados como las clases para la distribución de frecuencias. El procedimiento se muestra a continuación.

Paso 1: Construya una tabla como la siguiente

Tabla 2: Distribución de Frecuencias para los grupos sanguíneos de los estudiantes de ingeniería de la Universidad de El Salvador.

Clases	Conteo	Frecuencia f	Frecuencia relativa f_r
A			
AB			
B			
O			
	Total		

Paso 2: Cuente los datos y coloque los resultados en la segunda columna (conteo)

Paso 3: Traslade en forma numérica éstos resultados a la tercera columna (frecuencia)

Paso 4: Calcule la frecuencia relativa para cada una de las categorías usando la siguiente fórmula:

$$f_r = \frac{f}{n} \cdot 100\%$$

Paso 5: Calcule los totales para las columnas de frecuencia y frecuencia relativa.

La tabla completa se muestra a continuación.

Nota: las frecuencias relativas pueden escribirse tanto en valor decimal como en porcentaje, dependiendo de qué información quiera mostrar.

Tabla 3: Distribución de Frecuencias para la Variable Grupos Sanguíneos

Clases	Conteo	Frecuencia f	Frecuencia relativa (%) f_r
A	~~IIII~~	5	20
AB	IIII	4	16
B	~~IIII~~ II	7	28
O	~~IIII~~ IIII	9	36
	Total	***n*=25**	**100**

Con la información de las frecuencias podemos elaborar fácilmente el gráfico de barras, que no es más que una forma de mostrar gráficamente variables categóricas. Cada categoría es representada por un rectángulo o barra, de manera que la altura de cada barra sea proporcional a la frecuencia o frecuencia relativa correspondiente. En las figuras 1 y 2 se muestran éstos gráficos.

Figura 1: Gráfico de barras para la variable Grupo Sanguíneo

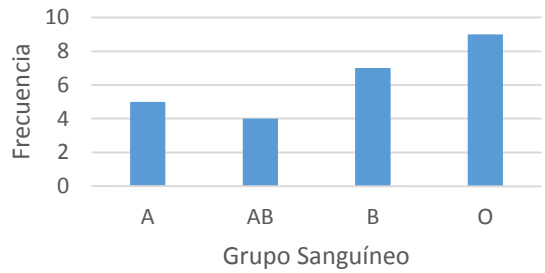
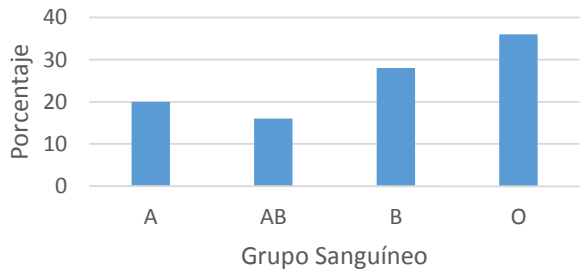


Figura 2: Gráfico de barras para la variable Grupo Sanguíneo.



Un aspecto muy importante a considerar cuando mostramos gráficamente el comportamiento de una variable es el propósito del gráfico. Para datos con una sola variable (univariados) generalmente el propósito es mostrar *la distribución de datos.* Para datos categóricos, como se menciona al inicio de esta sección, esto significa *mostrar cómo las observaciones se distribuyen a través de las distintas categorías posibles para la variable.*

Entonces ¿qué podemos concluir a partir de nuestros gráficos? Del gráfico resulta fácil ver que el tipo de sangre O ocurre con mayor frecuencia en los datos, más del doble que el tipo AB que es el que ocurre con menor frecuencia. El tipo B es el segundo con mayor ocurrencia entre los estudiantes.

Gráfico de Barras Comparativos.

Los gráficos de barras también pueden ser usados para comparar visualmente dos o más grupos. Esto se logra construyendo dos o más barras usando el mismo par de ejes horizontal y vertical.

Ejemplo 2: El artículo «2009 College Hopes & Worries Survey Findings» incluye un resumen de cómo 12,715 estudiantes de secundaria respondieron a la pregunta «Idealmente, ¿Qué tan lejos de su casa le gustaría que estuviera la Universidad a la que asistirá?» Los estudiantes respondieron seleccionando una de cuatro alternativas posibles para la distancia. El artículo también incluye las respuestas de 3007 padres de familia a la pregunta ¿Qué tan lejos de casa le gustaría que estuviera la Universidad a la que asistirán sus hijos? La información se muestra en la siguiente tabla.

	Frecuencia		**Frecuencia Relativa (%)**	
	Estudiantes	**Padres**	**Estudiantes**	**Padres**
Menos de 250 millas	4450	1594	35.00	53.01
De 250 a 500 millas	3942	902	31.00	30.00
De 500 a 1000 millas	2416	331	19.00	11.01
Más de 1000 millas	1907	180	15.00	5.99
Total:	**12715**	**3007**	**100%**	**100%**

Construya un diagrama de barras comparativo para la distancia ideal desde casa para estudiantes y padres de familia y comente sus aspectos más importantes.

Solución: Cuando se construye un gráfico de barras comparativo, usaremos las frecuencias relativas en lugar de las frecuencias absolutas para construir la escala del eje vertical, ya que esto nos permitirá hacer comparaciones que tengan sentido, aunque los tamaños de muestra sean diferentes. En la figura 3 se muestra un posible arreglo conocido como *gráfica de barras agrupado*, y en la figura 4 un arreglo conocido como *grafico de barras apilado*.

Figura 3: Gráfico de Barras Comparativo para la Distancia Ideal desde Casa.

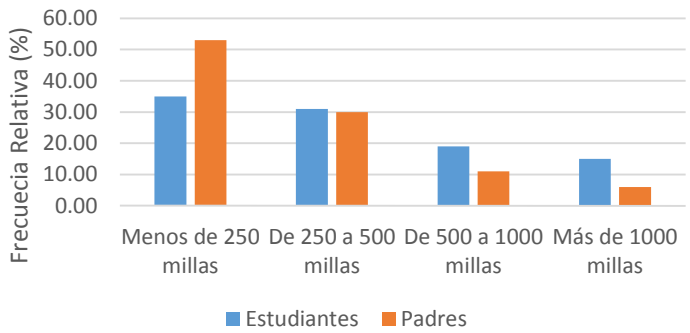
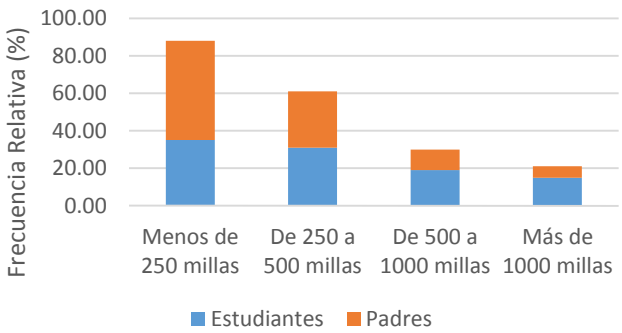
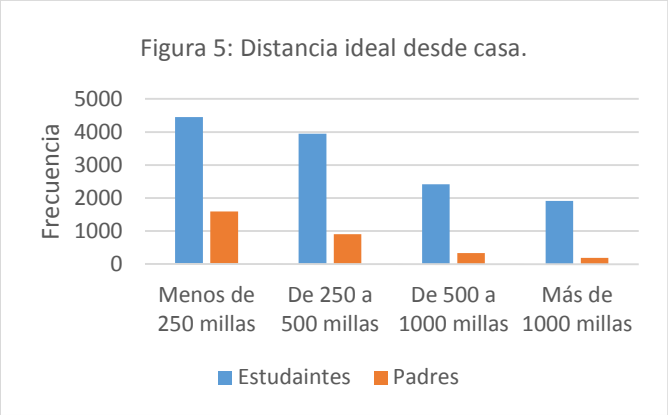


Figura 4: Gráfico de Barras Comparativo para la Distancia Ideal desde Casa



De los gráficos resulta fácil ver las diferencias entre estudiantes y padres de familia. Un alto porcentaje de padres de familia prefieren que la Universidad esté cerca de casa, y un alto porcentaje de estudiantes cree que la distancia ideal de casa es de más de 500 millas.

Nota adicional: para ver por qué es importante usar las frecuencias relativas en lugar de las absolutas, considere el siguiente diagrama incorrecto en el cual se han usado las frecuencias absolutas para comparar grupos de diferentes tamaños.



Debido a que hay más estudiantes que padres de familia participando en la encuesta (12,715 estudiantes y 3,007 padres de familia) el diagrama nos llevaría a conclusiones muy diferentes y erróneas en cuanto a las preferencias de estudiantes y padre de familia. Es decir, ya no parece que un alto porcentaje de los padres de familia prefieran que la universidad esté cerca de casa.

6. Distribuciones de Frecuencia y Gráficos para Variables Cuantitativas.

Una distribución de frecuencia es un arreglo de los datos por orden de magnitud. En ella se muestran los diferentes valores de un conjunto de datos y la frecuencia asociado con cada uno de ellos. Mostraremos como construir una distribución de frecuencia para variables cuantitativas. El procedimiento se muestra en el siguiente ejemplo.

Ejemplo 3: en la siguiente tabla se muestran las notas de 50 alumnos de Probabilidad y Estadística correspondientes a la primera evaluación. Construya una distribución de frecuencia para la nota de los estudiantes. Las notas se han multiplicado por 10 por facilidad.

75	89	57	88	61
90	79	91	69	99
83	85	82	79	72
78	73	86	86	86
80	87	72	92	81
98	77	68	82	78
82	84	51	77	90
70	70	88	68	81
78	86	62	70	76
89	67	87	85	80

Paso 1: Encuentre el mayor y el menor de los valores en la serie de datos. Para nuestro ejemplo estos valores son 99 y 51 respectivamente.

Paso 2: Calcule el Rango de los valores restando el menor valor del mayor. $R = \max - \min = 99 - 51 = 48$.

Paso 3: calcule el número de clases que usará. La cantidad $2^k = n$ se usa a menudo para estimar el número de clases conociendo el total de datos disponibles en nuestra muestra.

Entonces $2^k = 50$. Con $k = 5$ obtenemos 32 y con $k = 6$ obtenemos 64. Seleccionamos éste último valor por estar más cerca de 50. Entonces usaremos 6 clases.

Paso 4: determine el ancho de clases por medio de $c = R / \textit{Número de Clases}$. Así que el ancho de cada una de las clases será $c = 48/6 = 8$. Para asegurarnos de que las clases incluyan a todos los datos este resultado lo aproximaremos siempre al entero siguiente. Así que el ancho de clase que usaremos será de 9.

Paso 5: Construya el primer intervalo sumando al valor menor (51) el ancho de clase (9). Entonces nuestro primer intervalo irá desde 51 hasta 60. El segundo intervalo corresponde a valores desde $60-69$ y así sucesivamente hasta el último intervalo que corresponde a $96-104$. Finalmente, contamos cuantos de los datos corresponden a cada una de los intervalos y anotamos el resultado en la columna de frecuencias absolutas (f) y construimos una tabla como la siguiente en donde también se ha incluido la frecuencia relativa (f_r) y las marcas de clase que es el punto medio de cada una de las clases. Por ejemplo, la marca de clase para la cuarta clase es $\frac{78+86}{2} = 82$. Las marcas de clase suelen emplearse como valores representativos de los datos comprendidos en las clases. Por ejemplo, podemos decir que un valor representativo de los 20 datos comprendidos en la cuarta clase es 82.

Tabla 5: Distribución de frecuencia para las notas de la primera evaluación de Probabilidad y Estadística.

Límites de clase	f	f_r (en %)	Marcas de clase x_m
51 – 59	2	4	55
60 – 68	5	10	64
69 – 77	11	22	73
78 – 86	20	40	82
87 – 95	10	20	91
96 – 104	2	4	100
Total	50	100%	

¿Qué información nos proporciona una distribución de frecuencias? Organizar los datos de esta manera nos permite hacernos una idea general e inmediata del comportamiento de las notas de los alumnos. Por ejemplo, la nota más frecuente está entre 7.8 y 8.6 y un valor representativo de esta nota más frecuente es 8.2 que es la marca de clase. Hay dos estudiantes con notas sobresalientes entre 9.6 y 10.0, casi el doble de la nota que obtuvieron los dos alumnos con menor nota, pero solo representan un 4% de los estudiantes. Si consideramos que 6.0 es la nota requerida para aprobar el examen, podemos decir que la mayoría de estudiantes (48) aprobaron el examen, lo cual representa un 96% de aprobados. (recuerde que al inicio hemos multiplicado por 10 los datos, por lo que hay que tener en cuenta eso a la hora de obtener conclusiones).

Es importante mencionar que en este panorama general que nos hacemos con la tabla de frecuencia los datos individuales se pierden, no están disponibles para el usuario de la información.

Representación Gráfica de una Distribución de Frecuencias.

Para representar gráficamente una distribución de frecuencias usaremos el histograma de frecuencias, un gráfico muy parecido al gráfico de barras que construimos en la sección anterior. La diferencia más importante es que en el histograma, ambos ejes poseen escala ya que vamos a representar variables cuantitativas.

Iniciamos introduciendo el concepto de *limites reales de clase*. Éstos se obtienen restando 0.5 a los límites inferiores de clase y sumando 0.5 a los superiores, tal como se muestra en la tabla 6.

Los límites reales tienen la ventaja de representar como un solo intervalo el espacio donde se localizan las observaciones; es decir todos nuestros datos están comprendidos en el intervalo [50.5, 104.5], a diferencia de los límites de clases que son una serie de intervalos separados entre sí: [51,59]; [60,68];…; [96,104]. De esta manera, si consideramos las puntuaciones 86, 87 y 88, una puntuación en particular, por ejemplo 87, indica un nivel de conocimiento más cercano a 87 que a 86 u 88. Así que la puntuación 87 puede ser considerada como si se extendiera desde 86.5 hasta 87.5, como lo indica la siguiente figura.

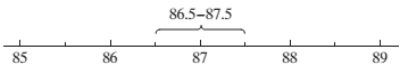
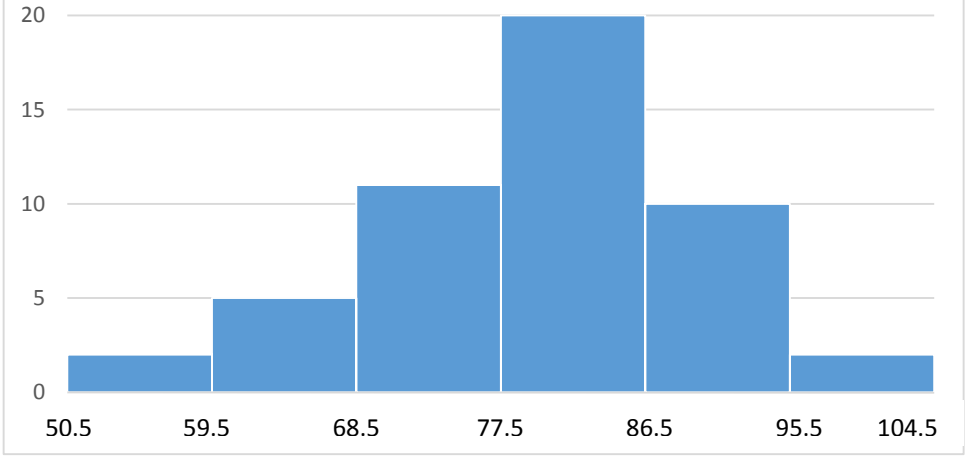


Tabla 6: Límites de clase y límites reales de clase

Límites de clase	Límites reales de clase	f
51 – 59	50.5 – 59.5	2
60 – 68	59.5 – 68.5	5
69 – 77	68.5 – 77.5	11
78 – 86	77.5 – 86.5	20
87 – 95	86.5 – 95.5	10
96 – 104	95.5 – 104.5	2

A partir de esta tabla construimos el histograma siguiente

Figura 6: Histograma de frecuencias para las notas de la primera evaluación de Probabilidad y Estadística.



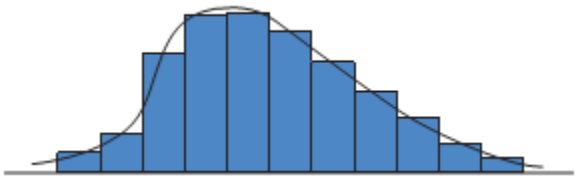
¿Qué información nos proporciona el histograma? Podemos obtener las mismas conclusiones que ya hemos mencionado anteriormente a partir de la distribución de frecuencias pero «de otra manera». Retomemos las ideas principales y explicaremos como obtenerlas a partir del gráfico:

- *La nota más frecuente está entre 7.8 y 8.6 y un valor representativo de esta nota más frecuente es 8.2*. Del histograma, es la barra con mayor altura, pero no olvide usar los límites de clase en lugar de los reales (sino diríamos que la nota más frecuente está entre 7.75 y 8.65, lo cual sería incorrecto ya que las notas no están reportadas hasta la centésima).
- *Hay dos estudiantes con notas sobresalientes entre 9.6 y 10.0, casi el doble de la nota que obtuvieron los dos alumnos con menor nota, pero solo representan un 4% de los estudiantes.* En el histograma esto se ve en la última barra, aunque a menudo, como en este caso, la lectura en el eje y resulta un poco difícil a simple vista, por lo que es una buena práctica incluir en cada barra su correspondiente frecuencia. Sin embargo, esto no representa una limitación del gráfico, ya que una información tan detallada por lo general no es necesaria, bastará con decir que hay «muy pocos estudiantes» con notas sobresalientes.
- *Si consideramos que 6.0 es la nota requerida para aprobar el examen, podemos decir que la mayoría de estudiantes (48) aprobaron el examen, lo cual representa un 96% de aprobados*. Del histograma a lo mejor no logramos tanta precisión, así que esta información podría quedar como: Si consideramos que 6.0 es la nota requerida para aprobar el examen, podemos decir que la mayoría de estudiantes aprobaron el examen (lo cual resulta evidente porque hay mayor cantidad de área a la derecha de 6.0).

Entonces, del histograma podemos obtener la misma información que de una distribución de frecuencias, pero a lo mejor con menos detalle. Sin embargo, hay un aspecto muy importante que revela el histograma: *la forma o perfil de distribución*, la cual nos proporciona un elemento más de análisis de una distribución de frecuencias. En los capítulos siguientes veremos que la forma de la distribución determina el método estadístico apropiado que debe usarse para analizar los datos.

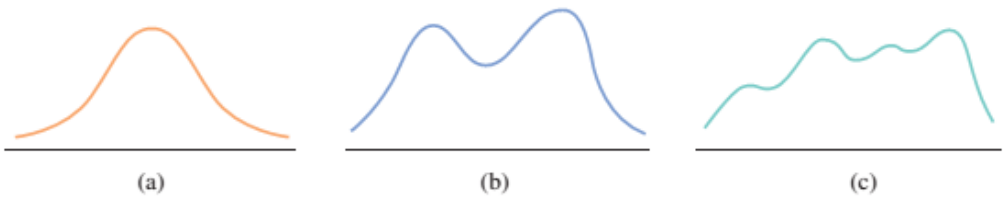
La forma general es una característica importante de un histograma. Al describir la forma resulta conveniente aproximar el histograma por una curva suavizada, tal como la muestra la siguiente figura

Figura 6: Histograma suavizado



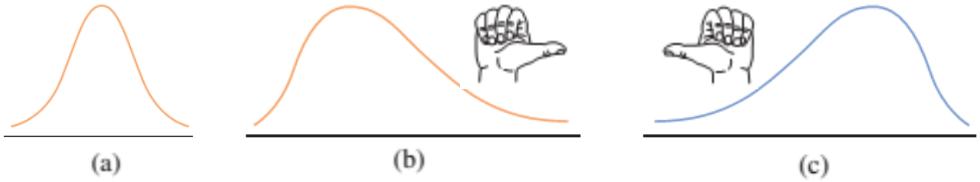
Una distribución puede tener muchas formas, pero para nuestros propósitos nos bastarán algunas de ellas y nos centraremos principalmente en:

- El número de picos o modas: Un histograma puede ser unimodal, si tiene un pico (a), bimodal, si tiene dos picos (b) y multimodal si tiene más de dos picos (c).

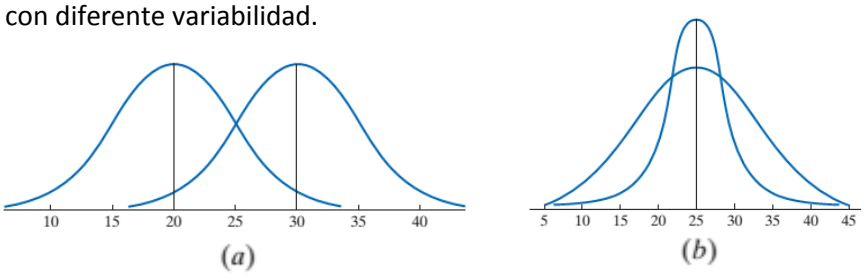


Una distribución bimodal se podría dar cuando los datos con los que disponemos provienen de dos poblaciones bastante diferentes. Imagine disponemos de una gran cantidad de datos que representan los tiempos de viaje desde Soyapango hasta San Salvador, para usuarios del SITRAMS y usuarios del sistema de transporte publico normal. Si construimos un histograma para estos datos combinados, posiblemente sea bimodal si los tiempos en el recorrido son significativamente diferentes entre ambos sistemas.

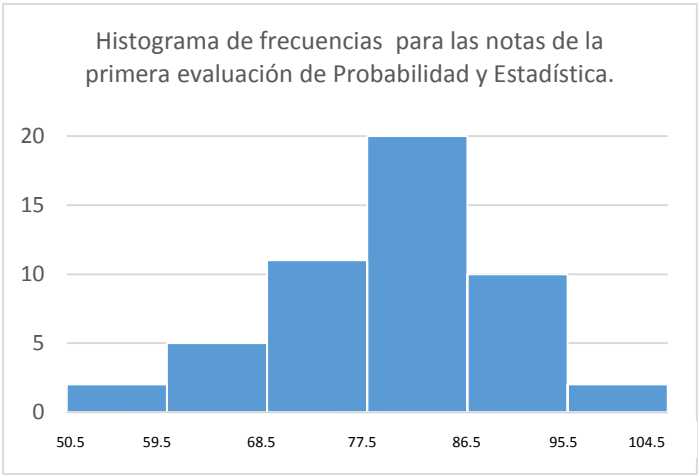
- La simetría: Un histograma es simétrico si existe una línea vertical tal que la porción del histograma a la izquierda de esta línea es similar a la porción que está a la derecha (a). Si esta línea no existe, entonces el histograma tendrá un sesgo a la derecha (b) o a la izquierda (c).



- La variabilidad: ¿se agrupan los datos alrededor de su valor representativo o se dispersan a lo largo del eje *x*? estas preguntas tienen que ver con la variabilidad de los datos en una distribución, que en el histograma se ve en el «ancho» que éste tiene. En (a) se muestran dos distribuciones que difieren en su valor central o típico pero tienen la misma variabilidad. En (b) se muestran dos distribuciones con diferente variabilidad.



Entonces, ¿qué podemos decir de nuestro histograma?



Nuestro gráfico es unimodal con una moda de 82 y sesgado a la izquierda, lo cual indica que la mayoría de puntuaciones fueron altas. Notamos además que hay poca variabilidad en los datos.

USO DE TECNOLOGÍA.

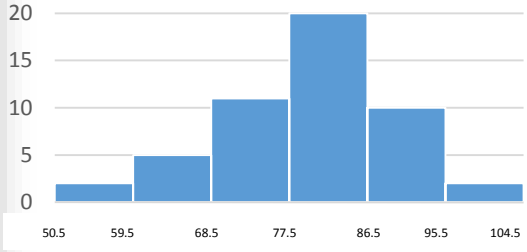
Existen numerosas herramientas que nos permitirán elaborar los gráficos que hasta este momento hemos estudiado. En Excel disponemos de tres alternativas para elaborar un histograma:

Una primera opción es hacer todo el procedimiento descrito anteriormente utilizando algunas de las funciones de Excel e insertar un diagrama de barras utilizando el asistente para gráficos (a).

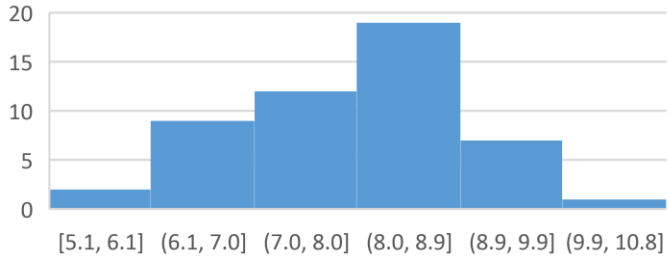
Otra alternativa es disponer todos los datos en una columna e insertar un histograma directamente (b).

Finalmente, podemos utilizar el complemento Análisis de Datos para generar el histograma (c)

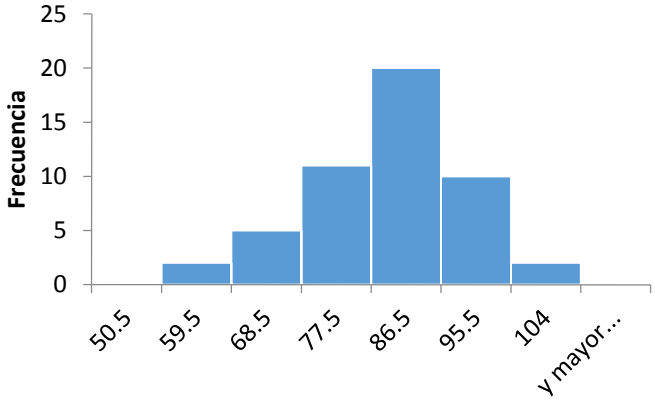
(a)



(b)



(c)



Note como los histogramas no son exactamente iguales, pero conservan las propiedades que hemos descrito anteriormente. Por lo general no se especifican los límites para las clases en el eje x, a menos que sea estrictamente necesario, ya que son generados automáticamente por Excel.

USO DE HISTOGRAMAS PARA COMPARAR DOS SERIES DE DATOS.

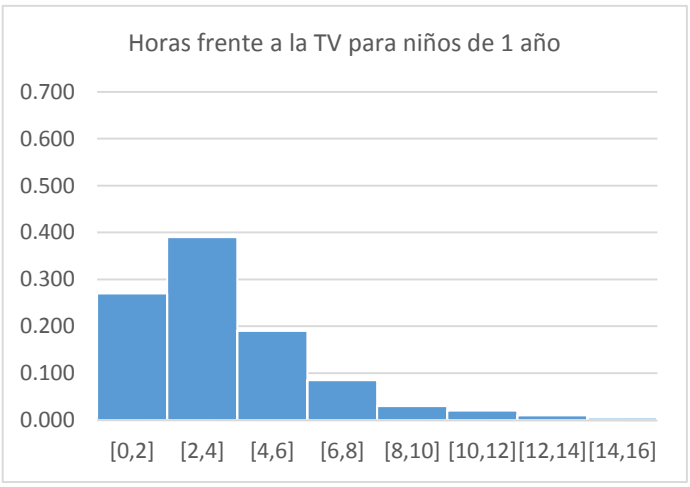
Si deseamos comparar dos grupos de datos, tenga en mente lo siguiente

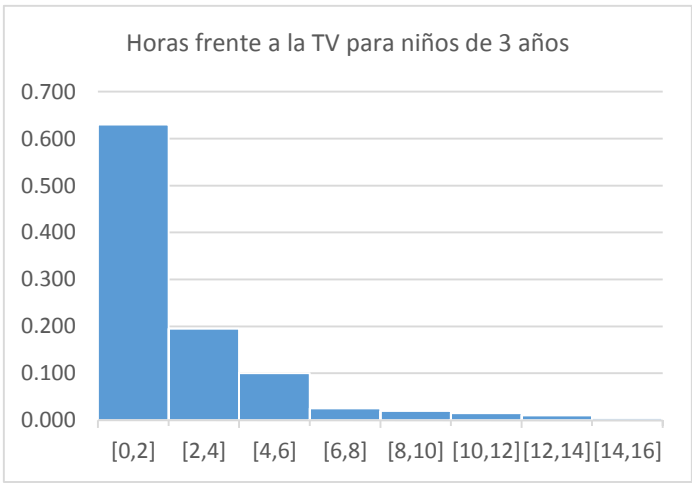
- Siempre use las frecuencias relativas para construir el histograma, ya que el número de observaciones en cada grupo podrían ser diferentes.
- Use las mismas escalas en ambos ejes para hacer que las comparaciones se vuelvan más fáciles.

Ejemplo 4: El artículo «Early Televisión Exposure and Subsequent Attention Problems in Children» (Pediatrics, April 2004) investiga los hábitos televisivos de los niños de EU. Los datos fueron obtenidos en un estudio a nivel nacional. En la siguiente tabla se muestran las frecuencias relativas aproximadas para el número de horas frente a la TV por día para niños de uno y tres años. Construya un histograma para cada variable y comente sus características más importantes.

TV Hours Per Day	Age 1 Year Relative Frequency	Age 3 Years Relative Frequency
0 to < 2	0.270	0.630
2 to < 4	0.390	0.195
4 to < 6	0.190	0.100
6 to < 8	0.085	0.025
8 to < 10	0.030	0.020
10 to < 12	0.020	0.015
12 to < 14	0.010	0.010
14 to < 16	0.005	0.005

Teniendo en mente las recomendaciones dadas se generaron en Excel los histogramas siguientes





De inmediato notamos que ambos histogramas tienen solamente un pico, con la mayoría de niños en ambos grupos de edades concentrados en pequeños intervalos frente a la TV. Por otra parte, ambos histogramas son sesgados a la derecha indicando un pequeño grupo de niños que ven mucha TV. La gran diferencia entre ambos gráficos está en el extremo

izquierdo de ambos gráficos. En esta parte vemos que hay una mayor proporción de niños de tres años en el intervalo de 0 a 2 horas que de un año. Un valor típico o representativo del número de horas frente a la TV para los niños de un año se ubica en el intervalo de 2 a 4 horas, mientras que para los niños de tres años este valor se ubica en el intervalo de 0 a 2 horas.

¿Y SI TENGO POCOS DATOS?

Con frecuencia dispondremos de pocos datos para nuestro análisis (menos de 30). En tal caso, no tiene sentido hacer un histograma ya que seguramente no resultará evidente la distribución de frecuencias de los datos.

En tales casos, resultará de mucha ayuda los diagramas de puntos y los diagramas de tallo y hoja.

Ejemplo 5: A continuación, se presentan los 12 mayores terremotos en la escala de Richter: 7.0 6.2 7.7 8.0 6.4 6.2 7.2 5.4 6.4 6.5 7.2 5.4

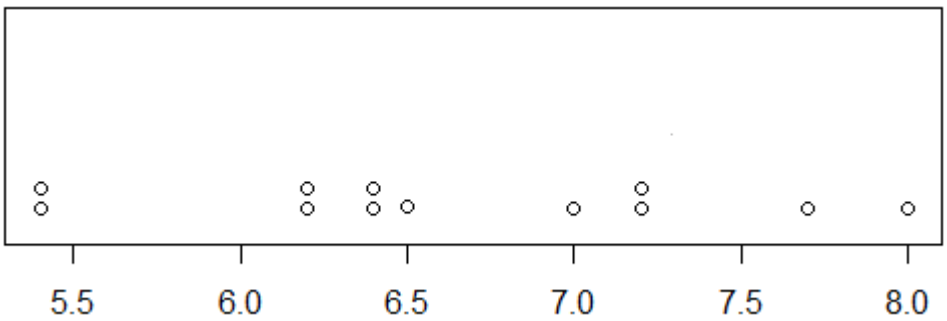
Un diagrama de tallo y hoja es una manera efectiva de resumir una serie pequeña de datos. cada número en el conjunto de datos se descompone en dos partes: el tallo y la hoja. El tallo es el primer dígito del número y la hoja el dígito final. Por ejemplo, el número 7.7 puede descomponerse como un 7 para el tallo y un 7 para la hoja. Al hacer este procedimiento para cada observación y teniendo en cuenta que el tallo debe ordenarse de manera descendente, obtenemos el siguiente diagrama:

5	4 4
6	2 4 2 4 5
7	0 7 2 2
8	0

Tallo: Unidades, hojas: décimas.

Del diagrama, rápidamente notamos que la distribución de datos es aproximadamente simétrica.

Otra alternativa es el diagrama de puntos que se construye colocando sobre una escala un punto correspondiente al valor del dato.



El inconveniente de este gráfico es que no resulta tan evidente como el primero que la distribución sea simétrica.

RESUMEN

Si bien una distribución de frecuencias resulta de mucha ayuda para averiguar algunas de las características más importantes de una serie de datos, una representación gráfica hace esta tarea más sencilla. El gráfico de barras es una manera muy popular de representar variables cualitativas, así como el histograma lo es para representar variables cuantitativas.

En el gráfico de barras, la frecuencia para cada categoría se representa por la altura de una barra construida sobre la categoría. De manera similar, en el histograma, las frecuencias para cada una de los intervalos de clase se representan por la altura de una barra que se construye sobre cada intervalo de valores.

Los datos generalmente se representan en el eje horizontal y las frecuencias (o frecuencias relativas) a lo largo del eje vertical. Las escalas deberán de ser seleccionadas de manera que el diagrama sea más ancho que alto, los ejes deberán tener sus respectivos rótulos que los identifiquen y deberá incluirse un título informativo principal.

Se pueden obtener diferentes formas en los gráficos al agrupar los datos de diferente manera y usando diferentes escalas en ambos ejes. Por estas razones los gráficos llevan a interpretaciones incorrectas de los datos (de manera intencional o no). Su tarea será ser siempre objetivo al comunicar los datos de manera clara, precisa e imparcial.

Finalmente, nos resultará de mucha ayuda describir una distribución de frecuencias y su representación gráfica en término de las características siguientes:

- Centro o valor típico
- La variabilidad o extensión en el eje horizontal
- La forma general
- Localización y número de picos
- La presencia de huecos

GRÁFICOS A EVITAR.

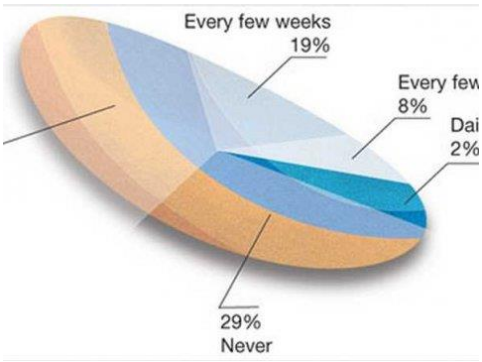
Esta sección se basa en una charla dictada por Karl W. Broman titulada «How to Display Data Badly», inspirada en el paper de 1984 de H. Wainer: «How to display data badly». Wainer fue el primero en hablar de los principios de la mala presentación de datos que, según Karl, con el uso creciente de Excel ha experimentado notables avances en los últimos años.

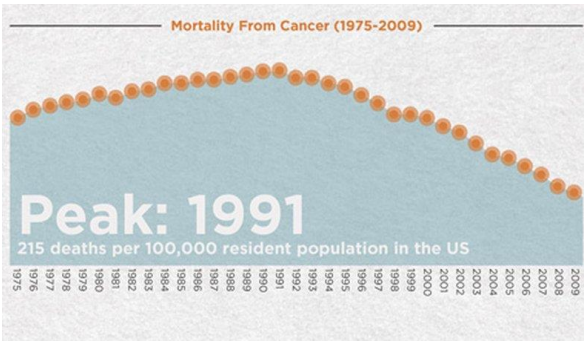
Principios Generales.

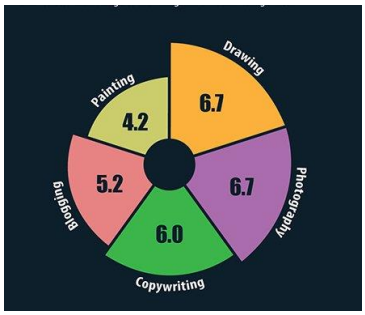
El objetivo de un buen gráfico es presentar la información de manera precisa y clara. Algunas reglas para hacer un mal gráfico son las siguientes:

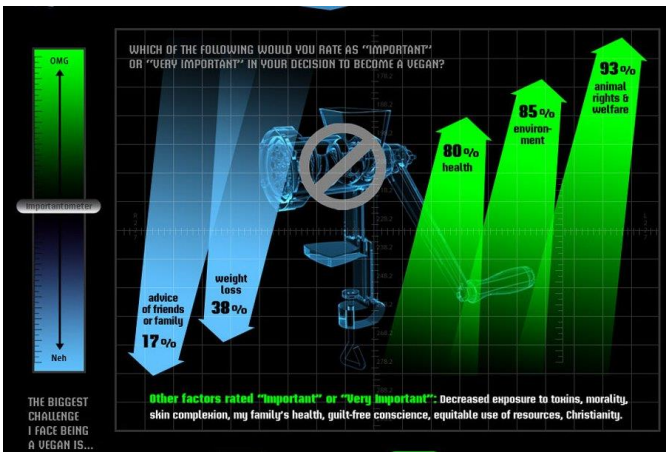
- Muestre la menor cantidad de información que le sea posible.
- Oculte lo que quiere mostrar (si es con basura gráfica ¡mejor!).
- Use gráficos en pseudo-3D y coloréelos generosamente.
- Haga un diagrama de pastel (de preferencia en 3D y a colores).
- Use una escala inadecuada.
- Ignore las cifras significativas

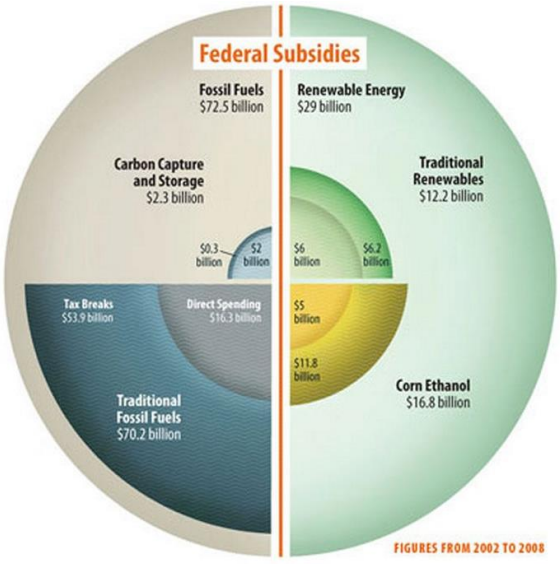
Algunos ejemplos de éstos y otros errores se muestran en la siguiente top list, que se hace con el objetivo de que el lector nunca los imite o supere.

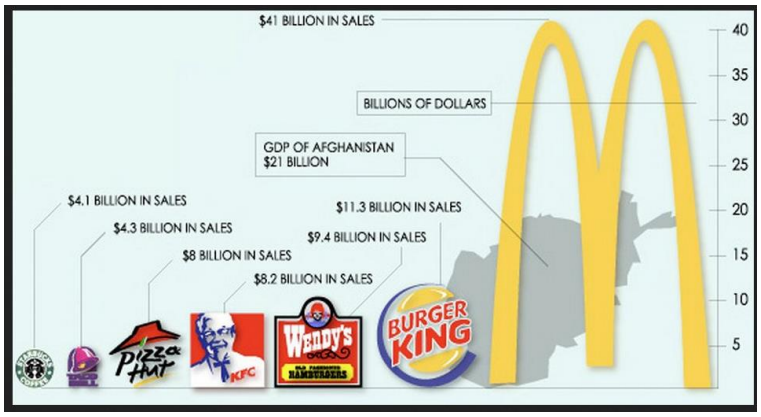


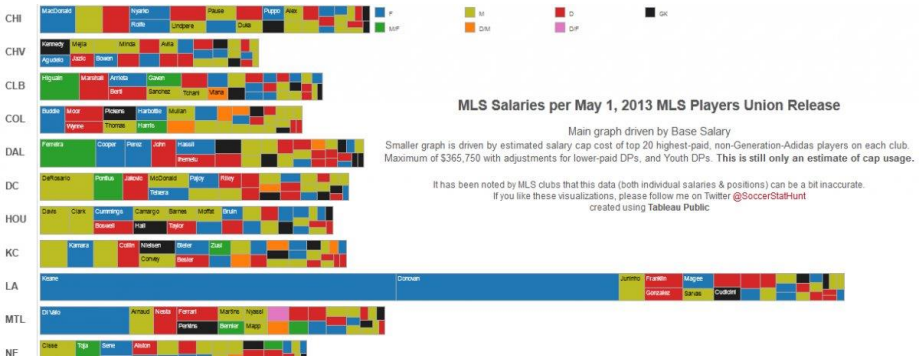


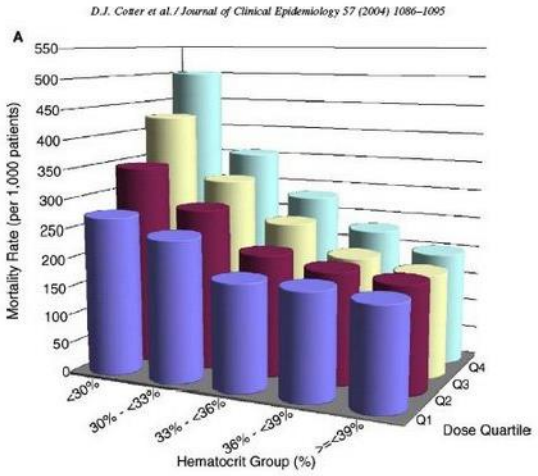


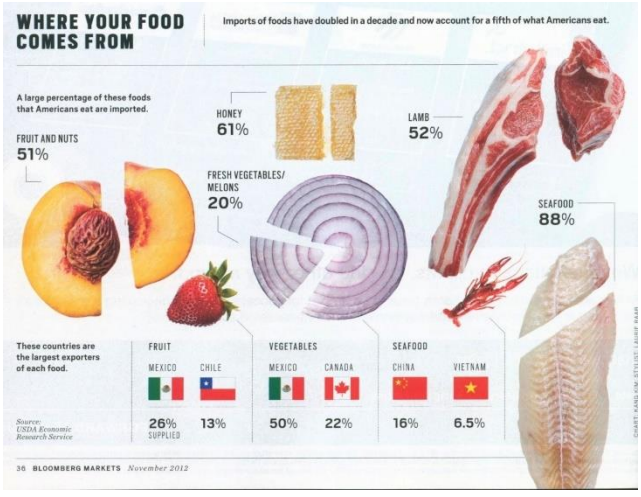



Federal Subsidies
Fossil Fuels
$72.5 billion
Renewable Energy
$29 billion
Carbon Capture and Storage
$2.3 billion
Traditional Renewables
$12.2 billion
$0.3 billion
$2 billion
$6 billion
$6.2 billion
Tax Breaks
$53.9 billion
Direct Spending
$16.3 billion
$5 billion
$11.8 billion
Traditional Fossil Fuels
$70.2 billion
Corn Ethanol
$16.8 billion
FIGURES FROM 2002 TO 2008


$41 BILLION IN SALES
BILLIONS OF DOLLARS
GDP OF AFGHANISTAN
$21 BILLION
$4.1 BILLION IN SALES
$4.3 BILLION IN SALES
$8 BILLION IN SALES
$8.2 BILLION IN SALES
$9.4 BILLION IN SALES
$11.3 BILLION IN SALES
40
35
30
25
20
15
10
5
Pizza Hut
KFC
Wendy's
BURGER KING


CHI
CHV
CLB
COL
DAL
DC
HOU
KC
LA
MTL
NF
MLS Salaries per May 1, 2013 MLS Players Union Release
Main graph driven by Base Salary
Smaller graph is driven by estimated salary cap cost of top 20 highest-paid, non-Generation-Adidas players on each club.
Maximum of $365,750 with adjustments for lower-paid DPs, and Youth DPs. This is still only an estimate of cap usage.
It has been noted by MLS clubs that this data (both individual salaries & positions) can be a bit inaccurate.
If you like these visualizations, please follow me on Twitter @SoccerStatHunt
created using Tableau Public


D.J. Cotter et al. / Journal of Clinical Epidemiology 57 (2004) 1086–1095
A
Mortality Rate (per 1,000 patients)
550
500
450
400
350
300
250
200
150
100
50
0
<30%
30% - <33%
33% - <36%
36% - <39%
>=<39%
Hematocrit Group (%)
Q1
Q2
Q3
Q4
Dose Quartile


WHERE YOUR FOOD COMES FROM
Imports of foods have doubled in a decade and now account for a fifth of what Americans eat.
A large percentage of these foods that Americans eat are imported.
FRUIT AND NUTS
51%
HONEY
61%
FRESH VEGETABLES/ MELONS
20%
LAMB
52%
SEAFOOD
88%
These countries are the largest exporters of each food.
FRUIT
MEXICO
CHILE
26%
SUPPLIED
13%
VEGETABLES
MEXICO
CANADA
50%
22%
SEAFOOD
CHINA
VIETNAM
16%
6.5%
Source: USDA Economic Research Service
36 BLOOMBERG MARKETS November 2012


TOP 5 ARTILLEROS
MESSI
301 GOLES
ZARRA
251 GOLES
CRISTIANO
244 GOLES
HUGO SÁNCHEZ
234 GOLES
RAÚL
228 GOLES
MESSI, 301
PRIMER JUGADOR 'TRICENTENARIO' EN LA LIGA
POR TEMPORADAS / GOLES
1
6
14
10
23
34
31
50
46
28
43
15*
2004-05
2005-06
2006-07
2007-08
2008-09
2009-10
2010-11
2011-12
2012-13
2013-14
2014-15
2015-16
* Hasta la jornada 24
57%
CAMP NOU
172
en 167 partidos
301
GOLES
43%
VISITANTE
129
en 166 partidos
RIVALES MÁS GOLEADOS
Atlético
21 GOLES
Valencia
16 GOLES
Sevilla
19 GOLES
Levante/Depor
15 GOLES
Osasuna
17 GOLES
Real Madrid
14 GOLES

Referencias:

1. Roxy Peck, **Statistics: Learning from Data**, 2015.

2. Coladarsi Theodore, et al. **Fundamentals of Statistical Reasoning In Education** 3rd ed. 2011.

3. Allan G. Bluman, **Elementary Statistics: a Step by Step Approach** . 7th ed. 2009

4. Nieves Antonio, **Probabilidad y Estadística para Ingeniería, un enfoque moderno**, 2010.