Probabilidad y Estadística Fundamental

Estadística descriptiva - Resumen y descripción de datos de una variable - Presentación tabular y gráfica de una variable

Profesor: Nicolás López

Universidad Nacional de Colombia

Contenido

Introducción a R

Este lenguaje de programación estadístico puede ser instalado localmente, o pueden ser usadas versiones cloud del lenguaje (https://posit.cloud/). Vamos a revisar el laboratorio 0 de programación.

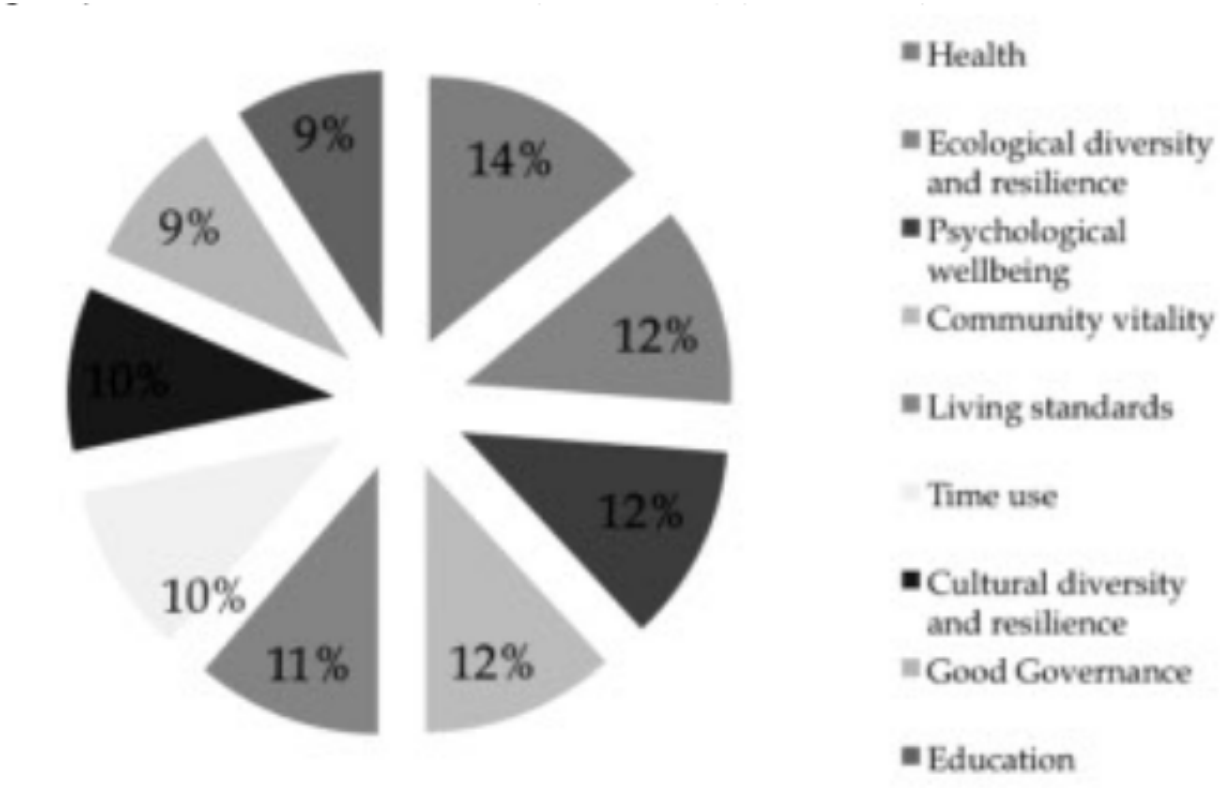
Introducción a R

Repasar el primer laboratorio de programación en R (ejecutarlo línea a línea y validar los comentarios).

Contenido

Happiness

Figura 1: Tomado de *World Hapiness Report*. Dominios en los cuales las personas felices disfrutan de suficiencia



¿Qué comentarios surgen al ver el gráfico?

Happiness

- ¿Qué pasó con los colores?
- ¿Cuál es el objetivo de "separar" la torta?
- ¡La suma no da 100 %!
- Si se ponen los *porcentajes* en las fracciones de la torta ¿no sería mejor hacer una tabla?
- Supongamos que quitamos los porcentajes en la torta. ¿se entiende realmente el gráfico? ¿cómo determino cuál porción del pie es "más grande" o "más pequeña"? ¿cuál sería el *valor agregado* del gráfico en un informe?

Si debe aclarar un gráfico con números,
utilice una tabla!

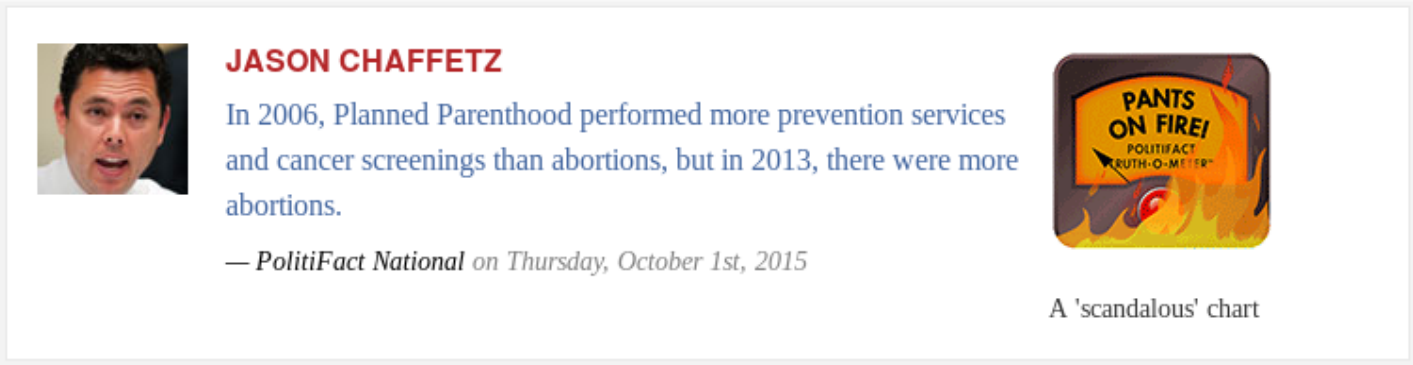
Happiness

- ¿Realmente es necesario "gastar tinta" en una tabla?. Todos los dominios considerados tienen casi el mismo porcentaje, puede bastar con decir: "Las nueve dimensiones consideradas contribuyen de manera similar a la suficiencia de las personas felices, así, todos los dominios son importantes" junto con un gráfico apropiado.
- Si se opta por un gráfico, casi siempre se prefiere un *gráfico de barras* a un *gráfico de torta*.

PolitiFact - Independent fact-checking journalism

"The goal of the Truth-O-Meter is to reflect the relative accuracy of a statement"

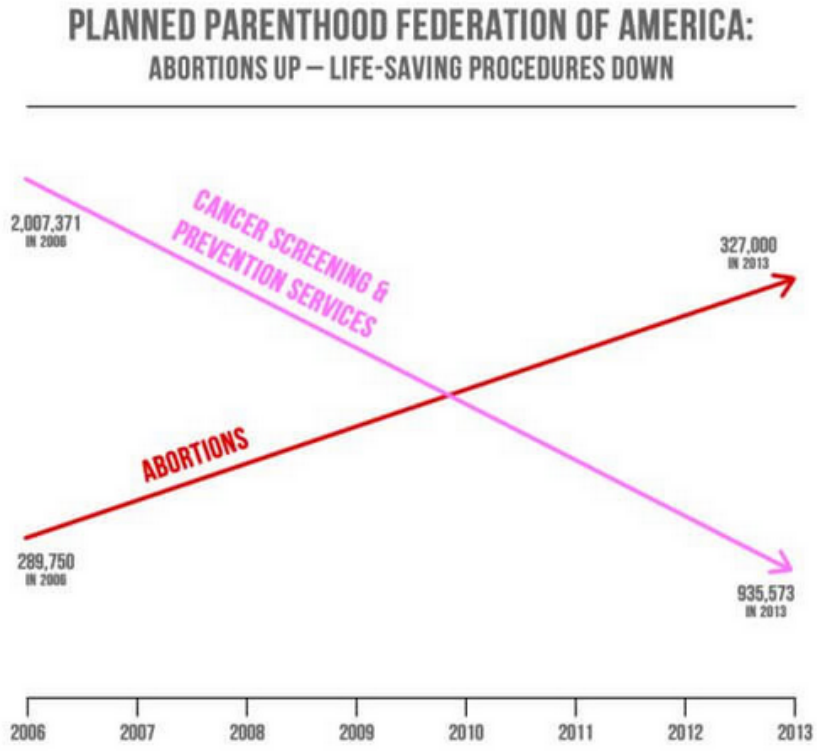
Figura 2: Tomado de *Politifact*.



La clasificación de las afirmaciones políticas de la página es: *True, Mosly True, Half True, Mostly False, False y Pants On Fire*, la última implica que *The statement is not accurate and makes a ridiculous claim*.

PolitiFact - Independent fact-checking journalism

Figura 3: Tomado de *Politifact*.



The numbers listed on the chart are based on actual statistics, but they are small and were hard to read during the televised hearings. The chart's most prominent feature, the much larger crossed arrows, suggests a conclusion that's flat wrong

Contenido

Distribución de los datos

Una vez recolectados y dispuestos en una base, los datos pueden resumirse en forma de tablas estadísticas y gráficas, las cuales pueden usarse para mostrar la **distribución** de los datos:

- ¿Qué valores de la variable han sido medidos?
- ¿Con qué frecuencia se presenta cada uno de los valores?

Lo que se busca es un **resumen más compacto de los datos** que la simple lista de observaciones.

Distribución de los datos

Se **sacrifica** la información individual para ganar **interpretabilidad** en los datos

Distribución de los datos

Ejemplo escolaridad

Se cuenta con el nivel educativo de una muestra de tamaño $n = 3000$ de la cual se muestran a continuación 30 observaciones:

```
[1] 1. < HS Grad 4. College Grad 3. Some College 4. College Grad
[5] 2. HS Grad 4. College Grad 3. Some College 3. Some College
[9] 3. Some College 2. HS Grad 3. Some College 2. HS Grad
[13] 2. HS Grad 4. College Grad 2. HS Grad 3. Some College
[17] 4. College Grad 1. < HS Grad 4. College Grad 3. Some College
[21] 4. College Grad 2. HS Grad 4. College Grad 4. College Grad
[25] 2. HS Grad 1. < HS Grad 2. HS Grad 2. HS Grad
[29] 2. HS Grad 2. HS Grad
```

¿Qué comentarios surgen al ver la lista de observaciones para la variable *Escolaridad*? (respecto a su distribución)

Distribución de los datos

Ejemplo escolaridad

Al revisar con cuidado, se observa que la escolaridad de estas 30 personas toma diferentes valores:

- `<HS Grad`. Menor de bachillerato.
- `HS Grad`. Bachillerato.
- `Some College`. Menor de universitario.
- `College Grad`. Universitario.

Esta es una variable **cualitativa ordinal**, como mencionamos la clase anterior.

Distribución de los datos

Ejemplo escolaridad

Preguntas

- ¿Representarán estas 4 categorías todos los posibles valores de la variable *Escolaridad*?
- ¿Cuál es le valor más frecuente de la variable *Escolaridad*? ¿cuál es el menos frecuente? ¿cómo se **distribuye** la variable?

Distribución de los datos

Ejemplo escolaridad

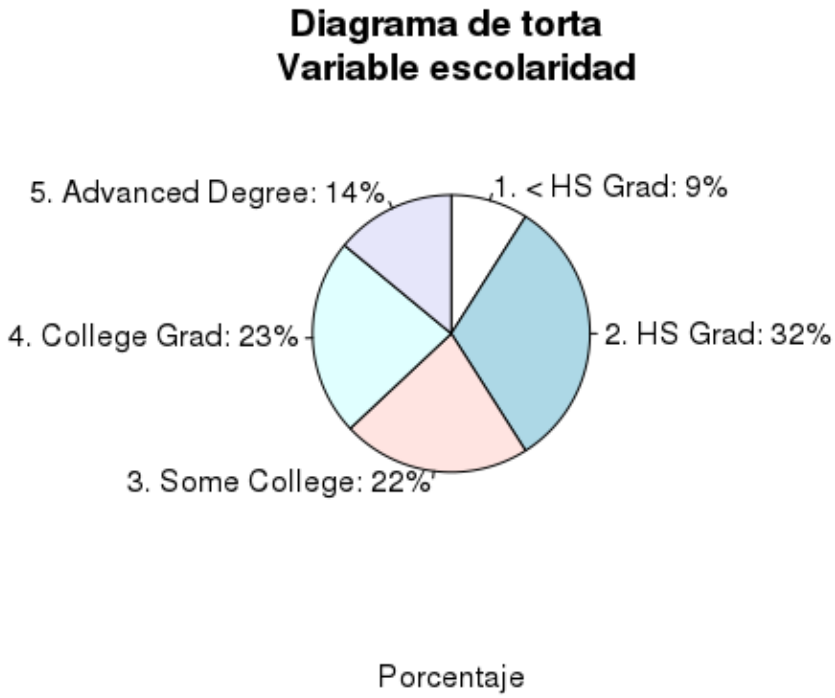
Respuestas

- Hay 3000 observaciones, puede haber personas con menor, o mayor nivel educativo. Requiero un análisis completo de la base de datos.
- Aún con **tan sólo 30 registros**, la lista no es clara: no se puede determinar cuál valor es más o menos frecuente.
- No se pueden determinar las características de los 3000 datos sin una descripción univariada de la variable.

Distribución de los datos

Ejemplo escolaridad

Figura 4: Diagrama de torta para la variable *Escolaridad*

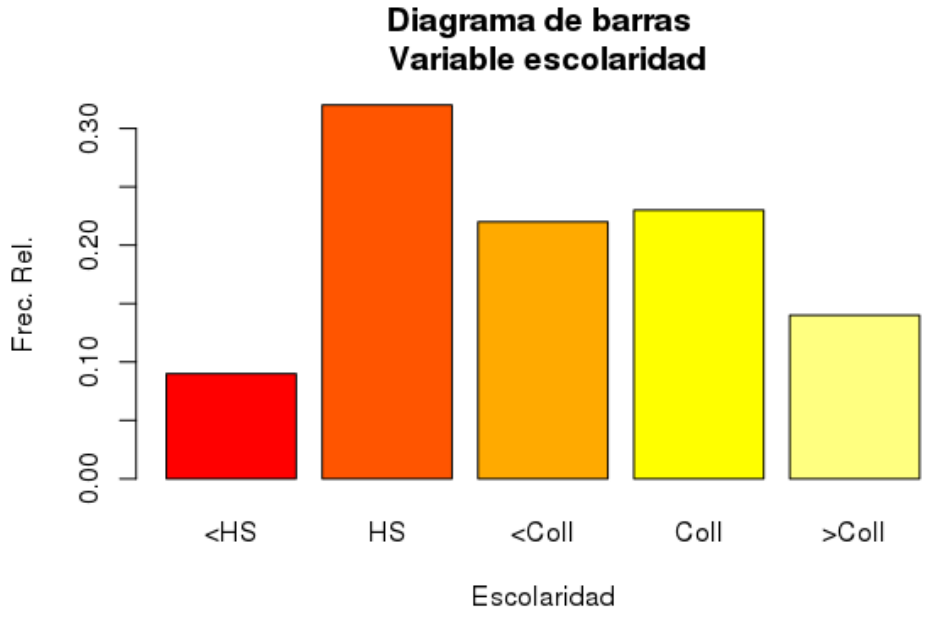


¿Qué comentarios surgen al ver el diagrama de torta? ¿cómo se caracteriza la **distribución de frecuencias porcentual** para la variable *Escolaridad*?

Distribución de los datos

Ejemplo escolaridad

Figura 5: Diagrama de barras para la variable *Escolaridad*



¿Qué comentarios surgen al ver el diagrama de barras? ¿cómo se caracteriza la **distribución de frecuencias relativas** para la variable *Escolaridad*? ¿cómo se diferencia del anterior gráfico?

Distribución de los datos

Ejemplo escolaridad

Cuadro 1: Distribución de frecuencias absolutas, relativas y acumuladas. Variable Escolaridad

Educación	Frec. Abs.	Frec. Rel	Frec. Acum.
1. $<$ HS Grad	268	0.09	0.09
2. HS Grad	971	0.32	0.41
3. Some College	650	0.22	0.63
4. College Grad	685	0.23	0.86
5. Advanced Degree	426	0.14	1.00

¿Qué comentarios surgen al ver la **distribución de frecuencias** (absoluta, relativa y acumulada) para la variable Escolaridad?

Distribución de los datos

Ejemplo edad

Ahora se muestran 270 observaciones para la variable *Edad*:

```
[1] 18 24 45 43 50 54 44 30 41 52 45 34 35 39 54 51 37 50 56 37 38
[22] 40 75 40 38 49 43 34 57 18 55 51 33 34 36 56 70 25 32 27 28 27
[43] 43 50 39 52 35 57 25 33 57 71 43 23 30 22 59 28 61 34 43 54 69
[64] 41 48 49 42 37 55 21 58 31 25 32 40 44 60 23 63 44 47 61 55 24
[85] 42 25 34 53 53 70 47 46 33 34 22 74 40 45 43 33 62 37 54 34 50
[106] 46 41 63 38 35 29 66 37 39 42 51 55 51 38 49 42 43 38 59 57 25
[127] 49 41 38 61 49 52 43 60 46 21 61 32 58 35 26 32 37 22 51 44 35
[148] 60 40 35 35 47 43 33 60 38 53 55 57 64 43 35 54 45 58 48 46 46
[169] 55 51 49 34 53 40 50 37 39 52 50 48 47 27 39 44 37 52 26 39 25
[190] 31 58 30 27 40 55 35 48 29 25 40 27 44 49 22 45 33 63 49 39 25
[211] 29 37 35 46 58 39 41 29 37 62 27 37 56 36 23 46 26 43 36 54 62
[232] 40 53 48 23 34 43 55 59 60 49 45 31 34 26 22 40 41 36 38 48 47
[253] 47 35 49 40 32 57 56 73 29 49 30 29 36 39 48 45 50 42
```

¿Qué comentarios surgen al ver la lista de observaciones para la variable *Edad?*

Distribución de los datos

Ejemplo edad

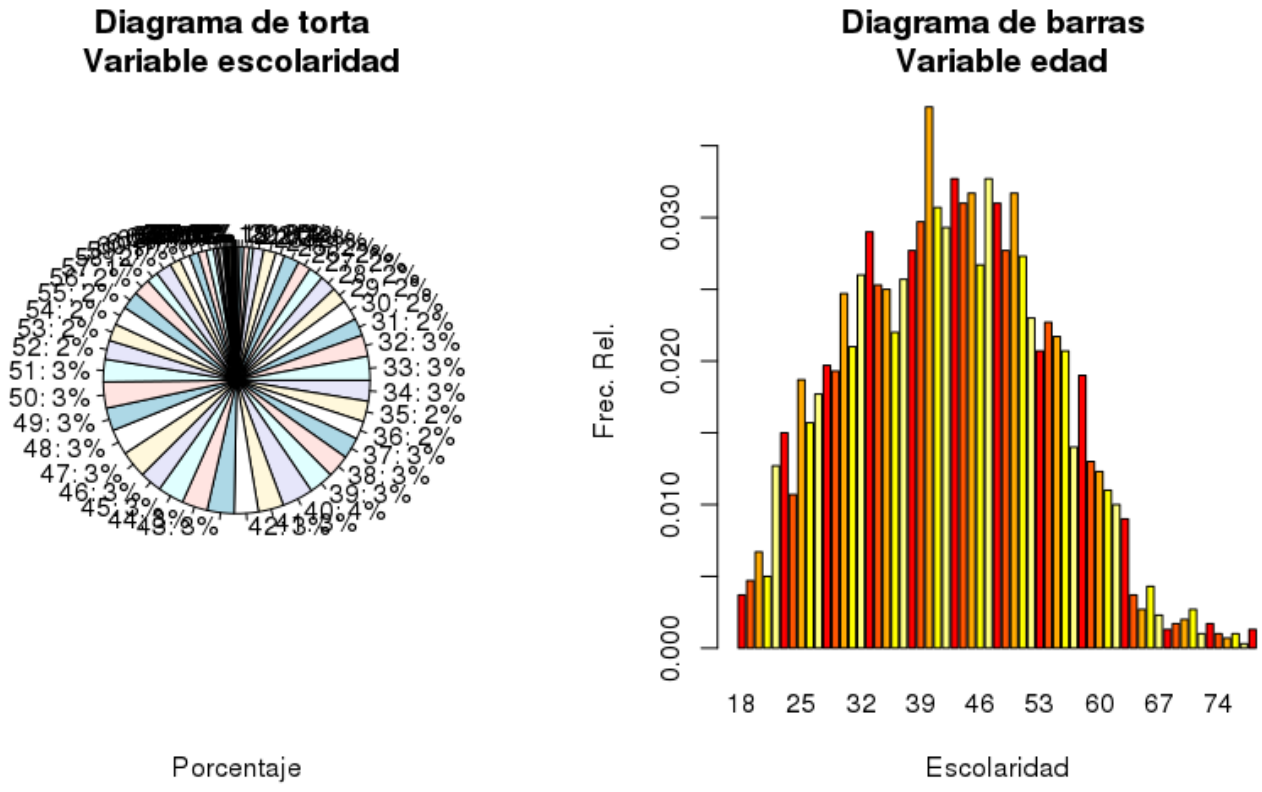
Al revisar con cuidado, se observa que la edad está dada en años cumplidos, por lo cual esta es una variable **cuantitativa (discreta) de razón**, como mencionamos la clase pasada:

- ¿Cómo describir la distribución de una variable de tipo cuantitativo?
- ¿Qué efecto tiene que la variable cuantitativa sea discreta o continua en la descripción estadística?

Distribución de los datos

Ejemplo edad

Figura 6: Diagrama de torta (izq.) y de barras (der.) para la variable *Edad*



Distribución de los datos

Ejemplo edad

Para la caracterización de la distribución de la variable *Edad*:

- El diagrama de torta divide demasiado el área del círculo, ocultando la **verdadera distribución** de la variable.
- El diagrama de barras parece hacer un mejor trabajo al ser más informativo que el diagrama de torta.
- Hacer una tabulación de la variable considerando cada edad pierde sentido, hay demasiadas categorías (61 posibles edades).

¿Cómo visualizaría la distribución de esta variable?

Distribución de los datos

Conclusión

- Se han presentado dos maneras de caracterizar la distribución univariada de los datos: tabularmente y gráficamente.
- Las gráficas y tablas deben ser construidas de manera diferente si la variable en cuestión es cualitativa o cuantitativa.
- En algunos casos, el escenario cuantitativo discreto puede llegar a verse como el cualitativo en la descripción gráfica y tabular de la distribución de la variable.
- En general, el escenario cuantitativo, principalmente para variables continuas, requiere un acercamiento diferente para encontrar la distribución de la variable de interés.

Ejemplo de aplicación

Índice de Masa Corporal IMC de la muestra de estudiantes

Recuerde que el peso de los 5 estudiantes de la muestra está dado por $y_1 = 50$, $y_2 = 59$, $y_3 = 64$, $y_4 = 45$ y $y_5 = 70$ y su talla en cms es $x_1 = 156$, $x_2 = 164$, $x_3 = 141$, $x_4 = 178$ y $x_5 = 161$. Si crea una nueva variable Z' como el estado de peso para cada estudiante (ver tabla abajo)

Figura 7: Tabulación del IMC (BMI). Tomado de *CDC*.

BMI	Weight Status
Below 18.5	Underweight
18.5 – 24.9	Healthy Weight
25.0 – 29.9	Overweight
30.0 and Above	Obesity

Caracterice la distribución de frecuencias para esta nueva variable.

Contenido

Presentación tabular

Datos cualitativos

Cuando la variable de interés es cualitativa, la **tabla estadística** lista las categorías consideradas junto a una medida de la frecuencia de cada valor. Usualmente se mide esta frecuencia en dos formas diferentes:

- Frecuencia absoluta: número de mediciones en cada categoría
- Frecuencia relativa: proporción de mediciones en cada categoría[1].

Adicionalmente, se puede calcular a partir de las frecuencias (absolutas o relativas) las **frecuencias acumuladas** (absolutas o relativas).

[1] Si se multiplica por 100, se obtiene un porcentaje.

Presentación tabular

Datos cuantitativos

Cuando la variable de interés es cuantitativa

- Discreta con pocas categorías: se procede con una tabulación similar al caso cualitativo.
- Discreta con bastantes categorías o continua: la variable se divide en **intervalos de clase** y se procede con una tabulación similar al caso cualitativo.

Presentación tabular

Datos cuantitativos

Al dividir el recorrido de la variable en intervalos, estos deben ser escogidos para que cada mediciones *caiga* en una sola clase:

1. Se define el número de clases k.
2. Se define el ancho de clase como el **rango** (diferencia entre el *máximo* y el *mínimo* valor de la variable) dividido entre número de clases[2].
3. Se definen los intervalos: empezando desde el mínimo valor de la variable se va adicionando el ancho de clase. Se incluye en el intervalo el límite inferior pero se excluye el superior.
4. Se mide la frecuencia de cada clase.

[2] Este ancho puede redondearse para comodidad en los siguientes pasos

Presentación tabular

Datos cuantitativos

Para el ejemplo de la variable *Edad*

1. Se define $k = 11$.
2. Se define el ancho de clase como:
   (80(edad máx) - 18(edad mín)) / 11 (no. de clases) $= 5.636 \approx 6$.
   Con lo cual 5.636 es el ancho mínimo de clase para obtener 11 intervalos de clase.
3. Se definen los subintervalos:
   - $[18, 24)$
   - $[24, 30)$
   - ...
   - $[72, 78)$
   - $[78, 84)$

   Si no se aproxima el ancho de clase, el último intervalo es cerrado a la derecha, esto para incluir el valor máximo de la variable.
4. Se mide la frecuencia de cada clase.

Presentación tabular

Datos cuantitativos

Cuadro 2: Distribución de frecuencias absolutas, relativas y acumuladas. Variable *Edad* en $k = 11$ intervalos de clase igualmente espaciados

Intervalo de edad	Frec. Abs.	Frec. Rel	Frec. Acum.
$[18, 24)$	143	0.048	0.048
$[24, 30)$	305	0.102	0.149
$[30, 36)$	453	0.151	0.300
$[36, 42)$	520	0.173	0.474
$[42, 48)$	552	0.184	0.658
$[48, 54)$	484	0.161	0.819
$[54, 60)$	333	0.111	0.930
$[60, 66)$	146	0.049	0.979
$[66, 72)$	43	0.014	0.993
$[72, 78)$	17	0.006	0.999
$[78, 84)$	4	0.001	1.000

Presentación tabular

Datos cuantitativos

Para una variable arbitraria, se tiene la siguiente representación de la tabla estadística general:

Figura 8: Tabla estadística general

Intervalos de clase	Marcas de clase	Frecuencias absolutas	Frecuencias relativas	Frecuencias relativas acumuladas
$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i/N$	$F_i = N_i/N$
$a_1 - a_2$	c_1	n_1	f_1	F_1
$a_2 - a_3$	c_2	n_2	f_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_k - a_{k+1}$	c_k	n_k	f_k	F_k

Note la incorporación de la **marca de clase**.

Presentación tabular

Datos cuantitativos

A tener en cuenta:

- Los intervalos de clase resultan en una mejor interpretabilidad a costo de pérdida de información.
- Los intervalos de clase podrían tener diferente ancho, pero generalmente tienen el mismo.
- Pocos intervalos de clase ocultan la distribución subyacente de los datos. Muchos intervalos también pueden presentar el mismo inconveniente.

Presentación tabular

Datos cuantitativos

A tener en cuenta:

- Entre 5 y 12 intervalos son **generalmente** aceptables y cuantos más datos haya, más clases se requieren.
- Hay diferentes reglas de referencia para determinar el número óptimo de intervalos k
  - Una recomendación muy general del número de clases según el tamaño de muestra está dada por: $n = 25 \rightarrow k = 6$. $n = 50 \rightarrow k = 7$. $n = 100 \rightarrow k = 8$. $n = 200 \rightarrow k = 9$. $n = 500 \rightarrow k = 10$[3].
  - Regla de Sturges y de la raíz (averiguar).

  Cualquiera que sea, debe estar acompañada de un **cuidadoso análisis** de los datos.

[3] Tomada de *Introducción a la probabilidad y estadística*

Contenido

Presentación gráfica

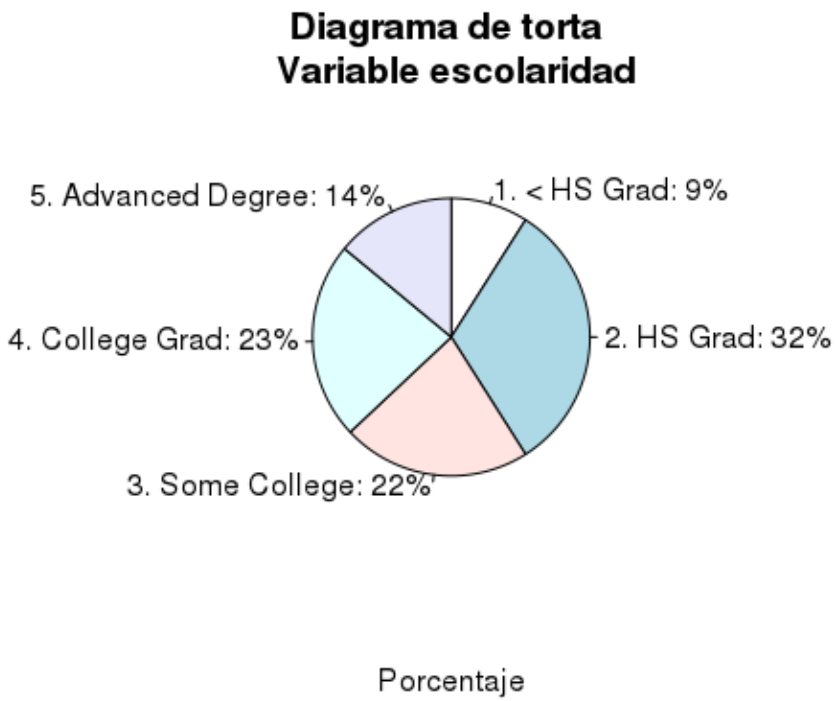
Note que en la última clase buscamos la distribución de los datos de forma completamente tabular. Un gráfico puede resumir de manera parsimoniosa características de nuestros datos.

Presentación gráfica

Datos cualitativos. Gráfico de torta

Usado para datos cualitativos. Un círculo se divide en segmentos dónde las áreas son proporcionales a los porcentajes de la variable en cuestión.

Figura 9: Diagrama de torta para la variable *Escolaridad*



Presentación gráfica

Datos cualitativos. Gráfico de torta - Ventajas

- ✓ Resume un conjunto grande de datos de forma visual.
- ✓ Ampliamente conocido y fácil de explicar.

Presentación gráfica

Datos cualitativos. Gráfico de torta - Desventajas

- ✗ Limitado a un número pequeño de categorías.
- ✗ Únicamente permite presentar porcentajes de un total.
- ✗ Para variables ordinales, visualizar el ordenamiento de las categorías en la torta no es claro.
- ✗ Difícil de interpretar: la dimensión de las porciones de la torta no es evidente, generalmente requiere adicionar los porcentajes para una mejor interpretación[4].
- ✗ Dado lo último, comparar las categorías entre si puede ser bastante complicado.
- ✗ Fácilmente distorsionable: al usar gráficos de torta en 3D, visualmente pueden manipularse las frecuencias observadas.
- ✗ Al comparar dos o más gráficos de torta, puede ser difícil diferenciar la distribución de las variables, mas aún si toman valores similares.
- ✗ Generalmente una tabla o un gráfico de barras son mejores opciones que el gráfico de torta.

[4] Esta dificultad es dada por la noción de área necesaria para dimensionar la frecuencia de cada categoría.

Presentación gráfica

Datos cualitativos y cuantitativos. Gráfico de barras

Usado para datos cualitativos o datos cuantitativos discretos con pocas clases. Las categorías usualmente se muestran en el eje horizontal y las frecuencias (absolutas o relativas) en el vertical.

- Las barras se separan entre ellas para enfatizar las distintas categorías.
- Las barras deben tener el mismo ancho.
- La longitud de cada barra es proporcional a la medida de la categoría (frecuencias, porcentajes o proporciones)

Presentación gráfica

Datos cualitativos y cuantitativos. Gráfico de barras - Ventajas

- ✓ Resume un conjunto grande de datos de forma visual.
- ✓ Conocido y fácil de explicar.
- ✓ Permite un gran número de categorías sin perder interpretabilidad, además de permitir observar el ordenamiento de las categorías en el caso ordinal.
- ✓ Fácil de interpretar[5], y a diferencia del diagrama de torta, permite no solo visualizar porcentajes, además, frecuencias absolutas y relativas.
- ✓ Dado lo último, comparar las categorías entre si es bastante simple.
- ✓ Adaptable a variables cuantitativas discretas.
- ✓✓✓ Al comparar dos o más gráficos de barras, las distribuciones de las variables subyacentes es bastante clara, más aún si se usan las frecuencias relativas.

[5] Esta ventaja es dada por la noción de altura necesaria para dimensionar la frecuencia de cada categoría.

Presentación gráfica

Datos cualitativos. Gráfico de torta - Desventajas

✗ "Fácilmente" distorsionable: al no realizar apropiadamente el gráfico, particularmente en la definición del eje de las frecuencias, visualmente pueden manipularse las frecuencias observadas.

Presentación gráfica

Datos cualitativos. Gráfico de torta - Desventajas

Obamacare: is a law enacted to ensure that all Americans have access to affordable health insurance

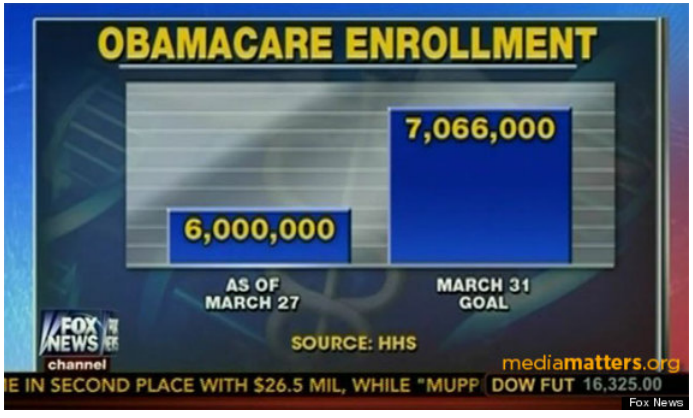


Figura 10: Gráfico sin eje de frecuencias

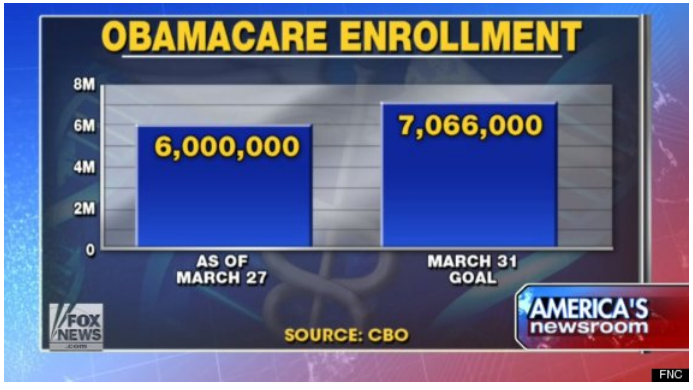


Figura 11: Gráfico originado en 0

Numero observado y esperado de inscritos al plan de seguro médico bajo la ley *Obamacare*. Tomado de *Snopes*.

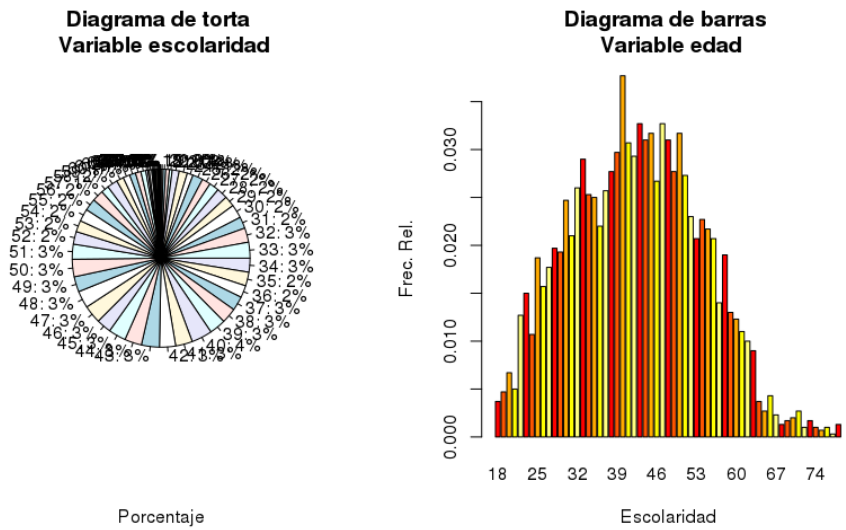
Presentación gráfica

Datos cualitativos y cuantitativos. Gráfico de barras

A tener en cuenta

- Cuándo se tienen variables cuantitativas discretas con bastantes categorías (como en el ejemplo de la variable *Edad*) es más común crear intervalos de clase que considerar los valores discretos como categorías de una variable cualitativa.

Figura 12: Diagrama de torta (izq.) y de barras (der.) para la variable *Edad*



- Para variables cuantitativas discretas, independiente si están en intervalos de clase o no, las barras adjacentes del diagrama de barras **no se tocan** entre ellas.

Presentación gráfica

Datos cualitativos y cuantitativos. Gráfico de barras

A tener en cuenta

- Es importante que las barras inicien en cero, tanto para frecuencias absolutas como relativas, esto para evitar una lectura errónea del gráfico.
- Una gráfica de barras en la que las barras están ordenadas de mayor a menor se denomina **gráfica de Pareto**.

Presentación gráfica

Datos cuantitativos. Tallo y hojas

Se determina el número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno[6]. A continuación se presentan los datos:

0 1 1 2 2 3 4 4 4 5 5 5 6 6 7 7 7 7 8 8 9 9 10 11 11 11 12 12 12 13 13 13 13 13 14 14 14 15 15 15 15 15 15 16 16 16 16 16 16 16 16 16 16 17 1717 18 18 18 18 19 19 20 20 21 21 22 22 23 23 24 24 24 24 25 25 26 26 27 28 28 29 30 30 35 40

¿Qué características tiene la distribución de los datos?

[6] Adaptado de *Fundamental Statistics for the Behavioral Sciences*

Presentación gráfica

Datos cuantitativos. Tallo y hojas. **Alternativa 1**

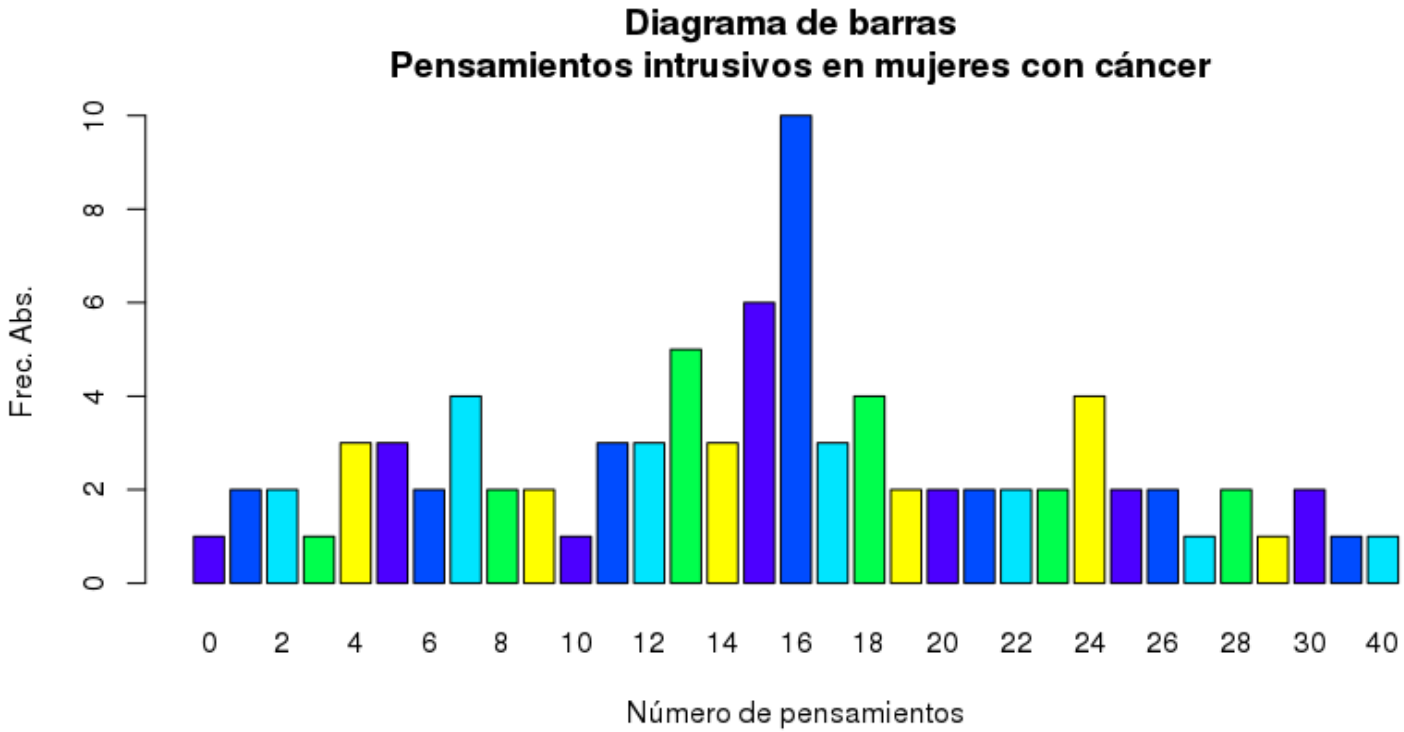
Cuadro 3: Distribución de frecuencias absolutas, relativas y acumuladas. Número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno

No. de Pensamientos Intrusivos	Frec. Abs.	Frec. Rel	Frec. Acum.
0	1	0.0116	0.0116
1	2	0.0233	0.0349
2	2	0.0233	0.0581
3	1	0.0116	0.0698
4	3	0.0349	0.1047
5	3	0.0349	0.1395
6	2	0.0233	0.1628
7	4	0.0465	0.2093
8	2	0.0233	0.2326
9	2	0.0233	0.2558
10	1	0.0116	0.2674
11	3	0.0349	0.3023
12	3	0.0349	0.3372
13	5	0.0581	0.3953
14	3	0.0349	0.4302
15	6	0.0698	0.5000
16	10	0.1163	0.6163
17	3	0.0349	0.6512
18	4	0.0465	0.6977
19	2	0.0233	0.7209
20	2	0.0233	0.7442
21	2	0.0233	0.7674
22	2	0.0233	0.7907
23	2	0.0233	0.8140
24	4	0.0465	0.8605
25	2	0.0233	0.8837
26	2	0.0233	0.9070
27	1	0.0116	0.9186
28	2	0.0233	0.9419
29	1	0.0116	0.9535
30	2	0.0233	0.9767
35	1	0.0116	0.9884
40	1	0.0116	1.0000

Presentación gráfica

Datos cuantitativos. Tallo y hojas. **Alternativa 2**

Figura 13: Gráfico de barras para el número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno



Presentación gráfica

Datos cuantitativos. Tallo y hojas. **Alternativa 3**

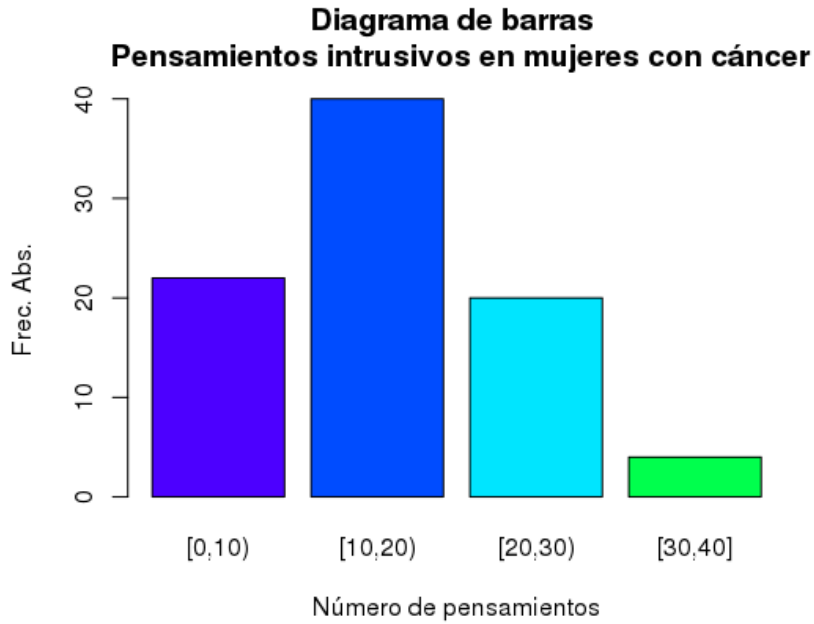
Cuadro 4: Distribución de frecuencias absolutas, relativas y acumuladas. Número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno en $k = 4$ intervalos de clase

Intervalo	Frec. Abs.	Frec. Rel	Frec. Acum.
$[0, 10)$	22	0.2558	0.2558
$[10, 20)$	40	0.4651	0.7209
$[20, 30)$	20	0.2326	0.9535
$[30, 40]$	4	0.0465	1.0000

Presentación gráfica

Datos cuantitativos. Tallo y hojas. **Alternativa 4**

Figura 14: Gráfico de barras para el número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno en $k = 4$ intervalos de clase



* Note que este gráfico no es más que la representación gráfica de la representación tabular de la frecuencia absoluta de la variable *Número de pensamientos intrusivos* en $k = 4$ intervalos de clase.

Presentación gráfica

Datos cuantitativos. Tallo y hojas

Podemos afirmar que:

1. Es evidente que las alternativas 1 y 2 no son muy informativas, en especial la 1.
2. Las alternativas 3 y 4 son más informativas, sin embargo, como se notaba anteriormente, se ha sacrificado información para ganar interpretabilidad con los intervalos de clase. ¿es esto necesario para este conjunto de datos?

Presentación gráfica

Datos cuantitativos. Tallo y hojas

Este gráfico presenta los datos usando los valores numéricos **reales** de cada punto de datos. Su elaboración esta dada como sigue:

- Divida cada segmento en dos partes: el tallo y las hojas.
- Ponga en lista los tallos en una columna, con una línea vertical a su derecha.
- Para cada medición, registre la parte de hoja en el mismo renglón con su tallo correspondiente.
- Ordene las hojas de menor a mayor en cada tallo.
- Dé una clave a su codificación de tallo y hoja.

Presentación gráfica

Datos cuantitativos. Tallo y hojas

Figura 15: Diagrama de tallo y hojas para el número de pensamientos intrusivos experimentados por mujeres recientemente diagnosticadas con cáncer de seno

```
> stem(thought)

  The decimal point is 1 digit(s) to the right of the |

  0 | 011223444
  0 | 5556677778899
  1 | 011122233333444
  1 | 555555666666666777888899
  2 | 001122334444
  2 | 55667889
  3 | 00
  3 | 5
  4 | 0

> stem(thought,scale=0.5)

  The decimal point is 1 digit(s) to the right of the |

  0 | 0112234445556677778899
  1 | 011122233333444555555666666666777888899
  2 | 00112233444455667889
  3 | 005
  4 | 0
```

Presentación gráfica

Datos cuantitativos. Tallo y hojas - Ventajas

- ✓ Conocido y fácil de explicar e interpretar.
- ✓ Son una alternativa a los gráficos de barras para variables cualitativas.
- ✓✓✓ Este gráfico no sacrifica información individual para obtener interpretabilidad de la distribución de los datos[7].

[7] Esto implica que con este gráfico se pueden recuperar fielmente los datos originales.

Presentación gráfica

Datos cuantitativos. Tallo y hojas - Desventajas

- ✗ Limitado a resumir visualmente un conjunto pequeño de datos.
- ✗ Al comparar dos o más gráficos de tallo y hojas, las distribuciones de las variables subyacentes está limitada al uso de las frecuencias absolutas.

Presentación gráfica

Datos cuantitativos. Histograma

Usado para variables cuantitativas continuas, el histograma es la versión continua del diagrama de barras, por lo cual:

- Las barras adyacentes **se tocan** entre ellas (para enfatizar la continuidad de la escala de medida).
- Su elaboración requiere la construcción de intervalos de clase y sus frecuencias (absolutas o relativas) correspondientes.

Presentación gráfica

Datos cuantitativos. Histograma

Además de la escolaridad y de la edad, se cuenta con el salario en dólares de la muestra de $n = 3000$ personas. Esta es una variable cuantitativa, que además es continua, por lo cual se procede con el histograma de la variable *Salario*.

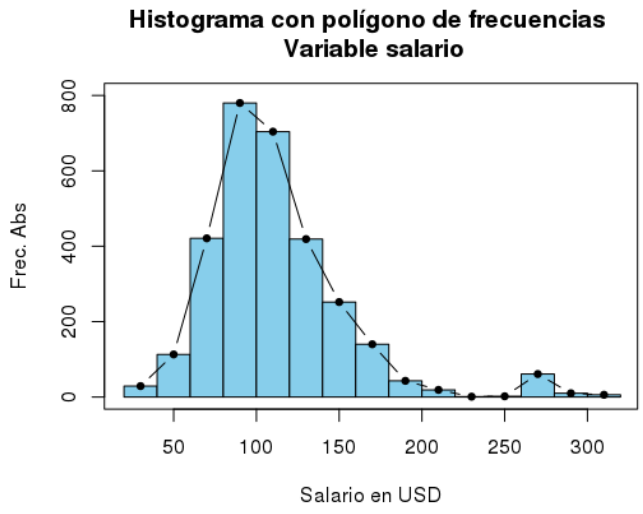


Figura 16: Histograma con **polígono de frecuencias** para la variable *salario*

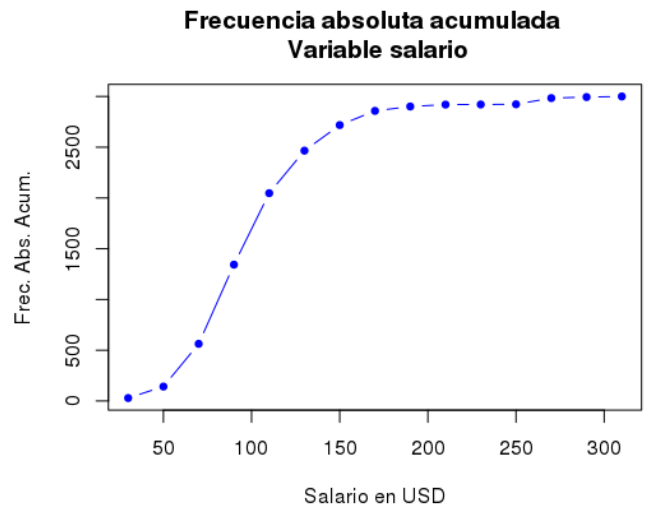


Figura 17: Frecuencia absoluta acumulada para la variable *Salario*

Presentación gráfica

Datos cuantitativos. Otros gráficos

Gráfica de líneas

Cuando una variable cuantitativa se registra en el tiempo a intervalos igualmente espaciados (por ejemplo diario, semanal o anual), el conjunto de datos forma una serie de tiempo. Los datos de una serie de tiempo se presentan en una **gráfica de líneas**.

Gráficas de puntos

Útil en pequeños conjuntos de datos. La **gráfica de puntos** apila los valores comunes de la variable de manera análoga a un diagrama de barras, pero con puntos.

Presentación tabular y gráfica

Conclusión

- Tanto variables cualitativas como cuantitativas son susceptibles a ser caracterizadas mediante la tabla estadística.
- Al categorizar una variable cuantitativa, se sacrifica el valor individual de cada observación, pero puede ganarse interpretabilidad en el análisis.
- Elementos clave en la categorización de una variable cuantitativa: intervalo y marca de clase, ancho de clase, rango de la variable.

Ejemplo de aplicación

Segundo laboratorio de programación en R

Realice el segundo laboratorio de programación en R.

Contenido

Interpretación

Descripción detallada de los resultados

- **Escalas:** Verificar las escalas horizontales y verticales, de manera que haya claridad respecto a lo que se mide.
- **Centro:** ¿Dónde preponderan los datos? [8] Si se comparan dos distribuciones, ¿están centradas en el mismo lugar?
- **Dispersión:** Respecto al punto medio, ¿qué tan dispersos se encuentran los datos?

[8] Es decir, cual es la *ubicación* de la distribución en la escala de medición.

Interpretación

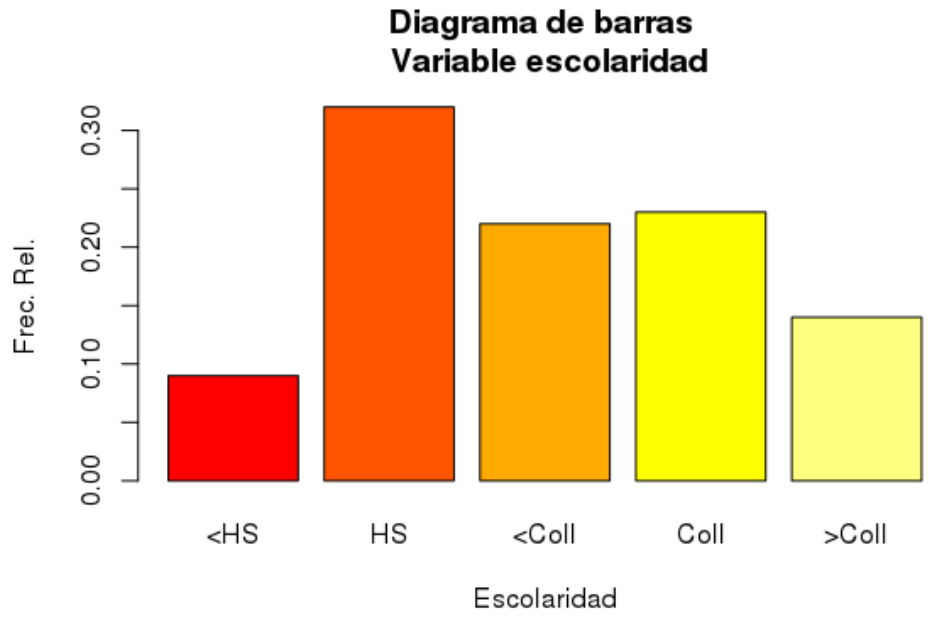
Descripción detallada de los resultados

- **Localización:** ¿Dónde están ubicados los datos? ¿cuáles valores son máximos y mínimos? ¿Hay otros valores de interés en la escala de medición que caractericen la distribución?
- **Forma:** ¿La distribución tiene un "pico", un punto que es más alto que cualquier otro? ¿Hay más de un pico? ¿Hay un número aproximadamente igual de mediciones a la izquierda y derecha del pico? ¿qué tan frecuentes son los valores alejados del valor central?
- **Atipicidades:** ¿hay mediciones mucho mayores o menores que todas las otras?.

Interpretación

Descripción detallada de los resultados

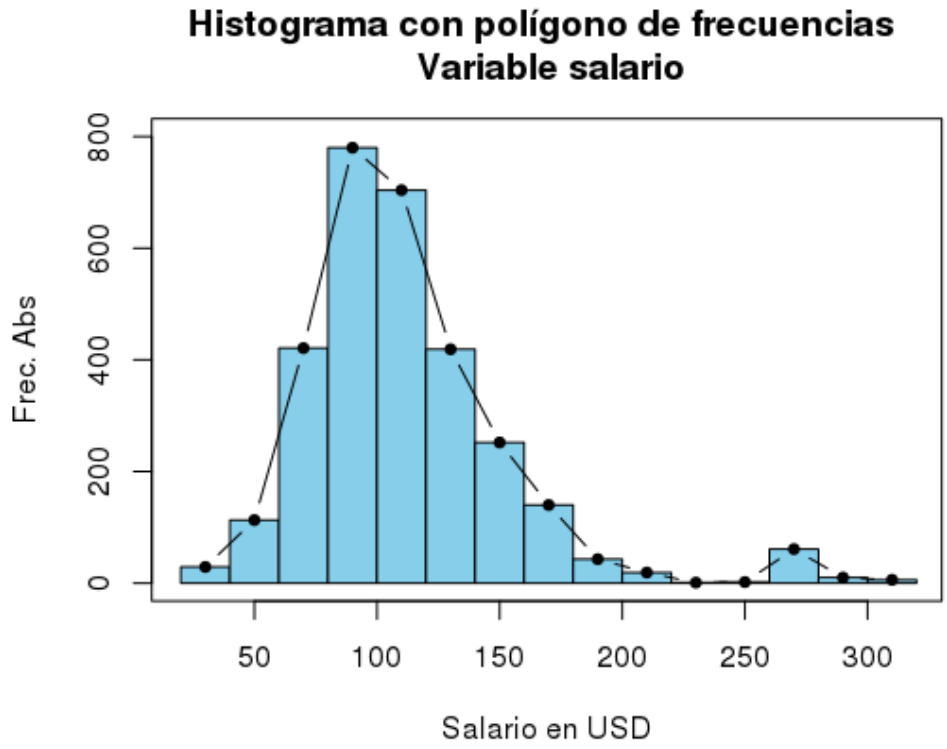
Figura 18: Diagrama de barras para la variable *Escolaridad*



Interpretación

Descripción detallada de los resultados

Figura 19: Histograma con polígono de frecuencias para la variable *Salario*



Interpretación

Descripción detallada de los resultados

Simetría

Una distribución es simétrica si los lados izquierdo y derecho de la distribución, cuando se divide en el valor medio, forman imágenes espejo.

Sesgo a derecha (o asimetría positiva)

Una distribución está sesgada a la derecha si una proporción más grande de las mediciones se encuentra a la derecha del valor pico.

Sesgo a izquerda (o asimetría negativa)

Una distribución está sesgada a la izquierda si una proporción mayor de las mediciones está a la izquierda del valor pico.

Interpretación

Descripción detallada de los resultados

Curtosis

La curtosis refleja el grado de concentración de los datos respecto a la parte central de las distribuciones con forma de campana, estas son:

1. Platicurticas (curtosis negativa).
2. Mesocurticas (curtosis cero).
3. Leptocurticas (curtosis positiva).

Entre mayor concentración se encuentre en los datos, más apuntado es el polígono de frecuencias correspondiente, implicando una mayor curtosis.

Interpretación

Descripción detallada de los resultados

Unimodal/bimodal

Una distribución es unimodal si tiene un pico; una distribución bimodal tiene dos picos. Las distribuciones bimodales representan a veces una combinación de dos poblaciones diferentes.

Presentación gráfica e interpretación

Conclusión

1. La presentación gráfica permite transmitir visualmente y de múltiples formas la distribución de los datos.
2. Se debe tener clara la tipología de las variables para una graficación efectiva.
3. Existen variadas características al describir la distribucipon de los datos: centro, dispersión, localización, forma y atipicidades.

Ejemplo de aplicación

Ejercicios grupales

Conforme grupos de 4 personas, cada grupo debe tener como nombre uno de los conceptos aprendidos en clase.

Ejercicios grupales

1.1 Según lo discutido en clase responda:

a. ¿Cuál es la principal diferencia entre la estadística y la matemática?
b. Explique 3 razones por las cuales se cuestiona la posibilidad inferencial en los sondeos electorales de las firmas encuestadoras del plebiscito por la paz en Colombia
c. ¿Es necesaria la inferencia estadística cuando se observa toda la población?

1.2 Responda Verdadero (V) o Falso (F) según corresponda:

a. En el estudio de la estadística se presenta un error inherente que en la práctica es despreciable
b. La característica común entre las 4 escalas de medida estudiadas (nominal, ordinal, intervalo y razón) es que entre las modalidades de respuesta se cumple la relación de igualdad o desigualdad
c. La escala de intervalo y de razón proveen el mismo nivel de sofisticación.

Ejercicios grupales

2.1 Identifique las unidades experimentales en las que se miden las siguientes variables. Adicionalmente, determine el tipo de variable involucrada y la escala de medición de la misma:

a. Tamaño del tumor cancerígeno de un paciente.
b. Intención de voto para las elecciones presidenciales.
c. Estadío del cáncer en un paciente.

2.2 Identifique cada una de las variables cuantitativas como discretas o continuas:

a. Número de accidentes en botes en un tramo de 50 millas del río.
b. Tiempo para completar un cuestionario.
c. Rendimiento en kilogramos de una cosecha de papas.

2.3 Un investigador médico desea estimar el tiempo de supervivencia de un paciente con cáncer después de un régimen particular de radioterapia:

a. ¿Cuál es la variable de interés para el investigador médico?
b. ¿La variable del inciso anterior es cualitativa, cuantitativa discreta o cuantitativa continua?
c. Identifique la población de interés para el investigador médico.

Ejercicios grupales

3.1 Responda Verdadero (V) o Falso (F) según corresponda **y explique**:

a. A diferencia del diagrama de tallo y hojas, el histograma permite recuperar los valores individuales de la variable de interés.

b. El número de intervalos seleccionados para la construcción de un histograma debe seleccionarse cuidadosamente y siempre de manera experta (es decir, por parte del investigador).

Ejercicios grupales

4.1 Se quiere estudiar el número de horas que emplean los estudiantes de PEF en transportarse diariamente. Por el gran volumen de estudiantes inscritos en la asignatura, se decide encuestar únicamente a 50 estudiantes. A continuación se muestra el conjunto de datos recolectados:

1, 2, 2, 3, 1, 3, 4, 2, 2, 1
1, 1, 2, 2, 3, 2, 2, 5, 2, 3
4, 1, 2, 1, 1, 2, 1, 2, 3, 1
2, 2, 2, 1, 3, 2, 3, 1, 2, 2
3, 2, 3, 2, 2, 3, 1, 2, 2, 2

a. ¿Es este conjunto de mediciones una población o una muestra?.
b. Identifique el tipo y la escala de al variable.
c. Realice la representación tabular de la variable. Tenga en cuenta el tipo de variable y su escala de medida.
d. Represente gráficamente el comportamiento de los datos. Tenga en cuenta la misma recomendación del numeral anterior.