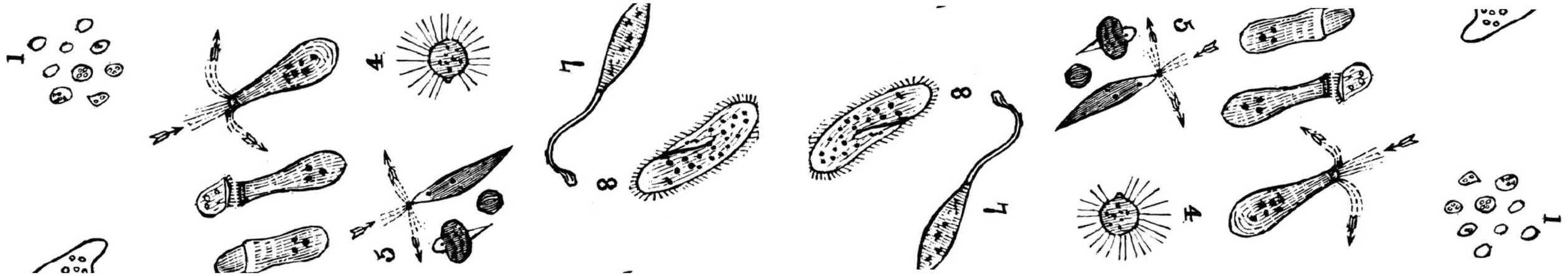
Histogramas y gráficos de densidad

www.castrolab.org
www.cbib.cl
www.ucdavischile.org

Eduardo Castro, PhD
Universidad Andrés Bello
26 de marzo de 2018

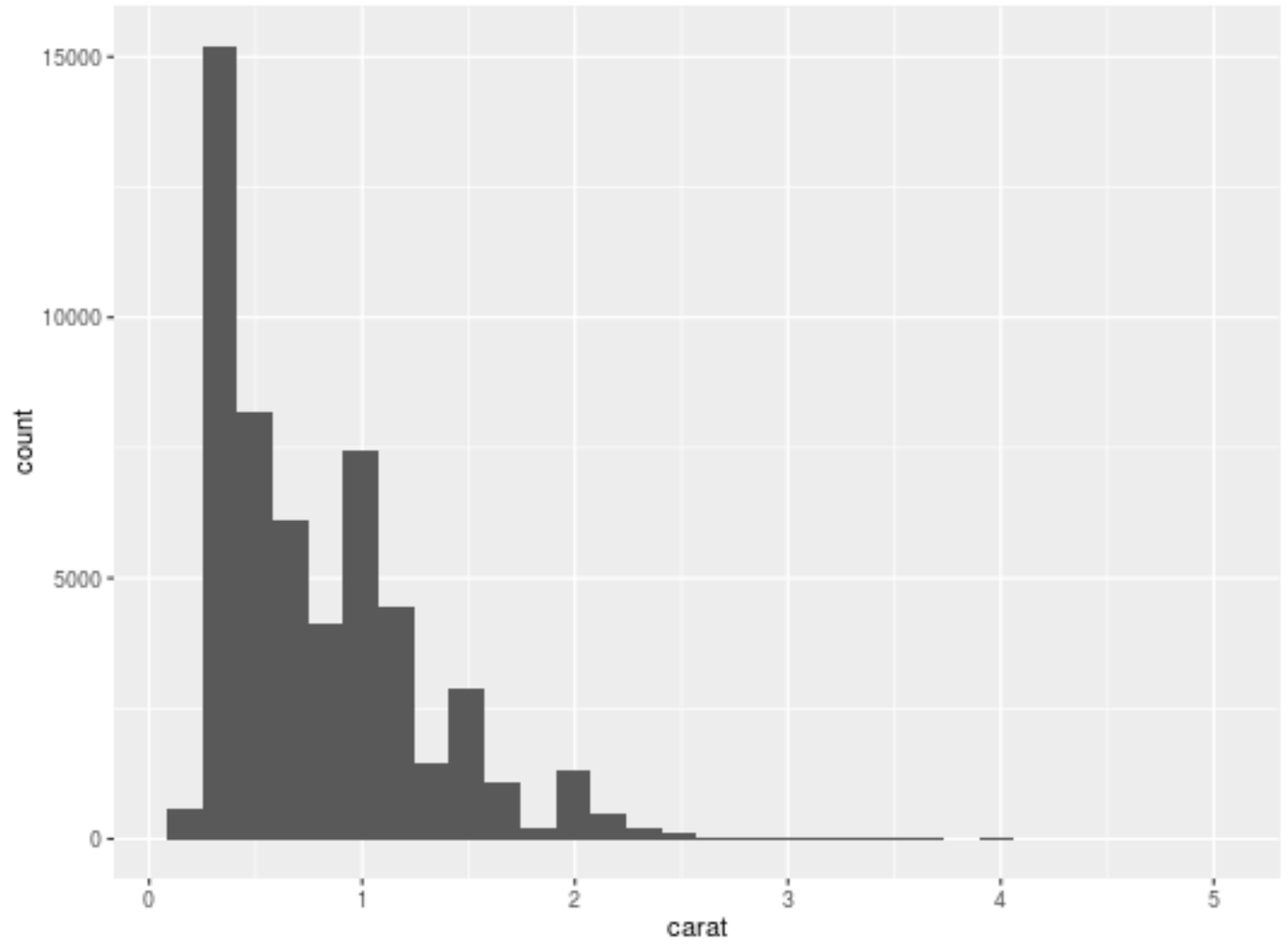


Revisar tarea

- Código y PDF/HTML del documento Tufte o Tufte handout

¿Qué es un histograma?

- Es un gráfico que muestra la frecuencia de un conjunto de **datos continuos y discretos**



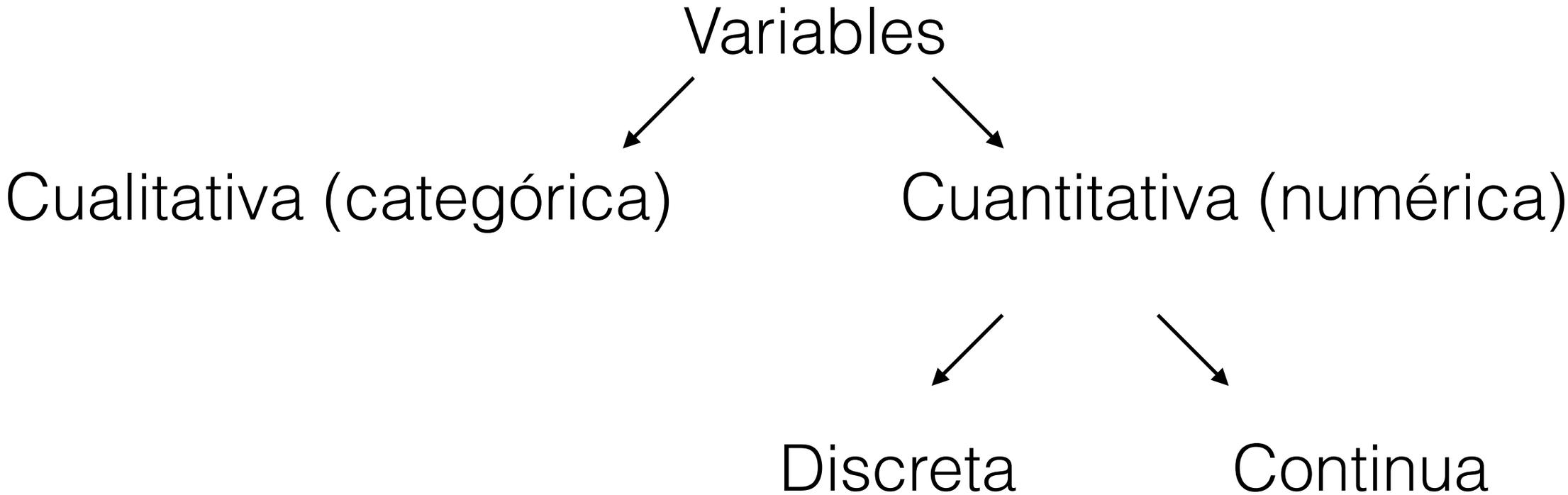
Datos continuos

diamonds

Filter

	carat	cut	color	clarity	depth	table	price	x	y	z
1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61.0	338	4.00	4.05	2.39
11	0.30	Good	J	SI1	64.0	55.0	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56.0	340	3.93	3.90	2.46
13	0.22	Premium	F	SI1	60.4	61.0	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54.0	344	4.35	4.37	2.71
15	0.20	Premium	E	SI2	60.2	62.0	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58.0	345	4.38	4.42	2.68
17	0.30	Ideal	I	SI2	62.0	54.0	348	4.31	4.34	2.68

Tipos de variables



*Si no se puede promediar es cualitativa

¿Para qué sirve?

- Al mostrarnos la distribución de los datos, nos da una idea sobre la “normalidad” de los datos, posibles valores atípicos, o algún tipo de sesgo.

¿Para qué sirve?

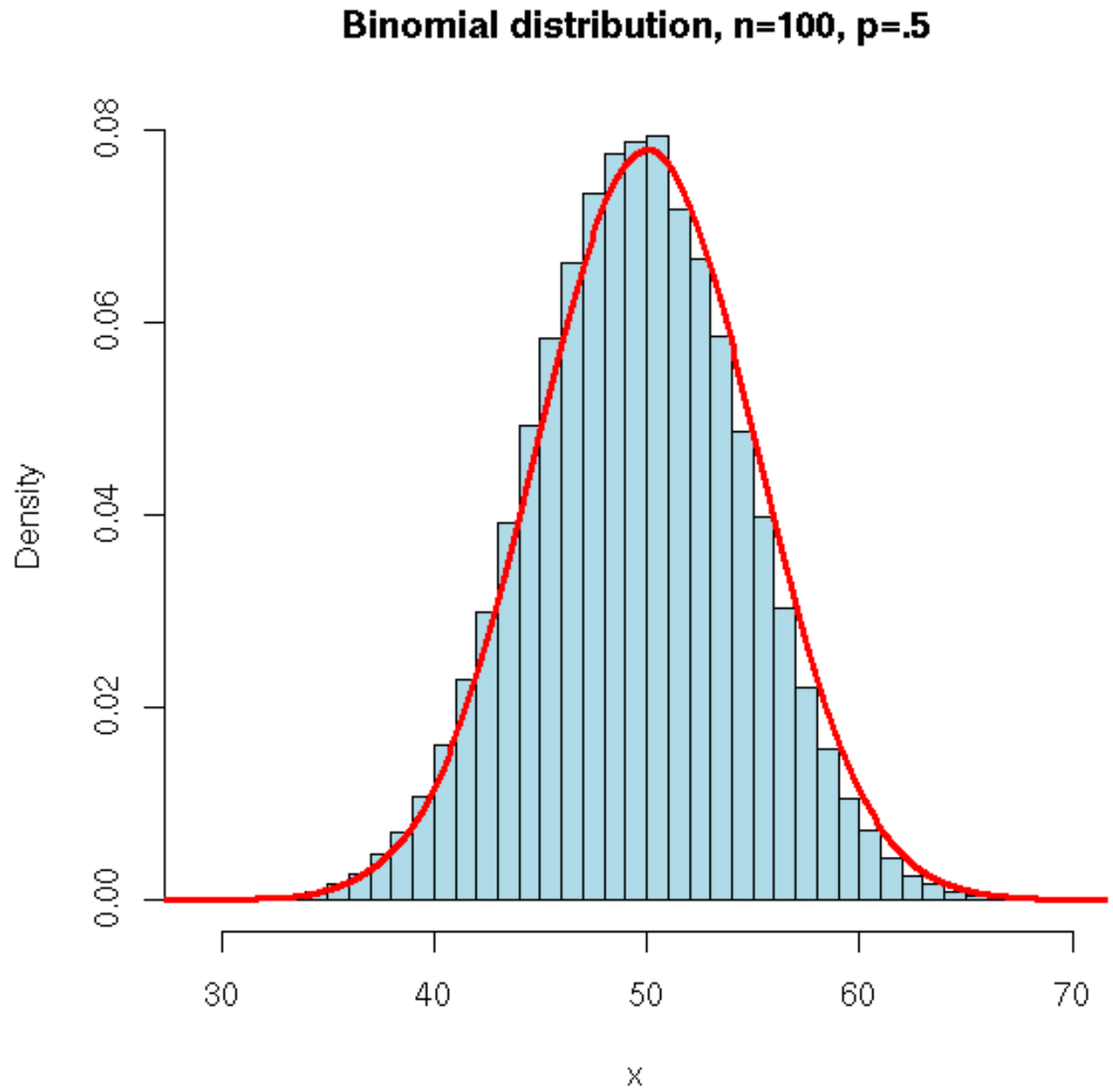
- Para aproximar la distribución de probabilidad de una variable al representar la frecuencia de observaciones dentro de ciertos rangos de valores

- Distribución de probabilidad = una función cuya integral en cualquier intervalo es la probabilidad de una variable

Pearson, K. (1895). "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **186**: 343–414

¿Para qué sirve?

- Ejemplo

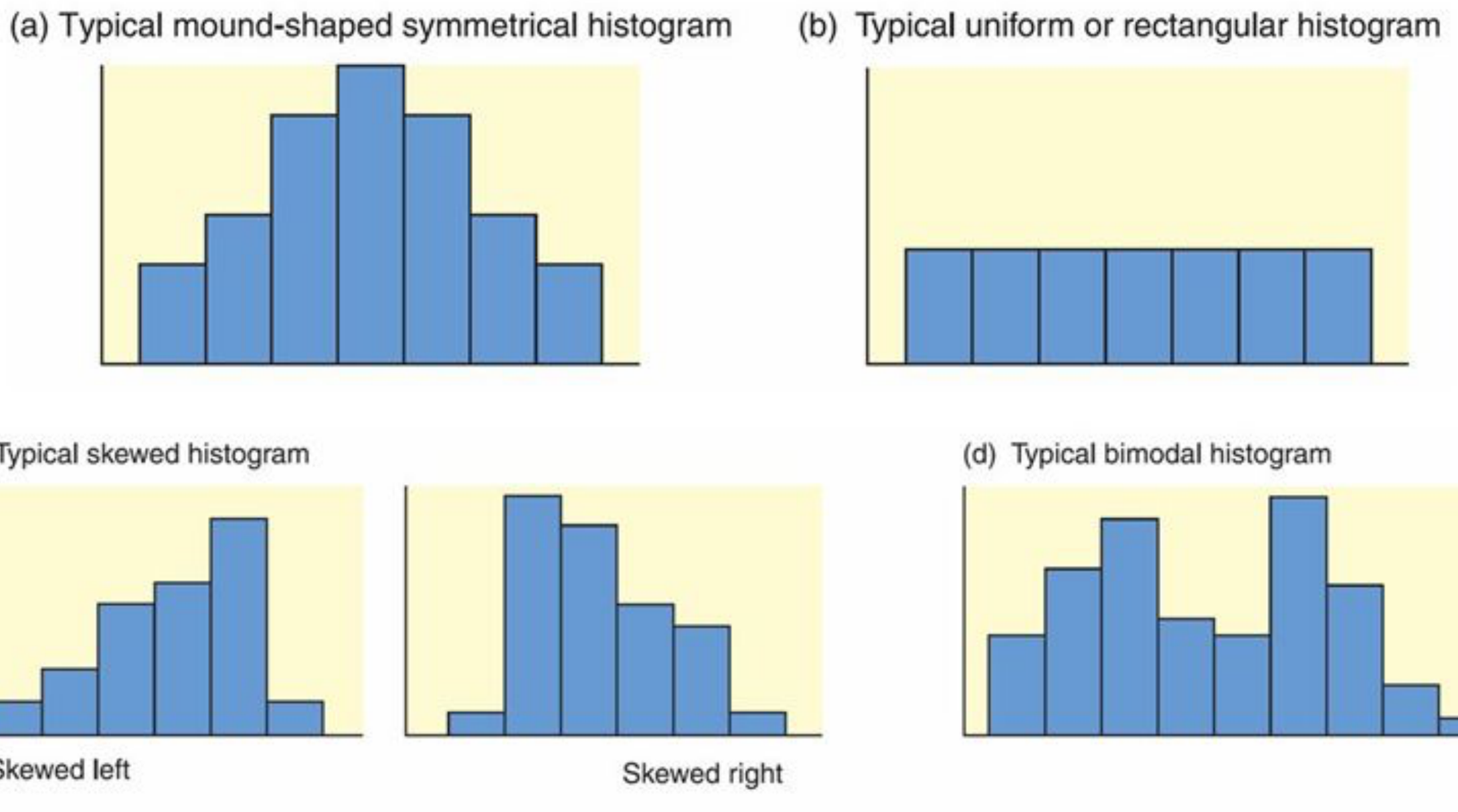


http://zoonek2.free.fr/UNIX/48_R/07.html

Tres tipos de histogramas

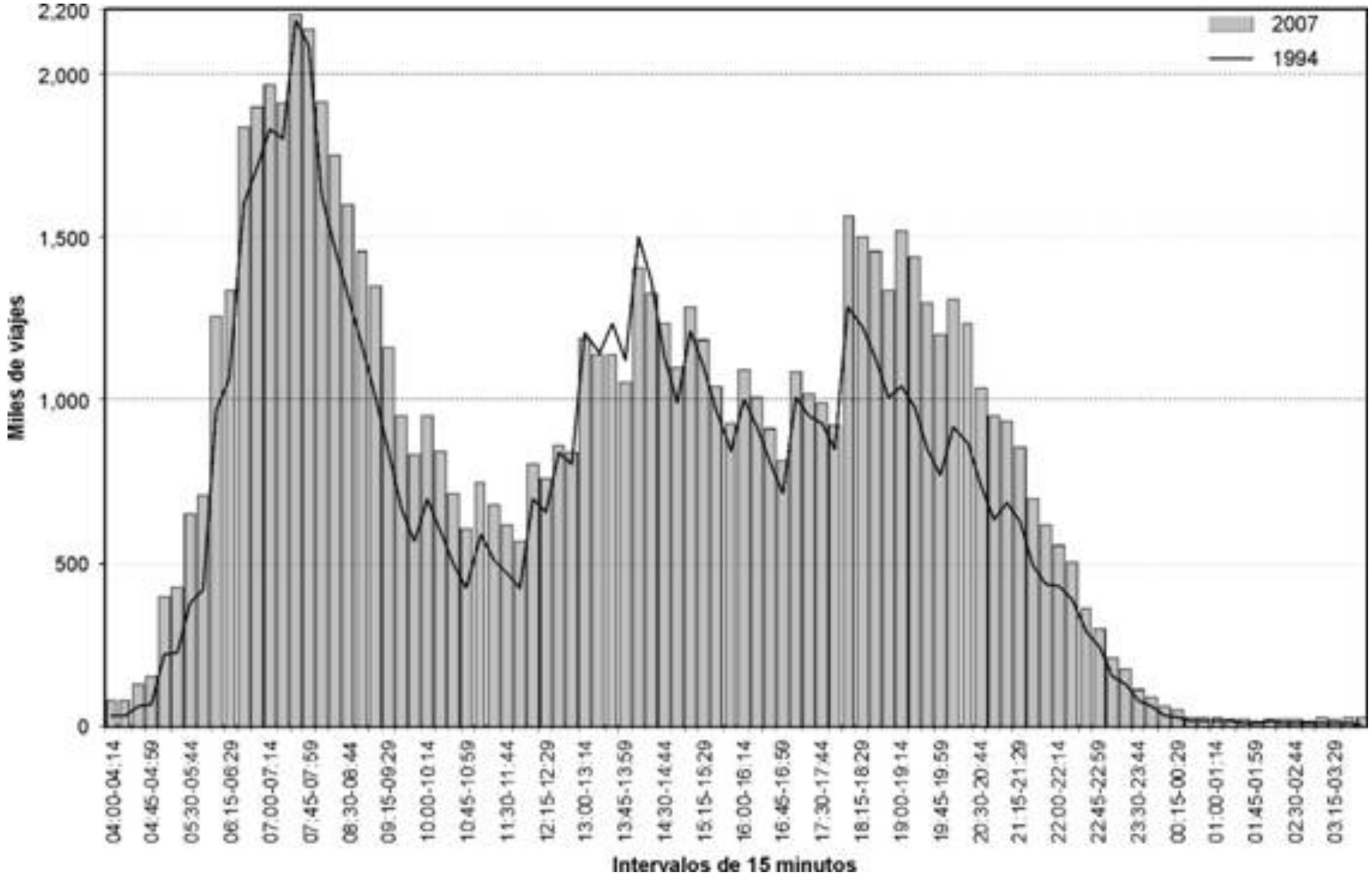
- De densidad = muestra proporciones en el eje Y
- De frecuencia = muestra el número de observaciones
- De frecuencia relativa = muestra el porcentaje de observaciones

Formas de distintas distribuciones - asimetría



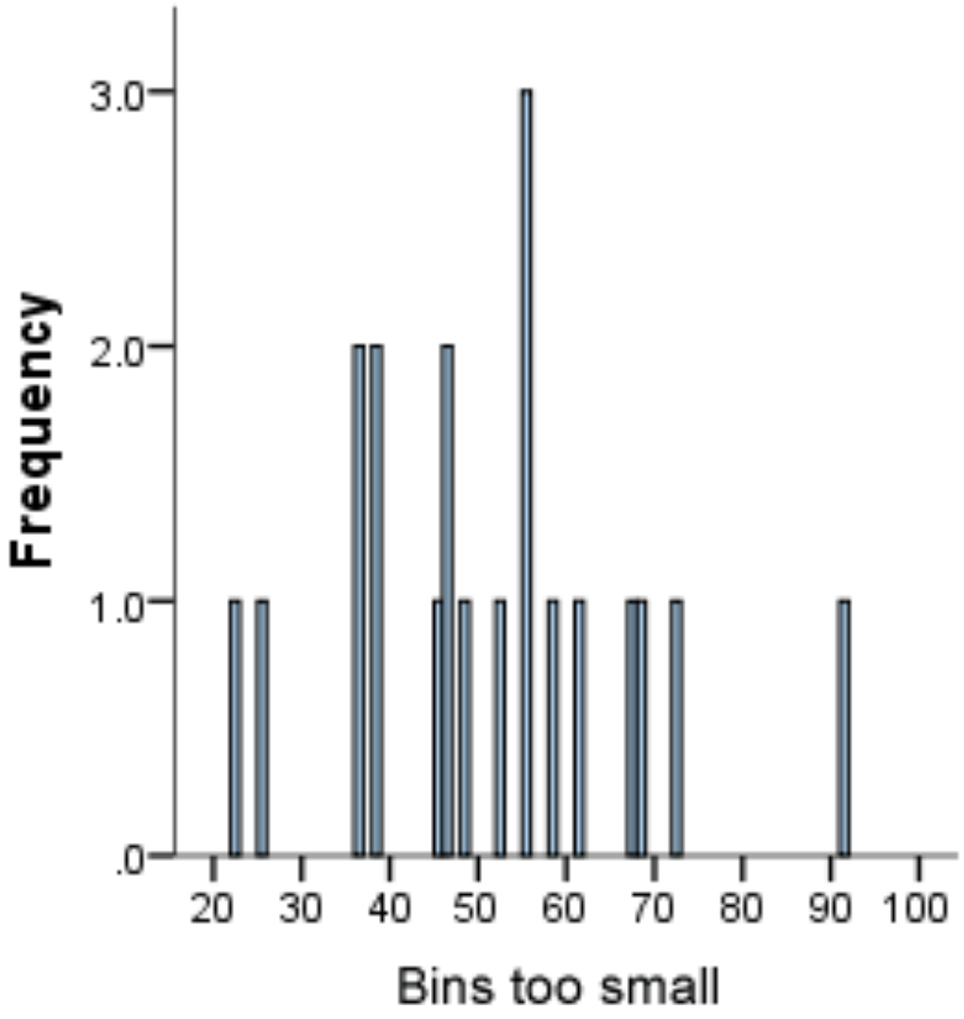
http://slideplayer.com/slide/6642460/

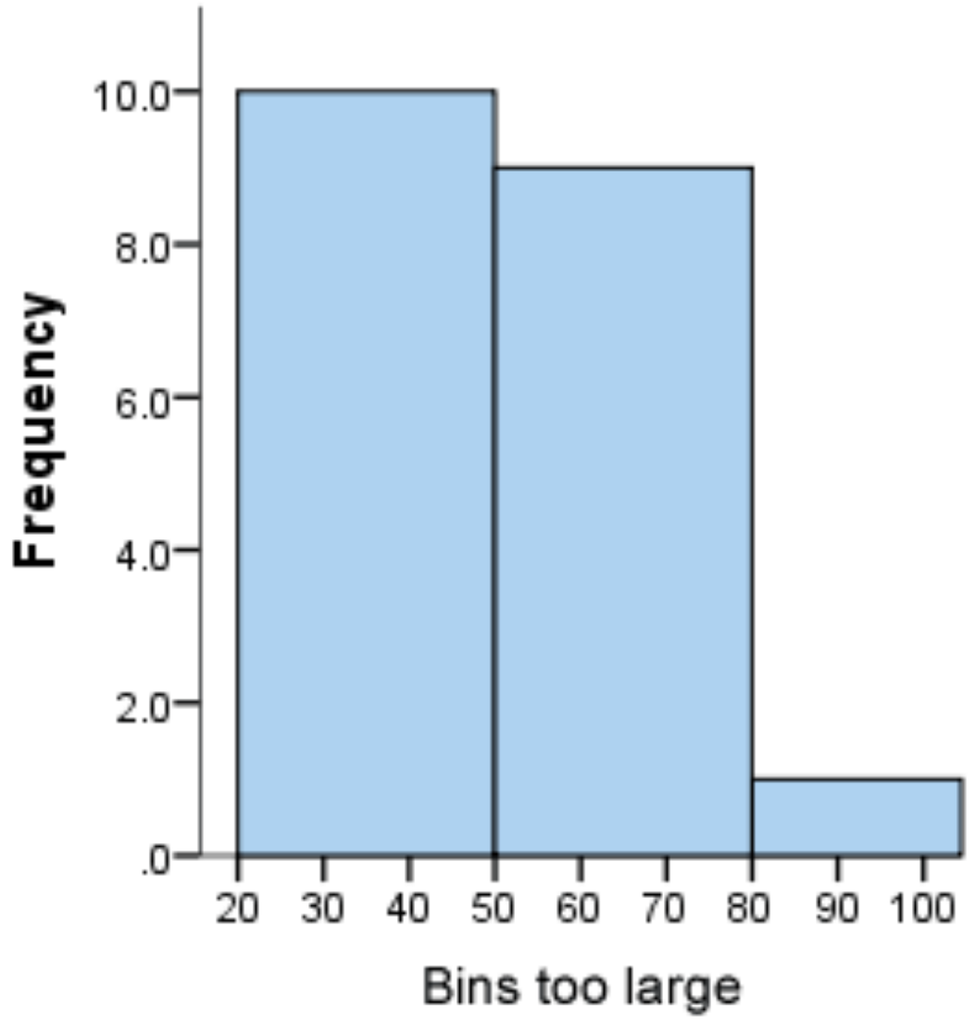
Formas de distintas distribuciones - asimetría



http://www.ub.edu/geocrit/sn/sn-487.htm

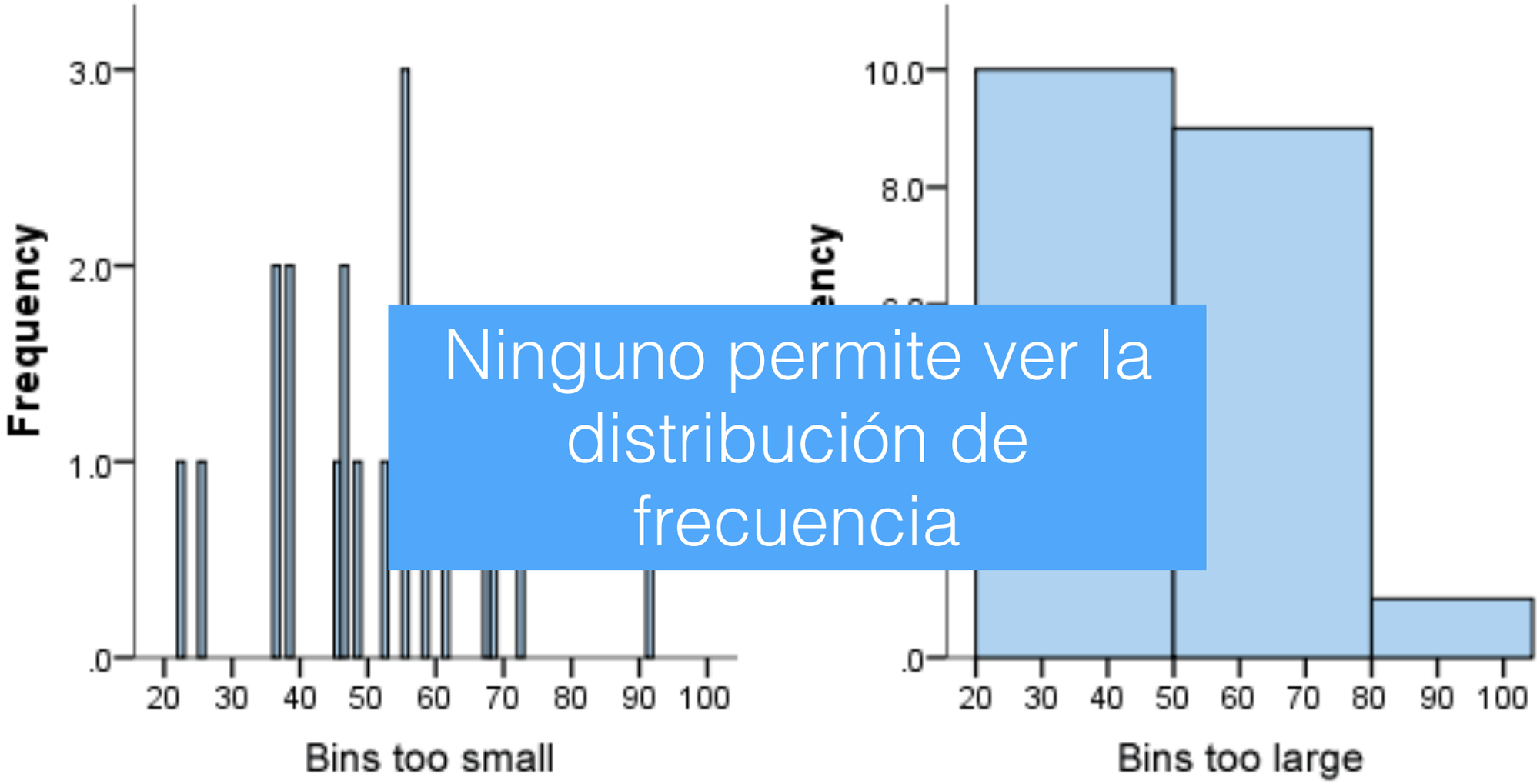
Binning - rangos de datos





https://statistics.laerd.com/statistical-guides/understanding-histograms.php

Binning - rangos de datos



https://statistics.laerd.com/statistical-guides/understanding-histograms.php

¿Cómo se calcula el número de bins?

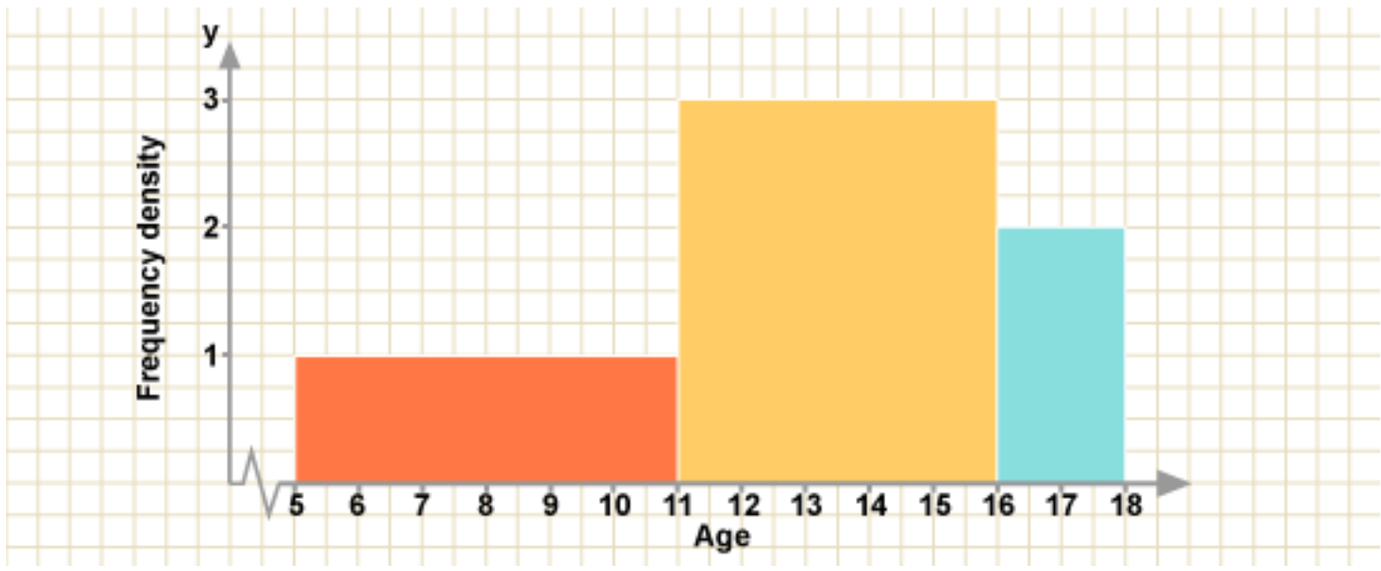
- Ejemplo: usando la fórmula de Sturges,

$$k = 1 + 3.322 \log10(n)$$

donde k es el número de bins y n es el número de observaciones

Sturges, H. A. (1926). "The choice of a class interval". Journal of the American Statistical Association: 65–66. doi:10.1080/01621459.1926.10502161

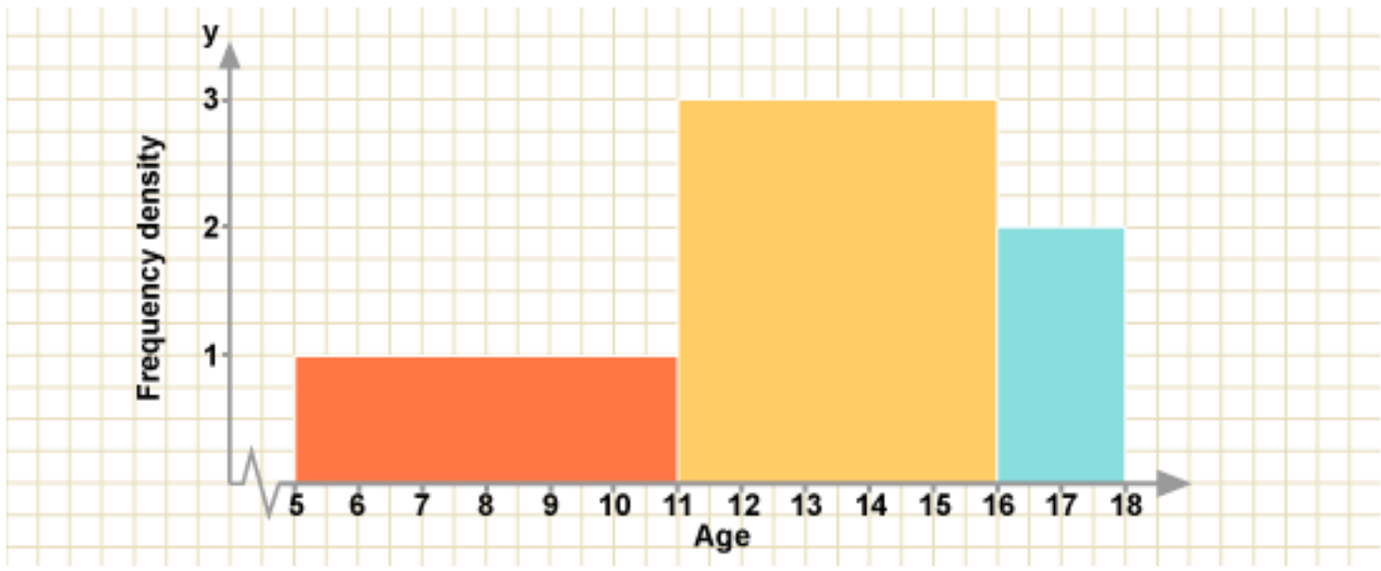
¿Cómo se compara un histograma? ¿Área o altura?



http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hirev3.shtml

¿Cómo se compara un histograma? ¿Área o altura?

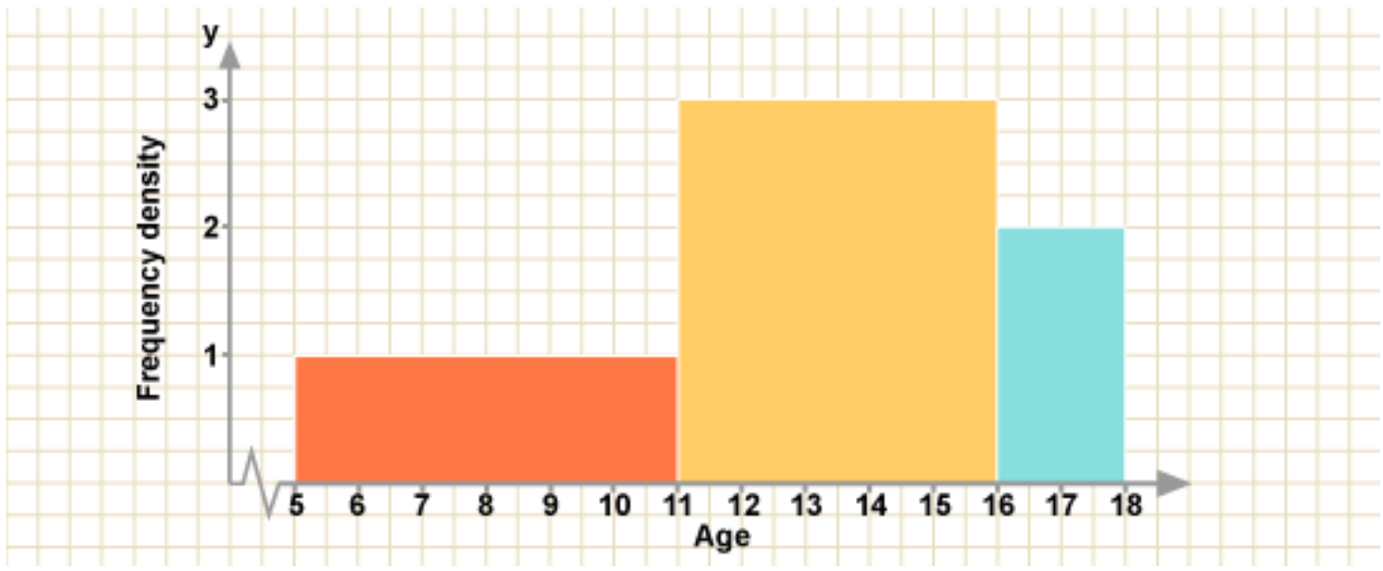
Por el área



http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hirev3.shtml

¿Cómo se compara un histograma? ¿Área o altura?

Por el área
Si los rangos son iguales da lo mismo



http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hirev3.shtml

Pregunta de prueba

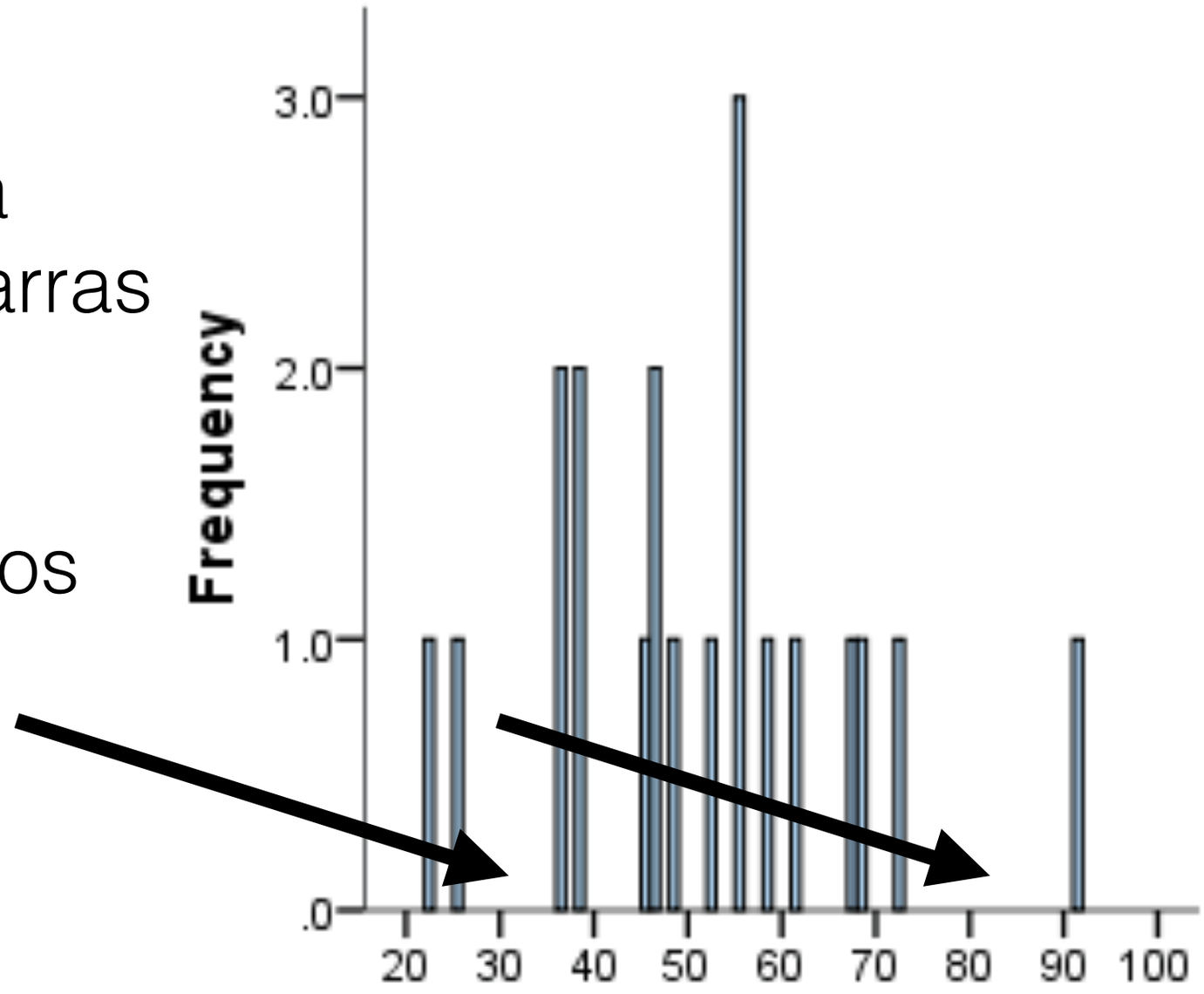
- ¿Un gráfico de barras y un histograma son lo mismo?

Pregunta de prueba

- ¿Qué son los espacios blancos?

  En un histograma nunca hay espacio entre las barras

  Si no hay una barra significa que no hay datos en ese bin o rango



Gráficos de densidad

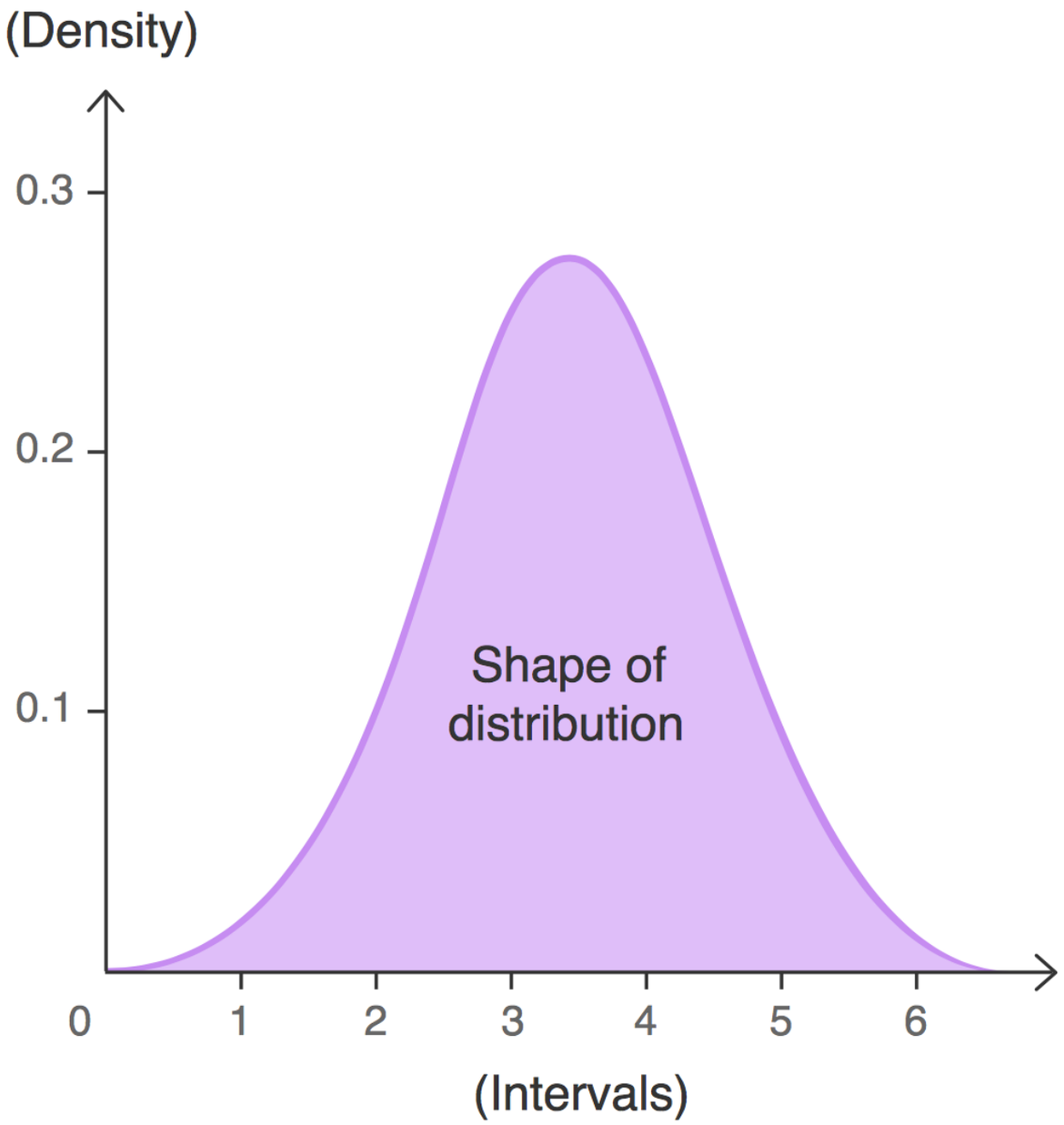
- ¿A qué nos referimos con densidad?
- Es la construcción de un estimado, basándose en datos observados, de una función de probabilidad no observable
- No podemos observar una probabilidad, pero la podemos estimar a partir de la frecuencia en que observamos un evento

Ejemplos de gráficos de densidad

- El área bajo la curva suma 1

- El área de un intervalo en x es la probabilidad

- Ventaja sobre histogramas, no son afectados por el número de bins

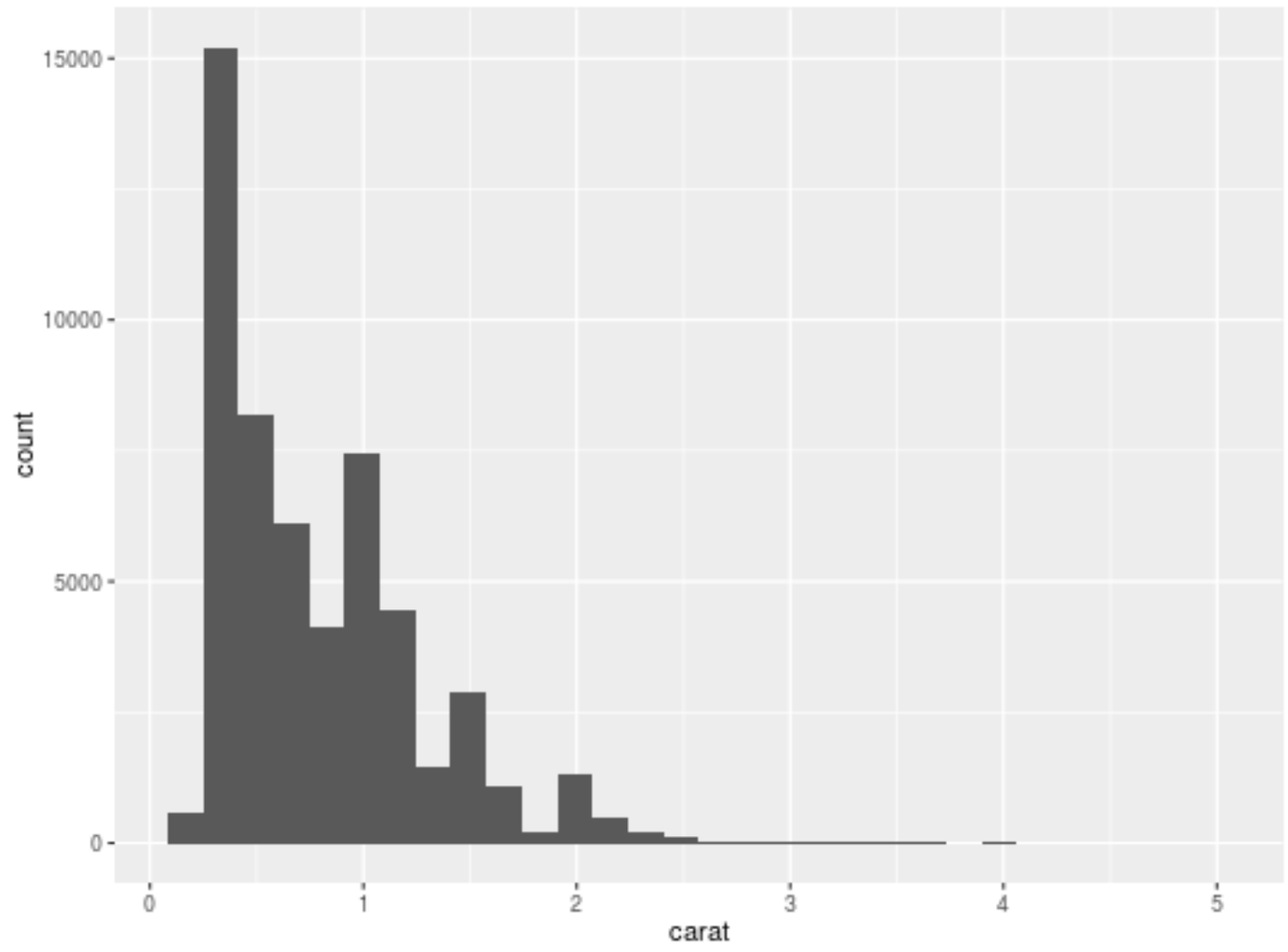


Gráficos de densidad

- Histogramas se hacían antes de que existieran computadores, ahora se hacen gráficos de densidad

Ahora a RStudio

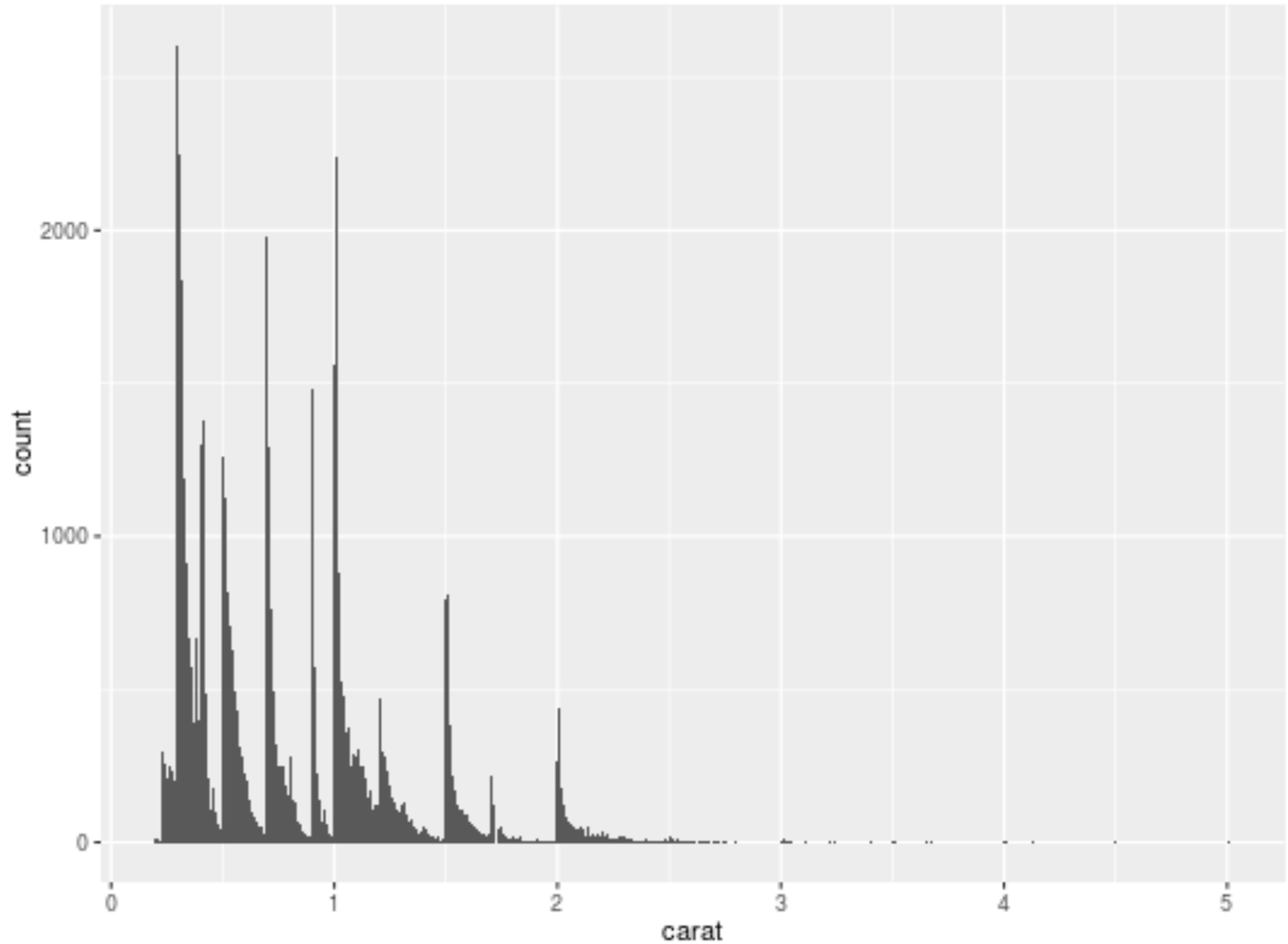
- library(ggplot2)
- data(“diamonds”)
- ggplot(diamonds, aes(carat)) + geom_histogram()



Efecto del ancho del bin

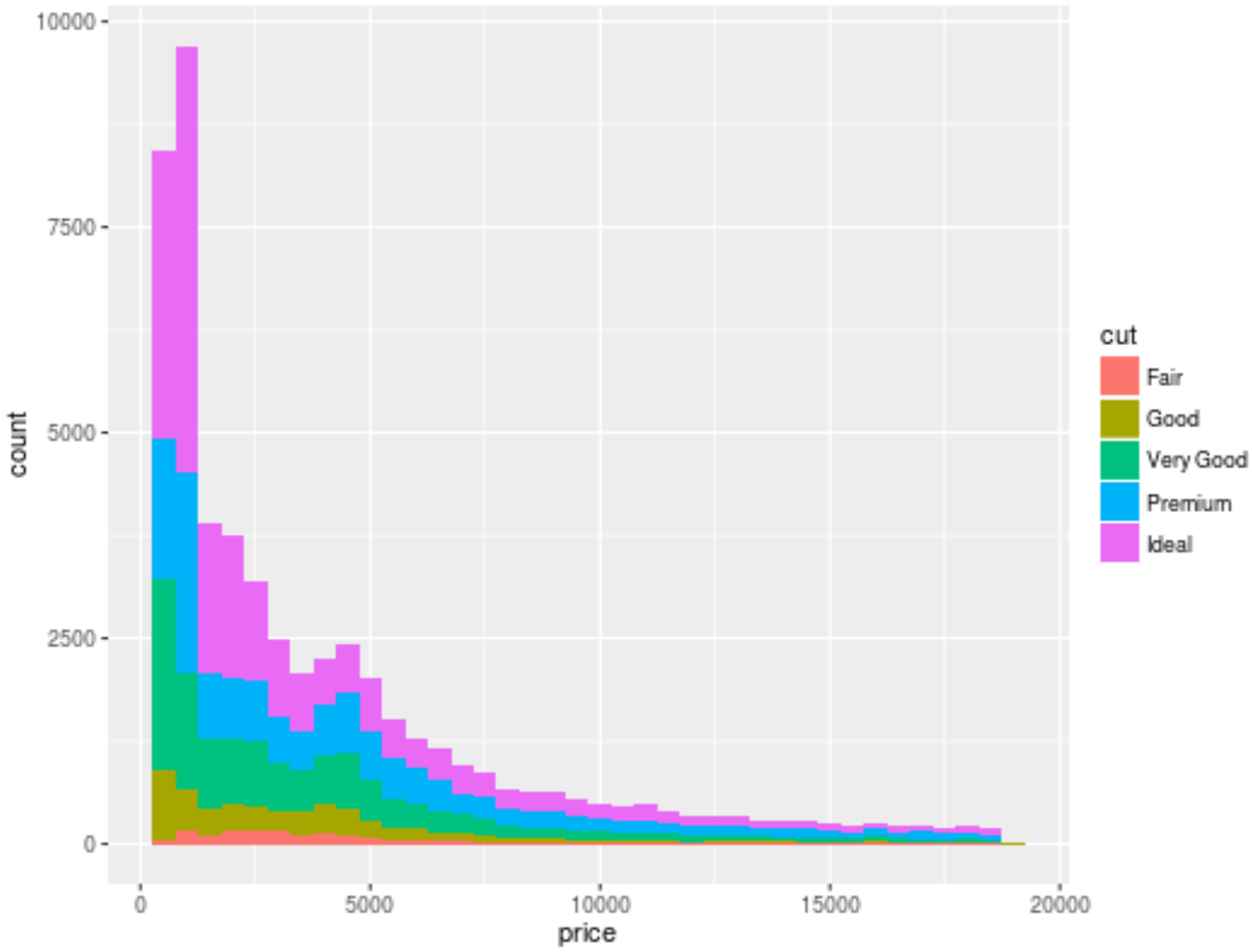
- ggplot(diamonds, aes(carat)) + geom_histogram(binwidth = 0.01)

- ggplot(diamonds, aes(carat)) + geom_histogram(bins = 200)



Destacando la composición de cada bin

- ggplot(diamonds, aes(price, fill = cut)) + geom_histogram(binwidth = 500)



Ahora probemos con geom_density

```
ggplot(diamonds, aes(carat)) + geom_density()
ggplot(diamonds, aes(carat)) + geom_density(adjust = 1/5)
ggplot(diamonds, aes(depth, fill = cut, colour = cut)) +
  geom_density(alpha = 0.1) +
  xlim(55, 70)
ggplot(diamonds, aes(carat, fill = cut)) +
  geom_density(position = "stack")
ggplot(diamonds, aes(carat, ..count.., fill = cut)) +
  geom_density(position = "fill")
```