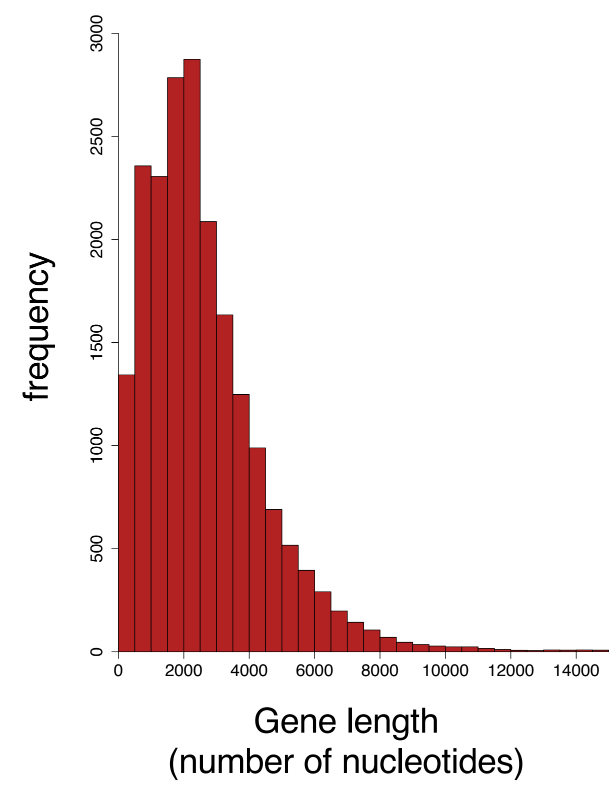
Gaining further insights into data and biological problems (experimental or observational)

Displaying numerical data in the form of frequency distributions: table and histograms & other visual aids to understand the characteristics of data.



Some raw data: Abundance of birds across species

Table 2.2-2 Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance
Greater roadrunner	1
Black-chinned hummingbird	1
Western kingbird	1
Great-tailed grackle	1
Bronzed cowbird	1
Great horned owl	2
Costa's hummingbird	2
Canyon wren	2
Canyon towhee	2
Harris's hawk	3
Loggerhead shrike	3
Hooded oriole	4
Northern mockingbird	5
American kestrel	7
Rock dove	7
Bell's vireo	10
Common raven	12
Northern cardinal	13
House sparrow	14
Ladder-backed woodpecker	15
Red-tailed hawk	16
Phainopepla	18
Turkey vulture	23
Violet-green swallow	23
Lesser nighthawk	25
Scott's oriole	28
Purple martin	33
Black-throated sparrow	33
Brown-headed cowbird	59
Black vulture	64
Lucy's warbler	67
Gilded flicker	77
Brown-crested flycatcher	128
Mourning dove	135
Gambel's quail	148
Black-tailed gnatcatcher	152
Ash-throated flycatcher	173
Curve-billed thrasher	173
Cactus wren	230
Verdin	282
House finch	297
Gila woodpecker	300
White-winged dove	625

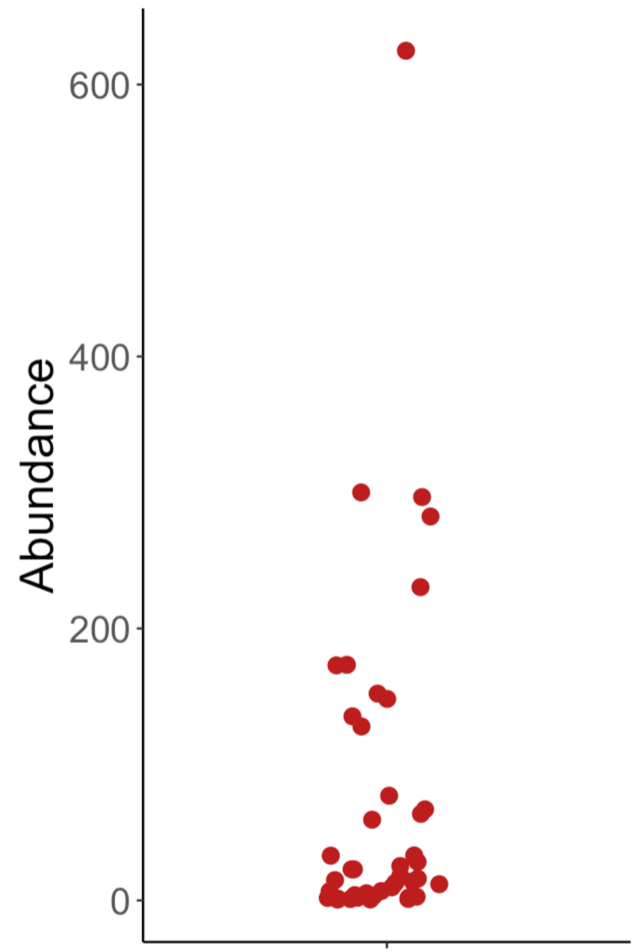



Total = 3215 individuals

Abundance of birds across species - plot of raw data

Table 2.2-2 Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance	Species	Abundance
Greater roadrunner	1	Turkey vulture	23
Black-chinned hummingbird	1	Violet-green swallow	23
Western kingbird	1	Lesser nighthawk	25
Great-tailed grackle	1	Scott's oriole	28
Bronzed cowbird	1	Purple martin	33
Great horned owl	2	Black-throated sparrow	33
Costa's hummingbird	2	Brown-headed cowbird	59
Canyon wren	2	Black vulture	64
Canyon towhee	2	Lucy's warbler	67
Harris's hawk	3	Gilded flicker	77
Loggerhead shrike	3	Brown-crested flycatcher	128
Hooded oriole	4	Mourning dove	135
Northern mockingbird	5	Gambel's quail	148
American kestrel	7	Black-tailed gnatcatcher	152
Rock dove	7	Ash-throated flycatcher	173
Bell's vireo	10	Curve-billed thrasher	173
Common raven	12	Cactus wren	230
Northern cardinal	13	Verdin	282
House sparrow	14	House finch	297
Ladder-backed woodpecker	15	Gila woodpecker	300
Red-tailed hawk	16	White-winged dove	625
Phainopepla	18		



Stripchart
"one dimensional scatter plot"

Displaying numerical data in the form of frequency distributions – the tabular (table) form

Table 2.2-2 Data on the abundance of each species of bird encountered during four surveys in Organ Pipe Cactus National Monument.

Species	Abundance	Species	Abundance
Greater roadrunner	1	Turkey vulture	23
Black-chinned hummingbird	1	Violet-green swallow	23
Western kingbird	1	Lesser nighthawk	25
Great-tailed grackle	1	Scott's oriole	28
Bronzed cowbird	1	Purple martin	33
Great horned owl	2	Black-throated sparrow	33
Costa's hummingbird	2	Brown-headed cowbird	59
Canyon wren	2	Black vulture	64
Canyon towhee	2	Lucy's warbler	67
Harris's hawk	3	Gilded flicker	77
Loggerhead shrike	3	Brown-crested flycatcher	128
Hooded oriole	4	Mourning dove	135
Northern mockingbird	5	Gambel's quail	148
American kestrel	7	Black-tailed gnatcatcher	152
Rock dove	7	Ash-throated flycatcher	173
Bell's vireo	10	Curve-billed thrasher	173
Common raven	12	Cactus wren	230
Northern cardinal	13	Verdin	282
House sparrow	14	House finch	297
Ladder-backed woodpecker	15	Gila woodpecker	300
Red-tailed hawk	16	White-winged dove	625
Phainopepla	18		

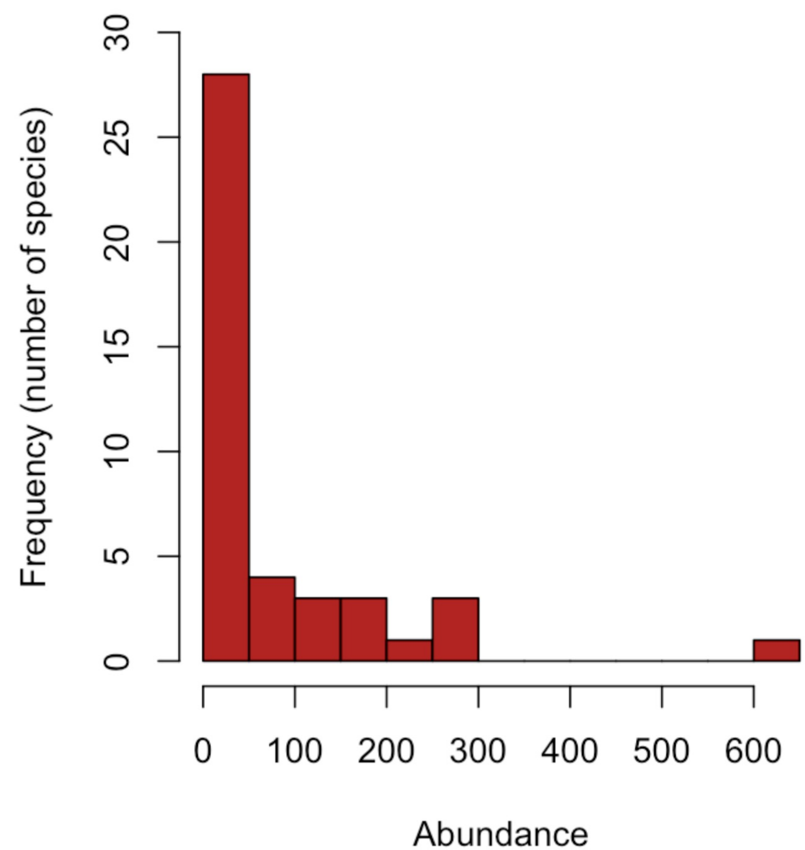
Table 2.2-3 Frequency distribution of bird species abundance at Organ Pipe Cactus National Monument.

Abundance	Frequency (Number of species)
0–50	28
50–100	4
100–150	3
150–200	3
200–250	1
250–300	2
300–350	1
350–400	0
400–450	0
450–500	0
500–550	0
550–600	0
600–650	1
Total	43

Displaying numerical data in the form of frequency distributions – from tabular to graphical form (histograms)

Table 2.2-3 Frequency distribution of bird species abundance at Organ Pipe Cactus National Monument.

Abundance	Frequency (Number of species)
0–50	28
50–100	4
100–150	3
150–200	3
200–250	1
250–300	2
300–350	1
350–400	0
400–450	0
450–500	0
500–550	0
550–600	0
600–650	1
Total	43



The formal definitions of frequency distributions

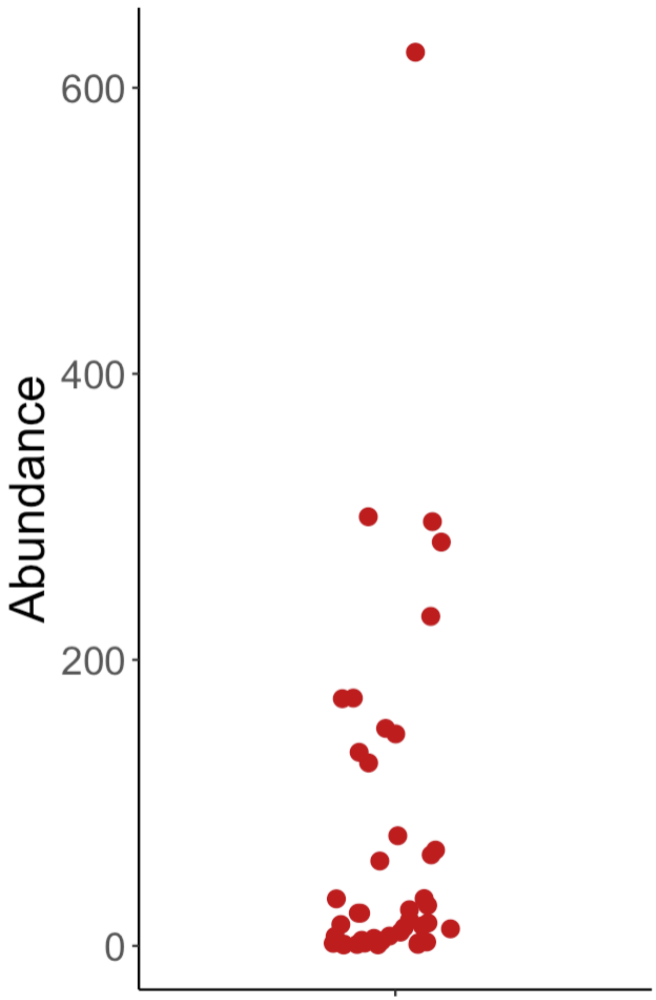
Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval of a quantitative variable (continuous or discrete).

The intervals must be *mutually exclusive* (each observation can only belong to one interval) and *exhaustive* (all observations must be included),

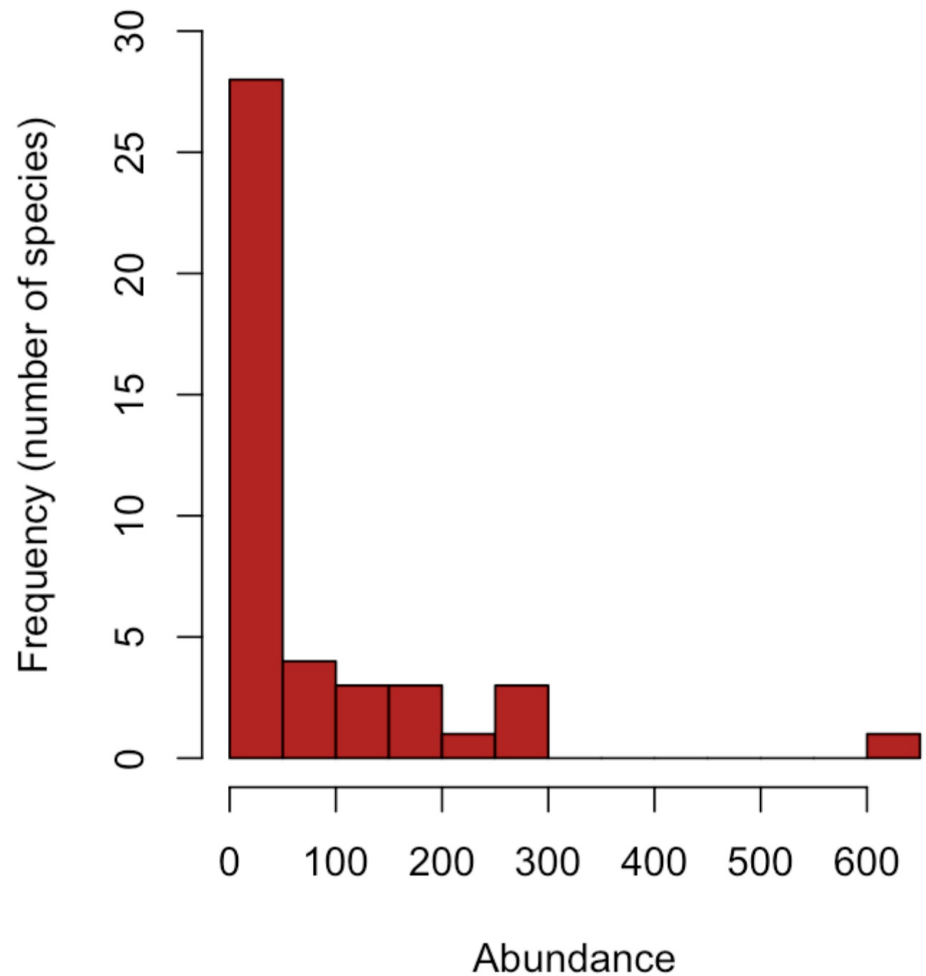
The interval size depends on the data being analyzed and the goals of the analyst.

- Adapted from: http://www.investopedia.com/terms/f/frequencydistribution.asp

Why frequencies and not the raw data?

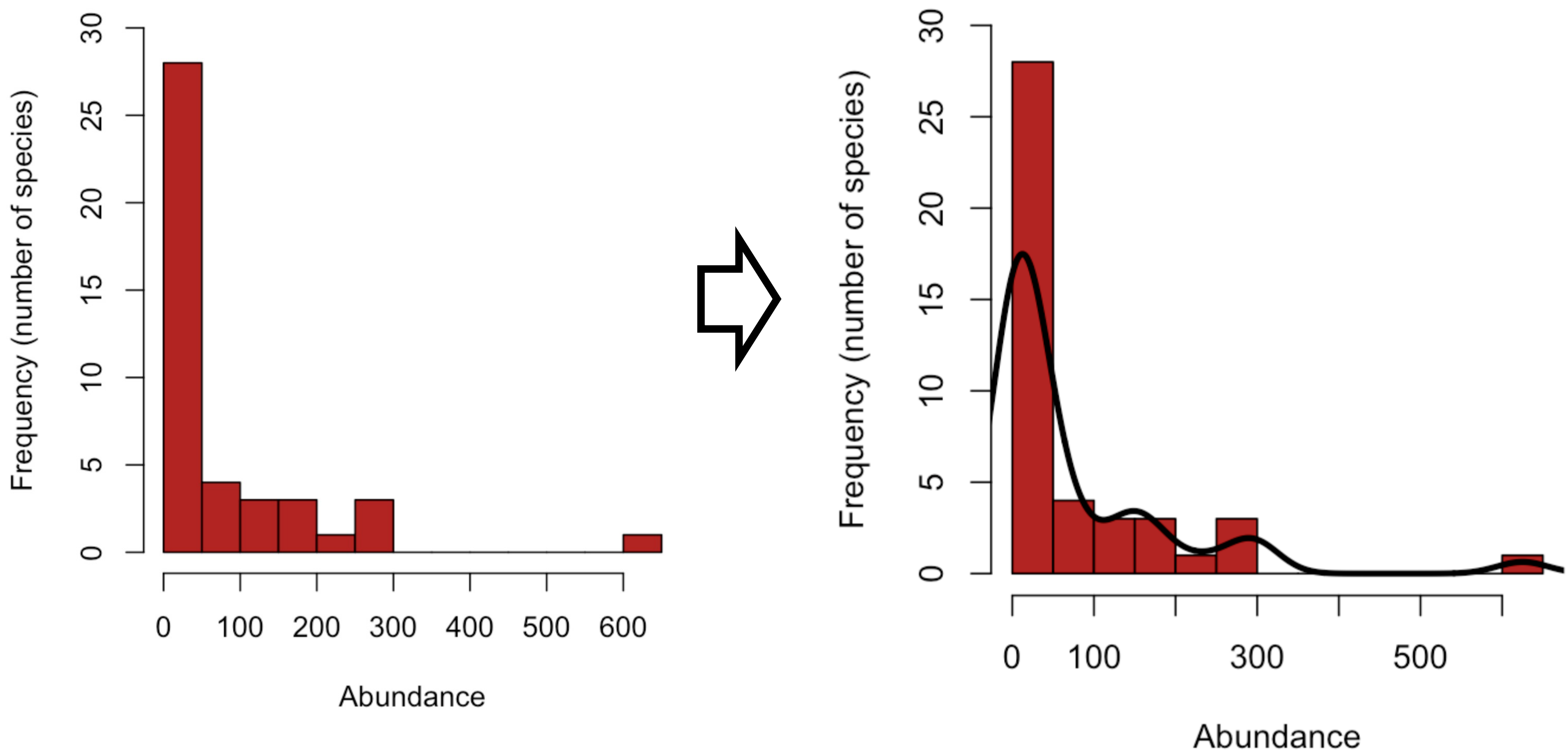


Stripchart

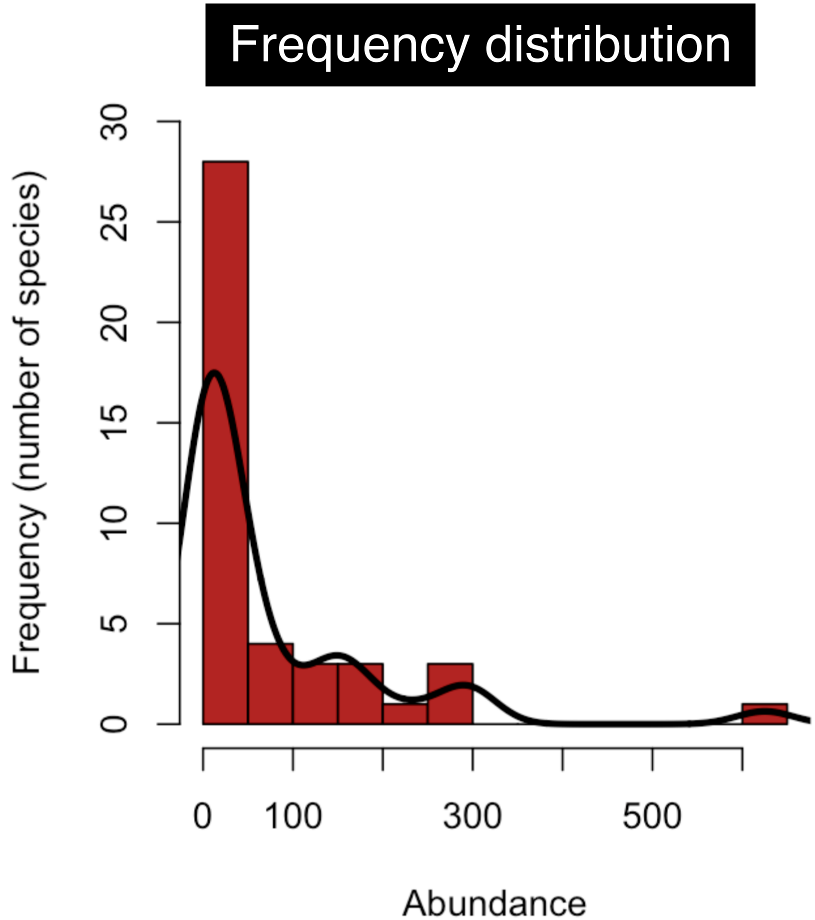


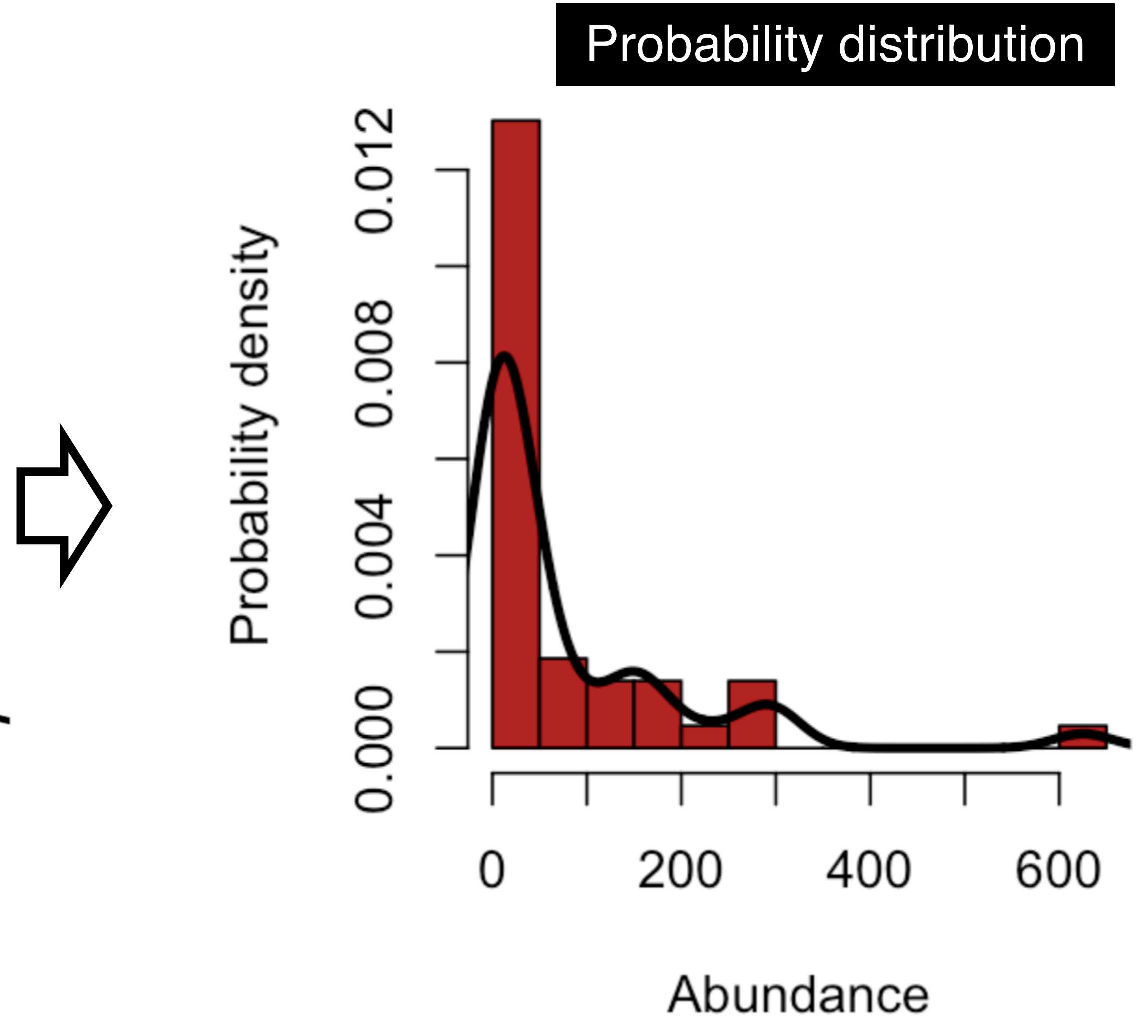
Histogram

Why frequencies and not the raw data?



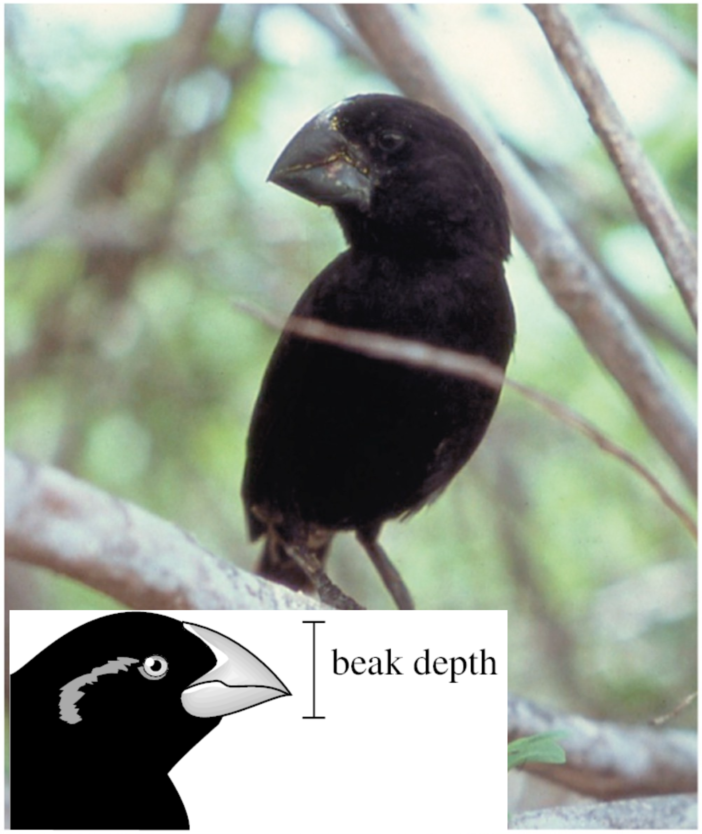
Why frequencies and not the raw data?





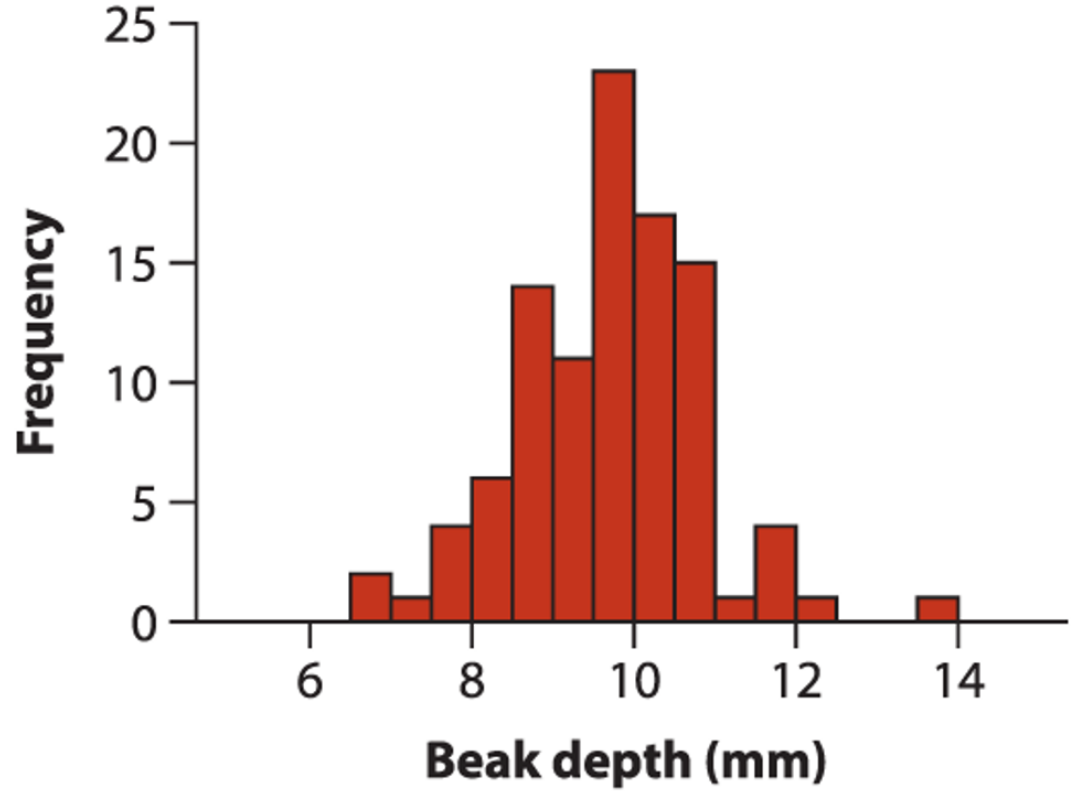
From frequencies to probabilities

Why frequencies and not the raw data?



The large-beaked ground finch on the Galápagos Islands.

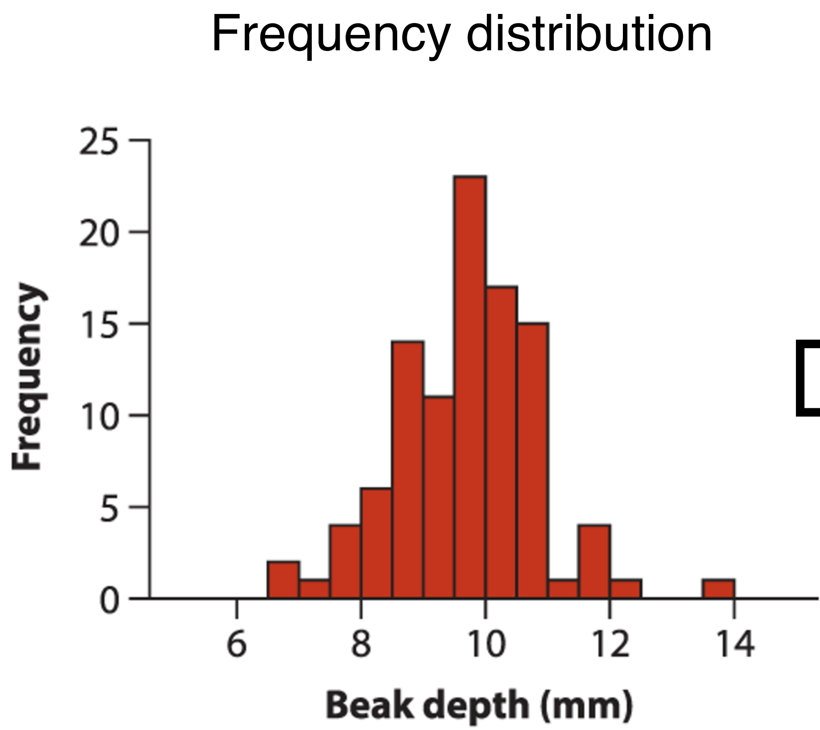
Geospiza magnirostris

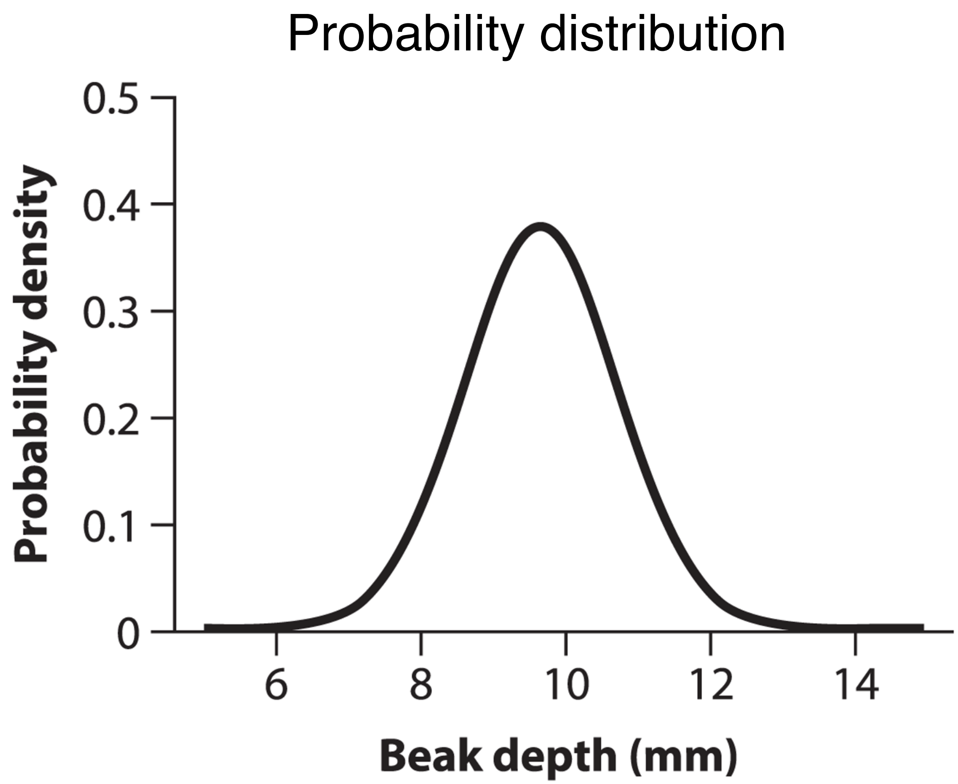


Total of 100 individual birds

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
Dolph Schluter

From frequencies to probabilities





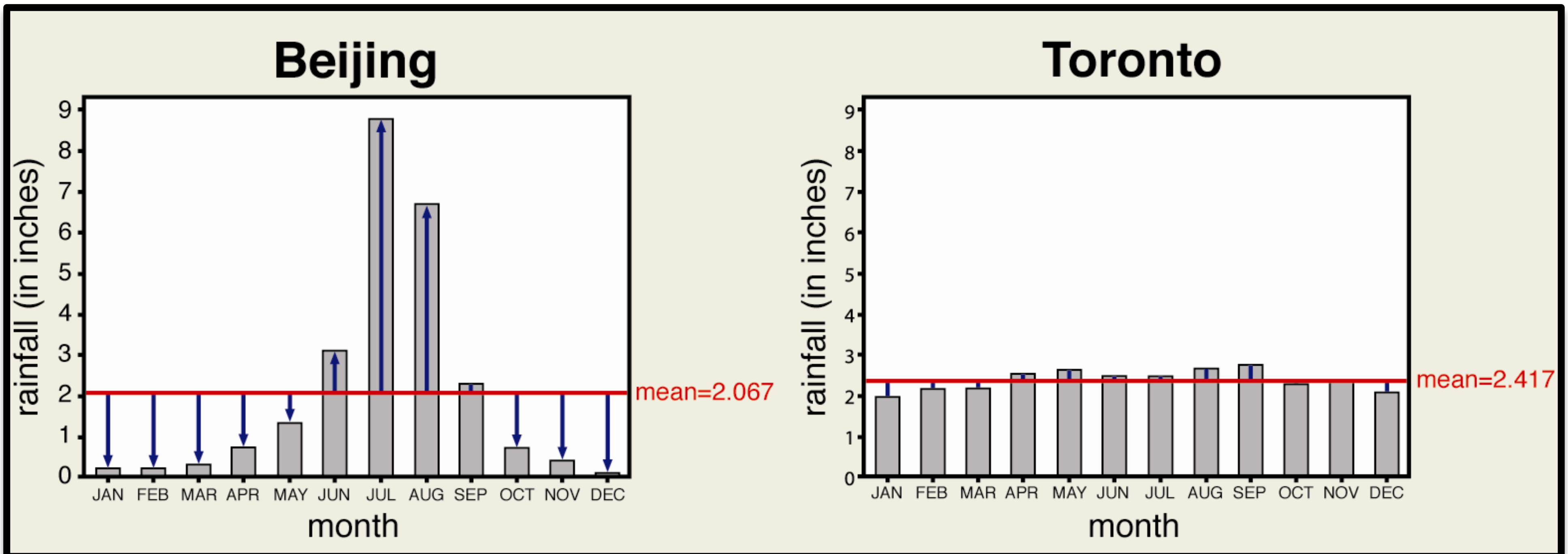
Frequency distributions are important because they describe shapes of numerical variables. Distributional shapes allow to determine proper population probability distributions for inferential statistics (i.e., use samples to estimate population parameters; & convey uncertainty)

Dolph Schluter

Variability in bar graphs (categorical) *versus* histograms (numerical)

Where does rain vary the most?

[the case of categorial variables]

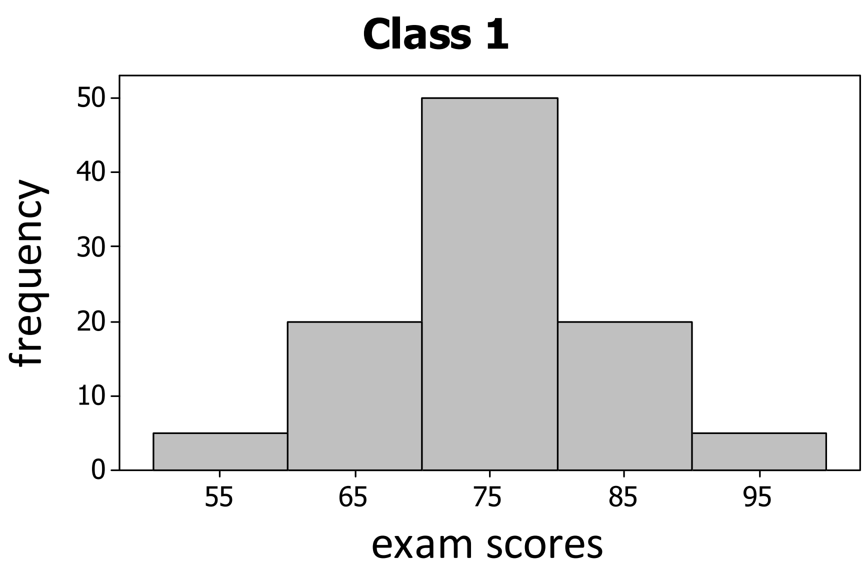


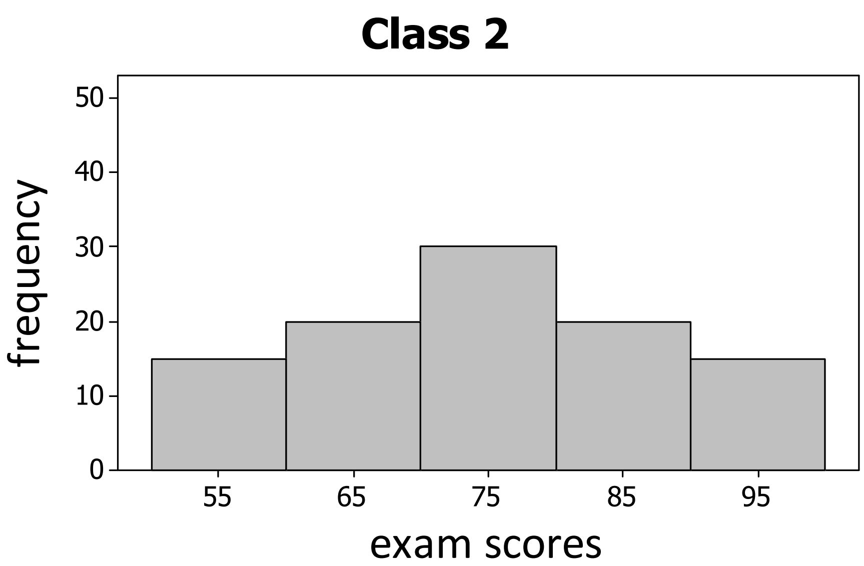
- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

Variability in bar graphs (categorical) versus histograms (numerical)

Which class has the most variation in exam scores?

[the case of continuous variables]

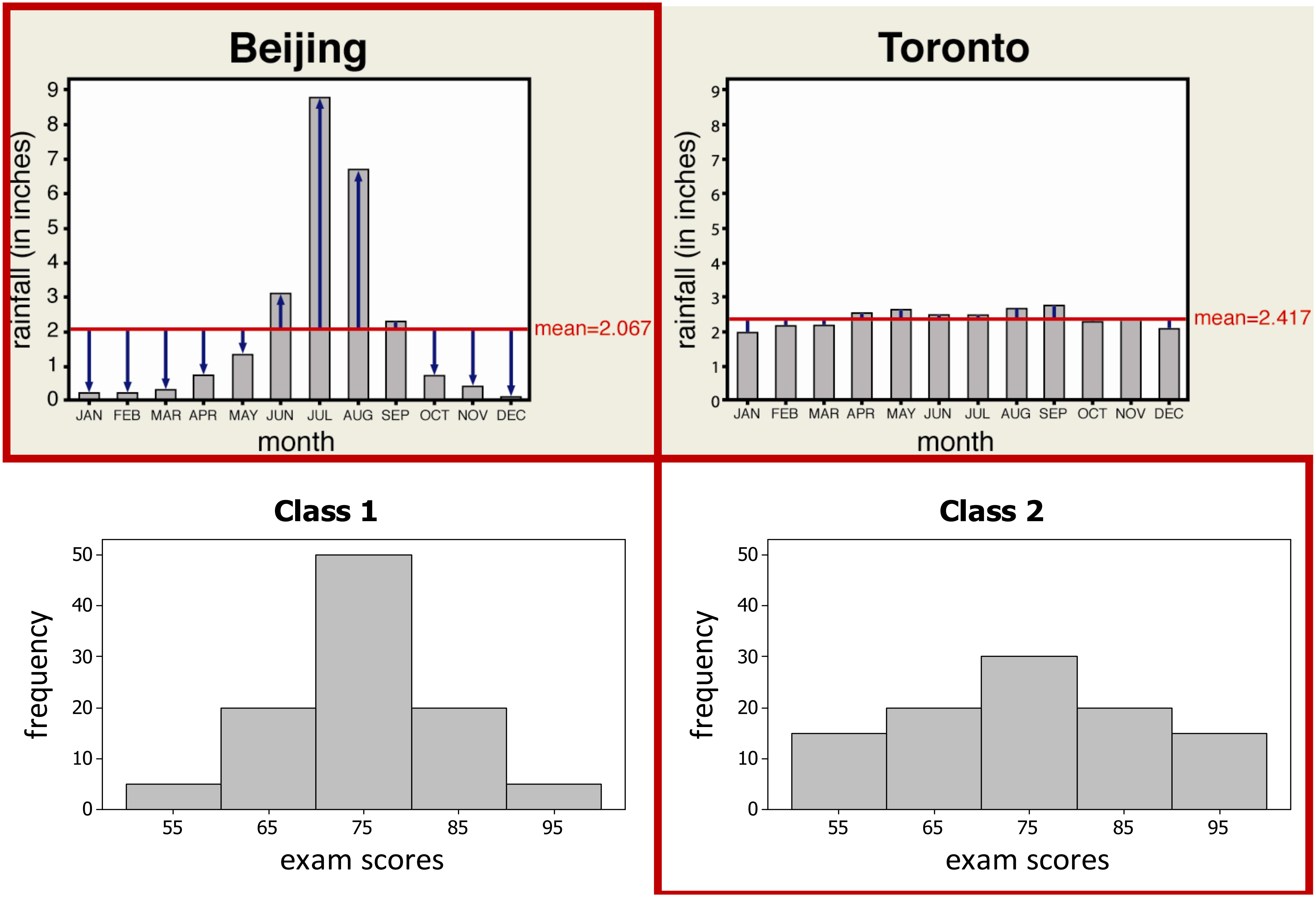




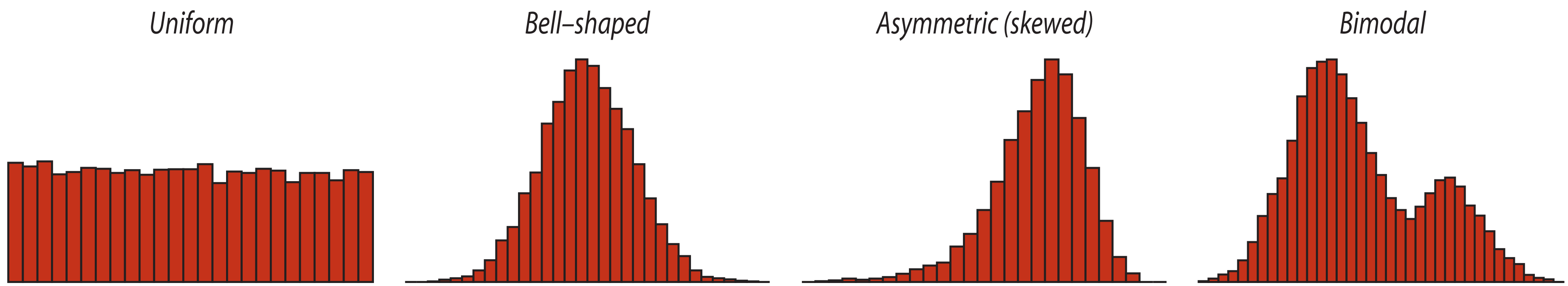
Note: scales (X and Y axis limits) are exactly the same

- Source: Cooper & Shore; Journal of Statistics Education (vol. 18, #2)

Variability in bar graphs (categorical) versus histograms (numerical) – where do data vary the most?



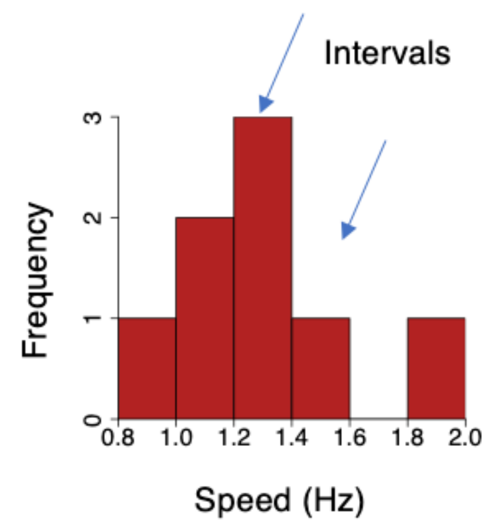
Frequency distributions are important because they describe shapes of numerical variables. Distributional shapes allow to determine proper population probability distributions for inferential statistics



Some possible shapes of frequency distributions.

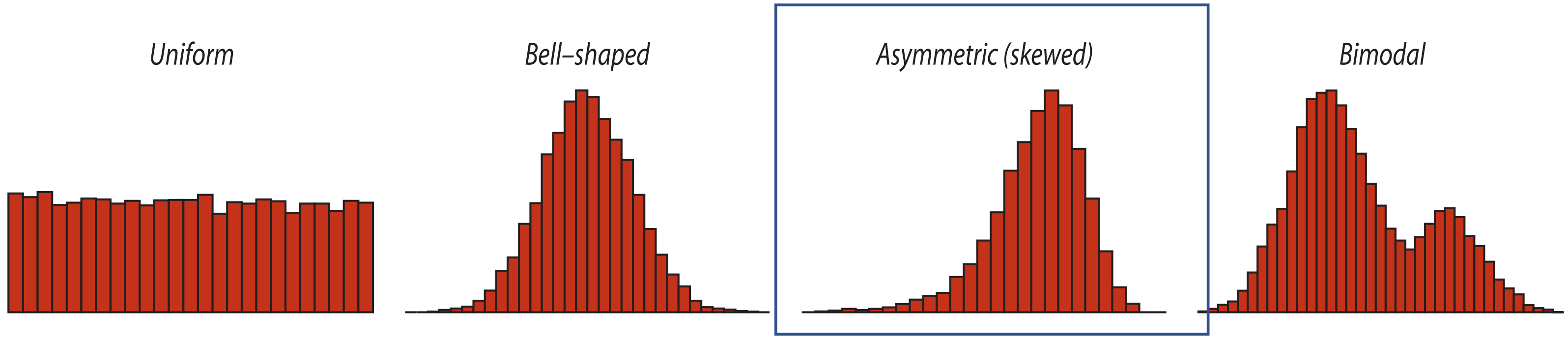
The ***mode*** is the **interval** corresponding to the highest peak in the frequency distribution. A distribution is said bimodal when it has two dominant peaks.

Mode is between 1.2 and 1.4



Skew refers to asymmetry in the shape of a frequency distribution for a numerical variable.

Frequency distributions are important because they describe shapes of numerical variables. Distributional shapes allow to determine proper population probability distributions for inferential statistics

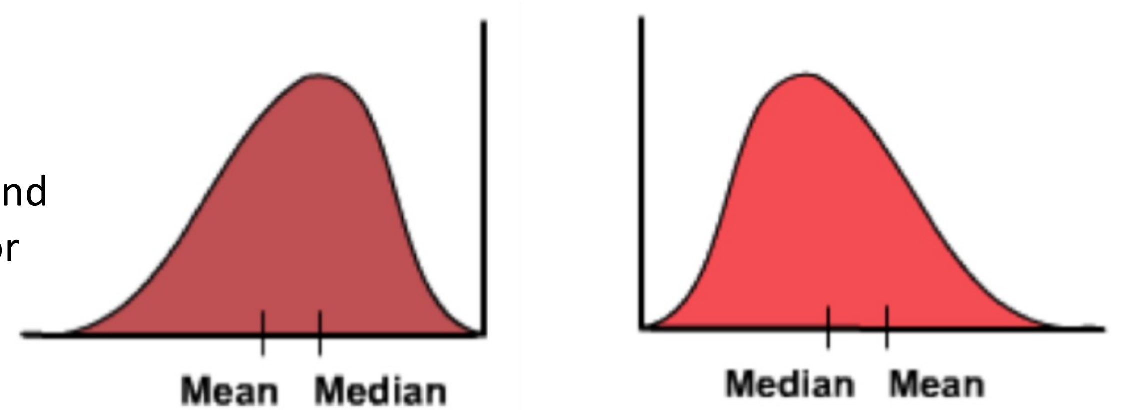


Asymmetric distributions can be either left or positive skewed.

The rule based on the relationship between the mean and median is particularly effective for large datasets (greater than 30 observations).

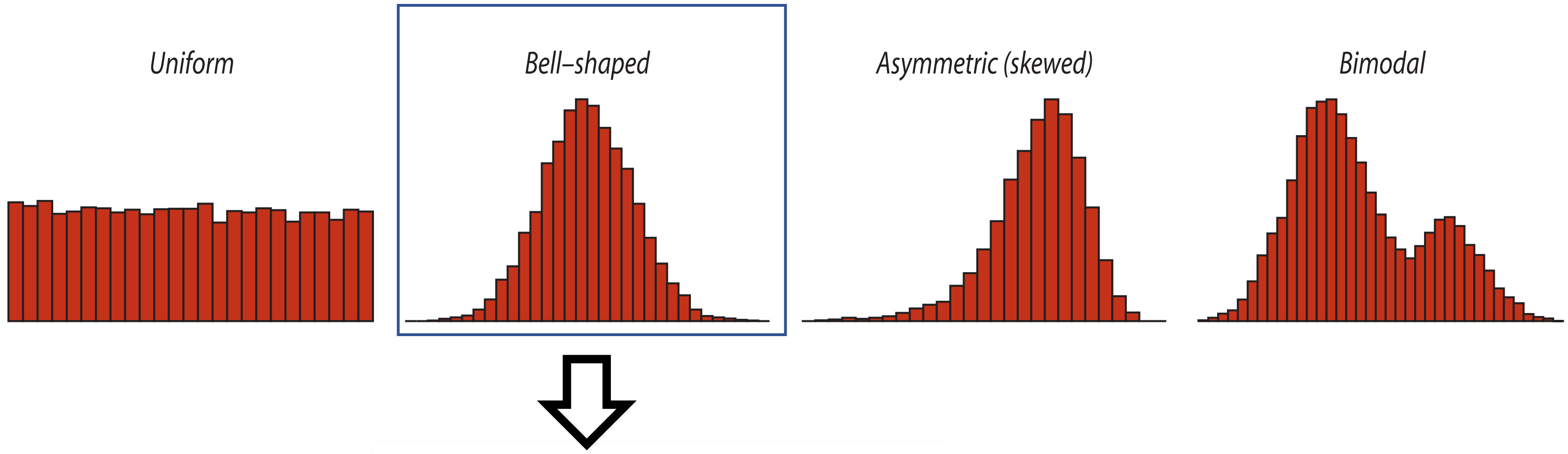
Left (or Negative) skewed

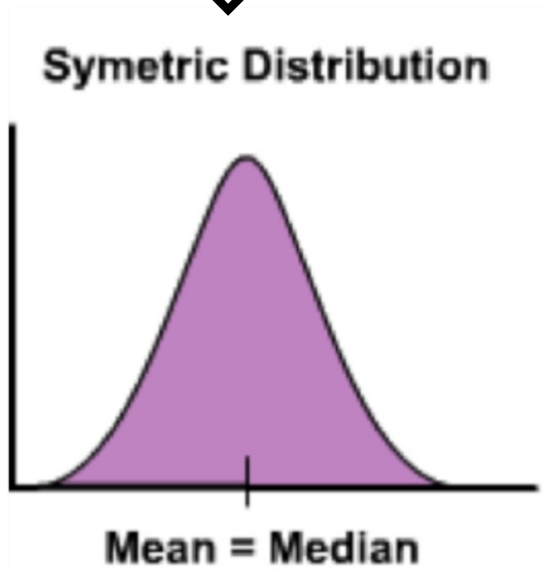
Right (or Positive skewed)



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company
Dolph Schluter

Frequency distributions are important because they describe shapes of numerical variables. Distributional shapes allow to determine proper population probability distributions for inferential statistics



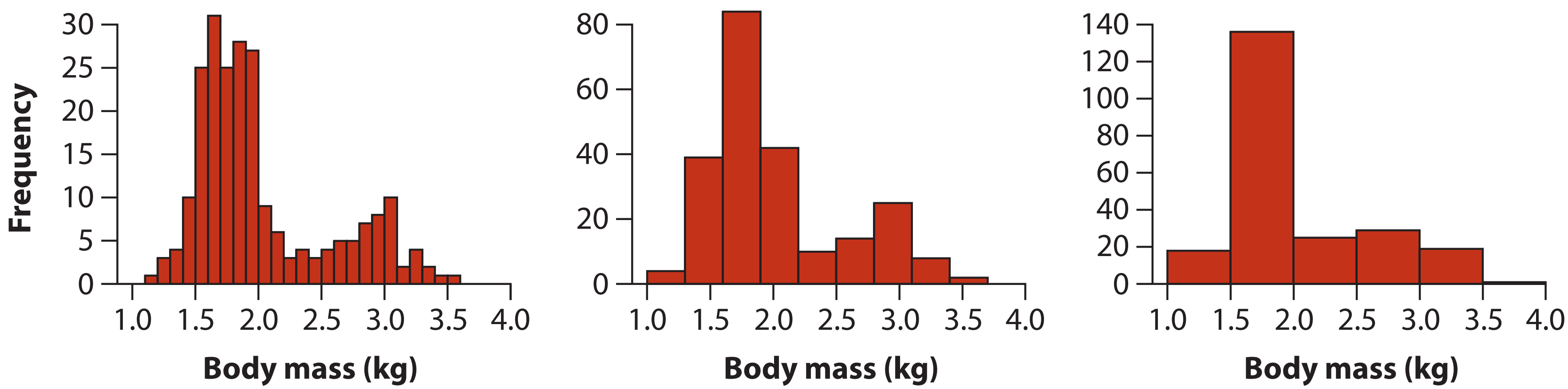


Let’s take a small break – 1 minute

Building a frequency distribution

How many intervals (classes of abundance) should be used?

No strict rules need to be imposed, but rather a number that best show patterns and exceptions in data.



Body mass of 228 female sockeye salmon sampled from Pick Creek in Alaska (Hendry et al. 1999). The same data are shown in each case, but the interval widths are different : 0.1 kg (left), 0.3 kg (middle), and 0.5 kg (right).

Remember that histograms are graphical representations of frequency distributions

Building a frequency distribution – How many intervals?

"Flying" paradise tree snake (*Chrysopelea paradisi*). To better understand how lift is generated, Socha (2002) videotaped glides (from a 10-m tower) of 8 snakes. Rate of side-to-side undulation was measured in hertz (number of cycles per second). The values recorded were:

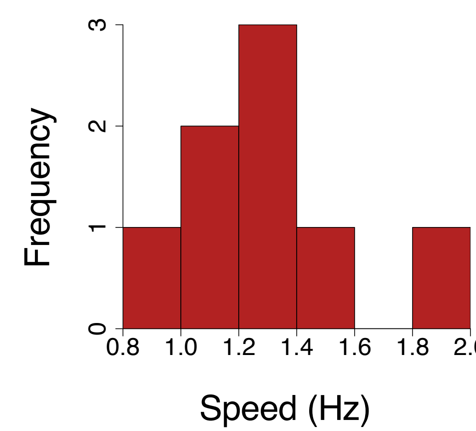
0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

No strict rules should be used, but rather a number that best show patterns and exceptions in data. Rules exist, however, example:

The Sturges' rule: number of intervals = $1 + \ln(n) / \ln(2)$,

For the snake data: $1 + \ln(8) / \ln(2) = 4$ classes.

NOTE: $1 + \ln(n) / \ln(2) = 1 + \log_2(n)$
(as often expressed in some sources).



Building a frequency distribution – The interval size

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, **2.0**

Snake data: 1 + ln(8) / ln(2) = 4 classes

Let's establish the speed intervals (let's say we decide on 4 intervals):

(max(value) - min (value)) / number of classes:

(2.0-0.9) / 4 = 0.275

NOTE: Intervals of frequency distributions are commonly referred to as "classes" as well

Remember

The intervals must be ***mutually exclusive*** (each observation can only belong to one interval) and ***exhaustive*** (all observations must be included), and the interval size depends on the data being analyzed and the goals of the analyst.

- Adapted from: http://www.investopedia.com/terms/f/frequencydistribution.asp

Building intervals

Let's establish the speed intervals: **0.9**, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, **2.0**

(max(value) - min (value)) / number of classes:

(**2.0**-**0.9**) / 4 = <u>0.275</u>

1st class: individuals with speeds between **0.900** and 1.175 (**0.900** + <u>0.275</u>)

Building intervals

Let's establish the speed intervals: **0.9**, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, **2.0**

(max(value) - min (value)) / number of classes:

(2.0-0.9) / 4 = 0.275

1st class: individuals with speeds between 0.900 and 1.175 (0.900 + 0.275)

2nd class: individuals with speeds between 1.175 and 1.450 (1.175 + 0.275)

Building intervals

Let's establish the speed intervals: **0.9**, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, **2.0**

(max(value) - min (value)) / number of classes:

(2.0-0.9) / 4 = <u>0.275</u>

1st class: individuals with speeds between 0.900 and 1.175 (0.900 + <u>0.275</u>)

2nd class: individuals with speeds between 1.175 and 1.450 (1.175 + <u>0.275</u>)

3rd class: individuals with speeds between 1.450 and 1.725 (1.450 + <u>0.275</u>)

Building intervals

Let's establish the speed intervals: **0.9**, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, **2.0**

(max(value) - min (value)) / number of classes:

(2.0-0.9) / 4 = <u>0.275</u>

1st class: individuals with speeds between 0.900 and 1.175 (0.900 + <u>0.275</u>)

2nd class: individuals with speeds between 1.175 and 1.450 (1.175 + <u>0.275</u>)

3rd class: individuals with speeds between 1.450 and 1.725 (1.450 + <u>0.275</u>)

4th class: individuals with speeds between 1.725 and 2.000 (1.725 + <u>0.275</u>)

Counting number of observations (frequencies)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

Let's use: left-closed & right-open [a,b)

Classes	Frequency
0.900 - 1.175	
1.175 - 1.450	
1.450 - 1.725	
1.725 - 2.000	

Intervals are either left-closed & right-open, e.g., 0.900 - 1.175 would contains snakes with rates between 0.9 Hz (included) and 1.175 Hz (not included) = [0.900,1.175).

OR left-open & right-closed, e.g., 0.900 - 1.175 would contains snakes with rates between 0.9 Hz (not included) and 1.175 Hz (included) = (0.900,1.175].

Counting number of observations (frequencies)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-closed & right-open [a,b)

Classes	Frequency
[0.900 - 1.175)	1
[1.175 - 1.450)	
[1.450 - 1.725)	
[1.725 - 2.000)	

Counting number of observations (frequencies)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-closed & right-open [a,b)

Classes	Frequency
[0.900 - 1.175)	1
[1.175 - 1.450)	**5**
[1.450 - 1.725)	
[1.725 - 2.000)	

Counting number of observations (frequencies)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-closed & right-open [a,b)

Classes	Frequency
0.900 - 1.175	1
1.175 - 1.450	5
1.450 - 1.725	1
1.725 - 2.000	

Counting number of observations (frequencies)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0 ?

left-closed & right-open [a,b)

Classes	Frequency
[0.900 - 1.175)	1
[1.175 - 1.450)	5
[1.450 - 1.725)	1
[1.725 - 2.000)	**???**

Counting number of observations (frequencies)

? 0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

Let’s try left-open & right-closed (a,b]

Classes	Frequency
(0.900 - 1.175]	???
(1.175 - 1.450]	
(1.450 - 1.725]	
(1.725 - 2.000]	

FAILED

Counting number of observations (frequencies)

Let's try a different number of classes (5) and interval size (0.275)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-closed & right-open [a,b)

Classes	Frequency
[0.900 - 1.175)	1
[1.175 - 1.450)	5
[1.450 - 1.725)	1
[1.725 - 2.000)	0
[2.000 - 2.275)	1

left-open & right-closed (a,b]

Classes	Frequency
(0.625 - 0.900]	1
(0.900 - 1.175]	0
(1.175 - 1.450]	5
(1.450 - 1.725]	1
(1.725 - 2.000]	1

It works, but the classes may not print well. They have too many decimals. We can change the number of classes to try to fix this issue (let's try 7 classes next).

Counting number of observations (frequencies)

Let's try a different number of classes (7) and interval size (0.2)

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

Let's use: left-closed & right-open [a,b)

Classes	Frequency
[0.8 - 1.0)	1
[1.0 - 1.2)	0
[1.2 - 1.4)	3
[1.4 - 1.6)	2
[1.6 - 1.8)	1
[1.8 - 2.0)	0
[2.0 - 2.2)	1

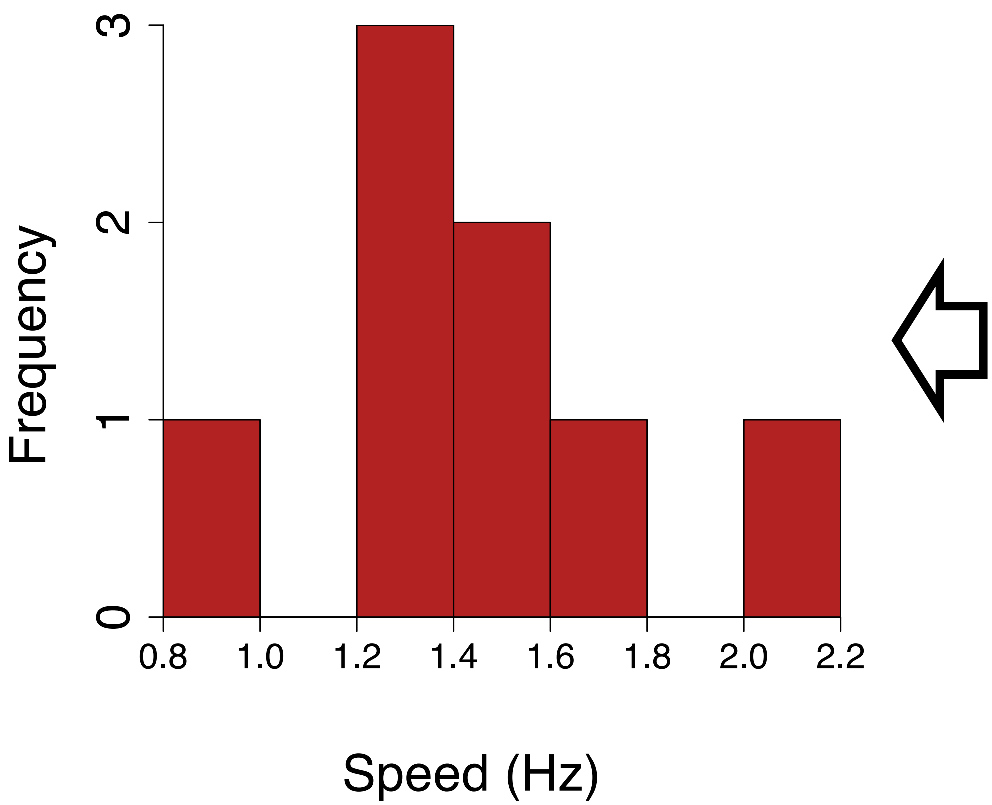
Total = 8

Note: some software may include 2.0 in this interval even though is opened. This may happen when the last values in the data fall here. (R does that)

✓

From frequency distribution tables to histograms

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0



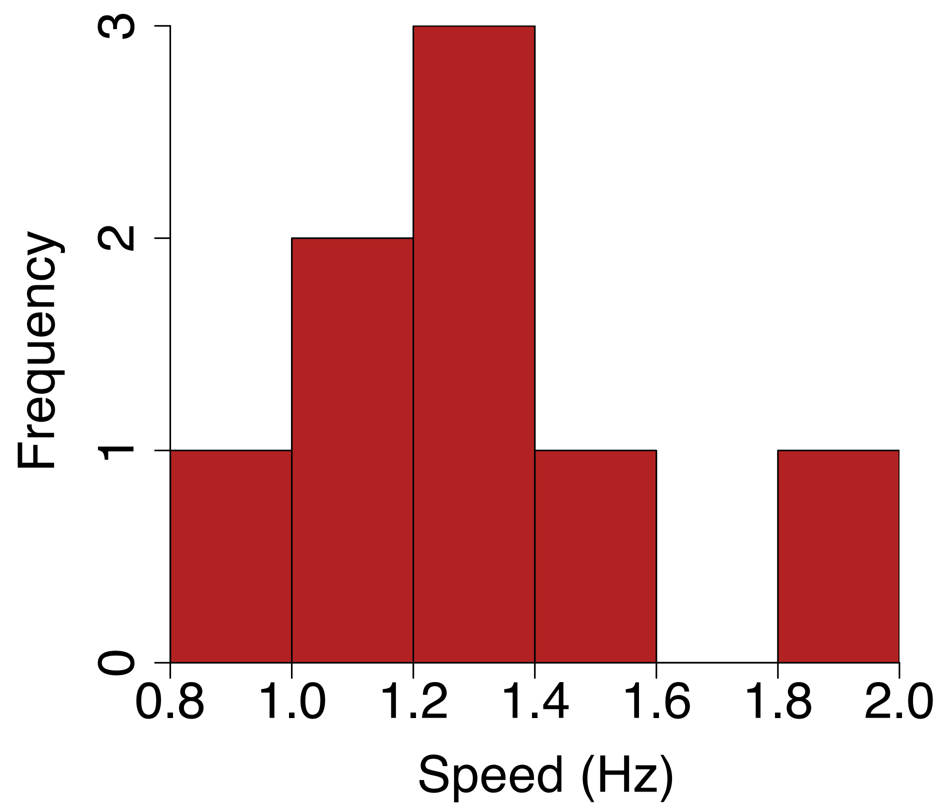
left-closed & right-open [a,b)	
Classes	Frequency
[0.8 - 1.0)	1
[1.0 - 1.2)	0
[1.2 - 1.4)	3
[1.4 - 1.6)	2
[1.6 - 1.8)	1
[1.8 - 2.0)	0
[2.0 - 2.2)	1

From frequency distribution tables to histograms

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-open & right-closed (a,b]

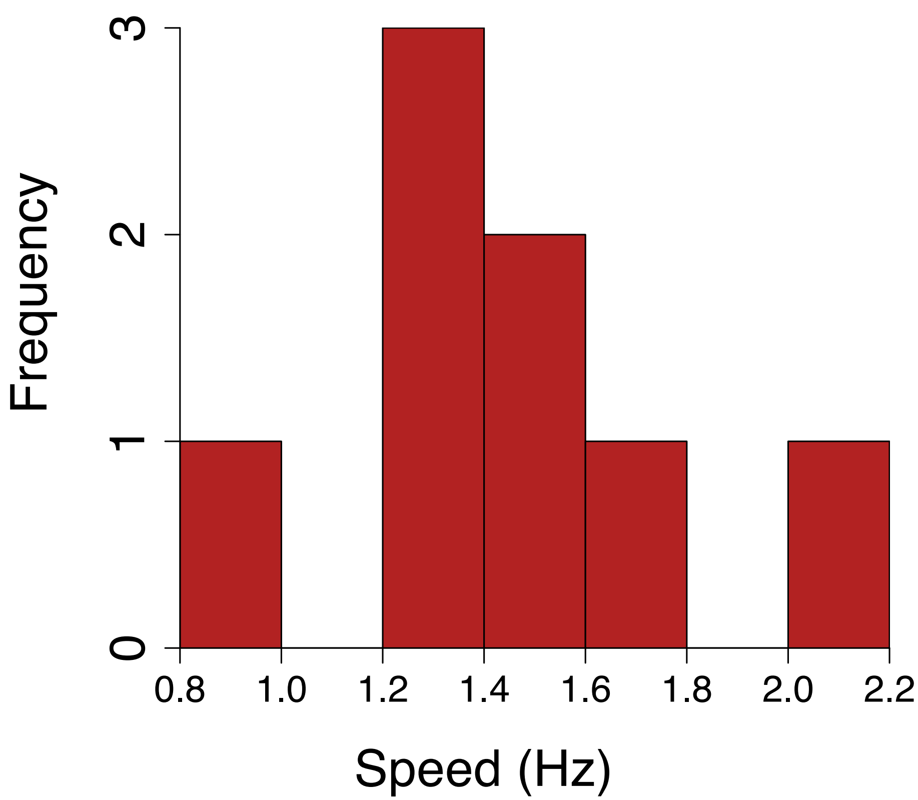
Classes	Frequency
(0.80 - 1.00]	1
(1.00 - 1.20]	2
(1.20 - 1.40]	3
(1.40 - 1.60]	1
(1.60 - 1.80]	0
(1.80 - 2.00]	1



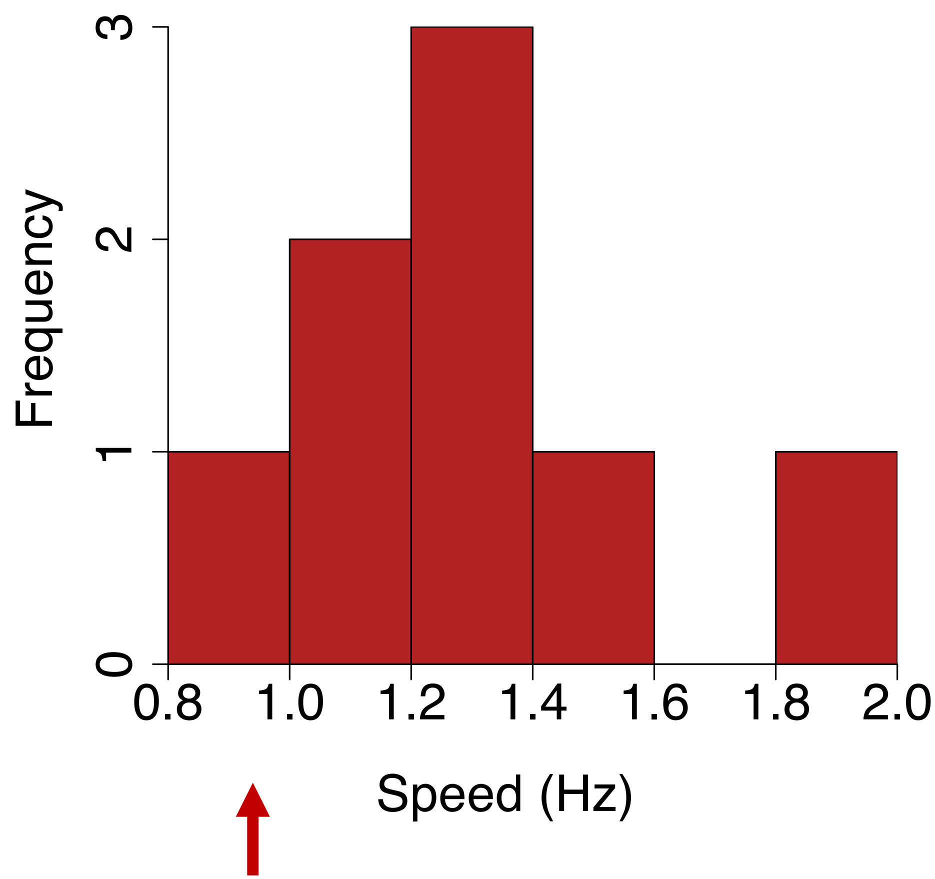
From frequency distribution tables to histograms

0.9, 1.2, 1.2, 1.3, 1.4, 1.4, 1.6, 2.0

left-closed & right-open

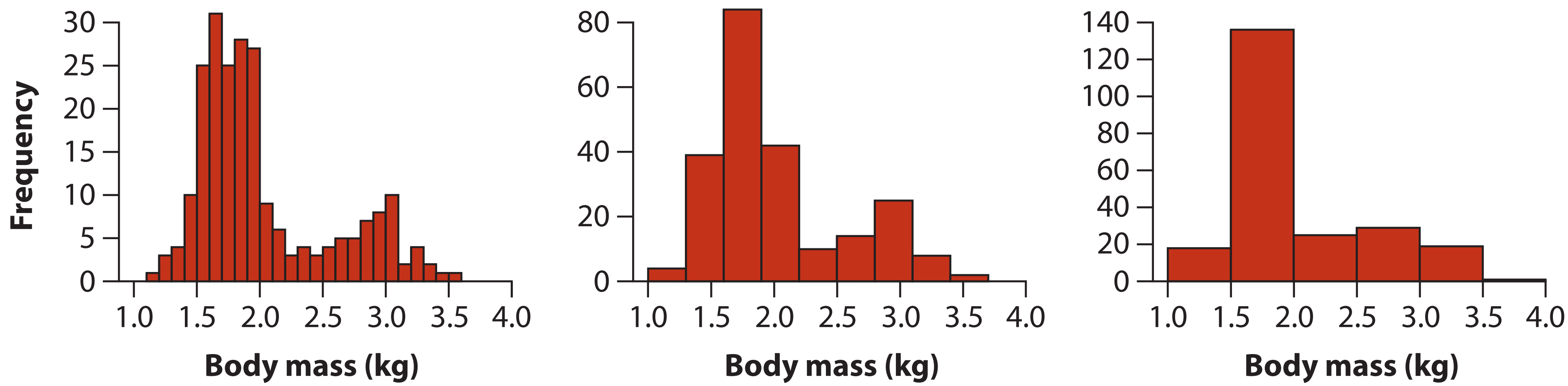


left-open & right-closed



Perhaps choose this one. Less intervals and only one empty interval.

Again, number of classes and interval size depend on the goal of the analyst & often based on trying different options



Body mass of 228 female sockeye salmon sampled from Pick Creek in Alaska (Hendry et al. 1999). The same data are shown in each case, but the interval widths are different : 0.1 kg (left), 0.3 kg (middle), and 0.5 kg (right).

Next lecture: describing data

Samples and populations are often made of lots of individual (observational) units and their associated information (observations, variables).

We need to be able to describe samples by summary statistics (mean, median, variance, etc) so that these summaries can serve as an estimate of the same summaries for their statistical populations.