INTRODUCTION TO DATA ANALYSIS

SUMMARY STATISTICS

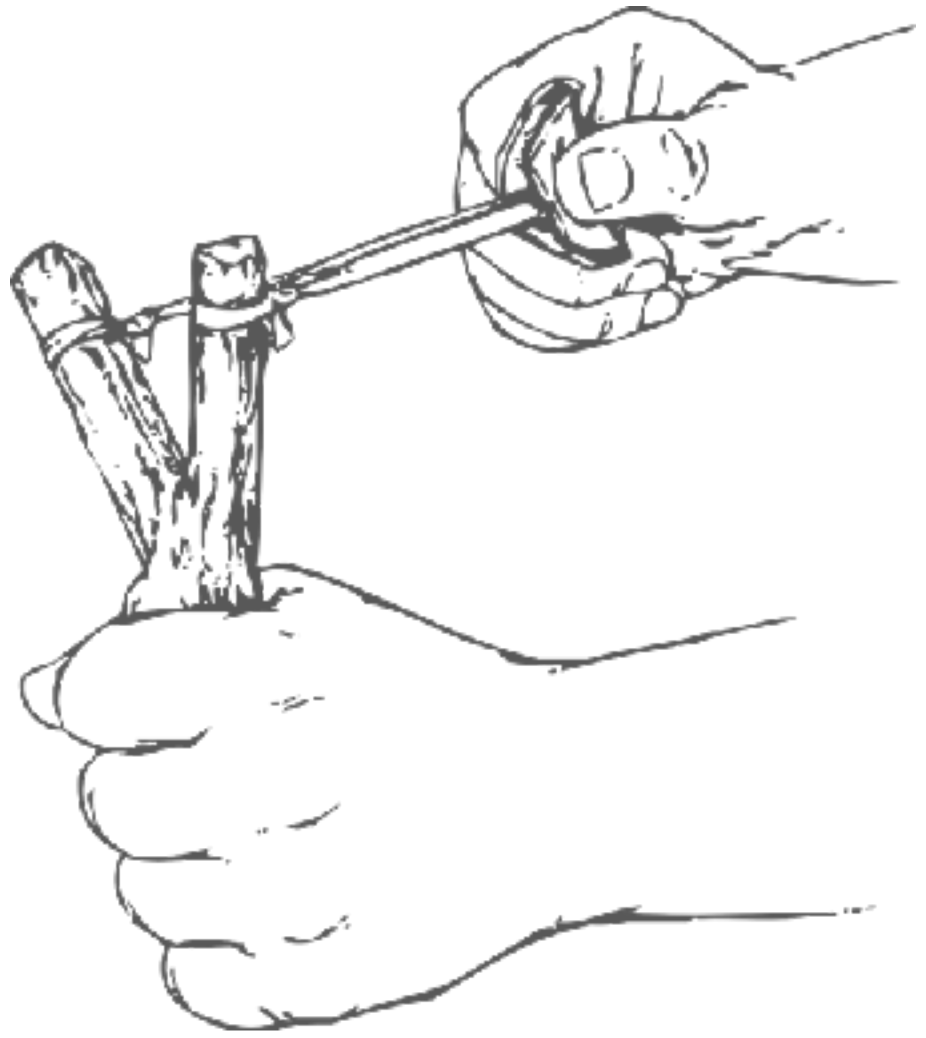
FINAL EXAM

- Friday February 7 2020 ::: 4-8pm
- 66/E33 & 66/E34
- no class at noon on that day

HOW (NOT) TO PERFORM OPTIMALLY IN THIS COURSE

- use the script, not the slides
- individual practice at home essential

LEARNING GOALS



- understand what a "summary statistic" is
- understand and be able to compute the following:
  - counts and frequencies for categorical data
  - measures of central tendency: mean, mode & median
  - measures of dispersion: variance, standard deviation & quantiles
  - bootstrapped confidence intervals for an estimate
  - co-variance & correlation

SUMMARY STATISTICS

- usually: what we analyze ≠ what we actually measured
- data observations are always already interpreted abstractions over a much richer reality
  - e.g., we record whether a coin landed heads or tails, not *where* it landed
- summary statistic: a single number that represent one aspect of the data
  - useful for communication about / understanding of the data at hand
  - e.g., counting observations of a particular type / calculating the mean of some numeric observations

SUMMARY STATISTICS

- usually: what we analyze ≠ what we actually measured
- data observations are always already interpreted abstractions over a much richer reality
  - e.g., we record whether a coin landed heads or tails, not *where* it landed
- summary statistic: a single number that represent one aspect of the data
  - useful for communication about / understanding of the data at hand
  - e.g., counting observations of a particular type / calculating the mean of some numeric observations

BIO-LOGIC JAZZ-METAL

- 102 participants from this course [THANKS FOR DOING THIS!]
- everybody got three 2-alternative forced-choice questions (in random order):

  "If you have to choose between the following two options, which one do you prefer?"

  1. Biology vs Logic
  2. Jazz vs Metal
  3. Mountains vs Beach

- no sane person would defend serious scientific hypotheses about this study, **but** the lecturer conjectures irresponsibly that a certain musical taste may be correlated with a particular preference for academic subjects

INSPECTING THE DATA

```
head(data_BLJM_processed)
```

```
## # A tibble: 6 x 3
##   submission_id condition response
##           <dbl> <chr>     <chr>
## 1           379 BM        Beach
## 2           379 LB        Logic
## 3           379 JM        Metal
## 4           378 JM        Metal
## 5           378 LB        Logic
## 6           378 BM        Beach
```

participant with ID 379 prefers:

- beaches over mountains
- logic over biology
- metal over jazz

COUNTING OBSERVATIONS

- functions `n`, `count`, and `tally` from `dplyr` package
  - caveats:
    - different versions of `dplyr` package implement `count` differently
    - several packages define a `count` function; use `dplyr::count` explicitly to be sure
- functions `table` and `prop.table` from base R

COUNTING OBSERVATIONS

- `n` works only in `mutate` and `summarize`
- `n` essentially counts rows (useful after grouping!)

```
data_BLJM_processed %>%
  group_by(condition) %>%
  summarise(nr_observation_per_condition = n()) %>%
  ungroup()
```

```
## # A tibble: 3 x 2
##   condition nr_observation_per_condition
##   <chr>                            <int>
## 1 BM                                 102
## 2 JM                                 102
## 3 LB                                 102
```

COUNTING OBSERVATIONS

- `count` and `tally` are wrappers around `n`
  - `count` implicitly groups/ungroups
  - `tally` does not tinker with existing grouping

```r
data_BLJM_processed %>%
  group_by(condition, response) %>%
  summarise(n = n())
```

```
## # A tibble: 6 x 3
## # Groups:   condition [3]
##   condition response      n
##   <chr>     <chr>     <int>
## 1 BM        Beach        44
## 2 BM        Mountains    58
## 3 JM        Jazz         64
## 4 JM        Metal        38
## 5 LB        Biology      58
## 6 LB        Logic        44
```

```r
data_BLJM_processed %>%
  # function`count` is masked by another package, must call explicitly
  dplyr::count(condition, response)
```

```
## # A tibble: 6 x 3
##   condition response      n
##   <chr>     <chr>     <int>
## 1 BM        Beach        44
## 2 BM        Mountains    58
## 3 JM        Jazz         64
## 4 JM        Metal        38
## 5 LB        Biology      58
## 6 LB        Logic        44
```

COUNTS OF CHOICE PAIRS

```
BLJM_associated_counts <- data_BLJM_processed %>%
  select(submission_id, condition, response) %>%
  pivot_wider(names_from = condition, values_from = response) %>%
  # drop the Beach-vs-Mountain condition
  select(-BM) %>%
  dplyr::count(JM,LB)
BLJM_associated_counts
```

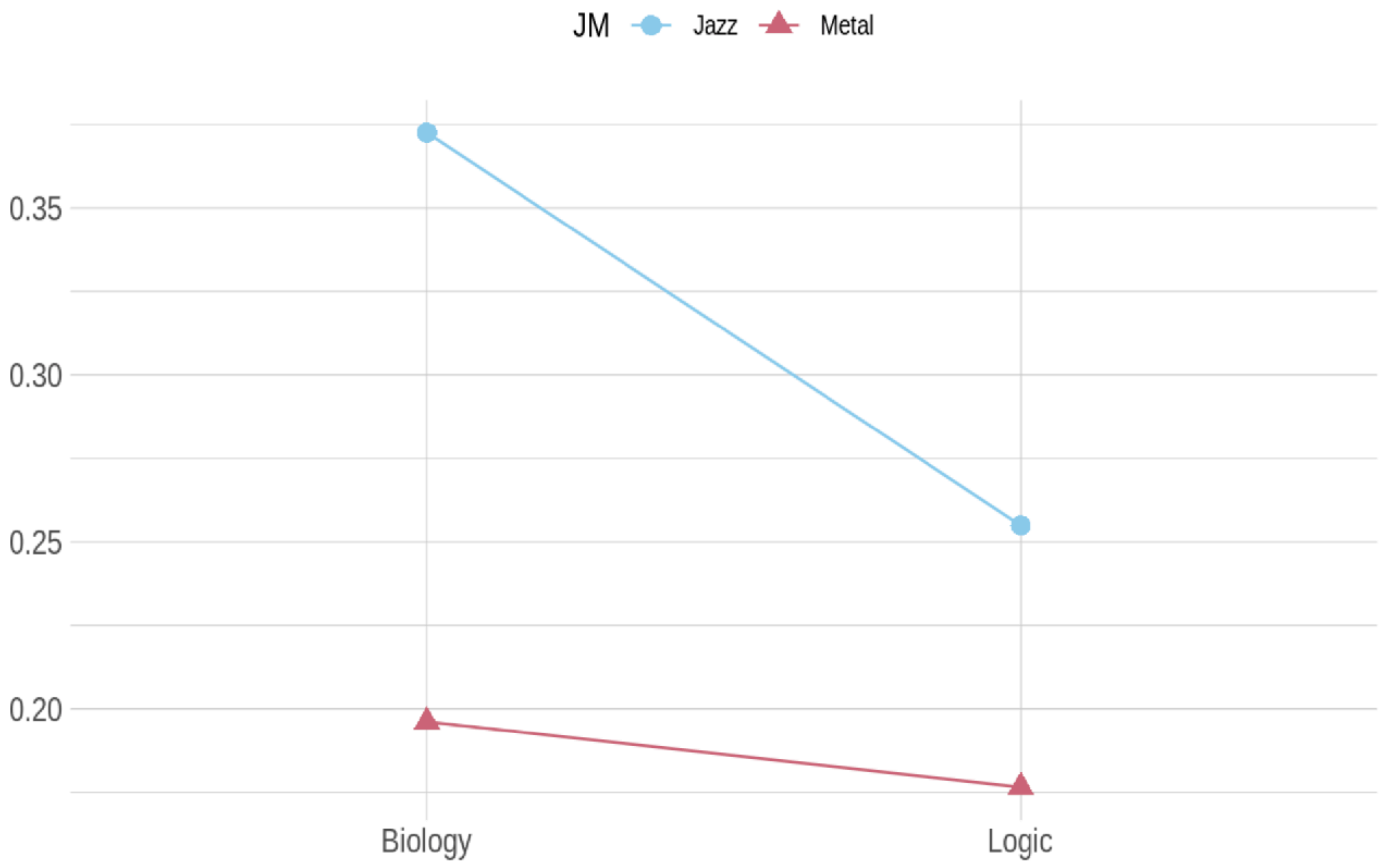
```
## # A tibble: 4 x 3
##   JM    LB          n
##   <chr> <chr>   <int>
## 1 Jazz  Biology    38
## 2 Jazz  Logic      26
## 3 Metal Biology    20
## 4 Metal Logic      18
```

PROPORTIONS OF CHOICE PAIRS

```
## # A tibble: 4 x 3
##   JM    LB          n
##   <chr> <chr>   <int>
## 1 Jazz  Biology    38
## 2 Jazz  Logic      26
## 3 Metal Biology    20
## 4 Metal Logic      18
```

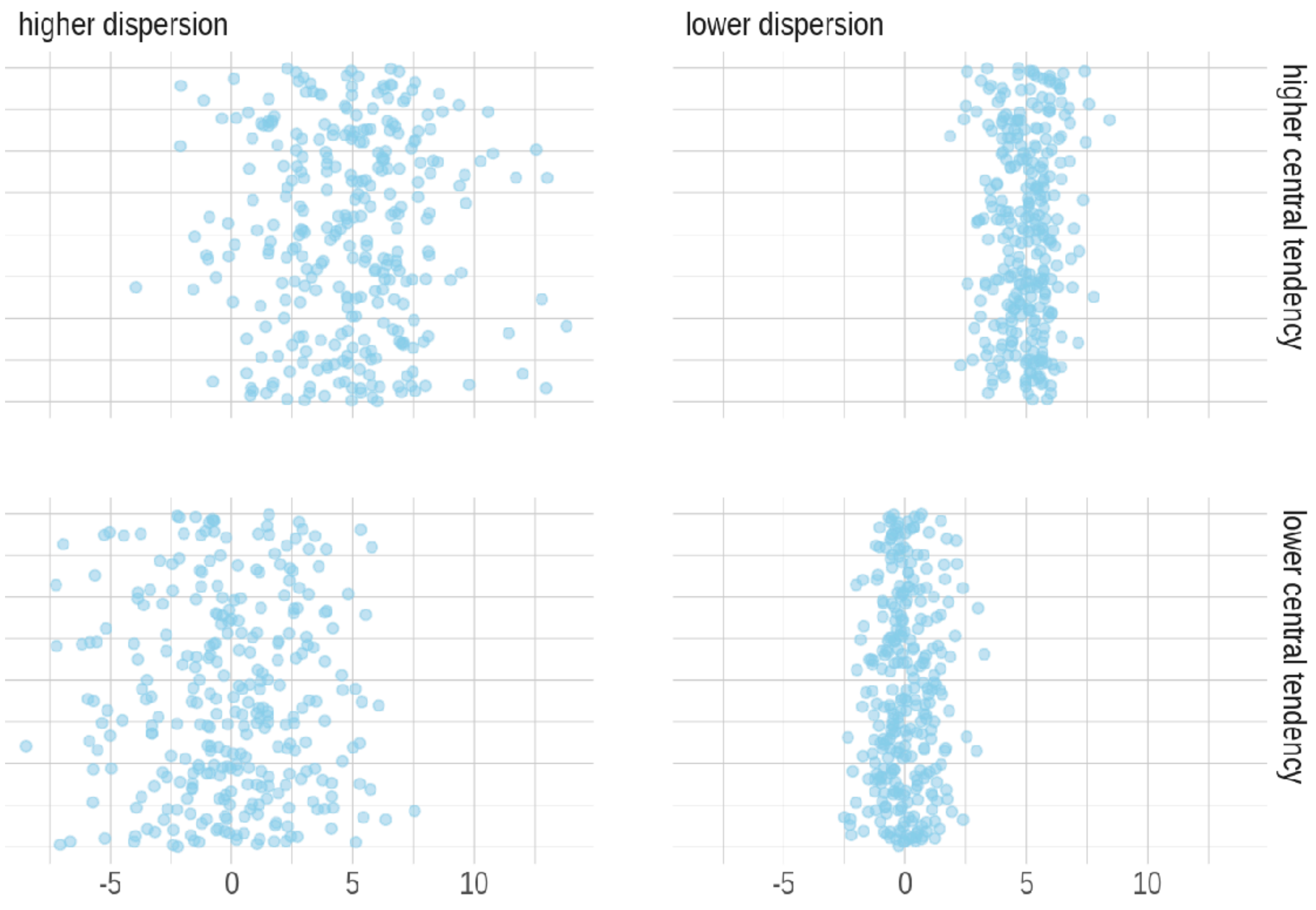
```
BLJM_associated_counts %>%
  # look at relative frequency, not total counts
  mutate(n = n / sum(n)) %>%
  pivot_wider(names_from = LB, values_from = n)
```

```
## # A tibble: 2 x 3
##   JM    Biology Logic
##   <chr>   <dbl> <dbl>
## 1 Jazz    0.373 0.255
## 2 Metal   0.196 0.176
```

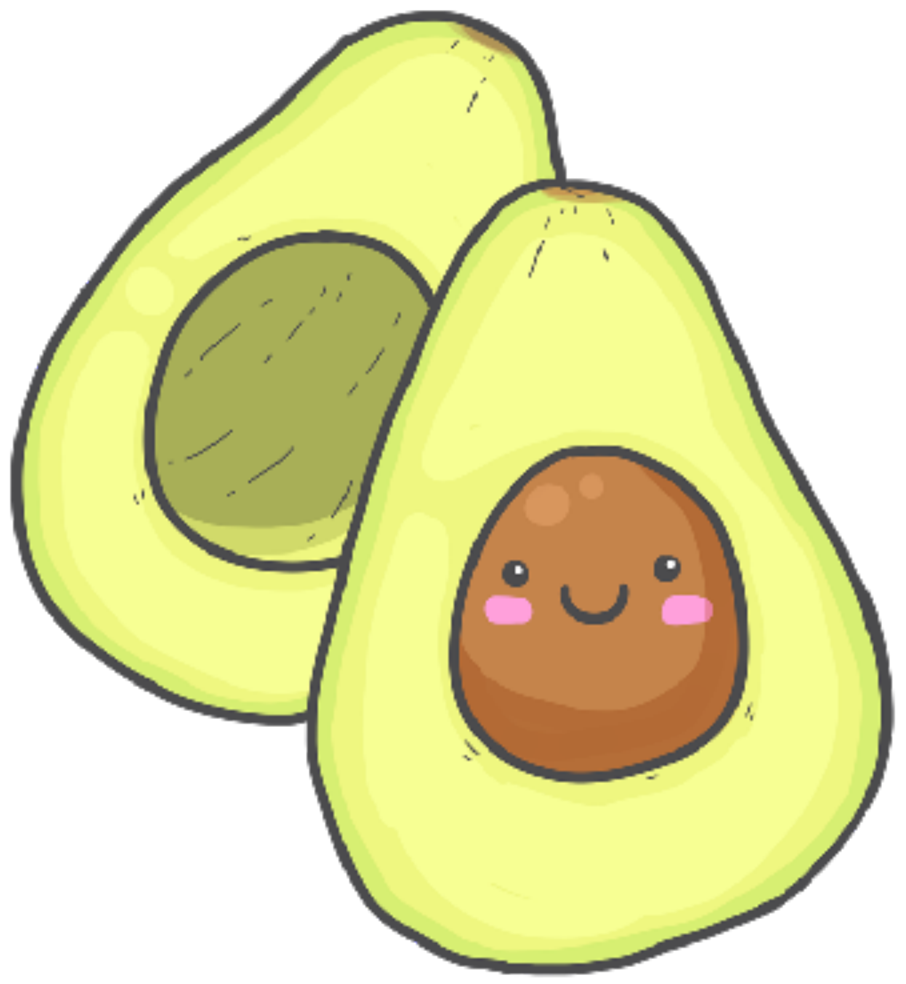

JM
Jazz
Metal
0.35
0.30
0.25
0.20
Biology
Logic

MEASURES OF CENTRAL TENDENCY & DISPERSION

- central tendency: where is "the center" of the data observations
- dispersion: how far are values distributed around "the center"



AVOCADO DATA



- data released by Hass Avocado Board (plucked from kaggle)

```
## # A tibble: 18,249 x 7
##    Date       average_price total_volume_sold small  medium large type
##    <date>             <dbl>             <dbl> <dbl>   <dbl> <dbl> <chr>
##  1 2015-12-27          1.33            64237. 1037.  54455.  48.2 conventional
##  2 2015-12-20          1.35            54877.  674.  44639.  58.3 conventional
##  3 2015-12-13          0.93           118220.  795. 109150. 130.  conventional
##  4 2015-12-06          1.08            78992. 1132   71976.  72.6 conventional
##  5 2015-11-29          1.28            51040.  941.  43838.  75.8 conventional
##  6 2015-11-22          1.26            55980. 1184.  48068.  43.6 conventional
##  7 2015-11-15          0.99            83454. 1369.  73673.  93.3 conventional
##  8 2015-11-08          0.98           109428.  704. 101815.  80   conventional
##  9 2015-11-01          1.02            99811. 1022.  87316.  85.3 conventional
## 10 2015-10-25          1.07            74339.  842.  64757. 113   conventional
## # … with 18,239 more rows
```

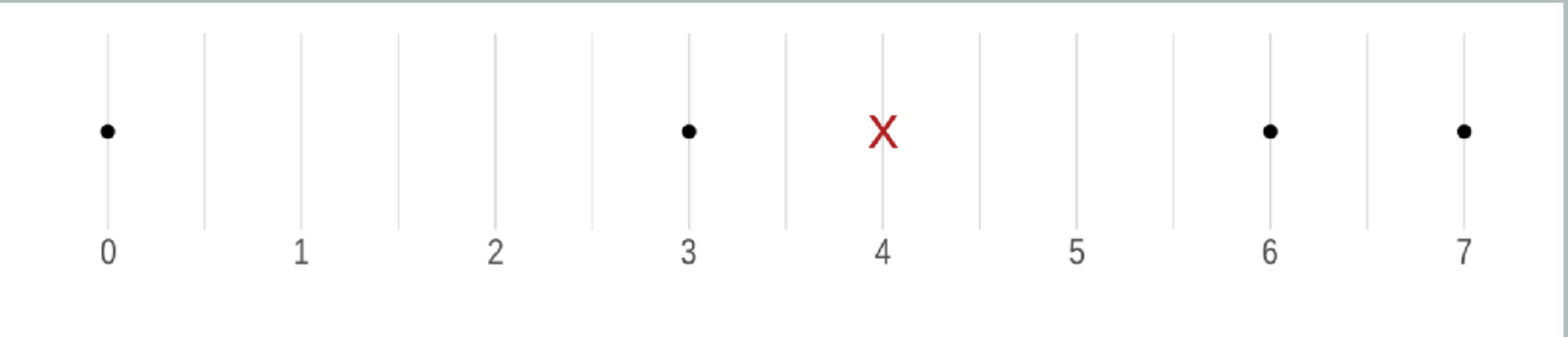
MEAN

If $\vec{x} = \langle x_1, \ldots, x_n \rangle$ is a vector of n observations with $x_i \in \mathbb{R}$ for all $1 \leq i \leq n$, the (arithmetic) **mean** of x, written $\mu_{\vec{x}}$, is defined as

$$\mu_{\vec{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i \, .$$

MEAN :: EXAMPLE

Example. The mean of the vector $\vec{x} = \langle 0, 3, 6, 7 \rangle$ is $\mu_{\vec{x}} = \frac{0+3+6+7}{4} = \frac{16}{4} = 4$. The black dots in the graph below show the data observations and the red cross indicates the mean. Notice that the mean is cleary *not* the mid-point between the maximum and the minimum (which here would be 3.5).



CALCULATING THE MEAN IN R

```
avocado_data %>%
  group_by(type) %>%
  summarise(
    mean_price = mean(average_price)
  )
```

```
## # A tibble: 2 x 2
##   type          mean_price
##   <chr>              <dbl>
## 1 conventional        1.16
## 2 organic             1.65
```

EXCURSION :: MEAN AS EXPECTED VALUE

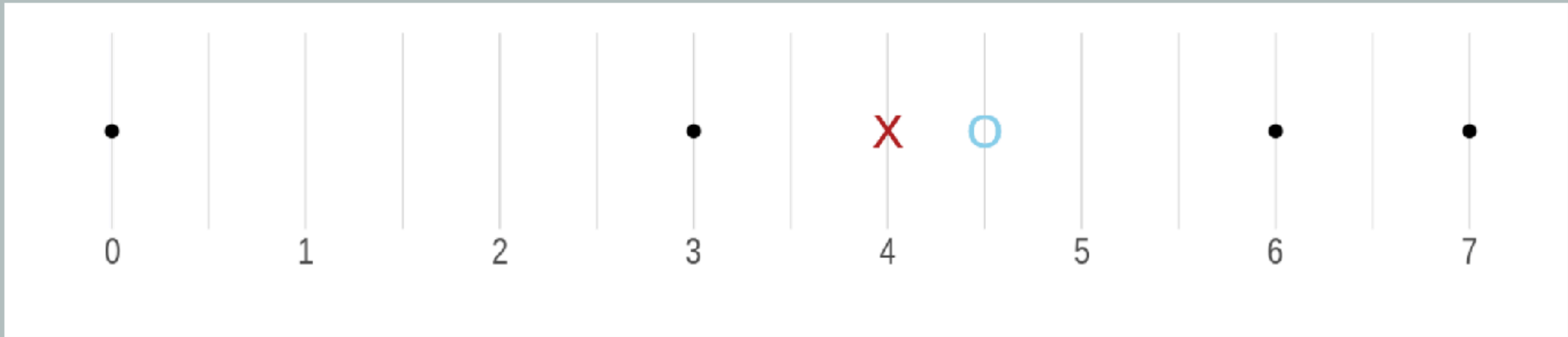
- the mean can be conceptualized also as the value you would expect to gain when you sample once from the observed data
- useful later to link this to the expected value of a random variable (but not important right now)

MEDIAN

If $\vec{x} = \langle x_1, \ldots, x_n \rangle$ is a vector of n data observations from an at least ordinal measure and if $\vec{x}$ is ordered such that for all $1 \leq i < n$ we have $x_i \leq x_{i+1}$, the **median** is the value x_i such that the number of data observations that are bigger or equal to x_i and the number of data observations that are smaller or equal to x_i are equal.

MEDIAN :: EXAMPLE

Example. The median of the vector $\vec{x} = \langle 1 = 0, 3, 6, 7 \rangle$ does not exist by the definition given above. However, for metric measures, where distances between measurements are meaningful, it is customary to take the two values "closests to where the median should be" and average them. In the example at hand, this would be $\frac{3+6}{2} = 4.5$. The plot below shows the data points in black, the mean as a red cross (as before) and the median as a blue circle



CALCULATING THE MEDIAN IN R

```
avocado_data %>%
  group_by(type) %>%
  summarise(
    mean_price   = mean(average_price),
    median_price = median(average_price)
  )
```

```
## # A tibble: 2 x 3
##   type         mean_price median_price
##   <chr>             <dbl>        <dbl>
## 1 conventional       1.16         1.13
## 2 organic            1.65         1.63
```

MEAN VS MEDIAN

- mean is more susceptible to outliers
- choice of mean vs. median is great for manipulation:
  - "How to mislead with statistics"

MODE

- the mode is the value that occurred most frequently in the data
- often not applicable to metric data (where each measurement, if fine-grained enough occurs only once)
- good for nominal and ordinal measures
- there is no built-in function in R to calculate the mode
  - caveat: function `mode` exists but is unrelated

VARIANCE

The variance $\mathrm{Var}(\vec{x})$ of a vector of metric observations $\vec{x}$ of length n is defined as the average of the squared distances from the mean:

$$\mathrm{Var}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\vec{x}})^2$$

VARIANCE :: EXAMPLE

$$\mathrm{Var}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\vec{x}})^2$$

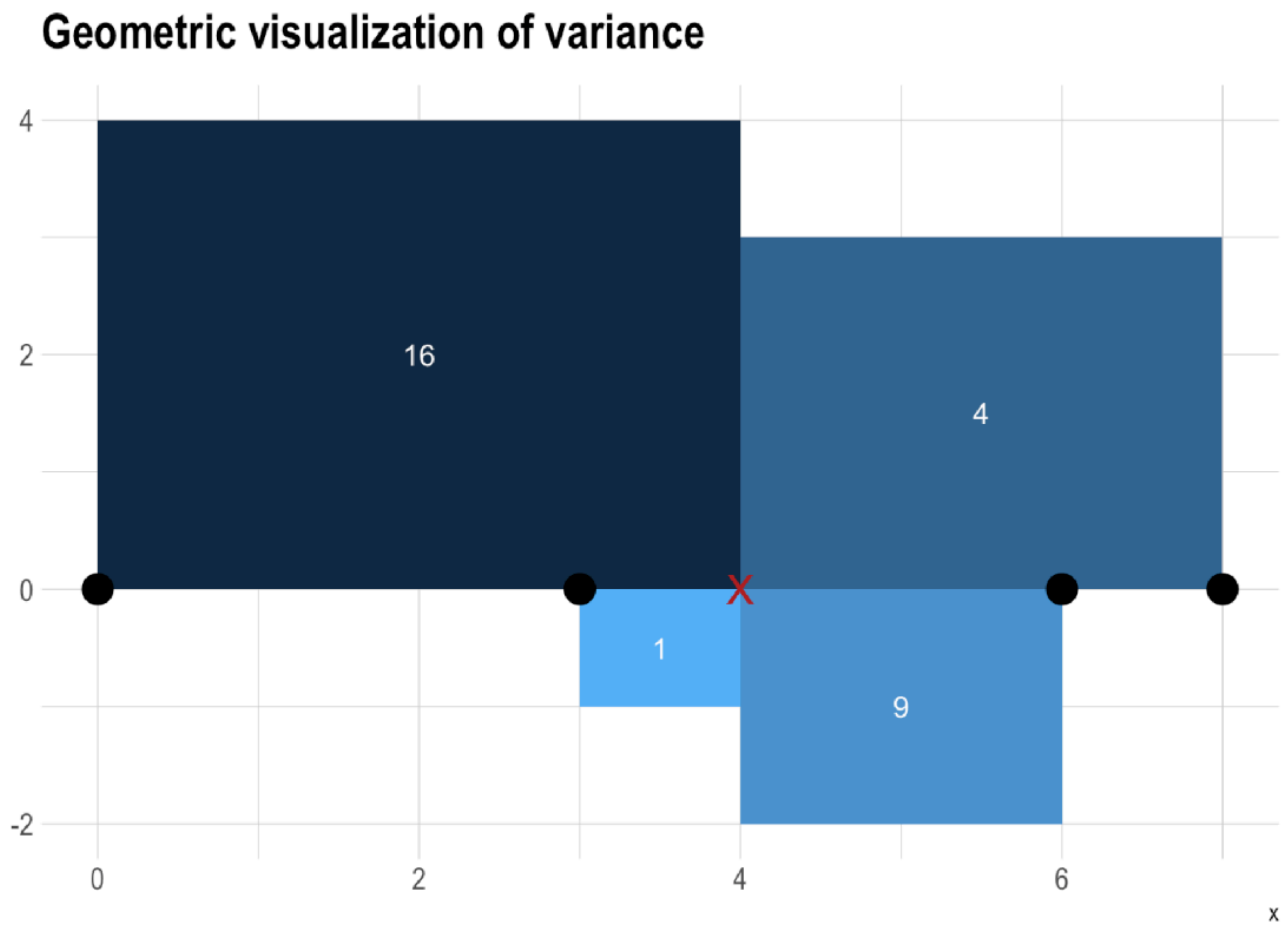
Example. The variance of the vector $\vec{x}$ $\langle 0, 3, 6, 7\rangle$ is computed as:

$$\mathrm{Var}(\vec{x}) = \frac{1}{4}\left((0-4)^2 + (3-4)^2 + (6-4)^2 + (7-4)^2\right) =$$

$$\frac{1}{4}(16+1+4+9) = \frac{30}{4} = 7.5$$

VARIANCE :: EXAMPLE

$$\mathrm{Var}(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\vec{x}})^2$$



$$\mathrm{Var}(\vec{x}) = \frac{1}{4}\left((0-4)^2 + (3-4)^2 + (6-4)^2 + (7-4)^2\right) =$$

$$\frac{1}{4}(16 + 1 + 4 + 9) = \frac{30}{4} = 7.5$$

VARIANCE :: BIASED AND UNBIASED ESTIMATORS

- biased estimator (unless mean is known)

$$\mathrm{Var}(\vec{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})^2$$

- unbiased estimator (if mean is estimated from data as well)

$$\mathrm{Var}(\vec{x}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})^2$$

- R's built-in function `var` calculates the unbiased estimator!

STANDARD DEVIATION

The standard deviation $\text{SD}(\vec{x})$ or numeric vector $\vec{x}$ is just the square root of the variance:

$$\text{SD}(\vec{x}) = \sqrt{\text{Var}(\vec{x})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{\vec{x}})^2}$$

VARIANCE & STANDARD DEVIATION :: EXAMPLE

```
avocado_data %>%
  group_by(type) %>%
  summarize(
    variance_price = var(average_price),
    stddev_price   = sd(average_price),
  )
```

```
## # A tibble: 2 x 3
##   type         variance_price stddev_price
##   <chr>                 <dbl>        <dbl>
## 1 conventional         0.0692        0.263
## 2 organic              0.132         0.364
```

QUANTILE

- the *k% quantile* is a value so that k% of the data are smaller

```r
quantile(
  # vector of observations
  x = avocado_data$average_price,
  # which quantiles
  probs = c(0.1, 0.25, 0.5, 0.85)
)
```

```
##  10%  25%  50%  85%
## 0.93 1.10 1.37 1.83
```

CONFIDENCE ESTIMATES VIA BOOTSTRAPPING

- variance & standard deviation tell us how far around the mean the data dwells
- they do not tell us how good our estimate of the mean is
- we can use bootstrapping, a special instance of resampling methods for this purpose

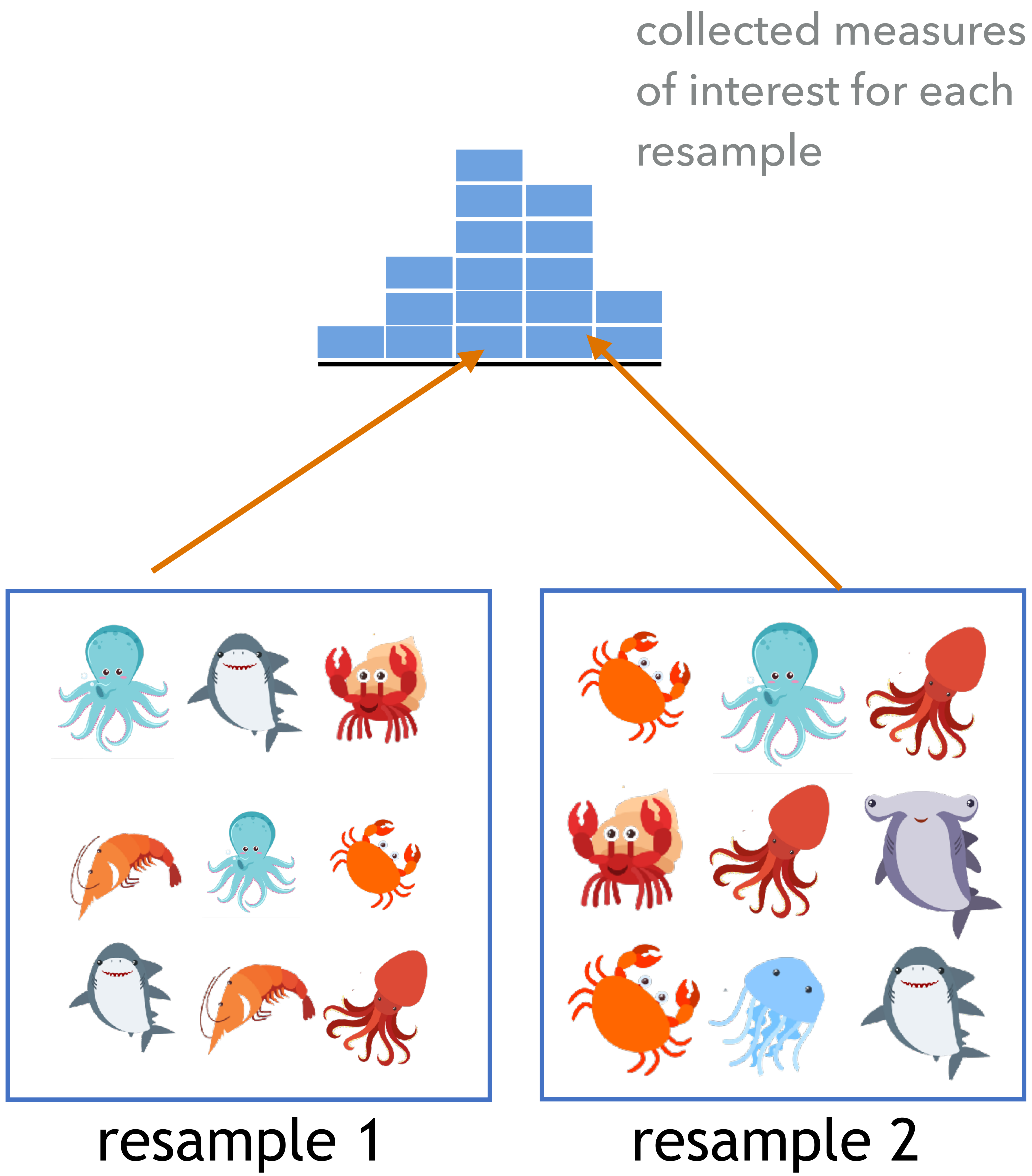
BOOTSTRAPPING 95 % CONFIDENCE INTERVALS FOR THE MEAN

An algorithm for constructing a 95% confidence interval of the mean of vector D of numeric data with length k looks as follows:

1. take k samples from D with replacement, call this D^{rep}
2. calculate the mean $\mu(D^{\text{rep}})$ of the newly sampled data
3. repeat steps 1 and 2 to gather r means of different resamples of D; call the result vector μ_{sampled}
4. the boundaries of the 95% inner quantile of μ_{sampled} are the bootstrapped 95% confidence interval of the mean

BOOTSTRAPPING 95 % CONFIDENCE INTERVALS

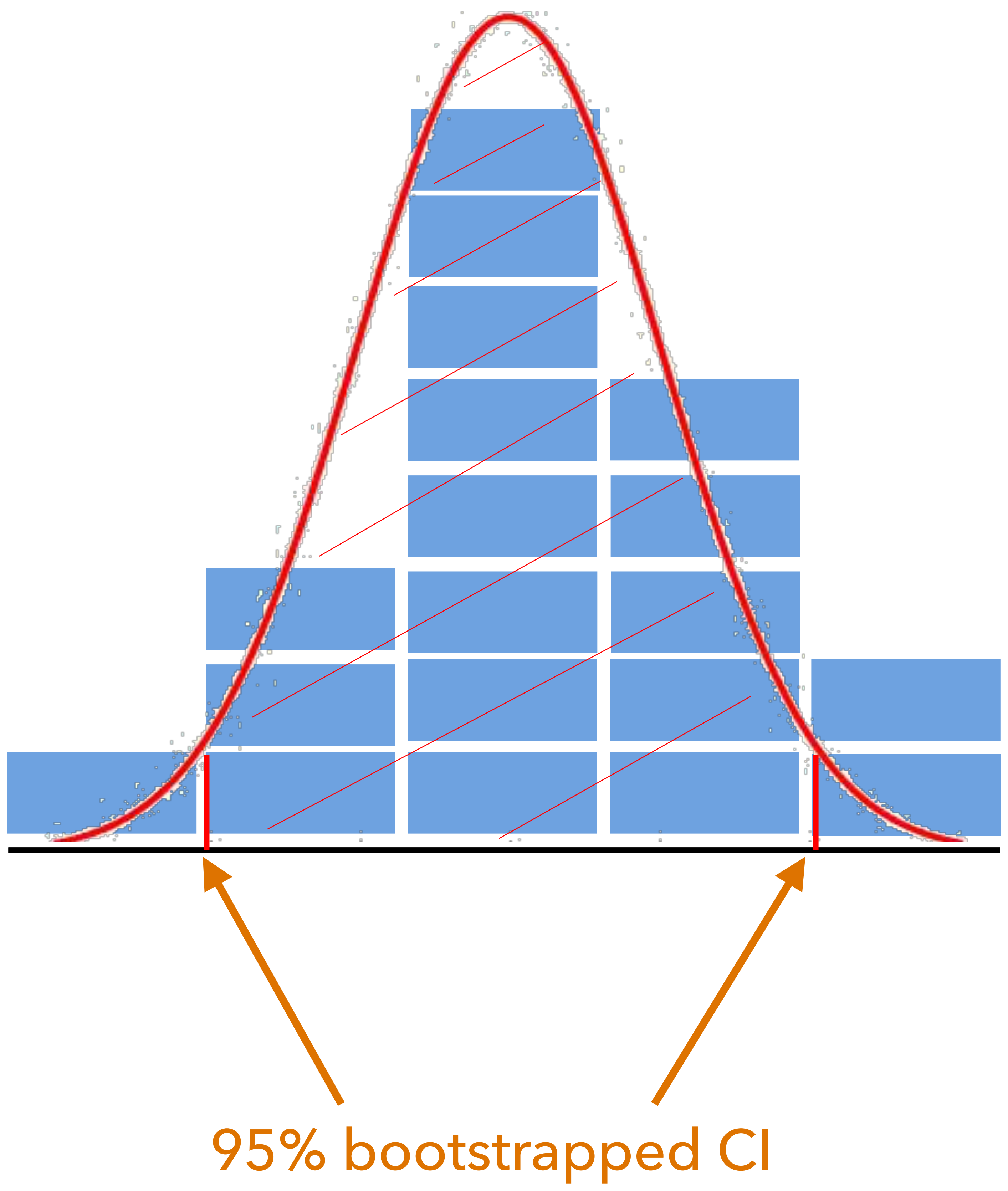
original data



- Fish: [Water vector created by brgfx - www.freepik.com](www.freepik.com)

BOOTSTRAPPING 95 % CONFIDENCE INTERVALS

original data



BOOTSTRAPPING IN R

```
## takes a vector of numbers and returns bootstrapped 95% ConfInt
## for the mean, based on `n_resamples` re-samples (default: 1000)
bootstrapped_CI <-  function(data_vector, n_resamples = 1000) {
  resampled_means <- map_dbl(1:n_resamples, function(i) {
      mean(sample(x = data_vector,
                  size = length(data_vector),
                  replace = T)
      )
    }
  )
  tibble(
    'lower' = quantile(resampled_means, 0.025),
    'mean'  = mean(data_vector),
    'upper' = quantile(resampled_means, 0.975)
  )
}
```

full data example

```
bootstrapped_CI(avocado_data$average_price)
```

```
## # A tibble: 1 x 3
##   lower  mean upper
##   <dbl> <dbl> <dbl>
## 1  1.40  1.41  1.41
```

partial data example

```
# first 300 observations of `average price` only
smaller_data = avocado_data$average_price[1:300]
bootstrapped_CI(smaller_data)
```

```
## # A tibble: 1 x 3
##   lower  mean upper
##   <dbl> <dbl> <dbl>
## 1  1.14  1.16  1.17
```

NESTED TIBBLES FOR GROUP SUMMARIES

```
avocado_data %>%
  group_by(type) %>%
  # nest all columns except grouping-column 'type' in a tibble
  # the name of the new column is 'price_tibbles'
  nest(.key = "price_tibbles") %>%
  # collect the summary statistics for each nested tibble
  # the outcome is a new column with nested tibbles
  summarise(
    CIs = map(price_tibbles, function(d) bootstrapped_CI(d$average_price))
  ) %>%
  # unnest the newly created nested tibble
  unnest(CIs)
```

```
## # A tibble: 2 x 4
##   type         lower  mean upper
##   <chr>        <dbl> <dbl> <dbl>
## 1 conventional  1.15  1.16  1.16
## 2 organic       1.65  1.65  1.66
```

NESTING TABLES

Species	S.L	S.W	P.L	P.W
setosa	5.1	3.5	1.4	0.2
setosa	4.9	3.0	1.4	0.2
setosa	4.7	3.2	1.3	0.2
setosa	4.6	3.1	1.5	0.2
setosa	5.0	3.6	1.4	0.2
versi	7.0	3.2	4.7	1.4
versi	6.4	3.2	4.5	1.5
versi	6.9	3.1	4.9	1.5
versi	5.5	2.3	4.0	1.3
versi	6.5	2.8	4.6	1.5
virgini	6.3	3.3	6.0	2.5
virgini	5.8	2.7	5.1	1.9
virgini	7.1	3.0	5.9	2.1
virgini	6.3	2.9	5.6	1.8
virgini	6.5	3.0	5.8	2.2

→

Species	Measurements			
setosa	S.L	S.W	P.L	P.W
	5.1	3.5	1.4	0.2
	4.9	3.0	1.4	0.2
	4.7	3.2	1.3	0.2
	4.6	3.1	1.5	0.2
	5.0	3.6	1.4	0.2
versi	S.L	S.W	P.L	P.W
	7.0	3.2	4.7	1.4
	6.4	3.2	4.5	1.5
	6.9	3.1	4.9	1.5
	5.5	2.3	4.0	1.3
	6.5	2.8	4.6	1.5
virgini	S.L	S.W	P.L	P.W
	6.3	3.3	6.0	2.5
	5.8	2.7	5.1	1.9
	7.1	3.0	5.9	2.1
	6.3	2.9	5.6	1.8
	6.5	3.0	5.8	2.2

i.e.,

Species	Measurements
setosa	<tibble [50x4]>
versi	<tibble [50x4]>
virgini	<tibble [50x4]>

UNNESTING NESTED TABLES

Species	Measurements
setosa	<tibble [50x4]>
versi	<tibble [50x4]>
virgini	<tibble [50x4]>

i.e.,

Species	Measurements			
	S.L	S.W	P.L	P.W
setosa	5.1	3.5	1.4	0.2
	4.9	3.0	1.4	0.2
	4.7	3.2	1.3	0.2
	4.6	3.1	1.5	0.2
	5.0	3.6	1.4	0.2
	S.L	S.W	P.L	P.W
versi	7.0	3.2	4.7	1.4
	6.4	3.2	4.5	1.5
	6.9	3.1	4.9	1.5
	5.5	2.3	4.0	1.3
	6.5	2.8	4.6	1.5
	S.L	S.W	P.L	P.W
virgini	6.3	3.3	6.0	2.5
	5.8	2.7	5.1	1.9
	7.1	3.0	5.9	2.1
	6.3	2.9	5.6	1.8
	6.5	3.0	5.8	2.2

→

Species	S.L	S.W	P.L	P.W
setosa	5.1	3.5	1.4	0.2
setosa	4.9	3.0	1.4	0.2
setosa	4.7	3.2	1.3	0.2
setosa	4.6	3.1	1.5	0.2
setosa	5.0	3.6	1.4	0.2
versi	7.0	3.2	4.7	1.4
versi	6.4	3.2	4.5	1.5
versi	6.9	3.1	4.9	1.5
versi	5.5	2.3	4.0	1.3
versi	6.5	2.8	4.6	1.5
virgini	6.3	3.3	6.0	2.5
virgini	5.8	2.7	5.1	1.9
virgini	7.1	3.0	5.9	2.1
virgini	6.3	2.9	5.6	1.8
virgini	6.5	3.0	5.8	2.2

COVARIANCE

- covariance measures the degree to which two associated measurements show similar deviation from their respective means

$$\text{Cov}(\vec{x}, \vec{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})\,(y_i - \mu_{\vec{y}})$$

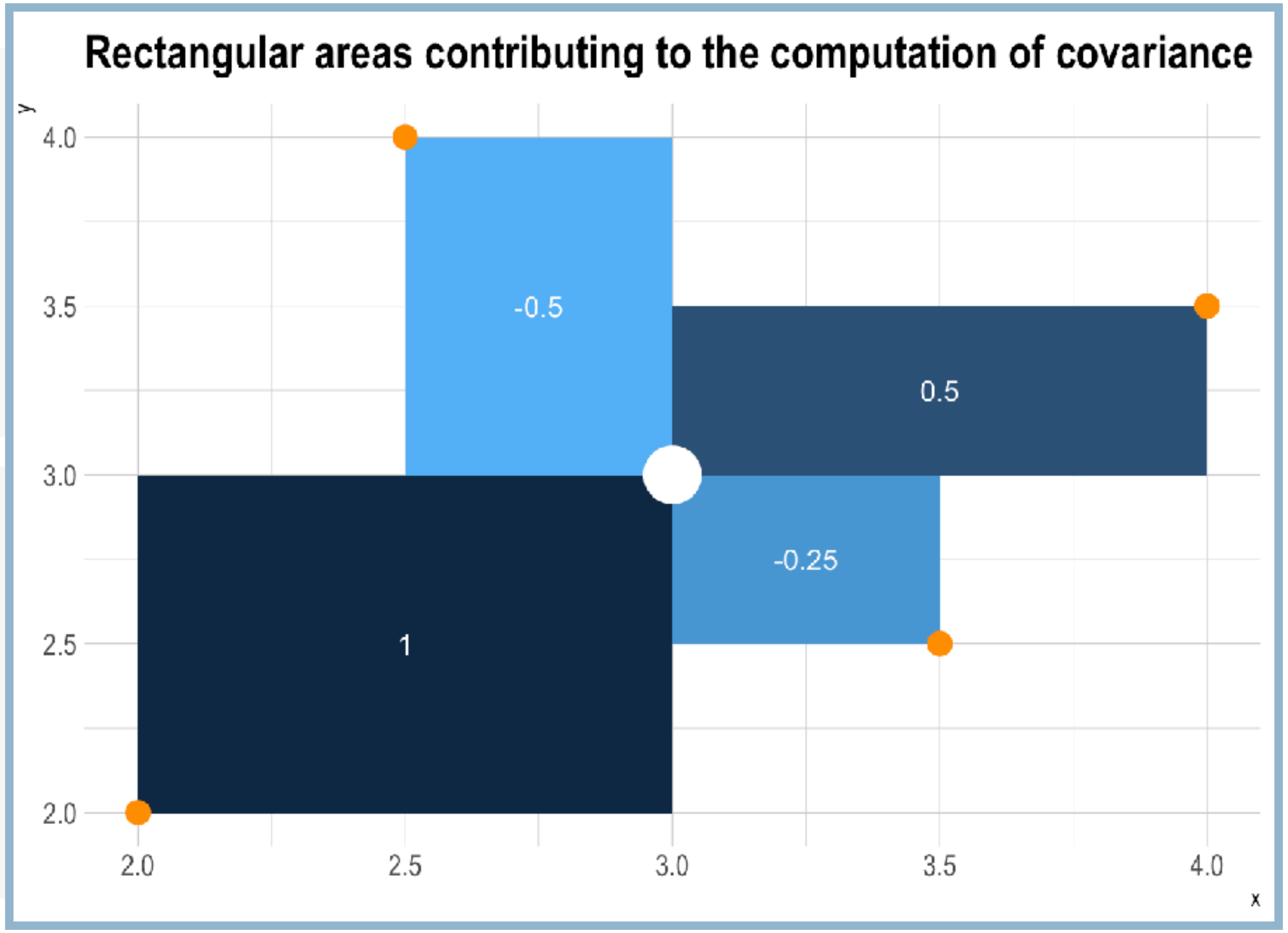
COVARIANCE :: EXAMPLE

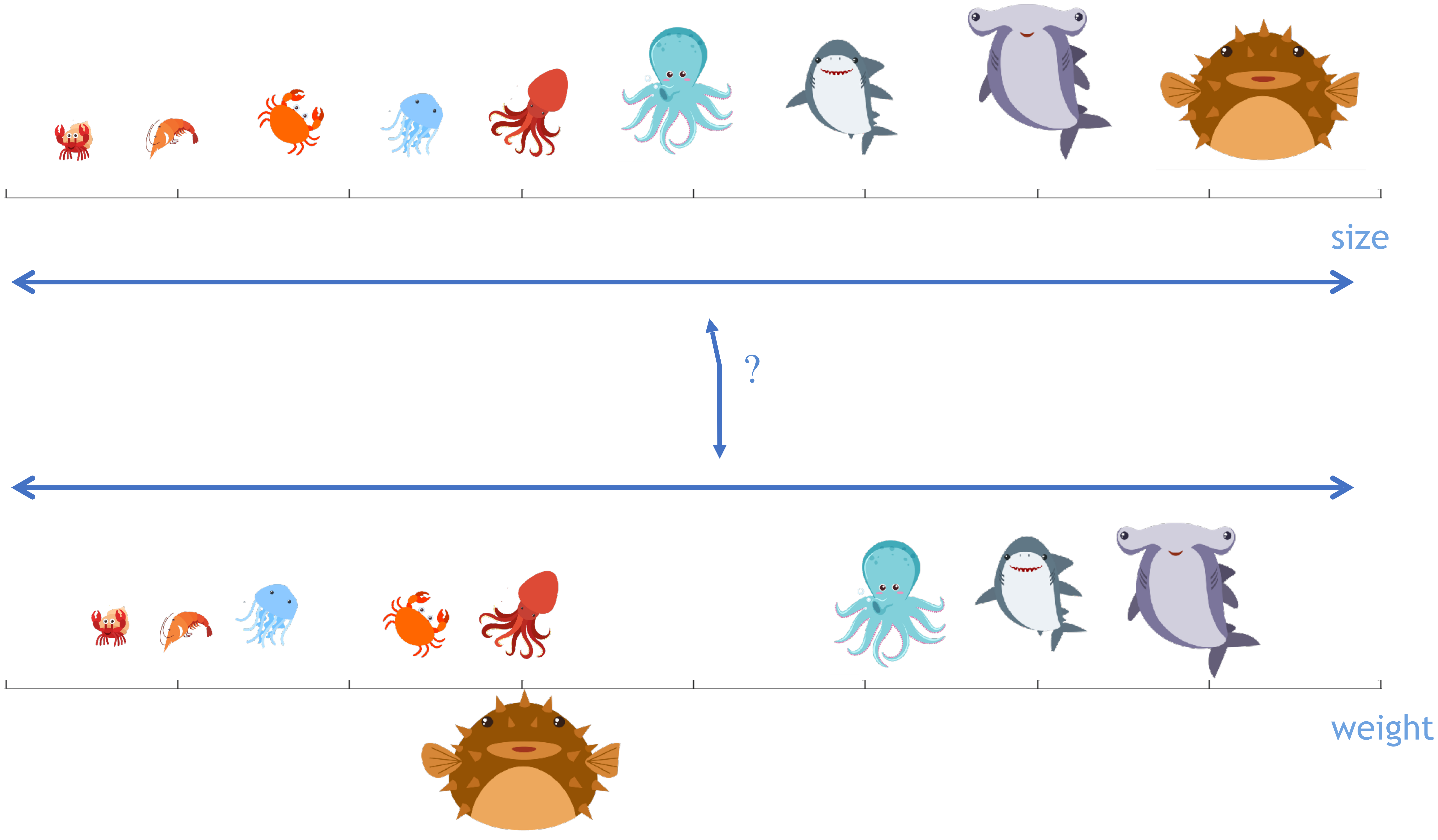
$$\text{Cov}(\vec{x}, \vec{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})\ (y_i - \mu_{\vec{y}})$$

```
contrived_example <-
  tribble(
    ~x,   ~y,
    2,    2,
    2.5,  4,
    3.5,  2.5,
    4,    3.5
  )
```

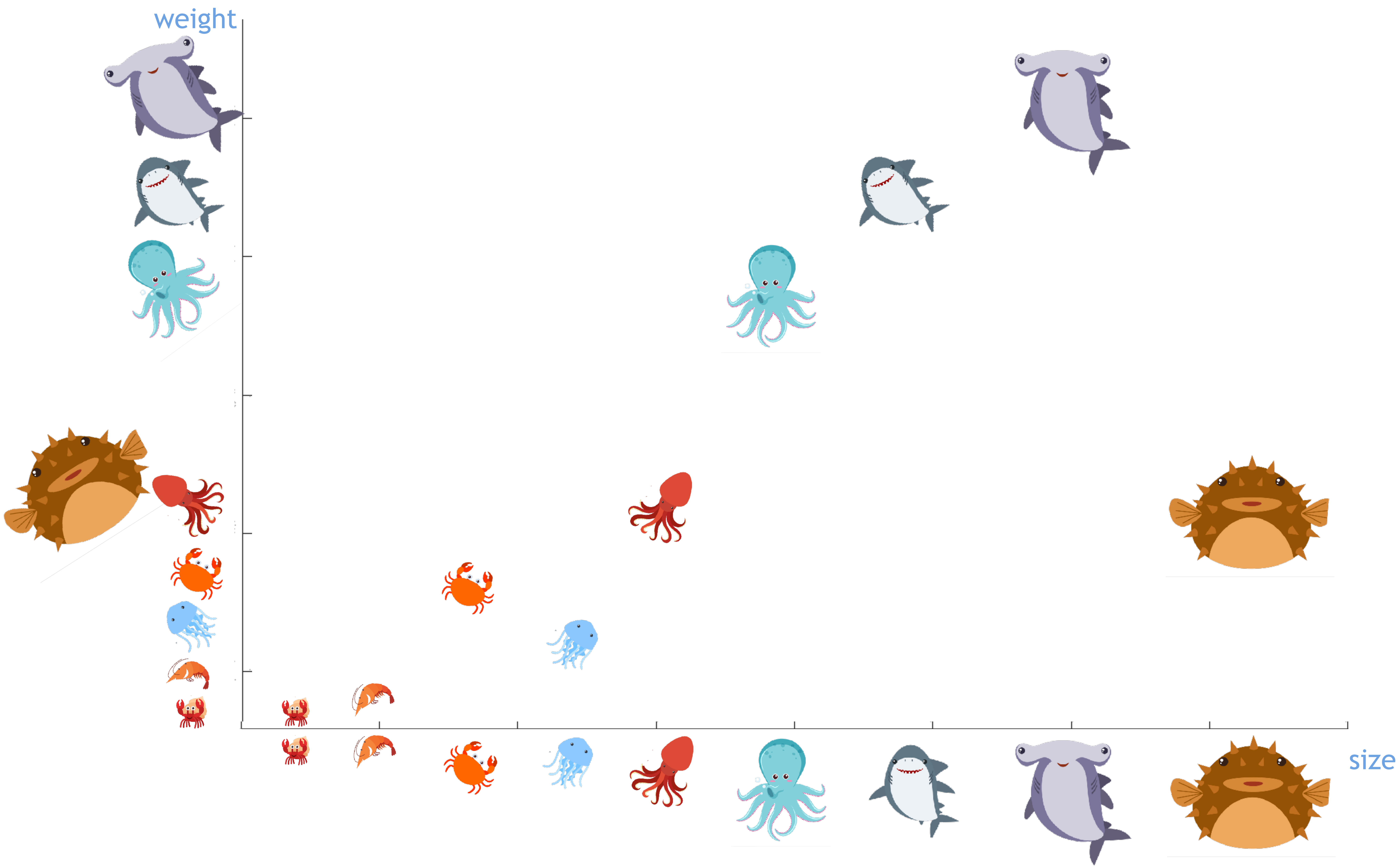
```
## # A tibble: 4 x 4
##       x     y area_rectangle covariance
##   <dbl> <dbl>          <dbl>      <dbl>
## 1   2     2             1          0.25
## 2   2.5   4            -0.5        0.25
## 3   3.5   2.5          -0.25       0.25
## 4   4     3.5           0.5        0.25
```

```
contrived_example <-
  contrived_example %>%
  mutate(
    area_rectangle =              (x-mean(x)) * (y - mean(y)),
    covariance     = 1/ (n()-1) * sum((x-mean(x)) * (y - mean(y)))
  )
contrived_example
```

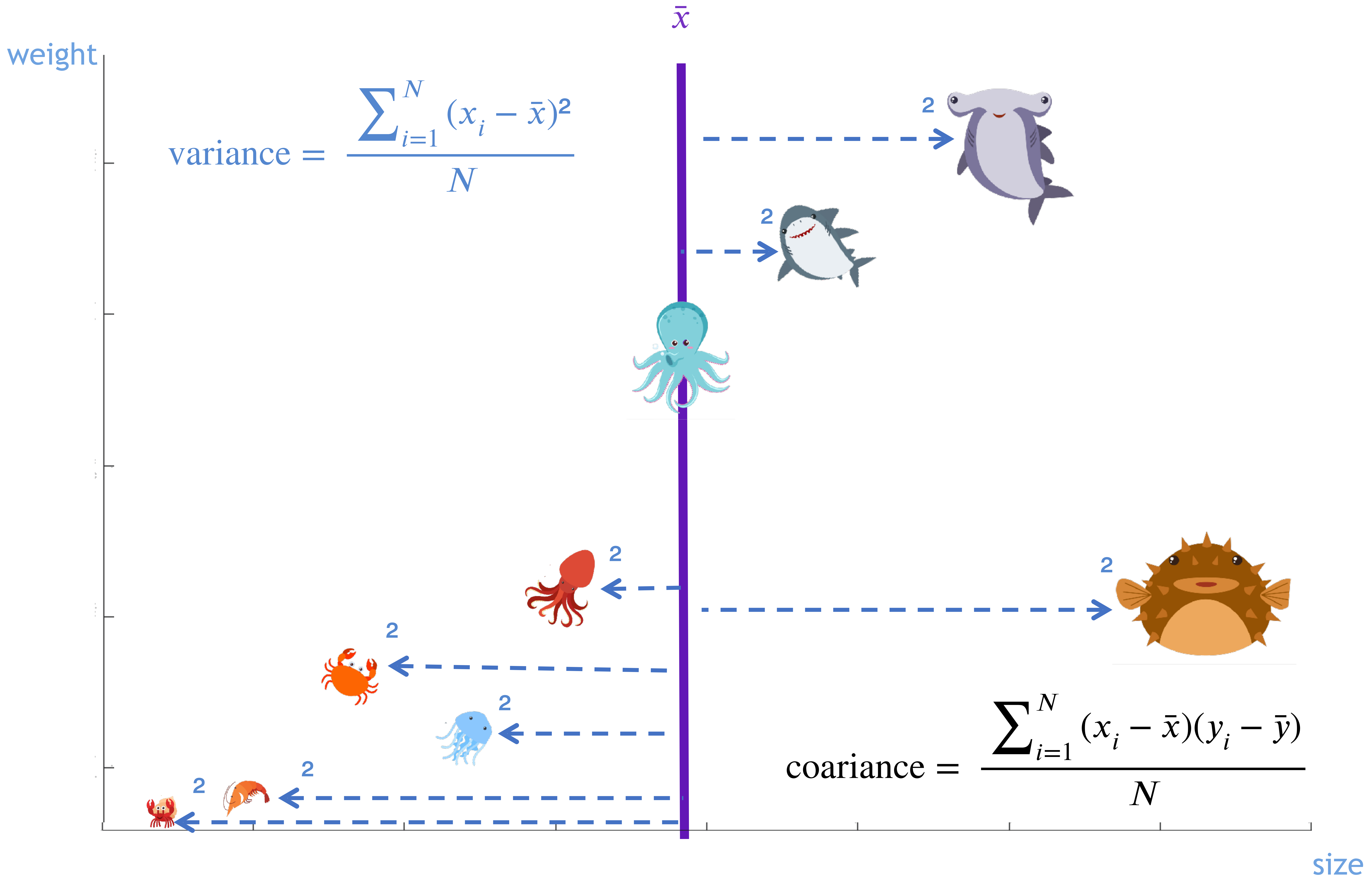




Maria Pershina • Jona Carmon

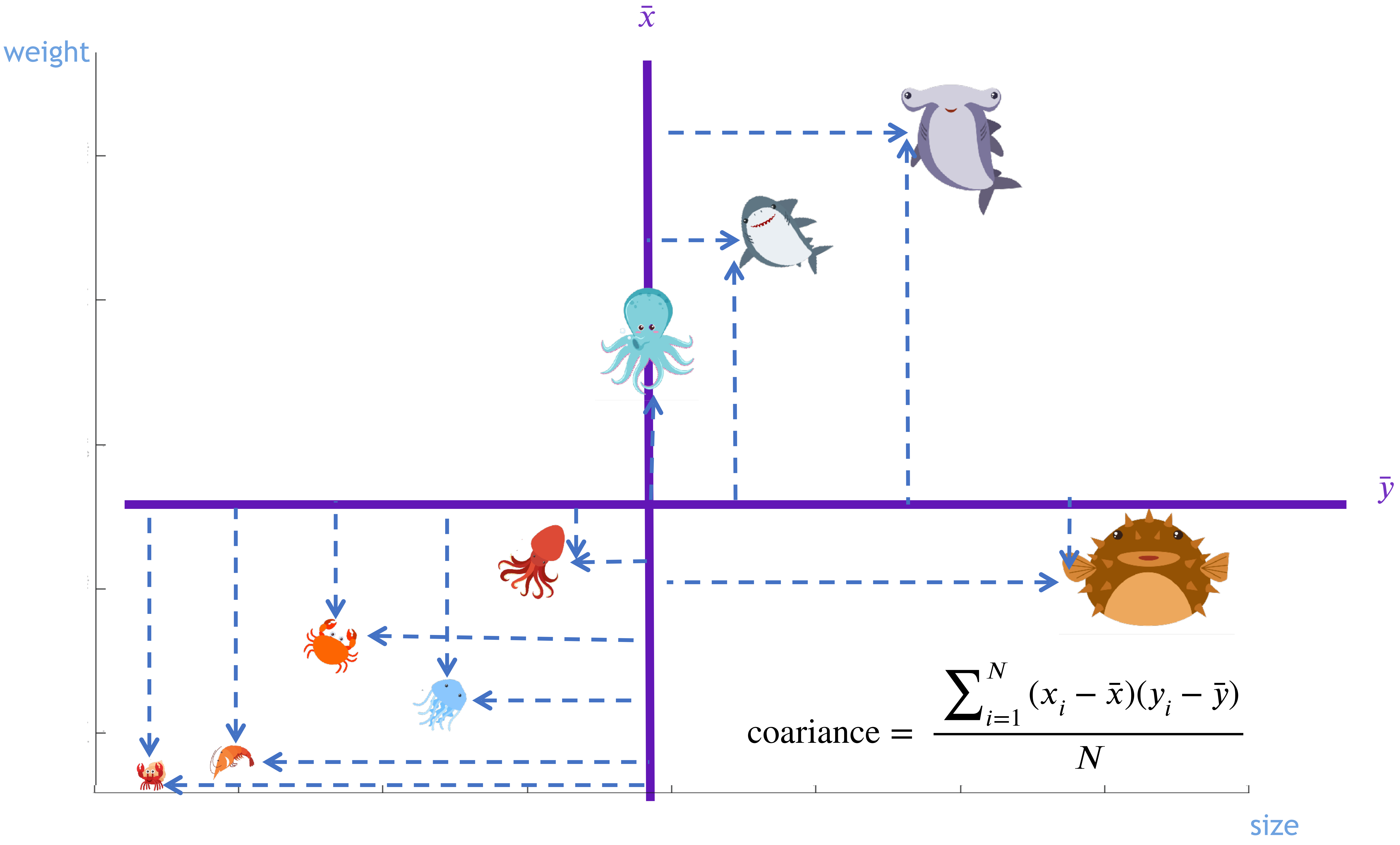


weight
size

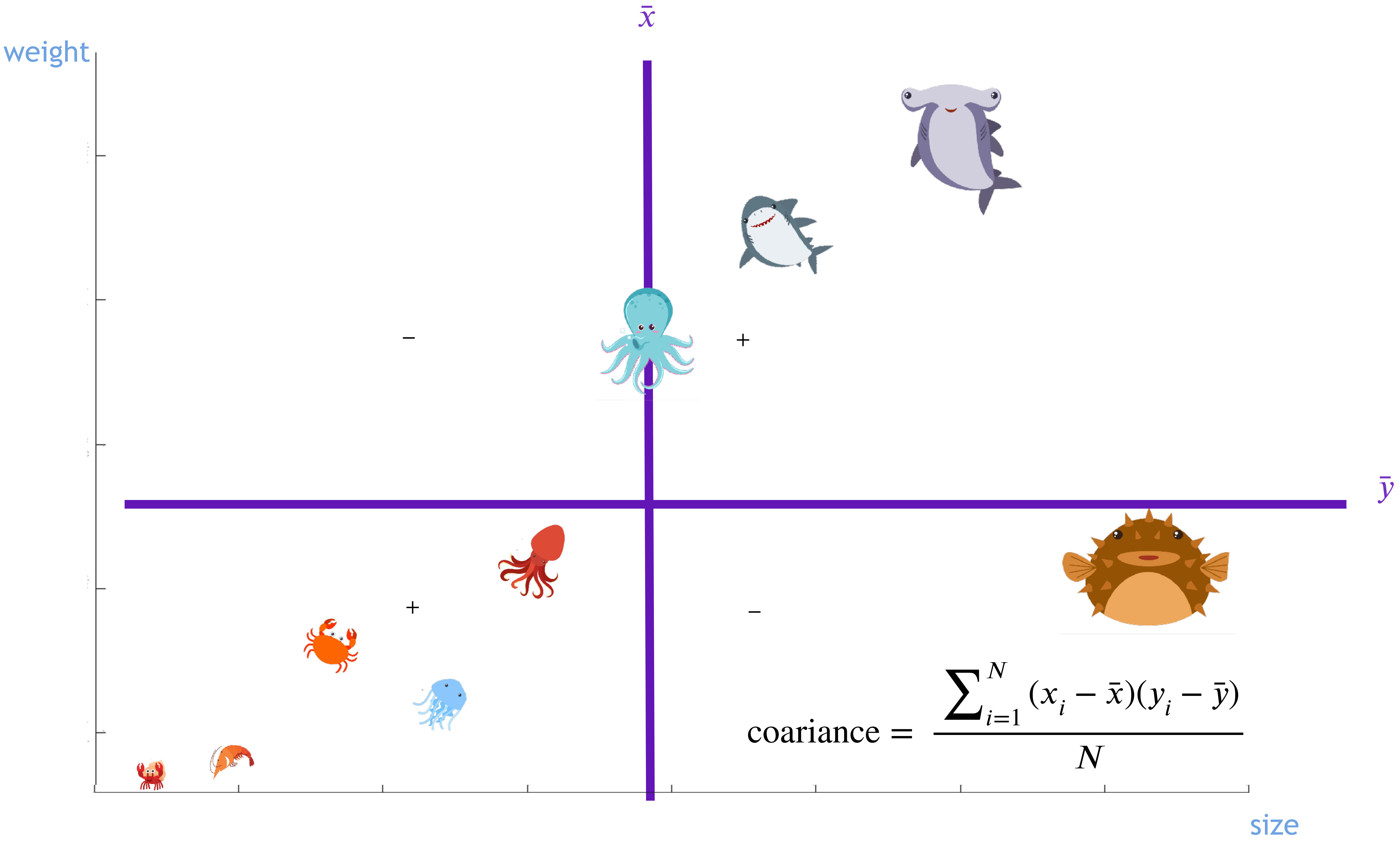


$$\text{variance} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}$$

$$\text{coariance} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$



$$\text{coariance} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$



$$\text{coariance} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

COVARIANCE :: INTERPRETATION

$$\text{Cov}(\vec{x}, \vec{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu_{\vec{x}})\ (y_i - \mu_{\vec{y}})$$

- summands are positive when x_i and y_i deviate "in the same direction" from their respective means
- positive (negative) covariance therefore reflects an overall tendency that that higher x_i, the higher (lower) y_i
- this is a descriptive property of the data, not an evidential indicator of a causal relation

COVARIANCE :: SCALE VARIANCE

- covariance is not invariant under positive linear transformation

```r
with(contrived_example, cov(x,y))
```

```
## [1] 0.25
```

```r
with(contrived_example, cov(x,1000 * y + 500))
```

```
## [1] 250
```

PRODUCT-MOMENT CORRELATION

- Bravais-Pearson product-moment correlation coefficient is defined as covariance standardized by std. deviations

$$r_{\vec{x}\vec{y}} = \frac{\text{Cov}(\vec{x}, \vec{y})}{\text{SD}(\vec{x})\ \text{SD}(\vec{y})}$$

CORRELATION :: EXAMPLE

- correlation is invariant under positive linear transformation

```
with(contrived_example, cor(x,y))
```

```
## [1] 0.3
```

```
with(contrived_example, cor(x,1000 * y + 500))
```

```
## [1] 0.3
```

CORRELATION :: EXAMPLE

```
with(avocado_data, cor(log(total_volume_sold), average_price))
```

```
## [1] -0.5834087
```

- negative correlation indicates an overall negative association: the higher total-volume-sold, the lower the average price

CORRELATION :: PROPERTIES & INTERPRETATION

- r lies in [-1;1]
- $r = 0$ indicates no correlation at all
- $r = 1$ indicates perfect positive correlation
- $r = -1$ indicates perfect negative correlation
- $r >= 0.5$ suggests noteworthy (pos.) correlation
- $r <= -0.5$ suggests noteworthy (neg.) correlation
- r^2 also interpretable as "variance explained" in a regression model (later)