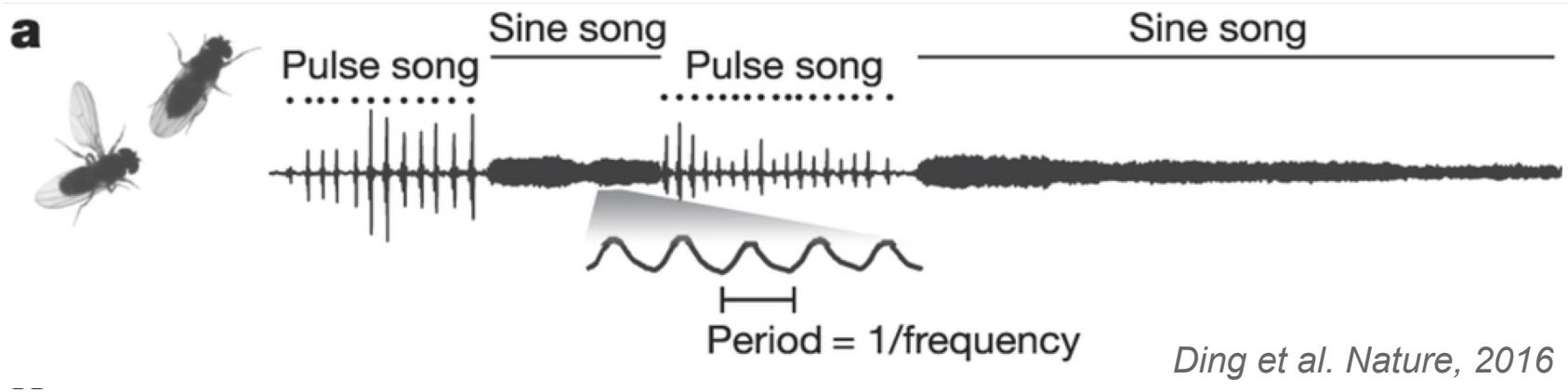
Data

The raw material of statistics **and biology**



Ding et al. Nature, 2016

fly male sine song

(frequencies in Hz)

Drosophila simulans (strain 5)

$$\begin{aligned} \textit{sim5} \quad = \{ \quad & 178.86, 174.13, 172.95, 172.71, 179.20, 175.24, \\ & 175.79, 176.27, 176.19, 176.07, 175.48, 167.09, \\ & 174.09, 174.31, 169.45, 176.58, 176.54, 175.31, \\ & 180.66, 178.74, 177.37, 175.31, 177.62, 174.91 \} \end{aligned}$$

Drosophila mauritiana (strain 29)

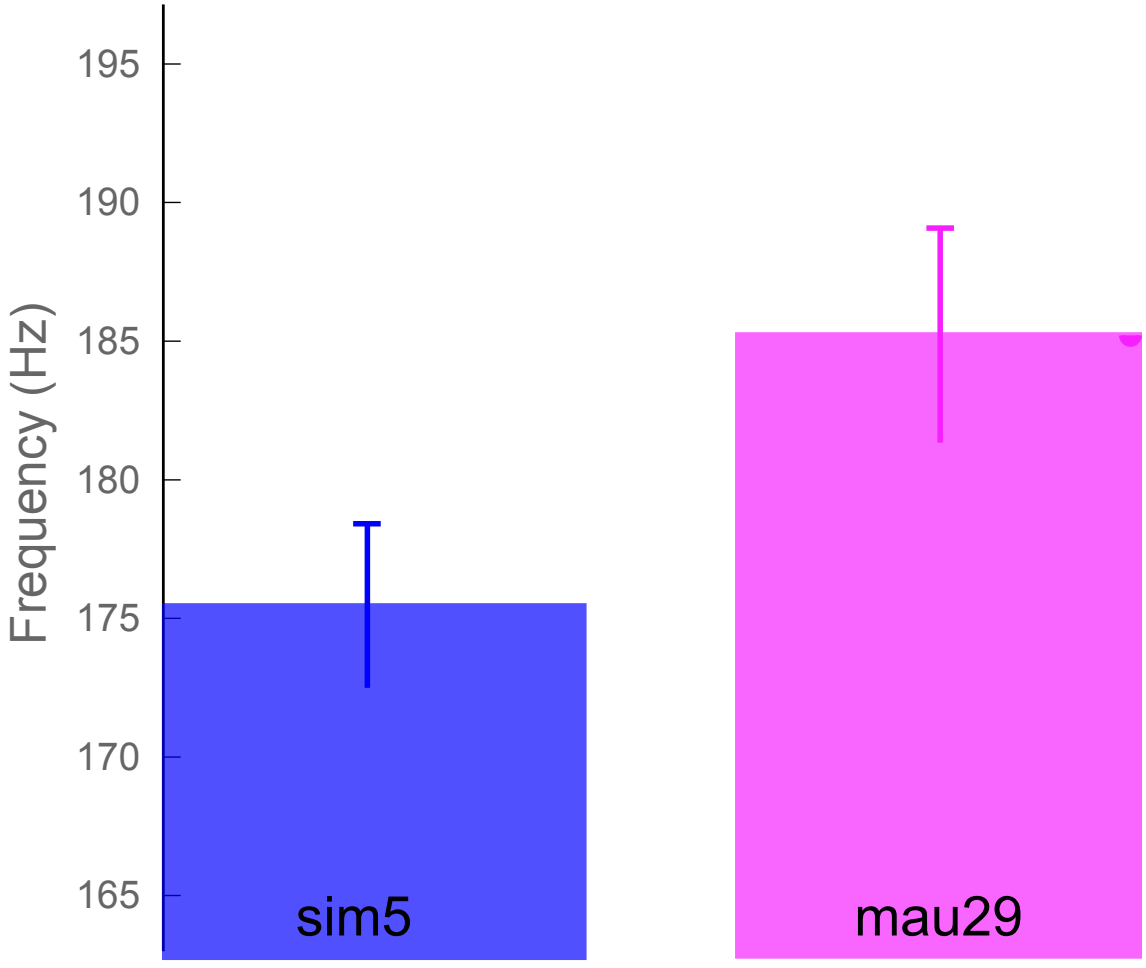
$$\begin{aligned} \textit{mau29} \quad = \{ \quad & 185.38, 187.66, 189.28, 176.37, 181.37, 181.25, \\ & 189.65, 189.34, 188.62, 184.70, 179.44, 186.31, \\ & 189.00, 185.12, 191.46, 187.72, 184.17, 187.45, \\ & 180.21, 181.92, 183.87, 181.16, 184.21, 188.15 \} \end{aligned}$$

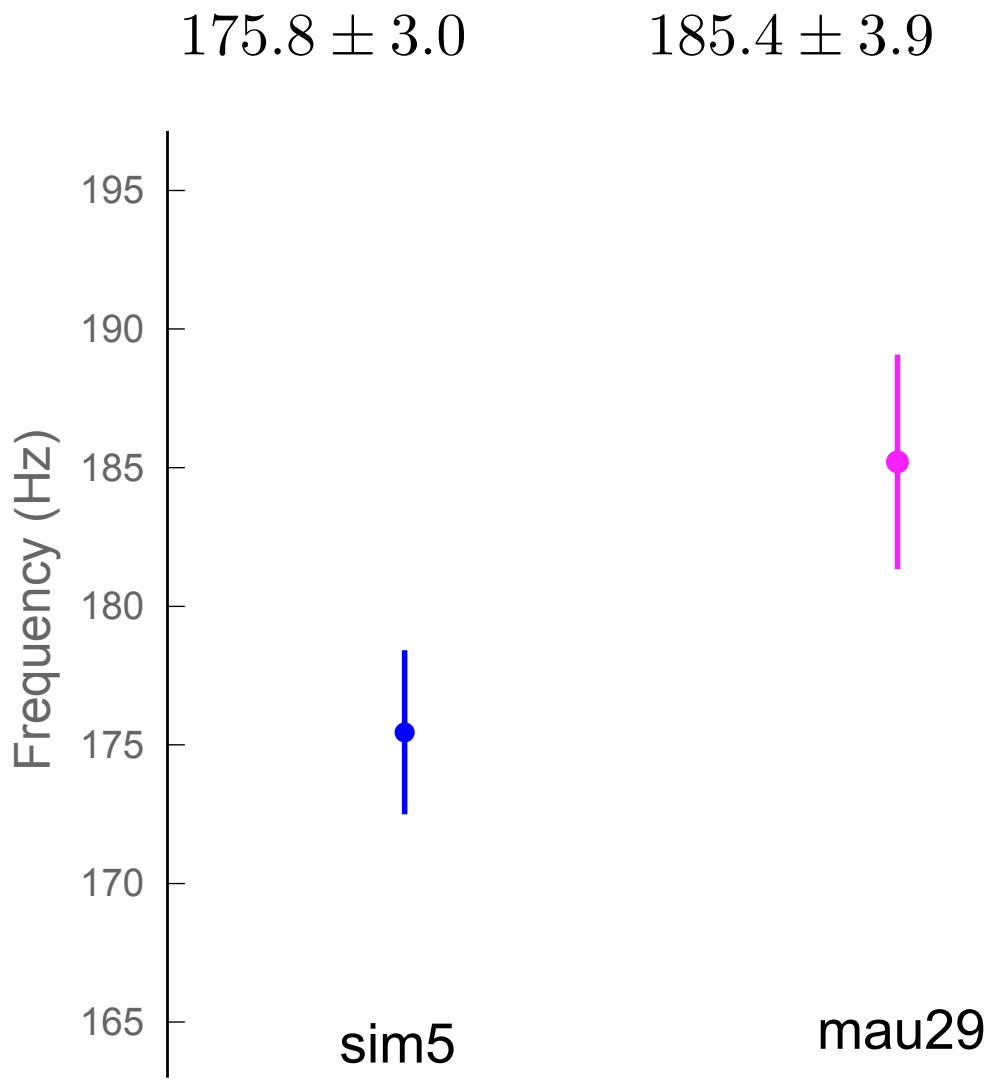
Ding *et al.*, Nature 2016

Show your data

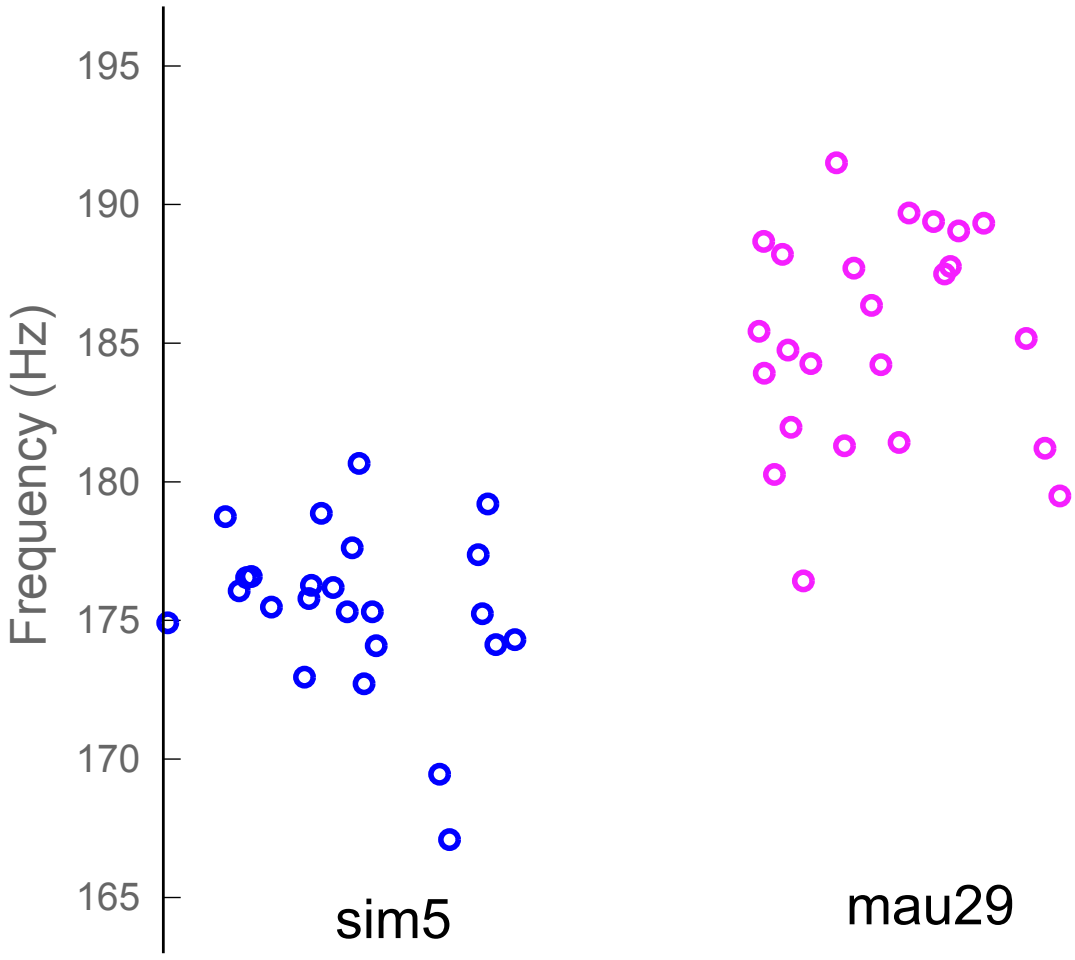
The way you present your data is crucial in making a point

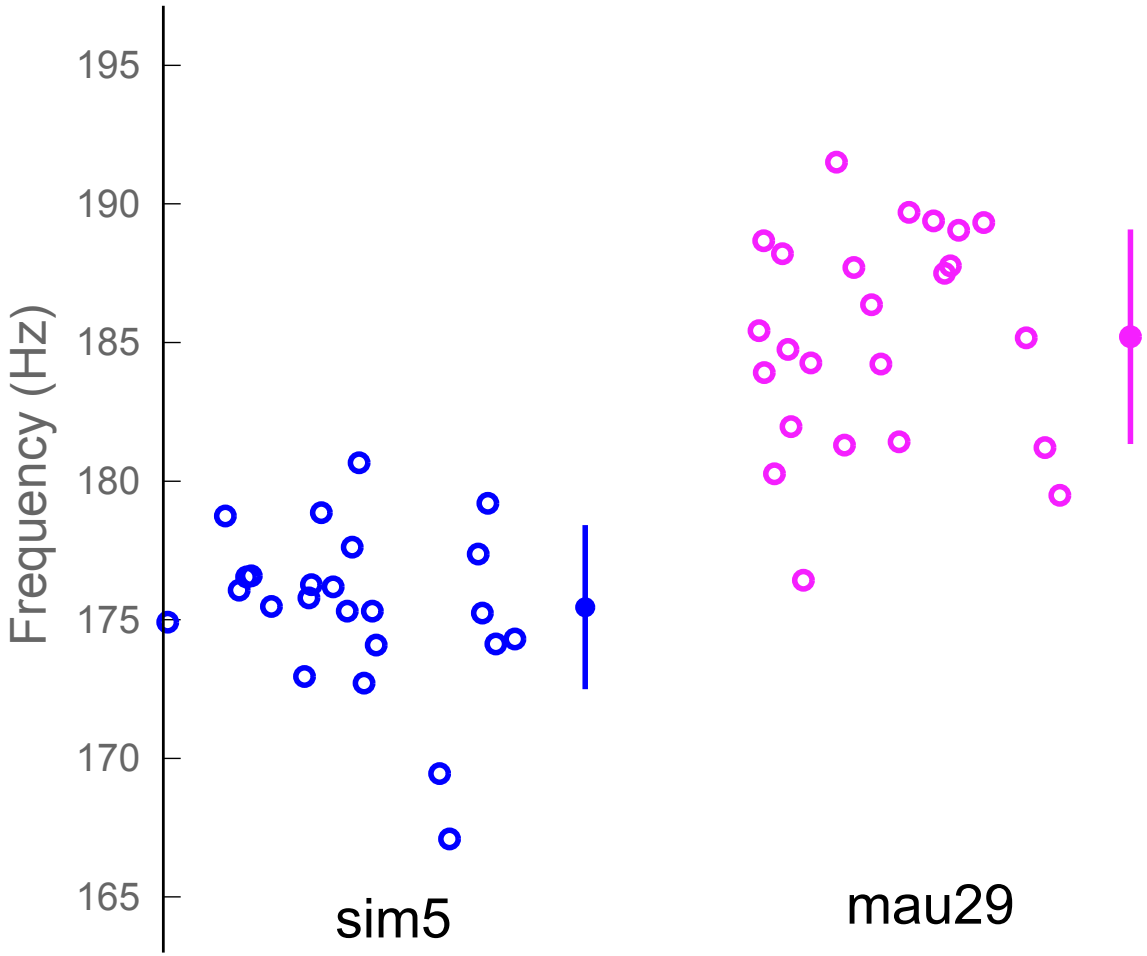
Summary statistics

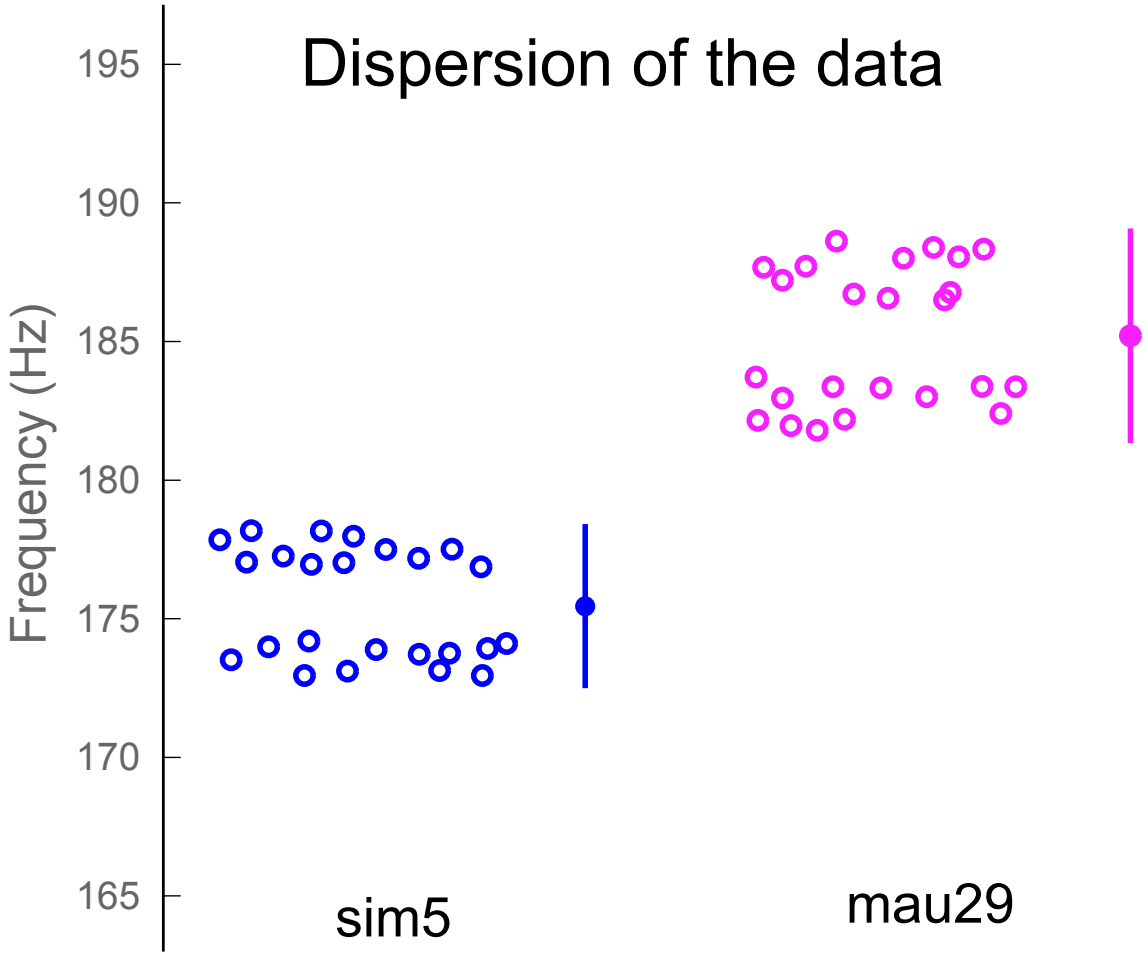


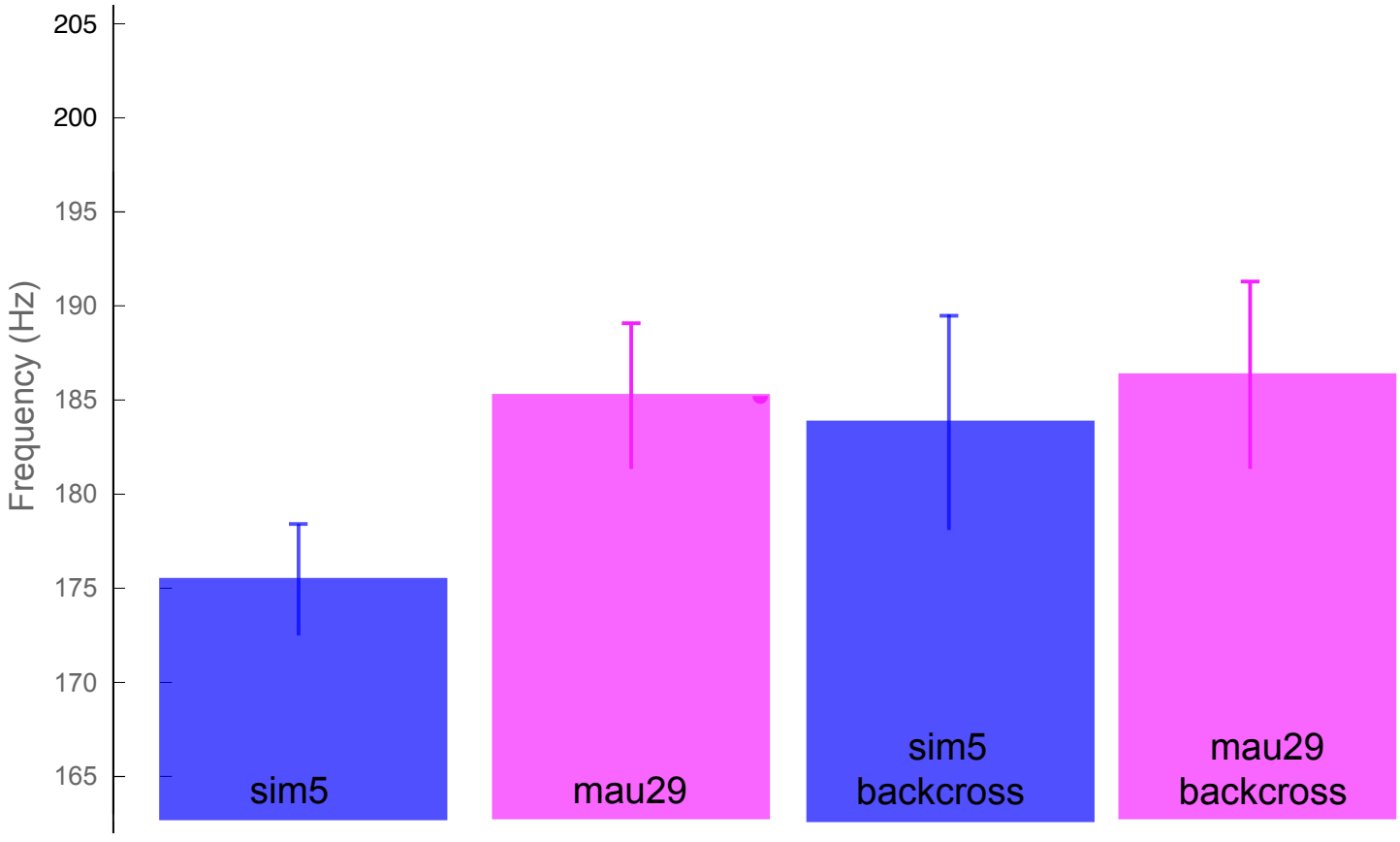

175.8 ± 3.0
185.4 ± 3.9
195
190
185
180
175
170
165
Frequency (Hz)
sim5
mau29

Beeswarm plot

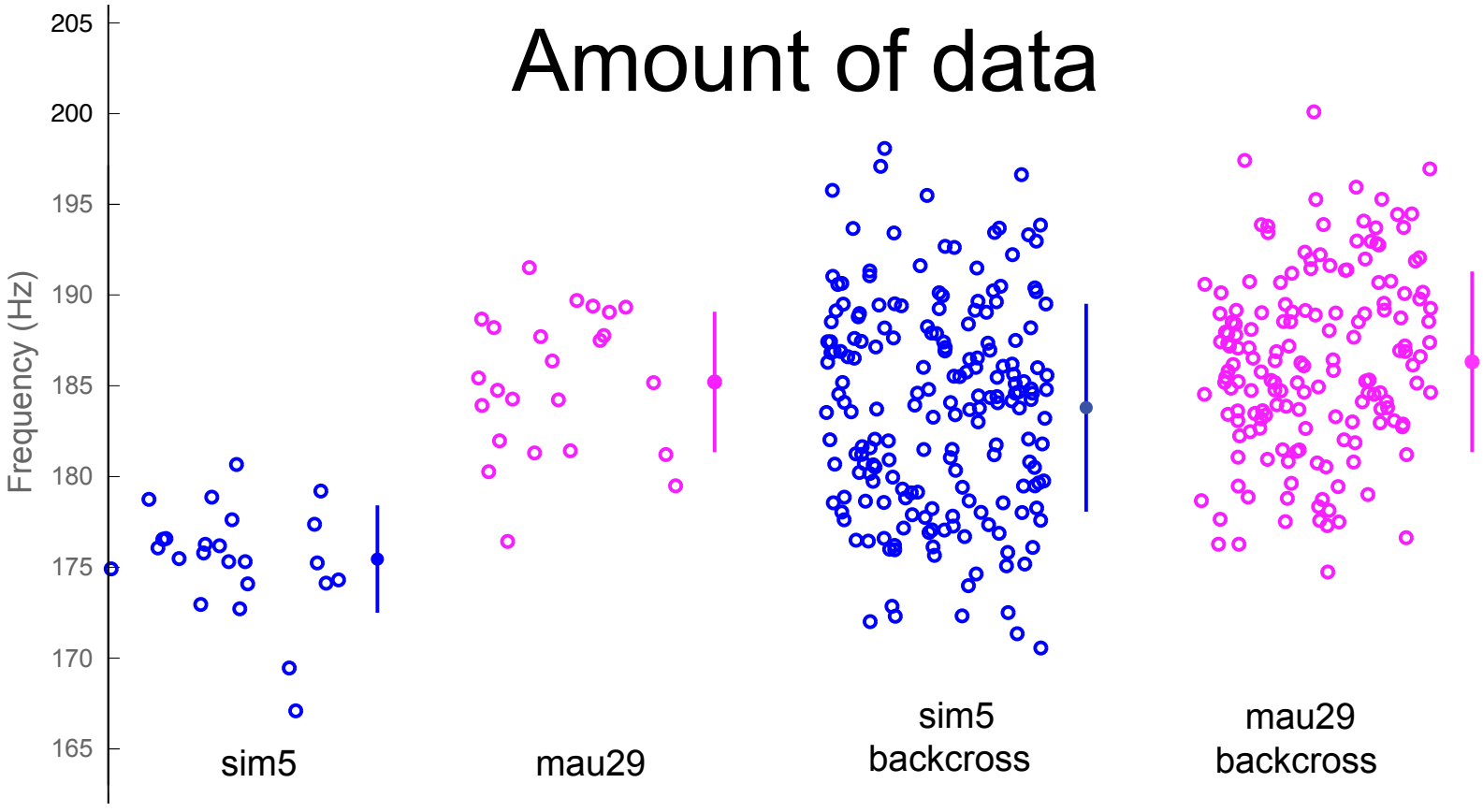



Frequency (Hz)
195
190
185
180
175
170
165
sim5
mau29


Dispersion of the data
Frequency (Hz)
195
190
185
180
175
170
165
sim5
mau29


Frequency (Hz)
205
200
195
190
185
180
175
170
165
sim5
mau29
sim5
backcross
mau29
backcross

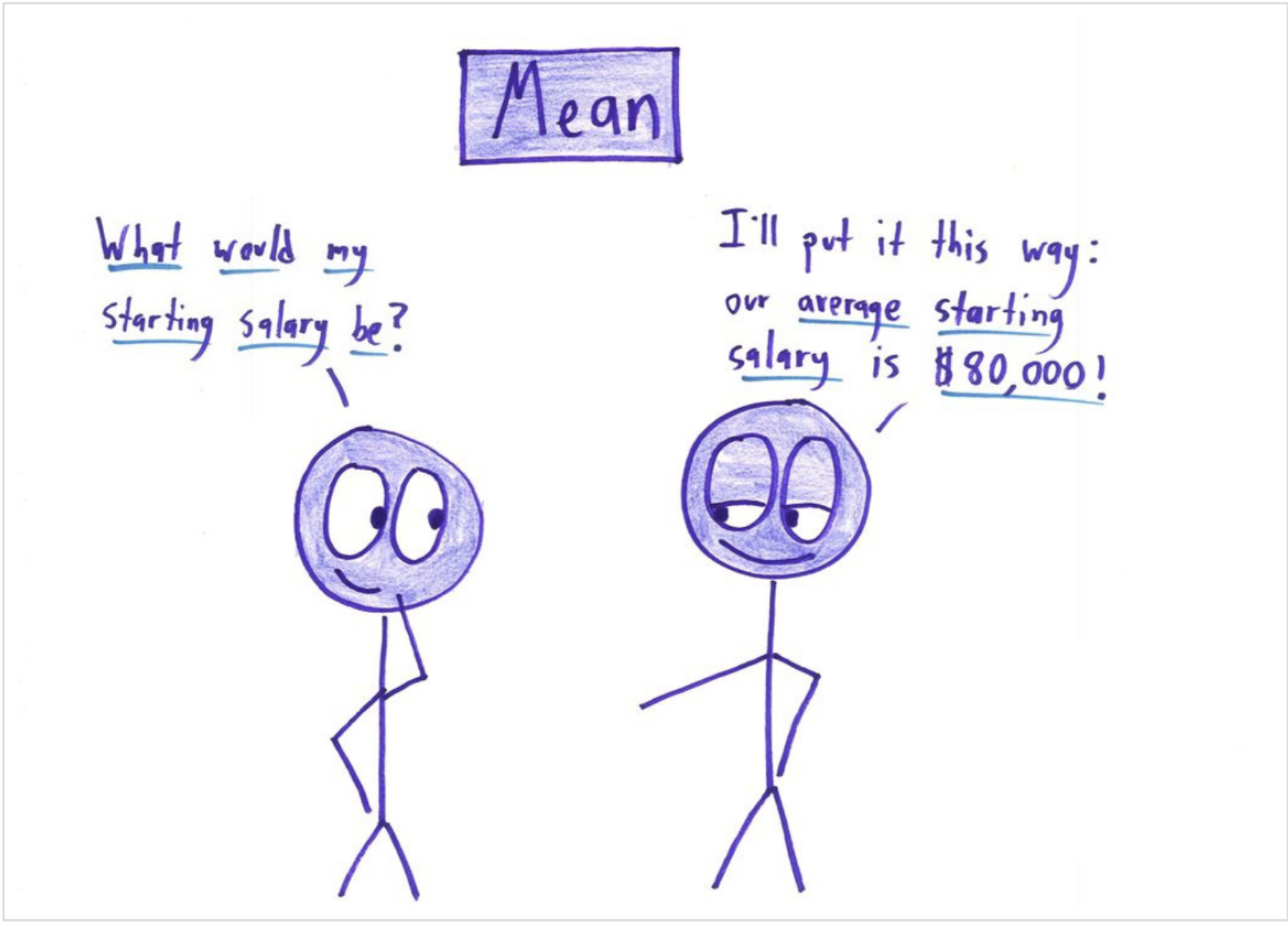
Amount of data



Any single statistic may fool you!

source: mathwithbaddrawings.com

Why Not to Trust Statistics



mean (μ) of a data set is found by adding all numbers and dividing by the number of values in the dataset

you → $ 30,000

all your coworkers
- $ 30,000
- $ 30,000
- $ 30,000
- $ 30,000
- $ 30,000
- $ 30,000

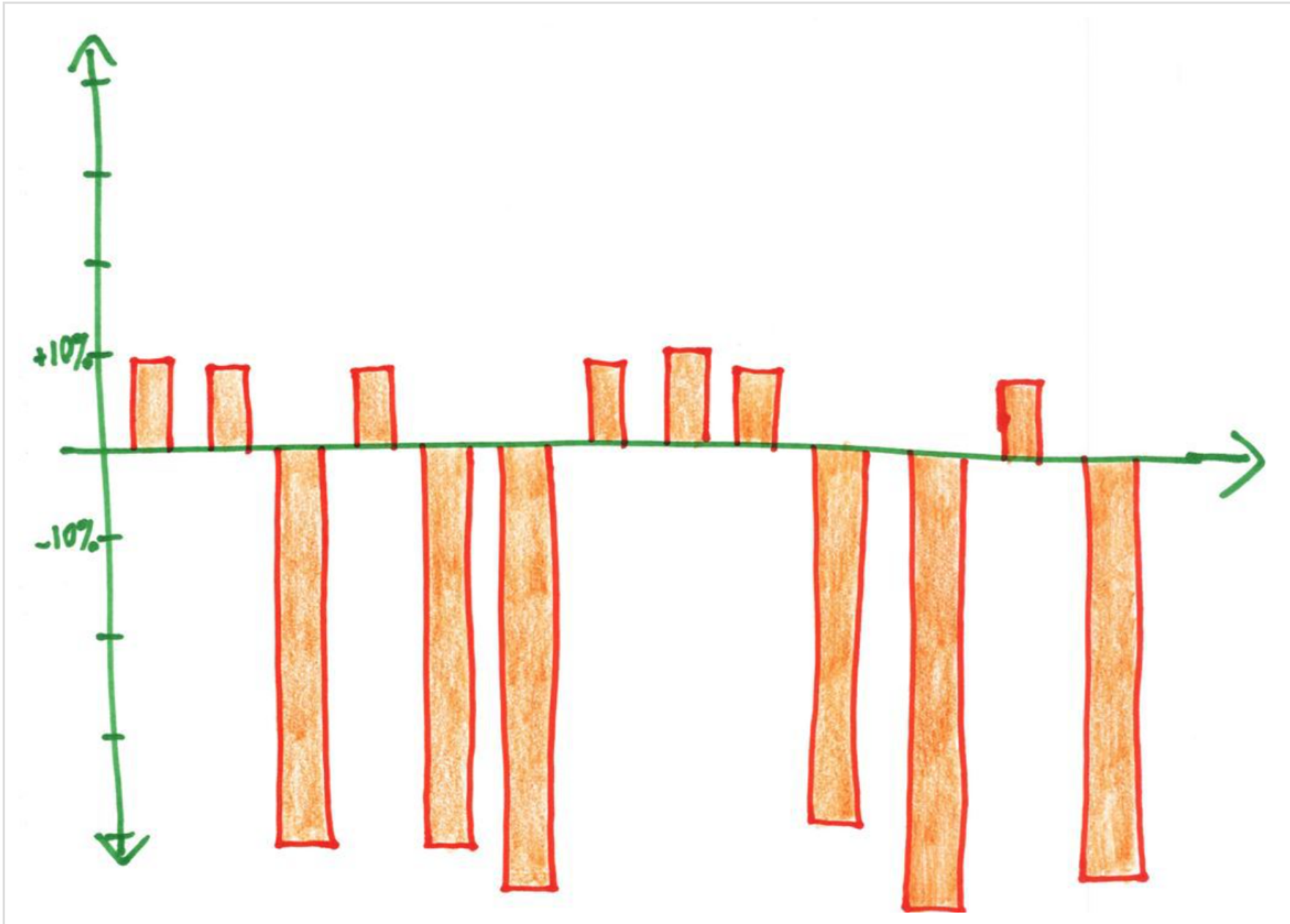
CEO's son → $ 430,000

Average: $80,000.

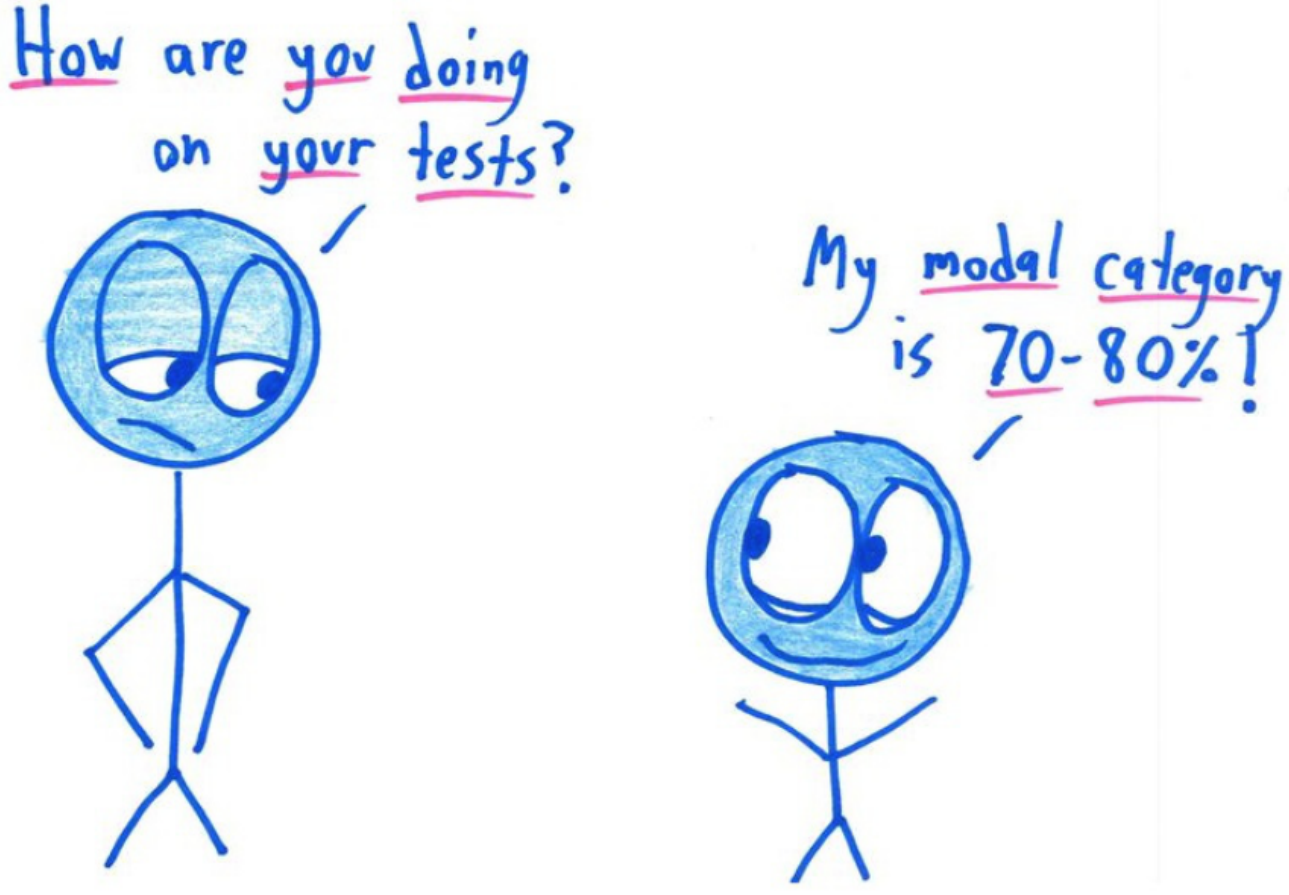
Median


So, why should I invest with you?
Well, not to brag, but my fund has a median gain of 8% per year!

median middle of a shorted list of numbers


+10%
-10%

Mode

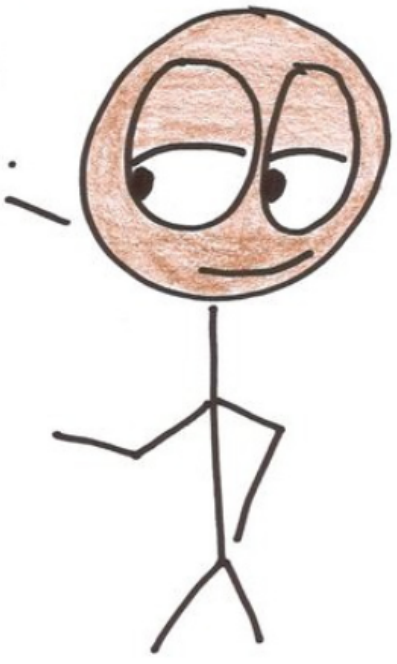


mode of a set of data values is the value that appears most often.

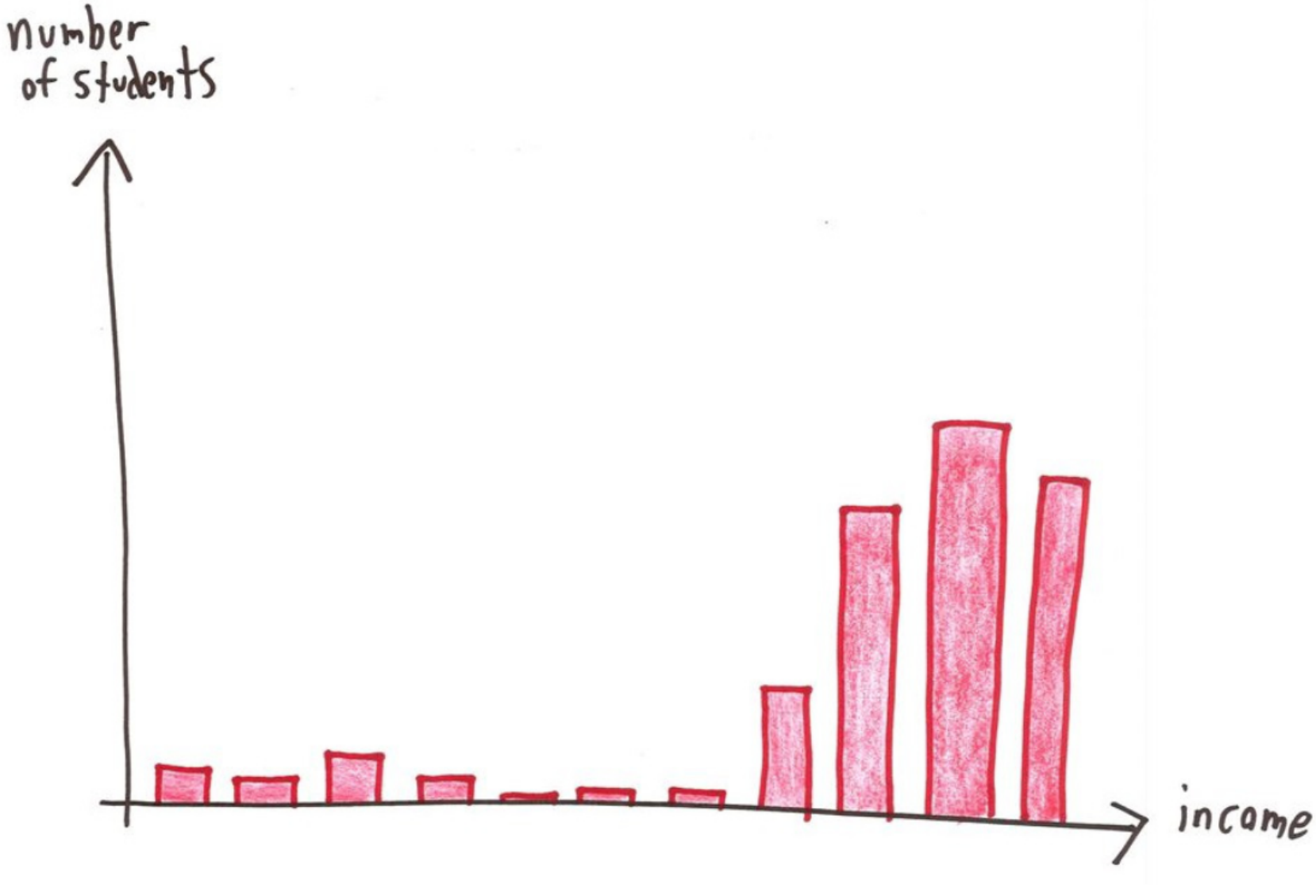
Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

Range

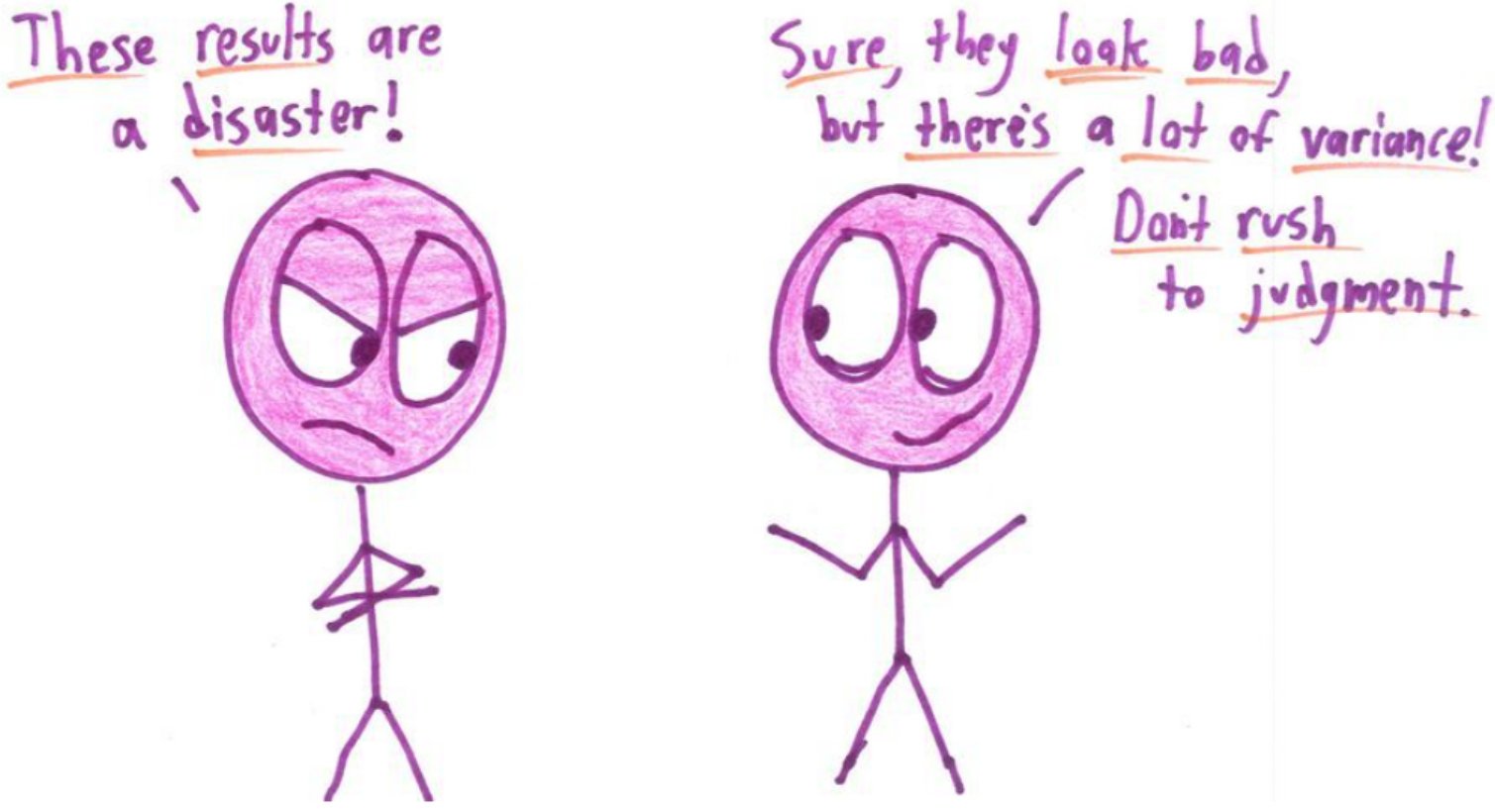
Our students come from a wide range of socioeconomic backgrounds...



range of a set of data values is the difference between the largest and smallest values


number
of students
income

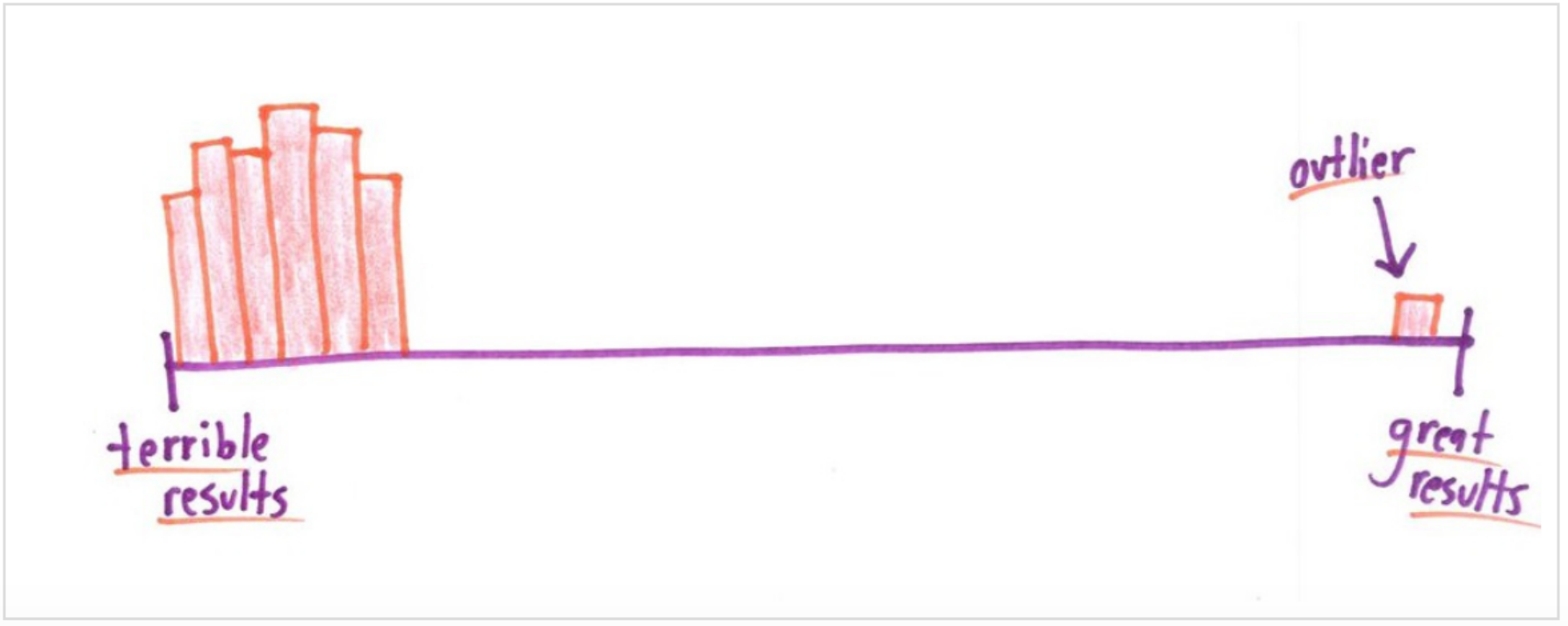
Variance

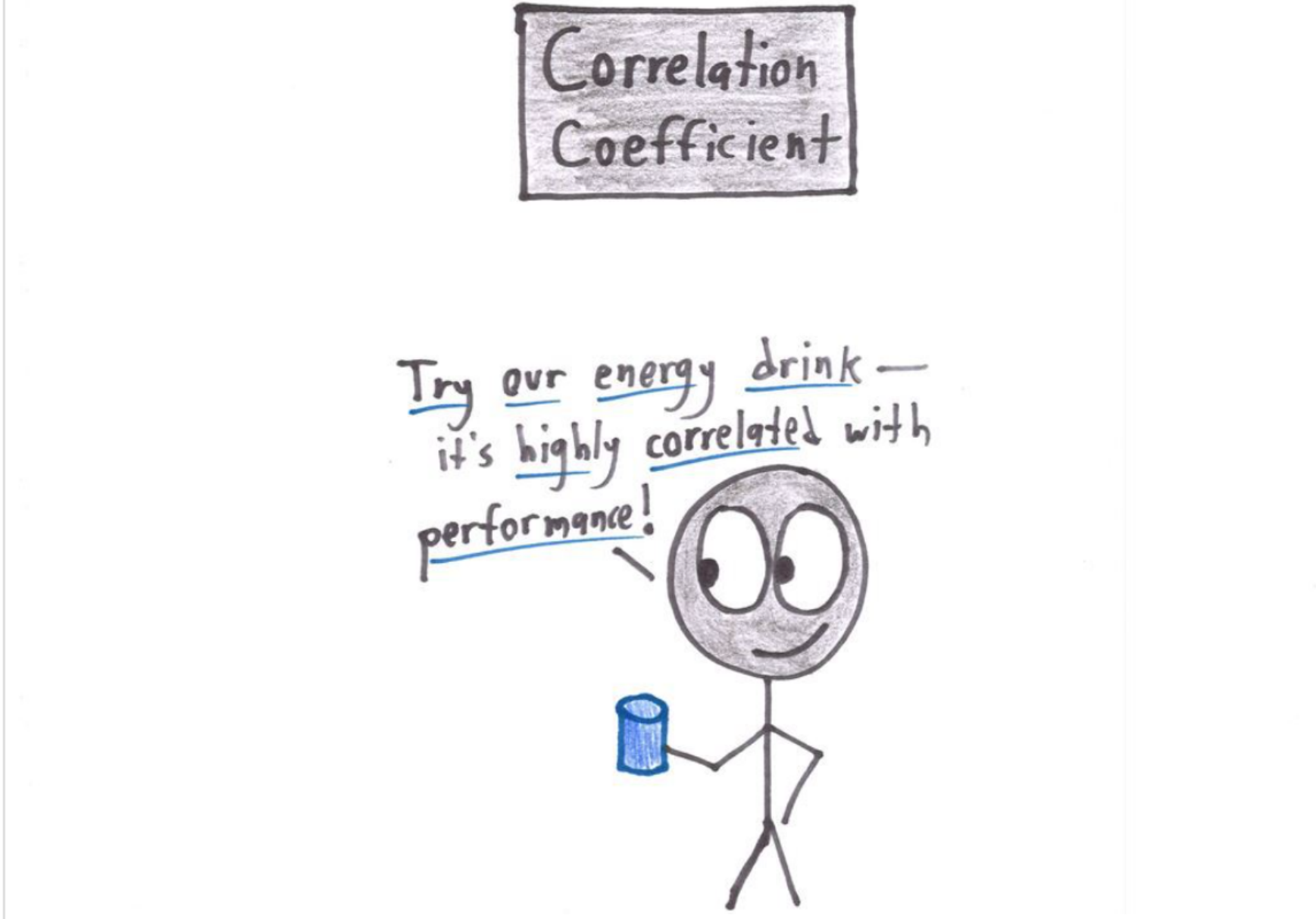

These results are a disaster!
Sure, they look bad, but there's a lot of variance!
Don't rush to judgment.

variance (σ^2) measures how far each number in the set is from the mean and therefore from every other number.

$$\sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$$

standar deviation $= \sigma$


outlier
terrible results
great results

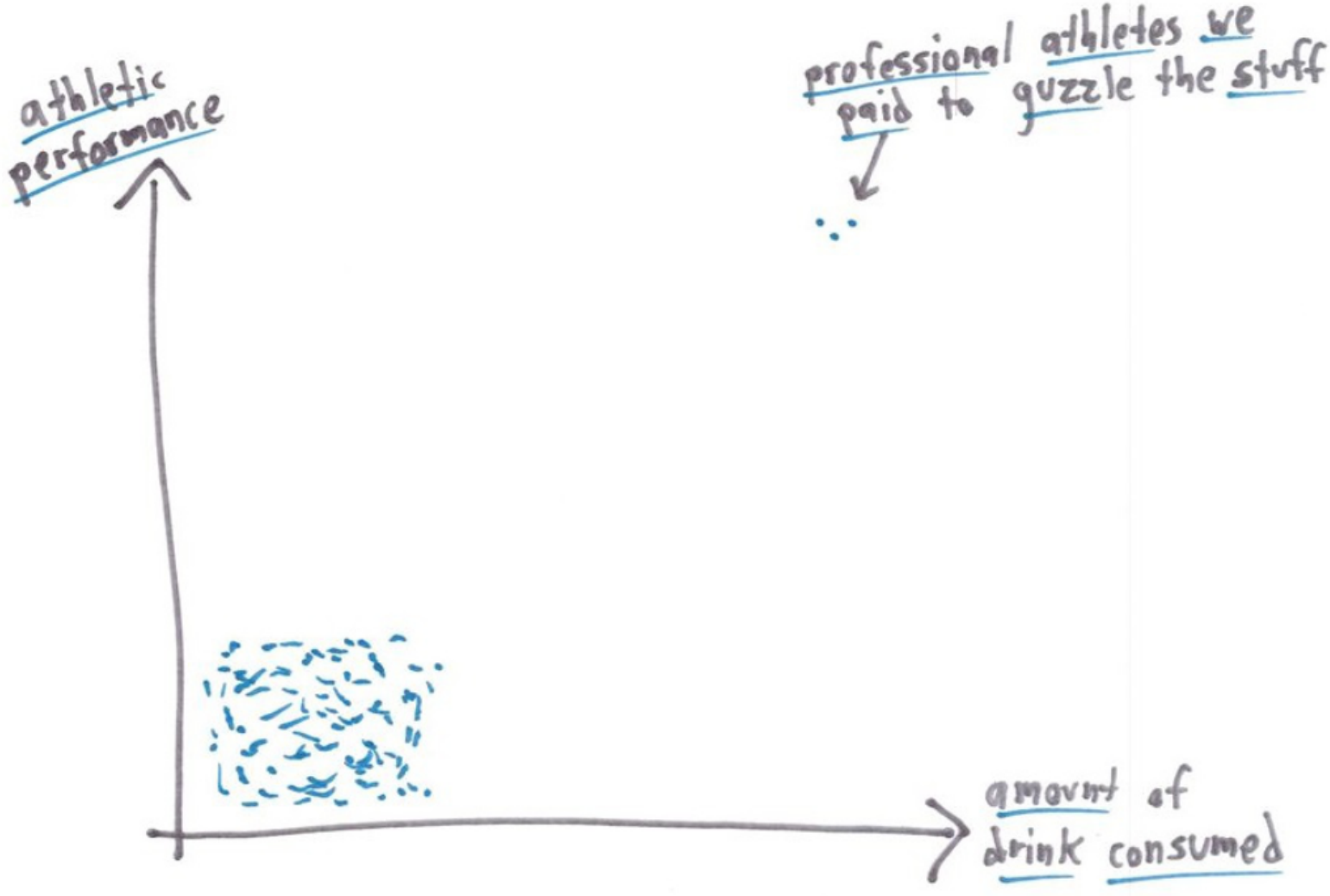

Correlation Coefficient
Try our energy drink — it's highly correlated with performance!

correlation coefficient

$$corr(X,Y) = \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X \sigma_Y}$$

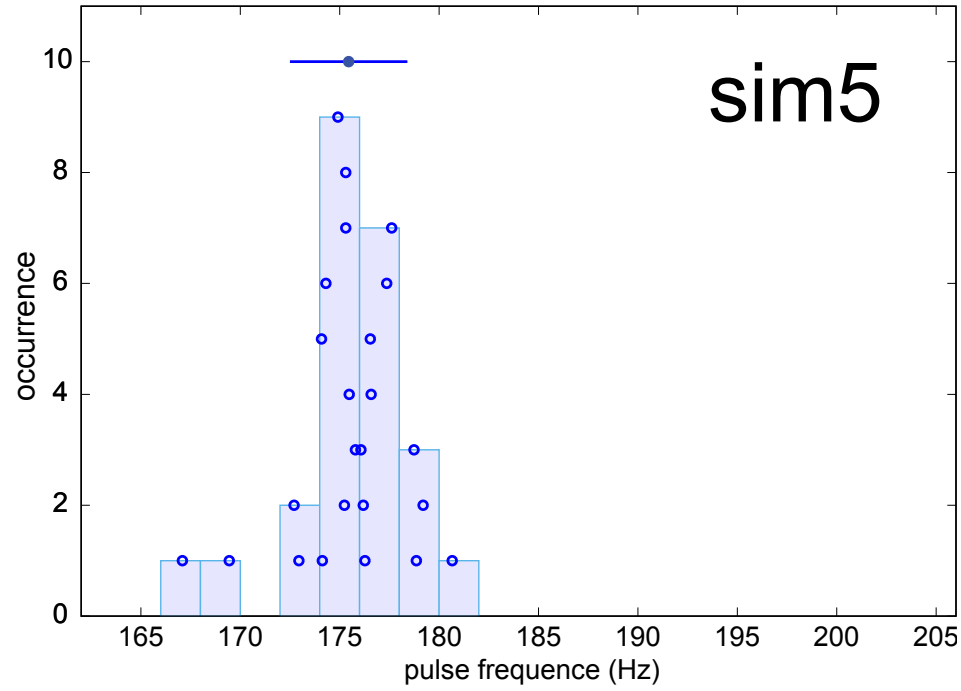
$$\mu = E(X) = \frac{1}{N}\sum_i x_i$$

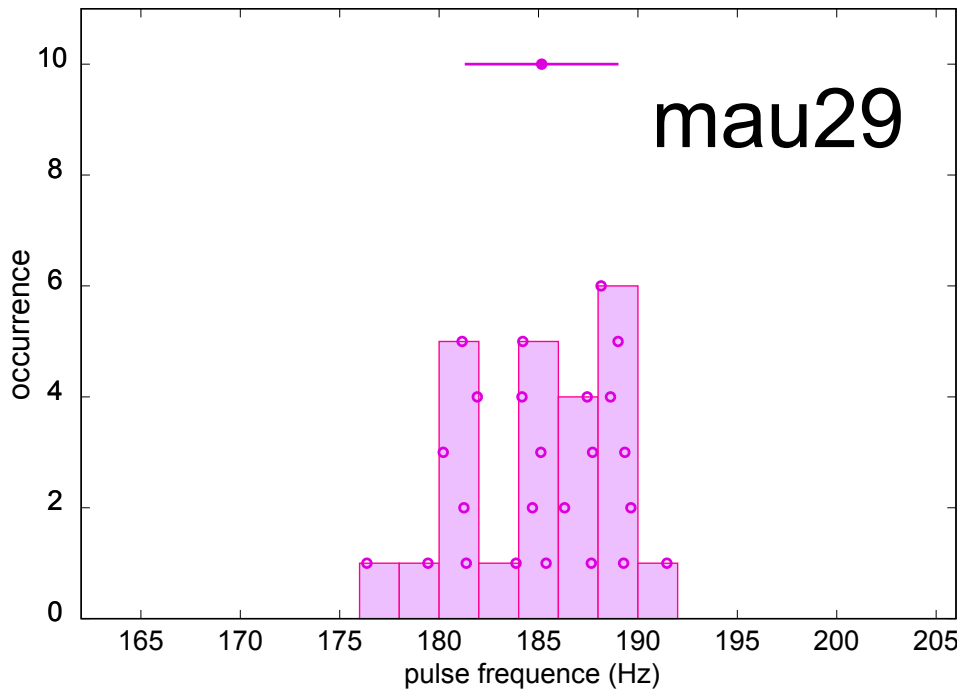
$$\langle f(X) \rangle = E(f(X)) = \frac{1}{N}\sum_i f(x_i)$$


athletic performance
professional athletes we paid to guzzle the stuff
amount of drink consumed

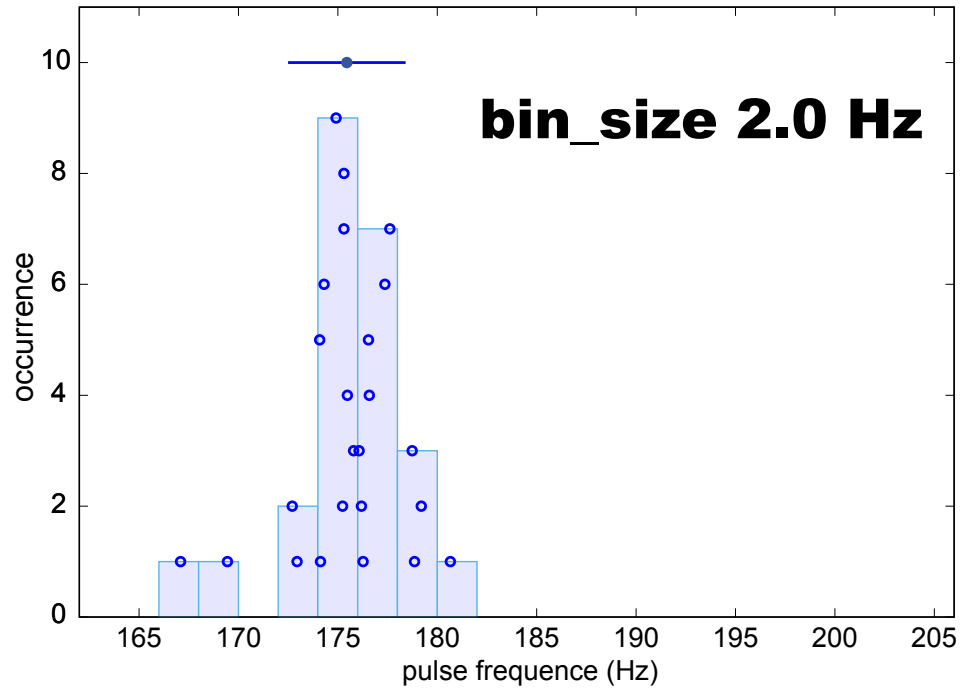
Histogram

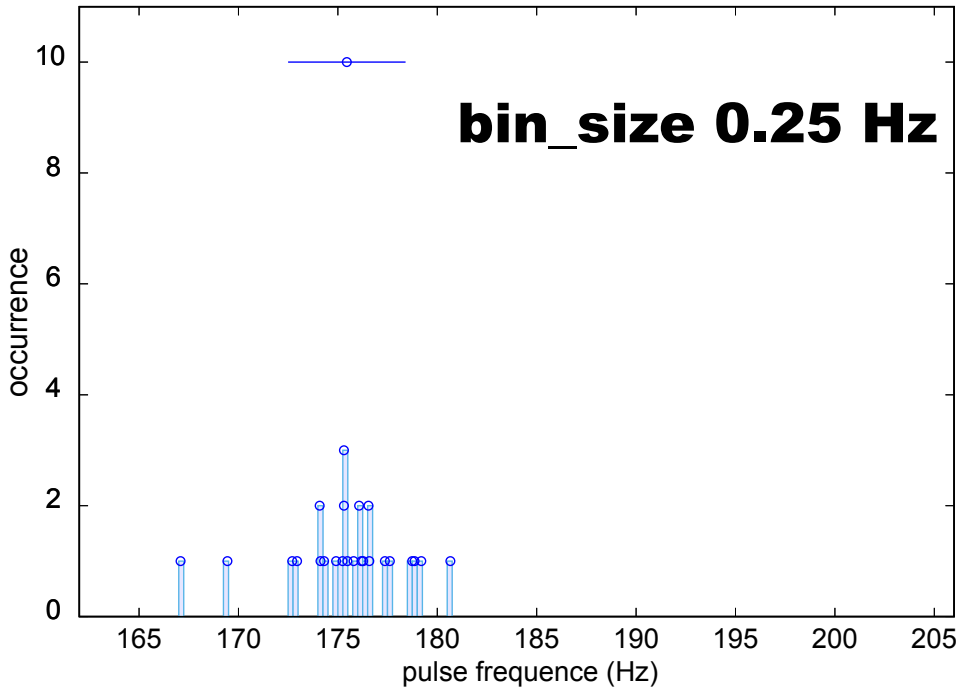
From a "set of data values" to a distribution of the data


sim5
occurrence
pulse frequence (Hz)
0
2
4
6
8
10
165
170
175
180
185
190
195
200
205


mau29
occurrence
pulse frequence (Hz)
0
2
4
6
8
10
165
170
175
180
185
190
195
200
205

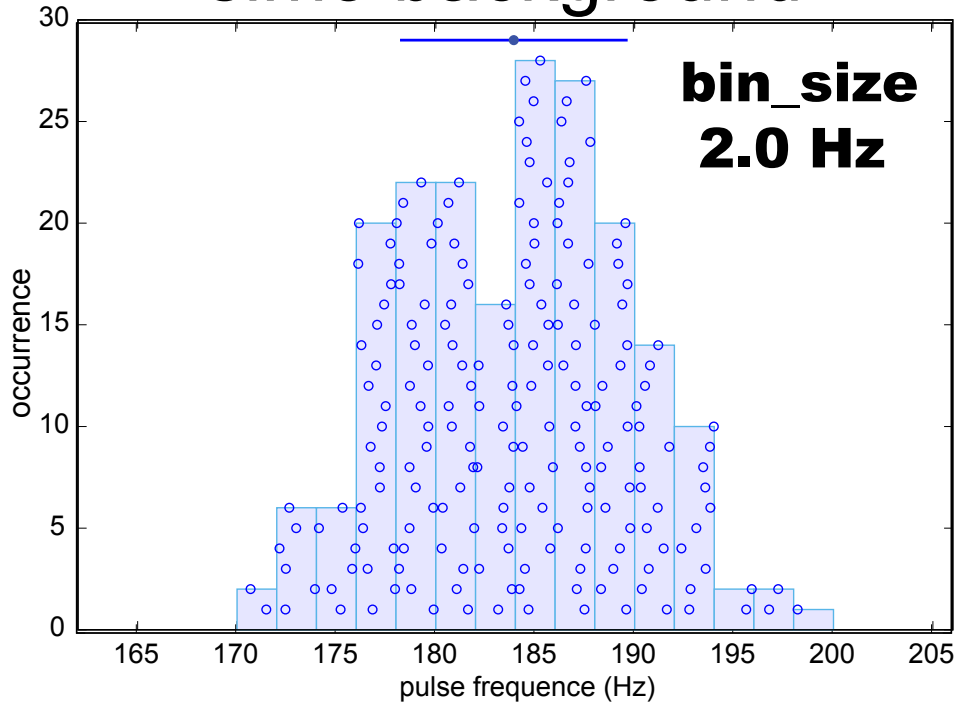
sim5

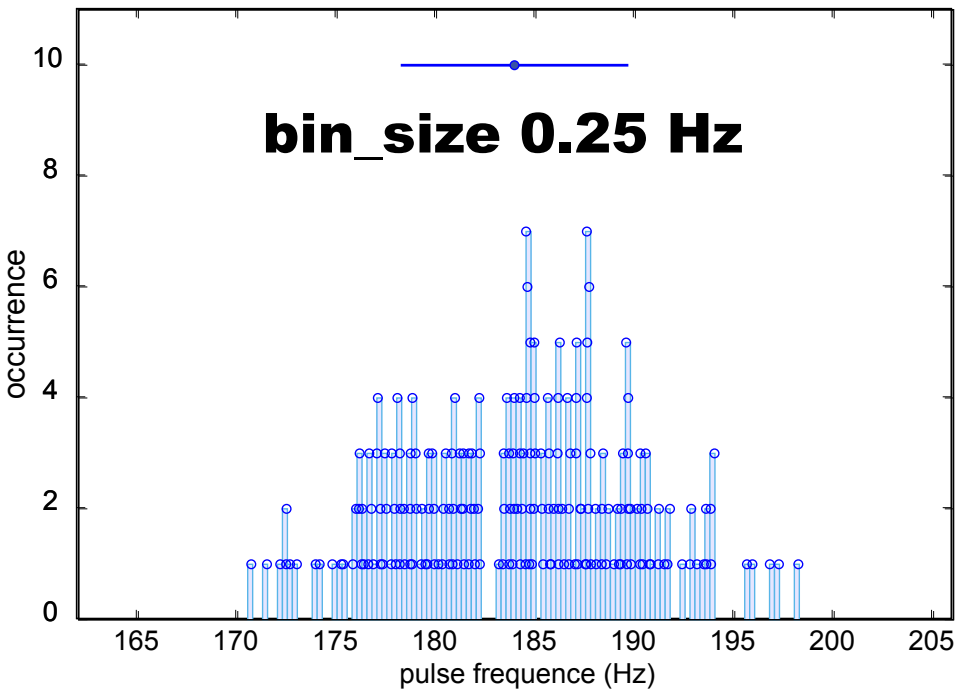




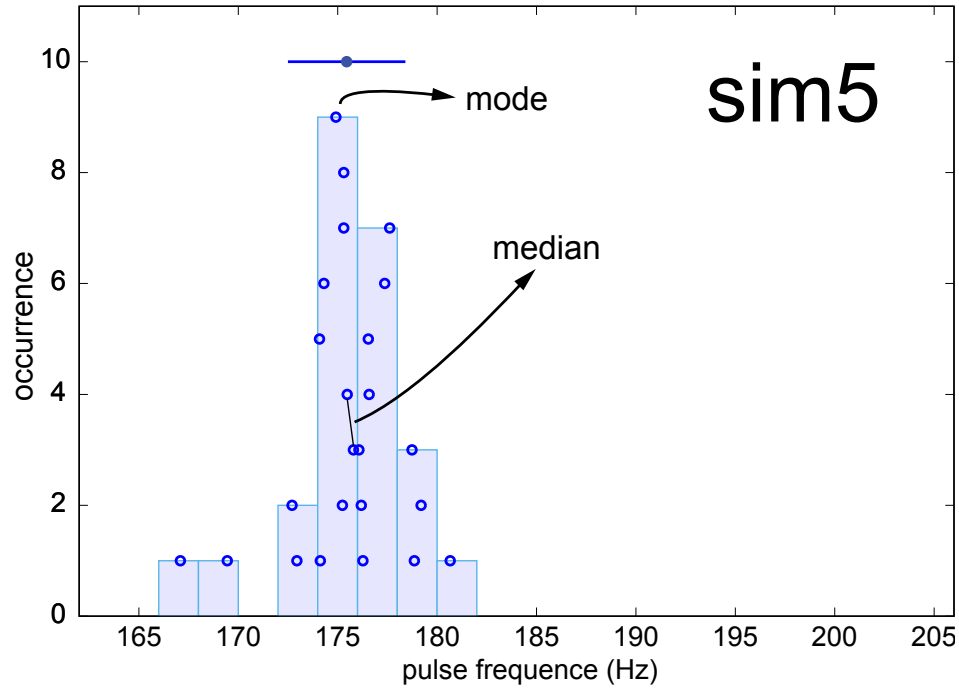
optimal number of bins $\sim \sqrt{N} = 5$

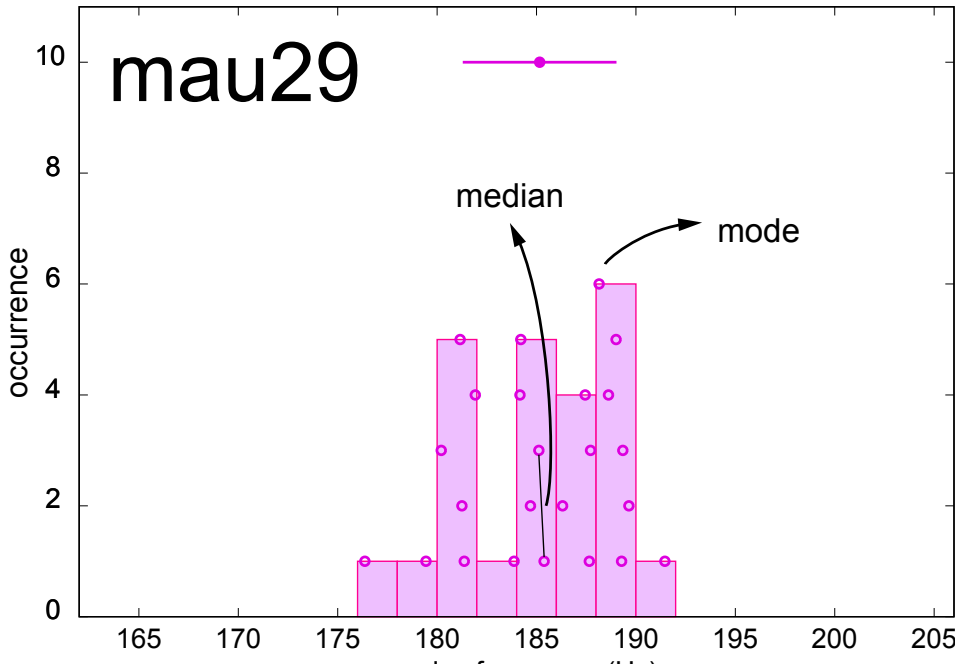
sim5 background





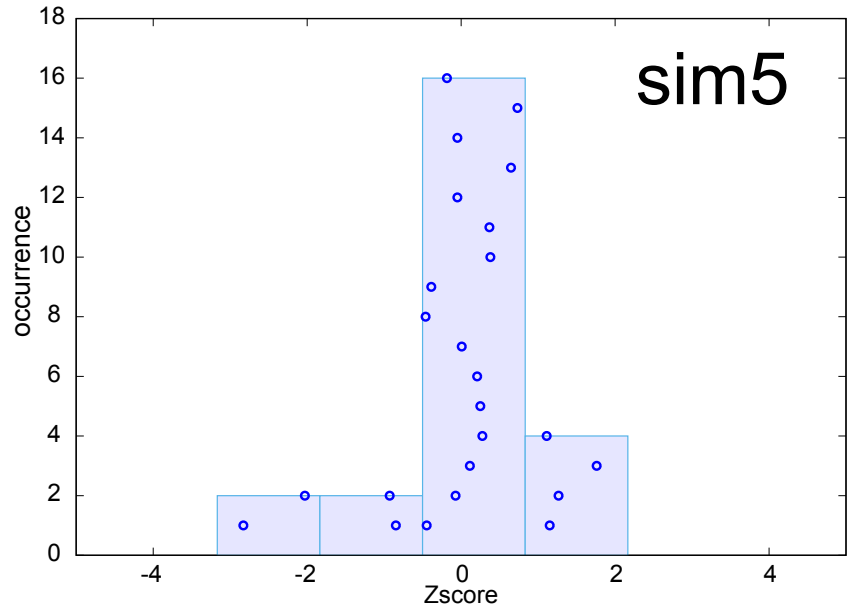
Median and Mode

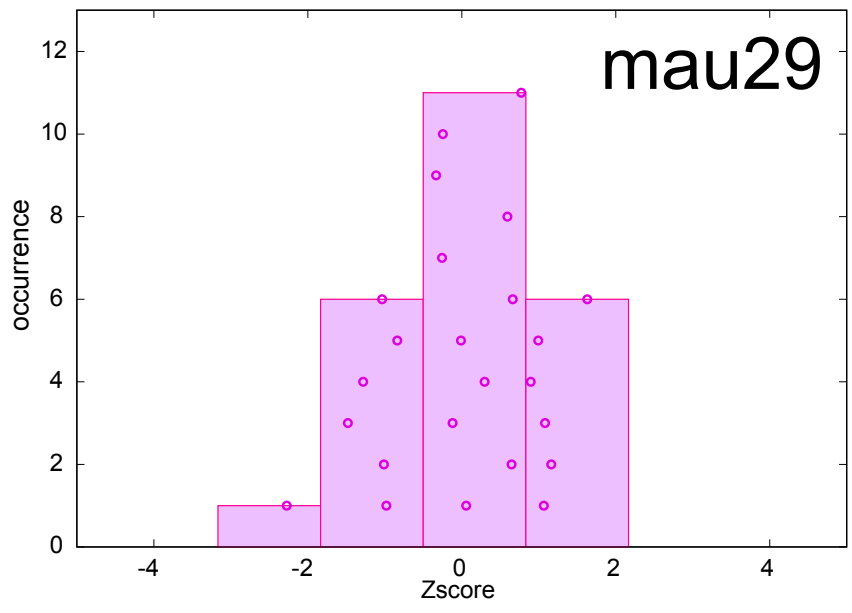




Dispersion (z-scores)

$$z_i = \frac{x_i - \mu}{\sigma} \qquad x_i = \mu + z_i\sigma$$





Random Variable

$$X$$

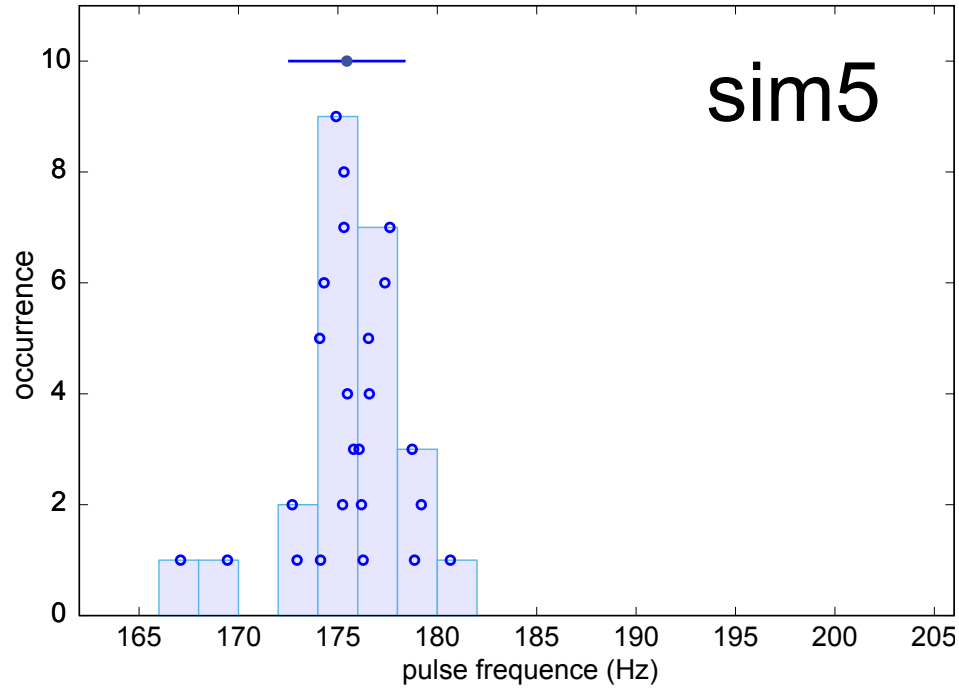
$$x_1, x_2, \ldots$$

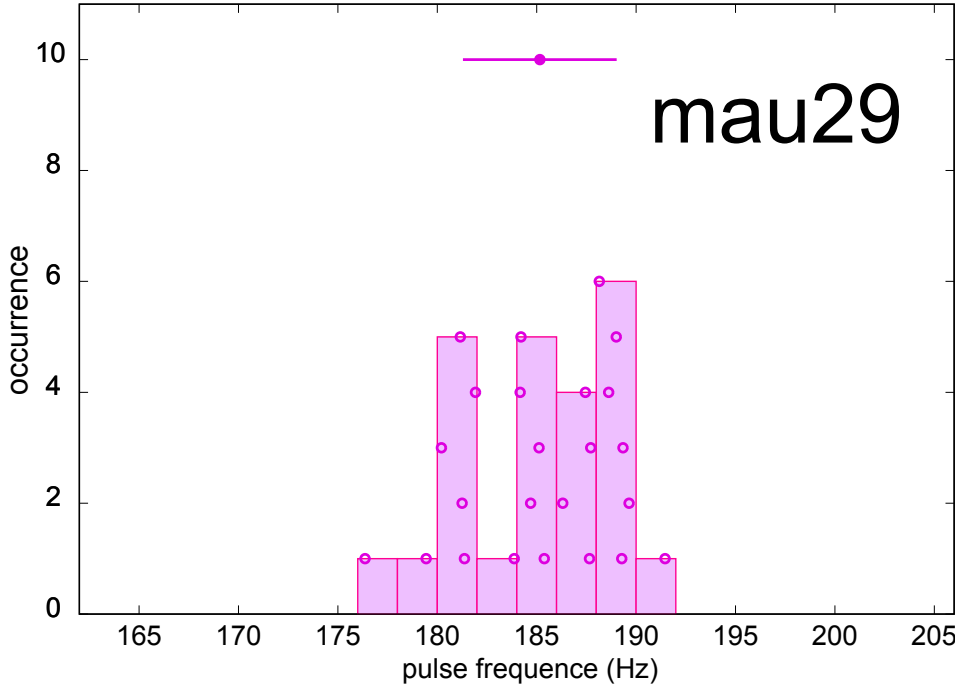
$$P(X = x_i) = p(x_i)$$

A random variable is the numerical result of a random experiment.

Pretty much any experiment in biology.

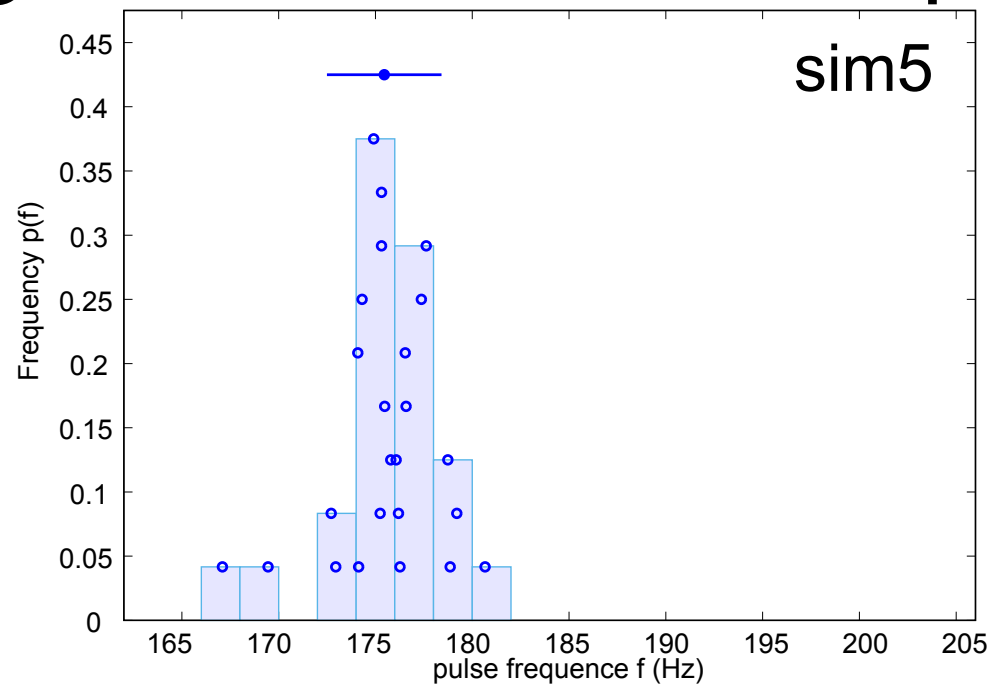
Histogram

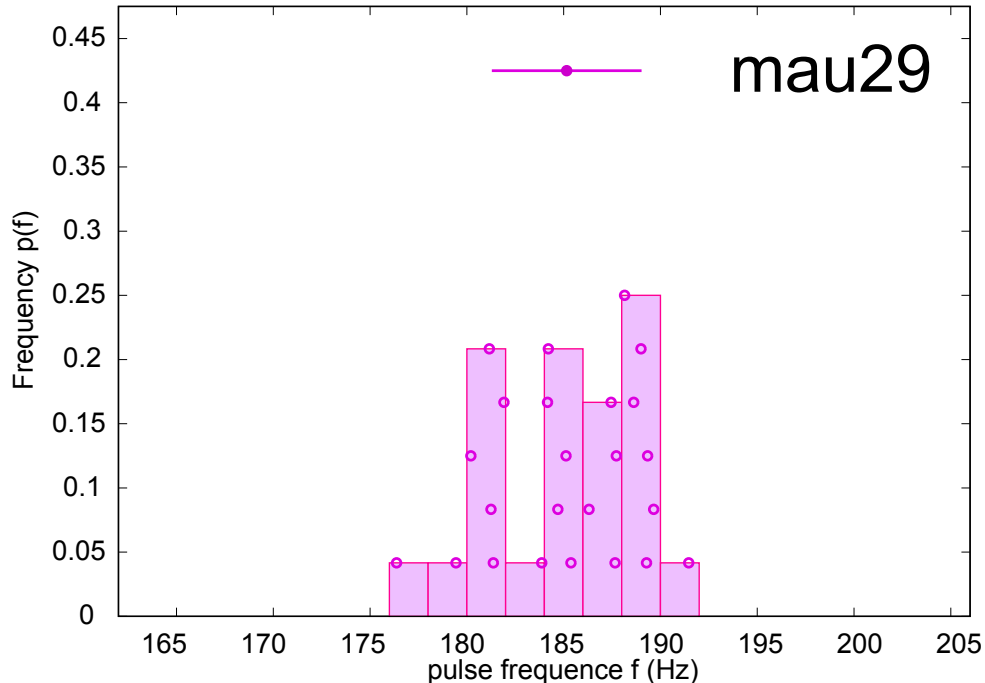

sim5
occurrence
pulse frequence (Hz)
0
2
4
6
8
10
165
170
175
180
185
190
195
200
205


mau29
occurrence
pulse frequence (Hz)
0
2
4
6
8
10
165
170
175
180
185
190
195
200
205

If we repeat the experiment a large-enough number of times we expect that the histogram of observed frequencies will represent the the actual probability distribution behind.

Histogram of observed frequencies





Summary Statistics

We have bined the frequencies and for each bin f, we have calculated the frequencies as the fraction of points N_f/N that have that frequency.

$$\langle f \rangle = \frac{1}{N}\sum_i f_i = \frac{1}{N}\sum_f N_f f = \sum_f \frac{N_f}{N} f = \sum_f p(f) f$$

The **mean (or average)** of a random variable X is

$$\langle x \rangle = E(X) = \mu = \sum_x x\, p(x).$$

Variance, Standard deviation

The **variance** is

$$\sigma^2 = Var(X) = E\left((X-\mu)^2\right) = \sum_x (x-\mu)^2\, p(x).$$

The **standard deviation** is

$$\sigma = \sqrt{\sigma^2}.$$

For any two random variables X and Y,

$$\begin{aligned} E(aX+b) &= aE(X)+b, \\ E(X+Y) &= E(X)+E(Y). \end{aligned}$$

As for their variances, we can say

$$Var(aX+b) = a^2 Var(X).$$

In general,

$$Var(X+Y) \neq Var(X)+Var(Y),$$

unless the two variables are independent, $p(x,y) = p(x)p(y)$.

Probability Density Function (PDF)

$$P(X = x) = p(x)$$

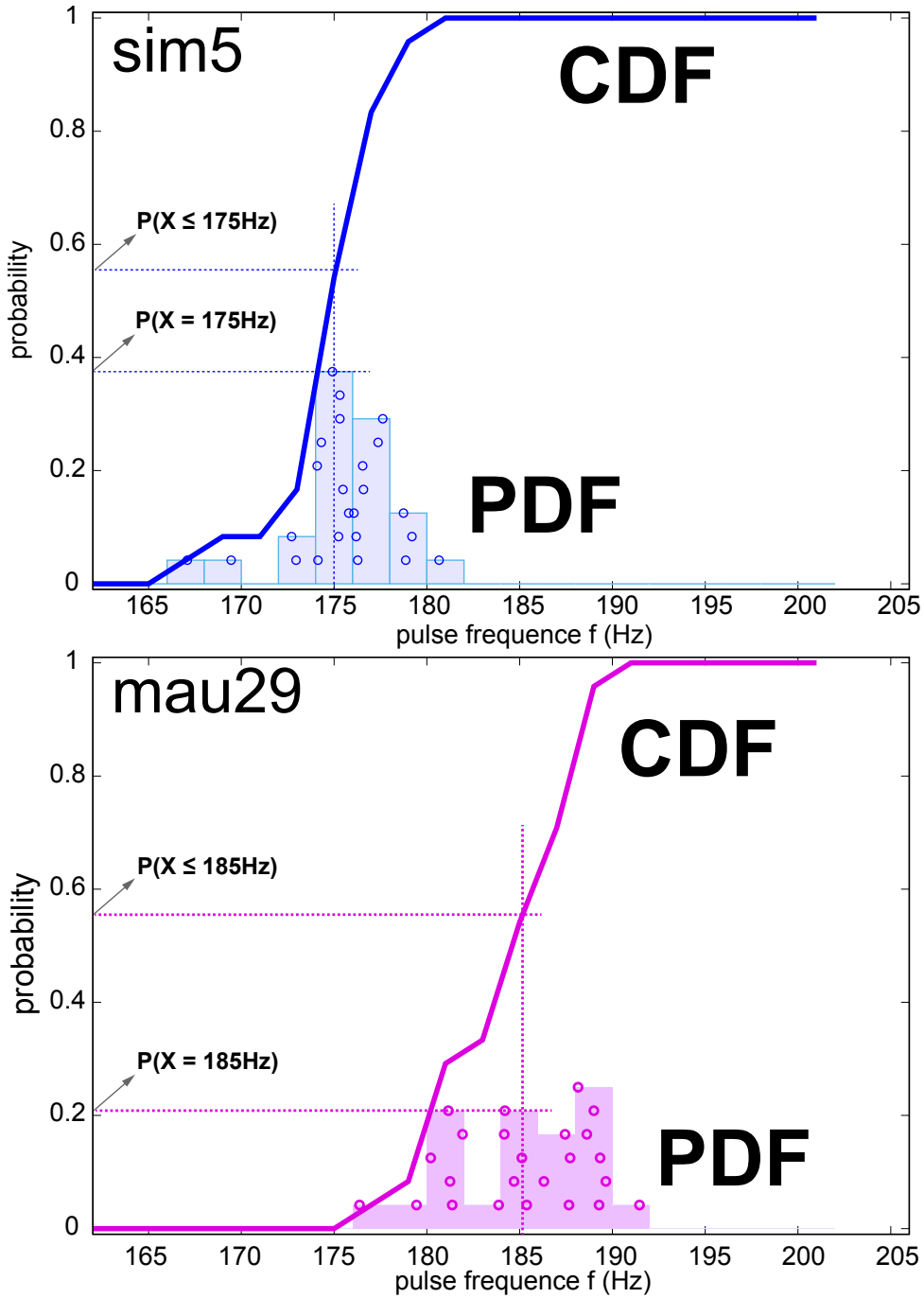
Cumulative Distribution Function (CDF)

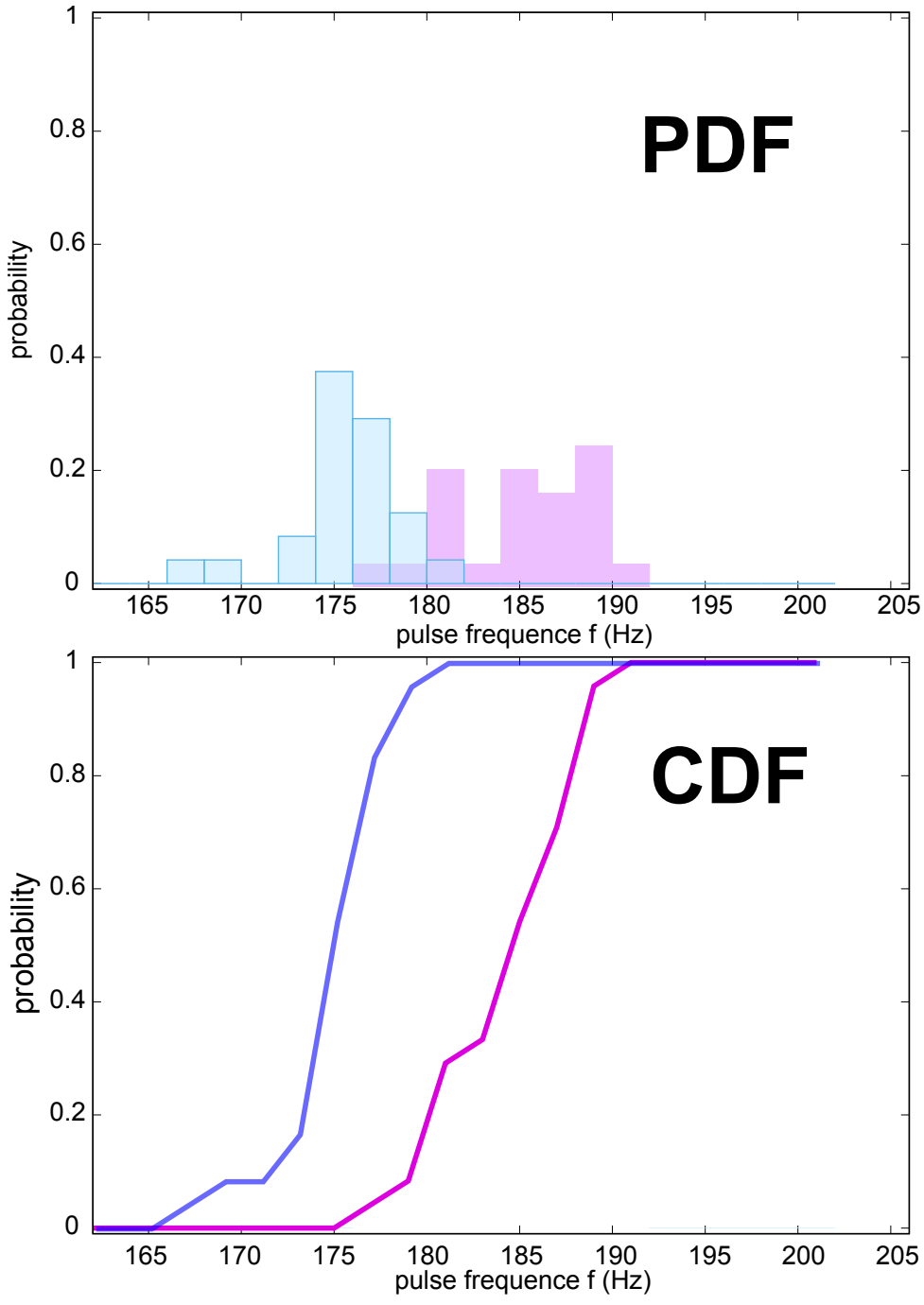
$$P(X \leq x) = P(x)$$

Relationships

$$P(x) = \sum_{y \leq x} p(y) \left(= \int_{-\infty}^{x} p(y) \right)$$

$$\frac{dP(x)}{dx} = p(x)$$

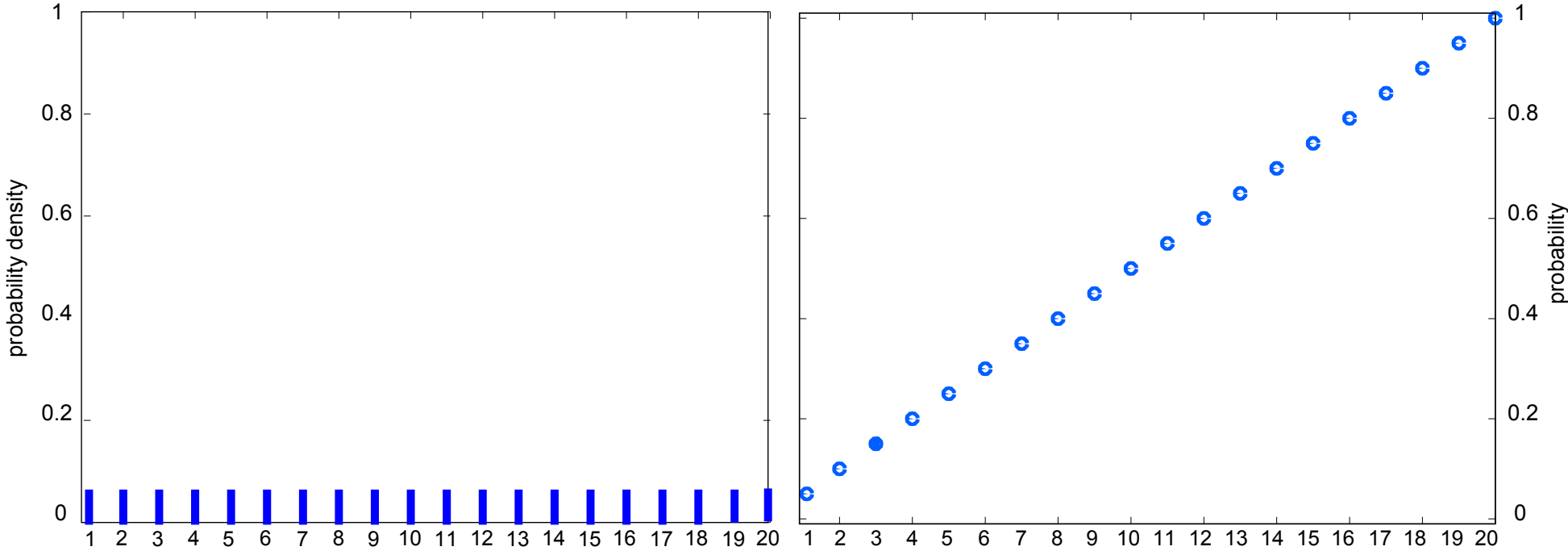

sim5
CDF
PDF
P(X ≤ 175Hz)
P(X = 175Hz)
probability
pulse frequence f (Hz)
mau29
CDF
PDF
P(X ≤ 185Hz)
P(X = 185Hz)
probability
pulse frequence f (Hz)


PDF
probability
1
0.8
0.6
0.4
0.2
0
165
170
175
180
185
190
195
200
205
pulse frequence f (Hz)
CDF
probability
1
0.8
0.6
0.4
0.2
0
165
170
175
180
185
190
195
200
205
pulse frequence f (Hz)

Some Useful Probability Distributions

Uniform Distribution

Distribution	range	PDF	mean	variance
uniform	$i = 1, 2, \ldots n$	$p(i) = \frac{1}{n}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$



All events are equally likely.

For instance, the spacial distribution of penguins.

We use uniform distributions as the maximum entropy distribution (less restricted) when we do not have any information about the distribution.

Bernouilli Distribution

Distribution	range	PDF	mean	variance
Bernoulli	$i = 0, 1$	$p(0) = (1-p)\ p(1) = p$	p	$p(1-p)$

We use the Bernoulli distribution when

- The result of each try is success or failure.
- Each try has the same probability of success, p; and failure, 1-p.
- Each try is independent from each other.

Examples:

1. Having a boy or a girl.
2. The success of a random clinical trial.
3. Carrying a mutation.

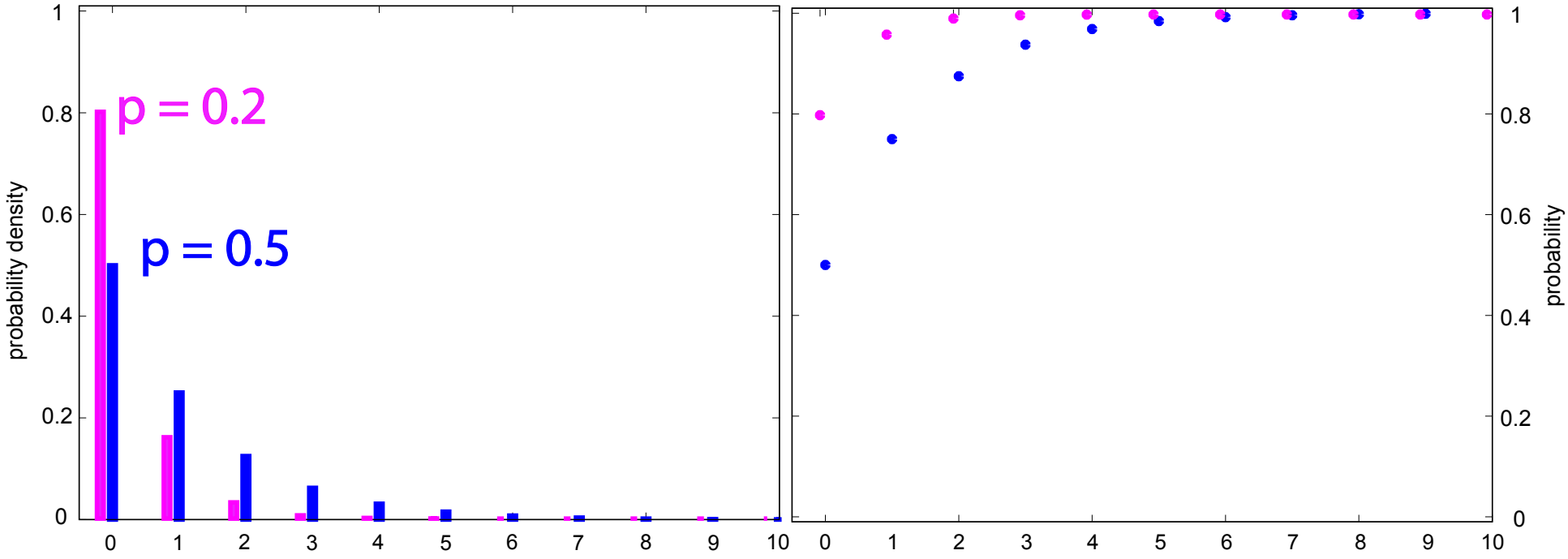
Geometric Distribution

Distribution	range	PDF	CDF	mean	varian
Geometric	$n = 1, 2, \ldots$	$p(n) = (1-p) * p^{n-1}$	$1 - p^n$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$

In a clinical trial, the probability of failure is p (and the probability of success is $1 - p$). We use the Geometric distribution to calculate the probability of having one success after n trials

$$P(\text{having 1 success after n trials}) = (1-p) * p^{n-1}.$$

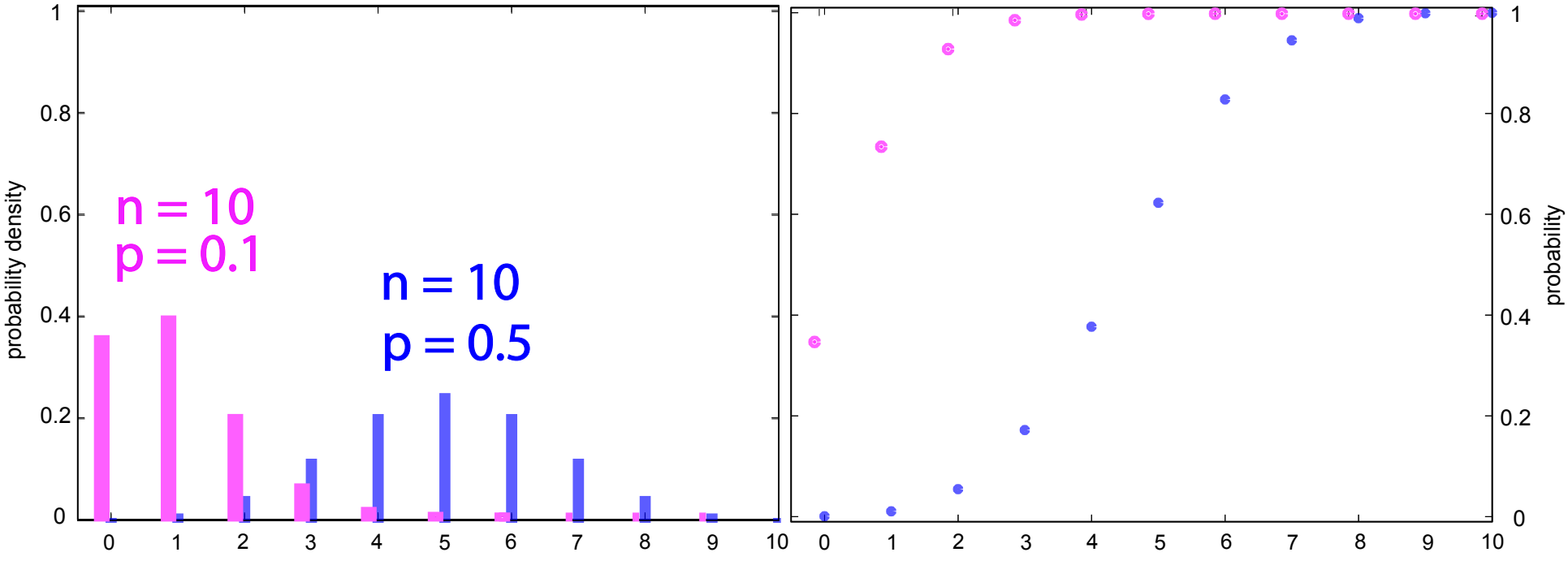
Geometric Distribution



Binomial Distribution

Distribution	range	PDF	mean	variance
Binomial	$i = 0, 1, 2, \ldots n$	$p(i) = \binom{n}{i} p^i (1-p)^{n-i}$	np	$np(1-p)$

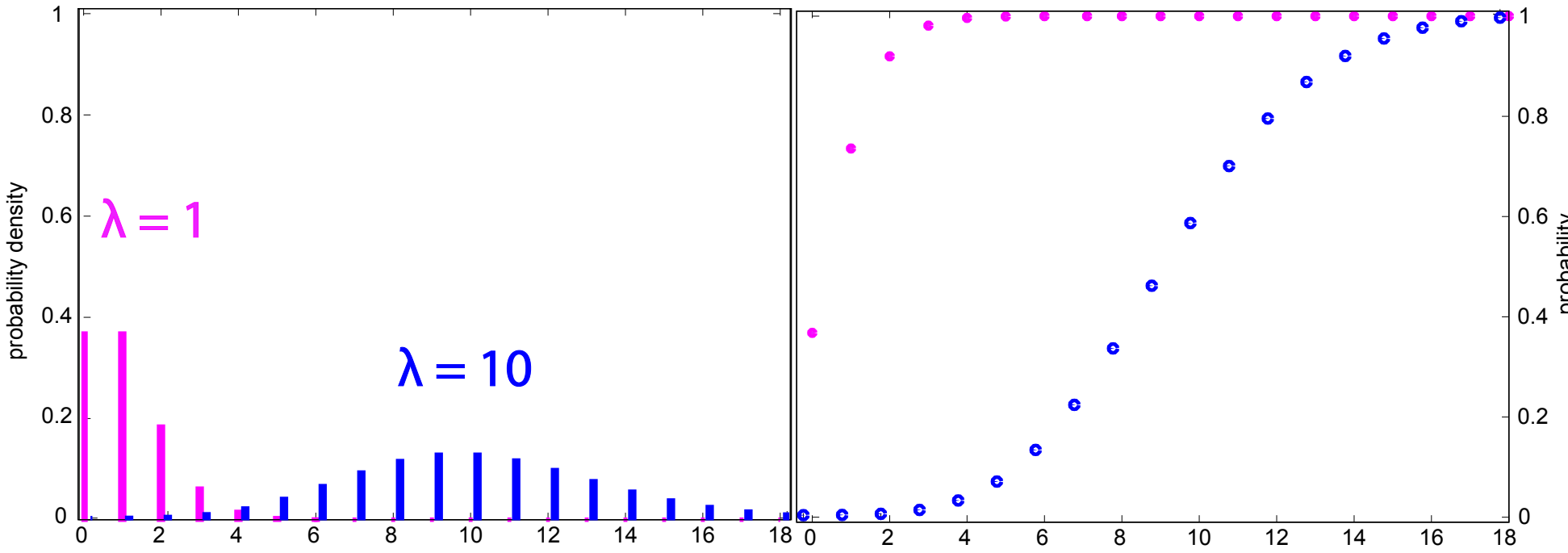
It is the sum of identical and independent Bernoulli random variables.



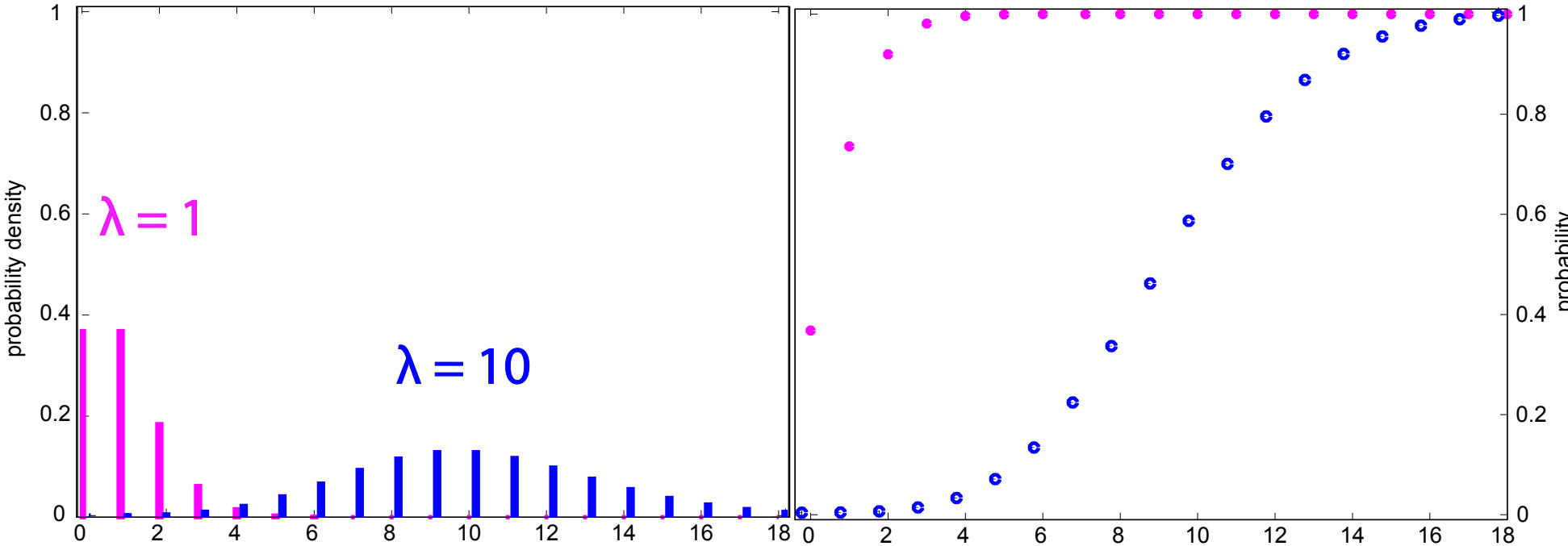
Poisson Distribution

Distribution	range	PDF	mean	variance
Poisson	$n = 0, 1, \ldots$	$p(n) = \lambda^n \frac{e^{-\lambda}}{n!}, \ (\lambda > 0)$	λ	λ

The Poisson distribution applies when counting the number of **rare but open ended** events.



Poisson Distribution

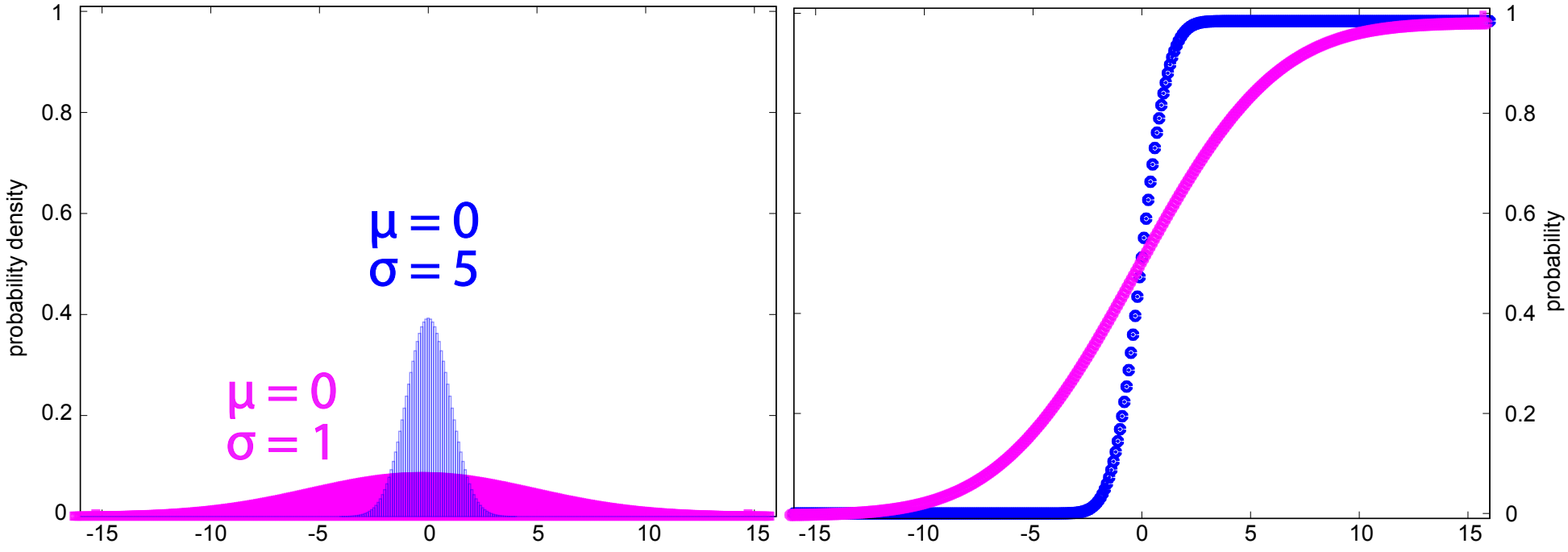


Examples in biology are

1. the number of offsprings of an individual;
2. the infectious rate of viruses;
3. the number of species extinct over time;
4. the number of nucleotide base substitutions in a gene over time;

Normal Distribution

Distribution	range	PDF	mean	variance
$N(\mu, \sigma)$	$-\infty < x < \infty$	$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2



Probability

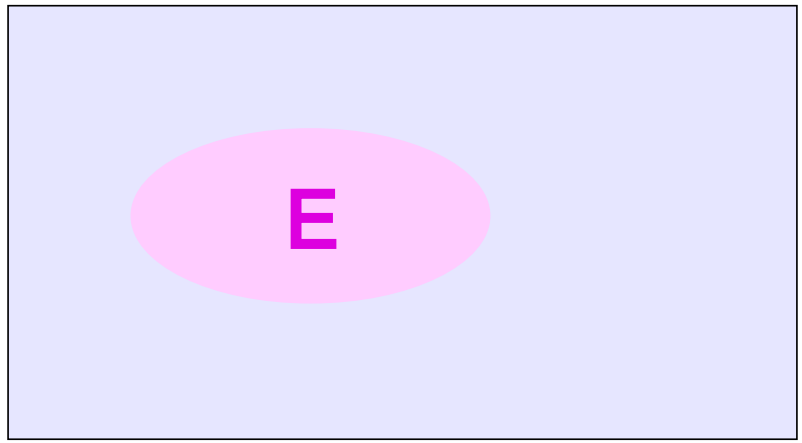
Cromosome segregation

probability of getting one of the two copies of your mom's chromosome 1.

$$0 \leq p \leq 1$$

$0 \leq P(\text{event}) \leq 1$ then $P(\text{complementary-event}) = 1 - P(\text{event})$

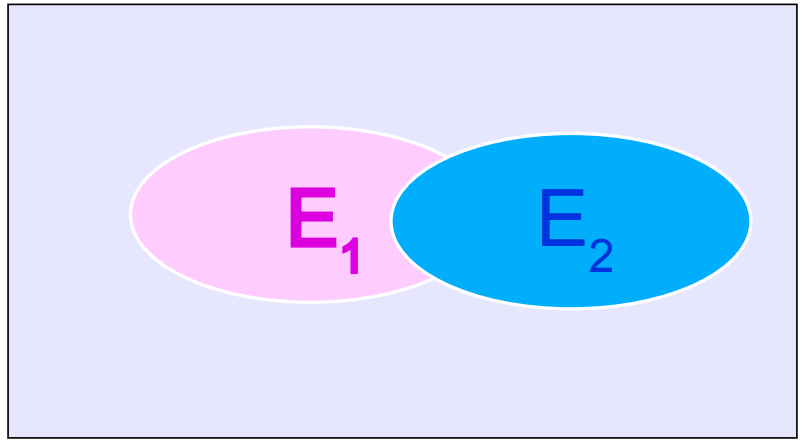
the sum rule

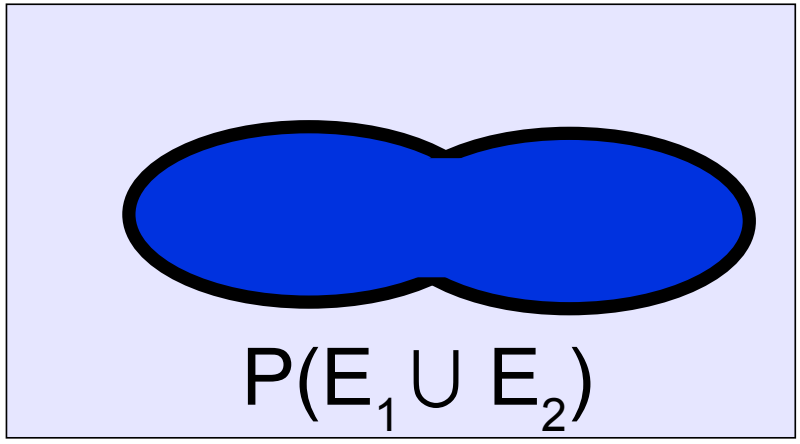


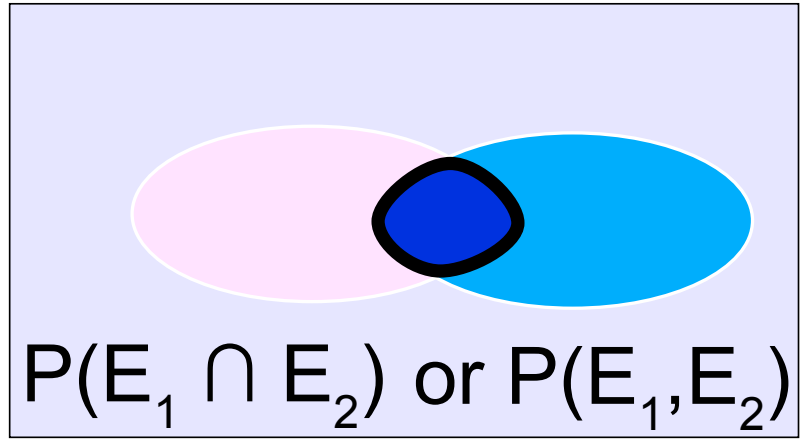
P(E) + P(not_E) = 1

$P(E_1 \cup E_2)$ either event happens,

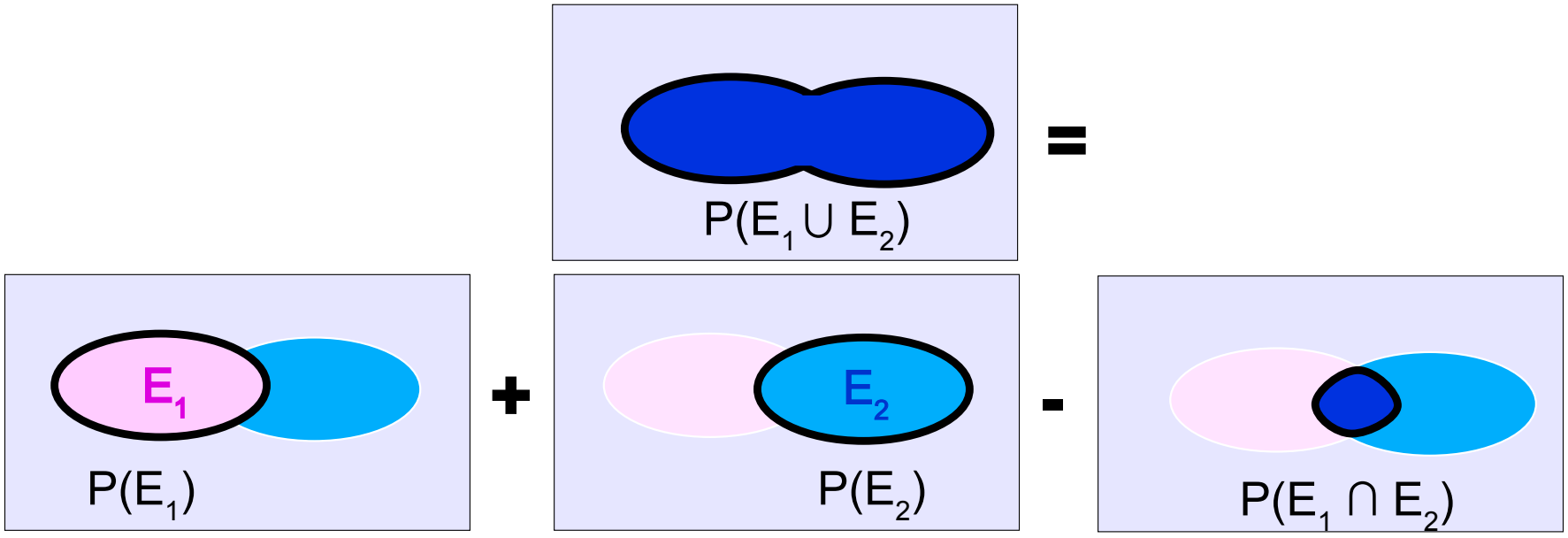
$P(E_1 \cap E_2)$ both events happen.





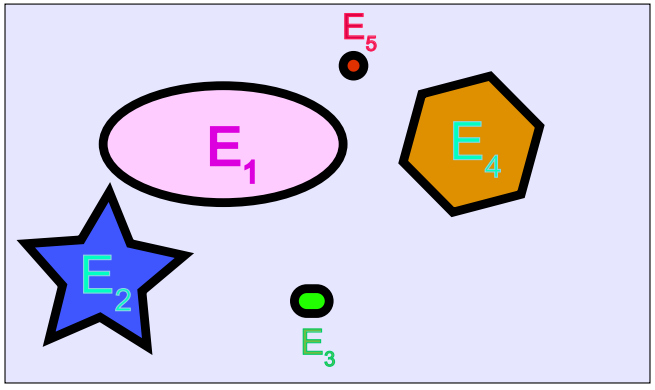


$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$



For N disjoint event, $P(E_i \cap E_j) = 0$, then

$$P(E_1 \cup \ldots \cup E_N) = P(E_1) + \ldots + P(E_N).$$



Your maternal and paternal chromosomes segregate **independently** from each other. Thus, if you mother is `Dd` and your father is `Dd`, the probability that you are `dd` is given by,

$$P(\texttt{dd}) = P(\texttt{d}_m \cap \texttt{d}_f) = P(\texttt{d}_m) * P(\texttt{d}_f) = 0.5 * 0.5 = 0.25.$$

Independent events,

$$P(E_1 \cap E_2) = P(E_1, E_2) = P(E_1) * P(E_2).$$

Conditional Probabilities

$$
\begin{aligned}
P(\mathsf{father\_chrX} \mid female) &= 1.0, \\
P(\mathsf{father\_chrX} \mid male) &= 0.0.
\end{aligned}
$$

By logic, Richard Cox in 1945 arrived to the following relationship between joint and conditional probabilities

$$P(A, B) = P(A \mid B) * P(B).$$

Because $P(A, B) = P(B, A)$ (by definition)

$$P(A, B) = P(B \mid A) * P(A).$$

The product Rule.

$$P(A, B) = P(A \mid B) * P(B) = P(B \mid A) * P(A).$$

- If A and B are independent then

$$P(A \mid B) = P(A),$$

and the product rule results in

$$P(A, B) = P(A) * P(B).$$

- It is obviously true in the cases above:

$$\begin{aligned} P(\text{female\&father\_chrX}) &= \\ P(\text{father\_chrX}|\text{female}) * P(\text{female}) &= \\ 1 * P(\text{female}) = 0.5. & \end{aligned}$$

$$\begin{aligned} P(\text{male\&father-chrX}) &= \\ P(\text{father\_chrX}|\text{male}) * P(\text{male}) &= \\ 0 * P(\text{male}) = 0. & \end{aligned}$$

If you rewrite the product rule, you get **Bayes' theorem**

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)}.$$

Very important point we will discuss more later. In general,

$$\boxed{P(B \mid A) \neq P(A \mid B).}$$

Marginal Probabilities

Marginalization allows us to calculate the probability of one event (A), when that event has different outcomes depending on some other even event (B) it always occurs with. If event B can have different disjoint outcomes $(B_1, \ldots, B_N)$ then,

$$P(A) = \sum_i P(A, B_i),$$

which using the sum rule can be rewritten as,

$$\boxed{P(A) = \sum_i P(A \mid B_i) P(B_i).}$$

Marginalization is a powerful device to deal with *nuisance* data.

Eye color is not a Mendelian trait, but depend on **many** ≥ 15 different genomic regions.

Assume, eye color depends on **two regions** in two different chromosomes:

1. **region A**, where allele aa alone provides blue color in 80% of the cases,
2. **region B** where bb alone provides blue color in 90% of the cases,
3. the presence of aa and bb provides a 100% chance of blue color.

Calculate the probability of my dog having blue eyes

$$P(\text{blue})$$

The nuisance variable = multi allele combination the color comes from.

- Assume that the prevalence of those alleles in a population is $P(\text{aa}) = 0.1$, and $P(\text{bb}) = 0.01$, and that they are independent,
- What if the two alleles are completely linked and $P(\text{aa} \cap \text{bb}) = 0.1$?

Marginalization also works for conditional probabilities,

$$\boxed{P(A \mid I) = \sum_i P(A, B_i \mid I) = \sum_i P(A \mid B_i \ I) * P(B_i \mid I).}$$

The probability of a conjecture given the data

Bayesian inference

Probability of that a randomly chosen human is the Pope is 1 in 7 billion

Francis is the Pope

$$P(F \mid H) = \frac{1}{7} 10^{-9}$$

Probability that Pope Francis is human?

$$P(H \mid F) = ?$$

Is the pope the pope? SR Eddy, DJC Makay, Nature 1996.

The probability of the data (D) given the Hypothesis (H)

$$P(D \mid H)$$

Is not the same as the probability of the Hypothesis given the data $P(H \mid D)$

$$P(H \mid F) = \frac{P(F \mid H)P(H)}{P(F \mid H)P(H) + P(F \mid A)P(A)}$$

A DNA fingerprint test has a great sensitivity

$$P(+ \mid \text{Inocent}) = 1 \text{ in a million}$$

In a city of 10 million, the probability being inocent after having tested positive is?