Introduction to Probability

MOL518, Lecture #9

Outline

- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

Outline

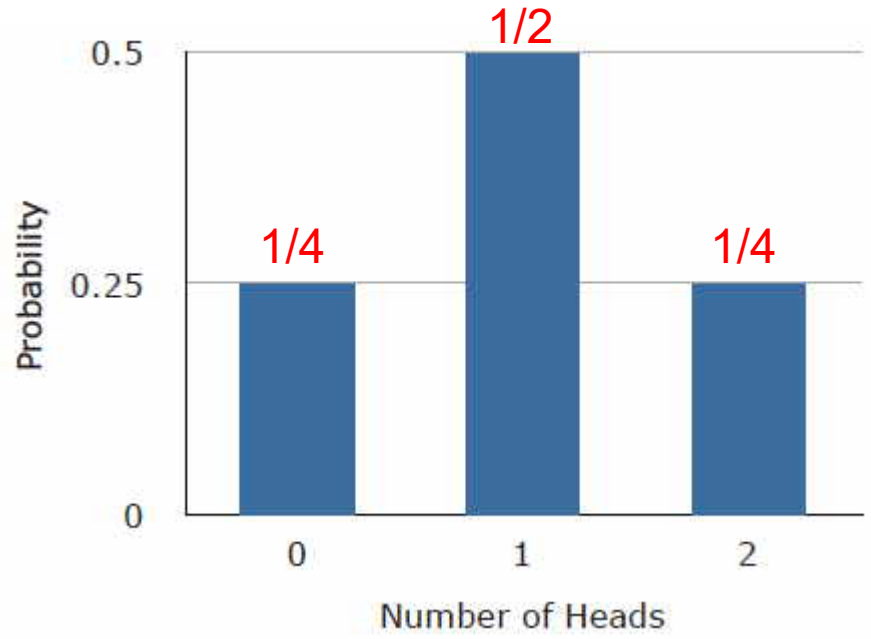
- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

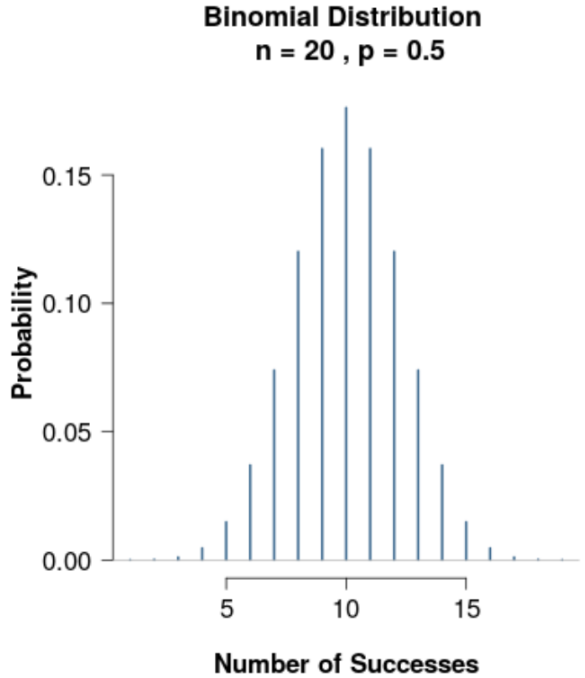
Coin flips are probabilistic in nature

- Assume we have a fair coin that can’t land on its edge
- Probability of heads is 50%
- P(heads) = 0.5
- P(tails) = 0.5

Repeated coin flips give us a binomial probability distribution

Outcome	First Flip	Second Flip
1	Heads	Heads
2	Heads	Tails
3	Tails	Heads
4	Tails	Tails





Key facts about the binomial distribution

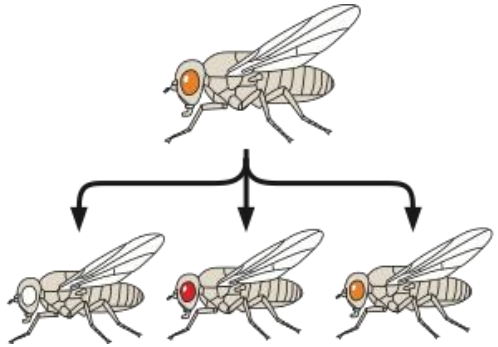
- Generalizing from counting coin flips:

- N is the number of trials
- Probability of success on each trial is π

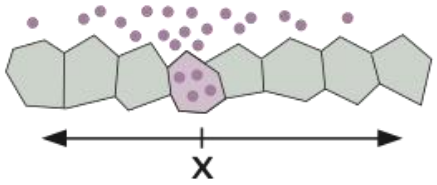
- Mean: $\mu = N\pi$
- Variance: $\sigma^2 = N\pi(1-\pi)$
- Standard deviation, σ, is $\sqrt{N\pi(1-\pi)}$

The four “Great Distributions” in Biology

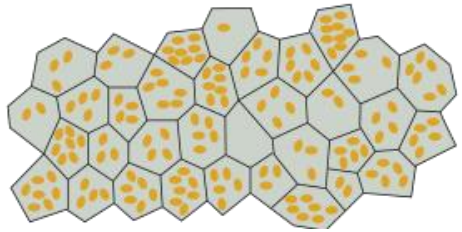
Allele segregation



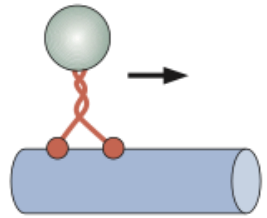
Diffusion



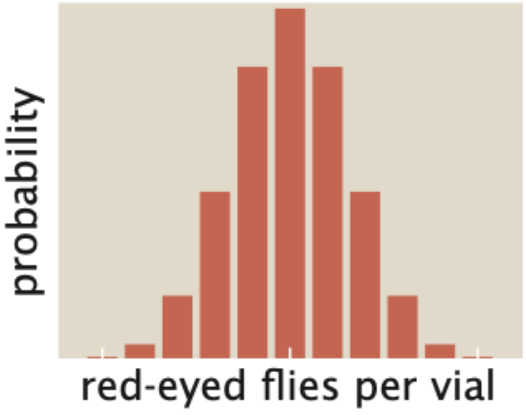
Viral Infection



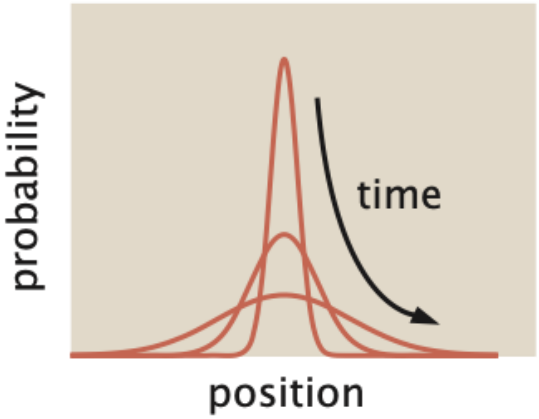
Wait times



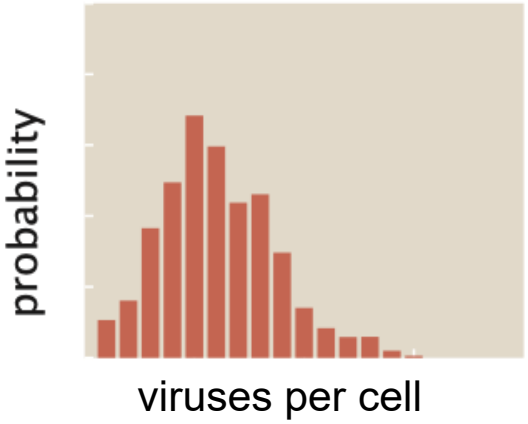
Binomial



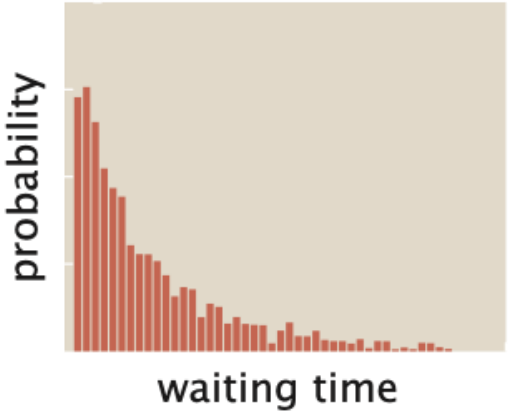
Normal



Poisson

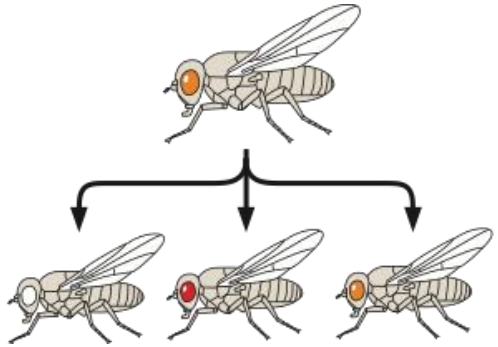


Exponential

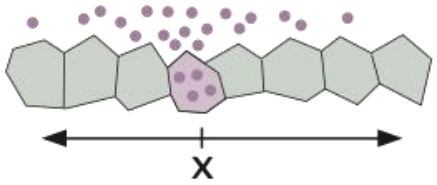


The four “Great Distributions” in Biology

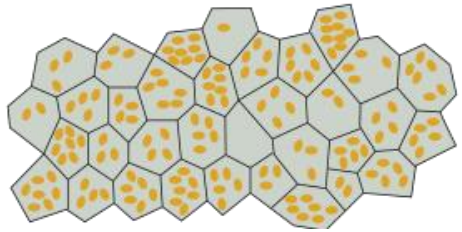
Allele segregation



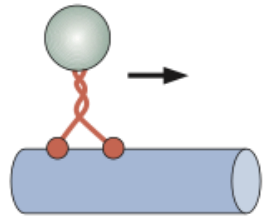
Diffusion



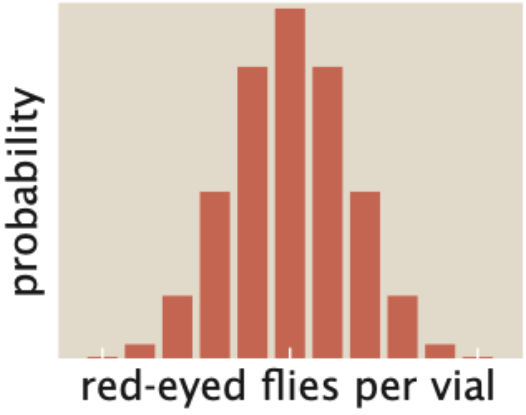
Viral Infection



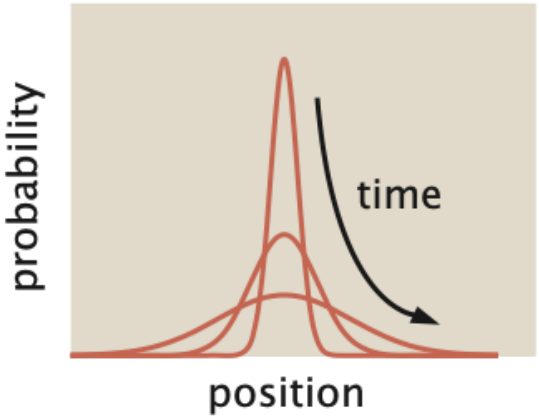
Wait times



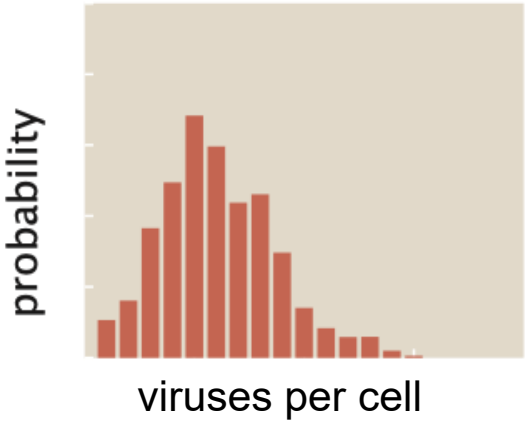
Binomial



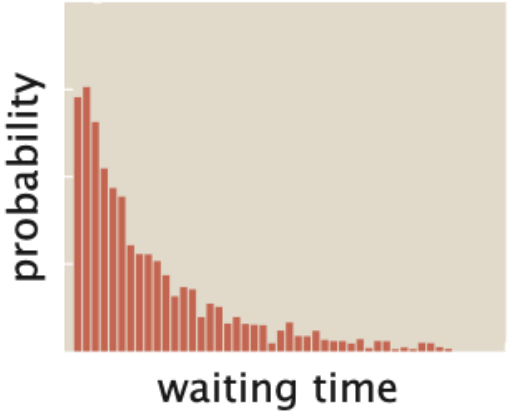
Normal



Poisson



Exponential



Outline

- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

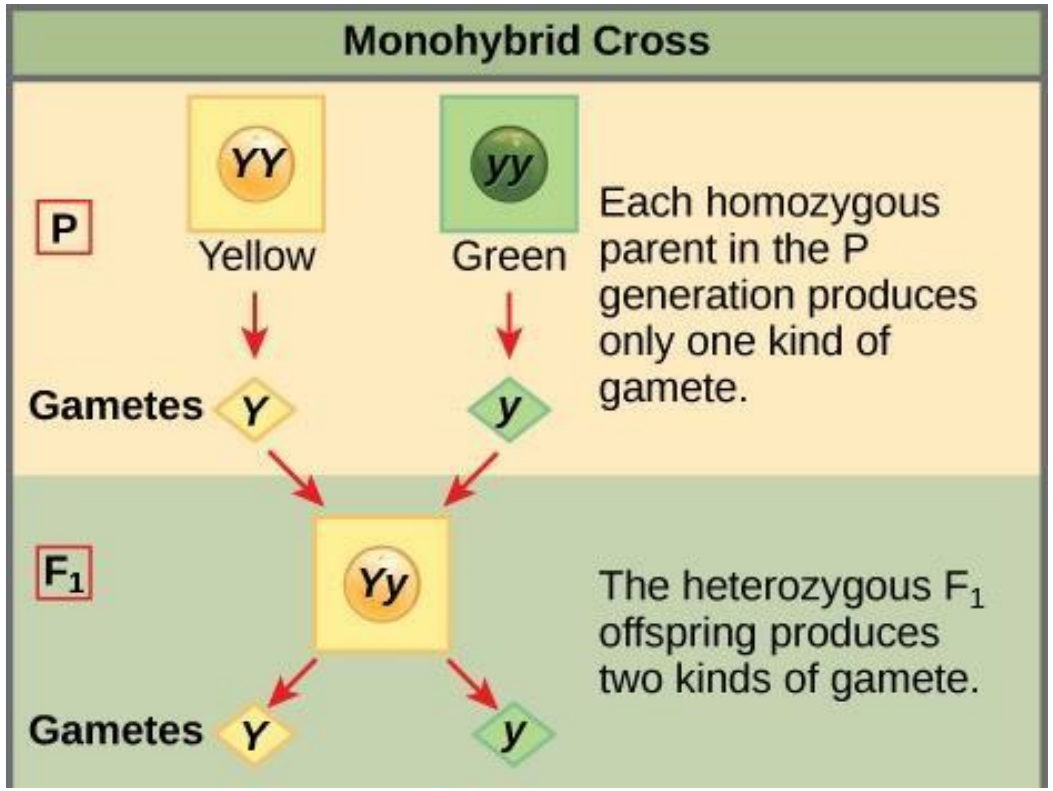
Mendel's observed a 3:1 ratio for the segregation of certain traits

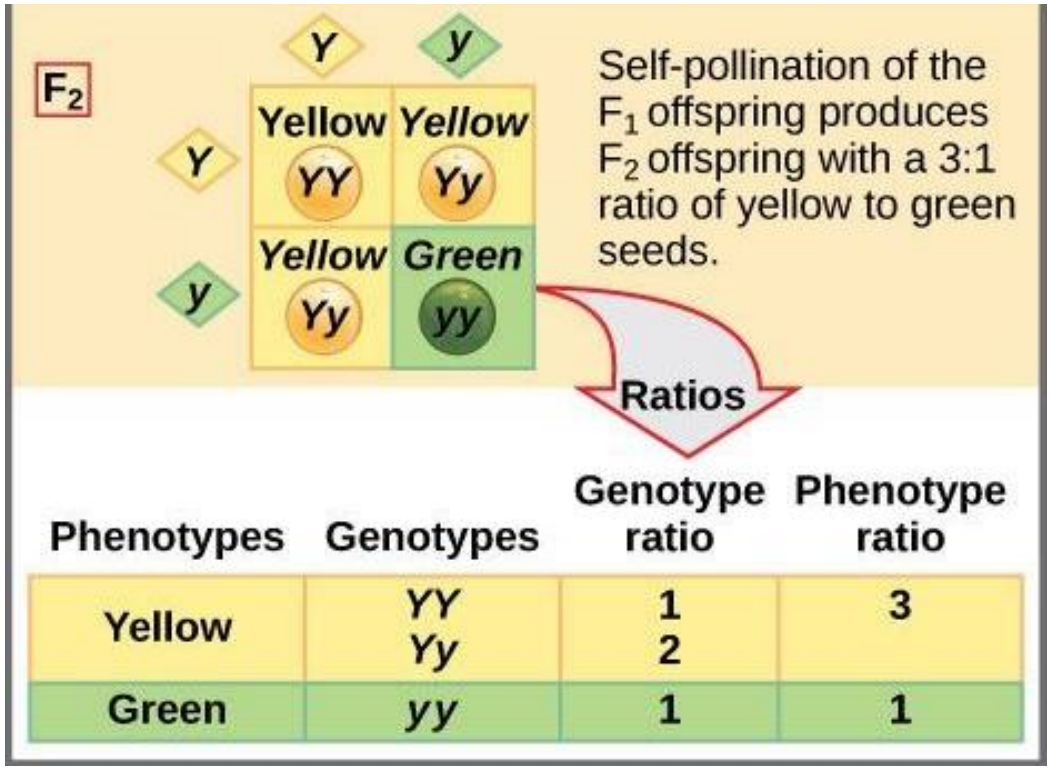
Expt 1: Form of seed. — From 253 hybrids 7,324 seeds were obtained in the second trial year. Among them were 5,474 round or roundish ones and 1,850 angular wrinkled ones. Therefrom the ratio 2.96:1 is deduced.

Expt 2: Color of albumen. — 258 plants yielded 8,023 seeds, 6,022 yellow, and 2,001 green; their ratio, therefore, is as 3.01:1.

Mendel, Gregor. 1866. Versuche über Plflanzen-hybriden. Verhandlungen des naturforschenden Ver-eines in Brünn, Bd. IV für das Jahr 1865, Abhand-lungen, 3–47; translated into English by William Bateson in 1901

Where does the 3:1 ratio come from?





Phenotypes	Genotypes	Genotype ratio	Phenotype ratio
Yellow	YY Yy	1 2	3
Green	yy	1	1

Source: https://raider.pressbooks.pub/biology1/chapter/10-punnett-squares/

This is true of many traits that Mendel studied

Trait	Dominant vs recessive	F_2 generations		Ratio
		Dominant	Recessive	
Flower colour		705	224	3.15 : 1
Seed colour		6022	2001	3.01 : 1
Seed shape		5474	1850	2.96 : 1
Pod colour		428	152	2.82 : 1
Pod shape		882	299	2.95 : 1
Flower position		651	207	3.14 : 1
Plant height		787	227	2.84 : 1

Source: https://old-ib.bioninja.com.au/standard-level/topic-3-genetics/34-inheritance/punnett-grids.html

Individual plants show much higher variability from the 3:1 ratio

	Experiment 1 Form of the Seed		Experiment 2 Color of the Albumen	
Plants	round	wrinkled	yellow	green
1	45	12	25	11
2	27	8	32	7
3	24	7	14	5
4	19	10	70	27
5	32	11	24	13
6	26	6	20	6
7	88	24	32	13
8	22	10	44	9
9	28	6	50	14
10	25	7	44	18

It's important to have a large enough sample size so that random fluctuations don't influence the overall result too much!

Mendel, Gregor. 1866. Versuche über Plflanzen-hybriden. Verhandlungen des naturforschenden Ver-eines in Brünn, Bd. IV für das Jahr 1865, Abhand-lungen, 3–47; translated into English by William Bateson in 1901

Outline

- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

Continuous vs. Discrete Probability Distributions

- Discrete distributions come from counting (coin flips, die rolls, sequencing reads), e.g.:
  - Heads, Heads, Tails, Tails, Heads, Tails
  - 1, 1, 0, 0, 1, 0
  - 1, 3, 4, 5, 2, 4, 1
- **Either integers or a countable number of outcomes**
- Continuous distributions come from measuring (fluorescence, luminescence, absorbance), typically non-integer amounts

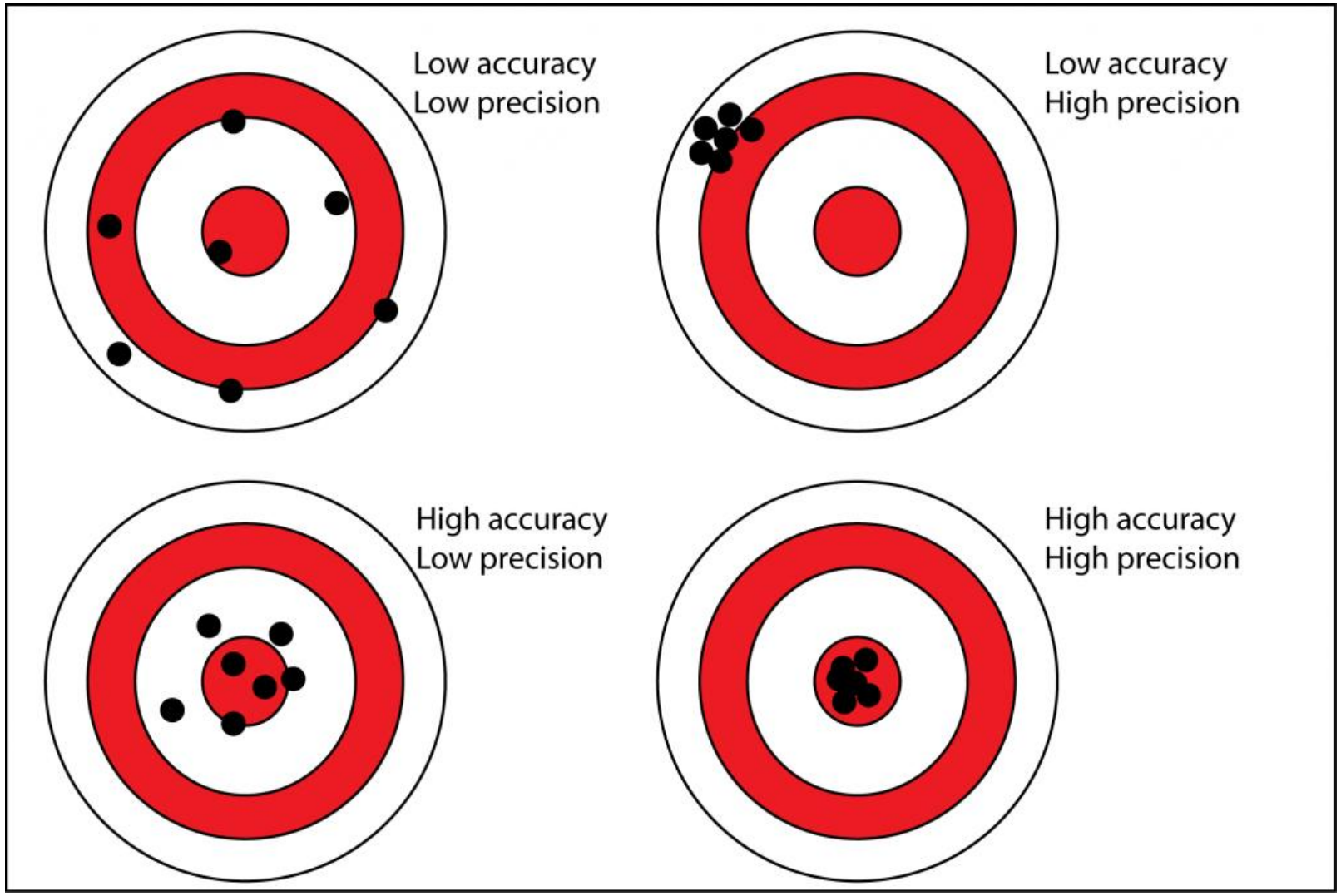
Biology is full of measurements

- Absorbance
- Luminescence
- Fluorescence
- [Radioactivity]
- Sequencing (will be covered by Mohamed Donia)
- Imaging (will be covered by Josh Shaevitz)

Measurements have error associated with them

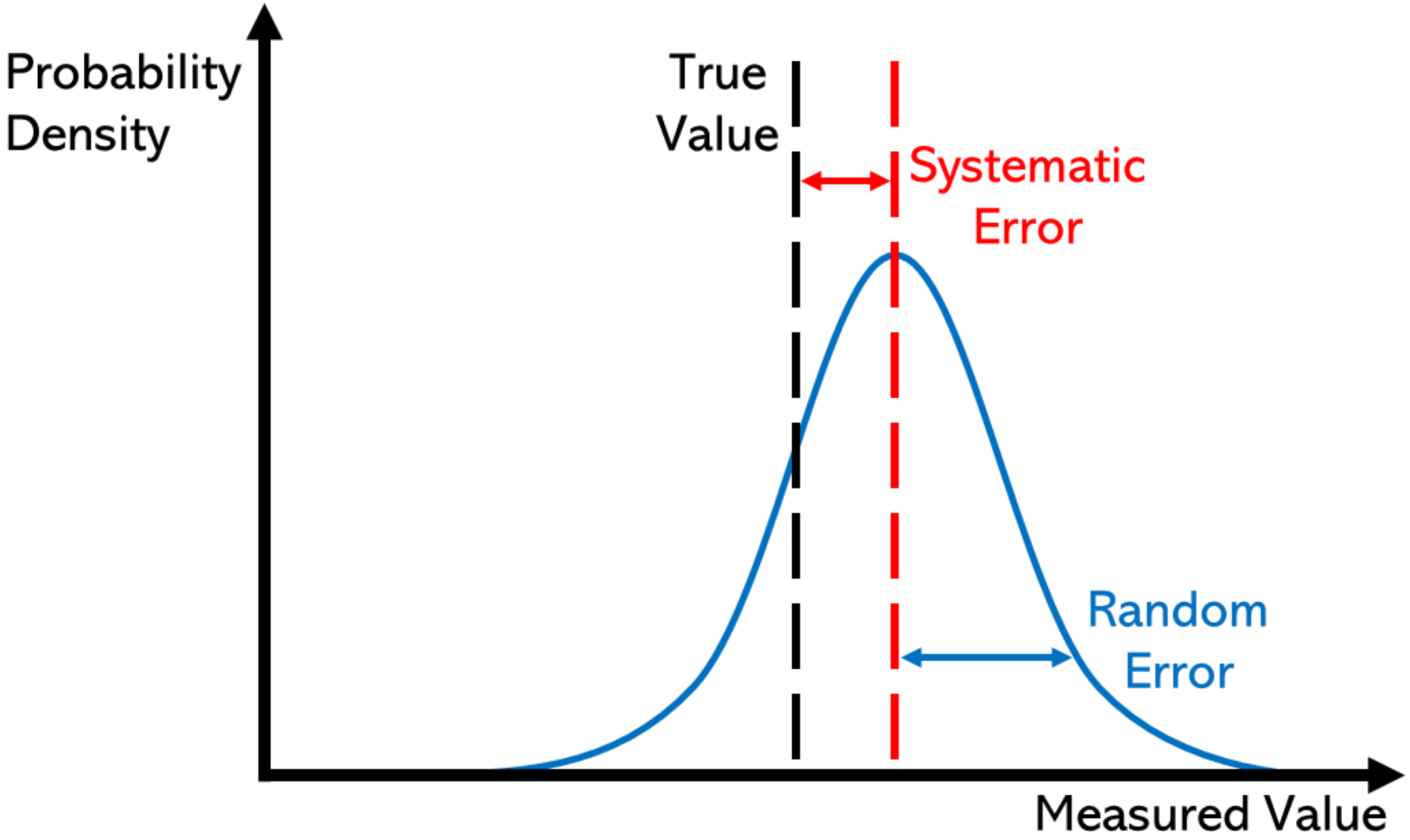
- Two types of measurement error: random and systematic
- Random error is always present, usually comes from the instrument itself
  - This is why we average across measurements!
- Systematic error is constant (or proportional to the true value)
- Some causes:
  - Improper calibration of the instrument (standards are important!!)
  - Environment influencing measurement (e.g. room lights for a fluorescent microscope, absorbance of a buffer)
- Systematic error is predictable
- This is why we often use a “blank” (for absorbance), or background-subtract (for fluorescence)

Two aspects of measurement error: precision and accuracy



Source: https://www.portaspecs.com/precision-and-accuracy/

Another way of visualizing systematic and random error



Source: https://en.wikipedia.org/wiki/Observational_error#/media/File:Measurement_distribution_with_systematic_and_random_errors.svg

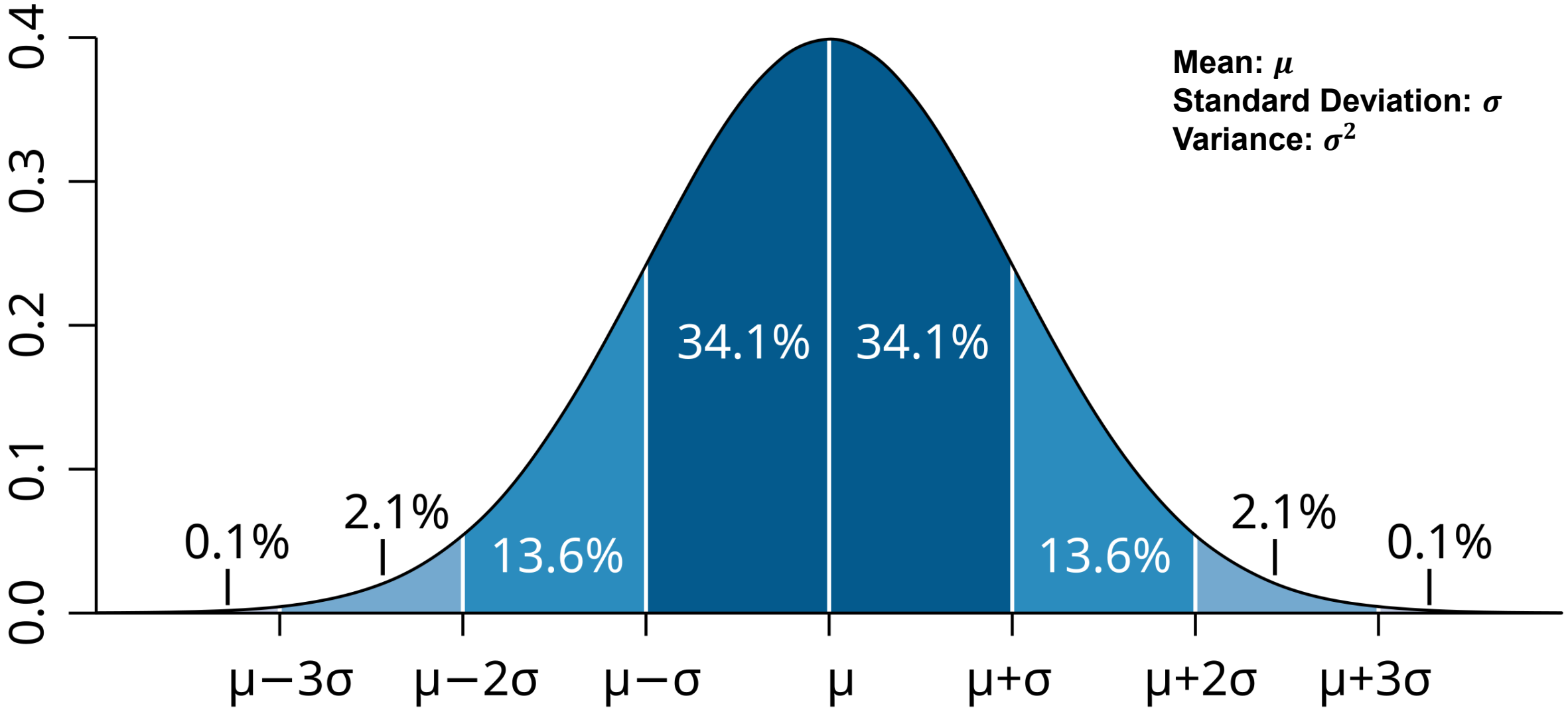
The sample mean often converges to the normal distribution

- This comes from the Central Limit Theorem in math

- Key assumptions:
  - Only true if we are taking the mean or the sum of a sample (e.g., averaging across replicates)
  - Large sample size
  - Samples come from the same distribution
  - Independence between observations

- In practice this lets us use the statistics of the normal distribution for many datasets

Properties of the normal distribution



Source: https://en.wikipedia.org/wiki/Normal_distribution#/media/File:Standard_deviation_diagram_micro.svg

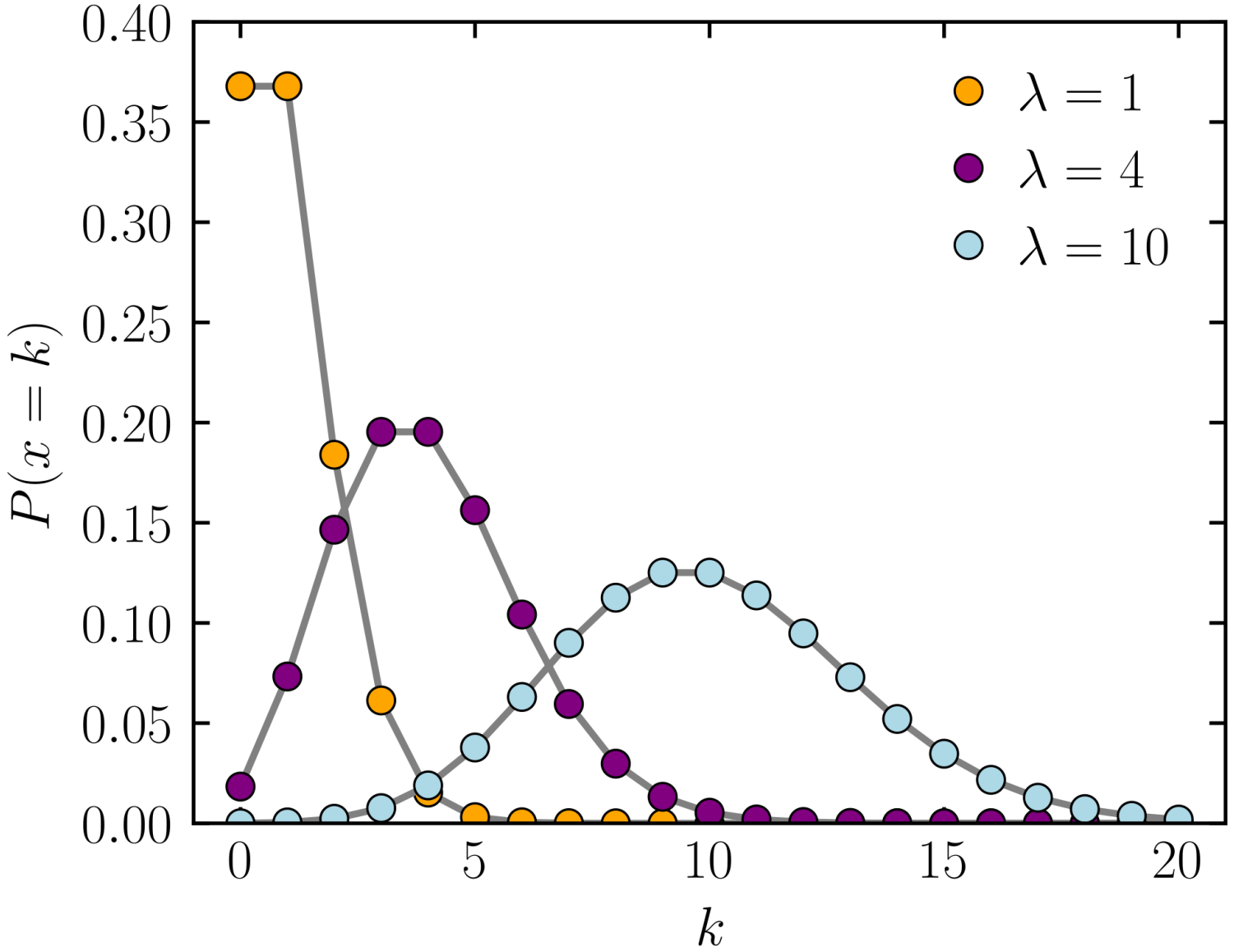
Outline

- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

Example: multiplicity of infection!

- Add 100 viral particles to a dish containing 100 cells
- Q: How many cells are infected by just one viral particle?
- A: only ~37%

The Poisson distribution has just a single parameter, λ



Both the mean and variance of the Poisson distribution are λ

- What does this mean?
- If you want consistency, use a low value of λ
- If you want almost every cell to be infected, use λ=3 or 4
- Trade-off as λ increases, you have more variability between cells

A more formal definition of the Poisson distribution

- Probability of a given number of events occurring in a fixed interval of time (or space), with two key assumptions:

1. Constant mean rate of occurrence λ

    - Would be broken if some cells are more susceptible to infection than others
    - Or, if some viral particles carry mutations that make them less infectious

2. Each event is independent of other events

    - Would be broken if viral infection causes a cell to become more (or less) susceptible to infection
    - Or, if a cell can influence its neighbors via immune signaling

- This is also called a Poisson process

The Poisson distribution is everywhere

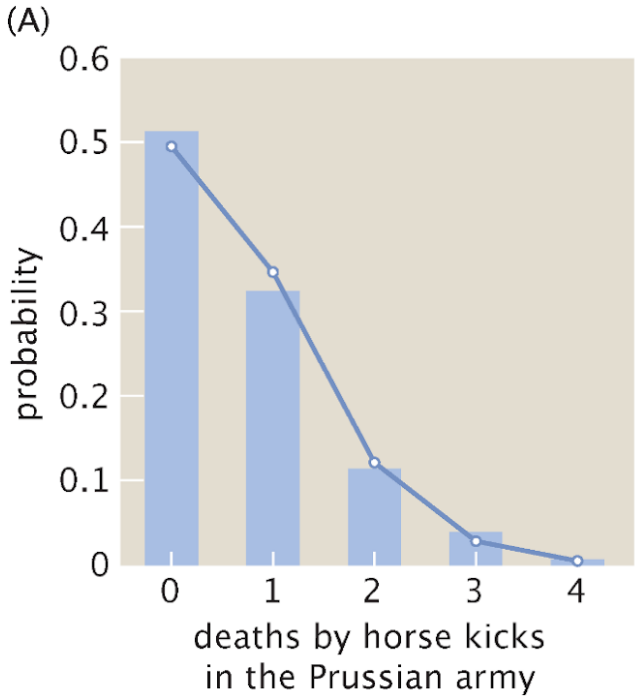
Table 4.1
Number of Deaths by Horsekicks in the Prussian Army from 1875-1894 for 14 Corps

Year	G*	I*	II	III	IV	V	VI*	VII	VIII	IX	X	XI*	XIV	XV	Total
1875								1	1				1		3
1876	2				1								1	1	5
1877	2						1	1			1		2		7
1878	1	2	2	1	1						1		1		9
1879				1	1	2	2		1			2	1		10
1880		3	2	1	1	1				2	1	4	3		18
1881	1			2	1			1		1					6
1882	1	2					1		1	1	2	1	4	1	14
1883			1	2		1	2	1		1		3			11
1884	3		1					1			2		1	1	9
1885							1			2		1		1	5
1886	2	1			1	1	1			1		1	3		11
1887	1	1	2	1			3	2	1	1		1	2		15
1888		1	1			1	1					1	1		6
1889			1	1		1	1			1	2	2		2	11
1890	1	2		2		1	1	2		2	1	1	2	2	17
1891				1	1	1		1	1		3	3	1		12
1892	1	3	2		1	1	3		1	1		1	1		15
1893		1				1		2			1	3			8
1894	1								1		1	1			4
Total	16	16	12	12	8	11	17	12	7	13	15	25	24	8	196

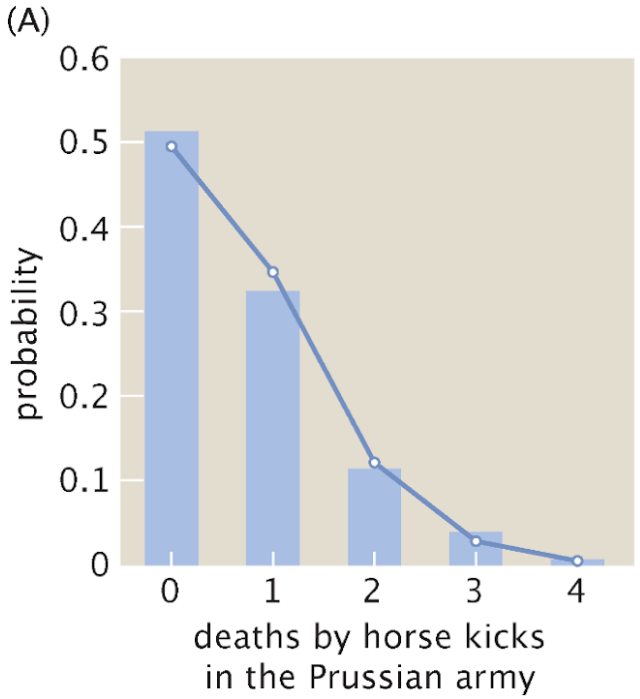
* G indicates Guard Corps
G,I,VI and XI Corps' organization differ from the others

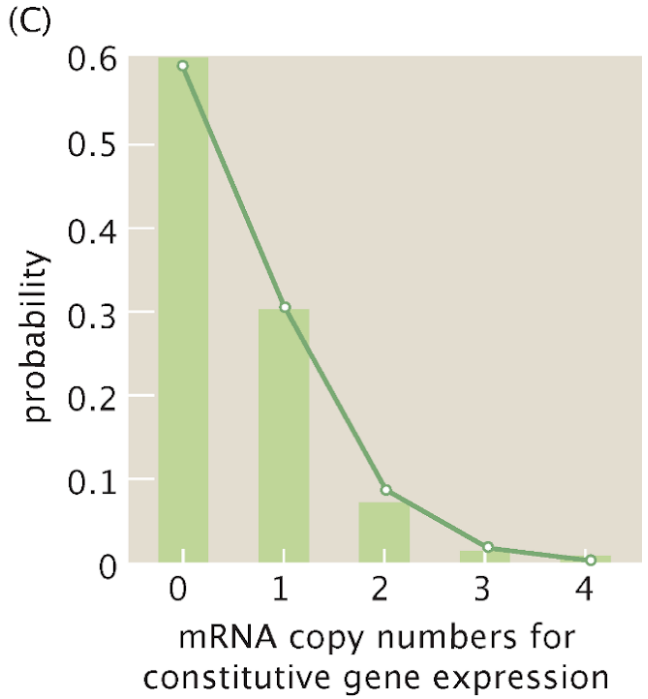
Source: https://link.springer.com/content/pdf/10.1007/978-1-4612-5098-2_5

The Poisson distribution is everywhere



The Poisson distribution is everywhere





An example from WWII

AN APPLICATION OF THE POISSON DISTRIBUTION

By R. D. CLARKE, F.I.A.

of the Prudential Assurance Company, Ltd.

READERS of Lidstone's *Notes on the Poisson frequency distribution* (*J.I.A.* Vol. LXXI, p. 284) may be interested in an application of this distribution which I recently had occasion to make in the course of a practical investigation.

During the flying-bomb attack on London, frequent assertions were made that the points of impact of the bombs tended to be grouped in clusters. It was accordingly decided to apply a statistical test to discover whether any support could be found for this allegation.

An area was selected comprising 144 square kilometres of south London over which the basic probability function of the distribution was very nearly constant, i.e. the theoretical mean density was not subject to material variation anywhere within the area examined. The selected area was divided into 576 squares of $\frac{1}{4}$ square kilometre each, and a count was made of the numbers of squares containing 0, 1, 2, 3, ..., etc. flying bombs. Over the period considered the total number of bombs within the area involved was 537. The expected numbers of squares corresponding to the actual numbers yielded by the count were then calculated from the Poisson formula:

Was London bombed randomly?

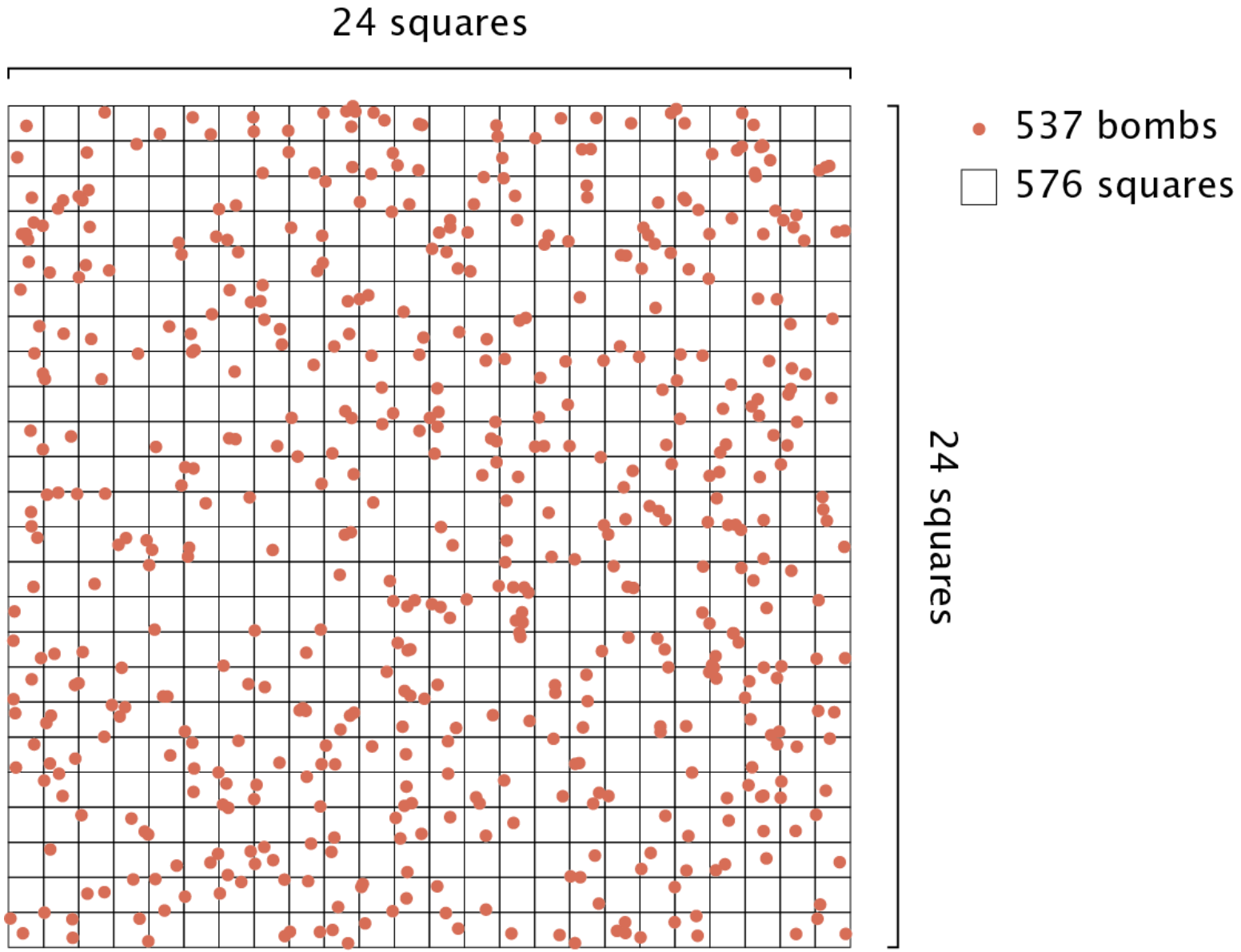
An area was selected comprising 144 square kilometers of south London.

The selected area was divided into 576 squares of ¼ square kilometer each, and a count was made of the number of squares containing 0, 1, 2, 3 …, etc. flying bombs.

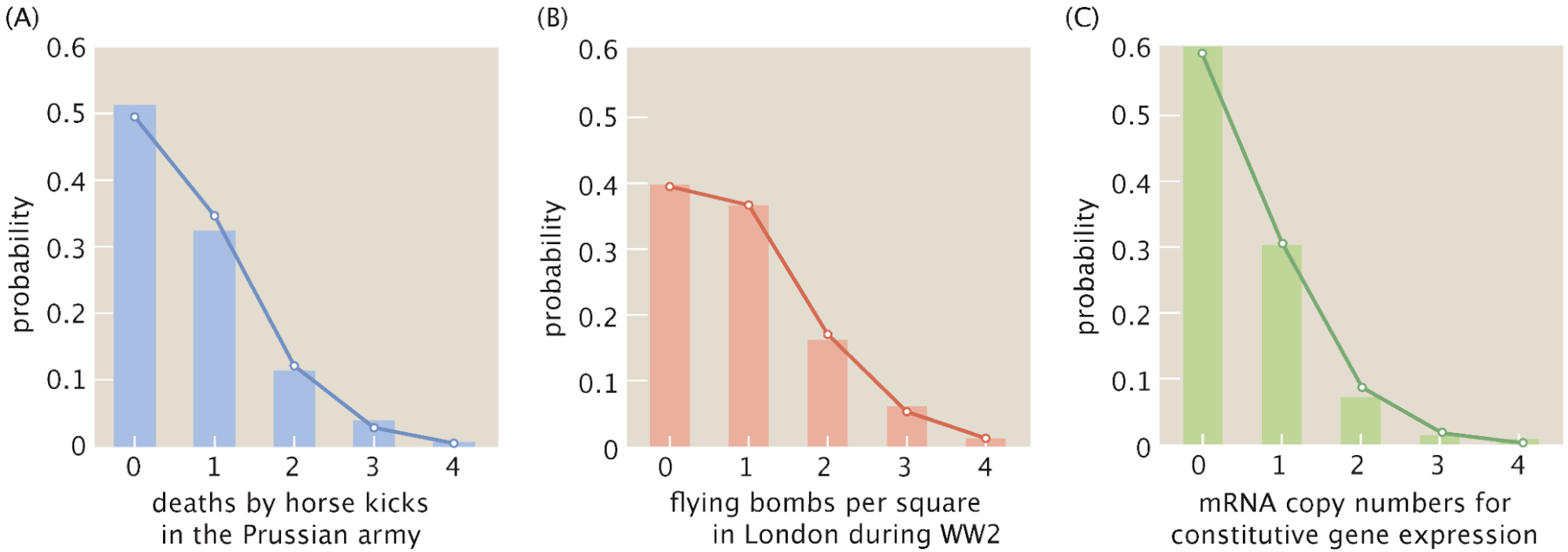
Over the period considered the total number of bombs within the area involved was 537.

No. of flying bombs per square	Expected no. of squares (Poisson)	Actual no. of squares
0 1 2 3 4 5 and over		
	576·00	576

Dividing London into a grid



The Poisson distribution is everywhere!



How general is the Poisson distribution?

- Useful when two sets of molecules, viruses, or cells are interacting, with the potential for multivalency, and when λ is not large

- When λ is large, the Poisson distribution converges to a normal distribution

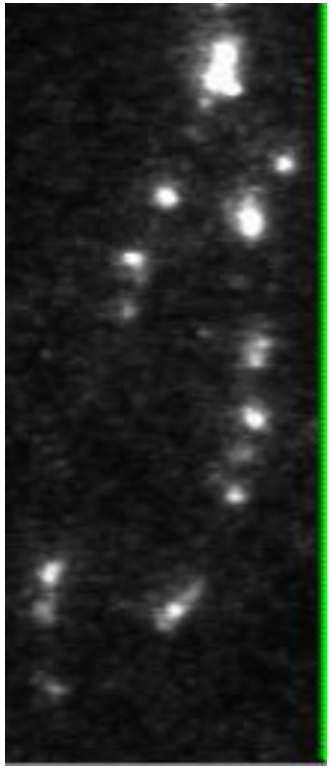
Other examples:

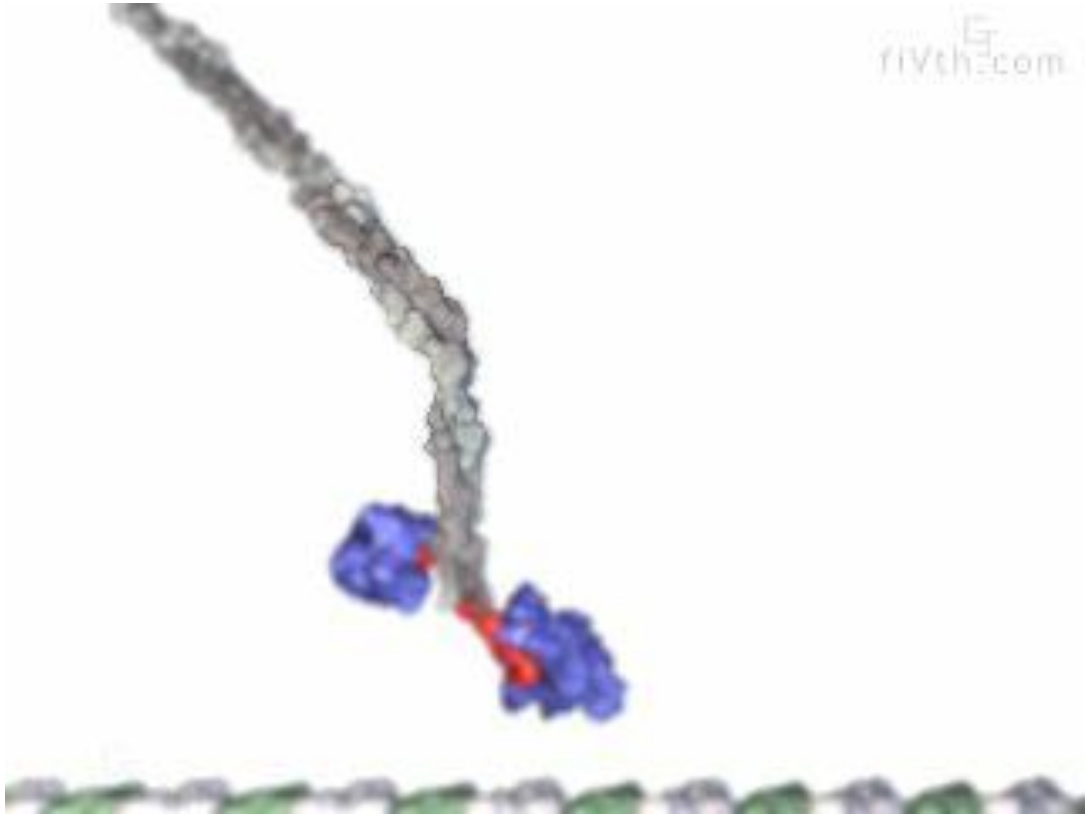
- Ligands binding to cells
- Partitioning cells into droplets for scRNA-seq
- Occurrence of mutations in cells
- Detection of fluorophores in a given volume using single-molecule imaging

Outline

- What is Probability? (Binomial distribution)
- Probability Distributions as Null Hypotheses in Biology
- The Normal Distribution
- The Poisson Distribution
- The Exponential Distribution

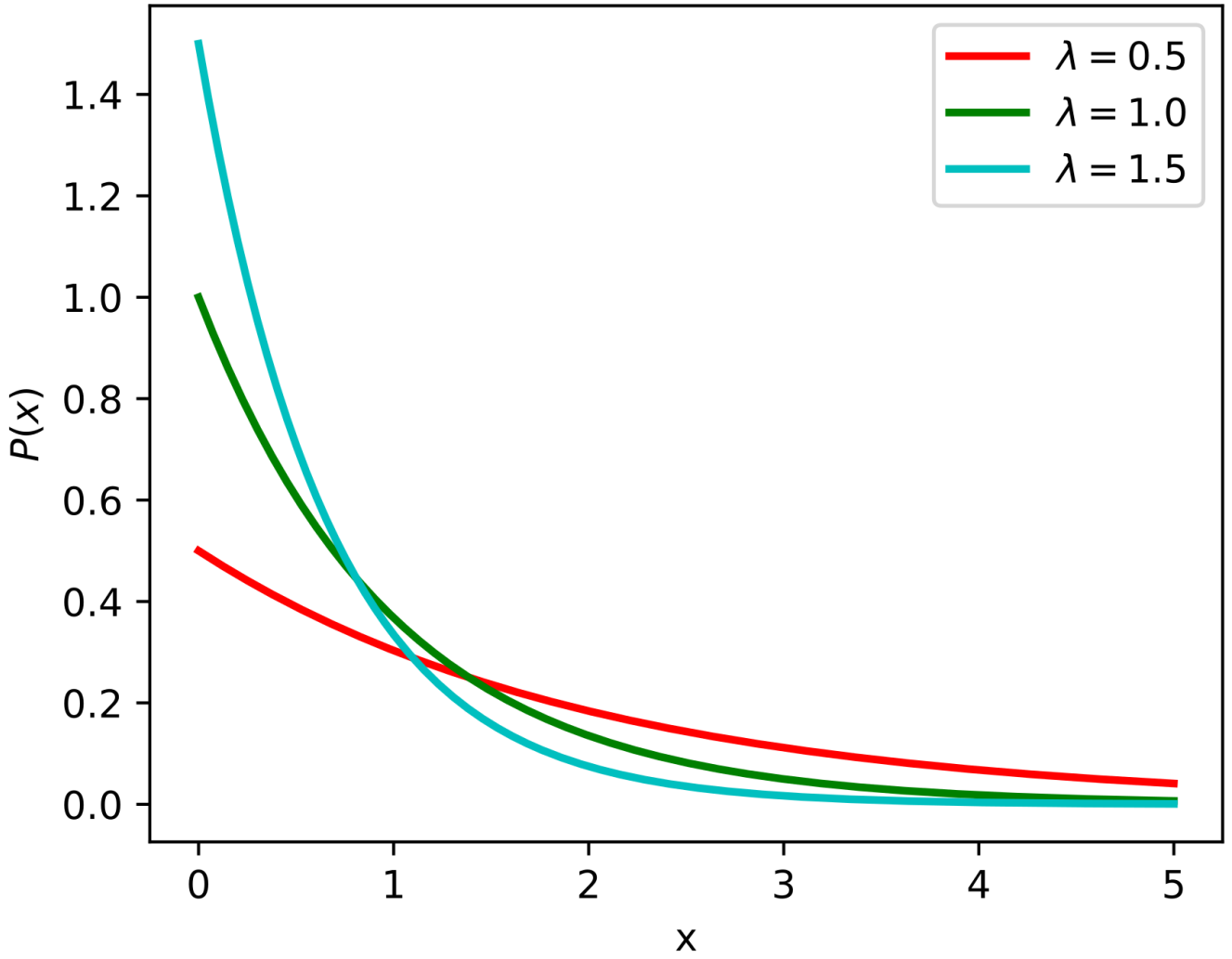
Example: molecular motor walking along a filament





Source: https://valelab4.ucsf.edu/external/moviepages/moviesMolecMotors.html

The exponential distribution describes the wait time between steps



Source: https://en.wikipedia.org/wiki/Exponential_distribution#/media/File:Exponential_distribution_pdf_-_public_domain.svg

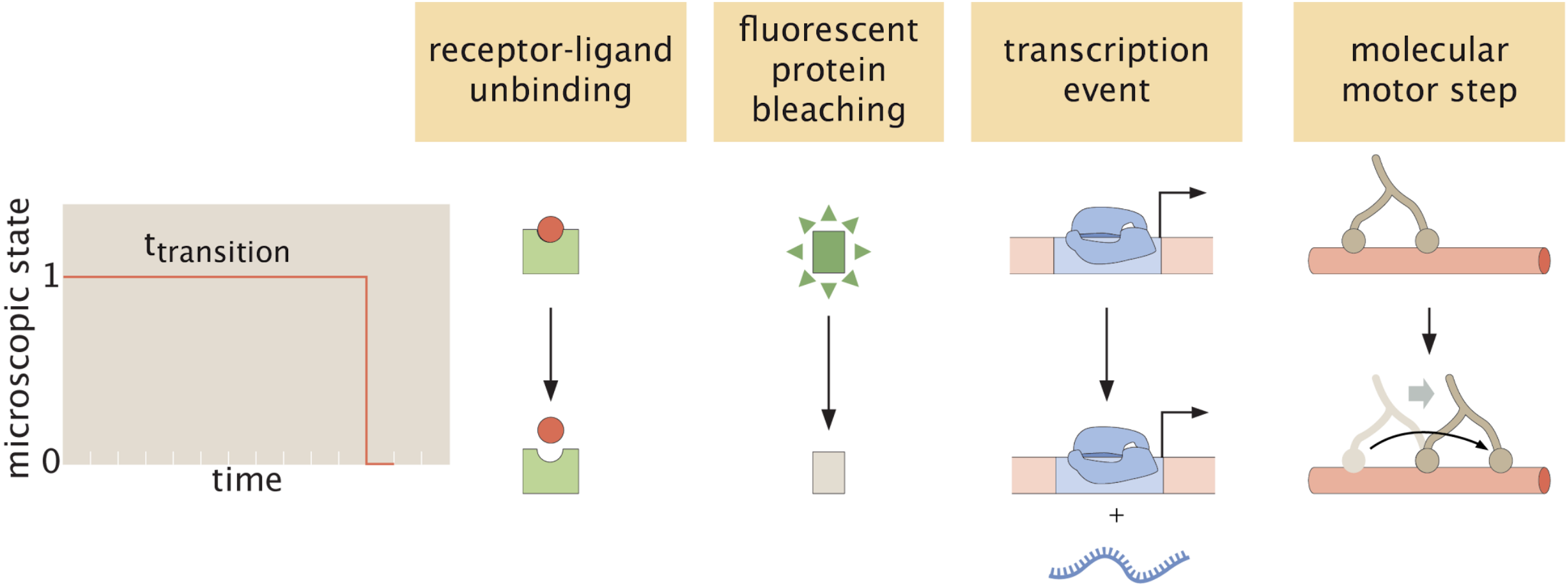
Math behind the exponential distribution

- The wait time between steps in a Poisson process:

- λ is the rate parameter, must be positive

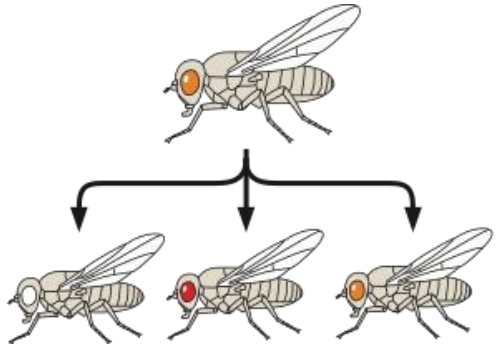
- Mean: $\mu = \frac{1}{\lambda}$
- Variance: $\sigma^2 = \frac{1}{\lambda^2}$
- Standard deviation, σ, is also $\frac{1}{\lambda}$

What is our null hypothesis for waiting times between molecular events?

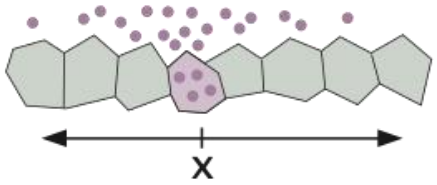


The four “Great Distributions” in Biology

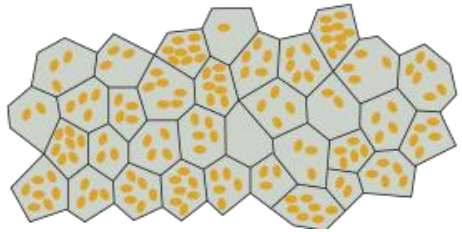
Allele segregation



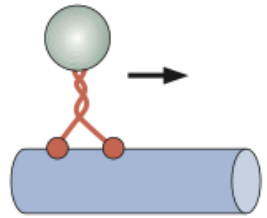
Diffusion



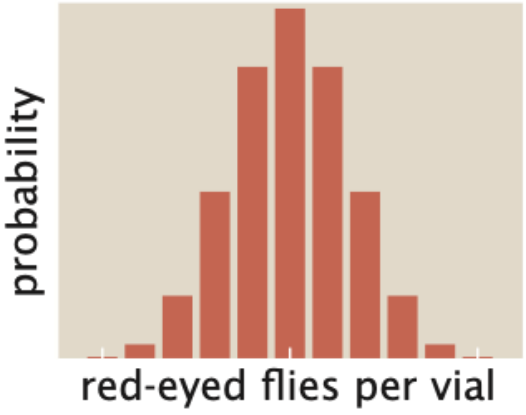
Viral Infection



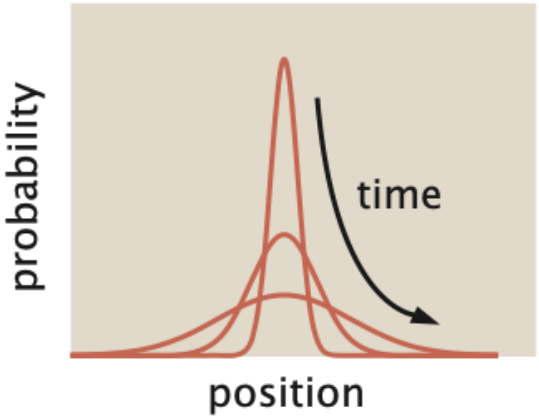
Wait times



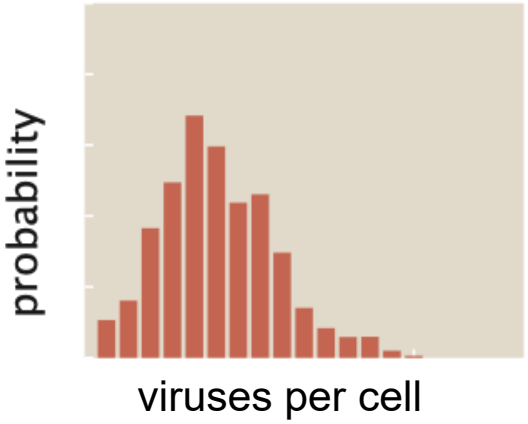
Binomial



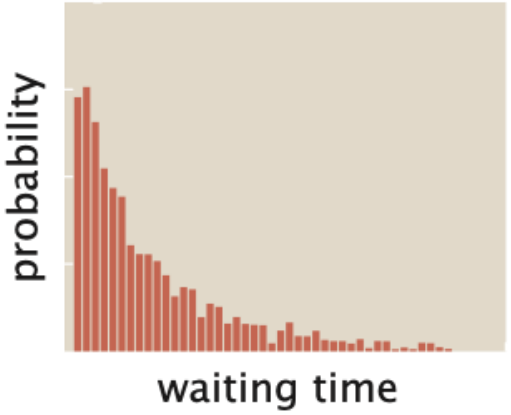
Normal



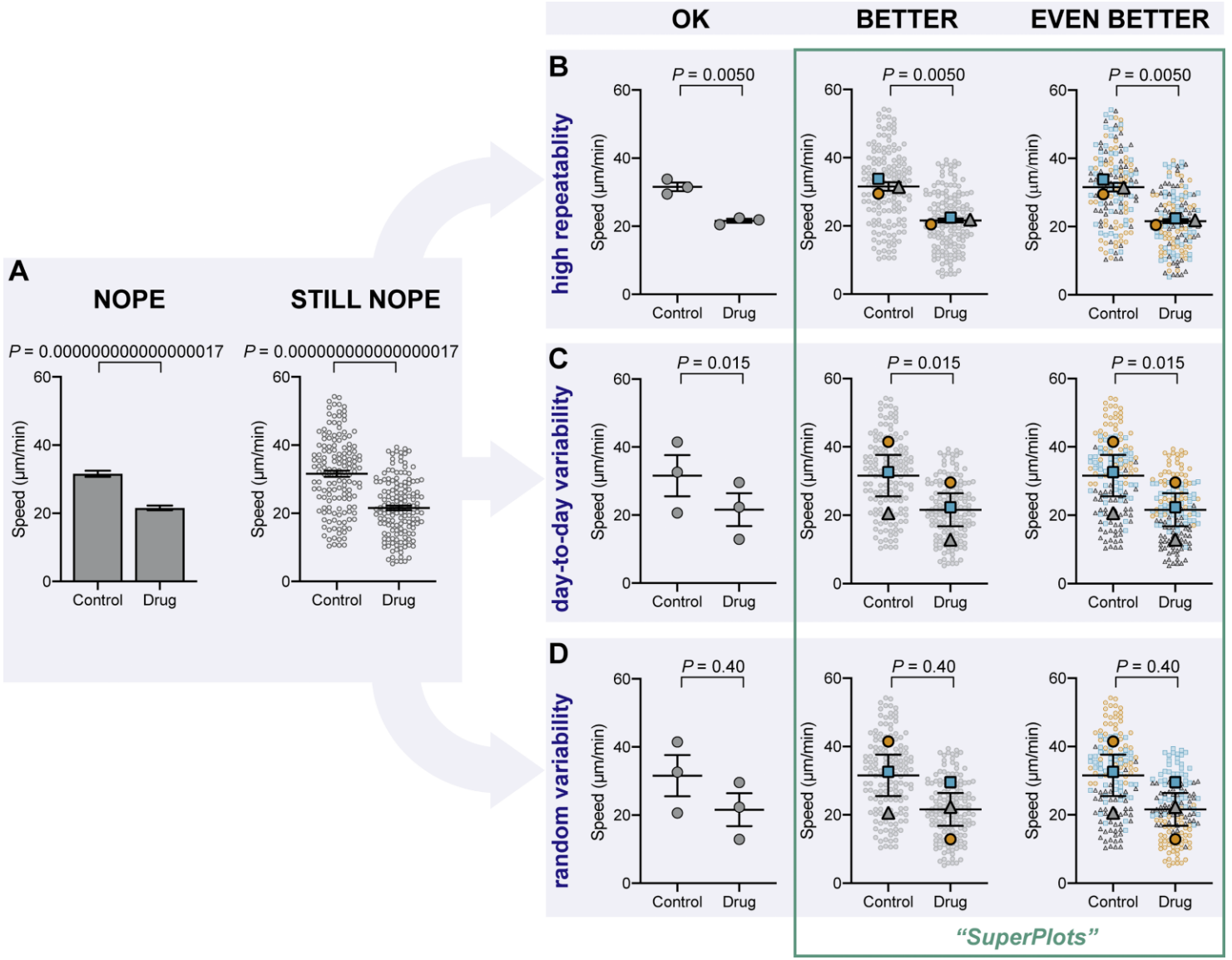
Poisson



Exponential



Reading for next time: SuperPlots



https://rupress.org/jcb/article/219/6/e202001064/151717/SuperPlots-Communicating-reproducibility-and

Extra Slides