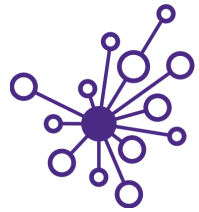
Knowledge and solutions for a changing world

Be boundless

Advancing data-intensive discovery in all fields



DISTRIBUTIONS & ERROR BARS

UW DIRECT

(Data Intensive Research Enabling Cutting-edge Tech)

https://uwdirect.github.io

Stéphanie Valleau

Chemical Engineering

Announcements

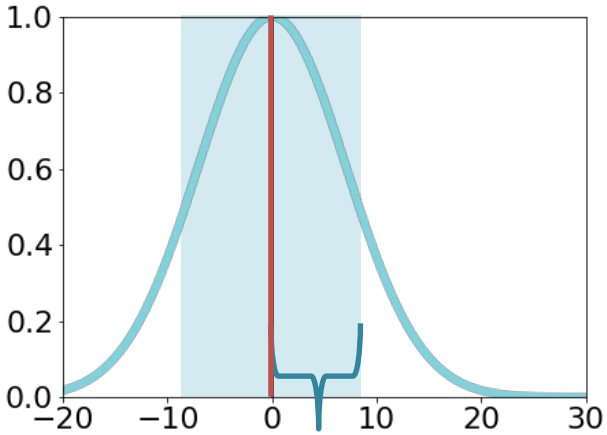
Ken Yashura from the Office for the Advancement of Engineering Teaching & Learning ET&L will be taking the **first 20min of lecture on Thursday 01/28/2021** to assess the class

- Please participate – **your feedback is fundamental to us!** It helps us understand what works and what does not work with the class
- We will not be present during the evaluation – your **feedback is confidential**
- Lecture will start right after the evaluation.

Review

- Population mean and standard deviation
- Population and Sample, *X*
- Sample mean and standard deviation
- Median, mode and variance
- Sampling Bias
- Sampling with/without replacement
- Bootstrapping (*we will see this in a few lectures*)

$$\mu = \sum_{i=1}^{N} \frac{X_i}{N}$$

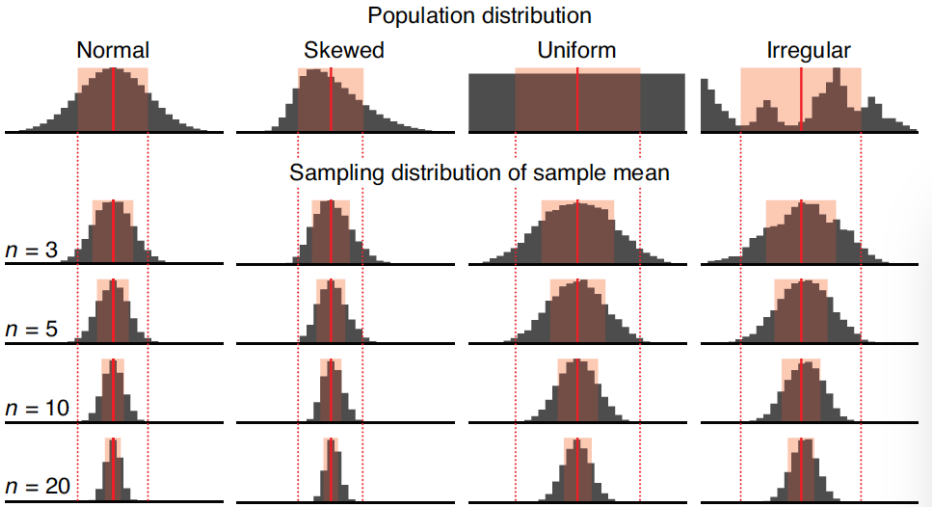


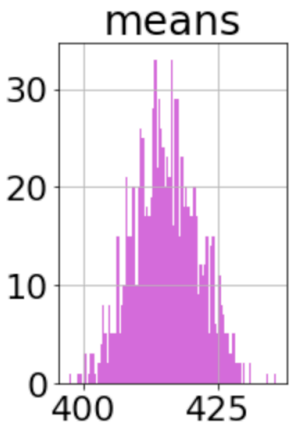
$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \mu)^2}{N}}$$

Review

The Central limit theorem

- As ***n*** increases, $\mu_{\bar{X}}$ decreases, i.e. we get **better and better estimates** of the population mean μ
- Thus big ***n*** makes $\mu \cong \mu_{\bar{X}}$
- As ***n*** increases, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$





Nature, 2013. Importance of being uncertain

Outline

- Continuous *vs* Discrete random variables
- Common PMF and PDF distributions and their uses
- Error bars
- The cumulative distribution function, CDF
- Sampling from distributions using the inverse CDF

Continuous random variable

A **continuous** random variable can take any value, bounded or unbounded

Some examples?

- You measure the weight of a new hydroscopic substance you have developed hourly for a week

Discrete random variable

A **discrete** random variable can only take certain values, e.g. integers

Some examples?

- Flip a count n times and count the number of heads. What is n?
- What is the range of values for the number of heads?
- Can you get $n/2$ heads? $n/3$ heads?

Continuous *vs* Discrete

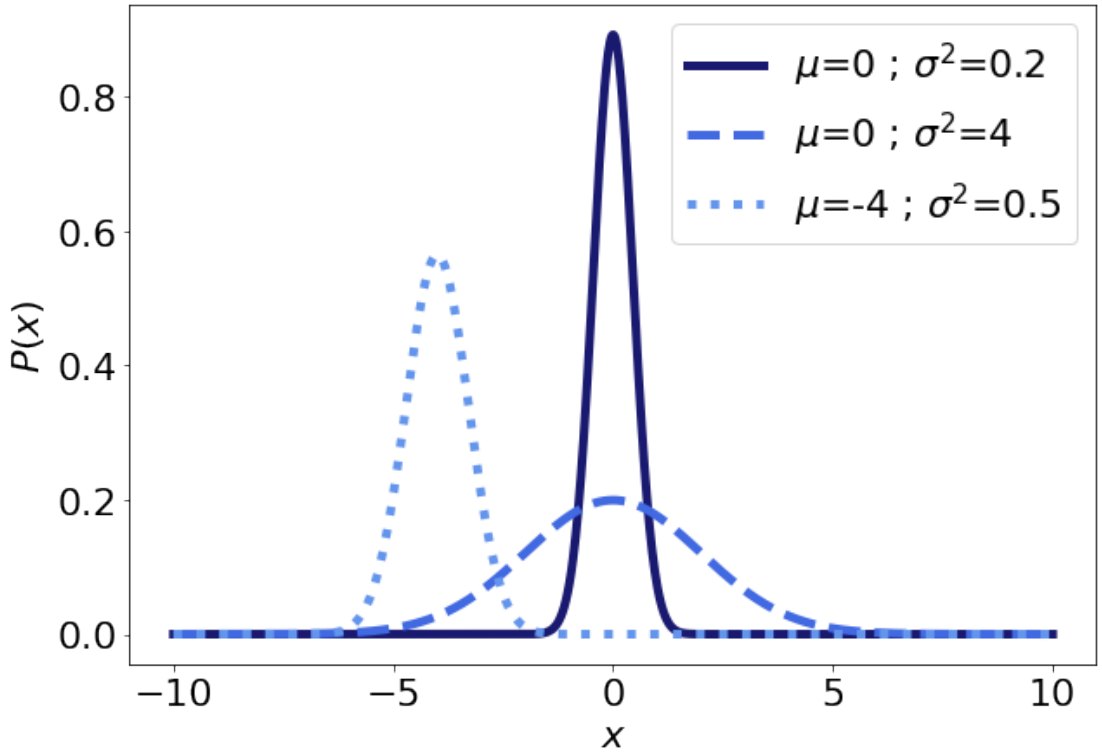
Discrete random variables have **discrete probability distributions**

Continuous random variables have **continuous probability distributions**

What is a probability distribution?

Probability distributions

For a given distribution, the **probability distribution function** *P*(X) is the probability of drawing the value *X* from the distribution



Normal distribution

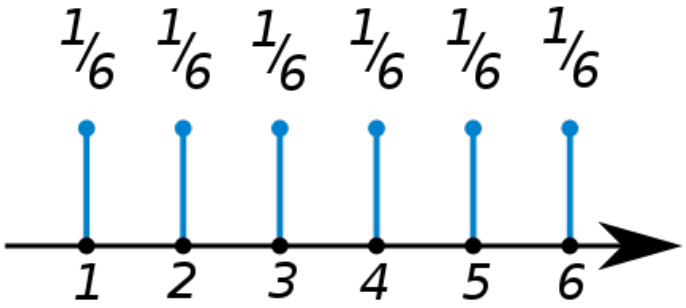
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Probability distributions

For **continuous** **random variables** the probability distribution is usually called the probability density function (**PDF**).

For **discrete** **random variables** the probability distribution is usually called the probability mass function (**PMF**).

What could this be a PMF of?



http://tinyurl.com/yattghb7

Discrete distrib. – Binomial

The **binomial distribution** can be used to describe the behavior of a random variable ***X*** **if**

- number of observations ***n*** is **fixed**
- each observation is **independent**
- the outcome of each observation is **boolean** - (“yes” or “no”, "success" or "failure”, “true” or “false”)
- **probability of "success" *p*** is the **same** for each outcome

Binomial coefficient

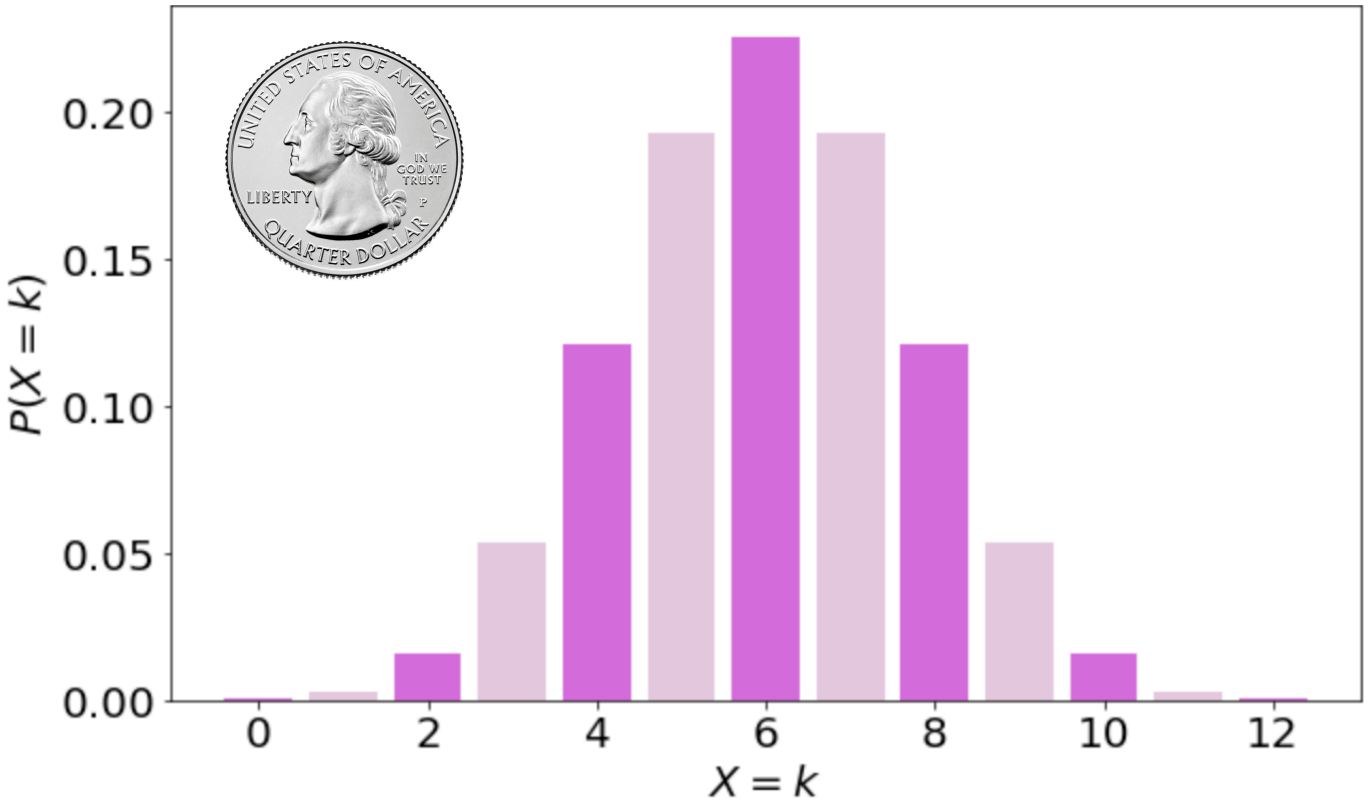
$$P(X = k) = P(n|k) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

Probability that $X = k$

Probability of failure $q = 1 - p$

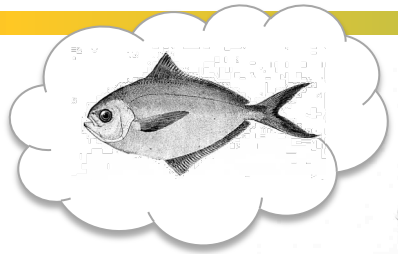
Binomial distribution

$$P(X = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$



n = 12
p = 0.5
q = (1 - p) = 0.5

Discrete distrib. – Poisson



The **Poisson distribution** can be used to describe the number of occurrences, ***k***, of an event - random variable ***X*** **- if**

- ***k*** number of times an event occurs in an interval – can take values 0, 1, 2, ...
- events occur independently
- average rate at which events occur is independent of any occurrences - usually assumed to be constant
- two events can't occur at exactly the same instant

Siméon Denis Poisson
1781 –1840

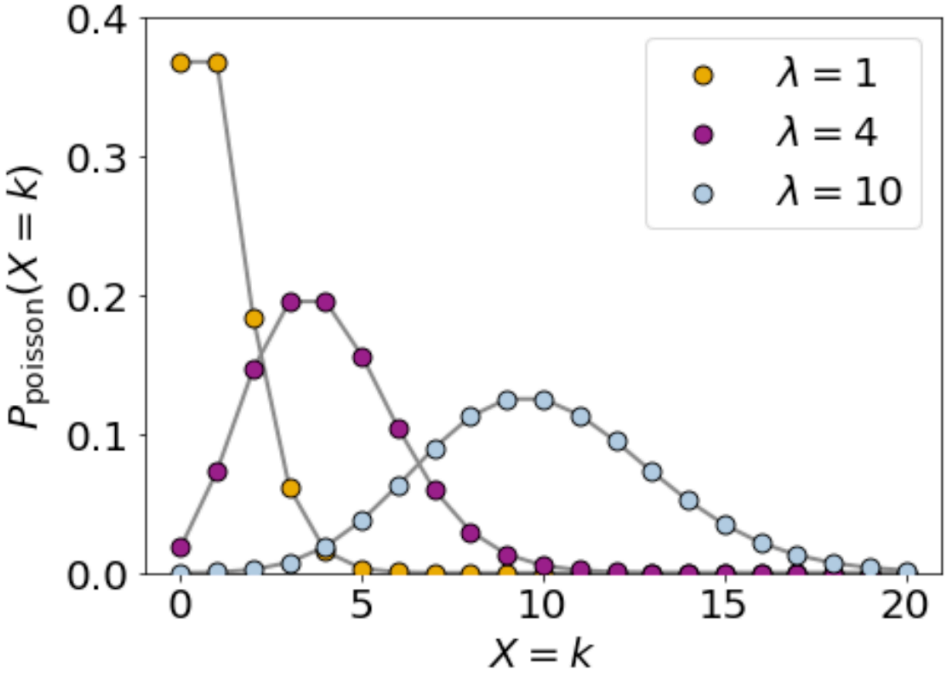
https://en.wikipedia.org/wiki/American_butterfish
https://en.wikipedia.org/wiki/Sim%C3%A9on_Denis_Poisson

Poisson distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$\lambda = \mu_{\text{poisson}} = E(X)$$

$$\lambda = \sigma^2_{\text{poisson}} = \text{Var}(X)$$



Applications

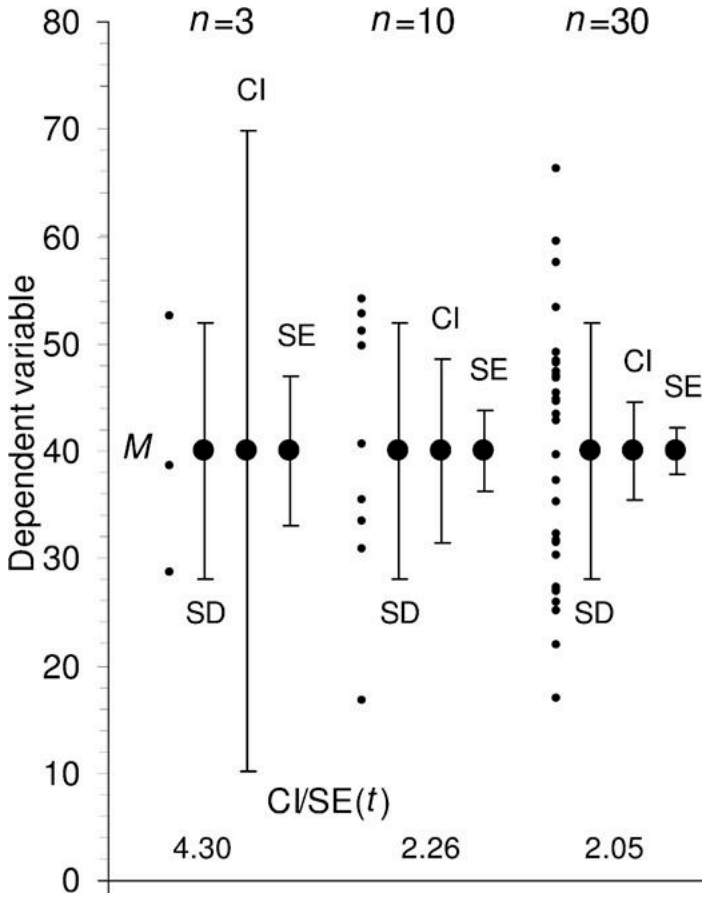
- number photons hitting a detector in a particular time interval (spectroscopy)
- number of mutations in given regions of a chromosome

Note: the Poisson distribution is the limit of the binomial distribution as $n \to \infty$

Distributions & Error bars

Error bars give you a measure of the **accuracy** of a measurement, calculation or inferred estimate of a quantity

- SD Standard deviation respect to the mean value (discrete distribution)
- SE / SEM Standard error of mean
- CI Confidence interval



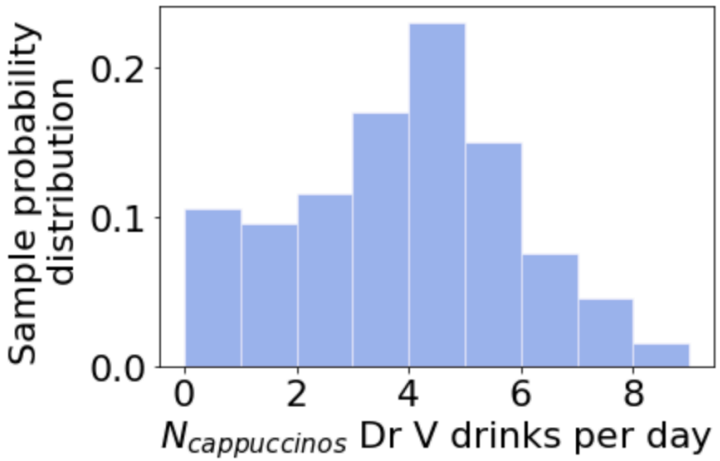
Error bars in experimental biology https://rupress.org/jcb/article/177/1/7/34602

Standard deviation

$$X_1 = \{x_0, x_1, \ldots, x_n\} \longrightarrow$$ Sample of number of cappuccinos Dr V drinks per day

Standard deviation

$$SD \equiv s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$



- descriptive error bar
- the smaller SD the better
- larger *n* means more accurate estimate of SD – get closer to SD of population

Standard error of mean

Samples of number of cappuccinos Dr V drinks per day

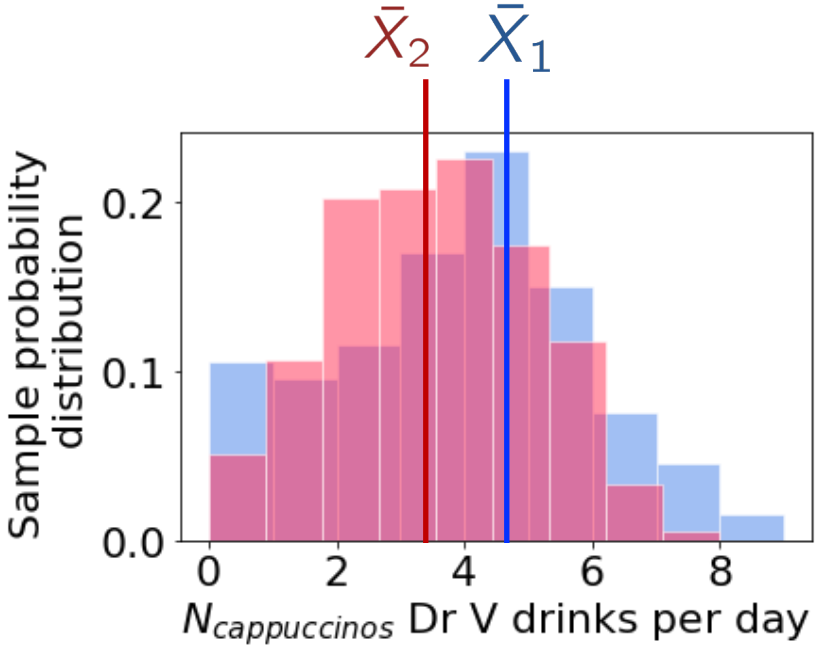
$$X_1 = \{x_0, x_1, \ldots, x_n\}$$
$$X_2 = \{x_0, x_1, \ldots, x_n\}$$

Standard error of mean

$$\sigma_{\bar{X}} \approx \frac{s}{\sqrt{n}}$$

Stdev. of sample

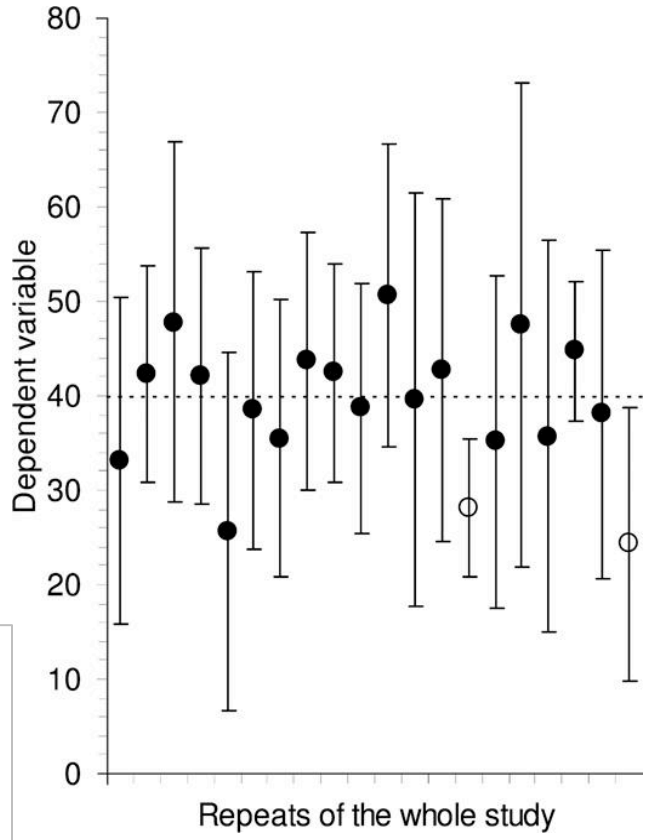
- inferential error bar
- estimate of variation between samples / uncertainty in mean



Confidence Interval (CI)

Confidence interval error bars tell us about the reliability of the measurement.

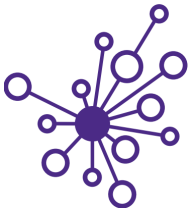
We will get back to confidence intervals after we talk about hypothesis testing and *p*-value.



Confidence intervals. Means and 95% CIs for 20 independent sets of results, each of size n = 10, from a population with mean μ = 40 (marked by the dotted line). In the long run we expect 95% of such CIs to capture μ; here 18 do so (large black dots) and 2 do not (open circles). Successive CIs vary considerably, not only in position relative to μ, but also in length. The variation from CI to CI would be less for larger sets of results, for example n = 30 or more, but variation in position and in CI length would be even greater for smaller samples, for example n = 3.

Error bars in experimental biology https://rupress.org/jcb/article/177/1/7/34602

Let's see PMFs in Jupyter!



Open the **L4_Distributions.ipynb** notebook

Continuous distrib. – Normal

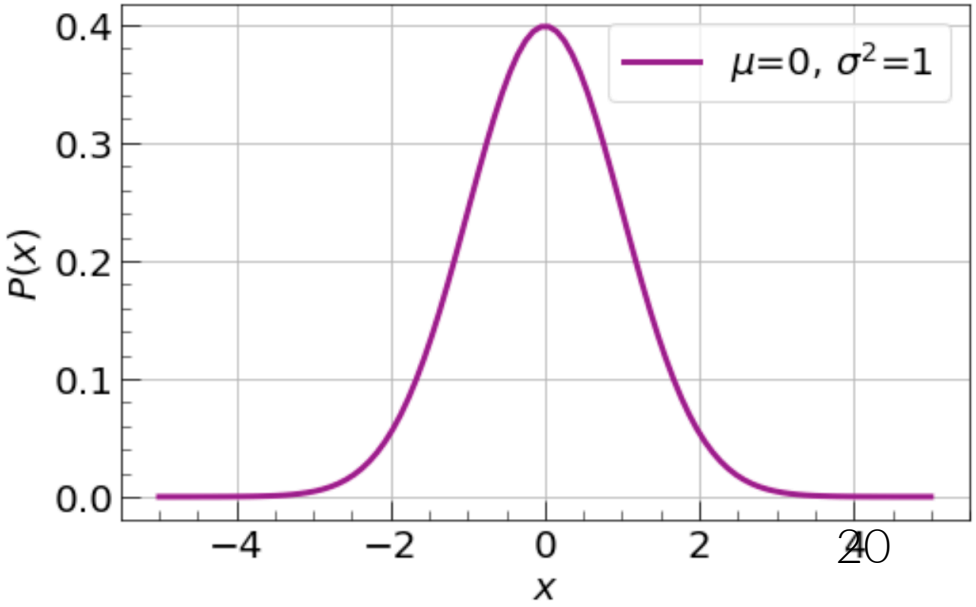
Normal distribution

$$P(x) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard normal distribution

mean = 0 & std. dev. = 1

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



Continuous distrib. – Gamma

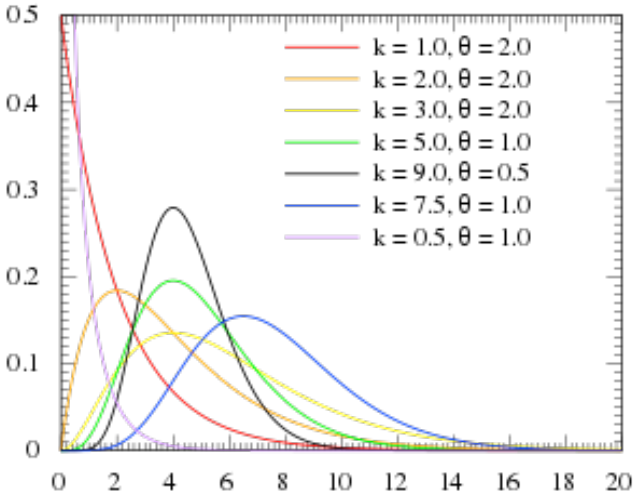
The **gamma distribution** describes the distribution of **waiting times** until the next *k*-th Poisson distributed occurrence

$$P(x) = \frac{x^{\alpha-1} e^{x/\theta}}{\Gamma(\alpha)\theta^{\alpha}}$$

$$\alpha = k \;\; ; \;\; \theta = 1/\lambda$$

shape

scale



E.g. Consider a spectroscopic experiment that collects a set of single molecule data: individual times between chemical reactions (i.e., the inverse of a reaction rate)

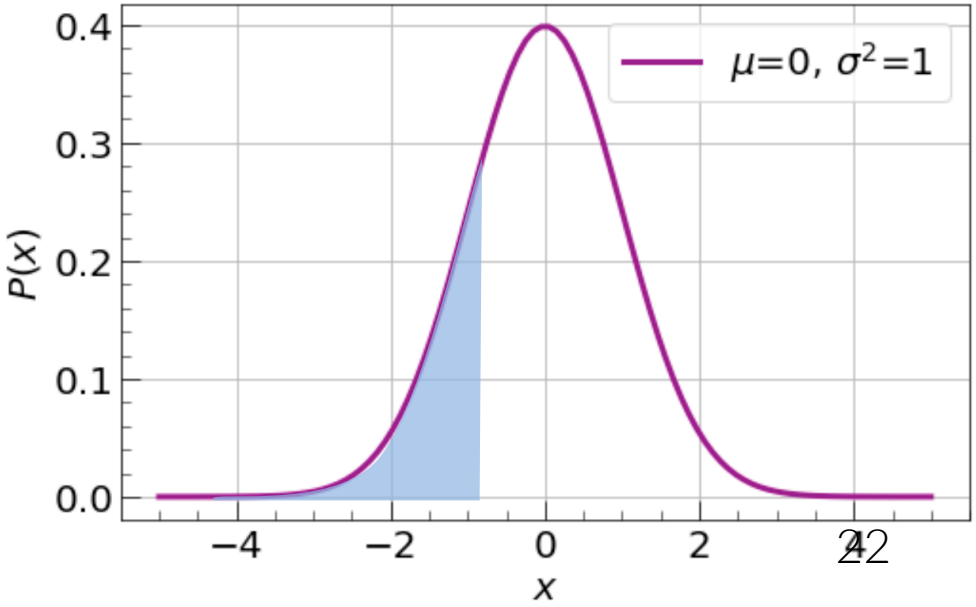
https://en.wikipedia.org/wiki/Gamma_distribution

Cumulative distribution function (CDF)

Probability distribution functions PDF, PMF: probability of **drawing a sample with a value *X***

Cumulative distribution function (CDF) is the probability of drawing a value from the distribution **less than *x*.**

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

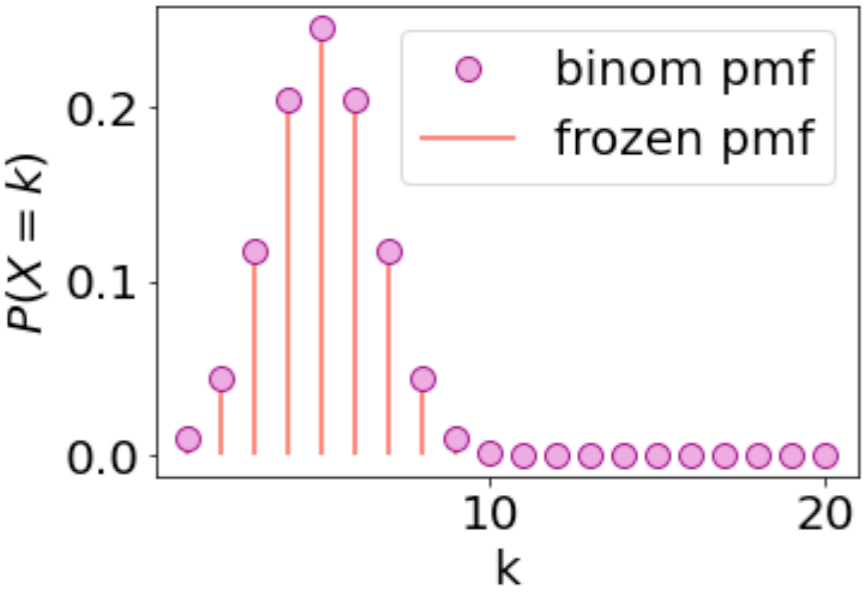


Binomial CDF

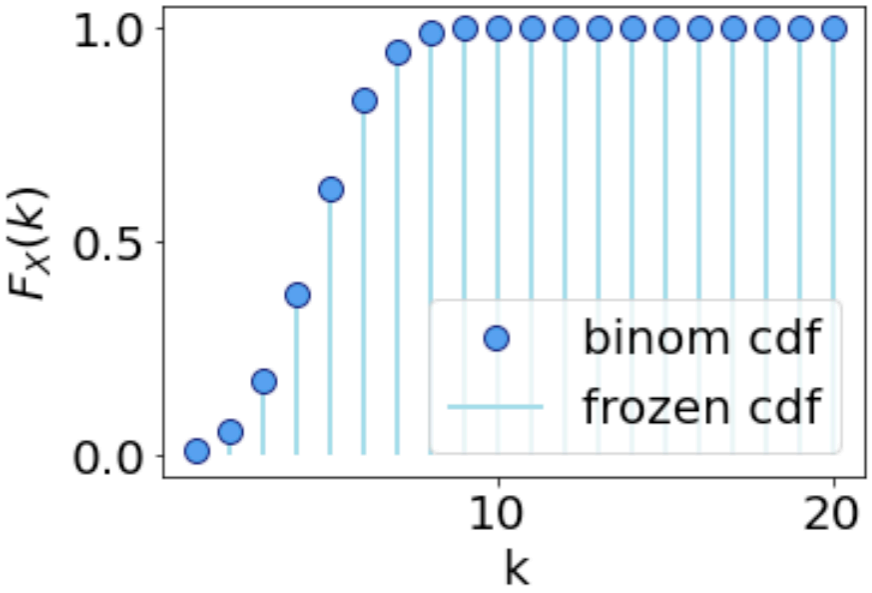
$$F_X(k) = \sum_{i=0}^{\lfloor k \rfloor} P(X = i)$$

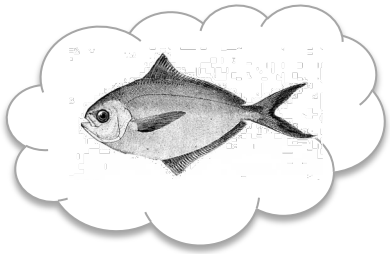
p = 0.5 ; n = 10

PMF



CDF

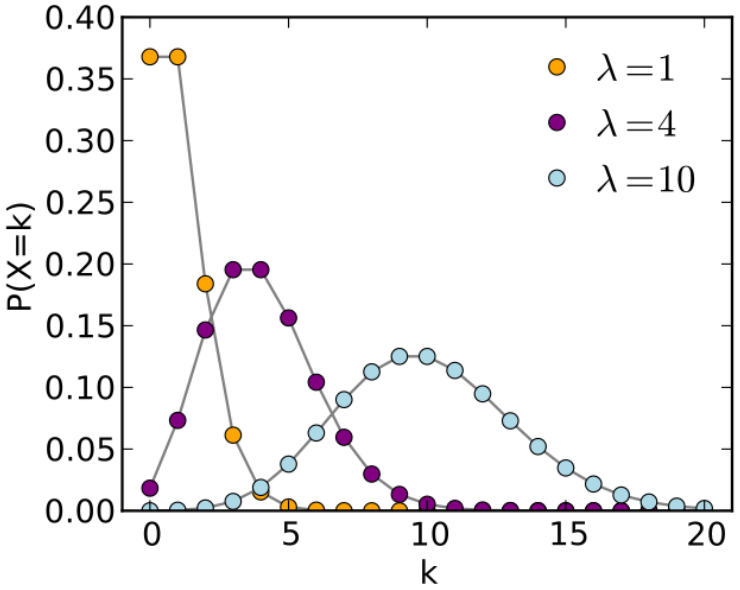




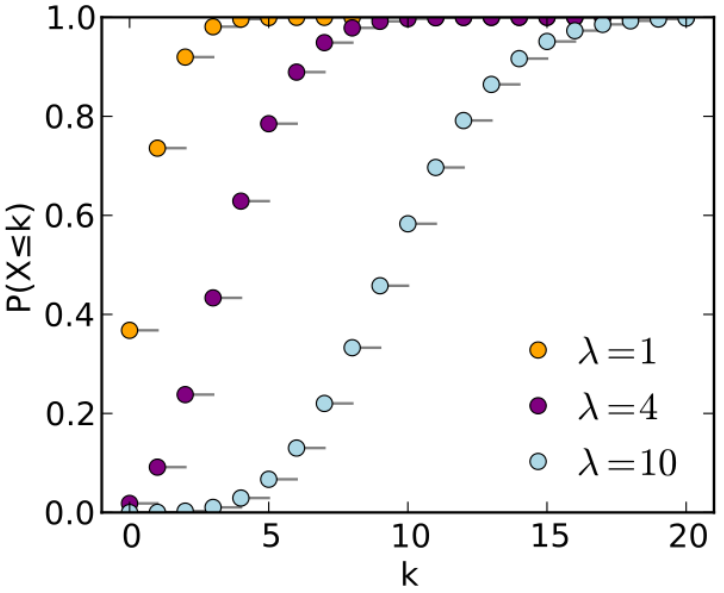
Poisson CDF

$$F_X(k) = \sum_{i=0}^{\lfloor k \rfloor} P(X = i)$$

PMF



CDF

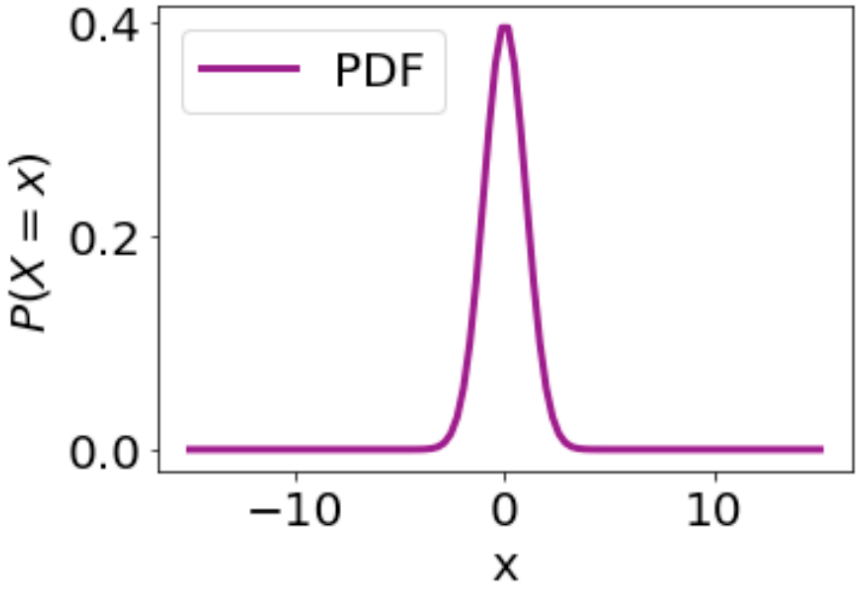


Cumulative distribution function (CDF)

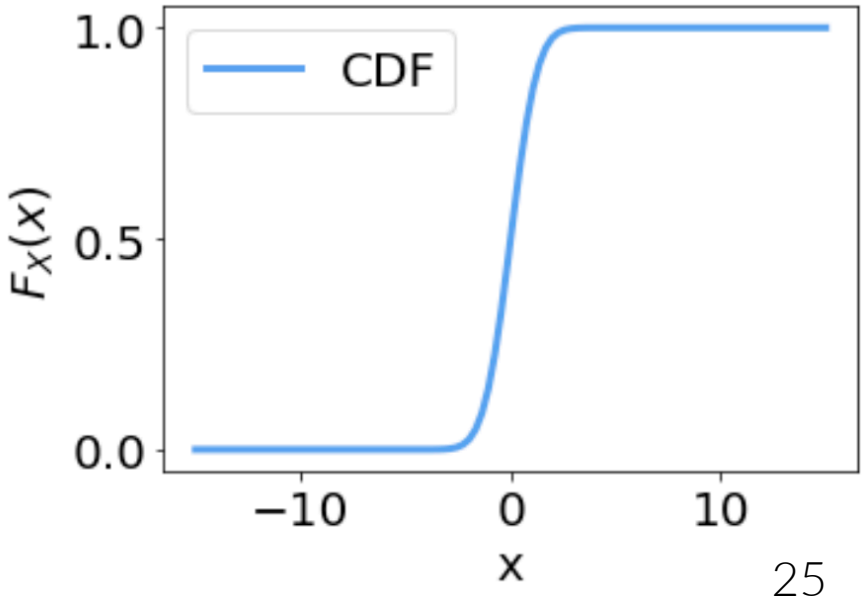
For a continuous distribution, F(x) is the integral of the PDF.

$$F_X(x) = \int_{-\infty}^{x} P(X = t)dt$$

PDF

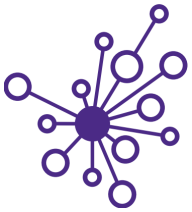


CDF



https://en.wikipedia.org/wiki/Normal_distribution

CDF's in Jupyter

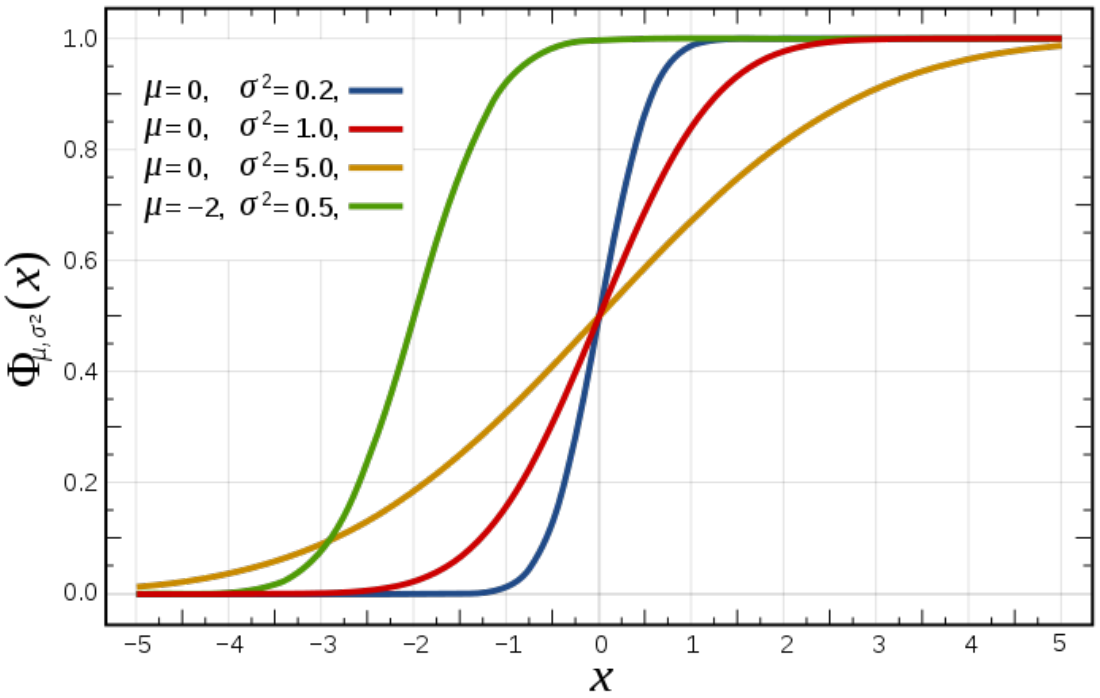


Let's go back to the notebook!

Cumulative distribution function (CDF)

Why is the CDF useful?

- E.g. survivability? What fraction of your mice are dead by day x?



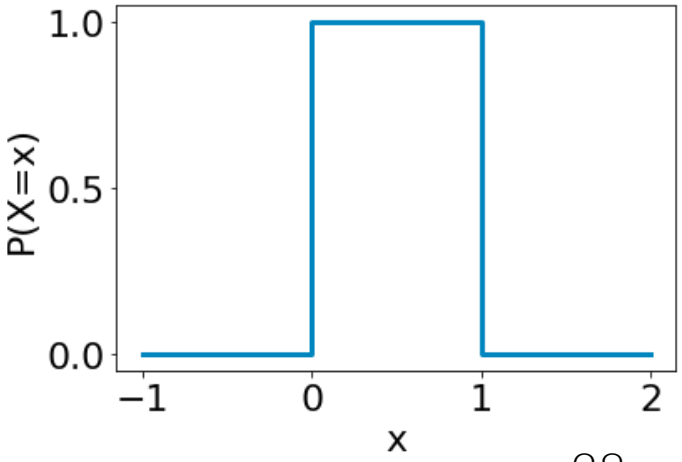
CDFs & distribution sampling

You can **sample** from a distribution **using its CDF** this approach is known as **inverse transform sampling**

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

$$F_X^{-1}(p) = x$$

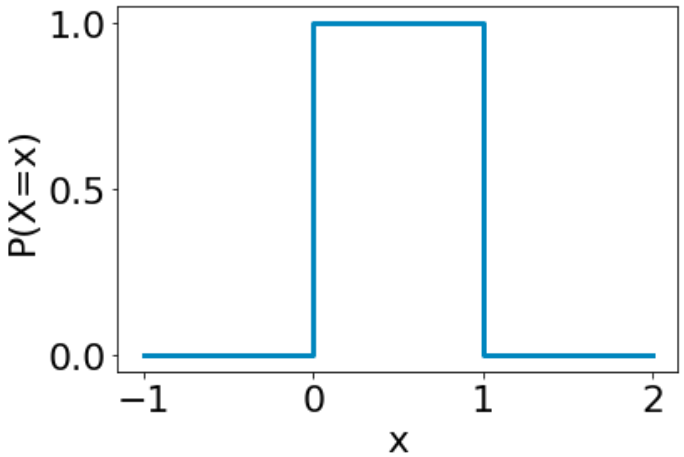
p is uniformly distributed – you can sample points from a uniform distribution and use the inverse CDF to sample points from the PDF



https://en.wikipedia.org/wiki/Probability_integral_transform

Inverse transform sampling

1. Generate a random number ***u*** from the **uniform distribution**
2. Find the **inverse of the CDF,** $F_X^{-1}(x)$
3. Compute $X = F_X^{-1}(u)$
4. X has distribution $F_X\ (x)$



Note: For continuous distributions, an analytical solution for the inverse CDF is not always possible (including normal)

Sampling from distrib. in Python!