Introduction to Statistics (Day 1/2)

Michelle Franc Ragsac, Ph.D. (*She/Hers/Siya*)
Postdoctoral Research Scholar — Amariuta Lab
Halicioglu Data Science Institute
mragsac@ucsd.edu

January 13th, 2026

Why should you care about probability and statistics?

01

Describe observations with a common quantitative language

What is the mean expression of my gene of interest across my five RNA-sequencing samples?

02

Make inferences about a population or a process from observations

What is the DNA mutation rate per base per generation?

03

Evaluate hypotheses systematically

Is the expression of my gene of interest greater in one condition versus the other condition?

04

Describe uncertainty (e.g., confidence intervals, probabilities)

How confident am I in a particular result from my RNA-sequencing experiment?

05

Make predictions about hypothetical scenarios or particular results

What is the probability that the expression of my gene of interest will be larger in one population than another?

06

Quantitatively describe relationships between variables

Is the expression of my gene of interest correlated with a particular genotype that is present in the population?

What are the topics of today’s lecture?

1. Descriptive Statistics
    a. Population versus Sampling from a Population
    b. Calculating the Mean, Variance, and Standard Deviation of a Dataset
    c. Precision and Accuracy
2. Palmer Penguins Dataset
3. Visualizing Data in Python

Distinction between populations and **sampling** from a population

When you observe nature or perform an experiment, you might be able to specify all subjects of interest

All babies born in Rady Children's Hospital last year

Every brewery that served beer in San Diego this year

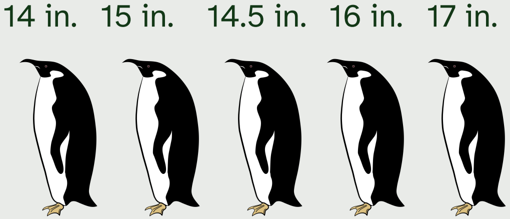
All of the individual penguins in the entire world

Sample of penguins in the population

However, it's unusual for entire populations to be small and easily defined, so we often take a **sample** of that population

Calculating the population mean, variance, & standard deviation

I'm interested in studying a hypothetical population of penguins living in **my pool**



Wingspan Lengths

Population Mean $\mu = \frac{\Sigma x_i}{n}$

What is the average wingspan within my population?

All of the individual penguins in **my pool**

Sample of penguins in the population

Population Variance $\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$

What is the spread or dispersion of wingspans within my population?

Population Standard Deviation $\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$

Estimating population characteristics by using a sample subset of observations

It’s often **impossible** to collect observations on **every member of a population**—*like all the penguins in the world*!

Sample Mean $$\bar{x} = \frac{\Sigma x_i}{n}$$

What is the average wingspan within the sample from the population?

All of the individual penguins in the entire world

Sample of penguins in the population

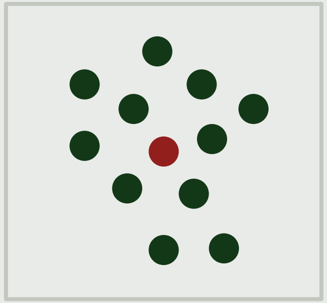
Sample Variance $$\sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

What is the spread or dispersion of wingspans within the sample from the population?

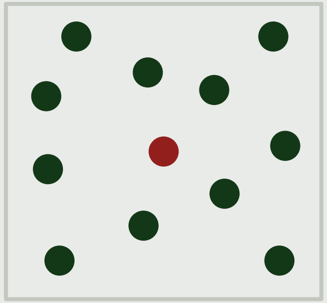
How is this calculation different?

We use $n - 1$ (Bessel’s Correction) instead of n to correct for bias in calculations as the sample mean tends to underestimate the true variability

Precision and accuracy measure observational error in a dataset



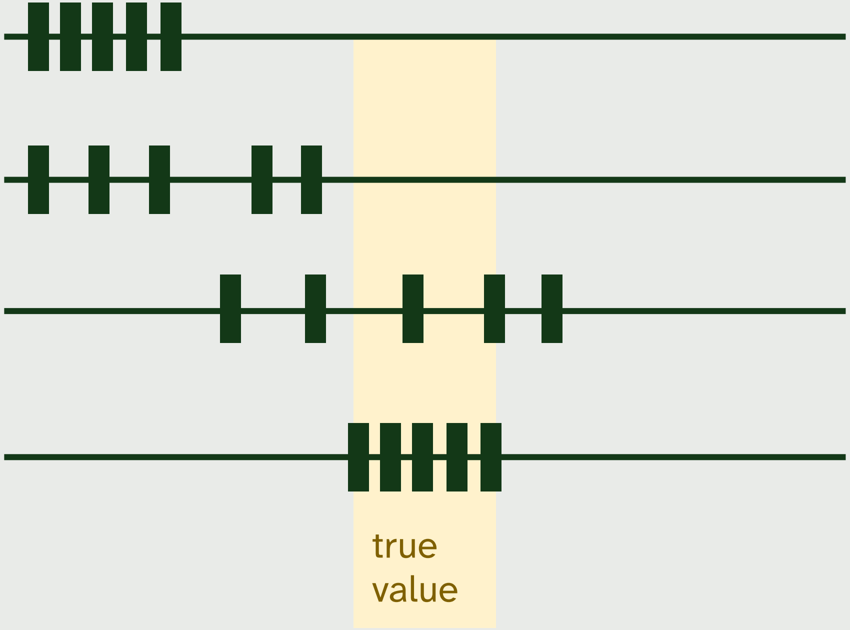
Dataset A



Dataset B

Precision evaluates how close measurements are to each other

Accuracy evaluates how close measurements are to the **true value**



precise, inaccurate

imprecise, inaccurate

imprecise, (*more*) accurate

precise, accurate

true value

The Palmer Penguin Dataset

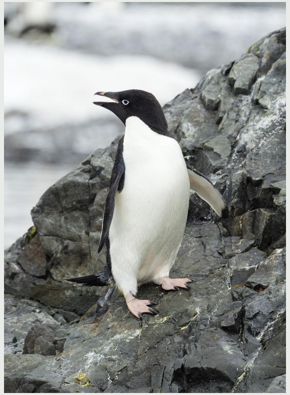
Kristen Gorman, Ph.D.
University of Alaska, Fairbanks
Assistant Professor of Marine Biology

Antarctic Penguin Data collected by Kristen Gorman, Ph.D. and the Palmer Station of the Antarctica Long Term Ecological Research Network

Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. https://doi.org/10.1371/journal.pone.0090081

Contains **individual penguin measurements** for three types of penguins

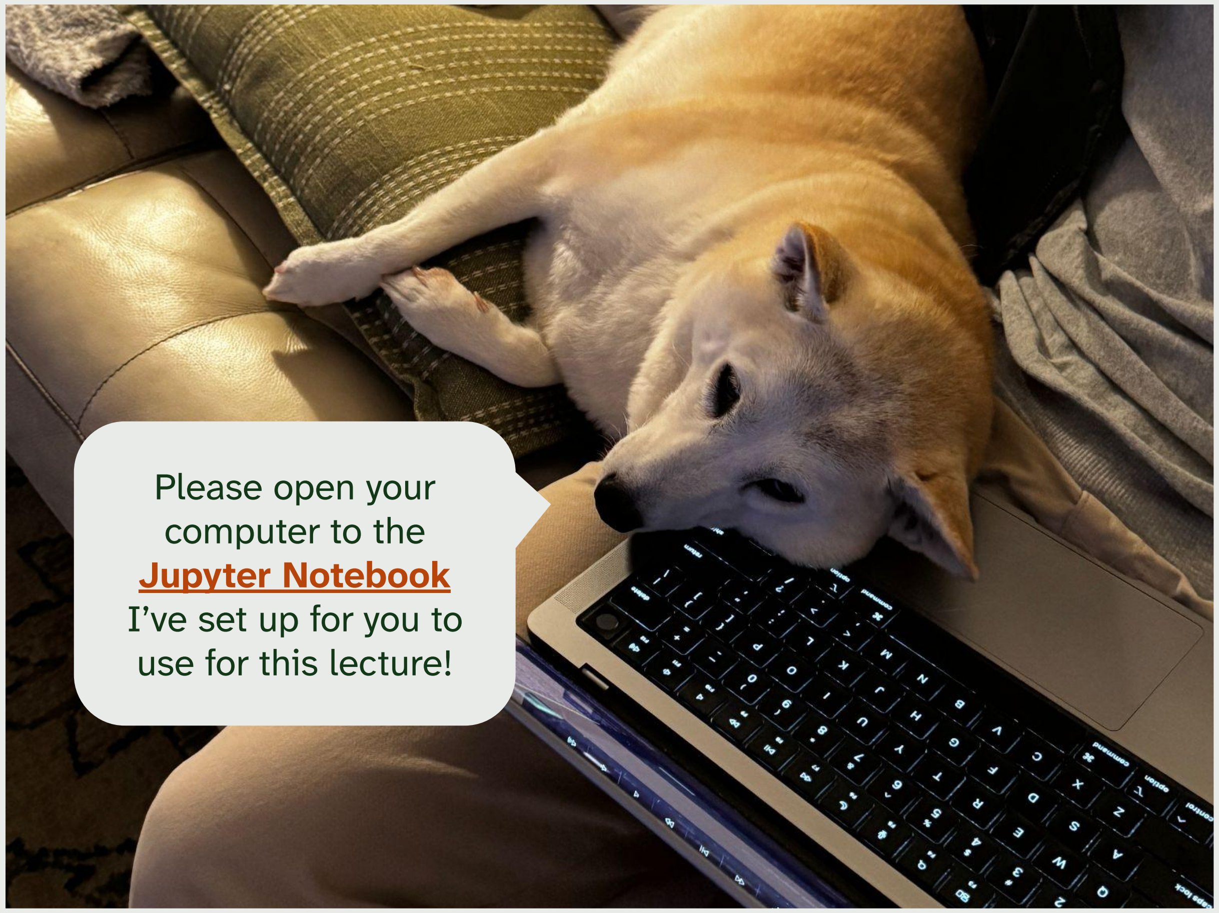
- Culmen (i.e., Penguin Beak) Length and Depth (mm)
- Flipper Length (mm)
- Body Mass (g)
- Island Name in the Palmer Archipelago of Antarctica (e.g., Dream, Torgersen, or Biscoe)
- Sex (e.g., Male or Female)



Adélie Penguin

Gentoo Penguin

Chinstrap Penguin



Please open your computer to the **Jupyter Notebook** I've set up for you to use for this lecture!

Interactive Exercise

In this exercise, we'll be using Python to calculate Descriptive Statistics on the Palmer Penguin dataset and perform some Data Visualization

Jupyter Notebook

You can open the Day 1 “*Interactive Lecture Notes*” .ipynb file if you’re interested in the *verbose* version of this exercise!

Leaving Interactive Exercise

Do you have any questions about what we covered in the **Jupyter Notebook**?
kashima
味自慢

Thank you, and see you again for Day 2!

!!

Module Split

Introduction to Statistics (Day 2/2)

Michelle Franc Ragsac, Ph.D. (*She/Hers/Siya*)
Postdoctoral Research Scholar — Amariuta Lab
Halicioglu Data Science Institute
mragsac@ucsd.edu

January 15th, 2026

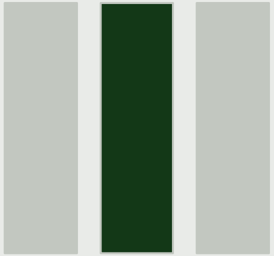
What are the topics of today's lecture?

1. Random Variables
   a. Discrete versus Continuous Variables
2. Common Probability Distributions**
   a. The Bernoulli and Binomial Distributions
   c. The Gaussian (*Normal*) Distribution
3. Central Limit Theorem
4. Confidence Intervals
5. Hypothesis Testing**
   a. T-tests and Z-tests

**NOTE

In today's lecture, we'll be hopping back and forth between the slides and the interactive Jupyter Notebook!

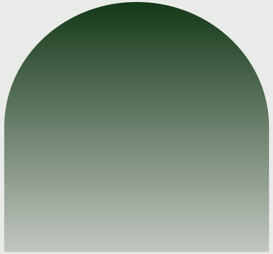
Random variables describe the outcome of random events



Discrete Random Variables can take on a countable number of discrete values

e.g., the number of times a coin lands on tails after being flipped 20 times

They are described by Probability Mass Functions (PMFs)



Continuous Random Variables can take on an infinite number of possible values

e.g., the height of various individuals within a particular clinical cohort

They are described by Probability Density Functions (PDFs)

$$X \sim Norm\left(\mu = 1, \sigma = 3\right)$$

Random Variables are named with **uppercase letters** (e.g., X)

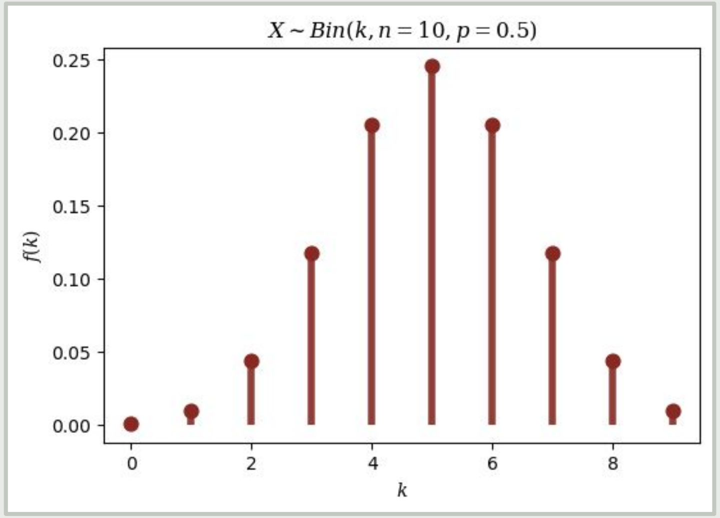
Random Variables have **associated distributions** that describe the probability of different event outcomes (e.g., $\mathrm{X} \sim \mathrm{Distr}$)

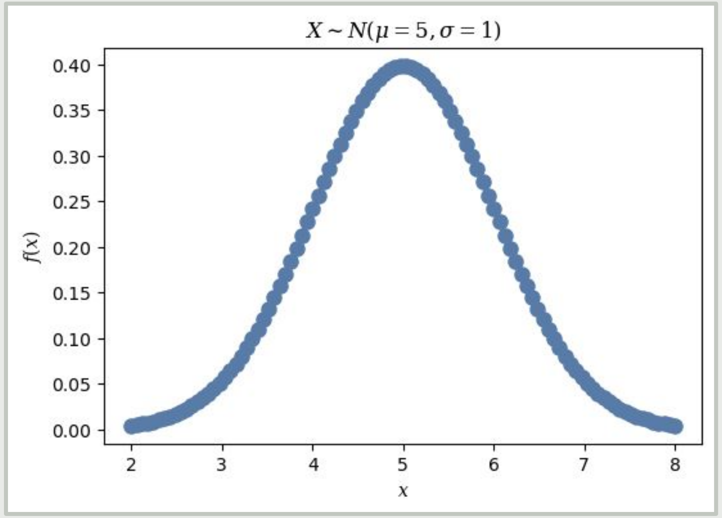
Observations of a Random Variable's Events are typically indicated with **lowercase letters** (e.g., x or x_i)

Probability distribution functions describe discrete and continuous random variables

Probability Mass Functions (PMFs)
for discrete random variables

Probability Density Functions (PDFs)
for continuous random variables





What distributions do we see in biological data?

Which distribution?	What does this distribution model?	What is an example in biology?
Binomial	Number of **successes** in a **fixed number of independent trials** where each trial has the same probability of success	Number of cells that express a particular marker out of a fixed number of cells
Poisson	Number of **times an event occurs** in a **fixed interval of time** when events happen independently at a constant average rate	Number of rare mutation events occurring in a defined stretch of DNA
Exponential	**Waiting time until the next event occurs** when events happen continuously and independently at a constant rate	Waiting time until a radioactive decay event happens, assuming a constant rate
Log-normal	Describes positive-valued, right-skewed data where the logarithm of a **variable follows the *Gaussian* distribution**	Gene Expression Levels; Biomarker Abundances
Gaussian *(Normal)*	A **symmetric, bell-shaped distribution centered at the mean** describing natural variation	Height; Blood Pressure

The binomial distribution describes the probability of success in Bernoulli trials

A **Bernoulli trial** is a random experiment with *two* possible outcomes (e.g., heads or tails)

$$\Pr\left(X=1\right)=p$$
$$\Pr\left(X=0\right)=1-p$$

Bernoulli PMF

$$f\left(k;p\right)=\begin{cases}p & \text{if } k=1\\ 1-p & \text{if } k=0\end{cases}$$
$$=p^k(1-p)^{1-k} \text{ for } k\in\{0,1\}$$

The probability of k successes from n Bernoulli trials → **Binomial Distribution**

Binomial PMF

$$f\left(k,n,p\right)=\binom{n}{k}p^k(1-p)^{n-k}$$

The Gaussian (*Normal*) distribution is a continuous probability distribution

The **Normal Distribution** is also called the **Gaussian distribution** after German mathematician Carl Friedrich Gauss (1777–1855)

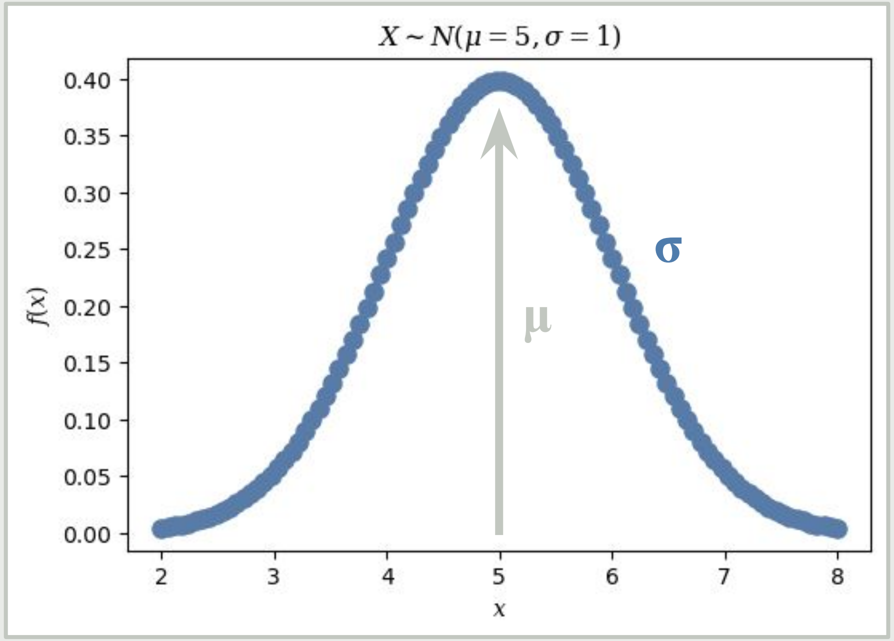
Normal PDF

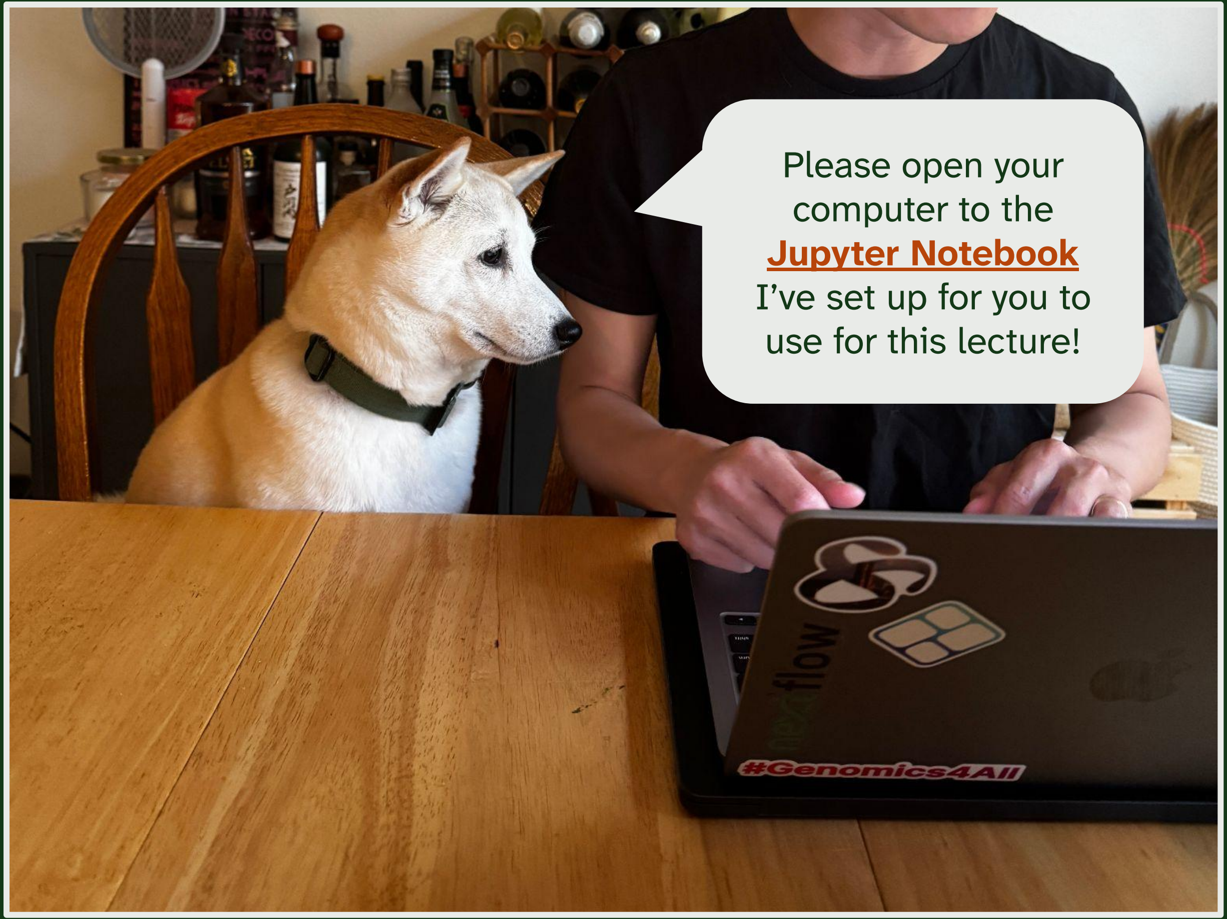
$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Standard Normal Distribution

Simplest case of the Normal distribution where $\mu = 0$ and $\sigma^2 = 1$

$$\varphi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$





Please open your computer to the **Jupyter Notebook** I’ve set up for you to use for this lecture!

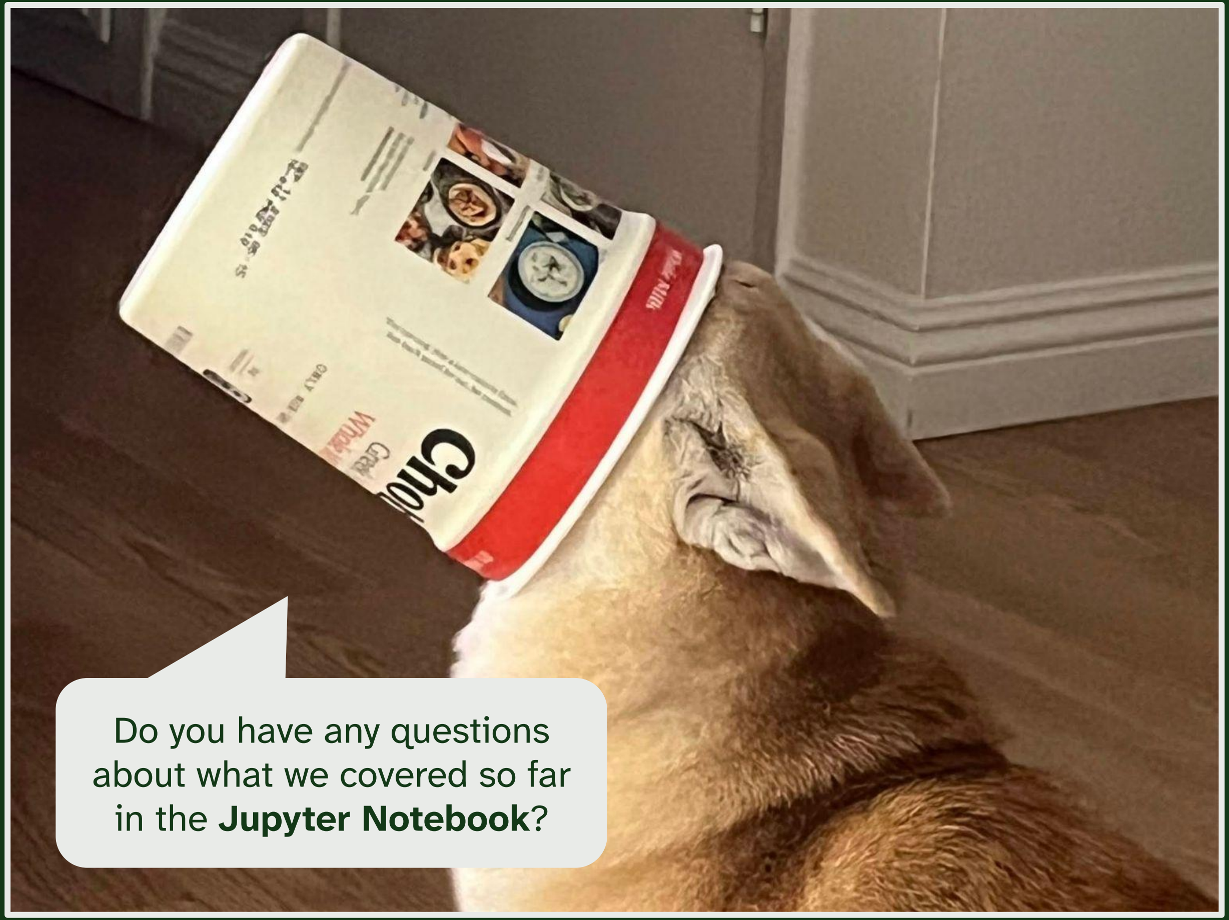
Interactive Exercise

For this portion of the exercise, we’ll be using Python to model and visualize the Binomial Distribution

Jupyter Notebook

You can open the Day 2 “*Interactive Lecture Notes*” .ipynb file if you’re interested in the *verbose* version of this exercise!

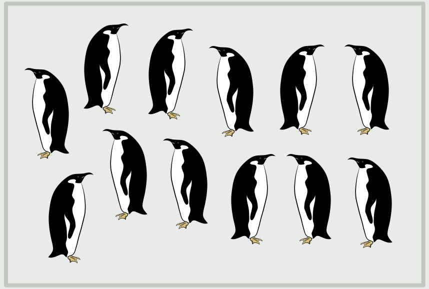
Taking a short break from the Interactive Exercise with some slides …



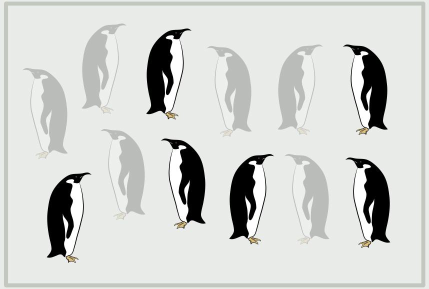
Do you have any questions about what we covered so far in the **Jupyter Notebook**?

The central limit theorem and sample means

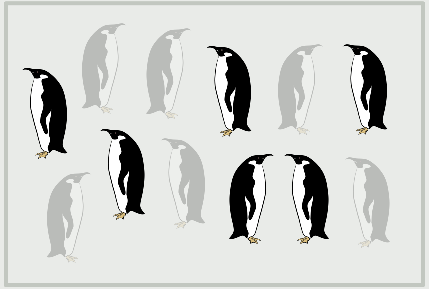
We want to take 3 random samples (X_1, X_2, X_3) of size $n = 6$ from a population *repeatedly*, then compute the sample mean ($\bar{X}$) within *each* random sampling



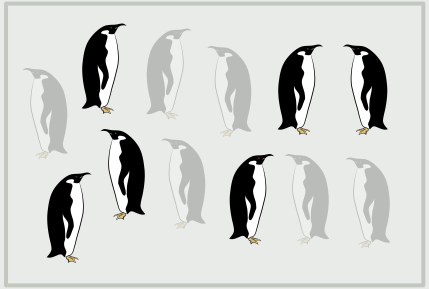
Population



Population – Sampling $i = 1$



Population – Sampling $i = 2$



Population – Sampling $i = 3$

Applies even if the starting population *does not follow* a Normal Distribution!

The Central Limit Theorem

If the sample size n is "*sufficiently large*", the sample mean will follow an **approximately Normal Distribution**

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

standard deviation
= standard error!

Standard Error $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

What is the variability of my statistic?
What is our uncertainty due to sampling?

Why is the central limit theorem useful?

We often *don't know* the true underlying distribution of our population *and* we only have a single sample.

➡

The **Central Limit Theorem** tells us what the **sampling distribution of the estimator** looks like!

01

Confidence Intervals

> “ How far might my sample estimate be from the true population parameter, just due to random sampling?

02

Hypothesis Testing

> “ Is the pattern we observe in our sample explained by random sampling variation alone?

03

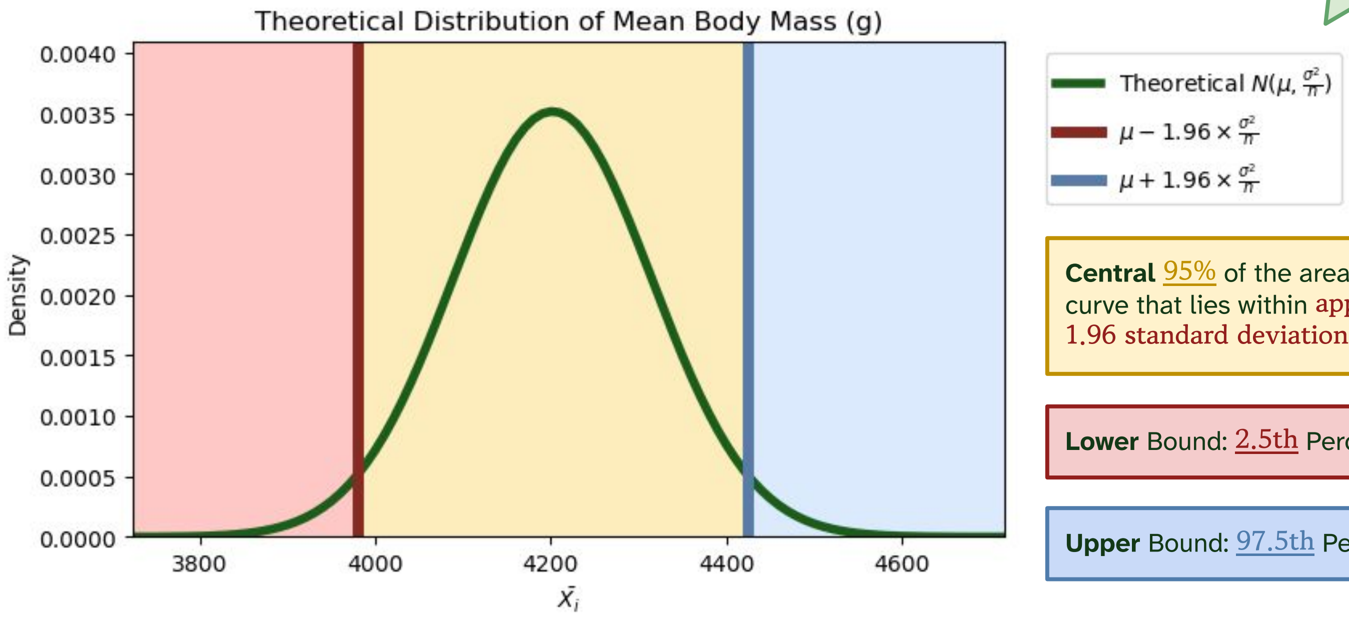
Measurement Precision

> “ If we repeatedly measure something many times under the same conditions, how much would our results vary?

Confidence intervals measure uncertainty around estimates

Relation to $p < 0.05$ Convention

Values *outside* the 95% confidence interval have $p < 0.05$, meaning they lie *in the most extreme* 2.5% on either tail



Central 95% of the area under the curve that lies within approximately 1.96 standard deviations of the mean

These are values that are consistent with the data!

Lower Bound: 2.5th Percentile

Upper Bound: 97.5th Percentile

These values are inconsistent and too extreme to be plausible under the **null hypothesis**!

Understanding the null (H_0) and alternative (H_A) hypotheses

Null Hypothesis, H_0

Assumption for evaluating how surprising the observed data is

Specific claim about a population parameter that has no effect, difference, or association

Alternative Hypothesis, H_A

Supported if observed data is inconsistent with the null hypothesis

Competing claim to the null that represents a meaningful effect, difference, or association

Hypothesis Testing Procedure

1. Compute a **test statistic** (e.g., sample mean)
2. Compute a p-value
3. Set an arbitrary **significance threshold** (typically $\alpha = 0.05$) **for the** p-value

   "The observed statistic is expected to occur less than 5% of the time under the null hypothesis"
4. If p-value < α, reject the null hypothesis
   If p-value > α, fail to reject the null hypothesis

Parameters for hypothesis testing

One-sample, **Two-sample**, and **Paired** tests describe the data structure and how variability is computed in our samples

One-sided and **Two-sided** tests determine the direction our evidence needs to follow to determine extreme evidence against the null

One-sample test

A single sample from one population is obtained,
then the sample statistic is compared to a known value

H_0 : Adelie penguins have the same body mass as the Palmer Station average
H_A : Adelie penguins have different body masses as the Palmer Station average

Two-sample test

Samples from two *different* populations are obtained,
then the sample statistic is compared *between* the samples

H_0 : Adelie and Chinstrap penguins have the same body mass
H_A : Adelie and Chinstrap penguins have different body masses

Paired test

Matched samples from two different populations are obtained
then the sample statistic compares the *difference* between matched samples

H_0 : Adelie penguins have the same body mass in winter and summer
H_A : Adelie penguins have different body masses in winter and summer

One-sided test

The alternative hypothesis evaluates if the mean is greater or less than some value

Two-sided test

The alternative hypothesis evaluates if the mean is different than some value

The importance of selecting a proper test statistic for hypothesis testing

z-statistic

Compares the Sample Mean versus the Population Mean

Relies on **independent observations** with **random sampling**

Assumes the data follows the **Normal Distribution** (or **C.L.T**)

Typically used with **large sample sizes** or when σ is known

t-statistic

Compares the Sample Mean versus the Population Mean

Relies on **independent observations** with **random sampling**

Assumes the data follows an **Approximately Normal Distribution** (or CLT)

Typically used with **small sample sizes** or when σ is not known (but can be estimated)

F-statistic

Evaluates the Ratio of Variances (i.e., Compares Multiple Means)

Relies on **independent observations**

Assumes *error* in the data follows a **Normal Distribution** and **Equal Variances** (i.e., homoscedasticity)

Typically used when comparing **≥2 sample means** (i.e., multiple groups)

X^2 statistic

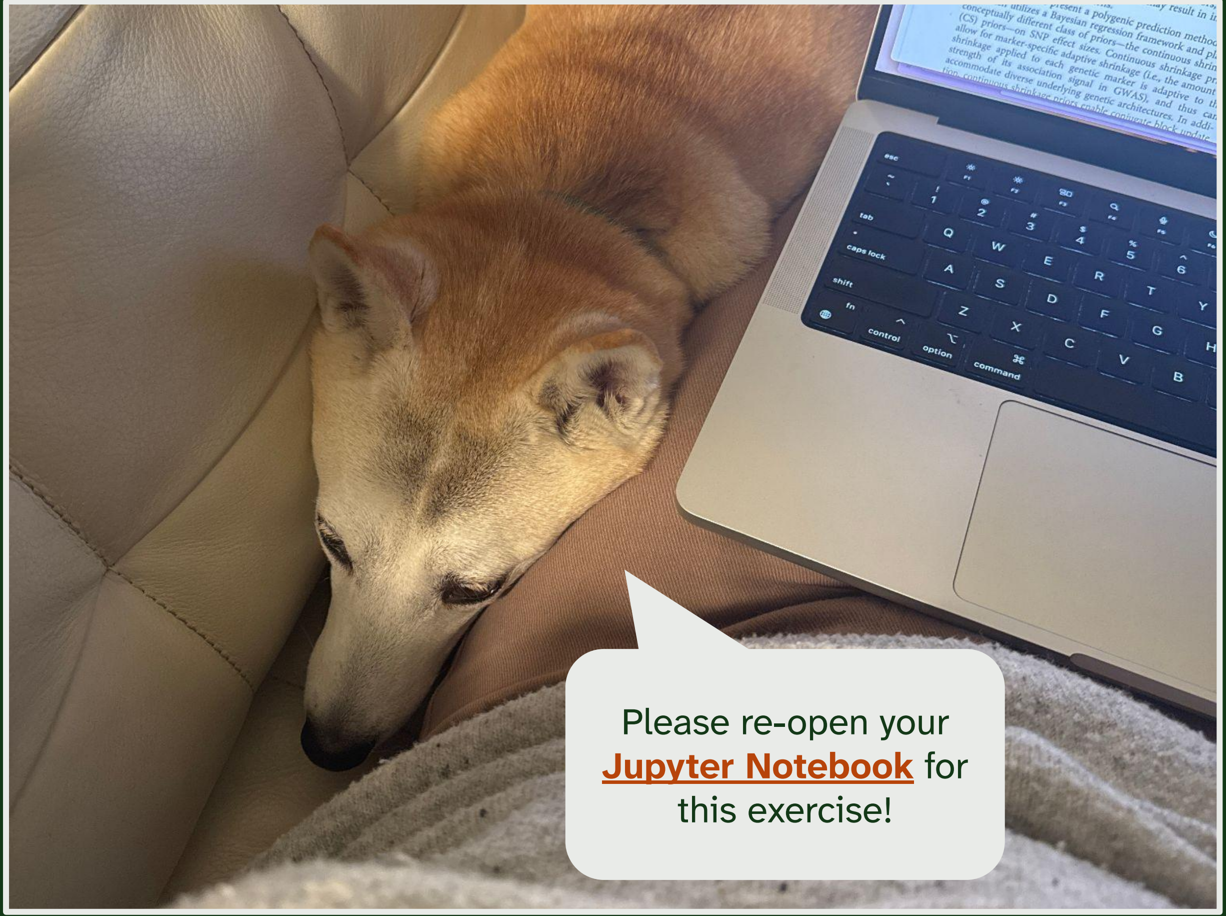
Compares Observed versus Expected Counts in the Data

Relies on **independent observations** and **count-based data that are sufficiently large**

Does *not* assume the data follows a Normal Distribution nor that the data has Equal Variances

Typically used when evaluating **categorical outcomes**

The **z-statistic** can be quite strict, so the **t-statistic** (i.e., **Student's t-Test**) is used more in practice!



Please re-open your **Jupyter Notebook** for this exercise!

Jumping back into the

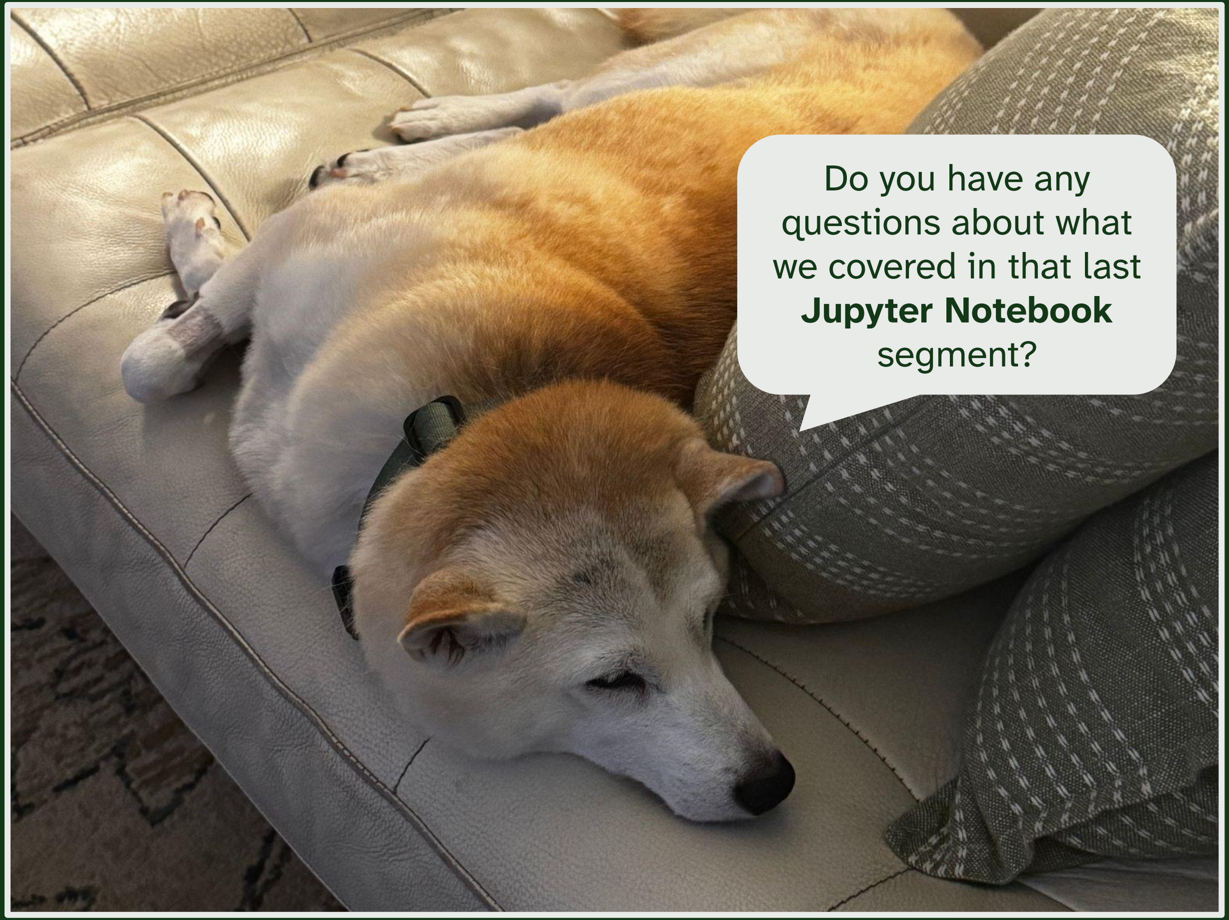
Interactive Exercise

In this segment, we'll be using Python to study the Central Limit Theorem and perform Hypothesis Testing

Jupyter Notebook

Re-open the Day 2 “*Interactive Lecture Notes*” .ipynb file from before!

Leaving Interactive Exercise


Do you have any questions about what we covered in that last **Jupyter Notebook** segment?

Thank you, and see you later in the Genome-Wide Association Studies (GWAS) module!

Frequently Asked Questions, Part I

Based on things Michelle found confusing with (*still*) learning statistics

What is the difference between statistics and probability?

Probability predicts outcomes *assuming we know the biology* about a process (e.g., a known mutation increases disease risk by 20%), whereas statistics works *backwards*. In this case, statistics uses patient data and information to predict risk.

Why is randomness important?

Randomness is important for reducing *bias*. For example, assigning samples in a pharmaceutical randomly helps assure that the differences observed are due to the drug being tested and not study group differences (e.g., age, sex, genetics, etc.)

What does a p-value actually represent?

A p-value is the probability of observing data at least this extreme *if the null hypothesis were true*.

For example, if a p-value of 0.03 is observed in a pharmaceutical study, it means that *if the drug truly had no effect*, then you would observe results that extreme only 3% of the time due to chance alone.

Does a small p-value mean the effect is important?

No. This is because a small p-value tells you that the *data being analyzed is inconsistent with the null hypothesis*, not that the difference between two groups is large, meaningful, or clinically useful.

A p-value *does not* tell you anything about how large the effect size is between groups, whether the effect matters clinically, or whether the effect will replicate in a different study group.

For example, as sample size increases, the standard error shrinks, which has the potential to make negligible effects statistically detectable!

There are different methods that can be used to study effect sizes, such as evaluating the standardized mean difference (Cohen's d) for continuous outcomes or evaluating odds ratios (ORs) for binary outcomes.

Frequently Asked Questions, Part II

Answers to more statistics-related questions and some random questions

Why are power calculations important?

Unfortunately, we were not able to cover power calculations in this module. *Power is the probability of correctly rejecting the null hypothesis when a meaningful effect truly exists.*

Power calculations are important because they help us avoid false negatives as low-powered studies often fail to detect any real effects! They also force you to design a proper study before seeing any data.

Does a 95% confidence interval mean there is a 95% chance the parameter is inside the interval?

Unfortunately, it does not. The 95% in a confidence interval describes the *long-run reliability of the method*, not the chance that the true value lies in a single computed interval.

Think of a confidence interval as a net designed to catch a fixed fish 95% of the time; once the net is cast, the fish is either inside or outside.

Whose dog is in your transition slides?

The photos are of my shiba inu, *Yuuki*. She is turning 9 years old this year!

When I TA'd for the course in 2020 and 2021, I used to bring her to campus or she would cameo in Zoom lectures during the COVID-19 shutdown, so I thought I would bring her back as an Easter Egg for these slides :-)

What software did you use to create this slide deck?

I used *Google Slides* with one of the free Templates available.

What is your research area in the Amariuta Lab?

I study *pediatric asthma* in Admixed American populations (e.g., Latino American). For my postdoctoral research, I am interested in developing a *multi-modal genetic risk model* for pediatric asthma to allow for prediction of asthma status and severity earlier than five years of age. Currently, diagnosis of pediatric asthma earlier than this age is difficult due to limitations in clinical diagnosis paradigms.

Helpful Resources

If you want more reading materials, here are some helpful websites to browse!

STAT 414: Introduction to Probability Course
from Pennsylvania State University
https://online.stat.psu.edu/stat414/
The notes for this course are well-written and helpful for explaining all of the concepts that we went over during this week's module.

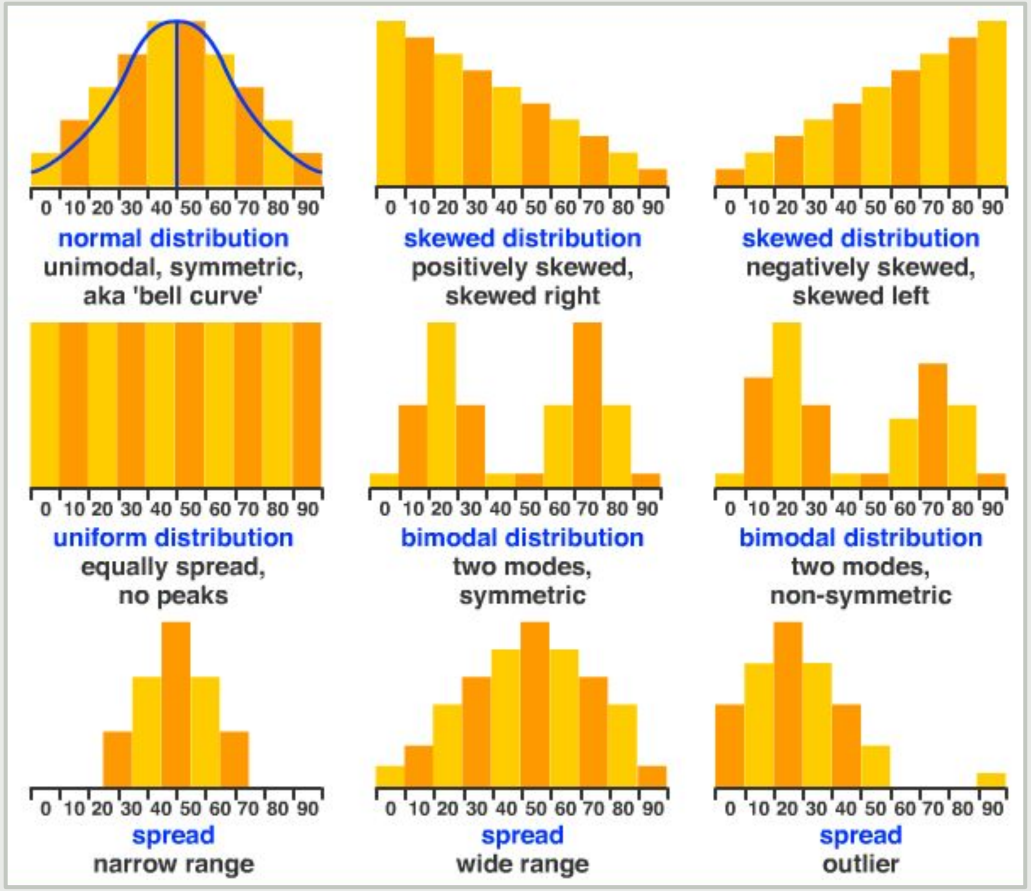
MATH 283: Statistical Methods in Bioinformatics Course
from University of California, San Diego
https://mathweb.ucsd.edu/~gptesler/283/calendar.html
This is one of the required courses for the Bioinformatics & Systems Biology Ph.D. program taught by Prof. Glenn Tesler. His lecture slides still serve as a reference for me today!

Statistics Resources from the LibreTexts Consortium
https://stats.libretexts.org/
LibreTexts is a multi-institution initiative to provide open, accessible textbooks for postsecondary education. They have multiple resources for statistics on their website!

Statistics Textbooks from the OpenStax Initiative
https://openstax.org/subjects/math#Statistics
OpenStax is a part of Rice University and it is an educational initiative to provide open, college-level educational resources that are peer-reviewed. They have several textbooks for statistics from their mathematics section.

Learn Statistics with Python Course from CodeAcademy
https://www.codecademy.com/learn/learn-statistics-with-python
If you're interested in more *applied* ways of learning statistics, I recommend this course from Codecademy! Their course goes over a lot of the same concepts we covered in our short module.

Introduction to Statistics Course from DataCamp
https://www.datacamp.com/courses/introduction-to-statistics
DataCamp is similar to Codecademy, but they tend to focus on fundamental concepts for Data Scientist jobs. They also have an introduction to statistics course in Python!



Different shapes of distributions

The figure here was taken from [Jenny Eather's mathematics website]

Information on error bars in scientific papers

Unfortunately, the meaning of error bars is **not consistent across papers** →
Read the figure caption to see how to interpret error bars in a particular manuscript!

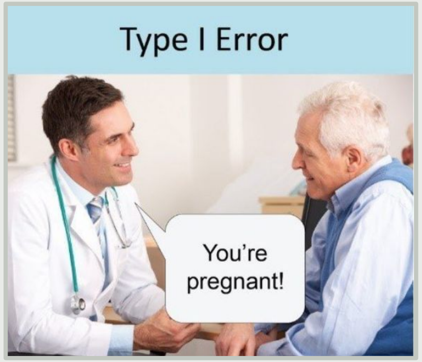
Common Uses of Error Bars in Figures

- Sample Mean ± 1 Standard Deviation
- Sample Mean ± 2 Standard Deviation
- 1 Standard Error of the Mean (S.E.M.)
- 2 Standard Error of the Mean (S.E.M.)
- 1.96 Standard Error of the Mean (S.E.M.) → 95% Confidence Interval

Differences between type I and type II error

When performing a hypothesis test, there is a chance that you either:

a. **incorrectly reject the null hypothesis** (type I error, or false positive)
b. **incorrectly fail *to* reject the null hypothesis** (type II error, or false negative)



In the context of sampling, type I errors represent cases where the random sample has ***overestimated*** the effect being studied by chance!

If $\alpha = 0.05$ and the null hypothesis is true, then there is a 5% chance that the test statistic will reject the null hypothesis incorrectly.

When your p-value is greater than your significance level (e.g., p-value > 0.05), there is a chance of type II error

- type II error can be caused by many factors, including small effect sizes, small sample sizes, or data variability
- You can estimate type II error rate (β) *before starting a study* with **power analyses**