Why Care About Statistics

- Core Component of Genomic Data Science
    - Genomic data science is built on three pillars: biology, computer science, and statistics.
    - Statistics is often overlooked or seen as the "third wheel".
- High-Profile Genomic Study Gone Wrong
    - A widely publicized study in *Nature Medicine* claimed genomic data could predict effective chemotherapy treatments.
    - The result was considered a breakthrough in personalized cancer therapy.
    - Statisticians Keith Baggerly and Kevin Coombes at MD Anderson attempted to reproduce the study. They found serious flaws in the analysis and couldn't replicate key results due to data and method issues.
    - Despite flawed statistics, clinical trials were launched based on the faulty findings. Patients were assigned chemotherapy treatments based on incorrect data analysis. This led to lawsuits and a major scandal involving Duke University.
    - The controversy prompted an Institute of Medicine report outlining new standards for genomic data science emphasizing reproducibility, transparency, and rigorous statistical modeling.

What Went Wrong

- Lack of Transparency
    - The data and code were not made available after publication.
    - Others trying to replicate the work couldn't access the raw or processed data or the statistical code.
    - This violated principles of reproducibility—others couldn't verify or reproduce the results.
    - The lead researchers were uncooperative, refusing to share data/code, further delaying detection of errors.
- Lack of Statistical Expertise
    - Incorrect and "silly" prediction rules were used (ex: formula with a random "-¼").
    - The models showed fundamental misunderstandings of probability and prediction.

    - Indicates those designing the statistical model lacked formal training in statistics.
- Poor Study Design
    - Samples were processed on different days, introducing batch effects (**batrifacts**).
    - This created confounding variables; technical artifacts mimicked or masked real biological effects.
    - The flawed design meant the study was set up to fail from the start.
- Unstable Predictions
    - Prediction models produced different outcomes on different days—even with the same data and code.
    - Randomness wasn't due to updated data or models, but to unstable and improperly managed statistical procedures. This made the model unsuitable for clinical use, especially for assigning treatments.
- Ultimately Reproducible—but Wrong
    - Statisticians at MD Anderson reconstructed the full analysis with the original data and code.
    - The analysis was technically reproducible, but the statistical methods were flawed. The core issue wasn't hidden data, but bad analysis from the beginning.

Central Dogma of Statistics

- Foundational Idea in Statistical Inference: understanding populations through sampling and inference.
- Goal: Learn About a Population Without Measuring All of It
    - Populations are often too large, costly, or impractical to measure in full.
    - Instead, we use probability to take a representative sample.
- Sampling as the Bridge Between Population and Analysis
    - A randomized sample allows us to gather data from a small subset.
    - This sample must represent the population well for the inference to be meaningful.
- Statistical Inference: Making Educated Guesses
    - We use inference to estimate characteristics of the population (e.g., proportions, averages) from the sample data.

    - Example: If 2 out of 3 sampled items are pink, we might infer the population is mostly pink.
- Accounting for Variability
    - Since the sample is only a subset, our estimates are uncertain.
    - We must quantify variability—understanding how much our estimate might differ from the true value.
- Importance of Knowing the Population
    - Inference is only valid if the population is clearly defined and unchanging.
    - If the population changes after sampling, the sample may no longer be representative.
- Real-World Example: Google Flu Trends
    - Initially effective, but predictions failed as search behaviors changed over time.
    - The population of interest shifted, breaking the link between sample and target.
- Summary of the Central Dogma
    - Start with a population → Use probability to draw a sample → Apply statistical inference to estimate population characteristics → Account for uncertainty and variability in those estimates.

Data Sharing Plans

- Why Data Sharing Matters
    - Lack of data sharing was a major issue in the original case study.
    - Transparent data sharing ensures reproducibility, verification, and collaboration in genomic research.
- Four Essential Components of a Complete Shared Dataset
    - Raw Data: unprocessed, original measurements (e.g., FASTQ, BAM files). What you receive in its "rawest available form"; no computation, summarization, or filtering applied.
    - Tidy Data: cleaned and structured for analysis.
        - One variable per column, one observation per row, one table per data type, use of linking IDs for multiple datasets (e.g., phenotype and genomic data)

    - Code Book: describes variables, units, and definitions in the tidy dataset. Critical for interpreting measurements (e.g., height in meters vs. feet) and preventing misunderstandings that can lead to major errors.
    - Recipe (Reproducible Processing Instructions): Details how raw data was transformed into tidy data.
        - Best done via scripts (e.g., R, Python) that can run without manual input.
        - If not scripted, must include a step-by-step written protocol:
            - List all software, parameters, versions, manual actions (e.g., Excel steps); avoid vague descriptions or skipped details; include recordings or screenshots if needed for GUI actions

- Resources
    - A pre-made data sharing plan template is available on GitHub for public use.
    - It expands on these principles with practical guidance for real-world projects.

Plotting Your Data

- Interactive Analysis
    - Most genomic data analysis should be done interactively, primarily through visualization.
    - Use plots to make large datasets understandable by summarizing them visually.
- "Make Big Data Small"
    - Summarize complex, massive datasets into a form that can be quickly visualized and interpreted.
    - Plotting helps identify patterns, problems, or features not obvious from raw data or summary statistics.
- Summary Statistics Can Be Deceptive
    - Example: Four datasets with identical statistics (slope, correlation, p-value) can have very different patterns when plotted.
    - Visualizing the data reveals outliers, nonlinear relationships, or anomalies that summaries hide.
- Show the Raw Data in Plots

    - Avoid basic bar plots with error bars only. Prefer plots that also show raw data points (better plots) to illustrate distribution and sample size. Raw data gives better context and prevents misleading interpretations.
- Plotting Replicates
    - Common in genomics to compare technical replicates (same sample, repeated runs).
    - Plot replicate 1 vs replicate 2 to assess correlation and reproducibility.
    - Be cautious of scale issues: 99% of the data might be clustered in a tiny region.
- Use Data Transforms (e.g., Log Transform)
    - Helps spread out tightly clustered data and makes patterns more visible.
    - Especially useful when plotting counts or expression levels.
- MA Plot (Bland-Altman) for Replicates
    - X-axis: Mean of two replicates; Y-axis: Difference between them.
    - Reveals where the replicates disagree and whether that disagreement depends on signal strength.
    - Commonly used in genomics; highlights variability in low vs. high expression genes.
- Avoid "Ridiculograms"
    - Visually appealing but uninformative plots, often complex network visualizations.
    - May appear in prestigious journals but fail to communicate real scientific insight.
    - Aim for clarity: good plots should be both informative and attractive.

Sample Size and Variability

- Sample Size Estimation
    - Naïve method: sample size (N) = Total budget / Cost per measurement
    - Better method: Use variability and power analysis to determine required sample size.
- Variability and Mean Differences
    - When comparing two groups (e.g., X vs. Y), variability around the mean affects how confidently we can say the means differ.
- Statistical Power: probability of detecting a true effect.
    - Depends on sample size (N), effect size (difference between means), and variability

    - Typical desired power is 80%
- Example with Power Analysis
    - Small sample (N = 10) → Low power (~18%)
    - Required sample (N = 64 per group) → Power reaches 80%
    - One-sided tests can require fewer samples if direction of effect is known.
- Power Curves
    - Power increases with larger effect sizes and larger sample sizes
    - Power is not fixed, it's a curve depending on parameters.

Three Types of Variability in Genomic Data

1. **Phenotypic Variability**: differences between groups (e.g., cancer vs. control).
    - The variability of interest for hypothesis testing.
2. **Measurement Error**: Inaccuracy or noise from the technology/platform used.
    - Varies by method (e.g., arrays vs. sequencing).
3. **Natural Biological Variability**: Inherent differences between individuals with similar traits.
    - Present even among identical healthy individuals.

Technology and Variability

- Newer technologies may reduce measurement error, but biological variability cannot be eliminated by better measurement tools.
- Low-variability genes remain low, and high-variability genes remain high regardless of measurement method.

Statistical Significance

- Determine if observed differences between groups are likely to be *real* (replicable), not due to random chance.
- Example Setup: Gene expression values for two groups (e.g., control vs. treatment). Some genes show clear differences, some don't; visual inspection helps but isn't enough.
- The t-Statistic: measures the difference between group means scaled by variability.

- Formula: $\frac{\overline{Y} - \overline{X}}{\sqrt{\frac{s_Y^2}{N} + \frac{s_X^2}{M}}}$
- Larger t-values → stronger evidence of a difference.

- p-Value: probability of observing a t-statistic as extreme or more extreme *if the null hypothesis is true*.
  - Calculated via methods like permutation testing: randomly shuffle group labels, recalculate t-statistic for each shuffle, compare real t-stat to the distribution from permutations.
  - p-value < 0.05 is often called "statistically significant", but this threshold is arbitrary and historically chosen (by Fisher).
  - A p-value is NOT: The probability the null hypothesis is true, the probability the alternative hypothesis is true, or a direct measure of evidence.
  - p-value = probability of observing such a difference under the null model. Smaller p-value → stronger evidence against the null.
- Issues in Practice:
  - Over-reliance on p-values can mislead.
  - The arbitrary 0.05 cutoff often leads to false claims of significance.
  - Misuse contributes to reproducibility problems in scientific research

Multiple Testing

- Multiple Testing Problem
  - In genomic studies, thousands or millions of tests (e.g., gene expressions, DNA variants) are performed simultaneously.
  - Standard hypothesis testing with p-values wasn't designed for many simultaneous tests.
- Uniform Distribution of P-values Under Null
  - If there is no real effect, p-values are uniformly distributed between 0 and 1.
  - About 5% of tests will have p-values < 0.05 by chance alone, even with no true effects.
- Example (Jelly Beans and Acne): Testing many jelly bean colors separately for association with acne leads to some false positives by chance. Reporting only the "significant" result (e.g., green jelly beans) is misleading due to multiple testing.

- Error Rates for Multiple Testing
    - **Family Wise Error Rate** (FWER): Probability of having *at least one* false positive among all tests. Very strict control.
    - **False Discovery Rate** (FDR): Expected proportion of false positives among the declared significant results. More liberal, allowing some false positives to increase power.

Family wise error rate:

$$\Pr(\#\text{ False Positives} \geq 1)$$

False discovery rate:

$$E\left[\frac{\#\text{ False positives}}{\#\text{ Total Discoveries}}\right]$$

- Interpretations Differ:
    - FWER control means very low chance of any false positive, but may miss true effects.
    - FDR control allows some false positives but provides a manageable error rate across discoveries.
- Example: 10,000 Genes
    - With p-value threshold 0.05 → expect ~500 false positives out of 550 significant results (mostly false).
    - Using FDR control → expect about 27.5 false positives among 550 discoveries (much lower error).
    - Using FWER control → very few (if any) false positives, but fewer discoveries.
- Avoiding False Conclusions
    - Don't "chase" p-values just below 0.05 by re-testing or changing analysis (p-hacking).
    - Specify analysis plans *before* looking at data to avoid bias.
    - Reporting negative or non-significant results is important to prevent publication bias.
    - Altering data analysis after seeing results can artificially create "significant" findings.

Study Design, Confounding, and Batch Effects

- Confounding: a variable related to both the independent variable and the outcome, potentially creating a false association.
    - *Example*: Shoe size appears associated with literacy, but age is the true confounder.
- Batch Effects: technical or procedural variables (ex: date of sample collection, assay changes, equipment) that correlate with biological groups can confound results.
    - *Example 1*: Gene expression differences between ethnic groups disappeared once adjusted for batch (collection year).

    - *Example 2*: Genetic studies and proteomic studies have shown false associations due to batch confounding (different technologies or sampling times used for different groups).
    - Common, affecting nearly all genomic technologies.
- Without accounting for confounders or batch effects, studies can report false associations that are actually artifacts.

How to Address Confounders and Batch Effects

- Randomization: Assign treatments or conditions randomly to break association between confounders and treatment groups.
    - Randomization helps ensure confounders are evenly distributed across groups, reducing bias.
- Stratification: Design experiments to balance known confounders across groups.
    - *Example*: In a mouse study with males and females treated over two weeks, ensure both sexes and both treatments are distributed across both weeks to avoid confounding by sex or date.
- Balanced Designs: Equal numbers in treatment and control groups help reduce confounding and improve statistical power.
- Replication

    Include technical replicates (multiple measurements of the same sample) to assess technology reliability.
    - Include biological replicates (samples from different individuals) to assess population variability.
- Controls: Use positive controls (known to produce effect) and negative controls (known to produce no effect) to verify experiment and technology validity.

Quiz

1. Which of the following are required for sharing a data set?

   A code book describing each variable and its values

   The raw data

   All of these options

   A tidy data set

2. Which of the following should be included in data tidying recipes?

   Sample size formulae

   Units of variables

   Explicit step-by-step instructions

   Preprocessed data

3. What is the central dogma of statistics?

   Using measurements on a probabilistically selected sample to infer knowledge about a population

   Using Bayes rule to calculate probabilities we care about

   That increased power comes with increased sample sizes

   Using measurements on a population to infer knowledge about a sample

4. Which of the following are types of variability in all genomic data?

   Measurement error

   Missing data variability

   Geographic variability

   Variation from changing technology

5. Which of the following will increase power in a statistical analysis?

   Increasing sample size

   Using a new technology

   Adjusting for confounders

   Increasing measurement variation

6. If 100 p-values are calculated on a data set with no signal, how many p-values would we expect to be less than 0.05 on average?

   20

   5

   50

   0.05

7. If we report 500 results as significant out of 10,000 tests while controlling the family-wise error rate at 5%, about how many false positives do we expect?

   0

   10

   200

   25

8. What is the most common confounder in genomics?

   Age

   Sex

   Genetic background

   Batch effects

9. Which of the following can be used to address potential confounders at the experimental design stage?

    Randomization

    Increasing sample size

    Using linear models

    Measuring DNA instead of RNA

10. Which of the following are benefits of making big data as small as possible as soon as possible?

    Reducing the data will reduce the number of hypothesis tests

    Smaller data sets are easier to share

    Reduced data will increase the power of statistical tests

    Smaller data sets will decrease false discovery rates