“Data don’t make any sense, we will have to resort to statistics.”
Upcoming deadlines

- URSA - April 29
- BLaST - May 31
Reading: Statistical Analysis of Data

Fast 5
Take out a paper and write at least 5 things you learned from the reading.
• Student’s t-test for comparing a mean value with a known, standard value
  • Degrees of freedom = N - 1

• Student’s t-test for comparing mean values from two different techniques
  • Degrees of freedom = N - 2

• Least squares analysis for predicting a concentration value for an unknown sample
1. Idea
2. Research
3. Hypothesis
4. Plan
5. Experiment
6. Analysis
7. Reflection
8. Sharing
Why Statistics?

How to quantify variability

From sample to population

The Gaussian Distribution

P Values

Significance

Confidence Intervals

T-Tests

Statistical Advice
Outline

Why Statistics?

From sample to population

How to quantify variability

The Gaussian Distribution

P Values

Significance

Confidence Intervals

T-Tests

Statistical Advice
Why Statistics?

• We tend to jump to conclusion.
  o A 3-year old girl told her male buddy: “You can’t become a doctor; only girls can become doctors.” To her this made sense, because the three doctors she knew were all women.
  o The ability to generalize from a sample to a population is hard wired into our brains and has even been observed in 8-month old babies (Xu & Garcia, 2008).

  Scientists need statistical rigor to avoid overly strong conclusions from limited data!
Why Statistics?

• We tend to be over-confident.

  o Question: diameter of the moon in miles?
  o Russo and Schoemaker (1989) tested more than 1000 people and reported that 99% of them were overconfident!

Scientist need statistics to avoid informal guesstimates!
Why Statistics?

• We see patterns in random data.

```
- - X - X - X X X - - X X X - X X - X X - - X X - - X X
X - - X - X X - - X X - - X - X - X - - X X X - - X X - -
X X X - - X X - X - X - X X X - - - - X - X - X X X - - - - X
- - X - X - - X X - X X - - X X X - - - - X X - - X - X -
- - X - X - X X - - - - X - - - - - - X - X - X - - - - X - X
- - X X X - X - X - - - - X X X X - X X X X - - - - - - X X - X X
X - - X X - - X X X X - X X - - X - - X X X X - X X X - - -
X - X - - - X X X X - - X - X X - X X X X - X X - - - - X -
X X X - - X X X X - X - X - X X - X X X X - X X - X X X
- - - - X X X - - X X X X - X - - X - X X X X - X X - - - - X
```

Table 1.1. Random patterns don’t seem random.

Table 1.1 represents 10 basketball players (1 per row) shooting 30 baskets each. An “X” represents a successful shot, and a “-” represents a miss. Is this pattern random? Or does it show signs of nonrandom streaks? Most people tend to see patterns, but in fact the arrangement is entirely random. Each spot in the table had a 50% chance of having an “X.”

Statistical rigor is needed to avoid being fooled by apparent patterns among random data!
Why Statistics?

- We want to draw conclusions about a population from a smaller dataset.

Statistics help you extrapolate from a particular set of data (your sample) to make a more general conclusion (about the population).
Why Statistics?

- Scientists need statistical rigor to avoid overly strong conclusions from limited data!
- Scientists need statistics to avoid informal guesstimates!
- Statistical rigor is needed to avoid being fooled by apparent patterns among random data!
- Statistics help you extrapolate from a particular set of data (your sample) to make a more general conclusion (about the population).
Probability vs Statistics

Deductive

• From theory to observation/confirmation

Inductive

• From observation to theory

<table>
<thead>
<tr>
<th>PROBABILITY</th>
<th></th>
<th>STATISTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>→</td>
<td>Specific</td>
</tr>
<tr>
<td>Population</td>
<td>→</td>
<td>Sample</td>
</tr>
<tr>
<td>Model</td>
<td>→</td>
<td>Data</td>
</tr>
<tr>
<td>General</td>
<td>←</td>
<td>Specific</td>
</tr>
<tr>
<td>Population</td>
<td>←</td>
<td>Sample</td>
</tr>
<tr>
<td>Model</td>
<td>←</td>
<td>Data</td>
</tr>
</tbody>
</table>

Probability: Given the information in the pail, what is in your hand?

Statistics: Given the information in your hand, what is in the pail?
From Sample to Population

• Statistical calculations generalize from sample to population.

  Quality Control: a factory makes lots of items (the population), but randomly selects a few items to test (the sample).

  Political Polls: a random sample of voters (the sample) is polled, and the results are used to make conclusions about the entire population of voters.

  Clinical Studies: patients included in a study are representative of other similar patients and the extrapolation from sample to population is made.
From Sample to Population

• What statistical calculations cannot do?
  o The population we really care about is more diverse than the population from which our data is sampled.
  o The subjects in the study are often not randomly sampled from a larger population.
  o The measured variable is a proxy for another variable we really care about.
  o Scientific (or clinical) conclusions require our looking at multiple outcomes, not just one.
• n-OF-1 Trials

Beware of extrapolating statistical conclusions too far!
How to quantify variability?

- Standard Deviation
- Variance
- Coefficient of Variation
- Interquartile Range
How to quantify variability?

- Standard Deviation
- Variance
- Coefficient of Variation
- Interquartile Range
How to quantify variability?

• Standard Deviation vs Standard Error of the Mean (SEM)

$$SEM = \frac{SD}{\sqrt{n}}$$

SD

Quantifies scatter

SEM

How precisely we know the true mean of the population

http://www.alivelearn.net/?m=201008
Outline

Why Statistics?

From sample to population

How to quantify variability

The Gaussian Distribution

Confidence Intervals

P Values

Significance

T-Tests

Statistical Advice
Probability Distributions

- J-shaped
- Normal
- Rectangular
- Bimodal
- Positive (right) skew
- Negative (left) skew
The Gaussian Distribution

- Why the Gaussian distribution is so central to statistical theory?
The Gaussian Distribution

Figure 10.1. Ideal Gaussian distributions.
The horizontal axis plots various values, and the vertical axis plots their relative abundance. The area under the curve represents all values in the population. The fraction of that area within a range of values tells you how common those values are. (Left) About two thirds of the values are within 1 SD of the mean. (Right) Slightly more than 95% of the values are within 2 SD of the mean.
Confidence Intervals

How precisely we know the population mean.

Doesn’t tell us about the spread of the values.

Depends on the scatter among the values as well as the sample size.

\[ w = t^{*} \cdot SD/\sqrt{n} \]
Confidence Intervals

- Why 95% confidence?
  - CI can be computed for any degree of confident.
  - By convention, 95% CIs are presented most commonly.

\[ w = t^* \cdot \frac{SD}{\sqrt{n}} \]
Confidence Intervals

• Larger n, narrower CI

Figure 12.1. The 95% CI does not contain 95% of the values, especially when the sample size is large.
Confidence Intervals

- Is a 99% CI wider or narrower than a 90% CI?
Outline

Why Statistics?

From sample to population

How to quantify variability

The Gaussian Distribution

P Values

Significance

T-Tests

Confidence Intervals

Statistical Advice
P Values, T-tests and Significance

• The body temperature example
  o Mean temperature (n = 130) : 36.82°C
  o 95% CI : 36.75 to 36.89°C
  o The null hypothesis : the average normal body temperature is 37.0°C.
  o The discrepancy (Δ) : 37.0 – 36.82 = 0.18°C

• The p-value:

  If the population mean is truly 37.0°C, what is the chance that in a sample with n = 130 the absolute value of Δ is 0.18°C or larger?

  \( p = 0.0000018 \)
P Values, T-tests and Significance

• Calculating the p-value

  o Calculate the discrepancy between the observed sample mean and the hypothetical population mean ($\Delta$)

  o Calculate the $t$ ratio: Divide $\Delta$ by SEM to account for variability and sample size

  o Computer programs do the calculation to find the p-value corresponding to the calculated $t$ ratio with the given degrees of freedom
P Values, T-tests and Significance

- Statistically Significant
- Clinically Important
Outline

Why Statistics?

From sample to population

How to quantify variability

The Gaussian Distribution

P Values

Significance

Confidence Intervals

T-Tests

Statistical Advice
Statistical Advice!

- Don’t forget the big picture!
  - Garbage In, Garbage Out!
  - Correlation or association does not imply causation!

![Image of Garbage Can and Ice Cream Chart]
Statistical Advice!

• Interpret p-values wisely!
  • P < 0.05 is not sacred.
  • Statistically significant does not mean scientifically important.
  • Not statistically significant does not mean no difference.
  • Published p values tend to be optimistic.

![P-Value Interpretation Diagram](https://xkcd.com/1478/)
Statistical Advice!

• Beware of multiple comparisons.
  • If you compute enough p values, some are likely to be small.

If you torture your data long enough, they will tell you whatever you want to hear!  
James L. Mills (1993)
It is easy to lie with statistics. It is hard to tell the truth without it.

Andrejs Dunkels
Lab today (April 20)

Lab: Experiments

Due next week:
CHEM 294 Research Proposal
CHEM 294/694 Research Proposal Feedback