Stats Lecture – Week 2 - Monday

Announcements:

- Quiz – 9:30-9:45am Tuesdays – will cover major concepts from Ch 1-3.
- Reading this week: by Tuesday Ch. 1, by Wednesday Ch 2-3.
- Seminar assignments posted on the doors to seminar this afternoon.
- Lab 1 Modeling: If you have NOT handed in your lab, last chance is tomorrow 9:30!

Today: Overview of this week's reading, and then Lab Preview.

1. Why statistics?
    a. Are patterns we see due to chance alone?
    b. How likely are they?
    c. Example – the scientific process….
        i. Go into nature, observe and wonder…"why are there more ants in the forest?
        ii. Formulate a testable hypothesis,
        iii. devise an experiment to test that hypothesis
        iv. ???
        v. ???
        vi. ???
        vii. ???

2. Probability & uncertainty – it's everywhere. …
    a. 40% chance of rain
    b. 100 year flood, e.g., usgs.gov
    c. Chance of winning, e.g., 1 in 4.5

3. Precision vs. accuracy (in stats)
    a. Precision: Agreement among many measurements, e.g., on same instrument.
        i. Contrast to "precision" for computer scientists….
    b. Accuracy: how close is the value to truth?

4. Terms
    a. Event – insect visits to a plant
    b. Outcome – 2 possibilities (prey capture, prey escape)
    c. Trial
    d. Probability -- #outcomes / # trials – 0.0, 1.0, 0.5
    e. Sample – can't measure all plants in a population. vs census (all)

5. Axiom: Sum i = 1,n P(Ai) = 1.0 Sum of all probabilities of all outcomes Ai = 1.0.
   Outcomes must be mutually exclusive & exhaustive
    a. Complex event P(A v B v C v D) = P(A) + P(B) + P(C) + P(D)
       EXAMPLE – P(ace) = 1/52 +1/52 +1/52 +1/52 = 4/52 = 1/13

    b. Shared (combined) events and conditional probabilities – WEDNESDAY

Ch. 2 – probability distributions.

1. Binomial – flip coin = n! / X!(n-X)! * p**x (1-p)**n-x
    a. Histogram showing probabilities of many binomial trials.
    b. Problems –
        i. very cumbersome when n large, or p small
        ii. must know number of attempts (number of seeds that didn't germinate)
2. Other distributions – WEDNESDAY….
3. The bell curve! Continuous variables….
    a. Mean
    b. Variance (standard deviation).
        i. 68% all values – 1 SD from mean,
        ii. 95% - 2 SD,
        iii. 99.7% - 3 SD.

Central Limit Theorem – regardless of the underlying distribution of a variable, the distribution of many, many means of that variable will be normally distribution.

Law of large numbers – Observations must be random, independent, and drawn from large population

Types of variables

- Discrete
- Continuous
- Categorical

Populations, samples, censuses –

Summary statistics

1. Arithmetic mean – Ybar - write this in summation notation
    a. Geometric Mean – for rates - GMy – always a little smaller than arithmetic mean
2. median, mode
3. Spread –
    c. sample variance, standard deviation, standard error – in graphs use SD or SE – convention by field.
    d. Range, quantiles, quartiles, box and whisker plot
    e. Coefficient of variation – unitless –standardized by the mean, comparable to other measures. JMP multiplies this by 100

The Lab:

Concepts you will need to understand to do the lab (the lab will help!):

1. Binomial Probability
2. Combining probabilities
3. Measures of Central Tendency: Mode, Median, Mean (summation notation)

   Mean (X-bar): A sample estimate of the population mean (μ)

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = (Y_1 + Y_2 + Y_3 + Y_4 + \ldots + Y_n)/n$$

4. Trial, sample, population
5. Measures of dispersion (spread):
   a. range
   b. sample variance: sum of each of the differences or deviations between each individual value and the mean value.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

   c. **Standard deviation:** The square root of the variance.
   d. **Standard error:** the standard deviation divided by the square root of the sample size (n): most commonly used estimate of variance around means for figures in scientific papers in biology.

$$SE = \frac{s}{\sqrt{n}} = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}{\sqrt{n}}$$

   e. **Coefficient of variation (CV).** Measures the variability of values in a sample relative to the magnitude of the sample mean ($\bar{x}$) – gives an index of population variability that is comparable across measurements and populations.

$$CV = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}{|\bar{x}|}$$

   f. Quantiles, quartiles

Tools you will need to use to do the lab:

1. Excel
    a. Enter equations using Excel functions to compute summary statistics
    b. Sort data
    c. Create a histogram
    d. Create a Bar Graph (either in JMP or Excel)
2. JMP
    a. Copy/paste data from Excel into a JMP table
    b. Run summary statistics in JMP
    c. Copy/paste JMP tables back into Excel (to make a nice table)
    d. Create a Bar Graph (either in JMP or Excel)

About graphs – and reporting them!

1. Dependent variable on Y-axis, independent on x-axis.
2. ALWAYS use a figure legend!
3. Know what kind of variables you have: continuous, categorical, discrete….
    a. Bar chart – categorical variables – example – male/female heights.
    b. Scatterplot - linear analysis.

On Wednesday – what you (should have) learned in this lab!