Outlier

Section Summary | Introduction to Statistics
**Probability**

## Basics of Probability

1. Inferential Statistics
    a. Using subpopulations to make inferences about the whole population
    b. There is bound to be uncertainty in our inferences
    c. Probability gives us the tools to measure this uncertainty and calculate the likelihood that our inferences are correct
    d. Hypothesis Testing: statisticians reject claims when there is a very small probability of it happening by pure chance
2. Probability basics
    a. *S* denotes a "sample space," or list of possible outcomes
    b. A capital letter (*A, B, C,…*) denotes a specific event, or subcollection of outcomes in *S*
    c. Simple event: an event that contains one outcome that cannot be broken down any further
    d. Compound event: an event that contains more than one outcome
    e. Complementary event: denoted by a capital letter with a bar, apostrophe, or lowercase *c* ($\bar{A}, A', A^c$), this is an event that contains all outcomes not in the original event ($\bar{A}$ contains all outcomes not in *A*)
    f. Union of 2 events (*A or B*): an event that contains all outcomes in *A*, *B*, or both (make sure not to double-count outcomes)
    g. Intersection of 2 events (*A and B*): an event that contains only the outcomes in both *A* and *B*
3. Calculating probability
    a. Probability of an event in a sample space where all outcomes are equally likely:
        i. P(*A*) = $\frac{number\ of\ outcomes\ in\ A}{number\ of\ equally{-}likely\ outcomes\ in\ S}$
        ii. P(*A*) is read "the probability of *A*" and sometimes denoted 'Pr(*A*)'
    b. Relative frequency approximation

i. You can calculate probabilities by running tests and seeing how many times an event occurs
ii. $P(A) = \frac{number\ of\ times\ A\ occured}{number\ of\ times\ test\ was\ run}$
iii. Law of large numbers: the more you run your procedure, the closer the relative frequency approximation approaches the true probability
iv. It is always better to run more tests
v. If you cannot run a large number of tests, be wary about the reliability of your results

4. Properties of probabilities
   a. Probabilities are always between 0 and 1
   b. A probability of 0 means the event is impossible
   c. A probability of 1 means the event is certain
   d. Rare event rule
      i. If, under a given assumption, the probability of an event happening is very small and the event occurs with a frequency that is different than what we would expect (from pure chance), the assumption is incorrect
      ii. A common cut-off for "small probabilities" is anything below 0.05
      iii. The cut-off you decide to use will depend on the stakes of your experiment
      iv. If it's really important for your hypothesis to be correct, lower that cut-off down to 0.01 or even 0.001
5. Venn diagram
   a. Helps visualize your sample space/events
   b. First, draw a box to represent the sample space
   c. Circles inside that box represent specific events
   d. Overlapping portions of different circles represent shared outcomes between those events
   e. You can highlight full circles, overlapping portions of circles, or even space outside circles to show the events of interest

## Rules of Probability

1. Complement Rule
    a. $P(A) + P(\bar{A}) = P(S) = 1$
    b. Thus, $P(\bar{A}) = 1 - P(A)$
    c. Given the probability of an event, we can find the probability of its complement
2. Addition Rule
    a. P(A or B) = P(A) + P(B) - P(A and B)
    b. P(A) and P(B) take care of all the outcomes in A and B, but you need to subtract P(A and B) so you don't double-count the overlapping outcomes
3. Mutually exclusive events
    a. Events that cannot happen at the same time are considered mutually exclusive
    b. P(A and B) = 0 for mutually exclusive events
    c. Addition Rule for mutually exclusive events: P(A or B) = P(A) + P(B)
4. Conditional Probability Rule:
    a. P(B|A) = "probability of B given A"
    b. P(B|A) = $\frac{P(B \text{ } and \text{ } A)}{P(A)}$
    c. This formula gives the proportion of outcomes in A that are also in B
    d. Confusion of the inverse: $P(B|A) \neq P(A|B)$
5. Multiplication rule for successive events
    a. "and" can refer to two events happening at the same time like in the intersection of events, or it can refer to 2 events happening over 2 trials, 1 after another, as in successive events
    b. For successive events: $P(A \text{ } and \text{ } B) = P(A) \times P(B|A)$
6. Tree diagrams
    a. These are good for visualizing successive event probabilities
    b. Nodes show different possible outcomes given previous outcomes

   c. Branches show the probability of the next node occurring given the previous events
   d. You can find the final probability of a certain node by multiplying all the branches that lead up to that node

7. Independent events
   a. Events are independent if the occurrence of 1 event does not affect the probability of occurrence for the other events
   b. Mathematically, event B is independent of event A if P(B|A) = P(B)
   c. Multiplication rule for independent events: $P(A \text{ } and \text{ } B) = P(A) \times P(B)$
   d. Sampling with or without replacement
      i. Without replacement:
         1. When you make your first selection, you leave that selection out so that it is not available in the second selection
         2. Selection probabilities are not independent
      ii. With replacement
         1. Put the first selection back in so it is available for the second selection
         2. Selection probabilities are independent
      iii. 5% rule for independent events: when sampling without replacement, if your sample size is less than 5% of the population size, you can approximate these events as independent
8. More than 2 events
   a. Addition rule for 3 events
      i. $P(A \text{ } or \text{ } B \text{ } or \text{ } C) = P(A) + P(B) + P(C) - P(A \text{ } and \text{ } B) - P(A \text{ } and \text{ } C) - P(B \text{ } and \text{ } C) + P(A \text{ } and \text{ } B \text{ } and \text{ } C)$
      ii. This can be extended to any number of events
   b. Multiplication rule for 3 events
      i. $P(A \text{ } and \text{ } B \text{ } and \text{ } C) = P(A) \times P(B|A) \times P(C|(A \text{ } and \text{ } B))$
      ii. This can also be extended to any number of events
9. Keywords
   a. "Not", "doesn't," and "other than" = complement rule
   b. "Or" = addition rule
   c. "Given" = conditional probability

d. “And” with one trial = intersection of venn diagram
e. “And” with multiple trials = multiplication rule for successive events

## Contingency Tables

1. What is a contingency table?
    a. It is a type of frequency distribution table
    b. Can be used to:
        i. Group data for 2 different variables
        ii. Find associated probabilities
        iii. Observe trends across one variable with respect to the other
    c. 1 variable is listed vertically
    d. The other variable is listed horizontally
    e. Cells along those rows and columns represent the classes of those variables
    f. You don't need the same number of classes for each variable, so the possibilities are endless
    g. Column, row, and table totals are listed in the margins
2. Probability from contingency tables
    a. Joint probability
        i. Joint probabilities are our "and"/intersection probabilities
        ii. Represented by each inner cell
        iii. 1 cell gives the number of members characterized by both that cell's column and that cell's row
        iv. Find the probability of picking a member from that column and row by dividing that cell by the table total in the bottom right
    b. Marginal probability
        i. A marginal probability is the probability of observing a value of one of the variables with no regard for the value of the other variable
        ii. "Total" squares in the margins represent marginal probabilities
        iii. Find this probability by dividing the total of that column/row by the table total in the bottom right

c. Conditional probabilities
    i. Find P(B|A)
    ii. Zoom in on the column/row of the given variable, A
    iii. Find the cell representing B within A's row/column
    iv. P(B|A) is found by dividing that cell by A's total
d. "Or"/Union probabilities
    i. Sum up the column/row totals of both variables
    ii. Subtract any overlapping cells
    iii. Divide the result by the table total in the bottom right
e. Probability of successive events
    i. With replacement (independent events): multiply the probabilities of each event
    ii. Without replacement (not independent):
        1. Calculate the probability of the first event
        2. Remove that event from the table, recalculating cell values and totals
        3. Calculate the probability of the second event
        4. Multiply both probabilities together