Midterm Note Review

Cheat sheet + explanations for 254 midterm

Overview

Summarizing Data + *Graphs*

- Types of Data
- Data Collection Methods
- Graphing Methods

Probability (Part 1-5)

- Definitions
- Notation
- Rules + Equations

R Code + Simulink

-
-
- ---

Summarizing Data + Graphs

Variable types

- Numerical
  - Continuous: Numerical Values that exist on a continuous range of values
  - Discrete: Numerical Values that can only take on discrete values (ex. integers)
- Categorical
  - Regular Categorical: Category's that cannot be ordered
  - ordinal: Categories with logical ordering (ex. Grade Level)

Variable Relationships

Variables can be either **dependant** or **independent** of each other.

If a variable is dependent on another variable, the dependency can be either positively associated or negatively associated. With a positive association, an increase in one variable equals an increase in the other, and the opposite occurring for negative associativity.

Explanatory Variables If two variables are related, one may be defined as the **explanatory** variable, and the other/s as **response** variable/s. The explanatory variable drives the value of the response variable, its a causal relationship. This relationship needs to be determined, since there is a possibility that both variables you are observing are both response variables to a different data point.

Data Collection

Probability

Definitions

Random Process: a process that generates a known set of possible outcomes, for which we cannot say with certainty what we will get

Random Experiment: A experiment that can be repeated indefinitely with different outcomes that can not be predicted beforehand.

Sample Spaces: A set of all possible outcomes of a random event, they do not all need to have the same probability

Probability: Probability is a measure of the outcome of a random process, there are two interpretations of probability.

- Frequentists Interpretation: Probability is the measure of frequency of an outcome occurring over an infinite number of events
- Bayesian Interpretation: (fill later - cool machine learning stuff comes from this)

  Probability can only be between $0 < x < 1$, **0** means the event will not occur, **1** means the event will occur

Probability Function: for a sample space Ω a probability function assigns each outcome ω a number $P(\omega)$ which defines the probability of ω occuring

- Rule 1: $0 \leq P(\omega) \leq 1$
- Rule 2: All probabilities must sum to one $\sum_{j=1}^{n} P(\omega_j) = 1$
- Rule 3: The events in the sample space Ω must be disjoint

Disjoint or Mutually exclusive outcomes: A set of outcomes that cannot occur at the same time in a random process, disjoint events are dependant

- A coin toss can not be heads **and** tails, cards can not be both ace and queen

Independence: The probability of one event is not affected by another, in general, if $P(A|B) = P(A)$ then events A and B are independent

Random Variables: a symbol whos value is determined as a result of a random experiment, it has an associated sample space, can be either discrete or continuous. Denoted with capital letters it is **a function of the outcomes of an experiment**

- For sample space Ω, X is a function with a domain of Ω and range of either all real numbers R, or a subset

$$X : \Omega \to R$$

Probability Mass Function: the P.M.F of a discrete random variable X is the function $p_X(x_i) = P(X = x_i)\ for\ -\infty < x_i < \infty$. can be given as a formula, graph, or table.

Expectation: For a random variable X with outcomes $x_1, ..., x_n$ with probabilities $P(X = 1), ..., P(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability

$$\mu = E(X) = x_1 \cdot P(X = x_1) + ... + X_k \cdot P(X = x_k) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Linear Combinations: Random variables can be combined in *Linear combinations* allowing their Expectation and Variance to be computed as a sum

$$E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$$

$$\sigma^2 = Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

Continuous Distributions: If you take the graph of a Probability Density Function, you can estimate the probability of an outcome as the area under the curve of the graph. For a continuous function this means that the probability of any exact value is 0 since your asking for the area under a point

Confidence Intervals: plausible range of values for a population parameter

Null Hypothesis: H_0 the status quo, a perspective of no change or no difference

Alternative Hypothesis: H_A Represents the research question that's being tested

p-values: the probability of observing data at least as favorable to the alternative hypothesis as our current dataset, if the null hypothesis were true. A low p-value indicates that the Alternative Hypothesis may be true and the null may be false

Notation

Probability Notation $P(A)$ = The probability of event A occurring

Basic Set Theory Element: $x \in S$ Means that x is in the set S

Subset: $A \subset S$ means that set A is a subset of S when all of A's elements are in S

Complement: A^C or $S - A$ means the complement of A in S, the set of all elements in S that are not in A

Union: $A \cup B$ is the set of all elements that are in ***at least** A or B

Intersection: $A \cap B$ is the set of all elements that are in **both** A and B

Empty Set: $\emptyset$, pretty self explanatory'

Disjoint: $A \cap B = \emptyset$, formula to show no common elements

Difference: $A - B$ signifies the set of elements in A that are not in B

Set Example

Rules and Equations

Addition Rule The likelihood that either event A **and/or** B occurs.

$$P(A \; or \; B) = P(A) + P(B) - P(A \; and \; B)$$

Note: if A and B are Disjoint then $P(A \; and \; B) = 0$

Multiplication Rule

For Independent Processes

$$P(A \; and \; B) = P(A) \times P(B)$$

Note: A and B must be two independent processes

General Rule

$$P(A \; and \; B) = P(A|B) \times P(B)$$

Note: This is for events that we believe are not independent, otherwise this is rearranged to the previous equation above

For Three Events

$$P(A \; and \; (B \; and \; C)) = P(A|(B \; and \; C)) \times P(B \; and \; C)$$

Complement Rule For finding the probability of the complement of A, since total probability should sum to one

$$P(A^C) = 1 - P(A)$$

Conditional Probability The probability that A occurs given B has occurred

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \quad P(B) \neq 0$$

Law of Total Probability Suppose a sample space Ω is partitioned in 3 disjoint events B_1, B_2, B_3. Then for any event A

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + P(A \text{ and } B_3)$$
$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)$$

Sum of Conditional Probabilities The sum of all conditional probabilities should always equate to 1

$$P(A_1|B) + ... + P(A_n|B) = 1$$

Complement Rule for Conditional Probability From equation above

$$P(A|B) = 1 - P(A^C|B)$$

Bayes' Theorem The events $B_1, ..., B_n$ form a partition of the sample space Ω if every sample point ω is in one and only one B_i. That is:

$$\Omega = \bigcup_{i=1}^{n} B_i$$

and $B_1, ..., B_n$ are pairwise disjoint, that is to say the intersection of all pairs B_i, B_j where $i \neq j$, is the empty set.

To find $P(B_i|A)$

1. Calculate
$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$
2. Then calculate $P(B_i|A)$ using
$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)}$$

Expectation of a Random Variable The sum of each outcome multiplied by its corresponding probability

$$\mu = E(X) = x_1 \cdot P(X = x_1) + ... + X_k \cdot P(X = x_k) = \sum_{i=1}^{k} x_i P(X = x_i)$$

Variability of Random Variables

$$\sigma^2 = Var(X) = \sum_{i=1}^{k} [x_i - E(X)]^2 P(X = x_i)$$

Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Linear Combinations of Random Variables For two independent variables

$$E(aX + bY) = a \cdot E(X) + b \cdot E(Y)$$

$$\sigma^2 = Var(aX + bY) = a^2 \times Var(X) + b^2 \times Var(Y)$$

Probability Distributions

Probability Distributions of Random Variables Represent the values of a Random variable and the likelihood of those values

Normal Distributions (Insert picture of normal distribution)

The normal distribution is denoted as $N(\mu, \sigma)$ where μ is the mean of the distribution, and σ is the standard deviation.

Z Scores A Z Score quantifies the number of standard deviations from the mean a value is in a distribution

$$Z = \frac{x - \mu}{\sigma} = \frac{observation - mean}{SD}$$

Note: if the distribution is normal, a Z Score can be used to calculate percentile

Comparing Data to Normal Distribution When comparing how data correlates to a normal distributions, you can utilize the equation below to map the data set to the probability that a subset of the data will occur

$$p_i = \frac{i - 0.5}{n}$$

Where i is the index of the data point and n is the total number of data elements in the data set

This can be called *"empirical quantiles"*, since it empirically evaluates the probability based on the entire set

p_i evaluates to a probability, you then use this value to calculate the Z Score of each data point and plot data point vs Z value, the closer to a linear distribution the data is, the more linear the data plot is.

Bernoulli Trial with two possible outcomes (success and failure)

$p = P(\text{success on one trial})$

$q = 1 - p = P(\text{failure on one trial})$

Bernoulli Random Variable A Bernoulli Variable can only take on two values, 1 and 0. It is the output of a Bernoulli Trial

$P(X = 1) = p$

$P(X = 0) = q$

Geometric Distributions If W is the waiting time for first success, which is the number of the trial on which the first success appeared

$$P(W = i) = q^{i-1} \cdot p$$

For $i = 1, 2, 3, ...$

The expectation is

$$E(W) = \mu = \frac{1}{p}$$

And the standard deviation is

$$\sigma = \sqrt{\frac{1 - p}{p^2}}$$

Combinatorics Combinatorics is an area of math concerning counting

Rule 1: Multiplication For a set of events where:

- Event 1 can be n_1 options
- Event 2 can be n_2 options
- ...
  - Event k can be n_k options

The total number of combinations of each event is

$k = n_1 \cdot n_2 \cdot ... \cdot n_k$

Rule 2: Permutations The number of permutations of a set defines the number of unique combinations of length k including unique orders that the set n can create

$$(n)_k = n(n-1)(n-2)...(n-k+1)$$

if k=n then

$$(n)_n = n!$$

Rule 3: Combinations The number of combinations is defined as the number of unique unordered subsets of a set, this can also be viewed as number of permutations regardless of order

$$\binom{n}{k} = \frac{(n)_k}{k} = \frac{n!}{k!(n-k)!}$$

Inference

Central Limit Theorem The distribution of a **sample mean** is well approximated by a normal distribution if: the events are independent, and the sample size is appropriate

$\bar{x} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$

$\frac{\sigma}{\sqrt{n}}$ is the standard error or SE

Confidence Intervals $\bar{x} \pm Z^* \times SE = (x_1, x_2)$ Z is the 'Z score' we want to get (% likely within from table)

Z = 2 for 95% confidence that a mean is within interval

Margin of Error In a confidence interval $Z^* \times SE$ is the **Margin of error**

p-value to find p-value you must first find the # of Standard Error away from the sample mean the null hypothesis is

$Z = \frac{\bar{x}-\mu_0}{SE} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ (s - sample variance, SE - Sample Error, n - sample size)

The p(x) is then calculated by looking at the table and finding the corresponding % chance that this variable would be observed