STAT1005 Week 4 Cheat Sheet

Independence

- Two events A and B are said to be independent if P(A|B) = P(A) and are dependent otherwise.
  - If A and B are independent, so too are A and B^c, A^c and B, and A^c and B^c
- If two events are independent, then P(A $\cap$ B) = P(A) x P(B)

Random Variables

- You can redefine sample spaces in terms of functions known as random variables
- A random variable is a function that assigns numbers with an attribute of a sample outcome
  - Eg, if X denotes the random variable, and s denotes a sample outcome, then X(s) = t, where t is a real number
- A random variable is a function whose domain is the sample space and whose range is some subset of real numbers
- **<u>Types of Random Variables</u>**
  - Discrete - countable, number of yellow smarties, etc. Finite or countably infinite values.
  - Continuous - uncountable, time, distance, etc

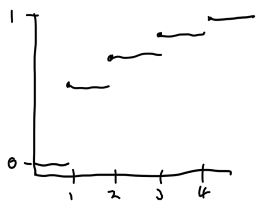
Discrete Probability Distributions

- Where the outcomes in S are assigned probabilities, these determine the probabilities associated with the values of any particular random variable X
- The probability distribution of X determines how the total probability of 1 is distributed among the possible values of X
- Usually expressed as P(X = x)
  - p(x) - probability that X takes on the value x
- Probability Function or Probability Mass Function
  - p(x) = P(X = x)
  - Properties of p(x)
    - 1. p(x) $\geq$ 0 for all x
    - 2. $\sum_x$ p(x) = 1

Bernoulli Random Variables

- Any random variable whose only possible values are 0 and 1 (2 possible outcomes, assigned as something happens or it doesn't)
- If X is a Bernoulli random variable then
  - $P(x; a) = \begin{cases} 1 \text{ with a probability of a} \\ 0 \text{ with a probability } 1 - a \end{cases}$
  - **<u>Bernoulli Distribution (or Trial)</u>**
    - When p(x) depends on a quantity that can take on one of a range of possible values, eg, a, the probability of "success", the each value of a defines a different probability distribution. This is a parameter of the distribution, and the collection of all probability distributions for different values of the parameter is called a family of probability distributions

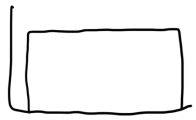
Cumulative Distribution Functions

- Used to look at the probability that the observed value of X will be at least, or at most x.
- The cumulative distribution function (cdf) F(x) of a discrete random variable X with pmf p(x) is defined for every number x by
  - $F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$
  - For any number x, F(x) is the probability that the observed value of X will be at most x
- For a discrete random variable X, the graph of F(x) is a step function: it jumps at every possible value of X and is flat between possible values
- 

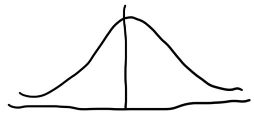
- P(a < X < b) = F(b) - F(a - 1)

Density Curves

- Uniform distribution



- Normal distribution



- **<u>Normal Calculations</u>**
  - Step 1.
    - State the problem in terms of the observed variable X. Draw a picture that shows the proportion you want in terms of cumulative proportions
  - Step 2.
    - Standardise X to restate the problem in terms of a standard Normal varible Z
  - Step 3.
    - Use Table A and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve

Parameters and Statistics

- Parameter - a number that describes some characteristic of the population (p)
- Statistic - a number that describes some characteristic of a sample, the statistic that estimates p is the sample proportion ($\hat{p}$)
- Statistical Inference - uses information from a sample to draw conclusions about a wider population
  - Need to be able to describe the sampling distribution of possible statistical values in order to perform statistical inference
  - **<u>The Law of Large Numbers</u>**
    - If we continue to take larger and larger samples, the statistic $\hat{p}$ is guaranteed to get closer to the parameter p
- **<u>Sampling Distribution</u>**
  - The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples os the same size from the population
  - **<u>Sampling Distribution of a Sample Proportion</u>**
    - As the sample size increases, the sampling distribution of $\hat{p}$ becomes approximately Normal. That is for large n, $\hat{p}$ has approximately a Normal distribution with mean p and standard deviation $\sqrt{p(1 - p)/n}$
      - $N = (p, \frac{p(1-n)}{n})$
    - Rule of thumb: can use the approximation when the sample size n is large enough so that both np and n(1 - p) are about 10+

Statistical Inference

- We can estimate the true proportion from a single sample of smarties, by attaching a measure of uncertainty
  - The uncertainty is based on the sampling distribution of samples of size n
- An approximate 95% (within 2σ) confidence interval for the true unknown proportion from a single sample of size n and which contains a sample proportion $\hat{p}$ is
  - $\hat{p} \pm 2\sqrt{\hat{p}(1 - \hat{p})/n}$
  - **<u>Steps</u>**
    - 1. Find $\hat{p}$
    - 2. Find $2\sqrt{\hat{p}(1 - \hat{p})/n}$
    - 3. $p \pm 2\sqrt{\hat{p}(1 - \hat{p})/n}$
    - 4. Then n x the two values, to find the practical range
- Point estimate - sample statistic $\hat{p}$ gives us point estimate of the population parameter p
- Interval estimate - tell us the range of plausible values the population parameter might take
- This is known as a confidence interval
  - point estimate ± margin of error