STAT1005 Week 5 Cheat Sheet

Probability Mass Functions (pmf's)

- Probability mass functions provide an over view of a random variables behaviour, where the x axis is the numerical value for the variable, and the y axis is the probability of that value

Expected Values

- The expected value is the average of a pmf, where the values included in the average are 'weighted' by their probabilities
- X - Discrete random variable
- D - set of possible values for X
- p(x) - PMF function
- E(X) - expected value, or mean value of X
  - $E(X) = \mu_x = \mu = \sum_{x \in D} x \cdot p(x)$
- ***Variance and Standard Deviation of X***
  - Var(X) - variance of X, or σ_x^2 or σ^2
    - $Var(X) = \sum_D [(x - \mu)^2 . p(x)] = E[(X - \mu)^2]$
  - SD(X) = standard deviation of X, or σ_x or σ
    - $\sigma_x = \sqrt{Var(x)}$
- ***Properties of Expectations***
  - If we are looking at the pmf of a function h(x), then
    - $E[h(X)] = \sum_D h(x).p(x)$
    - This is only valid if $\sum_D |h(x)|.p(x)$ is finite
  - For constants a and b,
    - $E(aX + b) = aE(X) + b$
      - Due to linearity of expectation
      - This can be extended to things like $E[X^2 - X] = E[X^2] - E[X]$
- ***Properties of Variance***
  - $Var(X) = E[(X - \mu)^2] = E(X^2) - \mu^2$
    - This is equivalent to $E[(X - \mu)^2] = E[X^2] - (E[X])^2$
  - When h(x) is linear (aX + b),
    - $Var(aX + b) = \sigma_{aX+b}^2 = a^2\sigma_x^2$ and $\sigma_{aX+b} = |a|\sigma_x$
    - $\sigma_{aX} = |a|\sigma_x$ and $\sigma_{X+b} = \sigma_X$

Binomial Distribution

- It is binomial if it meets the following requirements
  - 1. The experiment is made up of n smaller experiments called trials, and n is fixed at the start of the experiment (the same number of trial are conducted regardless of the results)
  - 2. Each trial can result in one of two of the same possible outcomes (usually denoted as a success (S) or failure (F) (Bernoulli trials)
  - 3. The trials are independent (no trial affects the outcome of any other trial)
  - 4. The probability of success is the same in every trial (denoted as p)
- n - number of trials
- X - number of successes from n trials
- The probability of the outcomes are denoted as
  - $y = \begin{cases} 1, \text{ with probability } p > 0 \\ 0, \text{ with probability } 1 - p \end{cases}$
- So, $p(y) = p^y (1 - p)^{1-y}$
- And, $X = \sum_{i=1}^{n} Y_i \sim Bin(n, p)$
- So we write $X \sim Bin(n, p)$ to show X is a binomial random variable, with n trials and success probability of p
- The pmf is written as b(x; n, p)
- $b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2 \dots n \\ 0 & \text{otherwise} \end{cases}$
- b(x; n, p) = {n trials consisting of x S's}.{probability of any particular such sequence}
- For X ~ Bin(n, p), the cdf is denoted by
  - $B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p) \qquad x = 0, 1, 2, \dots, n$
- E(X) for bin(x; n, p) = np
- Var(X) for bin(x; n, p) = np(1 - p) = npq, where q = 1 - p

Negative Binomial Distribution

- Is also looking at the number of successes from repeated Bernoulli trials, but wants to know how many trials it will take to reach the r^{th} success
- So number of trials is random, and the number of successes is fixed
- It is a negative binomial if it meets the following requirements
  - 1. The experiment consists of a sequence of independent trials
  - 2. Each trial results in success (S) or failure (F)
  - 3. The probability of success is constant in all trials (p)
  - 4. The trials are performed until a total of r successes has been observed
- The random variable X is the number of trials required to achieve the r^{th} success
  - Possible values of X are r, r +1, r + 2, ..., since it takes at least r trials to achieve r successes
- If the r^{th} success is to occur on the x^{th} trial, there must be (r - 1) successes in the first (x - 1) trials, and the probability of this is
  - $\binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)}$
- The pmf of the negative binomial random variable is;
  - $nb(x; r, p) = \binom{x-1}{r-1} p^{r}(1-p)^{x-r} \quad x = r, r+1, r+2, \ldots$
- Probabilities from a negative binomial distribution can be obtained from tables of binomial probabilities by making use of the following identity:
  - nb(x; r, p) = $\frac{r}{x}$ .b(r; x, p)
- E(X) = $\frac{r}{p}$
- Var(X) = $\frac{r(1-p)}{p^2}$

Confidence Intervals

- A level C confidence interval for a parameter has two parts:
  - 1. An interval calculated from the data, of the form:
    - estimate ± margin of error
  - 2. A confidence level, C, which is the proportion of times that the interval will capture the true parameter value in repeated samples. Or the success rate for the method
- For any level of confidence C, we can write,
  - $\hat{p} \pm z^*\sqrt{\hat{p}(1 - \hat{p})/n}$
    - z^* - the ciritcal value for the confidence level C
    - $\sqrt{p(1 - p)/n}$ - the standard error
    - z^* x standard error - the margin of error
  - This is based on the sampling distribution, which is only valid if np and n(1 - p) are both greater than 10.
- The actual values for the confidence intervals are based on normal distribution

Confidence level, C	90%	95%	99%
Critical value, z*	1.645	1.96	2.576

- Ideally you want a high confidence and a small margin of error
  - The margin of error gets smaller when
    - z^* gets smaller (lower confidence level C)
    - n gets larger - need to take 4 times as many observations to halve the margin of error

Choosing the Margin of Error

- The margin of error in the large-sample confidence interval for p is
  - $m = z^*\sqrt{\hat{p}(1 - \hat{p})/n}$
- Sample size for desired margin of error
  - The level C confidence interval for a population proportion p will have margin of error approximately equal to a specified value m when sample size is
    - $n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$
- p^* - is a guessed value for the sample proportion.
  - The margin of error will always be less than or equal to m if you take the guess p^* to be 0.5
  - (Round up)