STAT1005 Week 6 Cheat Sheet

Geometric Distribution

- Negative binomial with r = 1
  - Pmf: $nb(x; 1, p) = \begin{cases} (1-p)^{x-1}p, & x = 1, 2, 3, \ldots \\ 0 & \text{otherwise} \end{cases}$
- If we redefine X to be the number of failures, then
  - $nb(x; 1, p) = (1-p)^x p, \qquad x = 0, 1, 2, \ldots$
- $E[X] = 1/p$
- $Var(X) = \frac{(1-p)}{p^2}$

Hypergeometric Distribution

- Binomial distribution is the **exact** probability model for sampling **with** replacement from a finite dichotomous population, with an **approximate** probability model for sampling **without** replacement
- The hypergeometric distribution is the **exact** probability model for the number of successes when we sample **without** replacement
- 1. N is finite population to be sampled
- 2. Each individual is a success (S) or failure (F), are there are M successes in the population
- 3. n is the sample of individuals selected without replacement
- Random variable of interest is X = the number of sucesses (S) in the sample
  - To find the pmf:
    - $P(X = x) = h(x;n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
    - For integer x satifying $\max(0, n - N + M) \leq x \leq \min(n, m)$ *
- $E[X] = n.\frac{M}{N}$
- $Var(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)$

Poisson Distribution/Exponential Distribution

- Used where we count the number of successes in a particular region or interval of time
- A random variable X is said to have a Poisson distribution with parameter λ ($\lambda > 0$) if the pmf of X is:
  - $p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$
- If X has a Poisson distribution with parameter λ, we write that $X \sim Pois(\lambda)$, and it has mean and variance;
  - $E[X] = Var(X) = \lambda$
- A binomial with $n \to \infty$ and $p \to 0$ in such as way that $np \to$ a value $\lambda > 0$, then it tends toward a poisson distribution.
  - In practice, the approximation can be used if n > 50 and np > 5
- It is poisson if
  - Events occur randomly in time
  - Uniform rate
  - Independent
  - $P(\text{event in } (t, t + \delta t)) = \mu\delta t + o(\delta t)$
- The poisson distribution can be used as an approximate for a binomial distribution
  - Where $\lambda = E[X]$ of the binomial = np

Continuous Random Variables

- A random variable is continuous if:
  - Its possible values comprise either a single interval on the number line or a union of disjoint intervals, and
  - $P(X = c) = 0$ for any number c that is a possible value of X
- For continuous variables they might only be able to take discrete measurements, but we still treat them as continuous.
- When X is a continuous random variable, then the pdf of X is a function f(x) such that for any two numbers a and b with $a \leq b$,
  - $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - For f(x) to be a legitimate pdf, it must satisfy:
    - $f(x) \geq 0$ for all x
    - $\int_{-\infty}^{\infty} f(x)dx = 1$
- It does not matter if the upper or lower limit are included, the value will be the same
  - $P(a \leq X \leq b) = P(a < x < b)$

Uniform Distribution

- A continuous random variable X is said to have a uniform distribution on the interval [A, B] if the pdf of X is
  - $f(x; A, B) = \begin{cases} \frac{1}{B - A}, & A \leq x \leq B \\ 0, & \text{otherwise} \end{cases}$
  - We denote this by X ~ Unif[A, B]
- $E[X] = \frac{A + B}{2}$
- $Var(X) = \frac{(B - A)^2}{12}$

Continuous Numerical Variables

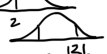
- Distribution symbols
  - μ - population mean
  - x - sample mean
  - σ^2 - (population) variance
  - s^2 - sample variance

Hypothesis Testing

- Confidence intervals are one of two common types of statistical inference
- Confidence intervals are used when the goal is to estimate a population parameter
- Test of significance is used when the goal is to assess the evidence provided by the data about some claim concerning the population
  - Make a claim (the null hypothesis) and test it against an alternative claim (the alternative hypothesis)
    - An outcome the would rarely happen if a claim were true is good evidence that the claim is not true.
- 1. Set up a null hypothesis, a claim we believe to be true
- 2. Set up an alternative hypothesis, a claim that challenges the null hypothesis
- 3. Start by assuming that the null hypothesis is true
- 4. Sampling distribution: If the null hypothesis is really true, then the proportion of heads in the sample ($\hat{p}$) will have a Normal distribution
- 5. Calculate the p-value, which is the probability, in either direction, of observing a value as large as what we actually observed, given the null hypothesis.

- **<u>P-Value and Statistical Significance</u>**
  - The probability, computed assuming H_0 is true, that the statistic would take a value as extreme as or more extreme than the one actually observed is called the p-value of the test.
    - The smaller the p-value, the stronger the evidence against H_0 provided by the data
      - Small p-values are evidence against H_0 because they say that the observed result is unlikely to occur when H_0 is true
    - Large p-values fail to give convincing evidence against H_0 because they say that the observed result could have occurred by chance if H_0 were true
  - Our conclusion in a significance test:
    - P-value small → reject H_0 → conclude H_A (in context)
    - P-value large → fail to reject H_0 → cannot conclude H_A (in context)
  - There is no rule for how small a p-value we should require in order to reject H - it's a matter of judgement and depends on the specific circumstances
    - We can compare the p-value to a fixed value that we regard as decisive called the significance level (generally 0.05, or 0.01)
    - When our p-value is less than the chosen significance level, we say that the result is statistically significant

Large Sample Tests for a Population Proportion

- State: What is the practical question that requires a statistical test?
- Plan: Identify the parameter, state null and alternative hypotheses, and choose the type of test that fits your situation
- Solve: Carry out the test in three phases:
  - 1. Check the conditions for the test you plan to use
  - 2. Calculate the test statistic
  - 3. Find the p-value
- Conclude: Return to the practical question to describe your result in this setting
- **<u>Significance Tests for a Proportion</u>**
  - Draw an SRS of size n from a large population with an unknown proportion of p successes. To test the hypothesis H : $p = p_0$, compute the z statistic
    - $z = (\hat{p} - p_0)/\ p_0(1 - p_0)/n$
  - In terms of variable Z having the standard Normal distribution, the approximate P-value for a test of H against
    - H : $p > p_0$ is $P(Z \geq z)$ 
    - H : $p < p_0$ is $P(Z \leq z)$ 
    - H : $p \neq p_0$ is $2xP(Z \geq |z|)$ 
  - Cautions about significance test:
    - Hypotheses always refer to the population, not to a particular outcome.
    - State H_0 and H_A in terms of population parameters
    - The hypotheses should express hopes and suspicions, it is not ethical to look at the data and then frame the hypotheses to fit
    - Failing to find evidence against H_0 means only that the data are not inconsistent with H_0, not that we have clear evidence that H_0 is true.
      - Only data that are inconsistent with H_0 provide evidence against H_0
    - There is no sharp border between "significant" and "not significant", only increasingly strong evidence as the p-value decreases.
      - P values are relatively arbitrary
    - How important an effect is depends on the size of the effect as well as on its statistical significance
      - Might not be practically important