STAT1005 Week 8 Cheat Sheet

Discrete and Continuous Probability Distributions

- ***Note:***
  - f(x) - continuous pdf
  - F(x) - continuous cdf
- ***Discrete Probability Distribution***
  - $P(X = x) = p(x)$
    - Properties of p(x)
      - 1. $p(x) \geq 0$ for all x
      - 2. $\sum_x p(x) = 1$
- ***Continuous Probability Distribution***
  - $P(a \leq X \leq b) = \int_a^b f(x)\,dx$
    - Properties of f(x)
      - 1. $f(x) \geq 0$ for all x
      - 2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$
  - For a random variable X with the following density function:
    - $f(x) = \begin{cases} kx + c, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$
    - To find the value of k find where $\int_a^b kx + c\,dx = 1$
  - $P(a \leq X \leq b) = F(b) - F(a)$
    - Same premise as for normal distributions
    - Different from the discrete case which is:
      - $P(a \leq X \leq b) = F(b) - F(a - 1)$
  - $F'(x) = f(x)$
    - Essentially, you convert from cdf to pdf by finding the derivative of F(x)
    - And you convert from pdf to cdf by finding the integral of f(x) for all values of x
  - ***Percentiles of a Continuous Distribution***
    - p - A number between 0 and 1 that represents a percentile. A percentile is the point at which 100p percent of values falls at or below. Eg, 0.5 represents the 50th percentile
    - $\eta(p)$ - is the function that gives the value for x that when put into F(x) = p. So when p = 0.5, F(x) = 0.5, which means that value of x is where 50% of the sum of the probabilities for the values of x between a and x equals 50%, which is the median.
    - $p = F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(y)\,dy$
      - Or if you can find the inverse of F(x)
        - *Note*: The inverse of a function (for a function f(x) represented by $f^{-1}(x)$) is the function that if you put the outcome of the function into, it will give you the value you originally put into the function to get that outcome.
        - $\eta(p) = F^{-1}(p)$
  - ***Expected Values***
    - The expected value or mean value of a continuous random variable X with pdf f(x):
      - $\mu = \mu_x = E(X) = \int_{-\infty}^{\infty} x.f(x)\,dx$
    - If X is a continuous random variable with pdf f(x) and h(x) is any function of x, then:
      - $E[h(x)] = \int_{-\infty}^{\infty} h(x).f(x)\,dx$
  - ***Variance and Standard Deviation***
    - The variance of a continuous random variable X with pdf f(x) and mean value μ is:
      - $\sigma_x^2 = Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 .f(x)\,dx$
    - The standard deviation of X is $\sigma_x = \sqrt{Var(X)}$
    - Also;
      - $Var(X) = E(X^2) - [E(X)]^2$

Normal Distribution

- Even when individual variables themselves are not normally distributed, sums and average of the variables will under suitable condition have approximately a normal distribution (Central Limit Theorum)
- A continuous random variable X is said to have a Normal (or Gaussian) distribution with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is:
  - $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$
  - Denoted as $X \sim N(\mu, \sigma^2)$
- Normal density is symmetric; the median and mean coincide
- The value of σ is the distance from μ to the inflection point of the curve
- Large values of σ yield density curves the are quite spread out, and so you may observe a value of X quite far from μ.
- ***Standard Normal Distribution***
  - To calculate $P(a \leq X \leq b)$ when $X \sim N(\mu, \sigma^2)$:
    - $\int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$
    - You can use specialised approximations for solve this
    - Otherwise use the standard normal distribution discussed previously
  - The normal distribution with parameter values $\mu = 0$ and, $\sigma = 1$ is called the standard normal distribution.
  - A random variable that has a standard normal distribution is called a standard normal random variable denoted as Z
    - The pdf of Z is:
      - $f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$
    - The cdf of Z denoted as $\Phi(z)$ is:
      - $P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$
- ***Log-Normal Distribution***
  - A continuous random variable X is said to have a log-normal distribution with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if:
    - $\ln(X) \sim N(\mu, \sigma^2)$
  - The pdf of X:
    - $f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$
  - The cdf of X:
    - $f(x;\mu,\sigma) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$
- ***Transformations of Random Variables***
  - Where $Y \sim N(\mu, \sigma^2)$:
    - $Pr(X \leq x) = Pr(e^Y \leq x) = Pr(Y \leq \ln(x))$
      - Where Pr is probability
  - It can be shown that:
    - $E(X^n) = e^{n\mu + \frac{1}{2}n^2\sigma^2}$
    - So that:
      - $E(X) = e^{\mu + \frac{\sigma^2}{2}}$
      - $Var(X) = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)$

Gamma Distribution/Exponential Distribution

- ***Exponential Distribution***
  - A continuous random variable X is said to have an exponential distribution if the pdf of X is:
    - $f(x;\lambda) = \lambda e^{-\lambda x}$, $x > 0$
      - The distribution of inter-arrival times in a Poisson Process with rate λ events per unit time
    - A cdf of:
      - $1 - e^{-\lambda x}$ $x > 0$
    - $E[X] = 1/\lambda$
    - $Var(X) = 1/\lambda^2$
- ***Gamma Distribution***
  - The gamma distribution is a generalisation of the exponential distribution
  - A continuous random variable X is said to have a gamma distribution if the pdf of X is:
    - $f(x;\alpha,\beta) = \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1}e^{-x/\beta}$ $x < 0$
      - Where $\Gamma(\alpha)$ is the gamma function
      - And $\alpha > 0$ and $\beta > 0$
      - The exponential distribution results from taking $\alpha = 1$ and $\beta = 1/\lambda$
      - When $\beta = 1$, X is said to have a standard gamma distribution
  - *Gamma Function*
    - For $\alpha < 0$, the gamma function $\Gamma(\alpha)$ is defined by:
      - $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx$
        - 1. For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
        - 2. For any positive integer n, $\Gamma(n) = (n - 1)!$
        - 3. $\Gamma(1/2) = \pi$

Hypothesis testing

- We use *confidence intervals* when our goal is to estimate a population parameter such as the true proportion p (categorical) or mean μ (quantitative) of a population
- We use *tests of significance* when the goal is to assess the evidence provided by the data about some claim concerning a population
- We make a claim (the null hypothesis - H_0) and test it against an alternative claim (the alternative hypothesis - H_A)
- We assess if the null hypothesis is really true, how likely would we be to observe an event as large as, or as small as, what we actually observed. (P-value)
- ***Hypothesis Testing About a Population Mean***
  - For a categorical proportion p, remember to use z values (Normal distribution)
  - For a quantitative variable X with population mean μ and population standard deviation σ, we want to test the hypotheses that:
    - H : $\mu = \mu_0$
    - H : $\mu > \mu_0$
  - The test statistic:
    - $T_{n-1} \sim t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
      - Where $\bar{x}$ and s are calculated from the sample data
      - This has a t-distribution with n - 1 degrees of freedom
  - The P-value for testing H_0 against:
    - $H_A: \mu > \mu_0$ is $P(T \geq t)$
    - $H_A: \mu < \mu_0$ is $P(T \leq t)$
    - $H_A: \mu \neq \mu_0$ is $2P(T \geq |t|)$ - this is used when H_0 wasn't set up as directional
  - Using R, the P-value is:
    - $P(T \geq t) = 1 - pt((\bar{x}-\mu)/(s/\sigma), (n-1))$
  - The level of significance is associated with a critical value.
    - An alternative way of doing the hypothesis test is to compare the value or t we calculated to the critical t-value associated with the level of significance we want to assess for.
      - If it is greater than t_{crit} we would reject H_0 ; if less, we would fail to reject H_0
- The P-value gives us the probability of rejecting the null hypothesis when in fact is it true

Comparing Two Population Means

- Population 1: $(\mu_1, \sigma_1^2) \rightarrow (\bar{x}_1, s_1^2)$
- Population 2: $(\mu_2, \sigma_2^2) \rightarrow (\bar{x}_2, s_2^2)$
- **Underlying Model**
  - Every test is based on a underlying model for the data generation process
  - In this case:
    - $Y_{ij} = \mu + g_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma_i^2)$, $j = 1, \ldots, n_i$
      - Where g_i is the effect of being in a certain group
      - And ϵ_{ij} is some random variance
- **Comparing Two Population Means**
  - 1. Must both be SRS's from two distinct populations.
    - a) So the samples must be independent
    - b) And measuring the same response variable for both samples
  - 2. Both must be normally distributed
    - The means and sd's of the population is unknown, but in practice it is enough that the distribution have similar shapes with no strong outliers
  - 3. For observations with moderate skewness and no outliers, the sum of the two sample sizes should be at least 15
  - 4. For observations from a population with strong skewness and no outliers, the sum of the two sample sizes should be at least 40
  - **Two-Sample t-Procedure**
    - ***Two-Sample t-Statistic***
      - We take variation into account and standardise the observed difference $\bar{x}_1 - \bar{x}_2$ by subtracting its mean, $\mu_1 - \mu_2$, and dividing the result by its standard error (two-sample t-statistic)
        - $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}}$
        - Where you don't know μ just use $(\bar{x}_1 - \bar{x}_2)$
      - The two-sample t-statistic has approximately a t distribution. It does not have exactly a t distribution if the populations are both exactly normal.
        - In practice, however, the approximation is very accurate
      - You can use software to use the statistic t with accurate critical values
      - Otherwise without software, use the statistic t with critical values from the t-distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$.
        - The significance test gives a P-value equal to or greater than the true P-value
    - ***Two-Sample Confidence Interval***
      - $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1}{n_1} + \frac{s_2}{n_2}}$
        - t* is the critical value for confidence level c for the t distribution with degrees of freedom from either software, or the smaller of $n_1 - 1$ and $n_2 - 1$