STAT1005 Data Analysis Condensed Cheat Sheet

Lisa Luff

5/28/2020

Qualitative/Numerical Data

Discrete Random Variables

Sample Analysis

- Mean : $(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- $E[X] = \mu_X = \sum_x x.p(x) = E[h(x)] = \sum h(x).p(x)$
  - Linearity of expectations
    - $E(aX+b) = aE(X) + b$
- In R:
  - Mean function
    $\rightarrow$ mean(x)
  - Equivalent function for $E[X]$
    $\rightarrow$ weighted.mean(y, p)
- Sample Standard Deviation: $s = \sqrt{\frac{1}{n-1}\sum^{n}(x_i - \bar{x})^2}$
- Standard Deviation: $s = \sqrt{\frac{1}{n}\sum^{n}(x_i - \bar{x})^2}$
- $Var(X) = \sum[(x-\mu)^2.p(x)] = E[(X-\mu)^2] = E[X^2] - E[X]^2$
  *Linearity of expectations
  - $Var(aX+b) = |a|\sigma_x$
- In R:
  - Standard Deviation function
    $\rightarrow$ sd(x)
  - Variance function
    $\rightarrow$ var(x)

Sample Probability Analysis

- Probability Mass Function (pmf): $p(x) = P(X = x)$
- Cumulative Density Function (cdf): $F(x) = P(X \leq x) = \sum p(y)$
  $\rightarrow P(a \leq x \leq b) = F(b) - F(a-1)$
- Confidence Interval: $\mu \pm Z * (\frac{s}{\sqrt{n}})$
  - 90% CI - Z* = 1.645

  - 95% CI - Z* = 1.96

  - 99% CI - Z* = 2.576

Continuous Random Variables

Sample Analysis

- Mean: $\bar{x} = E[X] = \int_{-\infty}^{\infty} x.f(x)dx$
- In R:
  - Mean function
    $\rightarrow$ mean(x)
- Standard Deviation: $\sigma = \sqrt{Var(X)}$
- $Var(X) = \int_{-\infty}^{\infty}(x-\mu)^2.f(x)dx$
- In R:
  - Standard deviation function
    $\rightarrow$ sd(x)
  - Variance function
    $\rightarrow$ var(x)

Sample Probability Analysis

- Probability Density Function (pdf): $f(x) = \int_a^b f(x)dx$
- Cumulative Density Function (cdf): $F(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(x)dx$
  $\rightarrow P(a \leq X \leq b) = F(a) - F(b)$
- Percentiles of a Continuous Distribution: $F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(y)dy$
- In R:
  - Find the nth percentile
    $\rightarrow$ quantile(x, $n^{th}p$)
- Confidence Interval: $\mu \pm Z * (\frac{\sigma}{\sqrt{n}})$
  - 90% CI - Z* = 1.645

  - 95% CI - Z* = 1.96

  - 99% CI - Z* = 2.576

Transforming Variables

- Transforming variables is when you rearrange the formula to find the value of a random variable X based on its probability.
  - 1. Need to use the cdf, so find the cdf if given a pdf
  - 2. Rearrange the cdf so that you have an equation that uses the value of F(X)/y to give X
  - 3. Enter the value into your new equation and now you know which random variable X will give you that value for y

Discrete Random Variables

Probability Estimations Using Distributions

Central Limit Theorum

- As sample size increases (number of trials), the distribution of sample means will move towards a Normal shape, regardless of the distribution of the actual observations

Normal Distribution

- Used when we know σ, otherwise use t distribution
  When np and $n(1-p) > 10$
- $z = \frac{x-\bar{\mu}}{\sigma}$
  $\rightarrow$ or if you don't know μ and σ use $\bar{x}$ and s
- The standard normal distribution has a mean of 0 and sd of 1
- In R:
  - Find p (probability)
    $\rightarrow$ pnorm(z) Or $\rightarrow$ pnorm(x, mean = a, sd = b)
  - Find the z score
    $\rightarrow$ qnorm(p)

T Distribution

- Used when we don't know σ
- Find t statistic: $T_{n-1} \sim t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
  - With n - 1 degrees of freedom (df)
- In R:
  - Find the probability
    $\rightarrow$ pt(x, df)
  - Find t
    $\rightarrow$ qt(p, df)
- In R:
  - Find p (probability) $\rightarrow$ pt(x, df)
  - Find t
    $\rightarrow$ qt(p, df)

Bernoulli Random Variable

- A random variable whose possible values are 0 and 1
- $P(x;a) = \begin{cases} 1 & a \\ 0 & (1-a) \end{cases}$
- Basis of the other distributions

Continuous Random Variables

Probability Estimations Using Distributions

Standard Normal Distribution

- To calculate $P(a \leq X \leq b)$ when $X \sim N(\mu, \sigma^2)$
  $\rightarrow z = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma}^2)} dx$
  - If $-\infty \leq \mu \leq \infty$, $\sigma > 0$
- Cdf denoted as: ϕ
- In R:
  - Find p (probability)
    $\rightarrow$ pnorm(z) OR $\rightarrow$ pnorm(x, mean = a, sd = b)
  - Find the z score
    $\rightarrow$ qnorm(p)

Uniform Distribution

- When $f(x;A,B) = \begin{cases} \dfrac{1}{A-B}, & A \leq x \leq B \\ 0, & otherwise \end{cases}$
- $E[X] = \frac{A+B}{2}$
- $Var(X) = \frac{(B-A)^2}{12}$
- In R:
  - Find p (probability)
    $\rightarrow$ punif(x, min = a, max = b)
  - Find x for a given p
    $\rightarrow$ qunif(p, min = a, max = b)

Log Normal Distribution

- If $\ln(X) = N(\mu, \sigma)$
- In R:
  - Find p (probability)
    $\rightarrow$ plnorm(x, meanlog = a, sdlog = b)
  - Find x for a given p
    $\rightarrow$ qlnorm(p, meanlog = a, sdlog = b)

Gamma Distribution

- Generalised exponential function
- If $\alpha = 1$ and $\beta = \frac{1}{\lambda}$ see exponential distribution
  - Exponential function results from $\alpha = 1$ and $\beta = \frac{1}{\lambda}$
- pdf: $f(x;\alpha,\beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{\frac{-x}{\beta}}$, x > 0
  - $\Gamma(\alpha)$ is the Gamma function
  - $\alpha > 0$, and $\beta > 0$, if $\beta = 1$ it is a standard Gamma distribution *Gamma function
  - $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$
  - $\Gamma(\frac{1}{2}) = \pi$
- In R:
  - Find p (probability)
    $\rightarrow$ pgamma(x, alpha, rate = beta, scale = 1/beta)
  - Find x for a given p
    $\rightarrow$ qgamma(p, alpha, rate = beta, scale = 1/beta)

Discrete Random Variables

Probability Estimations Using Distributions Continued

Binomial

- If you have S - success or F - failure
- Fixed n - number of independent trials
- How many successes in n number of trials?
- Exact for with replacement, approximate for without
- See hypergeometric if without replacement
- If $n > 50$ and $np > 5$ see Poisson
- pmf: $b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, .., n \\ 0 & otherwise \end{cases}$
- cdf: $B(x; n, p) = P(X \leq x) = \sum_{y=0}^{x} b(y; n, p)$, x = 1, 2, . . . , n
- $E[X] = np$
- $Var(X) = np(1-p) = npq$, where $q = 1 - p$
- In R:
  - Find p (probability) → pbinom(x, n, p)
  - Find x for a given p → qbinom(p, n, p)

Negative Binomial

- If you have S - success or F - failure
- Fixed r - number of successes
- How many trials will it take to reach r successes?
- If the r^th success occurs on the x^th trial, there must be (r - 1) successes in the first (x - 1) trials
- If $r = 1$, see geometric
- pmf: $nb(x; r, p) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & x = r, r^{r+1}, ... \\ 0 & otherwise \end{cases}$
- $E[X] = \frac{r}{p}$
- $Var(X) = \frac{r(1-p)}{p^2}$
- In R:
  - Find p (probability) → pnbinom(x, n, p, mu)
  - Find x for a given p → qbinom(p)

Continuous Random Variables

Probability Estimations Using Distributions Continued

Exponential Disribution

- Inter-arrival times in a poisson process with a rate of λ events per unit time
- pdf: $f(x; \lambda) = \lambda e^{-\lambda x}$, x > 0
- cdf: $1 - e^{-\lambda x}$, x > 0
- $E[X] = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- In R:
  - Find p (probability) → pexp(x, lambda)
  - Find x for a given p → qexp(p, lambda)

Discrete Random Variables

Probability Estimations Using Distributions Continued

Geometric

- Negative binomial where $r = 1$
- pmf: $nb(x; 1, p) = (1-p)^{x-1}p, x = 1, 2..$
- If we redfine x as the number of failures:
  - $nb(x; 1, p) = (1-p)^{x}p$, x = 0, 1, …
- $E[X] = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$
- In R:
  - Find p (probability)
    → pgeom(x, p)
  - Find x for a given p
    → qgeom(p)

Hypergeometric

- Binomial without replacement
- N is finite population to be sampled
- M is number of successes in the population
- pmf: $P(X = x) = h(x; n, M, N) = \frac{\binom{m}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$
- $E[X] = n\frac{M}{N}$
- $Var(X) = \frac{N-n}{N-1}n\frac{M}{N}(1 - \frac{M}{N})$
- In R:
  - Find p (probability)
    → phyper(x, m, n, k)
  - Find x for a given p
    → qhyper(p)

Poisson

- Number of successes without a time period
- It is poisson if:
  - Events occur randomly in time
  - Independently and at a uniform rate
- Can be used as an approximation for a binomial if $n > 50$ and $np > 5$
  - The $E[X]$ of the binomial is used for the value of *lambda*
- pmf: $p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, …
- $E[X] = Var(X) = \lambda$
- In R:
  - Find p (probability)
    → ppois(x, lambda)
  - Find x for a given p
    → qpois(p)

Discrete and Continuous Random Variables

Hypothesis Testing

- Set up hypotheses:
  $\rightarrow H_0 = \mu = \mu_0$
  $\rightarrow H_A = \mu \neq \mu_0$
- Find t statistic: $T_{n-1} \sim t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- Find the p-value:
  - $H_A : \mu \leq \mu_0 = P(T \leq t)$
  - $H_A : \mu \geq \mu_0 = P(T \geq t)$
  - $H_A : \mu \neq \mu_0 = 2P(T \geq |t|)$
- In R:
  - Find p-value
    → pt(t* , df = (n-1))
- Or you can find the t statistic, p-value and confidence interval in R
  - → t.test(Vector, μ_0, alternative = “less”“greater”“two.sided”, conf.level = X.XX)

Comparing Population Means

- Sample populations must be independent
- They must be Normally distributed (or similar shape with no significant outliers)
  - Moderate skewness: $n_1 + n_2 \geq 15$
  - Strong skewness: $n_1 + n_2 \geq 40$
- Two-sample t statistic: $t = \frac{(\bar{x}_1-\bar{x}_2)(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{N_1}+\frac{S_2^2}{N_2}}}$
- If you don't know μ, just use $\bar{x}$
- Two-Sample Confidence Interval: $(\bar{x}_1 - \bar{x}_2) \pm t * \sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$
- Think of it as a hypothesis test, where $H_0 = \bar{x}_1 - \bar{x}_2 = 0$, and, $H_A = \bar{x}_1 - \bar{x}_2 \neq 0$
- Both in R:
  - Find two sample t statistic and confidence interval
    → t.test(SampleVector, ComparisonVector, mu = 0, alternative = “less”“greater”“two.sided”, conf.level = X.XX)

Discrete Random Variables

Joint Probability Distribution

Sample Probability Functions

- Joint Probability Mass Function (pmf): $p(x, y) = P(X = x, Y = y)$
  - $\rightarrow$ Created by $p(x)p(y)$ for each square
- In R:
  - Joint pmf
    $\rightarrow$ joint(X, Y)
- Marginal Probability Mass Function (marginal pmf):
  $\rightarrow$ Marginal pmf of x: $p_x(x) = \sum_y P(x, y)$
  $\rightarrow$ Marginal pmg of y: $P_y(y) = \sum_x P(x, y)$
- Joint Cumulative Density Function (cdf): $F(x, y) = P(x \leq x, Y \leq y) = \sum_{u \leq x, v \leq y} p(u, v)$ for all x, y

Independence of Random Variables

- X and Y are independent if:
  $\rightarrow P(x, y) = P_x(x).P_y(y)$, for all values of x and y

Continuous Random Variables

Joint Probability Distribution

Sample Probability Functions

- Joint Probability Mass Function (pmf): $f(x, y) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dx dy$
- Calculate double integrals by:
  - 1. Inner integral: hold x constant, integrate over y
  - 2. Outer integral: Hold y constant, integrate over x
  - Doesn't matter which step is integrating for x or y
- Marginal Probability Mass Function (marginal pmf):
  $\rightarrow$ Marginal pmf of x: $p_x(x) = \int_{-\infty}^{\infty} f(x, y) dy$, for $-\infty \leq X \leq \infty$ $\rightarrow$ Marginal pmf of y: $P_y(y) = \int_{-\infty}^{\infty} f(x, y) dx$, for $-\infty \leq Y \leq \infty$
- Joint Cumulative Density Function (cdf): $F(x, y) = P(x \leq x, Y \leq y) = \sum_{u \leq x, v \leq y} p(u, v)$ for all x, y

Independence of Random Variables

- Independence of Random Variables
  - X and Y are independent if:
    $\rightarrow P(x, y) = P_x(x).P_y(y)$, for all values of x and y
  - AND the region of postive density has side parallel to the axis

Discrete Random Variables

Joint Probability Distribution Continued

Sample Probability Analysis

- $\mu_{h(x,y)} = E[h(x,y)] = \sum_x \sum_y h(x,y).p(x,y)$
- Covariance: $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$
  - $\rightarrow = \sum_x \sum_y (X-\mu_x)(Y-\mu_y)p(x,y)$
  - $\rightarrow = E[X,Y] - E[X]E[Y]$
- If X and Y move together, covariance will be positive
- If x and Y move away, covariance will be negative
- If X and Y don't have a strong relation, covariance will near 0
- In R:
  - Covariance
    $\rightarrow$ cov(table or dataframe of X, Y)
- Correlation: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$
- Correlation is a linear measure of the strength of the relationship between -1 and 1
  - Not being correlated doesn't mean there isn't a relationship, it might be non-linear
- In R:
  - Correlation
    $\rightarrow$ cor(table or dataframe of X, Y)

Continuous Random Variables

Joint Probability Distribution Continued

Sample Probability Analysis

- $\mu_{h(x,y)} = E[h(x,y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y).f(x,y)dxdy$
- Linearity of expectation:
  - $E[a_1h_1(X,Y) + a_2h_2(X,Y) + b] = a_1E[h_1(X,Y)] + a_2E[h_2(X.Y)] + b$
- If X and Y are independent:
  - $E[h(X,Y)] = E[f_x(X)f_y(Y)] = E[f_x(X)f_y(Y)]$
- Covariance: $Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$
  - $\rightarrow = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (X-\mu_x)(Y-\mu_y)f(X,Y)dxdy$
  - $\rightarrow = E[X,Y] - E[X]E[Y]$
- Properties of Covariance:
  - Cov(X, Y) = Cov(Y, X)
  - Cov(X, X) = Var(X)
  - For any random variable Z:
    - $\rightarrow$ $Cov(aX + bY + c, Z) = aCov(X,Z) + bCov(Y,Z)$
- Correlation: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$

Discrete Random Variables

Joint Probability Distribution Continued

Matched Pair Design

- Compare two populations in an experiment where they are paired
  - Case-control clinical trials
- Reduces two-sample t-test to a 1 sample
- More powerful (sensitive) test
- Pearson Correlation Coefficient
  - Covariance test for matched pair data
  - Strongly effected by outliers
    $\rightarrow r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$
  - Where:
    * $\rightarrow \bar{x} \sim \mu_x = n^{-1} \sum_{i=1}^{i=n} x_i$
    * $\rightarrow \bar{y} \sim \mu_y = n^{-1} \sum_{i=1}^{i=n} y_i$
    * $\rightarrow S_{X,Y} \sim Cov(X, Y)$
      $= \sum_{i=1}^{n} (X_i - \bar{x})(y_i - \bar{y})$
      · $\rightarrow = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$
    * $\rightarrow S_{X,X} \sim \sigma_x^2 = \sum_{i=1}^{n} (X_i - \bar{x})^2$
      · $\rightarrow = \sum_{i=1}^{n} x_i - n\bar{x}^2$
    * $\rightarrow S_{Y,Y} \sim \sigma_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$
      · $\rightarrow = \sum_{i=1}^{n} y_i - n\bar{y}^2$

- In R:
  $\rightarrow$ cov(table or dataframe of X, Y, paired = TRUE, method = 'pearson')
- Spearman's Correlation Coefficient
  - Based purely on the ranks of the data
    * From smallest to largest
  - Collected as numbers, then arranged in order
  - Not effected much by outliers
    $\rightarrow r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)}$
  - $\rightarrow d_i^2$ is the squared difference between the ranks of each category per variable
- In R:
  $\rightarrow$ cov(table or dataframe of X, Y, paired = TRUE, method = 'spearman')

Discrete Random Variables

Joint Probability Distribution Continued

Goodness-of-Fit Testing

- Tests is a probability model is an appropriate measure of the data collected, "close enough" to what we would expect to observe
- Chi-Squared Distribution: χ^2
  - Special case of Gamma distribution
    - When the degrees of freedom is ν, then $\alpha = \frac{\nu}{2}$, and $\beta = 2$
      $\rightarrow \chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$
  - Where:
    - o_i is the observed frequency
    - e_i is the expected frequency
    - $n = \sum_{i=1}^{k} n_i$
    - k - 1 degrees of freedom
- The size of the discrepency indicates how good a fit, to assess the size:
  - Find p-value
  - OR Compare to a critical value of α for χ^2_α
- In R:
  - Find p for χ^2
    $\rightarrow$ pchisq(χ^2_i, k-1)
  - Find the critical value of α
    $\rightarrow$ qchisq(α, k-1)
    OR chisq.test(Vector)

Continuous Random Variables

Joint Probability Distribution Continued

Goodness-of-Fit Testing

- See Goodness-of-Fit Discrete for more details
- Chi-Squared Distribution: χ^2
  $\rightarrow \chi^2_{k-1} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$
  - Where cell properties are given by: $\rightarrow$ $p_i = P(a_{i-1} \leq X \leq a_i) = \int_{a_{i-1}}^{a_i} f(x)dx$
- Find p-value or use χ^2_α, where the critical variable is found using $1 - \alpha$ (We want to find the top α percent)
- In R:
  - Find p for $\chi^2 \rightarrow$ pchisq(χ^2_i, k-1(df))
  - Find the critical value of $\alpha \rightarrow$ pchisq(α, k-1(df))
    OR chisq.test(Vector)

Discrete and Continuous Random Variables

Two-Way Contingency Tables

- For when there are multiple rows for multiple columns of information
- Multiple rows per column of information
  - I rows (I $\leq$ 2)
  - J columns

Testing for Independence

- Use χ^2 distribution
  - $\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}} \sim \chi^2_{(I-1)(J-1)}$
  - Where: $\rightarrow$ expected cell count $= \frac{\text{row total x column total}}{\text{expected count}}$
- In R:
  - Run the test
    - chisq.test(table or dataframe of X, Y)

Categorical/Qualitative

Nominal and Ordinal

Sample Analysis

- Sample Proportion: $\hat{p} \sim p$
- Standard Error: $\sqrt{p(1-p)/n}$

Calculating Sample Proportion

Defintions

- Ordered - Order of the outcomes matter. (DOG and GOD are different)
- Unordered - Order of the outcomes doesn't matter. (DOG and GOD are equivalent)
- With replacement - Even if chosen previously, equally likely to be chosen again
- Without replacement - Once chosen unable to be chosen again

Naive Definition of Probability

- All outcomes equally likely to occur
- $\hat{p} = P(A) = \frac{|A|}{|S|}$
- In R:
  - Find $P(A)$
    $\rightarrow$ Prob(A)

General Product Rule

- Based on naive definition
- Ordered outcomes
- With replacement
- n^k

Permutation

- Based on naive definition
- Ordered outcomes
- Without replacement
- $\hat{p} = P_{k,n} = \frac{n!}{(n-k)!}$
- In R:
  $\rightarrow$ permutations(n, r)

Combination

- Based on naive definition
- Unordered outcomes
- Without replacement
- $\hat{p} = \binom{n}{k} = \frac{P_{k,n}}{n!} = \frac{n!}{(n-k)!k!}$
- In R:
  $\rightarrow$ choose(n, r)

Non-Naive Definition of Probability

- Probability of an event is $P(A)$
- $P(S) = \sum_{i=1} P(a_i) = 1$

Operations in Set Theory

- Based on Non-Naive Definition
- In R:
  - Union
    $\rightarrow$ union(a, b)
  - Intersection
    $\rightarrow$ intersect(a, b)

De Morgan's Laws

- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

Additional Properties of Probability

- $P(A^c) = 1 - P(A)$
- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Property 1.
  - $P(S) = P(A \cup A^c) = P(A) + P(A^c) = 1$
- Property 2.
  - If $A \subseteq B$, then
  - $\rightarrow P(B) = P(A \cup (B \cap A^c)) = P(A) + P(B \cap A^c)$
  - $\rightarrow P(B) \geq P(A)$
- Property 3.
  - $P(B \cap A^c) = P(B) - P(A \cap B)$

Independence

- Two events are independent if
- $\rightarrow P(A \cap B) = P(A).P(B)$
- In R:
  - Test for independence
    $\rightarrow$ independent(X, Y)

Conditional Probability

- In R:
  - Find $P(A|B)$
    $\rightarrow$ prob(A, given = B)

Continuation of Set Theory

- $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Independence

- If $P(B|A) = P(B)$, the two are independent.

Multiplication Rule

- $P(A \cap B) = P(A|B).P(B)$
- $P(A \cap B \cap C) = P(C|A \cap B).P(A \cap B)$
- OR $P(A \cap B \cap C) = P(C|A \cap B).P(B|A).P(A)$

The Law of Total Probability

- $P(B) = \sum_{i=1}^{k} P(B|A_i).P(A_i)$

Bayes Theorum

- When $A_1, ..., A_k$ and mutually exclusive and exhaustive events
- $P(A_i|B) = \frac{P(B|A_i).(A_i)}{\sum_{i=1}^{k} P(B|A_i).P(A_i)}$

Statistical Inference

Law of Large Numbers

- As sample size increases, the mean observed will get closer and closer to the true mean.

Normal Distribution

- Can be used as an estimate when np and $n(1-p) \geq 10$
- N = (p, $\frac{p(1-p)}{n}$)
- In R:
  - Find p (proportion)
    $\rightarrow$ pnorm(z, $\hat{p}$, se)
  - Find the z score
    $\rightarrow$ qnorm(p)

Confidence Interval

- $\hat{p} = z * \sqrt{\hat{p}(1-\hat{p})/n}$
  - 90% CI - Z* = 1.645

  - 95% CI - Z* = 1.96

  - 99% CI - Z* = 2.576

Margin of Error

- m = $z * \sqrt{\hat{p}(1-\hat{p})/n}$
- Find n for a required margin of error
- $\rightarrow n = \frac{z*}{m}.p * (1 - p*)$
- Where p* is a guessed Value for the sample proportion
  - Margin of error will always $\leq m$ if $p* = 0.5$ (round up)

Hypothesis Testing

- Set up hypotheses:
  $\rightarrow H_0 = p = p_0$
  - If h_0 is correct $\hat{p} \sim N(p, \frac{p(1-p)}{n}) \rightarrow H_A = p \neq p_0$
- Find z statistic: $z = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n}}$
- Find p-value:
  - $H_A : p < p_0 = P(Z \leq z)$
  - $H_A : p > p_0 = P(Z \geq z)$
  - $H_A : p \neq p_0 = 2P(Z \geq |z|)$
- Find p-value in R:
  - pnorm(Z*)

Further Analysis

- If you treat the data as discrete, or utilise the data as a table, you can use the same analysis as for discrete variables for two-way contingency tables