Probability and Statistics Notes

December 2020

1 Measures of location

- Sample mean $\bar{x}$ is the mean of all observations
- Population mean μ is conceptual - this is the mean of the whole population: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ where N is the size of the population
- Median $\tilde{x}$ is the $\frac{n-1}{2}^{\text{th}}$ value if n is odd and the average of the $\frac{n}{2}^{\text{th}}$ and $\frac{n}{2}+1^{\text{th}}$ values if n is even
- Population median $\tilde{\mu}$ is the median of the population
- Our task is the make conclusions about μ and $\tilde{\mu}$ knowing just $\bar{x}$ and $\tilde{x}$
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{S_{xx}}{n-1}$ where $S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2$
- Population variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

2 Basic probability theory

- Conditional probability $P(A|B)$ means probability of A given B
- $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- For mutually exclusive and exhaustive events A_1 to A_i, $P(B) = P(B|A_1)P(A_1) + (B|A_2)P(A_2) + ... + P(B|A_i)P(A_i)$
- Bayes Theorem: $P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$
- Independence means that $P(A \cap B) = P(A)P(B)$

3 Discrete random variables

- For a given sample space $\mathcal{S}$ of an experiment, a random variable (rv) is a rule that associates a number with each outcome in $\mathcal{S}$
- A Bernoulli rv can only be 0 or 1
- rv's can be discrete or continuous

4 Probability distributions

- The probability distribution (pd) of a discrete rv is defined for every number x by $p(x) = P(X = x) = P(\text{all } s \in \mathcal{S} \text{ for which } X(s) = x)$ - this is the probability of observing value x when the experiment is performed
- The Bernoulli distributions are the family of distributions where $p(x; a) = \begin{cases} 1-\alpha & x = 0 \\ \alpha & x = 1 \\ 0 & \text{otherwise} \end{cases}$

5 Cumulative distribution function of a discrete random variable

- The cdp of a discrete rv x with probability distribution p is defined as $F(x) = P(X \leq x) = \sum_{y:y \leq x} p(y)$
- This is the probability that the observed value will be at most x

6 Expected value of a discrete random variable

- $E(X)$ or μ_X is the expected or mean value
- $E(X) = \mu_X = \sum_{x \in D} xp(x)$

7 Variance of a discrete random variable

- $V(X)$ or σ_X^2 is the variance
- $V(X) = \sum_{x \in D} (x - \mu)^2 p(x) = E[(X - \mu)^2]$
- The standard deviation is the square root of the variance
- Variance can also be obtained as $V(X) = E(X^2) - [E(X)]^2$

8 Binomial probability distribution

- A binomial experiment is defined as follows:
  1. The experiment consists of a sequence of n smaller experiments (trials).
  2. Each trial can result in one of two possible outcomes, which we denote by S (success) and F (failure).
  3. The trials are independent from each other.
  4. The probability of success is constant from trial to trial. We denote it by p.
- The binomial random variable X is the number of successes among the n trials
- The probability distribution for X, denoted by $b(X; n, p)$ is $b(X; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, ..., n \\ 0 & \text{otherwise} \end{cases}$
- If X is a binomial rv with n trials and probability p then
  - $E(X) = np$
  - $V(X) = np(1 - p)$

9 Poisson probability distribution

- $p(x;\lambda) = \frac{e^{\lambda}\lambda^{x}}{x!}$ for x = 0,1,2,...
- $E(X) = V(X) = \lambda$
- The Poisson distribution arises as the limit of the Binomial distribution when $n \to \infty$ and $p \to 0$ in such a way that $np \to \lambda > 0$ - in this case $b(X;n,p) \to p(x;\lambda)$
- A Poisson process is defined as:
  - There is an event, for example a radioactive pulse, that occurs at a certain rate α
  - The probability that the event occurs in a small time interval δt is $\alpha \cdot \delta t$
  - The number of events that occur during a time interval Δt is independent of the number of events that occurred prior to this time interval
  - $P_k(t) = \frac{e^{\alpha t}(\alpha t)^k}{k!}$ is the probability that k pulses will be received during an time interval of length t
  - $P_k(t)$ is a Poisson distribution with parameter $\lambda = \alpha t$

10 Continuous random variables

- A random variable is said to be continuous if its set of possible values is an entire interval of numbers, that is, if for some A ¡ B any number x between A and B is possible
- The probability distribution of X is a function $f(x)$ such that for any 2 numbers a and b with $a \leq b$, $P(a \leq X \leq b) = \int_a^b f(x)dx$
- Probability distributions always integrate to 1

11 Cumulative distribution function of continuous random variable

- Cumulative probability distribution $F(x) = \int_{-\infty}^{x} f(y)dy$
- $P(X > a) = 1 - F(a)$
- $P(a \leq X \leq b) = F(b) - F(a)$
- $F'(x) = f(x)$

12 Percentiles

- $p = F[\eta(p)] = \int_{-\infty}^{\eta(p)} f(y)dy$
- Probability to have $x \leq \eta(p)$ is p
- Median of continuous distribution, $\tilde{\mu}$, is the 50th percentile
- $\tilde{\mu}$ satisfies $F[\tilde{\mu}] = 0.5$ - half the area is on the right, half on the left

13 Expected value

- $E(x)$ is the expected or mean value
- $E(x) = \int_{-\infty}^{\infty} xf(x)dx$

14 Variance

- $V(x) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$

15 Normal distribution

- $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- The ; sign means parameterised by
- Expected value is μ
- Standard deviation is σ
- We can compute probabilities numerically or by looking up values in a statistical table
- A binomial distribution with $np > 10$ and $n(1-p) > 10$ can be approximated by a normal distribution with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$

16 Standard normal distribution

- Define $z = \frac{x-\mu}{\sigma}$
- z has a Standard Normal Distribution - can look up in statistical table
- Has cumulative probability distribution $\Phi(z)$

17 Joint probability distributions

- $p(x,y) = P(X = x \text{ and } Y = y)$
- If A is the set of all pairs (x,y), $P(X,Y) \in A = \sum_{(x,y)\in A} p(x,y)$
- For continuous random variables, $f(x,y) = P(X,Y) \in A = \int\int_A f(x,y)\,dx\,dy$
- $E[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y)p(x,y) & \text{discrete} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)\,dx\,dy & \text{continuous} \end{cases}$

18 Marginal probability distributions

- $p_x(x) = \sum_y p(x,y)$
- $p_y(y) = \sum_x p(x,y)$
- $f_x(x) = \int_{-\infty}^{\infty} f(x,y)dy$
- $f_y(y) = \int_{-\infty}^{\infty} f(x,y)dx$
- X and Y are independent if $p(x,y) = p_x(x)p_y(y)$

19 Covariance

- $\text{Cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] = \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$

20 Correlation coefficient

- Normalise the covariance $\rho_{x,y} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$
- $\rho_{x,y}$ is between -1 (perfect anti-correlation) and 1 (perfect correlation)

21 Central limit theorem

- States that in many situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution
- Normal distribution has the same mean as the sample mean
- $\sigma_z = \frac{\sigma}{\sqrt{n}}$ or $\sigma_z^2 = \frac{\sigma^2}{n}$
- What if the random variables are not independent or from different distributions?
  - Let random variables X_1 to X_n have mean values μ_1 to μ_n and variance σ_1^2 to σ_n^2
  - The linear combination $Y = \sum_{i=1}^{n} a_i X_i$ has mean $E(Y) = a_1\mu_1 + ... + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i$ and variances $V(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \text{Cov}(X_i, X_j)$
  - If the random variables are independent, the variance simplifies to $V(Y) = \sum_{i=1}^{n} a_i^2\sigma_i^2$

22 Parameter estimation

- θ is the symbol for a parameter, for example population average μ
- $\hat{\theta}$ is the symbol for a point estimator, for example sample average $\bar{x}$
- If the point estimator is unbiased, this means that $E(\hat{\theta}) = \theta$ for every possible θ
- $E(\hat{\theta}) - \theta$ is the bias of $\hat{\theta}$
- Sample average $\bar{x}$ and sample median $\tilde{x}$ are both unbiased estimators
- Precision of estimator $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$
- Standard error of $\bar{x} = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- If σ is unknown, we can estimate σ through the sample standard deviation: $\hat{\sigma} = s = \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$

23 Confidence intervals

- Assume $x_1, x_2, ..., x_n$ come from a normal distribution - this means $\bar{x}$ has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- If we know $\bar{x}$, this can be turned into a probability for μ, the population mean
- $p(\mu) = \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}} e^{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}}$
- The $100(1-\alpha)\%$ confidence interval for μ lies between $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

24 t - distribution

- If n is small, $s \neq \sigma$
- Must use t distribution
- If $x_1, x_2, ..., x_n$ come from a normal distribution, $T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a t distribution with n-1 degrees of freedom
- The t critical values can be used in the same way as normal
- t distribution leads to larger confidence intervals as it is wider

25 Bayesian statistics

- $m(x)$ is the probability distribution of x, called the Marginal Distribution - How likely is it to observe data x in general?
- $\pi(\mu)$ is the probability distribution of μ, called the Prior Distribution - What we believed about the the population mean before we started the test.
- $h(x, \mu)$ is the joint distribution of x and μ - How likely is it for both the data x to be observed and the mean to be μ.
- $f(x|\mu)$ is the conditional probability of x given μ, called the Likelihood - How likely was it to observe data x given we knew the mean distribution?
- $\pi(\mu|x)$ is the conditional probability of μ given x, called the Posterior Distribution - Probability distribution of possible population mean values given the data that has been observed.
- Bayes Theorem: $\pi(\mu|x) = \frac{f(x|\mu)\pi(\mu)}{m(x)}$
- $m(x) = \int f(x|\mu)\pi(\mu)d\mu$
- If $\pi(\mu)$ is distributed as $\mathcal{N}(\nu, \tau^2)$ and $f(x|\mu)$ is distributed as $\mathcal{N}(\mu, \sigma^2)$
  - We can use Bayes theorem to compute the posterior
  - $\pi(\mu|x)$ is distributed as $\mathcal{N}(\frac{\tau^2}{\sigma^2+\tau^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\nu, \frac{\tau^2\sigma^2}{\sigma^2+\tau^2})$
  - The posterior is a weighted average of the prior and new data we're measuring
  - The amount of shift depends on the confidence (variance) in prior and new data
- For n observations $x_1, x_2, ..., x_n$
  - $f(\text{All of the data points } |\mu) = \mathcal{N}(\mu, \frac{\sigma^2}{n})$
  - $\pi(\mu|x_1, x_2, ..., x_n)$ is distributed as $\mathcal{N}(\frac{\tau^2}{\frac{\sigma^2}{n}+\tau^2}\bar{x} + \frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2}\nu, \frac{\tau^2\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}+\tau^2})$

26 Bayesian prediction

- Given observations x_1 to x_n, we can predict the next observation x_{n+1} using Bayesian prediction
- $f(x_{n+1}|x_1, x_2, ..., x_n) = \int f(x_{n+1}|\mu)\pi(\mu|x_1, x_2, ..., x_n)\, d\mu$
- The Bayesian prediction is not only based on the most likely value for μ but on the whole distribution - this makes a difference when the distribution is not symmetrical as values lower or higher than μ may be more likely

27 Interval estimation

- We can define a credibility interval C with area $1-\alpha$ on the cumulative distribution of $\pi(\mu|x_1, x_2, ..., x_n)$
- $\pi(\mu|x_1, x_2, ..., x_n) = \mathcal{N}(\kappa, \lambda^2)$
- This can be standardised using variable $z = \frac{\mu-\kappa}{\lambda}$ which has a standard normal distribution
- $(\kappa - z_{\alpha/2}\lambda, \kappa + z_{\alpha/2}\lambda)$ is the $1-\alpha$ % credibility interval

28 Hypothesis testing

- Type 1 error: Rejecting the null hypothesis when it is in fact true - aka false positive - has probability α
- Type 2 error: Accepting the null hypothesis when it is in fact false - aka false negative - has probability β
- We generally try to err on the side of the null hypothesis and so aim to minimise α, the probability of a type 1 error
- Devising a test that minimises α
  - If we assume x comes from $\mathcal{N}(\mu, \sigma)$, we can say that $\bar{x}$ is distributed as $\mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$
  - The null hypothesis $\mu = \mu_0$ implies that $\bar{x}$ comes from a normal distribution with mean μ_0 and standard deviation $\frac{\sigma}{\sqrt{n}}$
  - If $\bar{x}$ lies within a selected confidence interval on the distribution $p_0(\bar{x})$, that is $\mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$, then it is likely that $\bar{x}$ comes from this old probability distribution p_0 and that $\mu = \mu_0$ - in this case we accept H_0
  - If $\bar{x}$ lies outside a certain interval on the distribution $p_0(\bar{x})$, it is unlikely that $\bar{x}$ comes from p_0 and probably $\mu \neq \mu_0$ - in this case we reject H_0
- We can use normalisation with variable $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ (which has a standard normal distribution) in order to define a rejection region using z cutoff values
- Tests can include both tails, or just one
- For a single tailed test, we can place a cutoff at z_α which separates off an area α on the standard normal distribution - in this case the probability that $z \geq z_\alpha$ is α
- We reject H_0 when $z \geq z_\alpha$
- The probability of a Type 1 error is α
- Probability of Type 2 error (null hypothesis accepted although it is false)
  - Assume the the true mean $\mu > \mu_0$
  - We can define $\tilde{z} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$
  - Then $\tilde{z} = z - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}$
  - $\beta = p(z \leq z_\alpha) = p(\tilde{z} \leq z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}}) = \phi(z_\alpha - \frac{\mu-\mu_0}{\sigma/\sqrt{n}})$
  - Knowing α, μ and μ_0 we can obtain β
  - The smaller α, the larger β
- What to do if standard deviation σ unknown?
  - For a large sample $n > 30$, use the sample standard deviation as a proxy - replace σ by sample standard deviation s - this means $z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a standard normal distribution
  - For a small sample $n < 30$, use the t distribution - this means $T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ has a t distribution with n-1 degrees of freedom

29 ANOVA

- ANOVA stands for analysis of variance
- Consider a number I of populations / medical treatments with means μ_1 to μ_I
- Each population / treatment 1 to I has the same standard deviation / variance σ/σ^2
- A number J measurements are taken from each of the I populations
- X_{ij} is the random variable that denotes the j^{th} measurement taken from the i^{th} population
- Mean square for populations (treatments) MSTr $= \frac{J}{I-1}\sum_{i=1}^{I}(\bar{X}_i - \bar{X})^2$
  - $\bar{X}_i = \frac{1}{J}\sum_{j=1}^{J}\bar{X}_{ij}$ is the mean for population i
  - $\bar{X} = \frac{1}{I}\sum_{i=1}^{I}\bar{X}_i$ is the grand mean
- Mean square error MSE $= \frac{1}{I}\sum_{i=1}^{I}S_i^2$
  - $S_i^2 = \frac{1}{J-1}\sum_{j=1}^{J}(X_{ij} - \bar{X}_i)^2$ is the variance for population i
  - MSE is the average population variance
- Null hypothesis: μ_1 to μ_I are equal $\longrightarrow \mathcal{E}(\text{MSTr}) = \mathcal{E}(\text{MSE}) = \sigma^2$
- Alternate hypothesis: At least 2 means are different $\longrightarrow \mathcal{E}(\text{MSTr}) > \mathcal{E}(\text{MSE}) = \sigma^2$
- The variable $f = \frac{\text{MSTr}}{\text{MSE}}$ follows an F distribution and is the test statsitics for a single factor ANOVA
  - The F distribution has 2 parameters, denoted by ν_1 and ν_2 which are both positive integers
  - A random variable that follows the F distribution is always non negative
  - ν_1 is the number of numerator degrees of freedom
  - ν_2 is the number of denominator degrees of freedom
  - For ANOVA $\nu_1 = I - 1$ and $\nu_2 = I \cdot (J - 1)$
- If $f \approx 1$ accept H_0
- If $f >> 1$ reject H_0
- Note that f can be smaller than 1 but this would be just due to chance fluctuations
- More precisely, if $f > F_{\alpha,\nu_1,\nu_2}$ reject H_0, otherwise accept H_0

30 Non parametric testing

- If the data does not follow a Gaussian distribution, we cannot use the t distribution
- An alternative is the Wilcoxon signed rank test
- We assume $\mu_0 = 0$ (or otherwise normalise data by looking at $x_1 - \mu_0$, $x_2 - \mu_0$, etc.)
- The null hypothesis is that $\mu = 0$
- We assume that the distribution is symmetric around the mean, and then test if the data is symmetric around 0
- Step 1: Rank the values by absolute value
- Step 2: Compute $s+$ which is the sum of the ranks of the values with a positive sign
- If H_0 is true, $p(s+)$ will follow a specific distribution
- Note the maximum value of $s+$ is $\frac{n(n+1)}{2}$ as this is the sum of the natural numbers from 1 to n
- We can test if $\mu > 0$, $\mu < 0$ or $\mu \neq 0$, and use a table to look up the critical value for any significance level α

31 Correlation

- Sample correlation coefficient $= r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$
- r is the estimator for the true population correlation coefficient ρ
- We can test whether ρ is greater, smaller or equal to 0
- The null hypothesis is that $\rho = 0$
- We look at test statistics $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ which follows a t distribution with $n-2$ degrees of freedom
  - n is the number of data points
- If $n > 30$, $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ tends towards the normal distribution

32 Linear Regression

- The aim of linear regression is to find β_0, β_1 and σ^2 such that $y = \beta_0 + \beta_1 x + \epsilon$ where ϵ is a normally distributed random variable with mean 0 and standard deviation σ
- By minimising the sum of squared deviations we can find an expression for $\hat{b}eta_1$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
- We can obtain $\hat{\beta}_0$ using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$
- Estimate of σ^2 is $\hat{\sigma}^2 = s^2 = \frac{\text{Sum of squared errors}}{\text{Degrees of freedom}} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$
  - This represents the noise strength
  - The sum of squared errors = SSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$
- How strong is the linear relationship?
  - The total sum of squares = SST $= \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$
  - Coefficient of determination $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$
    - This represents that proportion of observed variation in y that can be explained by the linear regression
    - r^2 is between 0 and 1
    - r^2 is the square of sample correlation coefficient r