|  |  |
| --- | --- |
| **Lecture 01: Data visualisation with ggplot** |  |
| **Importing data**  **Data types**  **Visualising Data**  **Tips** | It is available in the **DAAG** package or through this link<https://git.io/vhse1>.  **# install.packages("readr")**  library(readr)  ais = readr::read\_csv("https://git.io/vhse1")  To find out more about the ais data set use  **# install.packages("DAAG")**  help(ais, package = "DAAG")  %>% means then   1. Capitalise Boolean FALSE TRUE 2. When indexing, start at 1 and selection is inclusive (pok[2:4] is 2 to 4 inclusive). 3. With vectors, to select individuals, use pok[c("one","two")] 4. A Factor is a statistical data type used to store categorical variable 5. Vectors and matrices need to contain the same type (items). |

|  |  |
| --- | --- |
| **Lecture 02: Collecting data** |  |
| **Sample Surveys**  **Why not observe the whole population?**  **Survey sampling**  **Issues to consider:**  **Selection bias**  **Way you frame the question**  **Measurement bias**  **Bias**  **To combat, use a randomised controlled double- blind study**  **Sometimes we cannot do a controlled experiment, so rely on observational study**  **Misleading hidden confounders**  **Strategy for dealing with confounders**  **Controlling for confounding**  **Simpson's paradox**  **Main lesson** | * A **sample** is part of a **population** * A **parameter** is any summary number, like an average or percentage, that describes the entire population. * Usually a parameter cannot be determined exactly, but can only be **estimated** because we cannot survey all the population to get our parameter * A **statistic** can be computed from a sample, and used to **estimate** a parameter. * A **statistic** summarises what the researcher **knows**. A **parameter** is what the **researcher wants** to **know.** * When estimating a parameter, one major issue is **accuracy**: how **close** is the **estimated statistic** to the (unknown) **true parameter**?   Typical limitations   * Hard to observe the population * Not enough time * Not enough money  **Survey design**  * What survey design is appropriate for my study? * How survey will be conducted/implemented?   Sampling is the process of selecting a subset of observations from an entire population of interest so that characteristics from the subset (sample) can be used to draw conclusion or making inference about the entire population.  **Sampling procedure**   * What **sample size** is needed for my study? * How the design will affect the sample size? * Appropriate survey design provides the **best estimation** with high reliability at the lowest cost with the available resources.   When a selection procedure is biased, taking a larger sample DOES NOT help. This just repeats the basic mistake at a larger scale.  Can influence the response from a person, hence not independent. Avoid   * Recall bias (forgetful or prefer positive outlook) * Sensitive questions (who is asking the question) * Misinterpret the questions * Wording of question * Response bias (most people don't pick up phone calls from strangers) * Other attributes of the interview as a source of bias   Sample needs to be representative of the population.  **Bias** is any factor that favours certain outcomes or responses, or influences an individual's responses. Bias may be unintentional (accidental), or intentional (to achieve certain results).  When looking at data from a survey think about   1. **Selection bias / sampling bia**s: the sample does not accurately represent the population. Example: Attendees at a Star Trek convention may report that their favorite genre is science fiction. 2. **Non-response bias:** Certain groups are under-represented because they elect not to participate. Example: a restaurant may give each table a "customer satisfaction" survey with their bill. 3. **Measurement or designed bias:** Bias factors in the sampling method influence the data obtained. Example: a respondent may answer questions in the way she thinks the questioner wants her to answer.   Double blind means that the person administering the drug and the person receiving the drug don't know if the drug is placebo. The need for observational studies  * By necessity, many research questions require an **observational study**, rather than a controlled experiment. * For example, with a study on the effects of smoking, investigators cannot choose which subjects will be in the treatment group (smoking). Rather, they must **observe** medical results for the 2 groups. * Similarly, most **educational research** is based on observational studies. * The conclusions of observational studies require great care.   **\*\*\*\*Observational studies cannot establish causation**   * A good **randomised controlled experiment** can establish ***causation****,* an **observational study** can only establish ***association****.* * *An observational study may suggest causation, but it can't prove causation.*   Confounding occurs when the **treatment group** and **control group** differ by some **third variable** (other than the treatment) which **influences** the **response** that is studied.   * Confounders can be **hard to find**, and can **mislead about a cause** and **effect** relationship. * Confounding (or lurking) variables can be introduced into a randomised study if any of the **subjects drop out**, causing **selection bias** or **survivor bias**. Similarly, if **not all subjects keep taking the treatment or placebo**, we get the confounding of **adherers** and **non-adherers.**   Sometimes we can make the groups more comparable by dividing them into subgroups with respect to the confounder.  For example, if alcohol consumption is a potential confounding factor for smoking's affect on liver cancer, we can divide our subjects into 3 groups:   * heavy drinkers * medium drinkers * light drinkers.   This is called **controlling** for alcohol consumption.  We can control for confound by making 3 separate comparisons:   * heavy drinking: smokers vs non-smokers * medium drinking: smokers vs non-smokers * light drinking: smokers vs non-smokers   What are the limitations of this strategy?   * This strategy is limited by our ability to identify all confounders and then divide the study by the confounders. * This explains the long time to establish that smoking causes lung cancer. Researchers needed to control for factors such as health, fitness, diet, lifestyle, environment etc.   Observational studies with a confounding variable can lead to Simpson's paradox.   * Simpson's paradox occurs when there is a **clear trend** in **individual groups** of data that **disappears** when the **groups are pooled together**. * It occurs when relationships between percentages in subgroups are reversed when the subgroups are combined because of a confounding or lurking variable. * The association between a pair of variables (X,Y) (e.g. smoking and mortality rate) reverses sign upon conditioning of a **third variable Z**, regardless of the **value taken by Z**.   It was believed that smoking has a protective effect until we controlled the confounding variable (age) which then demonstrated the relationship between more cigarettes = more deaths. The problem that arose initially was grouping all ages together, which reversed the individual group trends.  As there are many more young women who smoked than older women, and as younger women are expected to live longer than older women, adding all the groups together makes smoking appear to be beneficial.      In age group 75+ and 65-74, the proportion of smokers to non-smokers was very low which may due to the smokers having already died. There were a lot more young smokers than old smokers (e.g. 60 young smokers and 1 old smoker). The old smokers die, which is less than the death of old non-smokers. There were more old people in non smokers compared to young non-smokers. Hence, the data is skewed towards young smokers, resulting in a lie. |

|  |  |
| --- | --- |
| **Lecture 3: Chi-squared tests** |  |
| **Explanatory variable**  **Response variable**  **Null hypothesis**  **Alternative hypothesis**  **P-Value**  **WTF is a chi square test? Outputs the test statistic and critical value**  **Genetic Linkage**  **No Linkage**  **WTF is a test statistic?**  **Test statistic (output is number)**  **Is 18 enough evidence for or against the null hypothesis? To determine this, simulate it.**  **Simulate**  **Is there a way to do it without simulation?**  **A X2 (chi-square) test!**  **Workflow: Chi squared goodness of fit test**  **Table for calculating the test statistic**  **No Linkage Model**  **Linkage Model**  **Linkage Model simulation**  **Calculate observed test statistic** | In an experimental study, the **explanatory variable** is the **variable** that is manipulated by the researcher. The x-axis or input (dependent on researcher).  The variable that varies depending on other variables. It is on the Y-axis.  The null hypothesis states that there is no relationship between the [two variables being studied](https://www.simplypsychology.org/variables.html) (one variable does not affect the other).  States that the independent variable did affect the dependent variable, and the results are significant in terms of supporting the theory being investigated (i.e. not due to chance).  Def 1: P-value states the level of significance/probability (between 0 and 1) necessary to disprove the null hypothesis. A p-value of less than 0.05 is statistically significant. Does not mean that it is true…  Def 2: P-value is the probability of obtaining a **sample** **as** or **more extreme than** the **observed sample** assuming the **null hypothesis is true**.  We use a chi square test on only **categorical** data (not numerical) and each category needs 5 or more values. We use X2 to see if we can reject the null hypothesis and maybe accept the alternative hypothesis. We want to check if the association between two variables is random or not.  The chi square test **outputs a p-value**, which we can use to determine whether there is a dependence relationship between 2 variables.  If P<0.05, we say that the variables are dependent on each other.  **Degree of Freedom** = (rows − 1) × (columns − 1) or # of values/outputs in the data set (sample size) - 1  rows and columns are our variables     * **Null hypothesis:** each of the phenotypes are equally likely. * **Alternative hypothesis:** the phenotypes are not equally likely.       A test statistic is a random variable that is calculated from sample data and used in a hypothesis test. You can use **test statistics** to **determine whether to reject the null hypothesis**. The test statistic compares your data with what is expected under the null hypothesis. The test statistic is used to calculate the p-value.    ei= expected outcomes (expected result to see if null hypothesis is true)  yi = observed data  Under the null hypothesis, the counts are uniformly distributed across the 4 categories.  Fixing the sample size at *n=400* we can **simulate data** assuming the null hypothesis is true.    As sample is a random function, we need to set the seed or else we get different results each time we run. X is the vector containing our population, size = size of the integers to choose. E.g. if X is 4, size = 2, we choose 2 out of 4.    cex means number indicating the amount by which plotting text and symbols should be scaled relative to the default. 1=default, 1.5 is 50% larger, 0.5 is 50% smaller, etc. par() specifies parameters.    sim\_test\_stats is where we initialise a vector where we add into it. Think arraylist.  **The above chart** shows the shape of the distribution of test statistics under the null hypothesis is true. This chart (our observed sample given NH is true) shows that our original test statistic of 18 is unlikely to occur in the graph so we should reject the null hypothesis, though we do not know if our test statistic is statistically significant (probability of obtaining a **sample** **as** or **more extreme than** the **observed sample** assuming the **null hypothesis is true)** . To find this out, we need to find the p-value.    This suggests that our original TS is statistically significant, so against null hypothesis.       * One categorical variable from a single population * Want to see if it follows a hypothesised distribution               In stats, as populations can be quite big, we do not use p. Instead, we use p hat which takes a random sample of a population and then on that sample population, we calculate p hat ( the probability of a certain event happening in a random sample).            X2 (4 is the four phenotypes, -1 is usual, but the last one is 1 as we needed to estimate the parameter of p hat). Hence, the coupling phase linkage model is correct.    \*\*\*\* you need to calculate the p-value as R is stupid and thinks that our df is 3 when it's supposed to be 2. |

|  |  |
| --- | --- |
| **Lecture 4: Goodness of fit tests** |  |
| **Goodness of fit tests for discrete distributions**  **Radiation exposure**  **PURPOSE: We want to test whether the random variable generating this data follows a Poisson distribution.**  **Poisson distribution**  **Chi-squared tests for discrete distributions**  **Radiation exposure**  **Hypothesis test**  **In R**  **R packages and functions** | The goal in biological dosimetry is to estimate the dose of **ionizing radiation**, absorbed by an exposed individual by using chromosome damage in peripheral lymphocytes.  When radiation exposure occurs, the damage in DNA is randomly distributed between cells producing chromosome aberrations (deviates from normal type). The outcome of interest is the number of aberrations observed. The number of aberrations typically follows a Poisson distribution, the rate of which depends on the dose.  The table below shows the number of chromosome aberrations from a patient exposed to radiation after the nuclear accident of Stamboliyski (Bulgaria) in 2011 (Puig and Weiß, 2020).    A **Poisson** random variable represents the probability of a given number of events occurring in a fixed interval (e.g. number of events in a fixed period of time) if these event occur independently with some known average rate λ per unit time.  If X is a Poisson random variable with rate parameter λ, the probability mass function is:    %>% pipe into        tilde means has the distribution of. E(X) expected value and Var(x) variance.    248 because sum of i\*yi = (0\*117+1\*94+2\*51+3\*15+7\*1) = 248  We expect number of counts is 5, which is part of assumption in the chi square test.      df is 2 because we have 4 categories (i column), hence 4-1 and we need subtract one again because we needed to estimate lambda hat.  # We do not say that we proved that the data is a Poisson distribution. We say that we do not reject the null hypothesis, meaning there is no evidence against the poisson distribution.      A: Because it does not realise that we estimated the lamda, hence the df is supposed to 2. |

|  |  |
| --- | --- |
| **Lecture 5: Measures of Performance** |  |
| **Breast Cancer**  **Test result vs actual status (note table order is important!)**  **Breast cancer**  **(FN is the worst)**  **Extra notation**  **Table with counts of outcomes**  **# Read pipe right to left, given D+, what is Probability of S- occurring.**  **# Put this in your formula cheat sheet for final exam**  **Breast cancer**  **Conditional probability**  **Bayes' Rule**  **Bayes' Rule**  **Example**  **Using Bayes' formula**  **NIPT**  **Problem with accuracy** | How **good** is the test for breast cancer? How you define **good**? If you test positive, what are the chances that you actually have breast cancer?      Let's formalise this. Let,      Prevalence = All Actual positive D+/all numbers    P(A|B) means inside the circle B, what part is A as part of B?          Non-invasive prenatal tests (NIPT) are an increasingly popular way of screening for chromosome conditions and advertise as having very high **accuracy**.  Taylor-Phillips, Freeman, Geppert, et al. (2016) perform a meta-analysis on the accuracy of non-invasive prenatal testing using cell-free DNA for detection of Down syndrome.  They broke their findings down by "general obstetric population" and "high risk population".  If prevalence ( all people who have disease/ whole population) is low, you will have a very high accuracy as most people(True negative) do not have the disease and an extremely few people (TP) have the disease.    For the general obstetric population, precision is 81.6% and high risk is 91.3%. The question is should the government fund this test for everyone or just for those in the high risk population. Since prevalence is only 0.4% for general obstetric, so gov should fund only high risk population. |

|  |  |
| --- | --- |
| **Lecture 6: Measures of risk** |  |
| **Asthma and hay fever**  **Hodgkin's disease and tonsillectomies**  **Notation**  **Prospective (or cohort study) studies**    **Prospective studies -- fictitious example**    **Prospective study -- Asthma**  **We originally fixed the number of people in R+ and R-, but at the end of study, we find column ending.**    **Retrospective (or case control) studies**    **Retrospective studies -- fictitious example**    **Retrospective study -- Hodgkin's disease**    **Estimating a population proportion**    **Application to prospective and retrospective studies**    **Prospective study**    **Retrospective study (see the difference with prospective)**    **Measures of risk**    **Relative risk (1 means the risk of having the disease in at risk group equals the risk of having the disease in at non-risk group. < 1 means if you are less times likely to have the disease if you are in at risk group. (better be at risk, than in non-risk group)**  **First point: this means if the probability of having the disease given you are not at risk is low, this implies that the relative risk is very high.**    **Relative risk -- interpretation**    **Relative risk -- prospective studies**  **Relative risk -- retrospective studies**    **Aspirin (relative risk)**  **Odds ratio ( a common way to measure risk)**  **Equivalent definitions of odds ratio**    **Odds ratio -- invariance**    **Odds ratio -- interpretation**    **Aspirin (odds ratio)**    **Standard errors and confidence intervals**    **Aspirin**  **For log odds-ratio interval, if it contains 0, this indicates there is no relationship.**  **For odds-ratio interval, if it contains 1, this indicates there is no relationship.**    **Hodgkin's disease**  **(This is a retrospective study because we select the diseased and the non-diseased)**  **The odds ratio of 2.93 means odds of a tonsillectomy patient having Hodgkin's disease are roughly 3 times the odds of a non-tonsillectomy patient having Hodgkin's. This result is significant at 5% level of significance.** | Prospective means we first choose a population disease free and see how they develop based on a certain factor.     * A **prospective study** was designed to assess the impact of sun exposure on skin damage in beach volleyball players. * During a weekend tournament, players from one team wore waterproof, SPF 35 sunscreen, while players from the other team did not wear any sunscreen. * At the end of the volleyball tournament players' skin from both teams was analyzed for texture, sun damage, and burns. * Comparisons of skin damage were then made based on the use of sunscreen. * The analysis showed a significant difference between the cohorts in terms of the skin damage.        * There is a suspicion that zinc oxide, the white non-absorbent sunscreen traditionally worn by lifeguards is more effective at preventing sunburns that lead to skin cancer than absorbent sunscreen lotions. * A **retrospective study** was conducted to investigate if exposure to zinc oxide is a more effective skin cancer prevention measure. * The study involved comparing a group of former lifeguards that had developed cancer on their cheeks and noses (cases) to a group of lifeguards without this type of cancer (controls) and assess their prior exposure to zinc oxide or absorbent sunscreen lotions. * This study would be **retrospective** in that the former lifeguards would be asked to recall which type of sunscreen they used on their face and approximately how often.     Suppose   * we have a large (but finite) population containing objects/individuals of two different types (say type 0 and type 1); * it is desired to determine or at least estimate the overall proportion of type 1 but it is not feasible to examine every object/individual.   If we can take a random sample from the population then we can use the sample proportion of type 1 as an estimate of the population proportion of type 1.  Extending this idea, consider two events *A* and *B*,   * If we can take a random sample from the whole population, we can estimate P(A)using the observed sample proportion with attribute A * If we can take a random sample from the **subpopulation** defined by B, we can estimate P(A|B) using the observed sample proportion (of the subpopulation) with attribute A.   In both kinds of study we have   * a population; * a subpopulation/attribute determined by a risk factor R+ (with complementary subpopulation/attribute R-); * an subpopulation/attribute determined by having/developing the disease D+ (with complementary subpopulation/attribute D-).   The labels "subpopulation" and "attribute" here are mathematically equivalent (they both mean event).  The main difference between prospective and retrospective studies are which (sub)populations we can sample from.   * In a prospective study we take **2** random samples:   + one from the **risk factor** group (subpopulation) R+   + another from the **non-risk** factor group R- * We then (wait to) see how many in each group develop the disease. * We can thus estimate P(D+|R+) as well as P(D-|R-). * \*\*\*\*\* We **CANNOT** however estimate **P(R+|D+)** or **P(R-|D-)** since we did not take random samples from the disease group. * In a retrospective study we take two random samples:   + one from the disease group (subpopulation) D+ and   + another from the non-disease group (subpopulation) D-. * We then (look back to) see how many in each group were exposed to the risk factor. * We can thus estimate P(R+|D+) as well as P(R-|D-). * We cannot however estimate P(D+|R+) or P(D-|R-) since we did not take random samples from the risk factor group   These are different ways to measure the association between a risk factor/treatment and the disease outcome.  How the data is **sampled** will greatly impact the ways in which these methods are applicable and interpretable.    If risk factor has no influence on disease, so RR is approximately equal to 1.                        x hat +- 1.96 \* SE(x hat) |

|  |  |
| --- | --- |
| **L7: Testing for homogeneity** |  |
| **Testing for homogeneity in 2 x 2 tables**  **Children’s TV**  **Test of homogeneity**  **Read down column for population (boys or girls)**  **Two way contingency table in R**  **Notation**  **Chi-squared test of homogeneity**  **Hypothesis testing workflow**  **Children’s TV**  **Testing for homogeneity in general tables**  **Example: Voters**  **A general two-way contingency table**  **Test of homogeneity in general two-way tables**  **Test of homogeneity**  **Degrees of freedom**  **Hypothesis testing workflow**  **# df is different. Test stat & p-value.**  **Example: Voters**  **In matrix, use byrow = FALSE to add/fill by columns first**  **Link with risk measures: Children’s TV** | In a study of the television viewing habits of children, a developmental psychologist selects a random sample of 100 boys and 200 girls of preschool age. Each child is asked which of the following TV programs they like best: Sesame Street or Play School. Results are shown in the table below.    **Is there any evidence that the viewing preferences of boys and girls are different?**   * Suppose that several samples are taken from two independent populations, each of which is categorised according to the same set of variables. * We want to test whether the probability distributions (proportions) of the categories are the same over the different populations.   In our children's TV example, we could consider the proportions of boys who prefer each show and (separately) the proportion of girls who prefer each show.            Because our degrees of freedom for test of homogeneity = (row-1)(column-1)        In R, pchisq(0.027, 1) is the left area under the curve of 0.027. To get our p-value, 1-pchisq(0.027, 1). |

|  |  |
| --- | --- |
|  |  |
|  |  |