Research Education Seminar Series

Lecture 3

Health
Northern NSW
Local Health District

RESEARCH!

Alex Stephens, PhD, Director of Research

Basic outline

1. Research Overview

- Overarching aims of the research office
- Why do we engage in research?
- Simplified research framework
- Research translation framework
- Evidence-based practice pyramid

2. Robust study design

- Epidemiological methods – study designs
- Key elements of study validity and critical appraisal
- Measures of association
- Determining sample size – power analysis

3. Analysis

- Basic biostatistical methods and analysis

4. Writing for research

- Writing grants, papers and scientific presentations

5. Ethics, governance and software

- Research ethics and governance
- Research software

All **data** are **hypothetical** and were generated for the purposes of demonstration and examples only.

Analysis of continuous (normally distributed) data

Independent (two) samples t-test

A very common way to assess a difference between two groups is to use a t-test. These are called independent or two samples t-tests.

The objective of the two sample t-test is to infer if there is a significant difference between two population means based on samples derived from the populations - **testing for the difference between two means**

Test statistic

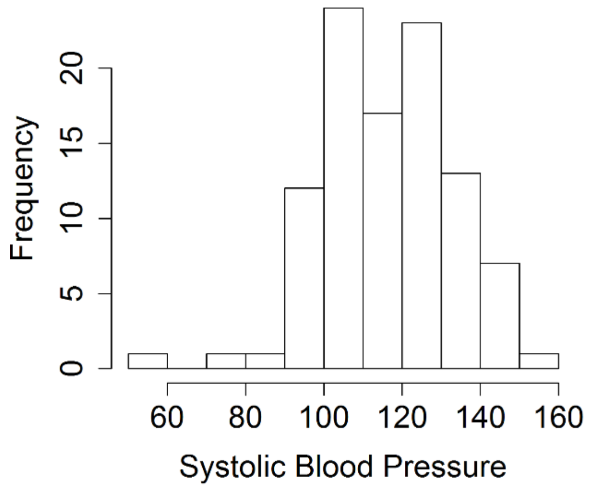
$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\bar{x}_1 - \bar{x}_2}}$$

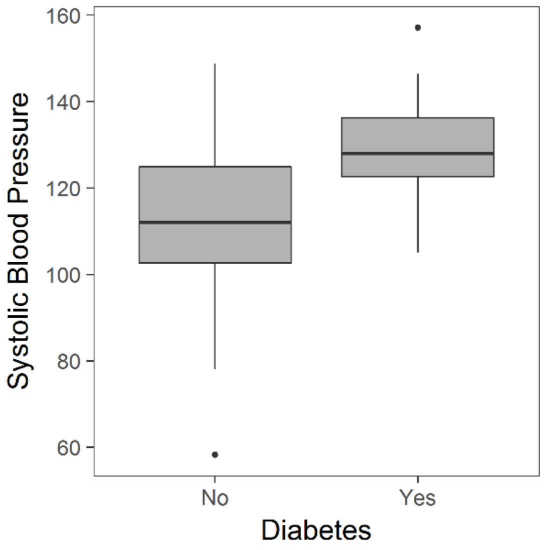
P-values are obtained by comparing the test statistic to critical values of the Student's t-distribution or is calculated more precisely using statistical software.

Perhaps more importantly, confidence intervals can be calculated for mean differences providing a range of plausible values of the true mean difference.

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha,df} \times SE_{\bar{x}_1 - \bar{x}_2}$$

Worked example in R – testing for a difference in systolic blood pressure between those with and without diabetes





```
        Two Sample t-test

data:  bp by Diab
t = -3.8869, df = 98, p-value = 0.0001847
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -23.123224  -7.492385
sample estimates:
mean in group 0 mean in group 1
       113.3022        128.6100
```

The estimated mean difference is -15.3 (mmHg), with a 95% confidence interval (CI) of -23.12 to -7.49. The interpretation of the 95% CI is that, in a long series of identical repeat experiments, the 95% CI will contain the true population difference on 95% of occasions (the true population difference is either in the interval or not). In this example, the mean difference is statistically significant as denoted by the 95% CI, which does not include the null value of 0, and the p-value which is less than 0.05.

Assumptions

- The samples are independent → the selection of observations in one sample does not influence the selection of observations in the other sample

- The populations from which the two samples were derived follow normal distributions → this in turn means that $\bar{x}_1 - \bar{x}_2$ follows a normal distribution and the properties of the normal distribution can be applied for statistical hypothesis testing – assess normality using histograms, P-P plots or Q-Q plots

- Ideally, the populations should have equal variances → known as homogeneity of variances

<u>What to do in the case of un-equal variances?</u>

The trick here is to use an adjustment that derives an approximate effective “degrees of freedom” for use in the t-test. This is termed Welch’s t-test and is produced by default when performing analysis in some statistical programs (such as SPSS).

One-way analysis of variance (ANOVA)

The purpose of the one-way analysis of variance is to determine if there are any significant differences between the means of three or more groups. It is termed analysis of **variance** as the procedure compares **two** estimates of the population variance through an *F*-ratio.

Stat's corner

You may recall from your introductory statistics classes that the process of ANOVA involves calculating several sources of variance in the form of sum of squares (SS). These include the Total SS, Regression (or Between) SS and Error SS. We know that the Error SS divided by the degrees of freedom (*df*) provides an estimate of the population variance (σ^2), often termed the mean square error (MSE)

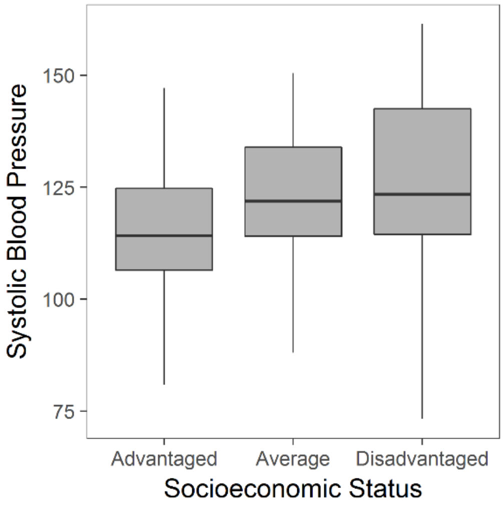
$$MSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 \Bigg/ \sum_{i-1}^{k}(n_i - 1)$$

Under the null hypothesis where all group population means are equal $\mu_1 = \mu_2 = \cdots = \mu_k = \mu$, the Regression SS divided by its *df*, the Mean Square Regression (MS(Reg)), also provides an unbiased estimate of σ^2 as follows

$$MS(Reg) = n\left[\sum_{i=1}^{k}(\bar{Y}_i - \bar{Y})^2 \Big/ k - 1\right]$$

If the groups are sampled from the same population (i.e. the null hypothesis), then the *F*-ratio should be close to one. Conversely, if the groups are sampled from populations with different means, the *F*-ratio will be larger than 1.

Worked example in R – testing for a difference in systolic blood pressure between socioeconomic status groups



```
                   Analysis of Variance Table

Response: bp
               Df  Sum Sq Mean Sq F value  Pr(>F)
as.factor(SES)  2  1991.2  995.62  4.0332 0.02077 *
Residuals      97 23945.0  246.86

Mean difference = 8.57 for average SES and 10.0 for
disadvantaged SES
```

MS(Reg) = 995.6, and MSE = 246.9. F ratio = 995.6/246.9 = 4.03. P-value = 0.021 i.e. significant effect – so a significant difference between, at least, two group means.

Assumptions

- The observations are independent
- The errors are normally distributed → this is inferred by normality of the residuals. Alternatively, can be inferred by assessing normality within each group.
- The variances of the data across the groups are equal (homogeneity of variances)

What to do when assumptions are violated?

- ANOVA can tolerate moderate departures from normality of errors. If a histogram of the residuals is clearly non-normal, apply a log transformation and see if this improves normality, or use a non-parametric test (e.g. Kruskal-Wallis test).
- ANOVA is fairly robust against violations of equal variances in balanced designs → so preferably, try to have equal sample sizes for each group
- If the design is unbalanced, and there is evidence to suggest heterogeneous variances, try a log transformation or use a non-parametric test that does not rely on distributional assumptions.

Post-hoc testing following a significant omnibus test

When ANOVA generates a statistically significant *F*-statistic, the convention is to follow the test by performing multiple pairwise comparisons through post-hoc testing to determine where the significant mean differences lie. Inflation of *type I* error (false positive) is a major concern when conducting multiple statistical tests and numerous adjustments to control type I error rates have been proposed → e.g. Bonferroni correction, Tukey's HSD. There is no consensus on which test is the most appropriate to use and people are free to read the literature and decide on which test they feel is most suitable.

In general, it is important to limit the number of comparisons made in an experiment and thought needs to be put into experimental design such that comparisons that are made are justifiable, and necessary to address study hypotheses.

Simple linear regression

The main objective of simple linear regression is to examine the relationship between an outcome (or dependent variable) and one predictor (or independent variable). The relationship is described by a mathematical model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Essentially, the observed value of the outcome for individual *i* (Y_i) is given by a linear combination of fixed-effects → the intercept (β_0) and the coefficient representing the effect of the independent predictor (β_1) – and the error term (ε_i). The relationship between the outcome and predictor variable is “linear” (i.e. a straight line) in simple linear regression.

The parameters in the regression model are most commonly estimated using ***least squares***, a method that minimises the sum of squared deviations from the fitted line (or fitted values) and observed values.

Stat’s corner

To estimate model parameters, firstly obtain an equation for the sum of squared deviations:

$$SS_{dev} = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2$$

Then differentiate with respect to β_0 and β_1 to obtain the least squares solutions, which, after a number of algebraic steps (particularly for b_1) are:

$$b_0 = \bar{Y} - b_1 \bar{X} \qquad b_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

The working model after estimation yields:

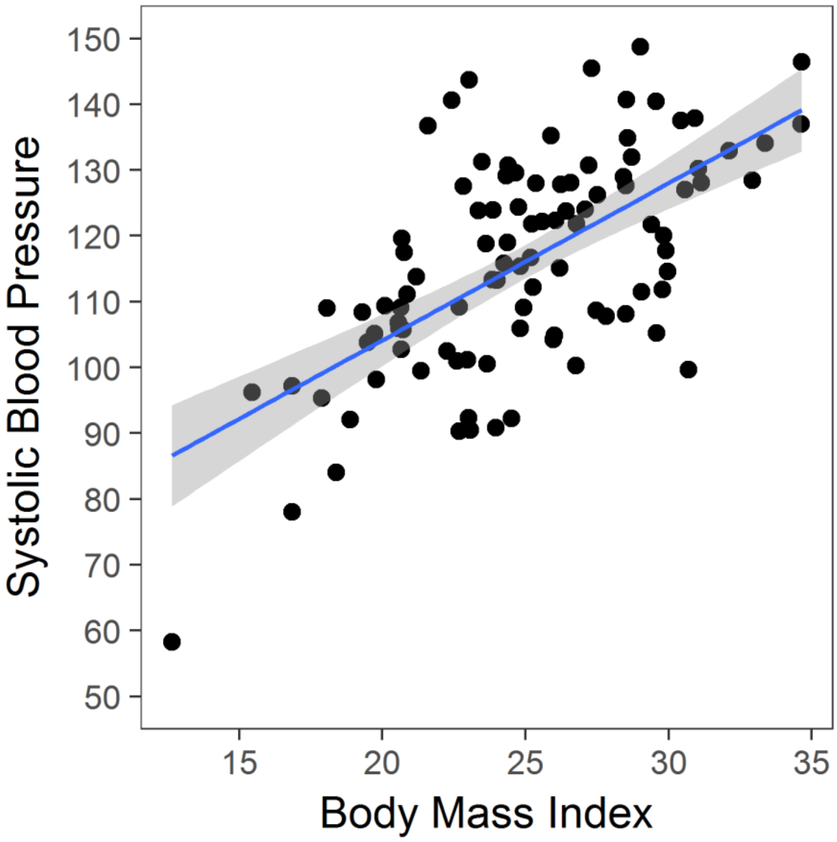
$$\hat{Y}_i = b_0 + b_1 X_i$$

Where *Y-hat* represents the fitted value. In a way analogous to ANOVA, we can partition the variance into the *MS(Reg)* and *MSE*, which are termed the *MS(Model)* and *MS(Residual)* in regression, as follows:

$$MS(Model) = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \Big/ \left(\#\ of\ predictors\ (p)\right)$$

$$MS(Residual) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \Big/ (n - p)$$

Worked example in R – the relationship between systolic blood pressure and body mass index (BMI)



```
                    Regression coefficients

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 116.2446     1.2859  90.399  < 2e-16 ***
BMI_C         2.4774     0.2924   8.472 2.47e-13 ***
```

The table (above) shows the estimated parameters. The intercept represents the expected (or estimated) systolic blood pressure at a BMI of 25 kg/m^2 (the trick here is to centre the variable on 25 such that a BMI of 25 is represented by 0 in the regression analysis). The estimated coefficient for BMI_C (centred BMI) is ~2.5, and indicates that, for every 1-unit increase in BMI, systolic blood pressure increases, on average, by 2.5 mmHg.

Multiple linear regression – the general linear model

In more complicated statistical analyses, sometimes our goal is to simultaneously estimate the effects of multiple predictors on a continuous outcome. A tool to achieve this is multiple linear regression, which is sometimes referred to as the general linear model. These models are simply an extension of simple linear regression by including additional predictors and take on the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

A major advantage of multiple regression is that effects of predictor variables (also termed covariates) are estimated while adjusting for all other predictors in the model. The estimated coefficients end up being a form of weighted averages of strata-specific estimates; with strata groups representing the different combinations of covariates (i.e. different covariate patterns). This is often referred to as adjusting or controlling for covariates or confounding factors.

Estimation of coefficients (parameters) in a multiple regression model is also achieved via least squares, and is derived using matrix calculations and the differentiation of a function of a vector quantity (a little more complex than simple linear regression!).

Worked example in R – the relationship between systolic blood pressure and diabetes status, SES and body mass index (BMI)

```
                Analysis of Variance Table

Response: bp
                Df  Sum Sq Mean Sq F value    Pr(>F)    
as.factor(Diab)  1  3749.3  3749.3 34.2885 6.781e-08 ***
as.factor(SES)   2  1985.2   992.6  9.0778 0.0002468 ***
BMI_C            1  9814.0  9814.0 89.7531 2.202e-15 ***
Residuals       95 10387.7   109.3                      

                 Regression coefficients

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      110.5242     1.8201  60.724  < 2e-16 ***
as.factor(Diab)1  13.9699     2.6223   5.327 6.68e-07 ***
as.factor(SES)1    9.2862     2.4872   3.734 0.000322 ***
as.factor(SES)2   13.5721     2.6322   5.156 1.37e-06 ***
BMI_C              2.2762     0.2403   9.474 2.20e-15 ***
```

Interpretation: The intercept is 110.52, which provides an estimate of mean systolic blood pressure when all covariates are set to reference levels (i.e. non-diabetic, advantaged SES, and a BMI of 25, remembering that BMI was centred to 25 and therefore 25 is the reference level). (Diab)1, (SES)1 and (SES)2 provide estimates of the mean difference in systolic blood pressure for diabetes, and average and disadvantaged SES, respectively. BMI_C is the estimated change in systolic blood pressure for every 1 unit change in BMI. How do these estimates compare to the estimates calculated using the individual (univariate) statistical methods above? → they are somewhat similar and change slightly due to adjustment for each other in the one model.

Categorical data analysis

Comparison of two or more proportions or categorical variables

A very common way of measuring outcomes is as counts of categorical variables or proportions derived from count data. Categorical data can be analysed as proportions (using a Z-test based on the normal approximation of the binomial distribution) or as the raw values themselves (the actual observed counts).

Analysis of observed counts can be achieved using the ***Chi-square test of independence***. The Chi-square test is a non-parametric test, in that it does not rely on the shape or form of the underlying distribution from which the data was derived. It also has the flexibility to include numerous levels within each of the categorical variables being compared (i.e. can handle larger than 2×2 tables). The test statistic for the Chi-square test of independence is:

$$\chi^2 = \sum {(O_i - E_i)^2}/{E_i}$$

Where

$$E_i = \frac{row\ total \times column\ total}{sample\ size}$$

The χ^2 test statistic *approximately (or asymptotically)* follows a chi-square distribution (χ^2_α) with $(r-1)(c-1)$ degrees of freedom where $r =$ the number of rows and $c =$ is the number of columns in the contingency table.

For example, let's explore diabetes status by socioeconomic status with the data as follows:

Observed counts

	Advantaged	Average	Disadvantaged	Total
Diabetic	7	8	5	20
Non diabetic	29	27	24	80
Total	36	35	29	100

Expected counts

	Advantaged	Average	Disadvantaged	Total
Diabetic	7.2	7	5.8	20
Non diabetic	28.8	28	23.2	80
Total	36	35	29	100

$$\chi^2 = 0.32345; \quad df = (2-1)(3-1) = 2$$

Right tail probability of a $\chi^2 = 0.32345$ at 2 degrees of freedom = 0.8507 (this is the *p*-value)

Worked example in R – the relationship between SES and diabetes status

```
             (R) Table of frequencies

                          SES
Diabetes Advantaged Average Disadvantaged
   No            29      27            24
   Yes            7       8             5

            Pearson's Chi-squared test

data:  M
X-squared = 0.32345, df = 2, p-value = 0.8507
```

When the sample size is small, say when at least one of the observed counts in the cells of a contingency table is less than 10, it's often useful to apply **Fisher's exact test** instead of the chi-square test, as the chi-square approximation tends to be poor. Fisher's exact test was originally devised for 2×2 tables, but, with modern computing power, can now handle larger tables (although, the calculation is likely to be based on some sort of approximation so the test is no longer exact!).

Applying Fisher's exact test for the diabetes and SES example (above) generates:

```
        Fisher's Exact Test for Count Data

data:  M
p-value = 0.9052
alternative hypothesis: two.sided
```

Comparison of two proportions using the Z-test

When the sample size and/or proportion are large enough (again, around a cell size of 10 or more in a 2×2 contingency table), it's possible to use the normal approximation of the binomial distribution to compare two proportions. The test statistic is as follows:

$$Z = \frac{p_1 - p_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

This test statistic squared (Z^2) is, in fact, mathematically equivalent to the 2×2 table chi-square statistic. The Z-statistic is used to compute a *p*-value, allowing statistical significance to be assessed. Furthermore, 95% CIs can be constructed providing an interval for the likely true value of the population difference in proportions.

Z-test in R – diabetes by overweight status

Using a Z-test to compare the prevalence of diabetes in normal and overweight individuals

```
                          Data

              Diabetes
Weight         Yes No |   Pr
  Normal        15 75 | 0.167
  Overweight    25 50 | 0.333

     2-sample test for equality of proportions
           without continuity correction

data:  N
X-squared = 6.1875, df = 1, p-value = 0.01287
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.29823519 -0.03509814
sample estimates:
   prop 1    prop 2
0.1666667 0.3333333
```

Binary logistic regression

Binary logistic regression belongs to the family of ***Generalized Linear Models***. For these models, the broad structure of multiple linear regression (the general linear model; described above) is extended to handle other types of outcomes (e.g. binary, counts, rates). The main requirement is that the outcomes have distributions that belong to the exponential family (e.g. binomial, Poisson and normal distributions). In the case of binary logistic regression, the outcomes are based on the binomial distribution, and is useful for analysing data where the study outcome is binary in nature (e.g. male versus female; “Yes” versus “No”; present versus absent).

Logistic regression has the general form:

$$logit(\pi_i) = \beta_0 + \beta_{1i}X_{1i} + \cdots + \beta_{pi}X_{pi}$$

Where

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

And the binomial (discrete) probability distribution is:

$$P(x|p,n) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

Therefore, the technique models a function of the probability of the event, and not the actual probability of the event. However, the logit can be back transformed to yield predicted probabilities. Logistic regression is useful for assessing the effects of multiple predictor variables, including continuous and categorical explanatory variables, on the likelihood (or probability) of a binary outcome. Maximum likelihood is used to estimate the model parameters (the β coefficients) and, when exponentiated, represent the estimated effects of individual predictor variables on the outcome in the form of odds ratios (ORs).

Worked example in R – effects of BMI and SES on the prevalence of diabetes

```
                  Logistic regression coefficients

 Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -2.00418    0.27352  -7.327 2.35e-13 ***
SES_FAverage        0.56635    0.34777   1.629  0.10341    
SES_FDisadvantaged  0.63331    0.34595   1.831  0.06715 .  
BMI_C               0.09063    0.02847   3.183  0.00146 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 374.07  on 399  degrees of freedom
Residual deviance: 359.94  on 396  degrees of freedom
AIC: 367.94
```

Exponentiation of the regression coefficients yields the odds ratios, which describes the nature (positive or negative) and strength of the association between the predictor variables and the outcome. In the example above, the regression coefficients for average SES, disadvantaged SES and BMI (a 1-unit increase) were 0.57, 0.63 and 0.09, respectively, which convert to odds ratios of 1.76, 1.88 and 1.09, respectively. Odds ratios are interpreted as the factor by which the odds of the outcome are changed (relative to the reference level) in the presence of the covariate. For example, in the illustration above, the odds of diabetes are 1.76 times greater in those with average SES compared to those with advantaged SES.

A note on diagnostics – “Goodness of fit”

When fitting binary logistic regression models (or any model for that matter), it’s important to assess how well the specified model explains or fits the data. For binary logistic regression, there are a variety of goodness of fit statistics. These include the ***Deviance*** and ***Pearson Chi-squared*** statistics, and the ***Hosmer-Lemeshow*** statistic. Deviance and Pearson Chi-squared statistics are asymptotically equivalent, but may differ slight when the sample size is small. If the model is specified correctly, both statistics are approximately chi-square (χ^2) distributed with *N*-*p* (sample size minus number of model parameters) degrees of freedom. The Hosmer-Lemeshow test is also commonly used, and works by dividing the data into groups, commonly deciles of predicted probabilities (or some other suitable number of groups), and performing a chi-square test on the observed and predicted values of the outcome across the groups. While these diagnostic goodness of fit statistics may be of some use, they are often underpowered, and it may be preferable to test model fit by thoroughly exploring the functional forms of the relationships between explanatory variables and the outcome (e.g. linear, non-linear, ordinal), and evaluating the addition of appropriate terms (covariates) in models.

Multinomial logistic regression

The concept of binary logistic regression can be extended to cater for more than 2 outcomes. This analysis is termed ***multinomial logistic regression***, and is based on the multinomial probability distribution, which gives the probability of observing y_1, y_2…& y_j outcomes (responses) from *n* independent observations of Y (this essentially means that the variable Y can take on J different outcomes, with y_1 to y_j denoting the numbers of times we observe each of the different outcomes when we sample Y *n* times).

The multinomial probability distribution is:

$$f(\boldsymbol{y}|n) = \frac{n!}{y_1!\,y_2!\ldots y_j!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_j^{y_j}$$

There are two main categories of multinomial logistic regression: ***nominal*** and ***ordinal***. In nominal multinomial logistic regression, it is assumed that there is no defined or natural order of the responses, and the analysis simply works by selecting a reference category for which the odds of all other categories can be generated (e.g. odds of outcome 2 relative to reference, say outcome 1; odds of outcome 3 relative to outcome 1, … odds of outcome j relative to outcome 1). On the other hand, when there is a natural order to the responses, ordinal multinomial logistic regression can be used. This method comes in a variety of forms (e.g. cumulative logit model, proportional odds model) depending on how we chose to order the responses. **Tip:** if you think your data would be best analysed using multinomial logistic regression, consult a statistician as the method is fairly sophisticated, and typically outside the comfort zone of most non-statisticians.

Poisson regression

When the outcome being assessed can be best captured in the form of count data, ***Poisson regression*** can be used to analyse the relationship between a set of predictor variables and the outcome. In this analysis, the underlying probability distribution for the count data, Y, is assumed to be the Poisson distribution which is written as:

$$f(y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

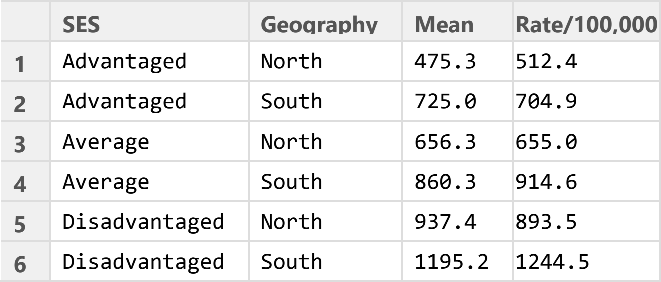
For this distribution, the expected value (or mean), $E(Y)$, is λ, and the variance, $Var(Y)$, is also λ. Therefore, this distribution has the property that as the mean increases, so does the variance.

Examples of outcomes that may be assumed to be Poisson distributed, and therefore would lend themselves nicely to Poisson regression analysis, include the number cancer cases by suburb, number of smoking related deaths by profession, and number of accidents at particular intersection by season (e.g. by spring, summer, autumn and winter). These outcomes are all counts. However, a natural question that arises when analysing count data is how do we account for the different populations at risk or size of exposure? For example, there might be more cancer cases in a particular area simply because there are more people living in the area (i.e. a greater exposure or larger population at risk). The way to account for varying levels of exposure is to express the magnitude of the outcome relative to the population at risk. For example, the rate per 100,000 persons, or rate per 1,000 years at risk. A major advantage of Poisson regression is that it is able to account for different populations at risk (by including a variable called the offset), effectively becoming a tool that can model rates.

Therefore, a useful application of Poisson regression is for the modelling of rates of outcomes/events. For example, Poisson regression might be used to model the rates of hip fractures by socioeconomic status and geography in New South Wales. Like logistic regression, Poisson regression is able to assess the effects of multiple predictor variables, including continuous and categorical explanatory variables, on count outcomes. Maximum likelihood is used to estimate the model parameters (the β coefficients) and, when exponentiated, represent the estimated effects of individual predictor variables on the outcome in the form of risk ratios or rate ratios.

Worked example in R – comparing the number of hip fractures by SES and whether you live in the north or south of New South Wales

Table: Mean number and rate per 100,000 of hip fractures by SES & north/south geography



	SES	Geography	Mean	Rate/100,000
1	Advantaged	North	475.3	512.4
2	Advantaged	South	725.0	704.9
3	Average	North	656.3	655.0
4	Average	South	860.3	914.6
5	Disadvantaged	North	937.4	893.5
6	Disadvantaged	South	1195.2	1244.5

```
   Poisson regression – without offset or dispersion parameter

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        6.241041   0.008789  710.08   <2e-16 ***
SES_FAverage       0.236809   0.009955   23.79   <2e-16 ***
SES_FDisadvantaged 0.576647   0.009354   61.64   <2e-16 ***
Geog_FSouth        0.286860   0.007017   40.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7887.0  on 99  degrees of freedom
Residual deviance: 1600.7  on 96  degrees of freedom
AIC: 2459.5
```

Exponentiation of the coefficients in Poisson regression yields the relative risk or rate ratio, which is a measure of the strength of the association between the predictive factors and the outcome. The relative risk or the rate ratio represents the factor by which the magnitude of the outcome is changed when the covariate is present. So, for the example data (above), the mean number of fractures is, on average, $e^{0.236809} = 1.27$ and $e^{0.576647} = 1.78$ times greater in those with average and disadvantaged SES, respectively, relative to those with advantaged SES. For south geography, the mean number of fractures is 1.33 times greater than those living in northern areas (lack of sunlight?).

To take into account different population sizes, natural log transformed population is included in the Poisson model (as the offset variable). In this version of Poisson regression, the outcome being modelled is actually the rate per unit population (i.e. per person).

```
  Poisson regression – with offset but no dispersion parameter

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.281006   0.008902 -593.22   <2e-16 ***
SES_FAverage        0.253616   0.009961   25.46   <2e-16 ***
SES_FDisadvantaged  0.563971   0.009352   60.30   <2e-16 ***
Geog_FSouth         0.330565   0.007017   47.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6456.846  on 99  degrees of freedom
Residual deviance:   99.151  on 96  degrees of freedom
AIC: 957.92
```

The regression coefficients above, where natural log population is included in the model as the offset, are similar to the non-offset model, with the exception of the intercept. The intercept relates to the magnitude of the outcome being modelled (which is the rate per person in this case and thus is a much smaller value than the actual mean number of fractures). Similarity of coefficients between offset and non-offset models suggests that the population structures are fairly similar between the strata groups defined by the levels of the predictor variables.

Overdispersion

As mentioned above, both the expected value and variance of the Poisson distribution is λ. This is a problem in cases where $E(Y) < Var(Y)$, so where the variance of a random variable, Y, is actually larger than the mean. This is termed overdispersion. Two common approaches are used to handle overdispersion. The first one is to fit an additional parameter, the dispersion parameter (ϕ), during estimation, which inflates that variance (i.e. $Var(Y) = \phi\lambda$). The second option is to use another type of regression termed ***negative binomial regression***, which has greater flexibility in the way the variance is handled. A common approach to check for overdispersion is to construct a ratio of the residual deviance to the residual degrees of freedom. Theoretically, they should be equal in a good fitting model (i.e. ratio = 1). However, ratios larger than 1 may reflect overdispersion (or could, in fact, reflect a poorly fitted model missing important predictor variables). If you think your data may be overdispersed, it's probably best to make an appointment with a statistician. Below is an example of the Poisson regression model fit to the example hip fracture data including both an offset variable and the dispersion parameter.

```
       Poisson regression – with offset and dispersion parameter

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -5.281006   0.009049 -583.57   <2e-16 ***
SES_FAverage        0.253616   0.010126   25.05   <2e-16 ***
SES_FDisadvantaged  0.563971   0.009507   59.32   <2e-16 ***
Geog_FSouth         0.330565   0.007133   46.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.033335)

    Null deviance: 6456.846  on 99  degrees of freedom
Residual deviance:   99.151  on 96  degrees of freedom
AIC: NA
```

Longitudinal and correlated data

The statistical methods we have covered so far have assumed statistical independence between observations or data points. However, it is fairly common to encounter situations where this is not the case, and statistical independence cannot reasonably be assumed. A paired design is a good example of this, where, in its simplest form, a pair of observations is taken from the same individual at two points in time with different exposures at each time point. If we term each pair of observations from the same individual as a cluster, this is an example where the covariate, or study variable of interest, is distributed within clusters. Such data can be analysed using a paired samples t-test, if the differences between the two measures (for each individual) follows a normal distribution.

Effects on covariates distributed within and between clusters

Taking into consideration the correlation among observations can be very powerful when covariates are distributed within clusters (i.e. the different levels of the covariate occur within clusters). For example, a study where the same subject receives the control treatment at one point in time and then the intervention at another point in time is a design where the covariate (treatment in this case) is distributed within clusters. Accounting for the correlation allows more of the variance to be explained, effectively splitting the error variation into the systematic difference between individuals and the remaining residual variation. Therefore, when the correlated data structure is taken into account, the variance (or standard errors) of the estimated effects associated with covariates distributed within clusters decrease, and effects are more precisely estimated.

In the case that all individuals within clusters are given the same treatment/exposure, but treatment/exposure is allowed to vary between clusters, the covariates are distributed between clusters (i.e. levels of the covariates are fixed within clusters but differ between clusters). In these situations, taking into account the correlated data structure (appropriately) increases the standard errors of estimated effects (compared to an analysis not accounting for the correlation) as the correlation within clusters effectively decreases the amount of independent information (e.g. think about the similarity between identical twins – i.e. highly correlated – and consider how independent their height measurements would be).

There are a number of study designs and data structures which involve some level of correlation among observations. Accordingly, there are a number of different statistical methods to analyse such data. These include:

- *Paired t-tests*
- *Summary measures approach for longitudinal data → area under the curve; maximum value; growth rate or linear regression coefficient*
- *Generalised estimating equations (GEE) → accounting for correlation in estimated effects (parameters) and standard errors*
- *Normal (linear) mixed (multilevel) models*
- *GEE and generalised linear mixed models (GLMM) for discrete data → binary and count data*

With the exception of paired t-tests and the summary measures approach, these statistical techniques are generally beyond the scope of non-statisticians, and I would recommend seeking the assistance of a statistician should you wish to undertake any of these analyses.

Kaplan-Meier survival analysis

Survival analysis seeks to study responses in the form of “time to failure” (e.g. time to death, onset of disease, or relapse) according to study covariates. An important aspect of survival analysis is that it deals with censoring, which is a mechanism whereby the outcome for an individual is missing. The most commonly encountered and handled form of censoring is right censoring, which is the situation where a subject is followed for a certain period of time before either leaving the study or the study ending, at which point we do not know if and when the event of interest occurs.

A common method for analysing survival curves, and differences between them, is Kaplan-Meier survival analysis. Using this method, we can obtain survival estimates at each time point where at least one observation fails.

The Kaplan-Meier estimator is:

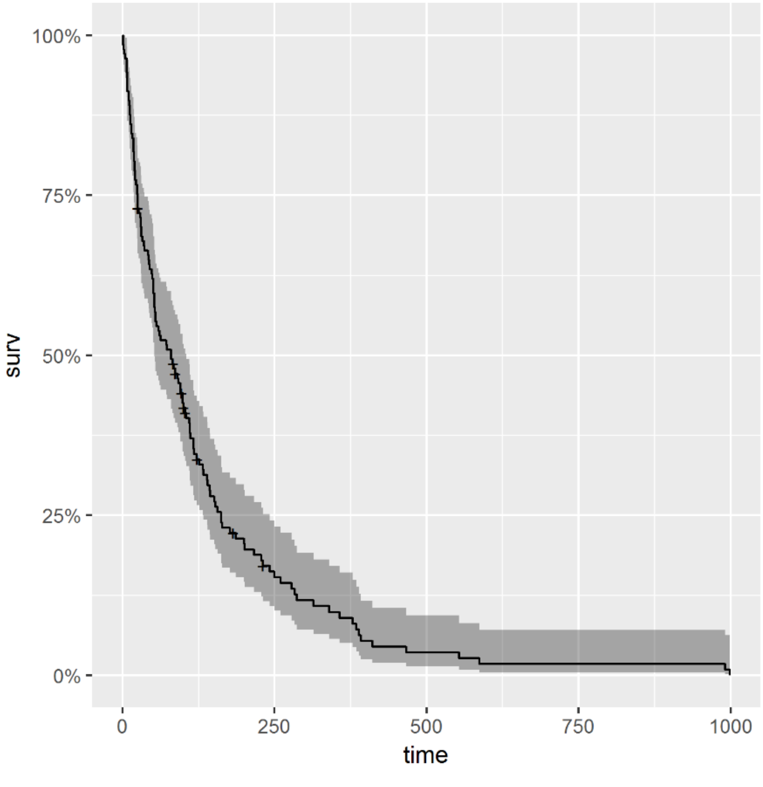
$$\widehat{S}\,(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Where d_i and n_i are the number of deaths and number of individuals at risk at time i.

Graphing survival estimates

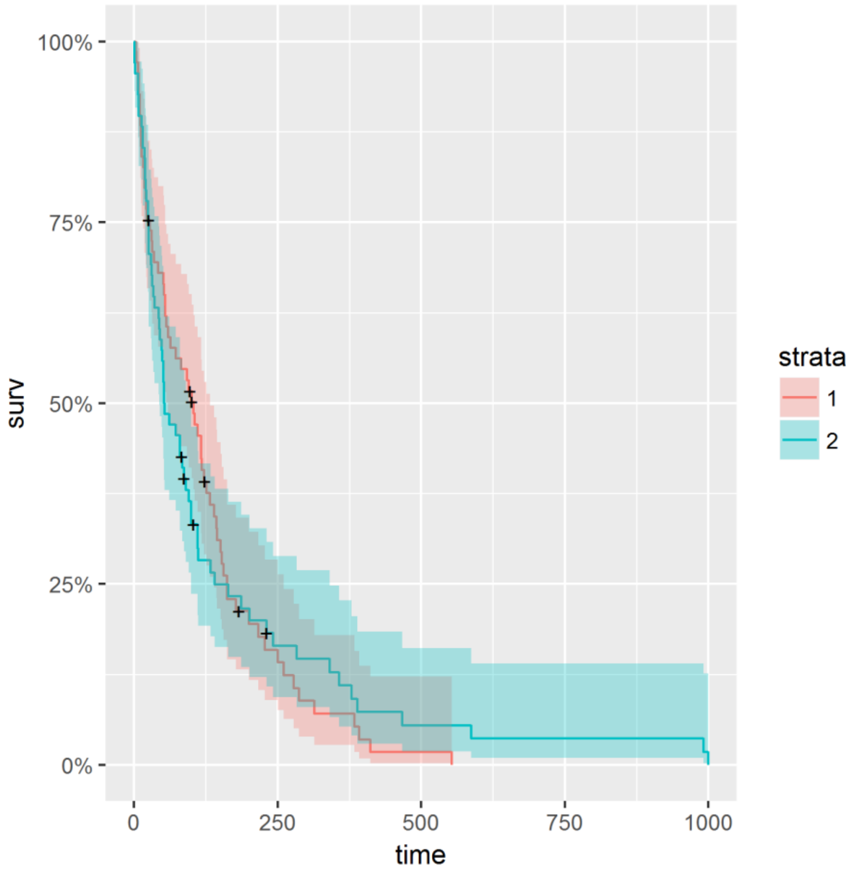
Kaplan-Meier estimates are often plotted to produce graphical displays of survivorship.

Example: % survival by time (days) following cancer diagnosis



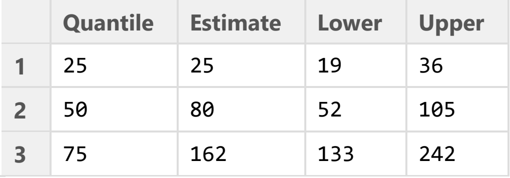
Two (or more) Kaplan-Meier survival curves can be assessed by plotting curves on a single graph

Example: % survival by time (days) following cancer diagnosis by treatment



The most common summary measure of survival data based on Kaplan-Meier survival curves is median survival, but sometimes quartiles (25th and 75th percentiles) of survival time are used.

Table: 25th, median (50th) and 75th percentile survival time estimates following cancer diagnosis, and lower and upper 95% confidence limits (in days)



	Quantile	Estimate	Lower	Upper
1	25	25	19	36
2	50	80	52	105
3	75	162	133	242

Statistical test for comparing survival curves – the log-rank test

Often, it is useful to assess whether survivorship between two groups is statistically different. There are several methods to do this, but a common way, that usually accompanies Kaplan-Meier graphical displays of survivorship, is the ***log-rank test***. The test is classified as non-parametric, and compares estimates of hazard functions between groups at each time an event is observed. The test compares the observed number of events in each group to the expected number of events based on the two groups having equal survival and hazard functions (i.e. equal rates of death in each group).

The log-rank statistic is calculated as:

$$Z = \frac{\sum_{j=1}^{J}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}$$

Where

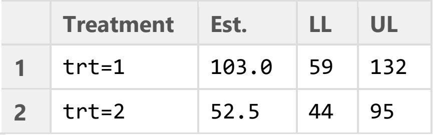
$$E_{1j} = \frac{O_j}{E_j} N_{1j}$$

$$V_j = \frac{O_j \left(N_{1j}/N_j\right)\left(1 - N_{1j}/N_j\right)\left(N_j - O_j\right)}{N_j - 1}$$

O_j = total number of events across both groups at period *j*
N_j = total number of individuals “at risk” (across both groups) at the start of period *j*
O_{j1} and O_{j2} are the number of observed events in each group, respectively, at *j*
N_{j1} and N_{j2} are the number “at risk” in each group, respectively, at *j*
E_{1j} and E_{2j} are the expected number of events in each group, respectively, at *j*
V_j is the variance at period *j*

Worked example in R – assessing equality of survival curves following cancer diagnosis by treatment

Table: Median survival time estimates (in days) with lower (LL) and upper (UL) 95% confidence interval limits following cancer diagnosis by treatment



	Treatment	Est.	LL	UL
1	trt=1	103.0	59	132
2	trt=2	52.5	44	95

Note the wide confidence intervals associated with each estimate of median survival time (low precision).

```
                Log-rank test for comparing survival curves

       N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 69       64     64.5   0.00388   0.00823
trt=2 68       64     63.5   0.00394   0.00823

 Chisq= 0  on 1 degrees of freedom, p = 0.928
```

Although there was a sizeable difference in median survival times between the two treatment groups, the confidence intervals associated with each estimate were large, indicating considerable variance (low precision), and, overall, the survival distributions were not significantly different according to the log-rank test.

Cox proportional hazards regression survival analysis

There are a variety of more sophisticated ways of analysing survival data, particularly those which take into account the effect of covariates on survival times. A common technique used for analysing survival data that caters for the individual effects of covariates is ***Cox proportional hazards regression***. In this method, the regression is not on survival times, but is specified to model the hazard function as follows:

$$h(t) = h_0(t) \times \exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

Where

$h(t)$ = the hazard
$h_0(t)$ = baseline hazard
$\exp(\beta_1 x_{i1} + \cdots + \beta_p x_{ip})$ = the model coefficients and subject-specific covariate values that modify the baseline hazard

Therefore, the hazard (instantaneous rate of failure) of one subject is a multiple of the hazard of another subject at any point in time (this does not mean the hazard is fixed over time, it certainly can vary).

Cox proportional hazards models are constructed in a way to conveniently convey the effects of covariates (e.g. age, sex, treatment …) on survival in the form of ***hazard ratios***. Exponentiation of estimated model parameters (the β coefficients) yields the hazard ratios for the covariates associated with the estimated parameters. For example, if the estimated parameter for a frailty covariate in a Cox regression model is 1.5, the hazard ratio is $e^{1.5} = 4.48$, which indicates that a subject that is frail (by whatever definition) is 4.48 times more likely to experience the event at any given time compared to someone that is not frail.

Worked example in R – melanoma survival time by sex, tumour thickness, age and site

```
  Cox proportional hazards regression estimated coefficients and hazard
                                  ratios

             coef      exp(coef) se(coef)  z       Pr(>|z|)
tt           0.115053  1.121933  0.044714  2.573   0.0101 *
sexFemale -0.522949  0.592770  0.277846 -1.882   0.0598 .
siteFace  -0.112545  0.893557  0.409518 -0.275   0.7835
siteNeck   0.203789  1.226039  0.426887  0.477   0.6331
siteScalp  0.290931  1.337672  0.388658  0.749   0.4541
age        0.007047  1.007072  0.007839  0.899   0.3687
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                              Hazard ratios

             exp(coef) exp(-coef) lower .95  upper .95
tt           1.1219     0.8913    1.0278     1.225
sexFemale    0.5928     1.6870    0.3439     1.022
siteFace     0.8936     1.1191    0.4004     1.994
siteNeck     1.2260     0.8156    0.5311     2.831
siteScalp    1.3377     0.7476    0.6245     2.865
age          1.0071     0.9930    0.9917     1.023
```

Only one covariate, tumour thickness (*tt*) is significant at a 95% confidence level with a hazard ratio of 1.12. The interpretation of this estimate is that, for every 1 mm increase in tumour thickness, the rate of death increases by a factor 1.12.