Yale SCHOOL OF MEDICINE

Basics of Stats, ROC curves and Survival Analysis Using R

Mercedeh Movassagh PhD

Class Assumptions: Basic R Programming Skills and Stats

If you do not have any R or statistics background or would like to learn more you can go to https://github.com/Mercedeh66/PR2022replicathon

Feel free to fork from this repository and learn a bit more on your own.

Plan For the Lecture and Lab Today

Hypothesis Testing

Student t-Test, Anova, Fisher and Chi-square test

What is Statistical Learning?

How do we estimate f ?

Linear Regression

ROC curves (Sensitivity vs Specificity)

Logistic Regression

R Lab

Break

Survival Analysis and Censored Data

Survival and Censoring Times

The Kaplan Meier Survival Curve

Log Rank Test

Hazard Function

Shrinkage for Cox Model

R Lab

T-test, ANOVA, Fisher’s Test and Chi-Square Test

T-test, ANOVA, Fisher’s Test and Chi-Square Test

Quick Review of the Null and Alternative Hypothesis

Null Hypothesis is a specific claim about the value of a population parameter. It is made for the purpose of argument which often embodies the skeptical point of view. (this is the statement that would make your observation interesting if it is **rejected**). It is often portrayed by H_0 .

The **Alternate Hypothesis** includes all other feasible values for the population parameter besides the value stated in the null hypothesis (The biologically/events that are more interesting than what is in the null hypothesis). You always hope that this is true. It is often portrayed by H_A .

Null distribution is sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true.

The P value is the probability of obtaining or showing data as greater difference than the null hypothesis.

The significance level α is the probability criterion for rejecting the null hypothesis. If P is less than or equal to α then the null hypothesis is rejected.

Type 1 error is when one rejects a true null hypothesis (**false positive**).

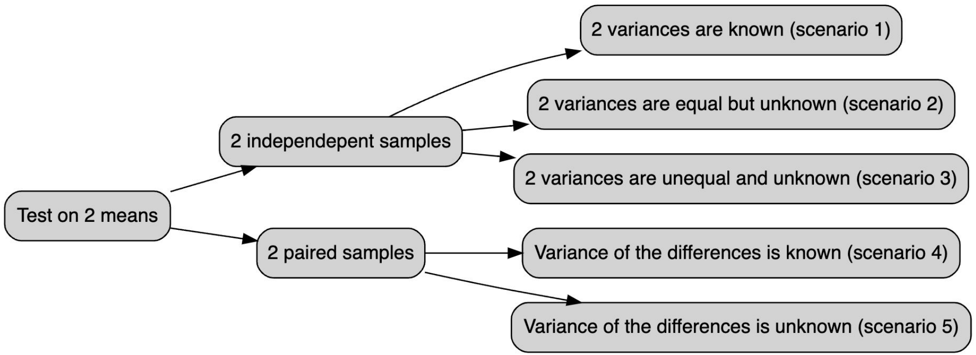
Type 2 error is when you fail to reject a true null hypothesis (**false negative**).

T-test

The Student's t-test for two samples is used to test whether **two populations** different in terms of a **quantitative/continuous** variable, based on the comparison of two samples drawn from these two groups.

To compare two samples, it is usual to compare a measure of central tendency computed for each sample. In the case of the Student's t-test, the **mean** is used to compare the two samples.

There are several Versions of T-test



Test on 2 means

2 independepent samples

2 paired samples

2 variances are known (scenario 1)

2 variances are equal but unknown (scenario 2)

2 variances are unequal and unknown (scenario 3)

Variance of the differences is known (scenario 4)

Variance of the differences is unknown (scenario 5)

$$z_{obs} = \frac{(\overline{x}_1 - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2{}_2}{n2}}}$$

- Number of observations in each sample *n1=n2*
- mean of samples x_1 group vs x_2 group.
- variances of both populations: σ for both groups
- Critical value can then be estimated using a statistical table from the Z observed.

[Reading Statistical Tables for T- test]

[T test Shiny App]

ANOVA

To compare means of three groups of more we use analysis of variance method, ANOVA. The question ANOVA aims to answer is there really a is there more variance among the same means that we would expect by chance alone. Through this we estimate an F ratio:

$$F = \frac{group\, mean\, square}{error\, mean\, square} = \frac{MS_{groups}}{MS_{error}}$$

Where $MS_{groups} = \frac{SS_{groups}}{df_{groups}}$

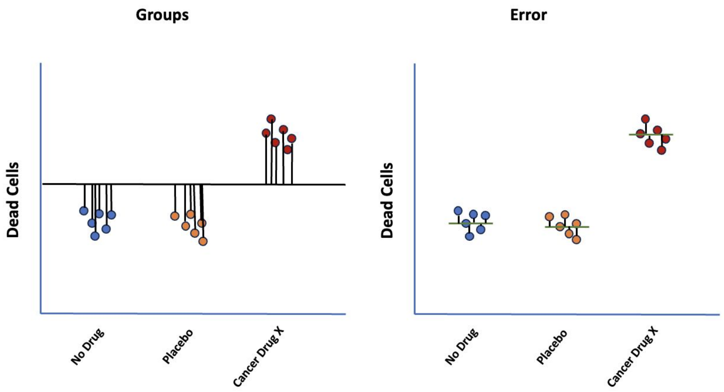
$SS_{groups} =$ group sum of squares $= \sum_i n_i \left(\bar{Y}_i - \bar{Y}\right)^2$

$df_{group} =$ group degrees of freedom which is always equal to number of groups (k) -1

Where $MS_{error} = \frac{SS_{error}}{df_{error}}$

$SS_{error} =$ Error sum of squares = $\sum_i \sum_j \left(Y_{ij} - \bar{Y}_i\right)^2$

$df_{error} = N - k$ Where N is the total number of data points in the all groups and k is the number of groups.



Fisher’s Test

Fisher’s exact test provides an exact P value for a test of association in a 2X2 contingency table. It is used where cell frequencies are low (n observation <10). For the most part this test examines the independence of two **categorical** variables with small expected values.

	Group X	Group Y	Row Total
Condition Z	a	b	a+b
Condition Q	c	d	c+d
Column Total	a+c	b+d	a+b+c+d (=n)

← **Contingency Table**

a is distributed as a hypergeometric distribution with a+c draws from a population and a+b success and c+d failures and can be estimated by the following:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!\,b!\,c!\,n!}$$

Chi-Square Test

Chi-square statistics is used on contingency tables to test whether the discrepancies are greater than expected by chance.It is used where cell frequencies are high (n observation >10). For the most part this test examines the independence **categorical** variables with larger expected values.

$$e = \frac{\sum_{row=1}^{r} * \sum_{col=1}^{c}}{grand\ total\ of\ samples}$$

$$\chi^2 = \frac{\sum (o-e)^2}{e}$$

$$df = (r-1)(c-1)$$

The df and X^2values can then be used to assess the critical value and reject or accept the null hypothesis.

What is Statistical Learning?

How Do We Estimate f ?

Non-parametric Method:

Do not make explicit assumptions about $f(\ )$ instead they seek an estimate of f that gets closest to the real data points. These methods have the capacity to accurately fit large ranges of very different forms of f. Very flexible models.

Major Disadvantage:

-They do not reduce the problem of estimating f to a small number of parameters so a very large number of observations is often required to obtain an accurate estimate of f.
-Interpertibily is also a problem in some cases.

Some examples are k nearest neighbors, Principal component analysis, smoothing splines and various form of MLs such as random forests and neural nets.

What is Statistical Learning?

Imagine you are trying to understand the relationship between cancer development and specific phenotypes like smoking, sugar consumption, and exercise level.

The questions you might have are:

Is there a relationship between cancer development and smoking?

How strong is the relationship between the features?

Are there associations between cancer development, sugar consumption and exercise levels?

How large is the association between each medium and cancer development?

How accurately can we predict cancer development using these features?

Is the relationship linear?

What is Statistical Learning?

We are trying to answer these questions with a set of statistical assumptions.

Basics of statistical learning:

In the setting mentioned cancer is the outcome variable Y (independent variable), and the smoking and other features are the input variables X (dependant variable) .

$$Y = f(X) + \epsilon$$

f is a fixed but an unknown function $X_1, \ldots, Xp$ and ϵ is a random error term, which is independent of X and has a mean zero.

What is Statistical Learning?

Why Estimate f ?

Prediction:

A set of X is readily available however Y cannot be easily obtained hence an estimate of the f will result in an estimate of Y which is very close to the prediction for Y. f is a black box.

$$\hat{Y} = \hat{f}(X),$$

Inference:

f is no longer a black box, association between Y and $X_1, \ldots, Xp$ is of interest.

How Do We Estimate f ?

How Do We Estimate f ?

Parametric Method:

1) Make an assumption on the functional form, or shape of f. An instance of a very simple linear model is:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

2) Use training data to fit and train the model. In this linear model this is used to estimate the parameters

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

The most common way to fitting the model is the least square which we will discuss in detail in a few slides.

Advantages of using the Parametric Method:

-Easy to interpret and easy to fit a simplified model on a known *f( )* versus an entirely arbitrary *f( )*.

Problems with using the Parametric Method:

-The model we choose usually does not match the true unknown form of f, less flexible.

Regression

Linear Regression

Most common type of regression is linear regression which draws a straight line through the data to predict the response variable Y. The relationship between the dependant and independant variable **really is linear for continuous variables**.

$$Y \approx \beta_0 + \beta_1 X + \epsilon$$

$$\begin{aligned} \beta_0 &= Intercpet \\ \beta_1 &= Slope \\ & Together\, they\, are\, known\, as\, the\, coefficients \end{aligned}$$

Residual Sum of Squares (*RSS*) and Least square means method (*LSMM*)

Assume $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_i$ is a predictor for Y based on the ith value of X. Then $e_i = y_{i-}\hat{y_i}$ represents the ith residual between the observed response and observed value predicted by the linear model. So,

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2$$

The least square means method uses sample means to minimize the RSS.

Linear Regression

How to assess the accuracy of a linear model?

Residual Standard Error (RSE):

The RSE is an estimate of the standard deviation of ϵ (the error term of the model, whatever we miss with the model in terms of variation in Y). RSE is the average amount that response (Y) will deviate from the true regression line. In other worlds it is an absolute measure of lack of fit in the model to the data.

Assuming, $Y \approx \beta_0 + \beta_1 X + \epsilon$

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

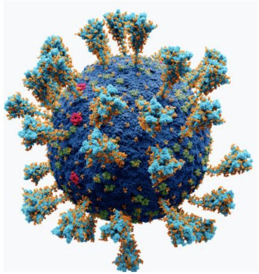
R^2 Statistic:

It is the proportion of variance explained and so it takes on a value between 0-1 and it is independence of the scale of Y.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Where, $TSS = \sum (y_i - \bar{y})^2$ is the total sum of squares and measures in response to Y.

Multiple Linear Regression



When trying to answer a biological or really any fundamental question, in practice we often have more than one predictor (X).

Assume we are trying to see if having a viral infection, exposure to x-ray, or eating a lot of sugar is useful in predicting the development of cancer (our response variable)?

Are all of these associated with cancer development or only some?

How well does the model fit the data?

How accurate are our predictions?

We can answer all of the above by using a multiple linear regression

Multiple Linear Regression

Assume we are trying to see if having a viral infection, exposure to x-ray, or eating a lot of sugar is useful in predicting the development of cancer (our response variable)?

AKA is there a relationship between the response and the predictors?

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

The larger the value of F statistics (larger than 1) the more compelling evidence that there is a relationship between response and the predictor.

Are all of these associated with cancer development or only some?

As you add the variables you check for the P values and how much improvement or decline they make.

Multiple Linear Regression

How well does the model fit the data?

R^2 or RSS of the model can be used for this.

How accurate are our predictions?

Use of confidence interval can be used to quantify the uncertainty surrounding the average (mean). In other words a 95% confidence interval contains the true value of $f(x)$.

In other words, for a 95% confidence interval you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.

Your desired confidence level is usually one minus the value α you used in your statistical test.

confidence interval = 1- α

What is the CI for α 0.001?

0.999

Logistic Regression

When the response variable **is qualitative (categorical -Yes, No)** instead of quantitative we use logistic regression to model the probability that Y belongs to a particular category. This method is used very commonly in dose response curves. The relationship between Y and X **is not linear** because it cannot fit between the interval of 0 and 1.

Instead of a normal distribution it assumes X has a binomial distribution (being either 0 or 1) :

$$\log - odds\,(Y)\; = \;\alpha\; +\; \beta X$$

Where, log - $odds$ refers to the natural log of the odds of Y.

$\alpha\; +\; \beta X$ Is the formula of a straight line with α as the intercept and β the slope.

To obtain the predicted values of Y we need to convert it to ordinary proportions.

$$\hat{Y} = \frac{e^{\log - odds\left(\hat{Y}\right)}}{1 + e^{\log - odds\left(\hat{Y}\right)}}$$

Receiver Operating Characteristic (ROC) Curve

A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:

True Positive Rate (TPR) is a synonym for recall and sensitivity and is therefore defined as follows:

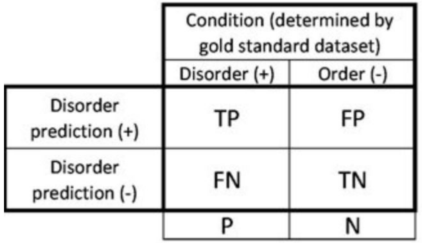
$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN}$$

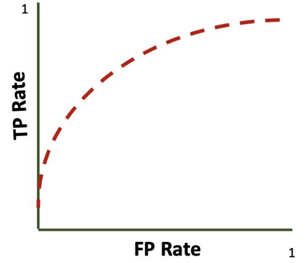
Specificity is defined as :

$$Specificity = \frac{TN}{TN + FP}$$

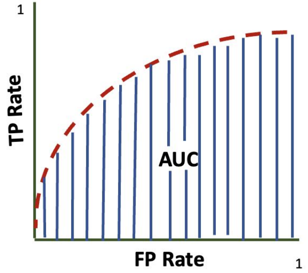


	Condition (determined by gold standard dataset)	
	Disorder (+)	Order (-)
Disorder prediction (+)	TP	FP
Disorder prediction (-)	FN	TN
	P	N

An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, hence increasing both False Positives and True Positives.



AUC stands for "Area under the ROC Curve." That is, AUC measures the entire two-dimensional area underneath the entire ROC curve.



References

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. “An Introduction to Statistical Learning.” https://link.springer.com/content/pdf/10.1007/978-1-0716-1418-1.pdf.

Whitlock, Michael C. n.d. “The Analysis of Biological Data.” Accessed July 19, 2023. https://marmamun.gov.np/sites/marmamun.gov.np/files/webform/pdf-the-analysis-of-biological-data-michael-c-whitlock-dolph-schluter-pdf-download-free-book-8cbddee.pdf.

R Lab Session **go to github**
https://github.com/Mercedeh66/HUSRL

You can clone:
git@github.com:Mercedeh66/HUSRL.git