Course: Univariate Statistics with R

Dr. Dishon Muloi – International Livestock Research Institute

> Like other occult techniques of divination, the statistical method has a private jargon deliberately contrived to obscure its methods from non-practitioners.
>
> — Attributed to G. O. Ashley

Which test to use?

It is almost impossible to get by in epidemiology without having at least a cursory understanding of what a null hypothesis, p-value, confidence interval means.

Statistical tests work by calculating a test statistic – a value that indicates how much the relationship between variables in your test differs from the null hypothesis of no relationship. Most, if not all, tests calculates a p-value (probability value) which estimates how likely it is that you would see the difference described by the test statistic if the null hypothesis of no relationship were true. If the p-value is less than 0.05, we reject the null hypothesis that there's no difference between the means and conclude that a **significant difference** does exist. If the p-value is larger than 0.05, we cannot conclude that a significant difference exists. (P value sceptic? https://www.nature.com/articles/d41586-019-00857-9). Statistical tests assume a null hypothesis of no relationship or no difference between groups.

To determine which statistical test to use, you need to know:

- Whether your data meets certain statistical assumptions (Independence of observations, Homoscedasticity and normality of data)
- The types of variables that you're dealing with.

If your data do not meet the assumptions of normality or homogeneity of variance, you may be able to perform a **nonparametric statistical test**.

Parametric tests

Prerequisite: data has to conform to the three assumptions. The most common types of parametric test include regression tests, comparison tests, and correlation tests.

Regression test: *used to test cause-and-effect relationships i.e., the effect of one or more continuous (or categorical for logistic regressions) variables on another variable.*

Test	Predictor variable	Outcome variable	Research question example
Simple linear regression	Continuous	Continuous	What is the effect of height on weight?
	1 predictor	1 outcome	
Multiple linear regression	Continuous	Continuous	What is the effect of height and age on weight?
	2 or more predictors	1 outcome	
Logistic regression	Continuous	Binary	What is the effect of weight on obesity incidence ?

Comparison tests: *used to investigate for differences between group means*

	Predictor variable	Outcome variable	Research question example
Paired t-test	Categorical	Quantitative	What is the effect of two different test prep programs on the average exam scores for students from the same class?
	1 predictor	groups come from the same population	
Independent t-test	Categorical	Quantitative	What is the difference in average exam scores for students from two different schools?
	1 predictor	groups come from different populations	
ANOVA	Categorical	Quantitative	What is the difference in average pain levels among post-surgical patients given three different painkillers?
	1 or more predictor	1 outcome	
MANOVA	Categorical	Quantitative	What is the effect of flower species on petal length, petal width, and stem length?
	1 or more predictor	2 or more outcome	

Correlation test: *used to check whether two variables are related without assuming cause-and-effect relationships. e.g How are latitude and temperature related?*

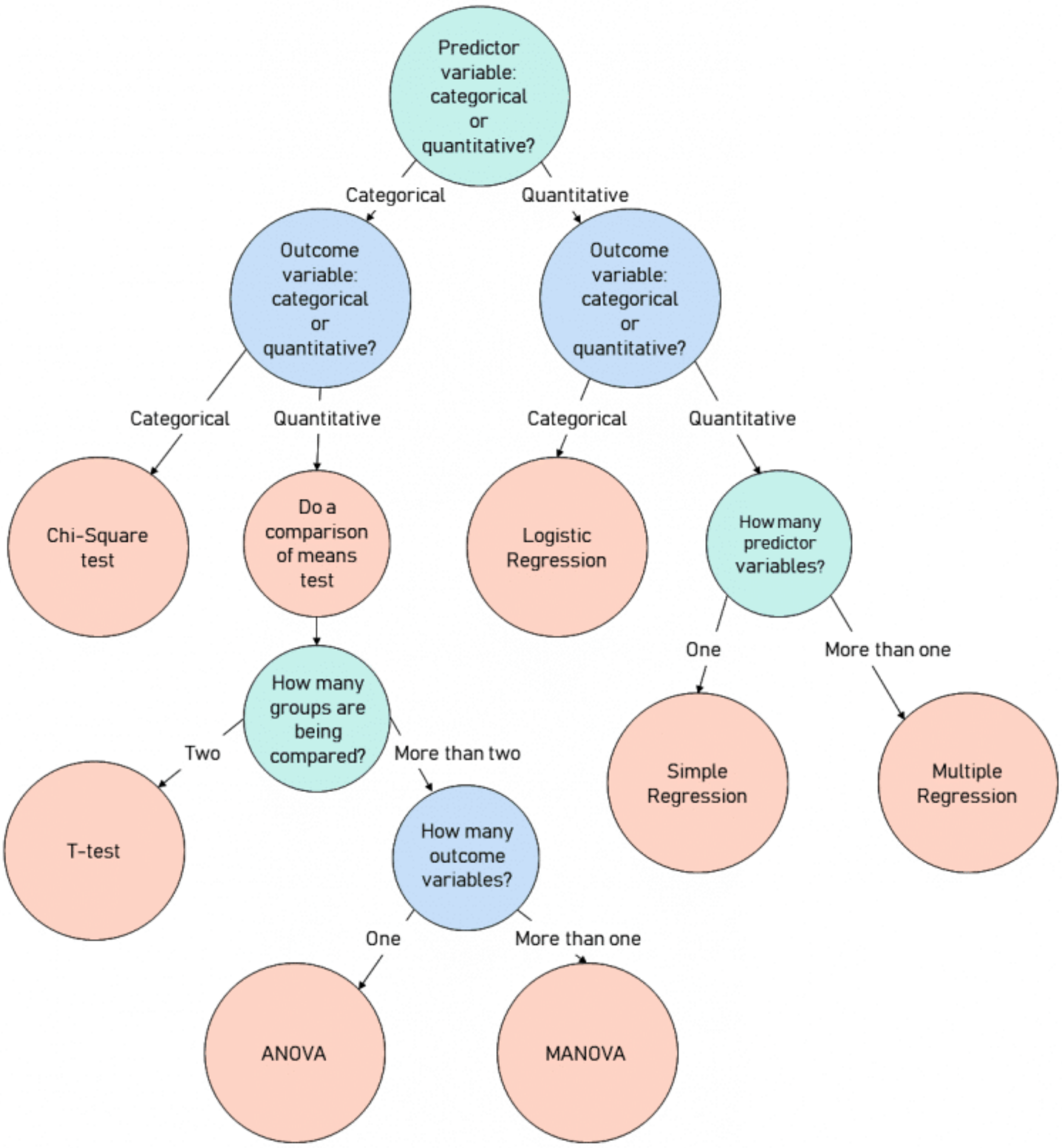
Non-parametric tests

Used when many assumptions about the data are not met. Always assess whether data are likely from a normal distribution using either Kolmogorov-Smirnov test, the Anderson-Darling test, or the Shapiro-Wilk test.

Test	**Predictor variable**	**Outcome variable**	**Use in place of**
Spearman's r	Quantitative	Quantitative	Pearson's r
Chi square test of independence	Categorical	Categorical	Pearson's r
Sign test	Categorical	Quantitative	One-sample t-test
Kruskal–Wallis H	Categorical	Quantitative	ANOVA
	3 or more groups		
ANOSIM	Categorical	Quantitative	MANOVA
	3 or more groups	2 or more outcome variables	
Wilcoxon Rank-Sum test	Categorical	Quantitative	Independent t-test
	2 groups	groups come from different populations	
Wilcoxon Signed-rank test	Categorical	Quantitative	
	2 groups	groups come from the same population	

There are two kinds of statistics, the kind you look up, and the kind you make up.
— Rex Stou

Choosing a statistical test


Predictor variable: categorical or quantitative?
Categorical
Quantitative
Outcome variable: categorical or quantitative?
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Categorical
Quantitative
Chi-Square test
Do a comparison of means test
Logistic Regression
How many predictor variables?
One
More than one
How many groups are being compared?
Two
More than two
Simple Regression
Multiple Regression
T-test
How many outcome variables?
One
More than one
ANOVA
MANOVA

```r
#T test
#Use Shapiro-Wilk normality test as described at: Normality Test in R. - Null
#hypothesis: the data are normally distributed - Alternative hypothesis: the data #are
not normally distributed.

lungcancer <- read.csv("Lungcancer.csv",header = T)

# Shapiro-Wilk normality test for Men's weights

with(lungcancer, shapiro.test(Weight[Sex == "Male"]))# p = 0.5

# Shapiro-Wilk normality test for Women's weights

with(lungcancer, shapiro.test(Weight[Sex == "Female"]))# p = 0.9


#the two p-values are greater than the significance level 0.05 implying that the
distribution of the data are not
#significantly different from the normal distribution. In other words, we can assume
the normality.


# use F-test to test for homogeneity in variances using var.test()

res.ftest <- var.test(Weight ~ Sex, data = lungcancer)
res.ftest

#The F test value is greater than the significance level alpha = 0.05 thus no
significant
#difference between the variances of the two sets of data. Therefore, we can use the
classic t-test witch assume equality of the two variances


t_test1 <- t.test(Weight ~ Sex, data = lungcancer)
t_test1

#The p-value of the test is 0.2858, which is less than the significance level #alpha =
0.05.
#We can conclude that men’s average weight is not significantly different from
#women’s average weight#
#look at the confidence interval! what can it tell regarding significance?

#Mann-Whitney-Wilcoxon

#suppose the data violated the assumptions we undertake a Mann-Whitney-Wilcoxon #Test
using the wilcox.test function

wilcox.test(Weight ~ Sex, data = lungcancer)

library(datasets)
head(mtcars)
str(mtcars)


#One way anova

#a one-way ANOVA on mpg (miles per gallon) being dependent on the gear (3, 4 and 5)

mtcars$gear <- as.factor(mtcars$gear)

#always plot your data first

boxplot(mpg~gear, data = mtcars)
fit <- aov(mpg ~ gear, data = mtcars)
summary(fit)

# p value is significant meaning different distribution hence we need to do a post
#hoc analysis
```

```r
posthoctukey <- TukeyHSD(fit)
posthoctukey
#plot the pairwise comparison

plot(posthoctukey,las=2,col="black",cex.axis=1,cex.lab =1)




#Kruskal wallis

#if the data was not normally distributed, we use a non-parametric test

fit_kruskal <- kruskal.test(mpg ~ gear, data = mtcars)
fit_kruskal

#post hoc test shoulr use Dunn’s pairwise tests or Nemeyi test

pairwise.wilcox.test(mtcars$mpg, mtcars$gear, p.adjust.method = "bonferroni")


#Chi-square test
#Chi-square test examines whether rows and columns of a
#contingency table are statistically significantly associated.

#Null hypothesis (H0): the row and the column variables of the contingency table are
independent.
#Alternative hypothesis (H1): row and column variables are dependent
#say you want to analyse the distribution of lung cancer cases by sex
#rows will be sex (the predictor) and column lung cancer incidence (the response)


#create a 2 by 2 contingency table

tab_canc_sex <- table(lungcancer$Sex,lungcancer$Lungcancer)
tab_canc_sex

chisq.test(tab_canc_sex)

#p value > that 0.05 thus there is no significant relationship between the two
categorical variables


#Fisher exact test
#the Fisher’s exact test is used when the sample is small
#same assumptions as for chi square

tab_canc_sex <- table(lungcancer$alcohol,lungcancer$Lungcancer)
fisher.test(tab_canc_sex)
```

```
#linear regression

#The mathematical formula of the linear regression can be written as y = b0 + b1*x +
e, where:
#b0 and b1 are known as the regression beta coefficients or parameters:
#b0 is the intercept of the regression line; that is the predicted value when x = 0
#b1 is the slope of the regression line.
#e is the error term (also known as the residual errors), the part of y that can be
explained by the regression model

#Create a scatter plot displaying wt (weight) vs mpg (miles per gallon) in mtcars
#dataset

library(ggplot2)
p_scatter <- ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point(color=c(“#00AFBB”),shape = 19,size =3)+
  labs(x=”mpg”, y = “wt”,title=”Relationship of mpg and wt”)+
  theme_bw()+
  theme(plot.title = element_text(color = “black”, size = 12,hjust = 0.5),
                               axis.text.x = element_text(color = “black”, size = 12),
                               axis.text.y = element_text(color = “black”, size = 12),
                               axis.title.x= element_text(color = “black”, size = 12),
                               axis.title.y = element_text(color = “black”, size =12))
p_scatter

#run a regression analysis using function lm

regress_mpg_wt <- lm(mpg~wt,data = mtcars)
regress_mpg_wt

# the estimated regression line equation can be written as follow: mpg = 49.65 + (-
5.344)*wt

#add regression line  on the scatter plot

p_scatter_regline <- p_scatter+geom_smooth(method = “lm”,color=”#00AFBB”)
p_scatter_regline

#try mpg and qsec
```

PS: For those keen on proceeding to intermediate level we will dive deeper into regression analysis.