Univariate analysis - Chi² JdG

Epi Lab color code

Software/Packages/Add-ins required

Software/Packages/Add-ins recommended

Description text

R code to copy/paste into console

R code to copy/paste into console that needs adjustment to your personal workspace

Websites where you can download requirements

Lab #5 requirements

- R - http://cran.r-project.org/bin/windows/base/
- R Studio - www.rstudio.com/ide/download/desktop
- Internet connection
- epicalc package (download info in code below)
- Internet connection

So last month we imported some SASVEPM congress attendance data and found that it needed some cleaning up. This month we start to analyse some of it for a bit of fun and this lab is focussed on two aspects - the first is an introduction to the use of a very cool package for epidemiologists called *"epicalc"* We use some of its summary functions as well as an attach function which helps in decreasing command input time in R. (as always, remember to hit TAB in R-Studio when typing in commands - it will help a lot!) The second aspect of the lab is to start analysis. In this case we are asking one question - Was there an association at the congress of academics attending and whether they presented a talk or not compared to non-academic attendees. We might be interested in this to determine whether a group of attendees like State officials are over represented in attendance but under represented in presenting - this may be something worthwhile knowing for the society when they advertise and call for papers for the next congress...

The code

```
# In this lab we are going to look at some very basic evaluation of data. Firstly we download the cleaned SASVEPM data which
#we performed last month If you haven't done that lab yet then go to http://www.elsenburg.com/vetepi/BPEL/
#BPEL_2014_08_EvalData_Clean1.pdf
#Here we download the cleaned data and put it into a data frame called 'sasvepmdataclean'
sasvepmdataclean<-read.csv('http://www.jdata.co.za/backpagelabs/backpagelabs_jdg_sasvepmclean.csv', header=T)
#to remind you of the data and its content
summary(sasvepmdataclean)
#for some reason - can't quite figure it out - R sometimes adds an additional X column with the same data as the id field into the data frame - so let's just
#remove that if it's there...if not don't worry
sasvepmdataclean$X<-NULL
#Before we get started we are just going to take a detour and use a very cool function in the epicalc package
#if you haven't installed the epicalc package yet then type this into your R console
install.packages("epicalc")
#activate the library after it has been installed (you can also tick it in R Studio's package window)
library("epicalc")
#note above how I needed to refer to the X column in the sasvepmdataclean data set by typing in sasvepmdataclean$X. This was of doing things is
#very pedantic and thorough but ultimately not necessary. In R there is a way to essentially attach your data frame that you are working with that it
#recognises column names without you having to refer to them explicitly. The epicalc package has a function that has simplified this so let's try it.
#first we see what is attached in your environment
search()
#these are all packages and data that is attached, so the epicalc package should be there
#now we want to attach sasvepmdataclean - the epicalc function is "use"
use(sasvepmdataclean)
#now try
search()
#again - you'll see a ".data" listed in the attached data and packages - this is your dataset you have now attached
#now instead of referring to say the participation field in sasvepmdataclean (sasvepmdataclean$participation) you could just type in
participation
#and get the same result
#ok to get back to our analysis - you'll remember that the summary command is a general one
summary(sasvepmdataclean)
#the epicalc package has some of its own summaries which are also useful - codebook is the first
codebook(sasvepmdataclean)
#this takes a look at each variable and does a frequency count for categorical variables (like most of the data in this example ) or measures of centrality
and spread for the continuous data - like id in our case - (which is meaningless) Another epicalc summary is:
summ(sasvepmdataclean)
#this seems more useful for continuous data so is not so worthwhile in our example. In this lab we are going to do some univariate analysis so we are
#going to try answer one question for now
##########Q1: Were presenters more likely to be from an academic institute?###########
#First we need to create a two by two table to evaluate all participation by whether the participant was an academic or not. To do this we use a table
#command.
table(participation,institute)
#note this gives us categories of all participation by all institutes - but some categories are poorly represented and our
```

Continued on next page

```
#question however was really just isolating the academic versus non academic - so let's classify all non academic participants into one category. to make it
#simpler I'm going to create a new column with this info in it.
sasvepmdataclean$q1<-ifelse(institute=="Academic","Academic","Non-Academic")
#this uses an if-else function - essentially we look at the value in th einstitute field and if its "Academic" we store "Academic in the new field or if not (else)
we store "Non-Academic" in the new field.
q1
#gives an error but typing in
sasvepmdataclean$q1
#works - why is this even though we have attached the data to our environment using the use() command? essentially when you attach data R will
#recognise only data that you have attached, adding new data in the attached data frame is not automatically attached. Running the following commands
#will detach all data, add the column we wanted and the attach the data again
zap()
#this removes all attached data and removes your datasets in the environment - try type in sasvepmdataclean  now and it should not work
#now to read in everything again
sasvepmdataclean<-read.csv('http://www.jdata.co.za/backpagelabs/backpagelabs_jdg_sasvepmclean.csv', header=T)
sasvepmdataclean$X<-NULL
sasvepmdataclean$q1<-as.factor(ifelse(sasvepmdataclean$institute=='Academic','Academic','Non-Academic'))
use(sasvepmdataclean)
codebook(sasvepmdataclean)
#note above I needed to first create the new column before I used the use() function. Also note I forced the new column into a factor class, if I did not do
#that it would have been a character class which differs from my other variables. This is what the table looks like prior to reclassification of the institute
#field
table(institute,participation)
#now reclassification making a 2X2 table
table(q1,participation)
#now we allocate the table to a variable we can use (not a necessary step but makes it easier in the long run)
q1table<-table(q1,participation)
#now we do a chi squared test of association on the table
chisq.test(q1table)
#you'll see the result but the chi squared test function does store the various elements of the test, so now we allocate the chi squared result to a variable
#and we look at each output at our leisure.
q1chitestresult<-chisq.test(q1table)
q1chitestresult$observed #the observed values (same as the input table)
q1chitestresult$expected #the expected values when your null hypothesis is true
q1chitestresult$p.value #so the actual p-value
q1chitestresult$statistic #the chi squared test statistic measuring the contrast between the observed and expected frequencies
q1chitestresult$parameter #degrees of freedom parameter
q1chitestresult$method #method used to calculate the statistic
#Try to reproduce the result but look at questions like: did specific Provinces present significantly more than others (although there may be some
#confounding there with the fact that certain academic institutes are only in certain Provinces:) Also were veterinarians over represented when giving
#presentations? Remember that if you want to use the epicalc use  function then when you are making  new columns if you wish to then zap()  the
#environment and start from scratch creating the columns you need prior to using the use() function.
```

The output

```
             participation
q1            Attendee Presenter
  Academic           5        14
  Non-Academic     160         5
```

The input and observed table

```
 Pearson's Chi-squared test with Yates' continuity correction

data:  q1table
X-squared = 84.3806, df = 1, p-value < 2.2e-16
```

The chi squared test result

```
             participation
q1             Attendee Presenter
  Academic      17.03804  1.961957
  Non-Academic 147.96196 17.038043
```

The expected values if null hypothesis was true

```
[1] 4.081285e-20
```

The actual p-value for the result

```
X-squared
 84.38062
```

The chi squared statistic

```
df
 1
```

The degrees of freedom

```
[1] "Pearson's Chi-squared test with Yates' continuity correction"
```

The chi squared method

The result

Attendees that are from an academic background were associated ($p<0.05$) with being presenters at the 2014 SASVEPM congress

References

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Virasakdi Chongsuvivatwong (2012). epicalc: Epidemiological calculator. R package version 2.15.1.0. http://CRAN.R-project.org/package=epicalc

Thompson, P. and Gummow, B. (2005). Epidemiology 752 Chapter 1: Statistical Building Blocks. Department of Production Animal Studies, University of Pretoria, Pretoria, South Africa