Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

Question 1 (40 points): Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1]Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multi-method Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) The χ^2 test statistic is calculated as follows:

Read in the data as a matrix.

```r
observed  <- matrix( c (14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)
```

Calculate the expected values, then calculate the difference between the observed and expected values for each sub-category. Calculate the contribution to the χ^2 statistic. (expected = number in class * number of outcomes / total number; difference = observed - expected; contribution = difference^2/expected)

For example, for the sub-category 'Upper Class' and 'Not Stopped':

Upper Class, Not Stopped	
observed	14
expected	13.5 = (27 * 21 / 42)
difference	0.5 = (14 - 13.5)
chi sq contribution	0.0185 = $(0.5\ )^2$ / 13.5

```r
ncols <- length(observed[1,])
nrows <- length(observed[,1])

# get totals
row_tots <- vector("double", nrows)
col_tots <- vector("double", ncols)

totals <- sum(observed)  # total number of observations

# calculate row and column totals, e.g, total for NotStopped, UpperClass, etc
for (i in 1:nrows) {row_tots[i] <- sum(observed[i, ])}
for (i in 1:ncols) {col_tots[i] <- sum(observed[, i])}

#get expected = row total * column total / total observations
expected <- observed
for (i in 1:nrows) {
  for (j in 1:ncols) {
    expected[i,j] <- row_tots[i] * col_tots[j] / totals
  }
}

# calculate difference between observed and expected
```

```r
o_e <- observed
for (i in 1:nrows) {
  for (j in 1:ncols) {
    o_e[i,j] <- (observed[i,j] - expected[i,j])^2 / expected[i,j]
  }
}

#calculate chi-squared value & degrees of freedom
chi_sq_val <- sum(o_e)
df = (nrows-1) * (ncols-1)
```

(b) Now calculate the p-value from the test statistic you just created (in `R`).[2] What do you conclude if $\alpha = 0.1$?

```r
p_value <- pchisq(chi_sq_val, df=df, lower.tail=FALSE)
alpha <- 0.1
```

```
The p-value is 15.02%, alpha is 10%
We cannot reject the null hypothesis that the two sets are from the
same population
1 observed cell(s) with less than 5 values
```

The observed and expected values are shown in Figure 1

The results of the builtin `R` chisq.test function are as follows:

```
Pearson's Chi-squared test

data:  observed
X-squared = 3.7912, df = 2, p-value = 0.1502
```

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

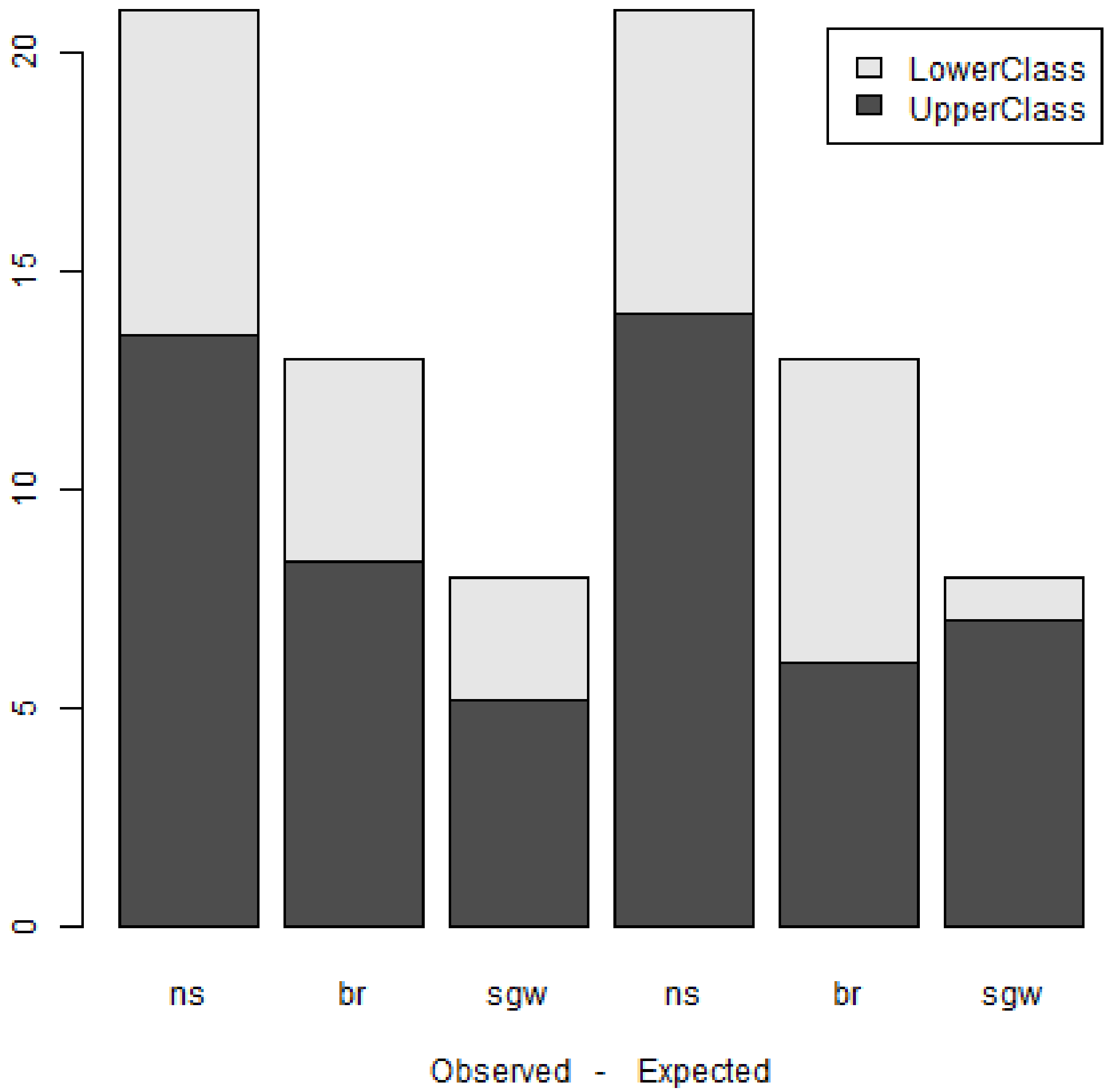


Figure 1: Observed vs Expected values for traffic stop. ns = Not Stopped; sgw = Stopped Given Warning; br = Bribe Requested

(c) The standardized residuals are set out in the table below:

Table 1: Standardised Residuals

	NotStopped	BribeRequested	StoppedGivenWarning
UpperClass	0.32	-1.64	1.52
LowerClass	-0.32	1.64	-1.52

(d) How might the standardized residuals help you interpret the results?

The biggest contribution to the residuals was from the ‘Bribe Requested’ variable - fewer upper class individuals were expected to hand over bribes. The difference between the two groups appears to be a combination of fewer upper class drivers being expected to hand over bribes and more of them being given a warning instead the opposite outcome occuring for lower class drivers.

Question 2 (40 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure 2 below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

Figure 2: Names and description of variables from Chattopadhyay and Duflo (2004).

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

[3] Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica.* 72 (5), 1409-1443.

(a) State a null and alternative (two-tailed) hypothesis.

Null The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.

Alternate The reservation policy does have an effect on the number of new or repaired drinking water facilities in the villages.

(b) Bivariate regression to test this hypothesis:.

Import the data.

The analyis used the builtin `R` function `lm` to investigate the relationship between the number of new or repaired drinking water facilities in the villages and the binary variable indicating whether the GP was reserved for women leaders or not.

```
water <- lm(water ~ reserved , data = policy)
```

This results in the following output:

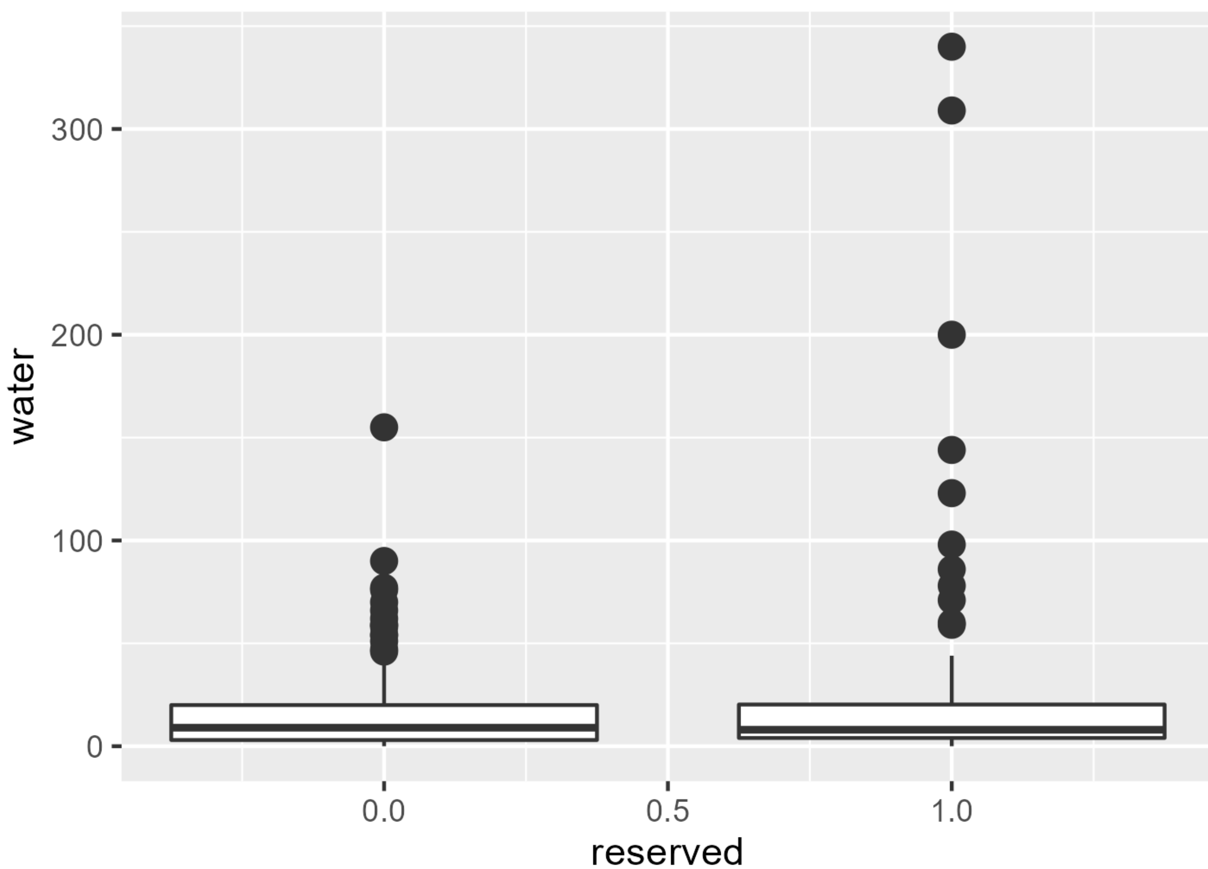
Table 2: Pearson Linear Regression - Water Reserved

	water
reserved	9.252^{**}
	(3.948)
Constant	14.738^{***}
	(2.286)
N	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493^{**} (df = 1; 320)

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

The

Figure 3: Boxplot of number of drinking water projects, grouped by reserved



The assumption in using a linear regression model is that each village is a separate case and each case is independent. However, in this study each GP is associated with two villages, so there is a risk that the values for each village are not independent.

(c) Interpret the coefficient estimate for reservation policy.

The model suggests that There are more outliers in the reserved=1 cohort

Appendix - Code

```
#####################################################################
# Imelda Finn, 22334657
# POP77003 - Stats I
# clear global .envir, load libraries, set wd
#####################################################################

# remove objects
rm(list=ls())

# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices"
    , "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1,
    TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package,
    character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# load necessary packages
lapply(c("ggplot2", "stargazer", "tidyverse", "stringr"),  pkgTest)

# function to save output to a file that you can read in later to your docs
output_stargazer <- function(outputFile, appendVal=TRUE, ...) {
  output <- capture.output(stargazer(...))
  cat(paste(output, collapse = "\n"), "\n", file=outputFile, append=appendVal)
}


# set working directory to current parent folder
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

######################
# Problem 1
######################

#Question 1 (40 points): Political Science

#The following table was created using the data from a study run in a major
# Latin American city.
```

```
# As part of the experimental treatment in the study, one employee of the research
# team was chosen to make illegal left turns across traffic to draw the attention
# of the police officers on shift. Two employee drivers were upper class, two were
# lower class drivers, and the identity of the driver was randomly assigned per
# encounter. The researchers were interested in whether officers were more or less
# likely to solicit a bribe from drivers depending on their class (officers use
# phrases like, ``We can solve this the easy way'' to draw a bribe).
# The table below shows the resulting data.


#& Not Stopped & Bribe requested & Stopped/given warning \\
#Upper class & 14 & 6 & 7 \\
#Lower class & 7 & 7 & 1 \\
observed <- matrix( c (14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# create data structure with named dimensions
cols <- c("NotStopped", "BribeRequested", "StoppedGivenWarning")
rows <- c("UpperClass", "LowerClass")

observed_df <- data.frame(observed, row.names = rows)
names(observed_df) <- cols
print(observed_df)

pairs(observed)

#\item [(a)]
#Calculate the $\chi^2$ test statistic by hand/manually\\

###------------------------------------ 0 start listing of code from here
ncols <- length(observed[1,])
nrows <- length(observed[,1])

# get totals
row_tots <- vector("double", nrows)
col_tots <- vector("double", ncols)

totals <- sum(observed) # total number of observations

# calculate row and column totals, e.g, total for NotStopped, UpperClass, etc
for (i in 1:nrows) {row_tots[i] <- sum(observed[i, ])}
for (i in 1:ncols) {col_tots[i] <- sum(observed[, i])}

#get expected = row total * column total / total observations
expected <- observed
for (i in 1:nrows) {
```

```r
  for (j in 1:ncols) {
    expected[i,j] <- row_tots[i] * col_tots[j] / totals
  }
}

# calculate difference between observed and expected
o_e <- observed
for (i in 1:nrows) {
  for (j in 1:ncols) {
    o_e[i,j] <- (observed[i,j] - expected[i,j])^2 / expected[i,j]
  }
}

#calculate chi-squared value & degrees of freedom
chi_sq_val <- sum(o_e)
df = (nrows-1) * (ncols-1)

cat(str_glue("The chi_squared statistic is {round(chi_sq_val,3)}"))
cat(str_glue("The chi_squared degrees of freedom is {df}"))

# plot of observed and expected values
png("obs_exp.png")
barplot(cbind(expected, observed ), legend.text = rows,
        names.arg = c("ns", "br", "sgw", "ns",  "br", "sgw"),
        xlab = "Observed   -    Expected")
dev.off()

#\item [(b)]
#Now calculate the p-value from the test statistic you just created R
# .\footnote{Remember frequency should be $>$ 5 for all cells, but let's calculate
# the p-value here anyway.}  What do you conclude if $\alpha = 0.1$?\\

p_value <- pchisq(chi_sq_val, df=df, lower.tail=FALSE)
alpha <- 0.1

# p > alpha, can't reject null
if (p_value > alpha ) txt <- "cannot " else txt <- ""

# todo - check rule/restriction
cells_under <- length(observed[observed<5]) + length(expected[expected<5])

cat(str_glue("The p-value is {round(p_value*100,2)}%, alpha is {alpha*100}%."))
cat(str_glue("We {txt}reject the null hypothesis that the two sets are from the\n same population."))
cat(str_glue("note: {cells_under} observed cell(s) with less than 5 values."))


# \item [(c)] Calculate the standardized residuals for each cell and put them in the table below.
```

```
z <- observed
for (i in 1:nrows) {
  row_prop<- (1 - (row_tots [i] / totals))
  for (j in 1:ncols) {
    col_prop<- (1-  (col_tots[j] / totals))
    z[i,j] <- (observed[i,j] - expected[i,j])  /sqrt (expected[i,j]* row_prop * col_prop)
  }
}
z_df <- data.frame(z, row.names = rows)
names(z_df) <- cols

print(z_df)

# output results for Zij values to .tex file
output_stargazer(z_df, outputFile="std_residuals.tex", type = "latex",
                 appendVal=FALSE,
                 title="Standardised Residuals",
                 digits=2,
                 summary = FALSE,
                 style = "apsr",
                 table.placement = "h",
                 label = "StandardisedResiduals",
                 rownames = TRUE
                 )


# check result
chisq.test(observed)

#https://www.rdocumentation.org/packages/stargazer/versions/5.2.3/topics/stargazer

# \item [(d)] How might the standardized residuals help you interpret the results?

#  fewer upper class individuals asked for bribes and more given warnings;
#  the contribution from lower class drivers expected to give bribes is nearly
#  equivalent to the contribution from upper class drivers getting warnings

# ################################################################################
# Problem 2
#####################

#Question 2 (40 points): Economics
#Chattopadhyay and Duflo were interested in whether women promote different policies
# than men.
```

```r
# Answering this question with observational data is pretty difficult due to potential
# confounding problems (e.g. the districts that choose female politicians are
# likely to systematically differ in other aspects too). Hence, they exploit a
# randomized policy experiment in India, where since the mid-1990s, 1/3 of
# village council heads have been randomly reserved for women. A subset of the data
# from West Bengal can be found at the following link:
#    \url{https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv}

# Each observation in the data set represents a village and there are two villages
# associated with one GP (i.e. a level of government is called "GP").
# Figure~\ref{fig:women_desc} below shows the names and descriptions of the variables
# in the dataset. The authors hypothesize that female politicians are more likely to
# support policies female voters want. Researchers found that more women complain about
# the quality of drinking water than men. You need to estimate the effect of the
# reservation policy on the number of new or repaired drinking water facilities
#in the villages.
# Names and description of variables from Chattopadhyay and Duflo (2004)
# 1 `GP`  Identifier for the Gram Panchayat   
# 2 `village`  identifier for each village
# 3 `reserved`  binary variable indicating whether the GP was reserved for women leaders or not
# 4 `female`  binary variable indicating whether the GP had a female leader or not
# 5 `irrigation` variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
# 6 `water`  variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

#\item [(a)] State a null and alternative (two-tailed) hypothesis.
# null: no diff in incidence of new or repaired drinking-water facilities
#  in the village since the reserve policy started
#   ie 'water' is independent of 'reserved'
# alternate: the incidence of new or repaired drinking-water facilities is
#  correlated to the reservation policy


policy <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
#write.csv(policy,"Data/policy.csv")
policy<-read_csv("Data/policy.csv")

summary(policy)
```

```
pairs(policy[4:7])

sum(policy$reserved)
sum(policy$female)


#\item [(b)] Run a bivariate regression to test this hypothesis in \texttt{R} (include your code!).

water <- lm(water ~ reserved , data = policy)
summary(water)

output_stargazer(water, outputFile="water_model.tex", type = "latex",
                 appendVal=FALSE,
                 title="Pearson Linear Regression - Water ~ Reserved",
                 style = "apsr",
                 label = "water_reserved"
)

p<- ggplot(policy, aes(reserved, water, colour=female, group_by(female)))
p + geom_jitter()
ggsave("resrvd_water.png")

p<- ggplot(policy, aes(reserved, water, group_by(reserved)))
p + geom_boxplot( outlier.size = 3, aes(group=reserved))
ggsave("resrvd_water_boxplot.png")


# there are more outliers in the reserved=1 cohort

# assumption is that each village is a separate case and each case is independent
# but, each GP relates to 2 villages - need to check for impact of combining villages


p<- ggplot(policy, aes(reserved, water, group_by(reserved)))
p + geom_boxplot( outlier.size = 3, aes(group=reserved)) +
  facet_wrap(policy$village)
ggsave("village_water_boxplot.png")


policy %>%
  group_by(reserved, village) %>%
  summarise(n = n(), sum_water = sum(water)) %>%
  mutate(prop_reserved = round(n / sum(n), 4), sum_water) %>% # mutate after our summarise to find the proportion
  arrange(desc(prop_reserved))

reserved_water_tab <- policy %>%
  group_by(reserved) %>%
```

```
  summarise(n = n(), sum_water = sum(water)) %>%
  mutate(prop_reserved = round(n / sum(n), 4), sum_water, prop_water_reserved =
           round(sum_water / sum(sum_water), 4)) %>% # mutate after our summarise to find the proportion
  arrange(desc(prop_reserved))

str(reserved_water_tab)

reserved_water_tab


sum(policy$water)


# todo document ignoring the paired village phenomenon??
# or combine the villages

combined_village_policy <- policy %>%
  group_by(GP) %>%
  mutate (sum_water = sum(water), sum_irrigation = sum(irrigation)) %>%
  select(GP, reserved, female, sum_water, sum_irrigation) %>%
  unique()


cvp <- lm(sum_water/2 ~ reserved, data = combined_village_policy)


summary(cvp)


one_village_policy <- policy %>%
  group_by(GP) %>%
  filter(village ==1)

two_village_policy <- policy %>%
  group_by(GP) %>%
  filter(village ==2)


#todo - work out how to bin data
chisq.test(x = one_village_policy$water, y = two_village_policy$water)
chisq_12 <- chisq.test(x = one_village_policy$water, y = two_village_policy$water)



t.test(x = policy$water, y = one_village_policy$water,var.equal = FALSE, conf.level = 0.1)

# Welch Two Sample t-test
```

```
#data:  policy$water and one_village_policy$water
#t = 0.78558, df = 392.33, p-value = 0.4326
#alternative hypothesis: true difference in means is not equal to 0
#10 percent confidence interval:
#  1.859858 2.568713
#sample estimates:
#  mean of x mean of y
#17.84161  15.62733

t.test(x = policy$water, y = two_village_policy$water,var.equal = FALSE, conf.
    level = 0.1)
#Welch Two Sample t-test#

#data:  policy$water and two_village_policy$water
#t = -0.61008, df = 279.55, p-value = 0.5423
#alternative hypothesis: true difference in means is not equal to 0
#10 percent confidence interval:
#  -2.670789 -1.757783
#sample estimates:
#  mean of x mean of y
#17.84161  20.05590


t.test(x = one_village_policy$water, y = two_village_policy$water,var.equal =
    FALSE, conf.level = 0.1)
#
#Welch Two Sample t-test
#
#data:  one_village_policy$water and two_village_policy$water
#t = -1.1805, df = 281.19, p-value = 0.2388
#alternative hypothesis: true difference in means is not equal to 0
#10 percent confidence interval:
#  -4.900404 -3.956738
#sample estimates:
#  mean of x mean of y
#15.62733  20.05590


village <- lm(water ~ village , data = policy)
summary(village)

output_stargazer(village, outputFile="village_water_model.tex", type = "latex"
    ,
                  appendVal=FALSE,
                  title="Pearson Linear Regression - Water ~ Village",
                  style = "apsr",
                  label = "water_village"
)

###--------------------------------------------------------------------------
```

```

water_female <- lm(water ~ female , data = policy)

summary(water_female)

plot(water)
#\item [(c)] Interpret the coefficient estimate for reservation policy.

with(policy, plot(water, reserved))
p<- ggplot(policy)
p+ geom_jitter(aes(reserved, water, colour=female))

p<- ggplot(policy, aes(reserved, water), colour=female) + geom_jitter()
p+ facet_wrap(vars(female))
```