Test of Hypothesis

2024-07-08

```
options(digits = 5) #set 5 significant figures for my workings
library(TeachingDemos)
```

Hypothesis Testing is a technique used to asses the statistically significance difference of models or populations.

Procedures in TOH

1. State parameter of interest
2. State the hypothesis
3. State the level of significance (L.O.S) and determine critical region
4. Compute test statistics
5. Compare critical region and the test statistics
6. Make conclusions based on comparison as to whether you'll reject or accept H_0

We'll check at some few scenarios of this

1. One sample Test for Mean

A. Variance is known

Lower Tail test

The hypothesis to be tested is $H_0 : \mu = \mu_0 \; vs \; H_1 : \mu \leq \mu_0$

When variance is known, a z-test is used to test hypothesis where $Z = \frac{\sqrt{n}(\bar{X}-\mu_0)^2}{\sigma} \sim \mathcal{N}(\prime, \infty)$

We reject H_0 when $Z < Z_{-\alpha}$, $Z_{-\alpha}$ is the tabulated test statistics and Z is the computed one

We also reject h_0 when p-value is less than the significance level

QST 1: Suppose the manufacturer claims that the mean lifetime of a light bulb is more than 10,000 hours. For a sample of 30 light bulbs, the mean lifetime turns out to be only 9,900 hours. Assume the population standard deviation to be 120 hours. At .05 significance level, can we reject the claim by the manufacturer?

$H_0 : \mu > 10,000 \; vs \; \mu \leq 10,000$

$z = \frac{\sqrt{30}(9900-10000)}{120}$

```
#sample mean
xbar=9900
#null hypothesis
mu0=10000
#population sd
sigma=120
#sample size
n=30
#computed test statistics
z=(sqrt(n)*(xbar-mu0))/(sigma);z
```

```
## [1] -4.5644
```

```
#tabulated test statistics
alpha=.05
(z.alpha=qnorm(alpha,lower.tail=T))
```

```
## [1] -1.6449
```

since $Z = -4.564355 < z_{-\alpha} = -1.644854$ we reject H_0 and thus the mean lifetime of bulbs is not greater than 10,000 hours

```
#using lower tail p_value to test hypothesis
(pval=pnorm(z,lower.tail = T))
```

```
## [1] 2.5052e-06
```

p-value=0.000002505166 which is less than the significance level .05, the we reject H_0

QST 1B:

Under same conditions as the problem above, can we reject the manufacturer's claim on the lifetime of light bulbs at .01 significance level?

```
#tabulated statistics
alpha=.01
(z.alpha<-qnorm(alpha,lower.tail = T))
```

```
## [1] -2.3263
```

$-4.564355 < -2.326348$ thus we still reject H_0

also the p-value is less than $\alpha = .01$, thus we reject H_0

QST 2: Suppose the following are the lifetime hours from a random sample of light bulbs. Assuming the population standard deviation to be 120 hours, can we reject a manufacturer's claim that the light bulbs last more than 10,000 hours at .05 significance level?

$H_0 : \mu > 10000$ *vs* $H_1 : \mu \leq 10000$

```
#load data in R
x<-scan("TOH_used_datasets/lightbulbs.txt",sep=" ")
xbar=mean(x)
mu0=10000
n=30
sigma=120
(z=(sqrt(n)*(xbar-mu0))/(sigma))
```

```
## [1] -4.006
```

```
(z.alpha=qnorm(.05,lower.tail = T))
```

```
## [1] -1.6449
```

```
pval<-pnorm(-4.006)
cat(z,"<",z.alpha,"\n",pval,"<",.05,"\n","reject null hypothesis")
```

```
## -4.006 < -1.6449
##  3.0878e-05 < 0.05
##  reject null hypothesis
```

Alternative solution

```r
install.packages("TeachingDemos")
library("TeachingDemos")
```

```r
(test=z.test(x,mu=mu0,stdev = sigma,alternative = "less"))
```

```
##
##  One Sample z-test
##
## data:  x
## z = -4.01, n = 30.0, Std. Dev. = 120.0, Std. Dev. of the sample mean =
## 21.9, p-value = 3.1e-05
## alternative hypothesis: true mean is less than 10000
## 95 percent confidence interval:
##    -Inf 9948.3
## sample estimates:
## mean of x
##    9912.2
```

```r
test$p.value
```

```
## [1] 3.088e-05
```

Upper Tail Test

The hypothesis tested is

$H_0 : \mu = \mu_0 \ vs \ \mu > \mu_0$

We reject H_0 when $Z > Z_\alpha$

QST1; Suppose the food label on a cookie bag states that there are at most 2 grams of saturated fat in a single cookie. In a sample of 35 cookies, it is found that there are 2.1 grams of saturated fat per cookie on average. Assume the population standard deviation to be 0.25 grams. At .05 significance level, can we reject the claim on food label?

We are testing the hypothesis $H_o : \mu \leq 2 \ vs \ H_1 : \mu > 2$

$Z = \frac{\sqrt{35}(2.1-2)}{0.25} \sim N(0,1)$

```r
xbar=2.1
n=35
mu0=2
sd=.25
(z=(sqrt(n)*(xbar-mu0))/sd)
```

```
## [1] 2.3664
```

```r
(z.test=qnorm(.05,lower.tail = F))
```

```
## [1] 1.6449
```

```r
(pnorm(-1.6449,lower.tail = F))
```

```
## [1] 0.95
```

since $2.3664 > 1.6449$ we reject H_0

Since our p-value$> \alpha$ we still reject H_0

QST 1B:

Under same conditions as the problem above, can we reject the claim on saturated fat as stated in the food label at .01 significance level?

```
(z.test=qnorm(.01,lower.tail = F))
```

```
## [1] 2.3263
```

$2.3263 > 1.6449$ thus we reject H_0

QST2: Suppose the following are the gram amount of saturated fat found in a random sample of cookies. Assuming the population standard deviation to be 0.25 grams, at .05 significance level, can we reject the claim in the food label that there are at most 2 grams of saturated fat per cookie?

$H_0 : \mu \leq 2 \; vs \; H_1 : \mu > 2$

```
x<-scan("TOH_used_datasets/cookies.txt")
xbar=mean(x)
n=length(x)
sd=.25
alpha=.05
mu0=2
(z<-(sqrt(n)*(xbar-mu0))/sd)
```

```
## [1] 0.41331
```

```
(qnorm(.05,lower.tail = F))
```

```
## [1] 1.6449
```

```
(pnorm(z,lower.tail = F))
```

```
## [1] 0.33969
```

$0.41331 < 1.6449$ and $0.33969 < 0.05$ we fail to reject H_0

Alternative

```
z.test(x,mu=mu0,stdev=0.25,alternative="greater")
```

```
##
##  One Sample z-test
##
## data:  x
## z = 0.413, n = 35.0000, Std. Dev. = 0.2500, Std. Dev. of the sample
## mean = 0.0423, p-value = 0.34
## alternative hypothesis: true mean is greater than 2
## 95 percent confidence interval:
##  1.948   Inf
## sample estimates:
## mean of x
##    2.0175
```

Two Tail Test

We test the hypothesis $H_0 : \mu = \mu_0 \; vs \; H_1 : \mu \neq \mu_0$

H_0 is rejected iff $|Z| > Z\frac{\alpha}{2}$

QST1: Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the population standard deviation to be 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

$H_0 : \mu = 15.4 \; vs \; H_o1 : \mu \neq 15.4$

If $|Z = \frac{\sqrt{(35(14.6-15.4)}}{2.5}| > Z_{1-\frac{\alpha}{2}} = Z_{0.025}$ we reject H_0

```
xbar=14.6
mu0=15.4
n=35
sd=2.5
(z<-(sqrt(n)*(14.6-15.4))/sd)
```

```
## [1] -1.8931
```

```
(z.test<-qnorm(1-0.05/2))
```

```
## [1] 1.96
```

Since $|1.8931| > 1.95$ we reject H_0

Alternative

To get the p-value of a two tail z-test, we double the p-value of the lower tail z-test

```
(pval=2*pnorm(z))
```

```
## [1] 0.058339
```

$0.058339 > 0.050$ thus rejection of H_0

Qst 1B:

Under same conditions as the problem above, can we reject the null hypothesis at .01 significance level that the mean penguin weight stays the same as last year?

reject $H_0 \; iff \; |Z| > Z_{1-\frac{0.01}{2}}$

```
qnorm(.01/2,lower.tail = F)
```

```
## [1] 2.5758
```

Alternative

```
(pval=2*pnorm(z,lower.tail = T))
```

```
## [1] 0.058339
```

QST 2 Suppose the following are body weight of King Penguins in kilograms found in a random sample within a colony. Assuming the population standard deviation to be 2.5 kg, at .05 significance level, can we reject the null hypothesis that the mean penguin weight is still 15.4 kg just like last year?

$H_0 : \mu = 15.4 \; vs \; H_1 : \mu \neq 15.4$

If $|Z = \frac{\sqrt{35}(14.775-15.4)}{2.5}| > Z_{0.025}$

```
x<-scan("TOH_used_datasets/penguins.txt")
xbar=mean(x)
mu0=15.4
n=length(x)
```

```r
sd=2.5
(z<-(sqrt(n)*(xbar-mu0))/sd)
```

```
## [1] -1.4799
```

```r
(qnorm(0.05/2,lower.tail = F))
```

```
## [1] 1.96
```

```r
(pval=2*pnorm(z,))
```

```
## [1] 0.1389
```

We reject H_0 since $|Z| = |1.4799| > Z_{0.025} = 1.96$ and $p - value = 0.1389 > \alpha = 0.05$

Alternative

```r
library("TeachingDemos")
z.test(x,mu=15.4,stdev=2.5,alternative="two.sided")
```

```
##
##  One Sample z-test
##
## data:  x
## z = -1.48, n = 35.000, Std. Dev. = 2.500, Std. Dev. of the sample mean
## = 0.423, p-value = 0.14
## alternative hypothesis: true mean is not equal to 15.4
## 95 percent confidence interval:
##  13.946 15.603
## sample estimates:
## mean of x
##    14.775
```

B. Variance is Unknown (student t-distribution)

Two Tail Test

Just like before the hypothesis tested is $H_0 : \mu = \mu_0 \; vs \; H_1 : \mu \neq \mu_0$

The test statistics is of the form:

$t = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \sim t(n-1)$, where $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$

We reject H_0 when $|t| > t_{\frac{\alpha}{2}}(n-1)$ and when $p - value < \alpha$

QST1: Suppose the mean weight of King Penguins found in an Antarctic colony last year was 15.4 kg. In a sample of 35 penguins same time this year in the same colony, the mean penguin weight is 14.6 kg. Assume the sample standard deviation to be 2.5 kg. At .05 significance level, can we reject the null hypothesis that the mean penguin weight does not differ from last year?

$H_0 : \mu = 15.4 \; vs \; H_1 : \mu \neq 15.4$

$t = \frac{\sqrt{35}(14.6-15.4)}{2.5}$

```r
xbar=14.6
mu0=15.4
sd=2.5
n=35
(t<-(sqrt(35)*(14.6-15.4))/sd)
```

```
## [1] -1.8931
```

```
(t.test<-qt(.05/2,df=n-1,lower.tail = F))
```

```
## [1] 2.0322
```

since $|1.8931| < 2.0322$ we fail to reject H_0

Alternative

```
(pval=2*pt(t,df=n-1))
```

```
## [1] 0.066876
```

since $.066876 > .05$ we fail to reject H_0

QST 1B:

Under same conditions as the problem above, can we reject the null hypothesis at .01 significance level that the mean penguin weight stays the same as last year?

```
(t.test<-qt(.01/2,n-1,lower.tail = F))
```

```
## [1] 2.7284
```

$1.8931 < 2.7284$ thus we fail to reject H_0

QST 2: Suppose the following are body weight of King Penguins in kilograms found in a random sample within a colony. Without knowledge of the population standard deviation, at .05 significance level, can we reject the null hypothesis that the mean penguin weight is still 15.4 kg just like last year?

$H_0 : \mu = 15.4 \; vs \; H_1 : \mu \neq 15.4$

```
x<-scan("TOH_used_datasets/penguins2.txt")
xbar=mean(x)
n=length(x)
sd=sd(x)
mu0=15.4
(t<-(sqrt(n)*(xbar-mu0))/sd)
```

```
## [1] -1.3943
```

```
(t_test<-qt(.05/2,n-1,lower.tail = F))
```

```
## [1] 2.0322
```

```
(pval<-2*pt(t,n-1))
```

```
## [1] 0.17227
```

$|1.3943| < 2.0322$ we fail to reject the null hypothesis and since $0.17227 > 0.05$ we fail to reject H_0

2. Hypothesis test for population proportions

Lower Tail Test

We test the hypothesis $H_0 : \mu > p_0 \; vs \; H_1 : \mu \leq p_0$

Let $X \sim f_X(x)$ where $f_X(x)$ is a certain distribution with mean and variance $\mu \; and \; \sigma^2$ respectively

Under Central Limit Theorem,

$\sum X_i \sim N(\mu n, \sigma^2 n)$

Thus our z-statistics is defined as: $Z = \frac{x-\mu n}{\sigma\sqrt{n}} = \frac{\bar{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Where p is the proportion/probability

QST 1: Suppose 60% of citizens voted in the last election. 85 out of 148 people in a telephone survey said that they voted in current election. At 0.5 significance level, can we reject the null hypothesis that the proportion of voters in the population is above 60% this year?

$$p(x = 85) = p(Z = \frac{85 - 148(0.6}{\sqrt{0.6(1-0.6) * 148}}) = \phi(\frac{85 - 148(0.6}{\sqrt{0.6(1-0.6) * 148}}) = \phi(\frac{\frac{85}{148} - 0.6}{\sqrt{0.6(1-0.6)/148}})$$

```
pbar<-85/148 #sample proportion
p0=.6 #hypothesis
n=148 #sample size
(z=(pbar-p0)/sqrt(p0*(1-p0)/n)) #test statistics
```

```
## [1] -0.6376
```

```
alpha=.05
(z.alpha<-qnorm(0.05,lower.tail = T))
```

```
## [1] -1.6449
```

Since $-0.6376 > -1.6649$ we fail to reject H_0

Alternative 1

```
pnorm(z)
```

```
## [1] 0.26187
```

Since $0.26187 > 0.05$ we fail to reject H_0

Alternative 2 :computes p-value directly

```
prop.test(85,148,0.6,alt="less",correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  85 out of 148, null probability 0.6
## X-squared = 0.407, df = 1, p-value = 0.26
## alternative hypothesis: true p is less than 0.6
## 95 percent confidence interval:
##  0.00000 0.63925
## sample estimates:
##       p
## 0.57432
```

QST 1B: Under same conditions as the problem above, can we reject the null hypothesis at.01 significance level that the voting percentage is above 60% this year?

```
qnorm(0.01,lower.tail = T)
```

```
## [1] -2.3263
```

$-0.6376 > -2.3263$ we fail to reject H_0, also our $p.value > \alpha$ thus we fail to reject H_0

Upper Tail

We test the hypothesis $H_0 \mu \leq p_o \; vs \; H_1 : \mu > p_o$

test statistics is $Z = \frac{\bar{b} - p_0}{\sqrt{p_0(1-p_0)/n}}$

We reject H_0 iff $Z > Z_{\frac{\alpha}{2}}$

QST 1: Suppose that 12% of apples harvested in an orchard last year was rotten. 30 out of 214 apples in a harvest sample this year turns out to be rotten as well. At .05 significance level, can we reject the null hypothesis that the proportion of rotten apples in harvest stays below 12% this year?

$Z = \frac{30/214 - 0.12}{\sqrt{0.12(1-0.12)/214}}$

```
pbar=30/214
p0=0.12
n=214
(z<-(pbar-p0)/sqrt(p0*(1-p0)/n))
```

```
## [1] 0.90875
```

```
(z.test<-qnorm(.05,lower.tail = F))
```

```
## [1] 1.6449
```

```
pnorm(z,lower.tail = F)
```

```
## [1] 0.18174
```

since $0.90875 < 1.6449$ *and* $0.18174 > 0.05$ we fail to reject null hypothesis

Alternative

```
prop.test(30,214,p=0.12,alt="greater",correct =F )
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  30 out of 214, null probability 0.12
## X-squared = 0.826, df = 1, p-value = 0.18
## alternative hypothesis: true p is greater than 0.12
## 95 percent confidence interval:
##  0.10563 1.00000
## sample estimates:
##       p
## 0.14019
```

Two Tail

The test hypothesis is $H_0 : \mu = p_0 \; vs \; H_1 : \mu \neq p_0$

test statistics is of the form $Z = \frac{\bar{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

We reject the null hypothesis if $|Z| > Z_{\frac{\alpha}{2}}$

QST 1: Suppose a coin toss turns up 12 heads out of 20 trials. At .05 significance level, can one reject the null hypothesis that the coin toss is fair?

$Z = \frac{12/20 - 0.5}{0.5(1-0.5)/20}$

```r
pbar=12/20
p0=.5
n=20
(z<-(pbar-p0)/sqrt(p0*(1-p0)/n))
```

```
## [1] 0.89443
```

```r
(z.test<-qnorm(.05/2,lower.tail = F))
```

```
## [1] 1.96
```

```r
(pval=2*pnorm(z,lower.tail=F))
```

```
## [1] 0.37109
```

Alternative

```r
prop.test(12,20,p=.5,alt="two.sided",correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  12 out of 20, null probability 0.5
## X-squared = 0.8, df = 1, p-value = 0.37
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.38658 0.78119
## sample estimates:
##   p
## 0.6
```

QST 1B:

Under same conditions as the problem above, can we reject the null hypothesis at .01 significance level that the coin is fair?

$Z = \frac{(\bar{p}-0.5)}{\sqrt{0.5(1-0.5)/20}}$, $\bar{p} = \frac{12}{20}$

```r
(z<-(pbar-p0)/sqrt(p0*(1-p0)/20))
```

```
## [1] 0.89443
```

```r
(z.test=qnorm(.01/2,lower.tail = F))
```

```
## [1] 2.5758
```

Since $0.89443 < 2.5758$ we fail to reject H_0

Alternative

```r
prop.test(12,20,p=.5,alt="two.sided",conf.level = .99,correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  12 out of 20, null probability 0.5
## X-squared = 0.8, df = 1, p-value = 0.37
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.32931 0.82087
## sample estimates:
```

```
##   p
## 0.6
```

Since $0.37 > 0.01$ we fail to reject H_0

Multinomial Goodness of Fit

A data population is called `multinomial` if its categorical and has been classified into a collection of discrete non-overlapping classes.

The chi-square goodness of fit test is used to test whether the observed events are similar to the expected ones. The null hypothesis is that the observed frequencies is equal to the expected frequencies.

$H_0 : f_i = e_i \; ve \; H_1 : f_i \neq e_i$, where

f_i - observed frequencies

e_i- expected frequencies

The test statistics is given as :

$Q = \sum_{i=1}^{n} \frac{f_i - e_i)^2}{e_i} \sim \chi^2(n-1)$

In the built-in data set survey, the column Smoke contains survey response about student smoking habit. As there are exactly four proper response in the survey: “Heavy”, “Regul” (regularly), “Occas” (occasionally) and “Never”, the Smoke data column is multinomial.

```
library(MASS)
levels(survey$Smoke)
```

```
## [1] "Heavy" "Never" "Occas" "Regul"
```

```
# using table to get frequencies of our qualitative data (categorical\factors)
(smoke_freq<-table(survey$Smoke))
```

```
##
## Heavy Never Occas Regul
##    11   189    19    17
```

Suppose the campus smoking statistic is as below. Determine whether it issupported by the sample data in survey at .05 significance level. Heavy 4.5% Never 79.5% Occas 8.5% Regul 7.5%

```
smoke.prob<-c(0.045,0.795,0.085,0.075)
chisq.test(smoke_freq,p=smoke.prob)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  smoke_freq
## X-squared = 0.107, df = 3, p-value = 0.99
```

Since p-value 0.991 is greater than the .05 significance level, we do not reject the null hypothesis that the sample data in survey supports the campus-wide smoking statistic.

Using the smoking habit data in survey, demonstrate how to conduct the Chi- squared goodness of fit test by computing the p-value with the textbook formula.

```
#observed frequencies
f<-table(survey$Smoke)
#expected frequencies
e<-smoke.prob*length(survey$Smoke)
```

```r
# difference of observed and expected frequencies
d<-f-e
#test statistics
q<-sum(d^2/e);q
```

```
## [1] 0.11121
```

```r
#degress of freedom
df<-length(f)-1
#p-value
(p_val<-pchisq(q,df=df,lower.tail=F))
```

```
## [1] 0.99046
```

Chi-squared Test of Independence

Two R.v are independent if the presence of one does not affect the other. Our null hypothesis is that the random variables are dependent on each other.

The chi-square goodness of fit test can also be used to test for this independence. The test statistics is given as:

$Q = \sum_{i=1}^{n} \frac{f_{ij}-e_{ij})^2}{e_{ij}}$, where;

f_{ij} - observed frequency in the i$^{\text{th}}$ category of x and in the j$^{\text{th}}$ category of y.

Consider the inbuilt data set of survey,the Smoke column records the students smoking habit, while the Exer column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

Let's tally the smoking habits against exercise level

```r
#contigency table of smoking habits aganist exercise level
(tbl<-table(survey$Smoke,survey$Exer))
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```r
chisq.test(tbl)
```

```
## Warning in chisq.test(tbl): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  tbl
## X-squared = 5.49, df = 6, p-value = 0.48
```

The warning is as a result of the small cells in our contingency table. To offset it, we should combine the 2$^{\text{nd}}$ and 3$^{\text{rd}}$ columns

```r
#combine columns
(ctbl<-cbind(
```

```
  tbl[,"Freq"],
  tbl[,"None"] + tbl[,"Some"]
))
```

```
##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8
```

```
#perform test
chisq.test(ctbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctbl
## X-squared = 3.23, df = 3, p-value = 0.36
```

Since $p - value > \alpha = 0.05$ we reject the null hypothesis implying that the random variables are independent of each other.

Using the student data of smoking and exercise in survey, demonstrate how to conduct the Chi-squared independence test by computing the p-value with the textbook formula.

```
# get tally for observed frequencies
(f<-table(survey$Smoke,survey$Exer))
```

```
##
##         Freq None Some
##   Heavy    7    1    3
##   Never   87   18   84
##   Occas   12    3    4
##   Regul    9    1    7
```

```
#get rowsums
(row_sum<-apply(f,1,sum))
```

```
## Heavy Never Occas Regul
##    11   189    19    17
```

```
#get columns sums
(col_sums<-apply(f,2,sum))
```

```
## Freq None Some
##  115   23   98
```

```
#get product of the 2 sums
p<-row_sum%*%t(col_sums);p
```

```
##       Freq None  Some
## [1,]  1265  253  1078
## [2,] 21735 4347 18522
## [3,]  2185  437  1862
## [4,]  1955  391  1666
```

```
#get expected frequencies by diving product with sample size
e<-p/nrow(survey);e
```

```
##         Freq    None    Some
```

```
## [1,]  5.3376  1.0675  4.5485
## [2,] 91.7089 18.3418 78.1519
## [3,]  9.2194  1.8439  7.8565
## [4,]  8.2489  1.6498  7.0295
```

```
#get difference of observed and expected frequencies
d<-f-e
#get test statistics
q<-sum(d^2/e)
#get df
df<-(nrow(f)-1)*(ncol(f)-1);df
```

```
## [1] 6
```

```
#get p-value
(p_val<-pchisq(q,df=df,lower.tail = F))
```

```
## [1] 0.47952
```

Here we are able to see better why we combine the second and 3rd columns... The reason being that some of the cells in the expected frequencies are less than 5. For better approximations using CLT (Central Limit Theorem), the expected frequencies(mean)>5

Just like before, using the *chisq.test*() let's combine this columns

```
#combined columns of the observed frequency
(cf<-cbind(
  f[,"Freq"],
  f[,"None"]+f[,"Some"]
))
```

```
##       [,1] [,2]
## Heavy    7    4
## Never   87  102
## Occas   12    7
## Regul    9    8
```

```
#combined columns for expected frequncies
(ce<-cbind(
  e[,"Freq"],
  e[,"None"]+e[,"Some"]
))
```

```
##         [,1]    [,2]
## [1,]  5.3376  5.6160
## [2,] 91.7089 96.4937
## [3,]  9.2194  9.7004
## [4,]  8.2489  8.6793
```

```
#get diffeerence
cd<-cf-ce
#get degrees of freedom
cdf<-(nrow(cf)-1)*(ncol(cf)-1)
#test statistics
cq<-sum(cd^2/ce);cq
```

```
## [1] 3.2507
```

```
#p-value
(p_val<-pchisq(cq,df=cdf,lower.tail = F))
```

```
## [1] 0.35456
```

Non-Parametric Testing methods

A test statistics is called non-parametric if it makes no assumption on the underlying data,

Non-parametric tests are more flexible, robust and applicable to qualitative data. However they are not as powerful as parametric tests with assume that data is quantitative, normally distributed and has a sufficiently large sample size

Sign Test

It is used to determine wheter a ppopulation with a binomial diatribution has equal chances of success or failure

```
print("hello world")
```

```
## hello world
```