This is the R Illustration for Unit3 Discrete Distributions and Fisher Exact Test.

Setting of the dataset: Calls to the New York Auto Club are possibly related to the weather, with more calls occurring during bad weather. This example illustrates descriptive analyses and simple linear regression to explore this hypothesis in a data set containing information on calendar day, weather, and numbers of calls.

Firstly, let’s download the dataset.

```r
library(foreign)
url <- "http://people.umass.edu/biep640w/datasets/ers.dta"
dat <- read.dta(file = url) # read.dta: Read Stata data
dat
head(dat)
```

```
## day calls fhigh flow high low rain snow weekday year sunday subzero
## 1 12069 2298 38 31 39 31 0 0 0 0 0 0
## 2 12070 1709 41 27 41 30 0 0 0 0 1 0
## 3 12071 2395 33 26 38 24 0 0 0 0 0 0
## 4 12072 2486 29 19 36 21 0 0 1 0 0 0
## 5 12073 1849 40 19 43 27 0 0 1 0 0 0
## 6 12074 1842 44 30 43 29 0 0 1 0 0 0
```

```
dim(dat) # 28*12
```

```
## [1] 28 12
```

```
## str(dat)
## summary(dat)
```

1. Binomial Distribution

```
## 1a. Probability Calculations
# Binomial(n, pi): Probability of exactly k events, Pr[X = k]
# Binomial(n=20, pi=.03), Prob[X=2] is:
dbinom(2, 20, 0.03)
```

```
```r
## [1] 0.09882967

# Binomial(n, pi): Probability of at most k events, Pr[X < k]
# Binomial(n=20, pi=.03) Prob[X <= 2]

pbinom(2, 20, 0.03)

## [1] 0.9789916

# Binomial(n=20, pi=.03) Prob[X < 2]

pbinom(1, 20, 0.03)

## [1] 0.880162

# Binomial(n=20, pi=.03) Prob[X >= 2]

1-pbinom(1, 20, 0.03)

## [1] 0.119838

# Binomial(n=20, pi=.03) Prob[X > 2]

1-pbinom(2, 20, 0.03)

## [1] 0.02100836

# 1b. Confidence Interval Estimation

binom.test(x=9, n=40, p = 0.25, alternative = "two.sided", conf.level = 0.90)

$conf.int

## [1] 0.1227117 0.3597924

attr("conf.level")

## [1] 0.9

# For a 0/1 variable (subzero) in the data set ers.dta

dat$subzero

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0

nsubzero <- length(dat$subzero)
nsubzero

## [1] 28

table(dat$subzero)

##
## 0 1
## 23 5
```
binom.test(x=5, n=nsubzero, p = 0.25, alternative = "two.sided", conf.level = 0.90)$conf.int

## [1] 0.07311377 0.33940197
## attr(,"conf.level")
## [1] 0.9

# 1c. Hypothesis Test for 1 Binomial
# Exact Test for 1 Binomial (Example: n=40, x=9, nullp = .25)
binom.test(x=9, n=40, p = 0.25, alternative = "two.sided", conf.level = 0.90)

##
## Exact binomial test
##
## data: 9 and 40
## number of successes = 9, number of trials = 40, p-value = 0.8556
## alternative hypothesis: true probability of success is not equal to 0.25
## 90 percent confidence interval:
## 0.1227117 0.3597924
## sample estimates:
## probability of success
## 0.225

# For a 0/1 variable (subzero) in the data set ers.dta
binom.test(x=5, n=nsubzero, p = 0.25, alternative = "two.sided", conf.level = 0.90)

##
## Exact binomial test
##
## data: 5 and nsubzero
## number of successes = 5, number of trials = 28, p-value = 0.5136
## alternative hypothesis: true probability of success is not equal to 0.25
## 90 percent confidence interval:
## 0.07311377 0.33940197
## sample estimates:
## probability of success
## 0.1785714

# 1d. Hypothesis Test for 2 Independent Binomials
# Normal approx Test for Equality of 2 Binomials (n1=49, x1=2, n2=40, x2=9)
# prop.test: Test of Equal or Given Proportions
prop.test(c(2,9),c(49,40),correct=FALSE)

## Warning in prop.test(c(2, 9), c(49, 40), correct = FALSE): Chi-squared approximation may be incorrect

##
## 2-sample test for equality of proportions without continuity correction
##
## data:  c(2, 9) out of c(49, 40)
## X-squared = 6.897, df = 1, p-value = 0.008634
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.32495169 -0.04341566
## sample estimates:
## prop 1 prop 2
## 0.04081633 0.22500000

prop.test(c(2, 9), c(49, 40), correct=FALSE)$p.value

## Warning in prop.test(c(2, 9), c(49, 40), correct = FALSE): Chi-squared approximation may be incorrect
## [1] 0.008633988

# For equality of event probability 0/1 variable (subzero) over 2 independent groups(sunday) in data set ers.dta
dat$subzero

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 0 0 1 0 0

table(dat$subzero)

##
## 0 1
## 23 5
dat$sunday

## [1] 0 1 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0

table(dat$sunday)

##
## 0 1
## 24 4

nsubzero <- length(dat$subzero)
nsubzero

## [1] 28

tsunday <- length(dat$sunday)
tsunday

## [1] 28
prop.test(c(5,4),c(nsubzero,nsunday),correct=FALSE)

## Warning in prop.test(c(5, 4), c(nsubzero, nsunday), correct = FALSE): Chi-
## squared approximation may be incorrect

## 2-sample test for equality of proportions without continuity
## correction
##
## data: c(5, 4) out of c(nsubzero, nsunday)
## X-squared = 0.13239, df = 1, p-value = 0.716
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.1564410 0.2278695
## sample estimates:
## prop 1 prop 2
## 0.1785714 0.1428571

prop.test(c(5,4),c(nsubzero,nsunday),correct=FALSE)$p.value

## Warning in prop.test(c(5, 4), c(nsubzero, nsunday), correct = FALSE): Chi-
## squared approximation may be incorrect
## [1] 0.7159691

2. Poisson Distribution

# 2a. Probability Calculations
# Poisson(mu): Probability of exactly k events, Pr[X = k]
# Pr[Poisson(mu=1.8) = 6]
dpois(6, 1.8)

## [1] 0.007808587

# Poisson(mu): Probability of at most k events, Pr[X <= k]
ppois(6, 1.8)

## [1] 0.9974306

# Poisson(mu): Probability of less than k events, Pr[X < k]
ppois(5, 1.8)

## [1] 0.989622
# Poisson(\(\mu\)): Probability of at least \(k\) events, \(Pr[X \geq k]\)
1-ppois(5, 1.8)  
## [1] 0.01037804

# Poisson(\(\mu\)): Probability of more than \(k\) events, \(Pr[X > k]\)
1-ppois(6, 1.8)  
## [1] 0.00256945

## 2b. Confidence Interval Estimation
poisson.test(x=2, alternative = "two.sided", conf.level = 0.90)$conf.int  
## [1] 0.3553615 6.2957936  
## attr(,"conf.level")  
## [1] 0.9

3. Central Hypergeometric Distribution

# Pr [Exactly \(a\) with exposure AND disease ]
# dhyper(x, m, n, k)
dhyper(x=2, m=4, n=255, k=23)  
## [1] 0.03829914

## x: the number of white balls drawn without replacement from an urn which contains both black 
## and white balls. Here \(x\) is the number of exposed people who got the disease.  
## m: the number of white balls in the urn. Here is number of people who got the disease.  
## n: the number of black balls in the urn. Here is number of people who didn't got the disease.  
## k: the number of balls drawn from the urn. Here is number of people who were exposed. 

# Pr [\(a\) or less with exposure AND disease ]
phyper(q=2, m=4, n=255, k=23)  
## [1] 0.99767

# Pr [\(a\) or more with exposure AND disease ]
1-phyper(q=1, m=4, n=255, k=23)  
## [1] 0.04062914
4. Fisher Exact Test for a Single 2x2 Table

# 4a. Smith, Delgado and Rutledge (1976) report data on ovarian carcinoma. Individuals had different numbers of courses of chemotherapy. The 5-year survival data for those with 1-4 and 10 or more courses of chemotherapy are: It is the same as HW4-Q3

Q4a <- data.frame(dead = c(21, 2), alive = c(2, 8))
Q4a

## dead alive
## 1 21 2
## 2 2 8

fisher.test(Q4a)

## Fisher's Exact Test for Count Data
## data: Q4a
## p-value = 0.0001255
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 3.819676 571.245711
## sample estimates:
## odds ratio
## 34.05494

# 4b.
Q4b <- data.frame(Diesease = c(9, 2), NoDisease = c(31, 47))
Q4b

## Disease NoDisease
## 1 9 31
## 2 2 47

# 2-sided:
fisher.test(Q4b, alternative = "two.sided")

## Fisher's Exact Test for Count Data
## data: Q4b
## p-value = 0.0108
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 1.263546 67.751190
## sample estimates:
## odds ratio
## 6.683415
# 1-sided:
fisher.test(Q4b, alternative = "greater")

## Fisher's Exact Test for Count Data
## data: Q4b
## p-value = 0.01001
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
## 1.547716  Inf
## sample estimates:
## odds ratio
## 6.683415

# 4c.
Q4c <- data.frame(Exposed = c(9, 31), Unexposed = c(2, 47))
Q4c

## Exposed Unexposed
## 1 9 2
## 2 31 47

# 2-sided:
fisher.test(Q4c, alternative = "two.sided")

## Fisher's Exact Test for Count Data
## data: Q4c
## p-value = 0.0108
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 1.263546 67.751190
## sample estimates:
## odds ratio
## 6.683415

# 1-sided:
fisher.test(Q4c, alternative = "greater")

## Fisher's Exact Test for Count Data
## data: Q4c
## p-value = 0.01001
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
## 1.547716  Inf
## sample estimates:
## odds ratio
## 6.683415
The approach and output is the same as 4c above.