Unit 4 – Categorical Data Analysis
Practice Problems (2 of 2)

Solutions

Before You Begin - R Users

__1. Be sure you have downloaded the dataset **`hersdata.Rdata`**

__2. The supplementary R illustration makes use of the following packages. Make sure you have done a one time installation of them, taking care to remember that R is case-sensitive.
`gmodels, DescTools, summarytools, tidyverse`

#1. *Source:* Rosenman RH, Friedman M, Straus R, Wurm M, Kositchek R, Hahn W and Werthessen NG (1964) A predictive study of coronary heart disease: the western collaborative group study. *Journal of the American Medical Association*, **189,** 113-120.

Note to class: This study was used in several data analysis illustrations in the book by Vittinghoff et al (Regression Methods in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models, 2nd edition, Springer, 2012).

This exercise gives you practice performing a **2 x K test of trend** for contingency table data.

The Western Collaborative Group Study (WCGS) was a prospective study of 3,154 men, all initially disease-free, who were followed for events of coronary heart disease (CHD). At the eight years follow-up mark, there were 257 events of CHD. Several potential predictors of CHD were of interest, including: age (at enrollment), cholesterol, systolic blood pressure, hematocrit, ECG status, smoking, and relative weight.

Age is a continuous variable, suggesting that it be modelled as such. However, the best way to model age is a question. Should it be treated as a linear predictor? Or modeled using a polynomial? Or modeled in some other way? To address this issue, an appropriate preliminary analysis might group study participants into age groups (a discrete predictor that is ordinal!). Following are the data:

		Age at Enrollment (years)					
CHD		35-40	41-45	46-50	51-55	55-60	Total =
	no	512	1036	680	463	206	2987
	yes	31	55	70	65	36	257
	Total =	543	1091	750	528	242	3154

1a. Perform a test of **general association**
H_O: The odds of event of CHD is independent of interval of age
H_A: The odds of event of CHD is associated with (differs with) interval of age

1b. Perform a test of **trend**
H_O: The odds of event of CHD is independent of interval of age
H_A: The odds of event of CHD is increases with interval of age

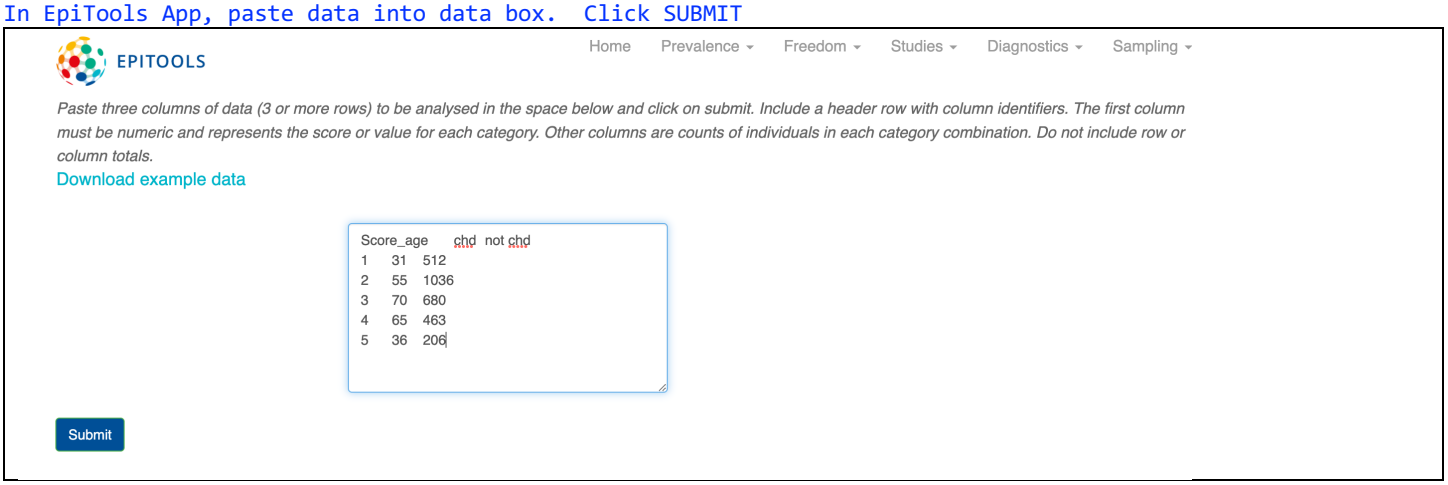
Online Apps Solution for BOTH questions #1A and #1B

EpiTools Chi Square Test for Trend
Link: https://epitools.ausvet.com.au/trend

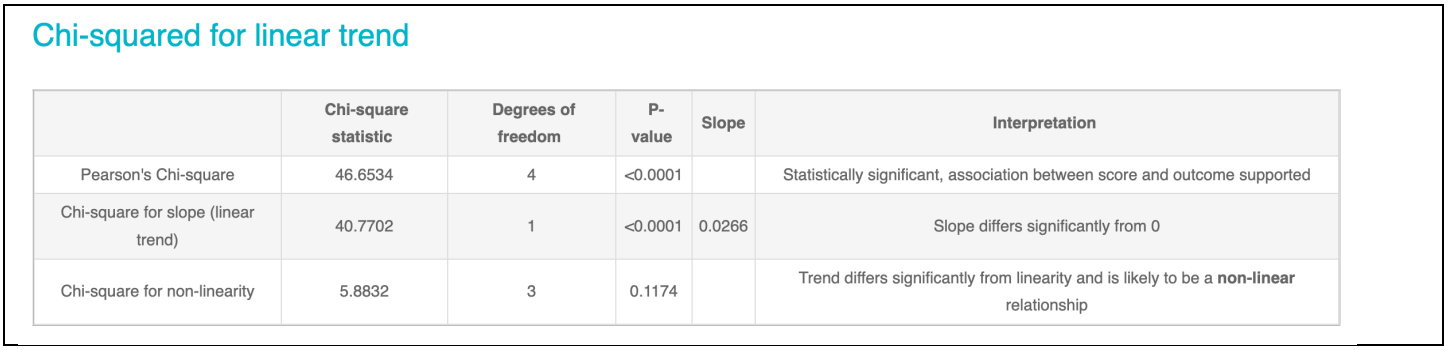
In Excel, enter data into 3 columns and provide headers

Score_age	chd	not chd
1	31	512
2	55	1036
3	70	680
4	65	463
5	36	206

In EpiTools App, paste data into data box. Click SUBMIT



You should then see:



Chi-squared for linear trend

	Chi-square statistic	Degrees of freedom	P-value	Slope	Interpretation
Pearson's Chi-square	46.6534	4	<0.0001		Statistically significant, association between score and outcome supported
Chi-square for slope (linear trend)	40.7702	1	<0.0001	0.0266	Slope differs significantly from 0
Chi-square for non-linearity	5.8832	3	0.1174		Trend differs significantly from linearity and is likely to be a **non-linear** relationship

Question #1A
H_0: No association of interval of age with event of CHD
H_A: General association of interval of age with event of CHD, two-sided
Chi square test (degrees of freedom = 4) = 46.65
p-value <<< .0001
Reject the null hypothesis. Assumption of the null hypothesis and application to the data has led to an extremely unlikely result. Conclude these data provide statistically significant evidence of an association.

Question #1B
H_0: No association of interval of age with event of CHD
H_A: Monotone increasing trend in event of CHD with increasing interval of age, one-sided

Chi square test (degrees of freedom = 1) = 40.77
p-value <<< .0001

Reject the null hypothesis. Assumption of the null hypothesis and application to the data has led to an extremely unlikely result. Conclude these data provide statistically significant evidence of a monotone increasing occurrence of event of CHD with advancing age.

R Solution

Create table

```
tableq1 <- as.table(rbind(c(512,1036,680,463,206),
                          c(31,55,70,65,36)))
dimnames(tableq1) <-list(CHD=c("No","Yes"),                    # Remember: order must be ordinal increasing
                         "Age (years)"=c("35-40", "41-45","46-50","51-55","55-60"))
cat("\nQuestion 1\nWCGS Data\n\n")

##
## Question 1
## WCGS Data

tableq1

##      Age (years)
## CHD   35-40 41-45 46-50 51-55 55-60
##   No    512  1036   680   463   206
##   Yes    31    55    70    65    36
```

Inspect Data

```
library(gmodels)                                                    # Attach package {gmodels}

cat("\nQuestion 1\nWCGS Data\n")
##
## Question 1
## WCGS Data

CrossTable(tableq1,prop.c=TRUE,prop.r=FALSE,prop.t=FALSE,prop.chisq = FALSE)     # Show column % only

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  3154
##
##
##              | Age (years)
##          CHD |     35-40 |     41-45 |     46-50 |     51-55 |     55-60 | Row Total |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##           No |       512 |      1036 |       680 |       463 |       206 |      2897 |
##              |     0.943 |     0.950 |     0.907 |     0.877 |     0.851 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------| Note:
##          Yes |        31 |        55 |        70 |        65 |        36 |       257 | Inspection of data shows that
##              |     0.057 |     0.050 |     0.093 |     0.123 |     0.149 |           | as age increases
## -------------|-----------|-----------|-----------|-----------|-----------|-----------| % with CHD is also increasing
## Column Total |       543 |      1091 |       750 |       528 |       242 |      3154 |
##              |     0.172 |     0.346 |     0.238 |     0.167 |     0.077 |           |
## -------------|-----------|-----------|-----------|-----------|-----------|-----------|
##
```

Q1a) Test of General Association

```
cat("\nQuestion 1a\nChi Square Test of General Association\nNull: No Association CHD w Interval Age\n\n")
##
## Question 1a
## Chi Square Test of General Association
## Null: No Association CHD w Interval Age

chisq.test(tableq1)
##
##  Pearson's Chi-squared test
##
## data:  tableq1
## X-squared = 46.653, df = 4, p-value = 0.000000001801
```

H_0: No association of interval of age with event of CHD
H_A: General association of interval of age with event of CHD, two-sided
Chi square test (degrees of freedom = 4) = 46.653
p-value <<< .0001
Reject the null hypothesis. Assumption of the null hypothesis and application to the data has led to an extremely unlikely result. Conclude these data provide statistically significant evidence of an association.

Q1b) Test of Trend

```
library(DescTools)                                                    # Attach package {DescTools}

cat("\nQuestion 1b\nChi Square Test of Trend\nNull: No Monotone Association CHD w Interval Age\n\n")
##
## Question 1b
## Chi Square Test of Trend
## Null: No Monotone Association CHD w Interval Age

DescTools::CochranArmitageTest(tableq1,alternative="increasing")
##
##  Cochran-Armitage test for trend
##
## data:  tableq1
## Z = -6.3852, dim = 5, p-value = 0.00000000008561       Z-score=-6.3852 -> Chi Square = [-6.3852]^2
## alternative hypothesis: increasing                                     = 40.77
```

H_0: No association of interval of age with event of CHD
H_A: Monotone increasing trend in event of CHD with increasing interval of age, one-sided

Chi square test (degrees of freedom = 1) = 40.77
p-value <<< .0001

Reject the null hypothesis. Assumption of the null hypothesis and application to the data has led to an extremely unlikely result. Conclude these data provide statistically significant evidence of a monotone increasing occurrence of event of CHD with advancing age.

#2. *Source: Triola MM and Triola MF. Biostatistics for the Biological and Health Sciences Boston: Pearson Addison Wesley, John Wiley, 2006. Chapter 10, Section 10-2. page 491.*

This exercise gives you practice performing a **chi square goodness of fit (GOF) test**.

Just briefly. Researchers suspect that people tend to self-report their weights lower than what they actually are and, in particular, tend to round down. If this is true, then one might expect to find that the last digits of self-reported weight are disproportionately often the digits "0", "1", "2", "3", "4" or "5". The following table are the values of the last digit of self-reported weights for a sample of n=80. Carry out an appropriately chi square goodness-of-fit (GOF) test to assess if there is statistically significant evidence in this sample of the suspected phenomenon of "rounding down" when self-reporting weight.

	Last Digit of Self-Reported Weight									
	0	1	2	3	4	5	6	7	8	9
Frequency	35	0	2	1	4	24	1	4	7	2

2a. What are the null (H_O) and alternative (H_A) hypotheses?
H_O: All ten (10) last digits of self-reported weight are equally likely (10% chance)
H_A: Not

2b. How many degrees of freedom does your chi square statistic have?
Degrees of freedom = (# intervals) - 1
= 9

2c. What is the value of your chi square statistic?
156.5

2d. What is the p-value?
<<< .0001

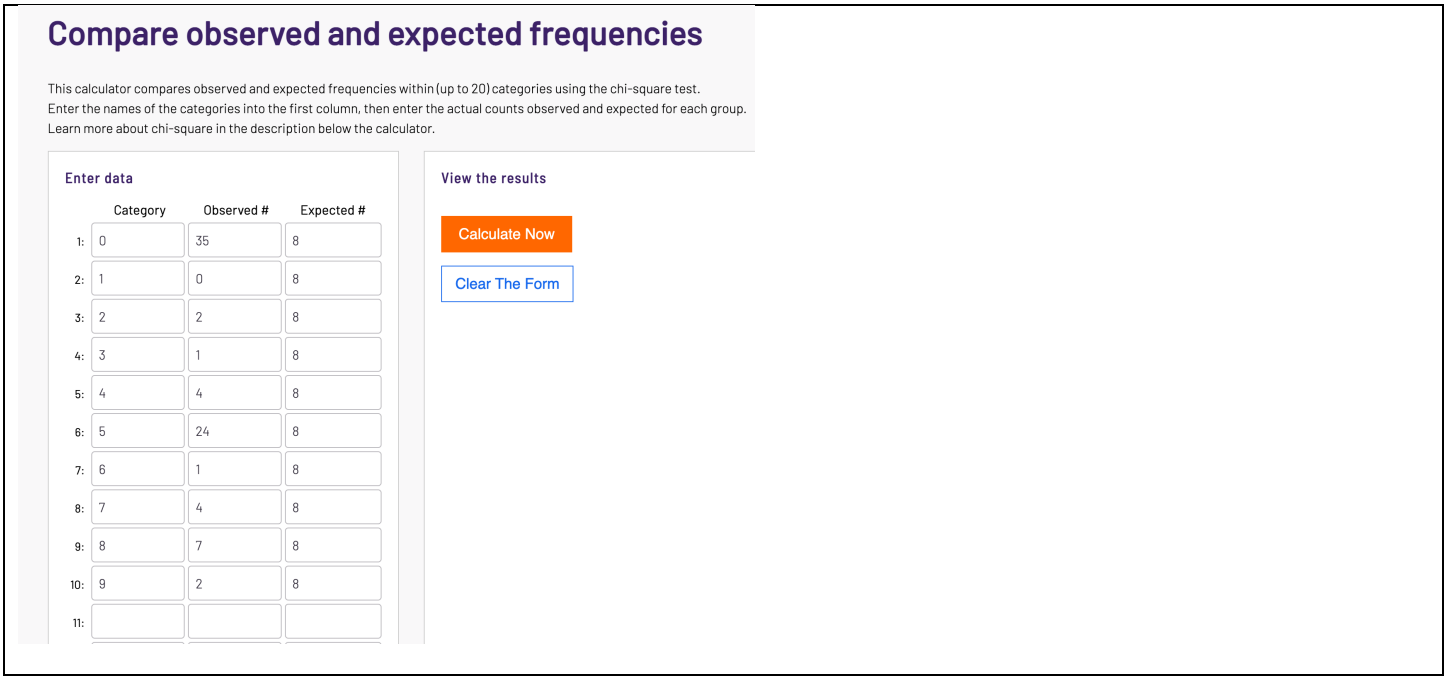
2e. In 1-2 sentences, what do you conclude?

Reject the null hypothesis. Assumption of the null hypothesis and application to the data has led to an extremely unlikely result. Conclude these data provide statistically significant evidence of non-equally likely last digits.

Online Apps Solution

GraphPad Chi Square Calculator
Link: https://www.graphpad.com/quickcalcs/chisquared1.Chi-square/

Under the null hypothesis of equality of proportions (10%) for each digit, the null hypothesis expected count for each digit is thus (n=80) x (10%) = 8. In GraphPad, enter observed counts and null hypothesis expected counts of 8. Click CALCULATE NOW.



Compare observed and expected frequencies

This calculator compares observed and expected frequencies within (up to 20) categories using the chi-square test.
Enter the names of the categories into the first column, then enter the actual counts observed and expected for each group.
Learn more about chi-square in the description below the calculator.

Enter data

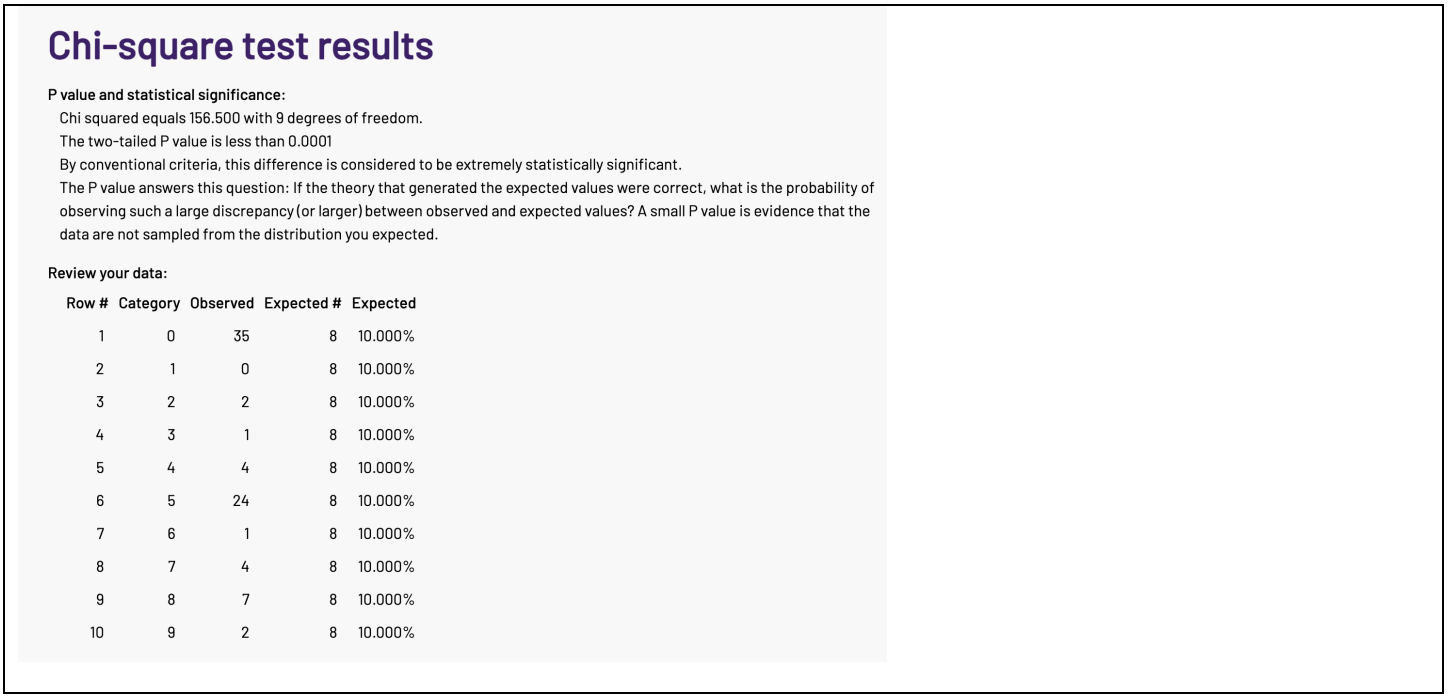
	Category	Observed #	Expected #
1:	0	35	8
2:	1	0	8
3:	2	2	8
4:	3	1	8
5:	4	4	8
6:	5	24	8
7:	6	1	8
8:	7	4	8
9:	8	7	8
10:	9	2	8
11:			

View the results

Calculate Now

Clear The Form

You should then see:



Chi-square test results

P value and statistical significance:

Chi squared equals 156.500 with 9 degrees of freedom.
The two-tailed P value is less than 0.0001
By conventional criteria, this difference is considered to be extremely statistically significant.
The P value answers this question: If the theory that generated the expected values were correct, what is the probability of observing such a large discrepancy (or larger) between observed and expected values? A small P value is evidence that the data are not sampled from the distribution you expected.

Review your data:

Row #	Category	Observed	Expected #	Expected
1	0	35	8	10.000%
2	1	0	8	10.000%
3	2	2	8	10.000%
4	3	1	8	10.000%
5	4	4	8	10.000%
6	5	24	8	10.000%
7	6	1	8	10.000%
8	7	4	8	10.000%
9	8	7	8	10.000%
10	9	2	8	10.000%

R Solution

Q2 - Create table

```
# Observed Counts, in order
digit <- c(0,1,2,3,4,5,6,7,8,9)
observed_n <- c(35,0,2,1,4,24,1,4,7,2)
expected_percent <- rep(.10,times=10)                          # rep( ) is the replicate function

# Inspect
q2_df <- data.frame(digit,observed_n,expected_percent)
cat("\nQuestion 2\nLast Digit of Self-Reported Weight (n=80)\n\n")
##
## Question 2
## Last Digit of Self-Reported Weight (n=80)

q2_df
##    digit observed_n expected_percent
## 1      0         35              0.1
## 2      1          0              0.1
## 3      2          2              0.1
## 4      3          1              0.1
## 5      4          4              0.1
## 6      5         24              0.1
## 7      6          1              0.1
## 8      7          4              0.1
## 9      8          7              0.1
## 10     9          2              0.1
```

Q2 – Chi Square GOF Test, minimal

```
chisquare_GOFTest <- chisq.test(observed_n, p = expected_percent)
cat("\nQuestion 2\nChi Square GOF Test\nNull: Equal Frequencies (10%) of Each Last Digit\n\n")
##
## Question 2
## Chi Square GOF Test
## Null: Equal Frequencies (10%) of Each Last Digit

chisquare_GOFTest
##
##  Chi-squared test for given probabilities
##
## data:  observed_n
## X-squared = 156.5, df = 9, p-value < 0.00000000000000022
```

Q2 - Chi Square GOF Test, "by hand" programmed directly

```
digit <- c(0,1,2,3,4,5,6,7,8,9)                                # Last digit
obs <- c(35,0,2,1,4,24,1,4,7,2)                                # observed counts
q2_full <- data.frame(digit,obs)                               # combine to make data frame
n_size <- sum(obs)                                             # total sample size
q2_full$exp <-rep(n_size/10,times=10)                          # null hypothesis expected counts

q2_full$ichisq <- ((q2_full$obs-q2_full$exp)^2)/(q2_full$exp)  # Component chi square
chisqvalue <- sum(q2_full$ichisq)                              # Chi Square (GOF)
dfvalue <- length(q2_full$ichisq) - 1                          # Degrees of Freedom
pvalue <- pchisq(chisqvalue,df=dfvalue,lower.tail=FALSE)       # p-value
pvalue <- round(pvalue,8)
```

```
# Output Stuff
cat("\nChi Square Goodness of Fit (Null: All Digits Equally Likely)\n\n")
##
## Chi Square Goodness of Fit (Null: All Digits Equally Likely)

q2_full[,c("digit","obs","exp","ichisq")]    # [  , c("var","var", "etc")] for ALL rows, selected vars
##    digit obs exp ichisq
## 1      0  35   8 91.125
## 2      1   0   8  8.000
## 3      2   2   8  4.500
## 4      3   1   8  6.125
## 5      4   4   8  2.000
## 6      5  24   8 32.000
## 7      6   1   8  6.125
## 8      7   4   8  2.000
## 9      8   7   8  0.125
## 10     9   2   8  4.500

cat("\nChi Square statistic =  ", chisqvalue)
## Chi Square statistic =   156.5

cat("\nDegrees of freedom = ", dfvalue)
## Degrees of freedom =  9

cat("\np-value =", pvalue)
## p-value = 0
```

Supplement - Learn R (nothing to turn in)

Practice with missing values, creating a 0/1variable and using the package {tidyverse}

Introduction to The Heart and Estrogen/Progestin Replacement Study (HERS)

Source:
*Hulley S, Grady D, Bush T, Furberg C, Herrington D, Riggs B and Vittinghoff E (1998). Randomized trial of estrogren plus progestin for secondary prevention of heart diessase in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. Journal of the American Medical Association, **280**(7), 605-613.*

In the HERS study, Hulley et al. (1998) sought to determine if exercise, a modifiable behavior, might lower the risk of diabetes in non-diabetic women who were at risk of developing the disease. The question is a complex one because there are many risk factors for diabetes. Moreover, the type of woman who chooses to exercise may be related in other ways to risk of diabetes, apart from the fact of her exercise habit. For example, women who exercise regularly are typically younger and have lower body mass index (BMI); these characteristics also confer a risk benefit with respect to diabetes. Finally, the benefit of exercise may be mediated through a reduction of body mass index. Vittinghoff, Glidden, Shiboski and McCullogh (2005) consider portions of this data in their 2005 text, Regression Methods in Biostatistics: Linear.Logistic, Survival and Repeated Measures Models (Springer). Their dataset has n=2,763 observations on 37 variables.

Here, we consider just 5 variables.

Data Dictionary - *Partial Listing*

Position	Variable	Variable Label	Type	Codes	# missing (NA)
1	HT	Hormone Therapy	character	"hormone therapy" "placebo"	None
2	drinkany	Current drinker	character	"no" "yes"	2
3	LDL	LDL Cholesteraol, mg/dl	numeric	Range: [36.8, 365.2]	11
4	SBP	Systolic, mm Hg	numeric	Range: [83.0, 224.0]	0
5	weight	Weight, kg	numeric	Range: [37.5, 132.0]	2

Practice #1 -

Reminder. R will not analyze a character variable. If you want to treat it as a "categorical variable", you must create a factor version.

Load hersdata.Rdata. Use **class( )** to check the datatype of the variable drinkany. Use **factor( )** to convert drinkany to factor type. Check.

```
load(file="hersdata.Rdata")

class(hersdata$drinkany)
## [1] "character"

hersdata$drinkany <- factor(hersdata$drinkany)

class(hersdata$drinkany)
## [1] "factor"
```

Practice #2 -

Use **library( )** to attach the package **{summarytools}**. Use the function **freq( )** in attached **{summarytools}** to produce a one way frequency table of drinkany. How many missing values are there?

```
library(summarytools)
freq(hersdata$drinkany)

## Frequencies
## hersdata$drinkany
## Type: Factor
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##          no   1680    60.848         60.848    60.803         60.803
##         yes   1081    39.152        100.000    39.124         99.928
##        <NA>      2                               0.072        100.000
##       Total   2763   100.000        100.000   100.000        100.000
```

There are 2 missing values.

Practice #3 -

Important. Whenever you create a new variable, always handle missing values explicitly (you'd be surprised what can happen when you don't!) Create a 0/1 numeric variable **drinkany01** that is coded as below; note that it handles missing values explicitly. Check:

drinkany01 = 0 if **drinkany == "no"**
1 if **drinkany == "yes"** and
NA if **drinkany == NA**

```
hersdata$drinkany01 <- NA                                    # Initialize to missing
hersdata$drinkany01[hersdata$drinkany=="no"] <- 0            # when drinkany=="no" assign 0 to drinkany01
hersdata$drinkany01[hersdata$drinkany=="yes"] <- 1           # when drinkany=="yes" assign 1 to drinkany01

table(hersdata$drinkany,hersdata$drinkany01,useNA="always")      # useNA choices "no", "ifany", "always"
##
##            0    1 <NA>
##   no    1680    0    0
##   yes      0 1081    0
##   <NA>     0    0    2
```

Practice #4 -

The functions **filter( )** and **select( )** are in the package **{dplyr}** which is a core package in **{tidyverse}. {dplyr}** is attached automatically when you attach **{tidyverse}**. **filter( )** and **select( )** make subsetting data very easy!

Create a subset of **hersdata** called **mytiny** as follows:
Include only the following variables: **HT, LDL,** and **SBP**
Include only the observations with: **weight > 125**

Show

```
library(tidyverse)

mytiny <- hersdata %>%                    # use hersdata.  THEN
          filter(weight > 125) %>%        # filter( ) to choose observations to keep.  AND THEN
          select(HT,LDL,SBP)              # select( ) to choose variables to keep

mytiny

##                HT   LDL SBP
## 1 hormone therapy 122.2 129
## 2         placebo 204.6 133
## 3         placebo 161.2 112
## 4         placebo 137.0 130
## 5         placebo 148.4 139
```

Save.

```
save(mytiny,file="mytiny.Rdata")
```