BIOSTATS 540 – Introductory Biostatistics
Fall 2022
Introduction to R
07 – 2 x 2 Table

Packages Used in this Illustration:
`{epiR}, {summarytools}, {ggplot2}`

#1. 2 x 2 Table – Data Wrangling

Background
This was a study of the performance of the dipstick test (clinical screen) for the detection of a UTI in a sample of 229 patients. True status of UTI present is defined by a positive culture test.

Source:
Woodward M. Epidemiology Study Design and Data Analysis, 2nd edition
Chapman & Hall/CRC, 2005. Example 2.15 on page 105

		Culture	
Dipstick Screen Result		Positive	Negative
	Positive	**a** = 84	**b** = 43
	Negative	**c** = 10	**d** = 92

1. Data Wrangling - create table with direct entry of counts

```
table1 <- as.table(rbind(c(84,43),c(10,92)))
dimnames(table1)<- list(
     Dipstick=c("Dipstick positive","Dipstick negative"),
     Culture=c("True positive","True negative"))
table1
```

```
##                   Culture
## Dipstick           True positive True negative
##   Dipstick positive            84            43
##   Dipstick negative            10            92
```

```
table2 <- as.table(rbind(c(27,95),c(44,443)))                    # we'll use this in Section 4 below
dimnames(table2)<- list(
     Cholesterol=c("High","Normal"),
     CHD=c("Yes","No"))
table2
```

```
##            CHD
## Cholesterol Yes  No
##      High    27  95
##      Normal  44 443
```

1. Data Wrangling - create dataframe of individual observations from a table

```
library(DescTools)
library(summarytools)

df.table1 <- Untable(table1)

ctable(df.table1$Culture,df.table1$Dipstick,
       prop='n',
       totals=FALSE)

## Cross-Tabulation
## Culture * Dipstick
## Data Frame: df.table1
##
## --------------- ---------- ------------------- -------------------
##                   Dipstick   Dipstick positive   Dipstick negative
##         Culture
##   True positive                             84                  10
##   True negative                             43                  92
## --------------- ---------- ------------------- -------------------

df.table2 <- Untable(table2)                                  # we'll use in Section 4 below
```

#2. 2 x 2 Table – Numerical Summaries

2. Numerical Summaries - basic**

```
library(gmodels)
library(summarytools)

with(df.table1,                                                          # using {gmodels}
     CrossTable(Culture,Dipstick,digits=2,
             prop.r=TRUE,
             prop.c=FALSE,
             prop.t=FALSE,
             prop.chisq=FALSE))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  229
##
##
##              | Dipstick
##      Culture | Dipstick positive | Dipstick negative |         Row Total |
## -------------|-------------------|-------------------|-------------------|
## True positive |                84 |                10 |                94 |
##               |              0.89 |              0.11 |              0.41 |
## -------------|-------------------|-------------------|-------------------|
## True negative |                43 |                92 |               135 |
##               |              0.32 |              0.68 |              0.59 |
## -------------|-------------------|-------------------|-------------------|
##  Column Total |               127 |               102 |               229 |
## -------------|-------------------|-------------------|-------------------|
##
##
```

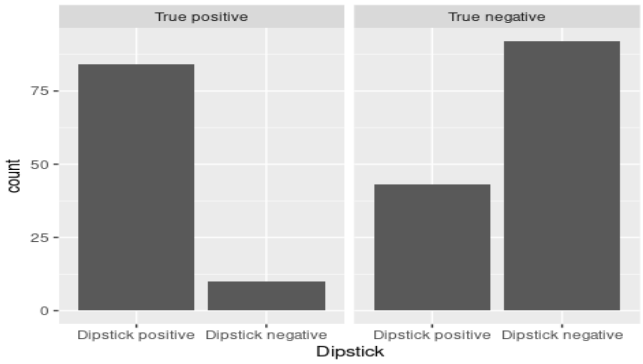
```
with(df.table1,                                                          # using {summarytools}
     ctable(x = Culture, y=Dipstick,
           prop = "r"))
```

```
## Cross-Tabulation, Row Proportions
## Culture * Dipstick
## Data Frame: df.table1
##
## --------------- ---------- ------------------- ------------------- --------------
##                   Dipstick   Dipstick positive   Dipstick negative          Total
##         Culture
##    True positive                    84 (89.4%)          10 (10.6%)    94 (100.0%)
##    True negative                    43 (31.9%)          92 (68.1%)   135 (100.0%)
##            Total                   127 (55.5%)         102 (44.5%)   229 (100.0%)
## --------------- ---------- ------------------- ------------------- --------------
```

#3. 2 x 2 Table – Visualization

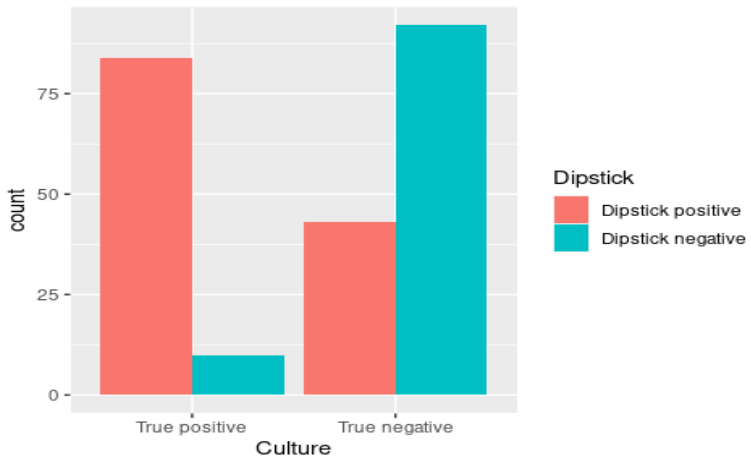
3. Data Visualization - basic facetted.

```
library(ggplot2)

ggplot(data=df.table1) +
   aes(x=Dipstick) +                                    # x= OUTCOME
   geom_bar(na.rm=T) +
   facet_wrap(~Culture)                                 # facet_wrap(~PREDICTOR)
```



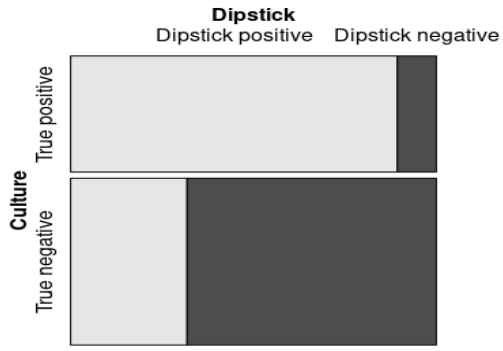
3. Data Visualization - basic grouped bar chart

```
library(ggplot2)


ggplot(data=df.table1) +
  aes(x = Culture) +                          # x = predictor var, must be factor
  aes(fill = Dipstick) +                      # y = outcome var, must be factor
  geom_bar(position = "dodge")                # choose either "dodge" or "stacked"
```

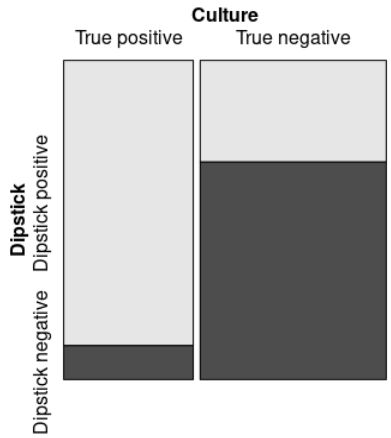


*3. **Data Visualization** - basic mosaic

```
library(vcd)

# Row percents total 100%
mosaic(data=df.table1,
       Dipstick ~ Culture)                                          # Outcome ~ Predictor
```

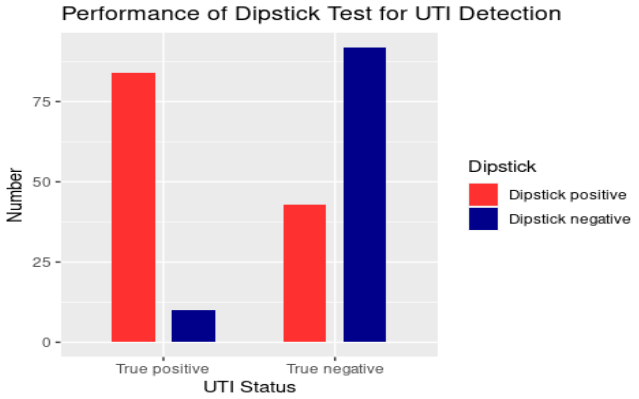



```
# Column percents total 100%
mosaic(data=df.table1,
       Dipstick ~ Culture,                                          # Outcome ~ Predictor
       direction=c("v","h"))
```



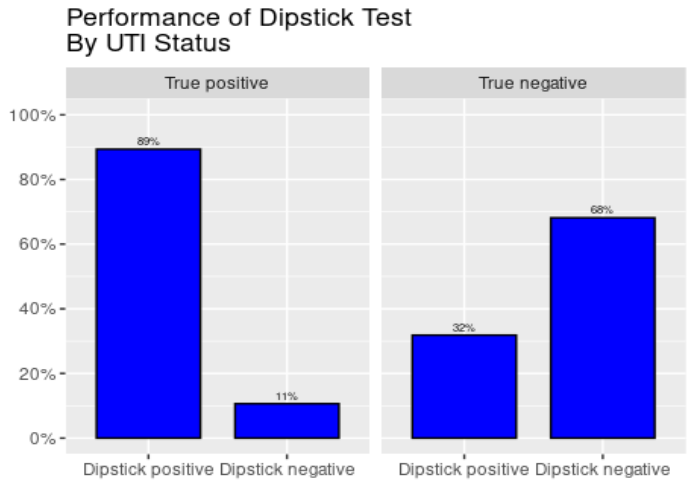
3. Data Visualization - grouped bar chart with aesthetics

```
library(ggplot2)

# choose your own colors
mycolors <- c("firebrick1","darkblue")               # c(dipstick pos, dipstick neg)

ggplot(data=df.table1) +
  aes(x=Culture) +                                        # x = predictor
  aes(y=..count..,fill=Dipstick) +                        # y = outcome

  geom_bar(width=0.5,                                     # bar width (default=0.9)
           position=position_dodge(0.7))   +              # space between (default=0.9)
  scale_fill_manual(values=mycolors) +

  ggtitle("Performance of Dipstick Test for UTI Detection") +
  xlab("UTI Status") +
  ylab("Number")
```



3. Data Visualization - grouped bar chart with display of percentages (and aesthetics!)

```
library(ggplot2)

ggplot(data=df.table1) +
   aes(x= Dipstick,                                    # outcome var
       group=Culture) +                                 # predictor var

   geom_bar(aes(y = ..prop..,                          # compute percentagles
                fill = factor(..x..)),
                stat="count",
                width=0.75,                           # default width = 0.9
                position=position_dodge(0.7),         # default is no space between bars
                color="black",
                fill="blue") +

  scale_y_continuous(labels = scales::percent,              # Plot as percents
                     limits=c(0,1),                             # Set range of Y-axis
                     breaks=c(0, .20, .40, .60, .80, 1)) +      # Set tick marks on Y axis

  geom_text(aes( label = scales::percent(..prop..),     # display percentagev values
                 y= ..prop.. ),
                 stat= "count",
                 vjust = -0.5,                        # negative to vjust UPWARD
                 size=2) +                            # size of font

   facet_wrap(~Culture) +                             # separate bar graphs by predictor


   ggtitle("Performance of Dipstick Test\nBy UTI Status") +
   xlab(" ") +
   ylab(" ")
```



#4. Estimation and Hypothesis Test of No Association

4. Estimation - diagnostic testing

```
library(epiR)
epi.tests(table1,conf.level=0.95)
```

```
##           Outcome +    Outcome -      Total
## Test +           84           43        127
## Test -           10           92        102
## Total            94          135        229
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------
## Apparent prevalence *                  0.55 (0.49, 0.62)
## True prevalence *                      0.41 (0.35, 0.48)
## Sensitivity *                          0.89 (0.81, 0.95)
## Specificity *                          0.68 (0.60, 0.76)
## Positive predictive value *            0.66 (0.57, 0.74)
## Negative predictive value *            0.90 (0.83, 0.95)
## Positive likelihood ratio              2.81 (2.17, 3.63)
## Negative likelihood ratio              0.16 (0.09, 0.28)
## False T+ proportion for true D- *      0.32 (0.24, 0.40)
## False T- proportion for true D+ *      0.11 (0.05, 0.19)
## False T+ proportion for T+ *           0.34 (0.26, 0.43)
## False T- proportion for T- *           0.10 (0.05, 0.17)
## Correctly classified proportion *      0.77 (0.71, 0.82)
## -------------------------------------------------------------
## * Exact CIs
```

4. Estimation and Hypothesis Testing - basic relative risk and odds ratio

```
library(DescTools)

OddsRatio(table2,conf.level=.95)
```

```
## odds ratio     lwr.ci     upr.ci
##   2.861483   1.687805   4.851321
```

```
RelRisk(table2, conf.level=.95)
```

```
## rel. risk    lwr.ci    upr.ci
##  2.449516  1.577514  3.749927
```

4. Estimation and Hypothesis Testing - more detail relative risk and odds ratio

```
library(summarytools)

with(df.table2,                                                    # Cohort Study - relative risk
     ctable(x = Cholesterol, y = CHD,
     prop = "r",
     RR=TRUE))

## Cross-Tabulation, Row Proportions
## Cholesterol * CHD
## Data Frame: df.table2
##
##
## ------------- ----- ------------ ------------- --------------
##                 CHD          Yes            No          Total
##   Cholesterol
##          High         27 (22.1%)    95 (77.9%)   122 (100.0%)
##        Normal         44 ( 9.0%)   443 (91.0%)   487 (100.0%)
##         Total         71 (11.7%)   538 (88.3%)   609 (100.0%)
## ------------- ----- ------------ ------------- --------------
##
## ----------------------------------
##  Risk Ratio   Lo - 95%   Hi - 95%
## ------------ ---------- ----------
##     2.45        1.58       3.79
## ----------------------------------

with(df.table2,                                                      # Case Control Study - odds ratio
     ctable(x = Cholesterol, y = CHD,
     prop = "c",
     OR=TRUE))

## Cross-Tabulation, Column Proportions
## Cholesterol * CHD
## Data Frame: df.table2
##
##
## ------------- ----- ------------- -------------- --------------
##                 CHD           Yes             No          Total
##   Cholesterol
##          High         27 ( 38.0%)    95 ( 17.7%)   122 ( 20.0%)
##        Normal         44 ( 62.0%)   443 ( 82.3%)   487 ( 80.0%)
##         Total         71 (100.0%)   538 (100.0%)   609 (100.0%)
## ------------- ----- ------------- -------------- --------------
##
## ----------------------------------
##  Odds Ratio   Lo - 95%   Hi - 95%
## ------------ ---------- ----------
##     2.86        1.69       4.85
## ----------------------------------
```

4. Estimation and Hypothesis Testing - test of Null: No association

```
# 2x2 Table
chisq.test(table2,correct=FALSE)                                   # without continuity correction

##
##  Pearson's Chi-squared test
##
## data:  table2
## X-squared = 16.246, df = 1, p-value = 0.00005561
```

```
chisq.test(table2)                                                    # with continuity correction (default)

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table2
## X-squared = 15, df = 1, p-value = 0.0001075



# Data frame of counts
chisq.test(df.table2$Cholesterol,df.table2$CHD, correct=FALSE)          # without continuity correction

##
##  Pearson's Chi-squared test
##
## data:  df.table2$Cholesterol and df.table2$CHD
## X-squared = 16.246, df = 1, p-value = 0.00005561

chisq.test(df.table2$Cholesterol,df.table2$CHD)                       # with continuity correction

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df.table2$Cholesterol and df.table2$CHD
## X-squared = 15, df = 1, p-value = 0.0001075



# Fisher exact test
fisher.test(table2,alternative="two.sided")                           # "two.sided", "greater", "less"

##
##  Fisher's Exact Test for Count Data
##
## data:  table2
## p-value = 0.0002049
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.614783 4.985252
## sample estimates:
## odds ratio
##    2.85516

fisher.test(table2)$estimate                                  #  good to know:  get the odds ratio

## odds ratio
##    2.85516
```