BIOSTATS 540 – Introductory Biostatistics
Fall 2022
Introduction to R
06 – One and Two Sample Inference

Dataset Used in this Illustration (right click to download):
`sepsis.Rdata`

Packages Used in this Illustration:
`{DescTools}, {stargazer}, {summarytools} {tidyverse}`

Tip for Hypothesis Testing

Alternative Hypothesis	R Code
`Two sided`	`, alternative="two.sided"`
`Right tail`	`, alternative="greater"`
`Left tail`	`, alternative="less"`

Tip for Confidence Intervals

If you want ...	R Code
`95% CI`	`Nothing you need to do ... this is default`
`90% CI`	`, conf.level = .90`
`... and so on`	`, conf.level = .FILLIN`

#1. One Sample – Continuous Outcome Normal Distribution Model

At a Glance

Numerical Summarization	`summary(outcome)` # Method 1
	
	`library(summarytools)` # Method 2
	`descr(df$outcome,`
	`  stats=c("n.valid", "mean", "sd", "med", "min", "max"),` # User chooses
	`  transpose=TRUE)`

Confidence Interval Estimation	`# Confidence Interval for mean`
	`t.test(outcome ~ 1, data=df, conf.level=.90)$conf.int` # Default is conf.level=.95
	
	`# Confidence Interval for variance`
	`library(DescTools)`
	`VarTest(df$outcome, conf.level=.90)$conf.int` # Default is conf.level=.95

Hypothesis Testing	`# One Sample t-test of mean`
	`t.test(outcome ~ 1, data=df, mu=nullmean)`
	
	`# One Sample t-test of variance`
	`library(DescTools)`
	`VarTest(df$outcome, sigma.squared=nullvariance)`

Examples.

```
# Z Test of mean:  Population variance/standard deviation are KNOWN

library(DescTools)
ZTest(sepsis$o2del,
        mu=1000,                                    # null hypothesis mean
        sd_pop=409,                                 # known population standard deviation sigma
        alternative="greater")                      # alternative: true mean > null mean
```

```
         One Sample z-test

data:  sepsis$o2del
z = 0.75478, Std. Dev. Population = 409, p-value = 0.2252        # Null mu=1000 v mu > 1000 is NOT rejected
alternative hypothesis: true mean is greater than 1000
95 percent confidence interval:
 971.9137      Inf
sample estimates:
mean of x
 1023.817
```

```
# T-test of mean:  Population variance/standard deviation NOT known

t.test(o2del~1,                                     # model formulation
       data=sepsis,                                 # data to use
       mu=1200,                                     # null hypothesis mean
       alternative="two.sided",                     # alterantive: true mean ≠ null mean
       conf.level=.90,                              # show 90% CI
       na.rm=TRUE)                                  # omit NA's (missing values)
```

```
         One Sample t-test

data:  o2del
t = -5.5773, df = 167, p-value = 0.00000009658              # 2 sided p << .0001.  Reject null (mu=1200)
alternative hypothesis: true mean is not equal to 1200
90 percent confidence interval:
  971.5676 1076.0665                                        # 90% CI does NOT contain null mu=1200
sample estimates:
mean of x
 1023.817
```

```
# Test of Variance
library(DescTools)                                  # You could use var.test( ) in {base}. I like this

VarTest(sepsis$o2del,
      sigma.squared=1600)                           # Null hypothesis variance (not SD!)
```

```
          One Sample Chi-Square test on variance

data:  sepsis$o2del
X-squared = 17498, df = 167, p-value < 0.00000000000000022      # 2 sided p << .0001 Reject null (sigma2 = 1600)
alternative hypothesis: true variance is not equal to 1600
95 percent confidence interval:
 136784.9 210324.6                                              # 95% CI does NOT contain null sigma2 = 1600
sample estimates:
variance of x
     167643.2
```

#2. One Sample – Discrete Outcome
Binomial Distribution Model

At a Glance

Numerical Summarization	`summary(outcome)` # Method 1 `library(summarytools)` # Method 2 `freq(df$outcome)` # Outcome must be factor

Confidence Interval Estimation	`# Confidence Interval for proportion – EXACT`
	`binom.test(x=#events,n=ntrials,conf.level=.90)$conf.int` # Default is conf.level=.95
	
	`# Confidence Interval for proportion – NORMAL APPROXIMATION`
	`prop.test(x=#events,n=ntrials,conf.level=.90)$conf.int` # Default is conf.level=.95

Hypothesis Testing	`# Hypothesis Test for Binomial Proportion – EXACT`
	`binom.test(x=#events,n=ntrials,p=nullp,`
	`alternative="less")` # "two.sided", "greater", "less"
	
	`# Hypothesis Test for Binomial Proportion – NORMAL APPROXIMATION`
	`prop.test(x=#events,n=ntrials,p=nullp,`
	`alternative="less")` # "two.sided", "greater", "less"

Examples.

```
# Binomial Proportion:  Exact Inference

library(tidyverse)


# For small to moderate sample size – For illustration I will obtain a small sample size = 25
temp <- sepsis %>%
       sample_n(25, na.rm=TRUE)

xevents <- sum(temp$treat, na.rm=TRUE)            # sum of 0/1 events gives x = xevents = # successes
ntrials <- sum(!is.na(temp$treat))                # sum of !is.na gives n = ntrials = # trials


binom.test(x=xevents,n=ntrials,p=.5)                          # Hypothesis Test (Null: p = .50)
```

```
         Exact binomial test

data:  xevents and ntrials
number of successes = 14, number of trials = 25, p-value = 0.69
alternative hypothesis: true probability of success is not equal to 0.5     # p=.69 do NOT reject null proportion =.50
95 percent confidence interval:
 0.3492816 0.7559763
sample estimates:
probability of success
                  0.56
```

```
# Binomial Proportion:  Normal Approximation

library(tidyverse)

xevents <- sum(sepsis$treat, na.rm=TRUE)            # sum of 0/1 events gives x = xevents = # successes
ntrials <- sum(!is.na(sepsis$treat))                # sum of !is.na gives n = ntrials = # trials


prop.test(x=xevents,n=ntrials,p=.5, correct=FALSE)    # Hypothesis Test (Null: p = .50)
```

```
         1-sample proportions test without continuity correction

data:  xevents out of ntrials, null probability 0.5
X-squared = 0.10769, df = 1, p-value = 0.7428                # pvalue = .74 do NOT reject null proportion =.50
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4466279 0.5381163                                         # 95% CI contains the null proportion = .50
sample estimates:
       p
0.4923077
```

#3. One Sample Paired – Continuous Outcome
Normal Distribution Model

Preliminary – Is your paired data wide or long?

Wide Data

For each studyid, the pre and post data are in the SAME row (horizontal)
e.g., pre = sbp1 and post = sbp2

```
  studyid sbp1 sbp2
1       1  120  115
2       2  140  138
```

Long Data

For each studyid, the pre and post data are each in their OWN/SEPARATE rows (vertical)
In long data, you have a variable that tells you occasion (pre v post)
and another variable that is the outcome

```
  studyid    visit sbp
1       1      pre 120
2       1     post 115
3       2      pre 140
4       2     post 138
```

At a Glance

Numerical Summarization	`* WIDE: Paired variables (e.g., pre and post) in WIDE format`
	`myvars <- c("prevar", "postvar")`
	`descr(df[myvars],`
	`stats=c("n.valid", "mean", "sd", "med", "min", "max"), # User chooses`
	`transpose=TRUE)`
	
	`* LONG: Paired variables (e.g., pre and post) are in LONG FORMAT`
	`library(summarytools)`
	`with(df,`
	`stby(data = outcomevar,`
	`INDICES = timevar, # timevar must be factor`
	`FUN = descr, stats = c("mean", "sd", "min", "med","max"), # User chooses`
	`transpose=TRUE))`

Confidence Interval Estimation	`# Confidence Interval for mean`
	`t.test(outcome ~ 1, data=df, conf.level=.90)$conf.int # Tip. Outcome = post - pre`
	
	`# Confidence Interval for variance`
	`library(DescTools)`
	`VarTest(df$outcome, conf.level=.90)$conf.int # Default is conf.level=.95`

Hypothesis Testing	`# One Sample t-test of mean`
	`t.test(outcome ~ 1, data=df, mu=nullmean)`
	
	`# One Sample t-test of variance`
	`library(DescTools)`
	`VarTest(df$outcome, sigma.squared=nullvariance)`

Examples.

```
# Paired Data Student t-Test:  WIDE

t.test(sepsis$temp0,sepsis$temp7, paired=TRUE,                  # data in WIDE
       var.equal=FALSE,
       na.rm=TRUE)
```

```
          Paired t-test

data:  sepsis$temp0 and sepsis$temp7
t = 13.144, df = 412, p-value < 0.00000000000000022          # p << .0001.  Null of equality pre/post is rejected
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.093632 1.478282                                           # 95% CI does NOT contain Null difference of 0
sample estimates:
mean of the differences
               1.285957
```

```
# Paired Data Student t-Test: LONG

library(tidyverse)

# paired t LONG requires sorted by id then by occasion nested in id
longdf <- longdf %>%
          arrange(id, hour)

# Now do paired t - LONG
t.test(temp ~ hour, data=longdf, paired=TRUE)
```

```
          Paired t-test

data:  temp by hour
t = 13.144, df = 412, p-value < 0.00000000000000022          # p << .0001.  Null of equality pre/post is rejected
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.093632 1.478282                                           # 95% CI does NOT contain Null difference of 0
sample estimates:
mean of the differences
               1.285957
```

#4. Two Independent Samples – Continuous Outcome
Normal Distribution Model

At a Glance

Numerical Summarization

```
* LONG:  data are in LONG format
by(df[ , c("outcomevar")],                                  # summarize only outcomevar
   df$groupvar,                                             # grouping variable
   summary)                                                 # use function summary in {base}

library(summarytools)
with(df,
   stby(data = outcomevar,
        INDICES = groupvar,                                          # groupvar must be factor
        FUN = descr, stats = c("mean", "sd", "min", "med","max"),    # User chooses
        transpose=TRUE))
```

Confidence Interval Estimation

```
* LONG:  data are in LONG format
# Confidence Interval for mean difference (group1 - group2)
t.test(outcome ~ groupvar, data=df, conf.level=.90)$conf.int
```

Hypothesis Testing

```
# Two Sample Test of Equality of Variances
var.test(outcome ~ groupvar, data=df,
      alternative = "two.sided")                            # "two.sided", "greater", "less"

# Two Sample Test of Equality of Means - UNEQUAL variances
t.test(outcome ~ groupvar, data=df,
    alternative="two.sided")                                # "two.sided", "greater", "less"

# Two Sample Test of Equality of Means - EQUAL variances
t.test(outcome ~ groupvar, data=df, var.equal=TRUE,
    alternative="two.sided")                                # "two.sided", "greater", "less"
```

Examples.

```
# Test of Equality of Variances

# REQUIRED:  group variable must be factor
sepsis$fatef <- factor(sepsis$fate,
                    levels=c(0,1),
                    labels=c("Alive", "Dead"))

var.test(o2del ~ fatef, data=sepsis)                    # Preliminary: test of vars

# Test of Equality of Means
t.test(o2del ~ fatef, data=sepsis,                      # t-test assuming equal var (provides CI, too)
        var.equal=TRUE)
```

```
        F test to compare two variances

data:  o2del by fatef
F = 0.91965, num df = 100, denom df = 66, p-value = 0.6975          #Okay to assume equal variances
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5846373 1.4175632
sample estimates:
ratio of variances
         0.9196542
```

```
        Two Sample t-test

data:  o2del by fatef
t = 2.5796, df = 166, p-value = 0.01076                              #Reject Null:  Equal means
alternative hypothesis: true difference in means between group Alive and group Dead is not equal to 0
95 percent confidence interval:
  38.40254 288.94124
sample estimates:
mean in group Alive  mean in group Dead
          1089.0910            925.4191
```

#5. Two Independent Samples – Discrete Outcome
Binomial Distribution Model

At a Glance

Numerical Summarization

```
table(df$discrete1,df$discrete2, useNA="always")                    # Method 1

library(summarytools)                                               # Method 2
with(df,
     ctable(rowvar, colvar,                                         # vars must be factor
     prop="n"),                                                     # User chooses "n", "r", "c"
     totals=TRUE)                                                   # use this if you want totals
```

Hypothesis Testing

```
# Fisher Exact Test of Equality of Proportions (NULL:  Odds Ratio = 1)
fisher.test(df$rowvar,df$colvar)

# Chi Square Test of Equality of Proportions – WITH continuity correction (default)
chisq.test(df$rowvar,df$colvar)

# Chi Square Test of Equality of Proportions – WITHOUT continuity correction
chisq.test(df$rowvar,df$colvar, correct=FALSE)
```

Example.

```
mytable <- table(sepsis$treat,sepsis$fate)                  # Use table( ) to create table
dimnames(mytable) <- list(
    Treatment=c("Untreated","Treated"),
    Fate=c("Alive","Dead"))

mytable
chisq.test(mytable,correct=FALSE)                           # large n, no correction needed
```

```
           Fate
 Treatment  Alive Dead
  Untreated   139   92
  Treated     140   84

          Pearson's Chi-squared test

data:  mytable
X-squared = 0.25959, df = 1, p-value = 0.6104                  # p-value = .61 Do NOT reject independence
```