Problem Set6 Key

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(summarytools)
```

```
Attaching package: 'summarytools'

The following object is masked from 'package:tibble':

    view
```

1 Problem 1

The weights of 15 randomly selected girls at birth from a single hospital in New York recorded an average of 3.09 kg and standard deviation of 0.29 kg. Suppose the hospital claimed that the average weight of girls at birth is higher than 3 kg.

a. List the parameter of interest and null and alternative hypotheses for this problem. [1 pt.]

The parameter of interest is the population/long-run average weight of girls at birth. The hypotheses are:

$$H_0 : \mu = 3; H_a : \mu > 3$$

b. Calculate the test statistic. [1pt.]

```
xbar <- 3.09
mu_null <- 3
sdsamp <- 0.29
sampsize <- 15

t<- (xbar-mu_null)/(sdsamp/sqrt(sampsize))
t
```

```
[1] 1.20196
```

c. Calculate the p-value. [1 pt.]

```
df <- sampsize - 1
pt(t,df=df,lower.tail=F)
```

```
[1] 0.1246582
```

d. Do we reject or fail to reject the null hypothesis? [1 pt.]

Since significance level is not provided, use 0.05 as the significance level. We fail to reject the null hypothesis.

e. Do we have sufficient evidence to support the hospital's claim? [1pt.]

No, we have insufficient evidence to support the local hospital's claim.

2 Problem 2

The file HINTS subset.csv contains a subset of the Health information National Trends Survey 6 (HINTS 6). The households were asked if they used health or wellness apps on their tablet or smartphone (column UsedHealthWellnessApps2) and the type of community they lived in (column PR_RUCA_2010). Suppose you want to test the claim that more than 60% of the US population have used health or wellness apps on their mobile devices.

```
hints<- read.csv("datasets/HINTS subset.csv")
glimpse(hints)
```

```
Rows: 500
Columns: 4
$ X                       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,~
$ HHID                    <int> 23004900, 23004313, 21011500, 23022289, 210041~
$ UsedHealthWellnessApps2 <chr> "Yes", "Yes", "Yes", "Yes", "No", "No", "No", ~
$ PR_RUCA_2010            <chr> "Metropolitan", "Metropolitan", "Metropolitan"~
```

a. List the parameter of interest and null and alternative hypotheses for this problem. [1 pt.]

The parameter of interest is the proportion of the US population who have used health or wellness apps on their mobile devices. The hypotheses are:

$$H_0 : \pi = 0.6; H_a : \pi > 0.6$$

b. Calculate the test statistic. [1pt.]

```
freq(hints$UsedHealthWellnessApps2)
```

```
Frequencies
hints$UsedHealthWellnessApps2
Type: Character

             Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
---------- ------ --------- -------------- --------- --------------
         No    186     37.20          37.20     37.20          37.20
        Yes    314     62.80         100.00     62.80         100.00
      <NA>      0                                0.00         100.00
      Total    500    100.00         100.00    100.00         100.00
```

```
sampsize <-500
phat <- 314/sampsize
pi_null <- 0.6
se <- sqrt(pi_null*(1-pi_null)/sampsize)

z <- (phat-pi_null)/se
z
```

```
[1] 1.278019
```

c. Calculate the p-value. [1 pt.]

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

```
[1] 0.1006213
```

d. Do we reject or fail to reject the null hypothesis? [1 pt.]

At a 0.05 significance level, we fail to reject the null hypothesis. e. Do we have sufficient evidence to support the claim regarding health and wellness app use? [1pt]

No, we have insufficient evidence to support the claim.

3. The data set in hflashes.csv contains data from a 14-year cohort study by Freeman et al (2001) that investigated the occurrence of hot flashes in 375 participants. Suppose you want to test if there is a difference between the variance of the baseline estradiol measurements (column estra) for current smokers and non-current smokers (column f1a).

```
hflashes <- read.csv("datasets/hflashes.csv")
glimpse(hflashes)
```

```
Rows: 375
Columns: 14
$ pt       <int> 3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 23, 24~
$ ageg     <int> 2, 3, 1, 1, 2, 3, 2, 2, 2, 3, 2, 1, 1, 1, 1, 3, 2, 1, 1, 1, 2~
$ aagrp    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0~
$ edu      <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1~
$ d1       <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0~
$ f1a      <int> 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1~
$ pcs12    <dbl> 56.80537, 59.18338, 57.73952, 55.83575, 55.89324, NA, 55.5009~
$ hotflash <int> 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0~
$ bmi30    <int> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ estra    <dbl> 106.710, 31.250, 13.410, 10.640, 24.060, 37.305, 26.320, 24.1~
$ fsh      <dbl> 3.005, 11.195, 14.545, 5.530, 9.780, 10.290, 7.960, 4.775, 7.~
$ lh       <dbl> 2.980, 5.760, 5.595, 2.260, 2.600, 3.395, 3.570, 2.095, 3.660~
$ testo    <dbl> 7.680, 11.930, 24.375, 8.280, 4.050, 8.275, 15.995, 12.340, 1~
$ dheas    <dbl> 61.225, 104.920, 117.450, 36.850, 11.165, 100.360, 76.780, 83~
```

a. List the parameter of interest and null and alternative hypotheses for this problem. [1 pt.]

The parameter of interest is the long-run difference in the variance of estradiol levels between smokers and non-smokers. The hypotheses are:

$$H_0 : \sigma_{ns}^2/\sigma_s^2 = 1; H_a : \sigma_{ns}^2/\sigma_s^2 \neq 1$$

b. Calculate the test statistic. [1pt.]

```
vartest <- var.test(estra~f1a,data=hflashes)
vartest
```

```
    F test to compare two variances

data:  estra by f1a
F = 1.2964, num df = 236, denom df = 137, p-value = 0.09462
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9558288 1.7372739
sample estimates:
ratio of variances
          1.296368
```

The test statistic value is t=1.2963676.

c. Calculate the p-value. [1 pt.]

The p-value is t=0.0946195.

d. Do we reject or fail to reject the null hypothesis? [1 pt.] We fail to reject the null hypothesis.

e. Comment on the strength/sufficiency of evidence against the null hypothesis. [1pt] There is insufficient evidence to claim that there is a difference in the variance of the estradiol levels between smokers and non-smokers.

4. The data set in hflashes.csv contains data from a 14-year cohort study by Freeman et al (2001) that investigated the occurrence of hot flashes in 375 participants. Suppose you want to test if there is a difference between the long-run average of the baseline estradiol measurements (column estra) for current smokers and non-current smokers.

a. Based on your answer in (3d) and (3e), is it recommended to assume equal variances between the two groups? [1 pt.]

It is recommended to assume equal variances.

b. List the parameter of interest and null and alternative hypotheses for this problem. [1 pt.]

The parameter of interest is the population/long-run difference in estradiol levels between smokers and non-smokers, The hypotheses are:

$$H_0 : \mu_{ns} - \mu_s = 0; H_a : \mu_{ns} - \mu_s \neq 0$$

c. Calculate the p-value. [1 pt.]

```
ttest <- t.test(estra~f1a,data=hflashes,var.equal=T)
ttest
```

```
    Two Sample t-test

data:  estra by f1a
t = -0.22103, df = 373, p-value = 0.8252
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
 -6.644836  5.301924
sample estimates:
mean in group 0 mean in group 1
       42.67546        43.34692
```

d. Do we reject or pt. to reject the null hypothesis? [1 pt.]

We fail to reject the null hypothesis.

e. Comment on the strength/sufficiency of evidence against the null hypothesis. [1pt]

There is insufficient evidence to support the hypothesis that there is a difference in estradiol levels between smokers and non-smokers in this population.

5. A study was performed to investigate the association between acetaminophen use during pregnancy and the children's risk of neurodevelopmental disorders. In a retrospective study that included 2,480,797 children, 185,909 children were exposed to acetaminophen and 2,294,888 were not. 5,912 children exposed to acetaminophen were diagnosed with autism, while 62,672 children were not exposed to acetaminophen. Suppose we were interested in whether there is a difference in proportion of children diagnosed with autism between the different exposures to acetaminophen.

a. List the parameter of interest and null and alternative hypotheses for this problem. [1 pt.]

The parameter of interest is the long-run difference in proportion of children diagnosed with autism between the different exposures to acetaminophen. The hypotheses are:

$$H_0 : \mu_{exposed} - \mu_{notexposed} = 0; H_a : \mu_{exposed} - \mu_{notexposed} \neq 0$$

b. Calculate the p-value. [1 pt.]

```
x1 <- 24
x2 <- 11
sampsize1 <- 44
sampsize2 <- 29
p_pool <- (x1+x2)/(sampsize1+sampsize2)

se <- sqrt(p_pool*(1-p_pool)*(1/sampsize1+1/sampsize2))

pi1 <- x1/sampsize1
pi2 <- x2/sampsize2



z <- (pi1-pi2)/se
z
```

```
[1] 1.39042
```

```
p <- pnorm(z,lower.tail=F)
p
```

```
[1] 0.08220063
```

c. Do we reject or fail to reject the null hypothesis? [1 pt.]
d. Comment on the strength/sufficiency of evidence against the null hypothesis. [1pt]
e. Is our result indicative of a causal relationship between acetaminophen exposure and autism prevalence in children? Why or why not? (Hint: Did we consider ALL variables that could affect both variables considered in this test?) [1pt]