

# Problem Set 5 Key

## 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.

- a. Which distribution should we use for the reliability coefficient of the confidence interval, the standard normal distribution or the t-distribution? [1pt.]

**i** Note

The t-distribution should be used to calculate the reliability coefficient of the confidence interval because the population variance is unknown.

- b. Construct a 95% confidence interval for the mean birth weight of girls in the population represented by this sample, assuming that the population is Gaussian/normally distributed. [2pts.]

The confidence interval can be calculated using  $\bar{x} \pm t_{1-\alpha/2}s/\sqrt{n}$ .

```
xbar <- 3.09
stdev <- 0.29
sampszie <- 15
df <- sampszie-1
alpha <- 1-0.95
relcoeff <- qt(p=1-alpha/2,df=df)

lower <- xbar - relcoeff*stdev/sqrt(sampszie)
lower
```

[1] 2.929403

```
upper <- xbar + relcoeff*stdev/sqrt(sampszie)
upper
```

[1] 3.250597

The 95% confidence interval for the mean birth weight of girls in the population represented by this sample is (2.93 kg, 3.25 kg)

- c. Is 3 kg a plausible value for the population mean? [1pt.]

**i** Note

Because 3 kg can be found in the interval, it is a plausible value for the population mean.

- d. Is this result generalizable to all girls born in the United States? Why or why not? [2 pts.]

**i** Note

No, the sample was from a single hospital in New York city, which might not be representative of all girls born in the entire country

## 2 Problem 2

The Wechsler's Adult Intelligence Scale was designed such that the population standard deviation of all IQ scores was maintained at 15.

- a. A random sample of 30 sophomores from a community college recorded an average Wechsler IQ score of 109.5. Construct a 90% confidence interval for the average Wechsler IQ score for the corresponding population. [2 pts.]

```
xbar <- 109.5
stdev <- 15
sampszie <- 30
alpha <- 1-0.9
relcoeff <- qnorm(p=1-alpha/2,mean=0,sd=1)

lower <- xbar - relcoeff*stdev/sqrt(sampszie)
lower
```

```
[1] 104.9954
```

```
upper <- xbar + relcoeff*stdev/sqrt(sampsize)
upper
```

```
[1] 114.0046
```

The 95% confidence interval for the mean Wechsler IQ score for the sophomores in the community college is (105, 114)

- b. Suppose you want to extend your study to calculate the average IQ of all the students from the community college such that the resulting 95% confidence interval has a margin of error of 2 points. How many students do you need to sample from the community college? [2 pts.]

```
E <- 2
alpha <- 1-0.95
sigma <- 15
z <- qnorm(p=1-alpha/2)

sampsize <- z^2*sigma^2/E^2
sampsize
```

```
[1] 216.0821
```

We need to sample 217 students from the community college to achieve a margin of error of 2 points in our 95% confidence interval for the average IQ of all the students from the community college

### 3 Problem 3

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).

- a. Construct a 95% confidence interval for the long run proportion of individuals who use health or wellness apps on their portable devices based on the whole sample. [2 pts.]

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

```
-- 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    3.2.1
v lubridate 1.9.4     v tidyr    1.3.1
v purrr    1.0.4
-- 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 become non-conflicting
```

```
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"~
```

We need the proportion of those who reported using health or wellness apps on their portable devices.

```
library(summarytools)
```

```
Warning: package 'summarytools' was built under R version 4.5.1
```

```
Attaching package: 'summarytools'
```

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

```
view
```

```
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

alpha <- 1-0.95
relcoeff <- qnorm(p=1-alpha/2)
se <- sqrt((phat*(1-phat))/sampsize)

lower <- phat - relcoeff*se
lower
```

[1] 0.5856343

```
upper <- phat + relcoeff*se
upper
```

[1] 0.6703657

The 95% confidence interval for the long-run proportion of individuals who use health or wellness apps on their portable devices is (0.59, 0.67).

- b. Construct a 95% confidence interval with continuity correction to compare the long run proportion of individuals who use health or wellness apps on their portable devices in the metropolitan and micropolitan areas. (Hint: Use count(dataset, PR\_RUCA\_2010, UsedHealthWellnessApps2) to get conditional counts for each area.) [3pts.]

```
count(hints, PR_RUCA_2010, UsedHealthWellnessApps2)
```

PR_RUCA_2010	UsedHealthWellnessApps2	n
1 Metropolitan	No	153
2 Metropolitan	Yes	289
3 Micropolitan	No	20
4 Micropolitan	Yes	13
5 Rural	No	2
6 Rural	Yes	6
7 Small town	No	11
8 Small town	Yes	6

In the metropolitan areas, 289 out of 442 used the apps, while in the micropolitan areas, 13 out of 33 used the apps.

### ⚠ Warning

The `prop.test()` can be used to implement the continuity correction. While the `prop.test()` can calculate the confidence intervals, it uses the pooled variance method in confidence intervals. It also uses a different distribution to calculate the reliability coefficient.

For this problem, I will use `prop.test()` to implement the continuity correction.

```
ptest <- prop.test(x=c(289,13), n=c(289+153,33), conf.level=0.95, correct=TRUE)  
ptest
```

```
2-sample test for equality of proportions with continuity correction  
  
data: c(289, 13) out of c(289 + 153, 33)  
X-squared = 7.8708, df = 1, p-value = 0.005024  
alternative hypothesis: two.sided  
95 percent confidence interval:  
 0.0711144 0.4486991  
sample estimates:  
 prop 1    prop 2  
0.6538462 0.3939394
```

The resulting 95% confidence interval for the long-run difference in proportion of individuals who use health or wellness apps on their portable devices between the metropolitan and micropolitan areas with the continuity correction is (0.0711, 0.4487).

## 4 Problem 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. Construct a 95% confidence interval that estimates the difference in average baseline follicle-stimulating hormone (FSH; column: fsh) measurements between those who did and did not experience hotflashes (column: hotflash). Assume unequal variances. [3pts.]

```
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, 3, 2, 1, 1, 1, 3, 2, 1, 1, 1, 2~
$ aagrp   <int> 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, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1~
$ d1      <int> 0, 0, 0, 1, 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, 1~
$ pcs12   <dbl> 56.80537, 59.18338, 57.73952, 55.83575, 55.89324, NA, 55.5009~
$ hotflash <int> 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0~
$ bmi30   <int> 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 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~
```

```
ttest <- t.test(hflashes$fsh ~ hflashes$hotflash,
                 data=hflashes,
                 var.equal=F)
ttest
```

Welch Two Sample t-test

```
data: hflashes$fsh by hflashes$hotflash
t = -3.1377, df = 151.77, p-value = 0.002046
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to
95 percent confidence interval:
-2.9014940 -0.6593356
```

```
sample estimates:  
mean in group 0 mean in group 1  
7.477043      9.257458
```

The 95% confidence interval for the difference in average baseline FSH measurements between those who did and did not experience hot flashes is (-2.9015, -0.6593)