Exam 1 - Selected Answers

```
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   3.5.2      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 becom
```

```
library(summarytools)
```

```
Attaching package: 'summarytools'

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

    view
```

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

1 Item 1

The variable hotflash includes the participant responses when asked if they experienced hot flashes in the past month (0=No, 1=Yes). What is the relative frequency of participants who reported to have experienced hot flashes in the past month?

```
freq(hflashes$hotflash)
```

```
Frequencies
hflashes$hotflash
Type: Integer

              Freq    % Valid    % Valid Cum.    % Total    % Total Cum.
         ----------- ------ --------- -------------- --------- --------------
           0    257     68.53           68.53      68.53           68.53
           1    118     31.47          100.00      31.47          100.00
        <NA>      0                                 0.00          100.00
       Total    375    100.00          100.00     100.00          100.00
```

1.1 Item 2

The variable ageg corresponds to the age group where the participants belong based on their age. Group 1: 35-39 years, Group 2: 40-44 years, and Group 3: 45-48 years. Which of the age groups had the highest representation in the study?

```
freq(hflashes$ageg)
```

```
Frequencies
hflashes$ageg
Type: Integer

              Freq    % Valid    % Valid Cum.    % Total    % Total Cum.
         ----------- ------ --------- -------------- --------- --------------
           1    133     35.47           35.47      35.47           35.47
           2    164     43.73           79.20      43.73           79.20
           3     78     20.80          100.00      20.80          100.00
        <NA>      0                                 0.00          100.00
       Total    375    100.00          100.00     100.00          100.00
```

1.2 Item 3

The variable fsh contains the baseline follicle-stimulating hormone (FSH) measurements for each participant. What is the mean/median/range/min/max baseline FSH measurement for this group of participants?

```
mean(hflashes$fsh)
```

```
[1] 8.03728
```

```
median(hflashes$fsh)
```

```
[1] 6.88
```

```
range(hflashes$fsh)
```

```
[1]  1.665 47.220
```

```
range(hflashes$fsh) |> diff()
```

```
[1] 45.555
```

1.3 Item

Suppose you randomly sample 50 participants from this cohort. What is the probability that
the average baseline FSH measurement for the sample is greater than 10? Assume the central
limit theorem holds for this scenario. (NOTE: The entire cohort is made up of 375 participants,
which is a finite population.)

```
n<- 50
N <- 375
mu <- mean(hflashes$fsh)
se <- sd(hflashes$fsh)/sqrt(n) * sqrt((N-n)/(N-1))

pnorm(9,mean=mu,sd=se,lower.tail=F)
```

```
[1] 0.04354046
```