**Problem Set 1 Solutions**

Answers are in red.

**Types of Data**. Decide if the following are examples of discrete or continuous data.

1. The number of deaths in the United States in a specific year

   a) Discrete, b) Continuous

2. The concentration of chlorine in a sample of water

   a) Discrete, b) Continuous

3. The length of time to recovery after a heart attack

   a) Discrete, b) Continuous

4. The number of adults hospitalized for respiratory disease during the summer of 2009 in Beijing

   a) Discrete, b) Continuous

**Frequency.** In this question, we examine the reported numbers of hospitalizations for cardiac events in the United States for each month in the period January 1991 to December 1992.

| Month 1991 | Number (Thousands) | Month 1992 | Number (Thousands) |
|---|---|---|---|
| January | 325 | January | 317 |
| February | 312 | February | 302 |
| March | 346 | March | 339 |
| April | 340 | April | 328 |
| May | 355 | May | 361 |
| June | 342 | June | 333 |
| July | 358 | July | 341 |
| August | 346 | August | 343 |
| September | 365 | September | 359 |
| October | 355 | October | 305 |
| November | 324 | November | 312 |
| December | 342 | December | 321 |

1. In the year 1991, what is the relative frequency of hospitalizations in September?

   365/(325 + 312 + 346 + 340 + 355 + 342 + 358 + 346 + 365 + 355 + 324 + 342)

   = 8.888%

2. Is the absolute frequency of hospitalizations in September 1991 greater than the absolute frequency of hospitalizations in September 1992?

   (a) Yes (b) No

   365 in 1991 compared to 359 in 1992.

3. Restricting to the year 1992, calculate the relative frequency of hospitalizations in September. Comparing this estimate with the relative frequency calculation in part (a), is the relative frequency of hospitalizations in September higher in 1991 or in 1992?

   (a) 1991 (b) 1992

   From previous part, we know that relative frequency is ~ 8.89% in 1991.

   In 1992, the relative frequency is

   359/(317 + 302 + 339 + 328 + 361 + 333 + 341 + 343 + 359 + 305 + 312 + 321)

   = 9.06%

   Even though the absolute frequency of hospitalizations in greater in September 1991 versus September 1992, the relative frequency is higher in 1992!

**Prevalent hypertension.** In this question, we use data from the NHLBI teaching data set. Our study population is the 4,434 participants in the Framingham Heart Study attending the first examination in 1956. Using the Framingham dataset, we explore data types, tables, and graphs in this question. We will examine the indicator of prevalent hypertension at exam 1 in 1956 (variable name: prevhyp1).

1. Prevalent hypertension at exam 1 is an example of what type of data?

   (a) nominal, (b) binary, (c) both a and b

2. How many individuals in the study population had prevalent hypertension at exam 1?

   1,430

```
. tabulate prevhyp1

  Prevalent |
hypertensio |
  n, exam 1 |      Freq.     Percent        Cum.
------------+-----------------------------------
         No |      3,004       67.75       67.75
        Yes |      1,430       32.25      100.00
------------+-----------------------------------
      Total |      4,434      100.00
```

3. What is the relative frequency of prevalent hypertension at exam 1?

   32.25%, from table above.

4. Among the individuals with prevalent hypertension at exam 1, how many are female?

   799

```
tabulate sex1 prevhyp1

           |       Prevalent
 Sex, exam | hypertension, exam 1
         1 |        No        Yes |     Total
-----------+----------------------+----------
      Male |     1,313        631 |     1,944
    Female |     1,691        799 |     2,490
-----------+----------------------+----------
     Total |     3,004      1,430 |     4,434
```

5. What percent of individuals with prevalent hypertension at exam 1 are female?

   55.8%

```
. display 799/1430
.55874126
```

6. Which graph would you use to summarize the distribution of the indicator for prevalent hypertension at exam 1 in the study population?

   (a) scatter plot, (b) histogram, (c) bar chart

**BMI at baseline.** In this question, we again use data from the NHLBI teaching data set to examine the continuous variable, body mass index (BMI). Our study population is the subset of participants in the Framingham Heart Study attending the first examination in 1956 with a non-missing BMI measure (4,417 participants out of 4,434).

1. To quickly examine the interquartile range for BMI at exam 1 in the study population, which graph would you use?

   (a) histogram, (b) boxplot, (c) scatter plot

2. We say an individual has high BMI at exam 1 if his BMI is greater than 25. How many individuals in the dataset have high BMI at exam 1?

   2,422

```
. gen bmihigh = .
(4434 missing values generated)

. replace bmihigh = 1 if bmi1 > 25 & bmi1 < .
(2422 real changes made)

. replace bmihigh = 0 if bmi1 <= 25
(1993 real changes made)

. tabulate bmihigh
```

| bmihigh | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 1,993 | 45.14 | 45.14 |
| 1 | 2,422 | 54.86 | 100.00 |
| Total | 4,415 | 100.00 | |

3. Out of the 4,415 participants with a BMI measurement at exam 1, what percent had high BMI at exam 1 (an individual has high BMI at exam 1 if his BMI is greater than 25).

   54.86%, from table above

4. Restricting to the population with BMI measurements at both exam 1 and exam 2, make a scatter plot of BMI at exam 1 (bmi1) versus BMI at exam 2 (bmi2). In general, higher BMI at exam 1 is associated with a _____ BMI at exam 2.

   (a) higher, (b) lower

**BMI over time.** In this question, we again use data from the NHLBI teaching data set to examine the continuous variable, body mass index (BMI). Our study population is the subset of participants in the Framingham Heart Study attending the first examination in 1956 with a non-missing BMI measure (4,415 participants out of 4,434).

1. What is the mean BMI at exam 1 in the study population?

   25.84616

```
. sum bmi1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        bmi1 |      4415    25.84616    4.101821      15.54       56.8
```

2. The median BMI at exam 1 is 25.45 in the study population. Comparing the mean and median, do these data suggest that the distribution of BMI at exam 1 is right skewed or left skewed?

   a) right skewed, b) left skewed

   The when the mean is larger than the median, the data is usually right skewed (note: this may not be true if you have one or two outlying points that inflate the mean substantially).

3. Is the mean BMI at exam 1 higher in males or females?

   a) Males, b) Females

```
bysort sex1: sum bmi1

-------------------------------------------------------------------------------
-> sex1 = Male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        bmi1 |      1939    26.16958    3.407115      15.54      40.38

-------------------------------------------------------------------------------
-> sex1 = Female

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        bmi1 |      2476    25.59288    4.557443      15.96       56.8
```

4. Should you compare the mode for BMI at exam 1 in males versus females?

   a) Yes, b) No

   Usually, comparing modes for continuous data is not informative.

5. Is the IQR for BMI at exam 1 larger in males or females?

   a) Males, b) Females

```
. by sex1, sort : centile bmi1, centile(25 75)

-------------------------------------------------------------------------------
-> sex1 = Male

                                                    -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
        bmi1 |    1939         25         23.97        23.79117       24.12
             |                 75         28.32        28.09284        28.5

-------------------------------------------------------------------------------
-> sex1 = Female

                                                    -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
        bmi1 |    2476         25         22.54           22.36       22.72
             |                 75         27.82        27.53037       28.06

. display 28.32 - 23.97
4.35

. display 27.82 - 22.54
5.28
```

**Now, for the remaining parts of this question, restrict your study population to the subset of participants with BMI measures at exam 1 and exam 2.**

6. What is the mean change in BMI from exam 1 to exam 2? Change in BMI is defined as BMI at exam 2 minus BMI at exam1. (Note: you need to generate this variable in Stata).

   0.068

```
. gen bmidiff = bmi2 - bmi1
(525 missing values generated)

. sum bmidiff

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     bmidiff |      3909    .0678306    1.801516      -10.5      10.43
```

7. What is the standard deviation of the change in BMI from exam 1 to exam 2?

   1.80, from summarize command above

8. What is the range of changes in BMI from exam 1 to exam 2?

   20.93

```
. di 10.43 - -10.5
20.93
```

9. Assuming that the empirical rule applies in this situation, we expect that 95% of individuals will have a change in BMI between exams 1 and 2 that lies within the interval _____.

   (-3.53, 3.67)

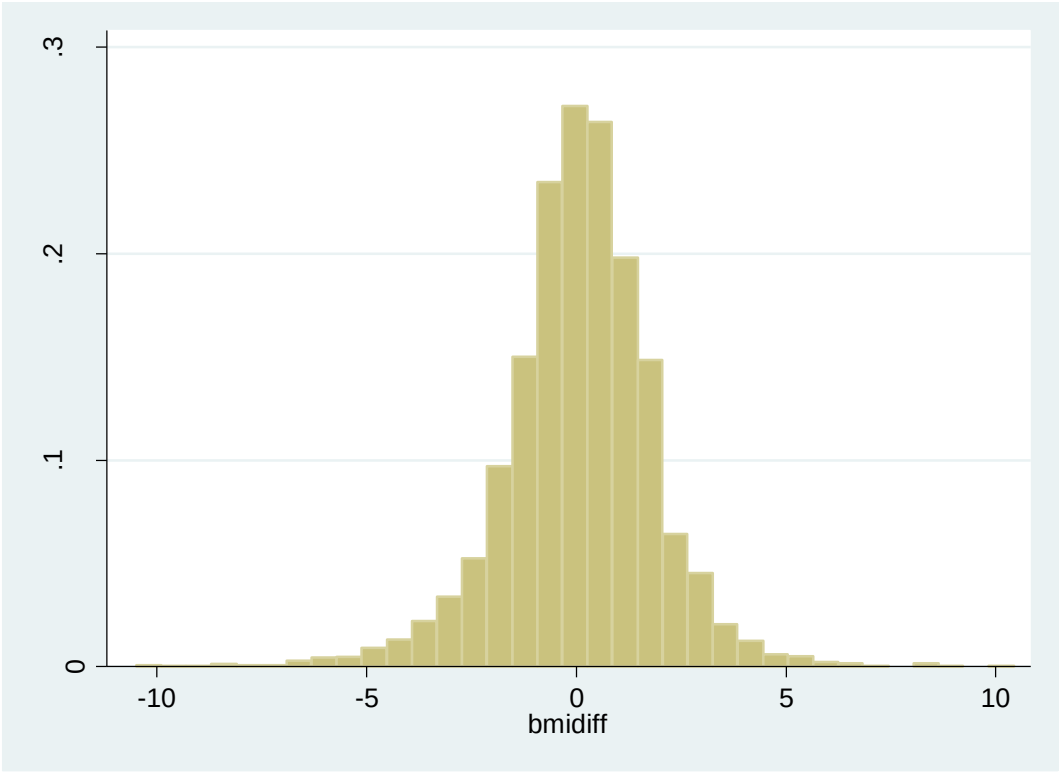
```
. di .0678306 - 2*1.801516
-3.5352014

. di .0678306 + 2*1.801516
3.6708626
```

10. Does it appear that the empirical rule can be used in the previous question, examining the change in BMI from exam 1 to exam 2?

    a) yes, b) no

    To use the empirical rule, the distribution of the variable should be symmetric and unimodal. Examining the histogram below, this seems to hold.


.3
.2
.1
0
-10
-5
0
5
10
bmidiff