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	Number	Month	Number
1991	(Thousands)	1992	(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)

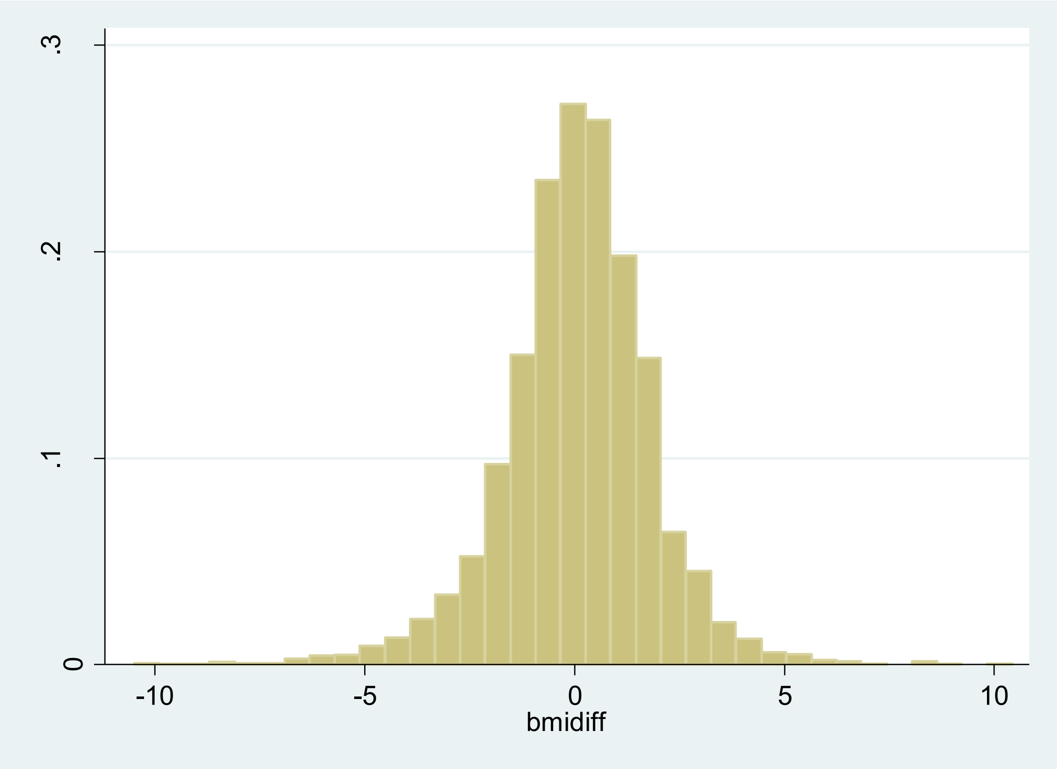
```
. di .0678306 - 2*1.801516
-3.5352014

. 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.


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

Problem Set 2 Solutions

BMI and CHD prevalence. The following table uses data from the NHLBI teaching data set and displays categories of body mass index for 4,415 participants in the Framingham Heart Study attending an examination in 1956 with non-missing values for body mass index. For each body mass index category, the table displays the number of subjects with existing Coronary Heart Disease (CHD) at that exam (**prevchd=1**)

Body Mass Index Category		Number of Subjects at 1956 Exam	Cases of CHD Diagnosed Prior to 1956
Under Weight	BMI < 18.5	57	0
Normal Weight	$18.5 \leq$ BMI < 25	1936	66
Overweight	$25 \leq$ BMI < 30	1848	90
Obese	BMI $\geq$ 30	574	38
Total		4415	194

1. What is the prevalence of obesity among the 4415 participants at the 1956 exam?

   574/4415 = 0.1300

2. What is the prevalence of CHD at the 1956 exam among the 4415 participants at the 1956 exam?

   194/4415 = 0.0439

3. What is the prevalence of CHD at the 1956 exam for each of the body mass index classes?

Under Weight Participants	**0/57 = 0.0**
Normal Weight Participants	**66/1936 = 0.0341**
Overweight Participants	**90/1848 = 0.0487**
Obese Participants	**38/574 = 0.0662**

Diabetes prevalence. Use Stata and the NHLBI data set to calculate the prevalence of diabetes among participants who attended and had non-missing data on diabetes at all three examinations. (**Hint: There were 3,206 such participants.**)

1. What is the prevalence of diabetes at the first exam (**diabetes1=1**)?

   58/3206 = 0.0181

2. What is the prevalence of diabetes at the second exam (**diabetes2=1**)?

   105/3206 = 0.0328

3. What is the prevalence of diabetes at the third exam (**diabetes3=1**)?

   251/3206 = 0.0783

```
tab1 diabetes* if diabetes1<. & diabetes2<. & diabetes3<.
```

BMI and hypertension prevalence. Use Stata and the BMI1 variable in the NHLBI data set to create the four categories of body mass index as defined in the first question.

1. What is the prevalence of hypertension (**prevhyp1=1**) at the 1956 exam for each of the body mass index classes?

Under Weight Participants	**6/57 = 0.1053**
Normal Weight Participants	**398/1936 = 0.2056**
Overweight Participants	**683/1848 = 0.3696**
Obese Participants	**336/574 = 0.5854**

Hypertension and high blood pressure. Use Stata to create a binary variable (**highbp1**) to represent the presence/absence of high blood pressure at the 1956 examination

```
gen highbp1=.
replace highbp1=1 if (sysbp1>=140 | diabp1 >= 90)
replace highbp1=0 if (sysbp1<140 & diabp1 < 90)
```

(**Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then they should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 $\geq$ 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 $\geq$ 90 then highbp1 =1, otherwise highbp1 should be missing.)**

1. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants with high blood pressure at the 1956 exam (highbp1=1)?

   106/1619 = 0.0655

2. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants without high blood pressure at the 1956 exam (highbp1=0)?

   88/2815 = 0.0313

Hypothetical life table. The table below lists the number of individuals alive at age *x*, for a hypothetical population in 1950-1952 and 1990-19992.

Number of Survivors out of 100,000 Live Births

Age	1950-1952	1990-1992
0	100,000	100,000
20	73,412	96,902
40	56,884	92,638
70	31,744	79,873

1. What is the probability of surviving from birth to age 20 in 1950-1952?

   0.73412

   73412/10000 = 0.73412

2. What is the probability of surviving from age 40 to age 70 in 1990-1992?

   0.86221

   79873/92638 = 0.86221

3. Define the absolute survival increase over the 40 year span as $p_1 - p_2$, where p_1 is the chance of surviving from age x to age x+n in 1990-1992 and p_2 is the chance of surviving from age x to age x+n in 1950-1952. Which age group has the greatest absolute survival increase?

   (a) 0 – 20 , (b) 20 – 40, **(c) 40 – 70**

   Survival: 1950 – 1952

   0 – 20: 73412/100000 = 0.73412
   20 – 40: 56884/73412 = 0.7748597
   40 – 70: 31744/56884 = 0.558048

   Survival: 1990 – 1992

   0 – 20: 96902/100000 = 0.9559968
   20 – 40: 92638/96902 = 0.9559968
   40 – 70: 79873/92638 = 0.8622056

   Absolute Survival Increase:

   0 – 20: (0.9559968 - 0.73412) = 0.239
   20 – 40: (0.9559968 - 0.7748597) = 0.181
   40 – 70: (0.8622056 - 0.558048) = 0.304

4. Define the relative survival increase over the 40 year span as $(p_1 - p_2)/p_2$, where p_1 is the chance of surviving from age x to age x+n in 1990-1992 and p_2 is the chance of surviving from age x to age x+n in 1950-1952. Which age group has the greatest relative survival increase?

   (a) 0 – 20 , (b) 20 – 40, **(c) 40 – 70**

   Relative Survival Increase:

   0 – 20: (0.9559968 - 0.73412)/0.73412 = 0.3022351
   20 – 40: (0.9559968 - 0.7748597)/0.7748597 = 0.2337676
   40 – 70: (0.8622056 - 0.558048)/0.558048 = 0.5450384

Problem Set 3

Probability and BMI. The following table uses data from the NHLBI teaching data set and displays the body mass index for 3,909 participants in the Framingham Heart Study with BMI measurements at the first two exams, in 1956 and in 1962.

BMI category	BMI ≤ 25, exam 2	BMI > 25, exam 2	Total
BMI ≤ 25, exam 1	1,492	278	1,770
BMI > 25, exam 1	249	1,890	2,139
Total	1,741	2,168	3,909

Assume a study participant has been randomly selected from this subset of 3,909 participants.

- Define **A** as the event that this participant has a high BMI at exam 1.
- Define **B** as the event that this participant has a high BMI at exam 2.
- Define **C** as the event that this participant has a low BMI at exam 2.

1. What is the probability of A?

   0.547

   2139/3909 = .54719877

2. What is the probability of B?

   0.555

   2168/3909 = .55461755

3. What is the probability of A and B?

   0.483

   1890/3909 = .48349962

4. Are A and B independent?

   No.

   P(A)P(B) = .54719877*.55461755 = .30348604 ≠ P(A and B) = .48349962

5. In a randomly selected participant, what is the probability that A and/or B occurs (namely, that the participant's BMI is high during at least one of the first two exams)?
   0.618

   P(A and/or B) = P(A) + P(B) – P(A and B) = .54719877 + .55461755 - .48349962 = .6183167

6. What is the probability that B occurs, given that A occurs?

   0.884

   P(B|A) = P(A and B)/P(A) = . 48349962/.54719877 = .88359047

7. What is the probability that C occurs, given that A occurs?

   0.116

   P(C|A) = 1 – P(B|A) = .11640953

Probability of age and smoking events. The following table uses data from the NHLBI teaching data set using the 4,434 participants in the Framingham Heart Study who attended the exam in 1956. Below, we show a table that contains the probabilities that a study participant falls into one of the eight categories defined by all possible combinations of the age categories and smoking status at the first exam in 1956. Recall that all study participants are between 30 and 70 years old.

Let A be the event that a randomly chosen study participant is a smoker at exam 1 in 1956.

Let B be the event that the person chosen is between 60 and 70 years old at exam 1.

Age, exam 1	Smoker, Exam 1	
	No	Yes
30-39	0.0519	0.0742
40-49	0.1554	0.2262
50-59	0.1809	0.1346
60-70	0.1200	0.0568

1. Eight categories representing age and smoking status groups are shown in the table above. Are these groups:

   (a) mutually exclusive, (b) exhaustive, (c) both

   Someone can only belong to one age group and one smoking status at any given time. Therefore, the eight categories are mutually exclusive. The sum of the probabilities in the above table is equal to 1, implying that these events are exhaustive (we also know that all participants in the study are between 30 and 70 years old and are either smokers or non-smokers, again implying the events are exhaustive).

2. What is the probability of A?

   0.4918

   P(A) = 0.0742 + 0.2262 + 0.1346 + 0.0568 = .4918

3. What is the probability of B^C, the complement of B?

   0.8232

   $P(B^C)$ = 1 – P(B) = 1 - 0.1200 - 0.0568 = 0.8232

4. What is the probability that a randomly selected individual is a non-smoker who is younger than 60 years old at exam 1?

   0.3882

   P(event) = 0.0519 + 0.1554 + 0.1809 = 0.3882

5. Are the events A and B independent?

   No

   P(A and B) = 0.0568 ≠ P(A)P(B) = 0.4918*(1-0.8232) = 0.08695024

Diagnostic Testing. Screening for prostate cancer in men is a controversial topic. One of the most common screening mechanisms is the PSA test (prostate antigen test). In a meta-analysis, Mistry and Cable (2003) report that the sensitivity of the PSA test is 72.1% and the specificity is 93.2%. In the United States, it is estimated that 16.1% of men will have prostate cancer at some point in their life (America Cancer Society 2012). Assume that the prevalence of prostate cancer among men ages 75 and older is 16.1%. We examine the properties of the PSA screening test in men ages 75 and older, using the sensitivity and specificity values above.

Mistry K. and Cable G. (2003). Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *The Journal of the American Board of Family Practice*, 16(2): 95-101.

1. What is the probability of a false negative test result?

   0.279

   P(T-|D+) = 1 – P(T+|D+) = 1 – sensitivity = 1 – 0.721 = 0.279

2. What is the probability of a false positive result?

   0.068

   P(T+ | D-) = 1 – P(T-|D-) = 1 – specificity = 1 – 0.932 = 0.068

3. What is the probability that a randomly selected man who is 75 years or older DOES NOT have prostate cancer, given that his PSA screening was positive?

   0.330

   Here we use Bayes' Theorem and simply substitute values for sensitivity and specificity

   $$P(D^+|T^+) = \frac{P(D^+)P(T^+|D^+)}{P(D^+)P(T^+|D^+) + P(D^-)P(T^+|D^-)}$$

   $P(D^+) = 0.161$ → $Pr(D^-) = 1 – 0.161 = 0.839$

   $P(T^+|D^+)$ = sensitivity = 0.721, $P(T^-|D^-)$ = specificity = 0.932 , $P(T^+|D^-)$ = 1 – specificity = 0.068

   $P(D^+)\ P(T^+|D^+) = 0.116081$

   $P(D^-)\ P(T^+|D^-) = 0.057052$

   $$P(D^+|T^+) = \frac{0.116081}{0.116081 + 0.057052}$$

   $P(D^+|T^+) = 0.67047299$

   $P(D^-|T^+) = 1 - P(D^+|T^+) = 0.329527$

4. What is the probability that a randomly selected man who is 75 years or older has prostate cancer, given that his PSA screening was negative?

   0.05432434

   $$P(D^+|T^-) = \frac{P(D^+)P(T^-|D^+)}{P(D^+)P(T^-|D^+) + P(D^-)P(T^-|D^-)}$$

   $$P(D^+|T^-) = \frac{0.161 * (1 - 0.721)}{0.161 * (1 - 0.721) + 0.839 * 0.932}$$

   $$P(D^+|T^-) = 0.05432434$$

Titanic Survival. The following table describes the survival status of passengers on the Titanic, stratified by Passenger Class (First, Second, or Third), Sex/Age (Child, Women, or Man), and Survival Status. The Frequency column indicates the number of passengers in each stratum. (For example there were 4 1st class women passengers who did not survive and 140 1st class women passengers who did survive). These data were obtained from the website anesi.com and refers to British Parliamentary Papers, Shipping. Casualties (Loss of the Steamship "Titanic"), 1912. cmd 6352 "Report of a Formal Investigation into the circumstances attending the foundering on the 15th April 1912 of the British Steamship "Titanic" of Liverpool after striking ice in or near Latitude 41 46 N., Longitude 50 14 W., North Atlantic Ocean, whereby loss of life ensued (London; His Majesty's Stationary Office, 1912) page 42.

Passenger Class	Age/Sex	Survival Status	Frequency
First	Child	Survived	6
First	Child	Did Not Survive	0
First	Women	Survived	140
First	Women	Did Not Survive	4
First	Man	Survived	57
First	Man	Did Not Survive	118
Second	Child	Survived	24
Second	Child	Did Not Survive	0
Second	Women	Survived	80
Second	Women	Did Not Survive	13
Second	Man	Survived	14
Second	Man	Did Not Survive	154
Third	Child	Survived	27
Third	Child	Did Not Survive	52
Third	Women	Survived	76
Third	Women	Did Not Survive	89
Third	Man	Survived	75
Third	Man	Did Not Survive	387

Use these data to calculate the cumulative incidence of surviving for each of the following groups of individuals:

1. All Women 296/402 = 0.74

2. All Children 57/109 = 0.52

3. All Women or Children 353/511 = 0.69

4. All First Class Passengers 203/325 = 0.62

BMI and Cumulative Incidence. The following table uses data from the NHLBI teaching data set and displays categories of body mass index (used in the previous homework assignment) for 4,415 participants in the Framingham Heart Study attending an examination in 1956 with non-missing values for body mass index. For each body mass index category, the table displays the number of subjects who died (**death=1**) during follow-up and the total person-years of follow-up (**timedeath**) until death or the end of the follow-up period (24 years). Assume that all deaths and time of death were recorded among the 4415 participants.

Body Mass Index Category		Number of Subjects at 1956 Exam	Number of Deaths	Total Person-Years
Under Weight	BMI < 18.5	57	18	1181.44
Normal Weight	18.5 ≤ BMI < 25	1936	571	40708.74
Overweight	25 ≤ BMI < 30	1848	691	37728.41
Obese	BMI ≥ 30	574	257	11254.52
Total		4415	1537	90873.11

1. What is the cumulative incidence of death among the 4415 participants at the 1956 exam?

   1537/4415 = 0.3481

2. What is the cumulative incidence of death during the 24 years of follow-up for each of the body mass index class?

   Under Weight Participants 18/57 = 0.3158

   Normal Weight Participants 571/1936 = 0.2949

   Overweight Participants 691/1848 = 0.3739

   Obese Participants 257/574 = 0.4477

3. What is the incidence rate of death among the 4415 participants during the 24 years of follow-up? (Express your answer as #deaths/(1000 person-years))

   1537/90873.11 = 16.9137 (deaths/(1000 person-years))

4. What is the incidence rate of death during the 24 years of follow-up for each of the body mass index classes? (Express your answers as #deaths/(1000 person-years))

   Under Weight Participants 18/1181.44 = 15.2356 (deaths/(1000 person-years))

   Normal Weight Participants 571/40708.74 = 14.0265 (deaths/(1000 person-years))

   Overweight Participants 691/37728.41 = 18.3151 (deaths/(1000 person-years))

   Obese Participants 257/11254.52 = 22.8353 (deaths/(1000 person-years))

BMI and CHD Incidence. Use Stata and the NHLBI data set to create a separate variable for each of the four categories of body mass index as defined in the previous question (for example, create a variable called “underwt” that equals 1 if a person’s BMI was less than 18.5 and 0 if a person’s BMI was ≥18.5). To create the separate variables for each of the four categories of body mass index, use the BMI1 variable in the NHLBI dataset. What is the incidence rate for developing CHD (anychd=1) during the 24-years of follow-up for participants in each of the body mass index categories? (Express your answers as #deaths/(1000 person-years)) Hint: Number of years a person was followed for CHD is recorded in the “timechd” variable in the NHLBI dataset.

1. Under Weight Participants 6/(1119.57py) = 5.36/(1000py)
2. Normal Weight Participants 415/(37324.05py) = 11.12/(1000py)
3. Overweight Participants 596/(32863.49py) = 18.14/(1000py)
4. Obese Participants 218/(9387.74) = 23.22/(1000py)

High Blood Pressure and CHD. Use Stata and the NHLBI data set to create the two categories of high blood pressure (**highbp1**).

```
generate highbp1=.
replace highbp1=1 if (sysbp1>=140 | diabp1 >= 90)
replace highbp1=0 if (sysbp1<140 & diabp1 < 90)
```

(**Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then they should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 $\geq$ 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 $\geq$ 90 then highbp1 =1, otherwise highbp1 should be missing.)**

1. What is the incidence rate for developing CHD (**anychd=1**) during the 24-years of follow-up for participants in each of the blood pressure categories? (Express your answers as #deaths/(1000 person-years))

   Participants with high blood pressure at the 1956 exam (highbp1=1)

   605/(25540.74py) = 23.69/(1000py)

   Participants without high blood pressure at the 1956 exam (highbp1=0)

   635/(55384.42py) = 11.47/(1000py)

Problem Set 4 Solutions

Titanic Survival Risk Ratios. The following table describes the survival status of passengers on the Titanic, stratified by Passenger Class (First, Second, or Third), Sex/Age (Child, Women, or Man), and Survival Status. The Frequency column indicates the number of passengers in each stratum. (For example there were 4 1st class women passengers who did not survive and 140 1st class women passengers who did survive). These data were obtained from the website anesi.com and refers to British Parliamentary Papers, Shipping. Casualties (Loss of the Steamship "Titanic"), 1912. cmd 6352 "Report of a Formal Investigation into the circumstances attending the foundering on the 15th April 1912 of the British Steamship "Titanic" of Liverpool after striking ice in or near Latitude 41 46 N., Longitude 50 14 W., North Atlantic Ocean, whereby 1oss of life ensued (London; His Majesty's Stationary Office, 1912) page 42.

Passenger Class	Age/Sex	Survival Status	Frequency
First	Child	Survived	6
First	Child	Did Not Survive	0
First	Women	Survived	140
First	Women	Did Not Survive	4
First	Man	Survived	57
First	Man	Did Not Survive	118
Second	Child	Survived	24
Second	Child	Did Not Survive	0
Second	Women	Survived	80
Second	Women	Did Not Survive	13
Second	Man	Survived	14
Second	Man	Did Not Survive	154
Third	Child	Survived	27
Third	Child	Did Not Survive	52
Third	Women	Survived	76
Third	Women	Did Not Survive	89
Third	Man	Survived	75
Third	Man	Did Not Survive	387

1. Use these data to calculate the Risk Ratio for surviving comparing "women or children" as the exposed group and "all other passengers" as the non-exposed group.

   RR = (353/511) / (146/805) = 3.81

2. Repeat this calculation for each passenger class.

   First Class: RR = (146/150) / (57/175) = 2.99

   Second Class: RR = (104/117) / (14/168) = 10.67

   Third Class: RR = (103/244) / (75/462) = 2.60

Incidence Rate Ratio Blood Pressure and CHD. The following table uses data from the NHLBI teaching data set and displays the blood pressure distribution for 4,434 participants in the Framingham Heart Study attending an examination in 1956. For each blood pressure category, the tables displays the number of subjects with existing Coronary Heart Disease (CHD) at that exam (Prevalent Cases of CHD) and also the number of new cases of CHD and the total amount of person-years of follow-up that was observed during a 24 year follow-up period **for those subjects who did not have CHD at the 1956 exam**. Follow-up for each subject began in 1956 and ended with the development CHD (fatal or non-fatal), death from another cause, loss to follow-up, or the end of the follow-up period (whichever came first).

Blood Pressure Category		Number of Subjects at 1956 Exam	Prevalent Cases of CHD	Number Developing CHD During Follow-up	Total Years of Follow-up
I	< 140 and DBP <90	2815	88	547	55,384.42
II	140 ≤ SBP < 160 or 90 ≤ DBP < 95	781	39	214	13,191.79
III	SBP ≥ 160 or DBP ≥95	838	67	285	12,348.94

1. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Groups II or III combined (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (499/ 25540.73py) / (547/55384.42py) = 1.98

2. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Group III (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (285/12348.94py) / (547/55384.42py) = 2.34

3. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Group II (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (214/13191.79py) / 547/384.42py) = 1.64

Risk Ratios, Odds Ratios, Rate Ratios for BMI, Death, and CHD. The following table displays categories of body mass index for the participants at the 1956 exam. (Note: 19 of the 4434 participants are excluded from this table because of missing data on bmi1.)

BMI Category	Range of BMI	Frequency
Underweight	0 ≤ bmi1 < 18.5	57
Normal Weight	18.5 ≤ bmi1 < 25	1936
Overweight	25 ≤ bmi1 < 30	1845
Obese	bmi1 ≥ 30	577

Use Stata to perform the following calculations.

Hint: All of the following questions ask you to compare obese subjects to normal weight subjects. Create a new binary variable using bmi1 which equals 1 if the person is obese and 0 if the person is normal weight. Anyone who is underweight or overweight should be missing a value for the new binary variable you create.

1. Calculate the 24-year Risk Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   RR = (259/577) / (571/1936) = 1.52

2. Calculate the 24-year Odds Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   OR = (259/318) / (571/1365) = 1.95

3. Calculate the 24-year Rate Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   RR = (259/11308.90py) / (571/40708.74py) = 1.63

4. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing obese subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Hint: Use the if/in tab options to restrict the sample to those without prevalent CHD at the 1956 exam (prevchd1 = = 0).

   Rate Ratio = (180/9387.74py) / (349/37324.05py) = 2.05

5. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing (obese or overweight) subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Rate Ratio = (686/42251.22py) / (349/37324.05py) = 1.74

6. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing underweight subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Rate Ratio = (6/1119.57py) / (349/37324.05py) = 0.57

Please use this information for the remainder of the assignment.

According to data from the CDC in 2010, 19.3% of adults age eighteen and older smoke cigarettes. In the year 2008, the incidence rate of lung cancer was 65.1 cases per 100,000 people per year.

Suppose you are conducting a lung cancer study in the United States, and you obtain a random sample of 2,000 adults (over 18 years of age) who do not have lung cancer. You plan to follow this study cohort over a period of 5 years and observe incident cases of lung cancer.

Smoking and the binomial distribution. Smoking status is an important predictor of lung cancer incidence. Therefore, as the study designer, it is important to think about baseline smoking rates in your study cohort. We first model the number of smokers in the study cohort using the binomial distribution, and assume that this cohort is representative sample from the US population. Use the binomial distribution to answer the parts below:

1. How many smokers would you expect to see in the study cohort, on average?

   386

   Let X denote number of smokers in the study cohort, X ~ Binomial(2000, 0.193).

   E(X) = n*p = 2000*0.193 = 386.

2. What is the standard deviation of the number of smokers in the study cohort?

   17.6

   Var(X) = n*p*(1-p) = 2000*0.193*(1-0.193) = 311.502

   sd(X) = √311.502 = 17.64942

3. What is the probability that you observe exactly 386 smokers?

   0.023

   ```
   . di binomialp(2000, 386, 0.193)
   ```

   ```
   .0225986
   ```

4. What is the probability that greater than or equal to 25% of the study population are smokers?

0.000

```
. di binomialtail(2000, 2000*0.25, 0.193)

2.403e-10
```

5. What is the probability that less than or equal to 20% of the study population are smokers?

   0.795

   ```
   . di binomial(2000, 2000*0.20, 0.193)

   .79487415
   ```

Smoking and the normal distribution. In the questions above, we modeled the number of smokers in the study cohort using the binomial probability model. Now, assume that the number of smokers in the study cohort follows a **normal distribution.**

1. What is the probability that you observe exactly 386 smokers?

   0

   Because the normal distribution is continuous, the probability of observing exactly 386 smokers is always 0.

2. What is the probability that less than 20% of the study population are smokers?

   0.786

   $X \sim \text{Normal}(386, 311.5)$

   $P(X < 2000*0.2) = P(X < 400)$

   $= P(Z < (400-386)/17.6)$

   $= P(Z < 0.795)$

   ```
   . di normal(0.795)
   ```

   ```
   .78669325
   ```

3. Do you think the normal model provides a reasonable approximation to the binomial model in this example?

   Yes

Lung cancer and the binomial distribution. You also need to carefully consider how many cases of lung cancer you expect to observe in your study over time. We first model the number of lung cancer cases observed in the first year using the **binomial distribution.**

1. What proportion of the study population would you expect, on average, to be diagnosed with lung cancer in the first year?

   0.000651

   65.1 cases/100,000 person-years = 0.000651

2. How many cases of lung cancer would we expect to observe in the first year?

   1.3

   X ~ Binomial(2000, 0.000651)

   E(X) = 2000*0.000651 = 1.302

3. What is the variance of the number of lung cancer cases observed in the first year?

   1.3

   Var(X) = 2000*0.000651*(1-0.000651) = 1.30152

4. Why would you expect the mean and variance to be similar in this example?

   (a) because the event is rare, (b) because the mean is close to 1, (c) because we are dealing with incidence rates, (d) both (a) and (b)

   Recall that the Poisson distribution is a good approximation to the binomial distribution when the event is rare. Further, the mean and variance of the Poisson distribution are the same. Therefore, when the event is rare, we would expect the variance and the mean to be approximately equal.

5. What is the probability that you observe more than 1 lung cancer case in the first year?

   0.374

   ```
   . di binomialtail(2000, 2, 0.000651)
   ```

   ```
   .37392011
   ```

6. What is the probability that you observe no lung cancer cases in the first year?

   0.272

   ```
   . di binomial(2000, 0, 0.000651)
   ```

   ```
   .27187198
   ```

Lung cancer and the Poisson distribution. Because lung cancer is a rare disease, we can model cases of lung cancer using the **Poisson distribution**, with incidence rate 65.1 cases per 100,000 person-years.

1. Using the Poisson distribution, what is the probability that you observe more than 1 lung cancer case in the first year?

   0.374

   ```
   . di poissontail(1.3, 2)
   ```

   ```
   .37317688
   ```

2. What is the expected number of lung cancer cases observed over the five year study period?

   6.51

   Y ~ Poisson(u)

   E(Y) = 2000*0.000651*5 = 6.51

3. What is the variance of the number of lung cancer cases observed over the five year study period?

   6.51 (same as part (p))

   var(Y) = E(Y) because Y ~ Poisson(6.51).

4. What is the probability that you observe more than 10 lung cancer cases over the five year period?

   0.067

   ```
   . di poissontail(6.51, 11)
   ```

   ```
   .06739806
   ```

5. What is the probability that you observe less than 5 lung cancer cases over the five year period?

   0.223

   ```
   . di poisson(6.51, 4)
   ```

   ```
   .22255574
   ```

The big picture. As the study designer, you can compile these facts together, along with the goals of the study, to decide if the sample size of 2,000 individuals is large enough and if the follow up period of 5 years is long enough. If you (or the funding committee) decides that the current study is not big enough to answer the questions of interest about incident lung cancer cases, you either need to find more funding to extend your study or increase your sample size, OR you will have to abandon the study.

In this assignment, we have learned about properties of common probability models in statistics. Namely, we learned that:

1. The _______ provides a good approximation to the binomial distribution when the event of interest is **rare** and the study population is large.

   (a) Poisson, (b) normal, (c) exponential

2. The _______ provides a good approximation to the binomial distribution when the event of interest is **not rare** and the study population is large.

   (a) Poisson, (b) normal, (c) exponential

Problem Set 5 Solutions

Conceptual Questions: Predictive and Confidence Intervals

1. True or False: Consider a random variable X. To construct a 95% predictive interval for X, all we need to know is the sampling distribution of X.

   True - predictive intervals use the distribution of a random variable X to make statements about where we expect the values of the random variables to lie.

2. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% predictive interval for the sampling mean $\bar{X} = \sum_{i=1}^{n} X_i$, all we need to know is the sampling distribution of X.

   True

3. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% confidence interval for the true mean of X, denoted μ, all we need to know is the Central Limit Theorem.

   False - the CLT is an asymptotic result. We also need to know n, the sample size!

Central Limit Theorem and Confidence Intervals According to the WHO Global Database on Anaemia, the mean hemoglobin levels among primary school children in Delhi were estimated at $\mu = 108$ g/L, with standard deviation $\sigma = 12.5$ g/L.

Source: `http://who.int/vmnis/anaemia/data/database/countries/ind_ida.pdf`

Suppose we took a random sample of 75 primary school children in Delhi. Denote the mean hemoglobin levels in this sample as $\bar{x}$. Throughout this question, assume that the sample size is large enough that the central limit theorem is applicable and that σ is known.

1. What is the expected value (mean) of $\bar{x}$?

   108

2. What is the standard deviation of $\bar{x}$?

   $12.5/\sqrt{75} = 1.443376$

3. Suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

   0.8341433

   $P(106 < \bar{x} < 110) = P\left(\frac{106-108}{1.44} < Z < \frac{110-108}{1.44}\right)$

   ```
   . di normal((110 - 108)/1.44) - normal((106 - 108)/1.44)
   .83513346
   ```

4. Suppose instead we repeatedly took random samples of size 25. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

   0.5762892

   $P(106 < \bar{x} < 110) = P\left(\frac{106-108}{12.5/\sqrt{(25)}} < Z < \frac{110-108}{12.5/\sqrt{(25)}}\right)$

   ```
   . di normal((110 - 108)/(12.5/5)) - normal((106 - 108)/(12.5/5))
   .5762892
   ```

5. Again, suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a mean less than $\bar{x} = 103$?

   0.0002660028

   $P(\bar{x} < 103) = P\left(Z < \frac{103-108}{12.5/\sqrt{75}}\right) = P(Z < -3.4641016)$

```
. di normal(-3.4641016)
.000266
```

6. If we repeatedly took samples of size 75, we would expect that, in 20% of the samples, $\bar{x}$ would be greater than _____?

   104.21478

   $P(\bar{x} > y) = P(Z > \frac{y-108}{12.5\sqrt{75}}) = 0.2 \rightarrow$

   $y = 108 + Z_{0.8} * 12.5/\sqrt{75}$

   ```
   . di 108 + invnormal(0.8)*12.5/sqrt(75)
   109.21478
   ```

7. After taking a sample of size 75, we found that the sample mean was $\bar{x} = 103$. Construct a two-sided 95% confidence interval for μ.

   (100.171, 105.829)

   ```
   . di 103 - invnormal(0.975)*12.5/sqrt(75)
   100.17104

   . di 103 + invnormal(0.975)*12.5/sqrt(75)
   105.82896
   ```

8. True or False. Based on the above interval, we can say that the probability that $\bar{x}$ lies in the interval is 0.95.

   False

   The interpretation of the confidence interval is that, if we repeatedly took samples of size 75, the true mean μ would lie in the confidence interval 95% of the time.

9. Suppose we were also interested in the mean of the highly right skewed indicator of iron absorption, ferritin. Compared to the relatively symmetrically distributed indicator hemoglobin, do you think a larger or smaller sample size would be required to apply the central limit theorem?

   Larger - The CLT is an asymptotic result - namely, you need a large sample size for it to work. However, how large is large enough is always an important question in statistics. You need much smaller sample sizes to invoke the CLT when the distribution of the variable of interest is symmetric and “well-behaved,” versus variables with highly skewed or multi-modal distributions (less “well-behaved”).

Hypothesis Testing with known Variance. Now, let's switch gears and assume that we didn't know the true population mean μ, and we only observed a sample of 25 school children in Delhi. Let's use the sample mean from this set of children to make inference about the true population mean μ. Assume that $\sigma = 12.5$ is known and that the sample size of 25 is large enough to use the Central Limit Theorem (i.e. base your inferences off of the normal distribution, not the t-distribution).

In this scenario, we can conduct a **one-sample Z-test** for inference about μ in a population with known variance σ^2. To test $H_0 : \mu = \mu_0$, we can use the test statistic:

$$Z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis, $Z^* \sim N(0, 1)$. So, we can use the standard normal distribution to calculate a p-value for this hypothesis test:

- For the one-sided test with alternative hypothesis $H_a : \mu > \mu_0$, we can calculate a p-value using the formula $p = P(Z > Z^*)$.
- For the one-sided test with alternative hypothesis $H_a : \mu < \mu_0$, we can calculate a p-value using the formula $p = P(Z \leq Z^*)$.
- For the two-sided test with alternative hypothesis $H_a : \mu \neq \mu_0$, we can calculate a p-value using the formula $p = 2 * P(Z < -|Z^*|)$.

Note: there is no command for directly conducting this one-sample Z-test in Stata. However, you can use the normal function in Stata to calculate the p-values.

1. In a sample of size 25, what is the value of the test statistic testing whether the mean hemoglobin level is equal to 108 g/L versus the alternative that it is not equal to 108 g/L, when $\bar{x} = 103$. Use a one-sample Z-test. What is the p-value?

   Test statistic:-2
   p-value: 0.04550026

   $H_0 : \mu = 108, H_A = \mu \neq 108$

   $Z = \frac{103-108}{12.5/\sqrt{25}} = -2$

   $Z \overset{H_0}{\sim} N(0, 1) \rightarrow P(Z < -2) = 0.023$

   ```
   . di normal(-2)*2
   .04550026
   ```

2. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is less than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

Test statistic:-2
p-value: 0.02275013

$H_0 : \mu = 108, H_A = \mu < 108$

$p = 2 * P(Z < -2)$

```
. di normal(-2)
.02275013
```

3. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is greater than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

Test statistic:-2
p-value: .9772498

$H_0 : \mu = 108, H_A = \mu > 108$

$p = P(Z > -2)$

```
. di 1 - normal(-2)
.9772498
```

Confidence intervals and testing with unknown variance. Suppose now that we are interested in the distribution of hemoglobin levels in Mumbai. We decide that it is unreasonable to extrapolate the Delhi results to Mumbai, and therefore the population standard deviation is unknown. We take a random sample of 15 children in Mumbai. The sample mean is $\bar{x} = 115$ g/L, with sample standard deviation $s = 10.2$ g/L. Assume hemoglobin levels in Mumbai are normally distributed (we could check this by looking at the distribution of hemoglobin levels in other similar populations).

1. Construct a two-sided 95% confidence interval for μ.

   (109.56446, 120.43554)

   ```
   . di 115 - invttail(24, 0.025)*10.2/sqrt(15)
   109.56446

   . di 115 + invttail(24, 0.025)*10.2/sqrt(15)
   120.43554
   ```

2. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.05$ level?
   (a) yes (b) no (c) not enough information

   Yes - 108 is not in the confidence interval.

3. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level?
   (a) yes (b) no (c) not enough information

   Not enough information - $\alpha = 0.01$ corresponds to a 99% confidence interval. This interval will be wider than the 95% confidence interval, and may or may not contain 108.

4. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.1$ level?
   (a) yes (b) no (c) not enough information

   Yes - $\alpha = 0.1$ corresponds to a 90% confidence interval, and this interval will always be narrower than the 95% confidence interval. Therefore, we know that 108 will not be in the 90% confidence interval and we can reject the null.

5. Conduct a one-sample t-test in Stata to test the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level.

```
. ttesti 15 115 10.2 108, level(99)

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      15         115    2.633629        10.2    107.1601    122.8399
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   2.6579
Ho: mean = 108                                   degrees of freedom =       14

   Ha: mean < 108               Ha: mean != 108               Ha: mean > 108
 Pr(T < t) = 0.9906         Pr(|T| > |t|) = 0.0187          Pr(T > t) = 0.0094
```

- What is your test statistic?

  $t = 2.658$

- Under the null hypothesis, the test statistic follows a t-distribution with how many degrees of freedom?

  24 degrees of freedom ($n - 1 = 25 - 1 = 24$).

- What is your p-value?

  $p = 0.0187$.

- What do you conclude?

  (a) reject the null hypothesis, (b) accept the null hypothesis, (c) fail to reject the null hypothesis

  Because $p = 0.019 > 0.01 = \alpha$, we fail to reject the null.

Nursing Home Study. Suppose that 500 residents of a large nursing home are screened for hypertension. All residents with above a specified level are labeled as having hypertension. The following table displays the results of this study.

	Hypertension	No Hypertension	Total
Male	100	100	200
Female	100	200	300

1. Which of the following measures of association would be the most appropriate to use to describe the finding in this study?

    a. Cumulative Ratio
    b. Incidence Rate Ratio
    c. Prevalence Ratio

2. What is the prevalence of hypertension among all residents?

    200/500 = 0.40

3. What is the prevalence odds ratio for having hypertension comparing male residents (exposed group) to female residents (non-exposed group)?

    OR = (100/100)/(100/200) = 2.0

4. Which of the following is the **least** likely explanation for the value for the odds ratio reported in the previous answer?

    a. Men have a higher risk of developing hypertension than women.
    b. Women who develop hypertension live longer after being diagnosed than men.
    c. Men who develop hypertension are more likely to be residents of a nursing home than women who develop hypertension
    d. Hypertension is harder to detect in women than in men.

5. Is reverse causation a plausible explanation for the association seen in the study?

    a. Yes
    b. No

Bed Occupancy in Two Hospitals. An investigator performs a bed-occupancy survey on July 1, 2012 at two hospitals (Hospital A and Hospital B). Hospital A has 100 available beds. Hospital B is larger with 200 available beds. Hospital A reports that 80 of its beds are occupied on the day of the survey (occupancy prevalence = 0.80).

1. The investigator reports a Prevalence Odds Ratio of 2.25 when comparing the occupancy prevalence of Hospital B to that for Hospital A. What is the occupancy prevalence for Hospital B on that day?

Prevalence odds in Hospital A:	**80/20=4**
Prevalence odds ratio for Hospital B vs. Hospital A:	**2.25**
Prevalence odds in Hospital B:	**X/4=2.25→X=9**
Occupancy prevalence for Hospital B:	**9/10=90%**

2. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. If the average length of stay is the same in each hospital then what is the incidence rate ratio for admission, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)?

a. 2.25
b. > 2.25
c. < 2.25
d. Cannot be determined from the information that is given

3. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. Furthermore, suppose that the incidence rate ratio for admission is 1.5, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)? Would the average length of stay for patients in Hospital B be?

a. Less than that for Hospital A
b. Equal to that for Hospital A
c. Greater than that for Hospital A
d. Cannot be determined from the information that is given

4. If the investigator presented these results at a local conference, then what should he described as the design of this study?

a. Case Report
b. Ecologic Study
c. **Cross Sectional Study**

Problem Set 6

Power and Sample Size. Recall (from the previous problem set) that according to the WHO Global Database on Anaemia, the mean hemoglobin levels among primary school children in Delhi were estimated at $\mu = 108$ g/L, with standard deviation $\sigma = 12.5$ g/L. Suppose you decide to go and conduct your own study. You aim to test whether hemoglobin levels in Mumbai are similar to the estimates from Delhi. For your calculations, you assume that the standard deviation of hemoglobin levels in Mumbai **is known** and is equal to that in Delhi. We aim to test:

$$H_0 : \mu_M \leq 108$$
$$H_A : \mu_M > 108$$

You plan to conduct this one-sided test at the $\alpha = 0.05$ level of signficance.

1. Which of the following will increase the power of a test: (a) increasing the type-I error, (b) increasing the sample size, (c) all of the above.

   (c)

2. If you randomly sample 30 children, what is the type I error of your test?

   0.05

3. If you randomly sample 30 children, what is the power of your test against the alternative $\mu_M = 112$?

   0.5429

```
. sampsi 108 112, sd1(12.5) n1(30) onesample onesided

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
 alternative m =      112
            sd =     12.5
 sample size n =       30

Estimated power:

         power =   0.5429
```

4. If the true mean levels of hemoglobin were 115 g/L, how many children do you need to sample to have 80% power (assume the standard deviation remains the same as Delhi)?

   20

```
. sampsi 108 115, sd1(12.5) power(.80) onesample onesided

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.8000
 alternative m =      115
            sd =     12.5

Estimated required sample size:

             n =       20
```

5. If the true mean levels of hemoglobin were 112 g/L, how many children do you need to sample to restrict your type II error below 10%?

   84

```
. sampsi 108 112, sd1(12.5) onesample onesided

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
 alternative m =      112
            sd =     12.5

Estimated required sample size:

             n =       84
```

Two Sample t-test. Schiff et al. (1990) investigated the effect of low doses of aspirin on women with mild pregnancy-induced hypertension. Forty-seven women hospitalized at 30-36 weeks' gestation because of mild pregnancy-induced hypertension were treated by a daily dose of either 100 mg aspirin or placebo. The 23 women who received aspirin had a mean arterial blood pressure of 111 mm Hg with a standard deviation of 8 mm Hg. The 24 women who received placebo had a mean arterial blood pressure of 109 mm Hg and a standard deviation of 8 mm Hg.

Source: Schiff, E., Barkai,G., Ben-Baruch G., and Mashiach, S., "Low-Dose Aspirin Does Not Influence the Clinical Course of Women with Mild Pregnancy-Induced Hypertension," Obstetrics and Gynecology, Vol. 76, November 1990, 742-744.

1. Conduct a two sample t-test with equal variances at α = 0.05 to test whether the population mean arterial blood pressure for women treated with aspirin is equal to the population mean arterial blood pressure for women treated with placebo. What is the value of the test statistic?

   Here, $s_p^2 = 64$. Now, $t = \frac{111-109}{\sqrt{\frac{64}{23}+\frac{64}{24}}} = 0.856763$

2. How man degress of freedom does your test statistic have?

   $23 + 24 - 2 = 45$

3. The p-value for this test is greater than 0.05. Based on this result can we conclude that the population mean arterial blood pressure for pregnant women treated with aspirin is equal to the population mean arterial blood pressure for pregnant women treated with placebo? A) Yes B) No

   B) No. We cannot conclude the two are equal. The most we can say is that we do not have enough evidence to conclude that the population mean arterial blood pressure for pregnant women treated with aspirin is different from the population mean arterial blood pressure for pregnant women treated with placebo

4. Based on this result, would you expect the 95% confidence interval for the difference in means to include zero? A)Yes B) No

   A) Yes. Since the p-value was greater than 0.05, the confidence interval will include zero. Remember there is a 1-1 correspondence between hypothesis testing and confidence intervals.

ANOVA. Consider a hypothetical clinical trial to examine the effects of 3 different drug doses (low, high, and placebo) on mean change systolic blood pressure among a random sample of patients with hypertension.

We use Stata to test the null hypothesis at the $\alpha = 0.05$ level that the three groups have equal population mean change in systolic blood pressure. The output (with some entries omitted) are presented below.

```
. oneway spb trt, tabulate

            |          Summary of SPB
        trt |        Mean   Std. Dev.       Freq.
------------+------------------------------------
    Placebo |   21.886666   1.1587349          15
   Low dose |   21.593333   1.1510659          15
  High dose |   23.506667   1.8771053          15
------------+------------------------------------
      Total |   22.328889   1.6413165          45

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      31.8564567      xx    15.9282283      xxx     0.0014
 Within groups      86.6760131      42     2.0637146
------------------------------------------------------------------------
    Total            118.53247      44    2.69391977

Bartlett's test for equal variances:  chi2(2) =   4.5868  Prob>chi2 = 0.101
```

1. What is the value of the F statistic?

   $\frac{15.9282283}{2.0637146} = 7.72$

2. How man degrees of freedom does the numerator of the F statistic have?

   3-1=2

3. Based on this result, we can conclude that we have evidence that there is a significant difference in the population mean change in systolic blood pressure between the high dose group and low dose group. A)True B) False

   B) False. We can only conclude that the three means are not equal. We do not know which pairs are not equal.

4. Now we wish to use a Bonferroni adjustment to compare the mean change in systolic blood pressure between all three pairs of groups. If we wish to set the overall level of significance at 0.05, what α must we use for each individual test?

   $\frac{0.05}{3} = 0.01667$

Experimental Study of Aspirin and Alzheimer's Disease. Suppose that an investigator plans to perform an experimental study on elderly subjects to examine if aspirin decreases the risk of developing Alzheimer's disease. Subjects will be randomized to receive either aspirin or a placebo and they are followed to record the development of this disease.

1. In an experimental study like this, equipoise refers to the fact that each subject has an equal chance of being assigned to either the aspirin or the placebo group.

   a. True
   b. False

2. In general, randomization tends to balance the distribution of factors that might influence the outcome, especially for small studies but unlikely for large studies.

   a. True
   b. False

3. One of the expected results of randomization is that the distributions of the risk factors among the cases of Alzheimer's disease that develop in the study will be approximately the same as among those who do not develop this outcome.

   a. True
   b. False – This question asks you to compare the distribution of risk factors for disease after randomization between those who develop disease and those who do not develop disease. However, randomization only balances the distribution of risk factors for disease between the exposed and non-exposed groups, not between those who develop disease and those who do not develop disease.

4. The purpose of randomizing some of the subjects to receive a placebo pill was to

   a. Blind the participants from knowledge of their treatment assignment so that such knowledge would not influence their compliance and outcomes
   b. Blind the participants from the purpose of the study so that such knowledge would not influence their compliance and outcomes
   c. Investigate if placebo pills decrease the risk of Alzheimer's disease among the elderly

5. Suppose the study used a blocked randomization scheme with a fixed block size. Suppose that the first twelve subjects are assigned to groups in the following order (A = aspirin, P = Placebo)

   A P P A A A P P P A P A

   Which of the following is a possible value for the fixed block size?

   a. Two
   b. Four
   c. Six
   d. None of the above

6. Because of potential side effects from aspirin, the investigator incorporates a run-in phase into the design of this study. The purpose of the run-in phase is to

   a. Collect risk factor information on all potential subjects to identify and enroll only high risk elderly subjects
   b. Give aspirin to all potential study subjects prior to randomization to determine compliance. Non-compliers may be less likely to comply with treatment assignment after randomization and therefore would not be enrolled in the study
   c. Monitor all subjects early after randomization for compliance to their assigned treatment and remove all non-compliers from the study at that point.
   d. Monitor all subjects assigned to the placebo group early after randomization to determine who are purchasing and taking aspirin.

Compliance in Experimental Study of Aspirin and Alzheimer's Disease.
Suppose that 500 subjects were randomized to the aspirin group and another 500 to the placebo. At the end of the study, compliance was assessed by asking all subjects if they took the assigned pill daily. The study reported the following results regarding compliance:

		Randomization Assignment	
		Aspirin	Placebo
Took assigned pill daily	Yes	400 (Group A)	450 (Group B)
	No	100 (Group C)	50 (Group D)

1. Which groups would be compared in an Intention-To-Treat Analysis to measure the effect of aspirin on the rate of developing Alzheimer's disease?

   a. Group A versus Group B
   b. (Group A + Group C) versus (Group B + Group D)
   c. Group A versus (Group B + Group C + Group D)

Problem Set 7 Solutions

Inference for proportions. According to Fergusson *et al.* (2012), acutely ill patients, including neonatal infants, often receive red blood cell transfusions. However, the consequences of the prolonged storage of red blood cells on health outcomes in premature infants are not well understood. In a double-blinded, randomized controlled trial, the authors looked at health outcomes in neonatal infants who underwent red blood cell transfusions, comparing the standard protocol (transfusions of blood stored for prolonged periods) with fresh blood transfusions (transfusions of blood store for less than seven days).

Specifically, the authors examined five outcomes listed below; as well as a composite outcome, defined as at least one of the five outcomes. In this question, we focus primarily on the composite outcome. The results of the study are shown in the following table:

	Standard	**Fresh**
Necrotizing enterocolitis	15	15
Intraventricular hemorrhage	11	18
Retinopathy of prematurity	26	23
Bronchopulmonary dysplasia	63	60
Death	31	30
Composite Outcome	100	99
Sample Size	189	188

A dataset `hw7.dta` (or `hw7.csv`) is also available on the course website, if you would rather not use the "immediate" commands in Stata.

Source: Dean A. Fergusson, MHA, PhD; Paul Hébert, MD, MHSc(Epid); Debora L. Hogan, BScN, BA, MScN; Louise LeBel, BScN; Nicole Rouvinez-Bouali, MD; John A. Smyth, LRCPSI; Koravangattu Sankaran, MBBS; Alan Tinmouth, MD, MSc(Clin Epi); Morris A. Blajchman, MD; Lajos Kovacs, MD; Christian Lachance, MD; Shoo Lee, MBBS, PhD; C. Robin Walker, MB,ChB; Brian Hutton, PhD; Robin Ducharme, HBSc; Katelyn Balchin, MSc; Tim Ramsay, PhD; Jason C. Ford, MD; Ashok Kakadekar, MD; Kuppuchipalayam Ramesh, MD; Stan Shapiro, PhD. (2012). Effect of Fresh Red Blood Cell Transfusions on Clinical Outcomes in Premature, Very Low-Birth-Weight InfantsThe ARIPI Randomized Trial. JAMA.

1. Construct a 95% confidence interval for the proportion of infants experiencing the composite outcome in the fresh red blood cell group, using the following methods:

   a) the exact binomial confidence interval

```
(45.3, 60.0)

. ci outcome if fresh==1, binomial

                                                         -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     outcome |        188     .5265957    .0364146        .4526364    .5997032
```

b) the Wilson confidence interval (which uses the normal approximation)

```
(45.5, 59.7)

. ci outcome if fresh==1, binomial wilson
                                                         ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     outcome |        188     .5265957    .0364146         .455408    .5967184
```

2. Is the normal approximation to the binomial appropriate in this setting?

   Yes.

3. Suppose you wanted to calculate a 95% confidence interval for infants experiencing intraventricular hemorrhage after receiving a fresh blood transfusion as well. Is the Wilson confidence interval still appropriate?

   Yes.

4. Estimate and construct a large-sample 95% confidence interval for the risk difference for experiencing the composite outcome for those with fresh blood versus the standard protocol blood. Calculate the risk difference as estimated proportion in fresh blood group minus estimated proportion in the standard blood group.

   -0.3 (-10.3, 9.8)

```
. prtest outcome, by(fresh)

Two-sample test of proportions                    0: Number of obs =      189
                                                  1: Number of obs =      188
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |   .5291005    .036308                       .4579382    .6002629
           1 |   .5265957   .0364146                       .4552244    .5979671
-------------+----------------------------------------------------------------
        diff |   .0025048   .0514227                      -.0982819    .1032915
             |  under Ho:   .0514228     0.05   0.961
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                                  z =   0.0487
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.5194         Pr(|Z| < |z|) = 0.9612          Pr(Z > z) = 0.4806
```

5. Use a two-sample test of proportions to determine whether there is a difference between fresh and standard groups at the α=0.05 level of significance. What is the test statistic? Null distribution? P-value? Conclusion?

   Z = 0.0487, Z ~ N(0, 1) under the null

   p = 0.9612

   (From risk difference problem above)

   no evidence that the risk difference is different from 0.

Contingency Tables. Continue using the Fergusson et al. (2012) clinical trial data to complete the following questions.

1. Estimate the odds ratio and a 95% confidence interval for experiencing the composite outcome for those with fresh blood versus standard protocol blood. Is there evidence of an association between blood group and the composite outcome (at the 0.05 level of significance).

   1.0 (0.7, 1.5)

   No evidence of an association.

```
. cs outcome fresh, or woolf

                 | fresh                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        99         100  |        199
        Noncases |        89          89  |        178
-----------------+------------------------+------------
           Total |       188         189  |        377
                 |                        |
            Risk |  .5265957    .5291005  |   .5278515
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0025048       |   -.1032915    .0982819
      Risk ratio |          .995266       |    .8222696    1.204659
 Prev. frac. ex. |          .004734       |   -.2046588    .1777304
 Prev. frac. pop |         .0023607       |
      Odds ratio |              .99       |    .6607011    1.483424 (Woolf)
                 +-------------------------------------------------
                               chi2(1) =     0.00  Pr>chi2 = 0.9612
```

2. Construct a 2x2 table for the composite outcome versus blood group. Are the expected cell counts large enough to conduct a Pearson Chi-square test?

   Yes.

3. Using the Pearson chi-square test, determine if there is an association between fresh versus standard blood and the composite outcome at the α=0.05 level of significance. What is your test statistic? Null distribution? P-value? Conclusion?

   X = 0.0024 ~ χ^2_1

   p = 0.961

   fail to reject the null – no evidence of an association.

Final Thoughts.

1. In the previous questions, we looked at three different tests of association: the Pearson Chi-square test, an odds ratio test, and a risk difference test. Are the results of these three tests consistent? Would you expect them to be?

   Yes, and yes.

2. Is there evidence of an association between blood group assignment and the composite outcome?

   No.

3. Think back to the Bonferroni correction from last week. If you were tasked with conducting hypothesis tests comparing the two blood groups for **each** of the 5 different outcomes, would you need to correct for multiple comparisons?

   Yes.

4. In this study, the authors state that they powered the study to detect an absolute difference of 15% in the two groups with 80% power, used a 2-sided test with α =0.05. After a few more adjustments, their final sample size calculation was 450.

   Now suppose you want to replicate the study using a different population. Given that the authors did not find an association in their data, you decide to increase the power and decrease the difference detected between standard and fresh groups. Using an equal number of infants in both groups, what is the total sample size needed in order to achieve 90% power, assuming that the proportion of infants experiencing the composite outcome in the standard group was 55% and 45% in the fresh blood group (again using a 2-sided test with α =0.05).

   1088

```
. sampsi 0.45 0.55

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                    and p2 is the proportion in population 2
Assumptions:

   alpha =   0.0500  (two-sided)
   power =   0.9000
      p1 =   0.4500
      p2 =   0.5500
   n2/n1 =   1.00

Estimated required sample sizes:

      n1 =      544
      n2 =      544
```

5. Consider a covariate, the clinical risk index for babies (CRIB), which was measured in the infants enrolled in the clinical trial. CRIB is usually associated with the composite outcome. From the baseline characteristics table in the Fergusson et al paper, we find that the median and IQR for CRIB is similar between the standard and fresh blood groups. This suggests that the distribution of CRIB is similar in both groups.

   True or False: Because the distribution of CRIB is similar betwen groups, the study investigators would not have gained any power to detect an effect by matching on CRIB score.

   False.

   Matching exploits correlation between CRIB score and outcome to gain power. If CRIB score is highly associated with the outcome, we could gain power by matching.

Randomized Clinical Trial versus Cohort Study

1. The benefit of a randomized clinical trial over an observational cohort study is that, in large enough samples, the groups are identical with respect to

   A. other extraneous factors that are associated with the outcome of interest

   B. factors that would make the results more generalizable to the larger population

   C. other factors related to the likelihood of participating in a study

   Choice A

Toxins and Parkinson's Disease Cohort Study

An investigator, Dr. Park, is interested in evaluating whether there is an association between exposures to toxins and risk of developing Parkinson's disease. She constructs a cohort of men and women that are living in her state in 1985 and every year, they are asked to complete a questionnaire about exposures at work and at home and whether they have been diagnosed with Parkinson's disease. The participants in Dr. Park's study contribute information on their changing toxin exposures over time for as long as they are residents in that state and for any year that they complete the questionnaire. As new people move into the state, they are enrolled in her cohort and remain in the cohort for as long as they are residents in the state and complete the questionnaire.

1. Should Dr. Park include people who reported that they had Parkinson's at the time that they were recruited?

   A. Yes

   B. No

   No, because they are not at risk of developing the disease.

2. Does Dr. Park need to be concerned about selection bias?

   A. Yes, if people who were at greater risk of developing Parkinson's were more likely to be exposed and were also more likely to participate in the study

   B. No, because it is a prospective cohort study so selection biases are not a concern

3. If those who are exposed to toxins are more likely to drop out of the study and more likely to develop Parkinson's disease, what effect with this have on the estimated relative risk compared to the true relative risk?

   A. No effect

   B. Biased towards the null

   C. Biased away from the null

4. Is this an open cohort or closed cohort? Why?

   A. Closed, because exposure at baseline is used for all follow-up

   B. Open, because loss to follow-up and competing risks are still a problem in this population-based study

   C. Closed, because risk ratios are the most appropriate measure to compare toxin levels and Parkinson's risk

   D. Open, because people can enter and exit the cohort over the follow up time of the study

5. True or False: Based on the data collected in her study, Dr. Park will be able to calculate absolute measures of Parkinson's disease incidence.

   True

6. Because Dr. Park conducted a prospective cohort study instead of a cross-sectional study, she can be less concerned about

   A. Confounding

   B. Bias

   C. Reverse causation

   D. Chance

Problem Set 8

Survey Sampling Design. You decide to conduct a survey to measure physical activity in Boston. You plan to collect information on the amount of physical activity per week, history of diabetes, weight, height, age, and gender.

There are seventeen distinct neighborhoods in Boston, with substantial differences in race/ethnicity, socioeconomic status, and population density. You expect to observe variability in physical activity indicators by neighborhood. Unfortunately there is no specific information about within-neighborhood heterogeneity (variance); therefore, you design the survey based on the assumption that variances of indicators are equal across neighborhoods.

The table below displays population data for Boston in 2010. Assume that these population numbers are still accurate today.

Neighborhood	**Population - 2010**
South End	34,669
Central	30,901
Fenway - Kenmore	40,898
South Boston	33,688
Charlestown	16,439
Allston - Brighton	74,997
West Roxbury	30,445
Roxbury	59,790
East Boston	40,508
Jamaica Plain	39,897
Back Bay - Beacon Hill	27,476
Hyde Park	31,813
North Dorchester	28,384
South Dorchester	59,949
Roslindale	32,589
Mattapan	34,616
Harbor Islands	535
Boston	617,594

Source: http://www.bostonredevelopmentauthority.org/PDF/ResearchPublications//PDPercentChange.pdf

Consider the following questions:

1. Suppose you decided to randomly sample 1,700 people from the city of Boston (call this Design 1). For any given individual in South Dorchester, what is the probability of being selected in the survey? What is this probability for an individual in Harbor Islands?

   South Dorchester: 0.002752617

   Harbor Islands: 0.002752617

P(individual i selected) = 1700/617594 = 0.002752617

2. What is likely the main challenge of implementing this survey?

   (a) SRS is inefficient, (b) SRS is difficult to implement in practice, (c) SRS does not necessarily sample people from each neighborhood

3. Now, you randomly sample 100 people within each neighborhood (Design 2). What is the probability of a random individual in South Dorchester being sampled? What is the probability of a randomly selected individual in Harbor Islands being sampled?

   South Dorchester: 0.0017

   Harbor Islands: 0.1869159

   P(individual selected who lives in S. Dorchester) = 100/59949 = 0.001668085

   P(individual selected who lives in Harbor islands) = 100/535 = 0.1869159

4. What kind of survey design is Design 2?

   (a) stratified sample, (b) cluster sample, (c) simple random sample

5. Consider an alternate design (Design 3). In each neighborhood the number of individuals sampled is proportional to the population size of the neighborhood. Assuming that the sample size is fixed at 1,700, would you expect Design 2 or Design 3 to provide more precise estimates?

   (a) Design 2, (b) Design 3

   Design 2 oversamples people in the small neighborhoods. By sampling fewer people in the smaller neighborhoods and more people in the larger neighborhoods, our design is closer to a simple random sample and is more efficient.

6. Once again, consider the probability of a random individual in South Dorchester being sampled; and the probability of a random individual in Harbor Islands being sampled. These probabilities are approximately the same as those in:

   (a) Design 1, (b) Design 2

7. Why might you want to use Design 2 compared to Design 3?

   (a) to increase precision, (b) if you wanted neighborhood-specific estimates, (c) both a and b

Recall Design 2 results in a higher variability of survey weights due to a disproportionately large number of individuals selected in smaller neighborhoods (decreasing precision). By oversampling within small neighborhoods, we could obtain neighborhood specific estimates.

8. Next, you decide you do not want to visit all 17 neighborhoods, so you randomly sample 10 neighborhoods. Within each neighborhood selected, you randomly sample 170 people. Call this Design 4. What is the probability of a random individual in South Dorchester being included in the survey? What is the probability of a random individual in Harbor Islands being included in the survey?

   South Dorchester: 0.001668085

   Harbor Islands: 0.1869159

   P(individual selected who lives in S. Dorchester)

   = P(S. Dorchester selected)P(individual selected|lives in S. Dorchester)

   = (10/17)(170/59949) = 0.001668085

   P(individual selected who lives in Harbor Islands) =

   = P(Harbor Islands selected)P(individual selected|lives in Harbor Islands)

   = (10/17)(170/535) = 0.1869159

9. A "self-weighting" design is a survey design for which every individual in the population has an equal probability of inclusion. Which survey designs are self-weighting (or approximately self-weighting)?

   (a) Design 1, (b) Design 2, (c) Design 3, (d) Design 4

10. Consider yet another design (Design 5), in which you again select 10 neighborhoods. Now, the probability of a neighborhood being included in the survey is proportional to its population size. Within each sampled neighborhood, you randomly sample 170 people. Would you expect Design 4 or Design 5 to provide more precise estimates?

    (a) Design 4, (b) Design 5

    To build some intuition, think about Harbor Islands. The probability of including a bunch of people from Harbor Islands is much higher in Design 4 compared to Design 5. Oversampling from Harbor Islands decreases our precision, because Harbor Islands represents such a small fraction of the population of Boston.

Other aspects of survey design. Thus far, we have examined sampling design, which is only one element of designing a survey. Building and testing the survey instrument/questionnaire, anticipating and preparing for non-response, and training field teams to conduct the survey are several other critical aspects that we have not even touched on! Consider the following:

1. When designing your survey, you are trying to decide whether to use a 2-page questionnaire with 15 simple questions about whether an individual exercises, how many times per week, and whether they have a history of diabetes, along with some other very basic demographics (call this Survey 1). Your colleague says you could get much better information if you used a 10 page questionnaire with a more complete medical history and history of exercise and physical activity (call this Survey 2).

   Krosnick (1991) discusses "satisficing" (satisfy + suffice) in surveys. To paraphrase this discussion, satisficing occurs when, rather than optimizing their responses to best reflect reality, survey respondents try to reduce the cognitive burden associated with the survey and consequently may select a survey response haphazardly or even arbitrarily.

   Which survey would be more susceptible to satisficing?

   (a) Survey 1, (b) Survey 2

   Source: Krosnick, J. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5: 236.

2. You decide to implement a door-to-door survey, where you randomly sample addresses within a neighborhood and train a field team to ask the survey questions at the selected households. (You have a complete listing of addresses in Boston).

   Your colleague says you should obtain a listing of all land-line telephone numbers in Boston (a listing of cell phone numbers is not available) and randomly sample numbers from this list and ask the questions over the phone. For sake of argument, assume that you get a 100% response rate for both modes (door to door and phone calls), and that you construct survey weights using the table above.

   Would you expect the door-to-door or land-line method to produce unbiased results? (Think about which sampling frame is likely to be more complete.)

   (a) Door-to-door, (b) landline

   Would you expect the door-to-door or land-line method to be cheaper to implement?

   (a) Door-to-door, (b) landline

Note: web-based surveys are also common. In order to minimize non-response, some surveys use multiple modes of response and follow-up (e.g. web + phone; or web + household follow-up).

3. Individuals can opt out of any part of your survey. High BMI and low physical activity are risk factors for diabetes, and you find that individuals with these characteristics are less likely to answer questions about history of diabetes (note that high BMI and low physical activity are risk factors for diabetes). In a complete case analysis, missing data is dropped, survey weights are recalculated, and data is analyzed assuming missing observations were never collected. In a complete case analysis, would you expect to obtain unbiased estimates of diabetes prevalence in Boston?
   (a) yes, (b) no

Case Control versus Cohort Study

1. One of the main advantages of case-control studies over cohort studies is that

   A. The investigator can identify exposed and unexposed people without concerns about selection bias.

   B. **The investigator does not need to wait a long time for cases to occur.**

   C. The investigator does not need to be concerned about issues of confounding.

   D. The investigator can examine whether a third factor modifies the relationship between exposure and outcome.

Alcohol Consumption and Cancer Study

Dr. Marks is interested in examining the association between alcohol consumption and a rare form of cancer. Since it is hard to identify cases, Dr. Marks designs a case-control study. Cases are identified from cancer treatment centers across the United States. Controls are selected on the day that the case is diagnosed with cancer from among the relatives of the cancer patient.

1. True or False: Dr. Marks should make sure to select the healthiest relative as the control to ensure that he will observe differences in risk between cases and controls.

   False: the controls are meant to represent the exposure distribution in the population giving rise to the cases, not to represent a level of outcome risk.

2. The main concern of using relatives of cases as the controls is that

   A. Some cases may not have relatives that drink alcohol.

   B. Some cases may have relatives that live far away so not all relatives will be available to participate.

   C. **Relatives of cases may be more likely to have levels of alcohol consumption that are more similar to the cases than the population that gave rise to the cancer cases.**

   D. The controls may develop other diseases and will not be available to participate in this study.

3. Based on the study description above, Dr. Marks conducted a

   A. **Density case control study**

   B. Nested case-control study

   C. Case-cohort study

   D. Retrospective cohort study

4. True or False: Using this design, a relative who served as a control cannot be included as a case if he later develops the cancer of interest.

   False

5. True or False: Using the data collected in this study, Dr. Marks will be able to estimate the rates for developing this cancer.

   False

Problem Set 2 Solutions

BMI and CHD prevalence. The following table uses data from the NHLBI teaching data set and displays categories of body mass index for 4,415 participants in the Framingham Heart Study attending an examination in 1956 with non-missing values for body mass index. For each body mass index category, the table displays the number of subjects with existing Coronary Heart Disease (CHD) at that exam (**prevchd=1**)

Body Mass Index Category		Number of Subjects at 1956 Exam	Cases of CHD Diagnosed Prior to 1956
Under Weight	BMI < 18.5	57	0
Normal Weight	18.5 ≤ BMI < 25	1936	66
Overweight	25 ≤ BMI < 30	1848	90
Obese	BMI ≥ 30	574	38
Total		4415	194

1. What is the prevalence of obesity among the 4415 participants at the 1956 exam?

   574/4415 = 0.1300

2. What is the prevalence of CHD at the 1956 exam among the 4415 participants at the 1956 exam?

   194/4415 = 0.0439

3. What is the prevalence of CHD at the 1956 exam for each of the body mass index classes?

Under Weight Participants	**0/57 = 0.0**
Normal Weight Participants	**66/1936 = 0.0341**
Overweight Participants	**90/1848 = 0.0487**
Obese Participants	**38/574 = 0.0662**

Diabetes prevalence. Use Stata and the NHLBI data set to calculate the prevalence of diabetes among participants who attended and had non-missing data on diabetes at all three examinations. (**Hint: There were 3,206 such participants.**)

1. What is the prevalence of diabetes at the first exam (**diabetes1=1**)?

   58/3206 = 0.0181

2. What is the prevalence of diabetes at the second exam (**diabetes2=1**)?

   105/3206 = 0.0328

3. What is the prevalence of diabetes at the third exam (**diabetes3=1**)?

   251/3206 = 0.0783

BMI and hypertension prevalence. Use Stata and the BMI1 variable in the NHLBI data set to create the four categories of body mass index as defined in the first question.

1. What is the prevalence of hypertension (**prevhyp1=1**) at the 1956 exam for each of the body mass index classes?

Under Weight Participants	**6/57 = 0.1053**
Normal Weight Participants	**398/1936 = 0.2056**
Overweight Participants	**683/1848 = 0.3696**
Obese Participants	**336/574 = 0.5854**

Hypertension and high blood pressure. Use Stata to create a binary variable (**highbp1**) to represent the presence/absence of high blood pressure at the 1956 examination

```
gen highbp1=.
replace highbp1=1 if (sysbp1>=140 | diabp1 >= 90)
replace highbp1=0 if (sysbp1<140 & diabp1 < 90)
```

(**Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then they should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 $\geq$ 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 $\geq$ 90 then highbp1 =1, otherwise highbp1 should be missing.)**

1. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants with high blood pressure at the 1956 exam (highbp1=1)?

   106/1619 = 0.0655

2. What is the prevalence of CHD (prevchd1=1) at the 1956 exam for participants without high blood pressure at the 1956 exam (highbp1=0)?

   88/2815 = 0.0313

Hypothetical life table. The table below lists the number of individuals alive at age *x*, for a hypothetical population in 1950-1952 and 1990-19992.

Number of Survivors out of 100,000 Live Births

Age	1950-1952	1990-1992
0	100,000	100,000
20	73,412	96,902
40	56,884	92,638
70	31,744	79,873

1. What is the probability of surviving from birth to age 20 in 1950-1952?

   0.73412

   73412/10000 = 0.73412

2. What is the probability of surviving from age 40 to age 70 in 1990-1992?

   0.86221

   79873/92638 = 0.86221

3. Define the absolute survival increase over the 40 year span as $p_1 - p_2$, where p_1 is the chance of surviving from age x to age x+n in 1990-1992 and p_2 is the chance of surviving from age x to age x+n in 1950-1952. Which age group has the greatest absolute survival increase?

   (a) 0 – 20 , (b) 20 – 40, **(c) 40 – 70**

   Survival: 1950 – 1952

   0 – 20: 73412/100000 = 0.73412
   20 – 40: 56884/73412 = 0.7748597
   40 – 70: 31744/56884 = 0.558048

   Survival: 1990 – 1992

   0 – 20: 96902/100000 = 0.9559968
   20 – 40: 92638/96902 = 0.9559968
   40 – 70: 79873/92638 = 0.8622056

   Absolute Survival Increase:

   0 – 20: (0.9559968 - 0.73412) = 0.239
   20 – 40: (0.9559968 - 0.7748597) = 0.181
   40 – 70: (0.8622056 - 0.558048) = 0.304

4. Define the relative survival increase over the 40 year span as $(p_1-p_2)/p_2$, where p_1 is the chance of surviving from age x to age x+n in 1990-1992 and p_2 is the chance of surviving from age x to age x+n in 1950-1952. Which age group has the greatest relative survival increase?

   (a) 0 – 20 , (b) 20 – 40, **(c) 40 – 70**

   Relative Survival Increase:

   0 – 20: (0.9559968 - 0.73412)/0.73412 = 0.3022351
   20 – 40: (0.9559968 - 0.7748597)/0.7748597 = 0.2337676
   40 – 70: (0.8622056 - 0.558048)/0.558048 = 0.5450384

Problem Set 3

Probability and BMI. The following table uses data from the NHLBI teaching data set and displays the body mass index for 3,909 participants in the Framingham Heart Study with BMI measurements at the first two exams, in 1956 and in 1962.

BMI category	BMI ≤ 25, exam 2	BMI > 25, exam 2	Total
BMI ≤ 25, exam 1	1,492	278	1,770
BMI > 25, exam 1	249	1,890	2,139
Total	1,741	2,168	3,909

Assume a study participant has been randomly selected from this subset of 3,909 participants.

- Define **A** as the event that this participant has a high BMI at exam 1.
- Define **B** as the event that this participant has a high BMI at exam 2.
- Define **C** as the event that this participant has a low BMI at exam 2.

1. What is the probability of A?

   0.547

   2139/3909 = .54719877

2. What is the probability of B?

   0.555

   2168/3909 = .55461755

3. What is the probability of A and B?

   0.483

   1890/3909 = .48349962

4. Are A and B independent?

   No.

   P(A)P(B) = .54719877*.55461755 = .30348604 ≠ P(A and B) = .48349962

5. In a randomly selected participant, what is the probability that A and/or B occurs (namely, that the participant's BMI is high during at least one of the first two exams)?
   0.618

   P(A and/or B) = P(A) + P(B) – P(A and B) = .54719877 + .55461755 - .48349962 = .6183167

6. What is the probability that B occurs, given that A occurs?

   0.884

   P(B|A) = P(A and B)/P(A) = . 48349962/.54719877 = .88359047

7. What is the probability that C occurs, given that A occurs?

   0.116

   P(C|A) = 1 – P(B|A) = .11640953

Probability of age and smoking events. The following table uses data from the NHLBI teaching data set using the 4,434 participants in the Framingham Heart Study who attended the exam in 1956. Below, we show a table that contains the probabilities that a study participant falls into one of the eight categories defined by all possible combinations of the age categories and smoking status at the first exam in 1956. Recall that all study participants are between 30 and 70 years old.

Let A be the event that a randomly chosen study participant is a smoker at exam 1 in 1956.

Let B be the event that the person chosen is between 60 and 70 years old at exam 1.

Age, exam 1	Smoker, Exam 1	
	No	Yes
30-39	0.0519	0.0742
40-49	0.1554	0.2262
50-59	0.1809	0.1346
60-70	0.1200	0.0568

1. Eight categories representing age and smoking status groups are shown in the table above. Are these groups:

   (a) mutually exclusive, (b) exhaustive, (c) both

   Someone can only belong to one age group and one smoking status at any given time. Therefore, the eight categories are mutually exclusive. The sum of the probabilities in the above table is equal to 1, implying that these events are exhaustive (we also know that all participants in the study are between 30 and 70 years old and are either smokers or non-smokers, again implying the events are exhaustive).

2. What is the probability of A?

   0.4918

   P(A) = 0.0742 + 0.2262 + 0.1346 + 0.0568 = .4918

3. What is the probability of B^C, the complement of B?

   0.8232

   P(B^C) = 1 – P(B) = 1 - 0.1200 - 0.0568 = 0.8232

4. What is the probability that a randomly selected individual is a non-smoker who is younger than 60 years old at exam 1?

   0.3882

   P(event) = 0.0519 + 0.1554 + 0.1809 = 0.3882

5. Are the events A and B independent?

   No

   P(A and B) = 0.0568 ≠ P(A)P(B) = 0.4918*(1-0.8232) = 0.08695024

Diagnostic Testing. Screening for prostate cancer in men is a controversial topic. One of the most common screening mechanisms is the PSA test (prostate antigen test). In a meta-analysis, Mistry and Cable (2003) report that the sensitivity of the PSA test is 72.1% and the specificity is 93.2%. In the United States, it is estimated that 16.1% of men will have prostate cancer at some point in their life (America Cancer Society 2012). Assume that the prevalence of prostate cancer among men ages 75 and older is 16.1%. We examine the properties of the PSA screening test in men ages 75 and older, using the sensitivity and specificity values above.

Mistry K. and Cable G. (2003). Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *The Journal of the American Board of Family Practice*, 16(2): 95-101.

1. What is the probability of a false negative test result?

   0.279

   P(T-|D+) = 1 – P(T+|D+) = 1 – sensitivity = 1 – 0.721 = 0.279

2. What is the probability of a false positive result?

   0.068

   P(T+ | D-) = 1 – P(T-|D-) = 1 – specificity = 1 – 0.932 = 0.068

3. What is the probability that a randomly selected man who is 75 years or older DOES NOT have prostate cancer, given that his PSA screening was positive?

   0.330

   $$P(D^+|T^+) = \frac{P(D^+)P(T^+|D^+)}{P(D^+)P(T^+|D^+) + P(D^-)P(T^+|D^-)}$$

   $P(D^+) = 0.161 \rightarrow Pr(D^-) = 1 – 0.161 = 0.839$

   $P(T^+|D^+)$ = sensitivity = 0.721, $P(T^-|D^-)$ = specificity = 0.932 , $P(T^+|D^-)$ = 1 – specificity = 0.068

   $P(D^+)\ P(T^+|D^+) = 0.116081$

   $P(D^-)\ P(T^+|D^-) = 0.057052$

   $$P(D^+|T^+) = \frac{0.116081}{0.116081 + 0.057052}$$

   $P(D^+|T^+) = 0.67047299$

   $P(D^-|T^+) = 1 - P(D^+|T^+) = 0.329527$

4. What is the probability that a randomly selected man who is 75 years or older has prostate cancer, given that his PSA screening was negative?

   0.05432434

   $$P(D^+|T^-) = \frac{P(D^+)P(T^-|D^+)}{P(D^+)P(T^-|D^+) + P(D^-)P(T^-|D^-)}$$

   $$P(D^+|T^-) = \frac{0.161 * (1 - 0.721)}{0.161 * (1 - 0.721) + 0.839 * 0.932}$$

   $$P(D^+|T^-) = 0.05432434$$

Titanic Survival. The following table describes the survival status of passengers on the Titanic, stratified by Passenger Class (First, Second, or Third), Sex/Age (Child, Women, or Man), and Survival Status. The Frequency column indicates the number of passengers in each stratum. (For example there were 4 1st class women passengers who did not survive and 140 1st class women passengers who did survive). These data were obtained from the website anesi.com and refers to British Parliamentary Papers, Shipping. Casualties (Loss of the Steamship "Titanic"), 1912. cmd 6352 "Report of a Formal Investigation into the circumstances attending the foundering on the 15th April 1912 of the British Steamship "Titanic" of Liverpool after striking ice in or near Latitude 41 46 N., Longitude 50 14 W., North Atlantic Ocean, whereby loss of life ensued (London; His Majesty's Stationary Office, 1912) page 42.

Passenger Class	Age/Sex	Survival Status	Frequency
First	Child	Survived	6
First	Child	Did Not Survive	0
First	Women	Survived	140
First	Women	Did Not Survive	4
First	Man	Survived	57
First	Man	Did Not Survive	118
Second	Child	Survived	24
Second	Child	Did Not Survive	0
Second	Women	Survived	80
Second	Women	Did Not Survive	13
Second	Man	Survived	14
Second	Man	Did Not Survive	154
Third	Child	Survived	27
Third	Child	Did Not Survive	52
Third	Women	Survived	76
Third	Women	Did Not Survive	89
Third	Man	Survived	75
Third	Man	Did Not Survive	387

Use these data to calculate the cumulative incidence of surviving for each of the following groups of individuals:

1. All Women 296/402 = 0.74

2. All Children 57/109 = 0.52

3. All Women or Children 353/511 = 0.69

4. All First Class Passengers 203/325 = 0.62

BMI and Cumulative Incidence. The following table uses data from the NHLBI teaching data set and displays categories of body mass index (used in the previous homework assignment) for 4,415 participants in the Framingham Heart Study attending an examination in 1956 with non-missing values for body mass index. For each body mass index category, the table displays the number of subjects who died (**death=1**) during follow-up and the total person-years of follow-up (**timedeath**) until death or the end of the follow-up period (24 years). Assume that all deaths and time of death were recorded among the 4415 participants.

Body Mass Index Category		Number of Subjects at 1956 Exam	Number of Deaths	Total Person-Years
Under Weight	BMI < 18.5	57	18	1181.44
Normal Weight	18.5 ≤ BMI < 25	1936	571	40708.74
Overweight	25 ≤ BMI < 30	1848	691	37728.41
Obese	BMI ≥ 30	574	257	11254.52
Total		4415	1537	90873.11

1. What is the cumulative incidence of death among the 4415 participants at the 1956 exam?

   1537/4415 = 0.3481

2. What is the cumulative incidence of death during the 24 years of follow-up for each of the body mass index class?

   Under Weight Participants 18/57 = 0.3158

   Normal Weight Participants 571/1936 = 0.2949

   Overweight Participants 691/1848 = 0.3739

   Obese Participants 257/574 = 0.4477

3. What is the incidence rate of death among the 4415 participants during the 24 years of follow-up? (Express your answer as #deaths/(1000 person-years))

   1537/90873.11 = 16.9137 (deaths/(1000 person-years))

4. What is the incidence rate of death during the 24 years of follow-up for each of the body mass index classes? (Express your answers as #deaths/(1000 person-years))

   Under Weight Participants 18/1181.44 = 15.2356 (deaths/(1000 person-years))

   Normal Weight Participants 571/40708.74 = 14.0265 (deaths/(1000 person-years))

   Overweight Participants 691/37728.41 = 18.3151 (deaths/(1000 person-years))

   Obese Participants 257/11254.52 = 22.8353 (deaths/(1000 person-years))

BMI and CHD Incidence. Use Stata and the NHLBI data set to create a separate variable for each of the four categories of body mass index as defined in the previous question (for example, create a variable called “underwt” that equals 1 if a person’s BMI was less than 18.5 and 0 if a person’s BMI was ≥18.5). To create the separate variables for each of the four categories of body mass index, use the BMI1 variable in the NHLBI dataset. What is the incidence rate for developing CHD (anychd=1) during the 24-years of follow-up for participants in each of the body mass index categories? (Express your answers as #deaths/(1000 person-years)) Hint: Number of years a person was followed for CHD is recorded in the “timechd” variable in the NHLBI dataset.

1. Under Weight Participants 6/(1119.57py) = 5.36/(1000py)
2. Normal Weight Participants 415/(37324.05py) = 11.12/(1000py)
3. Overweight Participants 596/(32863.49py) = 18.14/(1000py)
4. Obese Participants 218/(9387.74) = 23.22/(1000py)

High Blood Pressure and CHD. Use Stata and the NHLBI data set to create the two categories of high blood pressure (**highbp1**).

```
generate highbp1=.
replace highbp1=1 if (sysbp1>=140 | diabp1 >= 90)
replace highbp1=0 if (sysbp1<140 & diabp1 < 90)
```

(**Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then they should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 $\geq$ 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 $\geq$ 90 then highbp1 =1, otherwise highbp1 should be missing.)**

1. What is the incidence rate for developing CHD (**anychd=1**) during the 24-years of follow-up for participants in each of the blood pressure categories? (Express your answers as #deaths/(1000 person-years))

   Participants with high blood pressure at the 1956 exam (highbp1=1)

   605/(25540.74py) = 23.69/(1000py)

   Participants without high blood pressure at the 1956 exam (highbp1=0)

   635/(55384.42py) = 11.47/(1000py)

Problem Set 4 Solutions

Titanic Survival Risk Ratios. The following table describes the survival status of passengers on the Titanic, stratified by Passenger Class (First, Second, or Third), Sex/Age (Child, Women, or Man), and Survival Status. The Frequency column indicates the number of passengers in each stratum. (For example there were 4 1st class women passengers who did not survive and 140 1st class women passengers who did survive). These data were obtained from the website anesi.com and refers to British Parliamentary Papers, Shipping. Casualties (Loss of the Steamship "Titanic"), 1912. cmd 6352 "Report of a Formal Investigation into the circumstances attending the foundering on the 15th April 1912 of the British Steamship "Titanic" of Liverpool after striking ice in or near Latitude 41 46 N., Longitude 50 14 W., North Atlantic Ocean, whereby 1oss of life ensued (London; His Majesty's Stationary Office, 1912) page 42.

Passenger Class	Age/Sex	Survival Status	Frequency
First	Child	Survived	6
First	Child	Did Not Survive	0
First	Women	Survived	140
First	Women	Did Not Survive	4
First	Man	Survived	57
First	Man	Did Not Survive	118
Second	Child	Survived	24
Second	Child	Did Not Survive	0
Second	Women	Survived	80
Second	Women	Did Not Survive	13
Second	Man	Survived	14
Second	Man	Did Not Survive	154
Third	Child	Survived	27
Third	Child	Did Not Survive	52
Third	Women	Survived	76
Third	Women	Did Not Survive	89
Third	Man	Survived	75
Third	Man	Did Not Survive	387

1. Use these data to calculate the Risk Ratio for surviving comparing "women or children" as the exposed group and "all other passengers" as the non-exposed group.

   RR = (353/511) / (146/805) = 3.81

2. Repeat this calculation for each passenger class.

   First Class: RR = (146/150) / (57/175) = 2.99

   Second Class: RR = (104/117) / (14/168) = 10.67

   Third Class: RR = (103/244) / (75/462) = 2.60

Incidence Rate Ratio Blood Pressure and CHD. The following table uses data from the NHLBI teaching data set and displays the blood pressure distribution for 4,434 participants in the Framingham Heart Study attending an examination in 1956. For each blood pressure category, the tables displays the number of subjects with existing Coronary Heart Disease (CHD) at that exam (Prevalent Cases of CHD) and also the number of new cases of CHD and the total amount of person-years of follow-up that was observed during a 24 year follow-up period **for those subjects who did not have CHD at the 1956 exam**. Follow-up for each subject began in 1956 and ended with the development CHD (fatal or non-fatal), death from another cause, loss to follow-up, or the end of the follow-up period (whichever came first).

Blood Pressure Category		Number of Subjects at 1956 Exam	Prevalent Cases of CHD	Number Developing CHD During Follow-up	Total Years of Follow-up
I	< 140 and DBP <90	2815	88	547	55,384.42
II	140 ≤ SBP < 160 or 90 ≤ DBP < 95	781	39	214	13,191.79
III	SBP ≥ 160 or DBP ≥95	838	67	285	12,348.94

1. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Groups II or III combined (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (499/ 25540.73py) / (547/55384.42py) = 1.98

2. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Group III (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (285/12348.94py) / (547/55384.42py) = 2.34

3. What is the Incidence Rate Ratio for developing CHD for participants in Blood Pressure Group II (exposed group) compared to participants in the Blood Pressure Group I (non-exposed group)

   Rate Ratio = (214/13191.79py) / 547/384.42py) = 1.64

Risk Ratios, Odds Ratios, Rate Ratios for BMI, Death, and CHD. The following table displays categories of body mass index for the participants at the 1956 exam. (Note: 19 of the 4434 participants are excluded from this table because of missing data on bmi1.)

BMI Category	Range of BMI	Frequency
Underweight	0 ≤ bmi1 < 18.5	57
Normal Weight	18.5 ≤ bmi1 < 25	1936
Overweight	25 ≤ bmi1 < 30	1845
Obese	bmi1 ≥ 30	577

Use Stata to perform the following calculations.

Hint: All of the following questions ask you to compare obese subjects to normal weight subjects. Create a new binary variable using bmi1 which equals 1 if the person is obese and 0 if the person is normal weight. Anyone who is underweight or overweight should be missing a value for the new binary variable you create.

1. Calculate the 24-year Risk Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   RR = (259/577) / (571/1936) = 1.52

2. Calculate the 24-year Odds Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   OR = (259/318) / (571/1365) = 1.95

3. Calculate the 24-year Rate Ratio for death comparing obese subjects (exposed group, n=577) to normal weight subjects (non-exposed group, n=1936).

   RR = (259/11308.90py) / (571/40708.74py) = 1.63

4. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing obese subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Hint: Use the if/in tab options to restrict the sample to those without prevalent CHD at the 1956 exam (prevchd1 = = 0).

   Rate Ratio = (180/9387.74py) / (349/37324.05py) = 2.05

5. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing (obese or overweight) subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Rate Ratio = (686/42251.22py) / (349/37324.05py) = 1.74

6. Calculate the 24-year Rate Ratio for developing coronary heart disease comparing underweight subjects (exposed group) to normal weight (non-exposed group), excluding subjects with prevalent CHD at the 1956 exam.

   Rate Ratio = (6/1119.57py) / (349/37324.05py) = 0.57

Please use this information for the remainder of the assignment.

According to data from the CDC in 2010, 19.3% of adults age eighteen and older smoke cigarettes. In the year 2008, the incidence rate of lung cancer was 65.1 cases per 100,000 people per year.

Suppose you are conducting a lung cancer study in the United States, and you obtain a random sample of 2,000 adults (over 18 years of age) who do not have lung cancer. You plan to follow this study cohort over a period of 5 years and observe incident cases of lung cancer.

Smoking and the binomial distribution. Smoking status is an important predictor of lung cancer incidence. Therefore, as the study designer, it is important to think about baseline smoking rates in your study cohort. We first model the number of smokers in the study cohort using the binomial distribution, and assume that this cohort is representative sample from the US population. Use the binomial distribution to answer the parts below:

1. How many smokers would you expect to see in the study cohort, on average?

   386

   Let X denote number of smokers in the study cohort, X ~ Binomial(2000, 0.193).

   E(X) = n*p = 2000*0.193 = 386.

2. What is the standard deviation of the number of smokers in the study cohort?

   17.6

   Var(X) = n*p*(1-p) = 2000*0.193*(1-0.193) = 311.502

   sd(X) = √311.502 = 17.64942

3. What is the probability that you observe exactly 386 smokers?

   0.023

   ```
   . di binomialp(2000, 386, 0.193)
   ```

   ```
   .0225986
   ```

4. What is the probability that greater than or equal to 25% of the study population are smokers?

0.000

```
. di binomialtail(2000, 2000*0.25, 0.193)
```

```
2.403e-10
```

5. What is the probability that less than or equal to 20% of the study population are smokers?

   0.795

   ```
   . di binomial(2000, 2000*0.20, 0.193)
   ```

   ```
   .79487415
   ```

Smoking and the normal distribution. In the questions above, we modeled the number of smokers in the study cohort using the binomial probability model. Now, assume that the number of smokers in the study cohort follows a **normal distribution.**

1. What is the probability that you observe exactly 386 smokers? IMPORTANT !!!

   0

   Because the normal distribution is continuous, the probability of observing exactly 386 smokers is always 0.

2. What is the probability that less than 20% of the study population are smokers?

   0.786

   $X \sim \text{Normal}(386, 311.5)$

   $P(X < 2000*0.2) = P(X < 400)$

   $= P(Z < (400-386)/17.6)$

   $= P(Z < 0.795)$

   ```
   . di normal(0.795)
   ```

   ```
   .78669325
   ```

3. Do you think the normal model provides a reasonable approximation to the binomial model in this example?

   Yes

Lung cancer and the binomial distribution. You also need to carefully consider how many cases of lung cancer you expect to observe in your study over time. We first model the number of lung cancer cases observed in the first year using the **binomial distribution.**

1. What proportion of the study population would you expect, on average, to be diagnosed with lung cancer in the first year?

   0.000651

   65.1 cases/100,000 person-years = 0.000651

2. How many cases of lung cancer would we expect to observe in the first year?

   1.3

   X ~ Binomial(2000, 0.000651)

   E(X) = 2000*0.000651 = 1.302

3. What is the variance of the number of lung cancer cases observed in the first year?

   1.3

   Var(X) = 2000*0.000651*(1-0.000651) = 1.30152

4. Why would you expect the mean and variance to be similar in this example?

   (a) because the event is rare, (b) because the mean is close to 1, (c) because we are dealing with incidence rates, (d) both (a) and (b)

   Recall that the Poisson distribution is a good approximation to the binomial distribution when the event is rare. Further, the mean and variance of the Poisson distribution are the same. Therefore, when the event is rare, we would expect the variance and the mean to be approximately equal.

5. What is the probability that you observe more than 1 lung cancer case in the first year?

   0.374

   ```
   . di binomialtail(2000, 2, 0.000651)
   ```

   ```
   .37392011
   ```

6. What is the probability that you observe no lung cancer cases in the first year?

   0.272

   ```
   . di binomial(2000, 0, 0.000651)
   ```

   ```
   .27187198
   ```

Lung cancer and the Poisson distribution. Because lung cancer is a rare disease, we can model cases of lung cancer using the **Poisson distribution**, with incidence rate 65.1 cases per 100,000 person-years.

1. Using the Poisson distribution, what is the probability that you observe more than 1 lung cancer case in the first year?

   0.374

   ```
   . di poissontail(1.3, 2)
   ```

   ```
   .37317688
   ```

2. What is the expected number of lung cancer cases observed over the five year study period?

   IMPORTANT !!!

   6.51

   Y ~ Poisson(u)

   E(Y) = 2000*0.000651*5 = 6.51

3. What is the variance of the number of lung cancer cases observed over the five year study period?

   6.51 (same as part (p))

   var(Y) = E(Y) because Y ~ Poisson(6.51).

4. What is the probability that you observe more than 10 lung cancer cases over the five year period?

   0.067

   ```
   . di poissontail(6.51, 11)
   ```

   ```
   .06739806
   ```

5. What is the probability that you observe less than 5 lung cancer cases over the five year period?

   0.223

   ```
   . di poisson(6.51, 4)
   ```

   ```
   .22255574
   ```

The big picture. As the study designer, you can compile these facts together, along with the goals of the study, to decide if the sample size of 2,000 individuals is large enough and if the follow up period of 5 years is long enough. If you (or the funding committee) decides that the current study is not big enough to answer the questions of interest about incident lung cancer cases, you either need to find more funding to extend your study or increase your sample size, OR you will have to abandon the study.

In this assignment, we have learned about properties of common probability models in statistics. Namely, we learned that:

1. The _______ provides a good approximation to the binomial distribution when the event of interest is **rare** and the study population is large.

   [✓ (a) Poisson], (b) normal, (c) exponential

2. The _______ provides a good approximation to the binomial distribution when the event of interest is **not rare** and the study population is large.

   (a) Poisson, [✓ (b) normal], (c) exponential

Problem Set 5 Solutions

Conceptual Questions: Predictive and Confidence Intervals

1. True or False: Consider a random variable X. To construct a 95% predictive interval for X, all we need to know is the sampling distribution of X.

   True - predictive intervals use the distribution of a random variable X to make statements about where we expect the values of the random variables to lie.

2. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% predictive interval for the sampling mean $\bar{X} = \sum_{i=1}^{n} X_i$, all we need to know is the sampling distribution of X.

   True

3. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% confidence interval for the true mean of X, denoted μ, all we need to know is the Central Limit Theorem.

   False - the CLT is an asymptotic result. We also need to know n, the sample size!

Central Limit Theorem and Confidence Intervals According to the WHO Global Database on Anaemia, the mean hemoglobin levels among primary school children in Delhi were estimated at $\mu = 108$ g/L, with standard deviation $\sigma = 12.5$ g/L.

Source: `http://who.int/vmnis/anaemia/data/database/countries/ind_ida.pdf`

Suppose we took a random sample of 75 primary school children in Delhi. Denote the mean hemoglobin levels in this sample as $\bar{x}$. Throughout this question, assume that the sample size is large enough that the central limit theorem is applicable and that σ is known.

1. What is the expected value (mean) of $\bar{x}$?

   108

2. What is the standard deviation of $\bar{x}$?

   $12.5/\sqrt{75} = 1.443376$

3. Suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

   0.8341433

   $P(106 < \bar{x} < 110) = P\left(\frac{106-108}{1.44} < Z < \frac{110-108}{1.44}\right)$

   ```
   . di normal((110 - 108)/1.44) - normal((106 - 108)/1.44)
   .83513346
   ```

4. Suppose instead we repeatedly took random samples of size 25. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

   0.5762892

   $$P(106 < \bar{x} < 110) = P\left(\frac{106-108}{12.5/\sqrt{(25)}} < Z < \frac{110-108}{12.5/\sqrt{(25)}}\right)$$

   ```
   . di normal((110 - 108)/(12.5/5)) - normal((106 - 108)/(12.5/5))
   .5762892
   ```

5. Again, suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a mean less than $\bar{x} = 103$?

   0.0002660028

   $$P(\bar{x} < 103) = P\left(Z < \frac{103-108}{12.5/\sqrt{75}}\right) = P(Z < -3.4641016)$$

```
. di normal(-3.4641016)
.000266
```

6. If we repeatedly took samples of size 75, we would expect that, in 20% of the samples, $\bar{x}$ would be greater than ______?

   104.21478

   $P(\bar{x} > y) = P(Z > \frac{y-108}{12.5\sqrt{75}}) = 0.2 \rightarrow$

   $y = 108 + Z_{0.8} * 12.5/\sqrt{75}$

   ```
   . di 108 + invnormal(0.8)*12.5/sqrt(75)
   109.21478
   ```

7. After taking a sample of size 75, we found that the sample mean was $\bar{x} = 103$. Construct a two-sided 95% confidence interval for μ.

   (100.171, 105.829)

   ```
   . di 103 - invnormal(0.975)*12.5/sqrt(75)
   100.17104

   . di 103 + invnormal(0.975)*12.5/sqrt(75)
   105.82896
   ```

8. True or False. Based on the above interval, we can say that the probability that $\bar{x}$ lies in the interval is 0.95.

   False

   The interpretation of the confidence interval is that, if we repeatedly took samples of size 75, the true mean μ would lie in the confidence interval 95% of the time.

9. Suppose we were also interested in the mean of the highly right skewed indicator of iron absorption, ferritin. Compared to the relatively symmetrically distributed indicator hemoglobin, do you think a larger or smaller sample size would be required to apply the central limit theorem?

   Larger - The CLT is an asymptotic result - namely, you need a large sample size for it to work. However, how large is large enough is always an important question in statistics. You need much smaller sample sizes to invoke the CLT when the distribution of the variable of interest is symmetric and “well-behaved,” versus variables with highly skewed or multi-modal distributions (less “well-behaved”).

Hypothesis Testing with known Variance. Now, let's switch gears and assume that we didn't know the true population mean μ, and we only observed a sample of 25 school children in Delhi. Let's use the sample mean from this set of children to make inference about the true population mean μ. Assume that $\sigma = 12.5$ is known and that the sample size of 25 is large enough to use the Central Limit Theorem (i.e. base your inferences off of the normal distribution, not the t-distribution).

In this scenario, we can conduct a **one-sample Z-test** for inference about μ in a population with known variance σ^2. To test $H_0 : \mu = \mu_0$, we can use the test statistic:

$$Z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis, $Z^* \sim N(0,1)$. So, we can use the standard normal distribution to calculate a p-value for this hypothesis test:

- For the one-sided test with alternative hypothesis $H_a : \mu > \mu_0$, we can calculate a p-value using the formula $p = P(Z > Z^*)$.
- For the one-sided test with alternative hypothesis $H_a : \mu < \mu_0$, we can calculate a p-value using the formula $p = P(Z \leq Z^*)$.
- For the two-sided test with alternative hypothesis $H_a : \mu \neq \mu_0$, we can calculate a p-value using the formula $p = 2 * P(Z < -|Z^*|)$.

Note: there is no command for directly conducting this one-sample Z-test in Stata. However, you can use the normal function in Stata to calculate the p-values.

1. In a sample of size 25, what is the value of the test statistic testing whether the mean hemoglobin level is equal to 108 g/L versus the alternative that it is not equal to 108 g/L, when $\bar{x} = 103$. Use a one-sample Z-test. What is the p-value?

   Test statistic:-2
   p-value: 0.04550026

   $H_0 : \mu = 108, H_A = \mu \neq 108$

   $Z = \frac{103-108}{12.5/\sqrt{25}} = -2$

   $Z \overset{H_0}{\sim} N(0,1) \rightarrow P(Z < -2) = 0.023$

   ```
   . di normal(-2)*2
   .04550026
   ```

2. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is less than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

Test statistic:-2
p-value: 0.02275013

$H_0 : \mu = 108, H_A = \mu < 108$

$p = 2 * P(Z < -2)$

```
. di normal(-2)
.02275013
```

3. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is greater than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

Test statistic:-2
p-value: .9772498

$H_0 : \mu = 108, H_A = \mu > 108$

$p = P(Z > -2)$

```
. di 1 - normal(-2)
.9772498
```

Confidence intervals and testing with unknown variance. Suppose now that we are interested in the distribution of hemoglobin levels in Mumbai. We decide that it is unreasonable to extrapolate the Delhi results to Mumbai, and therefore the population standard deviation is unknown. We take a random sample of 15 children in Mumbai. The sample mean is $\bar{x} = 115$ g/L, with sample standard deviation $s = 10.2$ g/L. Assume hemoglobin levels in Mumbai are normally distributed (we could check this by looking at the distribution of hemoglobin levels in other similar populations).

1. Construct a two-sided 95% confidence interval for μ.

   (109.56446, 120.43554)

   ```
   . di 115 - invttail(24, 0.025)*10.2/sqrt(15)
   109.56446

   . di 115 + invttail(24, 0.025)*10.2/sqrt(15)
   120.43554
   ```

2. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.05$ level?
   (a) yes (b) no (c) not enough information

   Yes - 108 is not in the confidence interval.

3. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level?
   (a) yes (b) no (c) not enough information

   Not enough information - $\alpha = 0.01$ corresponds to a 99% confidence interval. This interval will be wider than the 95% confidence interval, and may or may not contain 108.

4. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.1$ level?
   (a) yes (b) no (c) not enough information

   Yes - $\alpha = 0.1$ corresponds to a 90% confidence interval, and this interval will always be narrower than the 95% confidence interval. Therefore, we know that 108 will not be in the 90% confidence interval and we can reject the null.

5. Conduct a one-sample t-test in Stata to test the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level.

```
. ttesti 15 115 10.2 108, level(99)

One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [99% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      15         115    2.633629        10.2    107.1601    122.8399
------------------------------------------------------------------------------
    mean = mean(x)                                                t =   2.6579
Ho: mean = 108                                   degrees of freedom =       14

   Ha: mean < 108               Ha: mean != 108                 Ha: mean > 108
 Pr(T < t) = 0.9906         Pr(|T| > |t|) = 0.0187          Pr(T > t) = 0.0094
```

- What is your test statistic?

  $t = 2.658$

- Under the null hypothesis, the test statistic follows a t-distribution with how many degrees of freedom?

  24 degrees of freedom $(n - 1 = 25 - 1 = 24)$.

- What is your p-value?

  $p = 0.0187$.

- What do you conclude?

  (a) reject the null hypothesis, (b) accept the null hypothesis, (c) fail to reject the null hypothesis

  Because $p = 0.019 > 0.01 = \alpha$, we fail to reject the null.

Nursing Home Study. Suppose that 500 residents of a large nursing home are screened for hypertension. All residents with above a specified level are labeled as having hypertension. The following table displays the results of this study.

	Hypertension	No Hypertension	Total
Male	100	100	200
Female	100	200	300

1. Which of the following measures of association would be the most appropriate to use to describe the finding in this study?

   a. Cumulative Ratio
   b. Incidence Rate Ratio
   c. Prevalence Ratio

2. What is the prevalence of hypertension among all residents?

   200/500 = 0.40

3. What is the prevalence odds ratio for having hypertension comparing male residents (exposed group) to female residents (non-exposed group)?

   OR = (100/100)/(100/200) = 2.0

4. Which of the following is the **least** likely explanation for the value for the odds ratio reported in the previous answer?

   a. Men have a higher risk of developing hypertension than women.
   b. Women who develop hypertension live longer after being diagnosed than men.
   c. Men who develop hypertension are more likely to be residents of a nursing home than women who develop hypertension
   d. Hypertension is harder to detect in women than in men.

5. Is reverse causation a plausible explanation for the association seen in the study?

   a. Yes
   b. No

Bed Occupancy in Two Hospitals. An investigator performs a bed-occupancy survey on July 1, 2012 at two hospitals (Hospital A and Hospital B). Hospital A has 100 available beds. Hospital B is larger with 200 available beds. Hospital A reports that 80 of its beds are occupied on the day of the survey (occupancy prevalence = 0.80).

1. The investigator reports a Prevalence Odds Ratio of 2.25 when comparing the occupancy prevalence of Hospital B to that for Hospital A. What is the occupancy prevalence for Hospital B on that day?

Prevalence odds in Hospital A:	**80/20=4**
Prevalence odds ratio for Hospital B vs. Hospital A:	**2.25**
Prevalence odds in Hospital B:	**X/4=2.25→X=9**
Occupancy prevalence for Hospital B:	**9/10=90%**

2. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. If the average length of stay is the same in each hospital then what is the incidence rate ratio for admission, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)?

a. 2.25
b. > 2.25
c. < 2.25
d. Cannot be determined from the information that is given

3. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. Furthermore, suppose that the incidence rate ratio for admission is 1.5, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)? Would the average length of stay for patients in Hospital B be?

a. Less than that for Hospital A
b. Equal to that for Hospital A
c. Greater than that for Hospital A
d. Cannot be determined from the information that is given

4. If the investigator presented these results at a local conference, then what should he described as the design of this study?

a. Case Report
b. Ecologic Study
c. **Cross Sectional Study**

Problem Set 6

Power and Sample Size. Recall (from the previous problem set) that according to the WHO Global Database on Anaemia, the mean hemoglobin levels among primary school children in Delhi were estimated at $\mu = 108$ g/L, with standard deviation $\sigma = 12.5$ g/L. Suppose you decide to go and conduct your own study. You aim to test whether hemoglobin levels in Mumbai are similar to the estimates from Delhi. For your calculations, you assume that the standard deviation of hemoglobin levels in Mumbai **is known** and is equal to that in Delhi. We aim to test:

This is one-sided

$$H_0 : \mu_M \leq 108$$
$$H_A : \mu_M > 108$$

You plan to conduct this one-sided test at the $\alpha = 0.05$ level of signficance.

1. Which of the following will increase the power of a test: (a) increasing the type-I error, (b) increasing the sample size, (c) all of the above.

   (c)

2. If you randomly sample 30 children, what is the type I error of your test?

   0.05

3. If you randomly sample 30 children, what is the power of your test against the alternative $\mu_M = 112$?

   0.5429

```
. sampsi 108 112, sd1(12.5) n1(30) onesample onesided

Estimated power for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
 alternative m =      112
            sd =     12.5
 sample size n =       30

Estimated power:

         power =   0.5429
```

4. If the true mean levels of hemoglobin were 115 g/L, how many children do you need to sample to have 80% power (assume the standard deviation remains the same as Delhi)?

   20

```
. sampsi 108 115, sd1(12.5) power(.80) onesample onesided

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.8000
 alternative m =      115
            sd =     12.5

Estimated required sample size:

             n =       20
```

5. If the true mean levels of hemoglobin were 112 g/L, how many children do you need to sample to restrict your type II error below 10%?

   84

```
. sampsi 108 112, sd1(12.5) onesample onesided

Estimated sample size for one-sample comparison of mean
  to hypothesized value

Test Ho: m =    108, where m is the mean in the population

Assumptions:

         alpha =   0.0500  (one-sided)
         power =   0.9000
 alternative m =      112
            sd =     12.5

Estimated required sample size:

             n =       84
```

Two samples here - aspirin vs placebo groups

Two Sample t-test. Schiff et al. (1990) investigated the effect of low doses of aspirin on women with mild pregnancy-induced hypertension. Forty-seven women hospitalized at 30-36 weeks' gestation because of mild pregnancy-induced hypertension were treated by a daily dose of either 100 mg aspirin or placebo. The 23 women who received aspirin had a mean arterial blood pressure of 111 mm Hg with a standard deviation of 8 mm Hg. The 24 women who received placebo had a mean arterial blood pressure of 109 mm Hg and a standard deviation of 8 mm Hg.

Source: Schiff, E., Barkai,G., Ben-Baruch G., and Mashiach, S., "Low-Dose Aspirin Does Not Influence the Clinical Course of Women with Mild Pregnancy-Induced Hypertension," Obstetrics and Gynecology, Vol. 76, November 1990, 742-744.

1. Conduct a two sample t-test with equal variances at α = 0.05 to test whether the population mean arterial blood pressure for women treated with aspirin is equal to the population mean arterial blood pressure for women treated with placebo. What is the value of the test statistic?

   Here, $s_p^2 = 64$. Now, $t = \frac{111-109}{\sqrt{\frac{64}{23}+\frac{64}{24}}} = 0.856763$

   Stata Command -- ttesti 23 111 8 24 109 8

2. How man degress of freedom does your test statistic have?

   $23 + 24 - 2 = 45$

3. The p-value for this test is greater than 0.05. Based on this result can we conclude that the population mean arterial blood pressure for pregnant women treated with aspirin is equal to the population mean arterial blood pressure for pregnant women treated with placebo? A) Yes B) No

   VERY IMPORTANT

   B) No. We cannot conclude the two are equal. The most we can say is that we do not have enough evidence to conclude that the population mean arterial blood pressure for pregnant women treated with aspirin is different from the population mean arterial blood pressure for pregnant women treated with placebo

4. Based on this result, would you expect the 95% confidence interval for the difference in means to include zero? A)Yes B) No

   A) Yes. Since the p-value was greater than 0.05, the confidence interval will include zero. Remember there is a 1-1 correspondence between hypothesis testing and confidence intervals.

Command for ANOVA = oneway

ANOVA. Consider a hypothetical clinical trial to examine the effects of 3 different drug doses (low, high, and placebo) on mean change systolic blood pressure among a random sample of patients with hypertension.

We use Stata to test the null hypothesis at the $\alpha = 0.05$ level that the three groups have equal population mean change in systolic blood pressure. The output (with some entries omitted) are presented below.

```
. oneway spb trt, tabulate

            |           Summary of SPB
        trt |        Mean   Std. Dev.       Freq.
------------+------------------------------------
    Placebo |   21.886666   1.1587349          15
   Low dose |   21.593333   1.1510659          15
  High dose |   23.506667   1.8771053          15
------------+------------------------------------
      Total |   22.328889   1.6413165          45

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      31.8564567     xx    15.9282283     xxx      0.0014
 Within groups      86.6760131     42     2.0637146
------------------------------------------------------------------------
    Total            118.53247     44    2.69391977

Bartlett's test for equal variances:  chi2(2) =   4.5868  Prob>chi2 = 0.101
```

1. What is the value of the F statistic?

   $\frac{15.9282283}{2.0637146} = 7.72$

2. How man degrees of freedom does the numerator of the F statistic have?

   3-1=2

3. Based on this result, we can conclude that we have evidence that there is a significant difference in the population mean change in systolic blood pressure between the high dose group and low dose group. A)True B) False

   VERY IMPORTANT B) False. We can only conclude that the three means are not equal. We do not know which pairs are not equal.

4. Now we wish to use a Bonferroni adjustment to compare the mean change in systolic blood pressure between all three pairs of groups. If we wish to set the overall level of significance at 0.05, what α must we use for each individual test?

   $\frac{0.05}{3} = 0.01667$

Experimental Study of Aspirin and Alzheimer's Disease. Suppose that an investigator plans to perform an experimental study on elderly subjects to examine if aspirin decreases the risk of developing Alzheimer's disease. Subjects will be randomized to receive either aspirin or a placebo and they are followed to record the development of this disease.

1. In an experimental study like this, equipoise refers to the fact that each subject has an equal chance of being assigned to either the aspirin or the placebo group.

    a. True
    b. False

2. In general, randomization tends to balance the distribution of factors that might influence the outcome, especially for small studies but unlikely for large studies.

    a. True
    b. False

3. One of the expected results of randomization is that the distributions of the risk factors among the cases of Alzheimer's disease that develop in the study will be approximately the same as among those who do not develop this outcome.

    a. True
    b. False – This question asks you to compare the distribution of risk factors for disease after randomization between those who develop disease and those who do not develop disease. However, randomization only balances the distribution of risk factors for disease between the exposed and non-exposed groups, not between those who develop disease and those who do not develop disease.

IMPORTANT

4. The purpose of randomizing some of the subjects to receive a placebo pill was to

    a. Blind the participants from knowledge of their treatment assignment so that such knowledge would not influence their compliance and outcomes
    b. Blind the participants from the purpose of the study so that such knowledge would not influence their compliance and outcomes
    c. Investigate if placebo pills decrease the risk of Alzheimer's disease among the elderly

5. Suppose the study used a blocked randomization scheme with a fixed block size. Suppose that the first twelve subjects are assigned to groups in the following order (A = aspirin, P = Placebo)

   A P P A A A P P P A P A

   Which of the following is a possible value for the fixed block size?

   a. Two
   b. Four
   c. Six
   d. None of the above

6. Because of potential side effects from aspirin, the investigator incorporates a run-in phase into the design of this study. The purpose of the run-in phase is to

   a. Collect risk factor information on all potential subjects to identify and enroll only high risk elderly subjects
   b. Give aspirin to all potential study subjects prior to randomization to determine compliance. Non-compliers may be less likely to comply with treatment assignment after randomization and therefore would not be enrolled in the study
   c. Monitor all subjects early after randomization for compliance to their assigned treatment and remove all non-compliers from the study at that point.
   d. Monitor all subjects assigned to the placebo group early after randomization to determine who are purchasing and taking aspirin.

Compliance in Experimental Study of Aspirin and Alzheimer's Disease.
Suppose that 500 subjects were randomized to the aspirin group and another 500 to the placebo. At the end of the study, compliance was assessed by asking all subjects if they took the assigned pill daily. The study reported the following results regarding compliance:

		Randomization Assignment	
		Aspirin	Placebo
Took assigned pill daily	Yes	400 (Group A)	450 (Group B)
	No	100 (Group C)	50 (Group D)

1. Which groups would be compared in an Intention-To-Treat Analysis to measure the effect of aspirin on the rate of developing Alzheimer's disease?

   a. Group A versus Group B
   b. (Group A + Group C) versus (Group B + Group D)
   c. Group A versus (Group B + Group C + Group D)

Problem Set 7 Solutions

Inference for proportions. According to Fergusson *et al.* (2012), acutely ill patients, including neonatal infants, often receive red blood cell transfusions. However, the consequences of the prolonged storage of red blood cells on health outcomes in premature infants are not well understood. In a double-blinded, randomized controlled trial, the authors looked at health outcomes in neonatal infants who underwent red blood cell transfusions, comparing the standard protocol (transfusions of blood stored for prolonged periods) with fresh blood transfusions (transfusions of blood store for less than seven days).

Specifically, the authors examined five outcomes listed below; as well as a composite outcome, defined as at least one of the five outcomes. In this question, we focus primarily on the composite outcome. The results of the study are shown in the following table:

	Standard	**Fresh**
Necrotizing enterocolitis	15	15
Intraventricular hemorrhage	11	18
Retinopathy of prematurity	26	23
Bronchopulmonary dysplasia	63	60
Death	31	30
Composite Outcome	100	99
Sample Size	189	188

A dataset `hw7.dta (`or `hw7.csv)` is also available on the course website, if you would rather not use the "immediate" commands in Stata.

Source: Dean A. Fergusson, MHA, PhD; Paul Hébert, MD, MHSc(Epid); Debora L. Hogan, BScN, BA, MScN; Louise LeBel, BScN; Nicole Rouvinez-Bouali, MD; John A. Smyth, LRCPSI; Koravangattu Sankaran, MBBS; Alan Tinmouth, MD, MSc(Clin Epi); Morris A. Blajchman, MD; Lajos Kovacs, MD; Christian Lachance, MD; Shoo Lee, MBBS, PhD; C. Robin Walker, MB,ChB; Brian Hutton, PhD; Robin Ducharme, HBSc; Katelyn Balchin, MSc; Tim Ramsay, PhD; Jason C. Ford, MD; Ashok Kakadekar, MD; Kuppuchipalayam Ramesh, MD; Stan Shapiro, PhD. (2012). Effect of Fresh Red Blood Cell Transfusions on Clinical Outcomes in Premature, Very Low-Birth-Weight InfantsThe ARIPI Randomized Trial. JAMA.

1. Construct a 95% confidence interval for the proportion of infants experiencing the composite outcome in the fresh red blood cell group, using the following methods:

   a) the exact binomial confidence interval

```
(45.3, 60.0)

. ci outcome if fresh==1, binomial

                                                     -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     outcome |        188    .5265957    .0364146        .4526364    .5997032
```

b) the Wilson confidence interval (which uses the normal approximation)

```
(45.5, 59.7)

. ci outcome if fresh==1, binomial wilson
                                                  ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
     outcome |        188    .5265957    .0364146         .455408    .5967184
```

2. Is the normal approximation to the binomial appropriate in this setting?

   Yes.

3. Suppose you wanted to calculate a 95% confidence interval for infants experiencing intraventricular hemorrhage after receiving a fresh blood transfusion as well. Is the Wilson confidence interval still appropriate?

   Yes.

4. Estimate and construct a large-sample 95% confidence interval for the risk difference for experiencing the composite outcome for those with fresh blood versus the standard protocol blood. Calculate the risk difference as estimated proportion in fresh blood group minus estimated proportion in the standard blood group.

   -0.3 (-10.3, 9.8)

```
. prtest outcome, by(fresh)

Two-sample test of proportions                     0: Number of obs =      189
                                                   1: Number of obs =      188
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |   .5291005    .036308                      .4579382    .6002629
           1 |   .5265957   .0364146                      .4552244    .5979671
-------------+----------------------------------------------------------------
        diff |   .0025048   .0514227                     -.0982819    .1032915
             |  under Ho:   .0514228     0.05   0.961
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                                  z =   0.0487
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.5194         Pr(|Z| < |z|) = 0.9612          Pr(Z > z) = 0.4806
```

5. Use a two-sample test of proportions to determine whether there is a difference between fresh and standard groups at the α=0.05 level of significance. What is the test statistic? Null distribution? P-value? Conclusion?

   Z = 0.0487, Z ~ N(0, 1) under the null

   p = 0.9612

   (From risk difference problem above)

   no evidence that the risk difference is different from 0.

Contingency Tables. Continue using the Fergusson et al. (2012) clinical trial data to complete the following questions.

1. Estimate the odds ratio and a 95% confidence interval for experiencing the composite outcome for those with fresh blood versus standard protocol blood. Is there evidence of an association between blood group and the composite outcome (at the 0.05 level of significance).

   1.0 (0.7, 1.5)

   No evidence of an association.

```
. cs outcome fresh, or woolf

                 | fresh                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        99         100  |        199
        Noncases |        89          89  |        178
-----------------+------------------------+------------
           Total |       188         189  |        377
                 |                        |
            Risk |  .5265957    .5291005  |   .5278515
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |        -.0025048       |   -.1032915    .0982819
      Risk ratio |          .995266       |    .8222696    1.204659
 Prev. frac. ex. |          .004734       |   -.2046588    .1777304
 Prev. frac. pop |         .0023607       |
      Odds ratio |              .99       |    .6607011    1.483424 (Woolf)
                 +-------------------------------------------------
                               chi2(1) =     0.00  Pr>chi2 = 0.9612
```

2. Construct a 2x2 table for the composite outcome versus blood group. Are the expected cell counts large enough to conduct a Pearson Chi-square test?

   Yes.

3. Using the Pearson chi-square test, determine if there is an association between fresh versus standard blood and the composite outcome at the α=0.05 level of significance. What is your test statistic? Null distribution? P-value? Conclusion?

   $X = 0.0024 \sim \chi^2_1$

   p = 0.961

   fail to reject the null – no evidence of an association.

Final Thoughts.

1. In the previous questions, we looked at three different tests of association: the Pearson Chi-square test, an odds ratio test, and a risk difference test. Are the results of these three tests consistent? Would you expect them to be?

   Yes, and yes.

2. Is there evidence of an association between blood group assignment and the composite outcome?

   No.

3. Think back to the Bonferroni correction from last week. If you were tasked with conducting hypothesis tests comparing the two blood groups for **each** of the 5 different outcomes, would you need to correct for multiple comparisons?

   Yes.

4. In this study, the authors state that they powered the study to detect an absolute difference of 15% in the two groups with 80% power, used a 2-sided test with α =0.05. After a few more adjustments, their final sample size calculation was 450.

   Now suppose you want to replicate the study using a different population. Given that the authors did not find an association in their data, you decide to increase the power and decrease the difference detected between standard and fresh groups. Using an equal number of infants in both groups, what is the total sample size needed in order to achieve 90% power, assuming that the proportion of infants experiencing the composite outcome in the standard group was 55% and 45% in the fresh blood group (again using a 2-sided test with α =0.05).

   1088

```
. sampsi 0.45 0.55

Estimated sample size for two-sample comparison of proportions

Test Ho: p1 = p2, where p1 is the proportion in population 1
                    and p2 is the proportion in population 2
Assumptions:

   alpha =   0.0500  (two-sided)
   power =   0.9000
      p1 =   0.4500
      p2 =   0.5500
   n2/n1 =   1.00

Estimated required sample sizes:

      n1 =      544
      n2 =      544
```

5. Consider a covariate, the clinical risk index for babies (CRIB), which was measured in the infants enrolled in the clinical trial. CRIB is usually associated with the composite outcome. From the baseline characteristics table in the Fergusson et al paper, we find that the median and IQR for CRIB is similar between the standard and fresh blood groups. This suggests that the distribution of CRIB is similar in both groups.

   True or False: Because the distribution of CRIB is similar betwen groups, the study investigators would not have gained any power to detect an effect by matching on CRIB score.

   False.

   Matching exploits correlation between CRIB score and outcome to gain power. If CRIB score is highly associated with the outcome, we could gain power by matching.

Randomized Clinical Trial versus Cohort Study

1. The benefit of a randomized clinical trial over an observational cohort study is that, in large enough samples, the groups are identical with respect to

   A. other extraneous factors that are associated with the outcome of interest

   B. factors that would make the results more generalizable to the larger population

   C. other factors related to the likelihood of participating in a study

   Choice A

Toxins and Parkinson's Disease Cohort Study

An investigator, Dr. Park, is interested in evaluating whether there is an association between exposures to toxins and risk of developing Parkinson's disease. She constructs a cohort of men and women that are living in her state in 1985 and every year, they are asked to complete a questionnaire about exposures at work and at home and whether they have been diagnosed with Parkinson's disease. The participants in Dr. Park's study contribute information on their changing toxin exposures over time for as long as they are residents in that state and for any year that they complete the questionnaire. As new people move into the state, they are enrolled in her cohort and remain in the cohort for as long as they are residents in the state and complete the questionnaire.

1. Should Dr. Park include people who reported that they had Parkinson's at the time that they were recruited?

    A. Yes

    B. No

    No, because they are not at risk of developing the disease.

2. Does Dr. Park need to be concerned about selection bias?

    A. Yes, if people who were at greater risk of developing Parkinson's were more likely to be exposed and were also more likely to participate in the study

    B. No, because it is a prospective cohort study so selection biases are not a concern

3. If those who are exposed to toxins are more likely to drop out of the study and more likely to develop Parkinson's disease, what effect with this have on the estimated relative risk compared to the true relative risk?

    A. No effect

    B. Biased towards the null

    C. Biased away from the null

4. Is this an open cohort or closed cohort? Why?

    A. Closed, because exposure at baseline is used for all follow-up

    B. Open, because loss to follow-up and competing risks are still a problem in this population-based study

    C. Closed, because risk ratios are the most appropriate measure to compare toxin levels and Parkinson's risk

    D. Open, because people can enter and exit the cohort over the follow up time of the study

5. True or False: Based on the data collected in her study, Dr. Park will be able to calculate absolute measures of Parkinson’s disease incidence.

   True

6. Because Dr. Park conducted a prospective cohort study instead of a cross-sectional study, she can be less concerned about

   A. Confounding

   B. Bias

   C. Reverse causation

   D. Chance

Problem Set 8

Survey Sampling Design. You decide to conduct a survey to measure physical activity in Boston. You plan to collect information on the amount of physical activity per week, history of diabetes, weight, height, age, and gender.

There are seventeen distinct neighborhoods in Boston, with substantial differences in race/ethnicity, socioeconomic status, and population density. You expect to observe variability in physical activity indicators by neighborhood. Unfortunately there is no specific information about within-neighborhood heterogeneity (variance); therefore, you design the survey based on the assumption that variances of indicators are equal across neighborhoods.

The table below displays population data for Boston in 2010. Assume that these population numbers are still accurate today.

Neighborhood	Population - 2010
South End	34,669
Central	30,901
Fenway - Kenmore	40,898
South Boston	33,688
Charlestown	16,439
Allston - Brighton	74,997
West Roxbury	30,445
Roxbury	59,790
East Boston	40,508
Jamaica Plain	39,897
Back Bay - Beacon Hill	27,476
Hyde Park	31,813
North Dorchester	28,384
South Dorchester	59,949
Roslindale	32,589
Mattapan	34,616
Harbor Islands	535
Boston	617,594

Source: http://www.bostonredevelopmentauthority.org/PDF/ResearchPublications//PDPercentChange.pdf

Consider the following questions:

1. Suppose you decided to randomly sample 1,700 people from the city of Boston (call this Design 1). For any given individual in South Dorchester, what is the probability of being selected in the survey? What is this probability for an individual in Harbor Islands?

   South Dorchester: 0.002752617

   Harbor Islands: 0.002752617

P(individual i selected) = 1700/617594 = 0.002752617

2. What is likely the main challenge of implementing this survey?

   (a) SRS is inefficient, (b) SRS is difficult to implement in practice, (c) SRS does not necessarily sample people from each neighborhood

3. Now, you randomly sample 100 people within each neighborhood (Design 2). What is the probability of a random individual in South Dorchester being sampled? What is the probability of a randomly selected individual in Harbor Islands being sampled?

   South Dorchester: 0.0017

   Harbor Islands: 0.1869159

   P(individual selected who lives in S. Dorchester) = 100/59949 = 0.001668085

   P(individual selected who lives in Harbor islands) = 100/535 = 0.1869159

4. What kind of survey design is Design 2?

   (a) stratified sample, (b) cluster sample, (c) simple random sample

5. Consider an alternate design (Design 3). In each neighborhood the number of individuals sampled is proportional to the population size of the neighborhood. Assuming that the sample size is fixed at 1,700, would you expect Design 2 or Design 3 to provide more precise estimates?

   (a) Design 2, (b) Design 3

   Design 2 oversamples people in the small neighborhoods. By sampling fewer people in the smaller neighborhoods and more people in the larger neighborhoods, our design is closer to a simple random sample and is more efficient.

6. Once again, consider the probability of a random individual in South Dorchester being sampled; and the probability of a random individual in Harbor Islands being sampled. These probabilities are approximately the same as those in:

   (a) Design 1, (b) Design 2

7. Why might you want to use Design 2 compared to Design 3?

   (a) to increase precision, (b) if you wanted neighborhood-specific estimates, (c) both a and b

Recall Design 2 results in a higher variability of survey weights due to a disproportionately large number of individuals selected in smaller neighborhoods (decreasing precision). By oversampling within small neighborhoods, we could obtain neighborhood specific estimates.

8. Next, you decide you do not want to visit all 17 neighborhoods, so you randomly sample 10 neighborhoods. Within each neighborhood selected, you randomly sample 170 people. Call this Design 4. What is the probability of a random individual in South Dorchester being included in the survey? What is the probability of a random individual in Harbor Islands being included in the survey?

   South Dorchester: 0.001668085

   Harbor Islands: 0.1869159

   P(individual selected who lives in S. Dorchester)

   = P(S. Dorchester selected)P(individual selected|lives in S. Dorchester)

   = (10/17)(170/59949) = 0.001668085

   P(individual selected who lives in Harbor Islands) =

   = P(Harbor Islands selected)P(individual selected|lives in Harbor Islands)

   = (10/17)(170/535) = 0.1869159

9. A "self-weighting" design is a survey design for which every individual in the population has an equal probability of inclusion. Which survey designs are self-weighting (or approximately self-weighting)?

   (a) Design 1, (b) Design 2, (c) Design 3, (d) Design 4

10. Consider yet another design (Design 5), in which you again select 10 neighborhoods. Now, the probability of a neighborhood being included in the survey is proportional to its population size. Within each sampled neighborhood, you randomly sample 170 people. Would you expect Design 4 or Design 5 to provide more precise estimates?

    (a) Design 4, (b) Design 5

    To build some intuition, think about Harbor Islands. The probability of including a bunch of people from Harbor Islands is much higher in Design 4 compared to Design 5. Oversampling from Harbor Islands decreases our precision, because Harbor Islands represents such a small fraction of the population of Boston.

Other aspects of survey design. Thus far, we have examined sampling design, which is only one element of designing a survey. Building and testing the survey instrument/questionnaire, anticipating and preparing for non-response, and training field teams to conduct the survey are several other critical aspects that we have not even touched on! Consider the following:

1. When designing your survey, you are trying to decide whether to use a 2-page questionnaire with 15 simple questions about whether an individual exercises, how many times per week, and whether they have a history of diabetes, along with some other very basic demographics (call this Survey 1). Your colleague says you could get much better information if you used a 10 page questionnaire with a more complete medical history and history of exercise and physical activity (call this Survey 2).

   Krosnick (1991) discusses "satisficing" (satisfy + suffice) in surveys. To paraphrase this discussion, satisficing occurs when, rather than optimizing their responses to best reflect reality, survey respondents try to reduce the cognitive burden associated with the survey and consequently may select a survey response haphazardly or even arbitrarily.

   Which survey would be more susceptible to satisficing?

   (a) Survey 1, (b) Survey 2

   Source: Krosnick, J. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, 5: 236.

2. You decide to implement a door-to-door survey, where you randomly sample addresses within a neighborhood and train a field team to ask the survey questions at the selected households. (You have a complete listing of addresses in Boston).

   Your colleague says you should obtain a listing of all land-line telephone numbers in Boston (a listing of cell phone numbers is not available) and randomly sample numbers from this list and ask the questions over the phone. For sake of argument, assume that you get a 100% response rate for both modes (door to door and phone calls), and that you construct survey weights using the table above.

   Would you expect the door-to-door or land-line method to produce unbiased results? (Think about which sampling frame is likely to be more complete.)

   (a) Door-to-door, (b) landline

   Would you expect the door-to-door or land-line method to be cheaper to implement?

   (a) Door-to-door, (b) landline

Note: web-based surveys are also common. In order to minimize non-response, some surveys use multiple modes of response and follow-up (e.g. web + phone; or web + household follow-up).

3. Individuals can opt out of any part of your survey. High BMI and low physical activity are risk factors for diabetes, and you find that individuals with these characteristics are less likely to answer questions about history of diabetes (note that high BMI and low physical activity are risk factors for diabetes). In a complete case analysis, missing data is dropped, survey weights are recalculated, and data is analyzed assuming missing observations were never collected. In a complete case analysis, would you expect to obtain unbiased estimates of diabetes prevalence in Boston?
   (a) yes, (b) no

Case Control versus Cohort Study

1. One of the main advantages of case-control studies over cohort studies is that

   A. The investigator can identify exposed and unexposed people without concerns about selection bias.

   B. **The investigator does not need to wait a long time for cases to occur.**

   C. The investigator does not need to be concerned about issues of confounding.

   D. The investigator can examine whether a third factor modifies the relationship between exposure and outcome.

Alcohol Consumption and Cancer Study

Dr. Marks is interested in examining the association between alcohol consumption and a rare form of cancer. Since it is hard to identify cases, Dr. Marks designs a case-control study. Cases are identified from cancer treatment centers across the United States. Controls are selected on the day that the case is diagnosed with cancer from among the relatives of the cancer patient.

1. True or False: Dr. Marks should make sure to select the healthiest relative as the control to ensure that he will observe differences in risk between cases and controls.

   False: the controls are meant to represent the exposure distribution in the population giving rise to the cases, not to represent a level of outcome risk.

2. The main concern of using relatives of cases as the controls is that

   A. Some cases may not have relatives that drink alcohol.

   B. Some cases may have relatives that live far away so not all relatives will be available to participate.

   C. **Relatives of cases may be more likely to have levels of alcohol consumption that are more similar to the cases than the population that gave rise to the cancer cases.**

   D. The controls may develop other diseases and will not be available to participate in this study.

3. Based on the study description above, Dr. Marks conducted a

   A. **Density case control study**

   B. Nested case-control study

   C. Case-cohort study

   D. Retrospective cohort study

4. True or False: Using this design, a relative who served as a control cannot be included as a case if he later develops the cancer of interest.

   False

5. True or False: Using the data collected in this study, Dr. Marks will be able to estimate the rates for developing this cancer.

   False

Problem Set 9 Solutions

Trends in Unemployment. We use publicly available data from the World Bank's website to examine national trends in unemployment percentages (percent unemployed in total labor force) in four countries: the United States, Great Britain, Japan, and Canada.

In this question, we examine unemployment trends over time using a correlation analysis, focusing primarily on the United States and Japan.

Use the dataset UnemploymentbyCountry.dta to answer this question. This dataset contains some missing data, but please just perform a complete case analysis within country (using all available data for a country and omitting any missing observations for the variables of interest within the country).

1. Calculate the Pearson and Spearman correlations between total unemployment and year for the United States and for Japan.

   United States:

Pearson:	-0.3415
Spearman:	-0.4517

   Japan:

Pearson:	0.8437
Spearman:	0.8162

   ```
   pwcorr unemployedtotal year if country == "United States"

   spearman unemployedtotal year if country == "United States"

   pwcorr unemployedtotal year if country == "Japan"

   spearman unemployedtotal year if country == "Japan"
   ```

2. Exclude all years after 2007 (recall that the financial collapse occurred in late 2008). Recalculate the correlations in question 1.

   United States:

Pearson:	-0.7557
Spearman:	-0.7733

   Japan:

Pearson:	0.8276
Spearman:	0.7998

```
pwcorr unemployedtotal year if country == "United States" & year < 2008

spearman unemployedtotal year if country == "United States" & year <
2008

pwcorr unemployedtotal year if country == "Japan" & year < 2008

spearman unemployedtotal year if country == "Japan" & year < 2008
```

3. Construct a scatter plot with year on the x-axis and with both unemployment in the United States and unemployment in Japan on the y-axis.

   Which pattern best describes the trend in unemployment in the United States between 1980 and 2010?

   (a) Linear (b) quadratic (c) cyclic (d) gradual non-linear increase

   Which pattern best describes the trend unemployment in the Japan between 1980 and 2010?

   (a) Linear (b) quadratic (c) cyclic (d) gradual non-linear increase

4. True or False: correlation analyses have the potential to mask important non-linear trends in data.

   True.

   Correlation analyses do not help us understand the cyclic pattern of unemployment in the United States. Looking at the data, we clearly see when the recession in the 80's happened, as well as the recent financial collapse. Correlation analyses are useful for looking at broad linear trends, but are usually not a last-stop.

5. Restricting to the United States, construct a scatterplot with year on the x-axis and with total unemployment; unemployment among women; and unemployment among men on the y-axis. Do there appear to be sex-specific differences in the unemployment trends?

   (a) Yes (b) No

Feel free to explore this data set further or go to the World Bank's website http://data.worldbank.org/ and construct the dataset for your own country!

Created from: World Bank, World Development Indicators and Global Development Finance.

Source: International Labour Organization, Key Indicators of the Labour Market database

Nonparametric Tests.

1. Suppose we wish to test a new treatment for dry, itchy eyes. We gather a group of eye patients, and, for each patient, we randomly treat one eye with the experimental treatment and one eye with the standard treatment. The outcome, eye relief, is a continuous measure, and the distribution of the differences in eye relief between eyes is normally distributed. Suppose we want to test whether the treatment is effective at improving eye relief. Which of the following tests are valid:

   A) Sign Test

   B) Signed-Rank Test

   C) Paired t-test

   D) All of the above

   IMPORTANT NOTE BELOW !

   The non-parametric tests (sign test and signed rank test) are still valid if the distribution of the differences is normally distributed. However, the non-parametric tests will not be as powerful as the paired t-test.

2. Suppose you are conducting a two-sided sign test. You have data from 8 paired samples, and you observe 1 positive sign and 7 negative signs. What is the p-value corresponding to the null hypothesis that there is no difference in median between the two groups?

   0.0704

   2*[Probability of observing 1 positive sign under the null hypothesis + Probability of observing 0 positive signs under the null hypothesis]

   = 2*[0.0313 + 0.0039]

   = 0.0704.

3. Walker et al. (1987) examined the characteristics children dying from sudden infant death syndrome. The sids.dta contains the age at death (in days) for a sample of 12 girls and 17 boys.

   Source: Walker, A.M., Jick, H., Perera, D.R., Thompson, R.S., and Knauss, T.A., “Diphtheria-Tetanus-Pertussis Immunization and Sudden Infant Death Syndrome,” American Journal of Public Health, Volume 77, August 1987, 945-951.

a. Using an appropriate non-parametric test with a 0.05 level of significance, test whether the median age at death is the same for boys and girls. Be sure to verify the assumptions of your test. What is the absolute value of your test statistic?

Use sign-rank for two-sample

0.044

b. What is your p-value?

0.9647

c. Based on this test, can you conclude that we do not have enough evidence to suggest that the median age at death is different between boys and girls?

Yes.

d. Would it be appropriate to use a two-sample t-test in this case?.

IMPORTANT

No.

Based on the histograms of age at death by sex it does not appear that age at death is normally distributed for boys or girls.

Confounding Multiple Choice

1. Select the best answer: In the presence of confounding,
    - A. The crude (unadjusted) results are not correct if the sample size is small.
    - B. The exposed and unexposed are exchangeable.
    - C. The results are not correct because there is a third factor associated with exposure and outcome that at least partially explains the results.
    - D. Stratification is not appropriate.

Coping and Outcomes after Surgery Study

Dr. Spencer is conducting a study to evaluate whether better coping styles are associated with improved outcomes after undergoing surgery. He is concerned that people with better coping styles have other healthy lifestyle factors and more social support that also impact post-surgery prognosis.

1. True or False: Because people cannot be randomized to a particular coping style, Dr. Spencer cannot test his hypothesis in a randomized controlled trial to assure that the correct counterfactual outcomes can be compared properly. Therefore, there is no way to evaluate this question.

False: Dr. Spencer can collect information on other healthy lifestyle factors to either stratify or model the results to account for potential confounders.

Stratification - helps to avoid confounding

2. Dr. Spencer chooses to draw a directed acyclic graph (DAG), which is useful because

   A. It prevents confounding so that stratification is not necessary.

   B. It helps the researcher evaluate the relationship between the factors under study to decide when stratification or statistical adjustment is appropriate.

   C. It addresses issues of chance and how that impacts the results of a stratified analysis.

3. True or False: In terms of examining whether there is a causal effect of exposure on outcome, Dr. Spencer wants to make sure that the risk of poor post-surgery prognosis is similar for all factors other than coping styles. This is important to achieve exchangeability to conclude that differences in coping styles is causing the differences in the risk of the outcome.

True

Obesity and Rate of Stroke

In the Framingham dataset that can be downloaded from the course website, we would like to examine the association between obesity and the rate of stroke. Since people with hypertension may have a higher body mass index and they are also at greater risk of a stroke, you may be concerned about confounding by hypertension.

1. Use the Framingham dataset and Stata to calculate the incidence rate ratio of stroke comparing obese participants (bmi1>=30) to all other participants. The variable for incident stroke in this dataset is “stroke” and number of years a person was followed for stroke is recorded in the “timestrk” variable in the NHLBI dataset.

1.76

```
generate obese = .
replace obese=0 if (bmi1<30.0)
replace obese=1 if (bmi1>=30.0 & bmi1<.)

. ir stroke obese timestrk

                 | obese                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
Incident Stroke  |        82         330  |        412
Time [years] to  |  10922.22    77530.79  |   88453.01
-----------------+------------------------+------------
                 |                        |
  Incidence rate |  .0075076    .0042564  |   .0046578
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |          .0032513      |    .0015626    .0049399
 Inc. rate ratio |          1.763857      |     1.36763    2.252718 (exact)
 Attr. frac. ex. |          .4330605      |    .2688079    .5560918 (exact)
 Attr. frac. pop |          .0861917      |
                 +-------------------------------------------------
                    (midp)   Pr(k>=82) =                     0.0000 (exact)
                    (midp) 2*Pr(k>=82) =                     0.0000 (exact)
```

2. Using the Framingham dataset and Stata, what is the incidence rate ratio of stroke among obese participants (bmi1>=30) compared to all other participants after adjusting for prevalent hypertension at visit 1 (prevhyp1)?

1.17

```
. ir stroke obese timestrk, by(prevhyp1)

Prevalent hypert |       IRR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
              No |   1.458798       .8408509   2.388242     11.37127 (exact)
             Yes |    1.09436       .8103538    1.46154     44.43501 (exact)
-----------------+-------------------------------------------------
```

```
           Crude |   1.763857       1.36763   2.252718              (exact)
    M-H combined |   1.168619      .9139881   1.494188
-----------------------------------------------------------------
 Test of homogeneity (M-H)    chi2(1) =      0.99  Pr>chi2 = 0.3191
```

3. Based on your results above, which option for reporting the association between obesity and stroke is best?

   A. The crude incidence rate ratio for the association between obesity and stroke.

   B. The incidence rate ratio for the association between obesity and stroke adjusted for prevalent hypertension using the Mantel-Haenszel formula.

   C. The incidence rate ratio for the association between obesity and stroke stratified by prevalent hypertension (ie calculate two incidence rate ratios – one among those with prevalent hypertension and one among those without prevalent hypertension).

   D. B or C

Problem Set 10

We use data from the Environmental Protection Agency (EPA) to track the BP oil spill in Louisiana between May and September 2010.

The oil well exploded on April 20, 2010, was capped in July and was declared dead in September 2010. The oil spill killed wildlife in the Gulf of Mexico and posed a significant public health risk to clean-up workers and residents of the Gulf Coast (Solomon and Janssen 2010). The EPA monitored air quality and took samples of sediment to measure the impact of the oil spill. For more information from the EPA, visit http://www.epa.gov/bpspill/sediment.html#understanding.

*Solomon G., Janssen S. (2010). Health Effects of the Gulf Oil Spill. *JAMA*. (doi:10.1001/jama.2010.1254)

Use the dataset `oilspill.dta` to answer the following questions.

Simple Linear Regression. In this example, we track changes in the amount of nickel found in sediment along the Louisiana coast between May and September 2010 (note that there is no data from July). Nickel is a metal that is found in sediment contaminated with oil. We model the amount of nickel as a function of time (month) using linear regression.

Fit a linear regression model with `nickel` as the outcome and `month` as the explanatory variable. Using indicator variables, model `month` as a categorical covariate (May = 5, June = 6, August = 8, September=9). Call this Model 1.

Assume the assumptions of linear regression are met for this model. You can make histograms of `nickel` by `month` to visually verify that the data does not appear to be skewed or any other evidence that would suggest a violation of the assumptions necessary to analyze this data using linear regression.

1. Does the amount of nickel in the soil tend to increase over the four month period?

   Yes

```
. xi: regress nickel i.month

i.month           _Imonth_5-9         (naturally coded; _Imonth_5 omitted)

      Source |       SS       df       MS              Number of obs =     248
```

```
-------------+------------------------------           F(  3,   244) =    7.39
       Model |  1593.53316     3  531.177721           Prob > F      =  0.0001
    Residual |  17536.6077   244  71.8713429           R-squared     =  0.0833
-------------+------------------------------           Adj R-squared =  0.0720
       Total |  19130.1408   247  77.4499629           Root MSE      =  8.4777

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imonth_6 |   1.740536   1.975111     0.88   0.379    -2.149906    5.630979
   _Imonth_8 |   4.265023   1.384627     3.08   0.002     1.537675     6.99237
   _Imonth_9 |   7.109776   1.581158     4.50   0.000     3.995315    10.22424
       _cons |   13.69724   1.113176    12.30   0.000     11.50458     15.8899
------------------------------------------------------------------------------
```

Examining the regression coefficients, there seems to be an increase in nickel over this period.

2. Examine the difference between the regression coefficients by month. Is it reasonable to assume that average amount of nickel increases linearly by month?

   Yes

3. Make a residual plot. Is there any evidence of outliers? Is there any evidence of heteroscedasticity?

   Yes and Yes

   ```
   rvfplot
   ```

   There are two obvious outliers and the variance appears lower at the lower fitted values (in May) versus the later months.

Now, assume the amount of nickel increases linearly by month, and the assumptions of linear regression continue to hold. Fit a model with `nickel` as the outcome and `month` modeled as a continuous explanatory variable. Call this model 2.

```
. regress nickel month

      Source |       SS       df       MS              Number of obs =     248
-------------+------------------------------           F(  1,   246) =   21.47
       Model |  1535.60852     1  1535.60852           Prob > F      =  0.0000
    Residual |  17594.5323   246  71.5224891           R-squared     =  0.0803
-------------+------------------------------           Adj R-squared =  0.0765
       Total |  19130.1408   247  77.4499629           Root MSE      =  8.4571

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
       month |   1.642664   .3545111     4.63   0.000     .9443998    2.340928
       _cons |   5.335122   2.646702     2.02   0.045     .1220335    10.54821
------------------------------------------------------------------------------
```

4. On average, nickel increases by _______ each month, from May to September.

   1.64

   This is simply the regression coefficient for month.

5. What is the 95% confidence interval for the average increase in nickel each month?

   (0.94, 2.34)

   This is the 95% confidence interval for the regression coefficient for month.

6. Given that the relationship between `month` and `nickel` appears to be linear, is it reasonable to use Model 2 to predict the amount of nickel in the soil in July 2010?

   Yes

   The month of July is within the range of the data.

7. Given that the relationship between `month` and `nickel` appears linear, is it reasonable to use Model 2 to predict the amount of nickel in the soil in October 2010?

   No.

   We would be extrapolating outside the range of our data.

8. What is the average amount of nickel in the soil during August 2010?

   - according to Model 1

     17.962263

```
. di 13.69724 + 4.26502

17.96226
```

OR

```
. lincom _cons +  _Imonth_8
```

```
 ( 1)  _Imonth_8 + _cons = 0

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   17.96226   .8234267    21.81   0.000     16.34033     19.5842
------------------------------------------------------------------------------
```

- according to Model 2

  18.476434

```
. di 5.335122 + 1.642664*8
```

OR

```
. lincom _cons + 8*month

 ( 1)  8*month + _cons = 0

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   18.47643   .5900412    31.31   0.000     17.31426    19.63861
------------------------------------------------------------------------------
```

9. True or False: Model 1 makes stronger modeling assumptions than Model 2.

   False

   Model 2 makes stronger assumptions than Model 1, because we model the relationship between month and average nickel level linearly.

Multiple Linear Regression. Again, we model the amount of nickel found in the sediment. Now, we model nickel as a function of both time (`month`) and location (`longitude`). The Gulf Coast in Louisiana is somewhat horizontal, so for simplicity we will ignore latitude in this problem.

1. Fit a linear regression model with `nickel` as the outcome and with `month` and `longitude` as explanatory covariates. Model `month` as a continuous variable, as in Model 2 from the previous question. Call this Model 3. Assume the assumptions of linear regression hold.

   Compare the adjusted R-squared from Models 2 and 3. Does the addition of `longitude` improve the adjusted R-square?

   Yes

   ```
   . regress nickel month longitude

         Source |       SS       df       MS              Number of obs =     248
   -------------+------------------------------           F(  2,   245) =   12.54
          Model |  1776.53846     2  888.269231           Prob > F      =  0.0000
       Residual |  17353.6024   245  70.8310301           R-squared     =  0.0929
   -------------+------------------------------           Adj R-squared =  0.0855
          Total |  19130.1408   247  77.4499629           Root MSE      =  8.4161

   ------------------------------------------------------------------------------
         nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -------------+----------------------------------------------------------------
          month |   1.678598   .3533309     4.75   0.000     .9826447    2.374552
      longitude |  -.7402628   .4013769    -1.84   0.066    -1.530852    .0503267
          _cons |  -62.10854   36.66326    -1.69   0.092    -134.3239    10.10687
   ------------------------------------------------------------------------------
   ```

   Adjusted R-squared is 0.086 for model 3, and is 0.077 in model 2.

2. Test whether the coefficient for `longitude` in the model is equal to 0 at the 0.05 level of significance.

   What is the coefficient?

   -.7402628

   The standard error of the coefficient?

   .4013769

   The test statistic?

-1.84

The distribution of the test statistic under the null?

t distribution with 245 degrees of freedom.

The p-value?

0.066

Your conclusion?

We have no evidence that, for a given month, average nickel levels have a linear relationship with longitude.

3. For each month, make a scatter plot with `longitude` on the x-axis and `nickel` on the y-axis to assess the linearity assumption. Conditional on month, does the relationship between nickel and longitude appear to be linear?

   No

In this example, our regression results show that we cannot assume a linear relationship between nickel levels and longitude, conditional on month. Incorporating location into statistical models is a whole genre of what is conveniently called spatial statistics. It is well beyond the scope of this course, but this is a brief introduction into “thinking spatially”. We need a more flexible model to reflect spatial heterogeneity.

More Multiple Regression

1. Examine the relationship between vanadium (another marker of oil in sediment) and nickel using a scatter plot. Does this relationship appear linear?

   Yes

   ```
   . twoway (scatter nickel vanadium)
   ```

2. How many outliers do you observe in this scatterplot?
   3

   As the investigator, it is important to determine the true cause of any observed outliers. If these are a result of error they should be removed, however outliers may also be indicators of a real but unusual occurrence that warrants further investigation.

   Fit a linear regression model with `nickel` as the outcome and `vanadium` as the explanatory covariate (Model 4).

```
regress nickel vanadium

      Source |       SS       df       MS              Number of obs =     248
-------------+------------------------------           F(  1,   246) =  736.50
       Model |  14340.3205     1  14340.3205           Prob > F      =  0.0000
    Residual |  4789.82031   246  19.4708143           R-squared     =  0.7496
-------------+------------------------------           Adj R-squared =  0.7486
       Total |  19130.1408   247  77.4499629           Root MSE      =  4.4126

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vanadium |   .4379549   .0161377    27.14   0.000     .4061692    .4697406
       _cons |   4.793981   .5407001     8.87   0.000     3.728989    5.858973
------------------------------------------------------------------------------
```

3. How much does nickel increase, on average, for a one mg/kg increase in vanadium?

   .4379549

4. How much does nickel increase, on average, for a 10 mg/kg increase in vanadium?

   4.379549

5. Add `month` to the linear regression model above to create Model 5. In Models 1 and 2 we saw that nickel levels seemed to increase each month. Does adding `month` to the model with `vanadium` as an explanatory covariate substantially improve the fit?

   No

```
. regress nickel vanadium month

      Source |       SS       df       MS              Number of obs =     248
-------------+------------------------------           F(  2,   245) =  370.45
       Model |  14376.2003     2  7188.10014           Prob > F      =  0.0000
    Residual |  4753.94057   245   19.403839           R-squared     =  0.7515
-------------+------------------------------           Adj R-squared =  0.7495
       Total |  19130.1408   247  77.4499629           Root MSE      =   4.405

------------------------------------------------------------------------------
      nickel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vanadium |    .446779   .0173678    25.72   0.000     .4125698    .4809882
       month |   -.270697   .1990685    -1.36   0.175    -.6628011     .121407
       _cons |   6.520048   1.379336     4.73   0.000     3.803179    9.236917
------------------------------------------------------------------------------
```

6. Think about the results from the various models. Compare the adjusted R-squared for Model 2 with that from Model 5. Does it appear that vanadium is a stronger predictor of nickel levels than month?

   Yes.

Pesticide Exposure and Cancer

Dr. Patel is interested in examining whether pesticide exposure is associated with a rare form of cancer. Since it is difficult to find enough cases to study, she decides to try and identify all cases in a large region. Since men and people younger than 50 are both more likely to be exposed to pesticides and more likely to develop this form of cancer, she obtains a random sample of people that did not develop the cancer living in the same region, matched to the case within 5 years of age and the same sex.

1. True or False: In this study, the matching eliminates the association between pesticide use and age.

   False: In a case-control study, matching eliminates the association between the confounder and the outcome, not the exposure.

2. First, Dr. Patel conducts an analysis that ignores the matching in the design and finds that the odds of developing cancer is 1.35 times greater among those exposed to the pesticide compared to those who were not exposed. If she then conducted another analysis that takes the matching into account, she would likely find an association of
   A) Less than 1.00
   B) 1.00
   C) Greater than 1.35
   D) Not enough information is provided.

   Since the first analysis did not account for the matching, it is biased toward the null and therefore, the adjusted analysis will be even farther from the null.

High Blood Pressure and Rate of Stroke

Use Stata and the NHLBI data set to create the two categories of high blood pressure (highbp1):

```
generate highbp1=.
replace highbp1=1 if (sysbp1>=140 | diabp1 >= 90)
replace highbp1=0 if (sysbp1<140 & diabp1 < 90)
```

(Note: There are no missing data on sysbp1 and diabp1. If data were missing on both sysbp1 and diabp1 then it should also be missing for highbp1. If data were missing on diabp1 only and sysbp1 >= 140 then highbp1 =1, otherwise highbp1 should be missing. Similarly, if data were missing on sysbp1 only and diabp1 >= 90 then highbp1 =1, otherwise highbp1 should be missing.)

1. What is the incidence rate ratio of stroke comparing those with high blood pressure to those without high blood pressure? Hint: The variable for stroke in the dataset is “stroke” and the number of years a person was followed for stroke is recorded in the “timestrk” variable.

   3.30

```
. ir stroke highbp1 timestrk

                 | highbp1                |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
Incident Stroke  |       255         160  |        415
Time [years] to  |  28887.76    59787.78  |   88675.54
-----------------+------------------------+------------
                 |                        |
  Incidence rate |  .0088273    .0026761  |     .00468
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Inc. rate diff. |         .0061511       |    .0049911    .0073112
 Inc. rate ratio |         3.298517       |    2.696321    4.045047 (exact)
 Attr. frac. ex. |         .6968334       |    .6291243    .7527841 (exact)
 Attr. frac. pop |         .4281748       |
                 +-------------------------------------------------
                     (midp)   Pr(k>=255) =                 0.0000 (exact)
                     (midp) 2*Pr(k>=255) =                 0.0000 (exact)
```

2. What is the incidence rate ratio (rounded to two decimal points) for the association between high blood pressure (highbp1) and the rate of stroke among men? Hint: Use the variable (sex1).

   2.27

3. What is the incidence rate ratio (rounded to two decimal points) for the association between high blood pressure (highbp1) and the rate of stroke among women?

   4.75

4. Conduct a test of homogeneity to evaluate whether the association between high blood pressure (highbp1) and the rate of stroke is different by sex. Based on this test, is there evidence that the difference between the sex-specific incidence rate ratios are more than just random sampling variability?

   [✓ A) Yes]
   B) No

```
. ir stroke highbp1 timestrk, by(sex1)

     Sex, exam 1 |       IRR       [95% Conf. Interval]   M-H Weight
-----------------+-----------------------------------------------------
            Male |   2.273711      1.697105    3.048007     29.75741
(exact)
          Female |    4.74629      3.537425    6.429301     21.94115
(exact)
-----------------+-----------------------------------------------------
           Crude |   3.298517      2.696321    4.045047
(exact)
    M-H combined |   3.323088      2.725029    4.052401
-----------------------------------------------------------------------
  Test of homogeneity (M-H)     chi2(1) =     12.84  Pr>chi2 = 0.0003
```

5. Based on these results, what are the options for properly reporting the association between high blood pressure (highbp1) and the rate of stroke?
   A) Sex-specific incidence rate ratios
   B) Pooled (Mantel-Haenszel) incidence rate ratio
   C) Standardized incidence rate ratio
   D) A or B
   [✓ E) A or C]

Confounding and Effect Modification

Dr. Smith conducts a randomized clinical trial to determine if aspirin reduces the risk of heart attack. Fifty percent of male patients are exposed and fifty percent of female patients are exposed. He finds an incidence rate ratio (IRR) of 0.75 comparing people who were assigned to take aspirin to those assigned to placebo. Among men, he finds that those assigned to take aspirin have an IRR=0.60 but among women, the IRR=0.95. Is gender a:

A) Confounder
B) Effect modifier
C) Both a confounder and an effect modifier

Problem Set 11

Death, Blood Pressure, and Age. We return to the Framingham cohort. We aim to assess the probability of death as a function of systolic blood pressure and age at baseline.

1. Fit a logistic regression model with death as your outcome and systolic blood pressure at baseline as the single predictor. (Assume "linearity on the logit-scale", i.e. the relationship between age and the probability of death is linear on the logit scale).

   Estimate the odds ratio of death for a one mmHg increase in systolic blood pressure and a corresponding 95% confidence interval.

   Odds ratio: 1.03017
   95% Confidence interval: 1.027   1.03335

```
. logistic death sysbp1

Logistic regression                             Number of obs     =      4434
                                                LR chi2(1)        =    410.44
                                                Prob > chi2       =    0.0000
Log likelihood = -2664.3807                     Pseudo R2         =    0.0715

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sysbp1 |    1.03017   .0016201    18.90   0.000        1.027     1.03335
       _cons |   .0099229   .0021315   -21.48   0.000     .0065133    .0151175
------------------------------------------------------------------------------
```

   Estimate the odds ratio of death for a ten mmHg increase in systolic blood pressure and a corresponding 95% confidence interval.

   Odds ratio: 1.346138
   95% Confidence interval: 1.305277   1.388277

```
. lincom sysbp1*10, or

 ( 1)  10*[death]sysbp1 = 0

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.346138   .0211704    18.90   0.000     1.305277    1.388277
------------------------------------------------------------------------------
```

2. Add age at baseline to the regression model above. (Assume linearity on the logit-scale holds).

   Estimate the odds ratio of death for a one mmHg increase in systolic blood pressure, conditional on age, and 95% confidence interval for this odds ratio.

   OR: 1.017406
   95% CI: 1.014031   1.020792

```
. logistic death sysbp1 age1

Logistic regression                               Number of obs   =       4434
                                                  LR chi2(2)      =     941.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -2398.8963                       Pseudo R2       =     0.1640

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sysbp1 |   1.017406   .0017247    10.18   0.000     1.014031    1.020792
        age1 |   1.104968   .0050887    21.67   0.000     1.095039    1.114987
       _cons |   .0003234   .0000929   -27.98   0.000     .0001842    .0005678
------------------------------------------------------------------------------
```

   Do you have evidence that age is a confounder of the relationship between systolic blood pressure and death? Considering the definition of a confounder, would you expect age to be a confounder?

   Yes, it appears that age is a confounder. Yes, we would expect age to be a confounder.

   Even though the change in the magnitude of the coefficient is small, when you consider a larger change in systolic blood pressure, such as 10 mmHg, we see that the odds ratio changes somewhat, depending on whether you condition on age.

```
. lincom 10*sysbp1, eform

 ( 1)  10*[death]sysbp1 = 0

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.188345   .0201448    10.18   0.000     1.149511    1.228491
------------------------------------------------------------------------------
```

3. Compare the model fit statistics between the model in question 1 and the model in question 2. True or false: Adding age to the model improves the model fit.

True

```
estat classification
```

Examining the classification statistics, the model with age seems to perform consistently as well or better than the model without age. We could also compare the ROC curves to further verify that the model with both age and systolic blood pressure is better than the model with just systolic blood pressure at predicting death.

Death, smoking, and age - We again use the Framingham cohort, but now aim to assess the probability of death as a function of smoking status.

1. Fit a logistic regression model examining the relationship between smoking and death. Test the hypothesis that smoking at baseline is associated with death during follow up at the 0.05 level of significance:
   - What is your test statistic

     1.61

   - What is the null distribution of your test statistic

     N(0,1)

   - What is the p-value

     0.107

   - What is your conclusion

     Fail to reject the null hypothesis

   - Estimate the odds ratio of death for smokers versus non-smokers, and a corresponding 95% confidence interval.

     OR: 1.106873

     95% CI: .9782858 1.252362

```
. logistic death cursmoke1

Logistic regression                               Number of obs   =       4434
                                                  LR chi2(1)      =       2.60
                                                  Prob > chi2     =     0.1070
Log likelihood = -2868.3016                       Pseudo R2       =     0.0005

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cursmoke1 |   1.106873   .0697412     1.61   0.107     .9782858    1.252362
       _cons |   .5110664   .0227584   -15.07   0.000     .4683519    .5576765
------------------------------------------------------------------------------
```

2. Use the logistic regression model in question 1 to estimate the probability of death in the cohort for non-smokers. For smokers.

   Non-smokers: .3382157

Smokers: .3613022

```
predict phat

table phat
```

3. Add age at baseline to the regression model in question 1. (Assume linearity on the logit-scale holds).

   - Estimate the odds ratio between smoking and death with 95% confidence interval, conditional on age at baseline.

     OR: 1.90

     95% CI: (1.64, 2.20)

   - Is there evidence of confounding by age?

     Yes, age is definitely a confounder.

```
. logistic death cursmoke1 age1

Logistic regression                               Number of obs   =       4434
                                                  LR chi2(2)      =     910.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -2414.4396                       Pseudo R2       =     0.1586

------------------------------------------------------------------------------
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cursmoke1 |   1.900338    .141002     8.65   0.000     1.643134    2.197802
        age1 |   1.133771   .0053181    26.77   0.000     1.123396    1.144243
       _cons |   .0006416   .0001671   -28.22   0.000     .0003851     .001069
------------------------------------------------------------------------------
```

4. Using the model in question 3, predict the probability of death for a 50 year old at baseline who is a smoker; and for a 50 year old who is a non-smoker.

   Smoker: 0.3936
   Non-smoker: 0.2546

```
. logit death cursmoke1 age1

Iteration 0:   log likelihood = -2869.6004
Iteration 1:   log likelihood = -2423.1073
Iteration 2:   log likelihood =  -2414.454
Iteration 3:   log likelihood = -2414.4396
Iteration 4:   log likelihood = -2414.4396

Logistic regression                               Number of obs   =       4434
                                                  LR chi2(2)      =     910.32
                                                  Prob > chi2       =     0.0000
Log likelihood = -2414.4396                       Pseudo R2         =     0.1586

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   cursmoke1 |   .6420315   .0741984     8.65   0.000     .4966053    .7874578
        age1 |   .1255496   .0046907    26.77   0.000      .116356    .1347431
       _cons |  -7.351551   .2604765   -28.22   0.000    -7.862076   -6.841027
------------------------------------------------------------------------------

. di invlogit(  -7.351551 + .1255496*50)
.25462966

. . di invlogit(  -7.351551 +  .6420315 + .1255496*50)
.39363942
```

5. Test for effect modification of the relationship between death and smoking status by age.
   - What is the absolute value of your test statistic?

     1.66

   - What is the null distribution of your test statistic

     N(0,1)

   - What is the p-value

     0.096

   - What is your conclusion

     Fail to reject the null. No evidence of effect modification.

   - Could you perform this same test using contingency table methods?

     No.

   - Examine the sign of the coefficients in the model. True or false: in this model, the differences between the log-odds of death between smokers and non-smokers decrease as age increases.

     True

```
. xi: logit death i.cursmoke1*age1
i.cursmoke1       _Icursmoke1_0-1     (naturally coded; _Icursmoke1_0 omitted)
i.cursmo~1*age1   _IcurXage1_#        (coded as above)

Iteration 0:   log likelihood = -2869.6004
Iteration 1:   log likelihood = -2423.2408
Iteration 2:   log likelihood =  -2413.104
Iteration 3:   log likelihood = -2413.0545
Iteration 4:   log likelihood = -2413.0545

Logistic regression                               Number of obs   =       4434
                                                  LR chi2(3)      =     913.09
                                                  Prob > chi2     =     0.0000
Log likelihood = -2413.0545                       Pseudo R2       =     0.1591

------------------------------------------------------------------------------
       death |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Icursmoke1_1 |   1.456121   .4962559     2.93   0.003     .4834771    2.428764
        age1 |   .1335346   .0068082    19.61   0.000     .1201907    .1468784
 _IcurXage1_1 |  -.0156416   .0094101    -1.66   0.096    -.0340851    .0028019
       _cons |  -7.787368   .3754011   -20.74   0.000     -8.52314   -7.051595
------------------------------------------------------------------------------
```

Remember – in the logistic regression models above, we have only looked at a few predictors of death; only considered logistic-linear relationship between the continuous covariates and probability of death; and only considered some interactions. Consequently, we cannot infer any causal relationship from the data above (only data associations). Further, if we were aiming to predict probability of death, a more complicated model would likely improve our predictions. When constructing models, there are always tradeoffs, for instance between complexity, precision, accuracy, and interpretation.

Additionally, in this analysis, we could use more information about the study participants if we considered time to death as our outcome (rather than just using death). "Time to event" analysis, or survival analysis, is the subject of next week's lectures!

Binary Exposure and Binary Outcome Model

1. When analyzing results from a case-control study with a binary exposure and binary outcome, which type of regression should be used?

   A) Logistic regression
   B) Linear regression
   C) Cox proportional hazards regression.

Obesity and Stroke Regression Model

1. Dr. Johnson decides to construct a Cox proportional hazards regression model to examine the association between obesity and stroke adjusting for age, gender and smoking status. The beta coefficient for obesity is 0.30. What is the proper interpretation of the association between obesity and stroke?

   A) Participants who are obese have 0.30 times the rate of stroke compared to participants who are not obese.
   B) After adjusting for age, gender and smoking status, participants who are obesity have 0.30 times the rate of stroke compared to participants who are not obese.
   C) After adjusting for age, gender and smoking status, participants who are obese have 1.35 times the rate of stroke compared to participants who are not obese.
   D) After adjusting for age, gender and smoking status, participants who are obese have 1.35 times the odds of stroke compared to participants who are not obese.
   E) Participants who are obese have 1.35 times the rate of stroke compared to participants who are not obese.

Propensity Score vs. Randomization

1. True/False: Using a propensity score to adjust for confounding in a longitudinal cohort study would yield identical results to those obtained from a randomized clinical trial because both methods control for all imbalances between the exposed and unexposed groups.

   False

Smoking and Death Logistic Regression, revisited. In lecture, the following logistic regression model for the association between smoking and death was presented:

> Log(P/(1-P)) = -7.5869 + 0.5522(CURSMOKE1) + 0.1181(AGE1)+ 0.7759(MALE) + 0.6386(HIGHBP1) + 1.5834(DIABETES1)

1. Using the above model, what is the odds ratio of death for a 50 year old man who does not smoke, has high blood pressure and does not have diabetes (ie AGE1=50, CURSMOKER1=0, MALE=1, HIGHBP1=1 and DIABETES=0) compared to a 50 year old woman who does not smoke, does not have high blood pressure and does not have diabetes (ie AGE1=50, CURSMOKER1=0, MALE=0, HIGHBP1=0 and DIABETES=0)?

   4.11443

2. Does the answer to the previous question change if different values are set for AGE1, MALE, HIGHBP1 and DIABETES1?

   No

3. What is this model's estimate for the odds ratio of death for a diabetic (DIABETES1=1) compared to a non-diabetic (DIABETES= 0), controlling for MALE, HIGHBP1 and CURSMOKE1)?

   4.871

4. What is this model's estimate for the odds ratio of death for a smoker (CURSMOKE1=1) compared to a non-smoker (CURSMOKE1= 0), controlling for MALE, HIGHBP1 and DIABETES)?

   1.73707

The following model contains the same the same risk factors listed in the previous model, except that it does not include age.

> Log(P/(1-P)) = -1.4584 + 0.1205(CURSMOKE1) + 0.6902(MALE) + 1.0304(HIGHBP1) + 1.81.44(DIABETES1)

5. What is this model's estimate for the odds ratio of death for a smoker (CURSMOKE1=1) compared to a non-smoker (CURSMOKE100), controlling for MALE, HIGHBP1 and DIABETES1).

   1.128

6. Based on these two models, what conclusion can you reach about AGE1 being a confounder, when estimating the effect of smoking on the odds of dying, once you control for MALE, HIGHBP1, and DIABETES1?

   a) AGE1 appears not to be a confounder because the Odds Ratio for CURSMOKE1 is different in the two models
   b) AGE1 appears to be a confounder because the Odds Ratio for CURSMOKE1 is different in the two models
   c) No conclusion can be reach because AGE1 is not included in the second model

Problem Set 12 Solutions

Time to death and Systolic Blood Pressure. For this problem set, we will return to the Framingham data set. We will examine time to death (in years), `timedth`, where the variable death, is the censoring indicator.

1. Out of the 500 people, how many died?

   161

2. Suppose we want to look at the effect of systolic blood pressure on time to death. Let's classify those people with a systolic blood pressure greater than 140 mmHg at exam 1 as having high systolic blood pressure.

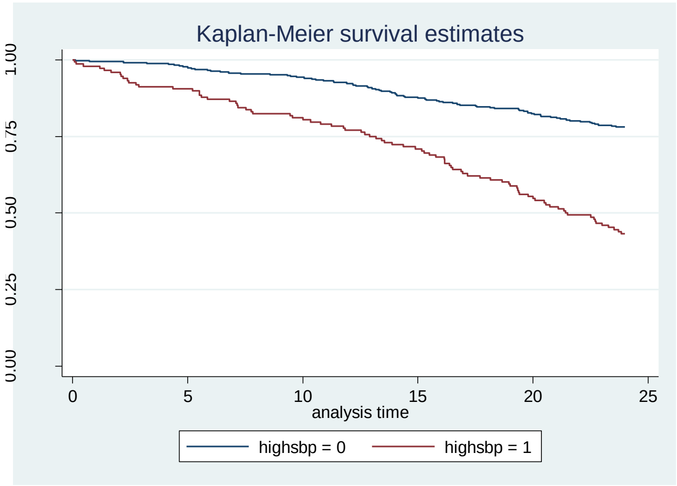
   - How many people have high systolic blood pressure?

     ```
     generate highsbp = 1 if sysbp1 > 140
     replace highsbp = 0 if sysbp1 <= 140 & sysbp1 < .
     tab highsbp
     ```

     148

Plot the Kaplan Meier estimates of the survival function for those do and do not have high blood pressure.

```
stset timedth, failure(death)
sts graph, by(highsbp)
```



- What is the probability of surviving beyond 2 years in the group with high blood pressure? (Hint: use the following command: sts list, by(highsbp)

  0.9595

- What is the probability of surviving beyond 2 years in the group without high blood pressure?

  0.9943

3. Conduct a log rank test to determine if the distributions of survival time differ between those with and without high systolic blood pressure. Use a 0.05 level of significance.

   - What is your test statistic?

     64.48
   - True or false: the test statistic is a random variable that follows a chi-square distribution with one degree of freedom under the null hypothesis.

     True

   - Is your p-value less than 0.05?

     Yes

   - This analysis tells us that we can significantly improve someone's life expectancy if we lower their systolic blood pressure from above 140 mmHg to below 140 mmHg. True or False?

     False. We are testing association, not causation. We can say that having a high (> 140 mmHg) systolic blood pressure is associated with a shorter life span compared to those without (<= 140 mmHG) high systolic blood pressure. There may be another variable that is associated with both death and high systolic blood pressure, which could confound the relationship between death and sbp, thereby preventing us from making causal statements.

Time to Death, Age, and Systolic Blood Pressure. For this question, use the full Framingham dataset (fhs.dta), posted on the webpage. Age is likely a confounder in the relationship between systolic blood pressure and time to death. In this problem, we will conduct two logrank tests – for older and younger subsets of the Framingham cohort – and compare the results.

1. First, let's examine some descriptive statistics, focusing on patients age 35-40.

```
tabulate highsbp death if age1 >= 35 & age1 <= 40, row

           |    Death indicator
   highsbp |        No        Yes |     Total
-----------+----------------------+----------
         0 |       584         78 |       662 
           |     88.22      11.78 |    100.00 
-----------+----------------------+----------
         1 |        56         12 |        68 
           |     82.35      17.65 |    100.00 
-----------+----------------------+----------
     Total |       640         90 |       730 
           |     87.67      12.33 |    100.00
```

- How many participants were between age 35-40 at baseline?

  730

  How many of those participants died during follow up?

  90

- Among participants age 35-40 at baseline, how many had high blood pressure.

  68

  What percent of patients with high blood pressure died?

  17.65

  What percent of patients without high blood pressure died?

  11.78

- How many participants were between age 65-70 at baseline?

  203

  How many of those participants died during follow up?

  174

2. Now, we look at patients age 65-70.

```
. tabulate highsbp death if age1 >= 65 & age1 <= 70, row

           |    Death indicator
   highsbp |        No        Yes |     Total
-----------+----------------------+----------
         0 |        13         68 |        81
           |     16.05      83.95 |    100.00
-----------+----------------------+----------
         1 |        16        106 |       122
           |     13.11      86.89 |    100.00
-----------+----------------------+----------
     Total |        29        174 |       203
           |     14.29      85.71 |    100.00
```

- Among participants age 65-70 at baseline, how many had high blood pressure?

  122

  What percent of patients with high blood pressure died?

  86.89

  What percent of patients without high blood pressure died?

  83.95

- Plot the survival curves for those age 35-40 and age 65-70. True or false: when stratifying by age, the difference between the survival curves is smaller than in the previous analysis, when we did not stratify.

  True.

3. Conduct a log rank test to determine if the distributions of survival time differ between those with and without high systolic blood pressure among **the subset of participants ages 35-40 at baseline**. Use a 0.05 level of significance.

```
sts test highsbp if age1 >= 35 & age1 <= 40, logrank

         failure _d:  death
   analysis time _t:  timedth


Log-rank test for equality of survivor functions

        |   Events         Events
highsbp |  observed       expected
--------+-------------------------
0       |        78          81.99
1       |        12           8.01
--------+-------------------------
Total   |        90          90.00

              chi2(1) =       2.18
              Pr>chi2 =     0.1394
```

- What is your test statistic?

  2.18

- What is the null distribution of your test statistic?

  Chi-square with 1 df

- Is your p-value less than 0.05?

  No

- What do you conclude?

  No evidence of a difference in the survival curves by high blood pressure among 35-40 year olds.

4. Conduct a log rank test to determine if the distributions of survival time differ between those with and without high systolic blood pressure among **the subset of participants ages 65-70 at baseline**. Use a 0.05 level of significance.

```
.
. sts test highsbp if age1 >= 65 & age1 <= 70, logrank

         failure _d:  death
   analysis time _t:  timedth


Log-rank test for equality of survivor functions

        |   Events          Events
highsbp |  observed       expected
--------+-------------------------
0       |        68          76.65
1       |       106          97.35
--------+-------------------------
Total   |       174         174.00

              chi2(1) =
              Pr>chi2 =     0.1859
```

- What is your test statistic?

  1.75

- What is the null distribution of your test statistic?

  Chi-square with 1 df

- Is your p-value less than 0.05?

  No

- What do you conclude?

  No evidence of a difference in the survival curves by high blood pressure among 65-70 year olds.

.