BIOS 662, Fall 2018

Homework 2

Assigned: Tuesday, September 4

Due: Tuesday, September 11

1. In the "Datasets" sub-folder of the "Homework materials" folder under "Resources" on the Sakai web site for this course, there is a dataset "HW2_SBP.txt". The data are systolic blood pressures of 40 women who had an MI (myocardial infarction, i.e. heart attack) less than two years after their blood pressure was measured and of 160 women who did not have an MI within two years. Each row in the dataset corresponds to one woman. The first observation in a row is 1 for those who had an MI within two years and 0 otherwise. The second observation is the systolic blood pressure.
   (a) Use R or SAS to draw a histogram and boxplot of systolic blood pressure for all 200 women (that is, do not separate those who did and did not have an MI).
   (b) Using the definition of percentile from the class notes, compute the 25th, 50th (i.e., median) and 75th percentiles.
   (c) Determine the IQR.
   (d) Find the largest observation $\leq 75^{\text{th}}$ percentile + 1.5 IQR and the smallest observation $\geq 25^{\text{th}}$ percentile $-$ 1.5 IQR (i.e., the extent of the "whiskers"). Based on these results, does the computed boxplot appear to agree with the definition of a boxplot from our notes? If not, investigate the discrepancy and report your findings.
   (e) Use a plot to compare the distribution of systolic blood pressure in those who had an MI against that of those who did not. Do blood pressures in the two groups appear to differ? If so, in what direction?

2. The dataset "HW2_PGE.txt" in the "Datasets" sub-folder contains the data in Table 3.20 of the textbook. The last value in each row is 1 for the patients with hypercalcemia and 0 otherwise.
   (a) Obtain the mean and standard deviation of plasma iPGE separately for patients with and without hypercalcemia. Do you think there is enough evidence to conclude that the means of the two groups differ? (Later in this course we will study more formal ways to compare the two means.)
   (b) Do part (c) of Problem 3.15 of the textbook.
   (c) The values for one patient appear to be particularly anomalous. Identify this patient. Suppose it was determined that there had been an

error in measuring the patient's serum calcium. Suggest a value for serum calcium that would be more consistent with the patient's plasma iPGE value and the pattern in the rest of the data.

(d) Without re-doing any of your calculations, what effect do you think changing the serum calcium value to the one you suggested would have on the means and standard deviations in the first part of this problem?

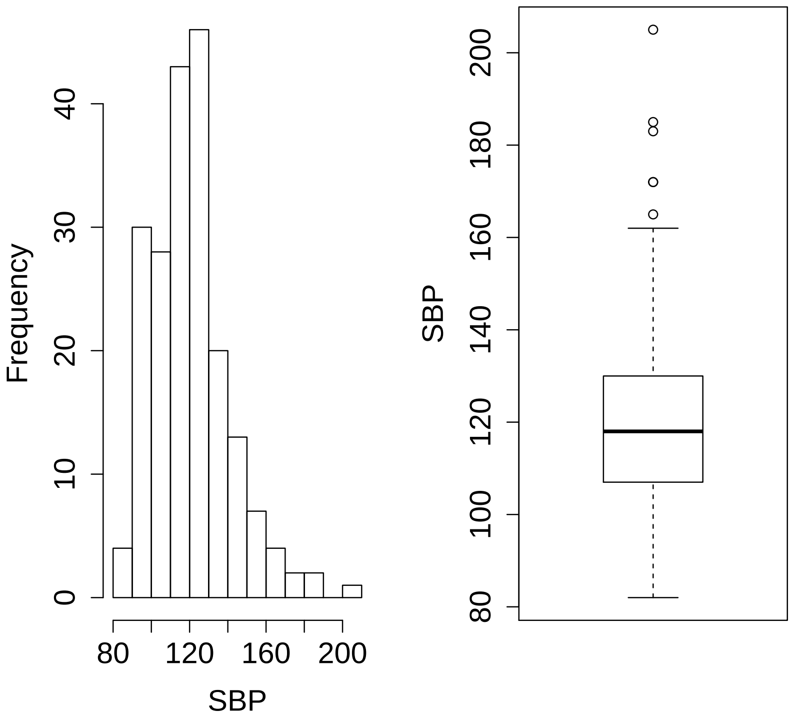
BIOS 662

Homework 2 Solution

September, 2018

Question 1

(a) Using the R functions hist() and boxplot(), we get the following output:



(b) Because $n = 200$ and $p = 0.25$, $np = 50$ is an integer, so the 25th percentile is given by

$$\hat{\zeta}_{0.25} = \frac{y_{(50)} + y_{(51)}}{2} = \frac{107 + 107}{2} = 107$$

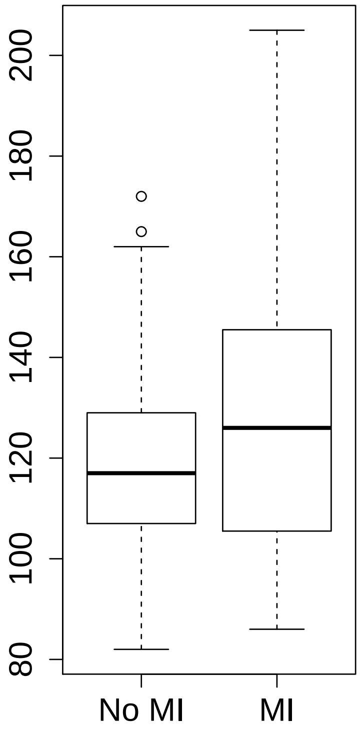
Similarly, one can show $\hat{\zeta}_{0.5} = 118$ and $\hat{\zeta}_{0.75} = 130$.

(c) Thus the IQR equals $130 - 107 = 23$.

(d) The 75th percentile + 1.5 IQR $= 130 + 1.5 \times 23 = 164.5$ and the largest observation less than this is 162. Likewise, the 25th percentile − 1.5 IQR $= 107 - 1.5 \times 23 = 72.5$ and the smallest observation greater than this is 82. Looking at the histogram we see that there are no outliers atthe lower end of the distribution, which is why in the boxplot there are no individual observations plotted below the lower whisker. The results agree with R exactly:

```
> boxplot(sbpall)$stats
     [,1]
[1,]   82
[2,]  107
[3,]  118
[4,]  130
[5,]  162
```

(e) The following figure shows side-by-side boxplots for the two groups. SBP tends to be higher in the MI group, with its median almost as high as the third quartile of the no MI group and its upper whisker extending substantially beyond the largest SBP in the no MI group.



Question 2

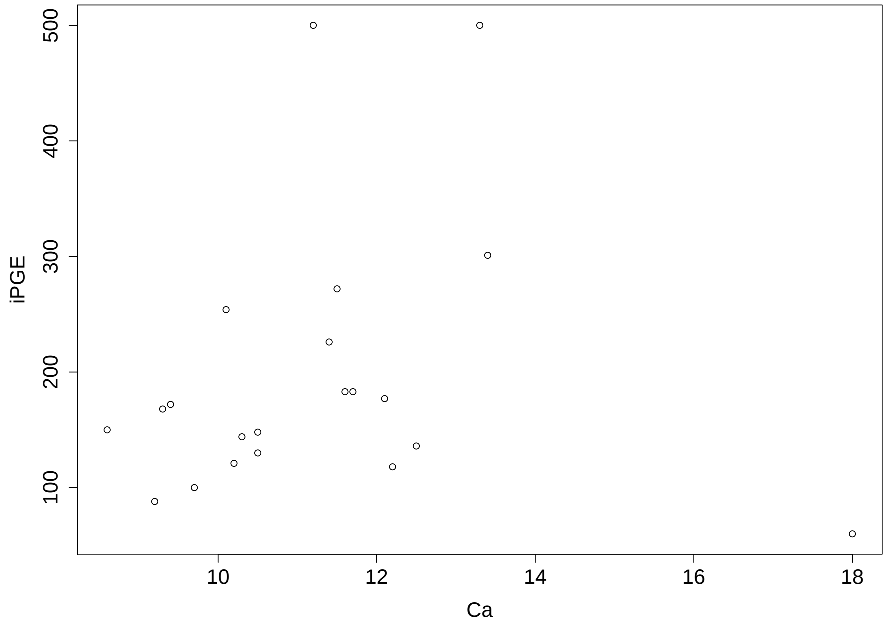
(a) $\bar{X}_{\text{Hypercalcemia}} = 2656/11 = 241.5; \quad \bar{X}_{\text{Normocalcemia}} = 1475/10 = 147.5.$

$s^2_{\text{Hypercalcemia}} = \frac{1}{11-1}\left(849988 - 11 \cdot 241.5^2\right) = 20868.47; \quad \text{so} \quad s = 144.46.$

$s^2_{\text{Normocalcemia}} = \frac{1}{10-1}\left(236749 - 10 \cdot 147.5^2\right) = 2131.83; \quad \text{so} \quad s = 46.17.$

The means seem to be substantially different. We'll see in the coming weeks that we need to use information about the standard deviations in order to decide whether the corresponding population means really do appear to differ.

(b) Below is a scatterplot of plasma iPGE against serum Ca.



If we ignore a few outliers, there is some evidence that higher plasma iPGE levels tend to be associated with higher serum Ca levels. There is substantial variability from person to person though, so if person A has higher serum Ca than person B it does not automatically follow that person A will have higher plasma iPGE than person B.

(c) Patient #11 has the highest serum Ca value among all patients yet has the lowest plasma iPGE level. A serum calcium level below 10 would be more in keeping with the tendency for lower plasma iPGE to be associated with lower serum Ca.

(d) Patients with serum calcium above 10.5 mg/dL are classified as hypercalcemic. If the serum calcium value for patient #11 is really below 10, then this patient would be classified as not having hypercalcemia and so would be moved from one group to the other. Because this patient has plasma iPGE level well below all the others in the

Hypercalcemia group, moving this patient out of the group would result in a larger mean plasma iPGE value for the group. By removing a value far from the mean, the standard deviation will decrease. The patient's plasma iPGE level is also lower than all values in the other group, so moving the patient into that group would lower its mean plasma iPGE value and because the newly added value is more extreme than other values in the group, the standard deviation will increase.

Suppose we change the serum calcium value for patient #11 from 18 to 10. Then the sample means and standard deviations of plasma iPGE change from

$\bar{X}_{\text{Hypercalcemia}} = 241.5; \quad \bar{X}_{\text{Normocalcemia}} = 147.5.$

$s_{\text{Hypercalcemia}} = 144.46; \quad s_{\text{Normocalcemia}} = 46.17$

to

$\bar{X}_{\text{Hypercalcemia}} = 259.6; \quad \bar{X}_{\text{Normocalcemia}} = 139.5.$

$s_{\text{Hypercalcemia}} = 138.43; \quad s_{\text{Normocalcemia}} = 57.13.$

BIOS 662, Fall 2018

Homework 3

Assigned: Tuesday, September 18

Due: Tuesday, September 25

1. This is a continuation of problem #2 from Homework 2 involving the dataset "HW2_PGE.txt". Use the data for all patients combined without regard to hypercalcemia status.
   (a) Use a plot to decide whether the distribution of serum calcium is approximately normal.
   (b) Calculate the sample mean and standard deviation for serum calcium, and construct a 95% confidence interval for the population mean serum calcium of such patients.
   (c) Suppose that the sample size is doubled (but yielding the same mean and standard deviation). Determine the percentage change in the width of the 95% confidence interval. Repeat assuming a sample three times the original size.
   (d) Use the bootstrap method to obtain a 95% confidence interval for the population mean serum calcium of such patients.
   (e) Calculate the sample median and obtain an exact 95% confidence interval for the population median serum calcium of such patients.

2. Problem 4.20 on page 111 of the textbook.

BIOS 662

Homework 3 Solution

September, 2018

Question 1

(a) A QQ plot of serum calcium is given on page 3. Most of the points lie close to a straight line but the point in the top right corner is way off the line. Recall that in part 2(c) of Homework 2 it was suggested that one of the calcium values was recorded incorrectly. It is the anomalous value that is in the top right corner. If this value is changed from 18 to 7.5 we get the second of the QQ plots on page 3. In that one all the points lie reasonably close to the line. Later in the semester we will look at a more formal test of normality.

For the rest of the question we will use the uncorrected calcium value.

(b) The population variance is unknown, so we have to handle this either as a small sample from the normal distribution (with unknown variance) or a "large" sample from an unknown distribution. Here $n = 21$, $\bar{Y} = 11.27$, $s^2 = 4.164$ and $s = 2.041$.

If this is a small sample from the normal distribution, a 95% CI for μ is:

$$\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 2.086 \times 2.041/\sqrt{21} = (10.34,\ 12.20).$$

If this is a large sample from an arbitrary distribution, using the Central Limit Theorem and Slutsky's Theorem a 95% CI for μ is:

$$\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 1.96 \times 2.041/\sqrt{21} = (10.40,\ 12.14).$$

The sample is a little too small to qualify as a large sample and if we leave the value of 18 uncorrected, the normality assumption is also questionable, so neither CI is very satisfactory here.

(c) If in (b) we assume the sample is from the normal distribution, the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.086 \times 2.041/\sqrt{21} = 1.86.$$

Doubling the sample size changes the degrees of freedom of t, so $t_{n-1,1-\alpha/2} = t_{41,0.975} = 2.02$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.02 \times 2.041/\sqrt{42} = 1.27.$$

This is a reduction of $100 \times (1.86 - 1.27)/1.86 = 31.5\%$.

Tripling the sample size, $t_{n-1,1-\alpha/2} = t_{62,0.975} = 2.00$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.00 \times 2.041/\sqrt{63} = 1.03.$$

This is a reduction of $100 \times (1.86 - 1.03)/1.86 = 44.7\%$.

If in (b) we assume the sample is from an arbitrary distribution, the width of the confidence interval is

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{21} = 1.75.$$

Now doubling the sample size changes just the $\sqrt{n}$ part and the width of the confidence interval becomes

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{42} = 1.23.$$

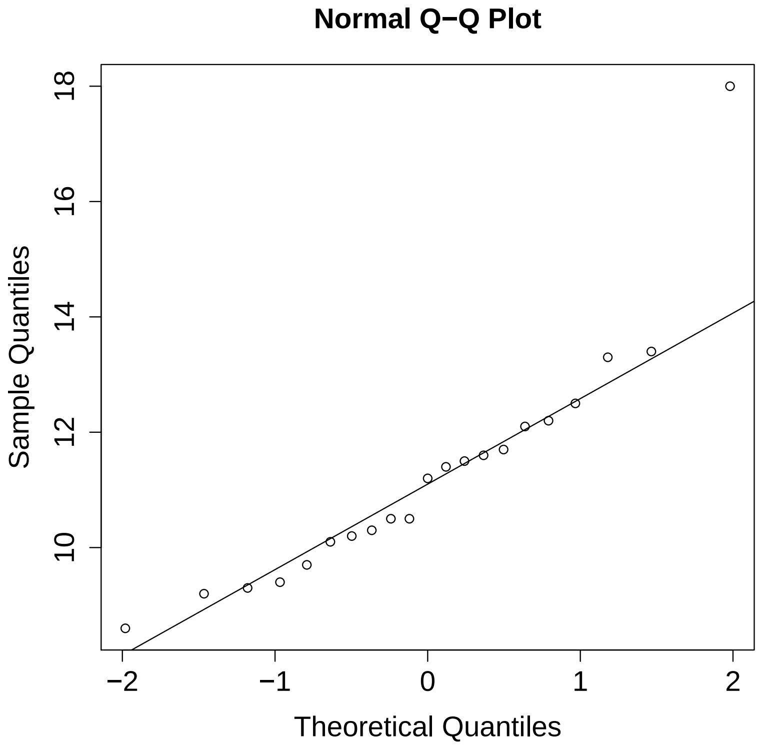
This is a reduction of $100 \times (1.75 - 1.23)/1.75 = 29.3\%$.

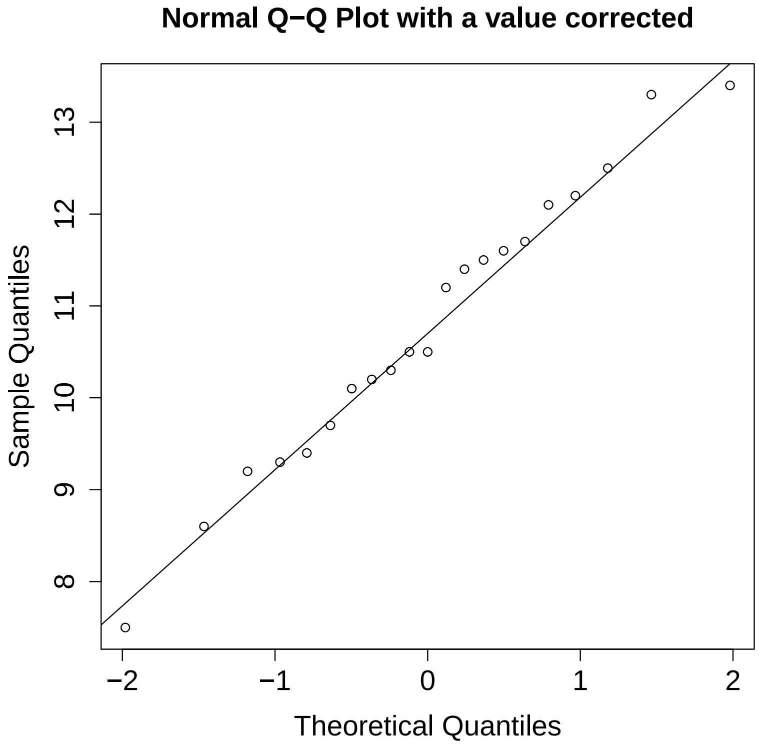
Tripling the sample size the width is $2 \times 1.96 \times 2.041/\sqrt{63} = 1.01$, which is a reduction of $100 \times (1.75 - 1.01)/1.75 = 42.3\%$.

In summary, although the term $t_{n-1,0.975}$ decreases somewhat with increasing sample size, the change in it is relatively small. Most of the change is because of the $1/\sqrt{n}$ part and even then the change is at the rate of $\sqrt{n}$.

(d) The R code I used is given on a subsequent page. (R does have a package, called "boot", for doing boostrapping. But I wanted to demonstrate the principles involved rather than just getting a confidence interval.)

For the particular value in set.seed this yielded 95% CI (10.48, 12.49). Different seeds will yield somewhat different confidence intervals. I tried several values and all gave intervals of the form with lower end 10.4X or 10.5X and upper end 12.4X).


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
18
16
14
12
10
−2
−1
0
1
2


Normal Q−Q Plot with a value corrected
Sample Quantiles
Theoretical Quantiles
13
12
11
10
9
8
−2
−1
0
1
2

R code for the bootstrap-t interval for 1(c):

```
# bootstrap-t interval
set.seed(64353)
mean.Ca<-mean(pge$Ca)
var.Ca<-var(pge$Ca)
n.Ca<-length(pge$Ca)
boots <- 500
zs <- matrix(0,1,boots)
for (jj in 1:boots){
   ysamp <- sample(pge$Ca,size=n.Ca,replace=T)
   zs[jj] <- (mean(ysamp)-mean.Ca)/sqrt(var(ysamp)/n.Ca)
   }
lower.t <- quantile(zs,.975)
upper.t <- quantile(zs,.025)

lower.y <- mean.Ca - lower.t*sqrt(var.Ca/n.Ca)
upper.y <- mean.Ca - upper.t*sqrt(var.Ca/n.Ca)

lower.y
upper.y
```

Programming bootstrap-t intervals in SAS is somewhat more complicated than in R. Below is SAS code, assuming that the data have been read in to a dataset called “pge”. For the particular seed used in “proc surveyselect”, this yielded 95% CI (10.58, 12.46).

```
proc means data=pge noprint;
  var Ca;
  output out=original mean=original_mean std=original_std n=n;

data original;
  set original;
one=1;  *** For later merging with bootstrap samples;
keep one original_mean original_std n;

proc surveyselect data=pge out=pge_samples seed=45921
    rep=500 sampsize=21 method=urs outhits;
*** rep - specifies the number of replicates
*** method=urs - "requests unrestricted random sampling, which is
***          selection with equal probability and with replacement."
*** outhits - when an observation is selected more than once, the
***          output dataset has a separate row for each occurrence,
***          rather than a single row plus a count of the number of
***          occurrences. ;
```

```
proc means data=pge_samples noprint;
   var Ca;
   by Replicate;
   output out=bootout mean=mean stderr=stderr;

data bootout;
   set bootout;
one=1;

data bootout;
   merge original bootout;
   by one;
zb=(mean-original_mean)/stderr;

proc univariate noprint;
   var zb;
   output out=outpctl pctlpre=P_ pctlpts= 2.5 97.5;

data outpctl;
   set outpctl;
one=1;

data outpctl;
   merge original outpctl;
   by one;

lower=original_mean - P_97_5*original_std/sqrt(n);
upper=original_mean - P_2_5*original_std/sqrt(n);

proc print data=outpctl;
  var lower upper;
```

(e) One way is to do the calculations “manually”, using the method on page 43 of the notes on “Point and Interval Estimation”. There are 21 patients in the dataset. So we need to find the largest r such that

$$\frac{1}{2^{21}} \sum_{i=0}^{r-1} \binom{21}{i} \leq \alpha/2$$

We can use R to find the largest k such that `sum(dbinom(0:k,21,0.5))`≤ 0.025 or, equivalently, such that `pbinom(k,21,0.5)`≤ 0.025.

Because `sum(dbinom(0:5,21,0.5)) = 0.0133` and `sum(dbinom(0:6,21,0.5)) = 0.0392`, $k = 5$ and thus $r-1 = k = 5$. Hence $r = 6$ and $n-r+1 = 21-6+1 = 16$.

A 95% CI for the median is $(X_{(6)}, X_{(16)})$. Sorting the serum calcium values, the 6th and 16th order statistics are 10.1 and 12.1, so the CI is $(10.1, 12.1)$.

Another way is to use SAS:

```
proc univariate cipctldf(type=symmetric);
   var ca;
```

```
                     Quantiles (Definition 5)

                    95% Conf. Limits   ------Order Statistics-----
Quantile  Estimate Distribution Free  LCL Rank UCL Rank  Coverage

100% Max      18.0
99%           18.0   .            .         .        .         .
95%           13.4  13.3         18.0      19       21     57.45
90%           13.3  12.2         18.0      17       21     83.84
75% Q3        12.1  11.4         13.4      12       20     96.03
50% Median    11.2  10.1         12.1       6       16     97.34
25% Q1        10.1   9.2         10.5       2       10     96.03
10%            9.3   8.6          9.7       1        5     83.84
5%             9.2   8.6          9.3       1        3     57.45
1%             8.6   .            .         .        .         .
0% Min         8.6
```

This also has the 95% CI for the median as $(10.1, 12.1)$. The 97.34 in the "Coverage" column is obtained as $100 \cdot (1 - 2 \times 0.0133) = 97.34$.

Question 2 – Problem 4.20 on page 111

(a) If $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

Here $\mu = 1.0$, $\sigma^2 = 9.0$ and $n = 9$, so $\bar{Y} \sim N(1, 9/9) = N(1, 1)$. That is, the sampling distribution of $\bar{Y}$ is normal with mean 1 and variance 1.

(b) Standardizing by subtracting the mean of $\bar{Y}$ and dividing by the standard error,

$$\begin{aligned}
\Pr[1 < \bar{Y} \leq 2.85] &= \Pr\left[\frac{1-1}{\sqrt{1}} < \frac{\bar{Y}-1}{\sqrt{1}} \leq \frac{2.85-1}{\sqrt{1}}\right] \\
&= \Pr[0 < Z \leq 1.85] = \Pr[Z \leq 1.85] - \Pr[Z \leq 1] \\
&= \Phi(1.85) - \Phi(1) = 0.9678 - 0.5 = 0.4678.
\end{aligned}$$

(c) Using properties on page 22 of the notes on "Statistical Inference: Populations and Samples", if $\bar{Y} \sim N(1, 1)$ and $W = 4\bar{Y}$, then $W \sim N(4 \times 1, 4^2 \times 1) = N(4, 16)$.

BIOS 662, Fall 2018

Homework 4

Assigned: Thursday, September 27

Due: Thursday, October 4

Instructions: For the problems below, confidence intervals and testing procedures should be done "by hand." You may use appropriate software such as R or SAS to estimate means and variances if these are needed. You should also feel free to use the software to verify any results. For problems involving testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution, the critical region, whether you reject the null, the p-value, and an interpretation of the results in a language suitable for investigators. (Get into the habit of supplying these, not just for this homework.) All tests should be performed at the $\alpha = 0.05$ significance level.

1. This is based on Problem 5.2 on page 142 of the textbook. "In data of Dobson et al. [1976], 36 patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford-Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford-Binet." The dataset "HW4_PKU.txt" contains data on the 21 pairs *not* listed in Problem 5.2, which is why numbering of pairs in the dataset starts with 16. For parts (a)–(d) assume IQ data are normally distributed.

   (a) State a suitable null and an alternative hypotheses with regard to these data.

   (b) Test the null hypothesis (using $\alpha = 0.05$).

   (c) Give a 95% confidence interval for the true effect of PKU on IQ.

   (d) State your conclusions.

   (e) What are your assumptions?

   (f) Now suppose we cannot assume normality and need to use the sign test. State the hypotheses, conduct the test and state your conclusions.

   (g) Discuss how and why your conclusions in parts (d) and (f) differ.

2. The following data concern the association between sodium chloride (salt) intake and hypertension. Fifteen hypertensive and twelve normotensive subjects were isolated for a week so that their sodium (Na^+) intakes could be measured accurately. The average daily (Na^+) intakes (in milligrams) are listed in the table below. Compare the average daily (Na^+) intake of the hypertensive subjects with that of the normal volunteers using an appropriate statistical test. Include a justification for the statistical test employed.

Hypertensive	Normal
1100	1000
1320	1220
1350	1300
1450	1400
1600	1555
1850	1600
1900	1780
1990	1780
2050	1900
2120	2020
2200	2350
2210	2375
2500	
2610	
2720	

BIOS 662

Homework 4 Solution

October, 2018

Question 1

The data come from pairs of children (a PKU case and his/her normal sibling). Because the children in a pair are siblings, they cannot be regarded as independent. So it is not appropriate to conduct two-sample tests. Instead, we conduct one-sample tests on the difference between the IQ test scores within each sibling pair.

(a) Let Y_i denote the IQ of the PKU case minus that of his/her normal sibling in pair i. If IQ is normally distributed, then the IQ of the differences is also normally distributed, so it may be reasonable to assume that $Y_1, \ldots, Y_n$ are iid with $Y_i \sim N(\mu, \sigma^2)$, for some μ and σ^2, where μ is the mean difference in IQ between a PKU case and his/her closest-age normal sibling.

Hypotheses: $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$. (A one-sided alternative may also be reasonable, if we think there is no chance the dietary therapy could be so effective as to reverse the direction of association.)

(b) Assuming that the Y_i are normally distributed but that σ^2 is unknown, we use a one-sample t-test. Here $n = 21$, so

$$C_\alpha = \{t : |t| > t_{n-1,1-\alpha/2}\} = \{t : |t| > t_{20,0.975}\} = \{t : |t| > 2.086\}$$

Now

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} = \frac{-6.05 - 0}{11.612/\sqrt{21}} = -2.39.$$

Because $|-2.39| = 2.39 > 2.086$, we reject H_0 and conclude that the dietary therapy does not eliminate the IQ gap between cases and their siblings.

Also $p = 2 \cdot \Pr(t_{20} \leq -2.39) = 0.027$.

Using R:

```
> t.test(iq.diff)

        One Sample t-test

data:  iq.diff
t = -2.3866, df = 20, p-value = 0.027
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -11.3335159  -0.7617221
sample estimates:
mean of x
-6.047619
```

Using SAS:

```
proc ttest;
   var iq_diff;

The TTEST Procedure

Variable:  iq_diff

  N        Mean     Std Dev     Std Err     Minimum     Maximum

 21     -6.0476     11.6124      2.5340    -33.0000     20.0000

    Mean       95% CL Mean        Std Dev      95% CL Std Dev

 -6.0476    -11.3335  -0.7617     11.6124      8.8842  16.7691

    DF    t Value    Pr > |t|

    20      -2.39      0.0270
```

(c) A 95% CI for μ is

$$\bar{Y} \pm t_{n-1,1-\alpha/2} \cdot s/\sqrt{n} = -6.05 \pm 2.086 \cdot 11.612/\sqrt{21} = (-11.33, -0.76).$$

This agrees with the results in R and SAS.

(d) Even with dietary therapy, on average a child with PKU has significantly lower IQ at age 4-6 than his or her closest-age normal sibling. The mean IQ of children with PKU who are on the dietary therapy in this study is 6.05 points lower than that of their normal siblings.

(e) Assumptions are that IQ data are normally distributed, that the difference between IQs of pairs all come from the same normal distribution (with the same mean and variance), and that the difference in IQ for any pair is independent of that for any other pair.

(f) Let $\zeta_{0.5}$ denote the median of the differences in IQ between children with PKU and their closest-age normal siblings. Then $H_0 : \zeta_{0.5} = 0$ and $H_A : \zeta_{0.5} \neq 0$.

To determine the critical region we need to find the largest $r_{\alpha/2}$ for which

$$\Pr[R \leq r_{\alpha/2} \mid H_0] = \frac{1}{2^n} \sum_{i=0}^{r_{\alpha/2}} \binom{n}{i} \leq \frac{\alpha}{2}$$

Using R

```
> 2*sum(dbinom(0:5,21,0.5))
[1] 0.0266037

> 2*sum(dbinom(0:6,21,0.5))
[1] 0.07835388
```

Confirming using the SIGN.test function:

```
> SIGN.test(iq.diff)

        One-sample Sign-Test

data:  iq.diff
s = 6, p-value = 0.07835
alternative hypothesis: true median is not equal to 0
```

So $r_{\alpha/2} = 5$ and thus $C_{0.05} = \{0, 1, 2, 3, 4, 5, 16, 17, 18, 19, 20, 21\}$

In this dataset, $r =$ (number of observations $> 0) = 6 \notin C_{0.05}$ so we cannot reject H_0 and we conclude that the data are consistent with the IQ of the PKU cases being similar to that of their normal siblings. Also, $p = 2 \cdot \Pr(r \leq 6) = 0.078 > 0.05$. Using R:

```
> 2*pbinom(6,21,0.5)
[1] 0.07835388
```

(g) In part (d) we rejected the null hypothesis that the mean difference is zero whereas in (f) we did not reject the null hypothesis that the median difference is 0. When the data are approximately normally distributed the t test can be more powerful than the sign test — the gain in power is because of the additional assumption (normality).

Question 2

Here there is no link between any particular hypertensive and normotensive subjects. So the two samples should be independent and thus two-sample tests should be used.

I argue below for the Wilcoxon test. If you make a reasonable argument for the assumptions of the t-test, it is okay to use it. Below is SAS code for the t-test and corresponding edited output. As with the Wilcoxon test, we would not reject H_0, which in this case is that the mean sodium intake is the same in the two groups.

```
proc ttest;
 class group;
 var sodium;

Variable:  sodium

group           N      Mean   Std Dev   Std Err   Minimum   Maximum

Hypertensive   15    1931.3     490.0     126.5    1100.0    2720.0
Normal         12    1690.0     430.0     124.1    1000.0    2375.0
Diff (1-2)            241.3     464.5     179.9

Method           Variances        DF    t Value    Pr > |t|

Pooled           Equal            25       1.34      0.1918
Satterthwaite    Unequal      24.744       1.36      0.1856

              Equality of Variances

Method      Num DF    Den DF    F Value    Pr > F

Folded F        14        11       1.30    0.6720
```

My preference is to use the Wilcoxon rank sum test here for the following reasons. First, based on the histograms in Figure 1, the sodium intakes in the two groups do not appear to be normally distributed. Second, based on the empirical distribution functions and the boxplots in Figure 1, the assumption of a location shift made by the rank sum test does seem to be plausible.

Let group 1 be the Hypertensive subjects and group 2 the Normal subjects, with sample sizes $n_1 = 15$ and $n_2 = 12$, respectively. Denote the corresponding distribution functions by F_1 and F_2. The null and alternative hypotheses are

$$H_0 : F_1(y) = F_2(y) \quad \text{and} \quad H_A : F_1(y + \Delta) = F_2(y)$$

for all y and some constant $\Delta \neq 0$.

Because n_1 and n_2 are both ≥ 12 the large sample approximation version of the test can be used. The test statistic is

$$Z = \frac{W_1 - E(W_1)}{\sqrt{V(W_1)}}$$

where W_1 is the sum of the ranks from the Hypertension group,

$$E(W_1) = \frac{n_1(N+1)}{2}$$

and

$$V(W_1) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2}{12N(N-1)} \sum_{i=1}^{q} t_i(t_i - 1)(t_i + 1)$$

where $N = n_1 + n_2 = 27$, q denotes the number of sets of ties, and t_i denotes the size of the i^{th} set of ties for $i = 1, \ldots, q$. At the $\alpha = 0.05$ level of significance, the critical region is

$$C_{0.05} = \{Z : |Z| > z_{0.975} = 1.96\}.$$

To compute W_1 we first get the ranks for the observed data, assigning the midrank in the case of ties, as in the table on the next page. There are three sets of ties, each with two tied observations, so $t_i = 2$ in each case.

Here $W_1 = 238$.

Also

$$E(W_1) = \frac{15 \times 28}{2} = 210$$

and

$$V(W_1) = \frac{12 \times 15 \times 28}{12} - \frac{12 \times 15}{12 \times 27 \times 26} \cdot \sum_{i=1}^{3} (2 \times 1 \times 3) = 419.62$$

so that $Z = (238 - 210)/\sqrt{419.62} = 1.367$ and hence we do not reject the null. The p-value is $2 * \Phi(-1.367) = 0.172$. Therefore, there is insufficient evidence from these data to suggest that there is a difference in sodium intake between normal and hypertensive individuals.

Verifying the results using R:

```
> wilcox.test(hypertensive,normal,exact=F,correct=F)

        Wilcoxon rank sum test

data:  hypertensive and normal
W = 118, p-value = 0.1717
alternative hypothesis: true location shift is not equal to 0
```

Hypertensive	Rank	Normal	Rank
1100	2	1000	1
1320	5	1220	3
1350	6	1300	4
1450	8	1400	7
1600	10.5	1555	9
1850	14	1600	10.5
1900	15.5	1780	12.5
1990	17	1780	12.5
2050	19	1900	15.5
2120	20	2020	18
2200	21	2350	23
2210	22	2375	24
2500	25		
2610	26		
2720	27		

Using SAS:

```
proc npar1way wilcoxon correct=no;
 class group;
 var sodium;
```

```
         Wilcoxon Scores (Rank Sums) for Variable sodium
                  Classified by Variable group

                      Sum of    Expected       Std Dev          Mean
group          N      Scores    Under H0      Under H0         Score
-------------------------------------------------------------------
Hypertensive   15      238.0       210.0     20.484516     15.866667
Normal         12      140.0       168.0     20.484516     11.666667

                      Average scores were used for ties.


   Wilcoxon Two-Sample Test

Statistic             140.0000

Normal Approximation
Z                      -1.3669
One-Sided Pr <  Z       0.0858
Two-Sided Pr > |Z|      0.1717
```

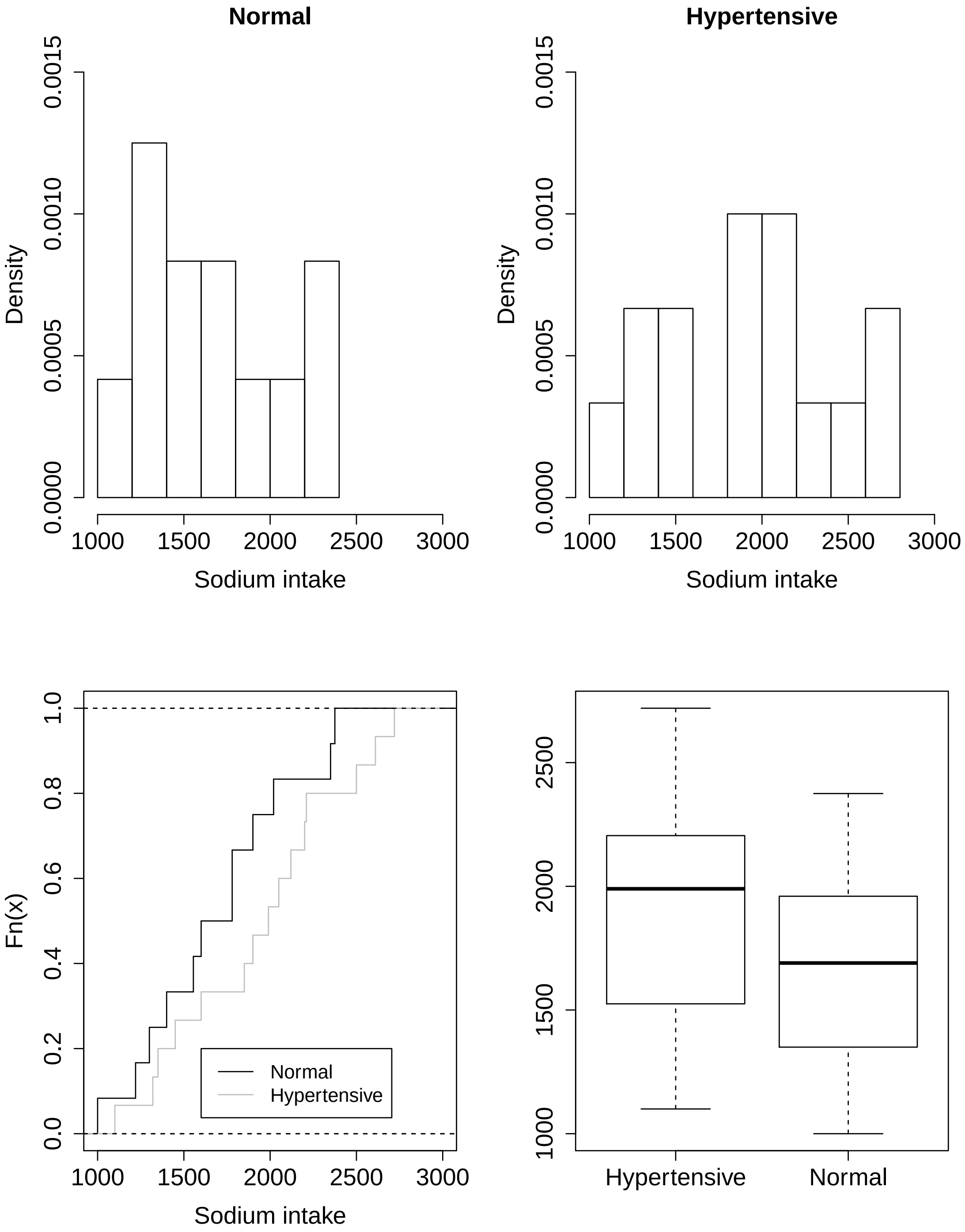


Figure 1: Histograms, EDFs, and boxplots for problem 2

BIOS 662, Fall 2018

Homework 5

Assigned: Tuesday, October 9

Due: Tuesday, October 16

1. This is a continuation of problem #2 from Homework 2 involving the dataset "HW2_PGE.txt". Use the Kolmogorov-Smirnov test to determine whether the distributions of plasma iPGE are the same for people with and without hypercalcemia.

2. Do Problem 6.5 on page 196 of the text. You do not need to do this "by hand" and so can ignore the parenthetical statement about needing to compute hypergeometric probabilities. (In my model solutions I will show how to do it "by hand" though.)

3. Problem 6.11(a)–(c) on page 197 of the text.

4. The file "HW5_Q4.txt" contains the data from Table 6.11 on page 199 of the text in versions suitable for SAS and R (both versions in one text file).
   (a) Verify that collapsing Table 6.11 over the smoking categories yields the table in Problem 6.13.
   (b) Calculate the odds ratio (and 95% confidence interval) for the association between coffee drinking and myocardial infarction, with and without taking into account smoking status. Do the calculations ignoring smoking status "by hand", confirming your results with SAS or R. (The calculations taking smoking status into account do not need to be done "by hand".)
   (c) Does smoking status confound the association between coffee drinking and myocardial infarction?

BIOS 662

Homework 5 Solution

October, 2018

Question 1:

You weren't asked to plot the empirical distribution functions. But it is instructive to see them (and consider ways to plot both in a single graph). The EDFs for the two groups of patients are given in Figure 1. The maximum difference between the two EDFs is indicated by an arrow. One way to obtain EDFs is to use the R function ecdf(...) and then plot the resulting object. To get R to include vertical lines in the plot, in the plot function use the option `verticals=TRUE`. (The default is `verticals=FALSE`.)

For the graph I used the function cumsum to obtain the EDFs "manually" and in the plot function used the option `type="s"` ("stair steps"). Here is my code:

```
ipge_h1<-c(0, 60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500, 1000)
ipge_h1c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,1,0))/11

ipge_h0<-c(0, 88, 100, 121, 130, 144, 148, 150, 168, 172, 254, 1000)
ipge_h0c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,0))/10

plot(ipge_h1,ipge_h1c,type="s",xlab="Plasma iPGE (pg/mL)",
   ylab="Empirical Distribution Functions F(y)",xlim=c(0,600),lty=2,
   cex.axis=1.25,cex.lab=1.25,cex.main=1.25,cex.sub=1.25)
lines(ipge_h0,ipge_h0c,lty=1,type="s")
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

Figure 2 is an alternative version of the plot created using the ecdf(...) function. I haven't been able to find how to suppress the horizontal dashed lines at 0 and 1, which overwrite the parts of the EDFs there.

```
f1=ecdf(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500))
f2=ecdf(c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))
plot(f1,verticals=TRUE,pch=NA,ylab="Empirical Distribution Functions F(y)"
  ,xlab="Plasma iPGE (pg/mL)",xlim=c(0,600),lty=2,cex.axis=1.25,
cex.lab=1.25,cex.main=1.25,cex.sub=1.25,ann=FALSE)
lines(f2,lty=1,verticals=TRUE,pch=NA)
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

We want to test

$$H_0 : F_1(x) = F_2(x) \text{ for all } x$$

versus

$$H_A : F_1(x) \neq F_2(x) \text{ for at least one } x$$

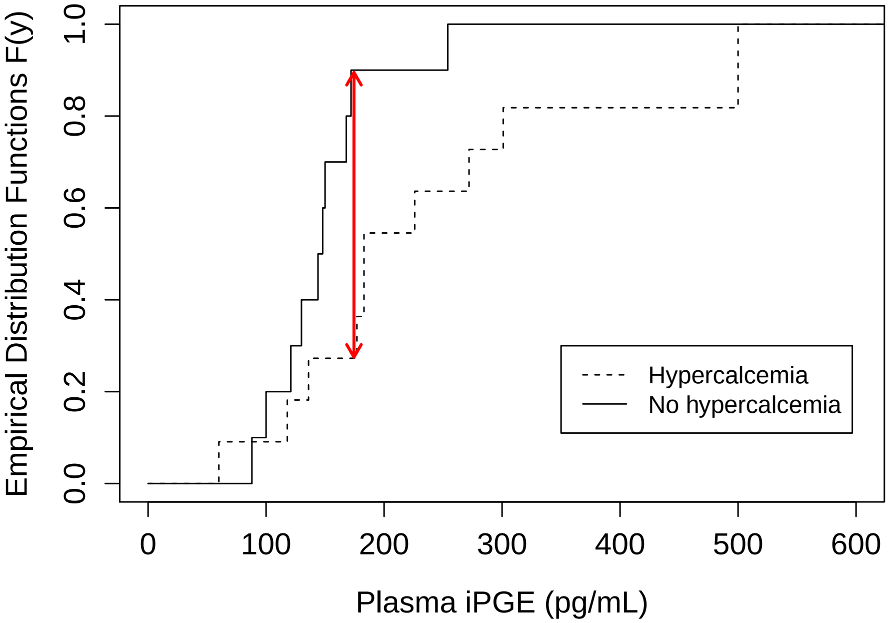


Figure 1: EDFs for problem 1

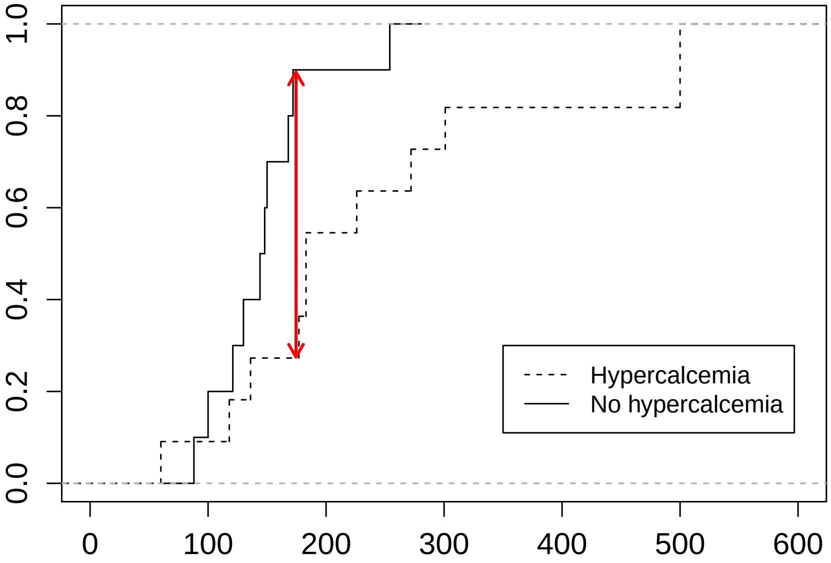


Figure 2: EDFs for problem 1 using ecdf() function

Here $D = \max_x |F_{1n}(x) - F_{2m}(x)| = 9/10 - 3/11 = 0.627$.

From the table on page 268 of the text, $C_{0.05} = \{\text{KS} : \text{KS} \geq 1.36\}$, where KS is defined as

$$\text{KS} = \sqrt{\frac{nm}{n+m}} D = \sqrt{\frac{10 \times 11}{10+11}} \times 0.627 = 1.4356.$$

Thus KS is in $C_{0.05}$ and so we conclude that the distributions of plasma iPGE differ for patients with and without hypercalcemia.

Using SAS to confirm this result and to obtain the p-value (the value for `KSa` is the large-sample approximation):

```
proc npar1way;
   var ipge;
   class hypercalcemia;
   exact ks;

        Kolmogorov-Smirnov Test for Variable iPGE
           Classified by Variable Hypercalcemia

                             EDF at    Deviation from Mean
Hypercalcemia       N       Maximum        at Maximum
----------------------------------------------------------
1                  11      0.272727         -0.990680
0                  10      0.900000          1.039034
Total              21      0.571429

       Maximum Deviation Occurred at Observation 13
             Value of iPGE at Maximum = 172.0

KS  0.3133    KSa  1.4356

Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2|      0.6273
Asymptotic Pr >  D     0.0324
Exact      Pr >= D     0.0154
```

Using R:

```
> ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),
+   c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))

        Two-sample Kolmogorov-Smirnov test

data:  c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500) and
    c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254)
D = 0.6273, p-value = 0.03242
alternative hypothesis: two-sided

Warning message:
In ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),  :
  cannot compute correct p-values with ties
```

Question 2: *Problem 6.5 on page 196 of the text.*

We want to compare the probability of 5-year survival for those with 1–4 courses of chemotherapy to those with ≥ 10 courses. Let $\pi_1 = \Pr[\text{dead}|\text{1–4 courses}]$ and $\pi_2 = \Pr[\text{dead}|\text{10+ courses}]$. Then

$$H_0 : \pi_1 = \pi_2 \text{ and } H_0 : \pi_1 \neq \pi_2.$$

Because the sample size is small we use Fisher's exact test.

To do it "by hand" in the way described in the class notes we first have to rearrange the table so that the row with the smaller row total is the first row and the column with the smaller column total is the first column. That is:

Courses	Alive	Dead	
≥ 10	8	2	10
1–4	2	21	23
Total	10	23	33

Setting $n_{11} = 0$ the table becomes:

Courses	Alive	Dead	
≥ 10	0	10	10
1–4	10	13	23
Total	10	23	33

$$\Pr[n_{11} = 0] = \frac{10!\,23!\,10!\,23!}{33!\,0!\,10!\,10!\,13!} = 0.0124.$$

Next, setting $n_{11} = 1$ the table becomes:

Courses	Alive	Dead	
≥ 10	1	9	10
1–4	9	14	23
Total	10	23	33

$$\Pr[n_{11} = 1] = \frac{10!\,23!\,10!\,23!}{33!\,1!\,9!\,9!\,14!} = 0.0883.$$

Similarly, $\Pr[n_{11} = 2] = 0.2384$, $\Pr[n_{11} = 3] = 0.3178$, $\Pr[n_{11} = 4] = 0.2290$, $\Pr[n_{11} = 5] = 0.0916$, $\Pr[n_{11} = 6] = 0.0201$, $\Pr[n_{11} = 7] = 0.0023$, $\Pr[n_{11} = 8] = 0.0001$, $\Pr[n_{11} = 9] < 0.0001$ and $\Pr[n_{11} = 10] < 0.0001$.

At this point $n_{21} = 0$ and we stop.

a	$\Pr[n_{11} = a]$	$\Pr[n_{11} \leq a]$	$\Pr[n_{11} \geq a]$
0	0.0124	0.0124	1.0000
1	0.0883	0.1006	0.9876
2	0.2384	0.3390	0.8994
3	0.3178	0.6569	0.6610
4	0.2290	0.8859	0.3431
5	0.0916	0.9775	0.1141
6	0.0201	0.9976	0.0225
7	0.0023	0.9999	0.0024
8	0.0001	1.0000	0.0001
9	<0.0001	1.0000	<0.0001
10	<0.0001	1.0000	<0.0001

The critical region for $H_A : \pi_1 \neq \pi_2$ is $C_{0.05} = \{n_{11} : n_{11} \in \{0, 6, 7, 8, 9, 10\}\}$. Because $n_{11} = 8$, we reject H_0 and, looking at the observed proportions dying within 5 years ($2/10 = 0.20$ and $21/23 = 0.91$), conclude that survival is more likely among those receiving at least 10 courses of chemotherapy. (Also, $p = 0.0001 < 0.05$.)

We confirm our answer using SAS:

```
data hw5_3;
   input chemo $1-5 status $7-12 count;
datalines;
c10p  alive  8
c10p  dead   2
c1to4 alive  2
c1to4 dead  21
;

proc freq data=hw5_3;
   tables chemo*status / norow nocol nopercent exact;
   weight count;
```

```
Chemo     Status

Frequency|alive   |dead    |  Total
---------+--------+--------+
c10p     |      8 |      2 |     10
---------+--------+--------+
c1to4    |      2 |     21 |     23
---------+--------+--------+
Total          10       23       33
```

```
       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        8
Left-sided Pr <= F         1.0000
Right-sided Pr >= F      1.255E-04

Table Probability (P)    1.230E-04
Two-sided Pr <= P        1.255E-04
```

Using R:

```
> fisher.test(matrix(c(8,2,2,21),nrow=2))

        Fisher's Exact Test for Count Data

data:  matrix(c(8, 2, 2, 21), nrow = 2)
p-value = 0.0001255
alternative hypothesis: true odds ratio is not equal to 1
```

Question 3: *Problem 6.11(a)-(c) on page 197 of the text.*

From the information given we can set up the table:

Usual church	Arteriosclerotic death		
attendance	Yes	No	
<1 per week	89	30,514	30,603
≥1 per week	38	24,207	24,245
Total	127		

Because the hypothesis seems to be that frequent church attendance is associated with "healthier" or "cleaner" living, the more frequent church attendance group is the "unexposed" or lower risk group.

Define $\pi_1 = \Pr[\text{arteriosclerotic death} \mid \text{church} <1 \text{ per week}]$

and $\pi_2 = \Pr[\text{arteriosclerotic death} \mid \text{church} \geq 1 \text{ per week}]$

(a) $\widehat{\mathrm{RR}} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2) = (89/30603)/(38/24245) = 1.8555$

(b) $\widehat{\mathrm{OR}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{89\times 24207}{38\times 30514} = 1.8580$

A 95% CI is $1.8580 \exp\left\{\pm 1.96\sqrt{\frac{1}{89}+\frac{1}{38}+\frac{1}{30514}+\frac{1}{24207}}\right\}$

So the confidence interval is (1.270, 2.717).

(c) $100(\widehat{\mathrm{OR}} - \widehat{\mathrm{RR}})/\widehat{\mathrm{RR}} = 100(1.8580 - 1.8555)/1.8555 = 0.13\%$

That is, in this setting in which the disease is rare, the percent error is just a small fraction of a percent.

We confirm parts (a) and (b) using SAS:

```
data;
  input church $1-5 arterio_death $7-9 count;
  datalines;
LT1pw Yes 89
LT1pw No 30514
GE1pw Yes 38
GE1pw No 24207
;

proc freq order=data;
  tables church*arterio_death / nopct nocol norow relrisk;
  weight count;
```

```
church     arterio_death

Frequency|Yes     |No      |  Total
---------+--------+--------+
LT1pw    |     89 |  30514 |  30603
---------+--------+--------+
GE1pw    |     38 |  24207 |  24245
---------+--------+--------+
Total         127    54721    54848

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.8580        1.2704        2.7174
Cohort (Col1 Risk)             1.8555        1.2696        2.7118
```

Question 4:

(a) *Verify that collapsing Table 6.11 over smoking categories yields the table in Problem 6.13.*

Using SAS on the dataset:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent;
   weight count;
```

yields the table in Problem 6.13:

```
Table of CupsCoffee by MIcase

CupsCoffee     MIcase

Frequency|Yes     |No      |  Total
---------+--------+--------+
GE5      |    152 |    183 |    335
---------+--------+--------+
LT5      |    335 |    797 |   1132
---------+--------+--------+
Total         487      980     1467
```

(b) *Calculate the odds ratio (and 95% confidence interval) for the association between coffee drinking and myocardial infarction, with and without taking into account smoking status. Do the calculations ignoring smoking status "by hand", confirming your results with SAS or R. (The calculations taking smoking status into account do not need to be done "by hand".)*

Ignoring smoking status, we use the data in the table above.

$\widehat{\text{OR}} = \frac{152\times797}{183\times335} = 1.9761$

A 95% CI is $1.9761\exp\left\{\pm1.96\sqrt{\frac{1}{152}+\frac{1}{335}+\frac{1}{183}+\frac{1}{797}}\right\}$

So the confidence interval is (1.5388, 2.5376).

Confirming this using SAS:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent relrisk;
   weight count;
```

```
Statistics for Table of CupsCoffee by MIcase

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                  Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.9761        1.5388        2.5376
```

Now using the Mantel-Haenszel method to take smoking status into account:

```
proc freq order=data;
   table Smoking*CupsCoffee*MIcase / norow nocol nopercent cmh;
   weight count;
```

```
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study   Method             Value  95% Confidence Limits
---------------------------------------------------------------
Case-Control    Mantel-Haenszel   1.3754    1.0505       1.8007
```

(c) *Does smoking status confound the association between coffee drinking and myocardial infarction?*

There is quite a substantial change in the odds ratio when smoking status is taken into account, decreasing from 1.976 to 1.375. Further evidence of the size of the change is that the latter is below the lower limit of the confidence interval for the former. (Note that this is not a formal test – these are both estimates rather than one being a hypothesized parameter.)

BIOS 662, Fall 2018

Homework 6

Assigned: Tuesday, November 6

Due: Tuesday, November 13

Instructions: For this homework, calculations need not be done "by hand." If you use software such as R or SAS to to the calculations, please include the code you used, not just the output. For all problems involving testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution, the critical region, whether you reject the null, the p-value, and an interpretation of the results in a language suitable for investigators. All tests should be performed at the $\alpha = 0.05$ significance level.

Peppermint extract is believed to have various medicinal properties. A study was conducted to investigate the effect of peppermint extract on triglyceride levels in rats. Fifty rats were randomly assigned to 5 groups, each having 10 rats. Those in groups 1 through 5 received, respectively, 0 (control group), 75, 150, 300 or 600 mg/kg of peppermint extract daily for three weeks. At the end of three weeks blood was drawn and assayed for various lipids, including triglyerides. The file "HW6_TRG.txt" in the "Datasets" sub-folder of the "Homework materials" section of the Sakai site contains data on triglyceride levels in the blood of the rats. Only those rats with non-missing triglyceride levels are included. There are three variables, ID, group and trg (triglycerides, in μg/dL).

1. The investigators' primary interest is in which groups (that is, which dosages of peppermint extract) differ from one another in terms of the effect on triglyceride levels. Conduct an appropriate statistical analysis of the data using a parametric ANOVA model. Include in your report: (a) an analysis plan, (b) results of your analysis, and (c) a brief conclusion in language suitable for the investigators. As part of your analysis, investigate whether a transformation of the data would be appropriate. If so, state what transformation should be used and check whether it improves the diagnostics, but conduct your analysis on the untransformed data (so as not to introduce an extra level of complication in the grading of this homework).
2. Two items of secondary interest are (i) whether the mean triglyceride level in the control group differs from that in the other 4 groups combined and (ii) whether there is a linear relationship between group number and triglyceride levels. Use your parametric ANOVA model to address these items.
3. Now use a linear regression model with peppermint extract dose as a continuous variable (actual dose in mg/dL, not group number). Provide an

estimate and associated confidence interval for how triglyceride levels change with dose of peppermint extract. Use the regression model to predict the mean triglyceride level for the control group. How does this compare with the sample mean for the control group? (For the purposes of this homework, it is *not* necessary to check the assumptions of the regression model.)

BIOS 662

Homework 6 Solution

November, 2018

Part 1

(a) Analysis Plan

Standard analysis of variance methodology will be used. Triglyceride level Y will be modeled by

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where the index $i = 1, 2, 3, 4, 5$ denotes the groups, receiving 0, 75, 150, 300, or 600 mg/kg of peppermint extract, respectively, the index j denotes the j^{th} rat within the i^{th} dosage group, μ_i denotes the population mean triglyceride level in the i^{th} dosage group, and the ε_{ij} are assumed to be independent and identically distributed as $N(0, \sigma^2)$. The primary interest is in pairwise comparisons between groups, to determine which groups differ from one another. The group sizes are unequal, varying from 7 to 10, so it is more appropriate to use the Scheffé or Bonferroni method to adjust for multiple comparisons. We'll use Scheffé's method for the primary analysis.

Standard ANOVA diagnostics will be used to assess the fit of the model above. In the event of violations of the assumptions of ANOVA (in particular, homogeneity of variance or normality), the Box-Cox family of transformations will be used to find the transformation of the data that minimizes the MSE. We recognize that the sample sizes are rather small so that only quite large departures from the assumptions are likely to be detectable.

(b) Analyses

Figure 1 has boxplots of the data for each group, with the individual points overlaid and Table 1 has corresponding summary statistics.

Group	N	Median	Mean	Std Dev
1	10	251.5	244.2	17.87
2	10	241.0	238.1	10.30
3	7	230.0	228.1	8.55
4	10	220.0	220.5	6.67
5	9	210.0	209.9	5.09

Table 1: Summary statistics for data from peppermint extract study

The boxplots, summary statistics and diagnostics for the ANOVA model suggest that the homogeneity of variance assumption is questionable. In particular, looking at the

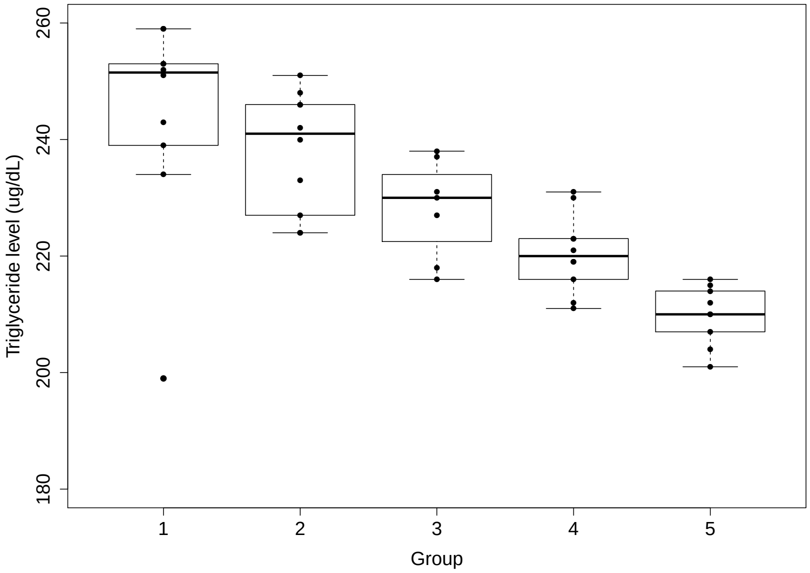


Figure 1: Triglyceride level by peppermint extract dosage group

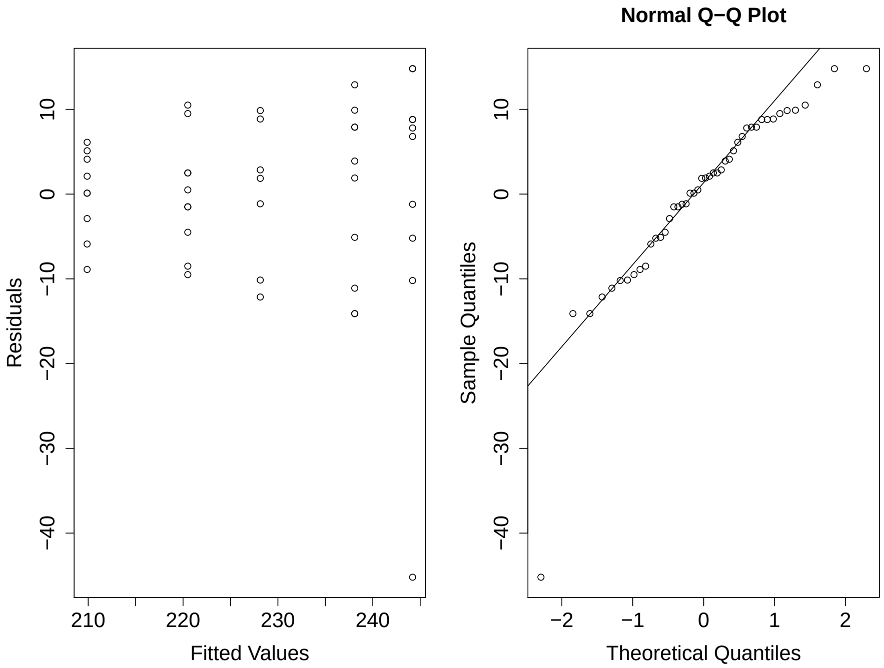


Figure 2: Residual plots from fitted ANOVA model

standard deviations in Table 1 and the residual plot (left panel of Figure 2) it appears that the variance decreases with increasing peppermint extract dose. The trend in the standard deviations is influenced substantially by a single observation, the 199 for rat R15. Omitting this observation reduces the standard deviation for group 1 from 17.86 to 8.70 (and increases that group's mean from 244.2 to 249.2). However, even without excluding this observation, the modified Levene test ($p = 0.38$, see below) does not reject the hypothesis of homogeneity of variance.

```
proc anova;
   class group;
   model trg = group;
   means group / hovtest=bf;

  Brown and Forsythe's Test for Homogeneity of trg Variance
       ANOVA of Absolute Deviations from Group Medians

                        Sum of        Mean
Source         DF      Squares      Square    F Value    Pr > F

group           4        307.6     76.9122       1.09    0.3759
Error          41       2902.5     70.7928
```

In the right panel of Figure 2, the QQ-plot for the residuals from the fitted model suggests that there is departure from normality in the tails of the distribution of the residuals. Pearson's correlation coefficient from the QQ-plot is 0.93 (see below). Here $n = 46$ and as $r = 0.93$ is smaller than the critical value of 0.97 for $n = 40$ on page 9 of the ”ANOVA, Part III” overheads, the normality assumption is questionable.

```
> group<-as.factor(group)
> fit<-aov(trg˜group)
> par(mfcol=c(1,2))
> plot(fit$fitted.values,fit$residuals)
> qq<-qqnorm(fit$residuals)
> qqline(fit$residuals)
> cor.test(qq$x,qq$y)

        Pearson's product-moment correlation

data:  qq$x and qq$y

sample estimates:
      cor
0.9255097
```

We are not given information about whether some of the rats were from the same litter so we don't have enough information to determine if the independence assumption is satisfied.

Because it appears that the normality assumption is violated and the sample size in each group is rather small for the Central Limit Theorem to help, we will use the Box-Cox method to investigate potential transformations. From the SAS code and output below, $\lambda = 0.6$ minimizes the MSE. The 95% confidence interval for λ extends from -3.0 to 4.2, indicating a large amount of uncertainty about the most appropriate value for λ. Because a square root transformation ($\lambda = 0.5$) is close to the optimal value, we'll try this to see whether it improves the normality of the residuals. Using this transformation, both parts of Figure 3 are very similar to those in Figure 2, Pearson's correlation coefficient from the QQ-plot is 0.92, little changed from the 0.93 using the untransformed data and the modified Levene test ($p = 0.44$) again does not indicate lack of homogeneity of variance.

(The results of these diagnostic checks are somewhat surprising. For this example I did not have the original data. I saw summary statistics in a journal article and generated data to yield similar summary statistics. I generated the residuals from the normal distribution, with different variances for the 5 groups, yet the diagnostics indicate that normality is questionable rather than homogeneity of variance. So, violation of one of the assumptions may manifest as violation of one of the other assumptions.)

```
data hw6;
   set hw6;

grp1=0; grp2=0;grp3=0;grp4=0;grp5=0;
if group=1 then grp1=1;
   else if group=2 then grp2=1;
   else if group=3 then grp3=1;
   else if group=4 then grp4=1;
   else if group=5 then grp5=1;


%boxcox(resp=trg,model=grp2 grp3 grp4 grp5,lopower=-2,hipower=2,
           npower=41,data=hw6);
```

```
 Box-Cox                     Root mean       0.95
  Power          Log          squared     Confidence
(lambda)     Likelihood        error       Interval
   0.0        -109.931        10.9113         *+
   0.1        -109.916        10.9079         *
   0.2        -109.905        10.9051         *
   0.3        -109.896        10.9030         *
   0.4        -109.890        10.9015         *
   0.5        -109.886        10.9007         *+
   0.6        -109.885        10.9005         <
   0.7        -109.887        10.9010         *
   0.8        -109.892        10.9021         *
   0.9        -109.899        10.9038         *
   1.0        -109.909        10.9062         *+
```

(The output above has been edited to show just part of the range of values of λ.)

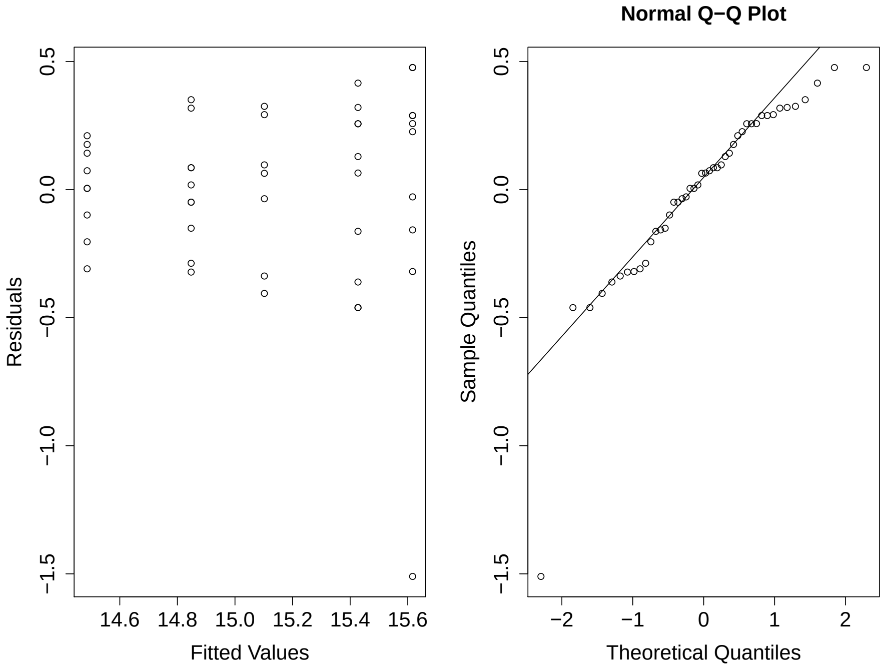


Figure 3: Residual plots from fitted ANOVA model after square root transformation

We now investigate pairwise differences in the means, using the untransformed data, with Scheffé's method for adjusting for the multiple comparisons. SAS code and output are presented below. Groups 1 and 2 differ significantly from groups 4 and 5 and group 3 differs significantly from group 5 but not from the other three groups. These comparisons are presented schematically in Figure 4.

Results using Bonferroni or Tukey's method are similar except that with those methods groups 1 and 3 are significantly different. The corresponding schematic is in Figure 5.

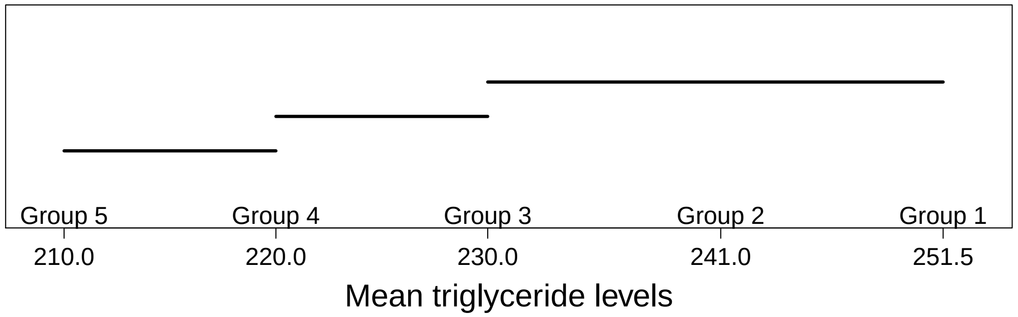


Figure 4: Using Scheffé's method; lines join groups that do not differ significantly

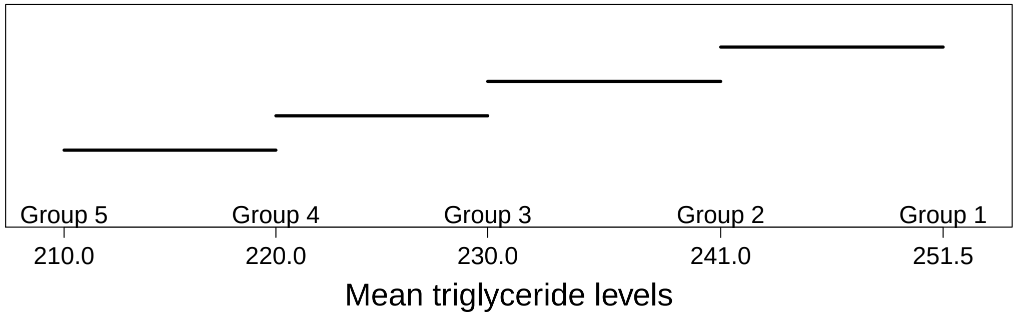


Figure 5: Using Bonferroni; lines join groups that do not differ significantly

```
proc anova data=hw6;
   class group;
   model trg = group;
   means group / scheffe;
```

```
Dependent Variable: trg

                             Sum of
Source             DF        Squares    Mean Square    F Value    Pr > F
Model               4    7144.55832     1786.13958      15.02    <.0001
Error              41    4876.74603      118.94503
Corrected Total    45   12021.30435


Scheffe's Test for trg

NOTE: This test controls the Type I experimentwise error rate, but it generally
has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha                        0.05
Error Degrees of Freedom       41
Error Mean Square         118.945
Critical Value of F       2.59997

Comparisons significant at the 0.05 level are indicated by ***.

              Difference
  group          Between    Simultaneous 95%
Comparison         Means    Confidence Limits

  1 - 2           6.100      -9.629   21.829
  1 - 3          16.057      -1.275   33.390
  1 - 4          23.700       7.971   39.429  ***
  1 - 5          34.311      18.151   50.471  ***
  2 - 1          -6.100     -21.829    9.629
  2 - 3           9.957      -7.375   27.290
  2 - 4          17.600       1.871   33.329  ***
  2 - 5          28.211      12.051   44.371  ***
  3 - 1         -16.057     -33.390    1.275
  3 - 2          -9.957     -27.290    7.375
  3 - 4           7.643      -9.690   24.975
  3 - 5          18.254       0.529   35.979  ***
  4 - 1         -23.700     -39.429   -7.971  ***
  4 - 2         -17.600     -33.329   -1.871  ***
  4 - 3          -7.643     -24.975    9.690
  4 - 5          10.611      -5.549   26.771
  5 - 1         -34.311     -50.471  -18.151  ***
  5 - 2         -28.211     -44.371  -12.051  ***
  5 - 3         -18.254     -35.979   -0.529  ***
  5 - 4         -10.611     -26.771    5.549
```

(c) Conclusions

Mean triglyceride levels appear to be included by dosage of peppermint extract, with higher doses associated with lower triglyceride levels. In comparing pairs of dosages, adjacent dosages generally do not have significantly different effects, though that may be because with the small sample sizes there is limited power to detect relatively small differences. Pairs of doages that are further apart generally do result in significant differences in triglyceride levels.

Part 2

Note that the problem specifically says to use your parametric ANOVA model to address these. This can be done using contrasts. For (i) we want to test

$$H_0 : \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 = 0 \quad \text{vs.} \quad H_A : \mu_1 \neq (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

or, equivalently,

$$H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \quad \text{vs.} \quad H_A : \mu_1 \neq (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

and one way to approach (ii) is to test

$$H_0 : -2 \cdot \mu_1 - 1 \cdot \mu_2 + 0 \cdot \mu_3 + 1 \cdot \mu_4 + 2 \cdot \mu_5 = 0$$

vs.

$$H_A : -2 \cdot \mu_1 - 1 \cdot \mu_2 + 0 \cdot \mu_3 + 1 \cdot \mu_4 + 2 \cdot \mu_5 \neq 0$$

In each case H_0 is of the form $\sum_{i=1}^{5} c_i \mu_i = 0$, with $\sum_{i=1}^{5} c_i = 0$.

Below is SAS code and corresponding output using contrasts to test these hypotheses. In both instances we reject H_0 and conclude that (i) the mean triclyceride level in rats on placebo differs significantly from that in rats given peppermint extract and (ii) there appears to be a linear relationship between group number and mean triclyceride level.

```
proc glm;
   class group;
   model trg = group;
   contrast 'Placebo vs. others' group 1 -0.25 -0.25 -0.25 -0.25;
   contrast 'Linear association' group -2 -1 0 1 2;

Dependent Variable: trg
                                 Sum of
Source                DF        Squares    Mean Square    F Value    Pr > F
Model                  4     7144.55832     1786.13958      15.02    <.0001
Error                 41     4876.74603      118.94503
Corrected Total       45    12021.30435

Contrast              DF    Contrast SS    Mean Square    F Value    Pr > F
Placebo vs. others     1    3129.040058    3129.040058      26.31    <.0001
Linear association     1    7117.919622    7117.919622      59.84    <.0001
```

How does using a contrast in an ANOVA model to test for a linear association differ from using a linear regression model? Below is SAS code and corresponding output from a regression model using group as the predictor variable. The F value is similar but not identical to that for the “Linear association” contrast. The difference arises because the ANOVA model fits 5 parameters (the 5 group means), leaving 41 degrees of freedom for the error whereas the regression model fits two parameters (intercept and slope), leaving 44 degrees of freedom for the error.

```
proc glm;
   model trg = group;

Dependent Variable: trg
                                  Sum of
Source              DF           Squares     Mean Square     F Value     Pr > F
Model                1        7094.02783      7094.02783       63.35     <.0001
Error               44        4927.27652       111.98356
Corrected Total     45       12021.30435
```

Part 3

Here we fit a regression model using actual dose rather than group number. The regression model is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X_i is the dose received by the i^{th} rat, Y_i is the triglyceride level of the i^{th} rat and the ε_i are iid $N(0,\sigma^2)$. SAS code and corresponding output are below. Dose is significantly associated with triglyceride levels. The estimated regression model is

$$\widehat{Y_i} = \widehat{\alpha} + \widehat{\beta} X_i.$$

From the output below, $\widehat{\beta} = -0.057$. That is, according to the model, mean triglyceride level decreases by 0.057 μg/dL for each 1 mg/kg increase in peppermint extract dose (or decreases by 5.7 μg/dL for each 100 mg/kg increase in peppermint extract dose). The corresponding 95% confidence interval is

$$\widehat{\beta} \pm t_{N-2,0.975}\text{SE}(\widehat{\beta}) = -0.057 \pm 2.02 \times 0.0076 = (-0.072, -0.041).$$

The predicted mean for the group with dose $= 0$ is $241.1 - 0.0566 \cdot 0 = 241$. Using SAS, the 95% confidence interval for the point on the line at dose $= 0$ is $(236.4, 245.8)$. This compares favorably with the actual mean for group 1, namely 244.2. On the other hand, omitting the outlier rat (R15) the estimate of the mean for dose $= 0$ becomes 243.1, with 95% confidence interval $(239.2, 247.0)$ whereas the mean for group 1 becomes 249.2.

(I hadn’t stated explicitly whether the regression should have been done with the observations in the dose $= 0$ group included or removed. The above results are when those observations are included. Fitting the model without the dose $= 0$ observations the predicted mean for dose $= 0$ is 238.6 and the 95% confidence interval for the point on the line at dose $= 0$ is $(233.8, 243.3)$.)

```
proc glm;
   model trg = dose / solution clparm;

Dependent Variable: trg
                                Sum of
Source               DF        Squares    Mean Square    F Value    Pr > F
Model                 1     6694.23245     6694.23245      55.29    <.0001
Error                44     5327.07190      121.06982
Corrected Total      45    12021.30435

R-Square     Coeff Var      Root MSE      trg Mean
0.556864      4.814019      11.00317      228.5652

                           Standard
Parameter     Estimate        Error  t Value  Pr > |t|    95% Confidence Limits
Intercept  241.1084873   2.34039305   103.02    <.0001   236.3917350  245.8252395
dose        -0.0565677   0.00760740    -7.44    <.0001    -0.0718994   -0.0412360
```

BIOS 662, Fall 2018

Homework 7

Assigned: Tuesday, November 13

Due: Tuesday, November 20

1. A case-control study is being designed to detect an odds ratio of 3 for bladder cancer associated with a certain medication that is used in about one person out of 50 in the general population. Suppose $\alpha = 0.05$ and that 100 cases and 100 controls are to be sampled for the study. What is the power to detect OR $= 3$? Would you recommend conducting such a study? If not, how many cases and controls would you recommend?

2. A cross-sectional study is being designed to investigate the association between a continuous exposure X and a continuous outcome Y. The data will be analyzed using a linear regression model

   $$Y = \alpha + \beta X + \epsilon.$$

   From a pilot study it seems reasonable to assume that $\epsilon \sim N(0, 10^2)$ and that $X \sim N(50, 8^2)$.

   (a) Using simulation, determine the sample size needed to detect $\beta = 0.20$ with power 0.90. Include a copy of the code you use for your simulation.

   (b) Confirm your answer using a sample size formula from the "Power and Sample Size, Part III" overheads.

BIOS 662

Homework 7 Solution

November, 2018

Question 1

First recast this as a two sample binary problem as in the class notes. Because bladder cancer is rare, the proportion of non-cases in the population is very close to 1. So, to a very good approximation, one person of 50 in the general population using the particular medication is the same as the proportion of controls who use the medication. That is, $\pi_2 = \Pr(\text{exposed}|\text{case}) = 1/50 = 0.02$. Also, as $\text{OR} = 3$, we have

$$\pi_1 = \frac{\pi_2 \text{OR}}{1 + \pi_2(\text{OR} - 1)} = 0.05769.$$

The power to detect $\pi_1 = 0.05769$ versus $\pi_2 = 0.02$ at the $\alpha = 0.05$ level of significance is 0.28. This result can be obtained using the formula in the class notes or by the following SAS code:

```
proc power;
   twosamplefreq
     refp   = 0.02
     pdiff  = 0.03769
     ntotal = 200
     power  = .;
```

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

               Fixed Scenario Elements

Distribution                          Asymptotic normal
Method                             Normal approximation
Reference (Group 1) Proportion                     0.02
Proportion Difference                           0.03769
Total Sample Size                                   200
Number of Sides                                       2
Null Proportion Difference                            0
Alpha                                              0.05
Group 1 Weight                                        1
Group 2 Weight                                        1


Computed Power

Power
0.280
```

We would not recommend conducting this study because the power is very low, so the chance of a type II error is too high. There is a probability of $1 - 0.28 = 0.72$ that we will fail to detect an OR as large as 3. Instead we would recommend 412 cases and 412 controls to have 80% power, or 551 cases and 551 controls to have 90% power. (Or a study with multiple controls per case, but even that would need many more than 100 cases to have reasonable power.) These results can be obtained using the formula in the class notes, as implemented in R in the function "ss_fleiss" defined in the class notes:

```
> ss_fleiss(0.02,0.05769,0.05,0.8)
[1] 411.4046

> ss_fleiss(0.02,0.05769,0.05,0.9)
[1] 550.2548
```

or by using proc power as follows:

```
proc power;
 twosamplefreq
   refp   = 0.02
   pdiff  = 0.03769
   ntotal = .
   power  = 0.8 0.9;
```

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

               Fixed Scenario Elements

Distribution                         Asymptotic normal
Method                            Normal approximation
Reference (Group 1) Proportion                    0.02
Proportion Difference                          0.03769
Nominal Power                                      0.9
Number of Sides                                      2
Null Proportion Difference                           0
Alpha                                             0.05
Group 1 Weight                                       1
Group 2 Weight                                       1


        Computed N Total

         Nominal    Actual        N
Index      Power     Power    Total
    1        0.8     0.801      824
    2        0.9     0.900     1102
```

Question 2

(a) Using either R or SAS, a key issue for any power/sample size simulation is to work out how to obtain the relevant p-value from the specific function or proc. In the SAS version below the dataset produced by proc reg for the first simulated dataset is printed to check where the p-value is. See the comment in the code. Similarly, in the R code the estimated coefficients and related test statistics are printed for the first simulated dataset.

Note that here for each dataset of size *n* we generate *n* values for X and for ε and then obtain *n* values for Y using $y_i = \alpha + \beta x_i + \varepsilon_i$. In this case any value can be used for α without affecting the results.

Because we want the sample size to achieve a specified power (0.9) we try various values of *n* until we find one that gives the appropriate power. The appropriate sample size is about 417. (I ran the simulation *before* using the formula in part (b) and will admit to being surprised at how well the simulation agrees with the result in (b).)

For a large sample size and/or large number of simulated datasets, the way the data are generated in the SAS example in the notes can cause out-of-memory errors. So I have included two different approaches to doing the simulation in SAS. The first one is like the example in the notes, with all the datasets being generated first and then the test of the regression coefficient being run on them. In the second approach the datasets are generated one at a time, with the test of the regression coefficient being run before the next dataset is generated. With this approach just one dataset of *n* observations is kept in memory at any time. The “end=eof”, the retain statement and the three lines beginning with “if eof then do;” are to pass an updated random number seed to the next iteration. See pages 20-21 of the “Random Number Generation” overheads.

R code for the simulation and the corresponding output:

```
alpha <- 0
beta <- 0.2

mysim <- function(seed0,n,nsims){
  set.seed(seed0)
  rejects <- 0
  for (ii in 1:nsims){
    e <- rnorm(n,0,10)
    x <- rnorm(n,50,8)
    y <- alpha + beta*x + e
    fit<-summary.lm(lm(y~x))
    coeffs<-fit$coefficients
# The next statement shows the output for the regression on the first simulated
# dataset; looking at the output explains why coeffs[2,4] is used below
    if (ii==1) print(fit$coefficients)
    if (coeffs[2,4]<0.05) rejects <- rejects + 1
    }
  print(paste("Sample size:",n,", estimated power:",rejects/nsims))
  }
```

```
mysim(19,417,10000)
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -1.0111480 3.21913633 -0.3141054 0.7535988457
x            0.2191917 0.06362334  3.4451458 0.0006289053
[1] "Sample size: 417 , estimated power: 0.902"
```

Using a different seed yields different estimates for the first dataset of the simulation and a very slightly different power estimate for this sample size:

```
> mysim(37,417,10000)
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 2.4377952 3.36466099 0.7245292 0.46914917
x           0.1508311 0.06653812 2.2668369 0.02391407
[1] "Sample size: 417 , estimated power: 0.9005"
```

SAS code using the first approach:

```
%macro epower(beta=,seed=,nsims=,n=);
%let sd_x=8; %let sd_e=10;

data simdata;
  %do i = 1 %to &nsims;
    i=&i;
    do j=1 to &n;
   x=50+rannor(&seed)*&sd_x;
   e=rannor(&seed)*&sd_e;
   y=&beta*x+e;
   output;
end;
  %end;

ods listing close;
ods output "Parameter Estimates"=params;
proc reg data=simdata;
   model y=x;
   by i;
   run;
quit;
ods output close;
ods listing;

*** The following code shows what is in the regression output    ;
*** in "params" dataset when i=1 and thus why "if variable=’x’" ;
*** is used below.                                               ;

proc print data=params;
   where i=1;
```

```
data params;
   set params;
if variable='x';

if Probt<0.05 then reject=1; else reject=0;

proc freq data=params;
   table reject;

%mend;

%epower(beta=0.2,seed=97461,nsims=10000,n=417);
```

SAS output from the first approach:

```
Obs i Model  Dependent Variable  DF  Estimate   StdErr  tValue  Probt

  1 1 MODEL1     y     Intercept  1  -1.16503  3.36072   -0.35 0.7290
  2 1 MODEL1     y     x          1   0.22517  0.06601    3.41 0.0007

The FREQ Procedure
                                   Cumulative    Cumulative
reject    Frequency     Percent     Frequency      Percent
-----------------------------------------------------------
     0         998        9.98           998         9.98
     1        9002       90.02         10000       100.00
```

SAS code using the second approach:

```
%let alpha=20; %let beta=0.20; %let sd_e=10;

data pvals; set _NULL_;

%macro rept(seed0=,reps=,sampsize=);
%do i=1 %to &reps;

proc iml;
   setnull=j(&sampsize,1,0);
create begindat from setnull[colname='zero'];
   append from setnull;
quit;

data tempsamp;
   set begindat end=eof;
   retain seed &seed0;

call rannor(seed,z);
call rannor(seed,etemp);
err=&sd_e*etemp;
x=8*z + 50;
y=&alpha + &beta*x + err;
```

```
if eof then do;
call symput('seed0',put(seed,best.));
end;

*** The following code is to check the data being generated ;
*** and look at where the p-value is in the output dataset. ;
%if &i=1 %then %do;

proc means data=tempsamp;
   var z x;

proc reg data=tempsamp outest=temp tableout;
   model y = x;

proc print data=temp;
   var _TYPE_ Intercept x;

%end;

proc reg data=tempsamp outest=regests tableout noprint;
   model y = x;

data regests;
   set regests;
if _TYPE_ NE 'PVALUE' then delete;
p_int=Intercept;
p_x=x;
keep p_int p_x;

data pvals;
   set pvals regests;

%end;
%mend rept;

%rept(seed0=421325,reps=1000,sampsize=417);

data pvals;
  set pvals;
if p_x le 0.05 then reject1=1;
   else reject1=0;

proc freq;
   table reject1;
```

Edited SAS output from the second approach, including PROC MEANS and PROC REG output for the first dataset generated:

```
The MEANS Procedure

Variable      N            Mean         Std Dev         Minimum         Maximum
-------------------------------------------------------------------------------
z           417      -0.0558561       1.0325162      -3.0516687       2.8930635
x           417      49.5531515       8.2601299      25.5866506      73.1445081
-------------------------------------------------------------------------------


The REG Procedure

Number of Observations Read         417
Number of Observations Used         417

                        Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1       18.53427        3.10077       5.98      <.0001
x             1        0.22700        0.06173       3.68      0.0003


Obs    _TYPE_    Intercept       x

 1     PARMS      18.5343     0.22700
 2     STDERR      3.1008     0.06173
 3     T           5.9773     3.67756
 4     PVALUE      0.0000     0.00027
 5     L95B       12.4391     0.10566
 6     U95B       24.6295     0.34833


The FREQ Procedure

                                     Cumulative    Cumulative
reject1    Frequency     Percent     Frequency      Percent
-------------------------------------------------------------
      0          95        9.50            95         9.50
      1         905       90.50          1000       100.00
```

(b) To use the formula from the "Power and Sample Size, Part III" overheads we need s_X and s_Y. Here $s_X = 8$. We are not given s_Y but can calculate it. Because α and β are constants and X and ε are independent:

$$\mathrm{Var}(Y) = \mathrm{Var}(\alpha + \beta \cdot X + \varepsilon) = \beta^2 \mathrm{Var}(X) + \mathrm{Var}(\varepsilon) = 0.2^2 \cdot 8^2 + 10^2 = 102.56.$$

So $s_Y = \sqrt{102.56} = 10.13$.

Letting $\widehat{\beta_1}$ be the alternative we are interested in being able to detect (0.2) we have

$$r = \frac{s_X}{s_Y} \widehat{\beta_1} = \frac{8}{10.13} \cdot 0.2 = 0.158$$

We should have the same power to test

$$H_A : \beta_1 = 0.2 \quad \text{against} \quad H_0 : \beta_1 = 0$$

as to test

$$H_A : \rho = 0.158 \quad \text{against} \quad H_0 : \rho = 0$$

Here

$$Z_0 = \frac{1}{2} \log \left(\frac{1+0}{1-0} \right) = 0$$

and

$$Z_1 = \frac{1}{2} \log \left(\frac{1+0.158}{1-0.158} \right) = 0.159$$

The sample size n to give us power 0.90 for testing $\rho = 0.159$ versus $\rho = 0$ is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(Z_{\rho_1} - Z_{\rho_0})^2} + 3 = \frac{(1.96 + 1.28)^2}{(0.159 - 0)^2} + 3 = 416.9$$

Rounding up, that gives a sample size of $n = 417$.

BIOS 662, Fall 2018
Homework 7

Assigned: Tuesday, November 27

Due: Tuesday, December 4

Calculations need not be done "by hand."

1. This question uses the data in Problem 15.5 on pages 654/5 of the textbook (available on Sakai in the dataset "HW8_Q1.txt", with the "Unknown" category of "Physical Status" coded as 0).

   Note that here the "sample" is the number of operations. These occur at a distinct point in time (and the outcome is vital status at 6 weeks after the operation). Although there is a 4-year time interval, in this particular problem that is essentially irrelevant in terms of incidence. The rate of interest is deaths per 100,000 operations rather than per year (or any other time period).

   SAS versions 9.3 and 9.4 have a procedure for direct and indirect standardization (`PROC STDRATE`). It uses somewhat different formulas for variance estimation from the ones covered in class but you are welcome to use it.

   (a) Calculate the crude death rate (that is, the overall rate, ignoring physical status) per 100,000 operations for halothane and cyclopropane. Are these two rates significantly different?

   (b) Using direct standardization (relative to the total study sample, not just those in the two specific treatment groups), calculate the standardized death rates for halothane and cyclopropane and test whether they are equal.

   (c) Using indirect standardization, with the total study sample as the reference population, calculate the standardized incidence ratio for halothane and test whether $\pi_{halothane}/\pi_{overall} = 1$.

   (Comment: The "total" numbers include those for halothane so the two are not independent, but that complication should be ignored.)

2. The dataset "HW8_SURV.txt" on the Sakai site contains data from a study investigating a new treatment for lung cancer. The variables in the dataset are ID (an identified for each participant), TIMEDEATH (time in days from randomization to death or censoring), DEATH (=0 for a censored observation, =1 for a death), AGE (in years) and GROUP (treatment group;

=1 placebo, =2 the new treatment). The new treatment is intended to be given in addition to usual care. Patients in the placebo group will also be receiving usual care, so the use of a placebo is ethical here.

(a) Compute and plot in the same graph the Kaplan-Meier (product limit) curves for the two treatment groups.

(b) Use the log-rank test to test whether the distribution of survival times is the same in the two groups.

(c) Now use the proportional hazards model to test whether the distribution of survival times is the same in the two groups. That is, use the p-value from the SAS or R output to determine whether the β coefficient differs significantly from 0.

(d) Using your model in part (c), estimate the hazard ratio for group 2 relative to group 1 and provide a 95% confidence interval for the true hazard ratio.

(e) Now include age in the proportional hazards model in part (c). Does age have a significant effect on survival? Does adjusting for age make a substantial difference to the estimate of the treatment effect?

(f) For the placebo group, estimate the median survival time, that is, the time at which $S(t) = 0.5$.

BIOS 662

Homework 8 Solution

December, 2018

Question 1

As mentioned in the question, the "sample" is the number of operations. These occur at a distinct point in time (and the outcome is vital status at 6 weeks after the operation). The rate of interest is deaths per 100,000 operations. We could think of it as "if 100,000 operations were performed in a year, how many people would die within 6 weeks of the operation". Our estimate would be the same if the 100,000 operations occurred over a longer or shorter period than a year. In the example on page 6 of the "Rates and Proportions" overheads, a person without diabetes is at risk of diabetes as long as he/she is being followed in the study. There the incidence of 0.033 can be thought of as the probability of a person becoming diabetic if followed for a year. But in the current problem it is just the short time immediately after the operation that is considered. It does not make sense to try to express this as risk over a year — the risk of death related to the operation decreases with time since the operation, so the risk in the first 6 weeks is unlikely to be representative of the risk over a longer period (such as a year) or even over a shorter period (such as in the first week).

(a) Let $I_{H,10^5}$ and $I_{C,10^5}$ denote the death rates per 100,0000 (per year) for halothane and cyclopropane, respectively.

$$\hat{I}_{H,10^5} = c \cdot \frac{\text{number of deaths}}{\text{number of operations using halothane}}$$

where $c = 100,000$. A corresponding formula holds for $I_{C,10^5}$.

So,

$$\hat{I}_{H,10^5} = 100,000 \cdot \frac{2,375}{146,200} = 1,624.5 \text{ deaths per 100,000 operations per year}$$

and

$$\hat{I}_{H,10^5} = 100,000 \cdot \frac{2,109}{68,169} = 3,093.8 \text{ deaths per 100,000 operations per year.}$$

To test whether the rates differ significantly, that is, $H_0 : I_{H,10^5} = I_{C,10^5}$ versus $H_0 : I_{H,10^5} \neq I_{C,10^5}$, the c terms drop out and we can use a two-sample test of proportions. The sample sizes are large and under H_0,

$$\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$

Using $\alpha = 0.05$, the critical region is $C_{0.05} = \{|z| > 1.96\}$. Here

$$\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0.0162 - 0.0309}{\sqrt{\frac{0.0162(1-0.0162)}{146200} + \frac{0.0309(1-0.0309)}{68169}}} = -19.8 < -1.96.$$

So we reject H_0 and conclude that the death rate when cyclopropane is used is significantly higher than when halothane is used.

(b) Using all the operations, the weights are given in Table 1, with w_i being the number of operations in physical status category i divided by the total number of operations.

Status	Operations	Weight
Unknown	69,239	0.162
1	185,919	0.435
2	104,286	0.244
3	29,491	0.069
4	3,419	0.008
5	21,797	0.051
6	11,112	0.026
7	2,137	0.005
Total	427,400	1.000

Table 1: Weights for direct standardization

Denote the standardized incidence rates by $I_{H_{adj},10^5}$ and $I_{C_{adj},10^5}$, where for $j \in \{\text{C,H}\}$,

$$\hat{I}_{j_{adj},10^5} = c \cdot \hat{p}_{j_{\text{adj}}} = c \cdot \frac{\sum_{k=1}^{K} w_k \hat{p}_{jk}}{\sum_{k=1}^{K} w_k}.$$

The values of $\hat{p}_{jk}$ are given in Tables 2 and 3 for halothane and cyclopropane, respectively. Using these, $\hat{I}_{H_{adj},10^5} = 100,000 \times 0.018091 = 1,809.1$ deaths per 100,000 operations per year and $\hat{I}_{C_{adj},10^5} = 100,000 \times 0.026843 = 2,684.3$ deaths per 100,000 operations per year.

To test $H_0 : I_{H_{adj},10^5} = I_{C_{adj},10^5}$ versus $H_A : I_{H_{adj},10^5} \neq I_{C_{adj},10^5}$, note that the constant c cancels out again and we have

$$Z = \frac{\hat{I}_{H_{adj},10^5} - \hat{I}_{C_{adj},10^5}}{\sqrt{\widehat{\text{Var}}(\hat{I}_{H_{adj},10^5} - \hat{I}_{H_{adj},10^5})}} \sim N(0,1).$$

Here

$$\widehat{\text{Var}}(\hat{I}_{H_{adj},10^5} - \hat{I}_{H_{adj},10^5}) = \frac{\sum_{k=1}^{K} w_k^2 \left(\widehat{\text{Var}}(\hat{p}_{Hk}) + \widehat{\text{Var}}(\hat{p}_{Ck})\right)}{\left(\sum_{k=1}^{K} w_k\right)^2}$$

with $\widehat{\mathrm{Var}}(\hat{p}_{Hk})$ and $\widehat{\mathrm{Var}}(\hat{p}_{Ck})$ given in Tables 2 and 3 for halothane and cyclopropane, respectively.

So $Z = (0.0181 - 0.0268)/\sqrt{4.95954 \times 10^{-7}} = -12.4 < -1.96$, and as in part (a) we reject H_0.

SAS code for direct standardization plus edited output:

```
proc stdrate data=hw7q2b refdata=hw7q2c method=direct
           effect stat=rate(mult=100000);
    population group=group event=deaths total=ops;
reference total=ops_total;
strata status / stats effect;
```

```
             Directly Standardized Rate Estimates
                    Rate Multiplier = 100000

         --------Study Population-------  -Reference Population-
         Observed  Population-     Crude  Expected   Population-
group      Events          Time      Rate    Events          Time
cyclo        2109         68169    3093.8   11472.6        427400
halo         2375        146200    1624.5    7731.9        427400

     Directly Standardized Rate Estimates
           Rate Multiplier = 100000

         -----------Standardized Rate----------
                   Standard      95% Normal
group    Estimate     Error   Confidence Limits
cyclo      2684.3   62.7938    2561.2    2807.4
halo       1809.1   37.6699    1735.2    1882.9


              Rate Effect Estimates (Rate Multiplier = 100000)

                                     Log
-------group------        Rate        Rate    Standard
   cyclo      halo       Ratio       Ratio       Error          Z    Pr > |Z|

  2684.3    1809.1      1.4838      0.3946      0.0313      12.60      <.0001
```

(c) In calculating the standardized incidence ratio the constant c again cancels out, so we can work with the proportions.

To test $H_0 : \pi_{\text{halothane}}/\pi_{\text{overall}} = 1$ versus $H_A : \pi_{\text{halothane}}/\pi_{\text{overall}} \neq 1$ we first calculate

$$s = \frac{\hat{p}_{\text{halothane}}}{\hat{p}_{\text{overall}}} = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} N_k m_k / M_k} = \frac{O}{E}.$$

If $\widehat{\mathrm{Var}}(O) = \sum_k n_k$ and $\widehat{\mathrm{Var}}(E) = \sum_k \left(\frac{N_k}{M_k}\right)^2 m_k$, then $\widehat{\mathrm{Var}}(s) = \dfrac{\widehat{\mathrm{Var}}(O) + s^2 \widehat{\mathrm{Var}}(E)}{E^2}$,

Status (k)	Weight (w_k)	Operations	Deaths	$\hat{p}_{Hk}$	$\widehat{\text{Var}}(\hat{p}_{Hk})$
Unknown	0.16200	23684	419	0.01769	0.000000734
1	0.43500	65936	125	0.00190	0.000000029
2	0.24400	36842	560	0.01520	0.000000406
3	0.06900	8918	617	0.06919	0.000007221
4	0.00800	1170	182	0.15556	0.000112272
5	0.05100	6579	74	0.01125	0.000001690
6	0.02600	2632	287	0.10904	0.000036912
7	0.00500	439	111	0.25285	0.000430332
Total	1.00000	146200	2375	—	—

Table 2: Halothane estimates for direct standardization

Status (k)	Weight (w_k)	Operations	Deaths	$\hat{p}_{Ck}$	$\widehat{\text{Var}}(\hat{p}_{Ck})$
Unknown	0.16200	10147	297	0.02927	0.000002800
1	0.43500	27444	91	0.00332	0.000000120
2	0.24400	14097	361	0.02561	0.000001770
3	0.06900	3814	403	0.10566	0.000024777
4	0.00800	681	127	0.18649	0.000222778
5	0.05100	7423	101	0.01361	0.000001808
6	0.02600	3814	476	0.12480	0.000028639
7	0.00500	749	253	0.33778	0.000298646
Total	1.00000	68169	2109	—	—

Table 3: Cyclopropane estimates for direct standardization

and under H_0, $Z = (s-1)/\sqrt{\widehat{\text{Var}}(s)} \sim N(0,1)$.

Using the data in Table 4, $O = 2,375$, $E = 2,695.12$, $s = 2,375/2,695.12 = 0.88$, $\widehat{\text{Var}}(O) = 2,375$ and $\widehat{\text{Var}}(E) = 848.27$.

So $\widehat{\text{Var}}(s) = \dfrac{2375 + 0.88^2 \cdot 848.27}{2695.12^2} = 0.00043$ and $Z = \dfrac{0.88-1}{\sqrt{0.00043}} = -5.73 < -1.96$.

Thus we reject H_0 and we conclude that the mortality rate for halothane is significantly less than the overall death rate.

SAS uses a somewhat different estimator for the variance. It uses

$$\widehat{\text{Var}}(s) = \frac{O}{E^2} = \frac{2375}{2695.12^2} = 0.00033$$

and this yields

$$Z = \frac{0.88-1}{\sqrt{0.00033}} = -6.57.$$

```
proc stdrate data=hw7q2 refdata=hw7q2 method=indirect stat=rate;
    population event=dth_halo total=ops_halo;
reference event=dth_total total=ops_total;
strata status / stats smr;
```

```
                    Standardized Morbidity/Mortality Ratio

Observed  Expected            Standard      95% Normal
  Events    Events        SMR     Error   Confidence Limits        Z  Pr > |Z|

    2375   2695.12     0.8812    0.0181    0.8458    0.9167    -6.57    <.0001
```

	Reference		Halothane			
Status (k)	m_k	M_k	n_k	N_k	$N_k m_k / M_k$	$\left(\frac{N_k}{M_k}\right)^2 m_k$
Unknown	1,378	69,239	419	23,684	471.36	161.23
1	445	185,919	125	65,936	157.82	55.97
2	1,856	104,286	560	36,842	655.68	231.64
3	2,135	29,491	617	8,918	645.62	195.23
4	590	3,419	182	1,170	201.90	69.09
5	314	21,797	74	6,579	94.77	28.61
6	1,392	11,112	287	2,632	329.71	78.10
7	673	2,137	111	439	138.25	28.40
Total	8,783	427,400	2,375	146,200	2695.12	848.27

Table 4: Counts for indirect standardization

Question 2

(a) Tables 5 and 6 give the calculations for the Kaplan-Meier curves for group 1 and group 2, respectively.

Figure 1 has the Kaplan-Meier survival function estimates for the two groups, plotted using the R code:

```
library("survival")

fit <- survfit(Surv(timedeath, death)~group,conf.type="none")

pdf("HW8_Surv.pdf",width=11,height=8.5)

plot(fit,xlab="Time (days)",ylab="S(t)",lwd=c(1,3),cex.axis=1.6,
   main="Kaplan-Meier estimates for the two groups",cex.lab=1.6,
 cex.main=1.6,cex.sub=1.6)

legend(425,1.0,c("New treatment","Placebo"),lwd=c(3,1),cex=1.6)

dev.off()
```

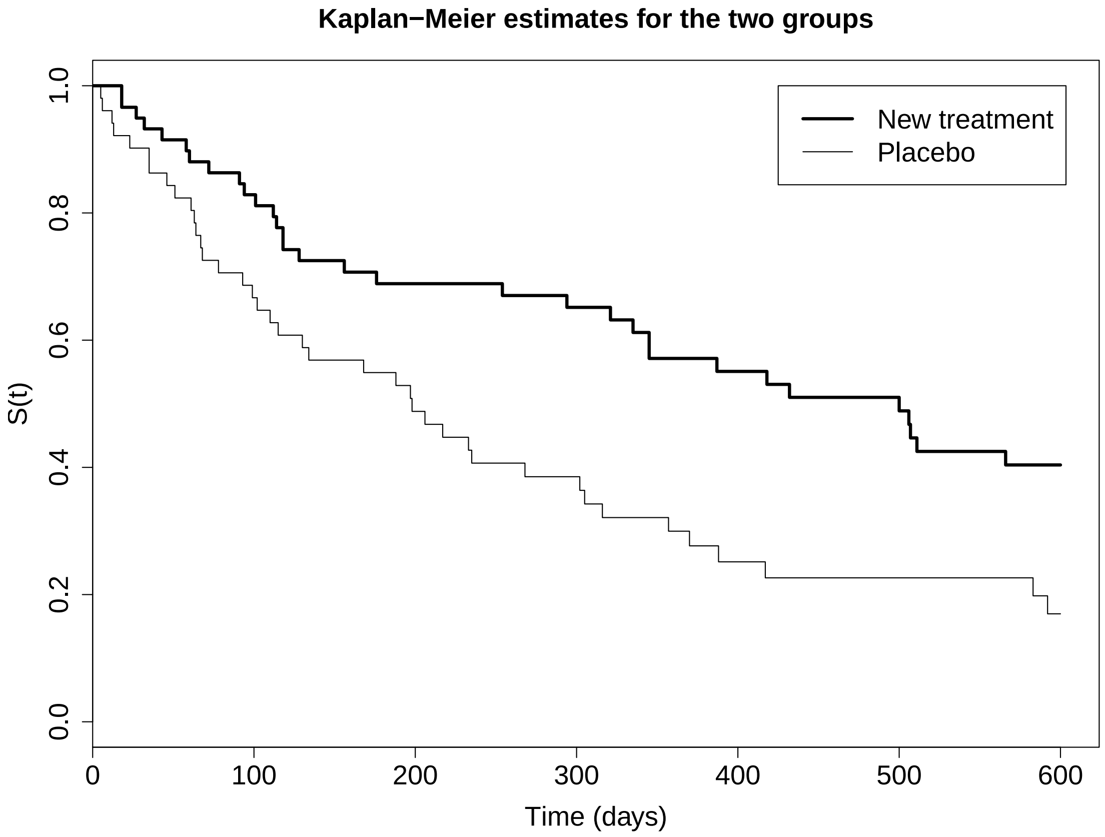


Figure 1: Calculation of Kaplan-Meier estimate for group 1

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$	$\hat{S}(t_{(j)})$
5	1	0	51	0.98039
6	1	0	50	0.96078
12	1	0	49	0.94118
13	1	0	48	0.92157
23	1	0	47	0.90196
35	2	0	46	0.86275
46	1	0	44	0.84314
51	1	0	43	0.82353
61	1	0	42	0.80392
63	1	0	41	0.78431
64	1	0	40	0.76471
67	1	0	39	0.74510
68	1	0	38	0.72549
78	1	0	37	0.70588
93	1	0	36	0.68627
99	1	0	35	0.66667
102	1	0	34	0.64706
110	1	0	33	0.62745
115	1	0	32	0.60784
130	1	0	31	0.58824
134	1	0	30	0.56863
168	1	1	29	0.54902
188	1	0	27	0.52869
197	1	0	26	0.50835
198	1	0	25	0.48802
206	1	0	24	0.46768
217	1	0	23	0.44735
233	1	0	22	0.42702
235	1	1	21	0.40668
268	1	0	19	0.38528
302	1	0	18	0.36387
305	1	0	17	0.34247
316	1	0	16	0.32106
357	1	1	15	0.29966
370	1	1	13	0.27661
388	1	0	11	0.25146
417	1	1	10	0.22632
583	1	0	8	0.19803
592	1	6	7	0.16974

Table 5: Calculation of Kaplan-Meier estimate for group 1

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$	$\hat{S}(t_{(j)})$
18	2	0	59	0.96610
27	1	0	57	0.94915
32	1	0	56	0.93220
43	1	1	54	0.91494
58	1	0	53	0.89768
60	1	0	52	0.88041
72	1	0	51	0.86315
91	1	0	50	0.84589
94	1	0	49	0.82863
101	1	0	48	0.81136
112	1	0	47	0.79410
114	1	0	46	0.77684
118	2	0	45	0.74231
128	1	2	43	0.72505
156	1	0	40	0.70692
176	1	1	39	0.68879
254	1	0	37	0.67018
294	1	2	36	0.65156
321	1	0	33	0.63182
335	1	0	32	0.61207
345	2	0	30	0.57127
387	1	0	28	0.55087
418	1	0	27	0.53046
432	1	1	26	0.51006
500	1	0	24	0.48881
506	1	0	23	0.46756
507	1	0	22	0.44630
511	1	0	21	0.42505
566	1	19	20	0.40380

Table 6: Calculation of Kaplan-Meier estimate for group 2

(b) Tables 7 and 8 have data for doing the log-rank test "by hand". We want to test $H_0 : S_1(t) = S_2(t)$ for all t against $H_A : S_1(t) \neq S_2(t)$ for at least one t. Under H_0, $X = (O_1 - E_1)^2/V_1 \sim \chi^2_1$. So the critical region is $C_{0.05} = \{X : X > \chi^2_{1,0.95} = 3.84\}$.

Using the data in Tables 7 and 8:

$$X = (O_1 - E_1)^2/V_1 = (40 - 27.8195)^2/16.7682 = 8.85 > 3.84.$$

So we reject H_0 and conclude that the new treatment tends to increase survival time in comparison with placebo. We confirm the result using SAS:

```
proc lifetest;
   time timedeath*death(0);
   strata group;
```

```
       Test of Equality over Strata
                                    Pr >
Test      Chi-Square      DF    Chi-Square

Log-Rank      8.8468       1      0.0029
```

Note that if one does the calculations using group 2 in X one obtains the same value for the statistic. In that case $O_2 = 32$ and $E_2 = 44.1805$.

(c) Let X be an indicator of being in group 2, that is $X = 0$ if in group 1 and $X = 1$ if in group 2. The model is

$$\log \lambda(t) = \log \lambda_0(t) + \beta X$$

We want to test $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$. We have not covered details of how the test is conducted and so will use just the output from SAS or R. Using SAS with the option "rl" in the "model" statement to obtain the "risk limits", that is, a 95% confidence interval for the hazard ratio, needed in part (d):

```
proc phreg;
   model timedeath*death(0) = group01 / ties=exact rl;
```

```
The PHREG Procedure

               Analysis of Maximum Likelihood Estimates

                    Parameter      Standard
Parameter    DF      Estimate         Error    Chi-Square    Pr > ChiSq

group01       1      -0.69719       0.23939        8.4820        0.0036

  Analysis of Maximum Likelihood Estimates

               Hazard      95% Hazard Ratio
Parameter       Ratio      Confidence Limits

group01         0.498       0.311       0.796
```

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
5	1	51	0	59	1	110	0.46364	0.24868
6	1	50	0	59	1	109	0.45872	0.24830
12	1	49	0	59	1	108	0.45370	0.24786
13	1	48	0	59	1	107	0.44860	0.24736
18	0	47	2	59	2	106	0.88679	0.48889
23	1	47	0	57	1	104	0.45192	0.24769
27	0	46	1	57	1	103	0.44660	0.24715
32	0	46	1	56	1	102	0.45098	0.24760
35	2	46	0	55	2	101	0.91089	0.49107
43	0	44	1	54	1	98	0.44898	0.24740
46	1	44	0	53	1	97	0.45361	0.24785
51	1	43	0	53	1	96	0.44792	0.24729
58	0	42	1	53	1	95	0.44211	0.24665
60	0	42	1	52	1	94	0.44681	0.24717
61	1	42	0	51	1	93	0.45161	0.24766
63	1	41	0	51	1	92	0.44565	0.24705
64	1	40	0	51	1	91	0.43956	0.24635
67	1	39	0	51	1	90	0.43333	0.24556
68	1	38	0	51	1	89	0.42697	0.24467
72	0	37	1	51	1	88	0.42045	0.24367
78	1	37	0	50	1	87	0.42529	0.24442
91	0	36	1	50	1	86	0.41860	0.24337
93	1	36	0	49	1	85	0.42353	0.24415
94	0	35	1	49	1	84	0.41667	0.24306
99	1	35	0	48	1	83	0.42169	0.24387
101	0	34	1	48	1	82	0.41463	0.24271
102	1	34	0	47	1	81	0.41975	0.24356
110	1	33	0	47	1	80	0.41250	0.24234
112	0	32	1	47	1	79	0.40506	0.24099
114	0	32	1	46	1	78	0.41026	0.24195
115	1	32	0	45	1	77	0.41558	0.24287
118	0	31	2	45	2	76	0.81579	0.47659
128	0	31	1	43	1	74	0.41892	0.24343
130	1	31	0	42	1	73	0.42466	0.24432
134	1	30	0	42	1	72	0.41667	0.24306
156	0	29	1	40	1	69	0.42029	0.24365
168	1	29	0	39	1	68	0.42647	0.24459
176	0	28	1	39	1	67	0.41791	0.24326
188	1	27	0	38	1	65	0.41538	0.24284

Table 7: First part of table for log-rank test

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
197	1	26	0	38	1	64	0.40625	0.24121
198	1	25	0	38	1	63	0.39683	0.23936
206	1	24	0	38	1	62	0.38710	0.23725
217	1	23	0	38	1	61	0.37705	0.23488
233	1	22	0	38	1	60	0.36667	0.23222
235	1	21	0	38	1	59	0.35593	0.22924
254	0	20	1	37	1	57	0.35088	0.22776
268	1	19	0	36	1	55	0.34545	0.22612
294	0	18	1	36	1	54	0.33333	0.22222
302	1	18	0	35	1	53	0.33962	0.22428
305	1	17	0	35	1	52	0.32692	0.22004
316	1	16	0	35	1	51	0.31373	0.21530
321	0	15	1	33	1	48	0.31250	0.21484
335	0	15	1	32	1	47	0.31915	0.21729
345	0	15	2	30	2	45	0.66667	0.43434
357	1	15	0	28	1	43	0.34884	0.22715
370	1	13	0	28	1	41	0.31707	0.21654
387	0	12	1	28	1	40	0.30000	0.21000
388	1	11	0	27	1	38	0.28947	0.20568
417	1	10	0	27	1	37	0.27027	0.19722
418	0	9	1	27	1	36	0.25000	0.18750
432	0	9	1	26	1	35	0.25714	0.19102
500	0	9	1	24	1	33	0.27273	0.19835
506	0	9	1	23	1	32	0.28125	0.20215
507	0	9	1	22	1	31	0.29032	0.20604
511	0	9	1	21	1	30	0.30000	0.21000
566	0	9	1	20	1	29	0.31034	0.21403
583	1	8	0	19	1	27	0.29630	0.20850
592	1	7	0	19	1	26	0.26923	0.19675
	40		32				27.8195	16.7682

Table 8: Second part of table for log-rank test

Equivalently, in R:

```
> coxph(Surv(timedeath, death)~group01)
> summary(coxph(Surv(timedeath, death)~group01))

Call:
coxph(formula = Surv(timedeath, death) ~ group)

  n= 110, number of events= 72

         coef exp(coef) se(coef)      z Pr(>|z|)
group -0.6972    0.4980   0.2394 -2.913  0.00358 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
group     0.498      2.008    0.3115    0.7961
```

In either case the p-value associated with the test of $\beta = 0$ is 0.0036. So we reject H_0 and we conclude that the time to death differs significantly between the two treatment groups, with the new treatment being better than placebo.

(d) The hazard ratio estimate is $\exp(\hat{\beta}) = \exp(-0.6972) = 0.498$. The 95% CI can be obtained from SAS or R output. Alternatively, as in Table 16.7 of the text, a 95% CI for β is given by $\hat{\beta} \pm 1.96 \cdot \mathrm{SE}(\hat{\beta}) = -0.6972 \pm 1.96 \cdot 0.2394 = (-1.166, -0.228)$. Taking antilogs gives the 95% CI for the hazard ratio as $(0.311, 0.796)$.

(e) The model is now

$$\log \lambda(t) = \log \lambda_0(t) + \beta_{group} X_{group} + \beta_{age} X_{age}$$

and the SAS code and output as below. The p-value for age is 0.0339, so age has a significant efect on survival. The associated hazard ratio is greater than one, so the hazard increases with age, that is, older age is associated with poorer survival probability. The hazard ratio for the treatment group variable has not changed substantially.

```
proc phreg;
   model timedeath*death(0) = group01 age / ties=exact rl;

The PHREG Procedure

                   Analysis of Maximum Likelihood Estimates

                 Parameter   Standard                                 Hazard
Parameter   DF    Estimate      Error   Chi-Square   Pr > ChiSq        Ratio

group01      1    -0.72071    0.24017       9.0047       0.0027        0.486
age          1     0.04922    0.02320       4.5012       0.0339        1.050
```

(f) Looking at Table 5, the time at which $\widehat{S}(t)$ is first ≤ 0.5 is at $t = 198$.

BIOS 662, Fall 2018
Midterm Examination

Assigned: Tuesday, October 23

Due: 11:59 PM on Thursday, November 1

Instructions:

The midterm exam is a take-home exam. It is due just before midnight on Thursday, November 1. Please put your completed exam in the BIOS662 mailbox in the Department of Biostatistics. (I will collect the exams from the mailbox *very* early on the Friday.)

Do not discuss the exam with anyone. If you need clarification on any question you may contact me by email (david_couper@unc.edu). The graders will not answer questions about the exam but you may continue to ask them more general questions during their office hours.

Please sign your name on the exam indicating that you did not receive assistance from anyone. Note that obtaining help from anyone other than me is an Honor Code violation.

You may use software to perform calculations but make sure that you include enough information in your answers that I can see what you have done. For instance, include the SAS or R statements you used, not just the output. For all problems involving statistical testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution (when this is relevant), whether you reject the null, and *an interpretation of the results in a language suitable for investigators.* You may do tests either by determining whether the test statistic falls in the critical region or by obtaining a p-value (you don't have to use both approaches). For any method you use, be sure to state and check the required assumptions. All tests should be performed at the $\alpha = 0.05$ significance level.

Datasets and manuscripts referred to in the questions are available on the Sakai web site, in the folder "Midterm materials" under "Resources". All data are fictitious, even when based on real studies.

1. Write and sign a statement confirming you have not obtained assistance from anyone.

2. The gestational age (GA) of an embryo or newborn infant is approximately the time since the mother's last menstrual period. GA is used to determine whether the pregnancy has reached full term (40 weeks). An infant born prior to 37 weeks GA is regarded as being premature or preterm. There has been some epidemiologic evidence for an association between a pregnant woman having periodontal disease and giving birth prematurely or for the infant's birthweight to be below normal.

   The MOTOR Study was a randomized clinical trial to investigate whether treating periodontal disease in pregnant women reduces the risk of preterm delivery and/or increases average birthweight. (The manuscript by Offenbacher *et al.* assigned in homework 1 provides information about the study. It is not necessary to consult that manuscript in order to complete this exam.)

   In MOTOR about 1,800 pregnant women with periodontal disease were randomized into two treatment groups. The "prenatal treatment" group received periodontal therapy early in pregnancy. The "post-partum treatment" group received periodontal therapy a few weeks after delivery.

   Two measures of GA at birth were available. The more accurate one was made using an ultrasound examination early in the pregnancy. This is given in weeks, calculated from days. So, for instance, 38.1429 corresponds to a GA of 38 weeks and 1 day. The other measure of GA was an estimate made when the infant was born and is in whole weeks.

   Various periodontal measurements were made around each tooth at baseline (early in the pregnancy, before randomization) and were repeated shortly after giving birth. For each woman, the measurements on the teeth were averaged to give a summary score for each type of measurement. One such measurement is probing depth, in millimeters, with larger values indicating more periodontal disease.

   The file "Midterm_BWT.dat" and corresponding SAS and R datasets contain data from a subset of the live births in the trial. The columns in the file are, respectively, ID (participant identifier), group (treatment group; 1 = prenatal, 2 = post-partum), rand_month (month in which the woman was randomized, with 1=January, 2=February, etc.), birth_month (month in which the woman gave birth), GA_ultra (GA estimated by ultrasound), GA_est (GA estimated at birth), ppnum (number of previous pregnancies, as a character variable, with ≥ 3 denoted as "3+"), PD_pre (average pocket depth at the time of randomization), PD_post (average pocket depth after delivery).

   If you detect any apparently incorrect data values, make a note of these, explain why you believe the data are incorrect, and set the values to missing. Regard a value as being incorrect only if it is clearly impossible. You may assume that there are no errors in the pocket depth variables. *Set incorrect values to missing.*

(a) Is the ultrasound version of GA approximately normally distributed?

(b) Do the means of the two gestational age variables differ?

(c) After taking into account any difference in the means (whether or not statistically significant), do the shapes of the distributions of the two gestational age variables differ?

(d) Classify both versions of gestational age into 3 intervals, $(0, 37)$, $[37, 40)$, and $[40, \infty)$. Determine how well the two versions agree and provide a 95% confidence interval for the true agreement.

(e) Is the number of women randomized in each month consistent with the number of days in each month?

(f) Without doing any additional tests, comment on how the distribution of the number of births each month compares with that of the number of women randomized each month.

For parts (g) and (h), dichotomize the ultrasound version of GA to define preterm delivery.

(g) Does the risk of preterm delivery vary monotonically with the number of previous pregnancies?

(h) Based on this study, is treating periodontal disease in pregnant women effective in terms of reducing the risk of prematurity?

The effect of the periodontal therapy on mean birthweight was smaller than the investigators had expected – there was not a statistically significant difference between the mean birthweights in the two treatment groups. One potential explanation for the lack of effect is that the periodontal therapy provided may not have been intensive enough to yield a substantial and sustained reduction in the amount of periodontal disease.

(i) Ignoring treatment group, is there a difference between the mean average pocket depth at randomization and the mean average pocket depth after delivery.?

(j) Did the mean change in average pocket depth differ between the two treatment groups?

(k) Based on the data on average pocket depth, discuss the effectiveness of the periodontal therapy and the consequences for the potential to affect birthweight.

3. The evidence on which the MOTOR Study was based includes data from case-control studies. The file "Midterm_CC.dat" contains data from one such study. Cases were defined as women who had given birth prematurely (< 37 weeks GA). Controls had full-term babies. The women had a periodontal examination soon after giving birth. Those who had moderate or severe periodontal disease (based on the investigators' criteria) were classified as "exposed" to periodontal disease and those who had no evidence of periodontal disease or just mild disease were classified as "unexposed". Age was considered to be a potential confounder of the association.

   The columns in the file are, respectively, ID (participant identifier), case (indicator of premature birth case status; 1 = case, 0 = control), exposed (indicator of periodontal disease status; 1 = moderate or severe periodontal disease, 0 = no more than mild periodontal disease), and age_group (the age group of the mother, with 1 representing the youngest age group and 3 the oldest). You may assume there are no errors in this dataset.

   First assume this was an unmatched case-control study.

   (a) Determine whether premature birth case status is associated with being exposed to periodontal disease.

   (b) Provide an estimate for a measure of the association between exposure and case status and give a 95% confidence interval for the true measure.

   (c) Repeat part (b) above, taking age group into account.

   (d) Does age group appear to be a confounder? Is the pooled estimate in part (c) a reasonable way to summarize the association here?

   The data were actually from an individually-matched case-control study, with one control per case, matching on age and number of previous pregnancies. The first 4 characters of the ID indicate the case-control pair (the case and the control in the pair have the same first 4 characters). The final character of the ID is 1 for the case in the pair and 0 for the control.

   (e) Repeat parts (a) and (b) above assuming an individually-matched case-control design.

   (f) Which of the estimates of the measure of association in (b), (c) and (e) is most appropriate? Justify your choice.

   (g) Discuss the strength of the evidence from this case-control study for an association between periodontal disease and preterm delivery.

BIOS 662, Fall 2018
Solution to Midterm Exam

Question 1

Question 2

Investigation of potentially incorrect data values: Below is a list of reasons for setting various values to missing, with the IDs of the corresponding women given in parentheses.

- A gestational age of 75 is impossible (GA_ultra for M2349).
- Months are numbered 1 through 12, so rand_month values of 0 (M1190) and 15 (M1722) are errors.
- Number of previous pregnancies cannot be negative, so values of -9 are errors (M1190 and M1410).

(a) A histogram, stem-and-leaf plot or boxplot suggests substantial skewness in the data, with a long tail towards lower values of GA. A QQ plot also shows substantial departure from normality See, for instance, Figure 1. To test for normality, use the Kolmogorov-Smirnov test. We need the Lilliefors version to adjust for having to estimate the mean and variance.

H_0 : GA_est is normally distributed; H_A : GA_est is not normally distributed.

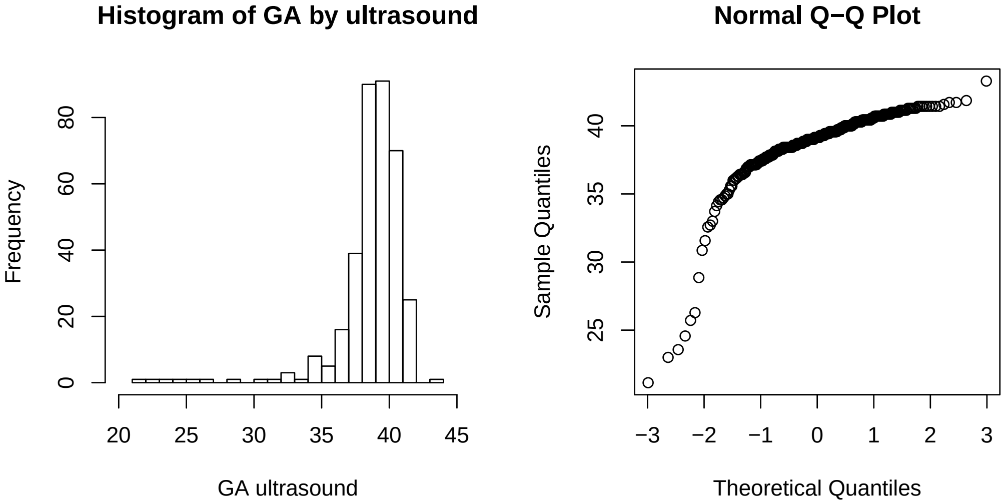
SAS gives the p-value for the Lilliefors version automatically; R needs the function lillie.test.

In the SAS output we see $D = 0.186$ with $p < 0.01$, so we reject H_0 and conclude that GA_est is not normally distributed.

```
proc univariate normal;
   var ga_ultra;
```

```
                   Tests for Normality

Test                  --Statistic---    -----p Value------

Shapiro-Wilk          W     0.698496    Pr < W     <0.0001
Kolmogorov-Smirnov    D     0.185894    Pr > D     <0.0100
Cramer-von Mises      W-Sq  3.729445    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  22.17517    Pr > A-Sq  <0.0050
```

Figure 1: Histogram and normal QQ plot for GA by ultrasound



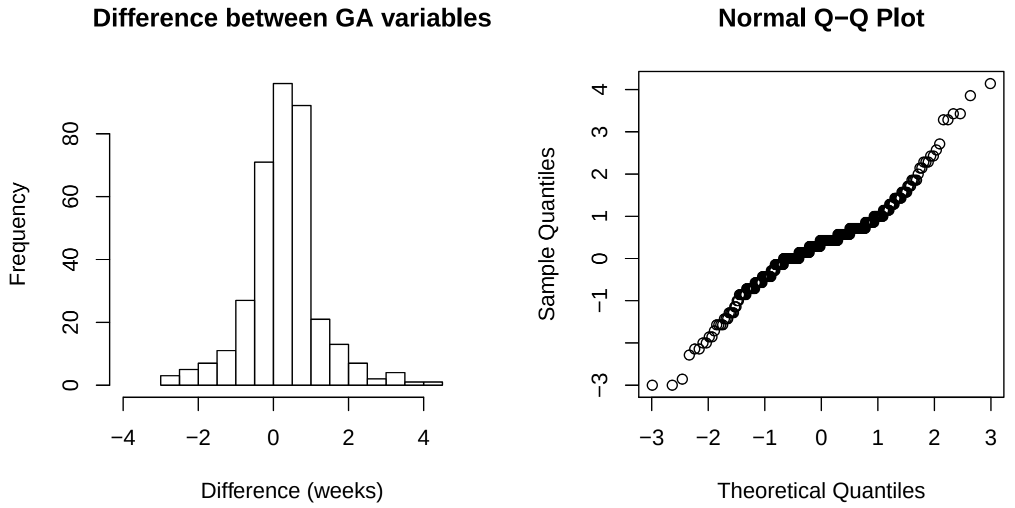
(b) Because the data are paired (each woman has gestational age estimated by each method), we don't have two independent samples. So let $Y_i = X_{1i} - X_{2i}$ where X_{1i} is the GA by ultrasound and X_{2i} the GA estimated at birth for woman i. Assume $E(Y_i) = \mu$ for all i. The observations on different women are independent. We want to test $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} \neq 0$. The histogram in Figure 2 shows that the distribution of differences is reasonably symmetric but the QQ plot suggests that it may not be reasonable to assume normality. The sample size is large, so we will use the CLT / Slutsky's Theorem to give a test using the Z statistic. (Using a t-test would also be reasonable.) The critical region is $C_{0.05} = \{Z : |Z| > 1.96\}$.

$$Z = \frac{\bar{Y} - 0}{s/\sqrt{n}} = \frac{0.3180}{0.960/\sqrt{358}} = 6.268 > 1.96.$$

So we reject H_0 and conclude that the mean GA by ultrasound differs from the mean GA estimated at birth. The sample means are 38.73 and 38.41 months respectively, so the mean GA by ultrasound appears to be larger than the mean GA estimated at birth. A t-test gives similar results (see SAS output below).

```
The TTEST Procedure

Variable:  ga_diff

   N        Mean     Std Dev
 358      0.3180      0.9601

   DF    t Value    Pr > |t|
   357      6.27      <.0001
```

Figure 2: Histogram and normal plot for differences between GA variables



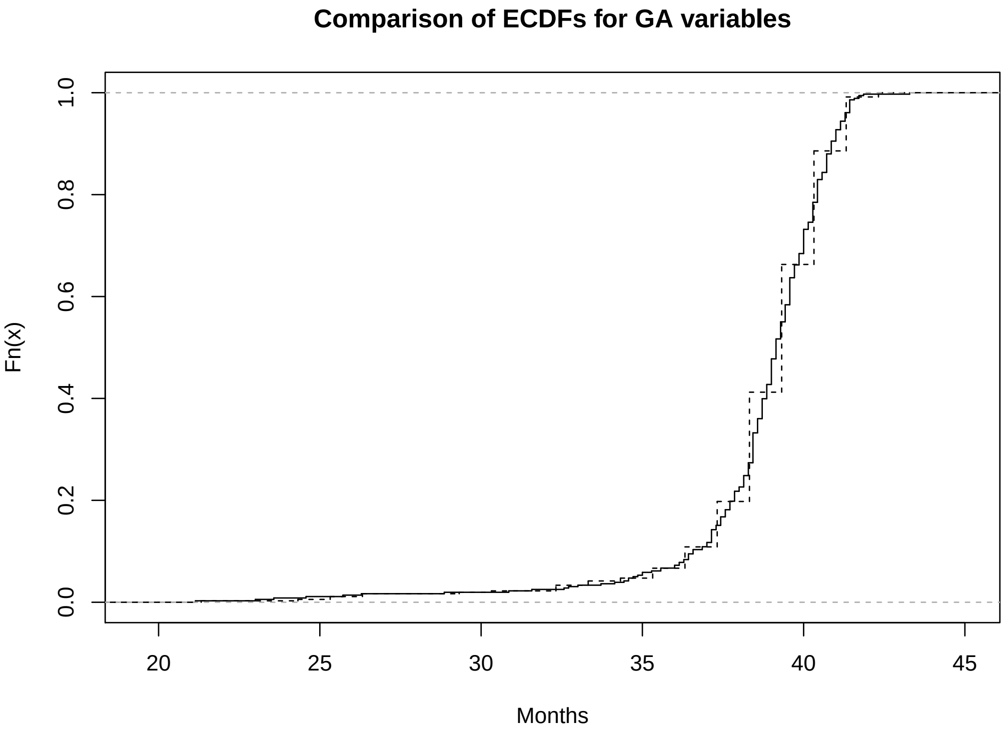
(c) Because the means differ by 0.3180 months, we add this to each GA_est observation. This eliminates the difference between the means. One way to test whether there are other differences between the distributions after eliminating the difference between the means is to use the Kolmogorov-Smirnov test. The KS test in this situation assumes independence of the two samples, but here the two measures are used for each infant and so the assumption of independence is violated. As we don't have a test that does not make the independence assumption we will use the KS test even though it is not ideal.

We want to test $H_0 : F_1(y) = F_2(y)$ for all y versus $H_A : F_1(y) \neq F_2(y)$ for at least one y, where F_1 and F_2 are the CDFs of the two GA variables. To run the KS test we first need to create a single variable that has the GA values of both types, along with an indicator of which are the GA_ultra and which the GA_est values. In the SAS code that follows these are the variables GA and GA_GROUP, respectively. The "mc" in the last line of the code is so that SAS doesn't take forever to run. The p-values for both the asymptotic and the exact versions of the test are < 0.05, so we reject the hypothesis that the two distributions have the same shape. However, looking at the plot in Figure 3, the only noticeable difference is that the step function for the GA_EST version has bigger steps because it is given in whole weeks whereas the other version is in weeks plus days. So, although the difference is statistically significant it isn't really a meaningful difference.

A test for equality of variances for the two variables does not reject the null hypothesis of equality (though here too the assumption of independence of the two samples is not met. The histogram in Figure 4 shows that the distribution of differences is reasonably symmetric.

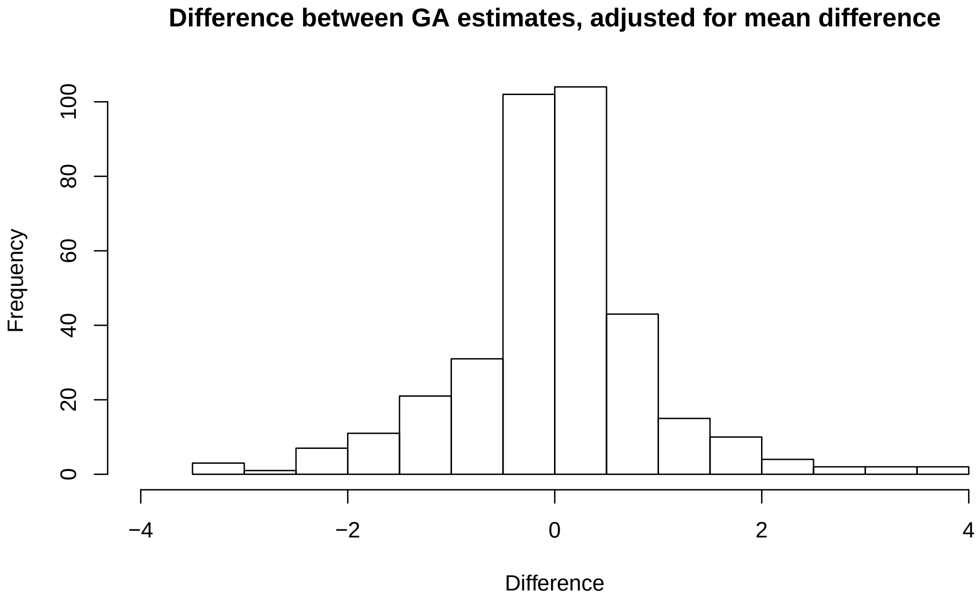
```
proc npar1way;
   class ga_group;
   var ga;
   exact ks / mc;

  Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2|                 0.1397
Asymptotic Pr >  D                0.0019

Monte Carlo Estimate
Exact Pr >= D                     0.0010
```

Figure 3: ECDFs for the GA variables after eliminating the difference in means



(d) The following table has the two versions of GA classified as stated. The observed proportion of agreement is $p_a = (33 + 170 + 91)/359 = 0.82$. As seen from the SAS output, the chance-corrected measure of agreement is $\kappa = 0.68$ and the associated 95% CI for the true agreement is $(0.61, 0.75)$. The agreement is reasonable, though not great.

Figure 4: Histogram of the difference between the GA variables after eliminating the difference in means



```
Table of ga_cat by ga_catp

ga_cat     ga_catp

Frequency|1. LT 37|2. 37-40|3. GE 40|  Total
---------+--------+--------+--------+
1. LT 37 |     33 |      6 |      0 |     39
---------+--------+--------+--------+
2. 37-40 |      6 |    170 |     30 |    206
---------+--------+--------+--------+
3. GE 40 |      0 |     22 |     91 |    113
---------+--------+--------+--------+
Total          39      198      121      358

Frequency Missing = 1


                      Kappa Statistics

Statistic          Value       ASE     95% Confidence Limits
------------------------------------------------------------
Simple Kappa      0.6826    0.0366       0.6108       0.7544
Weighted Kappa    0.7182    0.0335       0.6526       0.7838

Effective Sample Size = 358
Frequency Missing = 1
```

(e) If randomization is spread evenly across the calendar year, we would expect the proportion of women randomized in January to be 31/365, the proportion in February to be 28/365, etc. We need to conduct a goodness of fit test to see whether the distribution of the number of women randomized is consistent with this. This is similar to the example starting on page 33 of the overheads on "Categorical Data: Contingency Tables" with 12 probabilities rather than just 3.

$$H_0 : \pi_{\text{Jan}} = 31/365 = 0.084932, \pi_{\text{Feb}} = 28/365 = 0.076712, \ldots \pi_{\text{Dec}} = 31/365 = 0.084932$$

$$H_A : \text{at least one of the equalities is false.}$$

Using the SAS code and output below, $p = 0.0004$, so we reject H_0 and conclude that the proportion of women randomized each month is not consistent with the number of days in the month. Some slight variation from expected is likely to be because of the number of weekend days in a particular month. But we see that December, in particular, has a much lower percent randomized than expected just by the length of the month. Because of holidays in December, which means fewer working days and also potential participants having other priorities than being in a clinical trial, trials often struggle to randomize many participants in that month.

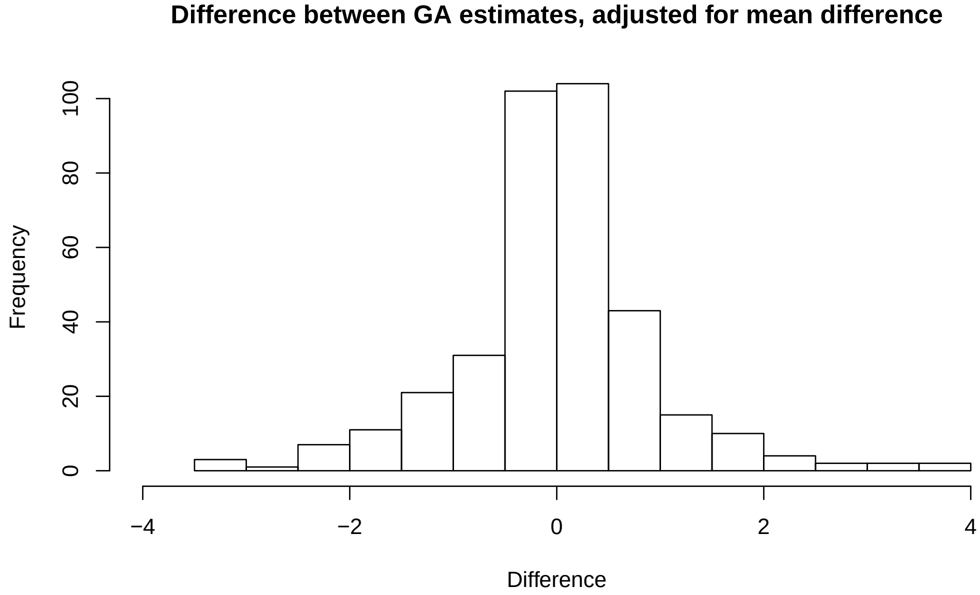
```
proc freq data=bw;
   table rand_month / testp=(8.4932 7.6712 8.4932 8.2192 8.4932 8.2192
        8.4932 8.4932 8.2192 8.4932 8.2192 8.4932);
```

```
The FREQ Procedure
                                            Test
rand_month    Frequency     Percent     Percent
-----------------------------------------------
         1          22        6.16        8.49
         2          17        4.76        7.67
         3          42       11.76        8.49
         4          24        6.72        8.22
         5          41       11.48        8.49
         6          35        9.80        8.22
         7          40       11.20        8.49
         8          22        6.16        8.49
         9          32        8.96        8.22
        10          39       10.92        8.49
        11          29        8.12        8.22
        12          14        3.92        8.49

     Chi-Square Test
for Specified Proportions
-------------------------
Chi-Square        33.4334
DF                     11
Pr > ChiSq         0.0004

Effective Sample Size = 357
Frequency Missing = 2
```

(f) Table 1 gives percentages of women randomized by month and live births by month. The percentage of birth each month is much more even than that of the number of women randomized. (I said not to do a test, but if one does the same test as in (e) for the births it yields $p = 0.5$.) That could be because the number of births in a month is not influenced substantially by the number of holidays in the month. (It is probably also because I had an error when I created the dataset – I used the month from each woman's date of birth rather than from her infant's date of birth.)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Randomized	6.2	4.8	11.8	6.7	11.5	9.8	11.2	6.2	9.0	10.9	8.1	3.9
Births	6.1	6.4	8.6	7.5	7.5	9.7	9.5	9.7	8.6	11.4	7.2	7.5

Table 1: Percentages of women randomized each month and births each month



(g) We use a χ^2 test of trend. Let ρ_i denote the probability of preterm delivery in previous pregnancy category i. We want to test $H_0 : \rho_0 = \rho_1 = \rho_2 = \rho_{3+}$ against $H_A : \rho_0 \leq \rho_1 \leq \rho_2 \leq \rho_{3+}$ or $\rho_0 \geq \rho_1 \geq \rho_2 \geq \rho_{3+}$ with at least one of the inequalities being strict. From the SAS output, $X^2_{\text{trend}} = (-0.0335)^2 = 0.0011$ with $p = 0.9733$. So we do not reject H_0 and conclude there isn't much evidence for a monotonic trend in the risk of preterm delivery with the number of previous pregnancies.

```
Table of preterm by ppnum

preterm     ppnum

Frequency|
Col Pct  | 0      | 1      | 2      |3+      |  Total
---------+--------+--------+--------+--------+
       0 |     57 |    117 |     72 |     28 |    274
         |  85.07 |  92.13 |  88.89 |  84.85 |
---------+--------+--------+--------+--------+
       1 |     10 |     10 |      9 |      5 |     34
         |  14.93 |   7.87 |  11.11 |  15.15 |
---------+--------+--------+--------+--------+
Total          67      127       81       33      308

Frequency Missing = 51

Cochran-Armitage Trend Test
---------------------------
Statistic (Z)       -0.0335
One-sided Pr <  Z    0.4867
Two-sided Pr > |Z|   0.9733
```

(h) Now we want to test whether there is an association between treatment group and preterm delivery. We use a χ^2 test of association or, equivalently, test whether the proportion of preterm deliveries is the same in the two treatment groups. We want to test H_0 : preterm delivery is independent of treatment group, versus H_A : preterm delivery is associated with treatment group. The χ^2 test yields $p = 0.14$, hence we do not reject the null hypothesis. From the SAS output we see that the preterm percentages in the two groups are 13.5 and 8.6, that is, the percentage is actually nominally higher in the prenatal treatment group. So there is no evidence that treatment of periodontal disease reduces the risk of preterm delivery.

```
group     preterm

Frequency|
Row Pct  |       0|       1|  Total
---------+--------+--------+
       1 |    148 |     23 |    171
         |  86.55 |  13.45 |
---------+--------+--------+
       2 |    171 |     16 |    187
         |  91.44 |   8.56 |
---------+--------+--------+
Total         319       39      358

Statistics for Table of group by preterm

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      2.2040    0.1376
```

(i) We have pocket depth measurements on each woman at two time points (apart from some missing values, which we exclude from this analysis). Because the two measurements on each woman are not independent, we cannot use a two-sample test. Instead, we calculate the difference and test whether the mean of the differences is zero, that is, $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} \neq 0$. Taking the difference as the pocket depth after delivery minus the pocket depth at baseline, the difference will be positive if pocket depth has increased (that is, periodontal disease has progressed) and negative if it has declined (that is, if there has been an improvement).

Even after omitting missing values, the sample size is large ($n = 286$), so we can rely on the CLT and Slutsky's Theorem to give a test using the Z statistic.

The critical region is $C_{0.05} = \{Z : |Z| > 1.96\}$.

Using the SAS output below, the test statistic is $z = (-0.0835 - 0)/\sqrt{0.2173/286} = -3.03 > 1.96$. The associated p-value is 0.002. (Using a one-sample t-test gives a very similar p-value – see the SAS output.)

So we reject H_0 and conclude that the mean average pocket depth changed from baseline to delivery. The estimate of mean change is -0.08mm. That is average pocket depth became slightly worse, even though about half of the participants had their periodontal disease treated in the interim.

```
N                          286    Sum Weights                286
Mean                -0.0835129    Sum Observations    -23.884703
Std Deviation       0.46619515    Variance            0.21733791

           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t  -3.02949    Pr > |t|    0.0027
```

(j) Now we do have two independent sets of measurements – the data from the two treatment groups (with the data within each group being the difference between the pocket depth measurements at the two time points, as in part (e)). The sample sizes in the two groups are still large ($n_1 = 61$ in the prenatal group and $n_2 = 62$ in the post-partum group), so we can again rely on the CLT and Slutsky's Theorem to give a test using the Z statistic. Here $H_0 : \mu_{\text{diff},1} = \mu_{\text{diff},2}$ versus $H_A : \mu_{\text{diff},1} \neq \mu_{\text{diff},2}$.

The critical region is again $C_{0.05} = \{Z : |Z| > 1.96\}$.

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(0.0130 - (-0.1662)) - 0}{\sqrt{\frac{0.4910^2}{132} + \frac{0.4284^2}{154}}} = 3.26$$

As $3.26 > 1.96$ we reject the null hypothesis and conclude there is a difference between the change in pocket depth in the prenatal treatment group compared with the post-partum treatment group. The associated p-value is 0.0011. Using a t-test yields similar results

```
The TTEST Procedure

Variable:  pd_chng

group            N        Mean     Std Dev

1              132      0.0130      0.4910
2              154     -0.1662      0.4284
Diff (1-2)              0.1792      0.4583


group         Method               Mean       95% CL Mean

1                                0.0130     -0.0716   0.0975
2                               -0.1662     -0.2344  -0.0980
Diff (1-2)    Pooled             0.1792      0.0721   0.2862
Diff (1-2)    Satterthwaite      0.1792      0.0710   0.2873

Method           Variances        DF    t Value    Pr > |t|

Pooled           Equal           284       3.30      0.0011
Satterthwaite    Unequal      262.16       3.26      0.0013
```

(k) In part (i) we saw that overall there is a very marginal improvement in the mean pocket depth. From part (j) we see that the change in pocket depth differs significantly between the two treatment groups. Looking at the point estimates and confidence intervals for the change within each group (in the SAS output in part (j), we see that pocket depth tended to get worse in the post-partum treatment group. For the prenatal treatment group the point estimate shows a small improvement in mean pocket depth, but the associated 95% confidence interval includes 0, so there is not a statistically significant (or clinically meaningful improvement. So, even if effective periodontal therapy does have the potential to reduce the risk of premature birth or low birthweight, the periodontal therapy given in this study does not appear to have had much effect on periodontal disease, but is better than leaving the periodontal disease untreated and so may have an effect on birth outcomes.

Question 3

(a) Below is a 2 × 2 table of exposure by case status along with the column percentages, that is the percentage in each exposure category among cases and among controls.

	Control	Case	Total
Exposed = 0	69	51	120
	79.3	58.6	
Exposed = 1	18	36	54
	20.7	41.4	
Total	87	87	174

Let π_1 be the probability of being exposed in the control group and π_2 the corresponding probability in the case group. We want to test $H_0 : \pi_1 = \pi_2$ versus $H_A : \pi_1 \neq \pi_2$. The sample size is large enough to use a χ^2 test of association. The critical region is $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,0.95} = 3.84\}$. From the SAS output below we see that $X^2 = 8.7 > 3.84$ and the corresponding p-value is 0.003. Thus we reject the null hypothesis and conclude that moderate to severe periodontal disease is associated with premature birth.

```
Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      8.7000    0.0032
```

(b) Because this is a case-control study, the appropriate measure of association is the odds ratio. From the 2 × 2 table we obtain

$$\widehat{\mathrm{OR}} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{69 \times 36}{51 \times 18} = 2.71.$$

This is confirmed by the SA output below, which gives the 95% CI as $(1.38, 5.30)$.

```
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                   Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel         2.7059        1.3823        5.2967
  (Odds Ratio)    Logit                   2.7059        1.3823        5.2967
```

(c) Using the Mantel-Haenszel method, from the SAS code and output below we obtain an estimated odds ratio (adjusted for age) of 3.05, with 95% CI $(1.48, 6.28)$.

```
proc freq;
   table age_group*exposed*case / norow nopercent cmh;

           Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                   Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel         3.0506       1.4819       6.2802
  (Odds Ratio)
```

(d) The adjusted odds ratio of 3.05 is reasonably similar to the unadjusted one of 2.71, so age group does not appear to be a substantial confounder of the association between periodontal disease and premature birth. Investigating whether age group is associated with both the exposure and the outcome will show that numbers of cases and controls are equal within each age group (because of the matching on age), so age and case status are not associated in this dataset. From separate 2×2 tables for the three age groups, we obtain estimated odds ratios of 2.23, 4.86 and 9.63, respectively. This suggests that the odds ratios are not homogeneous across the age groups and so it may not be appropriate to pool them using the Mantel-Haenszel estimator. (We have not covered a test of homogeneity of the odds ratios and I did not expect you to look for such a test. The Breslow-Day test for homogeneity does not reject the null hypothesis of homogeneity. This test is part of the output from the SAS code above.)

```
     Breslow-Day Test for
Homogeneity of the Odds Ratios
------------------------------
Chi-Square              1.8405
DF                           2
Pr > ChiSq              0.3984
```

(e) For this part we need the 2×2 table in a different form. To obtain this in SAS, we need each pair to be a single observation in the SAS dataset. One way to do this is to rename the exposure variables so that those for cases and controls are distinct, split the dataset into two, one consisting of cases, the other of controls, and then merge the two on the part of the ID that is common to the members of a pair. This yields the following table.

		Controls		
		$E = 0$	$E = 1$	Total
Cases	$E = 0$	44	7	51
	$E = 1$	25	11	36
		69	18	87

We want to test $H_0 : \pi_1 = \pi_2$ versus $H_A : \pi_1 \neq \pi_2$. We do so using McNemar's test statistic. Here $n_{12} + n_{21} = 7 + 25 = 32 > 30$, so we can use the χ^2 approximation. The critical region is $C_\alpha = \{\mathrm{M} : \mathrm{M} > \chi^2_{1,0.95}\} = \{\mathrm{M} : \mathrm{M} > 3.84\}$.

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(7 - 25)^2}{7 + 25} = 10.1 > 3.84.$$

So we reject the null hypothesis and conclude that moderate to severe periodontal disease is associated with premature birth. From the SAS output we see that the associated p-value is 0.0015.

```
Statistic (S)         10.1250
DF                          1
Asymptotic Pr >  S     0.0015
Exact      Pr >= S     0.0021
```

To estimate the odds ratio, whether one uses n_{12}/n_{21} or n_{21}/n_{12} depends on how the table is set up. Here $\widehat{\mathrm{OR}}_{\mathrm{M}} = n_{21}/n_{12} = 25/7 = 3.57$.

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})) \approx \frac{1}{n_{12}} + \frac{1}{n_{21}} = \frac{1}{25} + \frac{1}{7} = 0.183$$

So an approximate 95% CI on the log scale is $\log(3.57) \pm 1.96 \cdot \sqrt{0.183}$, that is $(0.435, 2.111)$. On the original scale this becomes $(e^{0.435}, e^{2.111}) = (1.54, 8.23)$.

(f) The estimate in (d) is most appropriate because it takes into account the matched case-control design. Although part (c) takes into account age group, it assumes frequency matching rather than individual matching and doesn't take into account the second matching factor (number of previous pregnancies).

(g) The estimated odds ratio of 3.57 in part (d) is fairly large, the lower end of the 95% confidence interval is well above 1 and the p-value from McNemar's test is substantially below 0.05, so the evidence from the matched case control study is strong. Evidence from a case-control study is relatively weak, for several reasons including concerns about the representativeness of the controls, potential differences in recall of exposure by cases versus controls, assessment of timing of exposure relative to becoming a case and, as in any observational study, the possibility that there may be unmeasured confounders. In this particular study it should be relatively easy to recruit appropriate controls (women giving birth in the same facility as the cases), exposure recall is not an issue for the main exposure (measured periodontal disease) but the timing of the exposure measurement is (after giving birth, so it is not known when the periodontal disease developed). Also, there may have been unmeasured confounders, such as smoking.