PSY 711/12 Exam

1. One measure of the spread, or variability, in a data set is the **interquartile range**, or IQR, which is the difference between the scores corresponding to the 75-th and 25-th percentiles. Imagine you had a data set containing 100 unique scores (i.e., no duplicate scores).

   (a) Describe a method for computing the *standard error* of the IQR.

   Answer: I do not know of any formula for computing the standard error of the IQR, therefore I would use a percentile bootstrap method. I would resample my original data – randomly, with replacement – to create a bootstrapped sample. Next, I would compute the IQR on the bootstrapped sample. Finally, I would repeat this procedure R times. The distribution of the bootstrapped IQRs is an estimate of the sampling distribution of IQR. The *standard deviation* of the bootstrapped IQRs is an estimate of the standard error of the IQR. For generating estimates of the standard error, usually setting $100 < R < 1000$ gives reasonably good results.

   (b) Describe a method for computing the 99% confidence interval for the IQR.

   Answer: In the previous question, let us suppose that $R = 1000$ and that we order the bootstrapped IQRs from smallest to largest. The 5th and 995th values in the ordered array would be the percentile-bootstrapped estimate of the 99% confidence interval. In other words, the IQR values corresponding to the 0.5 and 99.5 percentiles of the bootstrapped IQRs. This method of estimating the confidence interval has the advantage of being easy to describe and compute; it also does not assume that the sampling distribution is normal. However, the percentile bootstrap also has some disadvantages. Specifically, percentile bootstrapped confidence intervals can be biased, and they sometimes have poor coverage (i.e., can be too narrow or too wide). Other bootstrap methods (e.g., BCa) often provide better estimates of confidence intervals, although Wilcox recommends the percentile bootstrap for estimating confidence intervals for many robust statistics.

   (c) Describe a method of testing the null hypothesis that population IQR = 15, assuming that $\alpha = .01$.

   Answer: Because $\alpha = .01$, we can evaluate the null hypothesis that IQR = 15 by simply noting if the value of 15 falls within the 99% confidence interval estimated by the method described in the previous answer. If 15 falls within the confidence interval, then we *fail to reject* the null hypothesis that IQR=15; if 15 is outside the confidence interval, then we reject the null hypothesis in favor of the alternative (i.e., IQR $\neq$ 15).

2. `hills.time` is a numeric vector that contains the record times (in 1984) for 35 Scottish hill races. The total height (in feet) gained during each race is contained in the numeric vector `hills.climb`.

   (a) Calculate the classical (i.e., standard) 95% confidence interval of the mean record time.

   Answer: The R function `t.test` can be used to calculate confidence intervals. The following code shows that the 95% confidence interval of the mean is (40.68, 75.06):

```
> t.test(hills.time)

        One Sample t-test

data:  hills.time
t = 6.8424, df = 34, p-value = 7.09e-08
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 40.68613 75.06530
sample estimates:
mean of x
 57.87571
```

(b) Estimate the 95% confidence interval of the mean using the percentile bootstrap. Which confidence interval do you think is more accurate? Why?

Answer: I will use Wilcox's function trimpb while setting the amount of trimming to zero. The result will be a percentile bootstrap estimate of the 95% confidence for the (untrimmed) mean. The result is (42.5, 75.9), which is slightly narrower than, but remarkably similar to, the confidence interval calculated with the standard equations. I suspect that the bootstrapped confidence interval is more accurate, simply because the hills data are strongly positively skewed: the skew is obvious when the data are plotted with the `boxplot(hills.time)` command). This deviation from normality means that the assumptions underlying the standard/classical confidence interval calculation are not valid, and therefore the confidence interval is likely to be biased. The bootstrap estimate of the confidence interval does not rest on the normality assumption, and therefore should be more accurate than the standard method.

```
> trimpb(hills.time,tr=0,alpha=.05)$ci
[1] "The p-value returned by the this function is based on the"
[1] "null value specified by the argument null.value, which defaults to 0"
[1] "Taking bootstrap samples. Please wait."
[1] 42.53431 75.91811
```

(c) Use a bootstrap procedure to calculate the 95% confidence interval for the correlation (Pearson's r) between `hills.climb` and `hills.time`.

Answer: The following command uses the percentile bootstrap to estimate the 95% confidence interval for Pearson's r, which turns out to be (0.474, 0.963):

```
> pcorb(hills.climb,hills.time)
$r
[1] 0.8052392

$ci
[1] 0.4741237 0.9636232
```

Note that R's built-in function `cor.test` does not use a bootstrap procedure: Instead, it relies on the assumption that the data are distributed as bivariate normal variables.

3. The variable `ediff` is a numeric vector that contains the difference in energy intake (kJoules) in twelve women measured before and after menstruation (i.e., ediff = postIntake - preIntake).

(a) Calculate the 95% percent confidence interval for the mean of `ediff`.

Answer: The standard/classical method of computing the 95% confidence interval of the mean yields (-1357.29, 9.117). A method based on the percentile bootstrap yields (-1323.18, -166.36):

```
> t.test(ediff)

        One Sample t-test

data:  ediff
t = -2.1984, df = 10, p-value = 0.05258
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1357.299367     9.117548
sample estimates:
mean of x
-674.0909
```

```
> trimpb(ediff,tr=0,alpha=.05)$ci
[1] "The p-value returned by the this function is based on the"
[1] "null value specified by the argument null.value, which defaults to 0"
[1] "Taking bootstrap samples. Please wait."
[1] -1323.1818  -166.3636
```

(b) Evaluate the null hypothesis that some measure of central tendency – the mean, x% trimmed mean, median, M-estimator, etc. – is zero. Make sure your answer identifies the particular measure of central tendency that you are using, as well as your conclusions regarding the null hypothesis.

Answer: There are many ways you could answer this question. Here are a few:

```
> momci(ediff)
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1277.27273   -21.42857
> onesampb(x=ediff,est=mom,alpha=.05)
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1277.27273   -21.42857
> onesampb(x=ediff,est=onestep,alpha=.05)
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1277.27273   -58.97728
> onesampb(x=ediff,est=mean,alpha=.05,tr=0.2)
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1257.14286   -57.14286
> onesampb(x=ediff,est=median,alpha=.05)
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1200   -30
> trimpb(ediff,tr=0.2,alpha=.05)
[1] "The p-value returned by the this function is based on the"
[1] "null value specified by the argument null.value, which defaults to 0"
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -1307.14286   -57.14286

$p.value
[1] 0.002
```

The first two analyses show different ways of using a percentile bootstrap to estimate the 95% confidence interval of the modified onestep M-estimator (MOM). The other onesampb commands compute the percentile bootstrap confidence interval for the onestep M-estimator, the mean, and the median. The command trimpb computes the percentile bootstrap estimate of the confidence interval for the 20% trimmed mean. In each case, the confidence interval does not contain zero, and therefore we reject the null hypothesis that the measure of central tendency – i.e., the population mom, onestep M-estimator, mean, median, and trimmed mean – is zero.

(c) Defend your selection of the measure of central tendency that was used to answer the previous question.

Answer: There are two outliers that make the data highly negatively skewed, which is why the mean (-674) differs significantly from the median (-240). These outliers inflate the variance and widen the confidence interval. Furthermore, the mean of -674 is a poor measure of a "typical" value: only 3 of the 11 numbers are less than the mean. Therefore, I would use a robust measure of central tendency like the median, mom, timmed mean, etc. M-estimators like are particularly good in cases (like the current one) where the data are skewed, but the various robust measures yield similar results: Specifically, the confidence intervals of the robust measures are much narrower than the confidence interval of the mean.

4. The variables `math.male` and `math.female` contain the data from a hypothetical study that measured "math aptitude" in 1000 boys and girls in grade eight public schools in Ontario.

(a) Calculate the 95% confidence interval of the difference between the means. Use a t test to evaluate the null hypothesis of no difference between boys and girls.

```
> t.test(math.male,math.female)

        Welch Two Sample t-test

data:  math.male and math.female
t = 2.962, df = 1697.909, p-value = 0.003099
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9904972 4.8735028
sample estimates:
mean of x mean of y
  103.494   100.562
```

Answer: The 95% confidence interval of the difference between population means is (0.99, 4.87). The t-test was conducted using the assumption that the group variances differed: The difference between means was significant, $t(1697.91) = 2.96, p = 0.003$, indicating that the mean score was higher in boys than girls.

(b) Describe why the preceding analysis may be misleading.

Answer: The t test assumes that the data are distributed normally in both groups. However, the scores in both groups are strongly positively skewed, and so the normality assumption almost certainly is not valid in this case. Therefore the p value may be misleading. Furthermore, the fact that the data are skewed raises the possibility that the means do not represent "typical" scores. This idea is supported by the observation that the medians of the two groups are *very* similar (100 for males, 101 for females).

(c) Conduct an alternative analysis to evaluate the null hypothesis of no difference between "typical" math aptitude scores in boys and girls. Justify your analysis.

Answer: Here are some alternative analyses that use robust measures. The first compares trimmed means using the assumption that the difference between trimmed means will follow a t distribution. The second analysis use a percentile bootstrap to estimate the 95% confidence interval of the difference between trimmed means. The third and fourth analyses use a percentile bootstrap to estimate the difference between group mom's and onestep M-estimators, respectively. Note that only the difference between mom's is significant, and that result suggests that the "typical" score is *lower* in boys. Taken together, these analyses suggest that the significant difference found between group means reflects the influence of the subset of boys and girls who score very highly.

```
> yuen(math.male,math.female, tr = 0.2, alpha = 0.05)
$ci
[1] -2.3778088  0.5111422

$p.value
[1] 0.2051543

$dif
[1] -0.9333333

$se
[1] 0.736246

$teststat
[1] 1.267692

$crit
[1] 1.961947

$df
[1] 1197.698
> yuenbt(math.male,math.female, tr = 0.2, alpha = 0.05)
[1] "NOTE: p-value computed only when side=T"
[1] "Taking bootstrap samples. Please wait."
$ci
```

```
[1] -2.2675163  0.5472417

$test.stat
[1] -1.268115

$p.value
[1] NA
```

```
> pb2gen(math.male,math.female,alpha=0.05,nboot = 1999, est = mom)
```

```
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -3.536773 -0.486741

$p.value
[1] 0.008004002

$sq.se
[1] 0.6012606
```

```
> pb2gen(math.male,math.female,alpha=0.05,nboot = 1999, est = onestep)
```

```
[1] "Taking bootstrap samples. Please wait."
$ci
[1] -2.136523  0.697032

$p.value
[1] 0.3231616

$sq.se
[1] 0.5179252
```

5. A study measured the bulk resistivity of 15 silicon wafers manufactured in four different factories. The nominal resistivity of the wafers was 200 ohm cm. The data are stored in the data frame `silicon` and in the list `sil.list`.

   (a) Evaluate the null hypothesis that mean resistivity does not differ across factories using an analysis that assumes that the data are drawn from normal distributions that may differ in variance.

```
> oneway.test(resist~factory,data=silicon)

        One-way analysis of means (not assuming equal variances)

data:  resist and factory
F = 2.2069, num df = 3.000, denom df = 26.569, p-value = 0.1107
```

   Answer: Our test was not significant, $F(3/26.57) = 2.21,\ p = 0.11$, and so we fail to reject the null hypothesis that the population means do not vary across factories.

   (b) Next, evaluate the null hypothesis that mean resistivity does not differ across factories using an analysis that does not assume normality.

   Answer: The following analysis repeats the F test, but uses the percentile bootstrap to compute the p value. In other words, it does not assume that the data are distributed normally. Note that the test evaluates *means*, not trimmed means. The test is not significant, so we fail to reject the null hypothesis of no difference among factories:

```
> t1waybt(sil.list, tr = 0, nboot = 1999)
```

```
[1] "Taking bootstrap samples. Please wait."
[1] "Working on group  1"
[1] "Working on group  2"
[1] "Working on group  3"
[1] "Working on group  4"
$test
[1] 2.206895
```

```
$p.value
[1] 0.1360680
```

The test statistic in the previous analysis was F:

$$F = \frac{\text{MS}_{\text{BG}}}{\text{MS}_{\text{WG}}} = \frac{n \sum_{j=1}^{J} (\bar{Y}_j - \bar{Y}_{..})^2}{\frac{1}{N-J} \sum_{j=1}^{J} \left(\sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2 \right)} \tag{1}$$

which is an index of the *normalized* difference among group means. The following analysis computes the test statistic H:

$$H = (1/N) \sum_{j=1}^{J} n_j (\hat{\theta}_j - \bar{\theta})^2 \tag{2}$$

which is similar to F except for the fact that the it is not normalized by a measure of within-group variance. In this case, the test *is* significant, so we reject the null hypothesis of no difference among factories:

```
> b1way(sil.list,mean)
[1] "Taking bootstrap samples. Please wait."
[1] "Working on group  1"
[1] "Working on group  2"
[1] "Working on group  3"
[1] "Working on group  4"
$teststat
[1] 2.773787

$p.value
[1] 0.03171953
```

Wilcox recommends `b1way` over `t1waybt`.

(c) Finally, evaluate the null hypothesis that a more robust measure of "typical" resistivity does not differ across groups. Justify your analysis.

Answer: I would use an M-estimator because the data are skewed and the number of outliers appear to vary across groups. M-estimators have an advantage over trimmed means in that they criterion used to identify outliers can vary across groups. [1] Similar results are obtained when we use the mom or onestep estimator: in both cases, we fail to reject the null hypothesis of no difference among M-estimators from different factories:

```
> b1way(sil.list,mom)
[1] "Taking bootstrap samples. Please wait."
[1] "Working on group  1"
[1] "Working on group  2"
[1] "Working on group  3"
[1] "Working on group  4"
$teststat
[1] 1.074162

$p.value
[1] 0.07679466
> b1way(sil.list,onestep)
[1] "Taking bootstrap samples. Please wait."
[1] "Working on group  1"
[1] "Working on group  2"
[1] "Working on group  3"
[1] "Working on group  4"
```

[1] One *disadvantage* of M-estimators is that they sometimes cannot be computed when bootstraps are performed on small samples (because a bootstrapped sample has zero variance, which means that outliers cannot be defined). In such cases it is better to use trimmed means.

```
$teststat
[1] 1.374412

$p.value
[1] 0.06844741
```

6. In a between-subjects experimental design with J conditions, $n \times J$ subjects are assigned randomly to each condition (with the constraint that each condition has n subjects). In a within-subjects design, each of the n subjects is tested in all conditions. Consider a situation where we wanted to calculate an F statistic to evaluate the null hypothesis of no difference among the J means:

   (a) Describe and contrast the bootstrap methods that are appropriate for evaluating the null hypothesis for data collected with between-subjects and within-subjects experimental designs.

   Here is an example of an A+ answer taken from one of the exams.

   Answer: Both independent and dependent sample ANOVAs begin by centering the data in order to create an instance where we know the null to be true. In both ANOVA bootstrap, we compute the (trimmed) mean for each of our J groups. Then, for each group, we subtract the trimmed mean from each each individual score, which produces J Groups that ll have trimmed means of zero, thereby allowing us to create a bootstrap estimate of the sampling distribution when the null is true. Where the two bootstrapping procedures differ is in WHAT the bootstrap randomly samples. In the independent sample ANOVA, we construct bootstrapped samples by sampling each J group randomly with replacement (i.e., resample n scores for each J group). We then calculate an F* value on each bootstrapped sample. For the dependent sample ANOVA, the scores in each J group are likely to be correlated across each subject. Therefore, the bootstrap procedure should keep this relationship. To do this, we generate bootstrap samples on each of the n subject rather than create bootstrapped samples in each of the J conditions. If there are 30 participants with J scores, the bootstrap would randomly select a subject, with replacement, and the j scores associated with each subject would stay together. So to summarize, the methods are similar b/c they both center the data so that it creates a a bootstrap estimate of the sampling distribution when the null is true (note that the observed F score from the original data set is then compared to that distribution and we reject the null if our observed F is more extreme than F*(1-alpha/2)) The two procedures are different in that the independent method randomly selects scores from J conditions (with replacement) while the RM method randomly selects n participants with replacement.

   (b) In the case of a within-subjects design, what advantages and/or disadvantages would a bootstrap approach have over standard within-subjects (i.e., repeated-measures) ANOVA?

   Answer: The advantages of a bootstrap approach to RM ANOVA is that we are not required to make any assumptions about the sampling distribution. The standard RM ANOVA has three primary assumptions that need to be met in order to maintain an adequate level of Type 1 error and power. It assumes that the sampling distributions are (1) normally distributed, (2) have equal variances, and (3) have a variance-covariance matrix that is spherical. If any of these assumptions are violated, our observed F will not be distributed as an F with the adjusted degrees of freedom. Therefore, the bootstrap procedure allows us to test for group differences even when these assumptions are violated. That is, the main advantage is that you do not have to make any assumptions associated with your sampling distribution. Rather, you use the sample to create an empirical, rather than theoretical, distribution of your statistic. A potential disadvantage to this procedure is that is not robust to outliers. Furthermore, small n's could results in a larger representation of nonrepresentative data points. Furthermore, bootstrapping assumes that your sample is an accurate representation of the population, which may not be true.

   (c) Would the bootstrap-F methods protect you from the problems associated with data that do not follow a normal distribution? Explain.

   Answer: I guess it depends on the problems associated with your original data set. As stated earlier, bootstrapping does not deal with outliers. Therefore, if your original sample is plagued with abnormal data points, taking a bootstrap of the mean would not be beneficial. However, outliers can be dealt with using bootstrap when your sample statistic is a robust measure of central tendency. Note that a common assumption is not only that the statistic of interest be distributed in a particular fashion, but also that the observed F follow the F distribution. While certain robust measure (i.e., trimmed means, M-estimators) can make the data more normal, deviations from normality will have an influence on whether the observed F is distributed as an F with an adjusted df.