STATISTICAL METHODS II
IN-CLASS MIDTERM EXAMINATION I SOLUTIONS

This in-class examination covers the most important things we covered during the first third of the course. It is worth 15 points; the take-home is worth 85 points. Please make sure your answers are correct and make sense.

Good luck!

$\star\star\star$

Section: Distributions

For this section, let us assume

$$W \sim \mathcal{N}(\mu_W, \sigma^2 = 3)$$
$$X \sim \mathcal{N}(\mu = 5, \sigma^2 = 3)$$
$$Y \sim \mathcal{N}(\mu = 5, \sigma^2 = 1)$$
$$Z \sim \mathcal{N}(\mu = 4, \sigma^2 = 0.001)$$

and that I have a sample from each of the four distributions.

⋆⋆⋆

⟦1⟧1. I have a sample from W and a sample from X. Which test should I use to determine if the means of W and X are statistically different?

Solution: As we know that both random variables are Normally distributed, and since we know the variance of each random variable, we can use the z-test.

Had we not known the variance, we would use the sample to estimate it and then use a t-test. The t-test assumes the random variables are Normally distributed, but estimates the population variance from the sample. The z-test does not need to estimate the population variance, since it is known. ◇

⋆⋆⋆

⟦1⟧2. What is the distribution of $W - X$? Make sure you include the name, the expected value, and the variance.

Solution: We know that the sum (difference) of Normally distributed random variables is also Normally distributed. We also know that the mean of the difference is the difference of the means. Finally, as these two samples are independent, the variance of the difference is the sum of the variances (not the difference).

Thus,

$$W - X \sim \mathcal{N}(\mu = \mu_W - 5, \sigma^2 = 6)$$

◇

⋆⋆⋆

⟦1⟧3. I have a sample from one of the above distributions. Will it be easier to determine that the sample is not Y or that the sample is not X? Explain in a sentence.

Solution: The only difference between X and Y is that the variance of Y is smaller than that of X. As such, a sample from Y should be closer to $\mu = 5$ than a sample from X; X is more spread out. Thus, it will be easier to resolve the difference between our sample and Y than between our sample and X. ◇

⋆⋆⋆

Section: Tests

⟦3⟧4. I have a sample of GPA measurements taken from two groups, males and females. Which test should I like to use? Why? What are its three assumptions? How do we test the two assumptions *other than* independence (name the test, not the `R` function)?

Solution: I would like to use a parametric test like the t-test as parametric tests are uniformly more powerful than non-parametric tests (assuming the assumptions are met).

The three assumptions of the t-test are that the measures are independent, the measures have equal variances across the groups (which can be relaxed), and the measures are all Normally distributed within each group.

To test the Normality assumption, one would use a Shapiro-Wilk test (or an Anderson-Darling test or a Kolmogorov-Smirnov test or a slew of others). To test equality of variances, one would use Fisher's F test (or the Bartlett test or the Fligner test or a slew of others). ◇

⋆⋆⋆

⟦5⟧5. I have a sample of runs scored for major league baseball teams. My grouping variable is the division, so there are 6 groups (three divisions in two leagues). I would like to use an analysis of variance procedure to determine if there is any statistical difference between the number of home runs scored by division. Besides independence, what are the assumptions (is the assumption) of the analysis of variance procedure? What tests would I need to perform (name, not `R` code) to check the assumption(s)? What is the null hypothesis of the test(s)? Knowing that the number of runs scored is heavily right skewed (and, thus, the analysis of variance procedure would not be appropriate), what do I do next?

Solution: The analysis of variance procedure assumes that the population measurements in the groups are independent, that the population variances of the measurements across groups is equal, and that the population measurements within each group is Normally distributed.

I would use the Bartlett test or the Fligner test to test the equivariance assumption. I would use the Shapiro-Wilk test to test the Normality assumption.

In the equivariance tests, the null hypothesis is that the variances are equal (to a common variance). In the Normality tests, the null hypothesis is that the population measurements within the groups is Normally distributed.

Because the analysis of variance procedure is so powerful (compared to the non-parametric alternatives), I really would like to use the analysis of variance procedures. To do this, I must transform the data using a one-to-one and onto function (the transformation and its inverse are both functions over the range of the data). If I can find such a transformation that transforms the data into a set that meets the assumptions of analysis of variance, I can use that procedure.

If not, I will have to use the Kruskal-Wallis test, which is a non-parametric version of the analysis of variance procedure. ◇

⋆⋆⋆

R FUNCTIONS

⟦2⟧6. I have two variables, `height` is a measurement, `male` is a grouping variable. I want to create a basic boxplot. What is the command I would run in `R`? Make sure you write a command that would actually work.

Solution:

```
barplot(height ~ male)
```

You would need a `data=` parameter if the data was attached to a dataset. ◇

⋆⋆⋆

⟦1⟧7. I have two variables, `height` is a measurement, `eyeColor` is a variable that groups people into 5 groups. I need to use a non-parametric test to determine if the groups have the same median. What is the command I would run in `R`? Make sure you write a command that would actually work.

Solution:

```
kruskal.test(height ~ eyeColor)
```

You would need a `data=` parameter if the data was attached to a dataset. ◇

⋆⋆⋆

⟦1⟧8. Finally, With the same data in 7, I need to determine which groups are different from the others. What is the command I would run in R? Make sure you write a command that would actually work.

Solution:

```
kruskal(height, eyeColor)
```

There is no `data=` parameter available. Thus, you will need to wrap this in a `with( data=x,···)` function if the data was attached to a dataset. Also, I would need to load the `agricolae` library before this command would work.

```
with( data=x, kruskal(height, eyeColor) )
```

◇

The R functions coresponding to the tests mentioned above are

Test	Function	Package
Shapiro-Wilk test	shapiro.test()	stats
Anderson-Darling test	ad.test()	nortest
Kolmogorov-Smirnov test	ks.test()	stats
Fisher's F test	var.test()	stats
Bartlett test	bartlett.test()	stats
Fligner test	fligner.test()	stats

This table may be of help. Note that the `nortest` package must be installed from the web as it does not come with the standard distribution. Also note that `nortest` is a library of Normality tests.