EBOLA BREAKOUT IN WEST AFRICA

In Question 1, we have been given some data concerning the Ebola breakout in West Africa. The data includes the number of people who got infected as well as the number of deaths for Guinea, Liberia and Sierra Leone. We have been tasked with comparing the fatality rates of Ebola from the three countries.

a) In part A, we have been asked to find the 95% CI for the difference in fatality rates between Guinea and Liberia. As we want to compare the difference, we have to compute the 95% CI using the following formula:

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\frac{1}{2}a}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

Using the data in the question, we can compute the 95% CI, mathematically by doing the following:

$$\left(\frac{1327}{2164} - \frac{3145}{7635}\right) \pm 1.96\sqrt{\frac{\frac{1327}{2164}(1-\frac{1327}{2164})}{2164} + \frac{\frac{3145}{7635}(1-\frac{3145}{7635})}{7635}}$$

Which equates to 0.212974712 $\pm$ 0.1125421411, giving us a 95% CI of (0.1004, 0.3255). Using SAS, we can compute the 95% CI of the difference by using the FREQ procedure using the key word RISKDIFF. Looking at the results:

Column 1 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.6132	0.0105	0.5927	0.6337	0.5923	0.6338
Row 2	0.4119	0.0056	0.4009	0.4230	0.4008	0.4231
Total	0.4564	0.0050	0.4465	0.4662	0.4465	0.4663
Difference	0.2013	0.0119	0.1780	0.2246		
Difference is (Row 1 - Row 2)						

Table 1 – CI for Difference in Proportions for Guinea and Liberia

We can see the 95% CI of the difference in populations is 0.178, 0.2246 which are similar to the mathematically computed CI.

b) Looking at the results produced for the 95% CI of the difference in proportions, we can see that both end-points of the interval are positive; mathematically we got (0.1004, 0.3255) and using SAS's RISKDIFF we got (0.1780, 0.2246). This suggests strong evidence that Guinea had a significantly higher proportion of deaths from Ebola compared to Liberia.

c) In the second part of the question, we have been asked to run an appropriate test in order to see whether the fatality rates in Guinea and Sierra Leone are similar. I will employ the Z-Test to compare the two proportions. First, we have to state the null and alternative hypothesis:

$$H_0 : \hat{\pi}_1 = \hat{\pi}_2$$

$$H_1 : \hat{\pi}_1 \neq \hat{\pi}_2$$

In English, this means that the null hypothesis states that the proportion of deaths from Ebola are similar in Guinea and Sierra Leone. Meaning that the alternative hypothesis is that the proportions aren't the same. We can then compute the test statistic by using the following formula:

$$Z = \frac{|\,\hat{\pi}_1 - \hat{\pi}_2|}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

If we put the problem specific numbers into the formula we get:

$$Z = \frac{|\,\frac{1327}{2164} - \frac{1583}{7312}|}{0.30709(1-0.30709)(\frac{1}{2164}+\frac{1}{7312})} = 3113.243412$$

Since the test statistic is greater than the critical value, 3113.24 > 1.96, we reject H_0. There is sufficient evidence in the data to refuse the assertion that Guinea had a similar fatality rate to Sierra Leone during the Ebola outbreak.

To further strengthen our argument, we can employ SAS to compute the P-value. This is done by using the FREQ procedure using the keyword CHISQ to generate the p-value. Looking at the results we can see:

Statistic	DF	Value	Prob
Chi-Square	1	1235.0912	<.0001
Likelihood Ratio Chi-Square	1	1160.6242	<.0001
Continuity Adj. Chi-Square	1	1233.2274	<.0001
Mantel-Haenszel Chi-Square	1	1234.9608	<.0001
Phi Coefficient		0.3610	
Contingency Coefficient		0.3396	
Cramer's V		0.3610	

Table 2 – P-Value for the test of Two Proportions

Looking at the p-value generated, we can see it is very small, giving us concrete evidence in the data to reject the null hypothesis, therefore confirming the alternative hypothesis that the two fatality rates aren't similar.

CODE

```
/* Data for the CI Question */
DATA EbolaWstAfr;
INPUT Country $ Death $ COUNT;
DATALINES;
      Guinea YES 1327
      Guinea NO 837
      Liberia YES 3145
      Liberia NO 4490
      ;
RUN;
/* Use the FREQ procedure with RISKDIFF KEYWORK
      to get the 95% CI of the difference*/
PROC FREQ DATA=EbolaWstAfr ORDER=DATA;
      WEIGHT COUNT;
      TABLES Country * Death / RISKDIFF;
RUN;
/* Data for the Z-Test Question */
DATA EbolaWstAfr2;
INPUT Country $ Death $ COUNT;
DATALINES;
      Guinea YES 1327
      Guinea NO 837
      Sierra YES 1583
      Sierra NO 5729
      ;
RUN;
/* Use the FREQ procedure with CHISQ KEYWORK
      to get the 95% CI of the difference*/
PROC FREQ DATA=EbolaWstAfr2 ORDER=DATA;
      WEIGHT COUNT;
      TABLES Country * Death / CHISQ;
RUN;
```

FERTILISER

In question 2, we have been given the yields of two different brands of fertiliser used on each half of twelve plots. We have been tasked with finding out which brand of fertiliser produces a higher yield.

a) In order to test which brand of fertiliser produces the largest yield, we have to choose an appropriate test. Due to the nature of the data being in pairs, meaning that each plot has two values, one for each fertiliser. We can conclude that the appropriate test to run on the data is a paired T-Test.

b) In order to conduct a Paired T-Test, we have to first create a null and alternative hypothesis:

$$H_0 : \mu_{diff} = 0$$

$$H_1 : \mu_{diff} \neq 0$$

where $\mu_{diff} = \mu_{FertA} - \mu_{FertB}$

Translated into English, the null hypothesis is that there is no difference between the average yield of fertiliser A compare to fertiliser B. The Alternative Hypothesis states that there is a difference between the averages of the two fertilisers.

c) The first step in running a paired t-test is to compute the test statistic, this is done by using the following formula:

$$T = \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}}$$

If we put the problem specific numbers into the formula we get:

$$T = \frac{-4.1667 - 0}{10.9697/\sqrt{12}} = 1.315794616$$

Now that we have the test statistic, we can use the t-Distribution table to find the critical value. We first have to compute the Degrees of Freedom, which is n-1 = 12-1 = 11. With a significance level of 5%, we can conclude that the critical value is 2.2010.

Now, using SAS, we can compute the p-value from the paired t-test. Using the t-test procedure, using the keyword paired followed by the two variable that are being tested we get these results:

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-4.1667	-11.1365	2.8031	10.9697	7.7709	18.6251

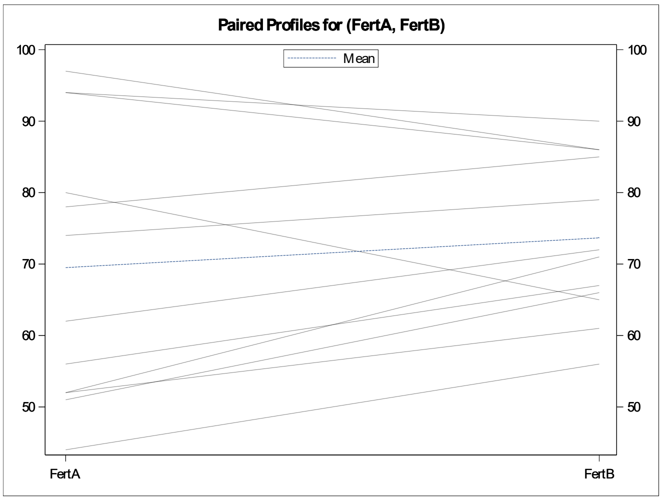
DF	t Value	Pr > |t|
11	-1.32	0.2150

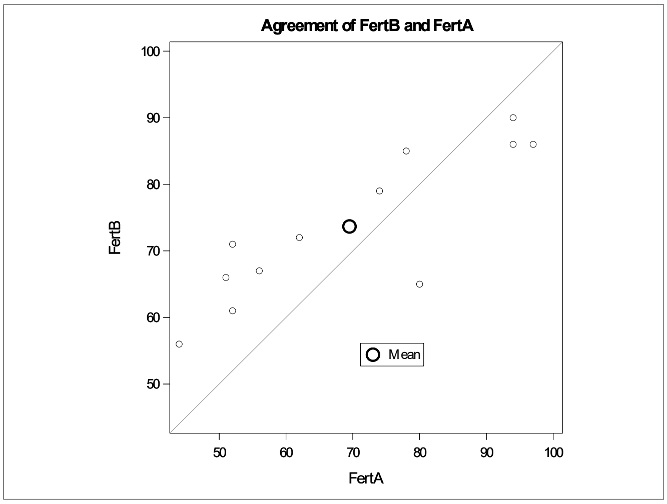
We can see from the results that the test statistic which we computed was the same as the one SAS has produced. Furthermore, we can see that the p-value is 0.2150.

From the various calculations we have produced, we can conclude that on average the fertilisers do produce the same yield. Firstly, this is due to the fact that the test statistic, 1.315794616 does not exceed the critical value of 2.2010. Meaning that we do not reject the null hypothesis, indicating that there is insufficient evidence to suggest that the average yield of plots differ using either fertiliser A or B.

If we look at the p-value computed by SAS, we can see that it isn't less than 0.05, as 0.215> 0.05, meaning that we cannot reject the null hypothesis at the 5% significance level.

d) Looking at some more results of the paired t-test, we can gain more insight to which fertiliser is better to use. The calculations above suggest that on average there is no difference between the two fertilisers but if we look at the paired profiles and agreement plots:





We can see that Fertiliser B, does seem to preform better. Looking at the paired profiles plot, we can see that FertB's mean is slightly better than FertA's. We can also see that FertB has more predictable results, with a range from 55-90 whereas FertA ranges from ~45 – 98. Looking at the Agreement of the two, we can see more data points in the FertB portion of the plot, with the mean also residing in the FertB portion. I therefore suggest that Fertiliser B would be the best choice.

CODE

```
/*Create Data Set for Analysis*/
data fert;
input Plot$ FertA FertB;
    datalines;
    1 56 67
    2 62 72
    3 74 79
    4 94 86
    5 52 71
    6 94 90
    7 97 86
    8 80 65
    9 78 85
    10 44 56
    11 52 61
    12 51 66
    ;
run;
/*Run Paired T Test with H0 = 0*/
proc ttest data=fert H0=0;
    paired FertA*FertB;
run;
```

OZONE

In question 3, we have been given a csv file with gain/loss in weight for two groups of rats. The experiment is trying to measure the effect the ozone. The scientist kept a group of 23 rats in an ozone free environment, which are the control group and another group which were exposed to the ozone.

In order to analyse the data to draw any conclusions, we have to first choose the appropriate test. As the population variance are unknown but known to be equal, we will use the two sample t-test or the unpaired t-test to evaluate the two groups.

Firstly, we shall state the null and alternative hypothesis for the test:

$$H_0 : \mu_{control} = \mu_{ozone}$$

$$H_1 : \mu_{control} \neq \mu_{ozone}$$

Which means that the null hypothesis states the average weight loss/gain is the same for both groups, whilst the alternative states that there is a difference.

The next step is to compute the test statistic, but for a two sample t-test we first have to compute the pooled estimate of the variance using the following formula:

$$S^2 = \frac{(n_1 - 1)S_x^2 + \ (n_2 - 1)S_y^2}{n_1 + n_2 - 2}$$

If we put the problem specific numbers into the formula we get:

$$S^2 = \frac{(23-1)(10.7768)^2 + (22-1)(19.0171)^2}{23+22-2} = 236.03998$$

Which we then use in to compute the test statistic, which is done using the following formula:

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})}}$$

If we put the problem specific numbers into the formula we get:

$$T = \frac{22.4261 - 11.0091}{15.36359268\sqrt{(\frac{1}{23} + \frac{1}{22})}} = 2.491886274$$

Which if we look up the critical value using the Degree of Freedom of v = $n_1 + n_2 - 2$, which is 43, we get 2.016. Comparing the test statistic and critical value we can see that 2.4918863 > 2.016, so we reject the null hypothesis. We can therefore conclude that there is evidence to suggest that the ozone does affect the weight of a rat.

Now, using SAS, we can look at the p-value:

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	43	2.49	0.0166
Satterthwaite	Unequal	32.918	2.46	0.0192

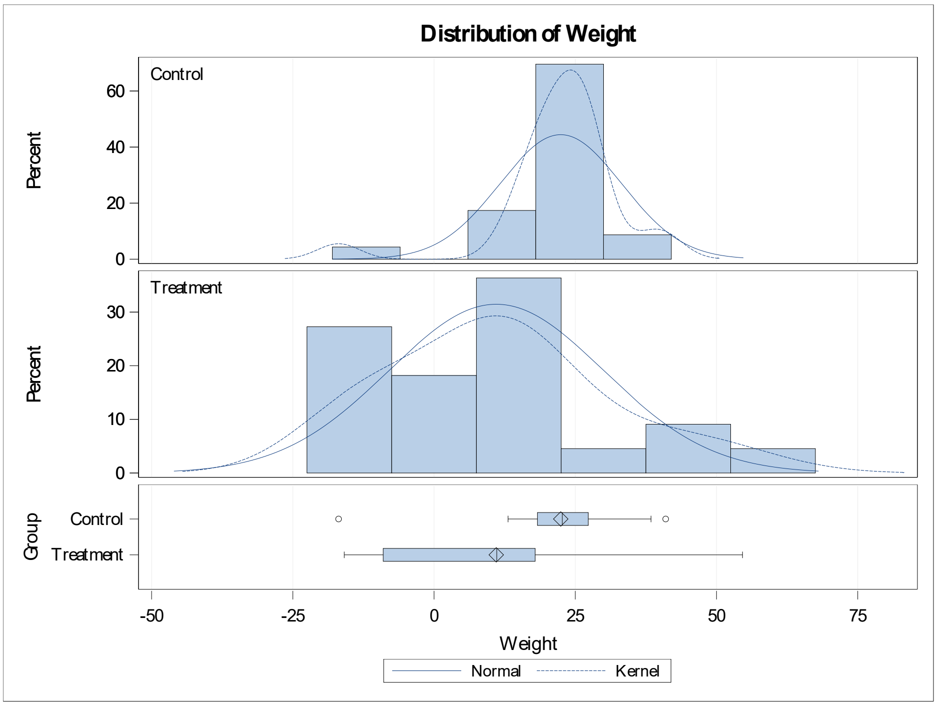
We can see that the pooled p value is 0.0166 which is less than 0.05, the significance level. We can therefore reject the null hypothesis. This means, that we have to accept the alternative hypothesis which states that the average gain/loss of the control group is not equal to that of the test group.

Looking at the descriptive statistics of the test:

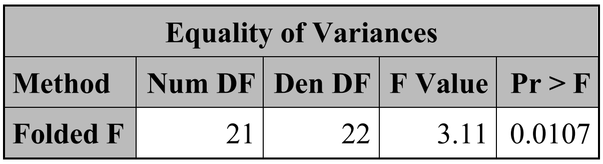
Group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Control		22.4261	17.7659	27.0863	10.7768	8.3347	15.2529
Treatment		11.0091	2.5774	19.4408	19.0171	14.6308	27.1767
Diff (1-2)	**Pooled**	11.4170	2.1772	20.6568	15.3636	12.6937	19.4660
Diff (1-2)	**Satterthwaite**	11.4170	1.9850	20.8489			

We can see that the difference in means from the two groups is 11.417, with the difference in 95% CI ranging from 2.1772 – 20.656.

Looking at the box plots of the two groups, we can see that the control group has a much smaller distribution than the treatment group. We can see that the Inter-quartile range of the treatment group is almost three times larger than the control group. The normal distributions of both groups are also interesting, we can see that the control group has thin tails and a high peak, whereas the treatment group has fat tails and a smaller peak.



Finally, if we look at the equality of variance, we can see:



Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	21	22	3.11	0.0107

That the assumption that the variance being equal is wrong, as the p value is less than 0.05 which does support the rejection of the null hypothesis.

CODE

```
/*Define path to csv file*/
FILENAME REFFILE '/folders/myfolders/5/data/ozr.csv';
/*Import data*/
PROC IMPORT DATAFILE=REFFILE DBMS=csv OUT=Ozone;
      GETNAMES=YES;
RUN;
/*Run Ttest on weight using the group class*/
PROC TTEST DATA=Ozone;
      CLASS Group;
      VAR Weight;
RUN;
```