STAT 500 Homework 4

Vu Thi-Hong-Ha; NetID: 851924086

September 18, 2020

1 Question 1

$Y \sim Poisson(\mu)$, then $E(Y) = \mu = Var(Y)$. If we write $Var(Y) = g(E(Y))$, then $g(x) = x$. The variance stabilizing transformation can be obtained in the form:
$h(y) \propto \int \frac{1}{\sqrt{g(x)}} dx = \int \frac{1}{\sqrt{x}} dx = 2\sqrt{x}$
Hence, $X = h(y) = \sqrt{Y}$ is a variance-stabilizing transformation for Y.

2 Question 2

(a)
Denote $Y_{control}$ and $Y_{bacilli}$ as the samples under control and those that received bacilli, respectively.
H_0: $\mu_{bacilli} - \mu_{control} = 0$
H_a: $\mu_{bacilli} - \mu_{control} < 0$
(i) Assume the two independent samples are from normal distributions with homogeneous variances:

The TTEST Procedure

Variable: time

treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Bacilli		58	242.5	117.9	15.4851	76.0000	598.0
Control		64	345.2	222.2	27.7747	18.0000	735.0
Diff (1-2)	Pooled		-102.7	180.4	32.6959		
Diff (1-2)	Satterthwaite		-102.7		31.7997		

treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Bacilli		242.5	211.5	273.5	117.9	99.6997	144.4
Control		345.2	289.7	400.7	222.2	189.3	269.1
Diff (1-2)	Pooled	-102.7	-Infty	-48.5015	180.4	160.1	206.5
Diff (1-2)	Satterthwaite	-102.7	-Infty	-49.8939			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	120	-3.14	0.0011
Satterthwaite	Unequal	97.807	-3.23	0.0008

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	63	57	3.55	<.0001

$p - value = 0.0011$. At the significance level of 5%, we reject the null that the two samples have the same mean survival times. We conclude that the infection with tubercle Bacilli tends to decrease survival times.
(ii) Permutation test based on computing t-statistics for each of 20,000 new random assignments of guinea pigs to treatment groups:

The FREQ Procedure

extreme_t	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	19975	99.88	19975	99.88
yes	25	0.13	20000	100.00

$p - value = 0.0013$. At the significance level of 5%, we reject the null that the two samples have the same mean survival times. We conclude that the infection with tubercle Bacilli tends to decrease survival times.
(b)

```
* (b) Wilcoxon rank-sum test;
proc npar1way data=pigs_surv wilcoxon;
    class treatment;
    var time;
    exact wilcoxon / alpha=.05 maxtime=20
                     MC N=20000 Seed=7892441;
run;
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable time Classified by Variable treatment					
treatment	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Control	64	4313.50	3936.0	195.052842	67.398438
Bacilli	58	3189.50	3567.0	195.052842	54.991379
Average scores were used for ties.					

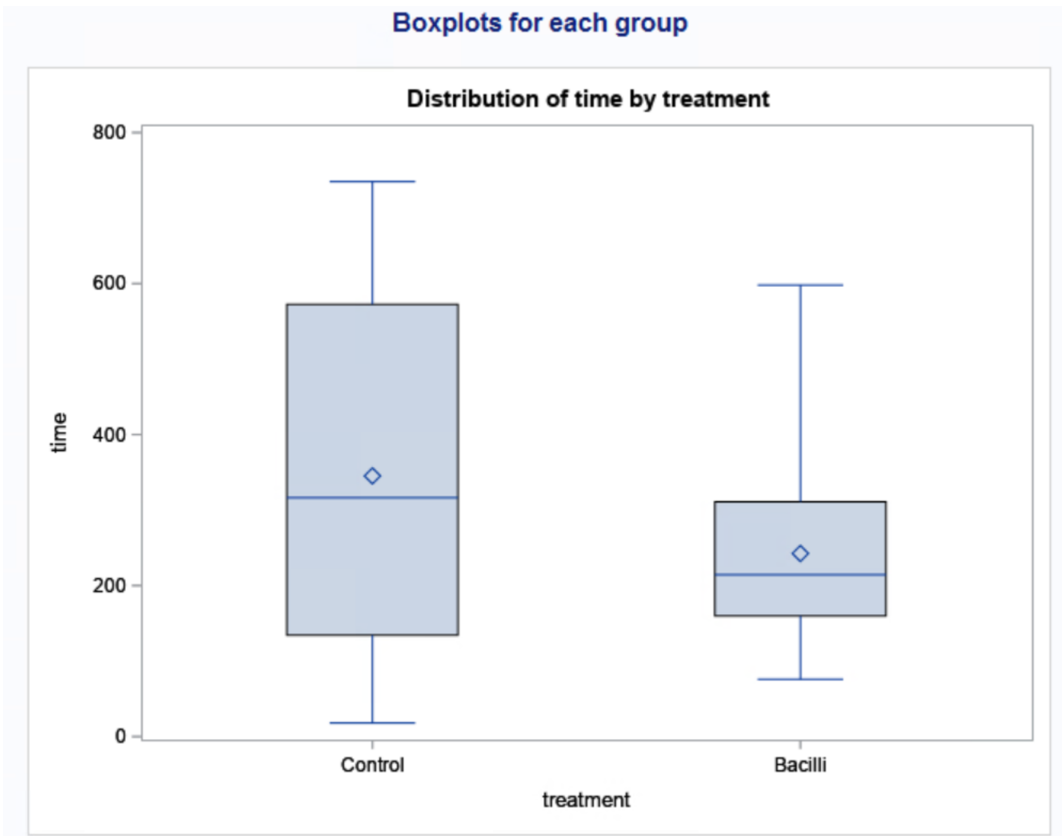
Wilcoxon Two-Sample Test					
				t Approximation	
Statistic (S)	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|
3189.500	-1.9328	0.0266	0.0533	0.0278	0.0556
Z includes a continuity correction of 0.5.					

Monte Carlo Estimates for the Exact Test					
Probability	Estimate	95% Confidence Limits		Samples	Seed
Pr <= S	0.0284	0.0261	0.0307	20000	7892441
Pr >= |S - Mean|	0.0552	0.0520	0.0583		

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
3.7457	1	0.0529

(i)
Sum of ranks for control group: 4313.50

Sum of ranks for bacilli group: 3189.50
(ii)
$p - value = 0.0266$
(iii)
At the significance level of 5%, we reject the null that the two samples have the same mean survival times. We conclude that the infection with tubercle Bacilli tends to decrease survival times.
(iv) The result from Wilcoxin rank-sum test agrees with the results from a(i) and a(ii).
(c)



From the box plots, we can see that the mean of the bacilli group is clearly lower than that of the control group. The result from the Wilcoxin test in part (b) agrees with the box plots.
(d) Assume the samples are independent and from normal distributions.
(i) The ratio of standard deviations:
$S_{control} = 222.2$; $S_{bacilli} = 117.9$
Ratio $= \dfrac{max\{S_{control}, S_{bacilli}\}}{min\{S_{control}, S_{bacilli}\}} = \dfrac{222.2}{117.9} = 1.884648$
With this ratio, the assumption of homogeneous variance does not make much impact on the test.
(ii) The folded F- test:
$F_{max} = \dfrac{max\{S^2_{control}, S^2_{bacilli}\}}{min\{S^2_{control}, S^2_{bacilli}\}} = \dfrac{222.2^2}{117.9^2} = 3.551898$
a $= n_{control} - 1 = 63$, b $= n_{bacilli} - 1 = 57$, so $F_{(a,b),1-\alpha/2} = F_{(63,57),0.975} \approx 1.5343$
Then $F_{max} > F_{(63,57),0.975}$. Therefore, we reject the null that the two samples have equal variances.
This result agrees with the F test carried out with t-test procedure in part (a)

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	120	-3.14	0.0011
Satterthwaite	Unequal	97.807	-3.23	0.0008

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	63	57	3.55	<.0001

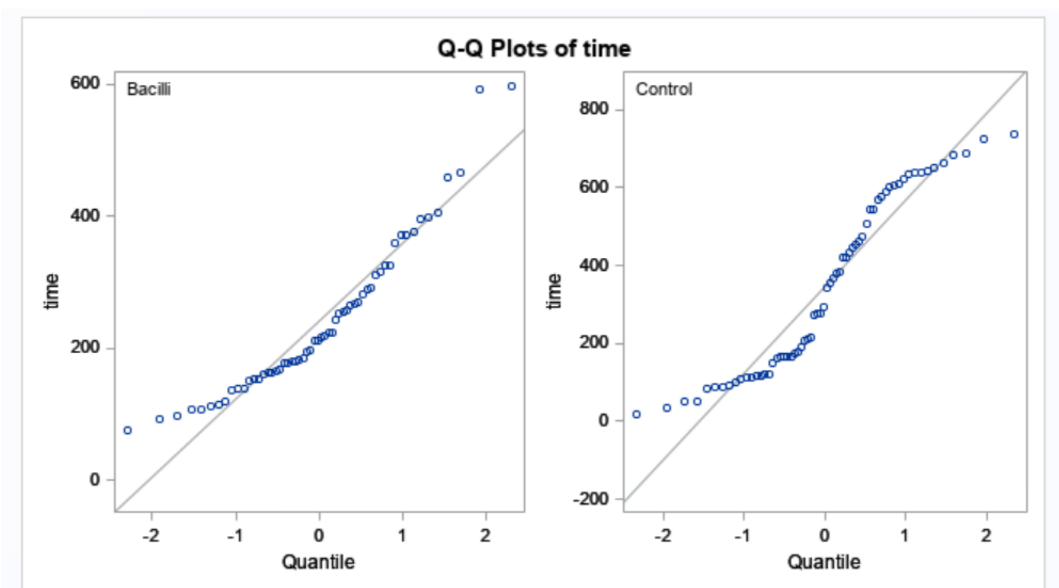
(iii) The Brown-Forsythe test:

The GLM Procedure

Dependent Variable: time

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	320913.593	320913.593	9.87	0.0021
Error	120	3903153.915	32526.283		
Corrected Total	121	4224067.508			

$p-value = 0.0021$. Therefore, we reject the null that the two samples have equal variances.
(e) Q-Q plots for the two samples:



The control group plot suggests that the sample has a light tail distribution. The bacilli group is not symmetric, and is skewed to the right.
(f)

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.910062	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.157089	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.323041	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	2.042673	Pr > A-Sq	<0.0050

$p-value = 0.0002$. Hence, we reject the null that the samples are following normal distribution.
(g) I would choose to carry out the Brown-Forsythe test because it is less sensitive to normality assumption like folded F-test, and we will not have to based on thresholds from previous study like ratio of sample standard deviations.

3 Question 3

(a) This is an experiment study as in this study treatments are assigned to experimental units and then the effect of the treatments on the experimental units are observed.
Treatment: build with extra insulation.
Experimental units: houses.
Observation units: annual gas assumption in MWh.
Replications: yes, each treatment has 10 replicates.
Randomization: yes.
(b) Assume the samples follow normal distributions with homogeneous variances. It is clear that the samples are independent. Let Y_{std} and Y_{extra} be the samples under standard insulation treatment and extra insulation treatment, respectively.
$H_0 : \mu_{std} = \mu_{extra}$
$H_a : \mu_{std} \neq \mu_{extra}$
$\bar{Y}_{extra} = 14.84$, $\bar{Y}_{std} = 17.04$; $S^2_{extra} = 2.1204^2 = 4.496096$; $S^2_{std} = 5.9367^2 = 35.24441$; $n_1 = n_2 = 10$.
$S_p = \sqrt{\dfrac{9 \times 4.496096 + 9 \times 35.24441}{18}} = \sqrt{19.87026} = 4.4576$
A 95% confidence interval is $((\bar{Y}_{std} - \bar{Y}_{extra}) \pm t_{n_1+n_2-2,1-\alpha/2} S_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}) \approx (2.2 \pm 2.10092 \times 4.4576 \times \sqrt{\dfrac{1}{9} + \dfrac{1}{9}}) = (-2.2147; 6.6147)$. If we construct such confidence interval for multiple times, 95% of the times those confidence intervals will contain the true mean difference.
(c) The 95% confidence interval does contain 0, so at 5% significance level, we fail to reject the null that the two group has equal means.
(d) $\alpha = 0.05$, then $z_{0.975} = 1.96$. Width of a 95% CI for the mean difference ≤ 1.
Consider $width = 1$. Then $1 = 2t_{2(n-1),0.975} S_p \sqrt{2/n}$.
Choose $n_0 = 8(\dfrac{1.96 \times 4.4576}{1})^2 = 610.67$. Choose $n_0 = 611$.
Iteration 1: $n_1 = 8(\dfrac{t_{611,0.975} \times 4.4576}{1})^2 = (\dfrac{1.96 \times 4.4576}{1})^2 = 610.67$.
Then the sample size for each group is 611.
(e)

```
proc power;
twosamplemeans
alpha = 0.05
meandiff = 1
stddev= 4.4576
npergroup = .
power = 0.975;
run;
```

The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Alpha	0.05
Mean Difference	1
Standard Deviation	4.4576
Nominal Power	0.975
Number of Sides	2
Null Difference	0

Computed N per Group	
Actual Power	N per Group
0.975	612

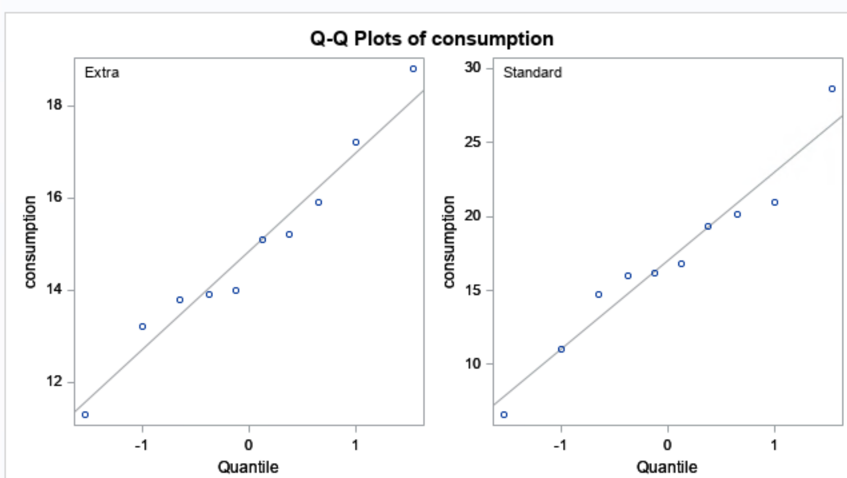
(f)
(i) The samples are clearly independent.
(ii) Check for homogeneous variance assumption:

The GLM Procedure

Dependent Variable: consumption

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	24.2000000	24.2000000	1.22	0.2843
Error	18	357.6680000	19.8704444		
Corrected Total	19	381.8680000			

As $p - value = 0.2843$, we fail to reject the null that the variances of the two samples are equal.
(iii) Check for normality assumption:



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.967632	Pr < W	0.8681
Kolmogorov-Smirnov	D	0.152376	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.042017	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.259332	Pr > A-Sq	>0.2500

Based on q-q plots and Shapiro-Wilk test which has p-value = 0.8681, we fail to reject the null that the samples follow normal distributions.

(g)

$H_0 : \mu_{std} - \mu_{extra} = 0$

$H_a : \mu_{std} - \mu_{extra} \neq 0$

Test statistic: $T = \dfrac{(\bar{Y_{std}} - \bar{Y_{extra}}) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \dfrac{2.2}{\sqrt{\frac{4.496096+35.24441}{10}}} = 1.1036$

$$d.f = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{S_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{S_2^2}{n_2})^2} = \frac{(\frac{4.496096}{10} + \frac{35.24441}{10})^2}{\frac{1}{9}(\frac{4.496096}{10})^2 + \frac{1}{9}(\frac{35.24441}{10})^2} = 11.2595$$

$t_{11.2595,0.975} \approx 2.195$

$p - value = 2 \times P(t_{d.f,0.975} > |T|) = 2 \times 0.1464 > 0.05$. At the significance level of 5%, we fail to reject the null that the two samples have equal means.

The TTEST Procedure

Variable: consumption

trt	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Extra		10	14.8400	2.1204	0.6705	11.3000	18.8000
Standard		10	17.0400	5.9367	1.8774	6.6000	28.6000
Diff (1-2)	Pooled		-2.2000	4.4576	1.9935		
Diff (1-2)	Satterthwaite		-2.2000		1.9935		

trt	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Extra		14.8400	13.3232	16.3568	2.1204	1.4585	3.8710
Standard		17.0400	12.7931	21.2869	5.9367	4.0835	10.8382
Diff (1-2)	Pooled	-2.2000	-6.3882	1.9882	4.4576	3.3682	6.5920
Diff (1-2)	Satterthwaite	-2.2000	-6.5754	2.1754			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	-1.10	0.2843
Satterthwaite	Unequal	11.259	-1.10	0.2928

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	7.84	0.0052

(h)

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable consumption Classified by Variable trt					
trt	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	10	124.0	105.0	13.228757	12.40
2	10	86.0	105.0	13.228757	8.60

Wilcoxon Two-Sample Test					
				t Approximation	
Statistic (S)	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|
124.0000	1.3985	0.0810	0.1620	0.0890	0.1781
Z includes a continuity correction of 0.5.					

Monte Carlo Estimates for the Exact Test					
Probability	Estimate	95% Confidence Limits		Samples	Seed
Pr >= S	0.0820	0.0781	0.0858	20000	7892441
Pr >= |S - Mean|	0.1650	0.1598	0.1701		

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
2.0629	1	0.1509

H_0 : two populations have the same distributions
H_a : two populations do not have the same distributions
Sum of ranks of standard group: 124.0
Sum of ranks of treatment group: 86.0
$p - value = 0.1781 > 0.05$. At the significance level of 5%, we fail to reject the null that the two samples have the same distribution.