Assignment 3

Statistical analyses of fruitfly data set

Name: Devosmita Chatterjee
Email Id: chatterjeedevosmita267@gmail.com

Statistical analyses of fruitfly data set

1 Introduction

The fruitfly data set consists of experimental data of five treatment groups. This is a two-way anova problem where the two factors are females and type.

2 Question 32a.

Table 1: This table presents the summary statistics for lifespan in each group.

	females	type	lifespan.Min.	lifespan.1st Qu.	lifespan.Median	lifespan.Mean	lifespan.3rd Qu.	lifespan.Max.
1	0	NA	37	47	62	63.56	75	96
2	1	pregnant	42	50	65	64.80	72	97
3	1	virgin	21	48	56	56.76	68	81
4	8	pregnant	35	56	65	63.36	77	86
5	8	virgin	16	32	40	38.72	47	60

Table 1 shows that there are significant differences in means between the different groups and the group (8,virgin) which mated the most has the lowest average lifespan.

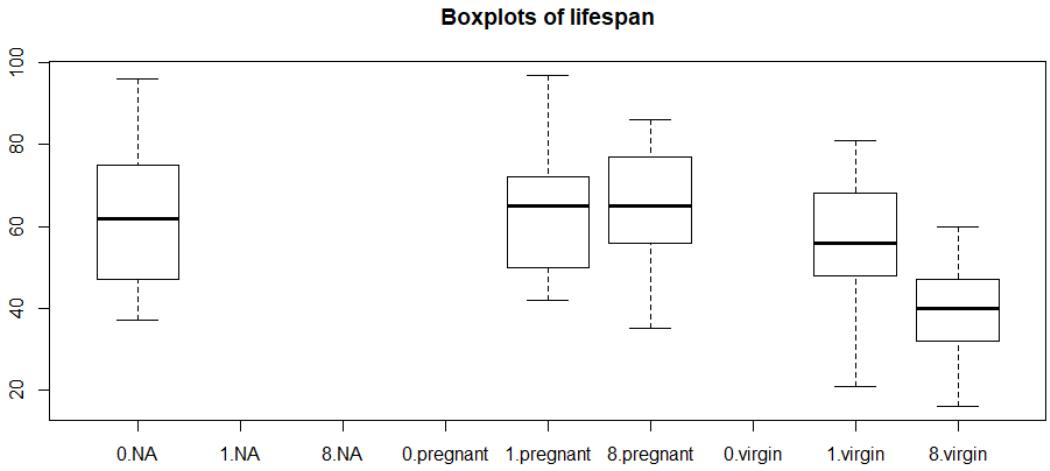

Figure 1: This figure shows the data in parallel boxplots of lifespan.

Also from figure 1, we see that the grouping (8,virgin) has the lowest mean. This concludes that increased reproduction decreases lifespan.

3 Question 32b.

Table 2: This table presents the summary statistics for sleep in each group.

	females	type	sleep.Min.	sleep.1st Qu.	sleep.Median	sleep.Mean	sleep.3rd Qu.	sleep.Max.
1	0	NA	2	14	18	21.56	33	50
2	1	pregnant	4	10	21	24.08	36	66
3	1	virgin	5	15	21	25.76	28	73
4	8	pregnant	1	14	23	25.16	26	83
5	8	virgin	4	12	20	20.76	30	40

Table 2 shows that there are no significant differences in means between the different groups.

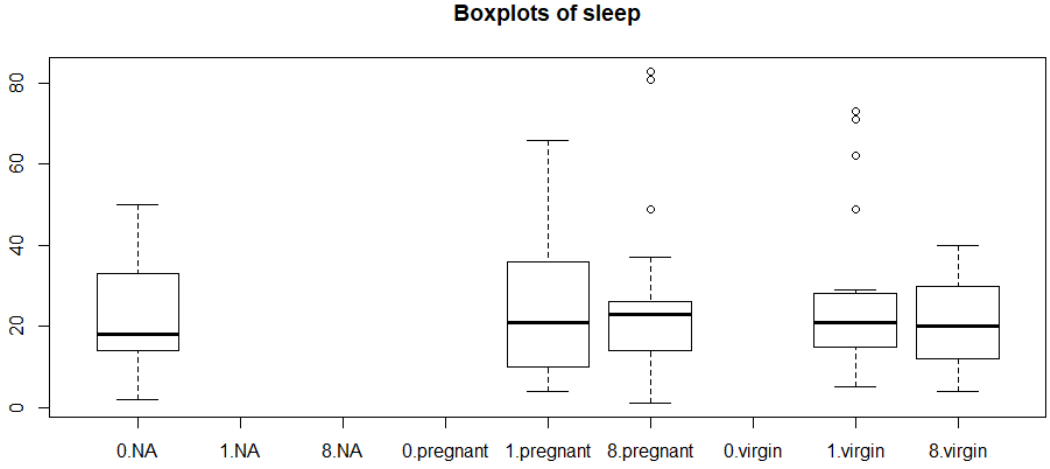

Figure 2: This figure shows the data in parallel boxplots of sleep.

From figure 2, we see that there are overlaps in the distributions of sleep and the observed mean differences are due to randomness.

4 Question 32c.

The scatterplot of lifespan versus thorax length can be seen in figure 3.

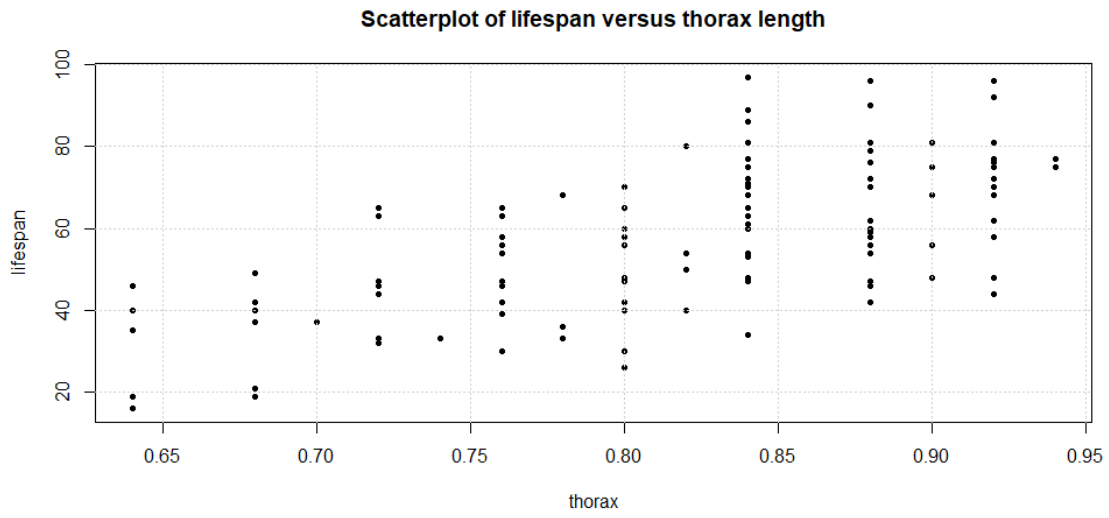

Figure 3: This figure shows the scatterplot of lifespan versus thorax length.

Table 3: This table presents the linear regression fit of thorax vs. lifespan.

Residuals:

Min	1Q	Median	3Q	Max
-28.415	-9.961	1.132	9.265	36.812

Coefficients:

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-61.05	13.00	-4.695	7.0e-06 ***
thorax	144.33	15.77	9.152	1.5e-15 ***

Residual standard error: 13.6 on 123 degrees of freedom
Multiple R-squared: 0.4051, Adjusted R-squared: 0.4003
F-statistic: 83.76 on 1 and 123 DF, p-value: 1.497e-15

We use linear regression fit of thorax vs. lifespan in table 3 and find a very significant P-value indicating that there is a relationship between the two variables- lifespan and thorax hence implying that thorax length is predictive of lifespan.

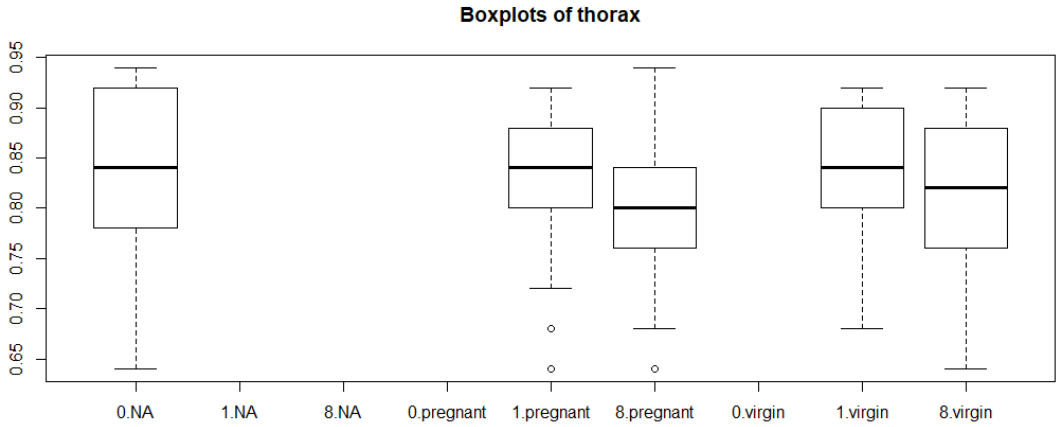

Figure 4: This figure shows the data in parallel boxplots of thorax.

We use boxplots shown in figure 4 to infer that the randomization balances thorax length between the groups.

5 Question 32d.

F-test

We present the F test to test for differences in longevity between the groups in table 4.

Table 4: This table presents the F test to test for differences in longevity between the groups.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
females	2	3542	1771	7.644	0.000748 ***
types	1	6675	6675	28.808	3.91e-07 ***
Residuals	121	28036	232		

We also show an interaction plot between type and females in figure 5.

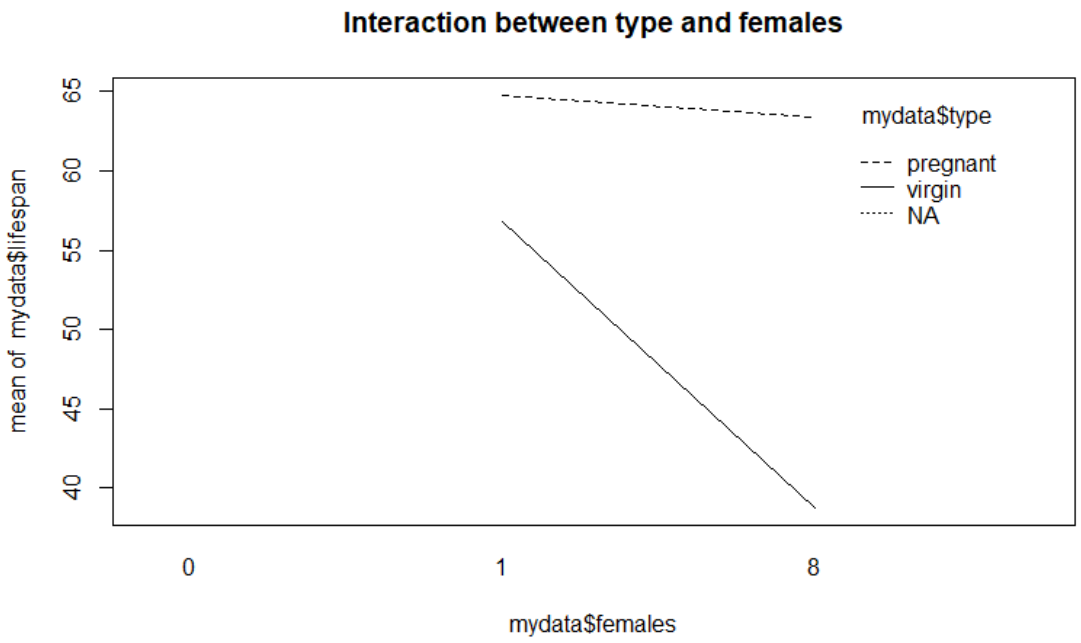

Figure 5: This figure shows the interaction plot between type and females.

Tukey method

If I independent samples $(y_{i1}, ..., y_{iJ})$ taken from $N(\mu_i, \sigma^2)$ have the same size J, then

$$Z_i = \bar{Y_{i.}} - \mu_i \sim N(0, \frac{\sigma^2}{J})$$

are independent. Consider the range of differences between Z_i:

$$R = max\{Z_1, ..., Z_I\} - min\{Z_1, ..., Z_I\}.$$

The normalised range has a distribution that is free from the parameter σ

$$\frac{R}{S_p/sqrtJ} \sim SR(I, df), df = I(J-1).$$

Tukey's 100(1-α)% simultaneous confidence interval is given by

$$T_{\mu_u - \mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm q_{I,df}(\alpha)\frac{s_p}{\sqrt{J}}.$$

We present the Tukey's method in table 5.

Table 5: This table presents the Tukey's method.

	diff	lwr	upr	p adj
pregnant-NA	8.17	-0.6774689	17.0174689	0.0767422
virgin-NA	-8.17	-17.0174689	0.6774689	0.0767422
virgin-pregnant	-16.34	-23.5639281	-9.1160719	0.0000012

The above results infer that there is a significant difference between the lifetime of the virgin group and the pregnant group.

Bonferroni method

Bonferroni method is a statistical test repeatedly applied to k independent samples of size n. The overall significance level α is obtained, if each single test is performed at significance level $\alpha_0 = \alpha/k$. Assuming the null hypothesis is true, the number of positive results is $X \sim Bin(k, \alpha_0)$. Thus for small values of α_0,

$$P(X \geq 1 \mid H_0) = 1 - (1 - \alpha_0)^k \approx k\alpha_0 = \alpha.$$

This gives Bonferroni's 100(1-α)% simultaneous confidence interval

$$B_{\mu_u - \mu_v} = \bar{y}_{u.} - \bar{y}_{v.} \pm t_{df}(\frac{\alpha}{2k}) s_p \sqrt{\frac{2}{J}}, 1 \leq u < v \leq I$$

where $df = I(J-1)$ and $k = \frac{I(I-1)}{2}$.

Here I = 3 and so we perform k = $\binom{I}{2}$ =3 t-tests to compare all pairs of means. We present the Bonferroni's method in table 6.

Table 6: This table presents the Bonferroni's method.

Groups	pregnant-NA	virgin-NA	virgin-pregnant
Diff	0.52	-15.82	-16.34

For the Bonferroni's method, we perform t-test on each of the three pairs and find no difference in means between the pregnant group and the NA group but there are differences in means between the other two groups- virgin-NA and virgin-pregnant.

6 Question 32e.

Kruskal-Wallis test

A nonparametric test, without assuming normality, for no treatment effect H_0: all observations are equal in distribution. Extending the idea of the rank-sum test, consider the pooled sample of size $N = IJ$. Let r_{ij} be the pooled ranks of the sample values y_{ij}, so that

$$\sum_i \sum_j r_{ij} = 1 + 2 + ... + N = \frac{N(N+1)}{2}$$

where the mean rank is $\bar{r}_{..} = \frac{(N+1)}{2}$. Kruskal-Wallis test statistic is given by

$$W = \frac{12J}{N(N+1)} \sum_{i=1}^{I} (\bar{r}_{i.} - \frac{N+1}{2})^2.$$

Reject H_0 for large W using the null distribution table.

We present the Kruskal-Wallis test in table 7.

Table 7: This table presents Kruskal-Wallis test.

data: lifespan by type
Kruskal-Wallis chi-squared = 24.4, df = 2, p-value = 5.029e-06

The above result infers that we find a difference between the virgin group and the other two groups.

7 Question 32f.

Table 8: This table presents the summary statistics for sleep in each type.

	type	sleep.Min.	sleep.1st Qu.	sleep.Median	sleep.Mean	sleep.3rd Qu.	sleep.Max.
1	NA	2	14.00	18.0	21.56	33.00	50
2	pregnant	1	12.00	22.5	24.62	28.75	83
3	virgin	4	13.25	20.5	23.26	28.75	73

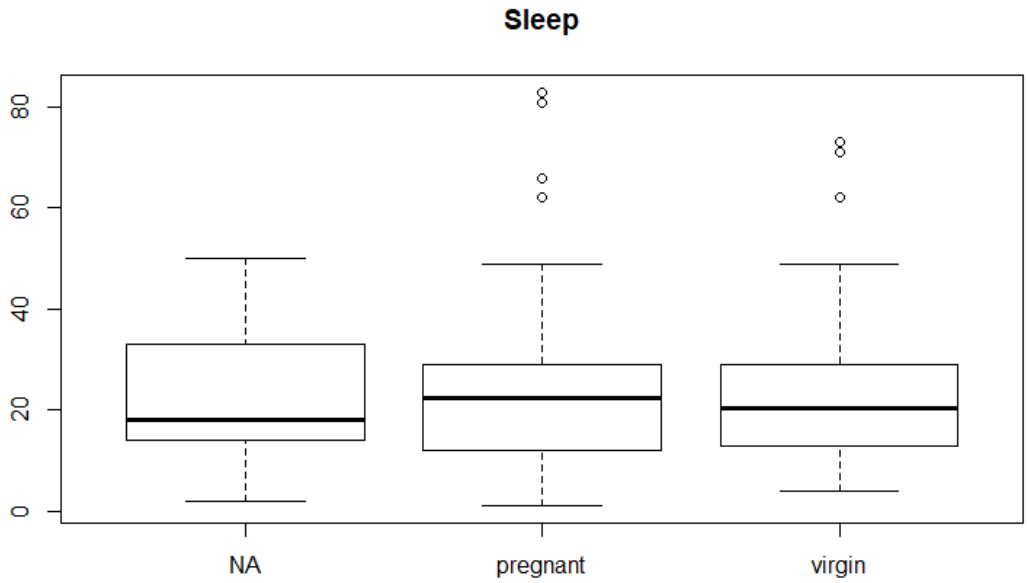

Figure 6: This figure shows the data in boxplots of sleep.

From table 8, we can see that there is a small difference between the average sleep when male fruitflies are paired with virgin females. We also infer the same thing when we visually look at the boxplots of figure 6.