

p8130_hw3_xj2249

xj2249

2019/10/24

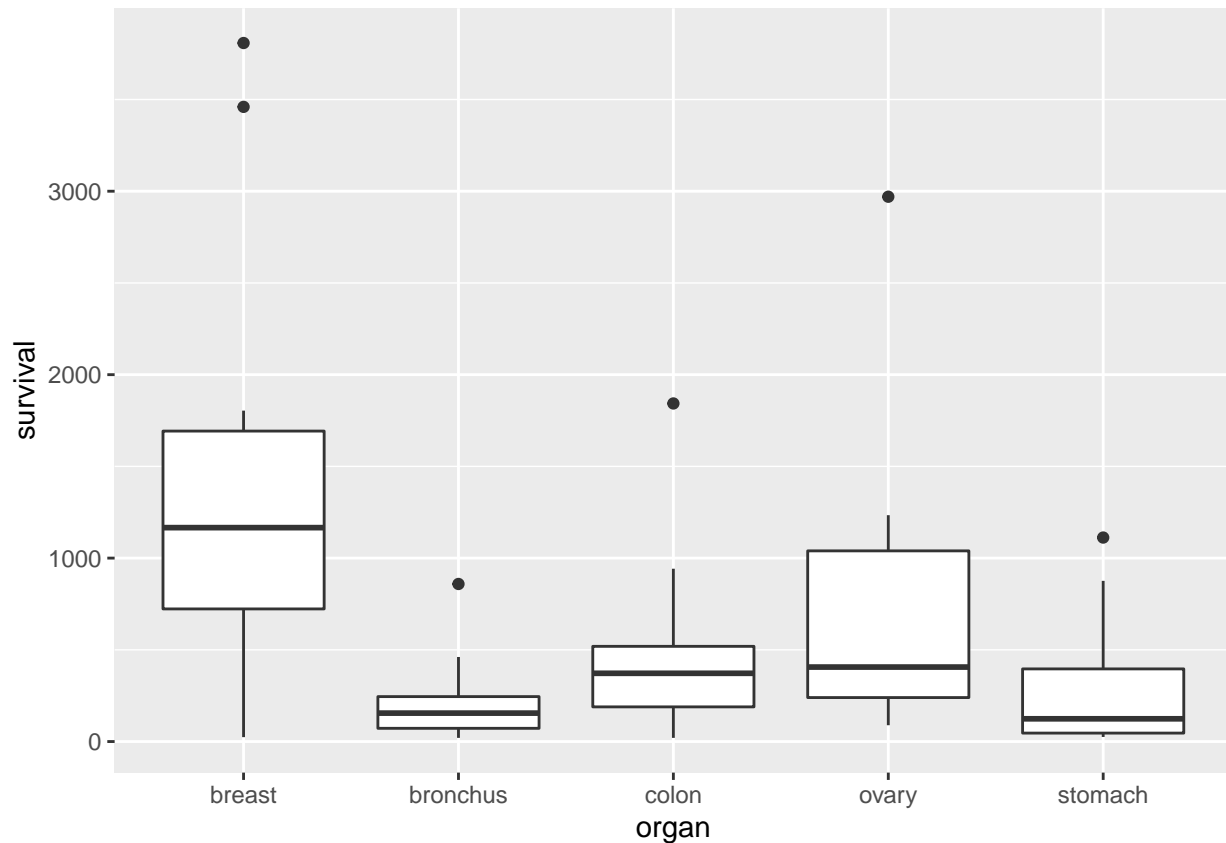
Problem2

a)

```
sur_df <- read_csv("./SurvCancer.csv") %>%  
  janitor::clean_names() %>%  
  mutate(organ = str_to_lower(organ),  
         organ = factor(organ))
```

```
## Parsed with column specification:  
## cols(  
##   SURVIVAL = col_double(),  
##   ORGAN = col_character()  
## )
```

```
# boxplot  
sur_df %>%  
  ggplot(aes(x = organ, y = survival)) +  
  geom_boxplot()
```



descriptive statistics

```
sur_df %>%
  group_by(organ) %>%
  summarise( n = n(),
             mean = mean(survival),
             median = median(survival),
             sd = sd(survival),
             IQR = IQR(survival),
             min = min(survival),
             max = max(survival)
           ) %>%
  knitr::kable()
```

organ	n	mean	median	sd	IQR	min	max
breast	11	1395.9091	1166	1238.9667	969.50	24	3808
bronchus	17	211.5882	155	209.8586	173.00	20	859
colon	17	457.4118	372	427.1686	330.00	20	1843
ovary	6	884.3333	406	1098.5788	799.75	89	2970
stomach	13	286.0000	124	346.3096	350.00	25	1112

b)

Hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \quad vs \quad H_1 : \text{not all means are equal}$$

Significance level: $\alpha = 0.01$

Assumptions: (1) Independence (2) equal variance (3) normality

Decision rule: Reject H_0 : if $F_{stats} > F_{4,59,1-\alpha/2} = F_{4,59,0.995} = 4.148$ Fail to reject H_0 : if $F_{stats} < F_{4,59,0.995}$

Interpretation: Since $F_{stats} = 6.433 > F_{4,59,1-\alpha/2} = F_{4,59,0.995} = 4.148$, we reject H_0 and conclude that there is a significant difference in average survival time among different cancer groups.

```
sur_aov <- aov(survival ~ organ, data = sur_df)
pander(sur_aov)
```

Table 2: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
organ	4	11535761	2883940	6.433	0.0002295
Residuals	59	26448144	448274	NA	NA

c) how to adjust 0.05 to 0.01

1) Bonferroni

```
pairwise.t.test(sur_df$survival, sur_df$organ, p.adj = 'bonferroni')
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: sur_df$survival and sur_df$organ
##
##      breast  bronchus  colon   ovary
## bronchus 0.00025 -      -      -
## colon    0.00608 1.00000 -      -
## ovary     1.00000 0.38575 1.00000 -
## stomach  0.00153 1.00000 1.00000 0.75283
##
## P value adjustment method: bonferroni
```

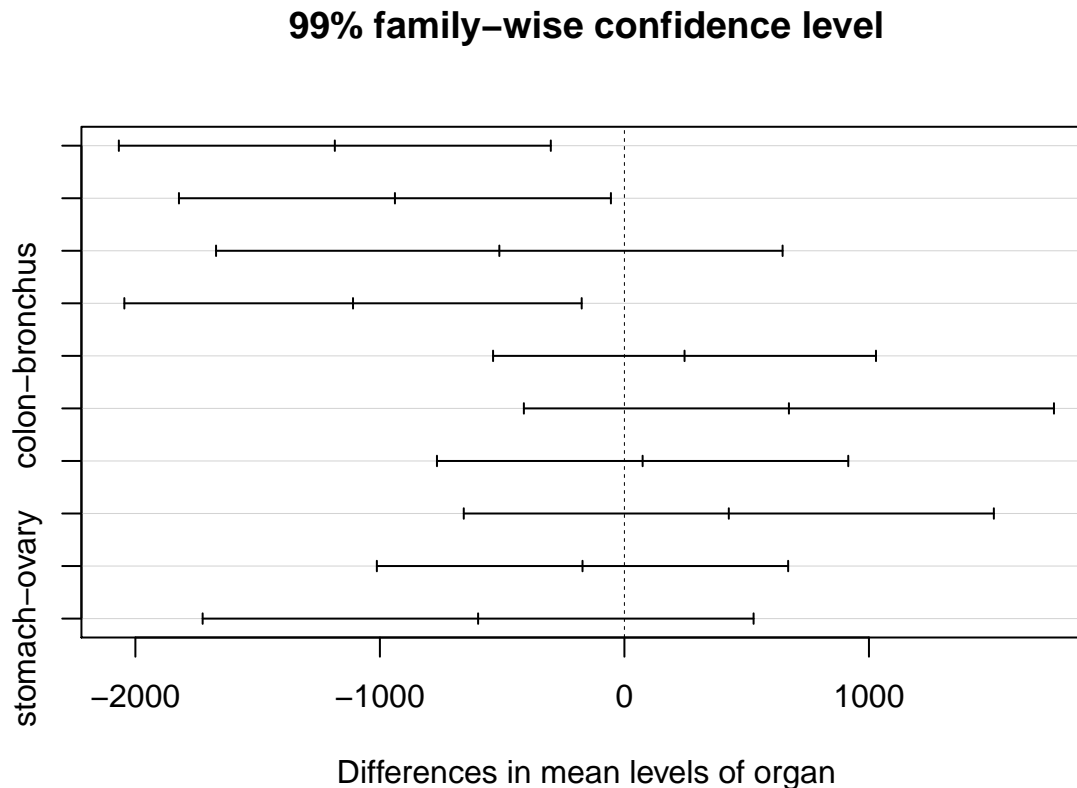
2) Tukey

```
TukeyHSD(sur_aov, conf.level = 0.99) ## 0.99???
```

```
## Tukey multiple comparisons of means
## 99% family-wise confidence level
##
## Fit: aov(formula = survival ~ organ, data = sur_df)
##
## $organ
##              diff              lwr              upr              p adj
## bronchus-breast -1184.32086 -2067.6073 -301.03446 0.0002385
## colon-breast    -938.49733 -1821.7837  -55.21093 0.0053072
```

```
## ovary-breast      -511.57576 -1670.0752  646.92367 0.5630900
## stomach-breast    -1109.90909 -2045.0583 -174.75983 0.0013962
## colon-bronchus     245.82353  -537.1262 1028.77324 0.8208402
## ovary-bronchus     672.74510  -411.1997 1756.68989 0.2271084
## stomach-bronchus    74.41176  -766.6111  915.43467 0.9981461
## ovary-colon        426.92157  -657.0232 1510.86636 0.6659115
## stomach-colon      -171.41176 -1012.4347  669.61114 0.9568289
## stomach-ovary      -598.33333 -1724.9413  528.27467 0.3772923
```

```
TukeyHSD(sur_aov, conf.level = 0.99) %>% plot()
```



3) Dunnett Test

```
glht(sur_aov, linfct = mcp(organ = "Dunnett")) %>% summary()
```

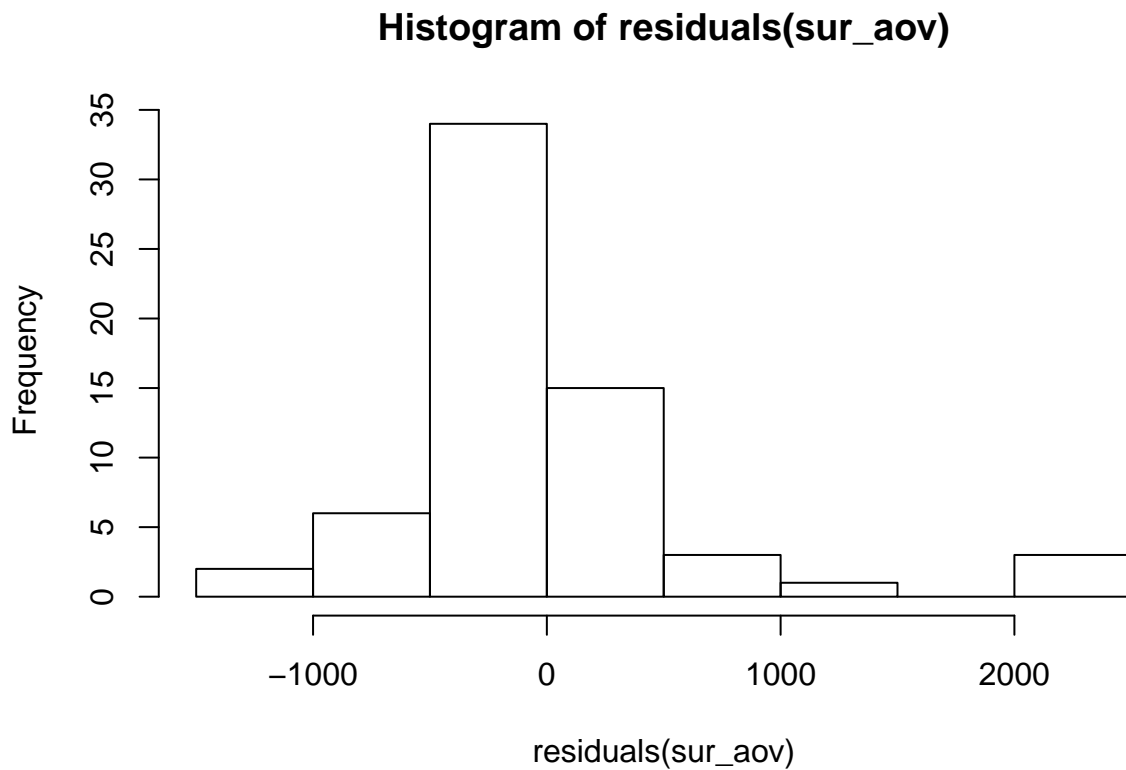
```
##
## Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = survival ~ organ, data = sur_df)
##
## Linear Hypotheses:
##              Estimate Std. Error t value Pr(>|t|)
## bronchus - breast == 0 -1184.3      259.1  -4.571  < 0.001 ***
## colon - breast == 0    -938.5      259.1  -3.622  0.00224 **
```

```
## ovary - breast == 0      -511.6      339.8 -1.506  0.36697
## stomach - breast == 0   -1109.9      274.3 -4.046 < 0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

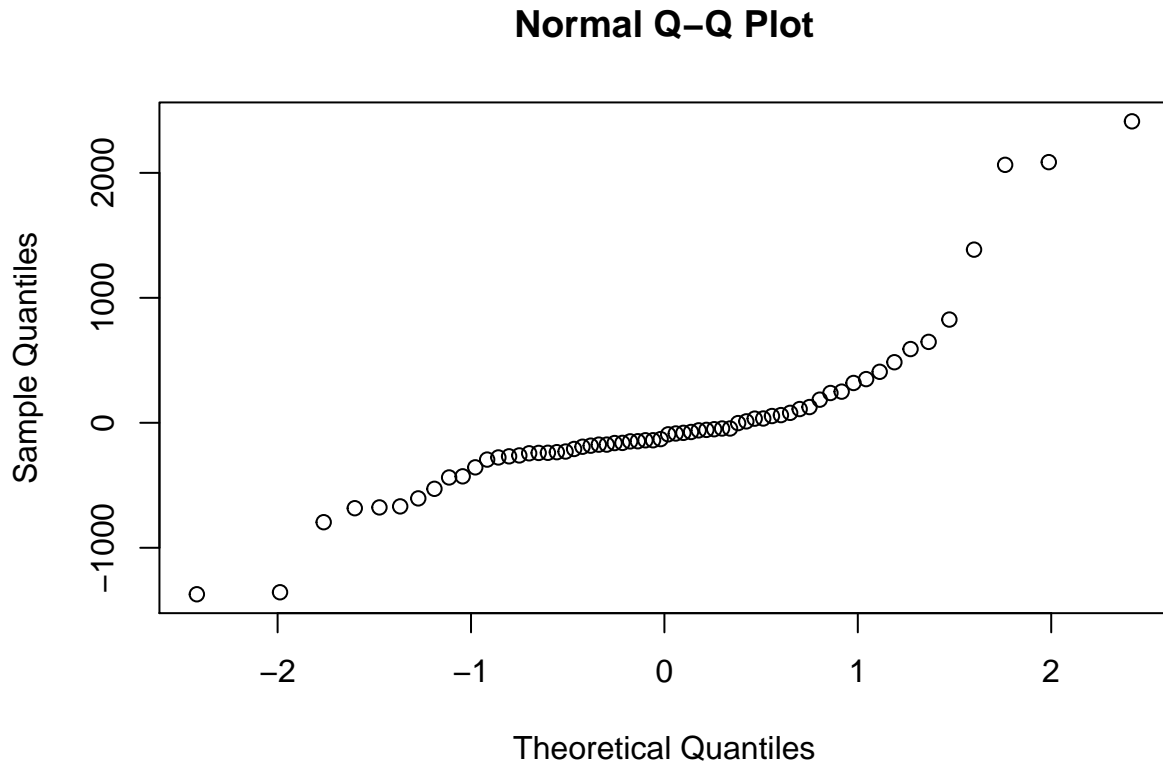
d)

i) check the normality assumption

```
# first, try a hist/density plot.
hist(residuals(sur_aov))
```



```
# check the normality (of residuals) assumption
qqnorm(residuals(sur_aov))
```



As the QQ-plot shows, the normality assumption is questionable. Therefore, we can Kruskal Wallis test to fix the problem.

ii) KW test

```
kruskal.test(survival ~ organ, data = sur_df) %>% pander()
```

Table 3: Kruskal-Wallis rank sum test: `survival` by `organ` The p-value is 0.005, at a significance level 0.01, we reject H_0 and conclude that there is a significant difference in average survival time among different cancer groups. The p-value of kw-test is 0.004798, much larger than that of the anova test(0.0002295), which shows that kw-test is harder to reject H_0 and it's more conservative and less powerful.

Test statistic	df	P value
14.95	4	0.004798 * *

Problem3