One-Way Analysis of Variance

STA 610 - Applied Statistics for Health Professions

Andrew DiLernia

## Learning Objectives

1. The One-Way ANOVA Model: Comparing Groups
2. Checking model assumptions
3. Obtaining and interpreting results of the ANOVA table
4. Conducting and interpreting the results of a post-hoc analysis

Let’s begin by loading some R packages for this activity using the code below. Note: if it is the first time you are using an R package, you may need to install it first using the install.packages() function.

*# Load necessary packages*  
library(tidyverse)  
library(ggthemes)  
library(flextable)  
library(broom)  
library(ggfortify)

Next, we can set default theme settings for plots, and load some functions to simplify table customization and creation using the code below.

*# Set ggplot theme for visualizations*  
theme\_set(ggthemes::theme\_few())  
  
*# Set options for flextables*  
set\_flextable\_defaults(na\_str = "NA")  
  
*# Load function for printing tables nicely*  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")

## GPA Example

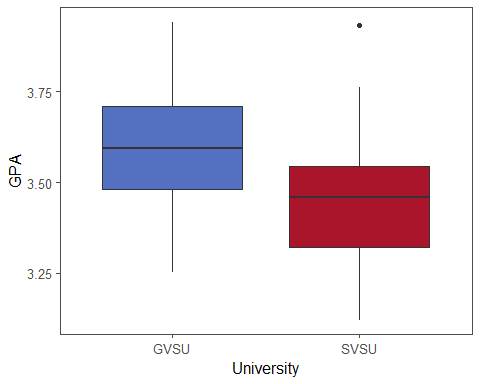
Consider how we can use the two-sample -test to compare the average quantitative response between two groups.

*# Importing fake GPA data*  
simGPA <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/simGPA.csv")  
  
*# Creating vector of colors*  
schoolColors <- setNames(c("#4E2583", "#5470c0", "#aa152b"),  
 c("AU", "GVSU", "SVSU"))

set.seed(1994)  
  
*# Printing random rows of data set*  
simGPA %>%   
 dplyr::filter(University %in% c("GVSU", "SVSU")) %>%   
 slice\_sample(n = 7) %>%   
 make\_flex()

| GPA | University |
| --- | --- |
| 3.66 | SVSU |
| 3.76 | GVSU |
| 3.66 | GVSU |
| 3.78 | GVSU |
| 3.44 | SVSU |
| 3.27 | GVSU |
| 3.65 | GVSU |

*# Creating side-by-side box plots*  
twoSampGG <- simGPA %>%   
 dplyr::filter(University %in% c("GVSU", "SVSU")) %>%   
 ggplot(aes(x = University, y = GPA, fill = University)) +   
 geom\_boxplot() +   
scale\_fill\_manual(values = schoolColors) +   
 theme(legend.position = "none")  
   
twoSampGG



We can also calculate summary statistics for each group.

*# Calculating descriptive statistics*  
quantStats <- simGPA %>%   
 dplyr::filter(University %in% c("GVSU", "SVSU")) %>%   
 group\_by(University) %>%   
 summarize(  
 Minimum = min(GPA, na.rm = TRUE),  
 Q1 = quantile(GPA, na.rm = TRUE, probs = 0.25),  
 M = median(GPA, na.rm = TRUE),  
 Q3 = quantile(GPA, na.rm = TRUE, probs = 0.75),  
 Maximum = max(GPA, na.rm = TRUE),  
 Mean = mean(GPA, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(GPA, na.rm = TRUE),  
 n = n()  
)  
  
*# Printing table of statistics*  
quantStats %>%   
 make\_flex(caption = "Summary statistics for student GPAs by university.")

*Table 2: Summary statistics for student GPAs by university.*

| University | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GVSU | 3.25 | 3.48 | 3.59 | 3.71 | 3.94 | 3.58 | 0.69 | 0.18 | 50 |
| SVSU | 3.12 | 3.32 | 3.46 | 3.55 | 3.93 | 3.45 | 0.81 | 0.17 | 50 |

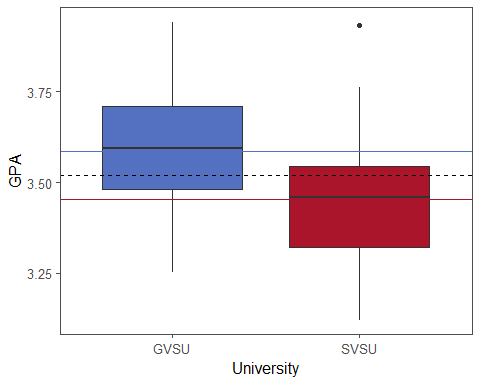
We can also provide a linear model interpretation of the test. Our model statement for the two-sample -test is given by

where

* is the high school GPA for the GVSU student in our sample
* is the high school GPA for the SVSU student in our sample
* is the average GPA of the ith university
* are the independent and normally distributed error terms

A formal statement of our hypotheses for the two-sample -test are:

where the average GPA of GVSU students and of SVSU students.



What if we had more than 2 groups?

## One-Way ANOVA

For a single quantitative response with more than 2 groups, we can use an **AN**alysis **O**f **VA**riance (ANOVA) model.

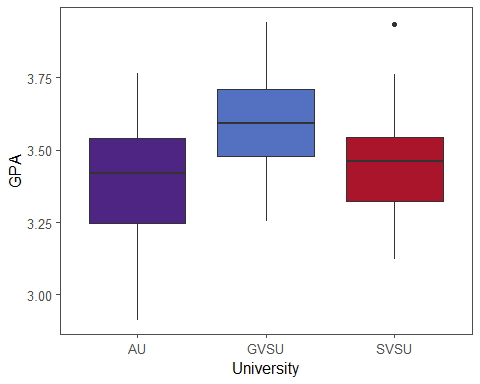
**Questions explored using ANOVA**:

1. Does the mean GPA differ between majors for students at GVSU? 🎓
2. Does the average cost of a textbook vary by department (i.e. Math, Biology, and Spanish)? 📚
3. Does the average blood-pressure differ for 4 different types of medications? 🩺💊

➡️ What are the response and explanatory variables for each of the example scenarios?

What if we wanted to compare the average GPA of GVSU, SVSU, and Ashland University?

*# Creating side-by-side box plots*  
threeSampGG <- simGPA %>%   
 ggplot(aes(x = University, y = GPA, fill = University)) +   
 geom\_boxplot() +   
scale\_fill\_manual(values = schoolColors) +   
 theme(legend.position = "none")  
   
threeSampGG



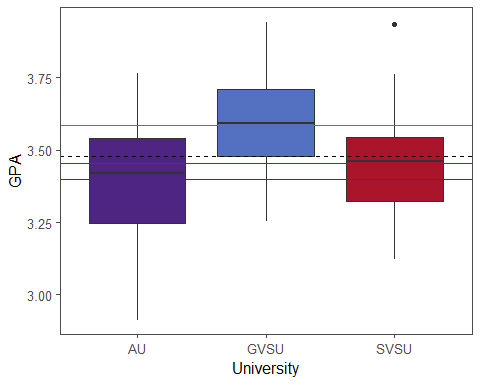
*# Calculating descriptive statistics*  
quant3Stats <- simGPA %>%   
 group\_by(University) %>%   
 summarize(  
 Minimum = min(GPA, na.rm = TRUE),  
 Q1 = quantile(GPA, na.rm = TRUE, probs = 0.25),  
 M = median(GPA, na.rm = TRUE),  
 Q3 = quantile(GPA, na.rm = TRUE, probs = 0.75),  
 Maximum = max(GPA, na.rm = TRUE),  
 Mean = mean(GPA, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(GPA, na.rm = TRUE),  
 n = n()  
)  
  
*# Printing table of statistics*  
quant3Stats %>%   
 make\_flex(caption = "Summary statistics for student GPAs by university.")

*Table 3: Summary statistics for student GPAs by university.*

| University | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| AU | 2.91 | 3.25 | 3.42 | 3.54 | 3.76 | 3.40 | 0.85 | 0.22 | 50 |
| GVSU | 3.25 | 3.48 | 3.59 | 3.71 | 3.94 | 3.58 | 0.69 | 0.18 | 50 |
| SVSU | 3.12 | 3.32 | 3.46 | 3.55 | 3.93 | 3.45 | 0.81 | 0.17 | 50 |

The **group means** formulation of the One-Way ANOVA model in general is:

* : response for the observation of the group
* : group specific means
* : independent and normally distributed error terms



To test if the group means are equal, a formal statement of the hypotheses for the group-means formulation of the One-Way ANOVA model is:

This test is conducted using an *F*-statistic and corresponding p-value from the ANOVA table.

The **group effects** formulation of ANOVA model in general is:

* : response for the observation of the group
* : mean response of all groups together
* : group effect of the group
* : independent and normally distributed error terms

For the group-effects approach, our hypotheses for testing if the group means are equal are:

Note that this is equivalent to testing:

Since the group-means formulation of the model is more commonly used, we will default to this approach moving forward.

## One-Way ANOVA Table

| Source | df | SS | MS | F.value | p.value |
| --- | --- | --- | --- | --- | --- |
| Model | *G-1* | SSGroups | MSGroups | *F* | p-value |
| Error | *n-G* | SSE | MSE |  |  |
| Total | *n-1* | SSTotal |  |  |  |

* : number of groups
* : total number of data points or observations

where stands for degrees of freedom, for sum of squares, and for mean square.

## One-Way ANOVA assumptions

The necessary assumptions / conditions to conduct a valid -test for a One-Way ANOVA model pertain to the errors, also called the residuals. The errors must:

1. Have the same standard deviation within each group (implies equal variances, also known as homoskedasticity)
2. Be normally distributed
3. Be independent

## Checking assumptions

### Equal variances

There are a few ways we can check the equal variances assumption, also called homoskedasticity.

1. Plot the residual by fitted values. If there is a fairly constant vertical spread going from left to right in the plot, this suggests that the homoskedasticity assumption is met.
2. Check if where is the largest standard deviation of any group and is the minimum standard deviation of any group. If this inequality is not true, then the assumption of equal variances is potentially violated.

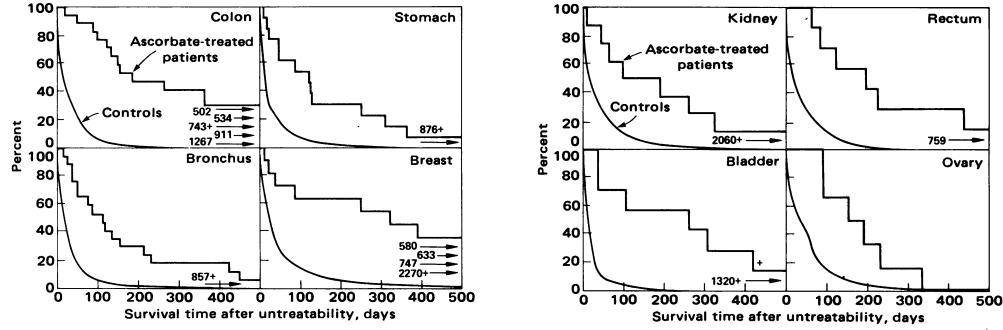
### [Normality](https://docs.google.com/document/d/1jVscyqIwKhuVfzlP60_8XNKoWQz7vvnN/edit#heading=h.1ksv4uv)

There are a few ways we can check the normality of the residuals.

1. Create a quantile-quantile plot ([QQ-plot](https://docs.google.com/document/d/1jVscyqIwKhuVfzlP60_8XNKoWQz7vvnN/edit#heading=h.35nkun2)) of the residuals. For Q-Q plots, the closer points fall along the diagonal line through the origin, the more closely the distribution of values resembles that of a normal distribution.
2. Plot a [histogram](https://docs.google.com/document/d/1jVscyqIwKhuVfzlP60_8XNKoWQz7vvnN/edit#heading=h.35nkun2) of the residuals. If the distribution of the residuals is unimodal, symmetric, and fairly bell-shaped, then the normality assumption is typically met.

## Survival Example

We consider a real-life example on the survivability of people diagnosed with cancer. Doctors in the 1970’s investigated if ascorbate, a form of Vitamin C, could extend the lives of people with cancer. Comparing patients with cancer who received ascorbate to patients with cancer who did not receive the supplement through a designed experiment, they found ascorbate extended the lives of the patients. From these results, they then wondered if the effect of ascorbate differed depending on the organ affected by the cancer. Considering only the participants who received the ascorbate, the doctors explored the survivability for 5 of the cancer types: stomach, bronchus, colon, ovary, and breast cancer.



*Figure 1. Kaplan-Meier plots displaying the percent of survivors at times after date of onset of terminal stage of ascorbate-treated patients with primary cancer of colon, stomach, bronchus, or breast, kidney, rectum, bladder, and ovary compared with that for matched controls (10 per ascorbate-treated patient).*

Data and figure from the article “[Supplemental Ascorbate in the Supportive Treatment of Cancer: Reevaluation of Prolongation of Survival Times in Terminal Human Cancer](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC336151/pdf/pnas00668-0476.pdf)” by Ewan Cameron and Linus Pauling, Proceedings of the National Academy of Sciences of the United States of America, Vol. 75, No. 9 (Sep., 1978), pp. 4538-4542.

A corresponding data set consists of 64 observations on the following 2 variables:

1. **Survival** - Survival time in days
2. **Organ** - Location or organ where cancer likely originated, being one of breast, bronchus, colon, ovary, or stomach

*# Loading survival data*  
cancerSurvival <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/CancerSurvival.csv")

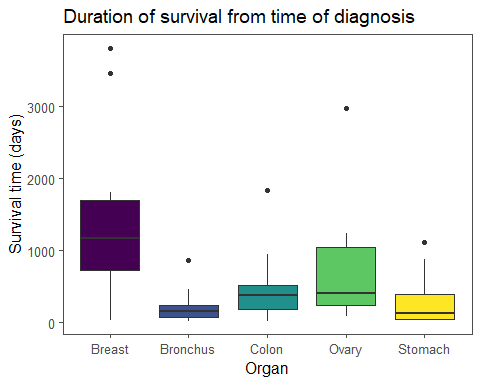
### Exploratory data analysis

➡️ Print 8 randomly selected rows from the survival data set, reproducing the table below.

*Table 4: Randomly selected rows from survival data set.*

| Survival | Organ |
| --- | --- |
| 151.00 | Bronchus |
| 776.00 | Colon |
| 719.00 | Breast |
| 146.00 | Stomach |
| 37.00 | Bronchus |
| 537.00 | Colon |
| 51.00 | Stomach |
| 519.00 | Colon |

*# Creating side-by-side box plots*  
cancerSurvival %>%   
 ggplot(aes(x = Organ, y = Survival, fill = Organ)) +   
 geom\_boxplot() +   
 labs(y = "Survival time (days)",  
 title = "Duration of survival from time of diagnosis") +  
 scale\_fill\_viridis\_d() +  
 theme(legend.position = "none")



➡️ Calculate summary statistics for each group, reproducing the table below.

*Table 5: Summary statistics for duration of survival from time of diagnosis by cancer type.*

| Organ | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Breast | 24.00 | 723.00 | 1,166.00 | 1,692.50 | 3,808.00 | 1,395.91 | 3,784.00 | 1,238.97 | 11 |
| Bronchus | 20.00 | 72.00 | 155.00 | 245.00 | 859.00 | 211.59 | 839.00 | 209.86 | 17 |
| Colon | 20.00 | 189.00 | 372.00 | 519.00 | 1,843.00 | 457.41 | 1,823.00 | 427.17 | 17 |
| Ovary | 89.00 | 239.75 | 406.00 | 1,039.50 | 2,970.00 | 884.33 | 2,881.00 | 1,098.58 | 6 |
| Stomach | 25.00 | 46.00 | 124.00 | 396.00 | 1,112.00 | 286.00 | 1,087.00 | 346.31 | 13 |

➡️ Based on the output obtained, which cancer type has the longest survival on average and what is the corresponding average survival time?

➡️ Which cancer type has the most variability in its survival times? Why?

➡️ What was the shortest duration of survival for any individual person, and what was the longest? Specify which type of cancer each person had.

### Fitting One-Way ANOVA model

We consider the One-Way ANOVA model for comparing the average survival in days across the five cancer types.

➡️ Provide a statement of the group means formulation of the One-Way ANOVA model.

➡️ Provide a formal statement of the hypotheses for the One-Way ANOVA.

➡️ What are the response and explanatory variables for this One-Way ANOVA?

We implement a One-Way ANOVA model for comparing the average survival time in days across the five cancer types using the code below.

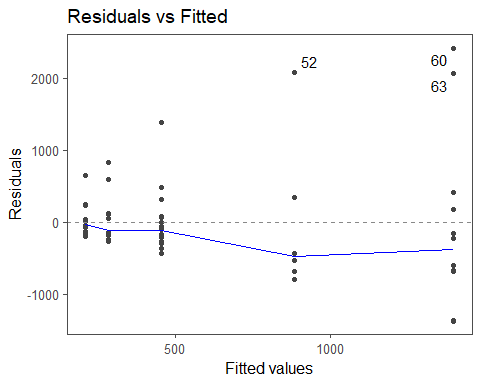
*# Fitting One-Way ANOVA model*  
survivalModel <- aov(Survival ~ Organ, data = cancerSurvival)  
  
*# Printing ANOVA table*  
survivalModel %>%   
 broom::tidy() %>%   
 make\_flex(caption = "ANOVA table for survival duration in days by cancer type")

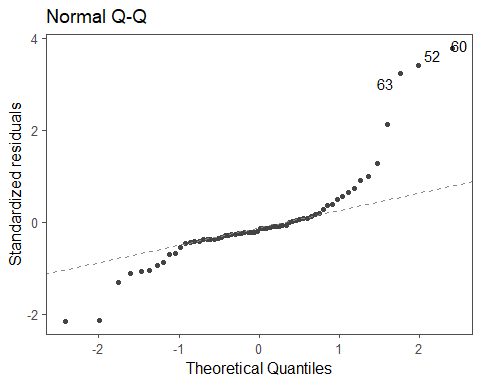
*Table 6: ANOVA table for survival duration in days by cancer type*

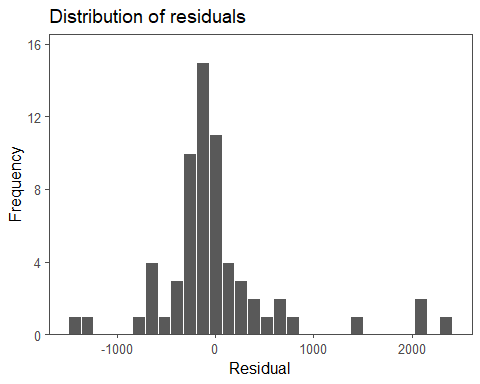
| term | df | sumsq | meansq | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| Organ | 4.00 | 11,535,760.52 | 2,883,940.13 | 6.43 | 0.00023 |
| Residuals | 59.00 | 26,448,144.48 | 448,273.64 | NA | NA |

### Checking assumptions

To conduct a valid test of the overall One-Way ANOVA model, we need to check that the model assumptions are met. We can obtain diagnostic plots for checking these assumptions using the code below.







➡️ State and check whether the model assumptions are met for the One-Way ANOVA, citing specific evidence from the obtained output to support your conclusions. Clearly convey each assumption, whether or not it is met, and why.

Even though model assumptions are not met here, for demonstration purposes we proceed with a One-Way ANOVA model for comparing the average survival in days across the five cancer types.

➡️ Provide the values of SSGroups, SSE, and SSTotal, and interpret them in the context of the problem.

➡️ From the model output, do the average survival times differ by cancer type?

➡️ Provide the value of the test statistic, p-value, and decision for the overall -test, citing specific evidence from the obtained output.

➡️ Provide our interpretation based on our stated decision in the context of the problem.

## Transformations

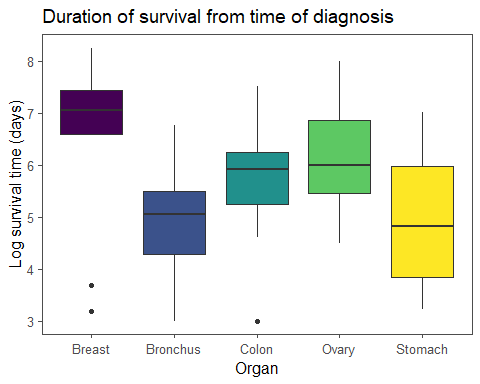
When a quantitative response variable is non-negative and unbounded, it is common that residuals for a One-Way ANOVA model are skewed-right. One solution in this setting is to transform the response variable, Survival, with common transformations being the square-root or logarithm.

The square-root and log-transformation are both monotonically increasing transformations, in that they preserve the ordering of values in a data set. Formally, a transformation is monotonically increasing if implies that . In this context, if person A survived 25 days and person B survived 49 days where , then the square-root of the survival of person A is still less than that of person B, where and . This follows for the log-transformed survival times as well, since where and .

Note that transformations of the response variable can help satisfy model assumptions, but come at a cost of interpretability. Due to this loss of interpretability, **transformations should not be used unless necessary**.

*# Calculating log-transformation of the survival times*  
cancerSurvival <- cancerSurvival %>%   
 dplyr::mutate(logSurvival = log(Survival))

*# Creating side-by-side box plots*  
cancerSurvival %>%   
 ggplot(aes(x = Organ, y = logSurvival, fill = Organ)) +   
 geom\_boxplot() +   
 labs(y = "Log survival time (days)",  
 title = "Duration of survival from time of diagnosis") +  
 scale\_fill\_viridis\_d() +  
 theme(legend.position = "none")



➡️ Fit a One-Way ANOVA model using a log transformation of the survival times, reproducing the ANOVA table below.

*Table 7: ANOVA table for survival duration in days by cancer type*

| term | df | sumsq | meansq | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| Organ | 4.00 | 24.49 | 6.12 | 4.29 | 0.0041 |
| Residuals | 59.00 | 84.27 | 1.43 | NA | NA |

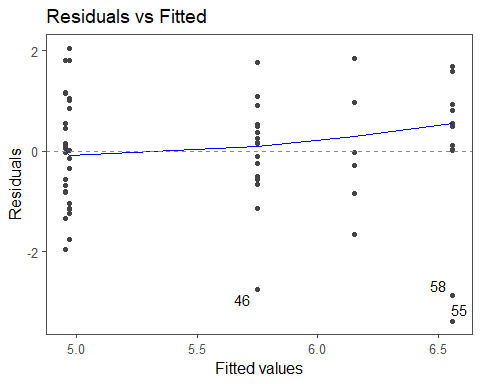
## Rechecking assumptions

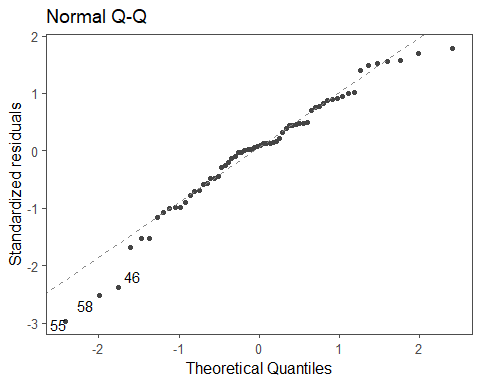
➡️ Reproduce the summary statistics and diagnostic plots below for the model fit using the log survival times.

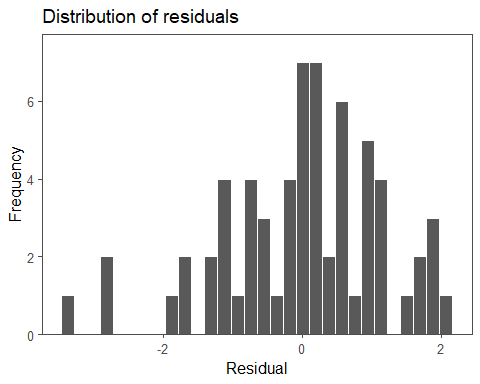
*# Calculating descriptive statistics*  
quantStats <- cancerSurvival %>%   
 group\_by(Organ) %>%   
 summarize(  
 Minimum = min(logSurvival, na.rm = TRUE),  
 Q1 = quantile(logSurvival, na.rm = TRUE, probs = 0.25),  
 M = median(logSurvival, na.rm = TRUE),  
 Q3 = quantile(logSurvival, na.rm = TRUE, probs = 0.75),  
 Maximum = max(logSurvival, na.rm = TRUE),  
 Mean = mean(logSurvival, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(logSurvival, na.rm = TRUE),  
 n = n()  
)  
  
*# Printing table of statistics*  
quantStats %>%   
 make\_flex(caption = "Summary statistics for log duration of survival from time of diagnosis by cancer type.")

*Table 8: Summary statistics for log duration of survival from time of diagnosis by cancer type.*

| Organ | Minimum | Q1 | M | Q3 | Maximum | Mean | R | s | n |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Breast | 3.18 | 6.58 | 7.06 | 7.43 | 8.24 | 6.56 | 5.07 | 1.65 | 11 |
| Bronchus | 3.00 | 4.28 | 5.04 | 5.50 | 6.76 | 4.95 | 3.76 | 0.95 | 17 |
| Colon | 3.00 | 5.24 | 5.92 | 6.25 | 7.52 | 5.75 | 4.52 | 1.00 | 17 |
| Ovary | 4.49 | 5.45 | 6.00 | 6.87 | 8.00 | 6.15 | 3.51 | 1.26 | 6 |
| Stomach | 3.22 | 3.83 | 4.82 | 5.98 | 7.01 | 4.97 | 3.80 | 1.25 | 13 |







➡️ State and check whether the model assumptions are met for the One-Way ANOVA using the log transformed survival times, citing specific evidence from obtained output to support your conclusions. Clearly convey each assumption, whether or not it is met, and why.

## Interpreting the results

➡️ Provide the values of SSGroups, SSE, and SSTotal, and interpret them in the context of the problem.

➡️ From the model output, do the average log survival times differ by cancer type?

➡️ Provide the value of the test statistic, p-value, and decision for the overall -test, citing specific evidence from the obtained output.

➡️ Provide our interpretation based on our stated decision in the context of the problem.

## Post-Hoc Analysis

Conditional on a statistically significant result for the overall -test for the One-Way ANOVA model, a post-hoc analysis should be conducted to investigate which groups differed in terms of their average response and by how much. A post-hoc is only appropriate if the -test was statistically significant.

### Post-Hoc Analysis: Log-transformed survival times

Let’s consider a post-hoc analysis for the survival times of patients with cancer. Note that we can use the exponentiation function in R, exp(), to undo the log-transformation. Mathematically, the exponential function, , is the inverse of the log function, . That is, , where is a numeric constant like .

*Table 9: Individual 95% confidence intervals for average log survival in days by cancer type.*

| Term | 2.5 % | 97.5 % |
| --- | --- | --- |
| OrganBreast | 5.84 | 7.28 |
| OrganBronchus | 4.37 | 5.53 |
| OrganColon | 5.17 | 6.33 |
| OrganOvary | 5.17 | 7.13 |
| OrganStomach | 4.30 | 5.63 |

*Table 10: Pairwise 95% confidence intervals for differences between cancer types in terms of average log survival in days.*

| term | contrast | null.value | estimate | conf.low | conf.high | adj.p.value |
| --- | --- | --- | --- | --- | --- | --- |
| Organ | Bronchus-Breast | 0.00 | -1.61 | -2.91 | -0.30 | 0.01 |
| Organ | Colon-Breast | 0.00 | -0.81 | -2.11 | 0.49 | 0.41 |
| Organ | Ovary-Breast | 0.00 | -0.41 | -2.11 | 1.30 | 0.96 |
| Organ | Stomach-Breast | 0.00 | -1.59 | -2.97 | -0.21 | 0.02 |
| Organ | Colon-Bronchus | 0.00 | 0.80 | -0.36 | 1.95 | 0.31 |
| Organ | Ovary-Bronchus | 0.00 | 1.20 | -0.40 | 2.79 | 0.23 |
| Organ | Stomach-Bronchus | 0.00 | 0.01 | -1.22 | 1.25 | 1.00 |
| Organ | Ovary-Colon | 0.00 | 0.40 | -1.20 | 2.00 | 0.95 |
| Organ | Stomach-Colon | 0.00 | -0.78 | -2.02 | 0.46 | 0.40 |
| Organ | Stomach-Ovary | 0.00 | -1.18 | -2.84 | 0.48 | 0.28 |

➡️ Provide and interpret the confidence interval for the average log survival of patients with breast cancer based on the log transformed output.

➡️ Provide and interpret the confidence interval for the average survival time in days of patients with breast cancer using the log transformed output.

*Note*: in this special case, we need to convert the log survival time back to original units of survival time. **This back-transformation is not needed unless we transformed our original response variable for the model.**

➡️ Based on the post-hoc output provided for the log-transformed model, which groups, if any, differ in terms of their average log survival time at the 5% significance level? Cite specific evidence from the output obtained.

➡️ Provide and interpret the confidence interval for the difference in the survival of patients with breast cancer compared to patients with bronchus cancer in terms of average log survival time.

➡️ Provide and interpret the confidence interval for the difference in the survival of patients with breast cancer compared to patients with bronchus cancer in terms of their average survival time in days.

➡️ What is the other confidence interval for a difference that would be valid to interpret here?