ANOVA\_activity

Mallikarjuna Mannem

2023-10-10

## Install packages

# Load necessary packages  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(broom)  
library(ggfortify)

# 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

# Importing fake GPA data  
simGPA <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/simGPA.csv")

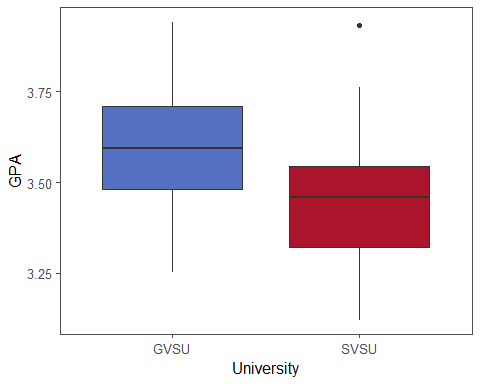
## Rows: 150 Columns: 2  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): University  
## dbl (1): GPA  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# 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 |

We can create side-by-side box plots to compare the GPAs of two the universities

# 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 university

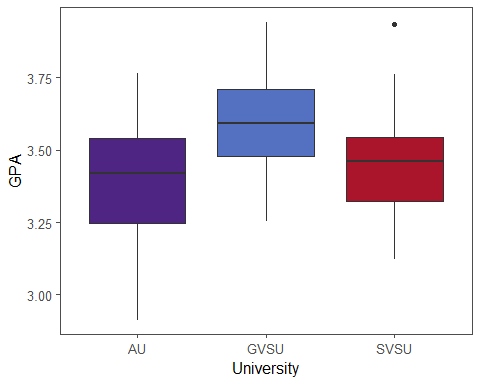
# Calculating descriptive statistics  
quantStats <- simGPA %>%   
 dplyr::filter(University %in% c("GVSU", "SVSU")) %>%   
 group\_by(University) %>%   
 summarize(  
 Minimum = min(GPA, na.rm = TRUE),  
 Q1 = quantile(GPA, na.rm = TRUE, probs = 0.25),  
 M = median(GPA, na.rm = TRUE),  
 Q3 = quantile(GPA, 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 consider the case in which we have a third university we want to compare.

# 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



we can calculate summary statistics

# 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 |

## survival example

let’s first import the data into R

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

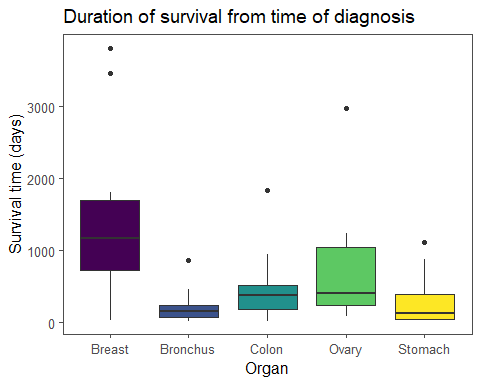
## Rows: 64 Columns: 2  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Organ  
## dbl (1): Survival  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

printing 8 randomly selected rows from the data set

set.seed(1994)  
  
# Printing random rows of data set  
cancerSurvival %>%  
 slice\_sample(n=8) %>%  
 make\_flex()

| 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")



Next, we calculate summary statistics for the survival of each cancer type

# Calculating descriptive statistics  
quant3Stats <- cancerSurvival %>%   
 group\_by(Organ) %>%   
 summarize(  
 Minimum = min(Survival, na.rm = TRUE),  
 Q1 = quantile(Survival, na.rm = TRUE, probs = 0.25),  
 M = median(Survival, na.rm = TRUE),  
 Q3 = quantile(Survival, na.rm = TRUE, probs = 0.75),  
 Maximum = max(Survival, na.rm = TRUE),  
 Mean = mean(Survival, na.rm = TRUE),  
 R = Maximum - Minimum,  
 s = sd(Survival, na.rm = TRUE),  
 n = n()  
)  
  
# Printing table of statistics  
quant3Stats %>%   
 make\_flex(caption = "Summary statistics for survival times in days by cancer type.")

Table 5: Summary statistics for survival times in days 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?**

Breast cancer,1395.91

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

breast cancer since R-value and s-value are more for breast cancer.

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

Shortest duration of survival is for the individuals with 20 days with Bronchus and Colon cancer. Longest is for the individual with breast cancer and survival duration of 3808 days.

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

The group effects formulation of ANOVA model in general is:

Response=Grand Mean +Group Effect +Error Term : is the survival time in days for the jth observation of the ith group $\_i : mean survival time of ith cancer type : independent and normally distributed error terms

###Hypotheses

Our hypotheses are given by

Ho: μ1 = μ2 = μ3 = μ4 = μ5 vs. Ha: μi ≠ μj for some i,j

where μ1 is the average duration of survival in days for breast cancer patients, μ2 for bronchus cancer, μ3 for colon cancer, μ4 for ovarian cancer, and μ5 for stomach cancer.

Alternative statement of hypotheses:

Our hypotheses are given by

Ho: μ1 = μ2 = μ3 = μ4 = μ5 vs. Ha: μi ≠ μj, for some i,j

where μi is the average duration of survival in days for the ith cancer type for i=1,2,…,5.

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

r: survival time in days e: cancer type

# 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

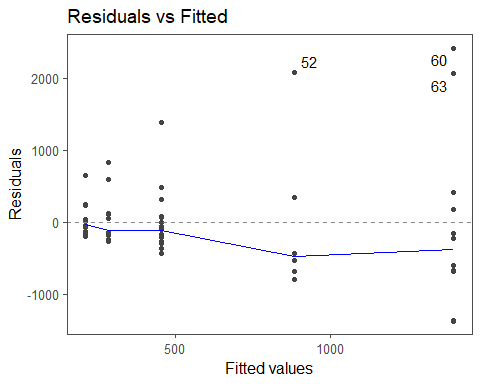
| term | df | sumsq | meansq | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| Organ | 4.00 | 11,535,760.52 | 2,883,940.13 | 6.43 | 0.00023 |
| Residuals | 59.00 | 26,448,144.48 | 448,273.64 | NA | NA |

* SSGroups = 11,535,760.52, this is the total variability in the survival times of the patients explained by their cancer type
* SSE = 26,448,144.48, this is the total variability in the survival times of the patients not explained by their cancer type

### Assumptions

1. **Have the same standard deviation within each group(implies equal variances, also known as homoskedasticity)**

autoplot(survivalModel, which = 1, label.repel = TRUE)@plots[[1]]



There is an outward fanning pattern in the residual by fitted value plot, indicating that the homoskedasticity assumption is violated.

**Ratio of standard deviations**

Smax = 1238.97

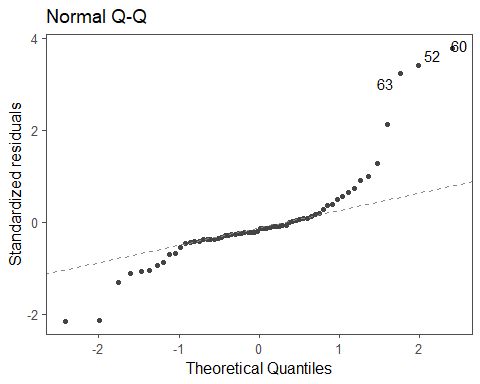
Smin = 209.86

1238.97 / 209.86 = 5.90 which is greater than 2, so this indicates that the homoskedasticity assumption is violated.

1. **Be normally distributed**

**QQ plot**

# Quantile-quantile (QQ) plot  
autoplot(survivalModel, which = 2, label.repel = TRUE)@plots[[1]]



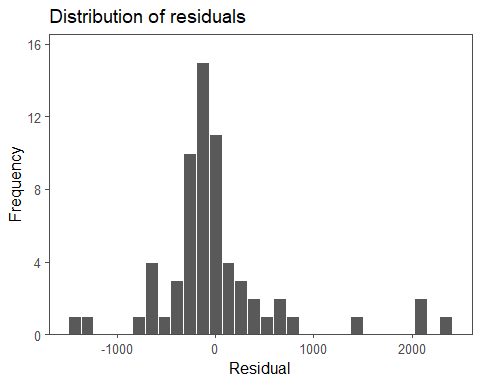
There are many points on the QQ plot that stray far from the diagonal line through the origin, indicating that the normality assumption is violated.

**Histogram of the residuals**

# Histogram of the residuals  
survivalModel %>%   
 broom::augment() %>%   
 ggplot(aes(x = .resid)) +   
 geom\_histogram(color = "white") +  
 scale\_y\_continuous(expand = expansion(mult = c(0, 0.10))) +  
 labs(title = "Distribution of residuals",  
 x = "Residual",  
 y = "Frequency")

## Warning: The `augment()` method for objects of class `aov` is not maintained by the broom team, and is only supported through the `lm` tidier method. Please be cautious in interpreting and reporting broom output.  
##   
## This warning is displayed once per session.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The distribution is right-skewed, indicating that the normality assumption is violated.

1. **Be independent**

since these are measurements for seperate individuals, we will consider this assumption to be sufficiently met.

The assumptions were violated, so truly the results of the ANOVA model are unavailable.

However, for demonstration purpose we will finish conducting the overall F-test.

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

**Test statistic**: F = 6.43

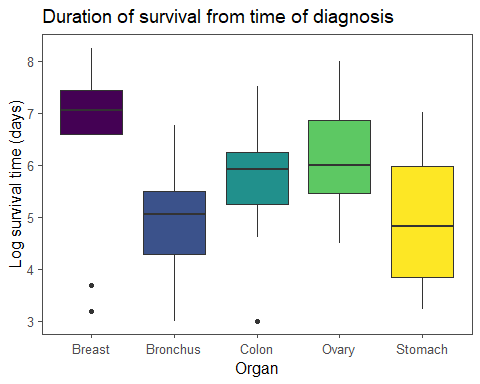
**p-value**: 0.00023

**Decision**: Reject Ho since the p-value is less than 0.05

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

Since the p-value is less than 0.05, we have sufficient evidence that at least two cancer types differ in their average survival time in days at the 5% significance level.

# 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*

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

Table 7: ANOVA table for log 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 |

**Checking assumptions for the log transformation model**

# 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 |

## Checking assumptions

**Homoskedasticity**

since 1.65/0.95 = 1.736842 is less than 2, and since the vertical spread is fairly constant in the residual by fitted value plot, the homoskedasticity assumption is met for the log-transformed model.

**normality**

since the points in the QQ plot do not stray far from the diagonal line, this indicates that the normality assumuption is met for the log-transformed model.

Moreover, since the histogram unimodal and only very slightly left-skewed, this together with the QQ plot indicates that the normality assumption is met for the log-transformed model.

Overall, the assumptions are now met, so the results of the ANOVA model for the log-transformed values are reliable.

**Test statistic**: F = 4.29

**p-value**: 0.0041

**Decision**: Reject Ho since the p-value is less than 0.05

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

Since the p-value is less than 0.05, we have sufficient evidence that at least two cancer types differ in their average survival time in days at the 5% significance level.

## Post-Hoc Analysis

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

confidence limits: (5.84,7.28)

We are 95% confident that the mean log survival time in days for breast cancer patients is between 5.84 and 7.28

**A better interpretation after having used a log-transformation**

We are 95% confident that the mean survival time in days for breast cancer patients is between 344 and 1451 days.

Next we obtain post-havoc output to compare pairs of groups

# Conducting post-hoc analysis using Tukey's HSD for pairwise differences  
pairwisePostHoc <- TukeyHSD(logSurvivalModel)  
  
# Printing post-hoc output  
pairwisePostHoc %>%   
 broom::tidy() %>%   
 make\_flex(caption = "Pairwise 95% confidence intervals for differences between cancer types in terms of average log survival in days.")

Table 9: 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 |

*From row 1):*

we have sufficient evidence that breast cancer patients have a longer log survival time in days on average, than bronchus cancer patients.

c limits: (-2.91, -0.30)

we are 95% confident that the mean log duration of survival in days for bronchus cancer patients is between 0.30 and 2.91 less than that of breast cancer patients

*From row 4):*

we have sufficient evidence that breast cancer patients have a longer log survival time in days on average, than stomach cancer patients.

C limits: (-2.97, -0.21)

we are 95% confident that the mean log duration of survival in days for stomach cancer patients is between 0.21 and 2.97 less than that of breast cancer patients.

**we will not have to do this back-transformation on exam**

**From row 1):**

we have sufficient evidence that breast cancer patients have a longer survival time in days on average, than bronchus cancer patients.

ca limits: log-scale (2.91, 0.30) —> —> (18.356799, 1.349859)

we are 95% confident that the mean duration of survival in days for breast cancer patients is between 1.3 and 18.3 times/ that of bronchus cancer patients.