

# Analysis of Scurvy Therapies from Medicaldata

KOMAL BHOSLE

## RStudio Link

<https://posit.cloud/spaces/603138/content/10306018>

## Library Calls

```
install.packages("medicaldata")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
install.packages("gt")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
install.packages("gtsummary")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
install.packages("ggpubr")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
install.packages("easystats")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
install.packages("openintro")
```

Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.4'  
(as 'lib' is unspecified)

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

```
library(tidyverse)
```

-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --

v forcats	1.0.0	v readr	2.1.5
v ggplot2	3.5.1	v stringr	1.5.1
v lubridate	1.9.3	v tibble	3.2.1
v purrr	1.0.2	v tidyr	1.3.1

-- Conflicts ----- tidyverse\_conflicts() --

x dplyr::filter() masks stats::filter()

x dplyr::lag() masks stats::lag()

i Use the conflicted package (<<http://conflicted.r-lib.org/>>) to force all conflicts to become

```
library(medicaldata)
library(gt)
library(gtsummary)
library(ggpubr)
library(easystats)
```

```
# Attaching packages: easystats 0.7.4
✓ bayestestR 0.15.3   ✓ correlation 0.8.7
✓ datawizard 1.0.2   ✓ effectsize 1.0.0
✓ insight 1.2.0     ✓ modelbased 0.11.0
✓ performance 0.13.0 ✓ parameters 0.25.0
✓ report 0.6.1     ✓ see 0.11.0
```

```
library(openintro)
```

```
Loading required package: airports
Loading required package: cherryblossom
Loading required package: usdata
```

```
Attaching package: 'openintro'
```

```
The following object is masked from 'package:gt':
```

```
sp500
```

```
library(DescTools)
```

```
Attaching package: 'DescTools'
```

```
The following object is masked from 'package:openintro':
```

```
cards
```

## Data Preparation

The Scurvy Therapy data set was from the medicaldata package of a clinical trial which was conducted by James Lind in 1747. The data set includes 12 sailors who participated in scurvy treatment trial. The aim was to assess the effectiveness of different treatments

on symptoms of scurvy. The variables of the data set are Treatment, Dosage, Gum\_Rot, Skin\_sores, Knee\_Weakness, Lassitude, Fit\_for\_Duty. To prepare the data set for analysis data wrangling was done by initially renaming the variables, and then changed the character and factor form of Study\_id and Treatment variables. Then the disease symptom factors were recoded to better applicable forms as None, Mild, Moderate and Severe, which were then arranged according to their level of effectiveness.

```
data("scurvy",
      package = "medicaldata")

scurvy_therapy <-
  scurvy |>
  rename(
    'Study_id' = study_id,
    'Treatment' = treatment,
    'Dosage' = dosing_regimen_for_scurvy,
    'Gum_Rot' = gum_rot_d6,
    'Skin_sores' = skin_sores_d6,
    'Knee_Weakness' = weakness_of_the_knees_d6,
    'Lassitude' = lassitude_d6,
    'Fit_for_Duty' = fit_for_duty_d6
  ) |>
  mutate(
    Study_id = as.character(Study_id),
    Treatment = as.factor(Treatment),
    Gum_Rot = recode(
      Gum_Rot,
      '0_none' = 'None',
      '1_mild' = 'Mild',
      '2_moderate' = 'Moderate',
      '3_severe' = 'Severe'
    ),
    Gum_Rot = factor(
      x = Gum_Rot,
      levels = c(
        "None",
        "Mild",
        "Moderate",
        "Severe"
      )
    ),
    Skin_sores = recode(
      Skin_sores,
```

```

    '0_none' = 'None',
    '1_mild' = 'Mild',
    '2_moderate' = 'Moderate',
    '3_severe' = 'Severe'
  ),
  Skin_sores = factor(
    x = Skin_sores,
    levels = c(
      "None",
      "Mild",
      "Moderate",
      "Severe"
    )
  ),
  Knee_Weakness = recode(
    Knee_Weakness,
    '0_none' = 'None',
    '1_mild' = 'Mild',
    '2_moderate' = 'Moderate',
    '3_severe' = 'Severe'
  ),
  Knee_Weakness = factor(
    x = Knee_Weakness,
    levels = c(
      "None",
      "Mild",
      "Moderate",
      "Severe"
    )
  ),
  Lassitude = recode(
    Lassitude,
    '0_none' = 'None',
    '1_mild' = 'Mild',
    '2_moderate' = 'Moderate',
    '3_severe' = 'Severe'
  ),
  Lassitude = factor(
    x = Lassitude,
    levels = c(
      "None",
      "Mild",

```

```

    "Moderate",
    "Severe"
  )
),
Fit_for_Duty = recode(
  Fit_for_Duty,
  '0_no' = 'No',
  '1_yes' = 'Yes'
),
Fit_for_Duty = factor(
  x = Fit_for_Duty,
  levels = c(
    "No",
    "Yes"
  )
)
)
)

```

## Univariate Analysis

Univariate analysis is to explore each variable and analyze them. To understand how often each category occurs in the data, we use frequency distribution table. In this table I have included the categories Gum\_Rot, Skin\_sores, Knee\_Weakness, Lassitude, Fit\_for\_Duty. The Treatment variable is the comparing category to all these symptom categories. The `tbl_summary` function is used to create frequency distribution table. In my method of data visualization, here I have used Bar plot and Dot plot to visualize the frequency of responses for single variable.

## Frequency Distribution Table

```

scurvy_therapy |>
tbl_summary(
  include = c(
    Gum_Rot,
    Skin_sores,
    Knee_Weakness,
    Lassitude,
    Fit_for_Duty
  ),
  by = "Treatment"
)

```

```

) |>
bold_labels() |>
italicize_levels() |>
as_gt() |>
tab_source_note(
  source_note = md(
    "Data from the `Scurvy` dataset (in medicaldata package)."
```

```

  )
) |>
tab_header(
  title = md(
    "**Symptom Rating and Fitness by Treatment Group**"
```

```

  ),
  subtitle = md("Based on 12 participants in the scurvy trial")
)
```

## First Distribution Plot

```

scurvy_therapy |>
  ggplot(
    mapping = aes(
      x = Gum_Rot,
      fill = Gum_Rot
    )
  ) +
  geom_bar(
    position = position_dodge()
  ) +
  labs(
    x = "Gum Rot Severity",
    fill = "Gum Rot",
    title = "Distribution of Gum Rot Severity Among Participants"
  ) +
  theme_classic()
```

## Symptom Rating and Fitness by Treatment Group

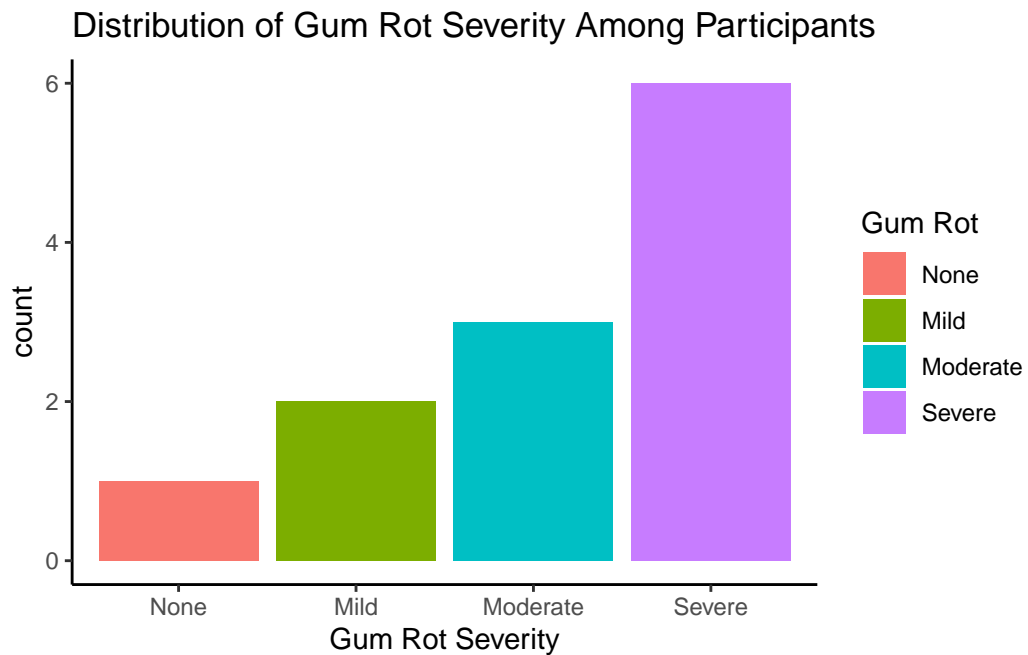
Based on 12 participants in the scurvy trial

Characteristic	cider N = 2 <sup>1</sup>	citrus N = 2 <sup>1</sup>	dilute_sulfuric_acid N = 2 <sup>1</sup>	purgative_m...
<b>Gum_Rot</b>				
<i>None</i>	0 (0%)	1 (50%)	0 (0%)	0 (0)
<i>Mild</i>	0 (0%)	1 (50%)	1 (50%)	0 (0)
<i>Moderate</i>	2 (100%)	0 (0%)	1 (50%)	0 (0)
<i>Severe</i>	0 (0%)	0 (0%)	0 (0%)	2 (10)
<b>Skin_sores</b>				
<i>None</i>	0 (0%)	1 (50%)	0 (0%)	0 (0)
<i>Mild</i>	1 (50%)	1 (50%)	0 (0%)	0 (0)
<i>Moderate</i>	1 (50%)	0 (0%)	0 (0%)	0 (0)
<i>Severe</i>	0 (0%)	0 (0%)	2 (100%)	2 (10)
<b>Knee_Weakness</b>				
<i>None</i>	0 (0%)	2 (100%)	0 (0%)	0 (0)
<i>Mild</i>	0 (0%)	0 (0%)	0 (0%)	0 (0)
<i>Moderate</i>	2 (100%)	0 (0%)	0 (0%)	0 (0)
<i>Severe</i>	0 (0%)	0 (0%)	2 (100%)	2 (10)
<b>Lassitude</b>				
<i>None</i>	0 (0%)	1 (50%)	0 (0%)	0 (0)
<i>Mild</i>	0 (0%)	1 (50%)	0 (0%)	0 (0)
<i>Moderate</i>	1 (50%)	0 (0%)	0 (0%)	0 (0)
<i>Severe</i>	1 (50%)	0 (0%)	2 (100%)	2 (10)
<b>Fit_for_Duty</b>	0 (0%)	1 (50%)	0 (0%)	0 (0)

<sup>1</sup>n (%)

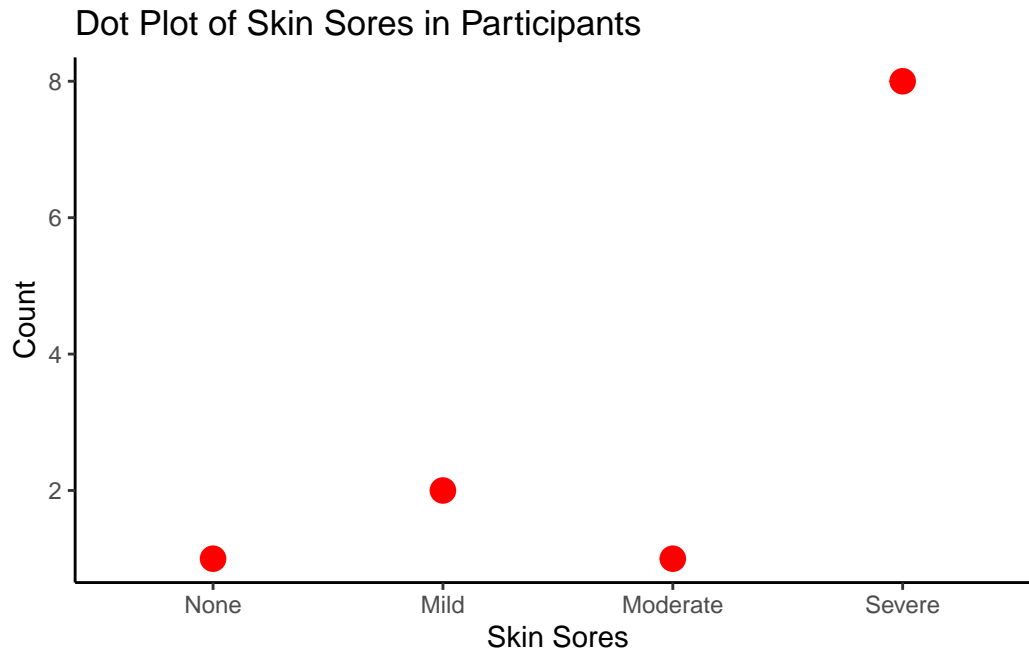
Data from the **Scurvy** dataset (in **medicaldata** package).





### Second Distribution Plot

```
scurvy_therapy |>
  count(Skin_sores) |>
  ggplot(
    mapping = aes(
      x = Skin_sores,
      y = n
    )
  ) +
  geom_point(
    size = 4,
    color = "red"
  ) +
  labs(
    x = "Skin Sores",
    y = "Count",
    title = "Dot Plot of Skin Sores in Participants"
  ) +
  theme_classic()
```



### Insights

The frequency distribution table gives overview of how participants are distributed across different categories. The first distribution plot the bar plot was used to visualize the frequency of variable of a single category. Here Gum\_Rot was the chosen variable and the bar plot represents that there were 6 participants who had severe Gum\_Rot where as there were 2 and 3 participants with mild and moderate Gum\_Rot respectively. In the second distribution plot which is dot plot for Skin sores variable represents the frequency of Skin sores in participants, which shows that there were 1 participant with no Skin sore and 1 with moderate skin sores, where as 2 participants had mild skin sores while 8 were found to have severe Skin sores.

### Bivariate Analysis

In Bivariate analysis we examine the relation between two variable at a time. This can be done by contingency tables, measures of association and bivariate plots. Contingency table helps to summarize the relationship between two categorical variables. For the scurvy\_therapy data set, we selected two pairs of variables- Fir\_for\_Duty vs Treatment(Dichotomous-Dichotomous) and Gum\_Rot vs Lassitude (Nominal-Nominal). The measure of association for each of these contingency tables are Odds Ratio and Goodman-Kruskal Gamma respectively. The bivariate plots selected were Bar plot for Fir\_for\_Duty vs Treatment, and Bubble plot for Gum\_Rot vs Lassitude.

## First Contingency Table

```
scurvy_treatment<-
  scurvy_therapy |>
  mutate(
    Treatment = if_else(Treatment == "citrus", "Citrus", "Other")
  )

scurvy_therapy_table <-
  scurvy_treatment |>
  tbl_cross(
    row = Fit_for_Duty,
    col = Treatment,
    percent = "row",
    label = list(
      Fit_for_Duty = "Fit for Duty",
      Treatment = "Treatment Received"
    )
  ) |>
  bold_labels() |>
  as_gt() |>
  tab_source_note(
    source_note = md(
      "Data from the `Scurvy` dataset(in **medicaldata** package)"
    )
  ) |>
  tab_header(
    title = md(
      "**2x2 Contingency Table of Scurvy Treatment by Fit for Duty**"
    ),
    subtitle = md(
      "Based on 12 participants in Scurvy Trial"
    )
  )
  )

scurvy_therapy_table
```

## 2x2 Contingency Table of Scurvy Treatment by Fit for Duty

Based on 12 participants in Scurvy Trial

	Treatment Received		Total
	Citrus	Other	
<b>Fit for Duty</b>			
No	1 (9.1%)	10 (91%)	11 (100%)
Yes	1 (100%)	0 (0%)	1 (100%)
<b>Total</b>	2 (17%)	10 (83%)	12 (100%)

Data from the Scurvy dataset(in **medicaldata** package)

## Second Contingency Table

```
scurvy_therapy |>
  tbl_cross(
    row = Gum_Rot,
    col = Lassitude,
    percent = "row",
    label = list(
      Gum_Rot ~ "Gum Rot",
      Lassitude ~ "Lassitude"
    ),
  ) |>
  bold_labels() |>
  italicize_levels() |>
  as_gt() |>
  tab_source_note(
    source_note = md(
      "Data from the `Scurvy` dataset(in medicaldata package)"
    )
  ) |>
  tab_header(
    title = md(
      "**2x2 Contingency Table of Gum Rot vs. Lassitude**"
    ),
    subtitle = md(
      "Based on 12 participants in Scurvy Trial"
    )
  )
```

## 2x2 Contingency Table of Gum Rot vs. Lassitude

Based on 12 participants in Scurvy Trial

	Lassitude				Total
	<i>None</i>	<i>Mild</i>	<i>Moderate</i>	<i>Severe</i>	
<b>Gum Rot</b>					
<i>None</i>	1 (100%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)
<i>Mild</i>	0 (0%)	1 (50%)	0 (0%)	1 (50%)	2 (100%)
<i>Moderate</i>	0 (0%)	0 (0%)	1 (33%)	2 (67%)	3 (100%)
<i>Severe</i>	0 (0%)	0 (0%)	0 (0%)	6 (100%)	6 (100%)
<b>Total</b>	1 (8.3%)	1 (8.3%)	1 (8.3%)	9 (75%)	12 (100%)

Data from the **Scurvy** dataset(in **medicaldata** package)

### Measures of Association

```
scurvy_fit_treatment <-  
  table(  
    scurvy_treatment$Fit_for_Duty,  
    scurvy_treatment$Treatment  
  )  
  
oddsratio(  
  x = scurvy_fit_treatment + 0.5  
)
```

Odds ratio	95% CI
-----	-----
0.05	[0.00, 1.79]

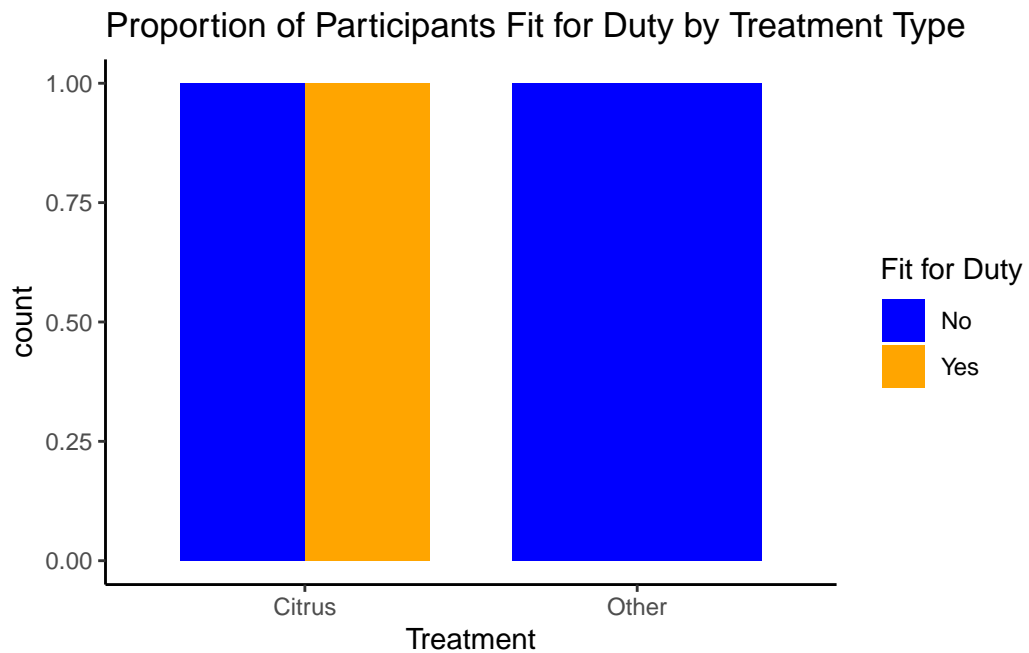
```
scurvy_gum_lassitude <-  
  table(  
    scurvy_therapy$Gum_Rot,  
    scurvy_therapy$Lassitude  
  )  
  
GoodmanKruskalGamma(  
  x = scurvy_gum_lassitude,  
  conf.level = 0.95  
)
```

```
) |>
  round(
    digits = 2
  )
```

```
gamma lwr.ci upr.ci
0.93   0.74   1.00
```

## First Bivariate Plot

```
scurvy_treatment |>
  group_by(
    Fit_for_Duty,
    Treatment
  ) |>
  summarise(
    n = n(),
    .groups = "drop"
  ) |>
  group_by(Treatment) |>
  ggplot(
    mapping = aes(
      x = Treatment,
      fill = Fit_for_Duty
    )
  ) +
  geom_bar(
    position = position_dodge(),
    width = 0.75
  ) +
  scale_fill_manual(
    values = c("blue", "orange")
  ) +
  labs(
    x = "Treatment",
    fill = "Fit for Duty",
    title = "Proportion of Participants Fit for Duty by Treatment Type"
  ) +
  theme_classic()
```



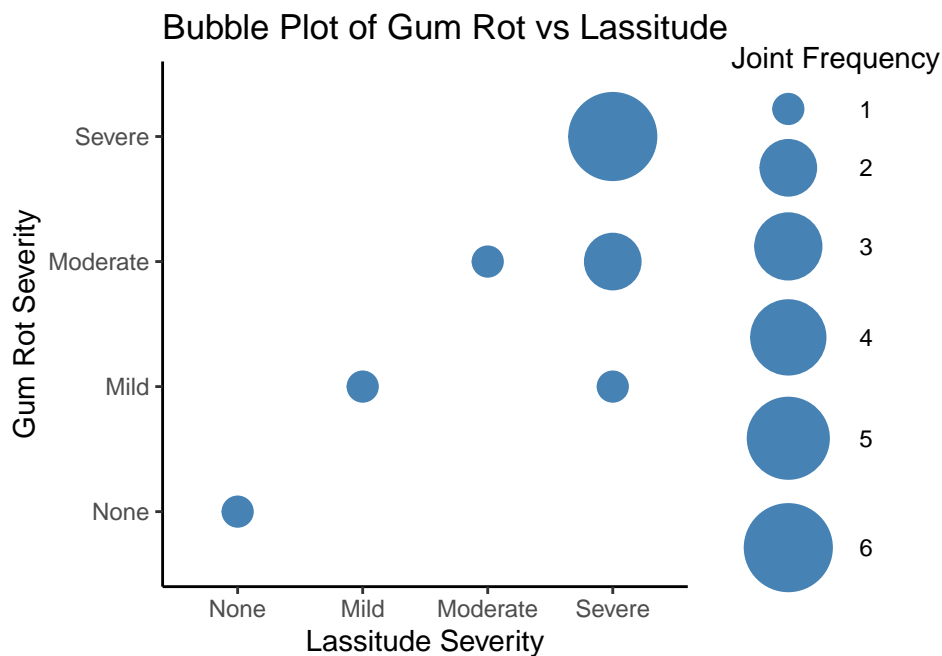
## Second Bivariate Plot

```
scurvy_therapy |>
  group_by(
    Gum_Rot,
    Lassitude
  ) |>
  summarize(
    n = n(),
    .groups = "drop"
  ) |>
  ggplot(
    mapping = aes(
      x = Lassitude,
      y = Gum_Rot,
      size = n
    )
  ) +
  geom_point(
    color = "steelblue"
  ) +
  scale_size_continuous(
```

```

    range = c(5, 15)
  ) +
  coord_fixed() +
  labs(
    x = "Lassitude Severity",
    y = "Gum Rot Severity",
    size = "Joint Frequency",
    title = "Bubble Plot of Gum Rot vs Lassitude"
  ) +
  theme_classic()

```



## Insights

The 2x2 contingency table showed how participants fitness for duty was effected based on whether they received citrus treatment or not. Most participants who received citrus treatment were fit for duty while those with other treatments were not fit. There was a pattern of co occurrence between scurvy symptoms in second contingency table where participants with severe gum rot has moderate to severe lassitude and if there was no or mild gum rot , it was likely to have no or mild lassitude. Odds ratio indicates that there is strong positive association between receiving citrus treatment and being fit fir duty. The Goodman-Kruskal Gamma of 0.93 shows positive association between Gum Rot Severity and Lassitude. The dodge bar plot of Fit for Duty and treatment type indicates that among the participants



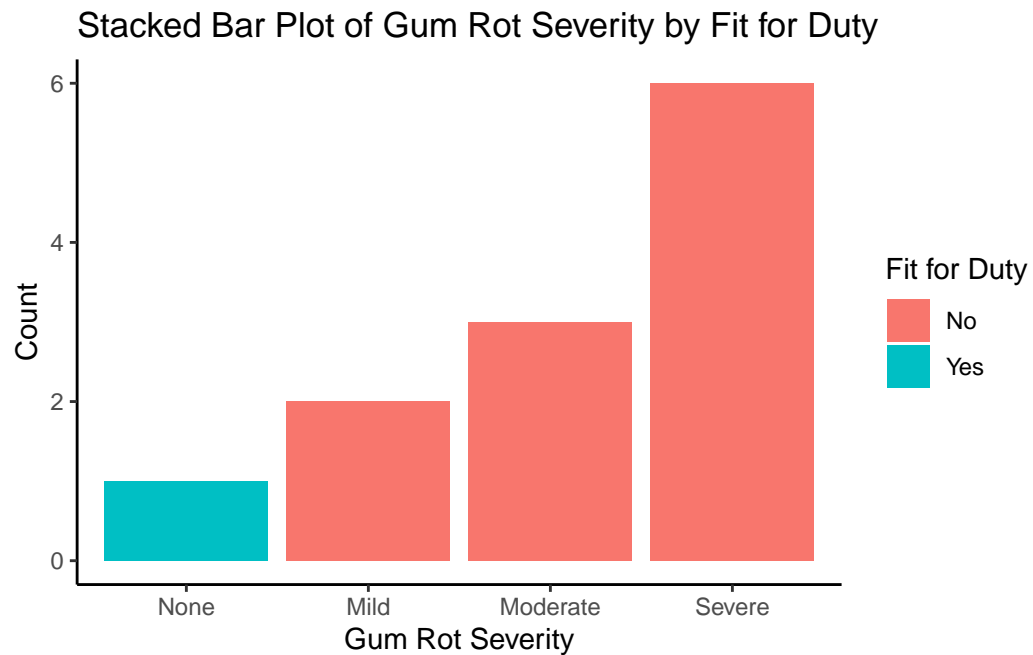
who were receiving citrus treatment were majorly Fit for Duty where as those who received other treatments were not fit for duty. The bubble plot for Gum\_Rot and Lassitude show a co-occurrence of symptoms, as we see in the bubble plot where bubble at severe-severe is the largest which is suggestive of symptoms of both variables showing association as observed from the Goodman-Kruskal Gamma.

## Multivariate Analysis

Multivariate analysis examines how the multiple variables interact between two variables. The first multivariate plot is the Stacked bar plot of Gum\_Rot and Fit\_for\_Duty variables with position stack. The second multivariate plot is scatter plot with facet\_grid of Knee Weakness vs Lassitude variable to that of Fitness for duty of participants. The function geom\_point is used to create scatter plot.

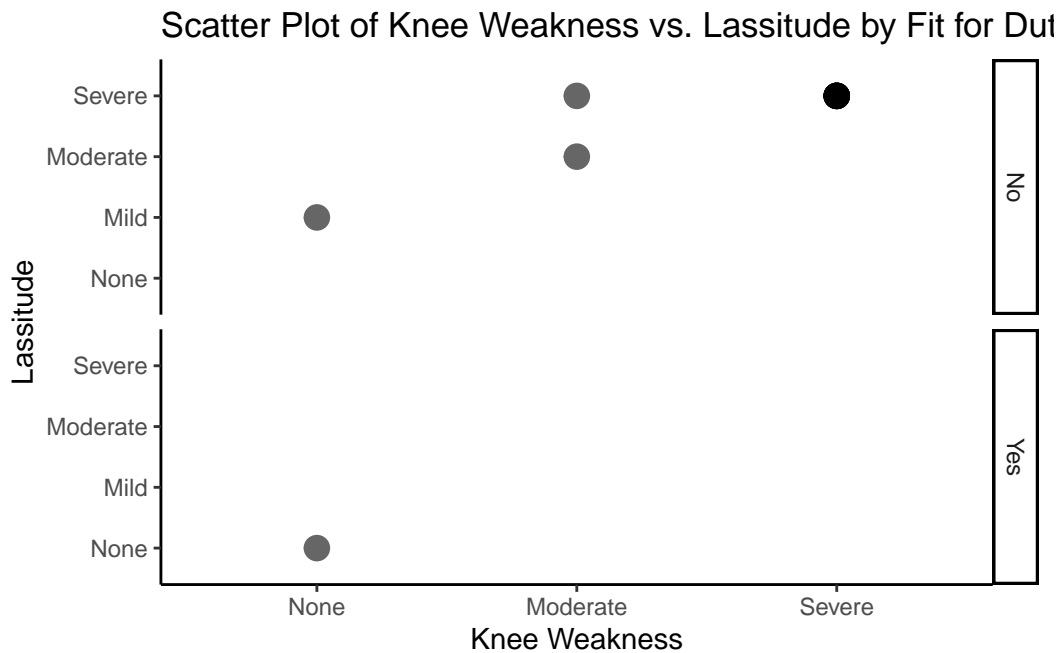
### First Multivariate Plot

```
scurvy_therapy |>
  ggplot(
    mapping = aes(
      x = Gum_Rot,
      fill = Fit_for_Duty
    )
  ) +
  geom_bar(
    position = "stack"
  ) +
  labs(
    x = "Gum Rot Severity",
    y = "Count",
    fill = "Fit for Duty",
    title = "Stacked Bar Plot of Gum Rot Severity by Fit for Duty"
  ) +
  theme_classic()
```



### Second Multivariate Plot

```
scurvy_therapy |>
  ggplot(
    mapping = aes(
      x = Knee_Weakness,
      y = Lassitude
    )
  ) +
  geom_point(
    size = 4,
    alpha = 0.6,
  ) +
  facet_grid(
    rows = vars(Fit_for_Duty)
  ) +
  labs(
    x = "Knee Weakness",
    y = "Lassitude",
    title = "Scatter Plot of Knee Weakness vs. Lassitude by Fit for Duty"
  ) +
  theme_classic()
```



### Insights

The stacked bar plot of Gum\_Rot and Fit for Duty highlighted how the increase in severity of Gum\_Rot will have impact on participants ability to be Fit\_for\_Duty. This shows a likely negative impact on Gum rot. The Scatter plot of Knee Weakness and Lassitude shows that if there is more severe weakness with severe lassitude is an indication of participant not being fit for duty. Where as individuals who are fit for duty would be experiencing mild symptoms for both Knee Weakness and Lassitude.

### Extra Credit

#### Bivariate Plot

The Bivariate plot has variables of Gum\_Rot and Knee Weakness which are from Scurvy\_therapy data set. A 2x2 contingency table is created followed by the Goodman-Kruskal Gamma Measure of Association. The Bivariate plot of Grouped bar plot is created.

```
Scurvy_therapy_table <-
scurvy_therapy |>
tbl_cross(
  row = Gum_Rot,
```

## 2x2 Contingency Table of Gum Rot vs. Knee Weakness

Based on 12 participants in Scurvy Trial

	Knee__Weakness				Total
	<i>None</i>	<i>Mild</i>	<i>Moderate</i>	<i>Severe</i>	
<b>Gum Rot</b>					
<i>None</i>	1 (100%)	0 (0%)	0 (0%)	0 (0%)	1 (100%)
<i>Mild</i>	1 (50%)	0 (0%)	0 (0%)	1 (50%)	2 (100%)
<i>Moderate</i>	0 (0%)	0 (0%)	2 (67%)	1 (33%)	3 (100%)
<i>Severe</i>	0 (0%)	0 (0%)	0 (0%)	6 (100%)	6 (100%)
<b>Total</b>	2 (17%)	0 (0%)	2 (17%)	8 (67%)	12 (100%)

Data from the **Scurvy** dataset(in **medicaldata** package)

```
col = Knee_Weakness,
percent = "row",
label = list(
  Gum_Rot ~ "Gum Rot",
  Knee_Weakness ~ "Knee_Weakness"
),
) |>
bold_labels() |>
italicize_levels() |>
as_gt() |>
tab_source_note(
  source_note = md(
    "Data from the `Scurvy` dataset(in **medicaldata** package)"
  )
) |>
tab_header(
  title = md(
    "**2x2 Contingency Table of Gum Rot vs. Knee Weakness**"
  ),
  subtitle = md(
    "Based on 12 participants in Scurvy Trial"
  )
)
```

Scurvy\_therapy\_table

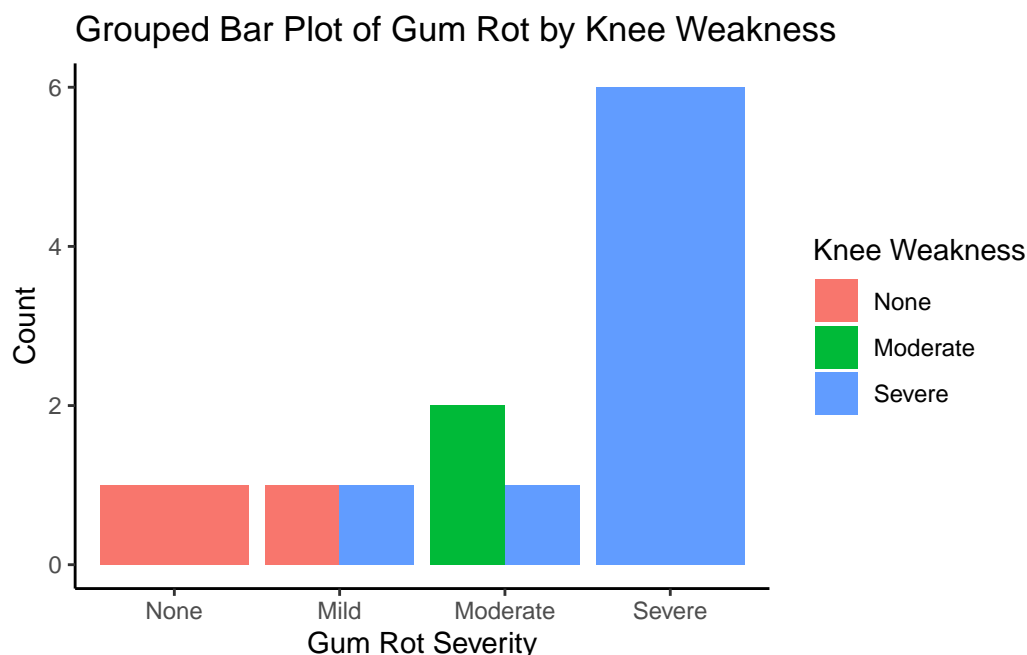
## Measure of Association

```
scurvy_gum_knee_Weakness <-  
  table(  
    scurvy_therapy$Gum_Rot,  
    scurvy_therapy$Knee_Weakness  
  )  
  
GoodmanKruskalGamma(  
  x = scurvy_gum_knee_Weakness,  
  conf.level = 0.95  
) |>  
  round(  
    digits = 2  
  )
```

```
gamma lwr.ci upr.ci  
0.88   0.62   1.00
```

## Bivariate Plot

```
scurvy_therapy |>  
  ggplot(  
    mapping = aes(  
      x = Gum_Rot,  
      fill = Knee_Weakness  
    )  
  ) +  
  geom_bar(  
    position = "dodge"  
  ) +  
  labs(  
    x = "Gum Rot Severity",  
    y = "Count",  
    fill = "Knee Weakness",  
    title = "Grouped Bar Plot of Gum Rot by Knee Weakness"  
  ) +  
  theme_classic()
```



### Insights

The contingency table of Gum\_Rot and Knee\_Weakness shows that it has positive association between severity of two conditions. The Goodman-Kruskal Gamma shows strong correlation with gamma 0.88, which implies that as the severity of Gum Rot increases, the severity of Knee Weakness also increases. The visualization by grouped bar plot between Gum rot severity and presence or absence of Knee Weakness indicates that with the increase in the severity of Gum Rot among participants, there is also increased severity of Knee Weakness. These symptoms are suggestive of occurring simultaneously.

### References

- 30 ggplot basics – The Epidemiologist R Handbook. (2020). Epirhandbook.com. [https://epirhandbook.com/en/new\\_pages/ggplot\\_basics.html](https://epirhandbook.com/en/new_pages/ggplot_basics.html)
- Kabacoff, R. (n.d.). Chapter 2 Data Preparation | Modern Data Visualization with R. rkabacoff.github.io. <https://rkabacoff.github.io/datavis/DataPrep.html#importing>
- Kabacoff, R. (2025). Chapter 4 Univariate Graphs | Modern Data Visualization with R. Github.io. <https://rkabacoff.github.io/datavis/Univariate.html#quantitative>
- Kabacoff, R. (2025). Chapter 5 Bivariate Graphs | Modern Data Visualization with R. Github.io. <https://rkabacoff.github.io/datavis/Bivariate.html#quantitative-vs.-quantitative>

Nahhas, R. W. (2024, June 25). Chapter 5 Descriptive statistics | An Introduction to R for Research. Bookdown.org. <https://bookdown.org/rwnahhas/IntroToR/descriptives.html>