TU Dortmund
Department of Statistics

**Case Studies I**

# Project 2: Comparison of K distributions

## Ants on the picnic blanket

Lecturers:
Prof. Dr. Paul Bürkner
Dr. Daniel Habermann
Lars Kühmichel

Author: Rahul Ramesh Vishwkarma
Matriculation Number.: 236862

Group Number.: 17

Group members: Md Jubirul Alam, Rafay Maqsood,
Syed Hassan Saqlain Tayyab

May 14, 2024

# Contents

# 1 Introduction

On holidays and sunny days at the park, the presence of ants can pose a challenge for outdoor picnickers, who often bring sandwiches to their picnic spot. One might wonder about the attraction of ants to certain types of bread, different toppings, and whether the butter is on top or not in their sandwiches. By understanding ants' preferences regarding bread types, toppings, and the existence of butter, one can take precautions against ants by avoiding such breads, toppings, or the use of butter. This may prevent their picnic from being ruined by ant infestations.

First, the assumptions of the ANOVA test are checked through descriptive statistical tools and the data collection method. Then, an ANOVA test is applied to the number of ants, in different groups formed by the types of bread in a sandwich to identify if the bread types affect the average number of ants attracted to it. Similarly, the same ANOVA test is conducted for the groups formed by different sandwich toppings to know whether the different toppings make a difference in the number of ants attracted to them on average or not. Secondly, the multiple testing problem is discussed by conducting a pairwise t-test for a different group variable or factor and then discussing how to solve such a problem.

The second section includes the description dataset and variables. In the third section, some statistical methods for hypothesis testing problems are discussed. In the fourth section, descriptive and inferential statistical tools are used to test the claim about the parameters of two groups of a population. Finally, the results from the analysis are summarized in the fifth section.

# 2 Problem Statement

## 2.1 Dataset and its quality

The given dataset is `SandwichAnts2` from the package Lock5Data (Robin Lock [aut (2021)). It includes 48 observations and four variables. Three variables are categorical, namely `Bread`, `Topping` and `Butter`, and one is a discrete variable, `Ants`. The dataset

`SandwichAnts2` is compiled as `Sandwich.sav` in `.sav` file. The data were collected by a university student, Dominic Kelly, by conducting an experiment. The experiment is conducted by counting the ants in the glass jar, which is placed over a piece of sandwich. The piece of sandwich is chosen randomly, and he also performed the experiment with all combinations of sandwiches with different types of bread, different toppings, and with or without butter, until he had two samples for each combination. There are no missing values in the dataset. However, the counting of ants may be incorrect as ants were wandering, so some of them may be missed or counted multiple times.

## 2.2 Explanation of variables

The variables `Bread`, `Topping` and `Butter` are categorical variables (or factors). `Bread` has four levels, which represent the types of bread, i.e. 'multigrain', 'rye', 'white', and 'wholemeal'. Similarly, `Topping` has three levels that correspond to sandwich toppings, i.e., 'ham pickles', 'peanut butter' and 'ham and pickles'. Then, `Butter` represents the existence of butter by levels 'with' and 'without' for butter being on the top of the sandwich or not, respectively. The dataset also contains a count variable, `Ants` which takes values in $\mathbb{N}_0$.

## 2.3 Goal of this project

The aim of this project is to first analyse if ants are attracted significantly by certain kinds of bread, different toppings, or the presence of butter. Then verifying the assumption of ANOVA and apply it to different groups formed by factors, `Bread` and `Topping`. Next, analyse the pairwise differences, in which we test whether two means of a response variable within two groups are equal or not.

# 3 Statistical Methods

## 3.1 Hypothesis Testing

A null hypothesis is a statement or a claim about a population parameter and denoted as $H_0$ (Paul Newbold (2019)). The $H_0$ is then maintained or rejected by a decision rule based on a test statistic. The test statistic is calculated using sample data. When $H_0$ is rejected, we are left with an alternative hypothesis and denoted by $H_1$. e.g., two-sided hypothesis testing for the mean of a distribution

$$H_0 : \mu = 16 \quad vs \quad H_1 : \mu \neq 16. \tag{1}$$

Here, $H_0$ claims the mean of the distribution is 16. There is also one-sided testing, which can be stated using a different alternative hypothesis as below;

$$H_0 : \mu = 16 \quad vs \quad H_1 : \mu > 16. \tag{2}$$

As the data is randomly sampled, our decision can be wrong about the hypotheses. The error caused by rejecting $H_0$ when it is true is known as an $Type - I$ error, and the probability of making an $Type - I$ error is denoted by $\alpha$. The $Type - II$ error is made when we do not reject $H_0$ when it is indeed false, and the probability of making $Type - II$ the error is denoted by $\beta$. The $\alpha$ is always considered to be the severe one, and we make our decision rule while keeping $\alpha$ to be some small constant value. The $\alpha$ is also known as *the significance level* and the popular choices for it are 0.05 or 0.01, but can vary and depend on the application. Alternatively, the decision rule can also be made using $p$-value. The $p$-value is defined as the probability of obtaining the value of the test statistic as or more extreme than the actual value under $H_0$.

## 3.2 Two sample t-test for equal unknown variances

Suppose we have two independent samples of size $n_1$ and $n_2$ of two random variables $X \sim \mathcal{N}(\mu_X, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$. Then, the test statistic $T$ follows $t$-distribution with $(n_1 + n_2 - 2)$ degrees of freedom under $H_0$, and given by (Michael Schomaker (2017)),

$$T = \frac{\bar{X} - \bar{Y}}{S}\sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \sim t_{(n_1+n_2-2)}. \tag{3}$$

Where $S_X^2$ and $S_Y^2$ are sample variances in the first sample of size $n_1$ and the second sample of size $n_2$, respectively. $S^2$ is the pooled sample variance computed as,

$$S^2 = \frac{(n_1 - 1)S_X^2 - (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}. \tag{4}$$

Let $t_{n_1+n_2-2,1-\frac{\alpha}{2}}$ be the $\frac{\alpha}{2}^{th}$ quantile of $t$-distribution, i.e., $P(T \geq t_{n_1+n_2-2,\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $T$-value be the value of the statistic $T$ computed from the sample, then the decision rule with $p$-value $= P(|T| \geq T\text{-value})$ at $\alpha$ is defined as below,

$$\text{Reject } H_0 : \text{ if } p\text{-value} < \alpha. \tag{5}$$

## 3.3 One-way Analysis of Variance (ANOVA)

Suppose we have $K$ independent random samples of size $n_1, n_2, ..., n_K$, and these samples are from $K$ populations with the total number of observations being $n$ (Paul Newbold (2019)). Let $x_{ij}$ be the $j$th observation in the $i$th group, $\bar{x_i}$ be the sample mean for the $i$th group, and $\bar{x}$ be the overall mean. Then, to measure variability, we calculate the sum of squares and mean sum of squares as follows:

$$\text{Within groups : } SSW = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2, \qquad MSW = \frac{SSW}{n-K}, \tag{6}$$

$$\text{Between groups : } SSG = \sum_{i=1}^{K} n_i(\bar{x_i} - \bar{x})^2, \qquad MSG = \frac{SSG}{K-1}, \tag{7}$$

$$\text{Total variability : } SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2. \tag{8}$$

Note also that $SST = SSW + SSG$. Now, we define the $F$ test statistic and its distribution under $H_0$, which follows the $\mathcal{F}$ (F-distribution) with degrees of freedom $(K-1)$ and $(n-K)$ as below,

$$F = \frac{MSG}{MSW} \sim \mathcal{F}_{(K-1,n-K)}. \tag{9}$$

Then $H_0$ claims that the means of these $K$ populations are the same, and $H_1$ claims that at least one group has a different mean. The testing problem is defined as below,

$$H_0 : \mu_i = \mu_{i'} \quad vs \quad H_1 : \mu_i \neq \mu_{i'} \quad ; \text{where, } 1 \leq i, i' \leq K \text{and } i \neq i'. \tag{10}$$

Additionally, we need to make the following assumptions:

1. Variances among the group are the same.
2. Each group in the population is normally distributed.

The decision rule with $p$-value $= P(F \geq F\text{-value})$ at significance level $\alpha$ is defined,

$$\text{Reject } H_0 : \text{ if } p\text{-value} < \alpha. \tag{11}$$

## 3.4 Multiple Testing Problem

In a multiple testing problem, we test $m$ null hypotheses $H_{01}, ..., H_{0m}$. Let us suppose all these hypotheses are true, and the probability of making $Type - I$ error for each hypothesis is $\alpha$, then we expect to falsely reject approximately $m \cdot \alpha$ hypotheses. Which is very high $Type - I$ errors.

### The Family-Wise Error Rate (FWER)

We know that $\alpha$ is also known as $Type - I$ error rate. The $FWER$ is the generalisation of it. If $V$ is the number of $Type - I$ errors at $\alpha$ for each hypothesis, then $FWER$ is defined as (James G. (2023)),

$$\begin{aligned} FWER(\alpha) &= P(V \geq 1) = 1 - P(V = 0) \qquad (12) \\ &= 1 - P\left(\bigcap_{j=1}^{m} \{ \text{ don't reject } H_0, \text{ when it is true } \}\right). \qquad (13) \end{aligned}$$

Since these $m$ tests are performed independently, we have,

$$FWER(\alpha) = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m. \qquad (14)$$

## 3.5 The Bonferroni Method

To control $FWER$, the Bonferroni method is applied whenever we have $m$ computed $p$-values. For this method, the form of hypotheses, the chosen test statistics, or whether the test was performed independently or not do not matter. Let $A_j$ be the event that we commit $Type - I$ error for the hypothesis $H_{0j}$, for $j = 1, 2, ..., m$ (James G. (2023)). Then $FWER$ at $\alpha$ is defined as follows,

$$\begin{aligned} FWER(\alpha) &= P(\{\text{falsely recjecting atleast one null hypothesis}\}) \\ &= P(\bigcup_{j=1}^{m} A_j) \leq \sum_{j=1}^{m} P(A_j) \qquad \text{(Since, P(A} \cup \text{B)} \leq \text{P(A) + P(B)).} \qquad (15) \end{aligned}$$

In the **Bonferroni method**, we set the significance level as $\alpha/m$ for each of these $m$ hypotheses, then we have $P(A_j) \leq \alpha/m$, and from Equation (13),

$$FWER(\alpha) \leq m \cdot \frac{\alpha}{m} = \alpha. \qquad (16)$$

This bounds $FWER$ at level $\alpha$. The decision rule for each individual hypothesis $H_{0j}$, for $j = 1, 2, ..., m$, in the multi-testing problem at $\alpha$ is defined as,

$$\text{Reject } H_0: \text{ if } p\text{-value} < \frac{\alpha}{m} \quad \text{or} \quad \text{if } m \cdot p\text{-value} < \alpha. \qquad (17)$$

## 3.6 Software

For our analysis, we used the statistical programming language R (2023-03-15 ucrt)(R Development Core Team (2020)). Apart from the base R package, several other packages are also used, such as *haven*(Hadley Wickham (2023)) for data importing, *dplyr*(Wickham *et al.* (2021)) for data manipulation, and *ggplot*(Hadley Wickham ORCID (2020)) for visualization.

# 4 Statistical analysis

## 4.1 Ants' attraction to sandwiches

### Validation of ANOVA assumptions

The sandwich is chosen randomly, and then a small piece of it is placed next to the ant hill, which implies that our observations are independent. From Table 3, for `Bread`, the highest variation, measured by standard deviation (*sd*), is in the group 'multigrain' with *sd* of 18.2 whereas, in other groups, the variation seems to be almost the same (see Figure 1). And we only have 12 observations in each group. Similarly, for `Topping`, *sd* does not deviate more than 3.02. Also, we only have 16 observations within each group. The variation of our estimator of variance, i.e., $sd^2$, is also high, and due to this high variance in the variance estimation we assume this assumption to be satisfied for further analysis. Lastly, the groups 'with' and 'without', appeared to have almost the same variation with a difference in *sd* of just 0.04.

From Figure 3, although we have very few data points, almost all points lie on the reference line for each bread type as well as for each topping, apart from a few points. In the case of Butter, most of the data points lie on the reference line for both groups. Hence, this explains that the distribution is normally distributed in each group of the factors. Since all the assumptions are validated, the ANOVA test can be conducted. First, for all the tests, we set the significance level $\alpha$ to 0.05. As `Bread` and `Topping` have more than 2 levels, we use the ANOVA test only for these two factors.

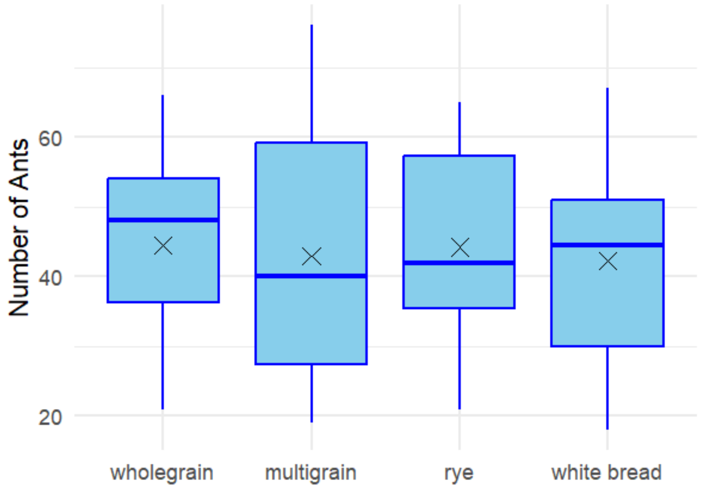
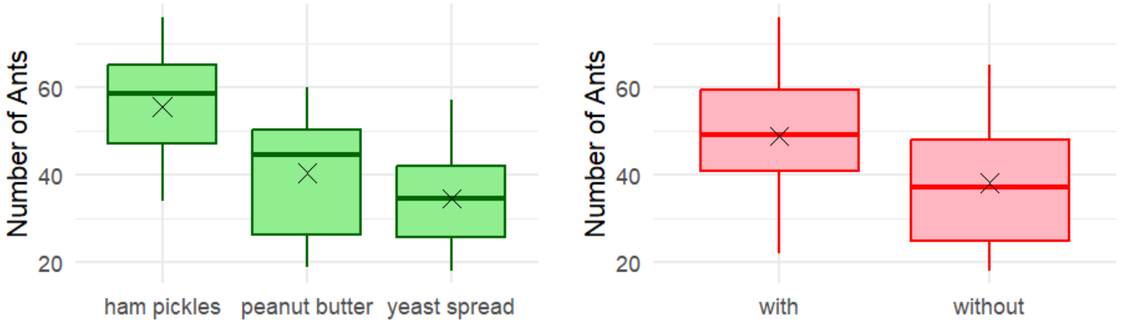


Figure 1: Box plot of `Ants` by `Bread`



(a) `Topping`

(b) `Butter`

Figure 2: Box Plots of `Ants` by the factors

### 4.1.1 ANOVA for bread types

For `Bread`, we defined $H_0$ as the number of ants within each group is the same on average, and $H_1$ as the number of ants in these groups is not the same on average. The testing problem is defined as follows,

$$H_0 : \mu_i = \mu_{i'} \quad vs \quad H_1 : \mu_i \neq \mu_{i'} \quad ;\text{ where, } 1 \leq i, i' \leq 4 \text{ and } i \neq i'. \tag{18}$$

Where $i = 1, 2, 3, 4$, which represent the bread types 'wholegrain', 'multigrain', 'rye' and 'white bread' respectively. From Table 1, $p$-value is 0.98267 and $p$-value $\not< \alpha$, which means we do not have sufficient statistical evidence against $H_0$ and we fail to reject $H_0$, . Hence, among these specified bread types, any one type of bread does not attract more or less ants than the other types on average.

### 4.1.2 ANOVA for different toppings

Now, for different types of topping, we defined the testing problem as below,

$$H_0 : \mu_i = \mu_{i'} \quad vs \quad H_1 : \mu_i \neq \mu_{i'} \quad ;\text{ where, } 1 \leq i, i' \leq 3 \text{ and } i \neq i'. \tag{19}$$

Here, these $\mu_1, \mu_2$, and $\mu_3$ are the means among the groups formed by different topping types, i.e., 'ham pickles', 'peanut butter' and 'yeast spread' respectively. The $p$-value, from Table 1, is 7.35e-05, which is less than $\alpha$. It suggests rejecting $H_0$ and implies that the average number of ants attracted to different toppings is not the same.

Table 1: ANOVA test for two factor variables

| factor | df | Sum Sq | Mean Sq | $F$-value | $p$-value |
|---|---|---|---|---|---|
| `Bread` | 3 | 41 | 13.5 | 0.055 | 0.983 |
| `Topping` | 2 | 3721 | 1860 | 11.85 | 7.35e-05 |

## 4.2 Pairwise t-test between different toppings

Since the average number of ants attracted to the different toppings is not the same, and it has more than 2 levels, we do a pairwise t-test only for the factor `Topping`. Since there are 3 levels, namely 'ham pickles', 'peanut butter' and 'yeast spread', for `Topping`, we have $\binom{3}{2} = 3$ testing problems,

$$H_{01} : \mu_1 = \mu_2 \quad vs \quad H_{11} : \mu_1 \neq \mu_2, \tag{20}$$

$$H_{02} : \mu_1 = \mu_3 \quad vs \quad H_{12} : \mu_1 \neq \mu_3, \tag{21}$$

$$H_{03} : \mu_2 = \mu_3 \quad vs \quad H_{13} : \mu_2 \neq \mu_3. \tag{22}$$

Where $\mu_1$, $\mu_2$ and $\mu_3$ are the average number of ants in the groups 'ham pickles', 'peanut butter' and 'yeast spread' respectively.

Table 2: Pairwise t-test for `Topping` and Two sample t-test for `Butter`

| factor | Hypothesis | adjusted $p$-value | pairs |
|---|---|---|---|
| `Topping` | $H_{01}$ | 0.00853 | 'ham pickles' & 'peanut butter' |
| | $H_{02}$ | 0.00005 | 'ham pickles' & 'yeast spread' |
| | $H_{03}$ | 0.63687 | 'peanut butter' & yeast spread |
| `Butter` | $H_{01}$ | 0.01232 | 'with' & 'without' |

First, we do a t-test of the groups 'ham pickles' and 'peanut butter' (c.f. Equation (20)). From Table 2, we have $p$-value = 0.02933 and $p$-value $< \alpha$ which infers to reject $H_{01}$. Hence, we conclude that the average number of ants attracted is not the same for sandwiches with toppings of 'ham pickles' and 'peanut butter'. Moreover, from Table 3, it can be concluded that, almost 15 more ants are attracted on average to the sandwiches with 'ham pickles' than the sandwiches with 'peanut butter'.

Secondly, we conducted the t-test for groups 'ham pickles' and 'yeast spread'. The $p$-value is 0.00052 and is very less than $\alpha$ (from Table 2). This indicates strong evidence against $H_{02}$, so we reject $H_{02}$. It also means that within the groups 'ham pickles' and 'yeast spread', the average number of ants attracted is different. And from Table 1, nearly 60% more ants are attracted on average to sandwiches with a topping of 'ham pickles' compared to 'yeast spread'.

Next, from the result of the t-test of the groups 'peanut butter' and 'yeast spread' and from Table 2, the $p$-value ($= 0.85308$) $\nless \alpha$ thus we do not have sufficient evidence to reject $H_{03}$. So, $H_{03}$ is maintained. Hence, on average, it does not matter if a sandwich is filled with 'peanut butter' or 'yeast spread' to the number of ants attracted to it.

## 4.3 The t-test for presence of butter

Now consider `Butter` and $\tilde{H}_0$ to be the means among the group formed by levels 'with' or 'without' being the same, and $\tilde{H}_1$ as the means among these groups are not the same. and defined the testing problem as,

$$\tilde{H}_0 : \mu_1 = \mu_2 \quad vs \quad \tilde{H}_1 : \mu_1 \neq \mu_2. \tag{23}$$

Where, $\mu_1$ and $\mu_2$ are the average number of ants in the groups 'with' and 'without' butter of the factor variable `Butter` respectively. From Table 2, the $p$-value is 0.015, which is less than $\alpha$. This indicates to reject $H_0$, and as a result and from Table 3, due to the presence of butter, around 10 more ants are attracted on average to the sandwiches.

# 5 Summary

The statistical hypothesis testing is conducted on the dataset `Sandwich.sav`, which is provided by the instructors of this course. The goal of this project is to understand the effect of bread types, different toppings, and the presence of butter in a sandwich on the average number of ants attracted to them.

The analysis begins with the validation of ANOVA assumptions for the study on ants' attraction to sandwiches. First, the observations are independent, as sandwiches are chosen randomly for the experiment. The analysis reveals varying estimated standard deviations (*sd*) across bread types due to the lower number of observations. For toppings, the *sd* remains consistent, with limited variation. Similarly, butter presence shows minimal *sd* differences between 'with' and 'without' groups. Then, using QQ-plot, the normality of the data in each group is justified.

The ANOVA test indicates no significant difference in ant attraction across types of bread, as $p$-value is bigger than $\alpha = 0.5$. However, for toppings, the $p$-value is smaller than $\alpha$ and suggests a significant difference in ant attraction among different toppings. Moreover, pairwise t-tests with small $p$-values, confirm significant differences between 'ham pickles' and 'peanut butter', as well as 'ham pickles' and 'yeast spread' toppings. Next, a t-test for butter presence reveals a significant difference, as $p$-values are less than $\alpha$, in ant attraction between sandwiches with and without butter. Specifically, sandwiches with butter attract, on average, 10 more ants than those without.

In conclusion, the study shows the importance of toppings and butter presence in attracting ants to sandwiches. While bread type may not significantly impact ant attraction, toppings and butter presence play crucial roles. These findings provide insights into ant behaviour and offer practical implications for sandwich preparation and picnic management.

# Bibliography

Hadley Wickham, Evan Miller. 2023. *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* R package version 2.5.4.

Hadley Wickham ORCID, Winston Chang. 2020. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* R package version 3.5.1.

James G., Witten D., Hastie T. Tibshirani R. 2023. *An Introduction to Statistical Learning: With Applications in R.* Springer.

Michael Schomaker, Christian Heumann. 2017. *Introduction to Statistics and Data Analysis, With Exercises, Solutions and Applications in R.* Springer.

Paul Newbold, William Carlson, Betty Thorne. 2019. *Statistics for Business and Economics.* Chapman Hall.

R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Robin Lock [aut, cre]. 2021. *Lock5Data: Datasets for "Statistics: UnLocking the Power of Data".* R package version 3.0.0.

Wickham, Hadley, François, Romain, Henry, Lionel, & Müller, Kirill. 2021. *dplyr: A Grammar of Data Manipulation.* R package version 1.0.6.

# Appendix

## A Additional Tables

Table 3: Summary of `Ants` by factors

| factor | levels | count | median | mean | sd |
|---|---|---|---|---|---|
| `Bread` | wholegrain | 12 | 48 | 44.5 | 14.6 |
| | multigrain | 12 | 40 | 43 | 18.2 |
| | rye | 12 | 42. | 44.2 | 13.4 |
| | white bread | 12 | 44.5 | 42.2 | 15.9 |
| `Topping` | ham pickles | 16 | 58.5 | 55.5 | 12.1 |
| | peanut butter | 16 | 44.5 | 40.4 | 14.2 |
| | yeast spread | 16 | 34.5 | 34.6 | 11.2 |
| `Butter` | with | 24 | 49 | 48.9 | 14.5 |
| | without | 24 | 37 | 38.1 | 14.1 |

# B Additional Figures

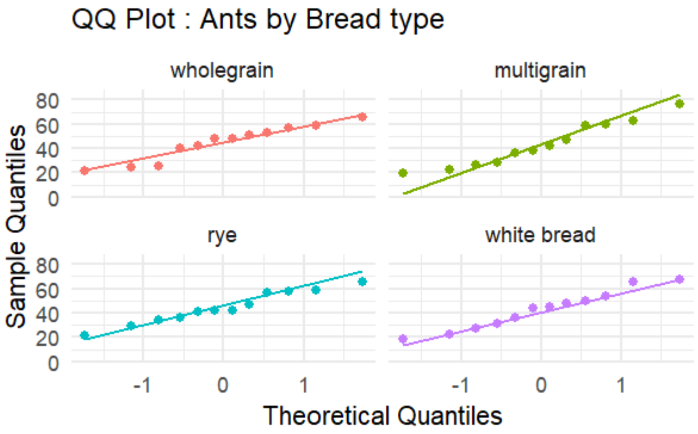
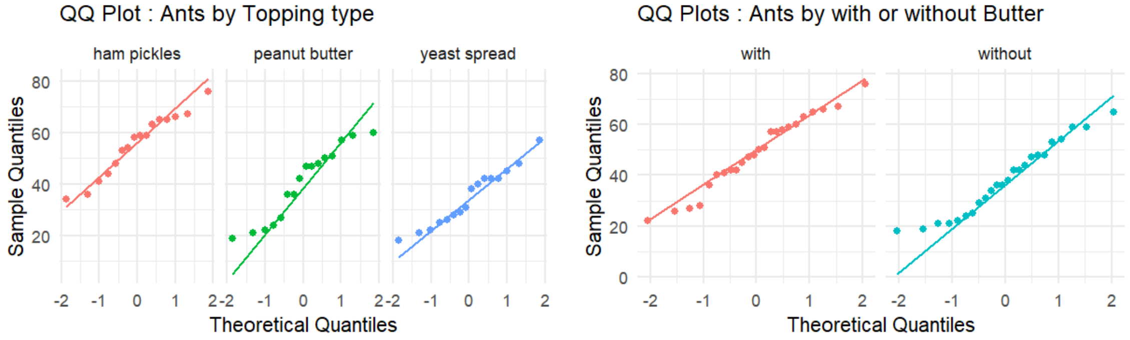


Figure 3: QQ-Plot of `Ants` by `Bread`



(a) `Topping`

(b) `Butter`

Figure 4: QQ-Plots of `Ants` by the factors