Formative exercise

B012345

February 2023

The Research Context

The German Breast Cancer Study Group recruited 510 women with node positive breast cancer into a cohort study in the late 1980's. The women were assessed for a number of possible risk factors (age, menopausal status, treatment with Tamoxifen, number of nodes involved, tumour grade and tumour size) and were then followed up for a median period of 5 years. At the end of follow-up, the participants' mortality status was determined (dead or alive). The Study Group was interested in determining whether any of these possible prognostic factors would be helpful in predicting mortality outcomes for this category of breast cancer patient in future.

Project Goals

The data for these 510 women can be found in the comma-separated values format file **Breast cancer data.csv**. Use these data to investigate the following questions, and write a report summarising your findings:

1. Do any of the risk factors measured appear to be associated individually with mortality at the end of follow-up?
2. What are the odds of dying in the following subgroups?
    - Tamoxifen-treated versus No Tamoxifen treatment
    - Tumour Grade 1 versus Tumour Grades 2 and 3

Data input

```
# Load packages and import data
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.3      v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(finalfit)
library(epitools)
brca<-read_csv("Breast cancer data.csv")
```

```
## Rows: 686 Columns: 8
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Menopause, Hormonal
## dbl (6): Subject, Age, Size, Grade, Nodes, Dead
##
## 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.
```

Data checking

```
# Quick inspection of the dataset
View(brca)
summary(brca)
```

```
##     Subject            Age          Menopause             Size
##  Min.   :   1.0   Min.   :21.00   Length:686         Min.   :  3.00
##  1st Qu.: 580.8   1st Qu.:46.00   Class :character   1st Qu.: 20.00
##  Median :1015.5   Median :53.00   Mode  :character   Median : 25.00
##  Mean   : 966.1   Mean   :53.05                      Mean   : 29.33
##  3rd Qu.:1340.5   3rd Qu.:61.00                      3rd Qu.: 35.00
##  Max.   :1819.0   Max.   :80.00                      Max.   :120.00
##      Grade           Nodes          Hormonal             Dead
##  Min.   :1.000   Min.   : 1.00   Length:686         Min.   :0.0000
##  1st Qu.:2.000   1st Qu.: 1.00   Class :character   1st Qu.:0.0000
##  Median :2.000   Median : 3.00   Mode  :character   Median :0.0000
##  Mean   :2.117   Mean   : 5.01                      Mean   :0.4359
##  3rd Qu.:2.000   3rd Qu.: 7.00                      3rd Qu.:1.0000
##  Max.   :3.000   Max.   :51.00                      Max.   :1.0000
```

Data preparation

```
# Create a working dataset with categorical variables converted to factors
# Change the levels of Mortality to Alive (0) and Dead (1)
brcaw <- brca %>%
  mutate(across(all_of(c("Dead","Menopause","Hormonal","Grade")), as_factor)) %>%
  mutate(Dead = fct_recode(Dead,"Alive"="0","Dead"="1"))
```

```r
# Transformation of Grade into a binary variable (1 vs 2 + 3)
brcaw <- brcaw %>%
  mutate(GradeBin = fct_collapse(Grade, low = c("1"),
                                 high = c("2","3")))
```

Exploratory Data Analysis

The dataset has 686 observations (patients) and 8 variables. We want to identify risk factors associated with mortality at the end of the follow-up period. Mortality is a binary categorical variable. We renamed its values as Alive (0) and Dead (1). Grade was transformed into a binary variable (1 vs 2 + 3) to increase power.

The numerical variables are:

1. Age: age of the patient (years)
2. Size: tumour size (mm)
3. Nodes: number of positive lymph nodes

The categorical variables are:

1. Menopause: pre- or postmenopausal status of the patient
2. Hormonal: whether tamoxifen hormone treatment was used
3. Grade: grade of the patient's tumour (1, 2, 3)

We first looked at the summary statistics for the whole dataset.

Graphical analysis

We then checked the distribution of each numerical variable using histograms for each Mortality subgroup. We created boxplots for Age, Nodes and Size to visualise the difference in centre and spread between Dead and Alive patients.

Numerical analysis

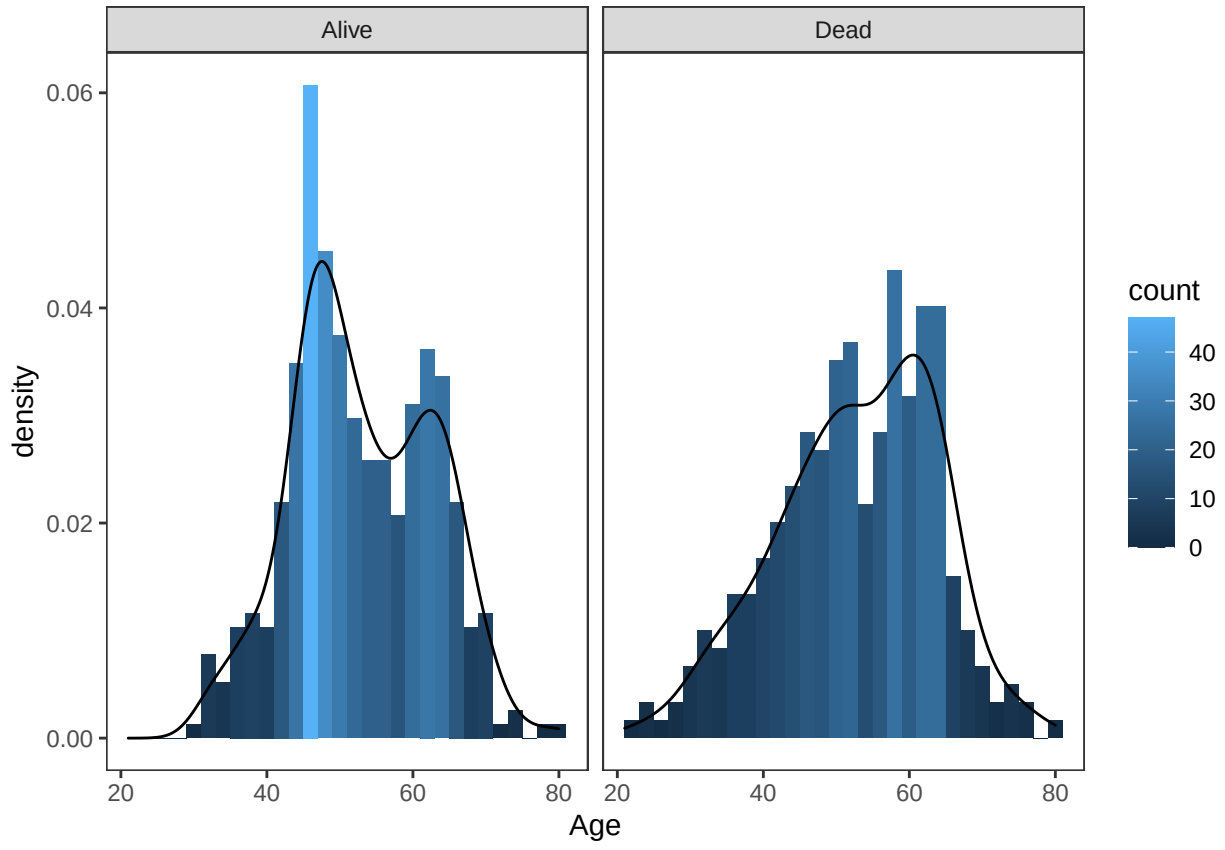
We investigated the association between mortality and the numerical variables. We looked at summary statistics (mean, median, and standard deviation) for each subgroup (Dead and Alive) of each numerical variable to get a first impression of the association with mortality status. We then investigated the association between mortality and the categorical variables (Menopause, Hormonal and Grade) by creating contingency tables for each variable.

```r
# Overview of the dataset summary statistics
summary(brcaw)
```

```
##     Subject            Age                Menopause        Size         Grade
##  Min.   :   1.0   Min.   :21.00   premenopausal :290   Min.   :  3.00   1: 81
##  1st Qu.: 580.8   1st Qu.:46.00   postmenopausal:396   1st Qu.: 20.00   2:444
##  Median :1015.5   Median :53.00                        Median : 25.00   3:161
##  Mean   : 966.1   Mean   :53.05                        Mean   : 29.33
##  3rd Qu.:1340.5   3rd Qu.:61.00                        3rd Qu.: 35.00
```

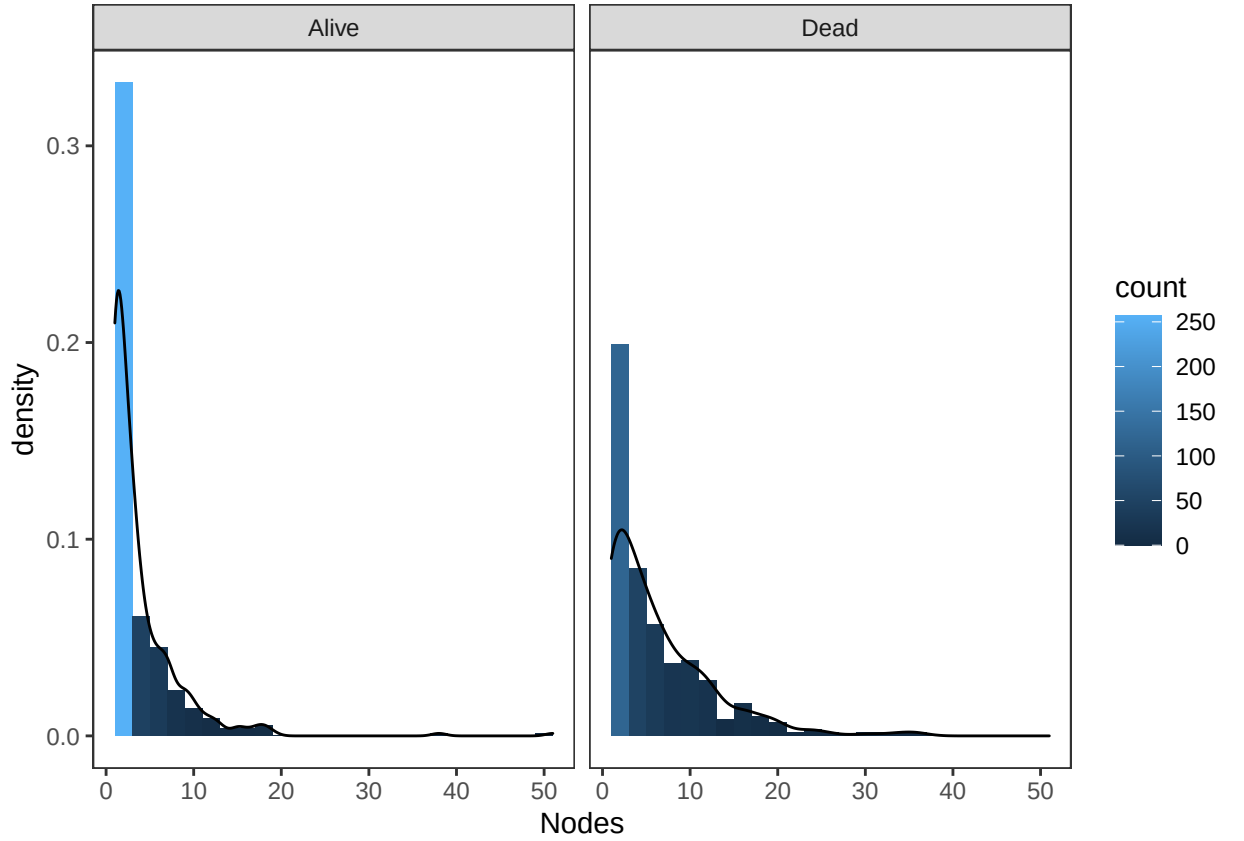
```
##  Max.   :1819.0   Max.   :80.00                          Max.   :120.00
##      Nodes                 Hormonal      Dead     GradeBin
##  Min.   : 1.00   no tamoxifen :440   Alive:387   low : 81
##  1st Qu.: 1.00   had tamoxifen:246   Dead :299   high:605
##  Median : 3.00
##  Mean   : 5.01
##  3rd Qu.: 7.00
##  Max.   :51.00
```

```
# Numerical variables
# Histogram for Age
brcaw %>%
  ggplot(aes(x=Age)) +
  geom_histogram(aes(y=..density.., fill=..count..),binwidth=2)+
  geom_density(aes(y=..density..)) +
  facet_wrap(~Dead) +
  theme_bw() +                          ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```

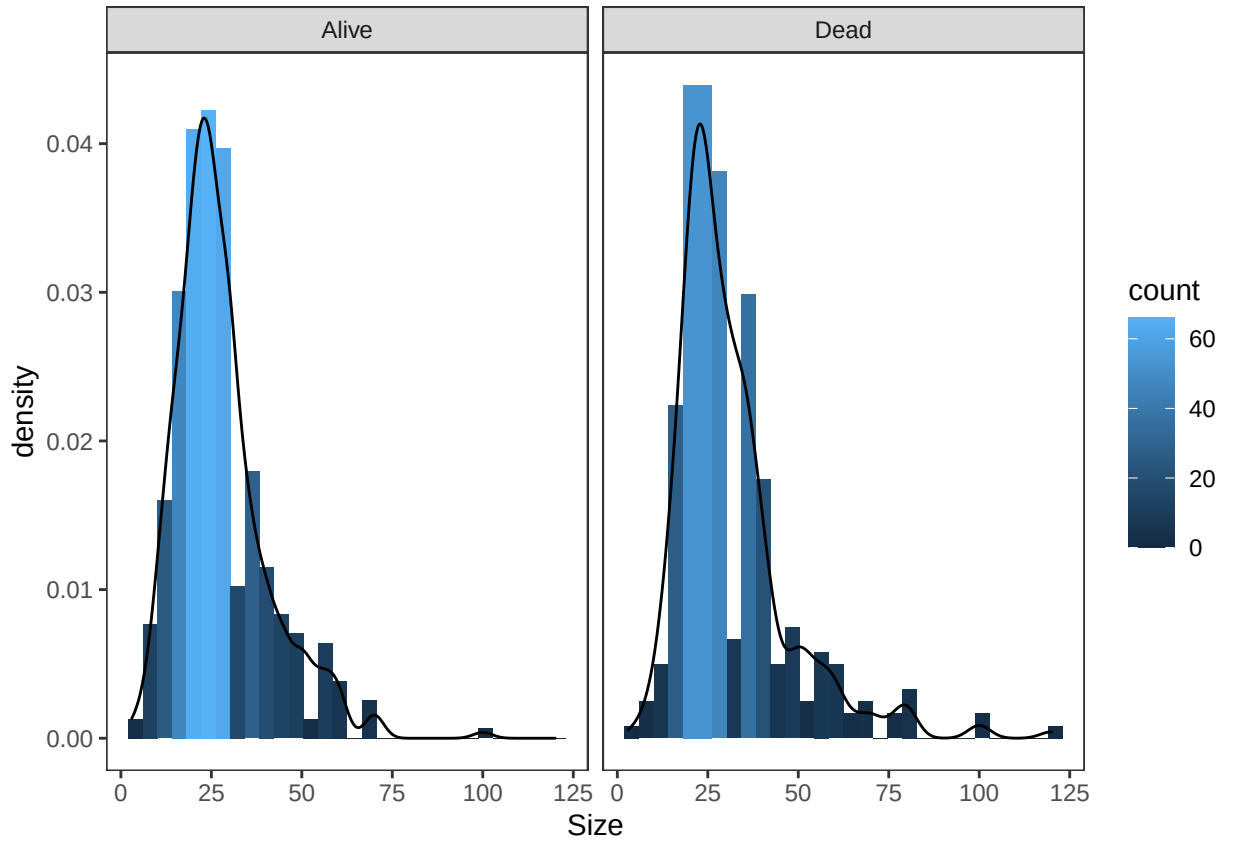



```
# Histogram for Nodes
brcaw %>%
  ggplot(aes(x=Nodes)) +
  geom_histogram(aes(y=..density.., fill=..count..),binwidth=2)+
  geom_density(aes(y=..density..)) +
  facet_wrap(~Dead) +
```

```
  theme_bw() +                          ## remove gray background
  theme(panel.grid=element_blank())   ## remove grid
```



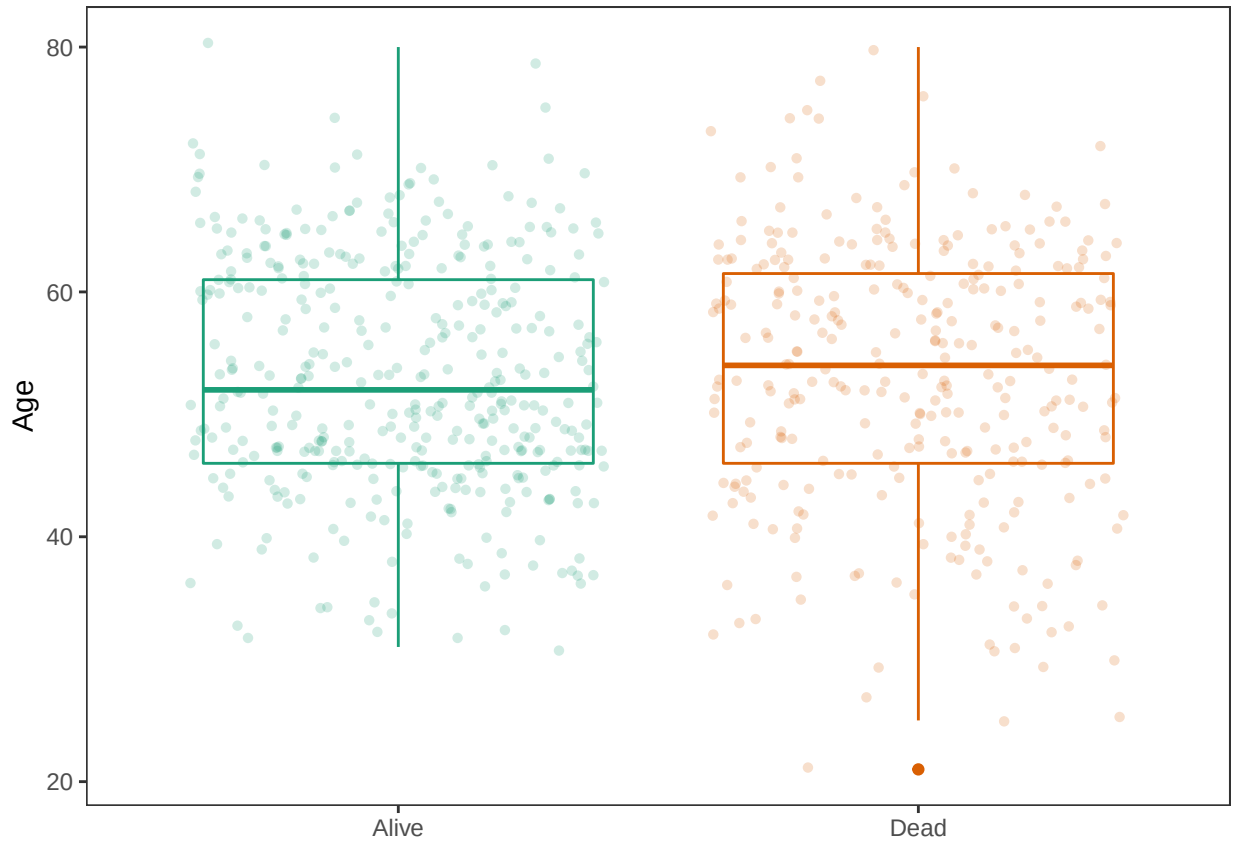

```
# Histogram for Size
brcaw %>%
  ggplot(aes(x=Size)) +
  geom_histogram(aes(y=..density.., fill=..count..),bins=30)+
  geom_density(aes(y=..density..)) +
  facet_wrap(~Dead) +
  theme_bw() +                          ## remove gray background
  theme(panel.grid=element_blank())   ## remove grid
```



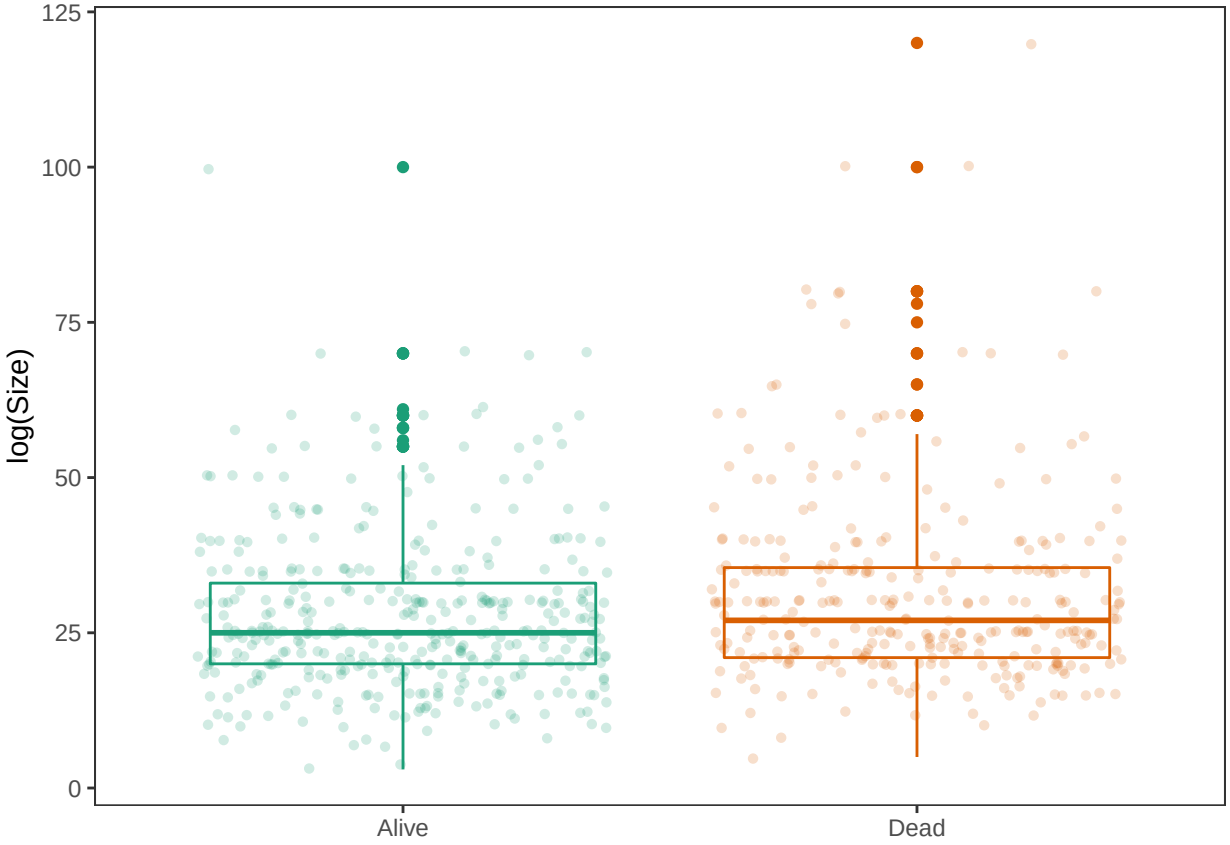

```
# Summary statistics per mortality status
brcaw %>%
  group_by(Dead) %>%
  summarize(meanAge=mean(Age), stdevAge=sd(Age), medianAge=median(Age),
            meanSize=mean(Size), stdevSize=sd(Size), medianSize=median(Size),
            meanNodes=mean(Nodes), stdevNodes=sd(Nodes),medianNodes=median(Nodes))
```

```
## # A tibble: 2 x 10
##   Dead  meanAge stdevAge media~1 meanS~2 stdev~3 media~4 meanN~5 stdev~6 media~7
##   <fct>   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Alive    53.1     9.52      52    27.7    12.8      25    3.84    4.58       2
## 2 Dead     53.0    10.9       54    31.5    15.8      27    6.52    6.14       5
## # ... with abbreviated variable names 1: medianAge, 2: meanSize, 3: stdevSize,
## #   4: medianSize, 5: meanNodes, 6: stdevNodes, 7: medianNodes
```

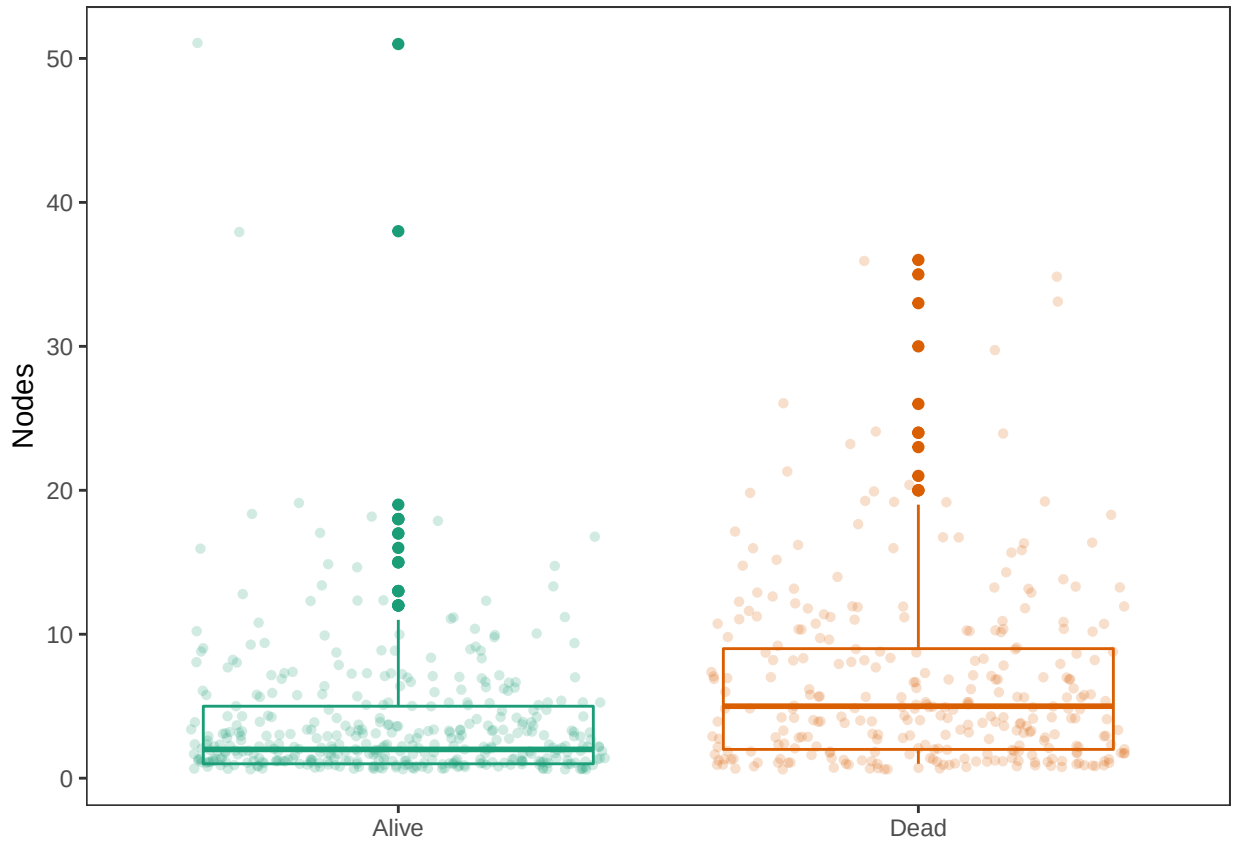
```
# Boxplot for Age
brcaw %>%
  ggplot(aes(x=Dead, y=Age, color=Dead)) +
  geom_boxplot() +
  scale_color_brewer(palette="Dark2") +
  geom_point(shape=16,alpha=0.2, position=position_jitter()) +
  labs(y='Age') +
  theme_bw() +
  theme(panel.grid=element_blank(),legend.position="none",axis.title.x=element_blank ())
```




```
# Boxplot for Size
brcaw %>%
  ggplot(aes(x=Dead, y=Size, color=Dead)) +
  geom_boxplot() +
  scale_color_brewer(palette="Dark2") +
  geom_point(shape=16,alpha=0.2, position=position_jitter()) +
  labs(y='log(Size)') +
  theme_bw() +
  theme(panel.grid=element_blank(),legend.position="none",axis.title.x=element_blank ())
```




```
# Boxplot for Nodes
brcaw %>%
  ggplot(aes(x=Dead, y=Nodes, color=Dead)) +
  geom_boxplot() +
  scale_color_brewer(palette="Dark2") +
  geom_point(shape=16,alpha=0.2, position=position_jitter()) +
  labs(y='Nodes') +
  theme_bw() +
  theme(panel.grid=element_blank(),legend.position="none",axis.title.x=element_blank ())
```




```
# Categorical variables
# Menopause
# Create a cross-tabulation with proportions and row totals
brcaw %>%
  select(Dead,Menopause) %>%
  summary_factorlist(dependent ="Dead",explanatory = "Menopause",column = F,total_col = T)
```

```
##      label         levels      Alive       Dead     Total
##  Menopause  premenopausal 171 (59.0) 119 (41.0) 290 (100)
##            postmenopausal 216 (54.5) 180 (45.5) 396 (100)
```

```
# Hormonal
# Create a cross-tabulation with proportions and row totals
brcaw %>%
  select(Dead,Hormonal) %>%
  summary_factorlist(dependent ="Dead",explanatory = "Hormonal",column = F,total_col = T)
```

```
##     label        levels      Alive       Dead     Total
##  Hormonal  no tamoxifen 235 (53.4) 205 (46.6) 440 (100)
##           had tamoxifen 152 (61.8)  94 (38.2) 246 (100)
```

```
# Grade
# Create a cross-tabulation with proportions and row totals
brcaw %>%
```

```
  select(Dead,GradeBin) %>%
  summary_factorlist(dependent ="Dead",explanatory = "GradeBin",column = F,total_col = T)
```

```
##     label levels      Alive       Dead     Total
##  GradeBin    low  63 (77.8)  18 (22.2)  81 (100)
##             high 324 (53.6) 281 (46.4) 605 (100)
```

Looking at the summary statistics for the whole dataset, 299 patients out of the 686 patients, had died by the end of the follow up, and 387 patients were still alive. No variable presented values outside of the expected range and there were no missing values in the dataset. We note that the Nodes variable had both a minimum and first quartile value of 1, indicating a high number of 1 values.

The summary statistics by mortality status showed that patients who have died by the end of the follow up had similar if not slightly higher age, larger tumour size and higher number of nodes than patients who were alive by the end of the follow up.

The histograms suggest a slight bimodal distribution for Age, a right skew for Size and a very strong right skew for Nodes.

For the categorical variables, contingency tables from the Exploratory Data Analysis suggested that more patients taking Tamoxifen were still alive by the end of the follow up than patients who did not take Tamoxifen; a higher percentage of post-menopausal women died than pre-menopausal women; and a large excess of patients died with a high grade tumour compared with patients with a low grade tumour.

Develop the analysis plan

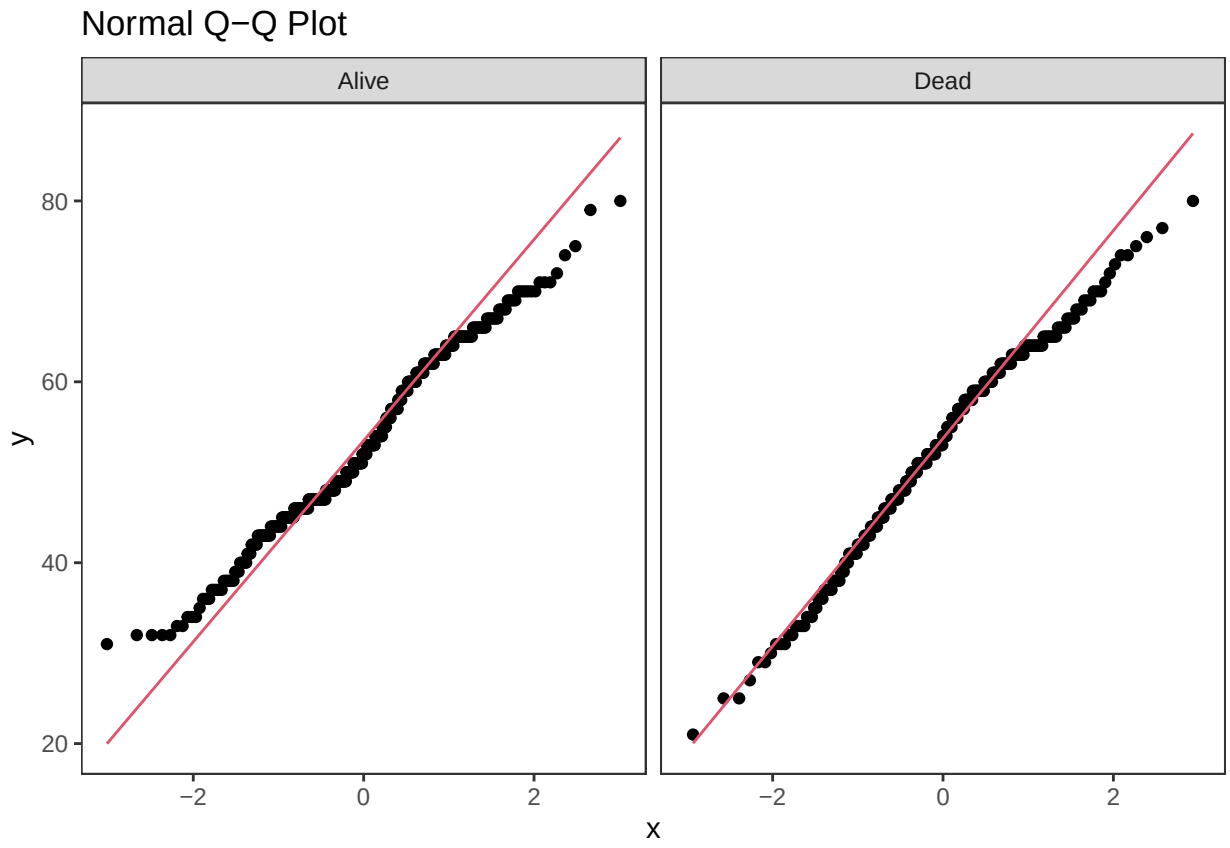
To determine whether the numerical variables are associated with mortality status, in addition to the boxplots from the exploratory data analysis, we plan to use t-tests comparing the mean of the variable between the two Mortality subgroups. The null hypothesis for each of these tests will be that there is no difference in means. We will perform two-sided tests, so our alternative hypothesis will be that there is a difference between the means. To determine whether the categorical variables are associated with mortality status, in addition to the contingency tables from the exploratory data analysis, we plan to use chi-square tests. The null hypothesis is that there is no association between the two categorical variables. In addition, an odds ratio was calculated for the mortality outcome given the hormonal tamoxifen treatment and for the mortality outcome given the tumour grade (high or low). The null hypothesis is that the odds ratio is equal to 1, meaning that there is no effect of the exposure on the outcome. For all tests, we will use a significance level of 5%.

Investigate the assumptions

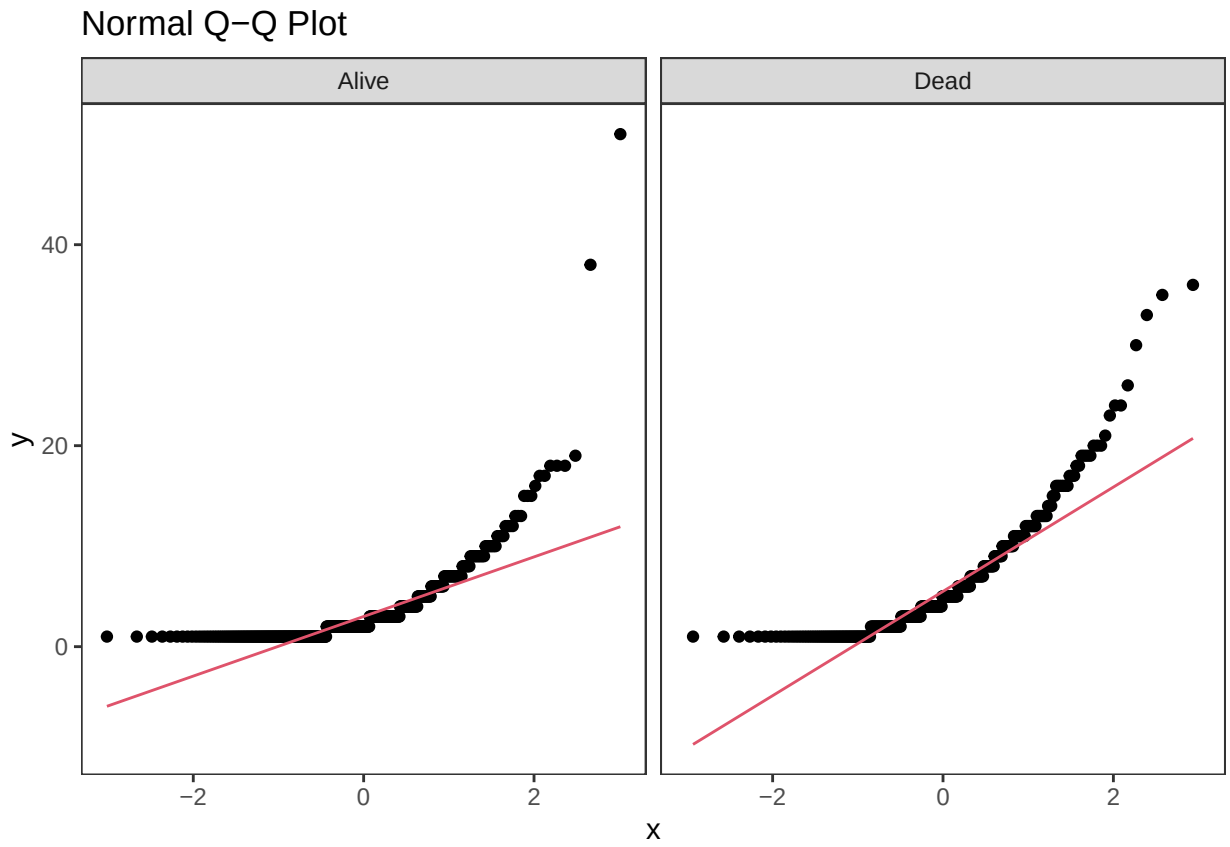
In addition to the histograms obtained in the graphical exploratory data analysis, we created QQ-plots for both Mortality subgroups for each numerical variable to check the assumption of normality of the t-tests. Since we saw in the Exploratory Data Analysis histograms that Size isn't normally distributed, we log-transformed this variable and checked whether the transformed variable is normally distributed. We also tested the assumption of homoscedasticity (equal variances) of the t-test with F-tests for Age and log(Size).

To check whether the Chi-square test could be used for all categorical variables, we looked whether all expected values were greater than 5.

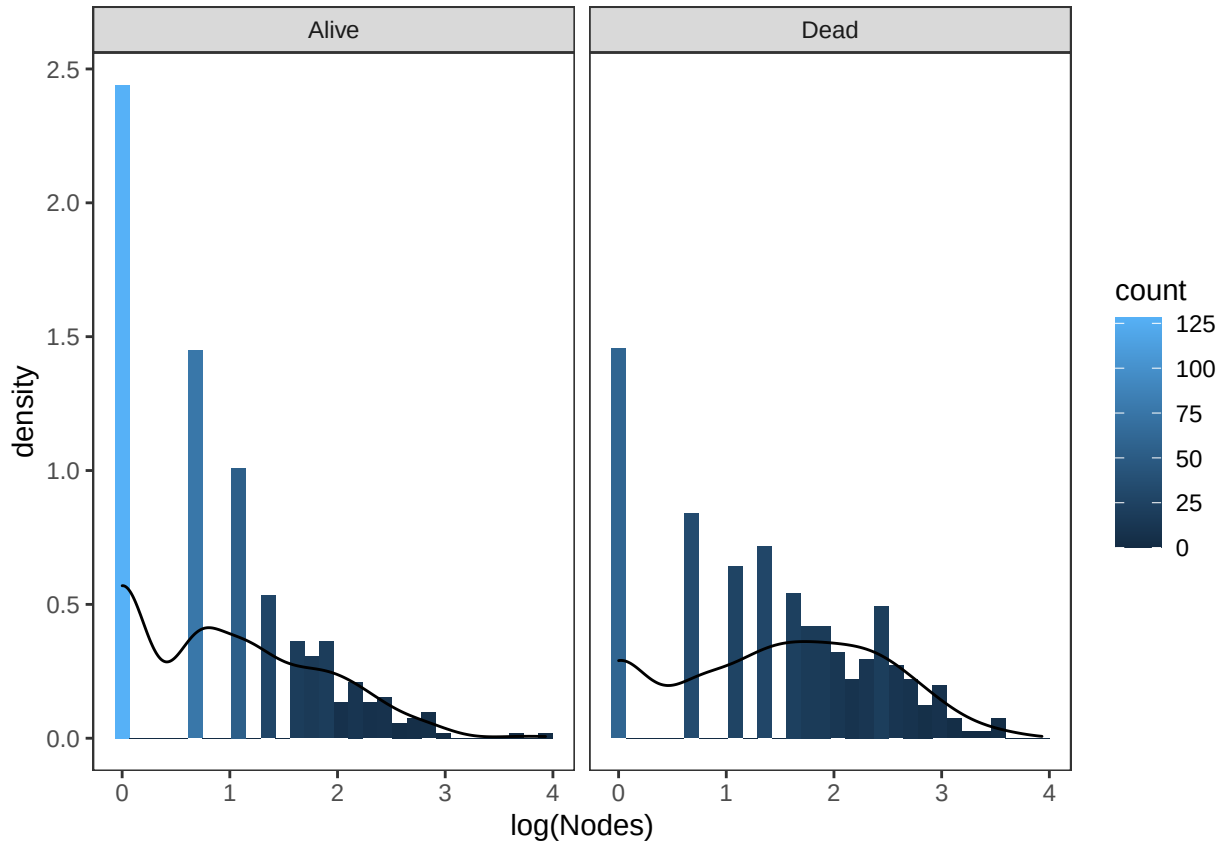
```r
# QQ-plot for Age
brcaw %>%
  ggplot(aes(sample=Age)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_wrap(~Dead) +
  labs(title="Normal Q-Q Plot") +    ## add title
  theme_bw() +                       ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```



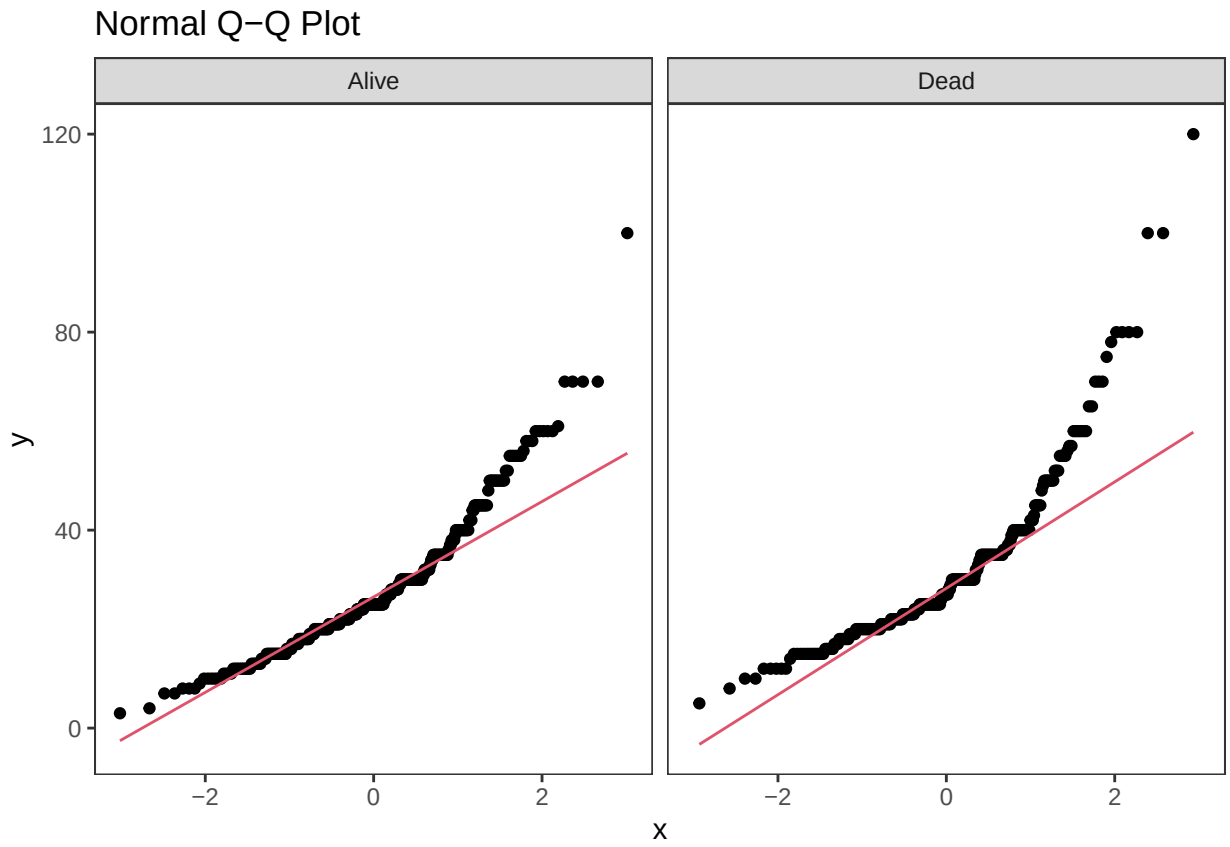

```r
# QQ-plot for Nodes
brcaw %>%
  ggplot(aes(sample=Nodes)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_wrap(~Dead) +
  labs(title="Normal Q-Q Plot") +    ## add title
  theme_bw() +                       ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```



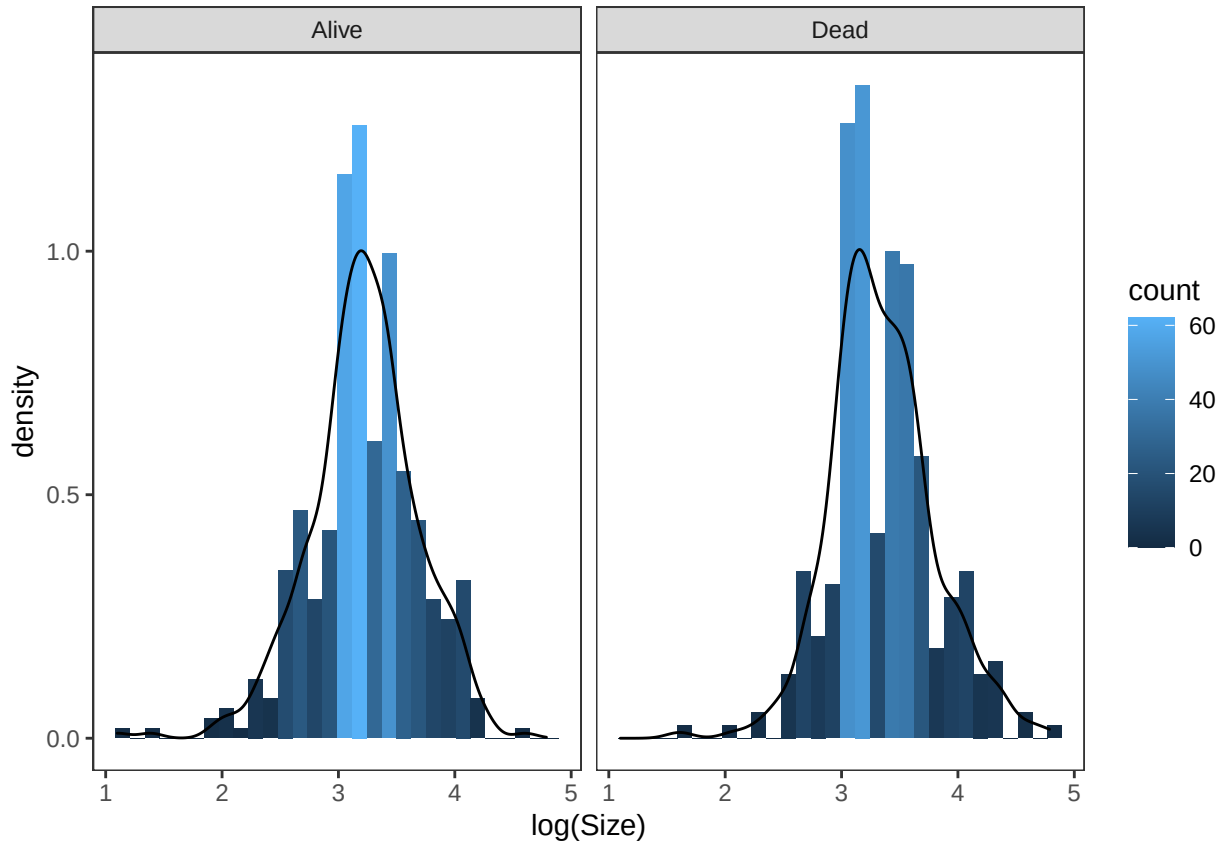

```
# Histogram for log(Nodes)
brcaw %>%
  ggplot(aes(x=log(Nodes))) +
  geom_histogram(aes(y=..density.., fill=..count..),bins=30)+
  geom_density(aes(y=..density..)) +
  facet_wrap(~Dead) +
  theme_bw() +                          ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```



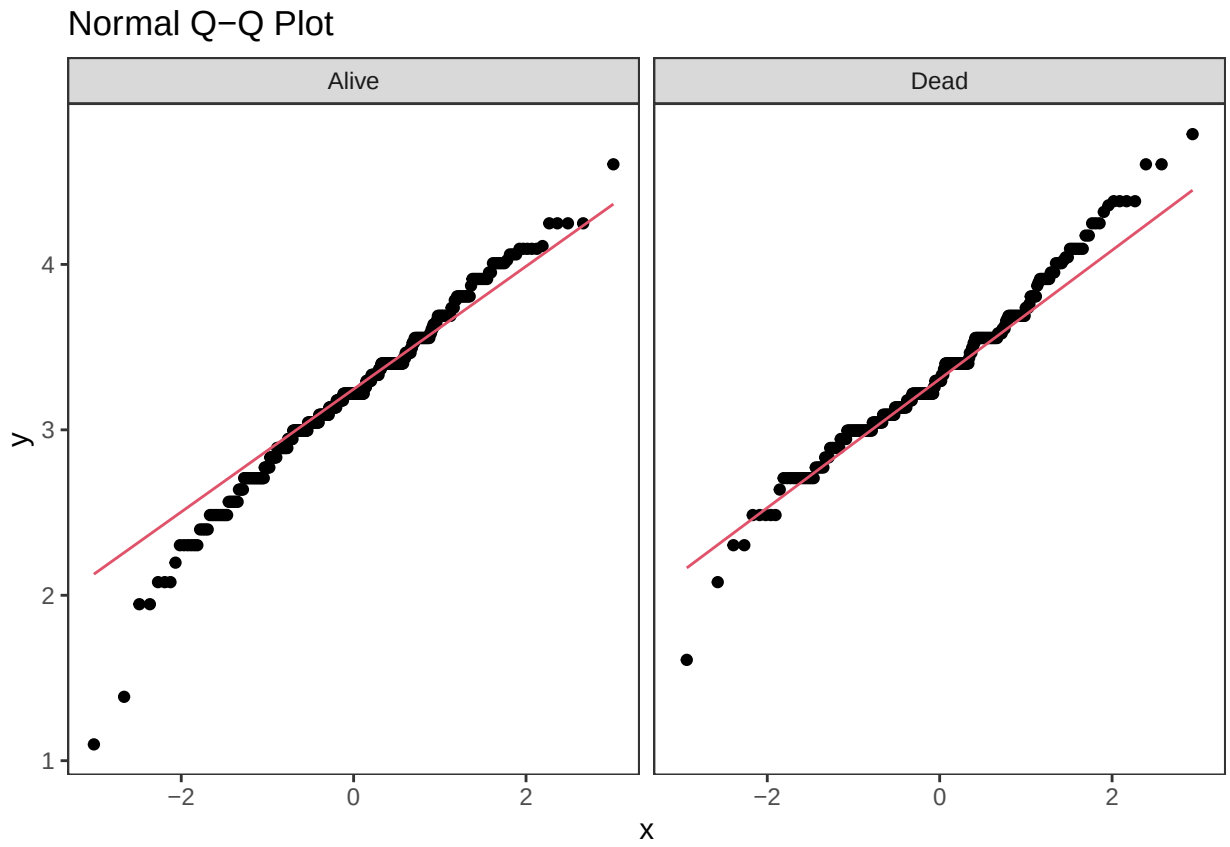

```
# QQ-plot for Size
brcaw %>%
  ggplot(aes(sample=Size)) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_wrap(~Dead) +
  labs(title="Normal Q-Q Plot") +    ## add title
  theme_bw() +                       ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```




```
# Histogram for log(Size)
brcaw %>%
  ggplot(aes(x=log(Size))) +
  geom_histogram(aes(y=..density.., fill=..count..),bins=30)+
  geom_density(aes(y=..density..)) +
  facet_wrap(~Dead) +
  theme_bw() +                          ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```




```
# QQ-plot for log(Size)
brcaw %>%
  ggplot(aes(sample=log(Size))) +
  stat_qq() +
  stat_qq_line(color=2) +
  facet_wrap(~Dead) +
  labs(title="Normal Q-Q Plot") +    ## add title
  theme_bw() +                       ## remove gray background
  theme(panel.grid=element_blank())  ## remove grid
```




```
# Test of equality of variances
brcaw %>%
  var.test(Age ~ Dead, ., alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  Age by Dead
## F = 0.7687, num df = 386, denom df = 298, p-value = 0.0153
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6196159 0.9508129
## sample estimates:
## ratio of variances
##          0.7687013
```

```
brcaw %>%
  var.test(log(Size) ~ Dead, ., alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  log(Size) by Dead
## F = 1.1019, num df = 386, denom df = 298, p-value = 0.3776
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.8881889 1.3629437
## sample estimates:
## ratio of variances
##           1.101896
```

```
# Categorical variables: expected values
brcaw %>%
  select(Menopause,Dead) %>%
  table() %>%
  chisq.test() %>%
  .$expected
```

```
##                 Dead
## Menopause          Alive     Dead
##   premenopausal  163.6006 126.3994
##   postmenopausal 223.3994 172.6006
```

```
# Hormonal: Perform Chi-square test
brcaw %>%
  select(Hormonal,Dead) %>%
  table() %>%
  chisq.test() %>%
  .$expected
```

```
##                Dead
## Hormonal          Alive     Dead
##   no tamoxifen  248.2216 191.7784
##   had tamoxifen 138.7784 107.2216
```

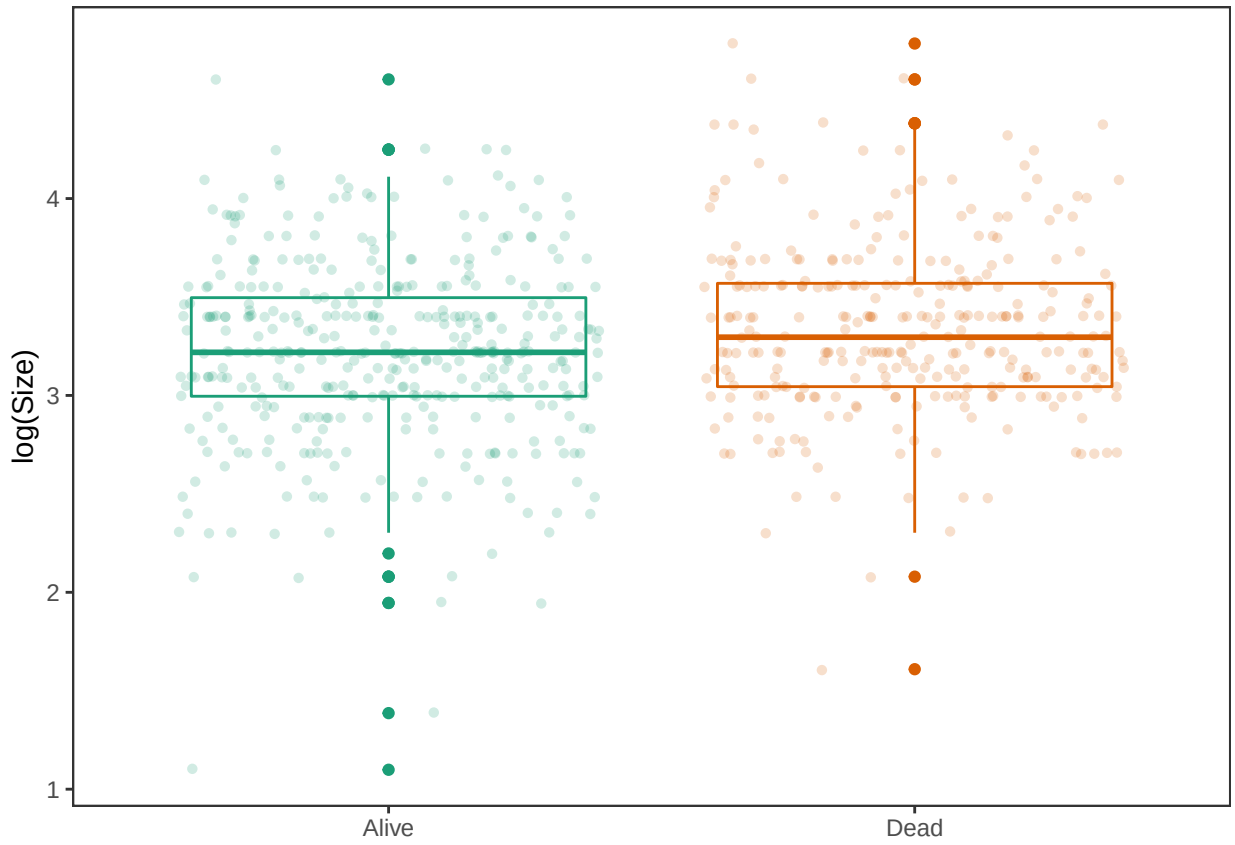
```
# GradeBin: Perform Chi-square test
brcaw %>%
  select(GradeBin,Dead) %>%
  table() %>%
  chisq.test() %>%
  .$expected
```

```
##         Dead
## GradeBin     Alive      Dead
##     low   45.69534  35.30466
##     high 341.30466 263.69534
```

Despite the bimodal appearance of Age in the histograms from the Exploratory Data Analysis, the Q-Q plots show that the distribution is "sufficiently normal" to perform a t-test. Both histograms and QQ-plots showed that Nodes and Size were not normally distributed. A log transformation was sufficient to make the Size distribution close to normal. However, that wasn't the case for Nodes, so we picked the non-parametric Wilcoxon test for this variable. It is worth noting that Nodes could also have been treated as a categorical variable given its skewed distribution and discrete nature. Given the similar spread in boxplot and standard deviation, and the not or only barely significant F-tests ($p>0.015$), we chose to perform Welch's t-test for each normally distributed variable (Age and log(Size)) as it allows for unequal variances. We performed a Wilcoxon's rank sum test for the Nodes variable. Since there were at least 5 patients in each cell of all contingency tables, we were able to test the significance of the association between Mortality and each categorical variable with Chi-square tests.

Carry out the analysis

```
# Numerical variables
# Boxplot for log(Size)
brcaw %>%
  ggplot(aes(x=Dead, y=log(Size), color=Dead)) +
  geom_boxplot() +
  scale_color_brewer(palette="Dark2") +
  geom_point(shape=16,alpha=0.2, position=position_jitter()) +
  labs(y='log(Size)') +
  theme_bw() +
  theme(panel.grid=element_blank(),legend.position="none",axis.title.x=element_blank ())
```




```
# t-tests
brcaw %>% t.test(.$Age~.$Dead,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  .$Age by .$Dead
## t = 0.10982, df = 595.04, p-value = 0.9126
## alternative hypothesis: true difference in means between group Alive and group Dead is not equal to 0
## 95 percent confidence interval:
##  -1.470397  1.644587
```

```
## sample estimates:
## mean in group Alive  mean in group Dead
##            53.09044            53.00334
```

```
brcaw %>% t.test(log(.$Size)~.$Dead,data=.)
```

```
##
##  Welch Two Sample t-test
##
## data:  log(.$Size) by .$Dead
## t = -3.7035, df = 654.92, p-value = 0.0002305
## alternative hypothesis: true difference in means between group Alive and group Dead is not equal to 0
## 95 percent confidence interval:
##  -0.1971727 -0.0605368
## sample estimates:
## mean in group Alive  mean in group Dead
##            3.218404            3.347259
```

```
# 95% CI for ratio of sizes (Alive/Dead)
exp(t.test(log(brcaw$Size)~brcaw$Dead,data=brcaw)$conf.int[1])
```

```
## [1] 0.8210488
```

```
exp(t.test(log(brcaw$Size)~brcaw$Dead,data=brcaw)$conf.int[2])
```

```
## [1] 0.9412591
```

```
# Wilcoxon test
brcaw %>% wilcox.test(.$Nodes~.$Dead,data=.,conf.int=TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  .$Nodes by .$Dead
## W = 40180, p-value = 3.346e-12
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -2.000058 -1.000074
## sample estimates:
## difference in location
##              -1.999919
```

```
# Categorical variables
# Menopause: Perform Chi-square test
brcaw %>%
  select(Menopause,Dead) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 1.1564, df = 1, p-value = 0.2822
```

```
# Hormonal: Perform Chi-square test
brcaw %>%
  select(Hormonal,Dead) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 4.1714, df = 1, p-value = 0.04111
```

```
# GradeBin: Perform Chi-square test
brcaw %>%
  select(GradeBin,Dead) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  .
## X-squared = 16.077, df = 1, p-value = 6.081e-05
```

```
# Calculate the Odds Ratio between Hormonal and Dead
brcaw %>%
  select(Hormonal,Dead) %>%
  count(Hormonal, Dead) %>%
  spread(Dead,n) %>%
  select(-Hormonal) %>%
  as.matrix() %>%
  epitab(.,method="oddsratio")
```

```
## $tab
##      Alive        p0 Dead        p1 oddsratio     lower     upper    p.value
## [1,]   235 0.6072351  205 0.6856187 1.0000000        NA        NA         NA
## [2,]   152 0.3927649   94 0.3143813 0.7089217 0.5157317 0.9744795 0.03694663
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

```r
# Calculate the Odds Ratio between GradeBin and Dead
brcaw %>%
  select(GradeBin,Dead) %>%
  count(GradeBin, Dead) %>%
  spread(Dead,n) %>%
  select(-GradeBin) %>%
  as.matrix() %>%
  epitab(.,method="oddsratio")
```

```
## $tab
##      Alive        p0 Dead         p1 oddsratio    lower    upper      p.value
## [1,]    63 0.1627907   18 0.06020067  1.000000       NA       NA           NA
## [2,]   324 0.8372093  281 0.93979933  3.035494 1.755454 5.248912 2.540017e-05
##
## $measure
## [1] "wald"
##
## $conf.level
## [1] 0.95
##
## $pvalue
## [1] "fisher.exact"
```

Interpretation

The t-test confirmed that age was not a significant risk factor (t=0.11,df=595.04,p=0.91) and the 95% confidence interval for the difference in means included 0 (95% CI: -1.47;1.64). Tumour size (t=-3.70,df=654.92,p=2.31x10-4) and number of nodes (W=40180,p=3.35x10-12) were both found to be significant risk factors as there was a significant difference in mean tumour size and in mean number of nodes between the two Mortality subgroups. The 95% CI for log(Size) is for the difference in tumour size between subgroups (Alive - Dead) in the log scale. This is equivalent to the 95% CI for log(Alive/Dead), the log of the ratio of sizes between the two subgroups. We can exponentiate the 95% CI to get the 95% CI for the ratio of sizes (95% CI: 0.821;0.941) and see that this doesn't include 1. Patients with more positive lymph nodes and larger tumours were more likely to die.

For the categorical variables, contingency tables from the Exploratory Data Analysis suggested that there were more patients taking Tamoxifen who were still alive by the end of the follow up (61.8%) than patients who did not take Tamoxifen (53.4%). The contingency table suggested that a higher percentage of post-menopausal women died (45.5%) than pre-menopausal women (41%). Grade showed a large excess of patients who died with a high grade tumour (46.4% with grade 2 or 3) compared with patients with a low grade tumour (22.2% with grade 1).

From the Chi-square tests, we see that hormonal treatment had a borderline significant association with mortality status (Chi-squared=4.171, df=1, p-value=0.041). Menopause status was not significantly associated with mortality status (Chi-squared=1.16, df=1, p=0.28). Grade had a significant association with mortality status (Chi-squared=16.08, df=1, p=6.08x10-5).

The odds of dying in those treated with tamoxifen was only 71% the odds of dying in the untreated group (OR: 0.71; p=0.037) and slightly significant. The confidence interval (95% CI: 0.52;0.97) is wide and its upper bound is close to 1. The odds-ratio analysis confirms that taking tamoxifen has a slight protective effect.

The odds of dying in those with a high tumour grade was 3 times the odds of dying in the group with low tumour grade (OR: 3.04; p=2.54x10-5) and highly significant. The confidence interval (95% CI: 1.76;5.25)

is wide but its lower bound is well above 1 and the upper bound goes up to over 5. The odds-ratio analysis confirms the very significant risk of dying with a higher grade tumour.

Conclusion

Overall, patients with more positive lymph nodes and larger tumours were more likely to die by the end of follow-up. Taking Tamoxifen seems to have a slight protective effect. A higher tumour grade led to significantly higher mortality. Age and menopause status didn't show any effect on mortality status.