Chapter 5

Summarizing Data

5.1 Introduction

A **random variable** is a characteristic or measurement that we have for each observation.

- **Categorical** $\longrightarrow$ places observations into one of several groups or categories
  - **Nominal** $\longrightarrow$ assumes groups are made up of named categories
  - **Ordinal** $\longrightarrow$ assumes groups are made up of ordered categories
- **Quantitative** $\longrightarrow$ assumes numerical values
  - **Continuous** $\longrightarrow$ assumes the variable is double measured
  - **Discrete** $\longrightarrow$ assumes the variable is countable (usually integer measured)

Which graphical and numerical summaries we select to employ will largely be determined by how we classify the random variable. Figure 5.1.1 and Figure 5.1.2 provide an outline of the next sections as well as a nice reference for graphical display and numerical summary selection, respectively.

Statisticians have promoted the use of graphics to make statistical analyses and theory more palatable for over 200 years; for example, below is an excerpt from The Statistical Breviary (**?**).

> *"For no study is less alluring or more dry and tedious than statistics, unless the mind and imagination are set to work or that the person studying is particularly interested in the subject; which is seldom the case with young men in any rank in life."*

In this chapter we will discuss graphical displays of data as well as the numerical summaries, or statistics, that help us describe what we see in the graph numerically.

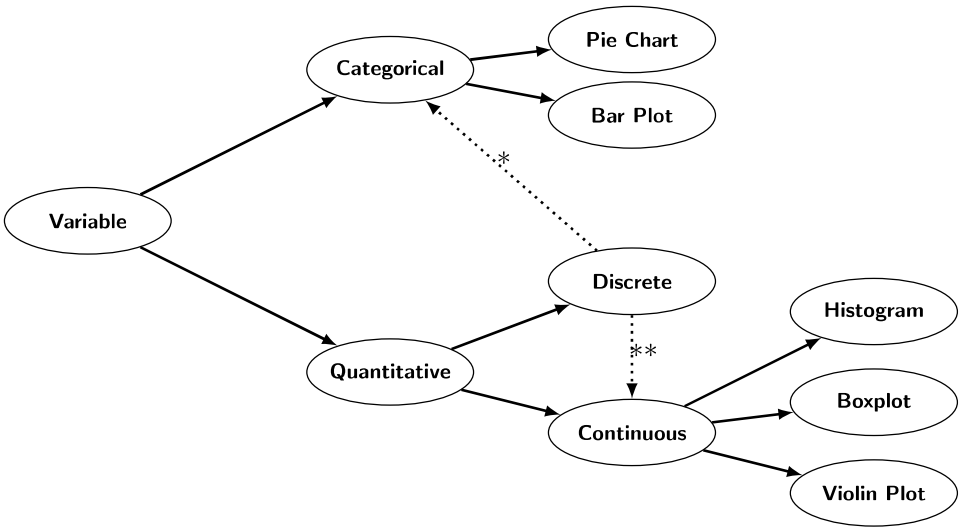

Figure 5.1.1: Flow chart for graphical display based on datatype. * and ** denote that we can treat a discrete variable as categorical when it has few observable values and as a continuous variable when it has very many observable values.

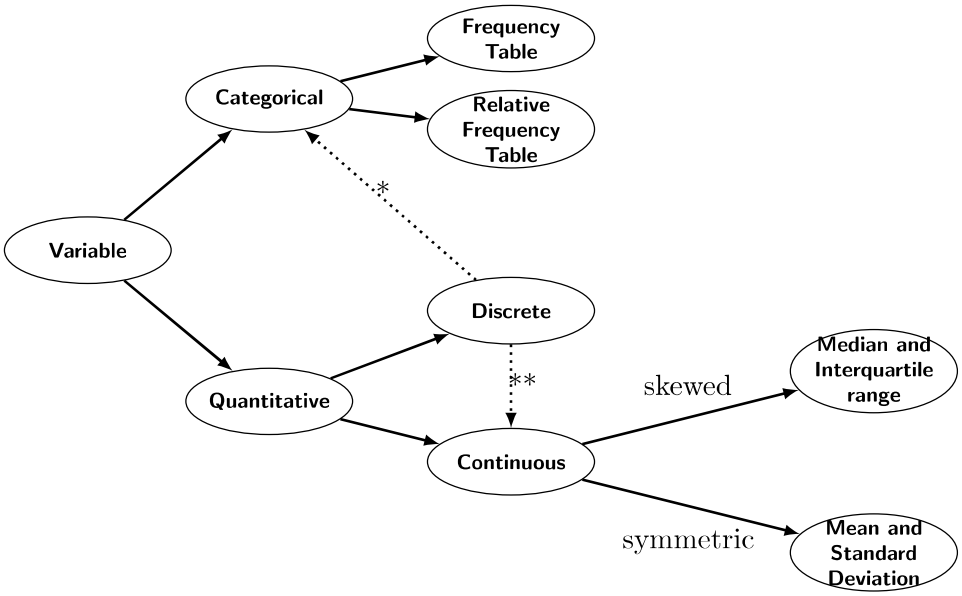

Figure 5.1.2: Flow chart for numerical summaries based on datatype. * and ** denote that we treat a discrete variable as categorical when it has few observable values and as a continuous variable when it has very many observable values.

5.2 Categorical Variables

5.2.1 Numerical Summary

Example 5.1. Recall the Ship1 data from Example 2.3. One of the variables recorded for each mouse was Ship1 status type, which had possible categories of "-/-", "+/-", "+/+".

Ship1 status is categorical because the values listed above identify nominal categories. In our data file, which contains these data, the following codings are used:

$$1 = \text{“-/-”}; \; 2 = \text{“+/-”}; \; 3 = \text{“+/+”}.$$

These codings are numerical, but they do not have a physical meaning; they simply keep track of which category is which. We can ask R for this mapping as follows.

```
> ship1dat<-read.table(file = "https://cipolli.com/students/data/ship1.txt",
+                      sep=",",header=TRUE)
> ship1dat$Ship1
[1] +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/+ +/- +/- +/- +/- +/-
[20] +/- +/- +/- +/- +/- +/- +/- +/- +/- +/- +/- +/- +/- +/- -/- -/- -/- -/- -/- -/-
[39] -/- -/- -/- -/- -/- -/- -/- -/- -/- -/-
Levels: -/- +/- +/+
> as.numeric(ship1dat$Ship1)
[1] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1
[40] 1 1 1 1 1 1 1 1
```

When we summarize this data, we will see the labels “-/-”, “+/-”, “+/+” in this order that is specified by their integer value. This is important because when we ask for numerical or graphical summaries by a factor's value it will be provided in that order. We can change the levels to change the labels or reorder the levels to change the order.

Definition 5.2. The distribution of a variable tells us what values the variable takes and how often it takes these values. A frequency or relative frequency table can be used to show the distribution of a categorical variable.

We will often attempt to summarize or describe the **distribution** of data. Our main approach for doing this will be to create and interpret graphical and numerical summaries of data.

Below, we create a frequency table and a relative frequency table for Ship1 status.

```
> table(ship1dat$Ship1) #frequency table
-/- +/- +/+
16  18  14
> prop.table(table(ship1dat$Ship1)) #relative frequency table
-/-       +/-       +/+
0.3333333 0.3750000 0.2916667
```

The relative frequency table provides the count of observations in each category while the relative frequency table provides the proportion of observations in each category. In general, we will prefer the relative frequency table because the proportion provides context of the observation with respect to all the data.

5.2.2 Graphical Summary – Bar Plots

A bar graph is a visual display used to depict the frequency or relative frequency tables that describe a categorical random variable.

Example 5.3. Recall the Ship1 data from Example 2.3. Both the frequency and relative frequency bar plots for the Ship1 observations are created in R and displayed in Figure 5.2.3

```
> par(mfrow=c(1,2)) #Graphics setting 1 row, 2 columns of plots
> #Frequency Bar Plot
> barplot(table(ship1dat$Ship1),              #which table to plot
+   main="Ship1 Status of Mice",              #title of graph
+   xlab="Ship1 Status", ylab="Frequency",    #x and y axes labels
+   col="lightblue")                          #bar colors
> abline(h=0)  #adds a line for the x-axis
> #Relative Frequency Bar Plot
> barplot(prop.table(table(ship1dat$Ship1)), #which table to plot
+   main="Ship1 Status of Mice",              #title of graph
+   xlab="Ship1 Status", ylab="Relative Frequency",   #x and y axes labels
+   col="lightblue")                          #bar colors
> abline(h=0)  #adds a line for the x-axis
```

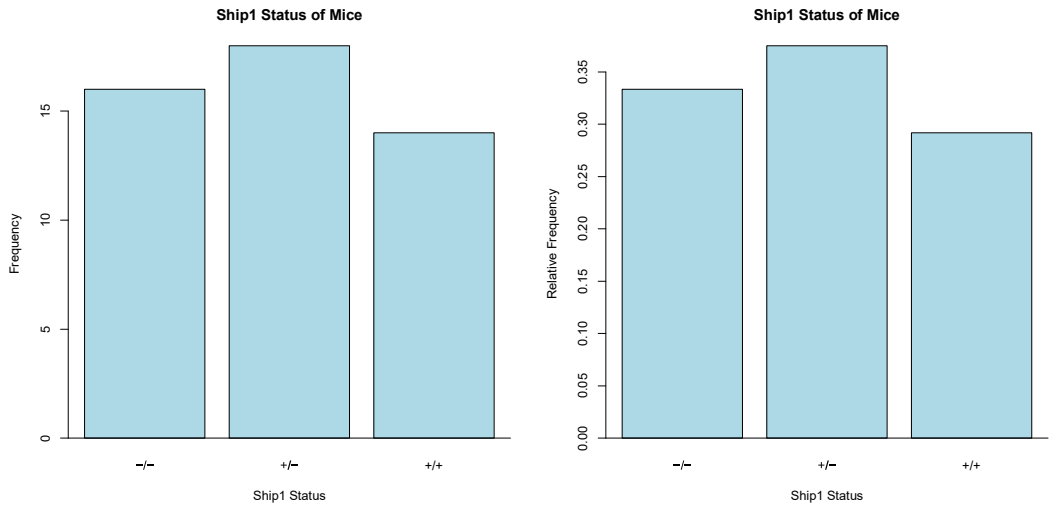

Figure 5.2.3: Ship1 data. Bar graphs of Ship1 status types for 48 mice. Left: Counts. Right: Proportions. Both figures were created using R.

Remark: The only difference in the figures above is the scale used for the vertical axis. The bar plot created using the frequency table has the frequency (count) on the y axis and the bar plot created using the relative frequency table has the relative frequency (proportion) on the y axis. Graphically, the difference is less important because we can assess how a count observations fits into all observations.

5.2.3 Graphical Summary – Pie Charts

The distribution of a categorical variable can also be shown using a pie chart.

Example 5.4. Recall the Ship1 data from Example 2.3. Both the frequency and relative frequency bar plots for the Ship1 observations are created in R and displayed in Figure 5.2.3The pie chart in Figure 5.2.3 is created in R using the code below.

```
> par(mfrow=c(1,1)) #Reset graphics setting
> labels<-paste(levels(ship1dat$Ship1)," (",  #labels
>   round(prop.table(table(ship1dat$Ship1))*100,1),"%)", #percentage
>   sep="") #don't separate text
> pie(table(ship1dat$Ship1), #which table to plot
+   labels=labels,          #add labels to pieces of the pie
+   main="Ship1 Status of Mice", #title of graph
)
```

Remark: We used paste to create the labels for each piece of the pie – the Ship1 status and the percentage of such observations.

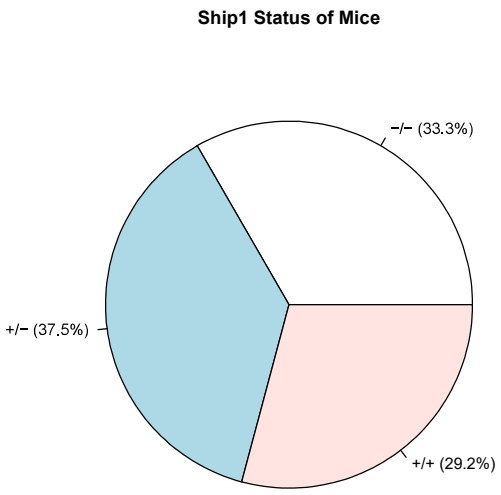

Figure 5.2.4: Ship1 data. Pie chart of Ship1 status types for 48 mice created using R.

Remark: Pie charts are generally only used when the category proportions add to 100% (i.e., so the angles add to 360 degrees).

5.2.4 Summary of R Commands

Table 5.2.1 summarizes the operators used in this section.

Operator	Functionality
table(...)	creates a frequency table
prop.table(...)	creates a relative frequency table
par(mfrow=c(r,c))	makes the graphics environment a $r \times c$ grid
barplot(...)	creates a barplot for a table of data
abline(h=0)	draws a horizontal line at $y = 0$
pie(...)	creates a pie chart for a table of data

Table 5.2.1: A list of basic operators for summarizing categorical data in R.

5.3 Continuous Quantitative Variables

5.3.1 Numerical Summary

In this section, we discuss statistics that describe how quantitative random variables are distributed – the center and spread. This section is only relevant for quantitative variables. We will avoid excessive hand calculations, relying on R whenever possible.

Definition 5.5. With a sample of observations $x_1, x_2, ..., x_n$, the **sample mean** $\overline{x}$ (pronounced "x-bar") is given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}.$$

In other words, the sample mean is the **average** of the n values $x_1, x_2, ..., x_n$. This is a measure of center thought of as the "balancing point" of a distribution.

Definition 5.6. With a sample of observations $x_1, x_2, ..., x_n$, the **sample standard deviation** s is given by

$$s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}.$$

The standard deviation is a measure of spread thought of as "an average distance from the mean."

Definition 5.7. The **median** M is the midpoint of a distribution. Half of the observations are smaller; half are larger. The median measures the "center" of a distribution.

Definition 5.8. The **first quartile** Q_1 is the median of the lower half of the observations. The **third quartile** Q_3 is the median of the upper half.

Definition 5.9. The **minimum** (Min) is the smallest observation in a dataset. The **maximum** (Max) is the largest.

Definition 5.10. The **5-number summary** of a dataset consists of these 5 values:

$$\text{Min}, Q_1, M, Q_3, \text{Max}.$$

Example 5.11. We calculate the numerical summaries for the frequencies of iNKT cells from the Ship1 data in R as follows.

```
> mean(ship1dat$iNKT)
[1] 0.8545833
> sd(ship1dat$iNKT)
[1] 0.3407748
> min(ship1dat$iNKT)
[1] 0.16
> quantile(x=ship1dat$iNKT,probs=0.25)
25%
0.615
> median(ship1dat$iNKT)
[1] 0.905
> quantile(x=ship1dat$iNKT,probs=0.75)
75%
1.1125
> max(ship1dat$iNKT)
[1] 1.61
> summary(ship1dat$iNKT)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1600  0.6150  0.9050  0.8546  1.1125  1.6100
> IQR(ship1dat$iNKT)
[1] 0.4975
```

We can also ask for the numerical summaries by group based on the factor Ship1 status.

```
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,mean)
-/-       +/-       +/+
0.5637500 0.9677778 1.0414286
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,sd)
-/-       +/-       +/+
0.3494162 0.2281268 0.2268804
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,min)
-/-  +/-  +/+
0.16 0.41 0.77
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,quantile,probs=0.25)
-/-    +/-    +/+
0.3175 0.8900 0.8825
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,median)
-/-  +/-  +/+
0.52 0.97 0.97
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,quantile,probs=0.75)
-/-    +/-    +/+
0.7175 1.1175 1.1575
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,max)
-/-  +/-  +/+
1.44 1.29 1.61
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,summary)
$`-/-`
```

Statistic	"-/-" Ship1 Status	"+/-" Ship1 Status	"+/+" Ship1 Status
Mean (SD)	0.56 (0.35)	0.97 (0.23)	1.04 (0.23)
Median (IQR)	0.52 (0.40)	0.97 (0.23)	0.97 (0.28)

Table 5.3.2: Ship1 data. Numerical summaries of the frequencies of iNKT cell populations by Ship1 status.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.1600  0.3175  0.5200  0.5637  0.7175  1.4400
$`+/-`
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.4100  0.8900  0.9700  0.9678  1.1175  1.2900
$`+/+`
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.7700  0.8825  0.9700  1.0414  1.1575  1.6100
> tapply(X=ship1dat$iNKT,INDEX=ship1dat$Ship1,IQR)
-/-    +/-    +/+
0.4000 0.2275 0.2750
```

These values gives us some comparative information across Ship1 types. We see that the measures of center, mean and median, are higher for the '+/+' and '+/-' Ship1 statuses compared to the '-/-' Ship1 status. We see the same pattern for spread, the standard deviation and interquartile range are lower for the '+/+' and '+/-' Ship1 statuses compared to the '-/-' Ship1 status, meaning the observations are less varied for those groups.

Table 5.3.2 provides measures of center and spread which provide empirical evidence that the frequencies of iNKT cell populations are higher for mice with "+/-" and "+/+" Ship1 status than mice with "-/-" Ship1 status. We also notice that the spread, or variability, of the frequencies of iNKT cell populations are lower for mice with "+/-" and "+/+" Ship1 status than mice with "-/-" Ship1 status.

While these statements are true in the sample, we can't simply extrapolate to the entire population of mice – we'll need to employ a statistical inference technique for that, which we'll cover in a later chapter. In the following subsections, we'll work on visualizing these differences.

5.3.2 Graphical Summary – Histograms

The distribution of a variable tells us (a) what values the variable takes and (b) how often it takes these values. Categorical variables take on just a few values (e.g., "-/-", "+/-", "+/+"). However, quantitative variables are numerical and can take on many values. Therefore, different graphical displays are needed.

Histograms appear similar to bar plots but they are quite different. A bar plot has a bar for each unique observation value, but a histogram has a bar for each collection of observations. This is why a histogram works well for continuous data that can take on infinitely many values – it is generated by creating observation categories by splitting up the number line. This discretization, however, can over or under simplify the distribution depending on some of our graphing choices. Because of this, we will generally plot a density line over the histogram. This approach ameliorates the discreteness of the histogram by plotting a smoothed version of the data continuously, without binning observations.

Example 5.12. Histograms and densities for the frequencies of iNKT cells from the Ship1 data are created using the base functionality of R are displayed in Figure 5.3.5, and created using the following R code.

```
> par(mfrow=c(1,2)) #Graphics setting 1 row, 2 columns of plots
> hist(ship1dat$iNKT, #which data to plot
+   #title of graph
+   main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+   xlab="Frequencies of iNKT Cells", ylab="Frequency",   #x and y axes labels
+   col="lightblue") #bar colors
> abline(h=0)  #adds a line for the x-axis
> hist(ship1dat$iNKT, #which data to plot
+   #title of graph
+   main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+   xlab="Frequencies of iNKT Cells",ylab="Density", #x and y axes labels
+   col="lightblue", #bar colors
+   probability=TRUE) #probability density, not frequency
> lines(density(ship1dat$iNKT),col="red",lwd=2) #add kernel density estimate
> abline(h=0) #adds a line for the x-axis
```

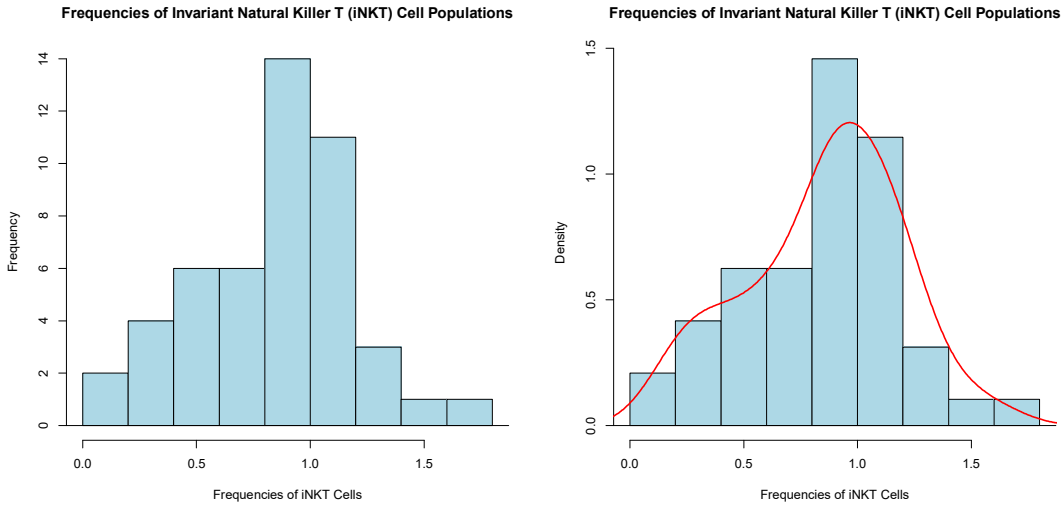

Figure 5.3.5: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using base functions in R. **Left:** Frequency (count); **Right:** Density with kernel density estimate superimposed in red.

Grouping observations creates another task for us – choosing bin width. Think about the bins as pixels on a screen – smaller bin widths yield more detail and larger bin widths yield less detail. Back in the day, we would have to define our own bins, count how many observations are in each bin and see if that provided a reasonable summary of the data. Now, we lean on R to do the work and we can ask for different bin widths if we decide we don't like what R selects automatically.

We can specify the bin-width used by the base functionality of R by specifying the `breaks` argument of `hist()`. Below, we created three histograms with varying bin-width which can be seen in Figure 5.3.6.

```
> par(mfrow=c(1,3))
> hist(ship1dat$iNKT, #which data to plot
+      #title of graph
+      main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+      xlab="Frequencies of iNKT Cells",ylab="Density", #x and y axes labels
+      col="lightblue", #bar colors
+      probability=TRUE, #probability distribution, not frequency
+      ylim=c(0,2.25), #set limits of y axis
+      breaks=seq(0,2,0.05)) #specify bins
> lines(density(ship1dat$iNKT),col="red",lwd=2) #add kernel density estimate
> abline(h=0) #adds a line for the x-axis
> hist(ship1dat$iNKT, #which data to plot
+      #title of graph
+      main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+      xlab="Frequencies of iNKT Cells",ylab="Density", #x and y axes labels
+      col="lightblue", #bar colors
+      probability=TRUE, #probability distribution, not frequency
+      ylim=c(0,2.25), #set limits of y axis
+      breaks=seq(0,2,00.2)) #specify bins
> lines(density(ship1dat$iNKT),col="red",lwd=2) #add kernel density estimate
> abline(h=0) #adds a line for the x-axis
> hist(ship1dat$iNKT, #which data to plot
+      #title of graph
+      main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+      xlab="Frequencies of iNKT Cells",ylab="Density", #x and y axes labels
+      col="lightblue", #bar colors
+      probability=TRUE, #probability distribution, not frequency
+      ylim=c(0,2.25), #set limits of y axis
+      breaks=seq(0,2,1)) #specify bins
> lines(density(ship1dat$iNKT),col="red",lwd=2) #add kernel density estimate
> abline(h=0) #adds a line for the x-axis
```

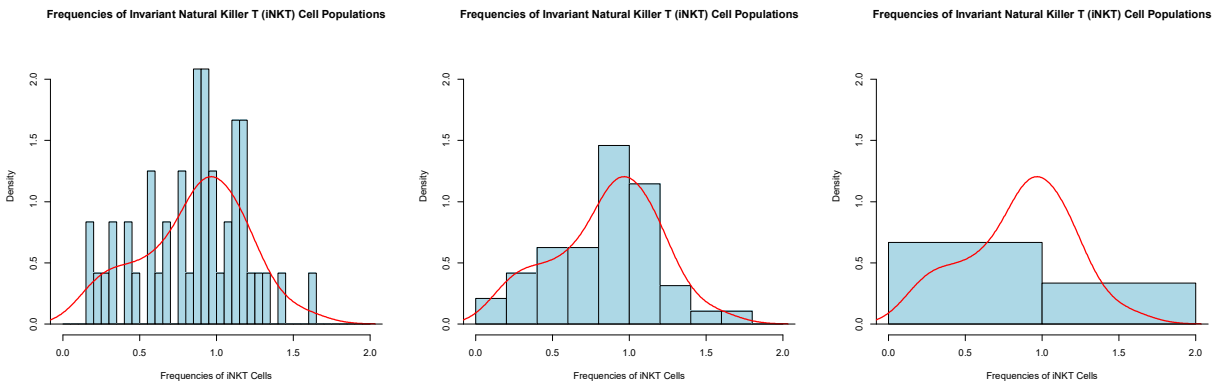

Figure 5.3.6: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using base functions in R. **Left:** Bin width of 0.05; **Center:** Bin width of 0.2; **Right:** Bin width of 1.

Our advice: When making a histogram in R, try the default settings first. If you do not like this,

use trial and error to get the appearance that best displays the distribution. There is no "correct" way to pick the interval widths, but there are certainly bad ways to do it.

We will lean on R to do all of the work, which makes constructing histograms easy. Interpreting them is more important. Histograms are used to show the distribution of recorded observations for a quantitative variable (like birth weight). If the observations we have are from a sample, then the histogram presents an impression of the underlying distribution for the variable in the population. Therefore, by interpreting characteristics we see in the histogram, we are interpreting what may be "going on" in the larger population of individuals.

Interpretation: We will focus on the following characteristics when we examine and describe histograms:

- Center: Where does the center of the distribution fall approximately?
  - mean ($\overline{x}$)
  - median ($\widehat{m}$)
- Spread: How much variation is in the distribution? How spread out is it? What is the range of possible values?
  - standard deviation (s)
  - variance (s^2)
  - interquartile range (IQR)
  - range
- Shape: What type of shape does the distribution have?
  - Left skewed – long left tail
  - Symmetric – symmetric
  - Right skewed – long right tail
  - Bimodal – two modes
- Deviations from the overall pattern (e.g., outliers, etc.)
  - An outlier is an individual observation that falls outside the overall pattern of the distribution.
    - **Symmetric:** An observation more than 3 standard deviations away from the mean
    - **Skewed:** An observation more than 1.5 IQR below Q_1 or above Q_3

Example 5.13. We can use the data from Anderson et al. (2015) to practice interpreting histograms by plotting histograms of the frequencies of iNKT by Ship1 status in Figure 5.3.7.

```
> par(mfrow=c(1,3))
> hist(ship1dat$iNKT[ship1dat$Ship1=="+/+"], #which data to plot
+   #title of graph
+   main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations
+        for Mice with +/+ Ship1 Status",
+   xlab="Frequencies of iNKT Cells", ylab="Frequency", #x and y axes labels
+   col="lightblue", #bar colors
+   probability=TRUE) #probability distribution, not frequency
```

```
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="+/+"]),col="red",lwd=2)
> hist(ship1dat$iNKT[ship1dat$Ship1=="+/-"], #which data to plot
+   #title of graph
+   main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations
+        for Mice with +/- Ship1 Status",
+   xlab="Frequencies of iNKT Cells", ylab="Frequency", #x and y axes labels
+   col="lightblue", #bar colors
+   probability=TRUE) #probability distribution, not frequency
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="+/-"]),col="red",lwd=2)
> hist(ship1dat$iNKT[ship1dat$Ship1=="-/-"], #which data to plot
+   #title of graph
+   main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations
+        for Mice with -/- Ship1 Status",
+   xlab="Frequencies of iNKT Cells", ylab="Frequency", #x and y axes labels
+   col="lightblue", #bar colors
+   probability=TRUE) #probability distribution, not frequency
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="-/-"]),col="red",lwd=2)
```

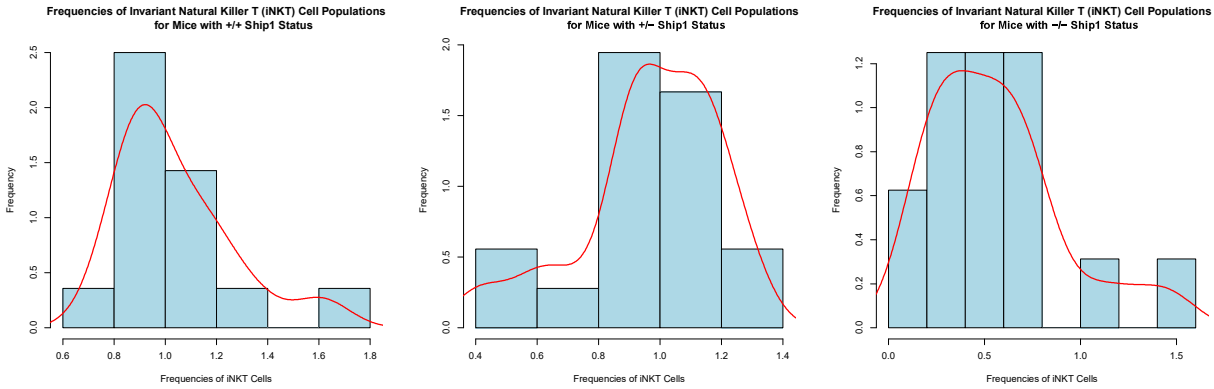

Figure 5.3.7: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using base functions in R. **Left:** "+/+" Ship 1 status; **Middle:** "+/-" Ship 1 status; **Right:** "-/-" Ship 1 status

Interpretation:

- **Center:**
  - The distribution of "+/+" is centered near 1
  - The distribution of "+/-" is centered near 1
  - The distribution of "-/-" is centered near 0.50
- Most of the iNKT measurements are between 0 and 1.75 (a rough range).
- **Shape**
  - The distribution of "+/+" is slightly right skewed

  - The distribution of "+/-" is slightly left skewed
  - The distribution of "-/-" is right skewed

- There are no obvious outliers.

Discussion: The histogram shows the distribution of the observed values (i.e., the iNKT observations for the 48 mice in the sample) by Ship1 status. If the sample is representative of a larger population (e.g., all mice), then the histogram may be "approximating" a smooth curve that describes this population represented by the superimposed density. We call this smooth curve the population density curve. This curve describes how the variable is distributed in the population; we will explore several distributions and their associated density curves in the next chapter.

Example 5.14. Here, we can plot and compare these data by putting all the histograms on the same plot. We, in Figures 5.3.8 (left) and 5.4.16 (left), that there is much overlap and the graph is difficult to read. Instead, we can look at the estimated kernel density estimates, in Figures 5.3.8 (right) and 5.4.16 (right), which are less cluttered and easier to read.

The plots in Figure 5.3.8 are created using base functions in `R` as below.

```
> library("grDevices") #for transparent coloring
> par(mfrow=c(1,2))
> #Left Plot
> hist(ship1dat$iNKT[ship1dat$Ship1=="+/+"], #which data to plot
+      #title of graph
+      main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations
+            by Ship1 Status",
+      xlab="Frequencies of iNKT Cells", ylab="Frequency", #x and y axes labels
+      col=adjustcolor("lightblue",alpha.f = 0.5), #bar colors, alpha controls transparency
+      probability=TRUE, #probability density, not frequency
+      xlim=c(0,2), #set domain
+      ylim=c(0,3)) #set range
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="+/+"]),col="blue",lwd=2)
> hist(ship1dat$iNKT[ship1dat$Ship1=="+/-"], #which data to plot
+      col=adjustcolor("pink",alpha.f = 0.5), #bar colors, alpha controls transparency
+      probability=TRUE, #probability distribution, not frequency
+      add=TRUE) #add to current plot
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="+/-"]),col="red",lwd=2)
> hist(ship1dat$iNKT[ship1dat$Ship1=="-/-"], #which data to plot
+      col=adjustcolor("lightgrey",alpha.f = 0.5), #bar colors, alpha controls transparency
+      probability=TRUE, #probability distribution, not frequency
+      add=TRUE) #add to current plot
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="-/-"]),col="darkgrey",lwd=2)
> legend("topright",legend=c("+/+","+/+","+/-","+/-","-/-","-/-"), #location and labels
+        lty=c(1,NA,1,NA,1,NA), #lty 1 is a line, NA means not a line
+        lwd=rep(2,6), # sets line width for lines
+        pch=c(NA,15,NA,15,NA,15), #uses point type 15 to show filling
+        col=c("blue",adjustcolor("lightblue",alpha.f = 0.5), #colors
+              "red",adjustcolor("pink",alpha.f = 0.5),
```

```
+              "darkgrey",adjustcolor("lightgrey",alpha.f = 0.5)))
> #Right Plot
> plot(density(ship1dat$iNKT[ship1dat$Ship1=="+/+"]),type="l",col="blue",lwd=2,
+      main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations
+            by Ship1 Status",
+      xlab="Frequencies of iNKT Cells", ylab="Frequency", #x and y axes labels
+      xlim=c(0,2), #set domain
+      ylim=c(0,3)) #set range
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="+/-"]),col="red",lwd=2)
> #add kernel density estimate
> lines(density(ship1dat$iNKT[ship1dat$Ship1=="-/-"]),col="darkgrey",lwd=2)
> legend("topright",legend=c("+/+","+/-","-/-"), #location and labels
+        lty=rep(1,3), #lty 1 is a line, NA means not a line
+        lwd=rep(2,3), # sets line width for lines
+        col=c("blue","red","darkgrey"))
```

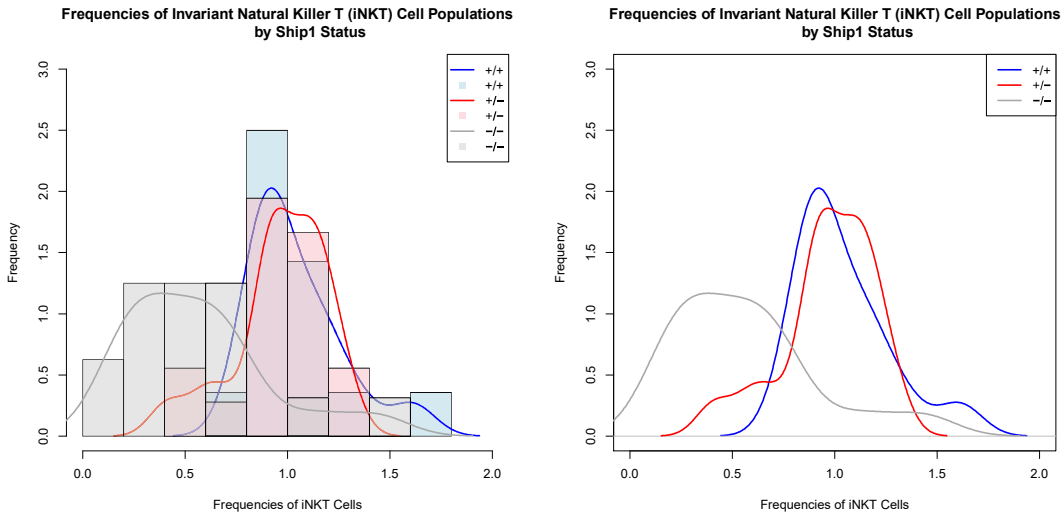

Figure 5.3.8: Ship1 data. Histograms of the frequency of iNKT populations by Ship1 status for 48 mice created using base functions in R. **Left:** Histograms and densities for the frequency of iNKT Ship1 populations by Ship1 status; **Right:** Densities for the frequency of iNKT Ship1 populations by Ship1 status

Interpretation:

- **Center:** The centers of the iNKT densities for mice with "+/-" and "+/+" Ship1 statuses are similar, while the center of the iNKT density for mice with "-/-" is lesser.
- Most of the iNKT measurements are between 0 and 1.75 (a rough range).
- **Shape**
  - The distribution of "+/+" is slightly right skewed
  - The distribution of "+/-" is slightly left skewed
  - The distribution of "-/-" is right skewed

- There are no obvious outliers.

This visual information is helpful to the researchers and matches the numerical summaries in Table 5.3.2 – if we want higher frequencies of iNKT cells in mice, having a Ship1 status with a "+" element appears to be of importance. This suggests that SHIP1 calibrates frequencies of iNKT cell populations in mice. If the behavior in mice is comparable to humans this might provide insight into where scientists can look to help fight against infections and enhance the immunity in humans. While we can see this difference visually, it is important to employ statistical methods that describe these differences. We will cover this in a later Chapter.

5.3.3 Graphical Summary – Boxplots

Definition 5.15. A **boxplot** is a graphical display that uses the 5-number summary.

- A central box spans the quartiles Q_1 and Q_3 – this represents the middle 50% of the data.
- A solid line within the box marks the median M.
- Lines extend from the box to the minimum and maximum values.

Example 5.16. Boxplots for the Ship1 data is shown in Figure 5.3.9, which are created using the base functionality of `R`, as follows.

```
> #Base
> par(mfrow=c(1,1)) #Reset graphics environment
> boxplot(ship1dat$iNKT, #which data to plot
+         #title of graph
+         main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+         xlab="", ylab="Frequencies of iNKT Cells", #x and y axes labels
+         col="lightblue") #box color
> #add points
> points(jitter(rep(1,length(ship1dat$iNKT)),amount = .2),ship1dat$iNKT,)
```

Remark: The `jitter()` function adds horizontal noise to the points so we can observe the observations more clearly than if they were plotted on one vertical line.

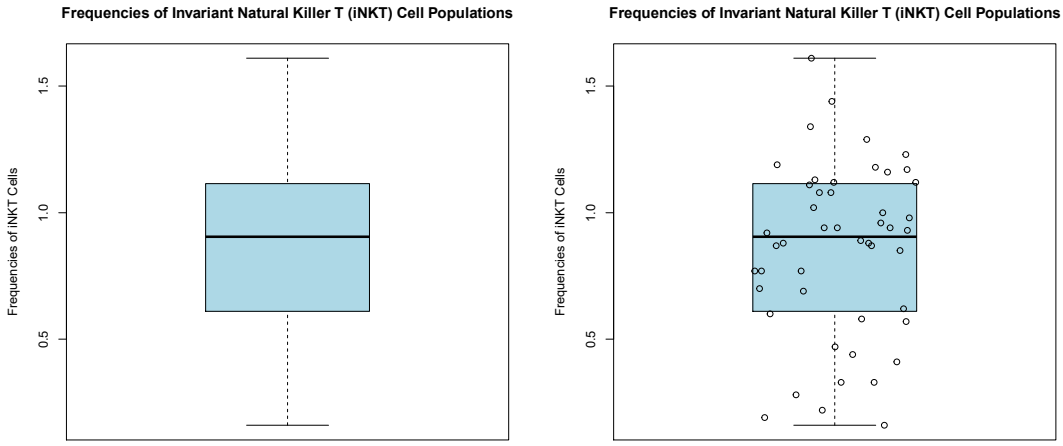

Figure 5.3.9: Ship1 data. Boxplot of the frequenices of iNKT cells for $n = 48$ mice created using base functions in R. **Left:** without points **Right:** with points (with horizontal noise).

Example 5.17. Just as with histograms, we may want to plot several boxplots together for each level of a factor variable so that we can visually display differences in the frequencies of iNKT cells across Ship1 deletion types. The boxplot in Figure 5.3.10 was created using base functions in R as follows.

```
> par(mfrow=c(1,1))
> boxplot(iNKT~Ship1, data=ship1dat, #frequencies of iNKT by Ship1 status
+       #title of graph
+       main="Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+       xlab="Frequencies of iNKT Cells", ylab="Frequency",   #x and y axes labels
+       col="lightblue") #bar colors
```

Remark: The `iNKT~Ship1` syntax tells R we want a boxplot of iNKT observations grouped by Ship1 status.

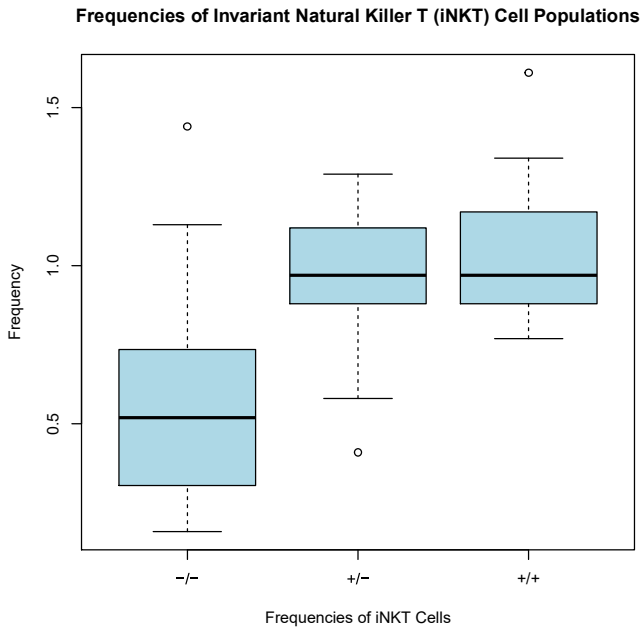

Figure 5.3.10: Ship1 data. Boxplot of the frequenices of iNKT cells for $n = 48$ mice created using base functions in R..

Remark: The default behavior in R is to mark outliers using points. For example, we can see there are among the frequencies of iNKT observations for each Ship1 status.

Discussion: Graphing boxplots side by side (as in Figures 5.3.10 and 5.4.18) allows us to **compare** two data distributions. For example,

- Which Ship1 deletion has a larger median frequency of iNKT cells?
- Which Ship1 deletion has more variability (spread) in its scores?

Statistical inference: What do these two samples say about the larger populations of female and male students? Are you prepared to conclude that

- Does Ship1 deletion decrease frequencies of iNKT cells?

Just as we saw with the histograms, if we want higher frequencies of iNKT cells in mice having a Ship1 status with a "+" element appears to be of importance. Note that here, the discussion is based on the median instead of the mean, though the story stays the same. With the boxplots, we can clearly see one outlier in each group – this is not something we were able to clearly visualize using histograms.

5.3.4 Summary of R Commands

Table 5.3.3 summarizes the operators used in this section.

Operator	Functionality
mean(...)	provides the mean for a vector
sd(...)	provides the standard deviation for a vector
min(...)	provides the minimum for a vector
quantile(...)	provides the specified percentile for a vector
median(...)	provides the median for a vector
max(...)	provides the max for a vector
summary(...)	provides the five number summary and mean for a vector
IQR(...)	provides the interquartile rang for a vector
hist(...)	creates a histogram for a vector of data
lines(...)	adds a line to a plot
density(...)	computes kernel density estimate for a vector of data
legend(...)	creates a legend for a plot
boxplot(...)	creates a boxplot for data
points(...)	adds points to a plot
jitter(...)	adds noise to a vector

Table 5.3.3: A list of basic operators for summarizing quantitative data in R.

5.4 Using ggplot2

All the plots of this chapter can also be created using the "ggplot2" package (Wickham, 2016) for R, which is a library that creates some of the most elegant and customizable graphs. While the base functions create publication quality graphics, we will find the extension provided by "ggplot2" functionality quite appealing.

There is much discussion in the statistics community over which is better – base graphing or "ggplot2" – with many statisticians on each side. The stance we take here is that both produce good plots, and each method of producing graphs has a place where it is better – whether better means easier or more aesthetically pleasing.

For readers wondering why there's an entire section about learning to plot graphs we just plotted using the base functionality, we discuss ggplot2 because it is often the case that the base graphics (and even the default ggplot) is not done. We will spend a lot of time (1) labeling axes, (2) changing text size, (3) adding unit labels, (4) specifying legends, (5) changing colors, (6) adding titles, (7) resizing, and countless other tweaks.

The one nice thing about learning both methods for graphing – base R and "ggplot2" – is that we will find that sometimes, our approaches to graphing take a ton of work or yield ugly graphs. In either case, having options allows us to think about what we tools we have and what we need to

change to make the plot better. Having options, makes us less likely to accept mediocre graphics. While learning to program in new ways takes effort and investment, creating more compelling graphics will yield more convincing work as pointed out by **?**.

5.4.1 Introduction

The first part of creating a graphic using ggplot2 is to pass in the data that we want to create the plot with and specify which variable maps onto the x and y axes as well as other visual properties. We can also specify `fill` which groups observations we may want to demonstrate in the plot. This is done as follows in "ggplot2."

```
> ggplot(data=dataframe, #Specify data to be used
+        #specify x and y and factor fill
+        aes(x=dataframe$x, y=dataframe$y, fill= dataframe$groupby))
```

Unlike base functionality, all the ggplots we will ask for will start with this call and we will add new aspects and customizations; see Table 5.4.4.

Operator	Functionality
+ geom_bar()	plots a bar plot
+ geom_histogram()	plots a histogram
+ geom_density()	plots a smoothed density estimate
+ geom_freqpoly()	plots an unsmoothed density estimate
+ geom_boxplot()	plots a boxplot
+ geom_jitter()	plots points with noise
+ geom_violin()	plots a violin plot

Table 5.4.4: How to create common types of plots using ggplot2.

We can specify axes labels and titles as described in Table 5.4.5.

Operator	Functionality
+ facet_grid()	splits graph pane by `fill` levels
+ coord_polar()	switches to polar coordinate
+ coord_flip()	switches x and y axes
+ theme_bw()	white background with grid lines
+ theme_classic()	white background without grid lines (similar to base)
+ theme_minimal()	minimal theme
+ geom_hline(...)	adds a horizontal line to the plot
+ scale_fill_brewer(...)	switches ggplot to a new color palette
+ scale_fill_manual(...)	manually specify colors by `fill` levels
+ geom_text(...)	add text annotations to a graph

Table 5.4.5: How to change some aspects plots created using ggplot2.

We can specify axes labels and titles as described in Table 5.4.6.

Operator	Functionality
+ xlab("...")	labels x axis
+ ylab("...")	labels y axis
+ ggtitle(title="...",subtitle="...")	Adds a title or subtitle

Table 5.4.6: How to label plots created using ggplot2.

In the following subsection we revisit the base plots created thus far and recreate them with "ggplot 2". We also introduce a new type of plot called a violin plot for quantitative data.

5.4.2 Graphical Summaries – Bar Plots

Revisiting Example 5.3: The ggplot2 bar plots are displayed in 5.4.11 and are created with the following `R` code.

```
> library(ggplot2)   #loads ggplot2 functionality
> library(gridExtra) #for arranging multiple plots
> ggdat<-data.frame(table(ship1dat$Ship1))
> colnames(ggdat)=c("Ship1","Count")
> p1<-ggplot(data=ggdat,aes(x=Ship1,y=Count)) + #tell ggplot which data to use
+   geom_bar(stat= "identity", #plot the count (no transformation needed)
+            color= "black", #bar outline color
+            fill= "lightblue")+ #bar colors
+   xlab( "Ship1 Status")+ #x axis label
+   ylab( "Frequency")+ #y axis label
+   ggtitle( "Ship1 Status of Mice") + #add title to plot
```

```
+   geom_hline(yintercept=0) +  #adds a line for the x-axis
+   theme_bw() #removes grey background
> #relative frequency bar plot
> ggdat<-data.frame(prop.table(table(ship1dat$Ship1)))
> colnames(ggdat)=c( "Ship1", "Proportion")
> p2<-ggplot(data=ggdat, aes(x=Ship1, y=Proportion)) + #tell ggplot which data to use
+   geom_bar(stat= "identity", #make y the realtive frequency (proportion)
+            color= "black", #bar outline
+            fill= "lightblue") + #bar colors
+   xlab("Ship1 Status")+ #x axis label
+   ylab("Relative Frequency")+ #y axis label
+   ggtitle( "Ship1 Status of Mice") + #add title to plot
+   geom_hline(yintercept=0)+ #adds a line for the x-axis
+   theme_bw() #removes grey background
> grid.arrange(p1,p2,ncol=2) #print plots side by side
```

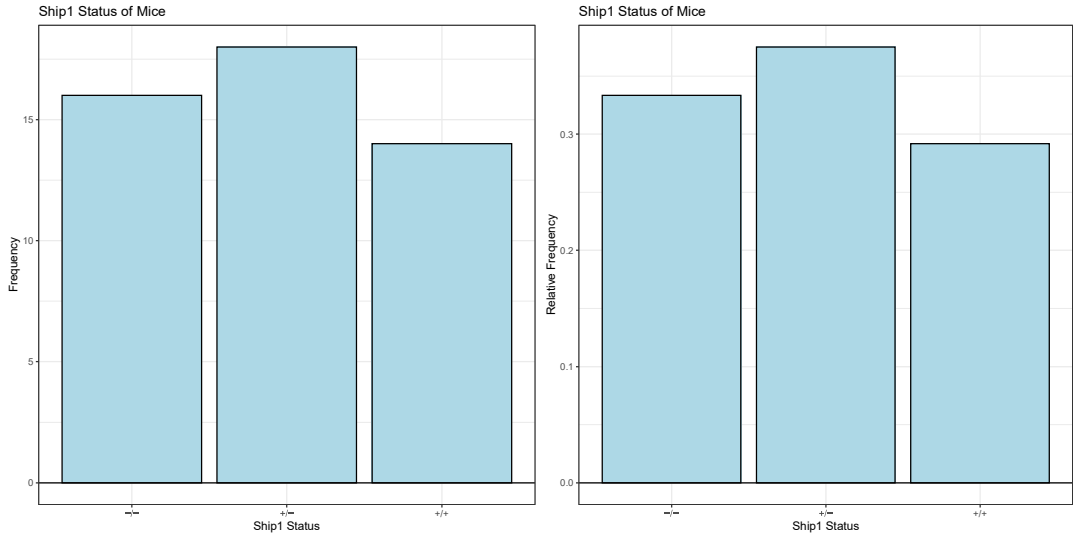

Figure 5.4.11: Ship1 data. Bar graphs of Ship1 status types for 48 mice. Left: Counts. Right: Proportions. Both figures were created using ggplot in R.

Using "ggplot2" we can also flip the coordinates easily as seen in Figure 5.4.12, created using the R code below.

```
> #We can flip these plots
> p1.flipped<-p1+ coord_flip()
> p2.flipped<-p2+ coord_flip()
> grid.arrange(p1.flipped,p2.flipped,ncol=2) #print both plots
```

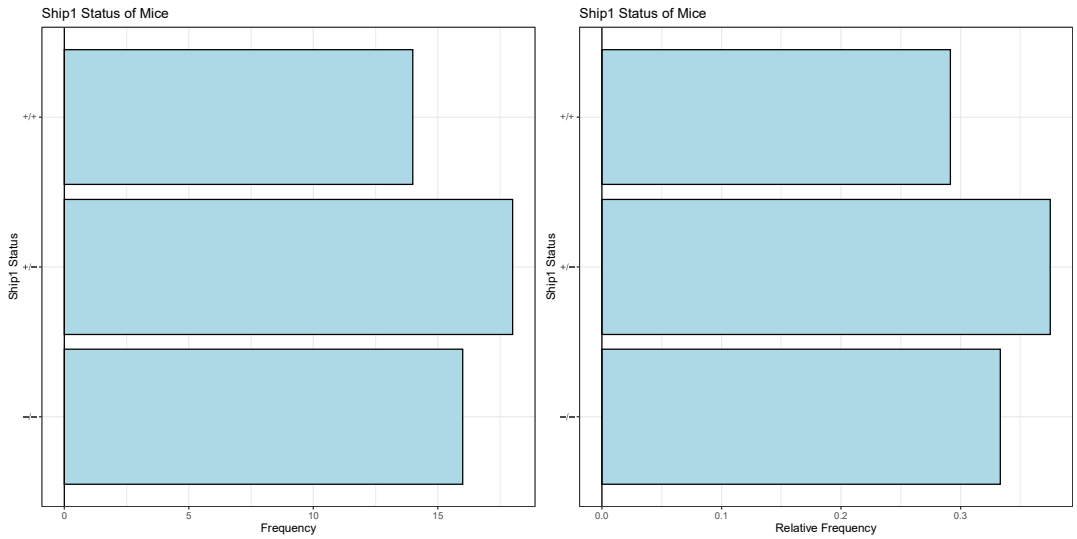

Figure 5.4.12: Ship1 data. Bar graphs of Ship1 status types for 48 mice. Left: Counts. Right: Proportions. Both figures were created using ggplot in R.

5.4.3 Graphical Summaries – Pie Charts

Revisiting Example 5.4: The ggplot2 pie chart is displayed in Figure 5.4.13 and created with the following R code.

```
> #Create Data for pie chart
> proportions<-(as.numeric(table(ship1dat$Ship1))/nrow(ship1dat))*100
> label.positions<-cumsum(proportions) - 0.5*proportions #Middle of each slice
> Ship1.levels<-levels(ship1dat$Ship1)
> (piechart.df<-data.frame(proportions,label.positions,Ship1.levels))
proportions label.positions Ship1.levels
1    33.33333        16.66667          -/-
2    37.50000        52.08333          +/-
3    29.16667        85.41667          +/+
> #Create Plot
> #tell ggplot to use the data frame created above
> p3<-ggplot(data=piechart.df, aes(x="",y=proportions,fill=Ship1.levels)) +
+   geom_bar(stat="identity")  + #uses the y-values specified above
+   coord_polar("y", start=0)  + #tells ggplot to convert to a pie chart
+   scale_fill_brewer(palette="Blues") + #gives a nice light blue color
+   xlab("")                   + #blank x axis label
+   ylab("")                   + #blank y axis label
+   ggtitle("Ship1 Status of Mice") + #add title to plot
+   theme_bw() + #removes grey background
+   geom_text(aes(y=label.positions,label = round(proportions,1))) #adds labels
> p3
```

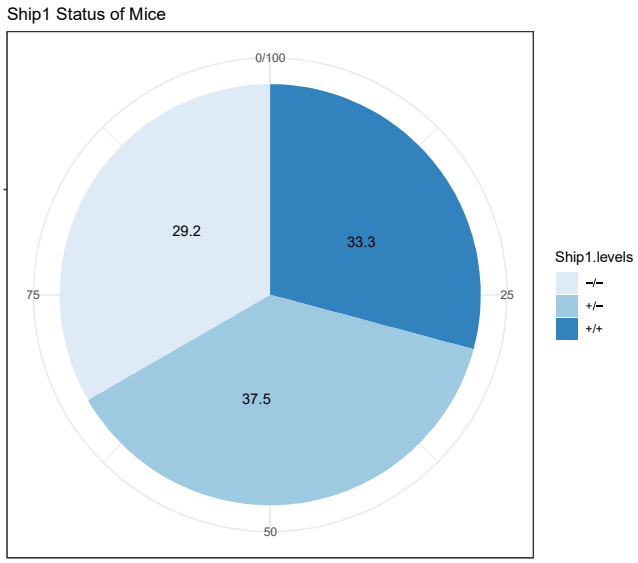

Figure 5.4.13: Ship1 data. Pie chart of Ship1 status types for 48 mice created using ggplot in R.

5.4.4 Graphical Summaries – Histograms

Revisiting Example 5.12: The ggplot2 histograms are displayed in Figure 5.4.14 and created with the following R code.

```
> p1<-ggplot(data=ship1dat, aes(x=iNKT)) + #which data to plot
+      geom_histogram(bins=10, #how many bins to use
+                       fill = "lightblue", color="black")  +  +      #color the histogram)
+      xlab("Frequencies of iNKT Cells")  + #x axis label
+      ylab("Frequency") + #y axis label
+      #add title to plot
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+      theme_bw() + #removes grey background
+      geom_hline(yintercept=0) #adds a line for the x-axis
> p2<-ggplot(data=ship1dat, aes(x=iNKT)) + #which data to plot
+      geom_histogram(aes(y = ..density..), #plots the density
+                     binwidth=density(ship1dat$iNKT)$bw, #sets bin width
+                     fill = "lightblue", color="black")  + +      #color the histogram
+      geom_density(fill="red", alpha = 0.2) +
+      xlab("Frequencies of iNKT Cells")  + #x axis label
+      ylab("Density") + #y axis label
+      #add title to plot
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+      theme_bw() + #removes grey background
+      geom_hline(yintercept=0) #adds a line for the x-axis
> grid.arrange(p1,p2,ncol=2) textcolorForestGreen#print both plots
```

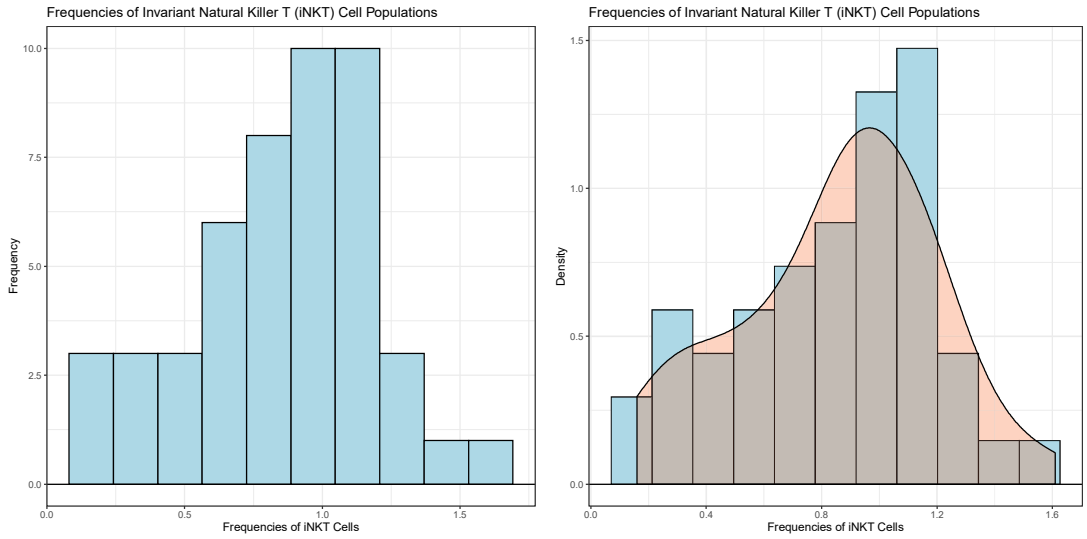

Figure 5.4.14: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using ggplot in R. **Left:** Frequency (count); **Right:** Density with kernel density estimate superimposed using a transparent red polygon.

These graphs are recreated using "ggplot" in R using the following code; these are displayed in Figure **??**,

```
> p1<-ggplot(data=ship1dat[ship1dat$Ship1=="+/+",], aes(x=iNKT)) + #which data to plot
+      geom_histogram(aes(y = ..density..), #plots the density
+                     binwidth=density(ship1dat$iNKT)$bw, #sets bin width
+                     fill = "lightblue", color="black")  + #color the histogram
+      geom_density(fill="red", alpha = 0.2) +
+      xlab("Frequencies of iNKT Cells") + #x axis label
+      ylab("Density") + #y axis label
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+              subtitle = "Mice with +/+ Ship1 Status") + #add title to plot
+      theme_bw() + #removes grey background
+      geom_hline(yintercept=0) #adds a line for the x-axis
> p2<-ggplot(data=ship1dat[ship1dat$Ship1=="+/-",], aes(x=iNKT)) + #which data to plot
+      geom_histogram(aes(y = ..density..), #plots the density
+                     binwidth=density(ship1dat$iNKT)$bw, #sets bin width
+                     fill = "lightblue", color="black")  + #color the histogram
+      geom_density(fill="red", alpha = 0.2) +
+      xlab("Frequencies of iNKT Cells") + #x axis label
+      ylab("Density") + #y axis label
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+              subtitle = "Mice with +/- Ship1 Status") + #add title to plot
+      theme_bw() + #removes grey background
+      geom_hline(yintercept=0)              #adds a line for the x-axis
> p3<-ggplot(data=ship1dat[ship1dat$Ship1=="-/-",], aes(x=iNKT)) + #which data to plot
+      geom_histogram(aes(y = ..density..), #plots the density
+                     binwidth=density(ship1dat$iNKT)$bw, #sets bin width
+                     fill = "lightblue", color="black") + #color the histogram
```

```
+       geom_density(fill="red", alpha = 0.2) +
+       xlab("Frequencies of iNKT Cells") + #x axis label
+       ylab("Density") + #y axis label
+       ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+               subtitle = "Mice with -/- Ship1 Status") + #add title to plot
+       theme_bw() + #removes grey background
+       geom_hline(yintercept=0) #adds a line for the x-axis
> grid.arrange(p1,p2,p3,ncol=3) #print all plots
```

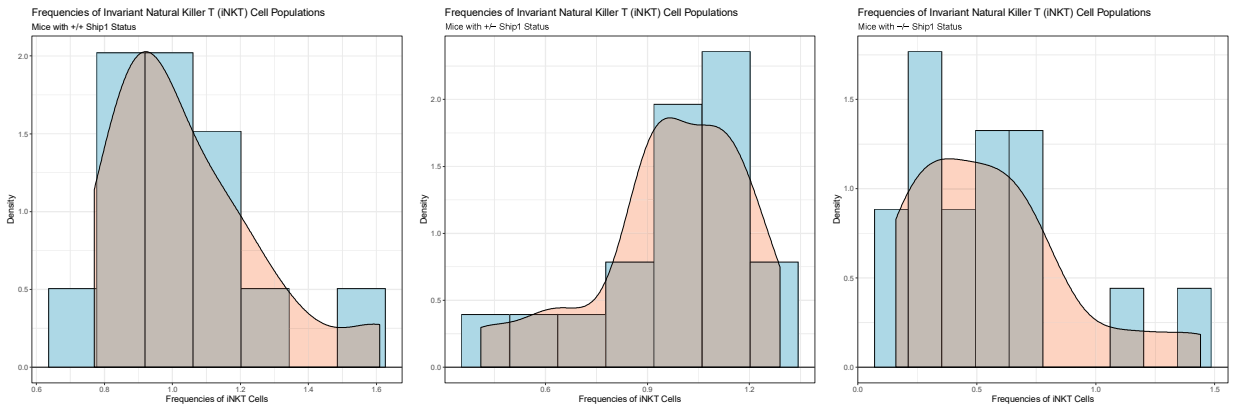

Figure 5.4.15: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using ggplot in R. **Left:** "+/+" Ship 1 status; **Middle:** "+/-" Ship 1 status; **Right:** "-/-" Ship 1 status

The plots in Figure 5.4.16 are created using "ggplot2" in R as below. We see that the "ggplot2" code is more concise. We can see that specifying the `fill` argument for `ggplot()` applies the subsequent commands like `tapply()`; i.e., a histogram with `fill` specified creates a histogram for each level of the factor specified to the `fill` argument.

```
> #Left Plot
> p1<-ggplot(data=ship1dat, aes(x=iNKT,fill=Ship1)) + #which data to plot
+       geom_histogram(aes(y = ..density..), #plots the density
+                      binwidth=0.2, #sets bin width
+                      color="black",alpha=0.25,
+                      position ="identity")+
+       geom_density(alpha = 0.2) +
+       xlab("Frequencies of iNKT Cells") + #x axis label
+       ylab("Density") + #y axis label
+       ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+               subtitle = "Mice with +/+ Ship1 Status") + #add title to plot
+       theme_bw() + #removes grey background
+       geom_hline(yintercept=0) + #adds a line for the x-axis
+       scale_fill_manual(values=c("red", "blue", "grey"))
> #Right Plot
> p2<-ggplot(data=ship1dat, aes(x=iNKT,fill=Ship1)) + #which data to plot
+       geom_density(alpha = 0.2) +
+       xlab("Frequencies of iNKT Cells")  + #x axis label
```

```
+      ylab("Density") + #y axis label
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations",
+              subtitle = "Mice with +/+ Ship1 Status") + #add title to plot
+      theme_bw() + #removes grey background
+      geom_hline(yintercept=0) + #adds a line for the x-axis
+      scale_fill_manual(values=c("red","blue","grey"))
> grid.arrange(p1,p2,ncol=2) #print both plots
```

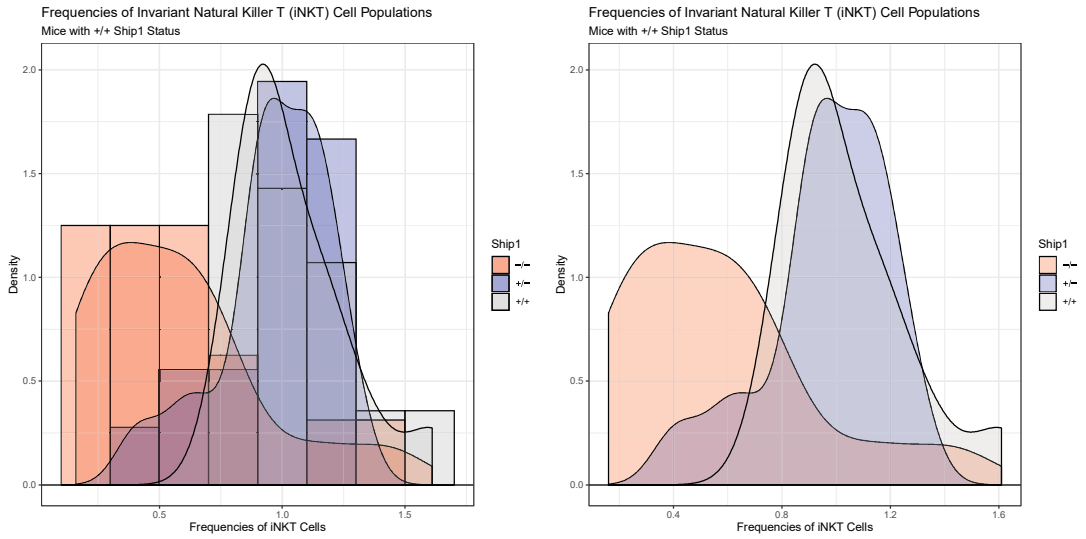

Figure 5.4.16: Ship1 data. Histograms of the frequency of iNKT Ship1 populations for 48 mice created using ggplot in R.

5.4.5 Graphical Summaries – Boxplots

The boxplots in Figure 5.4.17 were created using "ggplot2" in R as follows.

```
> p1<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+   geom_boxplot(fill="lightblue")+
+   xlab("") + #x axis label
+   ylab("Frequencies of iNKT Cells") + #y axis label
+   #add title to plot
+   ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+   theme_bw()  #removes grey background
> p2<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+   geom_boxplot(fill="lightblue")+
+   xlab("") + #x axis label
+   ylab("Frequencies of iNKT Cells") + #y axis label
+   #add title to plot
+   ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+   theme_bw()  #removes grey background
+   geom_jitter(position=position_jitter(0.2))
> grid.arrange(p1,p2,ncol=2)
```

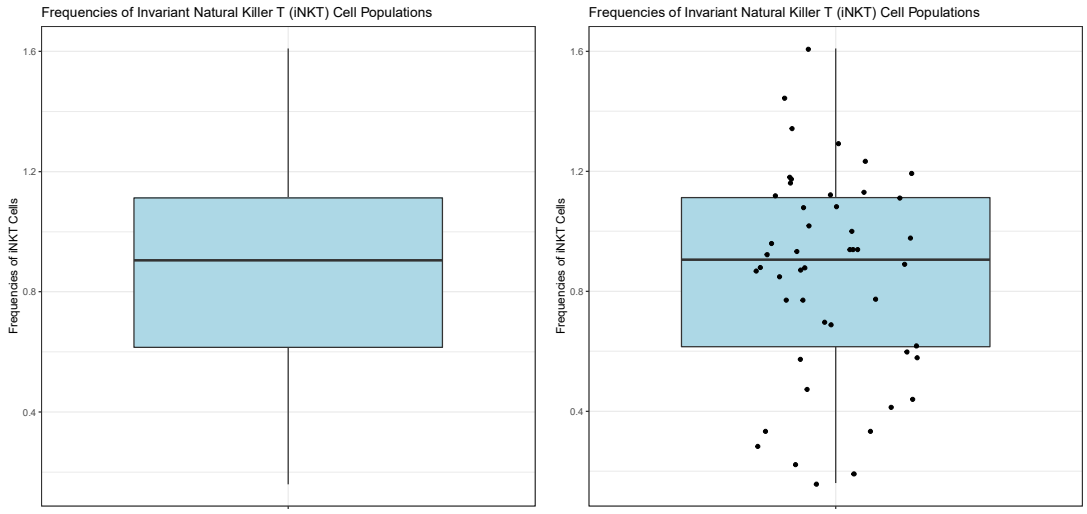

Figure 5.4.17: Ship1 data. Boxplot of the frequenices of iNKT cells for $n = 48$ mice created using ggplot2 in R. **Left:** without points **Right:** with points (with horizontal noise).

The boxplots in Figure 5.4.18 were created using "ggplot2" in R as follows.

```
> ggplot(ship1dat, aes(x=Ship1, y=iNKT)) +
+   geom_boxplot(fill="lightblue")+
+   xlab("Ship1")  + #x axis label
+   ylab("Frequencies of iNKT Cells") + #y axis label
+   #add title to plot
+   ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+   theme_bw() #removes grey background
```

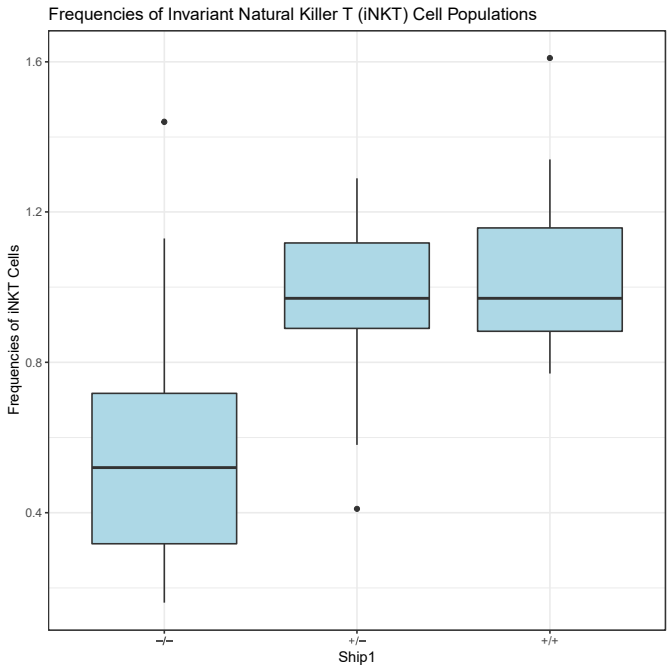

Figure 5.4.18: Ship1 data. Boxplot of the frequenices of iNKT cells for $n = 48$ mice created using ggplot2 in R.

5.4.6 Graphical Summaries – Violin Plots

Violin plots attempt to mix density plots and boxplots by mirroring a density to create a polygon. These show the shape of the distribution and we can add lines to the plots to denote the quartiles or superimpose a boxplot itself. These plots are especially useful when we have non-normal distributions.

The violin plots in Figure 5.4.19 were created using "ggplot2" in R as follows.

```
> p1<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+   geom_violin(fill="lightblue",
+               trim = FALSE,
+               draw_quantiles = c(0.25,0.5,0.75),
+               alpha = 0.5,
+               show.legend = FALSE)+
+   xlab("")  + #x axis label
+   ylab("Frequencies of iNKT Cells") + #y axis label
+   #y axis label
+   ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+   theme_bw()+ #removes grey background
+   geom_jitter(position=position_jitter(0.2))
> p2<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+   geom_violin(fill="lightblue",
+               trim = FALSE,
+               alpha = 0.5,
+               show.legend = FALSE)+
```

```
+   geom_boxplot(width = 0.25, fill="white") + #plot smaller boxplot inside violin
+   xlab("") + #x axis label
+   #y axis label
+   ylab("Frequencies of iNKT Cells") +
+   ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") + #add title to
+   theme_bw()  #removes grey background
> grid.arrange(p1,p2,ncol=2)
```

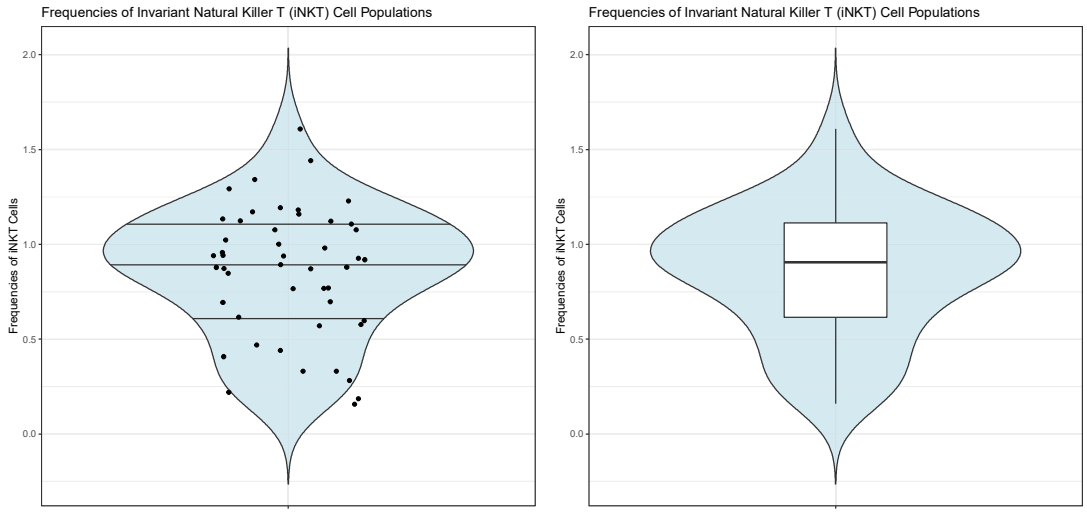

Figure 5.4.19: Ship1 data. Violin plot of the frequenices of iNKT cells for $n = 48$ mice created using ggplot2 in R.

The violin plots in Figure 5.4.20 were created using "ggplot2" in R as follows.

```
> p1<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+      geom_violin(fill="lightblue",
+                  trim = FALSE,
+                  draw_quantiles = c(0.25,0.5,0.75),
+                  alpha = 0.5,
+                  show.legend = FALSE)+
+      xlab("")  + #x axis label
+      ylab("Frequencies of iNKT Cells") + #y axis label
+      #add title to plot
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+      theme_bw()+  #removes grey background
+      geom_jitter(position=position_jitter(0.2))
> p2<-ggplot(ship1dat, aes(x="", y=iNKT)) +
+      geom_violin(fill="lightblue",
+                  trim = FALSE,
+                  alpha = 0.5,
+                  show.legend = FALSE)+
+      geom_boxplot(width = 0.25, fill="white") + #plot smaller boxplot inside violin
+      xlab("")  + #x axis label
```

```
+      ylab("Frequencies of iNKT Cells") + #y axis label
+      #add title to plot
+      ggtitle("Frequencies of Invariant Natural Killer T (iNKT) Cell Populations") +
+      theme_bw()  #removes grey background
> grid.arrange(p1,p2,ncol=2)
```

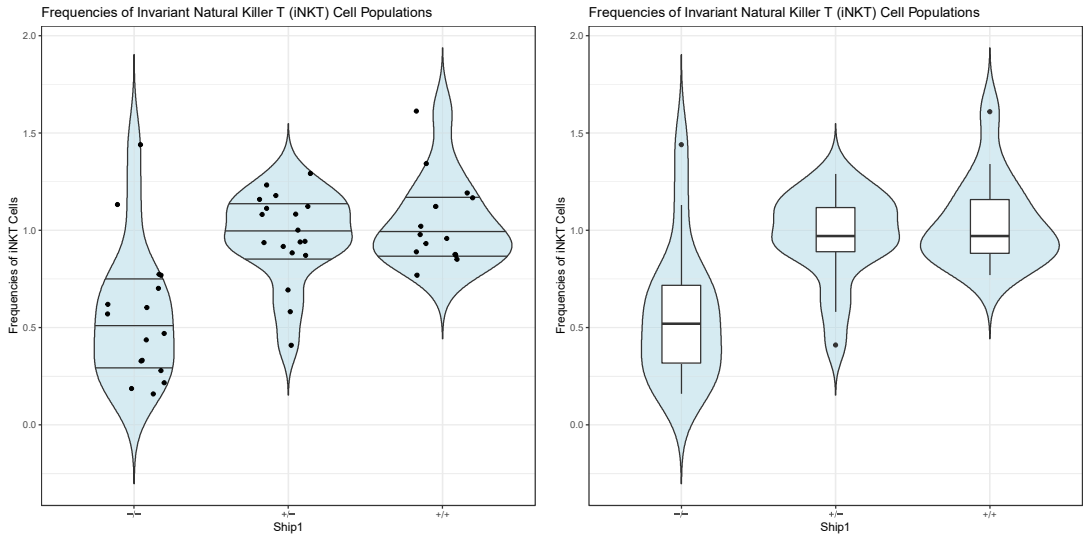

Figure 5.4.20: Ship1 data. Violin plot of the frequenices of iNKT cells for $n = 48$ mice created using ggplot2 in R.

From the violin plots, we can assess the center, spread, whether or not there are outliers as well as the shape of the distribution. These plots provide the analyses and empirical evidence of the histogram and boxplots together in one plot.