Statistics Chapter 3 Review - Exercises 5 and 7

Gun Death and Fish

Lucas Liona

Table of contents

Problem 5: Analysis of Gun Death Data

In 1989, Time magazine ran a 27-page article giving photographs of everybody who had been killed by a gun in the United States in a single week in May (464 people) and a little summary information for each. Professor Winston Cherry of the University of Waterloo compiled stem-and-leaf plots by categorizing the deaths into four classes and plotting the age at which each person had died (Units 2 | 3 = 23 years). Four people were not represented, as their ages were missing.

Data Entry and Preparation

First, let's load the data and create the necessary vectors:

```r
# Read the data
gun_deaths <- read.table("gundeath.txt", header = TRUE)

# Check the data structure
str(gun_deaths)
```

```
'data.frame':	460 obs. of  2 variables:
 $ type: chr  "accident" "accident" "accident" "accident" ...
 $ age : int  9 12 15 16 17 17 17 19 21 21 ...
```

```r
# Get a summary of the data
summary(gun_deaths)
```

```
     type                age       
 Length:460         Min.   : 2.00  
 Class :character   1st Qu.:24.00  
 Mode  :character   Median :33.00  
                    Mean   :37.17  
                    3rd Qu.:46.00  
                    Max.   :87.00  
```

```r
# Count total observations per type
table(gun_deaths$type)
```

```

    accident     homicide self.defense      suicide 
          21          195           22          222 
```

Stem-and-Leaf Plots

Now we can recreate the stem-and-leaf plots:

```r
# Split the data by type
accidents <- gun_deaths$age[gun_deaths$type == "accident"]
homicides <- gun_deaths$age[gun_deaths$type == "homicide"]
self_defense <- gun_deaths$age[gun_deaths$type == "self.defense"]
suicides <- gun_deaths$age[gun_deaths$type == "suicide"]

# Create stem-and-leaf plots one by one
cat("Homicides:\n")
```

Homicides:

```r
stem(homicides, scale = 0.5)
```

```
  The decimal point is 1 digit(s) to the right of the |

  0 | 2
  1 | 445666677778888888899999
  2 | 0000000000000011122222222223333333344444444455555555566666666667777+2
  3 | 000001111122222334444444445555566667777888888999
  4 | 0001122222244445566677788
  5 | 002244788
  6 | 01167
  7 | 9
```

```r
cat("\nAccidents:\n")
```

Accidents:

```r
stem(accidents, scale = 0.5)
```

```
  The decimal point is 1 digit(s) to the right of the |

  0 | 92567779
  2 | 1123378803447
```

```r
cat("\nSuicides:\n")
```

Suicides:

```r
stem(suicides, scale = 0.5)
```

```
  The decimal point is 1 digit(s) to the right of the |
```

```
  1 | 45566778888889999999
  2 | 00001111122222333444444444444555556666777777788888899
  3 | 000111111222333333333344445555556666677777788899
  4 | 000001112222344566677888
  5 | 01222223456666677899
  6 | 11122334445555555556677777799999
  7 | 0000011122334555667777899
  8 | 112346777
```

```r
cat("\nSelf-defense shootings:\n")
```

```
Self-defense shootings:
```

```r
stem(self_defense, scale = 0.5)
```

```
  The decimal point is 1 digit(s) to the right of the |

  2 | 12444689134455
  4 | 00667806
```

(a) Age Patterns in Different Types of Gun Deaths

Looking at the stem-and-leaf plots, we can observe several patterns about the ages at which people die from different types of gun deaths:

1. **Homicides**: Homicide victims are predominantly young adults, with the highest concentration in the late teens and twenties. The distribution is heavily skewed towards younger ages, with a peak around 18-25 years. We see a gradual decline in frequency with age, though homicides continue to occur across the age spectrum. The youngest victim is just 2 years old, highlighting that gun homicides affect even very young children.
2. **Accidents**: Accidental gun deaths appear more frequently among children, teenagers, and young adults. The distribution shows some concentration around the teenage years (15-19) and another cluster in the late 20s to early 30s. Accidents tend to affect younger people overall, with no accidental deaths recorded for people over 40 years old in this sample.

3. **Suicides**: The suicide distribution is quite different from the others and shows a bimodal pattern. There's a concentration among young adults (late teens through twenties), but also a substantial number of older adults, particularly in their 60s, 70s, and even 80s. This creates a much wider age distribution compared to homicides and accidents. The oldest person who died by firearm suicide was 87 years old.
4. **Self-defense**: Self-defense shootings involve primarily middle-aged adults, with most victims being in their 20s through 50s. The age distribution is more evenly spread across this range than the other categories, without the strong concentration in any particular age group seen in homicides or accidents.

(b) Better Arrangement of Plots

A better arrangement of the stem-and-leaf plots would be to align them horizontally rather than in a 2×2 grid, with the stems aligned so that the ages can be directly compared across all four categories. This would make it easier to compare the frequency of deaths at particular ages across the different types of gun deaths.

The ideal arrangement would be: - Place the stem values (age decades) in a single column on the left - Have four columns to the right, one for each type of death - This allows for direct visual comparison of the frequency and distribution of deaths by age across categories

(c) Five-Number Summaries and Box Plots

Let's compute the five-number summaries and create box plots:

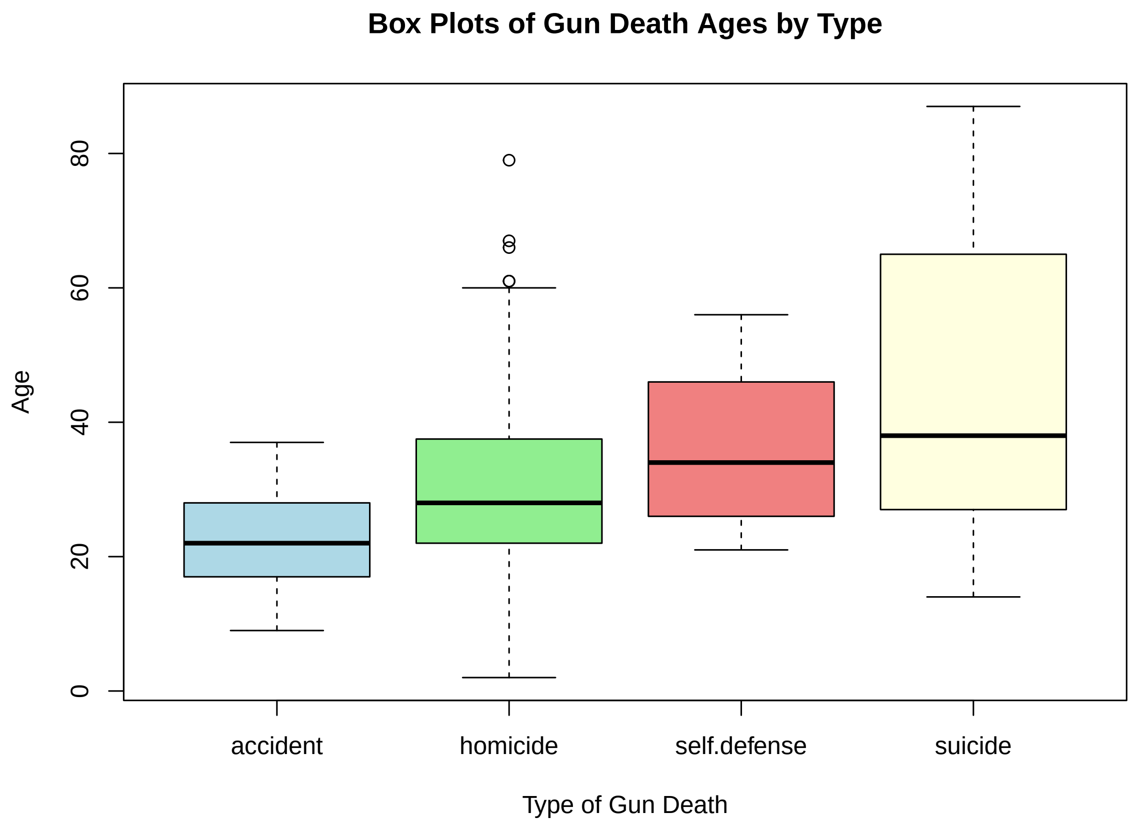
```r
# Five-number summaries
summary_homicides <- fivenum(homicides)
summary_accidents <- fivenum(accidents)
summary_suicides <- fivenum(suicides)
summary_self_defense <- fivenum(self_defense)

# Create a data frame to display summaries
summaries <- data.frame(
  Type = c("Homicides", "Accidents", "Suicides", "Self-defense"),
  Min = c(summary_homicides[1], summary_accidents[1], summary_suicides[1], summary_self_defen
  Q1 = c(summary_homicides[2], summary_accidents[2], summary_suicides[2], summary_self_defens
  Median = c(summary_homicides[3], summary_accidents[3], summary_suicides[3], summary_self_de
  Q3 = c(summary_homicides[4], summary_accidents[4], summary_suicides[4], summary_self_defens
  Max = c(summary_homicides[5], summary_accidents[5], summary_suicides[5], summary_self_defen
)
```

```r
# Display the summary table
knitr::kable(summaries, caption = "Five-Number Summaries of Gun Death Ages by Type")
```

Table 1: Five-Number Summaries of Gun Death Ages by Type

Type	Min	Q1	Median	Q3	Max
Homicides	2	22	28	37.5	79
Accidents	9	17	22	28.0	37
Suicides	14	27	38	65.0	87
Self-defense	21	26	34	46.0	56

```r
# Create box plots
boxplot(age ~ type, data = gun_deaths,
        main = "Box Plots of Gun Death Ages by Type",
        xlab = "Type of Gun Death",
        ylab = "Age",
        col = c("lightblue", "lightgreen", "lightcoral", "lightyellow"))
```



Important information hidden by the box plots:

The box plots hide several important aspects of the data that were visible in the stem-and-leaf plots:

1. The bimodal nature of the suicide age distribution is completely lost. The box plot shows the central tendency and spread but fails to reveal that there are two distinct concentrations of suicide deaths - one among young adults and another among elderly individuals.
2. The fine-grained details about clusters at specific ages are not visible. For example, the concentration of homicides in the 18-25 year range is smoothed over in the box plot.
3. The box plot doesn't clearly show the exact number of deaths at each age, which was visible in the stem-and-leaf display.
4. The presence of very young children in both the homicide and accident categories is less apparent in the box plots than in the stem-and-leaf plots.

Regarding the median gun-suicide age as a meaningful summary statistic:

The median gun-suicide age is not a particularly meaningful summary statistic because the distribution is bimodal. The median falls somewhere between the two modes and represents an age at which relatively fewer suicides occur compared to either the younger adult group or the elderly group. A single central measure like the median fails to capture the essential characteristic of this distribution - that it has two distinct peaks rather than a single central tendency.

Note: The quartile values calculated here may differ slightly from those in the textbook due to different calculation methods for quartiles in R versus the book definition.

(d) Shape of Combined Gun Deaths

If we were to construct a stem-and-leaf plot for all gun deaths combined (treating them as a single group), I would expect a distribution that:

1. Is dominated by the homicide pattern in the younger age ranges (teens to late 20s) since homicides are the most numerous category
2. Shows a second, smaller peak in the older age ranges (60s-80s) contributed mainly by suicides
3. Has a long right tail extending into the 80s due to the elderly suicide victims
4. Overall appears somewhat bimodal but with the first mode (young adults) being much more pronounced

Let's verify this by creating a combined stem-and-leaf plot:

```r
# Combined stem-and-leaf plot
cat("All Gun Deaths Combined:\n")
```

```
All Gun Deaths Combined:
```

```r
stem(gun_deaths$age, scale = 0.5)
```

```

  The decimal point is 1 digit(s) to the right of the |

  0 | 29
  1 | 24445555666666677777777788888888888888999999999999
  2 | 00000000000000000011111111111222222222222222223333333333333444444444+68
  3 | 000000000111111111111222222223333333333333444444444444444445555555555+24
```

```
4 | 00000000001111122222222222234444445556666666777777888888
5 | 000012222222344456666677788899
6 | 01111122334445555555556667777777999999
7 | 00000111223345556677778999
8 | 112346777
```

```
# Looking at the stem-and-leaf plot above, we can confirm our prediction:
# The combined plot shows a large peak in the late teens to 20s age range
# (from homicides and younger suicides), with a smaller secondary peak
# around the 65-70 age range (from the older suicide group)
```

(e) Applicability to Other Western Countries

I would not expect the patterns observed in this U.S. data to apply directly to other Western countries such as the United Kingdom, Canada, Australia, or New Zealand for several reasons:

1. **Gun ownership rates**: The U.S. has significantly higher rates of civilian gun ownership compared to these countries, which affects both the overall rate of gun deaths and the distribution across categories.
2. **Gun control legislation**: These other Western countries have stricter gun control laws, which particularly affect the accessibility of firearms to different age groups.
3. **Cultural differences**: Attitudes toward guns and their use in self-defense situations differ significantly across these countries.
4. **Homicide rates**: The U.S. generally has higher homicide rates than these comparison countries, which would affect the age distribution of victims.
5. **Mental health services**: Differences in accessibility and utilization of mental health services could affect suicide patterns across age groups.

I would expect these other Western countries to show: - Lower overall numbers of gun deaths - Relatively fewer homicides and accidents as a proportion of total gun deaths - A greater proportion of suicides among total gun deaths - Fewer deaths among younger age groups due to stricter access controls - Less pronounced age-specific patterns due to overall lower numbers

Problem 7: Analysis of Fish Dimensions

Figure 3 plots data on the dimensions of three species of fish (bream, pike, and perch) caught in Lake Laengelmavesi near Tampere in Finland. The variables measured are (maximal) length, (maximal) height, (maximal) width, and weight.

(a) Features in the Data

Looking at the individual plots, we can identify several interesting features:

1. **Dot Plots of Individual Variables (b-e)**:

   - **Weights (b)**: The three species show distinct weight distributions. Bream tend to be heavier than perch, while pike show the widest range of weights, including the heaviest specimens.
   - **Lengths (c)**: Pike tend to be the longest fish, while bream and perch have more overlapping length distributions.
   - **Widths (d)**: Bream appear to be the widest fish for their length, followed by perch, with pike being relatively narrow.
   - **Heights (e)**: The stem-and-leaf plots show that bream are generally taller (deeper-bodied) than the other species, with pike being the shallowest.

2. **Scatter Plots (f-h)**:

   - **Weight vs. Length (f)**: All three species show a strong positive relationship between length and weight, but with different trajectories. For a given length, bream tend to be heaviest, followed by perch, with pike being the lightest. This suggests different body shapes.
   - **Width vs. Length (g)**: Again, all species show positive relationships between length and width, but bream tend to be wider for their length than perch, with pike being the narrowest.
   - **Height vs. Length (h)**: This plot shows the most dramatic separation between species. Bream show much greater body depth (height) for their length compared to the other species, while pike are very shallow-bodied relative to their length.

3. **Overall Patterns**:

   - The data reveals that these three species have distinctly different body shapes.
   - Bream are deep-bodied, wide, and relatively short.
   - Pike are long, narrow, and shallow-bodied.
   - Perch have intermediate characteristics.
   - These morphological differences reflect their ecological niches and feeding behaviors. Pike are ambush predators with torpedo-shaped bodies for quick bursts of speed, while bream's deep body shape is suited for maneuverability and filter-feeding in slower waters.

(b) Best Variables for Discrimination

The height vs. length plot (h) provides the best discrimination between the three species. In this plot:

1. The three species form almost completely separate clusters with minimal overlap.
2. Bream occupy the upper portion of the plot (high height-to-length ratio).
3. Pike occupy the lower portion (low height-to-length ratio).
4. Perch occupy the middle, with only slight overlap with either of the other species.

This suggests that body depth (height) relative to length is the most distinctive morphological feature among these three species. The ratio of height to length effectively captures the fundamental difference in body shape - bream are deep-bodied, pike are elongated and slender, and perch have an intermediate body form.

The distinct separation in this plot would make it particularly useful for species identification or for studying the ecological significance of body shape in relation to habitat use and feeding strategies.

Conclusions

Both problems demonstrate the importance of looking beyond summary statistics to understand the full story in our data. In the gun deaths analysis, stem-and-leaf plots revealed critical patterns like bimodality that would be lost in simple summaries. Similarly, the fish dimension analysis showed how visualizing relationships between variables can expose meaningful biological differences. Statistical graphics provide insights that numbers alone cannot convey.