Summaries and visualization of distributions

Reflection on the last week

Objectives

Organizing your work

Descriptive Statistics

Characterizing centrality

Mean (*průměr*)

`mean(x)`

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}(\sum_{i=1}^{n} x_i)$$

Median (*medián*)

`median(x)`
- **Robust**, minimizes influence of outliers.

What are outliers? (*odlehlé hodnoty*)

- **Outliers** are data points that significantly differ from other observations.
- May indicate a measurement error, an exceptional observation, etc.

Characterizing centrality

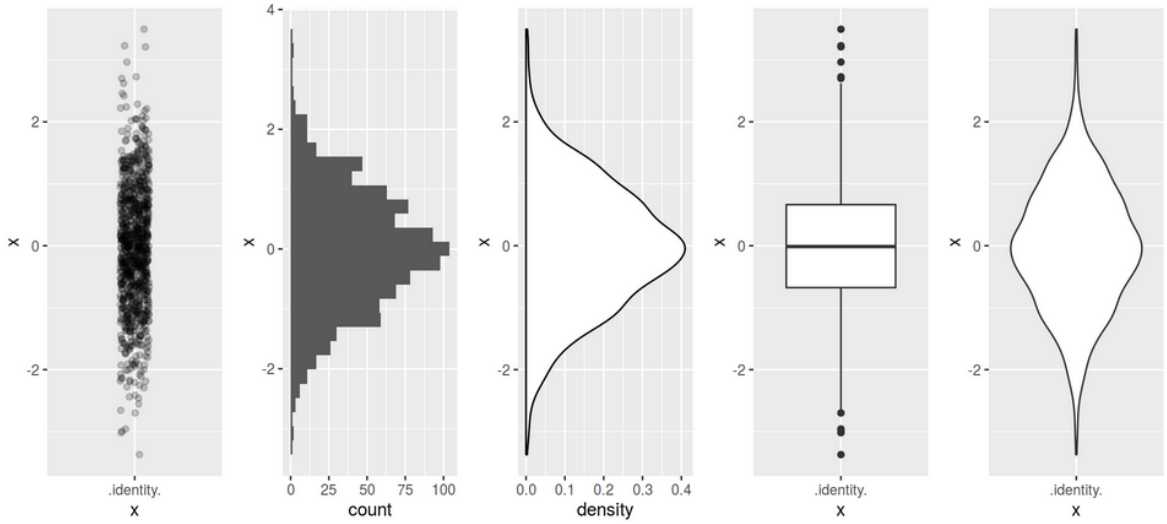


Figure 1: Various plots of a normal distribution

Characterizing dispersion and/or spread

Range (*rozpětí*)

`max(x) - min(x)` or `range(x)`

Variance and Standard deviation (*rozptyl* a *směrodatná odchylka*)

`sd(x)`

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

Interquartile range (midspread, IQR, *kvantil*, *mezikvartilové rozpětí*)

`IQR(x)`
- **Robust**, minimizes influence of outliers.

Characterizing dispersion and/or spread

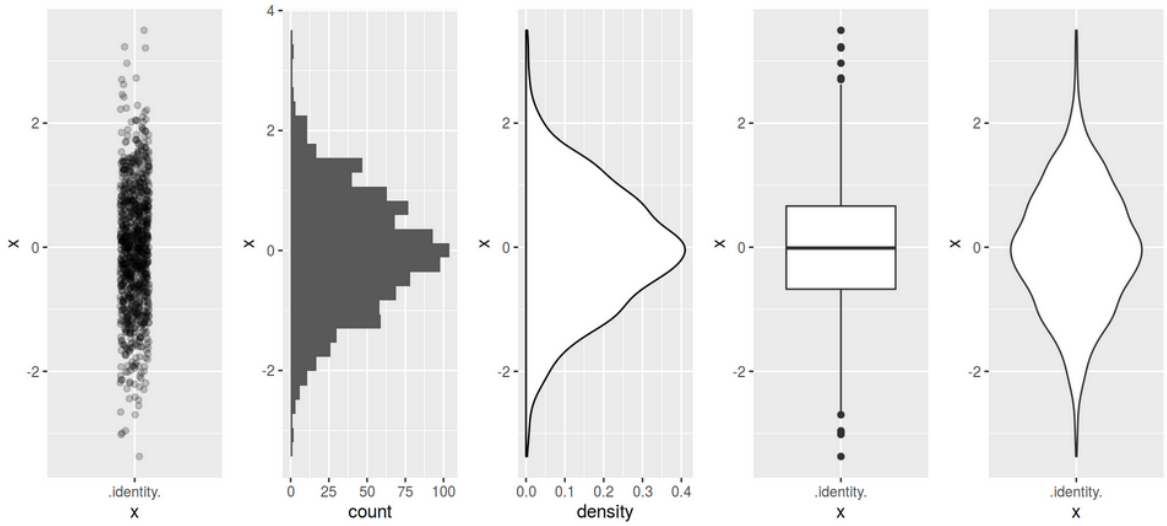


Figure 2: Various plots of a normal distribution

Exercise

- Start *RStudio*.
- Create a new project, save it somwhere you can find it.
- Use the dataset from the last lecture dartpoints.csv.
- Save it in your project directory.
- Load the data from the CSV file.
- Explore the dataset.
- Count mean and median **weight**, how do they differ?
- What is the range of the weights?
- What is the standard deviation of weights? What does it mean?
- Count the IQR. Compare it with standard deviation.
- Hints: `read.csv()`, `str()`, `colnames()`, `mean()`, `median()`, `range()`, `sd()`, `IQR()`, `summary()`

Solution

```
# DartPoints <- read.csv("dartpoints.csv")
colnames(DartPoints)
```

```
 [1] "Name"      "Catalog"   "TARL"      "Quad"      "Length"    "Width"
 [7] "Thickness" "B.Width"   "J.Width"   "H.Length"  "Weight"    "Blade.Sh"
[13] "Base.Sh"   "Should.Sh" "Should.Or" "Haft.Sh"   "Haft.Or"
```

```
DartPoints$Weight
```

```
 [1]  3.6  4.5  3.6  4.0  2.3  3.0  3.9  6.2  5.1  2.8  2.5  4.8  3.2  3.8  4.5
[16]  4.4  2.5  2.3  4.2  3.3  3.6  7.4  5.6  4.8  7.8  9.2  6.2  4.3  4.6  5.4
[31]  5.9  5.1  4.7  7.2  2.5  3.9  4.1  7.2 10.7 12.5 13.4 11.1  7.2 28.8 13.9
[46]  9.4  5.3  7.9  7.3 12.2  9.3 11.1 14.8 10.7 11.1 12.3 13.1  6.1  9.2  9.4
[61]  6.7 15.3 15.1  4.6  4.3 11.6 10.5  6.8  9.1  9.4  9.5 10.4  7.5  8.7  6.9
[76] 15.0 11.4  6.3  7.5  5.9  5.4  9.5  5.4  7.1  9.7 12.6 10.5  5.6  4.9  5.2
[91] 16.3
```

```
mean(DartPoints$Weight)
```

```
[1] 7.642857
```

```
median(DartPoints$Weight)
```

```
[1] 6.8
```

```
max(DartPoints$Weight) - min(DartPoints$Weight) # or range(DartPoints$Weight)
```

```
[1] 26.5
```

```
sd(DartPoints$Weight)
```

```
[1] 4.207088
```

```
IQR(DartPoints$Weight)
```

```
[1] 5.5
```

```
summary(DartPoints$Weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.300   4.550   6.800   7.643  10.050  28.800
```

Brainstorming

- Why do we visualize data?
- What elements does a *good* graph contain?
- How are these elements called?

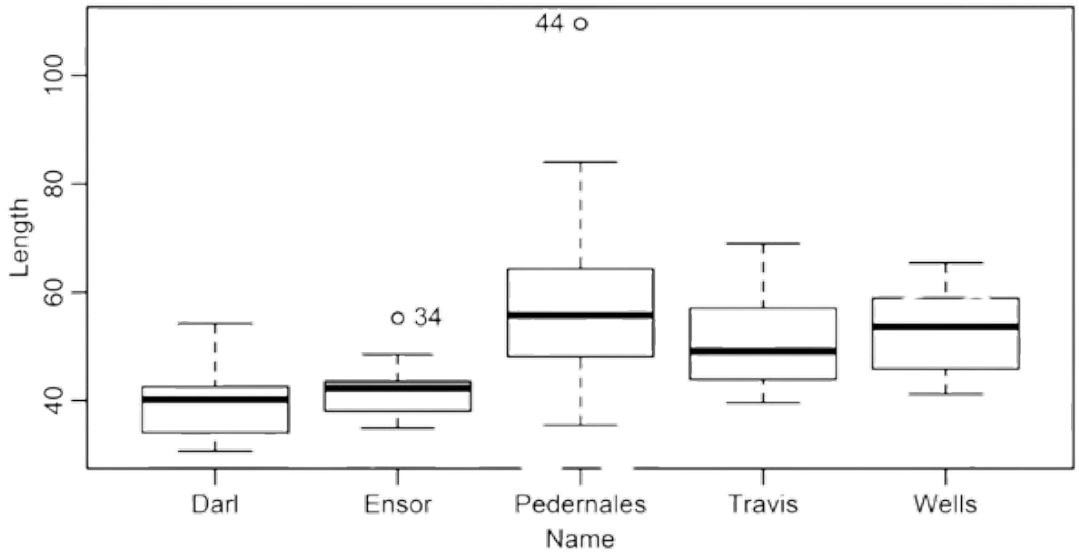


FIGURE 15 Box-and-whiskers plots for dart point lengths.

Figure 3: Boxplots from Carlson 2017

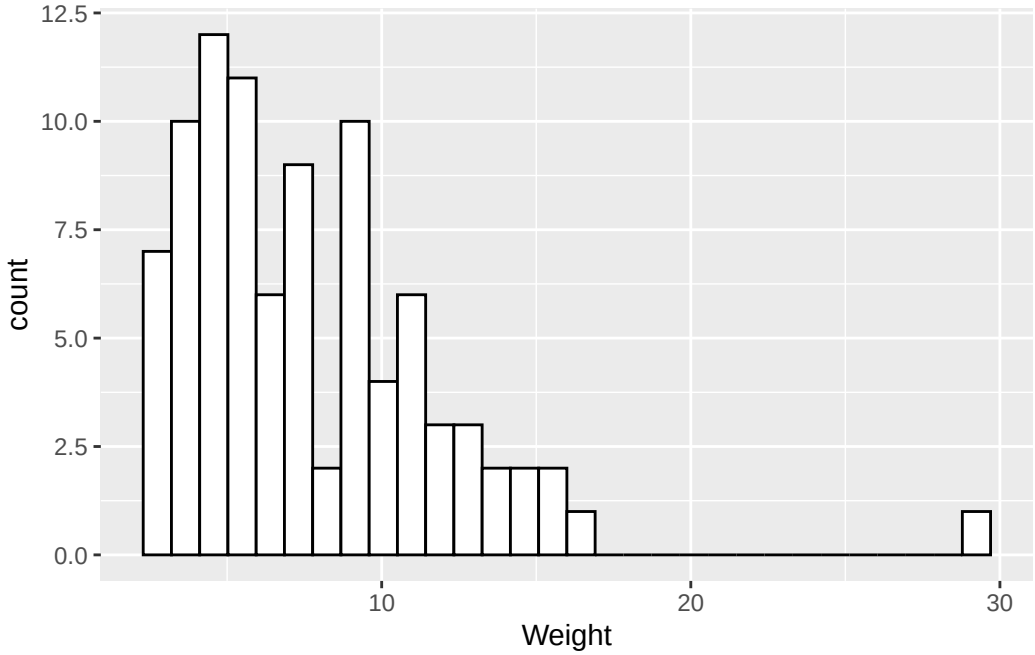
Plots for one variable

Histogram

- **Distribution** of values of a **quantitative** variable.

Distribution of dart point weights

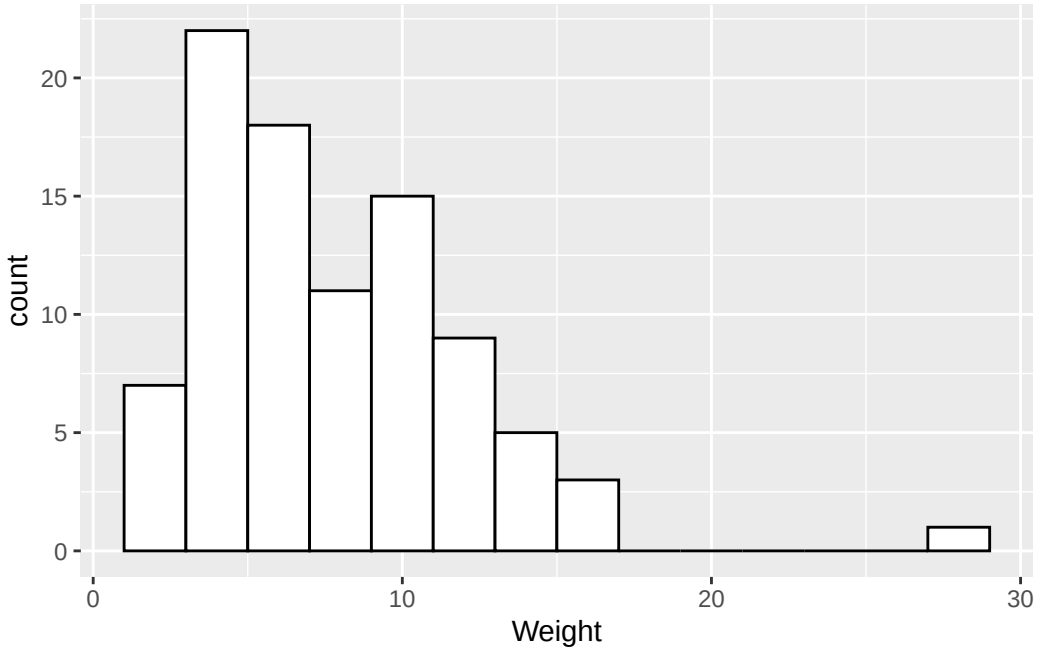
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histogram

- **Distribution** of values of a **quantitative** variable.

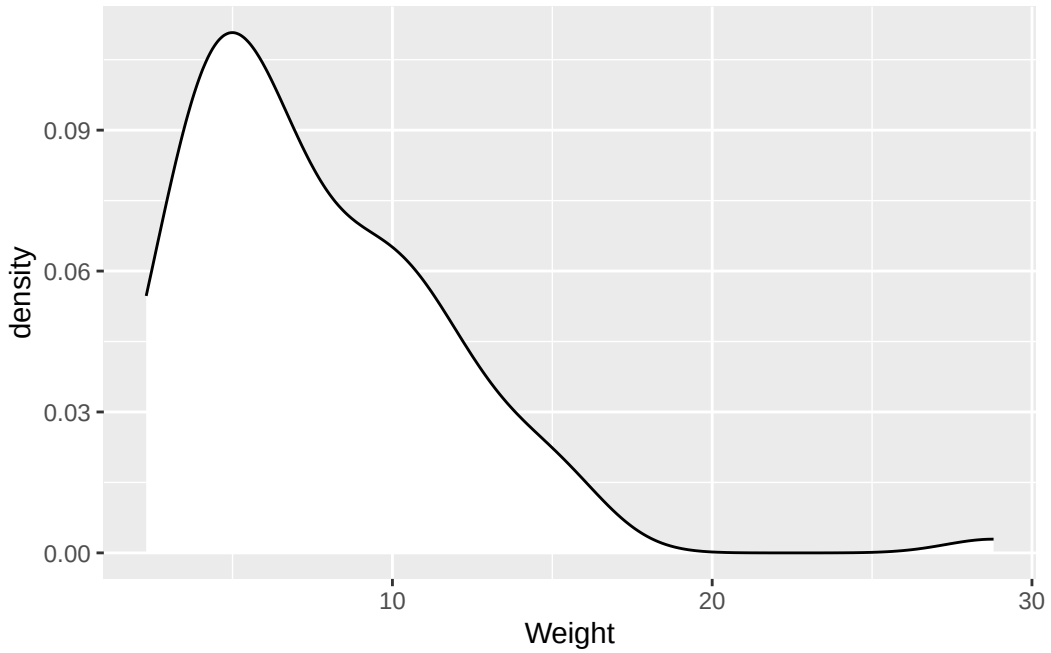
Distribution of dart point weights, one column (*bin*) equals 2g



Density plot

- **Distribution** of values of a **quantitative** variable

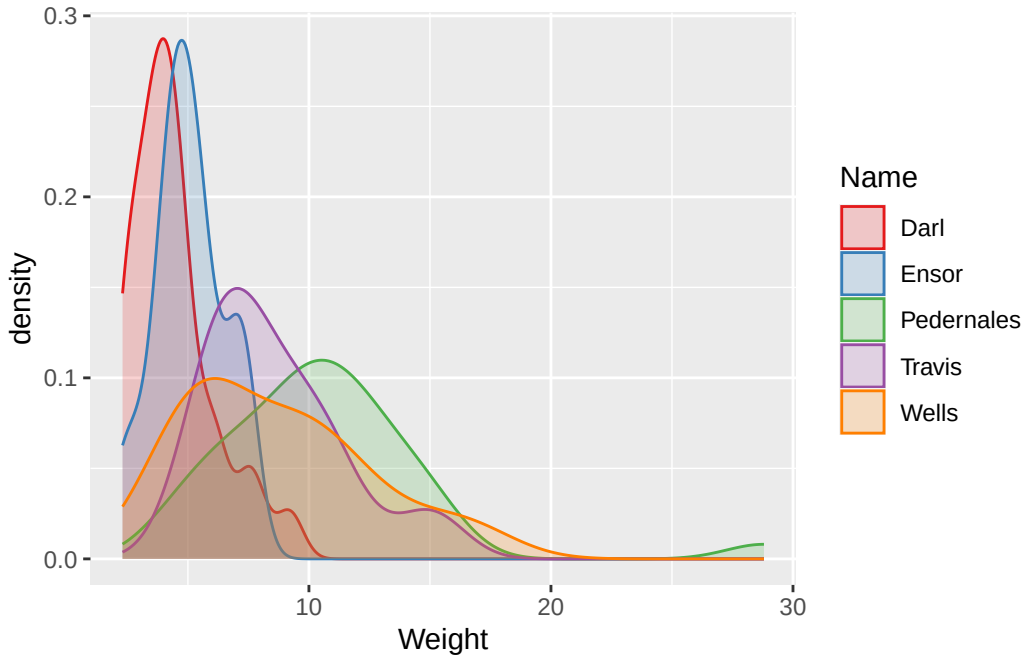
Distribution of dart point weights



Density plot

- **Distribution** of values of a **quantitative** variable, great for **comparisons**

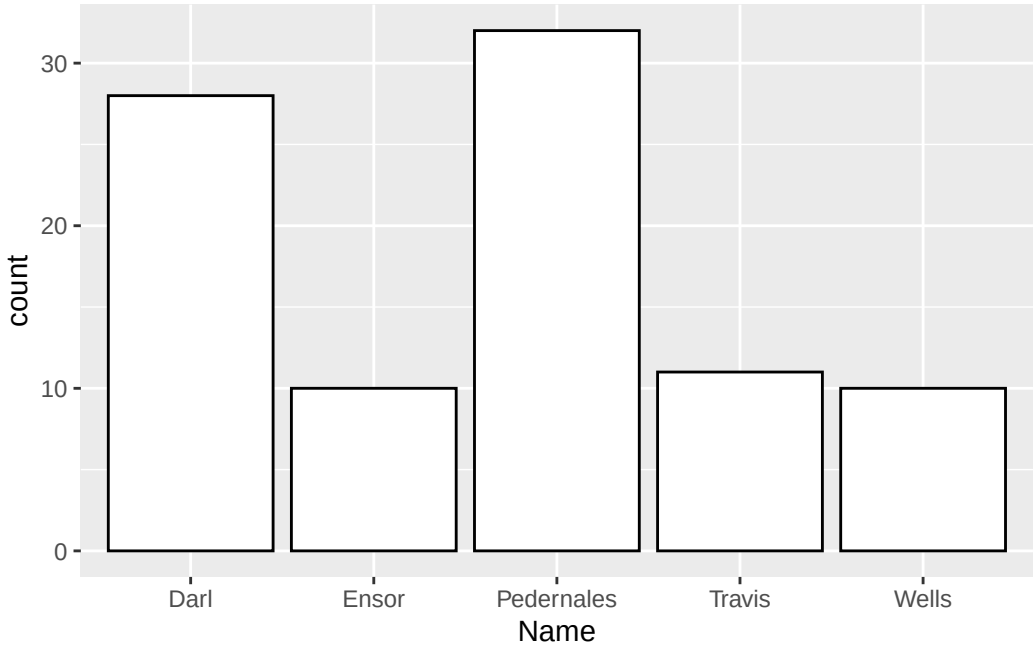
Distribution of different types of dart points by weight



Bar chart

- **Distribution** of values of a **qualitative** variable

Distribution of types of dart points



Plots in ggplot2 package

Exercises

Assignments

- Read Make a plot chapter in *Data Visualization* book by K. J. Healy.

Optional

- Go through *Visualize data* tutorials here.