Exploratory Data Analysis

uDASH Year 10 Work Experience Program: Day 2

Boris Beranger

October, 2024

UNSW
Data Science Hub

UNSW SYDNEY
Mathematics & Statistics

Introduction

Random Variables

Numerical summaries for one variables

Graphical Summaries for one variables

Summaries for two or more random variables

The steps involved in a Statistical Analysis

- What data to collect? This depends on what the research question is and who asks it.
- How to collect data in a clever way? Design of experiments, computer simulations.
- How to organize your data? In paper notebooks? files? data bases?
- How to describe your data? File format and size. File content. Statistical descriptive summaries.
- How to analyse data? Relationships, statistical inference

Data sets (I)

A data set (or dataset) is a collection of data (i.e., numbers, qualifiers, pieces of information).

In Statistics, data sets usually come from actual observations collected on cases, obtained by sampling a population (of such cases).

Most commonly, a data set corresponds to the contents of a single database table, or a single statistical rectangular table. Each row in the table corresponds to the observations (values) of a few variables (such as height and weight) on one given element (case) of that population. Each column of the table represents a particular variable.

The data set may comprise data for one or more members, corresponding to the number of rows, and called the sample size.

Data sets (II): Words you need to know

You should always clearly identify the:

- Population: the collection of all individuals or items or objects under consideration in a statistical study, usually determined by what we want to know.
- Cases: the members, objects, units, subjects or individuals from the population, from which information (i.e., data) is collected.
- IDs: the identification code of each case.
- Sample: a subset of the population.
- Sample size, n: number of cases/observations in the sample.
- Variable: a characteristic of the cases (measured, collected, recorded or counted).
- Number of variables, p: the total number of variable.

Example - A real Data Set

Example

The Tasmanian devil is a carnivorous marsupial. It was once native to mainland Australia and is now found in the wild only on the island state of Tasmania. This disappearance is usually blamed on dingoes, which are absent from Tasmania. Since the late 1990s, the devil facial tumour disease (DFTD) has drastically reduced the devil population and now threatens the survival of the species, which in 2008 was declared to be endangered.

Example - A real Data Set (II)

Example

A researcher decides to conduct a study about the population of Tasmanian devils (cases) living in Tasmania. She manages to capture 30 specimens in the wild (sample), and for each one (Specimen 1, Specimen 2, . . . , etc.) (IDs/labels) she records the following variables:

- weight,
- presence or absence of the disease,
- severity of the disease on a scale of 1 (mild) to 5 (extremely severe), and
- location where the animal was captured.

The sample size is $n = 30$ and the number of variables is $p = 4$.

Introduction

Random Variables

Numerical summaries for one variables

Graphical Summaries for one variables

Summaries for two or more random variables

Quantitative or Categorical?

Definition

The type of a variable is either categorical or quantitative.

- A qualitative or categorical variable places an individual into one of several categories.
- A quantitative variable takes numerical values, for which arithmetic operations such as averaging make sense.

Example

The variable **temperature** is a .... variable, whereas the variable **sex** (assigned at birth) is a .... variable.

Note: for a quantitative variable, it is important to give the units of measurements when relevant. For example, the units of temperature could be . . . .

Categorical random variables

Example

Here is a list of values that might be represented in a categorical variable:

- The roll of a six-sided die: possible outcomes are 1,2,3,4,5, or 6.
- Demographic information of a population: gender, disease status.
- The blood type of a person: A, B, AB or O.
- The political party that a voter might vote for, e.g. Green Party, Christian Democrat, Social Democrat, etc.
- The type of a rock: igneous, sedimentary or metamorphic.
- The identity of a particular word (e.g., in a language model): One of V possible choices, for a vocabulary of size V.

Quantitative random variables

We dissociate two types of quantitative random variables:

- Discrete random variables: can take only distinct, separate values.

Example

- Monthly production of cars
- Number of stars in the universe
- number of heads when flipping three coins

- Continuous random variables: can take any values in some interval (low, high).

Example

- Height of students in a class
- Volume of petrol sold in a specific period of time
- Commute time

Exercise

Exercise

How can we help wood surfaces resist weathering, when restoring historic wooden buildings? In a study of this question, wooden panels were prepared and then exposed to the weather. Here are some of the variables recorded:

1. Type of wood (yellow poplar, pine, cedar)
2. Water repellent (solvent-based, water-based)
3. Paint thickness (millimetres)
4. Paint colour (white, grey, light blue)
5. Weathering time (months)

Which of these variables are categorical, and which are quantitative?

Exercise

Exercise

How can we help wood surfaces resist weathering, when restoring historic wooden buildings? In a study of this question, wooden panels were prepared and then exposed to the weather. Here are some of the variables recorded:

1. Type of wood (yellow poplar, pine, cedar) categorical
2. Water repellent (solvent-based, water-based) categorical
3. Paint thickness (millimetres) quantitative
4. Paint colour (white, grey, light blue) categorical
5. Weathering time (months) quantitative

Introduction

Random Variables

Numerical summaries for one variables

Graphical Summaries for one variables

Summaries for two or more random variables

Numerical summaries for a categorical variable

We can summarise one categorical variable using table of frequencies which corresponds to

- list of the possible categories (even those not observed);
- together with their counts, percent or proportion of cases in each category.

Numerical summaries for a categorical variable (II)

Example

Consider the MATH1041 Class Survey data, and the survey question “Did you watch the Australian Open Women Tennis Final (yes/no)?”. Produce an appropriate numerical summary for this variable.

```r
load("MATH1041-2024T1.RData")
table(tennis.tv)
```

```
## tennis.tv
##   0   1
## 153  56
```

Numerical summaries for a categorical variable (III)

```
prop.table(table(tennis.tv)) * 100
```

```
## tennis.tv
##        0        1
## 73.20574 26.79426
```

```
tbl <- table(tennis.tv)
cbind(tbl, prop.table(tbl) * 100)
```

```
##   tbl
## 0 153 73.20574
## 1  56 26.79426
```

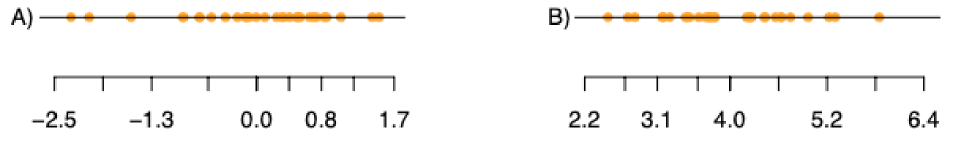
Numerical summaries for quantitative random variables

Is it a good idea to compute a table of frequencies or percentages for a quantitative variable? Why? Why not?

Numerical summaries for quantitative random variables

Is it a good idea to compute a table of frequencies or percentages for a quantitative variable? Why? Why not?

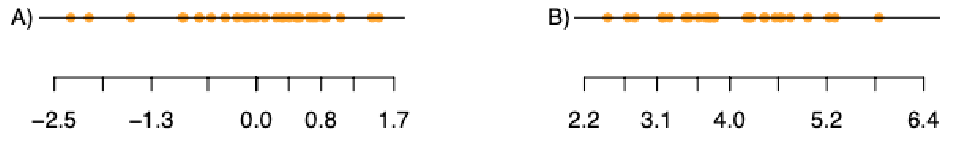
Try to describe / summarise / compare these two different data sets. (Pay attention to the x-axis values.)



Numerical summaries for quantitative random variables

Is it a good idea to compute a table of frequencies or percentages for a quantitative variable? Why? Why not?

Try to describe / summarise / compare these two different data sets. (Pay attention to the x-axis values.)



When summarising a **quantitative** variable numerically, we often make sure to include measures of:

- the location (where most of the data are), and
- the spread (or variability) of the data.

Numerical summaries for quantitative random variables (II)

A useful measure for location could be the mean. Mathematically, let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ represent the n observations (the data points), the mean is calculated as:

$$\text{mean}(\boldsymbol{x}) = \frac{x_1 + \cdots + x_n}{n}.$$

An alternative to the mean is the median, it corresponds to the `middle value`. If we sort the observations into $x_{(1)} \leq \cdots \leq x_{(n)}$, the median is computed as

$$\text{median}(\boldsymbol{x}) = x_{(n+1)/2} \text{ if } n \text{ is odd} \quad \text{and} \quad \text{median}(\boldsymbol{x}) = \frac{1}{2}\left(x_{n/2} + x_{(n+1)/2}\right) \text{ if } n \text{ is even.}$$

WARNING *The mean can be grossly affected by unusually large (or small) data values that tend to be quite far away from the bulk of the data.*

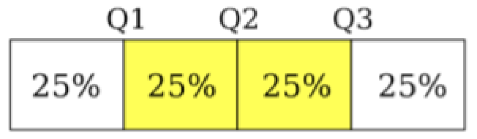
```
c(mean(unsw.satisfaction), median(unsw.satisfaction))
```

```
## [1] 4.507703 4.500000
```

Numerical summaries for quantitative random variables (III)

The median splits the data into two groups which contain roughly 50% of the data.

Quartiles divide the data into four groups which each contain roughly 25% (a quarter) of the data. Q_1 and Q_3 are known as the first and third quantile. Q_2 is just the median.



Q1 Q2 Q3

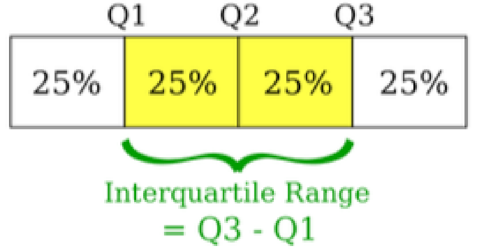
25% 25% 25% 25%

```r
quantile(unsw.satisfaction, c(0.25, 0.75))
```

```
## 25% 75%
##   4   5
```

Numerical summaries for quantitative random variables (IV)

A useful measure for spread could be the inter-quantile range (IQR).



Other useful measures are the standard deviation $\mathrm{sd}(\boldsymbol{x})$ and the variance $\mathrm{var}(\boldsymbol{x}) = \mathrm{sd}(\boldsymbol{x})^2$

We voluntarily omit the mathematical expressions, we will use R to calculate them directly.

Numerical summaries for quantitative random variables (V)

In R, the IQR, mean and variance are obtain by running

```
diff(quantile(unsw.satisfaction, c(0.25, 0.75)))
```

```
## 75%
##   1
```

```
c( sd(unsw.satisfaction), sd(unsw.satisfaction)^2, var(unsw.satisfaction))
```

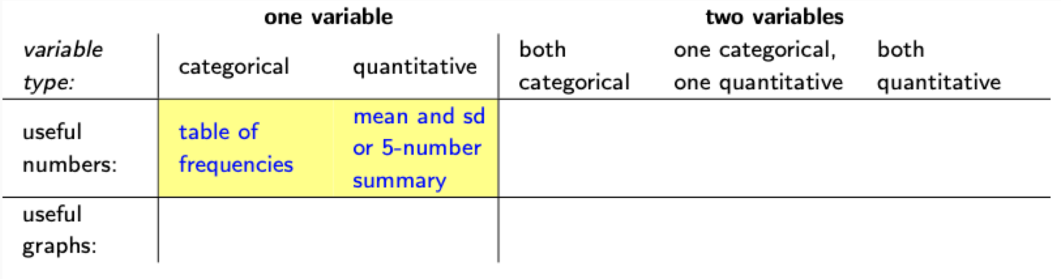
```
## [1] 1.038383 1.078239 1.078239
```

Another useful function is summary, which gives the following 5-number summary

```
summary(unsw.satisfaction)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.610   4.000   4.500   4.508   5.000   8.000
```

Numerical summaries for one variables



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary			
useful graphs:					

Introduction

Random Variables

Numerical summaries for one variables

Graphical Summaries for one variables

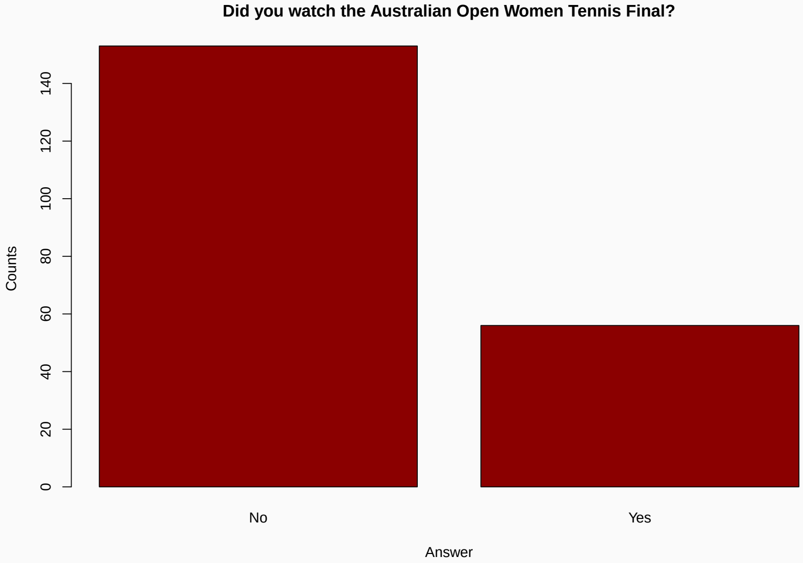
Summaries for two or more random variables

Graphical summaries for categorical random variables

Bar charts are used to summarise **categorical** random variables.

```
barplot(table(tennis.tv),
        names.arg=c("No", "Yes"),
        main="Did you watch the Australian Open Women Tennis Final?",
        xlab="Answer", ylab="Counts",col="darkred")
```

Graphical summaries for categorical random variables (II)



Graphical summaries for categorical random variables (III)

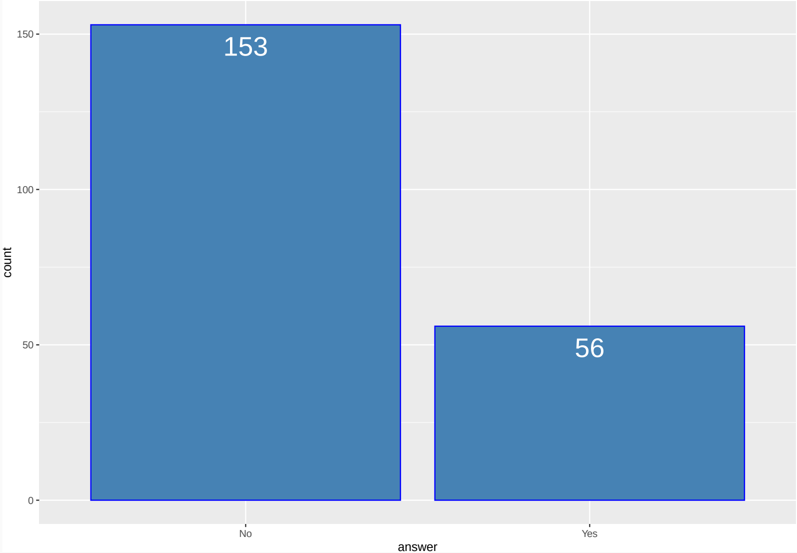
The previous plots were done in base R, now let's be fancy and use ggplot!

```
tennis <- data.frame(answer=c("No", "Yes"), count=as.vector(tbl))
head(tennis)
```

```
##   answer count
## 1     No   153
## 2    Yes    56
```

```
library(ggplot2)
p <- ggplot(data=tennis, aes(x=answer, y=count)) +
            geom_bar(stat="identity", color="blue",fill="steelblue")+
            geom_text(aes(label=count), vjust=1.6, color="white",
                      size=8.5)
p
```

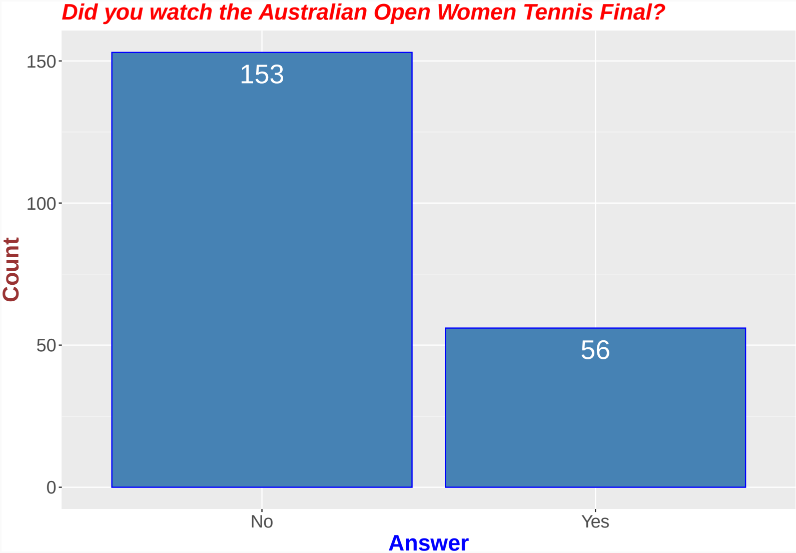
Graphical summaries for categorical random variables (IV)



Graphical summaries for categorical random variables (V)

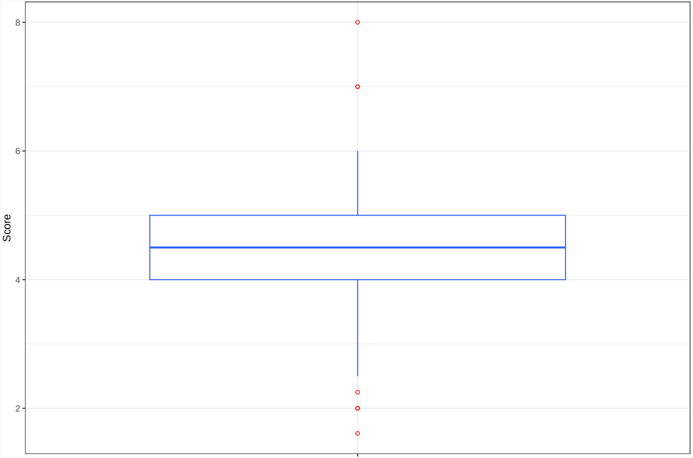
```
p +
  ggtitle("Did you watch the Australian Open Women Tennis Final?") +
  xlab("Answer") +
  ylab("Count") +
  theme(
        plot.title = element_text(color="red", size=20, face="bold.italic")
        axis.text = element_text(size=16),
        axis.title.x = element_text(color="blue", size=20, face="bold"),
        axis.title.y = element_text(color="#993333", size=20, face="bold")
        )
```

Graphical summaries for categorical random variables (VI)



Graphical summaries for quantitative random variables

Boxplots and histograms are used to summarise **quantitative** random variables.

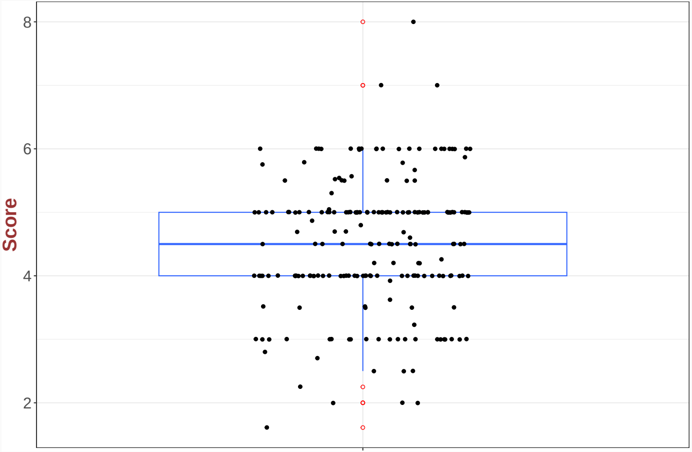


Graphical summaries for quantitative random variables (II)

```
sat <- data.frame(satisfaction = unsw.satisfaction)
p2 <- ggplot(data=sat, aes(x="", y=satisfaction)) +
             geom_boxplot(fill = "white", colour = "#3366FF",
                          outlier.colour = "red", outlier.shape = 1) +
             ylab("Score") +
             xlab("") +
             theme_bw()
p2
```

Note that the length of the whiskers is by default 1.5 times IQR, the argument coef controls that value.

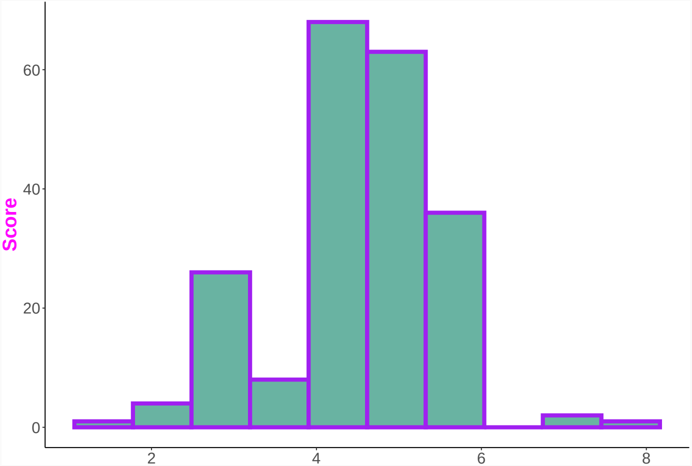
Graphical summaries for quantitative random variables (III)



Graphical summaries for quantitative random variables (IV)

```
p2 + geom_jitter(width = 0.2) +
     theme(
        axis.text = element_text(size=16),
        axis.title.y = element_text(color="#993333", size=20, face="bold")
     )
```

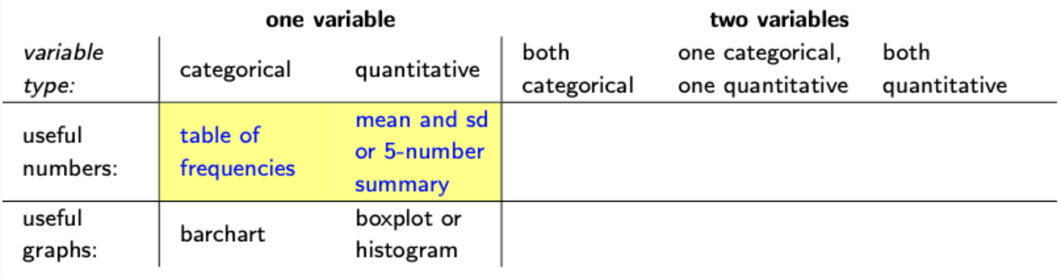
Graphical summaries for quantitative random variables (V)



Graphical summaries for quantitative random variables (VI)

```r
p3 <- ggplot(data=sat, aes(satisfaction)) +
             geom_histogram(fill = "#69b3a2", colour = "purple",
                            bins=10, size=2) +
             ylab("Score") +
             xlab("") +
             theme_classic() +
             theme(
                axis.text = element_text(size=16),
                axis.title.y = element_text(color="magenta", size=20, face=
     )
p3
```

Graphical Summaries for one variables



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary			
useful graphs:	barchart	boxplot or histogram			

Introduction

Random Variables

Numerical summaries for one variables

Graphical Summaries for one variables

Summaries for two or more random variables

Numerical summaries of two categorical random variables

In this case we use a **2-way table of frequencies**

```r
table(sex, stress)
```

```
##         stress
## sex      Not at all Slightly Moderately Very Extremely
##   Male            9       24         23    9         3
##   Female          9       38         61   18        10
```

For fancier table see e.g., `tabyl()`.

Numerical summaries of two categorical random variables

Adding proportions:

```
tbl2 <- table(sex, stress)
round(prop.table(tbl2, margin =2),3) # for column
```

```
##          stress
## sex        Not at all Slightly Moderately  Very Extremely
##   Male          0.500    0.387      0.274 0.333     0.231
##   Female        0.500    0.613      0.726 0.667     0.769
```

```
round(prop.table(tbl2, margin =1),3) # for row
```

```
##          stress
## sex        Not at all Slightly Moderately  Very Extremely
##   Male          0.132    0.353      0.338 0.132     0.044
##   Female        0.066    0.279      0.449 0.132     0.074
```

Numerical summaries of one categorical and one quantitative random variable

Here we compute a 5-number summary for each group/category.

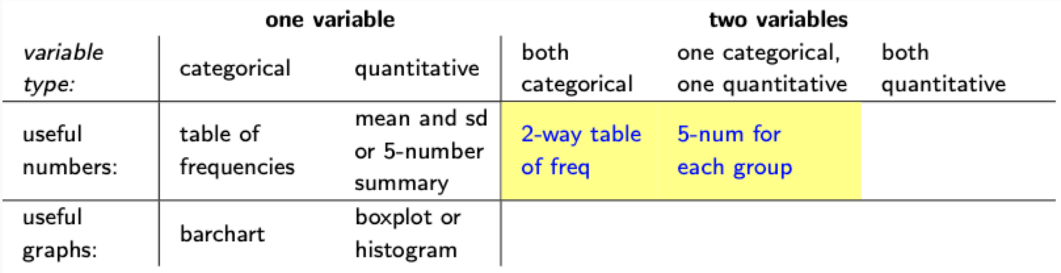
```
summary(hair.cost[sex=="Female"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    30.0    55.0   102.9   122.5   600.0       5
```

```
summary(hair.cost[sex=="Male"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    15.0    25.0    56.3    30.0  2000.0       5
```

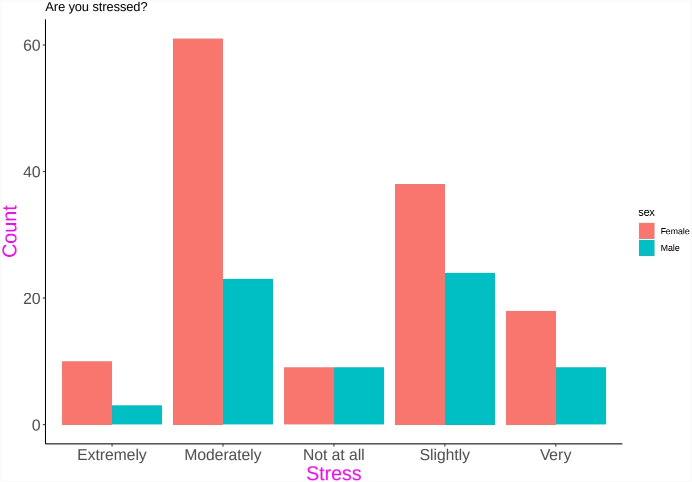
Numerical summaries with at least one categorical variable



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary	2-way table of freq	5-num for each group	
useful graphs:	barchart	boxplot or histogram			

Graphical summaries of two categorical random variables

When we have two categorical variables, we use a **clustered bar chart**

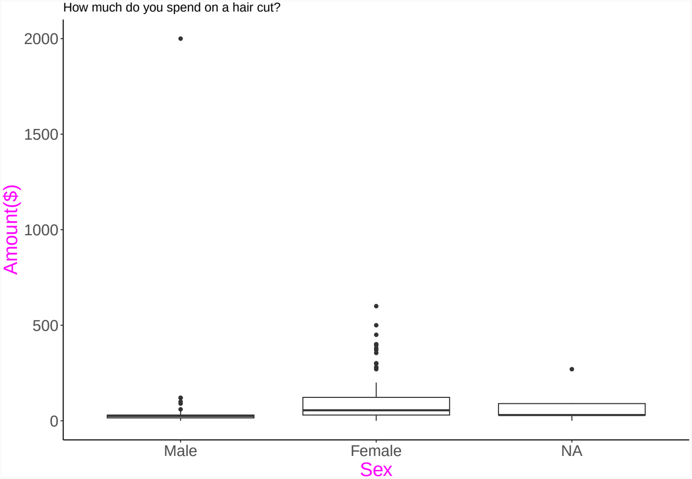


Graphical summaries of two categorical random variables (II)

```r
SS <- data.frame(stress=rep(colnames(tbl2), each=2),
                 sex = c(rep(rownames(tbl2), 5)),
                 count=as.vector(tbl2))
p4 <- ggplot(data=SS, aes(x=stress, y=count, fill=sex)) +
             geom_bar(stat="identity",
                      position="dodge") +
             ggtitle("Are you stressed?") +
             ylab("Count") +
             xlab("Stress") +
             theme_classic() +
             theme(
                axis.text = element_text(size=16),
                axis.title = element_text(color="magenta", size=20)
             )
```

Graphical summaries of one categorical and one quantitative random variable

In that case, we use **comparative boxplots**

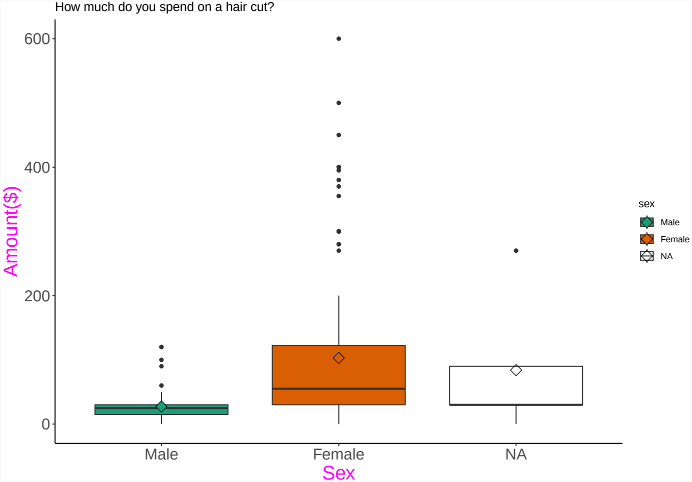


Graphical summaries of one categorical and one quantitative random variable (II)

```
SH <- data.frame(sex=sex, hair.cost=hair.cost)

p5 <- ggplot(data=SH, aes(x=sex, y=hair.cost)) +
            geom_boxplot() +
            ggtitle("How much do you spend on a hair cut?") +
            ylab("Amount($)") +
            xlab("Sex") +
            theme_classic() +
            theme(
               axis.text = element_text(size=16),
               axis.title = element_text(color="magenta", size=20)
            )
p5
```

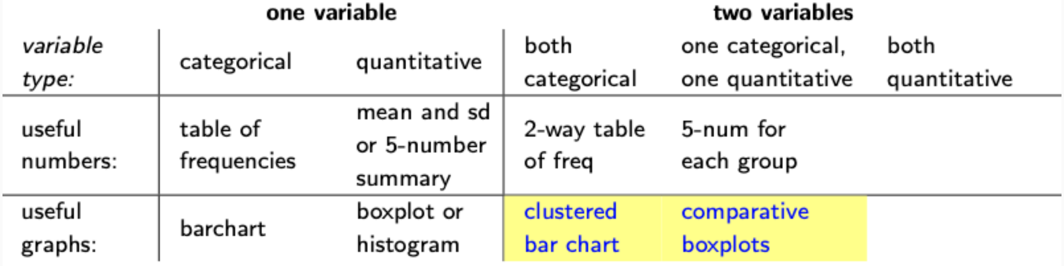
Graphical summaries of one categorical and one quantitative random variable (III)



Graphical summaries of one categorical and one quantitative random variable (IV)

```r
p5 <- ggplot(data=SH[-which.max(hair.cost),], aes(x=sex, y=hair.cost, fill=
             geom_boxplot() +
             ggtitle("How much do you spend on a hair cut?") +
             ylab("Amount($)") +
             xlab("Sex") +
             scale_fill_brewer(palette="Dark2") +
             theme_classic() +
             theme(
                axis.text = element_text(size=16),
                axis.title = element_text(color="magenta", size=20)
             )
p5 + stat_summary(fun=mean, geom="point", shape=23, size=4) # adding means
```

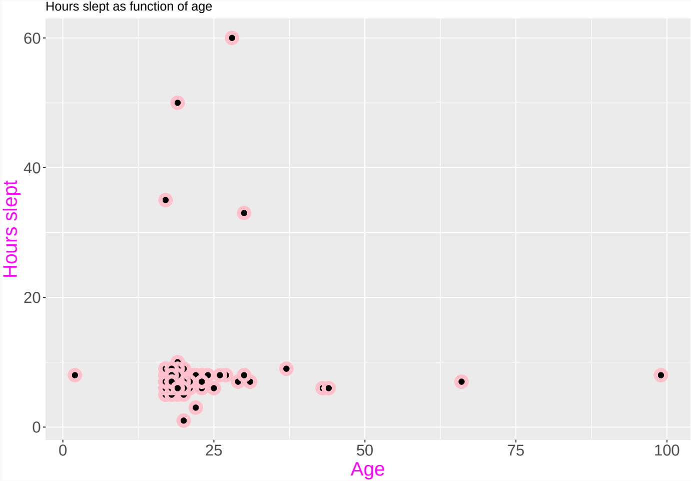
Graphical summaries with at least one categorical variable



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary	2-way table of freq	5-num for each group	
useful graphs:	barchart	boxplot or histogram	clustered bar chart	comparative boxplots	

Graphical summaries for two quantitative random variables

The easiest way to display **two quantitative random variables** is via a **scatterplot**.



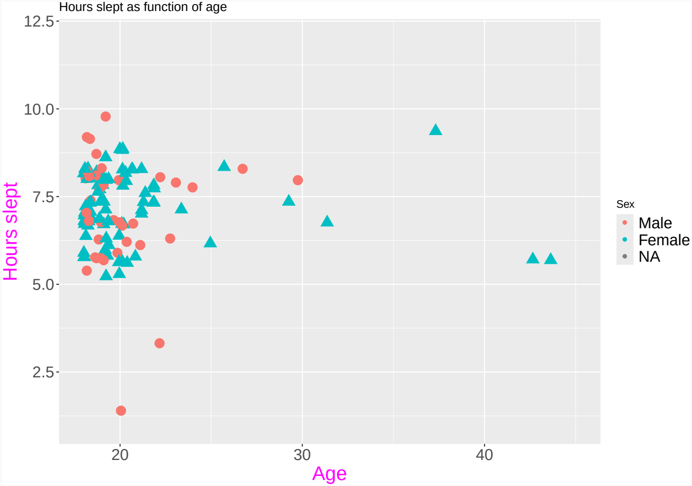
Graphical summaries for two quantitative random variables (II)

```
AH <- data.frame(age=age, hours=hours.sleep)

p6 <- ggplot(data=AH, aes(x=age, y=hours.sleep)) +
            ggtitle("Hours slept as function of age") +
            ylab("Hours slept") +
            xlab("Age") +
            theme(
               axis.text = element_text(size=16),
               axis.title = element_text(color="magenta", size=20)
            )
p6 + geom_point(shape = 21, colour = "pink", fill = "black", size = 3, str
```

Graphical summaries for two quantitative random variables (III)

The easiest way to display **two quantitative random variables** is via a **scatterplot**.

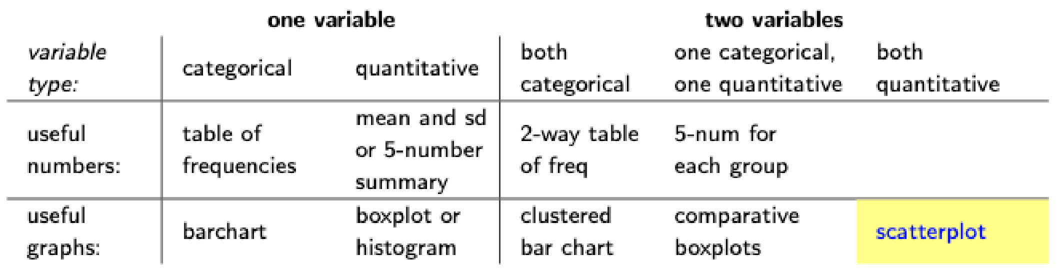


Graphical summaries for two quantitative random variables (IV)

```r
AH$sex <- sex

p6 + geom_jitter(aes(shape=factor(sex), colour = factor(sex), size=2)) +
     scale_x_continuous(limits = c(18, 45)) +
     scale_y_continuous(limits = c(1, 12)) +
     scale_colour_discrete("Sex") +
     guides(size = "none", shape = "none") +
     theme(legend.text = element_text(size=16))
```

Graphical summaries for two quantitative random variables



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary	2-way table of freq	5-num for each group	
useful graphs:	barchart	boxplot or histogram	clustered bar chart	comparative boxplots	scatterplot

Numerical summaries for two quantitative random variables

Correlation is a special form of dependence (association) **between two quantitative variables**. The correlation coefficient is a number that measures how two quantities are associated. More specifically, it measures both the strength and the direction of the linear statistical relationship between two quantitative variables.

We denote the (Pearson) Coefficient of correlation by r and for any data set

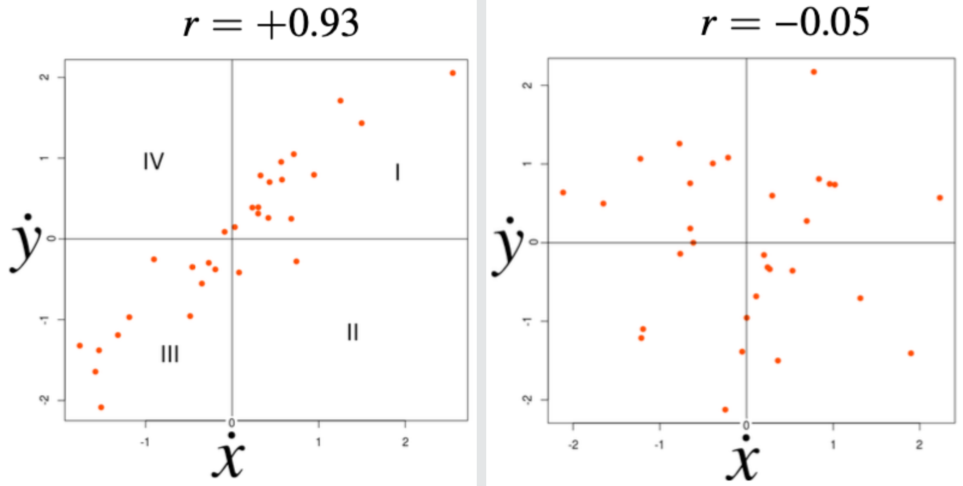
$$-1 \leq r \leq 1.$$

It is interpreted as follows:

- r close to 1 $\Leftrightarrow$ strong positive linear relationship.
- r close to -1 $\Leftrightarrow$ strong negative linear relationship.
- r close to 0 $\Leftrightarrow$ weak or non-existent linear relationship.

Numerical summaries for two quantitative random variables (II)

Example



Numerical summaries for two quantitative random variables (III)

Let's have some fun and play a `guess the correlation game`

In our example, we obtain the correlation using the `cor()` function.

First let's have a look at the `AH` `data.frame` that we created earlier:

```r
head(AH[,1:2])
```

```
##   age hours
## 1  43     6
## 2  23     8
## 3  22     7
## 4  NA     5
## 5  22     7
## 6  22     3
```

Numerical summaries for two quantitative random variables (III)

And now let's display the correlation:

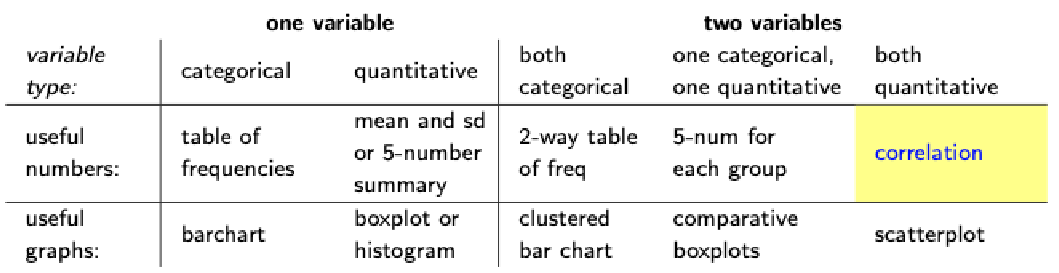
```
cor(AH[,1:2]) # Produces NA!!!
```

```
##       age hours
## age     1    NA
## hours  NA     1
```

```
cor(na.omit(AH)[,1:2])
```

```
##              age      hours
## age   1.00000000 0.07026975
## hours 0.07026975 1.00000000
```

Numerical summaries for two quantitative random variables



	one variable		two variables		
variable type:	categorical	quantitative	both categorical	one categorical, one quantitative	both quantitative
useful numbers:	table of frequencies	mean and sd or 5-number summary	2-way table of freq	5-num for each group	correlation
useful graphs:	barchart	boxplot or histogram	clustered bar chart	comparative boxplots	scatterplot