Report 1: Strava Study

Background

My primary means of commuting to and from college is by bike, and since January 2020 I have been regularly recording my cycling activities using the Strava app on my smartphone. Exporting my personal data from their website yielded a CSV file which contained details on 510 recorded activities. In this study I wanted to compare my cycling performance in two different years, and across the four seasons.

The app begins and ends recording data when manually started and stopped by the user. The GPS functionality on the smartphone allows the app to track the route followed and hence calculate the distance travelled in metres. The app records both the total *Elapsed Time* and the *Moving Time,* which subtracts the time the user is stationary (e.g. at traffic lights). The user's instantaneous speed is calculated as the time derivative of the GPS position, the highest value of which is recorded as the *Max Speed*, while the *Average Speed* is calculated as the total distance travelled divided by the *Moving Time*.

The Data

The raw as-obtained CSV file contained 83 variables, some of which were redundant (e.g. *Elapsed Time* appeared twice; *Distance* appeared once measured in kilometres and once in metres). Once imported into R, I created a new dataframe containing only several variables of interest. In order to simplify sorting by dates I converted the *Datetime* variable from "character" to "date" format.

The majority of entries dated from January 2020 to present, but there were 26 entries dating from October 2017 which did not contain entries in the *Average Speed* column. These were removed using the `na.omit()` function. There were also a small number of entries which were for non-cycling activities. These were omitted by again subsetting the dataframe to contain only entries of type "Ride". Finally, a glance through the data revealed three entries with anomalously low values of *Average Speed* (1.67, 2.58, 3.13 m/s). By checking these entries on the app it was determined I had forgot to stop recording for these entries when I had finished cycling, resulting in incorrect *Average Speed* values. These rows were removed by subsetting the entries for which *Average Speed* > 3.2.

Part 1: Normality, Normal Q-Q Plots and Histograms

To get a sense of the data and to check for normality, normal Q-Q plots and histograms were plotted for three variables: *Average Speed*, *Max Speed*, and *Moving Time* (Fig. 1). The Q-Q plot for *Average Speed* shows that the data adheres very closely to the normal distribution shape with only a small number of entries deviating from the line. These are above the line to the upper end and primarily below the line to the lower end, indicating that the distribution is somewhat long-tailed at both ends. In contrast, the data for *Max Speed* deviates significantly from the normal distribution. It has a unimodal shape and is skewed to the right. The deviations from the normal line lie above the line on both ends, indicating a short tail at the lower end of the data and a long tail at the upper end.

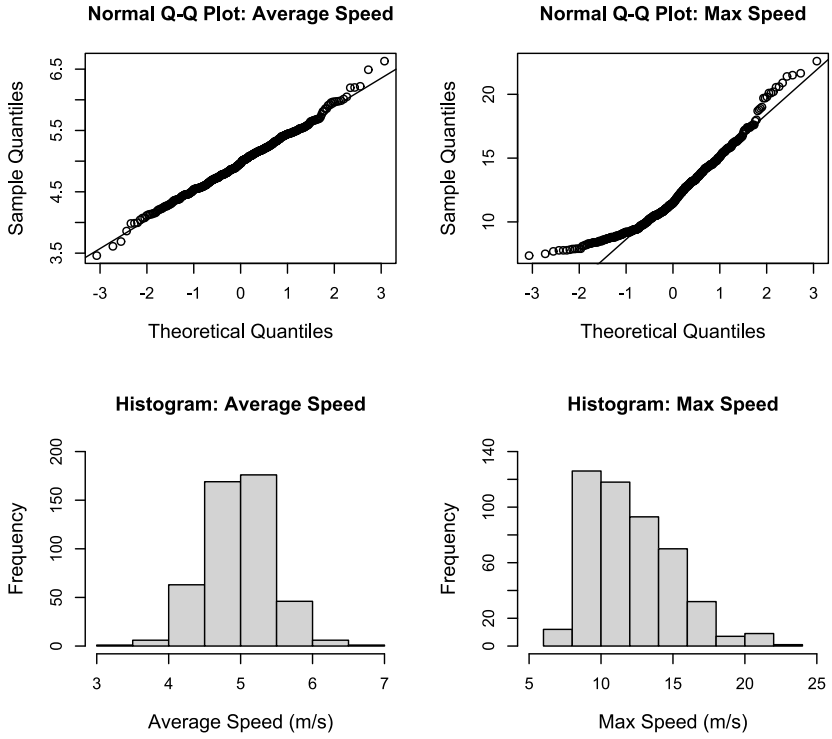

Fig. 1: Normal Q-Q plots and histograms for *Average Speed*, *Max Speed*, and *Moving Time*.

These informal observations were confirmed by conducting Shapiro-Wilk tests on the two sets of data:

	W	**p-value**	**Conclusion**
Average Speed	0.99612	0.3104	Data normally distributed
Max Speed	0.94452	3.181e-12	Data not normally distributed

Next, the average speeds recorded in 2020 and 2021 were compared. These subsets contained 175 and 282 observations respectively. Again, normal Q-Q plots and histograms were plotted (Fig. 2), box plots were plotted (Fig. 3) and Shapiro-Wilk tests performed to confirm the normality of the data:

	W	**p-value**	**Conclusion**
2020	0.98738	0.1187	Data normally distributed
2021	0.99525	0.5409	Data normally distributed

Using the `summary()` and `sd()` functions, the following statistics were obtained:

	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**	**s.d.**
2020	3.460	4.551	4.802	4.867	5.182	6.630	0.496
2021	3.689	4.759	5.048	5.035	5.326	6.490	0.430

Based on these data, which represent the entire population of my cycling activities in 2020 and 2021, I can conclude that my average speed was indeed higher in 2021 than 2020.

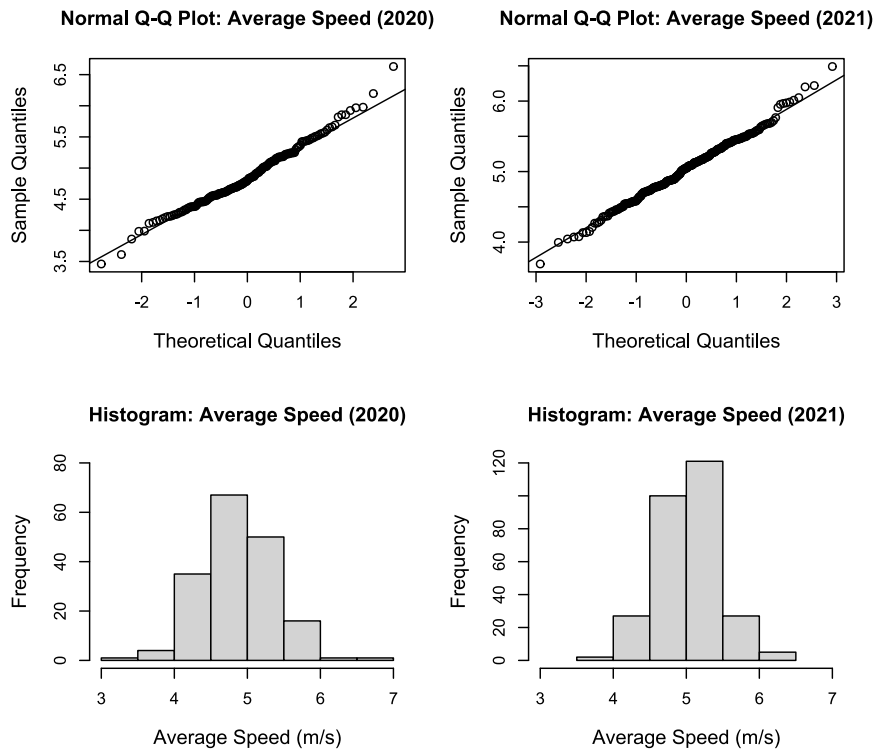

Fig. 2: Normal Q-Q plots and histograms for *Average Speed* observations in 2020 and 2021

The data was then subset into four new groups, one for each season of the year, and the average speeds data for each were plotted in boxplot (Fig. 4). At a glance, the median values for Spring, Autumn and Winter appear to be close in value, but the median value for Summer appears to be higher than the others. The distribution of data in the Autumn and Winter groups also appears to be similar. Using the `summary()` and `sd()` functions, the following statistics were obtained:

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	s.d.
Spring	4.411	4.772	4.878	4.944	5.141	5.667	0.317
Summer	4.375	4.978	5.208	5.157	5.460	5.684	0.385
Autumn	4.113	4.572	4.923	4.972	5.403	5.954	0.461
Winter	3.860	4.581	4.927	4.947	5.307	5.856	0.473

Performing a Shapiro-Wilk test on the season datasets yields the following results:

	W	p-value	Conclusion
Spring	0.95109	0.1673	Data normally distributed
Summer	0.92012	0.03493	Data not normally distributed
Autumn	0.96447	0.09236	Data normally distributed
Winter	0.98237	0.7765	Data normally distributed

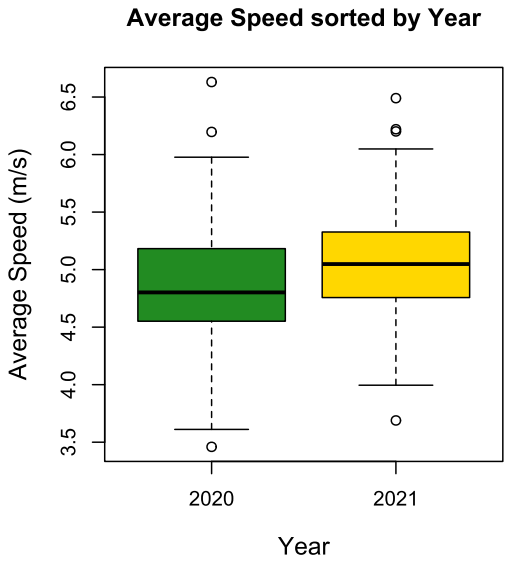

Fig. 3: Box plot of the average speed data recorded in 2020 and 2021

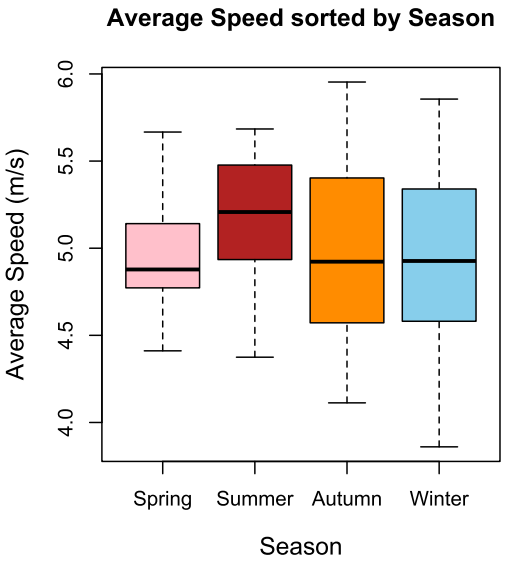

Fig. 4: Box plot of the average speed data sorted by season

In the case of Summer, the null hypothesis (i.e. that the data are normally distributed) must be rejected as the p-value falls below the significance level of $\alpha = 0.05$. If we chose a lower significance level of, e.g. $\alpha = 0.025$, we would instead accept the null hypothesis and conclude that the data is normally distributed. Decreasing α increases the critical value beyond which the test statistic is considered to be extreme, hence decreasing the risk of committing a Type I error, i.e. rejecting the null hypothesis when it is in fact true. However, it also increases the risk of committing a Type II error, denoted β, i.e. failing to reject the null hypothesis when it is in fact false, and hence reduces the "power of the statistical test", given by $1 - \beta$. A significance level of $\alpha = 0.05$ is commonly chosen as it is considered to be a good trade-off between the risks of committing both Type I and II errors.

Part 2: F-test

Question: Are the variances of my average speed in Autumn and Winter the same?

Null hypothesis: The variances of my average speed in Autumn and Winter are the same

Alternative hypothesis: The variances of my average speed in Autumn and Winter are not the same

```
        F test to compare two variances

data:  autumn$AverageSpeed and winter$AverageSpeed
F = 0.94806, num df = 56, denom df = 39, p-value = 0.8432
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5198014 1.6764472
sample estimates:
ratio of variances
         0.9480619
```

The output of the F test shows a p-value of 0.8432, much higher than the critical value of 0.025, indicating that the observed test statistic is not considered extreme for this distribution. Hence we accept the null hypothesis and conclude that, based on the recorded data, there is no evidence that the variances of my average speed in Autumn and Winter are not the same.

Part 3: t-test

Question: Are the means of my average speed in Autumn and Winter the same?

Null hypothesis: The means of my average speed in Autumn and Winter are the same

Alternative hypothesis: The means of my average speed in Autumn and Winter are not the same

```
        Two Sample t-test

data:  autumn$AverageSpeed and winter$AverageSpeed
t = 0.25682, df = 95, p-value = 0.7979
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1661730  0.2155554
sample estimates:
mean of x mean of y
 4.971774  4.947083
```

The output of the t-test shows a p-value of 0.7979, much higher than the critical value of 0.025, indicating that the observed test statistic is not considered extreme for his distribution. Hence we accept the null hypothesis and conclude that, based on the recorded data, there is no evidence that the means of my average speeds in Autumn and Winter are not the same.

Topics covered

- (T4) Normal distribution, Normal probability plot
- (T9) t-test
- (T10) F-test
- (T12) Confidence interval, p-value, critical value, significance level
- (T13) Type-I error, Type-II error, Power of a test
- (T15) Graphs
- (T20) Shapiro-Wilk test

Report 2: Football Study

Background

The English Premier League consists of 20 teams which play each other twice over the course of the competition. Ten matches are scheduled for each of the 38 rounds the competition runs for. Hence, each team plays one match per round and plays against each other team twice, once at home and once away. The English Championship runs in a similar format and consists of the 24 teams ranked immediately below the premier league team.

The Data

The two datasets for the results from the 2019 Premier League (*epl*) and Championship (*efl*) were obtained from GitHub[1] and were structured in the same format with six columns: *Round, Date, Team 1* (Home team)*, S1* (Home team score)*, S2* (Away team score)*, Team 2* (Away team).

Part 1: Chi-Square Test

Question: Are the team and number of goals scored independent of each other?

Null hypothesis: The team and number of goals scored are independent

Alternative hypothesis: The team and number of goals scored are not independent

A new dataframe with three variables (*Round*, *Team*, and *Goals*) named *goals_per_game was* created, containing the number of goals scored by each team in each match both at home and away. The entries in the *Goals* column were then renamed "1-2" if the value was 1 or 2, and "3+" if the value was greater than or equal to 3. These results were then compiled in a table and a chi-squared test performed, the output of which is shown below.

```
        Pearson's Chi-squared test

            data:  table
X-squared = 101.11, df = 38, p-value = 1.232e-07
```

The very low p-value indicates that the calculated test statistic is considered extreme for this distribution, hence we reject the null hypothesis and conclude that the team and number of goals scored are not independent. The data in the table was also used to create a bar chart (Fig. 1) which clearly shows a large difference between teams. At a glance, we can see that the team which the largest number of goalless matches was Norwich City FC, which failed to score in 20 out of their 38 matches. The team with the largest number of matches in which they scored 3 or more goals was Manchester City FC, while Crystal Palace FC was the only team which failed to score 3 or more goals in any match. The team with lowest number of goalless matches was the league winners, Liverpool FC.

[1] https://github.com/footballcsv/england/tree/master/2010s/2019-20

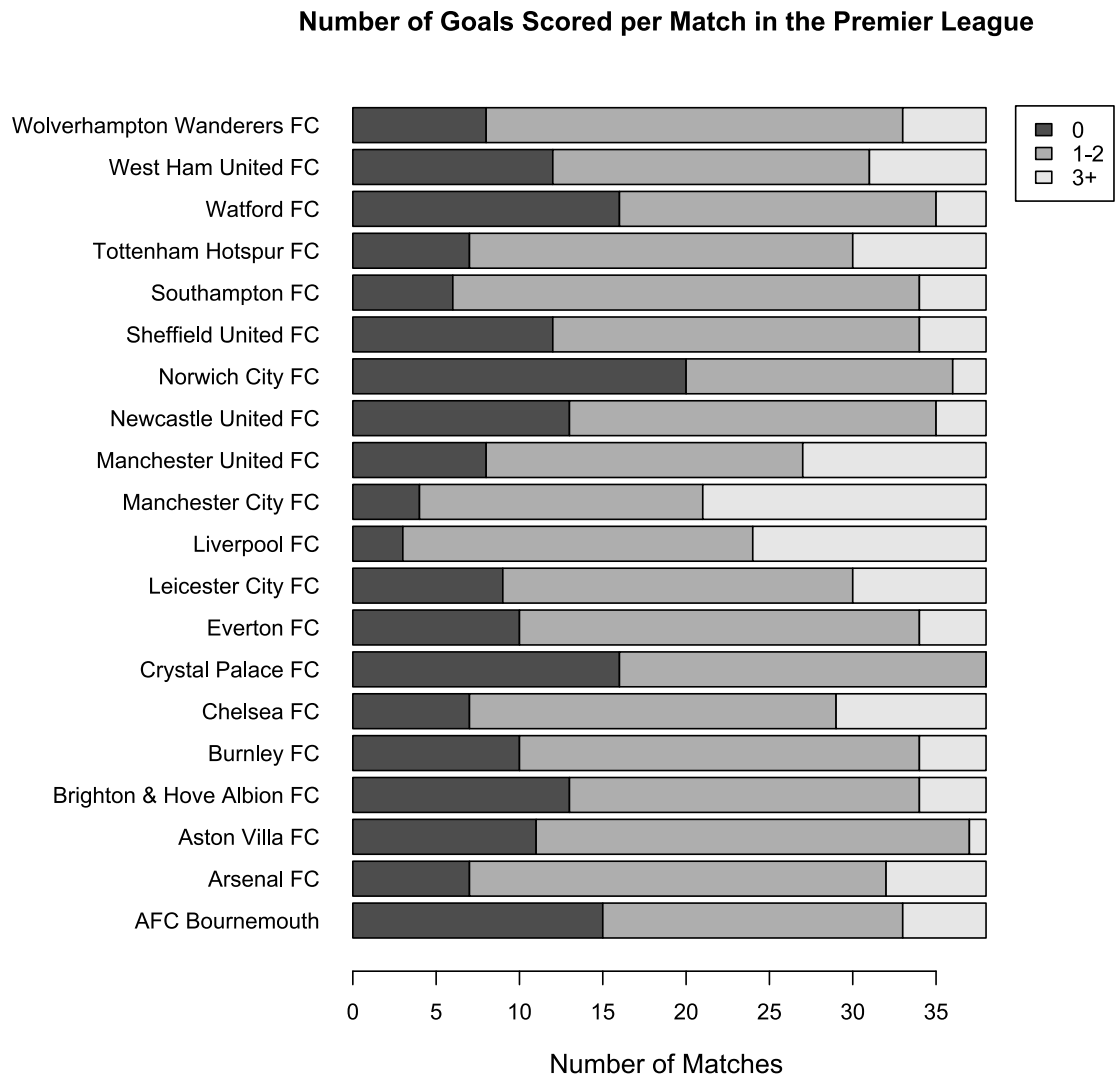

Fig. 1: A stacked bar chart displaying the number of goals scored per match by each team in the Premier League

Part 2: Z-Test

Question: Are the home teams equally likely to win/lose/draw in the Premier League and the Championship?

Null hypothesis: Home teams are equally likely to win/lose/draw in both leagues

Alternative hypothesis: Home teams are not equally likely to win/lose/draw in both leagues

It is a commonly-held belief that teams playing in their home stadium have a "home advantage" due to the presence of more fans, familiarity with the pitch, no long journey to the opponent's stadium, etc. In this section I wanted to check if the proportion of home team wins, losses and draws are the same in both the Premier League and the Championship. A new column named *home_result* was created in both the *epl* (Premier League) and *efl* (Championship) datasets, and the `ifelse()` function used to return a value of "Win" if S1 > S2, "Draw" if S1 < S2 and "Loss" if S1 == S2. The proportion of each result was determined by dividing the total number of each result, found using

the `sum()` function, by the total number of entries in each dataset, found using the `nrow()` function.

	X^2	p-value	95% Confidence Interval		Sample Estimates	
					Prem. League	Champ.
Win	1.0053	0.316	-0.03243737	0.10030215	0.4526316	0.4186992
Loss	0.013639	0.907	-0.06541215	0.05805229	0.3052632	0.3089431
Draw	1.022	0.312	-0.08858146	0.02807654	0.2421053	0.2723577

In each test the p-values are much larger than the than 0.05 so we fail to reject the null hypothesis in each instance, i.e. based on the 2019 data we conclude that there is no evidence that a particular outcome is more likely in one league over the other. We can also conclude that for the 2019 tournaments the "home advantage" belief held true, with a win being the most likely outcome for a home team in both leagues.

Part 3: Probability

In this section, I use the probability formulae learnt in module ST8001 to determine probability of certain outcomes in the 2019 Premier League.

Question: What is the probability that more than 30 goals are scored in a round?

```
epl$Goals <- epl$S1 + epl$S2
goals_per_round <- aggregate(epl$Goals, by = list(epl$Round), FUN = sum)
colnames(goals_per_round) <- c("Round", "Goals")

sum(goals_per_round$Goals > 30)/nrow(goals_per_round)
```

Answer: 0.3157895 → 31.6%

Question: What is the probability that no goals are scored in a match?

```
sum(goals_per_round$Goals > 30)/nrow(goals_per_round)
```

Answer: 0.05526316 → 5.5%

Question: What is the probability of a home team scoring more than one goal and losing?

```
sum(epl$home_result == "Loss" & epl$S1 > 1)/nrow(epl)
```

Answer: 0.02631579 → 2.6%

Question: What is the probability of an away team winning by more than two goals?

```
sum(epl$home_result == "Loss" & epl$S2 - epl$S1 > 2)/nrow(epl)
```

Answer: 0.06052632 → 6.1%

Question: What proportion of draws are goalless?

```
sum(epl$Goals == 0)/sum(epl$home_result == "Draw")
```

Answer: 0.2282609 → 22.8%

Question: Given that at least one goal has been scored, what is the probability of a draw?

```
p_goal = sum(epl$Goals > 0)/nrow(epl)
p_goal_draw = sum(epl$S1 == epl$S2 & epl$Goals > 0)/nrow(epl)
p_goal_draw/p_goal
```

Answer: 0.1977716 → 19.8%

Topics covered

- (T1) Random variable, probability, conditional probability
- (T8) z-test
- (T11) Chi-square test
- (T12) Confidence interval, p-value, critical value, significance level
- (T15) Graphs

Report 3: 2015 Boston Marathon study

Background

A marathon is a 40km road race which is popular with both professional and amateur athletes. The Boston marathon is the world's oldest annual marathon and attracts thousands of contestants each year.

The Data

The dataset contains the results of the 2015 Boston Marathon and was downloaded from Kaggle.[2] It contains 26598 observations of 25 variables which include the runners' names, city and country of origin, and the time at which they completed each 5km split during the race.

Part 1: Data Visualisation and Normality test

Firstly, the *Official Time* variable was converted from a character to numeric type using the *chron* library. The entries in the *Age* variable were then relabelled according to eight age groups and the *Official Time* variable was used to create a new *Time Group* variable with six categories. Bar charts were then plotted for both groups (Fig. 1). The age groups with the largest number of runners are the 41-45 and 46-50 groups. However the age groups with the largest number of runners with a finish time of less than three hours are the 31-35 and 26-30 groups (Fig. 2). At a glance, it is clear to see that most runners completed the marathon in between three and four hours. A sharp fall in the number of runners finishing in less than three hours is seen in the age groups above 31-35, despite having similar numbers of entrants.

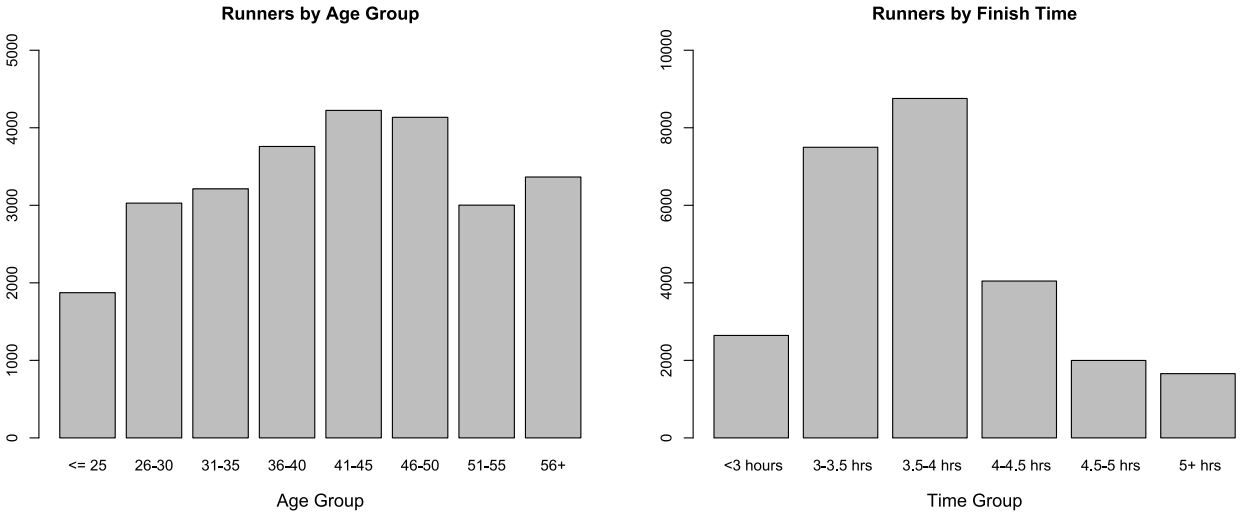

Fig. 1: Bar charts showing the number of runners sorted by age group and finish time

[2] https://www.kaggle.com/rojour/boston-results

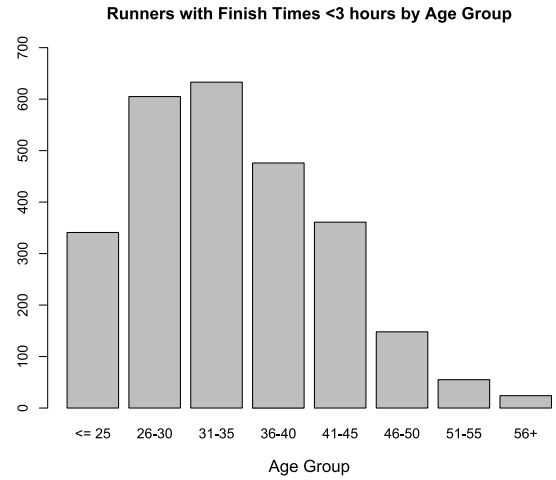

Fig. 2: Bar chart showing the number of runners with a finish time of less than 3 hours sorted by age group

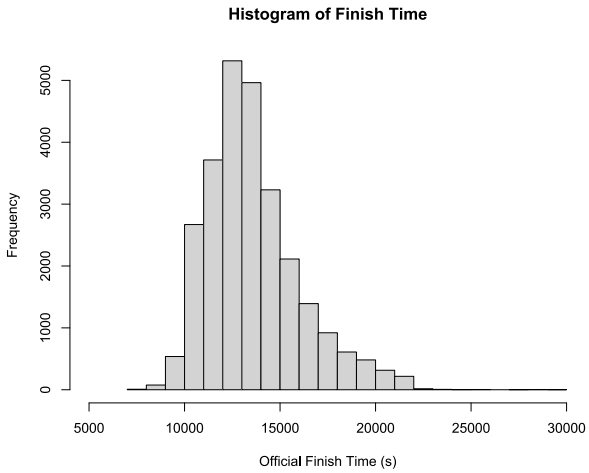

Fig. 3: Histogram of finish time (entire dataset)

A histogram of the finish times contained in the entire dataset (Fig. 3) reveals a unimodal distribution with a short-tailed lower end and long-tailed upper end. An Anderson-Darling normality test performed on this data yielded the following output:

```
        Anderson-Darling normality test

data:  data$Official.Time
A = 371.15, p-value < 2.2e-16
```

The extremely small p-value indicates the observed test statistic is extremely unlikely under the null hypothesis, hence we reject the null hypothesis and conclude that the data is not normally distributed.

Part 2: Chi-Square Test

Question: Are age group and finish time independent?

Null hypothesis: Age group and finish time are independent

Alternative hypothesis: Age group and finish time are not independent

A contingency table was printed (Fig. 4) and a chi-square test performed, the output of which is shown below:

```
        Pearson's Chi-squared test

data:  tb3
X-squared = 3462.9, df = 48, p-value < 2.2e-16
```

The extremely small p-value indicates the observed test statistic is extremely unlikely under the null hypothesis, hence we reject the null hypothesis and conclude that age group and finish time are not independent.

	<= 25	26-30	31-35	36-40	41-45	46-50	51-55	56+	Sum
<3 hours	341	605	633	476	361	148	55	24	2643
3-3.5 hrs	522	967	1057	1230	1352	1319	748	302	7497
3.5-4 hrs	448	713	795	1271	1486	1565	1278	1200	8756
4-4.5 hrs	234	308	310	366	550	666	566	1046	4046
4.5-5 hrs	182	229	212	229	280	254	181	431	1998
5+ hrs	146	205	205	187	195	183	174	362	1657
Sum	1873	3027	3212	3759	4224	4135	3002	3365	26597

Fig. 4: Contingency table with margins

Part 3: Z-test (Single Proportion)

According to one source, only 2.5% of runners in the Boston marathon finish with a time of less than 3 hours.[3]

Question: Does the data from the 2015 Boston marathon support the claim that 2.5% of runners finish with a time of less than 3 hours?

Null hypothesis: The proportion of runners finishing the 2015 Boston marathon is equal to 0.025

Alternative hypothesis: The proportion of runners finishing the 2015 Boston marathon is not equal to 0.025

```
       1-sample proportions test without continuity correction

data:  sub3 out of n, null probability 0.025
X-squared = 6041.2, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.025
95 percent confidence interval:
 0.09586782 0.10305981
sample estimates:
         p
0.09940597
```

The output shows that in the 2015 marathon, 9.94% of runners finished with a sub-3 hour time. Unsurprisingly then, the null hypothesis is rejected and we conclude that based on the data obtained from the 2015 marathon, there is no evidence to support the claim that only 2.5% of runners finish in less than 3 hours.

[3] https://www.runnersgoal.com/how-many-runners-can-run-a-marathon-in-under-4-hours/

Part 4: Z-Test (Two Proportions)

Question: Do the same proportion of runners finish in <3 hours in the <=25 and 26-30 age groups?

Null hypothesis: The proportion of runners finishing in <3 hours is the same in both age groups

Alternative hypothesis: The proportion of runners finishing in <3 hours is not the same in both age groups

```
        2-sample test for equality of proportions without continuity correction

data:  c(sub3_25, sub3_30) out of c(n_25, n_30)
X-squared = 2.4242, df = 1, p-value = 0.1195
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.040621242  0.004478771
sample estimates:
   prop 1    prop 2
0.1820609 0.2001321
```

The output shows that in the 2015 marathon, 18.2% and 20.0% of runners in the under 25 and 26-30 age groups completed the race in under three hours. However, the p-value indicates that there is a ~12% chance of observing these proportions in the sample (i.e. the 2015 marathon) under the null hypothesis (i.e. that the long-run proportions over many marathons are equal). Hence, we fail to reject the null hypothesis and conclude that there is no evidence that the proportion of runners finishing in under three hours is different in the under 25 and 26-30 age groups.

Topics covered

- (T1) Random variable, probability, conditional probability
- (T8) z-test (single proportion and two proportion)
- (T11) Chi-square test
- (T15) Graphs
- (T18) Anderson-Darling test