Mandatory assignment STK4011/9011 – Statistical Inference Theory

Department of Mathematics, University of Oslo

Autumn 2019

Submission deadline: *Friday 25 October 2019 14:30*, at the reception office on the 7th floor of Niels Henrik Abels Hus, Blindern.

Instructions: The written reports should be text-processed, preferably in LaTeX, and must be submitted as a single pdf-file. The submission must contain your name and course code. It is expected that you give a clear presentation with all necessary explanations, and that you write as concisely as possible. Remember to include all relevant plots and figures. These should preferably be placed in the text, close to the relevant subquestion. In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. The code should be placed in an appendix. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

Application for postponed delivery: If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (email: `studieinfo@math.uio.no`) well before the deadline. All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Exercise 1

Let $X_1, \ldots, X_n$ be i.i.d. random variables with positive density $f(x)$, distribution function $F(x)$ on some interval, and finite second moment. Set $Z_i = F(X_i)$ for $i = 1, \ldots, n$, and let $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n-1)} < Z_{(n)}$ be the order statistics.

(a) Show that $Z_{(i)}$ has density

$$g_i(z) = \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)} z^{i-1}(1-z)^{n-i}, \quad \text{on } (0,1).$$

Find the mean and the variance of $Z_{(i)}$.

(b) Suppose $n = 100$. Find the correlation between $Z_{(17)}$ and $Z_{(18)}$.

(c) Take for convenience $n = 2m+1$, and consider the median $M_n^0 = Z_{((n+1)/2)}$. Write down the density of $\sqrt{n}(M_n^0 - 1/2)$, say $f_n(x)$, and show that this $f_n(x)$ converges pointwise to the density of a mean zero normal with variance $1/4$. You may here use Stirling's formula, according to which $k!/\{k^k \exp(-k)\sqrt{2\pi k}\}$ tends to 1 as $k \to \infty$.

(d) Show that $X_{(i)}$ and $F^{-1}(Z_{(i)})$ must have the same distribution. Let M_n be the median of the X_i's, that is $M_n = X_{((n+1)/2)}$ with the odd choice of n. We'll use M_n as an estimator of the population median $\mu = F^{-1}(1/2)$. Show that the sample median M_n is consistent for the population median μ, i.e., that there is convergence in probability of M_n to μ. Also show that

$$\sqrt{n}(M_n - \mu) \xrightarrow{d} \mathrm{N}(0, \tau^2), \quad \text{with} \quad \tau^2 = \frac{1}{4f(\mu)^2}.$$

(e) Suppose the density $f(x)$ of the X_i's is symmetric around its centre point μ. Show that the mean $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ is consistent for μ. Give a formula for the approximate variance of M_n and compare it to the variance of $\bar{X}_n$. Illustrate this for the case of a normal distribution. Attempt to find a symmetric density where the sample median achieves a smaller variance than the sample mean.

Exercise 2

In love and football, to take just two examples, two minus one is a big number. In this exercise we'll deal with two-minus-one stuff in a statistical context. Suppose nature dictates that when it rains at Blindern, the amount of rain, measured in millimetres, follows a gamma distribution with parameter $\theta_0 = (a_0, b_0)$. The density of a gamma(a, b) distribution is

$$g(x;\theta) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx), \quad \text{for } x > 0, \tag{1}$$

and zero otherwise, for positive parameters a and b. Just as unknown to us as is the true parameter value, is the fact that the rain measurements have a density function in the class of gamma density functions. In our ignorance, therefore, we choose to model the rain data using an exponential distribution. In particular, sitting at our desk, we assume that the rain measurements, with sunny days deleted, are independent and come from a distribution with density

$$f(x;\lambda) = \lambda \exp(-\lambda x), \quad \text{for } x > 0, \tag{2}$$

and zero otherwise, for a positive parameter λ. In exercise (a) through (e) we we'll do statistics assuming the world is exponential, thereafter, in exercise (f) to (h), we'll play Laplace's demon and see how things map out in the true gamma state of affairs.

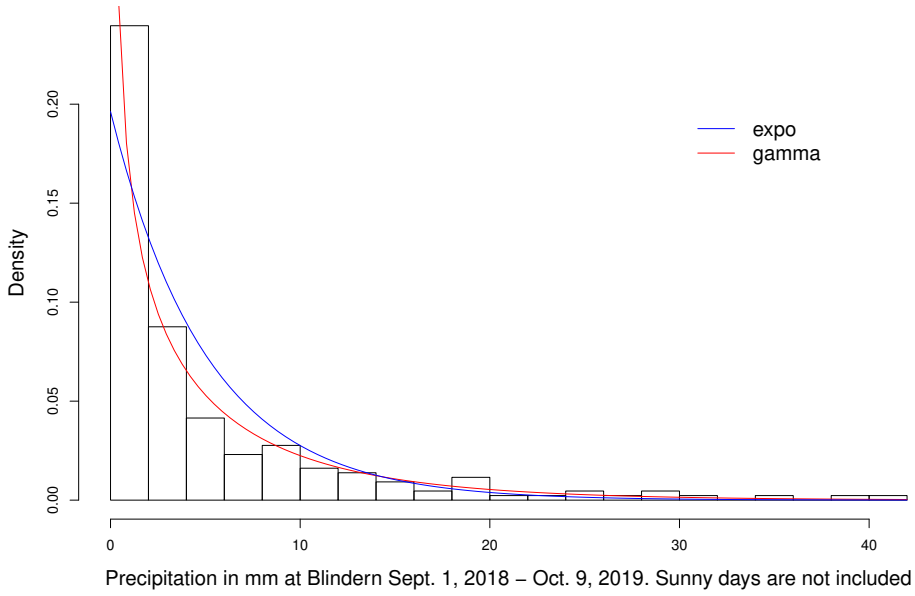

FIGURE 1. Rainfall measured in millimetres at Blindern weather station. Data from yr.no, extracted Oct. 10, 2019. See weather_blindern.txt for details.

(a) Let $X_1, \ldots, X_n$ be independent random variables, taken to stem from an exponential density of the form (2). Write down the log-likelihood function, and find the maximum likelihood estimator, say $\widehat{\lambda}_n$. Explain why $\widehat{\lambda}_n$ is unique, and compute its mean and variance under the model (i.e., under the assumption of exponentiality).

(b) Under our working assumption of exponentiality of the X_i's, find the Cramér–Rao lower bound for the variance of any unbiased estimator of λ. Construct an estimator $\widetilde{\lambda}_n = h(\widehat{\lambda}_n)$ that is unbiased for λ. Show that $\widetilde{\lambda}_n$ does not attain the Cramér–Rao lower bound, but that nevertheless, $\widetilde{\lambda}_n$ has the lowest variance, uniformly over $(0, \infty)$, among all unbiased estimators of λ.

(c) Reparametrise the $f(x;\lambda)$ model, that is, find a one-to-one function $\tau(\lambda) = \eta > 0$, so that an unbiased estimator of η, achieving the Cramér-Rao lower bound exists.

(d) By assuming the model in (2), we are in effect assuming that there is a true value of λ, denoted λ_0, such that the X_i are i.i.d. from $f(x, \lambda_0)$. Show that $\widehat{\lambda}_n$ is consistent for λ_0, and find the limiting distribution of $\sqrt{n}(\widehat{\lambda}_n - \lambda_0)$ as n tends to infinity.

(e) Find a function g so that $\sqrt{n}(g(\hat{\lambda}_n) - g(\lambda_0))$ converges in distribution to a standard normal. Use this to construct an approximate 95 percent confidence interval for λ_0 (still, under the model). Compare the confidence interval you found with an exact 95 percent confidence interval for λ_0.

(f) Now, we'll look at how our efforts in (a) through (d) appear to Laplace's demon. Recall that the Kullback–Leibler distance of a density f from a density g is

$$\mathrm{KL}(g, f) = \mathrm{E}_g \log(g(X)/f(X)) = \int \log\{g(x)/f(x)\}\, g(x)\, \mathrm{d}x.$$

Find the value of λ, which we'll call λ_{lf} (where the subscript lf stands for *least false*), that minimises $\mathrm{KL}(g(\cdot; \theta_0), f(\cdot; \lambda))$. Show that the estimator $\widehat{\lambda}_n$ you found in (a) is consistent for this value.

(g) Find the limiting distribution $\mathrm{N}(0, \kappa^2)$ of $\sqrt{n}(\widehat{\lambda}_n - \lambda_{\mathrm{lf}})$. Propose an estimator of κ^2 that does not require that you know the world is gamma. Show that your estimator is consistent for κ^2 (if it's not, look for another estimator).

(h) Go to the course site and download the data set weather_blindern.txt. Reproduce the plot in Figure 1. Provide an approximate 95 percent confidence interval for λ_0 assuming the world is exponential (e.g., using what you found in (d)). Use the variance estimator you found in (f) to construct an approximate 95 percent confidence interval for λ_{lf}, i.e., being agnostic about exponentiality. Are these two confidence intervals comparable? Still not knowing the world is really gamma(a, b), how would you construct an approximate confidence interval for a/b?

(i) Propose a more realistic model for the rain data.

Exercise 3

. . . summer time is the killing season/It's hot out this b.... that's a good enough reason, raps Curtis Jackson III, known as 50 cent, in a hit song from 2003. In this exercise we'll investigate whether this empirical insight, or rather a less violent version of it, holds true in Oslo in 2019. Crime data for Oslo is not openly available, at least I can't find any after some googling, so we'll have to settle with a proxy for crime, namely Oslo Police tweets. The user @oslopolitiops is an eager tweeter, and Twitter lets us freely download the 3200 last tweets of a user. Downloading these today (10 Oct. 2019), I obtained tweets dating back to May 1st 2019. In these tweets I searched for specific words ('pistol', 'kniv', 'slagsmål', and so on), indicating that something violent might be occurring in Oslo (I did not read all the tweets, so my count may be somewhat off. See the R-script make_twitterdata.r on the course page for details).

Before we proceed to a summer time & violence analysis, we'll make some rather bold assumptions about the twitter data to get to know the Poisson distribution. Assume that the violent tweet counts (the column `violence` in Table 1) are i.i.d. random variables from a Poisson distribution with expectation $\lambda > 0$. Denote these counts by $Y_1, \ldots, Y_n$, where $n = 163$, then

$$P_\lambda(Y = y) = \frac{\lambda^y}{y!} \exp(-\lambda), \quad \text{for } y = 0, 1, 2, \ldots.$$

(a) There are, fortunately, many days in our data where the counts of violence reporting tweets are zero. Find the maximum likelihood estimator, which we denote by $\widehat{\alpha}_n$, of $P_\lambda(Y = 0) = \exp(-\lambda)$. Compute its expectation and variance, and show by direct methods that $\widehat{\alpha}_n$ is consistent. Find also the Cramér–Rao lower bound of an unbiased estimator of $\exp(-\lambda)$.

(b) For $i = 1, \ldots, n$, let $Z_i = 1$ if $X_i = 0$, and $Z_i = 0$ otherwise. Consider $\widetilde{\alpha}_n = n^{-1} \sum_{i=1}^n Z_i$. Show that $\widetilde{\alpha}_n$ unbiased for $P_\lambda(Y = 0)$, and find the limit distributions

$$\sqrt{n}(\widehat{\alpha}_n - e^{-\lambda}) \xrightarrow{d} \mathrm{N}(0, \kappa_1^2),$$
$$\sqrt{n}(\widetilde{\alpha}_n - e^{-\lambda}) \xrightarrow{d} \mathrm{N}(0, \kappa_2^2),$$

with appropriate formulae for κ_1 and κ_2. Find an expression for the ratio κ_1^2/κ_2^2, called the *asymptotic relative efficiency*. Which of the two estimators do you prefer?

(c) With the aid of $\widetilde{\alpha}_n$, construct an estimator, say $\alpha_n^\star$, that is the best unbiased estimator of $\exp(-\lambda)$, in the sense that among all unbiased estimators, $\alpha_n^\star$ has the smallest variance uniformly over the parameter space. Explain why $\alpha_n^\star$ is unique.

(d) Go to the course website and download tweetcount_temp.txt. Compute the three point estimates for $P_\lambda(Y = 0)$.

(e) In order to test the summer time and violence hypothesis, we need to introduce covariates into our model. Let $x_1, \ldots, x_n$ for be the column `avg_temp` in Table 1. We are going to treat the x_i's as fixed and known covariates. Let $\theta = (\beta_0, \beta_1)$, and consider the Poisson regression model where $Y_1, \ldots, Y_n$ are independent with means

$$\lambda_i = \exp(\beta_0 + \beta_1 x_i), \quad \text{for } i = 1, \ldots, n. \tag{3}$$

Write down the log-likelihood function, the score functions, and provide an expression for the observed information matrix. Explain why assuming $n^{-1}\sum_{i=1}^n x_i^2 > \bar{x}_n^2$, with $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$, is sufficient for the maximum likelihood estimator $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1)$ to be unique in this model.

(f) Suppose we want to estimate a real valued function $\mu = \mu(\beta_0, \beta_1)$ of the parameters. This could for example be the probability of a zero count for a particular temperature x_0, in which case $\mu(\beta_0, \beta_1) = \exp\{-\exp(\beta_0 + \beta_1 x_0)\}$. Let $L(\delta, \mu)$ be a loss function that is convex in δ for every μ, and suppose that the estimator $\mu^\star$ is such that

$$R(\mu^\star, \mu) = \mathrm{E}\, L(\mu^\star, \mu) \le \mathrm{E}\, L(\delta, \mu) = R(\delta, \mu),$$

for every μ and every other estimator δ. Explain why $\mu^\star$ must be some function of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ you found in (e), that is, $\mu^\star = \mu^\star(\hat{\beta}_0, \hat{\beta}_1)$.

(g) Fit the Poisson regression model in (3) to the Oslo police twitter data using maximum likelihood. That is, take $Y_i = \texttt{violence}_i$ and $x_i = \texttt{avg\_temp}_i$ for $i = 1, \ldots, n$ in the notation of Table 1, and find the values of β_0 and β_1 maximising the log-likelihood function you found in (e). Provide an approximate 95 percent confidence interval for β_1. Do the insights of 50 cent in 2003 apply to Oslo in 2019?

(h) There is reason to believe, and a quick glance at Table 1 suggests, that an osloite's inclination towards violence is dependent on what day of the week it is. Consider therefore a model where we split the counts into seven groups, one for each day of the week. Associated with each weekday there is a parameter $\gamma_1, \ldots, \gamma_7$, where γ_1 is the Monday parameter, γ_2 the Tuesday parameter, and so on. We now label the counts $Y_{i,j}$, indicating that the i'th day in the data set is belongs to the j'th group, the group of Tuesdays, for example. Let $n_1, n_2, \ldots, n_7$ be the number of Mondays, Tuesdays, ..., Sundays in the data, so that $n = n_1 + \cdots + n_7$. The counts are taken as independent from

$$Y_{i,j} \mid x_i \sim \text{Poisson}(\gamma_j \exp(\beta x_i)), \quad \text{for } i = 1, \ldots, n.$$

This is now a model with eight parameters. Write down the full likelihood (or log-likelihood)

$$L_n(\beta, \gamma_1, \ldots, \gamma_7) = L_{n_1}(\beta, \gamma_1) \cdots L_{n_7}(\beta, \gamma_7),$$

and fit this model to the Oslo Police twitter data. Provide approximate confidence interval for each of them. Are Tuesdays and Fridays really different?

(i) Suppose now that the parameters $\gamma_1, \ldots, \gamma_7$ worked with in (h) are i.i.d. draws from an exponential distribution with mean $1/\eta$, so that our model is $Y_{i,j} \mid x_i, \gamma_j \sim \text{Poisson}(\gamma_j \exp(\beta x_i))$ independent, with $\gamma_j \sim \eta \exp(-\eta\gamma)$ independent for $j = 1, \ldots, 7$. Write down the full likelihood $L_n(\beta, \eta) = L_{n_1}(\beta, \eta) \cdots L_{n_7}(\beta, \eta)$, or the $\log L_n(\beta, \eta)$ version, and estimate the two parameters β and η. Supply these with 95 percent confidence intervals. How variable are the different weekdays, and how does the summer time & violence hypothesis fare under this model?

Appendix A. Police Twitter data

	date_dmy	day_of_week	total_tweets	violence	avg_temp
1	10-10-2019	Thursday	3	0	NA
2	09-10-2019	Wednesday	13	0	4.6
3	08-10-2019	Tuesday	9	0	3.4
4	07-10-2019	Monday	11	0	2.3
5	06-10-2019	Sunday	13	2	2.3
6	05-10-2019	Saturday	9	2	2.8
7	04-10-2019	Friday	7	1	3.7
8	03-10-2019	Thursday	5	1	5.0
9	02-10-2019	Wednesday	9	0	8.1
10	01-10-2019	Tuesday	10	0	8.5
11	30-09-2019	Monday	7	0	7.0
12	29-09-2019	Sunday	9	2	10.8
13	28-09-2019	Saturday	11	1	10.9
14	27-09-2019	Friday	9	0	9.6
15	26-09-2019	Thursday	5	1	10.4
16	25-09-2019	Wednesday	6	1	10.7
17	24-09-2019	Tuesday	12	1	10.5
18	23-09-2019	Monday	9	0	10.0
19	22-09-2019	Sunday	7	2	12.3
20	21-09-2019	Saturday	16	4	13.2
21	20-09-2019	Friday	13	1	12.1
22	19-09-2019	Thursday	9	0	8.1
23	18-09-2019	Wednesday	12	0	10.2
24	17-09-2019	Tuesday	8	1	9.8
25	16-09-2019	Monday	9	2	10.9
26	15-09-2019	Sunday	10	1	13.5
27	14-09-2019	Saturday	15	2	10.1
28	13-09-2019	Friday	7	0	12.1
29	12-09-2019	Thursday	9	0	12.7
30	11-09-2019	Wednesday	6	0	13.8
31	10-09-2019	Tuesday	5	0	14.9
32	09-09-2019	Monday	10	1	12.4
33	08-09-2019	Sunday	3	1	10.9
34	07-09-2019	Saturday	9	2	10.7
35	06-09-2019	Friday	9	0	10.2
36	05-09-2019	Thursday	10	0	11.6
37	04-09-2019	Wednesday	7	0	10.7
38	03-09-2019	Tuesday	5	1	12.5
39	02-09-2019	Monday	7	1	12.8
40	01-09-2019	Sunday	10	3	15.7
41	31-08-2019	Saturday	15	2	17.1
42	30-08-2019	Friday	7	0	15.3
43	29-08-2019	Thursday	13	1	17.6
44	28-08-2019	Wednesday	6	0	18.8
45	27-08-2019	Tuesday	4	0	19.5
46	26-08-2019	Monday	14	1	18.3
47	25-08-2019	Sunday	9	1	18.7
48	24-08-2019	Saturday	15	5	16.9
49	23-08-2019	Friday	14	2	16.8
50	22-08-2019	Thursday	11	2	12.8
51	21-08-2019	Wednesday	22	3	15.8
52	20-08-2019	Tuesday	6	0	15.5
53	19-08-2019	Monday	8	1	14.9
54	18-08-2019	Sunday	9	0	15.7
55	17-08-2019	Saturday	12	1	15.0
56	16-08-2019	Friday	10	1	15.6
57	15-08-2019	Thursday	12	1	14.2
58	14-08-2019	Wednesday	6	0	14.5
59	13-08-2019	Tuesday	4	1	16.7
60	12-08-2019	Monday	5	1	17.6
61	11-08-2019	Sunday	5	0	17.3
62	10-08-2019	Saturday	11	3	16.2
63	09-08-2019	Friday	10	1	18.2
64	08-08-2019	Thursday	12	1	17.7
65	07-08-2019	Wednesday	9	2	16.9
66	06-08-2019	Tuesday	8	1	16.1
67	05-08-2019	Monday	6	1	16.1
68	04-08-2019	Sunday	9	0	17.2
69	03-08-2019	Saturday	7	0	19.3
70	02-08-2019	Friday	11	0	20.3
71	01-08-2019	Thursday	11	4	18.5
72	31-07-2019	Wednesday	7	1	14.6
73	30-07-2019	Tuesday	5	0	14.3
74	29-07-2019	Monday	8	1	24.6
75	28-07-2019	Sunday	8	0	25.0
76	27-07-2019	Saturday	9	4	25.0
77	26-07-2019	Friday	5	1	25.3
78	25-07-2019	Thursday	6	1	21.6
79	24-07-2019	Wednesday	9	2	20.6
80	23-07-2019	Tuesday	10	0	20.3
81	22-07-2019	Monday	6	0	16.6
82	21-07-2019	Sunday	11	1	16.9
83	20-07-2019	Saturday	6	2	17.8
84	19-07-2019	Friday	8	0	16.1
85	18-07-2019	Thursday	6	0	16.5
86	17-07-2019	Wednesday	6	1	17.6
87	16-07-2019	Tuesday	6	2	16.2
88	15-07-2019	Monday	7	1	16.3
89	14-07-2019	Sunday	6	0	18.0
90	13-07-2019	Saturday	8	1	19.1
91	12-07-2019	Friday	10	4	18.9
92	11-07-2019	Thursday	9	0	19.7
93	10-07-2019	Wednesday	6	0	19.5
94	09-07-2019	Tuesday	7	0	17.4
95	08-07-2019	Monday	10	0	16.4
96	07-07-2019	Sunday	7	2	14.5
97	06-07-2019	Saturday	6	2	13.5
98	05-07-2019	Friday	8	2	15.7
99	04-07-2019	Thursday	9	2	12.9
100	03-07-2019	Wednesday	8	2	13.0
101	02-07-2019	Tuesday	8	2	14.3
102	01-07-2019	Monday	11	1	17.7
103	30-06-2019	Sunday	9	3	19.1
104	29-06-2019	Saturday	10	3	19.6
105	28-06-2019	Friday	15	0	18.7
106	27-06-2019	Thursday	14	0	17.3
107	26-06-2019	Wednesday	9	0	16.7
108	25-06-2019	Tuesday	7	1	15.4
109	24-06-2019	Monday	7	2	18.2
110	23-06-2019	Sunday	11	4	15.1
111	22-06-2019	Saturday	10	0	15.4
112	21-06-2019	Friday	11	2	14.8
113	20-06-2019	Thursday	11	1	15.6
114	19-06-2019	Wednesday	12	2	15.7
115	18-06-2019	Tuesday	9	1	15.5
116	17-06-2019	Monday	12	0	16.3
117	16-06-2019	Sunday	4	1	16.7
118	15-06-2019	Saturday	13	0	19.1
119	14-06-2019	Friday	11	1	15.0
120	13-06-2019	Thursday	6	0	11.7
121	12-06-2019	Wednesday	5	0	10.1
122	11-06-2019	Tuesday	11	1	15.3
123	10-06-2019	Monday	7	0	13.9
124	09-06-2019	Sunday	8	1	13.0
125	08-06-2019	Saturday	6	1	14.2
126	07-06-2019	Friday	13	0	15.4
127	06-06-2019	Thursday	13	0	18.9
128	05-06-2019	Wednesday	7	0	13.0
129	04-06-2019	Tuesday	15	2	13.7
130	03-06-2019	Monday	13	0	13.5
131	02-06-2019	Sunday	5	1	12.8
132	01-06-2019	Saturday	9	0	11.8
133	31-05-2019	Friday	10	1	12.1
134	30-05-2019	Thursday	5	1	9.2
135	29-05-2019	Wednesday	13	0	10.5
136	28-05-2019	Tuesday	11	3	10.3
137	27-05-2019	Monday	7	0	10.6
138	26-05-2019	Sunday	9	0	11.5
139	25-05-2019	Saturday	9	1	12.7
140	24-05-2019	Friday	4	0	10.2
141	23-05-2019	Thursday	2	0	11.6
142	22-05-2019	Wednesday	10	1	17.0
143	21-05-2019	Tuesday	9	2	18.7
144	20-05-2019	Monday	10	3	14.5
145	19-05-2019	Sunday	5	2	14.5
146	18-05-2019	Saturday	8	4	12.5
147	17-05-2019	Friday	10	4	17.4
148	16-05-2019	Thursday	13	1	16.3
149	15-05-2019	Wednesday	14	1	12.6
150	14-05-2019	Tuesday	11	1	10.5
151	13-05-2019	Monday	8	1	8.8
152	12-05-2019	Sunday	8	1	7.8
153	11-05-2019	Saturday	14	2	7.9
154	10-05-2019	Friday	4	0	8.7
155	09-05-2019	Thursday	16	1	6.6
156	08-05-2019	Wednesday	15	0	5.5
157	07-05-2019	Tuesday	7	3	4.9
158	06-05-2019	Monday	10	1	3.3
159	05-05-2019	Sunday	12	6	5.8
160	04-05-2019	Saturday	14	2	5.5
161	03-05-2019	Friday	11	0	4.0
162	02-05-2019	Thursday	10	1	6.8
163	01-05-2019	Wednesday	4	2	13.4

Table 1. Twitter data from the user @oslopolitiops. The average temperature is from yr.no. See the R script make_twitterdata.r on the course page for details.