Radboud Universiteit Nijmegen — Heyendaalseweg 135
Faculteit der Natuurwetenschappen, Wiskunde en Informatica — 6525 AJ Nijmegen

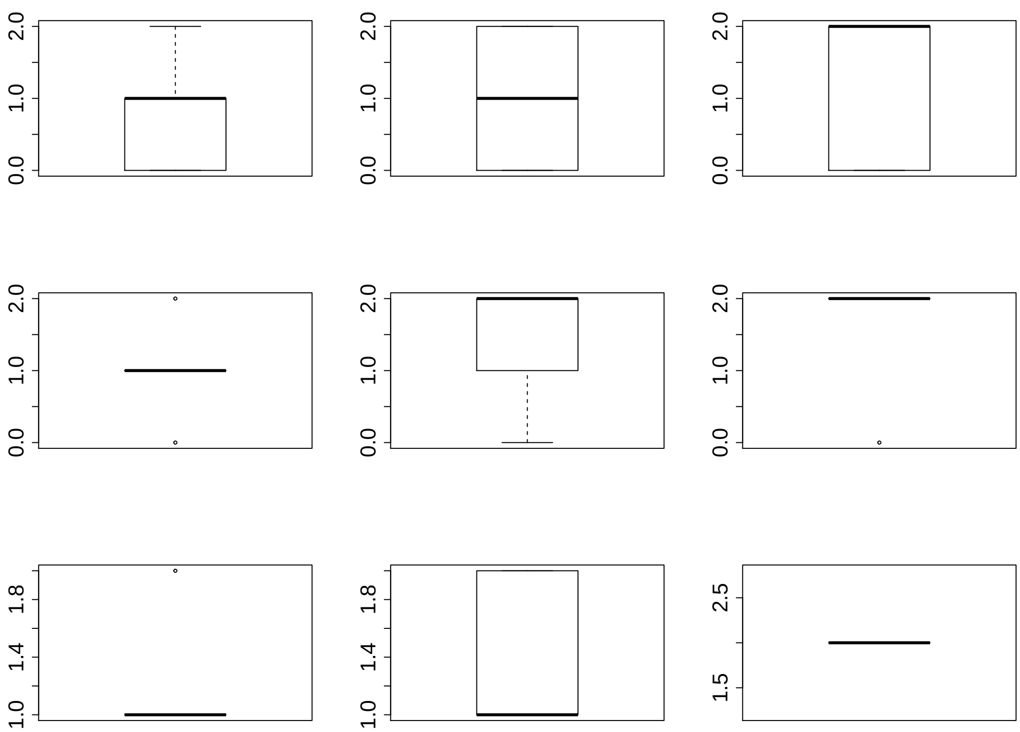
Tentamen — **3 november 2017**

Inleiding Statistiek — **14:45-17:45**

Het is toegestaan een (grafische) rekenmachine te gebruiken en het bijgeleverde formuleblad. Het is *niet* toegestaan een boek, aantekeningen, telefoons of apparaten met internetverbinding te gebruiken.

1. (a) Onderstaand ziet u boxplots van negen steekproeven van omvang $n = 7$ uit een $\mathrm{Bin}(2, \mathrm{p})$ verdeling. Hoeveel van deze steekproeven bevatten minstens 2 tweeën? U mag gebruiken dat het α-kwantiel x_α van een dataset gelijk is aan $x_{(k)}$ als $\alpha = \frac{k}{n+1}$. Motiveer uw antwoord!

   (b) Geef een zo goed mogelijke schatting voor p in de boxplot linksonder.



Z.O.Z.

2. Onderzoekers van de Radboud Universiteit waren nauw betrokken bij de recente waarneming van de samensmelting van twee neutronensterren. Deze waarneming vond plaats op 17 augustus 2017 met o.a. zwaartekrachtgolfdetectoren. Neem aan dat dit de enige waarneming in een tijdsbestek van drie maanden was en dat het aantal waarnemingen X per jaar een Pois(μ) verdeling heeft. U mag gebruiken dat het aantal waarnemingen Y in drie maanden dan een Pois($\mu/4$) verdeling heeft.

   (a) Laat zien dat de maximum likelihoodschatter voor μ gelijk is aan $\hat{\mu} = 4$ en schat hiermee $\mathbb{P}(X = 0)$.

   (b) Stel nu dat we met significantieniveau $\alpha = 0.05$ willen toetsen

   $$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$$

   Gebruik de likelihoodratio toetsingsgrootheid $\lambda(Y)$, en neem aan dat de verdeling onder H_0 met behulp van een χ^2-verdeling benaderd kan worden.

   Laat zien dat er een constante c bestaat zodat de nulhypothese verworpen wordt als

   $$\mu_0 - 4\log(\mu_0) > c,$$

   en benader de waarde van c.

   (c) Omdat de steekproef klein is, is het beter om Y zelf als toetsingsgrootheid te gebruiken. We willen opnieuw weten voor welke waarden van μ_0 de nulhypothese verworpen zal worden. Deze verzameling heeft de vorm $[0, c_1) \cup (c_2, \infty)$, waarbij $c_1 < 4 < c_2$. Bepaal een vergelijking voor c_1 en een vergelijking voor c_2.

3. De stochasten $X_1, \ldots, X_n$ zijn i.i.d. met kansdichtheidsfunctie

   $$f_\theta(x) = \begin{cases} 1/\theta & \theta \leq x \leq 2\theta, \\ 0 & \text{elders.} \end{cases}$$

   Hierbij is $\theta > 0$ een onbekende parameter. Schrijf $X_{(1)} = \min\{X_1, \ldots, X_n\}$, en $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

   (a) Bepaal de momentenschatter $\hat{\theta}_{mm}$ voor θ.

   (b) Laat zien dat de MSE van $\hat{\theta}_{mm}$ gelijk is aan $\dfrac{\theta^2}{27n}$.

   (c) Bepaal de likelihood functie $L(\theta; X)$ en schets deze. Bepaal vervolgens de maximum likelihood schatter $\hat{\theta}_{mle}$ voor θ.

   (d) De MSE van $\hat{\theta}_{mle}$ is gelijk aan $\dfrac{\theta^2}{2n^2 + 6n + 4}$. Welke van de twee schatters voor θ, $\hat{\theta}_{mm}$ or $\hat{\theta}_{mle}$, heeft uw voorkeur? Motiveer uw antwoord.

4. De discrete stochasten $X_1, \dots, X_n$ zijn i.i.d. met kansmassafunctie

$$f_p(k) = P_p(X = k) = \begin{cases} \binom{k+5}{k}(1-p)^6 p^k & k = 0, 1, 2, \dots \\ 0 & \text{elders.} \end{cases}$$

Hierbij is $p \in (0, 1)$ een onbekende parameter. Als a priori verdeling voor p nemen we de Beta(α, β) verdeling met $\alpha > 0$, $\beta > 0$.

(a) Laat zien dat de Bayes schatter T voor p gelijk is aan $\dfrac{\alpha + \sum_{i=1}^n X_i}{\alpha + \sum_{i=1}^n X_i + \beta + 6n}$.

(b) De maximum likelihood schatter voor p is gelijk aan

$$\hat{p}_n = \frac{\sum_{i=1}^n X_i}{6n + \sum_{i=1}^n X_i}.$$

Is het mogelijk de a priori verdeling zo te kiezen dat de Bayes schatter altijd gelijk is aan de maximum likelihood schatter? Zo ja, hoe moeten α en β gekozen worden?

5. Het KNMI heeft recent beweerd dat de kans op een kwakkelherfst meer is dan 30%. Een onderzoeker heeft op basis van historische data over een periode van 100 jaar vastgesteld dat er 23 kwakkelherfsten waren. Neem aan dat verschillende jaren onafhankelijk zijn en laat eventuele klimaatverandering buiten beschouwing.

(a) Als we de onderzoeksgegevens beschouwen als een rij enen en nullen (wel of geen kwakkelherfst), wat is dan de steekproefvariantie van deze rij?

(b) We willen de bewering van het KNMI toetsen met $\alpha = 0.05$. Formuleer het toetsingsprobleem en bepaal het kritieke gebied.

(c) Bereken de p-waarde behorend bij de gedane observatie en trek uw conclusie.

Hint: Bij de onderdelen 5b en 5c kunt u gebruik maken van de normale benadering.

EINDE

Puntenverdeling volgens onderstaande tabel.

Opgave	1	2	3	4	5	Gratis	Totaal
Punten	15	20	20	15	20	10	100

Indien u p punten scoort en de cijfers b_1 en b_2 voor de bonusopdrachten behaald hebt, dan is uw eindcijfer

$$\frac{2p + b_1 + b_2}{200}.$$

Formuleblad bij Statistiek

DIT BLAD IS NIET EEN SAMENVATTING OF OVERZICHT, EN DIENT SLECHTS ALS HULPMIDDEL.

Kansverdelingen

1. **Alternatieve of Bernoulli verdeling**: Alt(p) of Ber(p).

   $\mathrm{P}(X=1)=p$ en $\mathrm{P}(X=0)=1-p$. EX$=p$; Var $(X)=p(1-p)$.

2. **Binomiale verdeling**: Bin(n,p).

   $\mathrm{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ voor $k=0,1,\dots,n$. EX$=np$; Var $(X)=np(1-p)$.

3. **Geometrische verdeling** : Geo(p).

   $\mathrm{P}(X=k)=p(1-p)^{k-1}$ voor $k=1,2,\dots$. EX$=1/p$; Var $(X)=(1-p)/p^2$.

4. **Poisson-verdeling**: Pois(μ).

   $\mathrm{P}(X=k)=\dfrac{\mu^k}{k!}e^{-\mu}$ voor $k=0,1,\dots$. EX$=\mu$; Var $(X)=\mu$.

5. **Exponentiële verdeling**: Exp(λ).

   $f(x)=\lambda e^{-\lambda x}$ en $F(x)=1-e^{-\lambda x}$ voor $x\geq 0$. EX$=1/\lambda$; Var $(X)=1/\lambda^2$.

6. **Homogene of Uniforme verdeling** op $[a,b]$: hom$[a,b]$ of $U(a,b)$.

   $f(x)=\dfrac{1}{b-a}$ en $F(x)=\dfrac{x-a}{b-a}$ voor $a\leq x\leq b$. EX$=\frac{1}{2}(a+b)$; Var $(X)=\frac{1}{12}(b-a)^2$.

7. **Normale verdeling**: $N(\mu,\sigma^2)$.

   $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$ en $F(x)=\displaystyle\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2}dt$. EX$=\mu$; Var $(X)=\sigma^2$.

8. **Gamma verdeling**: $\Gamma_{\alpha,\lambda}$ $(\alpha,\lambda>0)$.

   $f(x)=\dfrac{x^{\alpha-1}\lambda^\alpha e^{-\lambda x}}{\Gamma(\alpha)}1_{[0,\infty)}(x)$ met $\Gamma(\alpha)=\displaystyle\int_0^\infty t^{\alpha-1}e^{-t}\,dt$. EX$=\alpha/\lambda$; Var $(X)=$ α/λ^2.

9. **Beta verdeling**: $B_{\alpha,\beta}$ $(\alpha,\beta>0)$.

   $f(x)=\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}1_{[0,1]}(x)$ met $B(\alpha,\beta)=\displaystyle\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt$.

   EX$=\dfrac{\alpha}{\alpha+\beta}$; Var $(X)=\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Covariantie en correlatie

1. Definitie covariantie: $\mathrm{Cov}(X,Y)=E(X-EX)(Y-EY)$.
   Eigenschappen: $\mathrm{Cov}(X,Y)=EXY-EX\cdot EY$, $\mathrm{Cov}(rX+s,tY+u)=rt\,\mathrm{Cov}(X,Y)$.

2. Definitie correlatiecoëfficiënt: $\rho(X,Y)=\dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$.

3. $\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)+2\,\mathrm{Cov}(X,Y)$.

Empirische verdelingsfunctie

Voor een dataset van n elementen: $F_n(x)=\dfrac{\text{aantal elementen in de dataset }\leq x}{n}$.

Centrale limietstelling

Als $X_1, X_2, \ldots$ een rij onafhankelijke stochasten is met alle dezelfde verdeling en met verwachting μ and variantie σ^2, dan geldt:

Centrale limietstelling:

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) = \mathrm{P}\left(Z \leq a\right),$$

waarbij Z een $N(0,1)$ verdeling heeft.

Locatie-schaal families

De locatie-schaal familie van dichtheden behorende bij een dichtheid f, wordt gegeven door

$$\left\{ f_{a,\lambda}(\cdot) = \frac{1}{\lambda} f\left(\frac{\cdot - a}{\lambda}\right) \mid a \in \mathbb{R}, \lambda > 0 \right\}.$$

Als $Z \sim f$ en $X = \lambda Z + a$, dan $X \sim f_{a,\lambda}$.

Kwantielen

Het α-onderkwantiel van een stochast X met verdelingsfunctie F wordt gedefinieerd als

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} \mid \mathrm{P}\left(X \leq x\right) \geq \alpha\}.$$

Als $X = \lambda Z + a$, dan $F_X^{-1}(\alpha) = \lambda F_Z^{-1}(\alpha) + a$. Voor een dataset $x_1, \ldots, x_n$ schatten we het α-onderkwantiel $\hat{x}_\alpha$ door $k = \lfloor \alpha(n+1) \rfloor$, $\eta = \alpha(n+1) - k$ en $\hat{x}_\alpha = (1-\eta)x_{(k)} + \eta x_{(k+1)}$.

Schatters

De onzuiverheid (bias) van een schatter T voor θ: $\quad ET - \theta$.
Als T_1 en T_2 zuivere schatters voor θ zijn, dan heet T_2 efficiënter dan T_1 als $\mathrm{Var}\,(T_2) < \mathrm{Var}\,(T_1)$.
De 'mean squared error' van een schatter T voor θ: $\quad \mathrm{MSE}(T) = E(T-\theta)^2$.
Eigenschap: $\mathrm{MSE}(T) = \mathrm{Var}\,(T) + (ET - \theta)^2$.

Gestandaardiseerd en gestudentiseerd steekproefgemiddelde

Als X_i een $N(\mu, \sigma^2)$ verdeling hebben voor $i = 1, \ldots, n$, en onafhankelijk zijn, dan hebben

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \text{en} \quad \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

een $N(0,1)$, respectievelijk $t(n-1)-$verdeling. Hierbij is $t(n-1)$ de t-verdeling met $n-1$ vrijheidsgraden. Verder geldt

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2(n-1).$$

Twee steekproeven

Bij de twee steekproeven t-toets:

Gelijke varianties: $S_p^2 = \dfrac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}\left(\dfrac{1}{n} + \dfrac{1}{m}\right)$.

Ongelijke varianties: $S_d^2 = \dfrac{S_X^2}{n} + \dfrac{S_Y^2}{m}$.

a	0	1	2	3	4	5	6	7	8	9
0.0	5000	4960	4920	4880	4840	4801	4761	4721	4681	4641
0.1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
0.2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
0.3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
0.4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
0.5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
0.6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
0.7	2420	2389	2358	2327	2296	2266	2236	2206	2177	2148
0.8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
0.9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
1.0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
1.1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
1.2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
1.3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
1.4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
1.5	0668	0655	0643	0630	0618	0606	0594	0582	0571	0559
1.6	0548	0537	0526	0516	0505	0495	0485	0475	0465	0455
1.7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
1.8	0359	0351	0344	0336	0329	0322	0314	0307	0301	0294
1.9	0287	0281	0274	0268	0262	0256	0250	0244	0239	0233
2.0	0228	0222	0217	0212	0207	0202	0197	0192	0188	0183
2.1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
2.2	0139	0136	0132	0129	0125	0122	0119	0116	0113	0110
2.3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
2.4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
2.5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0048
2.6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
2.7	0035	0034	0033	0032	0031	0030	0029	0028	0027	0026
2.8	0026	0025	0024	0023	0023	0022	0021	0021	0020	0019
2.9	0019	0018	0018	0017	0016	0016	0015	0015	0014	0014
3.0	0013	0013	0013	0012	0012	0011	0011	0011	0010	0010
3.1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
3.2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
3.3	0005	0005	0005	0004	0004	0004	0004	0004	0004	0003
3.4	0003	0003	0003	0003	0003	0003	0003	0003	0003	0002

Tabel 1: *Rechteroverschrijdingskans* $1 - \Phi(a) = P(Z \geq a)$ *van de* $N(0,1)$*-variabele* Z.

$m \backslash p$	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
∞	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Tabel 2: Right critical values $t_{m,p}$ of the t-distribution with m degrees of freedom corresponding to right tail probability p: $\mathrm{P}(T_m \geq t_{m,p}) = p$. The last row in the table are right critical values of the $N(0,1)$ distribution: $t_{\infty,p} = z_p$

m	$\alpha = 0.995$	0.99	0.975	0.95	0.90	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.3	12.8
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.1	13.3	14.9
5	0.41	0.55	0.83	1.15	1.61	9.24	11.1	12.8	15.1	16.8
6	0.68	0.87	1.24	1.64	2.20	10.6	12.6	14.4	16.8	18.5
7	0.99	1.24	1.69	2.17	2.83	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	92.0
70	43.2	45.4	48.8	51.7	55.3	85.5	90.5	95.0	100.4	104.2
100	67.3	70.0	74.2	77.9	82.4	118.5	124.3	129.6	135.8	140.2

Tabel 3: Right critical values $\chi^2_p(m)$ of the χ^2-distribution with m degrees of freedom corresponding to right tail probability p: $\mathrm{P}\left(\chi^2(m) \geq \chi^2_p(m)\right) = p$.

Volledige oplossingen:

1a Er zijn minstens twee tweeën precies als het 6e element $x_{(6)}$ in de geordende steekproef gelijk is aan 2. Het 6e element is ook precies het 3e kwartiel ($\alpha = \frac{3}{4}$), en dat is de bovenkant van de box in de boxplot. In 6 van de 9 plaatjes zit de bovenkant van de box bij 2.

1b In de boxplot zie je dat 0 niet voorkwam in de steekproef en dat 2 slechts een keer voorkwam. De som van de steekproef is dus $6 \cdot 1 + 2 = 8$. Dat is het aantal successen in 14 pogingen, dus het ligt voor de hand om p te schatten met $\hat{p} = \frac{8}{14} = \frac{4}{7}$.

2a Het aantal waarnemingen Y in drie maanden heeft een Pois($\mu/4$) verdeling, en we hebben één observatie, namelijk $Y_1 = 1$. De kansmassafunctie van Y is

$$\mathbb{P}(Y = k) = \frac{(\mu/4)^k}{k!} e^{-\mu/4},$$

dus de likelihoodfunctie is

$$L(\mu; Y) = \prod_{i=1}^{1} \mathbb{P}(Y = Y_i) = \frac{(\mu/4)^{Y_1}}{Y_1!} e^{-\mu/4} = \frac{\mu}{4} e^{-\mu/4}.$$

De log-likelihoodfunctie en haar afgeleide zijn derhalve

$$l(\mu; Y) = \log(\mu) - \log(4) - \frac{\mu}{4}; \qquad l'(\mu; Y) = \frac{1}{\mu} - \frac{1}{4}.$$

Gelijk stellen aan nul levert inderdaad $\hat{\mu} = 4$ als maximum likelihoodschatter. Hiermee kunnen we de kans op nul observaties in een jaar schatten:

$$\mathbb{P}_{\hat{\mu}}(X = 0) = e^{-\hat{\mu}} = e^{-4} \approx 0.018.$$

2b De toetsingsgrootheid is in dit geval

$$\lambda(Y) = \frac{\sup_\mu L(\mu; Y)}{L(\mu_0; Y)} = \frac{L(\hat{\mu}; Y)}{L(\mu_0; Y)}.$$

Het is equivalent om naar de log hiervan te kijken:

$$\log(\lambda(Y)) = l(\hat{\mu}; Y) - l(\mu_0; Y) = -1 - \left(\log(\mu_0) - \log(4) - \frac{\mu_0}{4}\right) = \frac{\mu_0}{4} - \log(\mu_0) + \log\left(\frac{4}{e}\right).$$

Onder H_0 heeft $2\log(\lambda(Y))$ bij benadering een $\chi^2(1)$-verdeling, met kritieke waarde 3.84. We moeten bepalen voor welke waarden van μ_0 deze waarde overschreden wordt, dus we verwerpen als

$$2\left(\frac{\mu_0}{4} - \log(\mu_0) + \log\left(\frac{4}{e}\right)\right) > 3.84.$$

Dit geeft

$$\mu_0 - 4\log(\mu_0) > 2 \cdot 3.84 - 4\log\left(\frac{4}{e}\right) \approx 6.13.$$

De numerieke oplossing hiervan is dat H_0 verworpen wordt als $\mu_0 \notin (0.23, 17.6)$.

2c Merk op dat het hier gaat om een tweezijdige toets, dus we zouden H_0 kunnen verwerpen als Y te klein is, maar ook als Y te groot is. In dit geval verwerpen we de nulhypothese als 1 in het kritieke gebied ligt, dus als $\mathbb{P}_{\mu_0}(Y \leq 1) \leq 0.025$ of als $\mathbb{P}_{\mu_0}(Y \geq 1) \geq 0.025$. De constanten c_1 en c_2 moeten zo gekozen worden dat de eerste ongelijkheid optreedt als $\mu_0 > c_2$ en de tweede als $\mu_0 < c_1$. Er geldt:

$$\mathbb{P}_{\mu_0}(Y \leq 1) = e^{-\mu_0/4} + \frac{\mu_0}{4} e^{-\mu_0/4}, \qquad \mathbb{P}_{\mu_0}(Y \geq 1) = 1 - e^{-\mu_0/4}.$$

Stellen we deze uitdrukkingen gelijk aan 0.025, dan vinden we als positieve oplossingen respectievelijk

$$\mu_0 \approx 22.3 \qquad \text{en} \qquad \mu_0 = -4\log(0.975) \approx 0.10,$$

Dus $c_1 \approx 0.10$ en $c_2 \approx 22.3$. We zouden H_0 dus verwerpen als $\mu_0 \notin (0.10, 22.3)$. Op basis van het feit dat we één waarneming hebben in drie maanden kunnen we dus niet zo heel veel zeggen over de frequentie, het kan variëren van 1 per 10 jaar tot 22 per jaar.

3a Merk op dat $X_1, \ldots, X_n$ een homogene verdeling op $[\theta, 2\theta]$ verdeling hebben. Daarom $\mathbb{E}_\theta X_1 = \frac{3}{2}\theta$. Om de momentenschatter te vinden, lossen we op $\mathbb{E}_\theta X_1 = \overline{X}_n$. Dit geeft $\hat{\theta}_{mm} = \frac{2}{3}\overline{X}_n$.

3b Merk op dat $\mathbb{E}_\theta \overline{X}_n = \mathbb{E}_\theta X_1$ en $\mathrm{Var}_\theta \mathrm{\overline{X}_n} = \mathrm{Var}_\theta \mathrm{X_1/n}$. Derhalve

$$\mathbb{E}_\theta \hat{\theta}_{mm} = \mathbb{E}_\theta\Big(\frac{2}{3}\overline{X}_n\Big) = \frac{2}{3} \cdot \mathbb{E}_\theta \overline{X}_n = \frac{2}{3} \cdot \mathbb{E}_\theta X_1 = \theta,$$

dus de momentenschatter is zuiver.

Vervolgens merken we op dat $\mathrm{Var}_\theta \mathrm{X_1} = \frac{\theta^2}{12}$, en dus

$$\mathrm{Var}_\theta \hat{\theta}_{\mathrm{mm}} = \mathrm{Var}_\theta\Big(\frac{2}{3}\mathrm{\overline{X}_n}\Big) = \frac{4}{9} \cdot \mathrm{Var}_\theta \mathrm{\overline{X}_n} = \frac{4}{9} \cdot \frac{\mathrm{Var}_\theta \mathrm{X_1}}{\mathrm{n}} = \frac{\theta^2}{27\mathrm{n}}.$$

Omdat de bias 0 is, is de MSE van $\hat{\theta}_{mm}$ gelijk aan deze variantie.

3c De definitie van de likelihood functie is

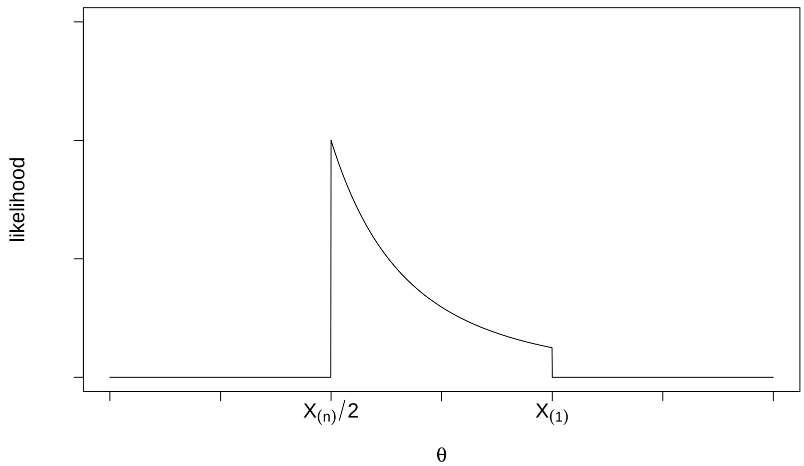
$$L(\theta; X) = \prod_{i=1}^{n} f_\theta(X_i).$$

Als θ zodanig is dat $X_i \notin [\theta, 2\theta]$, dan $f_\theta(X_i) = 0$ en dus is de likelihood functie dan ook 0. Als daarentegen $X_i \in [\theta, 2\theta]$ voor alle i, dan is de likelihood functie gelijk aan $1/\theta^n$. In het bijzonder geldt dan dat $\theta \leq X_{(1)} \leq X_{(n)} \leq 2\theta$, dus

$$L(\theta; X) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \frac{1}{2}X_{(n)} \leq \theta \leq X_{(1)} \\ 0 & \text{elders} \end{cases}$$

Het maximum wordt aangenomen in $\theta = X_{(n)}/2$, dus $\hat{\theta}_{mle} = X_{(n)}/2$. Zie ook bijgaande schets.

3d De MSE van $\hat{\theta}_{mle}$ gaat harder naar 0 dan de MSE van $\hat{\theta}_{mm}$, daarom verdient de maximum likelihood schatter de voorkeur. (Er geldt $\hat{\theta}_{mle} < \hat{\theta}_{mm}$ voor $n \geq 11$.)



Figuur 1: Schets van de likelihood functie

4a Eerst bepalen we de likelihood functie:

$$L(p; X) = \prod_{i=1}^{n} \binom{X_i + 5}{X_i} (1-p)^6 p^{X_i} = \left(\prod_{i=1}^{n} \binom{X_i + 5}{X_i} \right) (1-p)^{6n} p^{\sum_{i=1}^{n} X_i}.$$

De a posteriori dichtheid is evenredig met het product van de a priori dichtheid $\pi(p)$ (zie formuleblad) en de likelihood functie:

$$\begin{aligned} L(p; X) \cdot \pi(p) &\propto (1-p)^{6n} p^{\sum_{i=1}^{n} X_i} \cdot p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha-1+\sum_{i=1}^{n} X_i}(1-p)^{\beta+6n-1}. \end{aligned}$$

Hierin herkennen we de dichtheidsfunctie van een Beta$(\alpha + \sum_{i=1}^{n} X_i, \beta + 6n)$ verdeling. De Bayes schatter is de verwachting van de a posteriori verdeling, dus

$$T = \frac{\alpha + \sum_{i=1}^{n} X_i}{\alpha + \sum_{i=1}^{n} X_i + \beta + 6n}.$$

4b We moeten oplossen

$$\frac{\alpha + \sum_{i=1}^{n} X_i}{\alpha + \sum_{i=1}^{n} X_i + \beta + 6n} = \frac{\sum_{i=1}^{n} X_i}{6n + \sum_{i=1}^{n} X_i}.$$

Dit is equivalent met

$$\Big(6n + \sum_{i=1}^{n} X_i\Big) \cdot \Big(\alpha + \sum_{i=1}^{n} X_i\Big) = \Big(\alpha + \sum_{i=1}^{n} X_i + \beta + 6n\Big) \cdot \sum_{i=1}^{n} X_i,$$

en als we wat dingen wegstrepen blijkt dat moet gelden

$$\beta \sum_{i=1}^{n} X_i = \alpha \cdot 6n.$$

Omdat α en β strikt positief moeten zijn kan deze gelijkheid alleen optreden als $\alpha = \beta\bar{X}_n/6$. We kunnen α en β niet zodanig kiezen dat dit altijd zo is, omdat α en β niet mogen afhangen van de data. Het is immers een *a priori* verdeling.

5a De steekproefvariantie is

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{99}\sum_{i=1}^{100}\left(X_i - \frac{23}{100}\right) \\ &= \frac{1}{99}\left(23 \cdot (1 - \frac{23}{100})^2) + 77 \cdot (0 - \frac{23}{100})^2\right) = \frac{23 \cdot 0.77^2 + 77 \cdot 0.23^2}{99} \approx 0.179. \end{aligned}$$

5b Als model nemen we dat het aantal kwakkelherfsten Y een $\mathrm{Bin}(100, p)$ verdeling heeft. De hypothesen zijn

$$H_0 : p > 0.3 \qquad H_1 : p \leq 0.3.$$

De nulhypothese zal verworpen worden voor kleine waarden van Y, en de kans op ten onrechte verwerpen moet kleiner zijn dan α. We moeten de grootste waarde van k vinden waarvoor onder de nulhypothese geldt $P(Y \leq k) < 0.05$. Voor $p \geq 0.3$ geldt met behulp van de normale benadering

$$\begin{aligned} P_p(Y \leq k) \leq P_{p=0.3}(Y \leq k) &\approx P\left(N(0,1) \leq \frac{k + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \\ &= P\left(N(0,1) \leq \frac{k - 29.5}{\sqrt{21}}\right) \\ &= \Phi\left(\frac{k - 29.5}{\sqrt{21}}\right) \end{aligned}$$

Dit is gelijk aan 0.05 als $\frac{k-29.5}{\sqrt{21}} = \xi_{0.05} = -1.645$ (tabel), dus als $k = 29.5 - \sqrt{21} \cdot 1.645 \approx 21.96$. Om aan de veilige kant te zitten moeten we afronden naar beneden, dus het kritieke gebied is $\{0, 1, \ldots, 20, 21\}$.

5c De p-waarde is (opnieuw met de normale benadering en de tabel)

$$P_{p=0.3}(Y \leq 23) \approx \Phi\left(\frac{23 - 29.5}{\sqrt{21}}\right) \approx \Phi(-1.418) \approx 0.0798.$$

Er is dus niet voldoende grond om H_0 te verwerpen.