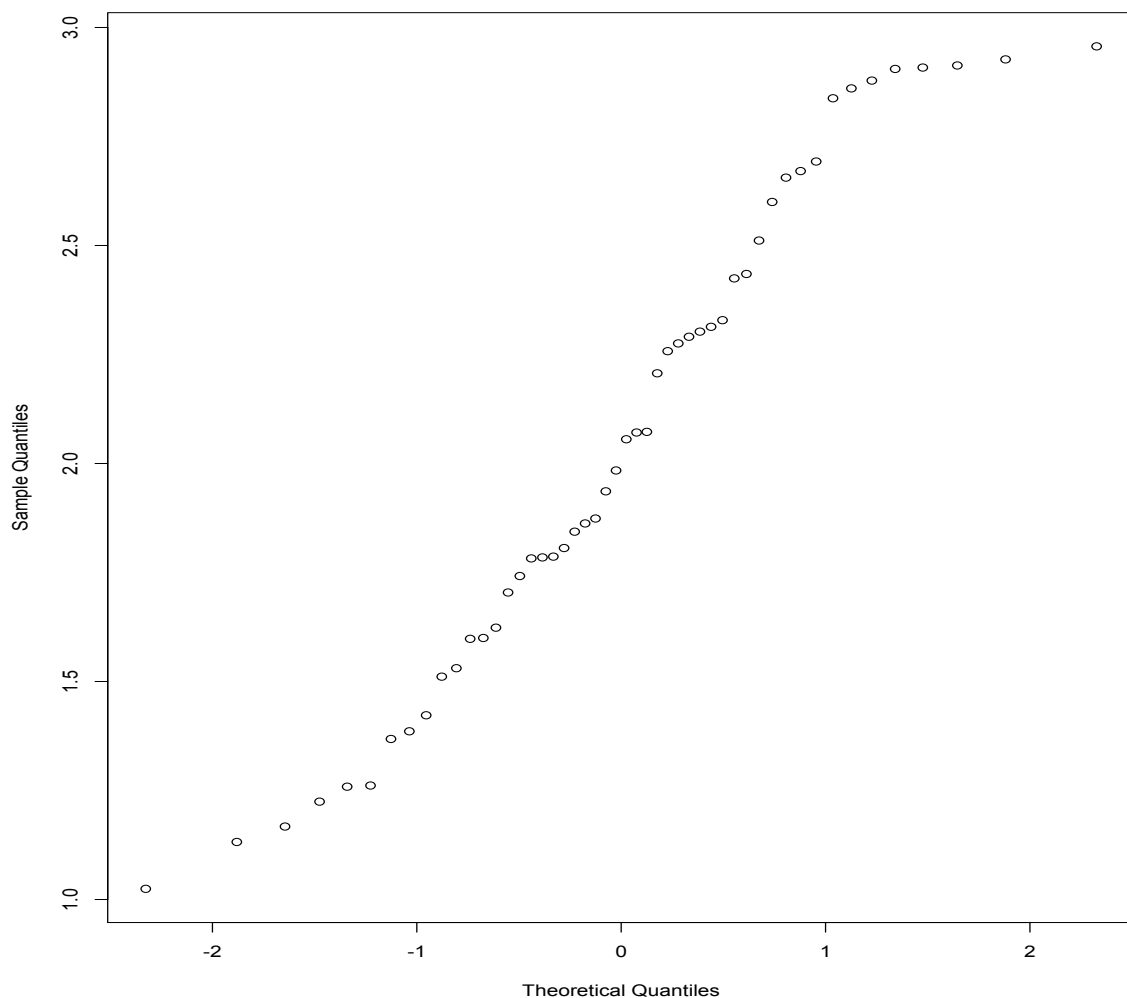


Het is toegestaan een rekenmachine te gebruiken en het bijgeleverde formuleblad. Het is *niet* toegestaan een boek of aantekeningen te gebruiken.

1. Van een dataset met 50 punten hebben we een QQ-plot gemaakt tegen de standaard normale kwantielen.



- (a) Welke van de volgende families van dichtheden past het beste bij deze dataset: de locatie-schaal familie van de homogene verdeling, de locatie-schaal familie van de normale verdeling of de locatie-schaal familie van de exponentiële verdeling? Licht uw antwoord duidelijk toe!
- (b) Maak een boxplot van deze data. Wees precies en geef relevante locaties van de elementen van uw boxplot!

2. Een bedrijf wil weten of de succeskans van een experiment afhangt van de temperatuur. Ze weten dat bij de huidige temperatuur de kans gelijk is aan $\frac{1}{2}$. Nu voeren ze $n = 25$ experimenten uit, en bij elk volgend experiment voeren ze de temperatuur een klein beetje op. Ze modelleren de succeskans p_i van experiment i als volgt:

$$p_i(b) = \frac{1}{1 + e^{-bi}},$$

voor een parameter $b \in \mathbb{R}$. De uitkomst modelleren we met de onafhankelijke stochasten $Y_1, \dots, Y_n$, waarbij

$$P(Y_i = 1) = 1 - P(Y_i = 0) = p_i(b).$$

- (a) Laat zien dat de log-likelihood $l(b; Y)$ voldoet aan

$$l(b; Y) = \sum_{i=1}^n -Y_i \log(1 + e^{-bi}) - (1 - Y_i) \log(1 + e^{bi}).$$

Laat ook zien dat de maximum likelihoodschatter $\hat{b}$ voldoet aan:

$$\sum_{i=1}^n i p_i(\hat{b}) = \sum_{i=1}^n i Y_i.$$

- (b) Een onderzoeker wil R gebruiken om b te schatten. Zij gebruikt hiervoor de volgende code, waarbij y de vector is met onze $n = 25$ meetpunten:

```
bhat1=-10 #je mag er vanuit gaan dat dit klein genoeg is
bhat2=10  #je mag er vanuit gaan dat dit groot genoeg is

N=10000
n=25

b=seq(from=bhat1, to=bhat2, length=N)
I=1:n

m=sum(-y*log(1+exp(-b[1]*I))-(1-y)*log(1+exp(b[1]*I)))
index=1

for (i in 2:N){
  f=sum(-y*log(1+exp(-b[i]*I))-(1-y)*log(1+exp(b[i]*I)))
  if (f>m){
    m=f
    index=i
  }
}

bhat=b[index]
```

Verklaar wat er in deze code gebeurt.

- (c) Na het runnen van de code blijkt $l(\hat{b}; Y) = -14.8$. Toets de hypothese $H_0 : b = 0$ tegen $H_1 : b \neq 0$. Gebruik de likelihood ratio toets bij een significantie van 5% en geef uw conclusie. Geef ook de p-waarde van deze toets.

3. Het bedrijf KolenCo wil een lading steenkool kopen, maar alleen als de calorische waarde van de lading groot genoeg is. We modelleren de gemeten calorische waarde van een monster steenkool uit de betreffende lading als een normale verdeling met verwachting μ en variantie σ^2 . KolenCo meet nauwkeurig de calorische waarde van 25 monsters, en vindt een gemiddelde van 23 MJ/kg, met een steekproefstandaarddeviatie van 2.0 MJ/kg. De lading moet met grote zekerheid minstens een gemiddelde calorische waarde hebben van 22 MJ/kg.
- (a) Stel de juiste nul-hypothese en alternatieve hypothese op voor KolenCo, en toets deze hypothese bij een significantie van 5%. Geef ook een (zo scherp mogelijke) onder- en bovengrens voor de p-waarde.
 - (b) Beschouw nu de toets $H_0 : \mu \geq \mu_0$ tegen $H_1 : \mu < \mu_0$. Voor welke verzameling $\mu_0 \in \mathbb{R}$ kunnen we H_0 *niet* verwerpen bij een significantie van 5%? Wat zou de betekenis van deze verzameling kunnen zijn?
4. Een groep artsen wil laten zien dat hun nieuwe behandeling voor een specifieke doelgroep patiënten bij meer dan 50% tot succes leidt. Noem p de kans op succes van de behandeling. Zij toetsen dus $H_0 : p \leq 0.5$ tegen $H_1 : p > 0.5$ bij een significantie van 5%. Noem n het aantal patiënten in hun steekproef.
- (a) Hoe groot moeten de artsen n kiezen om 90% zeker te zijn dat H_0 wordt verworpen als de echte kans op succes $p = 0.7$?
 - (b) De groep artsen behandelt $n = 50$ patiënten met de nieuwe behandeling, en heeft succes bij 30 patiënten. Een andere behandeling heeft bij een soortgelijk experiment bij $m = 80$ patiënten succes bij 40 patiënten. De artsen vinden dat dit duidelijk laat zien dat hun nieuwe behandeling een hogere succeskans heeft. Bent u het hiermee eens? Stel de relevante hypothese op en geef ook een p-waarde.

ER VOLGT NOG EEN SOM 5!

5. We zeggen dat een stochastische variabele N een *discrete Gamma* verdeling heeft met parameters $a \in [0, 1)$ en $k \in \{1, 2, \dots\}$, notatie $N \sim \text{DG}(a, k)$, als geldt dat

$$P(N = m) = \binom{m}{k} a^{m-k} (1-a)^{k+1} \quad \text{voor } m \in \{k, k+1, k+2, \dots\}.$$

Er geldt dus dat voor alle $a \in [0, 1)$ en $k \in \{1, 2, \dots\}$ dat

$$\sum_{m=k}^{\infty} \binom{m}{k} a^{m-k} (1-a)^{k+1} = 1.$$

- (a) Bewijs dat als $N \sim \text{DG}(a, k)$, dan geldt

$$E(N) = \frac{k+1}{1-a} - 1.$$

Hint: beschouw $E(N+1)$ en gebruik dat

$$(m+1) \binom{m}{k} = (k+1) \binom{m+1}{k+1}.$$

- (b) Nu beschouwen een variabele $X \sim \text{Bin}(N, p)$, waarbij $p \in [0, 1]$ vast is gekozen, en $N \in \{1, 2, \dots\}$ de onbekende parameter is. We kiezen als a priori verdeling voor N een $\text{Geo}(q)$ -verdeling met $q \in (0, 1]$, dus

$$P(N = m) = q(1-q)^{m-1} \quad \text{voor } m \in \{1, 2, \dots\}.$$

Laat zien dat de a posteriori verdeling van N (dus de verdeling van $N \mid X = x$) een discrete Gamma verdeling is, en laat zien dat de Bayes schatter $T(x)$ voor N gegeven wordt door

$$T(x) = \frac{x + (1-p)(1-q)}{p + q - pq}.$$

- (c) Stel $X \sim \text{Bin}(N, p)$ en $N = (1-q)/q$. Laat zien dat voor deze N geldt dat $E(T(X)) = N$. Bepaal ook de Mean Squared Error van $T(X)$ in dit geval als functie van p en q (u hoeft uw antwoord niet te vereenvoudigen).

Formuleblad bij Statistiek

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

Kansverdelingen

1. **Alternatieve of Bernoulli verdeling:** $\text{Alt}(p)$ of $\text{Ber}(p)$.

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

2. **Binomiale verdeling:** $\text{Bin}(n, p)$.

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

3. **Geometrische verdeling:** $\text{Geo}(p)$.

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

4. **Poisson-verdeling:** $\text{Pois}(\mu)$.

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

5. **Exponentiële verdeling:** $\text{Exp}(\lambda)$.

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

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

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

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

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

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

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

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

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

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

Covariantie en correlatie

1. Definitie covariantie: $\text{Cov}(X, Y) = E(X - EX)(Y - EY)$.

$$\text{Eigenschappen: } \text{Cov}(X, Y) = EXY - EX \cdot EY, \quad \text{Cov}(rX + s, tY + u) = rt \text{Cov}(X, Y).$$

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

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

Empirische verdelingsfunctie

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

Centrale limietstelling

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

Centrale limietstelling:

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

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

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

Als $X = \lambda Z + a$, dan $F_X^{-1}(\alpha) = \lambda F_Z^{-1}(\alpha) + a$. Voor een dataset $x_1, \dots, x_n$ 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 : $ET - \theta$.

Als T_1 en T_2 zuivere schatters voor θ zijn, dan heet T_2 efficiënter dan T_1 als $\text{Var}(T_2) < \text{Var}(T_1)$.

De ‘mean squared error’ van een schatter T voor θ : $\text{MSE}(T) = E(T - \theta)^2$.

Eigenschap: $\text{MSE}(T) = \text{Var}(T) + (ET - \theta)^2$.

Gestandaardiseerd en gestudentiseerd steekproefgemiddelde

Als X_i een $N(\mu, \sigma^2)$ verdeling hebben voor $i = 1, \dots, n$, 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

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

Twee steekproeven

Bij de twee steekproeven t -toets:

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

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

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 : $P(T_m \geq t_{m,p}) = p$. The last row in the table are right critical values of the $N(0, 1)$ distribution: $t_{\infty,p} = z_p$

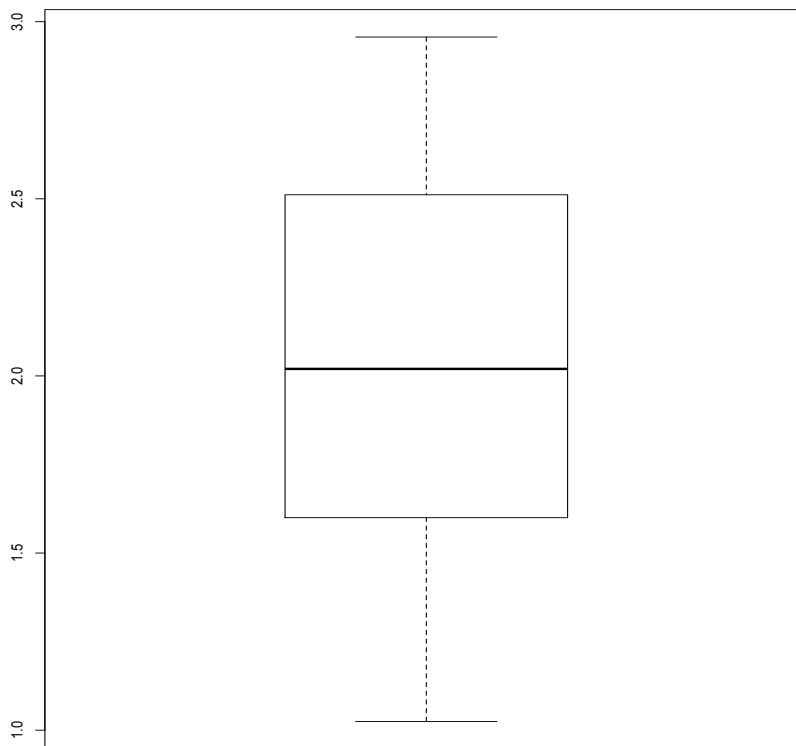
m	$\alpha = 0.995$	0.99	0.975	0.95	0.90	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.3	12.8
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.1	13.3	14.9
5	0.41	0.55	0.83	1.15	1.61	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 $\chi_p^2(m)$ of the χ^2 -distribution with m degrees of freedom corresponding to right tail probability p : $P(\chi^2(m) \geq \chi_p^2(m)) = p$.

Full solutions:

1a De staarten van de verdeling zijn aan beide kanten dunner dan die van de normale verdeling. Van de drie gegeven families heeft alleen de homogene verdeling deze eigenschap (de homogene verdeling heeft aan beide kanten dunnere staarten, de exponentiële verdeling heeft links een dunnere staart en rechts een zwaardere staart).

1b De mediaan ligt boven de 0, dat is ongeveer bij 2.0. Het onderste kwartiel kan op twee manieren gevonden worden: we kunnen een z vinden zodat $P(N(0,1) < z) = 0.25$: dit geeft $z = -0.675$ (ongeveer). Boven deze z vinden we de QQ-plot ter hoogte van ongeveer 1.6. Een andere methode is om datapunt nummer $(50 + 1)/4 = 12.75$ (ongeveer 13 dus) te vinden: die ligt op hoogte 1.6 (ongeveer). Dit is dus het eerste kwartiel. Het derde kwartiel kan op soortgelijke wijze gevonden worden. We vinden: 2.55. Dus de Inter Quartile Range is 0.95. De beneden whisker gaat dus tot ten laagste $1.6 - 1.5 * 0.95 = 0.18$. Het laagste punt ligt bij 1.05, boven de 0.18, en dat is dus de locatie van de beneden whisker. Naar boven toe krijgen we het hoogste punt voor de whisker: $2.55 + 1.5 * 0.95 = 3.98$. Het hoogste datapunt ligt ter hoogte 2.9. We hebben dus geenuitschieters. We krijgen het volgende plaatje:



2a We stellen eerst de log-likelihood op:

$$\begin{aligned}
 l(b; Y) &= \sum_{i=1}^n \log(f_b(Y_i)) \\
 &= \sum_{i=1}^n 1_{\{Y_i=1\}} \log(p_i(b)) + 1_{\{Y_i=0\}} \log(1 - p_i(b)) \\
 &= \sum_{i=1}^n -Y_i \log(1 + e^{-bi}) + (1 - Y_i) \log\left(\frac{e^{-bi}}{1 + e^{-bi}}\right) \\
 &= \sum_{i=1}^n -Y_i \log(1 + e^{-bi}) - (1 - Y_i) \log(1 + e^{bi}).
 \end{aligned}$$

Om dit te maximaliseren over b , hoeven we slechts de afgeleide naar b gelijk aan 0 te stellen:

$$\begin{aligned}
 \frac{dl}{db}(b) = 0 &\iff \sum_{i=1}^n \frac{iY_i e^{-bi}}{1 + e^{-bi}} - \frac{i(1 - Y_i)e^{bi}}{1 + e^{bi}} = 0 \\
 &\iff \sum_{i=1}^n \frac{iY_i}{1 + e^{bi}} + \frac{iY_i e^{bi}}{1 + e^{bi}} - \frac{ie^{bi}}{1 + e^{bi}} = 0 \\
 &\iff \sum_{i=1}^n iY_i = \sum_{i=1}^n ip_i(b).
 \end{aligned}$$

2b De code moet het minimum vinden van de functie $l(b)$. We kiezen een grid van $N = 10000$ punten tussen -10 en 10 , rekenen de log-likelihood uit in al die punten en onthouden waar het maximum zit. In de `for`-loop geldt altijd dat `m` de maximale waarde tot dan toe is, en `index` de locatie van dat maximum tot dan toe. Aan het eind van de loop is `b[index]` dus de locatie van het globale maximum over $[-10, 10]$, met een nauwkeurigheid van 0.0001.

2c Uit het gegeven blijkt dat $l(\hat{b}) = -14.8$. Ook geldt dat

$$l(0) = \sum_{i=1}^n -Y_i \log(2) - (1 - Y_i) \log(2) = -n \log(2) = -25 \cdot \log(2) = -17.3.$$

Dus

$$2l(\hat{b}) - 2l(0) = 5.$$

We gebruiken nu dat het model onder H_0 geen parameters meer bevat, en dat het model onder H_1 precies 1 parameter bevat. Dit betekent dat twee keer de log-likelihood ratio toetsingsgrootte onder H_0 een $\chi^2(1)$ -verdeling heeft, in goede benadering. De p-waarde is dus in goede benadering

$$P(\chi^2(1) > 5) = P(Z^2 > 5) = 2P(Z > 2.24) = 0.025.$$

We verwerpen dus H_0 : er is bewijs dat de succes kans afhangt van de temperatuur.

3a We kiezen $H_0 : \mu \leq 22$ en $H_1 : \mu > 22$. We gebruiken de t -toets:

$$T = \sqrt{n} \frac{\bar{x}_n - 22}{s_n} = 2.5.$$

Onder (de meest ongunstigste parameter in) H_0 geldt dat $T \sim t(24)$. We verwerpen voor grote (sterk positieve) waarden van T . De kritieke waarde $t(24)_{0.05} = 1.711 < 2.5$, dus we verwerpen H_0 ten gunste van H_1 . Merk op dat de p -waarde gelijk is aan $P(t(15) > 2)$. Nu geldt dat $P(t(15) > 2.492) = 0.01$ en $P(t(15) > 2.797) = 0.005$. We zien dus dat $0.01 < p\text{-waarde} < 0.005$, en omdat $2.492 \approx 2.5$, ligt de p -waarde net onder 0.01.

3b We gebruiken de t -toets:

$$T = \sqrt{n} \frac{\bar{x}_n - \mu_0}{s_n} = \frac{5}{2}(23 - \mu_0).$$

We verwerpen H_0 voor sterk negatieve waarden van T . Aangezien onder (de meest ongunstigste parameter in) H_0 geldt dat $T \sim t(24)$, verwerpen we als $T < -t(24)_{0.05} = -1.711$. Dit betekent dat we niet verwerpen als

$$\frac{5}{2}(23 - \mu_0) \geq -t(24)_{0.05} \iff \mu_0 \leq 23 + \frac{2}{5}t(24)_{0.05} = 23.7.$$

De verzameling $(-\infty, 23.7]$ is een boven-betrouwbaarheidsinterval voor μ : dit betekent dat we dankzij de data tamelijk zeker zijn dat de gemiddelde calorische waarde van de lading steenkool lager is dan 23.7 MJ/kg.

4a We gebruiken de volgende toetsingsgrootheid: zij X het aantal patiënten dat met succes wordt behandeld, dus $X \sim \text{Bin}(n, p)$. De artsen willen de hypothese $H_0 : p \leq 0.5$ verwerpen ten gunste van $H_1 : p > 0.5$. Zij verwerpen H_0 als $X \geq k$, voor een kritieke waarde k . Deze kritieke waarde moet voldoen aan:

$$\sup_{p \in H_0} P_p(X \geq k) = P_{p=0.5}(X \geq k) \leq 0.05.$$

We gebruiken nu de normale benadering om k als functie van n te bepalen:

$$\begin{aligned} P_{0.5}(X \geq k) &= P_{0.5}\left(X \geq k - \frac{1}{2}\right) \\ &= P_{0.5}\left(\frac{X - 0.5n}{0.5\sqrt{n}} \geq \frac{k - \frac{1}{2} - 0.5n}{0.5\sqrt{n}}\right) \\ &\approx P\left(N(0, 1) \geq \frac{k - \frac{1}{2} - 0.5n}{0.5\sqrt{n}}\right) \\ &= 0.05. \end{aligned}$$

Dit betekent dat met $\alpha = 0.05$ en $P(N(0, 1) \geq z_\alpha) = \alpha$, er geldt dat

$$k = 0.5n + 0.5z_\alpha\sqrt{n} + \frac{1}{2}.$$

Nu gebruiken we dat ook moet gelden dat de kans op het verwerpen van H_0 gelijk is aan $\beta = 0.9$, als $p = 0.7$:

$$P_{0.7}(X \geq k) = 0.7 \iff P\left(N(0, 1) \geq \frac{k - \frac{1}{2} - 0.7n}{\sqrt{0.21n}}\right) = 0.9.$$

Hieruit volgt

$$-0.2n + 0.5z_{\alpha}\sqrt{n} = z_{1-\beta}\sqrt{0.21n}.$$

Hieruit volgt dat $n = 49.7$, ofwel we kiezen $n = 50$.

4b Stel X is het aantal patiënten dat baat heeft bij de nieuwe behandeling, dus $X \sim \text{Bin}(n, p)$ met $n = 50$, en stel Y is het aantal patiënten dat baat heeft bij de andere behandeling, dus $Y \sim \text{Bin}(m, q)$, met $m = 80$. We toetsen $H_0 : p \leq q$ tegen $H_1 : p > q$. We schatten p en q met $\hat{p} = 0.6$ en $\hat{q} = 0.5$. De beste manier is om uit te gaan van $p = q$ (worst case van H_0). Noem $p_0 = p = q$. Dan volgt dat

$$\text{Var}(\hat{p} - \hat{q}) = \frac{p_0(1-p_0)}{n} + \frac{p_0(1-p_0)}{m}.$$

We schatten p_0 met $(30 + 40)/(50 + 80) = 7/13$. Onze toetsingsgrootte wordt

$$T = \frac{\hat{p} - \hat{q}}{\sqrt{\hat{p}_0(1-\hat{p}_0)(\frac{1}{n} + \frac{1}{m})}} = 1.113.$$

De p-waarde van de toets is dus $P(N(0,1) > 1.113) = 0.13$. Deze p-waarde is te groot om H_0 te verwerpen. De artsen hebben dus nog geen overtuigend bewijs dat de nieuwe behandeling beter werkt dan de andere behandeling. De volgende toetsingsgrootte, waarbij de variantie van $\hat{p}$ en $\hat{q}$ apart worden berekend, wordt ook goed gerekend:

$$T = \frac{\hat{p} - \hat{q}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{m}}} = 1.123.$$

De p-waarde hier is ook 13%, en de conclusie is hetzelfde.

5a

$$\begin{aligned} E(N+1) &= \sum_{m=k}^{\infty} (m+1) \binom{m}{k} a^{m-k} (1-a)^{k+1} \\ &= \sum_{m=k}^{\infty} (k+1) \binom{m+1}{k+1} a^{m-k} (1-a)^{k+1} \\ &= (k+1) \sum_{m=k}^{\infty} \binom{m+1}{k+1} a^{m+1-(k+1)} (1-a)^{(k+2)-1} \\ &= \frac{k+1}{1-a} \sum_{m=k+1}^{\infty} \binom{m}{k+1} a^{m-(k+1)} (1-a)^{k+2} \\ &= \frac{k+1}{1-a}. \end{aligned}$$

Dit laat zien dat

$$E(N) = \frac{k+1}{1-a} - 1.$$

5b De a posteriori kansverdeling voor N , noem het $\pi(m)$, is proportioneel met het product van de likelihood en de a priori. Bedenk dat alleen factoren met een m erin van belang zijn! Dus

$$\begin{aligned}\pi(m) &\propto \binom{m}{x} (1-p)^{m-x} p^x \cdot q(1-q)^{m-1} 1_{\{m \geq x\}} \\ &\propto \binom{m}{x} ((1-p)(1-q))^m 1_{\{m \geq x\}}.\end{aligned}$$

Hieruit volgt dat $N \mid X = x \sim \text{DG}((1-p)(1-q), x)!$ Voor de Bayes schatter hoeven we nu alleen nog de verwachting uit te rekenen:

$$T(x) = E(N \mid X = x) = \frac{x+1}{1 - (1-p)(1-q)} - 1 = \frac{x + (1-p)(1-q)}{p + q - pq}.$$

5c Er geldt

$$E(T(X)) = \frac{Np + (1-p)(1-q)}{p + q(1-p)} = N \iff (1-p)(1-q) = q(1-p)N \iff N = \frac{1-q}{q}.$$

Dus geldt voor deze N dat

$$\text{MSE}(T) = \text{Var}(T) = \frac{\text{Var}(X)}{(p + q - pq)^2} = \frac{Np(1-p)}{(p + q - pq)^2} = \frac{p(1-p)(1-q)}{q(p + q - pq)^2}.$$