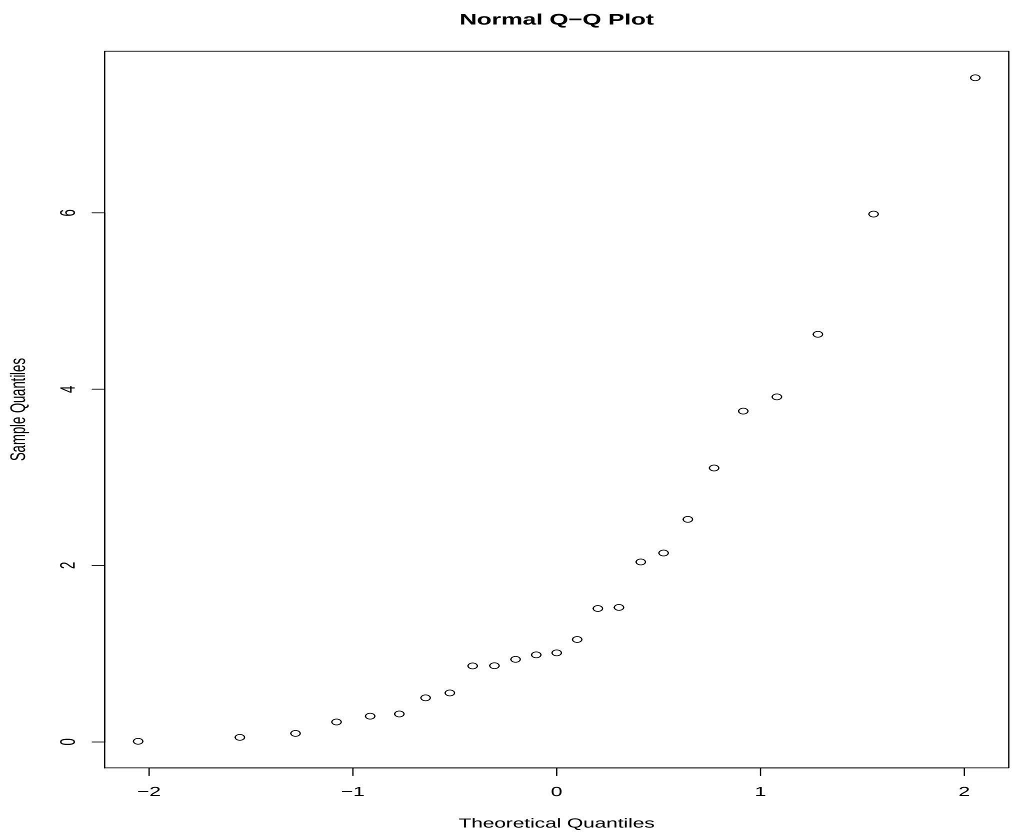
Radboud University Nijmegen	Heyendaalseweg 135
Faculty FNWI	6525 AJ Nijmegen
Tentamen	**25 oktober 2016**
Inleiding Statistitiek	**8:30 - 11:30**

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

1. We hebben een normale QQ-plot gemaakt van een dataset met 25 punten.



(a) Deze dataset is een trekking uit een exponentiële verdeling met parameter λ. Geef een schatting voor λ, en licht uw schatting toe.

(b) Maak een boxplot van deze data. Werk met een redelijke mate van precisie!

2. We hebben een ijkgewicht een aantal keer gemeten 2 jaar geleden. Dit leverde de metingen $X_1, \ldots, X_n$ op, met $n = 10$. Nu meten we hetzelfde gewicht weer, maar dit keer $m = 20$ keer. Dit levert de metingen $Y_1, \ldots, Y_m$ op. We modelleren beide sets metingen als en steekproef uit een normale verdeling, met $X_i \sim N(\mu_1, \sigma^2)$ en $Y_j \sim N(\mu_2, \sigma^2)$. Hierbij is σ bekend, en gelijk aan 0.05g (dit wordt bepaald door onze weeginstrument). We vinden $\bar{X}_n = 100.02$g en $\bar{Y}_m = 99.96$g.

   (a) Bepaal de maximum likelihood schatter voor (μ_1, μ_2). Geef ook de maximum likelihood schatter voor de parameter in het kleinere model

   $$\{(\mu_1, \mu_2) \in \mathbb{R}^2 \mid \mu_1 = \mu_2\},$$

   waar we er dus vanuit gaan dat beide steekproeven uit dezelfde normale verdeling komen.

   (b) We willen de nul-hypothese toetsen dat $\mu_1 = \mu_2$, tegen $\mu_1 \neq \mu_2$. Voer de likelihood ratio toets uit met een significantie van 5%, en geef de p-waarde. U mag gebruiken dat voor getallen $a_1, \ldots, a_n, b \in \mathbb{R}$ geldt dat

   $$\sum_{i=1}^{n}(a_i - b)^2 = \sum_{i=1}^{n}(a_i - \bar{a}_n)^2 + n(b - \bar{a}_n)^2.$$

   (c) Een onderzoekster schrijft de volgende functie in R:

   ```
   p=function(mu1,mu2){
   L= (sum((x-mu1)^2) + sum((y-mu2)^2) - sum((x-mean(x))^2)
         - sum((y-mean(y))^2))/sigma^2
   return(1-pchisq(L,df=2))
   }
   ```

   Hier geldt dat x $= (X_1, \ldots, X_n)$, y $= (Y_1, \ldots, Y_m)$ en sigma$=0.05$. Zij krijgt het volgende resultaat:

   ```
   > p(100,100)
   [1] 0.0007465858
   ```

   Wat voor conclusie trekt u hieruit? Licht uw antwoord duidelijk toe!

3. We bekijken een veld met klaverbladen. In dit veld hebben de plantjes twee, drie of vier blaadjes. We vinden op een klein deel van het veld $N_2 = 20$ plantjes met 2 blaadjes, $N_3 = 160$ plantjes met 3 blaadjes en $N_4 = 40$ plantjes met 4 blaadjes.

   (a) We willen laten zien dat het gemiddelde aantal blaadjes van een klaverblad groter is dan 3. Stel de juiste nul-hypothese en alternatieve hypothese op, en toets deze hypothese bij een significantie van 5%. Geef ook een (zo goed mogelijke) indicatie van de p-waarde. Hint: definieer X_i als het aantal blaadjes van plant i. Wat is de steekproefstandaarddeviatie van $X_1, \ldots, X_n$?

   (b) Noem p_2 de kans op twee blaadjes, en p_4 de kans op 4 blaadjes. Toets de hypothese $H_0 : p_2 = p_4$ tegen $H_1 : p_2 \neq p_4$. Verzin een geschikte toetsingsgrootheid en geef de p-waarde van uw toets, eventueel met behulp van een normale benadering. U mag gebruiken dat als $p_2 = p_4 = p$ (en dus $p \leq \frac{1}{2}$), dan geldt

   $$\mathrm{Cov}(N_2, N_4) = -np^2.$$

4. We bekijken twee ladingen computerchips, en onderwerpen een steekproef uit beide ladingen aan een test. Bij de steekproef uit lading A bleken 5 van de 1000 geteste chips niet in orde, en bij de steekproef uit lading B bleken 6 van de 500 geteste chips niet in orde.

   (a) Gebruik de likelihood ratio toets om de hypothese te toetsen dat de kans om onze kwaliteitstest te falen voor chips uit beide ladingen gelijk is, bij een significantie van 5%, en geef de p-waarde.

   (b) We verwachten dat voor lading B zo'n 1.5% van de chips niet aan de test kan voldoen. We willen bewijzen bij een significantie van 5% dat de kans op falen bij lading B kleiner is dan 2%. Hoeveel chips moeten we testen van lading B om 90% zeker te zijn dat we de nulhypothese $H_0 : p_B \geq 0.02$ kunnen verwerpen, als we aannemen dat eigenlijk $p_B = 0.015$? Gebruik de normale benadering!

5. We tellen het aantal vossen in stukken bos van 1 vierkante kilometer, en krijgen zo een steekproef $X_1, \ldots, X_n$, waarbij $X_i \sim \text{Pois}(\mu)$. We gebruiken als a priori verdeling voor μ een exponentiële verdeling met parameter λ.

   (a) Laat zien dat de Bayes schatter T voor μ gegeven wordt door

$$T = \frac{1 + \sum_{i=1}^n X_i}{\lambda + n}.$$

   (b) Laat zien dat

$$\text{MSE}(T; \mu) = \frac{\lambda^2 \mu^2 + (n - 2\lambda)\mu + 1}{(n + \lambda)^2}.$$

   (c) Definieer $T_1 = \bar{X}_n$ (dit is de maximum likelihood schatter). Bepaal $\text{MSE}(T_1; \mu)$, en laat zien dat voor μ klein genoeg,

$$\text{MSE}(T_1; \mu) < \text{MSE}(T; \mu).$$

   Waarom moeten er ook $\mu > 0$ zijn waarvoor geldt dat $\text{MSE}(T_1; \mu) > \text{MSE}(T; \mu)$?

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 $k=0,1,\ldots,n$. EX$=np$; Var$(X)=np(1-p)$.
3. **Geometrische verdeling** : Geo(p).

   $\mathrm{P}(X=k)=p(1-p)^{k-1}$ voor $k=1,2,\ldots$. 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 $k=0,1,\ldots$. 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$(X)=\alpha/\lambda^2$.
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: $\mathrm{P}\left(T_m \geq t_{m,p}\right) = p$. 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

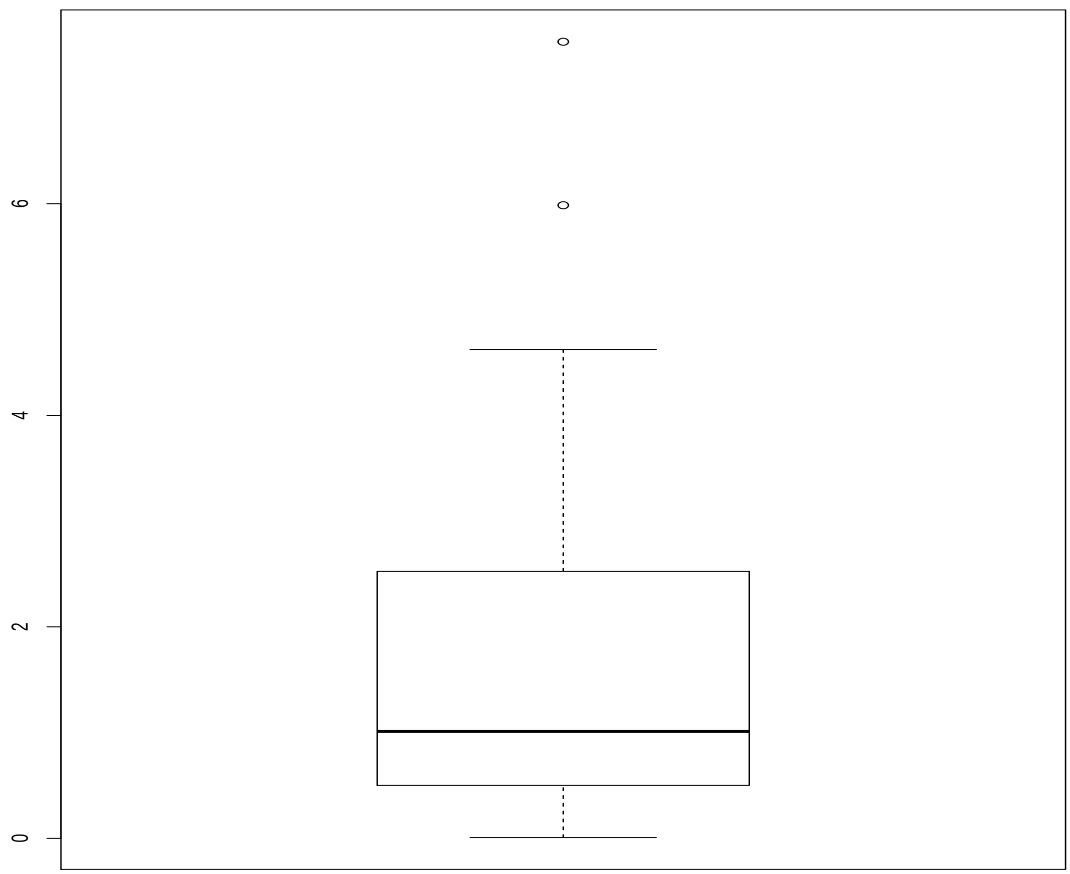
Tabel 3: Right critical values $\chi^2_p(m)$ of the χ^2-distribution with m degrees of freedom corresponding to right tail probability p: $\mathrm{P}\left(\chi^2(m) \geq \chi^2_p(m)\right) = p$.

Full solutions:

1a We bepalen de mediaan van de verdeling: deze ligt boven 0, en is ongeveer gelijk aan 1. Dus er moet ongeveer gelden dat $\mathbb{P}(\text{Exp}(\lambda) > 1) = 0.5$, ofwel

$$e^{-\lambda} = 1/2 \Leftrightarrow \lambda = \log(2) = 0.7.$$

1b



2a We weten dat voor een normale steekproef, de MLE voor de parameter gegeven is door het gemiddelde, dus $\hat{\mu}_1 = \bar{X}_n$ en $\hat{\mu}_2 = \bar{Y}_m$. Dit volgt direct uit de log-likelihood:

$$l(\mu_1; X) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2.$$

Als $\mu_1 = \mu_2$, dan zijn beide steekproeven getrokken uit dezelfde normale verdeling, dus dan krijgen we dat

$$\hat{\mu} = \frac{n\bar{X}_n + m\bar{Y}_m}{n+m} = 99.98\text{g}.$$

2b We weten al dat de log-likelihood gegeven wordt door

$$l(\mu_1, \mu_2; X, Y) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_1)^2 - \frac{m}{2}\log(2\pi) - m\log(\sigma) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j - \mu_2)^2.$$

Als we invullen $(\mu_1, \mu_2) = (\bar{X}_n, \bar{Y}_m)$, dan krijgen we

$$l(\hat{\mu}_1, \hat{\mu}_2; X, Y) = -\frac{n+m}{2}\log(2\pi) - (n+m)\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_i - \bar{Y}_m)^2\right).$$

Onder H_0 is de MLE voor μ_1 en μ_2 gelijk aan

$$\hat{\mu}_0 = \frac{n\bar{X}_n + m\bar{Y}_m}{n+m}.$$

Dus

$$l(\hat{\mu}_0, \hat{\mu}_0; X, Y) = -\frac{n+m}{2}\log(2\pi) - (n+m)\log(\sigma) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_i - \hat{\mu}_0)^2\right).$$

Onze toetsingsgrootheid L, twee keer de log-likelihood ratio, is dus gelijk aan

$$\begin{aligned}
L &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i - \hat{\mu}_0)^2 + \sum_{j=1}^{m}(Y_i - \hat{\mu}_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 - \sum_{j=1}^{m}(Y_i - \bar{Y}_m)^2\right) \\
&= \frac{1}{\sigma^2}\left(n(\bar{X}_n - \hat{\mu}_0)^2 + m(\bar{Y}_m - \hat{\mu}_0)^2\right) \\
&= \frac{1}{\sigma^2}\frac{nm}{n+m}(\bar{X}_n - \bar{Y}_m)^2 \\
&= 9.6.
\end{aligned}$$

De p-waarde is gelijk aan

$$\mathbb{P}(\chi^2(1) > 9.6) = 2\mathbb{P}(N(0,1) > 3.10) = 0.0010.$$

Dit is kleiner dan 5%, dus we verwerpen de nul-hypothese: er is significant bewijs dat het ijkgewicht een andere massa heeft gekregen.

2c De functie toetst de hypothese $H_0 : (\mu_1, \mu_2) = (100, 100)$ tegen $H_1 : (\mu_1, \mu_2) \neq (100, 100)$ door middel van de likelihoodratio toets. De functie berekent twee keer de log-likelihoodratio L, en bepaalt dan de p-waarde door de kans te geven dat een $\chi^2(2)$ verdeling groter is dan L. Dit is de juiste p-waarde, aangezien H_0 twee parameters minder heeft dan het volledige model. Aangezien de p-waarde kleiner is dan 5%, verwerpen we H_0.

3a Definieer $\mu = \mathbb{E}(X_1)$ en $n = 220$. We zien dat $\bar{X}_n = \frac{1}{n}\sum_{k=2}^{4} kN_k = 3.09$. Dan geldt dus

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\sum_{k=1}^{3} N_k(k - \bar{X}_n)^2 = 0.2658.$$

Nu beschouwen we $X_1, \ldots, X_n$ als een iid steekproef die niet normaal verdeeld is, en toetsen we de nul-hypothese $H_0 : \mu \leq 3$ tegen $H_1 : \mu > 3$ door middel van de toetsingsgrootheid

$$T = \sqrt{n} \cdot \frac{\bar{X}_n - \mu}{S_X} = 2.62.$$

Onder (het slechtste geval van) de nul-hypothese, heeft T in benadering een $N(0,1)$ verdeling. De p-waarde is dus $\mathbb{P}(N(0,1) > 2.62) = 0.0044$, dus we verwerpen H_0. We hebben dus significant bewijs dat het gemiddeld aantal blaadjes groter is dan 3.

3b Het is logisch om naar $N_4 - N_2$ te kijken, maar we willen deze toetsingsgrootheid standaardizeren. Dit doen we door te delen door een geschatte standaarddeviatie van $N_4 - N_2$, onder de aanname dat $p_2 = p_4 = p$:

$$\mathrm{Var}(N_4 - N_2) = \mathrm{Var}(N_2) + \mathrm{Var}(N_4) - 2\mathrm{Cov}(N_2, N_4) = 2np.$$

Uiteraard is p onbekend, dus die moeten we schatten door $\hat{p} = \frac{1}{2}(N_2 + N_4)/n$. We krijgen dus als toetsingsgrootheid

$$T = \frac{N_4 - N_2}{\sqrt{N_2 + N_4}} = 2.58.$$

Dit heeft onder H_0 in goede benadering een $N(0,1)$-verdeling, en we verwerpen H_0 voor zowel sterk positieve als sterk negatieve waarden van T, dus we vinden een p-waarde gelijk aan

$$2\mathbb{P}(N(0,1) > 2.58) = 0.0098.$$

Dit is kleiner dan 5%, dus we verwerpen de nul-hypothese.

4a De log-likelihood om $X = 5$ defecte chips te zien uit $n = 1000$ geteste chips, met kans op falen gelijk aan p_A, wordt gegeven door

$$l_A(p_A, X) = \log \binom{n}{X} + X \log(p_A) + (n - X)\log(1 - p_A).$$

De maximum likelihood schatter is gegeven door $\hat{p}_A = X/n$. Hetzelfde kunnen we concluderen voor lading B, met $Y = 6$ uit $m = 500$ gefaalde chips en kans p_B. Onder H_0 geldt dat $p_A = p_B = p_0$, en uiteraard schatten we p_0 door $\hat{p}_0 = (X + Y)/(m + n)$. We vinden dus voor twee keer de log-likelihood ratio:

$$\begin{aligned} L &= 2 \cdot (l_A(X/n, X) + l_B(Y/m, Y) - l_A(\hat{p}_0, X) - l_B(\hat{p}_0, Y)) \\ &= 2X\log(X/n) + 2(n - X)\log(1 - X/n) + 2Y\log(Y/m) + 2(m - Y)\log(1 - Y/m) \\ &\quad - 2(X + Y)\log((X + Y)/(m + n)) - 2(m + n - X - Y)\log(1 - (X + Y)/(m + n)) \\ &= 2.096. \end{aligned}$$

De p-waarde is dus

$$\mathbb{P}(\chi^2(1) > 2.096) = 2\mathbb{P}(N(0,1) > 1.45) = 0.147.$$

We kunnen de nul-hypothese dus niet verwerpen: er is geen bewijs dat lading A verschillend is van lading B.

4b Kies $p_0 = 0.02$, $p_B = 0.015$, $\alpha = 0.05$ en $\beta = 0.1$. We gebruiken als toetsingsgrootheid Y, het aantal chips dat faalt van de m gekozen chips. We verwerpen als $Y \leq k$, waarbij geldt dat

$$\mathbb{P}_{p_0}(Y \leq k) \approx \alpha.$$

Dit betekent dat

$$\begin{aligned}\mathbb{P}_{p_0}(Y \leq k) &= \mathbb{P}_{p_0}\left(\frac{Y - mp_0}{\sqrt{mp_0(1-p_0)}} \leq \frac{k - mp_0}{\sqrt{mp_0(1-p_0)}}\right) \\ &\approx \alpha.\end{aligned}$$

Dus

$$\frac{k - mp_0}{\sqrt{mp_0(1-p_0)}} = \xi_\alpha = -1.645 \Longleftrightarrow k = mp_0 + \sqrt{mp_0(1-p_0)}\,\xi_{1-\alpha}.$$

Nu moet gelden dat als de kans op falen gelijk is aan $p_B = 0.015$, dan moeten we verwerpen met kans $1 - \beta = 0.9$. Dus

$$\begin{aligned}\mathbb{P}_{p_B}(Y \leq k) &= \mathbb{P}_{p_B}\left(\frac{Y - mp_B}{\sqrt{mp_B(1-p_B)}} \leq \frac{k - mp_B}{\sqrt{mp_B(1-p_B)}}\right) \\ &\approx 1 - \beta.\end{aligned}$$

Dit impliceert dat

$$\frac{k - mp_B}{\sqrt{mp_B(1-p_B)}} = \xi_{1-\beta} = 1.282.$$

Dus

$$m(p_0 - p_B) + \sqrt{mp_0(1-p_0)}\,\xi_\alpha = \sqrt{mp_B(1-p_B)}\,\xi_{1-\beta}.$$

Hieruit volgt dat

$$m = \frac{\left(\sqrt{p_B(1-p_B)}\,\xi_{1-\beta} - \sqrt{p_0(1-p_0)}\,\xi_\alpha\right)^2}{(p_0 - p_B)^2} = 5961.546.$$

Dit betekent dus dat we 5962 chips moeten testen uit lading B!

5a De a posteriori verdeling $\pi(\mu)$ is proportioneel aan de prior en de likelihood, dus

$$\begin{aligned}\pi(\mu) &\propto e^{-\lambda\mu} e^{-n\mu} \prod_{i=1}^{n} \mu^{X_i} \\ &\propto e^{-(\lambda+n)\mu} \mu^{\sum_{i=1}^{n} X_i}.\end{aligned}$$

Dit laat zien dat a posteriori, μ een $\Gamma_{\sum_{i=1}^{n} X_i + 1, \lambda+n}$ verdeling heeft. De Bayes schatter T is de verwachting van deze verdeling, dus

$$T = \frac{\sum_{i=1}^{n} X_i + 1}{\lambda + n}.$$

5b We weten dat

$$\mathrm{MSE}(T; \mu) = \mathrm{Var}_\mu(T) + (\mathbb{E}_\mu(T) - \mu)^2.$$

Dus

$$\begin{aligned}\mathrm{Var}_\mu(T) &= \frac{\mathrm{Var}_\mu(\sum_{i=1}^n X_i)}{(n+\lambda)^2}\\ &= \frac{n\mu}{(n+\lambda)^2}\end{aligned}$$

en

$$\mathbb{E}_\mu(T) = \frac{1+n\mu}{n+\lambda}.$$

Dit impliceert dat

$$\begin{aligned}\mathrm{MSE}(T;\mu) &= \frac{n\mu}{(n+\lambda)^2} + \left(\frac{1+n\mu}{n+\lambda} - \mu\right)^2\\ &= \frac{\lambda^2\mu^2 + (n-2\lambda)\mu + 1}{(n+\lambda)^2}.\end{aligned}$$