Oppsummering

Torstein Fjeldstad
Institutt for matematiske fag, NTNU
08.04.2019

I dag

— Andre statistikk-kurs
— Oversikt over pensum

Naturlege neste kurs i statistikk

1. TM4265 Stokastisk modellering (H2019)
   - sannsynsrekning
   - rekne på prosesser som utviklar seg i tid
2. TMA4255 Anvendt statistikk (V2020)
   - statistisk inferens
   - multippel lineær regresjon, forsøksplanlegging
3. TMA4267 Lineære statistiske modeller (V2020)
   - Som TMA4255, men noko meir matematisk
4. TMA4268 Statistisk læring (V2020)
   - Regresjon, ikkje-linearitet, klassifikasjon
   - Fokus på programmering i R

DEL 1: Sannsynsrekning

Sannsyn

Stokastisk forsøk : eit eksperiment der resultatet er underlagt tilfeldigheter,

Utfallsrom S : mengda av moglege resultat i eit stokastisk forsøk.

Utfall x : eit element i utfallsrommet S.

Ein **sannsynsmål** på eit utfallsrom S er ein reell funksjon definert på hendingane i S slik at:

— $0 \leq P(A) \leq 1$ for alle $A \subset S$.

— $P(S) = 1$

— Dersom $A_1, A_2, \ldots$ er parvis disjunkte (dvs. $A_i \cap A_j = \emptyset$ for alle $i \neq j$), så er

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Rekneregler for sannsyn

Additiv regel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Komplementærsetninga: $P(A) = 1 - P(A^c)$

Betinga sannsyn: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Multiplikasjonssetninga: $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$

Uavhengighet: A, B uavhengige $\Leftrightarrow P(A|B) = P(A)$

$$P(A \cap B) = P(A)P(B)$$

Bayes regel: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

Lova om total sannsyn: $B_1, B_2, \ldots, B_n$ partisjon av S

$$P(A) = \sum_{i=1}^{n} P(B_i \cap A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Stokastiske variabel og sannsynsfordeling

Ein stokastisk variabeler ein reell funksjon, X, definert på eit utfallsrom.
Eigenskapar til X:

- $f(x) \begin{cases} \text{punktsannsyn} & \text{diskret} \\ \text{sannsynstettleik} & \text{kontinuerleg} \end{cases}$
- $F(x) = P(X \leq x)$ same tolkning for båe diskret og kontinuerleg tilfelle

Forventningsverdi og varians

— Forventningsverdi (tolkning: gjennomsnitt av uendeleg mange realisasjonar)

$$E(g(X)) = \begin{cases} \sum_x g(x)f(x) & \text{diskret} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{kontinuerleg} \end{cases}$$

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

— Varians (tolkning: måler spredning/variasjon)

$$Var(X) = E((X - E(X)^2) = E(X^2) - (E(X))^2$$

$$Var(aX + b) = a^2 Var(X)$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

— Kovarians (tolkning: kan tolke forteiknet)

$$Cov(X, Y) = E((X - E(X))(Y - E(Y)))$$

X, Y uavhengige $\Rightarrow Cov(X, Y) = 0$

Lineærkombinasjoner

La $Y = \sum_{i=1}^{n} a_i X_i + b$. Da er

$$\mathrm{E}(Y) = \sum_{i=1}^{n} a_i \mathrm{E}(X_i) + b,$$

$$\mathrm{Var}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i a_j \mathrm{Cov}(X_i, X_j).$$

Kombinatorikk

$$n! = 1 \cdot 2 \cdots n \text{ er antall måter å permutere (ordne) } n \text{ elementer.}$$

$$_nP_r = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdots (n-r+1) \text{ er antall ordnede utvalg når } r \text{ elementer}$$

velges blant n uten tilbakelegging.

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{_nP_r}{r!} = {_nC_r} \text{ er antall ikke-ordnede utvalg når } r \text{ elementer velges}$$

blant n uten tilbakelegging.

Viktig diskrete fordelingar

Binomisk fordeling: Trekning frå urne **med** tilbakelegging

— n uavhengige forsøk

— suksess eller fiasko i kvart forsøk med konstant sannsyn for suksess p

— X er talet på suksessar

Multinomisk fordeling: Som binomisk, men k mogleg utfall i kvart forsøk

Hypergeometrisk fordeling: Trekning frå urne **utan** tilbakelegging

— N kuler totalt, k raude og $N - k$ gule. X er talet på raude du trekk

Negativ binomisk fordeling: same situasjon som for binomisk, men trekk inntil k suksessar ($k = 1$ gir geometrisk fordeling). X er talet på trekningar

Poissonfordeling: X er talet på hendingar på intervallet $[0, t]$ i ein poissonprosess

1 Noen diskrete sannsynlighetsfordelinger

Binomisk fordeling

$$f(x) = b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

$$\mathrm{E}(X) = np, \quad \mathrm{Var}(X) = np(1-p), \quad M_X(t) = (pe^t + 1 - p)^n.$$

Poissonfordeling

$$f(x) = p(x; \mu) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \ldots.$$

$$\mathrm{E}(X) = \mu, \quad \mathrm{Var}(X) = \mu, \quad M_X(t) = e^{\mu(e^t - 1)}.$$

Hypergeometrisk fordeling

$$f(x) = h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, \ldots, \min(k, n).$$

$$\mathrm{E}(X) = \frac{nk}{N}, \quad \mathrm{Var}(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N} \cdot \left(1 - \frac{k}{N}\right).$$

$M_X(t)$ har ikke et pent uttrykk.

Geometrisk fordeling

$$f(x) = g(x; p) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots.$$

$$\mathrm{E}(X) = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2}, \quad M_X(t) = \frac{pe^t}{1-(1-p)e^t} \text{ for } t < -\ln(1-p).$$

Negativ-binomisk fordeling

$$f(x) = b^*(x; k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad x = k, k+1, \ldots.$$

$$\mathrm{E}(X) = \frac{k}{p}, \quad \mathrm{Var}(X) = k\frac{1-p}{p^2}, \quad M_X(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^k \text{ for } t < -\ln(1-p).$$

Spesialtilfelle: $k = 1$ gir geometrisk fordeling.

Binomisk fordeling

$$P(X \leq x)$$

		p											
n	x	.05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
2	0	.902	.810	.722	.640	.562	.490	.360	.250	.160	.090	.040	.010
	1	.998	.990	.978	.960	.938	.910	.840	.750	.640	.510	.360	.190
3	0	.857	.729	.614	.512	.422	.343	.216	.125	.064	.027	.008	.001
	1	.993	.972	.939	.896	.844	.784	.648	.500	.352	.216	.104	.028
	2	1.000	.999	.997	.992	.984	.973	.936	.875	.784	.657	.488	.271
4	0	.815	.656	.522	.410	.316	.240	.130	.062	.026	.008	.002	.000
	1	.986	.948	.890	.819	.738	.652	.475	.313	.179	.084	.027	.004
	2	1.000	.996	.988	.973	.949	.916	.821	.688	.525	.348	.181	.052
	3	1.000	1.000	.999	.998	.996	.992	.974	.938	.870	.760	.590	.344
5	0	.774	.590	.444	.328	.237	.168	.078	.031	.010	.002	.000	.000
	1	.977	.919	.835	.737	.633	.528	.337	.187	.087	.031	.007	.000
	2	.999	.991	.973	.942	.896	.837	.683	.500	.317	.163	.058	.009
	3	1.000	1.000	.998	.993	.984	.969	.913	.813	.663	.472	.263	.081
	4	1.000	1.000	1.000	1.000	.999	.998	.990	.969	.922	.832	.672	.410
6	0	.735	.531	.377	.262	.178	.118	.047	.016	.004	.001	.000	.000
	1	.967	.886	.776	.655	.534	.420	.233	.109	.041	.011	.002	.000
	2	.998	.984	.953	.901	.831	.744	.544	.344	.179	.070	.017	.001
	3	1.000	.999	.994	.983	.962	.930	.821	.656	.456	.256	.099	.016
	4	1.000	1.000	1.000	.998	.995	.989	.959	.891	.767	.580	.345	.114
	5	1.000	1.000	1.000	1.000	1.000	.999	.996	.984	.953	.882	.738	.469
7	0	.698	.478	.321	.210	.133	.082	.028	.008	.002	.000	.000	.000
	1	.956	.850	.717	.577	.445	.329	.159	.063	.019	.004	.000	.000
	2	.996	.974	.926	.852	.756	.647	.420	.227	.096	.029	.005	.000
	3	1.000	.997	.988	.967	.929	.874	.710	.500	.290	.126	.033	.003
	4	1.000	1.000	.999	.995	.987	.971	.904	.773	.580	.353	.148	.026
	5	1.000	1.000	1.000	1.000	.999	.996	.981	.938	.841	.671	.423	.150
	6	1.000	1.000	1.000	1.000	1.000	1.000	.998	.992	.972	.918	.790	.522
8	0	.663	.430	.272	.168	.100	.058	.017	.004	.001	.000	.000	.000
	1	.943	.813	.657	.503	.367	.255	.106	.035	.009	.001	.000	.000
	2	.994	.962	.895	.797	.679	.552	.315	.145	.050	.011	.001	.000
	3	1.000	.995	.979	.944	.886	.806	.594	.363	.174	.058	.010	.000
	4	1.000	1.000	.997	.990	.973	.942	.826	.637	.406	.194	.056	.005
	5	1.000	1.000	1.000	.999	.996	.989	.950	.855	.685	.448	.203	.038
	6	1.000	1.000	1.000	1.000	1.000	.999	.991	.965	.894	.745	.497	.187
	7	1.000	1.000	1.000	1.000	1.000	1.000	.999	.996	.983	.942	.832	.570

Viktige kontinuerlege fordelingar

Normalfordeling: $X \sim n(x; \mu, \sigma)$

— lineærkombinasjon av uavhengige normalfordelte stokastiske variabler er normalfordelt

— $X \sim n(x; \mu, \sigma) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \sim n(z; 0, 1)$

Eksponensialfordeling: X tid til første hending/tid mellom etterfølgande hendingar i ein poissonprosess

2 Noen kontinuerlige sannsynlighetsfordelinger

Normalfordeling (Gaussfordeling)

$$f(x) = n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < \infty.$$

$$\mathrm{E}(X) = \mu\,, \quad \mathrm{Var}(X) = \sigma^2\,, \quad M_X(t) = e^{\mu t + \sigma^2 t^2/2}.$$

$$F(x) = P(X \le x) = \Phi(\frac{x-\mu}{\sigma}).$$

Eksponentialfordeling

$$f(x) = \frac{1}{\beta}e^{-x/\beta}\,, \quad x \ge 0.$$

$$\mathrm{E}(X) = \beta\,, \quad \mathrm{Var}(X) = \beta^2\,, \quad M_X(t) = \frac{1}{1-\beta t} \text{ for } t < \frac{1}{\beta}.$$

Kommentar: Ofte vanlig å bruke parameteren $\lambda = 1/\beta$.

Standard normalfordeling

$$\Phi(z) = P(Z \leq z)$$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.7	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Kritiske verdier i standard normalfordelingen

$$P(Z > z_\alpha) = \alpha$$

α	z_α
.2	0.842
.15	1.036
.1	1.282
.075	1.440
.05	1.645
.04	1.751
.03	1.881
.025	1.960
.02	2.054
.01	2.326
.005	2.576
.001	3.090
.0005	3.291
.0001	3.719
.00005	3.891
.00001	4.265
.000005	4.417
.000001	4.753

Transformasjon av stokastiske variablar

Situasjon: $X \sim f(x)$
Mål: finne fordelinga til $Y = u(X)$ (merk: $X = w(Y)$)

$$g(y) = \begin{cases} f(w(y)) & \text{diskret} \\ f(w(y))|w'(y)| & \text{kontinuerleg} \end{cases}$$

Nokre samanhenger mellom fordelingar

- Hypergeometrisk $\approx$ binomisk dersom N er stor i forhold til k
- Binomisk $\approx$ poisson viss n stor og p liten
- Binomisk $\approx$ normal viss n stor

Ordningsvariablar

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval med $X_i \sim f(x_i)$

Maksimum:

$$V = \max(X_1, X_2, \ldots, X_n)$$

$$\begin{aligned} F_V(v) &= P(\max(X_1, X_2, \ldots, X_n) \leq v) \\ &= \prod_{i=1}^{n} P(X_i \leq v) = (F_X(v))^n \end{aligned}$$

Minimum:

$$U = \min(X_1, X_2, \ldots, X_n)$$

$$\begin{aligned} F_U(u) &= P(\min(X_1, X_2, \ldots, X_n) \leq u) \\ &= 1 - \prod_{i=1}^{n} [1 - P(X_i \leq u)] = 1 - (1 - F_X(u))^n \end{aligned}$$

Momentgenererande funksjon

Definisjon: $M_X(t) = E(e^{tX})$
Rekneregler:

- $M_{aX}(t) = M_X(at)$
- $M_{a+X}(t) = aM_X(t)$
- $X_1, X_2, \ldots, X_n$ uavhengige

$$M_{X_1+X_2+\ldots X_n}(t) = M_{X_1}(t)M_{X_2}(t)\ldots M_{X_n}(t)$$

Eigenskapar:

- $M_X^{(r)}(0) = E(X^r)$
- Fordelinga til X og Y er like viss $M_X(t) = M_Y(t)$ for alle t

DEL 2: Statistikk

Sentralgrenseteoremet

Viss $\bar{X}$ er gjennomsnittet av eit tilfeldig utval av storleik n tatt frå ein populasjon med forventningsverdi μ og varians $\sigma^2 < \infty$ vil

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to n(z; 0, 1) \qquad \text{når } n \to \infty.$$

Estimator

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval med $X_i \sim f(x_i; \theta)$
Mål: ynskjer å anslå verdien til θ (og seie noko om tilhøyrande uvisse)

Observator: funksjon av stokastiske variablar

Estimator: ein observator $\widehat{\theta}$ som nyttes til å estimere verdien til den ukjende parameteren θ

Sannsynsmaksimering

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval $X_i \sim f(x_i;\theta)$

1. Definer rimelighetsfunksjonen

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

2. Ofte er det enklare å sjå på log-rimelighetsfunksjonen

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$$

3. Maksimer $l(\theta)$ (ofte ved å derivere og setje lik null)

Eigenskapar til estimatorar

— Ynskjer ein forventningsrett (eng: unbiased) estimator

$$E(\widehat{\theta}) = \theta$$

— For to forventningsrette estimatorar $\widehat{\theta}_1, \widehat{\theta}_2$ vil me velge den med lågast varians:

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}) \Rightarrow \text{velg } \widehat{\theta}_1$$

χ^2-fordeling (kjikvadratfordeling)

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}\,, \quad x \geq 0\,, \quad \nu = 1, 2, \ldots.$$

$$\mathrm{E}(X) = \nu\,, \quad \mathrm{Var}(X) = 2\nu\,, \quad M_X(t) = \left(\frac{1}{1-2t}\right)^{\nu/2} \text{ for } t < \frac{1}{2}.$$

Kommentar: Dersom $X_1, \ldots, X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2 har vi at

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \quad \text{er } \chi^2\text{-fordelt med } n-1 \text{ frihetsgrader,}$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 \text{ og } \bar{X} \text{ er uavhengige.}$$

t-fordeling (Student t-fordeling)

$$f(x) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)\sqrt{\pi\nu}} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}\,, \quad -\infty < x < \infty.$$

$$\mathrm{E}(X) = 0 \text{ hvis } \nu \geq 2\,, \quad \mathrm{Var}(X) = \frac{\nu}{\nu - 2} \text{ hvis } \nu \geq 3\,, \quad M_X(t) \text{ eksisterer ikke.}$$

Spesialtilfeller: $\nu = 1$ gir Cauchyfordelingen.
$\nu = \infty$ gir Normalfordelingen.

Kommentar: Dersom Z er standard normalfordelt og V er χ^2-fordelt med ν frihetsgrader og Z og V er uavhengige, har vi at

$$\frac{Z}{\sqrt{V/\nu}} \quad \text{er } t\text{-fordelt med } \nu \text{ frihetsgrader.}$$

Spesielt gir dette at dersom $X_1, \ldots, X_n$ er uavhengige og normalfordelte med forventning μ og varians σ^2 har vi at

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \quad \text{er } t\text{-fordelt med } n-1 \text{ frihetsgrader,}$$

der $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Fordeling til lineærkombinasjoner

La $X_1, \dots, X_n$ være uavhengige variabler.

Dersom X_i er normalfordelt med forventning μ_i og varians σ_i^2 vil $Y = \sum_{i=1}^n a_i X_i$ være normalfordelt med forventning $\sum_{i=1}^n a_i \mu_i$ og varians $\sum_{i=1}^n a_i^2 \sigma_i^2$.

Dersom X_i er binomisk fordelt med parametre m_i og p vil $Y = \sum_{i=1}^n X_i$ være binomisk fordelt med parametre $\sum_{i=1}^n m_i$ og p.

Dersom X_i er Poissonfordelt med parameter μ_i vil $Y = \sum_{i=1}^n X_i$ være Poissonfordelt med parameter $\sum_{i=1}^n \mu_i$.

Dersom X_i er χ^2-fordelt med ν_i frihetsgrader vil $Y = \sum_{i=1}^n X_i$ være χ^2-fordelt med $\sum_{i=1}^n \nu_i$ frihetsgrader.

Konfidensintervall

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval $X_i \sim f(x_i; \theta)$
Mål: ynskjer konfidensintervall $[\widehat{\theta}_L(X_1, X_2, \ldots, X_n, \widehat{\theta}_U(X_1, X_2, \ldots, X_n)]$
slik at

$$P(\widehat{\theta}_L(X_1, X_2, \ldots, X_n \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)]) = 1 - \alpha$$

Generell framgangsmåte for konfidensintervall

1. Estimator for θ, $\hat{\theta}$ (t.d. SME)
2. La $Z = h(\hat{\theta}, \theta)$ der $h(\cdot, \cdot)$ er ein funksjon s.a. Z har ei kjend fordeling.
3. Har då
$$P(z_{1-\alpha/2} \leq h(\hat{\theta}, \theta) \leq z_{\alpha/2}) = 1 - \alpha$$
4. Løys ulikskapane (mhp. θ) kvar for seg og finn eit uttrykk med θ i midten
$$P(\hat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \hat{\theta}_U(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$
5. Et $(1 - \alpha) \cdot 100$ % konfidensintervall for θ er
$$\left[\hat{\theta}_L(X_1, X_2, \ldots, X_n), \hat{\theta}_U(X_1, X_2, \ldots, X_n)\right]$$

Hypotesetesting

	H_0 riktig	H_1 riktig
Forkast H_0	Type I-feil	Ok
Ikkje forkast H_0	Ok	Type II-feil

Ide: vi må vere "sikre" før me påstår at H_1 er rett. Me velg signinifikansnivået α liten og krev

$$P(\text{Type I-feil}) = P(\text{Forkast } H_0 \text{ når } H_0 \text{ er riktig}) \leq \alpha$$

$$\beta = P(\text{Type II-feil}) = P(\text{Ikkje forkast } H_0 \text{ når } H_1 \text{ er riktig})$$

Generell framgangsmåte

Situasjon: $X_1, X_2, \ldots, X_n$ tilfeldig utval med $X_i \sim f(x_i; \theta)$.

1. Ynskjer å teste:
   a) $H_0 : \theta = \theta_0$ mot $H_1 : \theta > \theta_0$
   b) $H_0 : \theta = \theta_0$ mot $H_1 : \theta < \theta_0$
   c) $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$
2. Estimator for θ; $\widehat{\theta}$
3. La $Z = h(\widehat{\theta}, \theta_0)$, der $h(\cdot, \cdot)$ er ein funksjon s.a. Z har ei kjend fordeling under H_0
4. Bestem eit forkastningskriterium (antar Z stor når $\widehat{\theta}$ stor)
   a) Forkast H_0 dersom $Z > k$
   b) Forkast H_0 dersom $Z < k$
   c) Forkast H_0 dersom $Z < k_l$ eller $Z > k_u$

   der k bestemmes frå kravet

$$P(\text{Forkast H}_0 \text{ når H}_0 \text{ er riktig}) \leq \alpha$$

5. Sett inn tal og konkluder

p-verdi

Ein *p*-verdi er det lågaste signifikansnivået α slik at observert verdi for observatoren gjev at me skal forkaste H_0. Det vil seie, forkast H_0 dersom *p*-verdien er *mindre* enn α.

Teststyrke

Styrken til ein test er sannsynet for å forkaste H_0 gitt at ein spesifikk alternativ hypotese er sann.

Eksamen august 2016 oppgåve 2

Oppgave 2

Harald vurderer hotellferie i Barcelona på seinsumaren. Ein kan antake at prisen på eit hotellrom kan beskrivast ved ei normalfordeling med forventning μ Euro per natt og standardavvik 20 Euro. Ein kan og antake ein kronekurs på 9.2 kroner per Euro.

Harald meiner forventa pris er 100 Euro, medan ein ven seier at det har vorte dyrare. Dei er einige om at standardavviket er kjend og lik 20 Euro. Harald samler inn data for å undersøkje om forventa pris har vorte høgare enn 100 Euro. Han ringjer rundt på staden og samlar inn prisdata, $x_1, \ldots, x_{20}$, for 20 hotellrom. Han får at $\bar{x} = 120$ Euro.

b) Formuler undersøkinga som ein hypotesetest.

Nytt antakinga om normalfordelte data og tala ovanfor til å gjennomføre hypotesetesten på signifikansnivå 0.05.

Eksamen august 2016 oppgåve 2

c) Vi antek så at sann forventning er lik 110 Euro.

Rekn ut teststyrken.

Kor mange data måtte Harald ha samla inn for å få ein teststyrke på 0.95?

Eksamen august 2016 oppgåve 2

c) Vi antek så at sann forventning er lik 110 Euro.
Rekn ut teststyrken.
Kor mange data måtte Harald ha samla inn for å få ein teststyrke på 0.95?

Teststyrke:

$$\begin{aligned} P(\text{forkast } H_0 | H_1 \text{ sann}) &= 1 - P(\text{ikkje forkast } H_0 | H_1 \text{ sann}) \\ &= 1 - P(\text{type II feil}) \\ &= 1 - \beta \end{aligned}$$

H_0	Value of Test Statistic	H_1	Critical Region
$\mu = \mu_0$	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$; σ known	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu = \mu_0$	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $v = n - 1$, σ unknown	$\mu < \mu_0$ $\mu > \mu_0$ $\mu \neq \mu_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$; σ_1 and σ_2 known	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$z < -z_\alpha$ $z > z_\alpha$ $z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{1/n_1 + 1/n_2}}$; $v = n_1 + n_2 - 2$, $\sigma_1 = \sigma_2$ but unknown, $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t' = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$; $v = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$, $\sigma_1 \neq \sigma_2$ and unknown	$\mu_1 - \mu_2 < d_0$ $\mu_1 - \mu_2 > d_0$ $\mu_1 - \mu_2 \neq d_0$	$t' < -t_\alpha$ $t' > t_\alpha$ $t' < -t_{\alpha/2}$ or $t' > t_{\alpha/2}$
$\mu_D = d_0$ paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$; $v = n - 1$	$\mu_D < d_0$ $\mu_D > d_0$ $\mu_D \neq d_0$	$t < -t_\alpha$ $t > t_\alpha$ $t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$

Lineær regresjon

Situasjon: observert $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Modell:

$$Y_i | x_i \sim n(y_i; \alpha + \beta x_i, \sigma)$$

Kan tilpasse α, β, σ

— Minste kvadraters metode (kun α, β)

— Sannsynsmaksimering

La $Y_1, \ldots, Y_n$ være uavhengige variabler med samme varians σ^2 og forventningsverdier

$$\mathrm{E}(Y_i) = \alpha + \beta x_i \,, \quad i = 1, \ldots, n.$$

Minste kvadratsumsestimatorene for α og β er da

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \,, \qquad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

og en forventningsrett estimator for σ^2 er gitt ved

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2.$$

Dersom i tillegg $Y_1, \ldots, Y_n$ er normalfordelte vil

$$\frac{(n-2)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

være χ^2-fordelt med $n-2$ frihetsgrader. Det kan også vises at $(n-2)S^2/\sigma^2$ er uavhengig av $\hat{\alpha}$ og $\hat{\beta}$.

Prediksjon lineær regresjon

Merk at det er skilnad på følgande:

- Forventa respons $\mu_{Y|x_0}$
- Ein ny observasjon Y_0 (gitt x_0)