ST0103 Brukerkurs i statistikk Høst 2016

Forelesning 27, 23.11.2016

Repetisjon

Eksamen 19. desember 2016

Oppgave 1

I forbindelse med undersøkelse av en mineralforekomst blir det boret n hull. For hvert hull er det sannsynlighet p for at mineralet blir funnet. Anta at antall hull, X, der mineralet blir funnet, er binomisk fordelt med parametre n og p.

a) Hvilke forutsetninger må være oppfylt for at X skal være binomisk fordelt? Finn $P(3 \leq X \leq 4)$ når $n = 20$ og $p = 0{,}3$.

b) Finn en tilnærmet verdi for $P(50 \leq X \leq 75)$ når $n = 200$ og $p = 0{,}3$.

Anta nå at p er ukjent. En ny serie på m hull blir boret. Antall hull, Y, med forekomst av mineralet i den nye serien antas å være uavhengig av X og binomisk fordelt med suksessannsynlighet p. To estimatorer for p blir foreslått, $\hat{p} = (X + Y)/(n + m)$ og $\tilde{p} = \frac{1}{2}(X/n + Y/m)$.

c) Hvilken av de to estimatorene vil du foretrekke? Det er tilstrekkelig at du undersøker egenskapene til $\hat{p}$ og $\tilde{p}$ for $n = 20$ og $m = 10$.

Binomisk forsøksrekke

Binomisk eksperiment: Et eksperiment som består i gjentatte forsøk med følgende egenskaper:

1. Det er n identiske uavhengige forsøk.
2. Hvert forsøk har to mulige utfall, ofte kalt suksess og fiasko.
3. P(suksess)=p, P(fiasko)=q, $p + q = 1$.
4. Den binomiske tilfeldige variabelen x er antallet suksessfulle utfall som inntreffer, og x kan anta enhver heltallsverdi fra 0 til n.

Oppgave 2

Et bestemt mønster av nukleotider opptrer på en DNA-sekvens som en poissonprosess med intensitet (rate) $\lambda = 0{,}03$, slik at antall ganger mønsteret forekommer på en DNA-sekvens av lengde l (målt i antall tusen basepar) er poissonfordelt med parameter λl.

a) Finn sannsynligheten for at mønsteret forekommer 2 eller flere ganger på en DNA-sekvens av lengde 100. Finn den betingede sannsynligheten for at mønsteret forekommer nøyaktig 2 ganger gitt at det forekommer 2 eller flere ganger.

b) Begrunn kort hvorfor avstanden fra et mønster til neste sted mønsteret forekommer er eksponentielt fordelt med parameter λ (dvs. med forventningsverdi $1/\lambda$). Hva er sannsynligheten for at denne avstanden er større enn 50?

Vi måler avstanden fra første sted mønsteret forekommer til andre sted, avstanden fra andre til tredje sted, og avstanden fra tredje til fjerde sted mønsteret forekommer. Fra egenskapene til en poissonprosess følger det at disse tre avstandene er uavhengige (du skal ikke vise dette).

c) Finn sannsynligheten for at den minste av de tre avstandene er større enn $1/\lambda$.

Poissonfordelingen
X = ...

- Antall fisker Kristine får i løpet av en viss *tid* med kontinuerlig dorgefiske
- Antall kunder som ankommer et legekontor i løpet av en viss *tid*
- Årlig antall påkjørte elger på en veistrekning av en viss *lengde*
- Antall produksjonsfeil i et tekstilstykke med et bestemt *areal*
- Antall hvite blodceller i en blodprøve av et visst *volum*

$$
\begin{aligned}
P(X = x) &= \frac{\mu^x}{x!} e^{-\mu} \text{ for } x = 0, 1, 2, \ldots \\
E(X) &= \mu \\
Var(X) &= \mu
\end{aligned}
$$

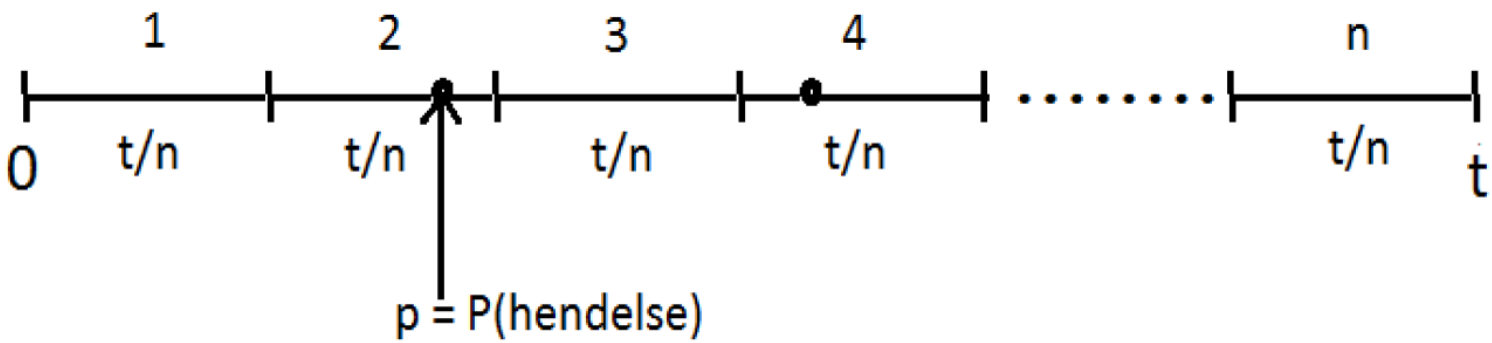
Ofte er $\mu = \lambda t$ når det telles antall hendelser i en tid t.
Da er $\lambda =$ antall hendelser pr. tidsenhet.

Begrunnelse for bruk av Poissonfordelingen

Hvis X er binomisk med stor n og liten p,
der $E(X) = np = \mu$ er "moderat",
er X tilnærmet Poisson-fordelt med parameter μ.

(Bra tilnærming hvis $n > 10, p < 0.1$)

Poissonprosess



- La X være antall hendelser i tiden fra 0 til t.
- Del intervallet i n like store deler, hver med lengde t/n.
- I hvert intervall er $P(\text{hendelse}) = p$
- Da antas X binomisk med n og p og $E(X) = np = \mu$,
  og dermed er X tilnærmet Poisson-fordelt med parameter μ

Videre:
La $\lambda =$ antall hendelser pr. tidsenhet.
Da kan antas $\mu = \lambda t$

5.6 Eksponentialfordelingen

Regel 5.10 (Eksponentialfordeling) Ventetiden T i en poissonprosess er eksponentialfordelt med parameter λ. Ventetiden har sannsynlighetstetthet, fordelingsfunksjon, forventning og varians lik

$$f(t) = \lambda e^{-\lambda t}, \qquad F(t) = 1 - e^{-\lambda t} \quad \text{for } t > 0$$

$$\mu = E(T) = \frac{1}{\lambda}, \qquad \mathrm{Var}(T) = \frac{1}{\lambda^2}$$

Oppgave 3

Konsentrasjonen (målt i mM) av et stoff som dannes i en reaksjon er normalfordelt med forventningsverdi 9,0 og standardavvik 2,0. Reaksjonen gjennomføres to ganger. Konsentrasjonene blir X første gang og Y andre gang, og X og Y er uavhengige.

a) Finn $P(X > 10{,}0)$ og $P\left(\frac{1}{2}(X+Y) > 10{,}0\right)$.

En analysemetode er avhengig av at konsentrasjonen etter første reaksjon er større enn 10,0, og av at konsentrasjonen etter andre reaksjon er større enn konsentrasjonen etter første reaksjon.

b) Finn sannsynligheten for dette – det vil si $P(10{,}0 < X < Y)$. (Vink: Det kan være til hjelp å skissere hendelsen i xy-planet.)

Regel 6.2 (Estimering av μ) Utvalgets gjennomsnitt er vår beste gjetning på populasjonens forventningsverdi. Den naturlige estimatoren for μ er derfor

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\overline{X}) = \mu, \qquad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}, \qquad \mathrm{SE}(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Sentralgrenseteoremet

Regel 5.18 (Sentralgrenseteoremet) La $X_1, X_2, \ldots, X_n$ være uavhengige variabler fra samme sannsynlighetsfordeling med forventning μ og standardavvik σ. Da er

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) \qquad \text{tilnærmet Normal}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Regel 5.19 La $X_1, X_2, \ldots, X_n$ være uavhengige variabler fra samme sannsynlighetsfordeling med forventning μ og standardavvik σ. Da er summen

$$X_1 + X_2 + \cdots + X_n \qquad \text{tilnærmet Normal}(n\mu, \sqrt{n}\sigma) \qquad [5.6]$$

Oppgave 4

John James Strutt (Lord Rayleigh) publiserte i 1894 *On an anomaly encountered in determinations of the density of nitrogen gas* (*Proc. R. Soc. Lond.* 1894 **55**, s. 340–344). I artikkelen beskriver han hvordan han i en serie på 8 forsøk fra november 1893 til januar 1894 produserte nitrogengass fra nitrogenholdige forbindelser og veide innholdet av denne gassen i en kolbe med kjent volum. De 8 massene ble målt til 2,30143, 2,29890, 2,29816, 2,30182, 2,29869, 2,29940, 2,29849 og 2,29889 g.

Anta at målingene er uavhengige og kommer fra samme normalfordeling. Det oppgis at gjennomsnittet av de 8 målingene er 2,29947 g, og at utvalgsstandardavviket er 0,00138 g.

a) Finn et 95 %-konfidensintervall for standardavviket i normalfordelingen.

I en annen serie på 7 forsøk fra desember 1893 til februar 1894 veide Strutt nitrogengass utvunnet fra luft (i motsetning til i de 8 nevnte forsøkene, der gassen ble produsert fra nitrogenholdige forbindelser). Massene ble målt til 2,31017, 2,30986, 2,31010, 2,31001, 2,31024, 2,31010 og 2,31028 g.

Anta at også disse målingene er uavhengige (også fra forrige forsøksserie) og kommer fra en normalfordeling. Det oppgis at gjennomsnittet av de 7 målingene er 2,31011 g, og at utvalgsstandardavviket er 0,00014 g.

b) Test nullhypotesen som sier at forventningsverdien for målt masse ved de to metodene er like, mot alternativet som sier at forventningsverdiene er ulike. Bruk signifikansnivå 0,001. Hvilke forutsetninger må gjelde for at testmetoden skal være riktig? Diskuter kort om forutsetningene ser ut til å holde.

6.3.4 Konfidensintervall for σ

Regel 6.10 (Konfidensintervall for σ^2) Det tilfeldige intervallet

$$\left[\frac{(n-1)S^2}{\chi_{\alpha/2}}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}}\right]$$

er et $100(1-\alpha)$ % konfidensintervall for variansen σ^2. Når vi finner et bestemt punktestimat s^2, kan vi beregne tallverdier for intervallgrensene. Et konfidensintervall for standardavviket finner vi ved å ta kvadratroten av begge intervallgrensene ovenfor. Det er en forutsetning at målingene er tilnærmet normalfordelte.

Regel 5.22 (Fordeling til S^2)

$$Y = \frac{(n-1)\cdot S^2}{\sigma^2} \qquad \text{er kjikvadratfordelt med } n-1 \text{ frihetsgrader}$$

Kritiske verdier i χ^2-fordelingen

$$P(\mathcal{X}^2 > \chi^2_{\alpha,\nu}) = \alpha$$

$\nu \backslash \alpha$	.995	.990	.975	.950	.050	.025	.010	.005
1	.000	.000	.001	.004	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	11.070	12.833	15.086	16.750
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319

Definisjon 8.1. Interpolert varians

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(Vi ser at dette er et vektet gjennomsnitt mellom S_1^2 og S_2^2.)

For å teste om det er forskjell i forventningene kan vi teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } \mu_1 \neq \mu_2$$

Da er testobservatoren for den såkalte **to-utvalgs T-test**:

$$T = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Student-t-fordelt med $df = n_1 - n_2 - 2$ frihetsgrader *hvis H_0 gjelder.*

Vi forkaster H_0 hvis $T < -t_{\alpha/2, n_1+n_2-2}$ eller $T > t_{\alpha/2, n_1+n_2-2}$.

Interpolert utvalgsvarians er definert ved $S_{\mathrm{p}}^2 = ((n-1)S_x^2 + (m-1)S_y^2)/(n+m-2)$, der S_x^2 og S_y^2 er utvalgsvariansene. Med våre tall får S_{p} verdien

$$s_{\mathrm{p}} = \sqrt{\frac{7 \cdot 0{,}00138^2 + 6 \cdot 0{,}00014^2}{8+7-2}} = 0{,}00102,$$

og T får verdi

$$t = \frac{2{,}29947 - 2{,}31011}{0{,}00102\sqrt{1/8 + 1/7}} = -20{,}2.$$

Vi forkaster altså nullhypotesen, og påstår at forventningsverdien for målt masse ved de to metodene er ulike.

Testmetoden forutsetter at de to utvalgene, som skal være uavhengige, består av uavhengige observasjoner fra hver sin normalfordeling, som begge skal ha samme varians. Om den siste forutsetningen er oppfylt her, er tvilsomt, siden utvalgsstandardavvikene er så forskjellige (det ene er ti ganger større enn det andre).

(Lord Rayleighs forsøk bidrog til oppdagelsen av argon, som utgjør omtrent 0,93 % av lufta. Når nitrogen ble utvunnet fra luft, skjedde det ved at oksygen ble fjernet. Det ble antatt at nesten all gassen som var igjen var nitrogen. Men argon, som er tyngre enn nitrogen, var også igjen, og blandingen fikk større tetthet enn ren nitrogen.)