STK2120 våren 2012

- Generelle inferensmetoder
- Spesielle modeller/metoder
- Bruk av R
  - Vil ikke bli testet på kommandoer, men må forstå generelle utskrifter

Generelle inferensmetoder

- Estimering
  - Maksimum likelihood
- Konfidensintervaller
  - Normaltilnærming
  - Bootstrapping
- Hypotesetesting
  - Likelihood ratio test
  - Normaltilnærming
  - Bootstrapping

Spesielle modeller og metoder

- Variansanalyse
- Regresjon
  - Lineær
  - Ikke-lineær
  - Logistisk
- Kategoriske data/føyningstest

Maksimum likelihood/Sannsynlighetsmaksimering

- $y_1, \ldots, y_n \overset{uif}{\sim} f(y;\theta)$
- $L(\theta;\mathbf{y}) = f(y_1, \ldots, y_n;\theta) = \prod_i f(y_i;\theta)$
- $\log L(\theta;\mathbf{y}) = \sum_i \log f(y_i;\theta)$
- $\hat{\theta} = \text{argmax}_\theta L(\theta;\mathbf{y})$
  - Konsistent, asymptotisk effisient Cramer-Rao!
  - Analyttiske løsninger for lineære/Gaussiske modeller
    Ett-/to- utvalgs modeller, variansanalyse, lineær regresjon
  - Generelt: Numerisk optimering

Egenskaper til MLE

- For stor n:

$$\hat{\theta} \approx N(\theta, I(\hat{\theta})^{-1}) \approx N(\theta, J(\hat{\theta}; \mathbf{y})^{-1})$$

Observert informasjon

$$J(\theta; \mathbf{y}) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log L(\theta; \mathbf{y})$$

$$I(\theta) = E[J(\theta; \mathbf{y})] \quad \text{Alltid pos. (semi)definit}$$

Fisher-informasjonen (matrise!)

- Normaltilnærming: $\hat{\theta}_j \pm z_{\alpha/2}\mathrm{SE}(\hat{\theta}_j)$.
  - $\mathrm{SE}(\hat{\theta}_j)$: Normaltilnærming ($\sqrt{I(\theta)^{-1}_{jj}}$) eller bootstrapping

7 Maksimum likelihood metoden

FORMELSAMLINGEN!
NB!

Anta at $X_1, X_2, \ldots, X_n$ har simultan punktsannsynlighet/sannsynlighetstetthet $f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ er en parametervektor (skalar hvis $p = 1$). Vi antar at $f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$ tilfredsstiller visse deriverbarhetsbetingelser.

(*a*) Gitt observerte verdier $X_i = x_i$; $i = 1, \ldots, n$; er likelihood-funksjonen $\text{lik}(\boldsymbol{\theta}) = f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$ og loglikelihood-funksjonen $l(\boldsymbol{\theta}) = \log(\text{lik}(\boldsymbol{\theta}))$.

(*b*) Maksimum likelihood *estimatet* er den verdien av $\boldsymbol{\theta}$ som maksimerer $\text{lik}(\boldsymbol{\theta})$ eller ekvivalent maksimerer $l(\boldsymbol{\theta})$. Hvis vi erstatter de observerte x_i-ene med de stokastiske X_i-ene, får vi maksimum likelihood *estimatoren*.

(*c*) Maksimum likelihood estimatet $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ er en løsning av ligningene $s_j(\boldsymbol{\theta}) = 0$; $j = 1, \ldots, p$; der $s_j(\boldsymbol{\theta}) = (\partial/\partial\theta_j) l(\boldsymbol{\theta})$ er score-funksjonene.
Vektoren av scorefunksjoner er $\boldsymbol{s}(\boldsymbol{\theta}) = (s_1(\boldsymbol{\theta}), \ldots, s_p(\boldsymbol{\theta}))^T$.

(*d*) Den observerte informasjonsmatrisen $\bar{\boldsymbol{J}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{J}_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta})$.
Den forventede informasjonsmatrisen (eller Fishers informasjonsmatrise) $\bar{\boldsymbol{I}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{I}_{ij}(\boldsymbol{\theta}) = \text{E}[\bar{J}_{ij}(\boldsymbol{\theta})]$.
For uavhengige og identisk fordelte observasjoner har vi at $\bar{\boldsymbol{I}}(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$ der $\boldsymbol{I}(\boldsymbol{\theta})$ er forventet informasjon til en observasjon.

(e) Når ligningene i punkt (c) ikke har en eksplisitt løsning, kan vi finne maksimum likelihood estimatet ved å bruke Newton-Raphsons metode:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\boldsymbol{J}}^{-1}(\boldsymbol{\theta}^{(s)})\boldsymbol{s}(\boldsymbol{\theta}^{(s)})$$

, ved å bruke Fishers scoringsalgoritme:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta}^{(s)})\boldsymbol{s}(\boldsymbol{\theta}^{(s)}),$$

eller ved passende modifikasjoner av disse.

(f) Når vi har "tilstrekkelig mye" data, er $\hat{\theta}_i$ tilnærmet normalfordelt med forventning θ_i og med varians lik det i-te diagonalelementet til $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$. Kovariansen mellom $\hat{\theta}_i$ og $\hat{\theta}_j$ er tilnærmet lik element (i, j) i $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$. Vi kan estimere varianser/kovarianser ved å sette inn $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ i $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$ eller i $\bar{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})$.

Numerisk optimering

- Sentrale begreper
  - Likelihood $L(\boldsymbol{\theta};\mathbf{y}) = f(\mathbf{y};\boldsymbol{\theta})$
  - Skår funksjonen $s(\boldsymbol{\theta};\mathbf{y}) = \frac{\partial}{\partial\boldsymbol{\theta}} \log L(\boldsymbol{\theta};\mathbf{y})$
  - Observert informasjon $J(\theta;\mathbf{y}) = -\frac{\partial}{\partial\theta\theta^T} \log L(\boldsymbol{\theta};\mathbf{y})$
  - Forventet (Fisher) informasjon $I(\boldsymbol{\theta}) = E[J(\boldsymbol{\theta};\mathbf{y})]$
- Newton-Raphson

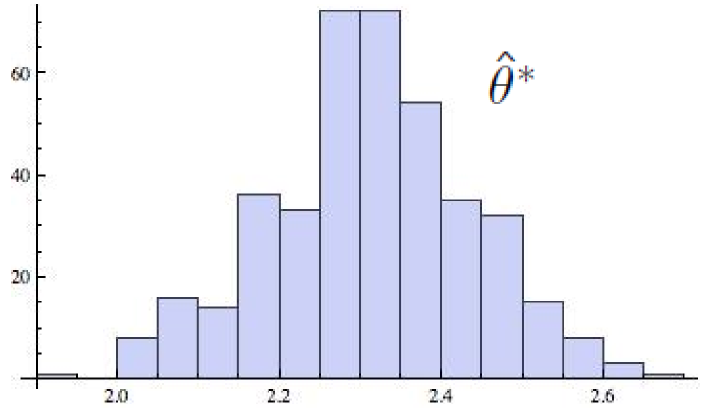
$$\boldsymbol{\theta}^{s+1} = \boldsymbol{\theta}^s + J(\boldsymbol{\theta}^s;\mathbf{x})^{-1}s(\boldsymbol{\theta};\mathbf{x})$$

Eller nlm i R...

Bootstrapping

- Av interesse: Egenskaper til $\hat{\theta} = \hat{\theta}(\mathbf{y})$ *ved gjentatt bruk av denne*
- Bootstrap idé: Simuler $\hat{\theta}^* = \hat{\theta}(\mathbf{y}^*)$ der $\mathbf{y}^* = (y_1^*, ..., y_n^*)$ er bootstrap simuleringer av $\mathbf{y}$.
- Ikke-parametrisk bootstrapping: Trekk $y_1^*, ..., y_n^*$ *med tilbakelegging* fra $\{y_1, ..., y_n\}$.
- Parametrisk bootstrapping: Anta $y_i \sim f(y; \theta)$. Simuler $y_i^* \sim f(y; \hat{\theta})$
- Forventningsskjevhet, usikkerhet, konfidensintervaller
- Egenskaper: STK4170

'Bootstrapping og resampling'



FORMELSAMLINGEN!
NB!

6 Bootstrapping

Anta fordelingen til data $\boldsymbol{X}$ er beskrevet ved en fordelingsfunksjon F. La $\theta = \theta(F)$ være en egenskap ved F som estimeres ved $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$.

(a) Bootstrapping-idéen er å tilnærme egenskapene til $\hat{\theta}$ ved å anta at et estimat $\widehat{F}$ på F er den sanne fordelingsfunksjonen.

(b) Bootstrap estimering av skjevhet til $\hat{\theta}$:

$$b_{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\theta_b^* - \theta(\widehat{F})$$

(c) Bootstrap estimering av standardavvik til $\hat{\theta}$:

$$\sqrt{\mathsf{E}^{\hat{F}}[(\hat{\theta}(\boldsymbol{X}^*) - \mathsf{E}^{\hat{F}}[\hat{\theta}(\boldsymbol{X}^*)])^2]}$$

(d) Standard bootstrap konfidensintervall:

$$(\hat{\theta} - \overline{\delta}, \hat{\theta} - \underline{\delta})$$

der $\underline{\delta}$ og $\overline{\delta}$ er nedre og øvre $\alpha/2$ kvantil i bootstrap fordelingen til $\Delta = \hat{\theta} - \theta$.

Hypotesetesting

Antar data $y_1, ..., y_n \overset{uif}{\sim} f(y;\theta)$
Ønsker å teste $H_0 : \theta \in \Omega_0$ mot $H_a : \theta \in \Omega_a$
Prosedyre

- Spesifiser en test-observator
- Bestem et forkastningsområde for gitt signifikansnivå
- Beregn test-observator og forkastningsområde numerisk og konkluder
  - Hvis test-observator i forkastningsområde, forkast H_0 på det gitte signifikansnivå
  - Ellers, konkluder med at det ikke er grunnlag i data for å forkaste H_0 på det gitte signifikansnivå.
    Merk: Dette er *ikke* det samme som å påstå at H_0 er riktig!
- Ofte vanlig å rapportere P-verdi som angir hvor mye bevis det ligger i data.
- Merk: Bør skille mellom *statistisk signifikant* og *praktisk signifikant*
  Ved mye data kan en ende opp med å forkaste $H_0 : \theta = \theta_0$ selv om $\hat{\theta}$ er svært lik θ_0.

Likelihood ratio test

Antar data $y_1, \ldots, y_n \overset{uif}{\sim} f(y;\theta)$
Ønsker å teste $H_0 : \theta \in \Omega_0$ mot $H_a : \theta \in \Omega_a$

- Neyman-Pearson: $H_0 : \theta = \theta_0$ mot $H_a ; \theta = \theta_a$

Av alle tester
har LR størst styrke
(minst type-II-feil)

$$LR = \frac{L(\theta_0; \mathbf{y})}{L(\theta_a; \mathbf{y})}$$

Enkle hypoteser

optimal testobservator

Forkast når LR liten!

- Generell likelihood ratio

$$LR = \frac{\max_{\theta \in \Omega_0} L(\theta; \mathbf{y})}{\max_{\theta \in \Omega} L(\theta; \mathbf{y})}, \quad \Omega = \Omega_0 \cup \Omega_a$$

- $-2 \log LR \overset{H_0}{\approx} \chi^2_{\text{df}}$, $\text{df} = |\Omega| - |\Omega_0|$
- P-verdi: $\Pr(\chi^2_{\text{df}} > -2 \log LR)$
- Ofte: LR må beregnes numerisk.

Gjelder når n er stor

Variansanalyse

- Enveis variansanalyse

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \sum_i \alpha_i = 0$$

- $H_0 : \alpha_i = 0, \quad F = \frac{SSTr/(I-1)}{SSE/I(J-1)}$

Forkast når F stor!

Ingen gruppe-effekter
Alle Yij fra samme
fordeling

NB! F-fordeling for forholdet mellom
to kji-kvadrat

- Tukey's metode

Simultane konfidensintervall for kontrastene,
for å finne hvilke grupper som skiller seg ut

Studentifisert rangfordeling

1 Enveis variansanalyse

FORMELSAMLINGEN!
NB!

Anta at $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$; $j = 1, 2, \ldots, J_i$; $i = 1, 2, \ldots, I$; der ϵ_{ij}-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Da har vi at:

(a) Den totale kvadratsummen $SST = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{..})^2$ kan skrives som $SST = SSE + SSTr$ der

$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2$ er kvadratsummen for feil eller kvadratsummen innen ("within") grupper

$SSTr = \sum_{i=1}^{I} J_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$ er kvadratsummen for behandling eller kvadratsummen mellom ("between") grupper

(b) SSE og $SSTr$ er uavhengige

(c) $MSE = SSE / [\sum_{i=1}^{I} (J_i - 1)]$ er en forventningsrett estimator for σ^2.
SSE/σ^2 er kji-kvadratfordelt med $\sum_{i=1}^{I} (J_i - 1)$ frihetsgrader.

(d) Hvis alle α_i-ene er lik null, er $SSTr/\sigma^2$ kji-kvadratfordelt med $I - 1$ frihetsgrader

(e) Hvis $J_i = J$ for $i = 1, \ldots, I$, så er
$\max_{i_1, i_2} |(\bar{Y}_{i_1 \cdot} - \mu_{i_1}) - (\bar{Y}_{i_2 \cdot} - \mu_{i_2})| / \sqrt{MSE/J}$
fordelt som den studentifiserte variasjonsbredde med parametere I og $I(J-1)$.

- Toveis variansanalyse

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}$$

$$H_0 : \alpha_i = 0 \quad F = \frac{SSA/(I-1)}{SSE/IJ(K-1)}$$

$$H_0 : \beta_j = 0 \quad F = \frac{SSB/(J-1)}{SSE/IJ(K-1)}$$

$$H_0 : \delta_{ij} = 0 \quad F = \frac{SSAB/(I-1)(J-1)}{SSE/IJ(K-1)}$$

2 Toveis variansanalyse

FORMELSAMLINGEN!
NB!

Anta at $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}\,;\ k = 1, \ldots, K\,;\ j = 1, \ldots, J\,;\ i = 1, \ldots, I\,;$ der ϵ_{ijk}-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Da har vi at:

(a) Den totale kvadratsummen $SST = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \bar{Y}_{\dots})^2$ kan skrives som $SST = SSA + SSB + SSAB + SSE$ der

$$SSA = JK \sum_{i=1}^{I} (\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\dots})^2$$

$$SSB = IK \sum_{j=1}^{J} (\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\dots})^2$$

$$SSAB = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\dots})^2$$

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \bar{Y}_{ij\cdot})^2$$

(b) SSA, SSB, $SSAB$ og SSE er uavhengige

(c) $MSE = SSE/IJ(K-1)$ er en forventningsrett estimator for σ^2. SSE/σ^2 er kji-kvadratfordelt med $IJ(K-1)$ frihetsgrader.

(d) Hvis alle α_i-ene er lik null, er SSA/σ^2 kji-kvadratfordelt med $I-1$ frihetsgrader

(e) Hvis alle β_j-ene er lik null, er SSB/σ^2 kji-kvadratfordelt med $J-1$ frihetsgrader

(f) Hvis alle γ_{ij}-ene er lik null, er $SSAB/\sigma^2$ kji-kvadratfordelt med $(I-1)(J-1)$ frihetsgrader

3 Blokkforsøk (toveis variansanalyse uten gjentak)

Anta at $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\,;\ j = 1,\ldots,J\,;\ i = 1,\ldots,I\,;$ der ϵ_{ij}-ene er uavhengige og $N(0,\sigma^2)$-fordelte. Da har vi at:

(a) Den totale kvadratsummen $SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{..})^2$ kan skrives som $SST = SSA + SSB + SSE$ der

$$SSA = J\sum_{i=1}^{I}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SSB = I\sum_{j=1}^{J}(\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2$$

(b) SSA, SSB og SSE er uavhengige

(c) $MSE = SSE/[(I-1)(J-1)]$ er en forventningsrett estimator for σ^2. SSE/σ^2 er kji-kvadratfordelt med $(I-1)(J-1)$ frihetsgrader.

(d) Hvis alle α_i-ene er lik null, er SSA/σ^2 kji-kvadratfordelt med $I-1$ frihetsgrader

(e) Hvis alle β_j-ene er lik null, er SSB/σ^2 kji-kvadratfordelt med $J-1$ frihetsgrader

Lineær regresjon

- Modell $Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$
- Antagelser
  - $E[\varepsilon_i] = 0$
  - $Var[\varepsilon_i] = \sigma^2$
  - Uavhengighet
  - Normalfordelte
- Estimering
  - $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$
  - Forventningsrett
  - $\mathrm{COV}(\widehat{\beta}) = \sigma^2[\mathbf{X}^T\mathbf{X}]^{-1}$
  - $\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_i (y_i - \hat{y}_i)^2$
  - $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}, \mathbf{H} = \mathbf{X}[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T$

Egenskaper til $\widehat{\beta}$

Hatt-matrisen
Prediksjon, leverage
residualer

- Konfidensintervaller
  $\widehat{\beta}_j \pm t_{\alpha/2;n-k-1} s_{\widehat{\beta}_j}$

F-test for model utility
R^2 multippel korrelasjonskoeff.

5 Multippel lineær regresjon

FORMELSAMLINGEN! NB!

Anta at $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{i,k} + \epsilon_i\,;\ i = 1, 2, \ldots, n\,;$ der x_{ij}-ene er kjente tall og ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte.
På matriseform kan vi skrive modellen som $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, der $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ og $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^T$ er henholdsvis n- og $k+1$-dimensjonale vektorer, og $\mathbf{X} = \{x_{ij}\}$ er en $n \times (k+1)$-dimensjonal matrise. Vi har at:

(a) Minste kvadraters estimator for $\boldsymbol{\beta}$ er $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

(b) La $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_k)^T$. Da er $\hat{\beta}_j$-ene normalfordelte og forventningsrette, og

$$\mathrm{Var}(\hat{\beta}_i) = \sigma^2 c_{ii} \quad \text{og} \quad \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$$

der c_{ij} er element (i, j) i $(k+1) \times (k+1)$ matrisen $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

(c) La $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{i,k}$, og sett $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Da er $S^2 = \mathrm{SSE}/(n-(k+1))$ en forventningsrett estimator for σ^2, og $(n-(k+1))S^2/\sigma^2 \sim \chi^2_{n-(k+1)}$. Videre er S^2 og $\hat{\boldsymbol{\beta}}$ uavhengige.

(d) La $S^2_{\hat{\beta}_i}$ være den variansestimatoren for $\hat{\beta}_i$ vi får ved å erstatte σ^2 med S^2 i formlen for $\mathrm{Var}(\hat{\beta}_i)$ i punkt b). Da er $(\hat{\beta}_i - \beta_i)/S_{\hat{\beta}_i} \sim t_{n-(k+1)}$.

Logistisk regresjon

- Respons $Y_i \in \{0, 1\}$.
- $Y_i \sim \text{Binom}(1, p(x_i))$

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- Numerisk optimering for å finne ML-estimater
- Egenskaper/konfidensintervall ved normaltilnærming (eller bootstrapping)
- Eksempel på *Generaliserte lineære modeller*, tema i STK3100.

Analyse av kategoriske data

- Gruppering av data i kategorier, data er antall innen hver kategori
- Sentral fordeling: Multinomisk fordeling
- Enveis gruppering
- Toveis gruppering
  - Test av homogenitet
  - Test av uavhengighet

En-veis gruppering

- En populasjon, utvalg på n, N_i antall i kategori i
- Antar $(N_1, \ldots, N_k) \sim \text{Multinom}(n, p_1, \ldots, p_k)$
- $H_0 : p_i = p_{i0}, i = 1, \ldots, k$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} \overset{H_0}{\approx} \chi^2_{k-1}$$

- $H_0 : p_i = \pi_i(\theta), i = 1, \ldots, k$

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i(\hat{\theta}))^2}{n\pi_i(\hat{\theta})} \overset{H_0}{\approx} \chi^2_{k-1-m}$$

- $\hat{\theta}$ er ML-estimat.
- Kan brukes til testing av fordelingsantagelser

To-veis gruppering

- Testing av homogenitet
  - I populasjoner, utvalg n_i fra populasjon i, n_{ij} fra kateg. j
  - $(N_{i1}, ..., N_{iJ}) \sim \text{Multinom}(n_i; p_{i1}, ..., p_{iJ}), i = 1, ..., I.$
  - $H_0 : p_{ij} = p_j$
  - Pearson's χ^2 test, $df = (I - 1) * (J - 1)$
- Testing av uavhengighet
  - 1 populasjon, utvalg n, n_{ij} fra kateg. (i, j)
  - $(N_{11}, ..., N_{ij}, ..., N_{IJ}) \sim \text{Multinom}(n; p_{11}, ..., p_{ij}, ..., p_{IJ}).$
  - $H_0 : p_{ij} = p_{i.} * p_{.j}$
  - Pearson's χ^2 test, $df = (I - 1) * (J - 1)$

FORMELSAMLINGEN!
NB!

4 Tabelldata og kji-kvadrattester

(a) Anta at $(N_1, \ldots, N_k)$ er multinomisk fordelt med sannsynligheter p_i, der $\sum_{i=1}^{k} N_i = n$ og $\sum_{i=1}^{k} p_i = 1$.
Hvis $p_i = \pi_i(\boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, og $\hat{\boldsymbol{\theta}}$ er maksimum likelihood estimatoren for $\boldsymbol{\theta}$, så er

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}$$

tilnærmet kji-kvadratfordelt med $k - 1 - m$ frihetsgrader når $E_i = n\pi_i(\hat{\boldsymbol{\theta}}) \geq 5$ for (nesten) alle i

(b) Homogenitetstesting: Anta at for $I = 1, \ldots, I$ er $(N_{i1}, \ldots, N_{iJ})$ uavhengige og multinomisk fordelt med sannsynligheter p_{ij}, der $\sum_{j=1}^{J} p_{ij} = 1$.
Hvis $p_{1j} = \cdots = p_{Ij}$, så er

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

tilnærmet kji-kvadratfordelt med $(I-1)(J-1)$ frihetsgrader når $E_{ij} = (N_{i\cdot}N_{\cdot j})/N_{\cdot\cdot} \geq 5$ for (nesten) alle i, j

(c) Uavhengighetstesting: Anta at $(N_{11}, \ldots, N_{1J}, N_{21}, \ldots, N_{2J}, \ldots, N_{I1}, \ldots, N_{IJ})$ er multinomisk fordelt med sannsynligheter p_{ij}, der $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$.
Hvis $p_{ij} = p_{i\cdot}p_{\cdot j}$ for alle i, j, så er

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

tilnærmet kji-kvadratfordelt med $(I-1)(J-1)$ frihetsgrader når $E_{ij} = (N_{i\cdot}N_{\cdot j})/N_{\cdot\cdot} \geq 5$ for (nesten) alle i, j

Veien videre

- STK2120 dekker de *generelle* prinsipper.
- Kan takle mange ulike situasjoner (også mange vi ikke har diskutert!)

Men...

STK3100 - Innføring i generaliserte lineære modeller

STK4010 - Asymptotisk teori

STK4020 - Bayesiansk statistikk

STK4030 - Moderne dataanalyse

STK4040 - Multivariabel analyse

STK4050 - Statistiske simuleringer og numeriske beregninger

STK4060 - Tidsrekker

STK4080 - Forløpsanalyse

STK4130 - Estimeringsteori

STK4140 - Forsøksplanlegging

STK4150 - Miljøstatistikk - romlig statistikk

STK4160 - Statistisk modellvalg

STK4170 - Bootstrapping og resampling

+ + +