Kompendium over testteorien

i

Teoretisk statistik for økonomer

Version 2

Udarbejdet af Simon Reusch

Maj 2000

Indhold

Binomialfordelingen

Test i binomialfordeling (p.317-318):
H_0: $\theta = \theta_0$

Signifikanssandsynlighed (*Beregnes eksakt ved opslag i Binomial-tabel hvis muligt*):

H_1: $\theta > \theta_0 \rightarrow p_1 = P(X \geq x) \approx 1 - \phi\left(\frac{x - ½ - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right)$

H_1: $\theta < \theta_0 \rightarrow p_2 = P(X \leq x) \approx \phi\left(\frac{x + ½ - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}\right)$

H_1: $\theta \neq \theta_0 \rightarrow p = 2\min\{p_1, p_2\}$

Sammenligning af 2 binomialfordelinger (p.566-568):
H_0: $\theta_A = \theta_B$

Signifikanssandsynlighed:
Baseres på at fordelingen af den ene binomialfordeling, betinget summen af dem begge, er hypergeometrisk fordelt X - hyp($N=n_A+n_B$, $M=n_A$, $n=x_A+x_B$), hvor X
er den observerede værdi af X_A.

H_1: $\theta_A > \theta_B \rightarrow p_1 = P(X \geq x) \approx 1 - \phi\left(\frac{x - ½ - n\frac{M}{N}}{\sqrt{n\frac{M}{N}(1-\frac{M}{N})(\frac{N-n}{N-1})}}\right)$

H_1: $\theta_A < \theta_B \rightarrow p_2 = P(X \leq x) \approx \phi\left(\frac{x + ½ - n\frac{M}{N}}{\sqrt{n\frac{M}{N}(1-\frac{M}{N})(\frac{N-n}{N-1})}}\right)$

H_1: $\theta_A \neq \theta_B \rightarrow p = 2\min\{p_1, p_2\}$

Sammenligning af k binomialfordelinger (p.584-586):
H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$
H_1: Mindst to θ'er forskellige

Hypotesen afprøves ved en kontingenstabelanalyse, hvor der specifikt udføres et homogenitetstest. Dette test bygger på en sammenligning af k multinomialfordelinger, hvor binomialfordelingen jo netop er et specialtilfælde af multinomialfordelingen.

$X_i \sim Bin(n_i, \theta_i^{bin}) \sim M(n_i, \theta_{i1}^M = \theta_i^{bin}, \theta_{i2}^M = 1 - \theta_i^{bin})$

Se mere under "Sammenligning af k multinomialfordelinger"

Poissonfordelingen

Test i poissonfordeling (p.317-318):
H_0: $\lambda = \lambda_0$

Eksempel:
Bemærk, at validiteten af nedenstående tests er betinget af, at der testes i én periode. Betragt opgaveformuleringen "Erfaringsmæssigt omkommer 35 personer årligt ved trafikulykker. Rådet for større færdselssikkerhed iværksætter en kampagne for at mindske antallet af dræbte i trafikken. I de to efterfølgende år dør gennemsnitligt 27 personer pr. år. Havde kampagnen en signifikant effekt?"
For korrekt at anvende nedenstående testørrelser skal man sammenligne antal hændelser i hele observationsperioden med den totale forventning. Her iagttages således 54 dræbte.
H_0: $\lambda = 70$ (Hændelsesintensiteten over en periode på de betragtede 2 år.)
H_1: $\lambda < 70$

Signifikanssandsynlighed (*Beregnes eksakt ved opslag i poisson-tabel hvis muligt*.):

H_1: $\lambda > \lambda_0 \rightarrow p_1 = P(X \geq x) \approx 1 - \phi\left(\frac{x - ½ - \lambda_0}{\sqrt{\lambda_0}}\right)$

H_1: $\lambda < \lambda_0 \rightarrow p_2 = P(X \leq x) \approx \phi\left(\frac{x + ½ - \lambda_0}{\sqrt{\lambda_0}}\right)$

H_1: $\lambda \neq \lambda_0 \rightarrow p = 2 \min\{p_1, p_2\}$

Sammenligning af 2 poissonfordelinger (p.564-565):
H_0: $\lambda_A = \lambda_B$

Signifikanssandsynlighed:
Baseres på, at fordelingen af den ene poissonfordeling, betinget summen af dem begge, er binomialfordelt X - bin(n=x_A+x_B, p=½), hvor X er den observerede værdi af X_A.

H_1: $\lambda_A > \lambda_B \rightarrow p_1 = P(X \geq x) \approx 1 - \phi\left(\frac{x - ½ - 0{,}5n}{\sqrt{n0{,}5(1-0{,}5)}}\right)$

H_1: $\lambda_A < \lambda_B \rightarrow p_2 = P(X \leq x) \approx \phi\left(\frac{x + ½ - 0{,}5n}{\sqrt{n0{,}5(1-0{,}5)}}\right)$

H_1: $\lambda_A \neq \lambda_B \rightarrow p = 2 \min\{p_1, p_2\}$

Sammenligning af k poissonfordelinger (p.558-563):
H_0: $\lambda_1 = \lambda_2 = \ldots = \lambda_k$
H_1: Mindst to λ'er forskellige

Testet bygger på at fordelingen af k poissonfordelinger betinget med summen af dem, er multinomialfordelt $X_1, X_2, ..., X_k \sim M(n, \theta_1, \theta_2, ..., \theta_k)$, idet $\theta_i = \frac{\lambda_i}{\sum \lambda}$.

Se mere under test i multinomialfordelingen.

Test for multiplikativitet i 2-dimensional poissonstruktur (p.593-601):

Model: $X_{ij} \sim Ps(\lambda_{ij})$, hvor λ_{ij} er hændelsesintensiteten i den *ij*'te celle.

$H_0 : \lambda_{ij} = \gamma\delta_i\varepsilon_j \ \forall\, i \in \{1,2,...,I\}, j \in \{1,2,...,J\}$, hvor γ er en niveaufaktor (=n), δ_i rækkesandsynligheden og ε_j søjlesandsynligheden.

$H_1 : Ej\ H_0$

Teststørrelse:

$Q = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(X_{ij} - n\hat{\delta}_i\hat{\varepsilon}_j\right)^2}{n\hat{\delta}_i\hat{\varepsilon}_j} \sim \chi^2((I-1)(J-1))$, hvor *I* er antal rækker og *J* er antal søjler i poissonstrukturen, og $\hat{\delta}_i = \frac{x_{i.}}{n}$ og $\hat{\varepsilon}_j = \frac{x_{.j}}{n}$ er middelrette estimater for række- og søjlesandsynligheder.

Signifikanssandsynlighed: $p = P(Q > q)$

Hypergeometrisk fordeling

Test i hypergeometrisk fordeling (p.318-319):

H_0: $\theta = \theta_0 = \frac{M_0}{N}$

Signifikanssandsynlighed (*Beregnes eksakt hvis muligt*.):

H_1: $\theta > \theta_0 \rightarrow p_1 = P(X \geq x) \approx 1 - \phi\left(\frac{x - \frac{1}{2} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)(\frac{N-n}{N-1})}}\right)$, idet $\theta_0 = \frac{M_0}{N}$

H_1: $\theta < \theta_0 \rightarrow p_2 = P(X \leq x) \approx \phi\left(\frac{x + \frac{1}{2} - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)(\frac{N-n}{N-1})}}\right)$, idet $\theta_0 = \frac{M_0}{N}$

H_1: $\theta \neq \theta_0 \rightarrow p = 2\min\{p_1, p_2\}$

Sammenligning af 2 hypergeometriske fordelinger:

Bemærk: pensum giver ingen formler til denne test. Man kan dog ud fra pensum (p.278) udlede følgende approksimative test mod et dobbeltsidet alternativ:

H_0: $\theta_A = \theta_B$

H_1: $\theta_A \neq \theta_B$

Signifikanssandsynlighed:

$p \approx 2 * \left[1 - \phi\left(\frac{\left|\hat{\theta}_A - \hat{\theta}_B\right|}{\sqrt{\left(\frac{\hat{\theta}_A(1-\hat{\theta}_A)}{n_A - 1}\right)\left(\frac{N_A - n_A}{N_A - 1}\right) + \left(\frac{\hat{\theta}_B(1-\hat{\theta}_B)}{n_B - 1}\right)\left(\frac{N_B - n_B}{N_B - 1}\right)}}\right)\right]$, hvor $\hat{\theta}_A = \frac{x_A}{n_A}$ og $\hat{\theta}_B = \frac{x_B}{n_B}$

Mod enkeltsidet alternativ kan signifikanssandsynligheden defineres på vanlig vis.

Normalfordelingen

U-test for middelværdi i NF med kendt varians σ^2 (p.304)

H_0: $\mu = \mu_0$

$$U = \frac{\overline{X} - \mu_0}{\sigma_0 / \sqrt{n}}$$

Signifikanssandsynligheder findes ved opslag i U-fordelingen

H_1: $\mu > \mu_0 \rightarrow$ $p = P(U>u) = 1 - P(U<u) = 1 - \Phi(u)$

H_1: $\mu < \mu_0 \rightarrow$ $p = P(U<u) = \Phi(u)$

H_1: $\mu \neq \mu_0 \rightarrow$ $p = 2\,(P(U > |u|)) = 2\,(1 - P(U < |u|)) = 2\,(1-\Phi(|u|))$

T-test for middelværdi i NF med ukendt varians σ^2 (p.306)

Den empiriske varians s^2 anvendes som estimat for variansen.

H_0: $\mu = \mu_0$

$$T = \frac{\overline{X} - \mu_0}{s / \sqrt{n}} \sim T(n-1)$$

Signifikanssandsynligheden angives evt. i interval efter opslag i T-fordelingen.

H_1: $\mu > \mu_0 \rightarrow$ $p = P(T > t) = 1 - P(T < t)$

H_1: $\mu < \mu_0 \rightarrow$ $p = P(T < t)$

H_1: $\mu \neq \mu_0 \rightarrow$ $p = 2\,(P\,(T > |t|)) = 2\,(1 - P(T < |t|))$

Q-test for variansen i NF (p.309-311)

H_0: $\sigma^2 = \sigma_0^2$

$$Q = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1)$$

Signifikanssandsynligheden angives evt. i interval efter opslag i χ^2-fordelingen.

H_1: $\sigma^2 > \sigma_0^2 \rightarrow p_1 = P(Q > q) = 1 - P(Q < q)$

H_1: $\sigma^2 < \sigma_0^2 \rightarrow p_2 = P(Q < q)$

H_1: $\sigma^2 \neq \sigma_0^2 \rightarrow p = 2 * \min\{p_1, p_2\}$

Sammenligning af 2 normalfordelinger X og Y

Det kan testes separat hvorvidt varianserne σ_X^2 og σ_Y^2 er ens. Uanset resultatet kan man teste hvorvidt niveauet (middelværdien μ_X og μ_Y) i de to fordelinger er ens. (Her skelner pensum mellem 3 tilfælde:

A. Fordelingerne har ukendte varianser, som kan antages at være ens.
B. Fordelingerne har ukendte varianser, som kan antages at være forskellige.
C. Begge fordelinger X og Y har kendte varianser σ_X^2 og σ_Y^2)

Såfremt hypoteser om ens varianser og ens middelværdier begge accepteres kan det antages at de variable X og Y følger samme normalfordeling, hvor μ og σ^2 estimeres på grundlag af samtlige $n=n_x+n_y$ observationer.

<u>Test for ens varians i 2 normalfordelinger (p.323-326)</u>
H_0: $\sigma_X^2 = \sigma_Y^2$

Ved **ensidet** test defineres teststørrelsen således:

$V = \frac{s_x^2}{s_y^2} \sim F(f_1 = n_{tæller}-1, f_2 = n_{nævner}-1)$, *hvor f_1 er antal frihedsgrader i tælleren, f_2 i nævneren.*

Signifikanssandsynligheden angives evt. i interval efter opslag i F-fordelingen.
H_1: $\sigma_X^2 > \sigma_Y^2 \rightarrow p = P(V > v) = 1- P(V < v)$
H_1: $\sigma_X^2 < \sigma_Y^2 \rightarrow p = P(V < v)$

Ved **dobbeltsidet** test defineres teststørrelsen således:

$$V' = \frac{\max\{s_x^2, s_y^2\}}{\min\{s_x^2, s_y^2\}}$$

Signifikanssandsynligheden angives evt. i interval efter opslag i $F(f_1 = n_{tæller}-1, f_2 = n_{nævner}-1)$, *hvor f_1 er antal frihedsgrader i tælleren, f_2 i nævneren..*
H_1: $\sigma_X^2 \neq \sigma_Y^2 \rightarrow p = 2 * P(V' > v') = 2 * (1- P(V' < v'))$

<u>Test for ens middelværdi i 2 NF med ukendt varians, som kan antages ens (tilfælde A, p. 327)</u>
Ved at foretage test for ens varianser i 2 NF kan det afgøres hvorvidt varianserne kan antages ens. En fælles varians S_f^2 kan herefter sammenvejes således:

$$S_f^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

H_0: $\mu_X = \mu_Y$

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{S_f^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim T(n_X + n_Y - 2)$$

Signifikanssandsynligheden angives evt. i interval efter opslag i T-fordelingen.
H_1: $\mu_X > \mu_Y \rightarrow p = P(T > t) = 1 - P(T < t)$

$H_1: \mu_X < \mu_Y \rightarrow p = P(T < t)$
$H_1: \mu_X \neq \mu_Y \rightarrow p = 2\,(P\,(T > |t|\,)) = 2\,(1 - P(T < |t|\,))$

Test for ens μ i 2 NF med ukendt varians, som kan antages forskellig (tilfælde B, p. 321-322))

Hvis man får forkastet en hypotese om ens varianser i 2 NF angiver pensum at man ikke kender den eksakte fordeling af differensen $\overline{X} - \overline{Y}$ (Fisher-Behrens problemet). Man må derfor nøjes med nedenstående approksimative test til at vurdere en hypotese om at niveauet i fordelingerne er ens. Der er principielt intet problematisk ved at vurdere om to fordelinger har samme niveau selvom deres varianser er forskellige (det fremgår f.eks. enkelt af tilfælde A.)

$H_0: \mu_X = \mu_Y$

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}}$$

Signifikanssandsynligheder findes ved opslag i U-fordelingen

$H_1: \mu_X > \mu_Y \rightarrow p = P(U>u) = 1 - P(U<u) = 1 - \Phi(u)$
$H_1: \mu_X < \mu_Y \rightarrow p = P(U<u) = \Phi(u)$
$H_1: \mu_X \neq \mu_Y \rightarrow p = 2\,(P(U > |u|\,)) = 2\,(1 - P(U < |u|\,)) = 2\,(1-\Phi(\,|u|\,))$

Test for ens middelværdi i 2 NF med kendt varians (tilfælde C, p. 320)

$H_0: \mu_X = \mu_Y$

$$U = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$$

Signifikanssandsynligheder findes ved opslag i U-fordelingen

$H_1: \mu_X > \mu_Y \rightarrow p = P(U>u) = 1 - P(U<u) = 1 - \Phi(u)$
$H_1: \mu_X < \mu_Y \rightarrow p = P(U<u) = \Phi(u)$
$H_1: \mu_X \neq \mu_Y \rightarrow p = 2\,(P(U > |u|\,)) = 2\,(1 - P(U < |u|\,)) = 2\,(1-\Phi(\,|u|\,))$

Test for middelværdi ved parvist samhørende observationer (p.344-347)

Såfremt en stikprøve er indsamlet med parvist samhørende observationer af to forskellige variable kan det undersøges om middelværdien af de fordelinger er ens. Dette gøres ved at betragte differensen mellem de to variable $D_i = X_i - Y_i$ mellem de *n* talpar.

Man kan herefter teste om middelværdien af denne differens kan antages at have en given værdi:

$H0: \mu_D = \mu_{0D}$

$T_D = \dfrac{\overline{D} - \mu_0}{S_D / \sqrt{n}}$ *~ T(n-1), idet* $\overline{D}$ *og* S_D *er hhv. gennemsnittet og spredning på differenserne.*

Signifikanssandsynligheden angives evt. i interval efter opslag i T-fordelingen.

$H_1: \mu_D > \mu_{0D} \rightarrow p = P(T > t) = 1 - P(T < t)$
$H_1: \mu_D < \mu_{0D} \rightarrow p = P(T < t)$
$H_1: \mu_D \neq \mu_{0D} \rightarrow p = 2\,(P\,(T > |t|\,)) = 2\,(1 - P(T < |t|\,))$

Sammenligning af flere normalfordelinger $X_1, X_2, .. , X_k$

Bartlett's test for ens varians i flere normalfordelinger (p.334-335)

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

H_1: Mindst en varians er forskellig fra de andre

De empiriske varianser S_j^2 estimeres på vanlig vis i de k normalfordelinger.
Den fælles vægtede varians S_I^2 - *variansen indenfor fordelingerne* - beregnes herefter således:

$$S_I^2 = \frac{1}{n-k}\sum_{j=1}^{k}(n_j - 1)S_j^2 \text{ , hvor } n = n_1 + n_2 + \ldots + n_k$$

Bartlett's teststørrelse:

$$B = (n-k)\ln(S_I^2) - \sum_{j=1}^{k}(n_j - 1)\ln(S_j^2) \sim \chi^2(\text{k-1})$$

Signifikanssandsynligheden angives evt. i interval efter opslag i χ^2-fordelingen:
p = P(B>b) = 1 - P(B<b)

Test for ens middelværdi i flere normalfordelinger (p.336-337)

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: Mindst en middelværdi er forskellig fra de andre

Gennemsnittene $\overline{X}_j$ estimeres på vanlig vis for middelværdierne i de k normalfordelinger.
Man betragter *variationen mellem gennemsnittene* udtrykt ved

$$S_M^2 = \frac{1}{k-1}\sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X})^2 ,$$

hvor $\overline{X}$ er det sammenvejede gennemsnit, defineret som $\overline{X} = \frac{1}{n}\sum_{j=1}^{k} n_j \overline{X}_j$.

V-teststørrelse:

$$V = \frac{S_M^2}{S_I^2} \sim \text{F(k-1, n-k)}$$

Signifikanssandsynligheden angives evt. i interval efter opslag i F-fordelingen.
p = P(V > v) = 1 - P(V < v)

Estimation af parametre, der beskriver én fælles normalfordeling

Såfremt hypoteser om ens varianser og ens middelværdier begge accepteres kan det antages at de variable $X_1 \ldots X_k$ følger samme normalfordeling, hvor μ_T og σ_T^2 så estimeres på grundlag af samtlige $n = \sum_{i=1}^{k} n_i$ observationer. Denne estimation kan let foretages ved at udnytte at

μ_T estimeres ved $\bar{x}_T = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j$

σ_T^2 estimeres ved $s_T^2 = \frac{Q_T}{n-1} = \frac{Q_I + Q_M}{n-1} = \frac{(n-k)s_I^2 + (k-1)s_M^2}{n-1}$

Multinomialfordelingen

Model: $X_1, X_2, ..., X_J \sim M(n, \theta_1, \theta_2, ..., \theta_J)$

Test i multinomialfordelingen (p.526-536):

$H_0 : \theta_1 = \theta_{10}, \theta_2 = \theta_{20}, ..., \theta_J = \theta_{J0}$ (Dvs. at multinomialfordelingen følger en given sandsynlighedsfordeling. Ofte testes specifikt om sandsynlighedsfordelingen er en *ligefordeling* på de J udfald.)

$H_1 : Ej\ H_0$

Teststørrelse:

$Q = \sum_{i=1}^{J} \frac{(X_i - n\theta_{i0})^2}{n\theta_{i0}} \sim \chi^2(J - r - 1)$, hvor *J* er antal kategorier og *r* er antal restriktioner i den parametriske multinomialfordeling.

Signifikanssandsynlighed:

$p = P(Q > q)$

Bemærk at *antal restriktioner* er relevant når multinomialfordelingen bruges til numerisk kontrol af om et datasæt følger en given fordeling. Hvis det er normalfordelingen der kontrolleres estimeres 2 parametre μ og σ^2,følgelig fås r=2. Ved binomial-, poisson- eller eksponentialfordeling fås r=1.

Sammenligning af 2 multinomialfordelinger (p.579-584):

$H_0 : \theta_{1j} = \theta_{2j}\ \forall\ j \in \{1,2,...,J\}$

$H_1 : Ej\ H_0$

Teststørrelse:

$Q = \sum_{i=1}^{2} \sum_{j=1}^{J} \frac{(X_{ij} - n_i\hat{\theta}_j)^2}{n_i\hat{\theta}_j} \sim \chi^2((2-1)(J-1))$, hvor *J* er antal kategorier, og $\hat{\theta}_j = \frac{x_{.j}}{n}$ er et middelret estimat for sandsynligheden for det j'te udfald under H_0.

Signifikanssandsynlighed:

$p = P(Q > q)$

Sammenligning af I multinomialfordelinger (p.579-584):

$H_0 : \theta_{ij} = \theta_{.j}\ \forall\ i \in \{1,2,...,I\}, j \in \{1,2,...,J\}$

$H_1 : Ej\ H_0$

Teststørrelse:

$Q = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(X_{ij} - n_i\hat{\theta}_j)^2}{n_i\hat{\theta}_j} \sim \chi^2((I-1)(J-1))$, hvor *I* er antal multinomialfordelinger og *J* er antal kategorier, og $\hat{\theta}_j = \frac{x_{.j}}{n}$ er et middelret estimat for sandsynligheden for det j'te udfald under H_0.

Signifikanssandsynlighed:

$p = P(Q > q)$

Stikprøveteorien

Pensum skelner mellem
- Simpel tilfældig udvælgelse
- Stratificeret udvælgelse (herunder generelt, proportional all. samt optimal all. (kursorisk))

samt mellem
- Alternativ variation (Binær variabel, man estimerer andel/antal af mærkede enheder.)
- Generel variation (Alle andre typer variable end binær. Man estimerer variablens niveau.)

"Alternativ variation" er naturligvis bare et specialtilfælde af "Det generelle tilfælde".

Simpel tilfældig udvælgelse med alternativ variation

Den sande andel $\theta = \frac{M}{N}$, andel af mærkede enheder M i hele population N.

Et estimat for denne andel er $\hat{\theta} = \frac{x}{n}$, dvs. andel mærkede enheder x i stikprøven n.

X ~ Hyp.geo (N,M,n), $E(X) = \frac{M}{N}n$, $Var(x) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(\frac{N-n}{N-1}\right)$

For andelen θ gælder at $E(\hat{\theta}) = \theta$, og $Var(\hat{\theta}) = \frac{N}{N-1}\frac{\theta(1-\theta)}{n}(1-f)$, som estimeres ved det mid-

delrette skøn $\hat{Var}(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n-1}(1-f)$, hvor udvalgsbrøken $f = \frac{n}{N}$

Antallet af mærkede i populationen Nθ estimeres ved $N\hat{\theta}$. $Var(N\hat{\theta}) = N^2 Var(\hat{\theta})$

Konfidensinterval for θ: $\hat{\theta} \pm u_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\theta})}$

Hvis N er stor, og n er lille er *endelighedskorrektionen* (1-f) ≈ 1.

Bestemmelse af stikprøvestørrelse, som tilfredsstiller et givet præcisionskrav:
Hvis det er krævet at $Var(\hat{\theta}) \le \sigma_0^2$, gælder at stikprøvestørrelsen bestemmes som:

mindste tal for hvilket $n \ge \frac{\theta(1-\theta)}{\sigma_0^2 + \theta(1-\theta)/N}$

Hvis det er krævet at et konfidensinterval højest har bredden L_0, kan det udnyttes at $\sigma_0^2 = \left(\frac{L_0}{2u_{1-\alpha/2}}\right)^2$

θ skal bestemmes ved et kvalificeret skøn, evt. ved at sætte θ = ½, hvilket sikrer maksimal populationsvarians, og dermed det mest forsigtige skøn over stikprøvestørrelsen.

Simpel tilfældig udvælgelse, det generelle tilfælde

Som estimat for populationsgennemsnittet anvendes stikprøvegennemsnittet $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Som estimat for populationsvariansen τ^2 anvendes stikprøvevariansen $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Som estimat over variansen på gennemsnittet $\bar{x}$ anvendes $\hat{Var}(\bar{x}) = \frac{s^2}{n}(1-f)$

Et approx. konfidensinterval for det sande gennemsnit $\overline{X}$ er $\bar{x} \pm u_{1-\alpha/2}\sqrt{\hat{Var}(\bar{x})}$

Populationstotalen estimeres ved $N\bar{x}$.

Variansen på populationstotalen er $\hat{Var}(N\bar{x}) = N^2\hat{Var}(\bar{x})$

Bestemmelse af stikprøvestørrelse, som tilfredsstiller et givet præcisionskrav:

Hvis det er krævet at $Var(\bar{x}) \le \sigma_0^2$, gælder at stikprøvestørrelsen bestemmes som:

mindste tal for hvilket $n \ge \frac{\tau^2}{\sigma_0^2 + \tau^2/N}$, hvor τ^2 er et kvalificeret bud på populationsvariansen, som s^2 jo er vores skøn over.

Hvis det er krævet at et konfidensinterval højest har bredden L_0, kan det udnyttes at $\sigma_0^2 = \left(\frac{L_0}{2u_{1-\alpha/2}}\right)^2$

Stratificeret udvælgelse, alternativ variation

Populationen opdeles i *j* strata.

Den sande andel i et stratum $\theta_j = \frac{M_j}{N_j}$, andel af mærkede enheder M_j i stratum N_j.

Et estimat for denne andel er $\hat{\theta}_j = \frac{x_j}{n_j}$, dvs. andel mærkede enheder x_j i stikprøvestratum n_j.

Variansen på andelen i det enkelte stratum kan estimeres ved det middelrette skøn

$\hat{Var}(\hat{\theta}_j) = \frac{\hat{\theta}_j(1-\hat{\theta}_j)}{n_j - 1}(1-f_j)$, hvor udvalgsbrøken $f_j = \frac{n_j}{N_j}$

Som skøn over andel mærkede θ i populationen anvendes

$\hat{\theta} = \sum_{j=1}^{m} W_j\hat{\theta}_j$, hvor stratumvægten $W_j = \frac{N_j}{N}$

Et middelret skøn over variansen på estimatet for populationsgennemsnittet er

$$\hat{Var}(\hat{\theta}) = \sum_{j=1}^{m} W_j^2\hat{Var}(\hat{\theta}_j) = \sum_{j=1}^{m} W_j^2\frac{\hat{\theta}_j(1-\hat{\theta}_j)}{n_j - 1}(1-f_j)$$

Et approx. konfidensinterval for den sande andel θ er $\hat{\theta} \pm u_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\theta})}$

Populationstotalen estimeres ved $N\bar{x}$.

Variansen på populationstotalen er $\hat{Var}(N\bar{x}) = N^2\hat{Var}(\bar{x})$.

Stratificeret udvælgelse, det generelle tilfælde

Populationen opdeles i *j* strata.
Som estimat for populationsgennemsnittet i det j. stratum anvendes stikprøvegennemsnittet

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}$$

Som estimat for populationsvariansen i det j. stratum anvendes stikprøvevariansen

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

Variansen på gennemsnittet i det enkelte stratum kan estimeres ved

$$\hat{Var}(\bar{x}_j) = \frac{s_j^2}{n_j}(1 - f_j)$$

Som skøn over populationsgennemsnittet $\bar{X}$ anvendes

$\bar{x} = \sum_{j=1}^{m} W_j \bar{x}_j$, hvor stratumvægten $W_j = \frac{N_j}{N}$

Et middelret skøn over variansen på estimatet for populationsgennemsnittet er

$$\hat{Var}(\bar{x}) = \sum_{j=1}^{m} W_j^2 \hat{Var}(\bar{x}_j) = \sum_{j=1}^{m} W_j^2 \frac{s_j^2}{n_j}(1 - f_j)$$

Et approx. konfidensinterval for det sande gennemsnit $\bar{X}$ er $\bar{x} \pm u_{1-\alpha/2}\sqrt{\hat{Var}(\bar{x})}$
Populationstotalen estimeres ved $N\bar{x}$.
Variansen på populationstotalen er $\hat{Var}(N\bar{x}) = N^2 \hat{Var}(\bar{x})$

Proportional allokering

Bemærk: Alle formler herunder er formuleret i det generelle tilfælde. Hvis man analyserer en binær variabel kan man blot erstatte: $\bar{x} = \hat{\theta}$ og $\text{var}(\bar{x}) = \text{var}(\hat{\theta})$.

Når man foretager en stratificeret udvælgelse, er det naturligvis af interesse at minimere $Var(\bar{X})$.
En proportional allokering udnytter *variationen mellem strata* til at mindske variansen på estimatoren. (*variationen mellem strata* består i forskellige niveauer af en kontinuert variabel eller forskellige andele med et givet karakteristika hvis variablen er binær).
Dvs. at $Var(\bar{X}_{prop}) \le Var(\bar{X})$.

Den proportionale allokering er kendetegnet ved ens udvalgsbrøk for alle strata $f_j = \frac{n_j}{N_j} = \frac{n}{N}$.

For hvert stratum skal således udvælges $n_j = n\frac{N_j}{N} = nW_j$ enheder.

Bemærk herunder: Ved alternativ variation haves $\tau_j^2 = \frac{N_j}{N_j - 1}\theta_j(1-\theta_j) \approx \theta_j(1-\theta_j)$.

Bestemmelse af stikprøvestørrelse, som tilfredsstiller et givet **præcisionskrav**:
Hvis det er krævet at $Var(\bar{x}) \le \sigma_0^2$, gælder at stikprøvestørrelsen bestemmes som:

mindste tal for hvilket $n \ge \frac{\sum_{j=1}^{m} W_j \tau_j^2}{\sigma_0^2 + \frac{1}{N}\sum_{j=1}^{m} W_j \tau_j^2}$, hvor τ_j^2 er populationsvariansen, som s_j^2 jo er vores skøn over (derfor kan man i mangel af bedre anvende s_j^2 som estimat for τ_j^2.)

Hvis det er krævet at et konfidensinterval højest har bredden L_0, kan det udnyttes at $\sigma_0^2 = \left(\frac{L_0}{2u_{1-\alpha/2}}\right)^2$

Optimal allokering

Bemærk: Alle formler herunder er formuleret i det generelle tilfælde. Hvis man analyserer en binær variabel kan man blot erstatte: $\bar{x} = \hat{\theta}$ og $\text{var}(\bar{x}) = \text{var}(\hat{\theta})$.

Givet at variansen indenfor stratummet er ens for vores strata, så sikrer den proportionale allokering at variansen på estimatoren minimeres. Men variansen indenfor strata er sjældent ens, og disse variansforskelle udnyttes af den optimale allokering til at minimere variansen på estimatoren i forhold til den proportionale allokering. Dvs. at $Var(\bar{X}_{opt}) \le Var(\bar{X}_{prop}) \le Var(\bar{X})$.

Bemærk herunder: Ved alternativ variation haves $\tau_j^2 = \frac{N_j}{N_j - 1}\theta_j(1-\theta_j) \approx \theta_j(1-\theta_j)$.

For hvert stratum skal således udvælges $n_j = n\frac{W_j \tau_j}{\sum_{i=1}^{m} W_i \tau_i}$ enheder, hvor τ_j^2 er populationsvariansen, som s_j^2 jo er vores skøn over (derfor kan man i mangel af bedre anvende s_j^2 som estimat for τ_j^2.)

Bestemmelse af stikprøvestørrelse, som tilfredsstiller et givet **præcisionskrav**:
Hvis det er krævet at $Var(\bar{x}) \le \sigma_0^2$, gælder at stikprøvestørrelsen bestemmes som:

mindste tal for hvilket $n \ge \frac{\left(\sum_{j=1}^{m} W_j \tau_j\right)^2}{\sigma_0^2 + \frac{1}{N}\sum_{j=1}^{m} W_j \tau_j^2}$, hvor τ_j^2 er populationsvariansen, som s_j^2 jo er vores skøn over (derfor kan man i mangel af bedre anvende s_j^2 som estimat for τ_j^2.)

Hvis det er krævet at et konfidensinterval højest har bredden L_0, kan det udnyttes at $\sigma_0^2 = \left(\frac{L_0}{2u_{1-\alpha/2}}\right)^2$