STOCKHOLMS UNIVERSITET VT2025

Statistiska institutionen
Dan Hedlin

Tentamen i statistik och dataanalys III, ST2101

2025-03-19

Skrivtid: 08.00-12.00
Godkända hjälpmedel: miniräknare, språklexikon samt en medhavd A4-sida på vilken studenten kan ha skrivit eller kopierat vad som helst, på vardera sidan

Tentamen består av tio uppgifter. För full poäng på en uppgift 8-10 krävs tydliga, utförliga och väl motiverade svar och lösningar

Lägg märke till att det finns ett kort formelblad i slutet på tentan

Betygsgränser:
A: 90-100
B: 80-89
C: 70-79
D: 60-69
E: 50-59
Fx: 40-49
F: 0-39

Uppgift 1-7 ska bara besvaras med endera av bokstäverna a, b, c eller d. Motivering behövs inte. Se formler i slutet på den har tentamen för definitioner. Den maximala poängen är 35, fem för vardera av uppgift 1-7.

Uppgift 1.
Nedan följer fyra påståenden. Vilket av dem är *inte* korrekt?

a) Inklusionssannolikheten för objekt *k*, π_k, är sannolikheten att objektet ingår i urvalet *s*, dvs $\Pr(k \in s)$.
b) Att dra överurval (oversampling) syftar ofta på att man väljer att ha större urvalsstorlek i ett stratum än den optimala urvalsstorleken.
c) Om man drar ett OSU utan återläggning med storleken *n*, då har objekt *k* inklusionssannolikheten $\pi_k = N/n$.
d) För att ett urval ska vara ett sannolikhetsurval (slumpmässigt urval) ska alla objekt i rampopulationen ha positiv inklusionssannolikhet, dvs $\pi_k > 0$ för alla $k \in U$.

Uppgift 2.
Vilket av följande fyra påståenden är *inte* korrekt?

a) Statistikmyndigheten SCB (Statistiska centralbyrån) publicerar statistik över människors levnadsförhållanden, som bland annat inkluderar genomsnittlig längd och vikt hos de som bor i Sverige.

b) Skattningar av hur mycket skog i skogskubikmeter det finns i Sverige ingår i Sveriges officiella statistik. (Skogskubikmeter är mått på virkesvolym.)
c) Brottsförebyggande rådet har ett personregister över hela befolkningen i Sverige där det framgår för varje person om hon/han har varit utsatt för brott under året som gått.
d) Opinionsundersökningar är en typ av statistiska urvalsundersökningar.

Uppgift 3.
Ett OSU utan återläggning dras ur befolkningsregistret. Man vill skatta andel som har BMI > 25. Målpopulationen är personer som är bosatta i Sverige. Det är känt att befolkningsregistret innehåller personer som inte bott i Sverige på mer än tre år. Vilket av följande fyra påståenden om *den här undersökningen* är *inte* korrekt?

a) Det finns ingen skillnad mellan ram- och målpopulation.
b) I det här fallet omfattar ramen och rampopulationen samma objekt.
c) Undertäckningen består av personer som bor i Sverige men som inte finns i befolkningsregistret.
d) Övertäckningen består bland annat av de personer som finns befolkningsregistret men som inte bott i Sverige på mer än tre år.

Uppgift 4.
Vilket av a), b), c) eller d) är *inte* korrekt?

a) Beslutet om urvalsstorlek kan inte avgöras med enbart statistisk teori; man måste också väga in hur undersökningens resultat ska användas.
b) En undersökning om en heterogen population kräver större urval än om populationen är relativt homogen, om allt annat är lika.
c) Antag att en undersökning går ut på att skatta hur många elever i skolår 7-9 som har sett våld mellan skolelever. Variansen för skattningen kan mycket väl bli mindre om man drar ett OSU utan återläggning på 800 elever än om man drar ett klusterurval med 12 skolor med sammanlagt 1000 elever, om vi antar att vi inte kommer att få något bortfall och att ramen är perfekt.
d) Tänk på två målpopulationer, den ena med storleken $N = 500$, den andra $N = 500\,000\,000$, och förutsätt att σ_y^2 är densamma. Man kommer att använda Horvitz-Thompsonestimatorn för att skatta totalen $\mathcal{T}_y$. Ett OSU utan återläggning med $n = 50$ (10%) från den mindre populationen kommer att ge lägre varians än ett OSU utan återläggning med $n = 500$ (0.0001%) från den större populationen, om vi antar att vi inte kommer att få något bortfall och att ramen är perfekt.

Uppgift 5.
Anta att man drar ett OSU utan återläggning med storleken *n* ur en population med storleken *N*. Låt $w_k = \frac{N}{n}$. Vilket av följande påståenden *är korrekt*?

a) $\sum_{k \in s} w_k = n$
b) $\sum_{k \in s} w_k = \frac{N}{n}$
c) $\sum_{k \in s} w_k = \frac{n}{N}$

d) $\sum_{k \in s} w_k = N$

Uppgift 6.
Vilket av a), b), c) eller d) är *inte* korrekt?

a) Klusterurval betyder att man först delar in populationen i grupper och sedan drar man ett urval av dessa grupper.
b) Stratifierat urval betyder att man först delar in populationen i grupper och sedan drar man ett urval i en mindre andel av dessa grupper.
c) I systematiskt urval ordnar man populationen i någon bestämd ordning och bestämmer steglängden i. Man drar slumpmässigt ett första objekt bland objekten $1, 2, \dots, i$. Vi kan beteckna det dragna objektet med t. Därefter dras systematiskt, ej slumpmässigt, objekt $t + i, t + 2i, t + 3i, \dots, t + (n - 1)i$.
d) Två val som måste göras när man utformar ett stratifierat urval är vilken eller vilka urvalsdesigner som ska användas och hur urvalet ska fördelas på stratum.

Uppgift 7.
Vilket av följande påståenden är *inte* korrekt?

a) När man använder formeln för optimal allokering för stratifierat OSU är det inte ovanligt att formeln resulterar i $n_h > N_h$ för ett stratum, dvs. att urvalsstorleken är större än populationsstorleken i det stratumet.
b) Den mängd urval som kan dras med systematiskt urval är en delmängd av den mängd urval som kan dras med OSU utan återläggning.
c) Den mängd urval som kan dras med enstegsklusterurval är en delmängd av den mängd urval som kan dras med OSU utan återläggning.
d) Sannolikheterna för att ingå i ett stratifierat OSU kan vara lika eller olika inom ett stratum.

Uppgift 8. (30 poäng)

En universitetslektor tänker skatta andelen studenter som har läst kursboken (och inte bara t.ex. sett videor med samma innehåll som kursboken). Hon avgränsar en målpopulation med 400 studenter och drar ett OSU utan återläggning, $n = 100$. 60 av dem svarar att de har läst kursboken, 40 att de inte har det.

a) Skatta andelen i målpopulationen som har läst kursboken.
b) Bilda ett 95% konfidensintervall för andelen du kom fram till i a) med Waldintervallet.
c) Bilda ett 95% konfidensintervall andelen du kom fram till i a) på ett bättre sätt än med Waldintervallet.
d) Antag att 100 av 100 studenter säger att de har läst kursboken. Vad blir konfidensintervallen i b) och c)?

Uppgift 9. (20 poäng)

Ett företag i mäklarbranschen vill dra ett stratifierat OSU för att skatta medelvärdet av det belopp som personer som söker bostad kan tänka sig lägga på en bostadsrätt, mätt i kronor per kvadratmeter. De tänker sig två stratum: 1) innerstad+närförort 2) alla andra platser. De vill ha en skattning för dels alla tillsammans, dels för varje stratum för sig.

Stratum	Gissat värde på σ_h^2	N_h	Gissat värde på $\sum_{k \in U_h} y_k$ i tusen kronor
1	4	10	620
2	16	100	4 100

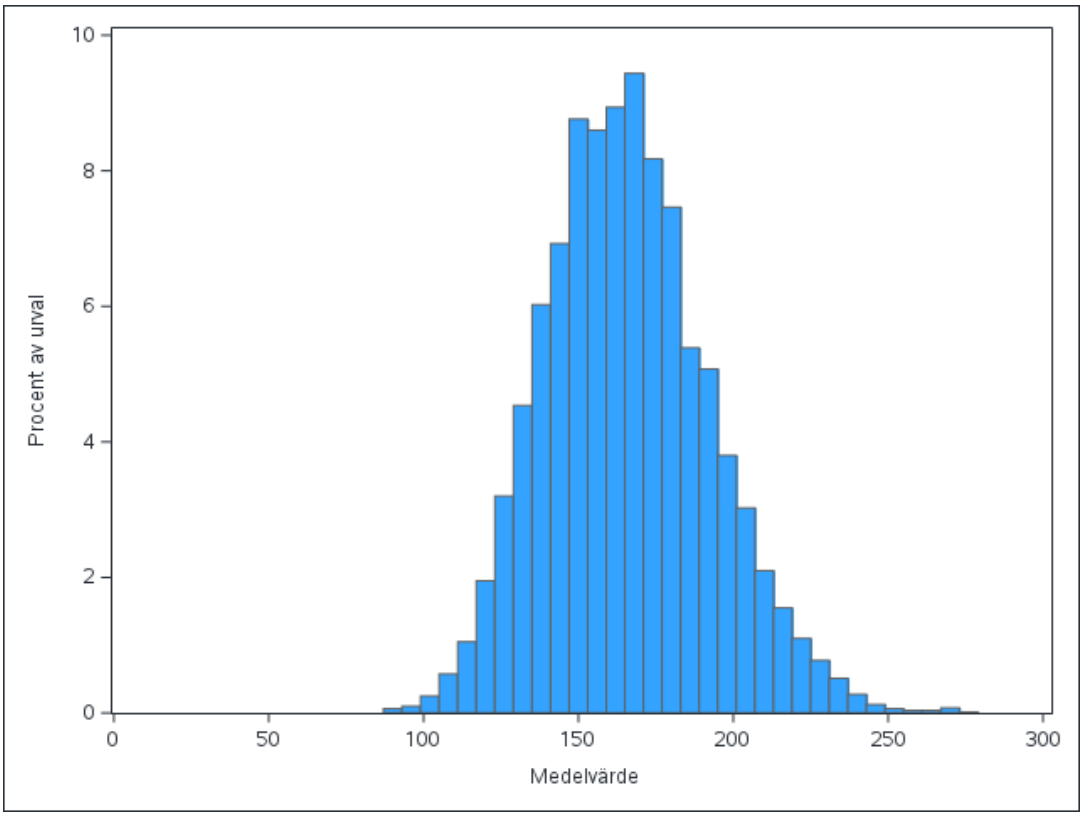
a) Urvalsstorleken är satt till $n = 20$. Fördela dessa tjugo till stratum 1 och 2 med den formel i bilagan som kallas optimal allokering
b) Vad syftar ”optimal” på i det här sammanhanget?
c) Företaget vill samtidigt ha separata skattningar för varje stratum, och båda ska ha ett 95% konfidensintervall med halva bredden 2. Beräkna urvalsstorlekar som uppfyller det kravet. Du behöver inte ta hänsyn till ändlighetskorrektion.

Uppgift 10. (15 poäng)

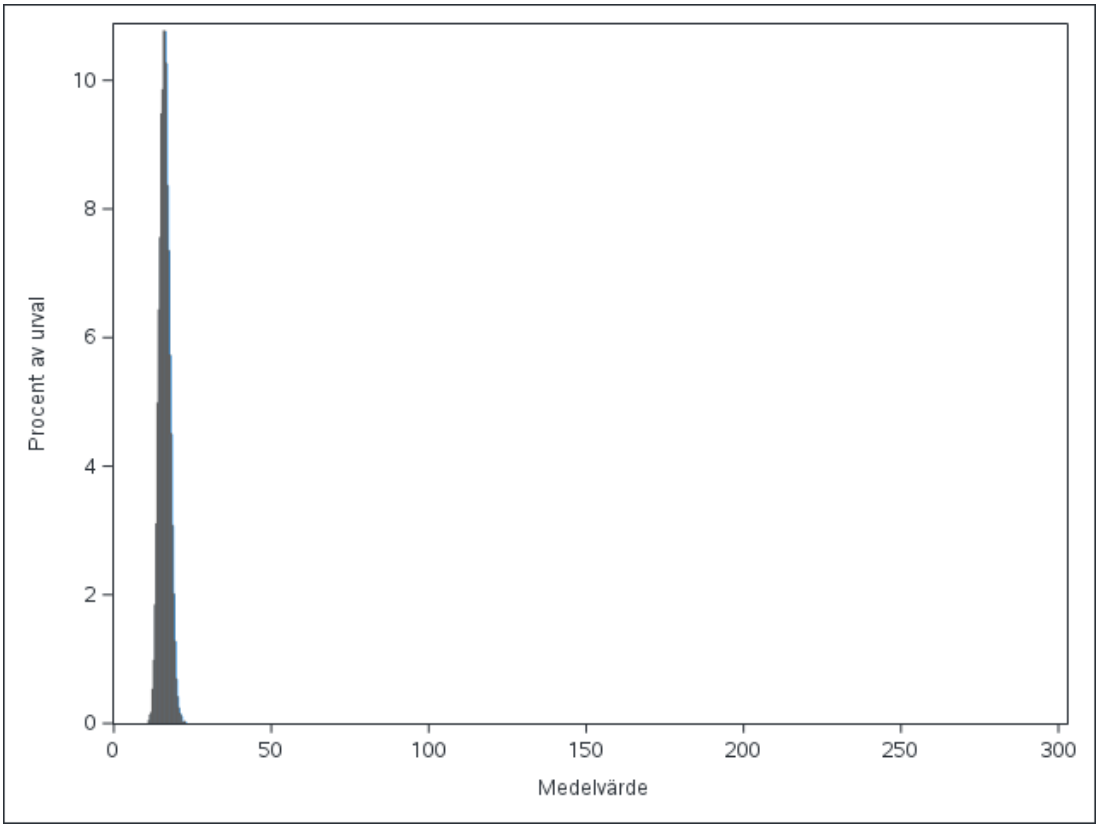
Tre grafer nedan visar urvalsfördelningen (samplingfördelningen) för OSU utan återläggning ur tre populationer. Alla tre populationerna innehåller 5000 objekt. Vi har i vår dator en fullständig lista samtliga objekt i alla tre populationerna. För att undersöka hur urval och estimation fungerar har vi med datorns hjälp dragit 8000 urval ur varje population, och för varje urval har stickprovsmedelvärdet noterats. Varje urval har storleken n =150.

a) Rangordna estimatorns varians från den population som har minst varians till den som har störst. Notera att en motivering av typen ”formen på histogrammet visar att a) har minst varians” är inte tillräcklig; du måste förklara varför formen (om du vill hänvisa till den) visar på mindre/större varians. Om du landar i att det inte går att avgöra, förklara vilken ytterligare information som behövs för att kunna avgöra frågan.
b) Vad beror den till synes märkliga formen av urvalsfördelningen för population c på?

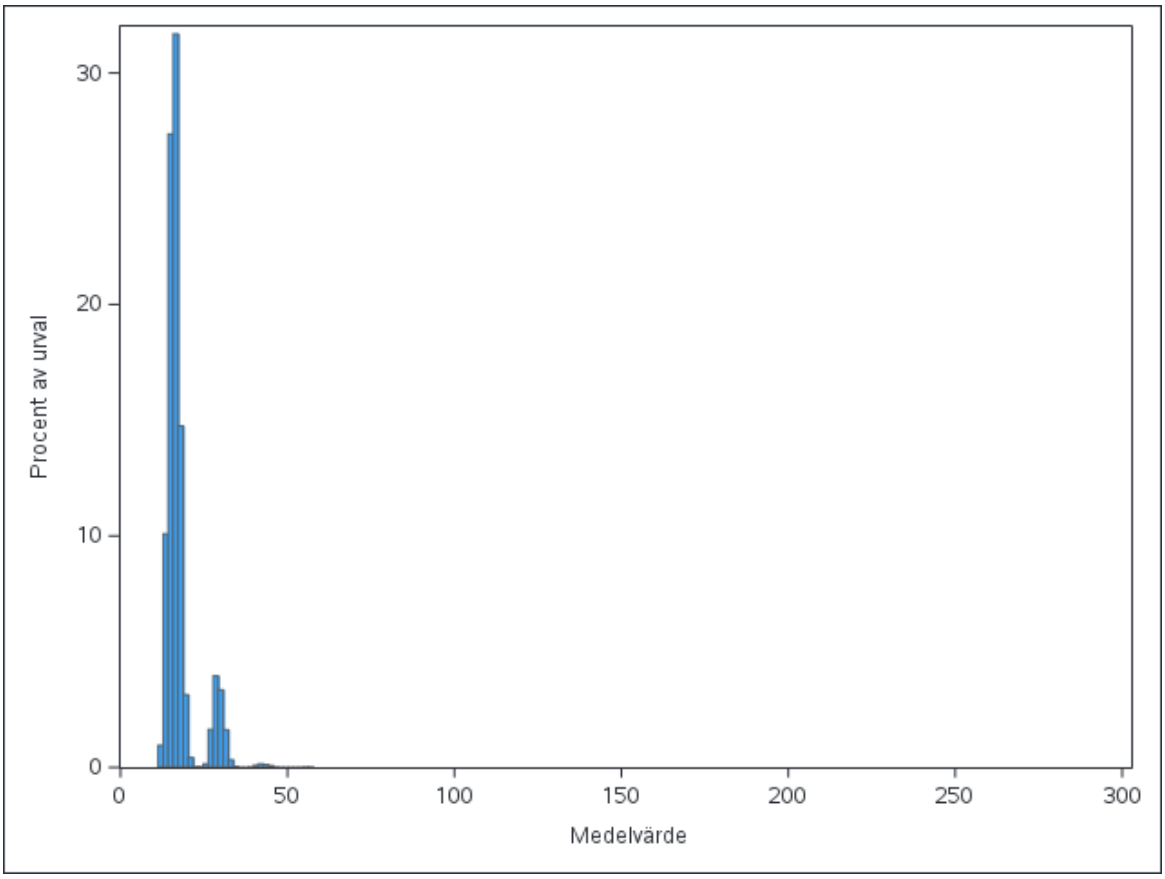
Population a



Population b



Population c



Formler

Populationsparametrar

Population $U = \{1,2, \dots, N\}$ av storlek N

Medelvärde: $\mu_y = \frac{1}{N}\sum_{k=1}^{N} y_k = \frac{1}{N}\sum_{k \in U} y_k$

Specialfall där y är en binär variabel, dvs $y_k = 0$ eller 1. Andel/proportion: $p_y = \frac{1}{N}\sum_{k \in U} y_k$.
Antal $A = \sum_{k \in U} y_k = Np_y$.

Total/populationstotal: $\mathcal{T}_y = \sum_{k \in U} y_k = N\mu_y$

Populationsvarians: $\sigma_y^2 = \frac{1}{N-1}\sum_{k \in U}(y_k - \mu_y)^2$

Populationsvarians för andel (variabel på nominalskala): $\sigma_y^2 \approx p_y(1 - p_y)$

Estimatorer/skattningar av populationsparametrar samt variansskattningar

Urval/stickprov $s = \{1,2, \dots, n\}$ av storlek n

Medelvärde: $\bar{y}_s = \frac{1}{n}\sum_{k\in s} y_k,\ \hat{V}(\bar{y}_s) = \left(1-\frac{n}{N}\right)\frac{\hat{\sigma}_y^2}{n}$

Total: $\hat{\mathcal{T}} = N\bar{y}_s, \hat{V}\left(\hat{\mathcal{T}}\right) = N^2\left(1-\frac{n}{N}\right)\frac{\hat{\sigma}_y^2}{n}$

Andel: $\hat{p}_y = \frac{1}{n}\sum_{k\in s} y_k, \hat{V}\left(\hat{p}_y\right) = \left(1-\frac{n}{N}\right)\frac{\hat{p}_y(1-\hat{p}_y)}{n-1}$

Antal: $\hat{A} = N\hat{p}_y, \hat{V}\left(\hat{A}\right) = N^2\left(1-\frac{n}{N}\right)\frac{\hat{p}_y(1-\hat{p}_y)}{n-1}$

Horvitz-Thompson-estimatorn för total

$\hat{\mathcal{T}}_{HT} = \sum_{k\in s}\frac{y_k}{\pi_k} = \sum_{k\in s} w_k y_k$, där π_k är inklussionssannolikheten för objekt *k*

Ett approximativt 95% konfidensintervall för OSU och medelvärde av en kontinuerlig variabel

$\bar{y}_s \pm 1.96\cdot\sqrt{\hat{V}(\bar{y}_s)} = \bar{y} \pm 1.96\cdot\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{\sigma}_y^2}{n}}$

Ett approximativt 95% konfidensintervall för OSU och andel m.a.p. en diskret variabel

$\hat{p}_y + \frac{\frac{2}{n}(1-2\hat{p}_y) \pm 2\sqrt{\hat{V}(\hat{p}_y)+\frac{1}{n^2}}}{1+\frac{4}{n}}$ Istället för 1.96 har vi här använt 2

Urvalsstorlek för OSU

$n = \left(1.96\frac{\sigma_y}{e}\right)^2$, där *e* är halva konfidensintervallet (95%). Kan användas för skattning av både medelvärde och andel, diskret eller kontinuerlig.

Stratifierat OSU, med *L* stratum

$\hat{\mathcal{T}} = N_1\bar{y}_1 + \cdots + N_L\bar{y}_L = \hat{\mathcal{T}}_1 + \cdots + \hat{\mathcal{T}}_L$, där t.ex. N_1 är antal objekt i stratum 1 i populationen och $\bar{y}_1$ är stickprovsmedelvärdet i stratum 1

Estimator för populationsmedelvärdet: $\bar{y}_{str} = \frac{1}{N}(N_1\bar{y}_1 + \cdots + N_L\bar{y}_L)$

$\hat{V}(\bar{y}_{str}) = \frac{1}{N^2}\left(N_1^2\hat{V}(\bar{y}_1) + \cdots + N_L^2\hat{V}(\bar{y}_L)\right)$

Proportionell allokering: $n_h = n\cdot\frac{N_h}{\sum_{j=1}^{L} N_j}$

Optimal allokering: $n_h = n\cdot\frac{N_h\sigma_h}{\sum_{j=1}^{L} N_j\sigma_j}$

Regressionsestimatorn för OSU

Estimator för populationsmedelvärdet av *y*: $\hat{\mu}_{reg} = \bar{y} + \hat{\beta}(\mu_x - \bar{x})$, där μ_x är populationsmedelvärdet av *x* och $\bar{x}$ är stickprovsmedelvärdet, och

$\hat{\beta} = \frac{\sum_{k\in s}(y_k-\bar{y})(x_k-\bar{x})}{\sum_{k\in s}(x_k-\bar{x})^2}$

$\hat{V}(\hat{\mu}_{reg}) = \left(1 - \frac{n}{N}\right) \cdot \frac{\hat{\sigma}_y^2}{n}\left(1 - \left(\widehat{Cor}(x,y)\right)^2\right)$, där $\widehat{Cor}(x,y) = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x \hat{\sigma}_y}$ och

$\hat{\sigma}_{x,y} = \frac{1}{n-1}\sum_{k \in s}(x_k - \bar{x})(y_k - \bar{y})$

Alternativ variansskattning: $\hat{V}(\hat{\mu}_{reg}) = \left(1 - \frac{n}{N}\right) \cdot \frac{\left(\sum_{k \in S}(y_k - \bar{y})^2 - \hat{\beta}^2 \sum_{k \in S}(x_k - \bar{x})^2\right)/(n-2)}{n}$

Estimator för medelvärdet i en redovisningsgrupp ("domain")

$\bar{y}_d = \frac{1}{n_d}\sum_{k \in s_d} y_k$, där s_d är de objekt i urvalet som hör till redovisningsgruppen och n_d är antalet av dem.

$\hat{V}(\bar{y}_d) = \left(1 - \frac{n}{N}\right)\frac{\hat{\sigma}_{yd}^2}{n_d}, \qquad \hat{\sigma}_{yd}^2 = \frac{\sum_{k \in s_d}(y_k - \bar{y}_d)^2}{n_d - 1}$

Alternativt:

$\hat{V}(\bar{y}_d) = \left(1 - \frac{n_d}{N_d}\right)\frac{\hat{\sigma}_{yd}^2}{n_d}$