STOCKHOLMS UNIVERSITET
Statistiska institutionen
Michael Carlson (examinator)

HT2022
2023-02-28

# Tentamen i Undersökningsmetodik (4,5 hp)

## Kurs: Regressionsanalys och undersökningsmetodik

**2023-02-16**

# Lösningförslag

**Uppgift 1.**

$U =$ population bestående av $N = 1500$ jurister.

$S =$ OSU u.å. draget från $U$ ($S \subset U$), stickprovsstorlek $n = 200$.

Låt $y_k = 1$ om jurist nummer $k$ svarar Ja och $y_k = 0$ om $k$ svarar Nej.

Låt $P = \frac{1}{N}\sum_{k\in U} y_k =$ andel i populationen som svarar Ja.

Andelen $P$ är okänd och skattas med $\hat{p} = \frac{1}{n}\sum_{k\in S} y_k =$ andel Ja-svar i stickprovet $S$.

Utöka tabellen med marginalsummorna (i fetstil nedan):

| | Ja | Nej | Total |
|---|---|---|---|
| Kvinnor | 34 | 36 | **70** |
| Män | 12 | 118 | **130** |
| Total | **46** | **154** | **200** |

a) Punktskattning och ett 90% konfidensintervall för $P$:

Skattning av $P$: $\hat{p} = \frac{\sum_{k\in S} y_k}{n} = \frac{34+12}{200} = \frac{46}{200} = \mathbf{0.23}$

Skattning av $V(\hat{p})$: $\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = \left(1 - \frac{200}{1500}\right)\frac{0.23\cdot 0.77}{30} = \mathbf{0.00077129}$

Standardfel: $SE(\hat{p}) = \sqrt{\hat{V}(\hat{p})} = \sqrt{0.00077129} = \mathbf{0.027772}$

Felmarginal 90%: $z_{0.05} \cdot SE(\hat{p}) = 1.6449 \cdot 0.40012 = \mathbf{0.045692}$

95% KI för $P$: $= \mathbf{0.23 \pm 0.045692}$ eller $(\mathbf{0.18432}\,;\mathbf{0.27568})$

eller avrundat: $\mathbf{0.23 \pm 0.046}$ eller $(\mathbf{0.184}\,;\mathbf{0.276})$

b) Punktskattning och ett 90% konfidensintervall för $A$ = *antalet* i $U$ som svarar Ja.

Enklast: Eftersom $A = NP$ och därmed $\hat{A} = N\hat{p}$ så räcker det att multiplicera punktskattningen och felmarginalen och/eller intervallgränserna i a) ovan med $N = 1500$:

Skattning av $A$: $\hat{A} = N\hat{p} = 1500 \cdot 0.23 = \mathbf{345}$

90% KI för $A$: $N \cdot \text{felmarginal} = 1500 \cdot 0.045692 = 68.538 \Rightarrow \boxed{\mathbf{345 \pm 68.538}}$

eller $(N \cdot \text{nedre gräns} \,;\, N \cdot \text{övre gräns}) =$

$= (1500 \cdot 0.18432 \,;\, 1500 \cdot 0.27568) = \boxed{(\mathbf{276.48} \,;\, \mathbf{413,52})}$

c) Fel 1: Urvals- och undersökningsenheter var jurister, inte advokatbyråer. I stickprovet kan det alltså förkomma jurister från samma advokatkontor. Extremfall är om samtliga som svarade Ja kommer från en och samma advokatbyrå eller att inga av de tillfrågade ens arbetar på advokatbyråer.

Fel 2: Alla jurister är inte anställda i advokatbyråer, troligen finns det jurister både i populationen och i stickprovet som arbetar inom andra typer av arbetsplatser (företag eller myndigheter).

Fel 3: Vi vet inte hur populationen $U$ är definierad och avgränsad gentemot populationen av samtliga jurister (i Sverige?). Det enda vi kan uttala oss om inferensmässigt är just populationen $U$ och de objekt som hade en positiv sannolikhet att dras till stickprovet. Generaliseringar till grupper utanför $U$ är inte vetenskapligt korrekt.

Fel 4: är en förlängning av Fel 3. Könsfördelningen tog flera upp som ett potentiellt problem. Inte helt fel men vi vet inget om detta. Det kan mycket väl vara så att andelen kvinnliga jurister i populationen $U$ av 1 500 jurister är ca 35 %. Vi vet ännu mindre om vad som gäller i hela Sverige.

NOT: Flera tog upp att det endast är ett urval och att man inte kan veta hur det är för resten av juristerna. Men det är ju det som statistik handlar om att med viss felmarginal kunna säga något om hela populationen med ledning av ett fåtal dvs. urvalet (latin: *pars pro toto*). I och med att det är ett väntevärdesriktigt förfarande så kan man ändå säga att förfarandet är representativt.

d) Givet: $n = 250$, bortfallet var $n_b = 50$ och antalet svarande var $n_s = 250 - 50 = 200$. Extremfallen är att antingen så är samtliga i bortfallet Nej-svarare eller så är de alla Ja-svarare, detta ger

$$\hat{p}_{\min} = \frac{46 + 0}{250} = \mathbf{0.184} \qquad \hat{p}_{\max} = \frac{46 + 50}{250} = \mathbf{0.384}$$

Möjliga värden med noll bortfall skulle ligger alltså i intervallet $\boxed{(\mathbf{18.4\,\%\,;\,38.4\,\%})}$.

NOT: Längden på intervallet är exakt 0.20 vilket är bortfallsandelen = 50/250, se KD sid 358.

*Poststratifiering* brukar kunna användas för att justera för skeva urval. Man väljer en eller flera lämpliga stratifieringsvariabler och stratifierar OSU-stickprovet i efterhand. Man använder sedan kända populationstotaler $N_h$ och stratumvikter $W_h$ i beräkningarna. Lämplig variabel i detta fall kanske är just kön då upplevelsen av sexuella trakasserier förmodligen uppfattas olika och om ena könet är över- eller underrepresenterat blir det ”fel”. Frågan är dock om det är en skev fördelning i detta stickprov men det vet vi inget om utifrån den givna beskrivningen ($N_M$ och $N_{Kv}$ är inte givna).

e) Ett 95 % konfidensintervallets längd får inte överskrida 3.92 %-enheter vilket är samma sak som att säga att $B =$ felmarginalen högst får vara $3.92\%/2 = 1.96$ %-enheter, dvs.

max felmarginal: $B = z_{0.025} \cdot D = 1.96 \cdot D \leq 0.0196$

max standardfel: $D = \frac{B}{1.96} \leq \frac{0.0196}{1.96} = 0.01$

max varians: $D^2 = V(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{P(1-P)}{n} \approx \frac{P(1-P)}{n} \leq 0.01^2 = 0.0001$

Notera att ändlighetskorrektionen approximeras bort ovan eftersom vi inte vet hur stort $N$ är för hela Sverige. N antas vara stort och om $N \to \infty$ gäller att $\left(\frac{N-n}{N-1}\right) \to 1$.

Sedan, eftersom vi måste garantera att variansen inte överstiger $0.0001$ så sätts $P = 0.5$ (maximal varians). Insättning ger

$$\frac{P(1-P)}{n} = \frac{0.25}{n} = \frac{1}{4n} \leq 0.0001 \iff 4n \geq 10\,000 \iff \boldsymbol{n \geq 2\,500}$$

NOT: Det är inte helt fel att använda $P = 0.5$ för kvinnorna. Om man tittar på tabellen som gavs ser man att 34 av 70 kvinnor svarade Ja på frågan, dvs. 48.6 %. Används detta värde blir svaret $n \geq 2\,499$ vilket är ok. Om man däremot utgick ifrån skattningen $\hat{p} = 23.0$ % för hela populationen får man räkna med poängavdrag, den skattningen gäller ju för män och kvinnor tillsammans.

**Uppgift 2.**

$U$ = population bestående av $N = 121$ projekt.

$S$ = OSU u.å. med stickprovsstorlek $n = 10$.

$x_k$ = total kostnad i mkr för projekt $k = 1, \dots, 121$.

$y_k$ = total vinst i mkr för projekt $k = 1, \dots, 121$.

$\tau_x = 965$ mkr = total kostnaden för samtliga 121 projekt.

Från tabellen avläser vi

$$\sum_{k\in S} y_k = 157 \quad \sum_{k\in S} y_k^2 = 2\,635 \quad \sum_{k\in S} x_k = 80 \quad \sum_{k\in S} x_k^2 = 784 \quad \sum_{k\in S} x_k y_k = 1\,382$$

a) Det är troligt att små projekt genererar små vinster och stora kostsamma projekt genererar större vinster. Om det är ett linjärt eller icke-linjärt samband är osäkert men det är nog värt att undersöka om en kvot- eller regressionsskattning kan fungera (båda metoderna utgår ifrån att det finns ett linjärt samband).

Det verkar också rimligt att anta att ett projekt som inte medfört några kostnader dvs. $x = 0$ inte heller har genererat några vinster/besparingar, dvs. $y = 0$. Detta talar för att regressionslinjen passerar genom origo och att en kvotskattning vore bättre.

Små projekt med låga vinster kanske varierar mindre i $y$-led jämfört med stora kostsamma projekt som varierar mer i $y$-led. Detta skulle i så fall gå att se i ett spridningsdiagram om det finns ett s.k. trattmönster. Om så är fallet är en kvotskattning bättre.

Sammantaget verkar en kvotskattning vara det bättre alternativet men det är osäkert.

b) Skatta total vinst $\tau_y = \sum_{k\in U} y_k$ och standardfelet med kvot- eller regressionsestimatorn.

Alternativ 1: Kvotskattning

Skattning av kvoten:
$$\hat{R} = \frac{\sum_{k\in S} y_k}{\sum_{k\in S} x_k} = \frac{157}{80} = \frac{\bar{y}}{\bar{x}} = \frac{15.7}{8.0} = \mathbf{1.9625}$$

Punktskattning:
$$\hat{\tau}_{\text{kvot}} = \hat{R} \cdot \tau_x = 1.9625 \cdot 965 = \boxed{\mathbf{1\,893.8125 \approx 1894}}$$

Beräkna summan:
$$\sum_{k\in S}\left(y_k - \hat{R}x_k\right)^2 = \sum_{k\in S} y_k^2 - 2\hat{R}\sum_{k\in S} x_k y_k + \hat{R}^2 \sum_{k\in S} x_k^2$$
$$= 2635 - (2 \cdot 1.9625 \cdot 1382) + (1.9625^2 \cdot 784) = \mathbf{230.1525}$$

Variansskattning:
$$\hat{V}(\hat{\tau}_{\text{kvot}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}\left(y_k - \hat{R}x_k\right)^2}{n-1}\right)$$
$$= 121^2 \cdot \left(1 - \frac{10}{121}\right) \cdot \frac{1}{10} \cdot \left(\frac{230.1525}{9}\right) = \mathbf{34\,346.42}$$

Standardfel:
$$se(\hat{\tau}_{\text{kvot}}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{34\,346.42} = \boxed{\mathbf{185.33 \approx 185}}$$

Alternativ 2: Regressionsskattning

Lutningskoefficient: $b = \frac{n\sum x_k y_k - (\sum x_k)(\sum y_k)}{n\sum x_k^2 - (\sum x_k)^2} = \frac{10 \cdot 1382 - 80 \cdot 157}{10 \cdot 784 - 80^2} = 0.875$

Punktskattning: $\hat{\tau}_{\text{reg}} = N\hat{\mu}_{\text{reg}} = N\left(\bar{y} + b(\mu_x - \bar{x})\right)$

$= 121\left(15.7 + 0.875 \cdot (7.9752 - 8)\right) = \boxed{\mathbf{1897.075 \approx 1897}}$

Stickprovsvarianser: $s_y^2 = \frac{2635 - 157^2/10}{10 - 1} = 18.9 \qquad s_x^2 = \frac{784 - 80^2/10}{10 - 1} = 16$

Beräkna summan: $\sum_{k\in S}(y_k - \bar{y})^2 - b^2\sum_{k\in S}(x_k - \bar{x})^2 = (n-1)s_y^2 - b^2(n-1)s_x^2$

$= 9 \cdot 18.9 - 0.875^2 \cdot 9 \cdot 16 = \mathbf{59.85}$

Variansskattning: $\hat{V}(\hat{\tau}_{reg}) = N^2\left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \left(\frac{\sum_{k\in S}(y_k - \bar{y})^2 - b^2\sum_{k\in S}(x_k - \bar{x})^2}{n-2}\right)$

$= 121^2 \cdot \left(1 - \frac{10}{121}\right) \cdot \frac{1}{10} \cdot \left(\frac{59.85}{8}\right) = 10\,048.07$

Standardfel: $se(\hat{\tau}_{\text{reg}}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{10\,048.07} = \boxed{\mathbf{100.24 \approx 100}}$

c) Skatta den totala vinsten $\tau_y = \sum_{k\in U} y_k$ och ett 95% konfidensintervall med HT-estimatorn.

Punktskattning: $\hat{\tau}_{\text{HT}} = N\bar{y} = 121 \cdot \frac{157}{10} = \boxed{\mathbf{1\,899.7 \approx 1897}}$

Stickprovsvarians: $s_y^2 = \frac{2635 - 157^2/10}{10 - 1} = 18.9$

Variansskattning: $\hat{V}(\hat{\tau}_{\text{HT}}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_y^2}{n} = 121^2 \cdot \left(1 - \frac{10}{121}\right) \cdot \frac{18.9}{10}$

$= 25\,384.59$

Standardfel: $se(\hat{\tau}_{\text{reg}}) = \sqrt{\hat{V}(\bar{y})} = \sqrt{25\,384.59} = \boxed{\mathbf{159.33 \approx 159}}$

d) Vi kan se ett hyfsat linjärt samband i början men också att de tre observationerna med störst $x$-värden ”drar ner” linjen vilket kan tyda på ett icke-linjärt samband. Det är tveksamt om linjen går genom origo, linjen tycks snarare skära $y$-axeln vid en punkt mellan 5 och 10 vilket betyder att projekt som inte kostar något i genomsnitt genererar en positiv vinst (vilket är konstigt och bör undersökas närmare). Det finns ett svagt trattmönster åtminstone i början men de tre observationerna visar en minskad spridning för de största x-värdena.

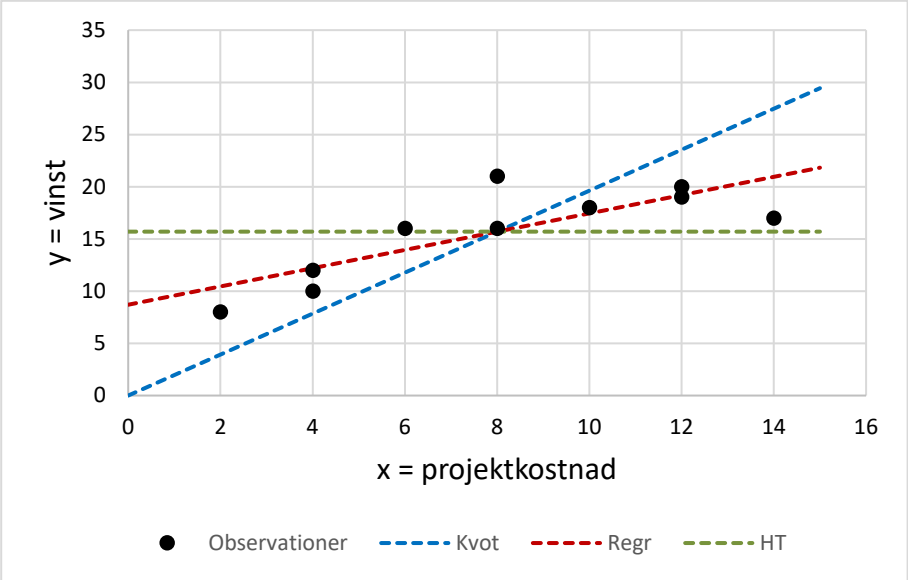
Slutsats: Förutsättningarna är inte riktigt uppfyllda här vare sig för kvot- eller regressionsskattning. Sammantaget verkar det kanske vara bättre med en regressionsskattning då interceptet är skiljt från noll men vi har ändå (vissa) problem med heteroskedasticitet och (allvarligare) problem med icke-linjäritet. Men vi ska också notera att det är ett relativt litet stickprov och att det är svårt att dra definitiva slutsatser.

NOT 1: En sammanställning av resultaten visar att det inte blir så stor skillnad i punktskattningarna men att regressionsskattning ger lägst standardfel och kvotskattning ger störst:

| | Kvot | HT | Regression |
|---|---|---|---|
| Punktskattning | 1894 | 1897 | 1897 |
| Standardfel | 185 | 159 | 100 |

Möjligen är den bakomliggande regressionsmodellen mer flexibel med två parametrar (lutning och intercept) och anpassar sig bättre till just detta datamaterial jämfört med kvotskattning med endast en parameter (kvoten, lutningen).

NOT 2: För att ”se” detta kan man rita in de skattade regressionslinjerna i diagrammet (ingick inte i uppgiften) och studera hur avstånden till respektive linje påverkar slutresultaten vilket i och för sig kan vara lite svårt för ett otränat öga:




Det man ska titta efter är avstånden (egentligen de kvadrerade avstånden) till respektive linje och hur de skiljer sig mellan metoderna. Dessa avstånd är ju residualerna $y_k - \hat{y}_k$ dvs. skillnaden mellan observerat värde och predicerat värde enligt modellen. Observera att ”modellen” för HT-estimatorn helt enkelt är $\hat{y}_k = \bar{y}$, för kvotestimatorn är det $\hat{y}_k = \hat{R}x_k$ och för regressionsestimatorn är det $\hat{y}_k = a + bx_k$.

NOT 3: Vi ser en tendens till icke-linjäritet, att det finns en böjning nedåt för stora $x$. Om det hade ingått i kursen så hade vi kunnat pröva *polynomregression* t.ex. en kvadratisk modell enligt $\hat{y}_k = a + b_1x_k + b_2x_k^2$. Då hade vi kunnat pressat ner standardfelet till ca 63.

**Uppgift 3.**

Population bestående av $N = 1000$ jordbruk indelade i $L = 4$ strata (storleksklasser). Uppgifter om $N_h$ = antal, $\mu_h$ = medelvärde och $\sigma_h$ = standardavvikelse för $y$ = skörd i ton ges för respektive stratum och för populationen som helhet ($N$, $\mu_y$ resp. $\sigma_y$) för året 2020. Dessutom ges stratumstorlek, medelvärde och varians per stratum i stickprovet år 2022.

a) Ett stratifierat urval lönar sig om stratumen är olika med avseende på nivå, dvs. de har markant olika medelvärden $\mu_h$. Att det var så år 2020 kan vi se i den givna tabellen. Observationerna inom strata ska dessutom likna varandra, dvs. standardavvikelserna $\sigma_h$ (eller varianserna $\sigma_h^2$) ska vara små jämfört med hela populationens standardavvikelse $\sigma_y$. I den givna tabellen ser vi att stratumen 1-3 har standardavvikelser som är mindre än populationens standardavvikelse $\sigma_y$, endast $\sigma_4$ är större.

Sammantaget och om inga större förändringar har skett sedan 2020 ska ett stratifierat urval för år 2022 ge en bättre skattning än ett vanligt OSU. Om man dessutom ska särredovisa varje stratum så lönar det sig att stratifiera urvalet från start istället för att post-stratifiera efteråt.

b) En optimal allokering utifrån 2020 års data ges av:

| $h$ | $N_h$ | $\sigma_h$ | $N_h\sigma_h$ | $\frac{N_h\sigma_h}{\sum_i N_i\sigma_i}$ | $n \cdot \frac{N_h\sigma_h}{\sum_i N_i\sigma_i}$ | **Avrundat** |
|---|---|---|---|---|---|---|
| 1 | 200 | 450 | 90 000 | 0,084986 | 8,4986 | **8** |
| 2 | 420 | 700 | 294 000 | 0,277620 | 27,7620 | **28** |
| 3 | 250 | 1 400 | 350 000 | 0,330500 | 33,0500 | **33** |
| 4 | 130 | 2 500 | 325 000 | 0,306893 | 30,6893 | **31** |
| | 1 000 | | 1 059 000 | 1 | 100 | **100** |

NOT 1: Proportionell allokering är *inte* rätt svar men delpoäng ges om det åtminstone gjordes rätt:

| $h$ | $N_h$ | $\frac{N_h}{N}$ | $n \cdot \frac{N_h}{N}$ |
|---|---|---|---|
| 1 | 200 | 0,20 | **20** |
| 2 | 420 | 0,42 | **42** |
| 3 | 250 | 0,25 | **25** |
| 4 | 130 | 0,13 | **13** |
| | 1 000 | 1 | **100** |

Observera att en optimal allokering drar fler från stratum 3 och 4 och färre från stratum 2 jämfört med en proportionell allokering. Detta beror förstås på att spridningen (standardavvikelserna) är jämförelsevis större i stratum 3 och 4 och mindre i stratum 2; genom att dra fler från dessa kan vi alltså minska standardfelet för skattningen $\hat{\tau}_{\text{str}}$.

c) Beräkna stratumvikterna $W_h = N_h/N$ och summera de viktade medelvärdena:

| $h$ | $N_h$ | $W_h$ | $\bar{y}_h$ | $W_h\bar{y}_h$ |
|---|---|---|---|---|
| 1 | 190 | 0.19 | 425 | 80.75 |
| 2 | 410 | 0.41 | 1 620 | 664.2 |
| 3 | 260 | 0.26 | 3 010 | 782.6 |
| 4 | 140 | 0.14 | 5 040 | 705.6 |
| Summa | 1 000 | 1.0 | - | **2 233.15** |

$$\hat{\tau}_{\text{str}} = N\bar{y}_{\text{str}} = N \cdot \sum_{h=1}^{L}[W_h\bar{y}_h] = 1\,000 \cdot 2\,233.15 =$$

$$= \boxed{\mathbf{2\,233\,150}\text{ (ton)}} \quad \text{eller} \quad \boxed{\mathbf{2\,233}\text{ (kiloton)}}$$

NOT: Punktskattningen ovan blir givetvis densamma även med proportionell allokering; detta eftersom stickprovsstorlekarna $n_h$ (allokeringen) inte ingår i beräkningen/definitionen av $\hat{\tau}_{\text{str}}$.

Beräkna sedan standardfelet och felmarginalen:

| $h$ | $n_h$ | $s_h^2$ | $N_h^2$ | $1-\frac{n_h}{N_h}$ | $s_h^2/n_h$ | $N_h^2(1-\frac{n_h}{N_h})\frac{s_h^2}{n_h}$ |
|---|---|---|---|---|---|---|
| 1 | 8 | 176 400 | 36 100 | 0.95790 | 22 050 | 762 489 000 |
| 2 | 28 | 656 100 | 168 100 | 0.93171 | 23 432 | 3 669 942 214 |
| 3 | 33 | 2 016 400 | 67 600 | 0.87308 | 61 103 | 3 606 300 848 |
| 4 | 31 | 6 760 000 | 19 600 | 0.77857 | 218 065 | 3 327 664 516 |
| Summa | 100 | - | - | - | - | **11 366 396 579** |

Varians: $\hat{V}(\hat{\tau}_{\text{str}}) = \sum_{h=1}^{4}\left[N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}\right] = 11\,366\,395\,579$

Standardfel: $SE(\hat{\tau}_{\text{str}}) = \sqrt{\hat{V}(\hat{\tau}_{\text{str}})} = \sqrt{11\,366\,396\,579} = 106\,613$

Felmarginal 90%: $z_{0.05} \cdot SE(\hat{\tau}_{\text{str}}) = 1.6449 \cdot 106\,613 = 175\,368$

90% KI för $\mu_y$: $\boxed{\mathbf{2\,233\,150 \pm 175\,368}\text{ (ton)}}$ eller $\boxed{\mathbf{2\,233 \pm 175}\text{ (kiloton)}}$

NOT: Med proportionell allokering och stickprovsstorlekarna $n_h$ enligt noten i b)-uppgiften ovan och med allting annat lika ovan får man istället

Standardfel: $SE(\hat{\tau}_{\text{str}}) = \sqrt{\hat{V}(\hat{\tau}_{\text{str}})} = \sqrt{15\,958\,629\,000} = 207\,796$

90% KI för $\tau_{\text{str}}$: $\mathbf{2\,233\,150 \pm 207\,796}$ (ton) eller $\mathbf{2\,233 \pm 208}$ (kiloton)

d) Inklusionssannolikhet är sannolikheten att ett givet objekt dras till urvalet. Brukar betecknas med $\pi_k$ där indexet $k$ avser objekt nummer $k$. Inklusionssannolikheterna kan vara lika för alla $k \in U$ som i ett vanligt OSU u.å.; då är $\pi_k = n/N$. Eller så kan de var olika.

I ett stratifierat OSU gäller att man drar ett OSU u.å. från varje stratum med oberoende mellan strata. Detta medför att objekt som tillhör samma stratum har samma inklusionssannolikhet men mellan stratum kan det skilja sig åt beroende på olika stratumstorlekar () och hur allokeringen till stickprovet har gjorts. Låt $\pi_{k\in U_h}$ beteckna inklusionssannolikheten för samtliga objekt i stratum nummer $h$. Då gäller att $\pi_{k\in U_h} = n_h/N_h$:

| $h$ | $N_h$ | $n_h$ | $\pi_{k\in U_h}$ |
|---|---|---|---|
| 1 | 190 | 8 | $8/190 =$ **0.04211** |
| 2 | 410 | 28 | $28/410 =$ **0.06829** |
| 3 | 260 | 33 | $33/260 =$ **0.12692** |
| 4 | 140 | 31 | $31/140 =$ **0.22143** |

NOT 1: Om man allokerar proportionellt mot stratumstorleken $N_h$ får man att samtliga objekt i hela populationen $U$ får samma inklusionssannolikhet oavsett vilket stratum objektet tillhör:

$$n_h = n \cdot \frac{N_h}{N} \implies \pi_{k\in U_h} = \frac{n_h}{N_h} = \frac{n}{N}$$

vilket är inklusionssannolikheten vid ett vanligt OSU u.å.. I detta får fall $\pi_{k\in U_h} = \pi_k = \frac{100}{1000} = 0.1$.

NOT 2: Notera att förhållandet mellan ändlighetskorrektionerna (fpc) i c) ovan och inklusionssannolikheterna kan skrivas $1 - \frac{n_h}{N_h} = 1 - \pi_k$.

e) Anta att man skattar en populationsparameter $\theta$ med $\hat{\theta}$. $\theta$ kan vara någon av parametrarna som vi har behandlat på kursen, $\mu$, $\tau$, $\sigma^2$, $P$ eller $A$, eller någon annan valfri parameter. Eftersom skattningen $\hat{\theta}$ är baserad på ett slumpmässigt draget stickprov så är den en slumpvariabel. Om väntevärdet $E(\hat{\theta})$ är lika med värdet på $\theta$ så säger man att $\hat{\theta}$ är en väntevärdesriktig skattning för $\theta$. Dvs. om

$$Bias(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta = 0$$

så är skattningsmetoden väntevärdesriktig (*unbiased*).

Vi vet att HT-skattningen för $\tau$ är väntevärdesriktig (se F4 sid 7). Skattningarna $\hat{\tau}_h$ inom respektive stratum är HT-skattningar så dessa är väntevärdesriktiga för $\tau_h$. Skattningarna vägs ihop linjärt med $W_h = N_h/N$ som vikter till en skattning $\hat{\tau}_{str}$ för $\tau$. Alltså är $\hat{\tau}_{str}$ väntevärdesriktig för $\tau$. Formellt:

$$\begin{aligned} E(\hat{\tau}_{str}) &= E(W_1\hat{\tau}_1 + W_2\hat{\tau}_2 + W_3\hat{\tau}_3 + W_4\hat{\tau}_4) \\ &= W_1E(\hat{\tau}_1) + W_2E(\hat{\tau}_2) + W_3E(\hat{\tau}_3) + W_4E(\hat{\tau}_4) \\ &= W_1\tau_1 + W_2\tau_2 + W_3\tau_3 + W_4\tau_4 = \tau \ \text{[se F4 sid 22]} \end{aligned}$$

$$\implies Bias(\hat{\tau}_{str}) = E(\hat{\tau}_{str}) - \tau = \tau - \tau = 0 \implies \hat{\tau}_{str} \text{ är väntevärdesriktig.}$$

**Uppgift 4. (20p)**

KOMMENTAR: Följande är förslag på svar, mer än vad som anges här kan nästan säkert sägas.

a) Icke-systematiskt fel:

- Urvalsfel = fel/avvikelse som beror på att stickprovet endast är en delmängd av målpopulationen, ofta ett slumpmässigt fel som går mäta och kontrollera.

Systematiska fel: risk för att det kan uppstå skevheter och icke-representativa stickprov pga. av över- och underrepresentation av olika grupper, typiskt svårt att kvantifiera effekterna.

- Täckningsfel = uppstår när rampopulationen inte är identisk med målpopulationen; objekt som borde finnas i ramen saknas, objekt som inte ingår i målpopulationen är med.
- Bortfallsfel = uppstår då vissa valda objekt avstår från att svara eller inte alls deltar.
- Mätfel = felet som uppstår då respondenten medvetet eller omedvetet svarar fel, ej sanningsenliga data.
- Bearbetningsfel = fel som uppstår pga. av hanteringen av data, tex. felräkningar, fel i formler eller program, mänskliga misstag

b) Se kurslitteratur och föreläsningsanteckningar. (5p)

c) Se föreläsningsanteckningar F9 sid 24-29

d) CASM = *Cognitiv Aspects in Survey Methodology* – en modell som beskriver den psykologiska processen när respondenter ska besvara en fråga. Förenklat:

Förstå frågan ⇒ Finna svar ⇒ Bedöma svaret ⇒ Avge svar

En faktafråga och attitydfrågor. Det räcker om du kan peka på någon väsentlig skillnad. (5p)
Faktafråga – när det finns ett objektivt sett sant svar – ett faktum
Attitydfråga – när det handlar om åsikter

Anta att $X$ är en stokastisk variabel med väntevärde $E(X) = \mu$ och varians $Var(X) = \sigma^2$. Då gäller att väntevärdet och variansen för linjärkombinationen $a + bX$ där $a$ och $b$ är konstanter är

$$E(a + bX) = a + b\mu \qquad Var(a + bX) = b^2\sigma^2 = (b\sigma)^2$$

Se t.ex. undervisningsmaterial från Statistikens grunder.

Enligt definitionen är $\tau_y = N\mu_y$ vilket skattas med $\hat{\tau}_y = N\bar{y}$. Notera att $\tau_y$ och $\hat{\tau}_y$ är linjärkombinationer ($a = 0$ och $b = N$) av $\tau_y$ resp. $\bar{y}$ och därmed är

$$E(\hat{\tau}_y) = E(N\bar{y}) = N \cdot E(\bar{y}) = N\mu_y = \tau_y$$

$$V(\hat{\tau}_y) = N^2 V(\bar{y}) = N^2 \left(\frac{N-n}{N-1}\right)\frac{\sigma_y^2}{n} = \left(\frac{N-n}{N-1}\right)\frac{N^2\sigma_y^2}{n} = \left(\frac{N-n}{N-1}\right)\frac{\sigma_{Ny}^2}{n}$$

Notera att $\sigma_{Ny}^2$ endast är en beteckning för $N^2\sigma_y^2$ som införs för att göra notationen lite enklare.

$$N^2\sigma_y^2 = N^2\frac{\sum_{k\in U}(y_k - \mu_y)^2}{N} = \frac{\sum_{k\in U}N^2(y_k - \mu_y)^2}{N} = \frac{\sum_{k\in U}(Ny_k - N\mu_y)^2}{N}$$
$$= \frac{\sum_{k\in U}(Ny_k - \tau_y)^2}{N} = \sigma_{Ny}^2$$

Kravet är nu att $V(\hat{\tau}_y)$ högst får vara lika med $D_\tau^2$ med avseende på $n$ vilket ger

$$V(\hat{\tau}_y) = \left(\frac{N-n}{N-1}\right)\frac{\sigma_{Ny}^2}{n} \le D_\tau^2 \iff [\text{enligt formelsamlingen}] \iff n \ge \frac{N\sigma_{Ny}^2}{D_\tau^2(N-1) + \sigma_{Ny}^2}$$
$$= [\text{byt ut } \sigma_{Ny}^2 \text{ mot } N^2\sigma_y^2] = \frac{N^3\sigma_y^2}{D_\tau^2(N-1) + N^2\sigma_y^2}$$