Instuderingsfrågor / instuderingsområden Statistisk översiktskurs

Listan är avsedd som en instuderingshjälp när kursmaterialet gås igenom och som tentaförberedelse. Listan följer i stort ordningen på föreläsningarna. Listan täcker inte allt innehåll men ska förhoppningsvis kunna fungera som en instuderingshjälp och täcker de centrala delarna av kursinnehållet (dock inte R-programmeringsdelen, men tentan innehåller inte R-programmeringsfrågor).

Vissa frågor är lite mer av ”räknetyp”, några av dessa kan tas upp på föreläsning 13. Maila lärare om det är någon speciell fråga du vill ska tas upp eller om du tycker någon fråga är oklar.

Frågor av räkneövningstyp har också gjorts på Ö1-Ö3 och nedan finns flera hänvisningar till räkneövningarna. Datorlabbar 2-4 innehåller mindre teoridelar som kan vara relevanta för vissa frågor.

Frågor av typen ”beskriv” behöver inte besvaras med långa svar, det viktiga är få med det centrala i att svara på den ställda frågan.

---

1. Beskriv kort skillnaden mellan deskriptiv statistik och inferens.

2. Ibland refereras till välstrukturerade data, eller “tidy data” på engelska, vad avses med begreppet?

3. Kunna beskriva och exemplifiera skillnaden mellan kategoriska variabler och numeriska variabler, och mellan diskreta och kontinuerliga numeriska variabler, svara på frågor om vilken variabeltyp en viss variabel har (exv. bilmärken på en parkering, antal bilar på olika parkeringar, mängd flingor i ett paket flingor, antal flingpaket, kön, ålder, partitillhörighet, betyg på skala A-F, betyg på skala 1-5, betyg på skala 1-6, en variabel som är 1 om ett land inte har kust och annars 0, osv.)

4. Kunna redogöra för på vilken skalnivå en variabel är, exv. temperatur i grader Celsius; antal barn i olika familjer; variabeln ”färg” (exv. gul, grön, blå); betyg enligt betygsskalan UG - G – VG, osv.

5. Skillnaden mellan simultan fördelning, marginell fördelning och betingad fördelning (definition och exempel finns i labb 2, kapitel 3; även något på F2).

6. Följande tabell visar mobilmärken bland 100 besökare i ett köpcentrum, uppdelat på åldersgrupp:

	Åldersgrupp		
Märke	18-30	31-50	50+
Samsung	10	10	10
IPhone	30	10	0
Casio	0	20	10

A. Ta fram den marginella fördelningen, i procentandelar, för variabeln telefonmärke

B. Ta fram den marginella fördelningen, i andelar, för variabeln åldersgrupp

C. Ta fram fördelningen för telefonmärke betingat på ålder

D. hur stor andel av alla individer som har en Samsung tillhör åldersgruppen 18-30?

7. Göra enklare beskrivningar av fördelningar av data (skev åt höger, skev åt vänster, unimodal/bimodal, etc.) (IMS kap 5, F3).

8. Vilseledande statistik, areaprincipen (F2-F3). Exv. om en graf presenteras och y-axeln inte börjar på noll eller om areaprincipen inte följs i ett cirkel- eller stapeldiagram, vad är problemen?

9. Om en graf visas och information såsom namn och enhet på axlarna eller titel saknas, kunna komplettera grafen. Resonera kring om en graf är vilseledande.

10. Resonera om hur medelvärde och median förhåller sig till varandra om man får en given fördelning plottad (se exv. F2, s. 33).

11. Beskriv vad skillnaden är mellan lägesmått och spridningsmått.

12. Beräkna de lägesmått och spridningsmått vi gick igenom på F3 och Ö1, för exempeltalsekvenser. Om vi frågar om **kvartiler och interkvartilavstånd** kommer vi följa metoden vi använt på F3. Om vi frågar om **varians och standardavvikelse** kommer uppgifterna att vara av liknande omfattning som på Ö1 (ungefär n=3 till n=6).

För medel, median, varians och standardavvikelse kan du enkelt kolla dina svar i R. **Skapa en vektor x** med din sekvens och använd sen mean(), median(), var() och sd(). Om du vill kolla dina svar för kvartiler måste du använda följande uttryck i R: quantile(x, probs = c(0.25, 0.5, 0.75), type = 2)

Här är tre talsekvenser, ta fram typvärde, medel, kvartiler (Q1, Q2, Q3), median (=Q2), kvartilavstånd (IQR) och variationsbredd:

{0, 1, 1, 2, 2, 1, 0, 1} typv: 1, medel=1, Q1: 0.5, Q2: 1, Q3: 1.5, IQR: 1, var-bredd: 2

{0, 1, 1, 2, 18, 1, 0, 1} typv: 1, medel=3, Q1: 0.5, Q2: 1, Q3: 1.5, IQR: 1, var-bredd: 18

{0, 1, 1, 18, 18, 1, 0, 1} typv: 1, medel=5, Q1: 0.5, Q2: 1, Q3: 9.5, IQR: 9, var-bredd: 18

Ta fram varians och standardavvikelse för sekvenserna 1: {-1,0,1}, 2: {-2,0,2} och 3: {-1,0,0,1} (svar: samma som på Ö1.1-Ö1.3)

13. Frågor av följande typ:

- Vi har data för ett antal individer och om de tar en viss medicin eller inte, samt om de är pensionärer eller inte. Ge ett exempel på hur dessa data kan presenteras.

- För 500 företag har vi data på antal anställda. Hur kan denna information presenteras/visualiseras för att vi på ett överskådligt sätt ska få en sammanfattande bild av antal anställda på dessa företag?

14. I en liten kommun med 105 invånare bor 100 personer som tjänar mellan 20 och 50 tusen kronor per månad och 5 personer som tjänar 1 miljon kronor var per månad. Resonera kring vilket mått, medelvärde eller median, som bäst sammanfattar hur mycket de flesta i kommunen tjänar?

15. Kunna redogöra kortfattat för skillnaden mellan en observationsstudie och en experimentell studie, samt vilken av dessa typer som i typfallet är mest lämplig för att studera kausala samband.

16. Vad skiljer tvärsnittsdata och paneldata?

17. Frågor av typen: ”Forskare vid Karolinska institutet och Stockholms Universitet använder data från Socialstyrelsens register. Är dessa data primärdata eller sekundärdata?”

18. Begreppen population och urval/stickprov (bla. F4).

19. Beskriv varför vi (i statistiska studier) vill arbeta med slumpmässiga urval från en population, snarare än andra typer av urval (IMS kap 2, F4, F5, F11).

20. En firma blir ombedd att ta fram ett nationellt medelvärde för andel vuxna individer (18+) som sopsorterar. Firman hittar inte tillförlitliga nationella data och bestämmer sig för att intervjua kommunansvariga, som ska ha tillförlitliga data för sina kommuner. Firman tänker sig att ta fram andelen per kommun, sedan beräkna andelen för landet som helhet, och då ta hänsyn till befolkningstalen per kommun. Firman kollar med en statistiker som anger att ett urval av 50 slumpvis utvalda kommuner krävs för att få fram en tillförlitlig skattning.

Av olika anledningar kommer fysiska kommunbesök krävas och då firman är baserad i Stockholm och har en låg budget tänker sig de ansvariga att jobba avståndsmässigt utåt från Stockholm, med start i de närmaste kommunerna, tills man nått 50 kommuner. Metoden skulle täcka både stora och små kommuner, vilket firman anser är en tillräckligt bra metod.

Ange ett problem med firmans datainsamlingsmetod. Motivera ditt svar.

21. Begreppen övertäckning och undertäckning.

22. Översiktligt kunna skilja på olika typer av urval (IMS kap. 2, F4).

23. Vad är slump, vad är en slumpvariabel, vad är sannolikhet, vad menas med oberoende observationer? (SDM kap. 12, F5)

24. Sannolikhetsberäkningar liknande de vi gjorde på Ö1 (F5, SDM kap. 12, Ö1)

Exempel:

Vad är sannolikheten att få tre sexor på tre tärningskast? (1/216)

Vad är sannolikheten att inte få tre sexor på tre tärningskast? (215/216)

På väg till jobbet får du rött ljus med 50% sannolikhet. Vad är sannolikheten att du får rött ljus minst en gång av fyra? (15/16)

Vad är sannolikheten att du får rött ljus fem dagar i rad? (1/32)

(Det går också bra att svara med decimaltal).

Vilka antaganden bygger dessa beräkningar på?

25. Man behöver inte kunna förklara alla begrepp relaterade till Centrala Gränsvärdessatsen (IMS kap. 13; F5) men det centrala budskapet är viktigt: Fördelningen av medelvärden, från upprepade slumpmässiga urval från en viss population (oberoende observationer), kan (under vissa antaganden) approximeras med en normalfördelning. Detsamma gäller för andelar. Detta är vad som gör att vi kan använda normalfördelningen för att dra slutsatser (göra inferens) gällande populationen, från urval/stickprov, om andelar och medelvärden (om exv. röstandel för ett visst parti i en befolkning (F5-F7), tid för mobilsurfande/dag (F7, Ö2), tid i medel att få hem en pizza, etc.) oberoende av hur data i sig är fördelade.

(Också när vi gör inferens för regression (F8-F10) använder vi samma teori – Centrala Gränsvärdessatsen.)

26. Vad är ”success-failure condition”?

27. Begreppen punktskattning, standardfel, felmarginal, konfidensintervall (F6, se också läsinstruktionen).

28. Ta fram konfidensintervall för andelar resp. för medelvärde av numeriska variabler, för olika konfidensnivåer, med användande av standardnormalfördelningstabellen (F6, flera exempel på Ö2).

29. Tolka konfidensintervall (F6, s. 29).

30. Antag att socialdemokraterna (S) fick 40% av rösterna i ett val. I en korrekt genomförd opinionsundersökning med ett slumpmässigt urval av 1500 röstberättigade, något år senare, har S fått 44%. Du är skeptisk och undrar om andelen som sympatiserar med S verkligen har ökat. Ställ upp en nollhypotes och alternativhypotes (F7, Ö2). Beskriv hur du skulle kunna utföra hypotestestet.

31. På s. 28-30 på föreläsning 6, tänk dig att du använder konfidensnivån 90%. Får du ett smalare eller bredare konfidensintervall? Förklara varför, i ord.

32. Vilket konfidensintervall är bredast, ett 80%-igt eller ett 95%-igt? Förklara varför.

33. F7, s. 6. Gå igenom vilka hypoteser som är ensidiga och tvåsidiga (också Ö2, uppgift 5).

34. Vad menas med signifikansnivå respektive p-värde? (F7)

35. Formulera nollhypotes och alternativhypotes från påståenden av typen: ”Inflationen är lägre än två procent”, ”parti M har över 18% väljarstöd”, ”mobilsurfande är skilt från 3.5 h/dag”.

36. Gäller de hypoteser vi diskuterar i kursen populationen eller urvalet? (F7)

37. Gå igenom och formulera och utför hypotestest för den typ av problem som diskuteras på F7, s. 24-26 och Ö2 uppgift 6-8, inklusive val av signifikansnivå (om denna inte är given), användande av standardnormalfördelningstabellen och begreppen testvariabel, observerat värde (på testvariabeln), kritiskt värde. Beskriva om/när nollhypotesen förkastas.

38. Kommentera utefter spridningsdiagram (liknande de data som plottas på F8, s. 13) om korrelationen är hög eller låg, positiv eller negativ. Kunna beskriva när korrelation är ett lämpligt mått och när korrelation inte är ett lämpligt mått (exempel på F8 och Ö3).

39. Vi såg i ett spridningsdiagram på F8-F9 att det finns ett visst samband mellan lägenhetsstorlek och hyresnivå i Uppsala. Variablerna är korrelerade. Vad händer med värdet på korrelationskoefficienten mellan de två variablerna om du byter enhet på lägenhetsstorlek, från kvadratmeter till (som i USA) kvadratfot? (IMS kap 7, F8)

40. Se figur 7.11 i boken. Vad är problemet med att använda begreppet korrelation, för de variabelsamband som visas?

41. Beskriva skillnaden mellan korrelation och kausalitet och om/hur dessa två koncept förhåller sig till varandra (F8).

42. Vad avses med skensamband (spurious correlation på engelska)?

43. Vad är skillnaden mellan korrelation och regression?

44. När vi för regression pratar om en populationsmodell (populationssamband) respektive en skattad modell, vad är skillnaden?

45. Kunna tolka ett regressionsresultat från enkel linjär regression (koefficientskattning, R2, t-värde, p-värde) med bas i en regressionsutskrift / R-utskrift (F8-F9; inlämningsuppgiften, del 4, fråga 4.4.)

46. Tolka regressionskoefficienter (intercept, lutning) i enkel linjär regression (exempel på F8, s. 6-9) och i multipel linjär regression (exempel lutningskoefficienter F10, s. 13-17). Vad betyder ”ceteris paribus” i detta sammanhang? (F10; inlämningsuppgift, uppgift 4.5).

47. Förklara i ord vad R2 är (F9) är och vad problemet är med att använda (”vanliga”) R2 i multipel linjär regression (F10).

48. Du har skattat en enkel linjär regression mellan y och x och fått fram intercept = 40 och lutningskoefficient 5. Skatta y för x-värdet 4. Skatta y för x-värdet -8. Skatta y för x-värdet 0.

49. Du har skattat en enkel linjär regression mellan y (bensinförbrukning i liter per 100 km) och x (en bils vikt i ton) och fått fram b_0 = interceptet = 3 och lutningskoefficient = b_1 =3. Modellen skattades med data som innehöll bilar som vägde mellan 500 kg och 1500 kg.

- Definiera variabler och skriv upp populationsmodellen (F9)
- Skriv upp den skattade linjens ekvation
- Prediktera förbrukningen för en bil som väger 1000 kg. Glöm inte enheter.
- Du vill prediktera förbrukning för en bil som väger 3 ton. Är modellen lämplig? (kap 7.2.4)

50. Antag att du har skattat en regressionsmodell för sambandet mellan lägenhetsstorlek och hyra. Lägenheterna i datasetet som använts har storlekar mellan 30 och 120 kvadratmeter. Du vill nu använda din skattade modell för att skatta hyran för en lägenhet av storleken 200 kvadratmeter. Finns något problem med att göra en sådan skattning? Förklara. (boken kapitel 7.2.4)

51. I en regression med individlängd i cm (x-variabel) och kaloriintag i kcal (y-variabel), två variabler som du har slumpmässigt utvalda data på från Sveriges vuxna befolkning, har du skattat en lutningskoefficient 10 kcal/cm, med standardfelet 1 kcal/cm. Innan du började analysen formulerade du nollhypotesen att det bland Sveriges vuxna inte finns något samband mellan variablerna och alternativhypotesen att det finns ett samband (positivt eller negativt).

Är det skattade sambandet signifikant på 95%-nivån, dvs har vi en t-statistika (ett t-värde) över 1.96?

Ta fram ett 95%-igt konfidensintervall för skattningen. Glöm inte enheter.

Kan du förkasta nollhypotesen? Formulera ett textsvar.

52. Vad är risken med att skatta en regressionsmodell när modellens antaganden inte stämmer, mer specifikt när det inte finns ett underliggande linjärt förhållande mellan förklaringsvariabel och responsvariabel i populationen? (F10, s. 21; F9, s. 15).

53. Antag att vi vill analysera sambandet mellan två variabler men att sambandet inte är linjärt. Vi vill helst använda linjär regression. Vad skulle vi då kunna göra? (tips: jämför din graf med BNP/capita och barnadödlighet i inlämningsuppgiften, uppgift 4.2, med grafen på F10, s. 22). Se diskussion på F10, s. 21.

54. Avgöra om residualplottar, liknande de på sid 111 i IMS kap 7 (F9, s. 15), signalerar att en linjär regressionsmodell är lämplig eller inte.

55. Ställ upp nollhypotes och alternativhypotes för om en lutningskoefficient i ett populationssamband är noll eller inte. Utför hypotestestet och formulera ett textsvar (F9, F10, Ö3 uppgift 4, 6). Vi kan behöva välja signifikansnivå (om denna inte är given), och också använda standardnormalfördelningstabellen och begreppen testvariabel, observerat värde (på testvariabeln) och kritiskt värde. Beskriva om/när nollhypotesen förkastas.

56. Vad är en indikatorvariabel (dummyvariabel) och hur används sådana variabler i regressionsanalys? Ge ett exempel.

57. Förklara i ord när multikollinearitet föreligger och problem vi kan få när vi har multikollinearitet.

58. Beskriv fördelar och eventuella nackdelar eller utmaningar med att använda registerdata (F11).

59. Beskriv, utan matematiska formler, olika feltyper vi kan ha i statistiska studier. Om det underlättar, använd en exempelstudie som du hittar på. Tips: fem olika huvudtyper diskuterades i samband med undervisningen (F11).

60. På 60- och 70-talet var svarsfrekvensen på olika enkätundersökningar från exempelvis SCB mycket hög. På senare år är svarsfrekvensen betydligt lägre, det finns en markant nedgång i svarsfrekvens. Beskriv några olika orsaker till bortfall. Beskriv även, utifrån statistiska resonemang, ett problem som nedgången har lett till (F11).

61. En svensk tidning inriktad på åldersgruppen 30-50 år skapar en databas med e-postadresser, till individer som själva registrerar sig för att kunna vara med och svara på enkäter om fenomen och frågor som är relevanta för den vuxna befolkningen som helhet, exv. om attityd till husdjur, politik och annat. Beskriv, med bas i föreläsning 11, två problem med det urval som skapas, jämfört med om vuxna individer valts slumpvis från SCBs Register över totalbefolkningen (RTB).

62. Antag att Universitet U:s studenter och lärare gör en lista med alla (svenska) e-postkontakter de har, det blir 10.000 unika e-postadresser. Vi får också en lista med 5.000 svenska adresser från en tidning. Vi mailar de 15.000 adresserna och får 3.000 svar på följande fråga: Vem vill du se som statsminister: A. Moderaternas kandidat, B. Socialdemokraternas kandidat, C. Någon annan.

Beskriv minst två statistiska problem/feltyper i denna undersökning (F11).

63. I samband med undervisningen gick vi igenom sju olika kvalitetskriterier som måste redovisas i all officiell statistik. Även andra statistikprodukter (ex. icke-officiell) och statistikproducenter (ex. privata aktörer) bör redovisa hur det ligger till med dessa.

a) Lista och beskriv kortfattat de sju kriterierna.

b) Välj ut två av kriterierna och resonera kring om det finns en konflikt mellan dem, dvs. en kvalitetsförbättring för den ena kan medföra en kvalitetsförsämring för den andra (F11).