Instuderingsfrågor / instuderingsområden Statistisk översiktskurs

Listan är avsedd som en instuderingshjälp när kursmaterialet gås igenom och som tentaförberedelse. Listan följer i stort ordningen på föreläsningarna.

Vissa frågor är i blått och är lite mer av ”räknetyp”, några av dessa kommer tas upp på föreläsning 11 eller på föreläsning 12. Maila Ulf och Anders om det är någon speciell fråga du vill ska tas upp. Frågor av räkneövningstyp har också gåtts igenom på Ö1-Ö3.

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, F2.)

8. Resonera om hur medelvärde och median förhåller sig till varandra om man får en given fördelning plottad (se exv. F2, s. 42).

9. Beskriv vad skillnaden är mellan lägesmått och spridningsmått.

10. Beräkna de lägesmått och spridningsmått vi gick igenom på F2 och Ö1, för exempeltalsekvenser. Om vi frågar om **kvartiler och interkvartilavstånd** kommer vi följa metoden vi använt på F2. Om vi frågar om **varians och standardavvikelse** kommer uppgifterna att vara av samma omfattning som på Ö1, inte större.

Det kommer finnas en formelsamling med uttrycken på varians och standardavvikelse på tentan.

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)

11. Frågor av följande typ:

- Vi har data för ett antal individer och om de tagit 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?

12. 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?

13. 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.

14. 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.

15. Vad skiljer tvärsnittsdata och paneldata?

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

17. Begreppen population och urval/stickprov.

18. Beskriv varför vi (i statistiska studier) vill arbeta med slumpmässiga urval från en population, snarare än andra typer av urval (IMS kap. 2, F3, F4).

19. Begreppen övertäckning och undertäckning.

20. Översiktligt kunna skilja på olika typer av urval (IMS kap. 2, F3).

21. Vad är slump, vad är en slumpvariabel, vad är sannolikhet, vad menas med oberoende observationer? (SDM kap. 12, F4)

22. Sannolikhetsberäkningar liknande de vi gjorde på Ö1 (F4, 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?

23. Man behöver inte kunna förklara alla begrepp relaterade till Centrala Gränsvärdessatsen (IMS kap. 13; F4) 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. Detta är vad som gör att vi kan använda normalfördelningen för att dra slutsatser (göra inferens), från urval/stickprov, om proportioner och medelvärden (om exv. röstandel för ett visst parti i en befolkning, andel korrekta underskrifter i en ”population av underskrifter” (labb 3), betyg för filmer (labb 3), tid i medel att få hem en pizza, etc.) oberoende av hur data i sig är fördelade.

24. Frågor av typen: ta fram standardnormalvärdet (”z-score”) för datapunkten x_i, om medelvärdet är μ och standardavvikelsen σ.

25. Varför tar vi fram standardnormalvärden? (IMS kap. 13, 16, F4, F5, F6)

26. 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, antingen ”formellt” eller förklarat i ord (F5, F6). Beskriv ett sätt för hur du skulle kunna testa dina hypoteser.

27. Vad är typ I och typ II-fel? Ge ett påhittat exempel från forskningsvärden (F6, del 2.)

28. Antag att du genomför en undersökning. Du gör allt rätt, men din alternativa hypotes är felaktig. Kan din studie resultera i ett typ I-fel? Förklara.

29. Vad är en metastudie?

30. Vad är ”the replication crisis?”

31. Vad menas med statistik signifikans?

32. Vad är poängen med att simulera data med antagandet att nollhypotesen stämmer? Ge ett exempel.

33. Vad menas med p-värde?

34. Hur kan vi använda p-värde i hypotestest?

35. Vad är en punktskattning? Vad är standardfel?

36. Vad är ett konfidensintervall? Hur ska det tolkas? Ge ett exempel

37. Hur går det till när man beräknar konfidensintervall med bootstrap?

38. Vad är ”the success-failure condition”? Vad kan vi göra om kravet inte är uppfyllt?

39. Totalt 1000 slumpmässigt valda vuxna personer får frågan: ”är du ensamstående?” 470 svarar ”ja”. Beräkna ett 95% konfidensintervall.

(På tentan kommer du ha tillgång till formler för standardfel och konfidensintervall (finns bla. på F5, s. 37-38))

40. Kommentera utefter spridningsdiagram (liknande de data som plottas på F7, 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.

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

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

43. Översiktligt förklara, exv. med ett eget ritat spridningsdiagram, vad minsta kvadratmetoden består i (inga uträkningar alls behövs).

44. Vad är skillnaden mellan korrelation och regression?

45. Kunna tolka ett regressionsresultat från enkel linjär regression (koefficientskattning, R2, t-värde, p-värde) med bas i en R-utskrift (IMS, kap 24; F7-F8; inlämningsuppgiften, del 4, fråga 4.4.)

46. Antag att vi vill analysera sambandet mellan två variabler men att sambandet inte är linjärt. Vi vill helst använda linjär regression. Vad kan vi då göra? (tips: jämför graferna med BNP/capita och barnadödlighet i inlämningsuppgiften, uppgift 4.2, med F7, s. 5.)

47. Vad är skillnaden mellan populationsmodell (populationssamband) och en skattad modell?

48. Ställ upp en nollhypotes och en alternativhypotes för om en lutningskoefficient i ett populationssamband, är noll eller inte (F8, IMS kap. 24).

49. Tolka intercept och lutningskoefficient i en enkel linjär regressionsmodell.

50. Avgör om residualplottar, liknande de på sid 111 i IMS (F8, s. 15), signalerar att en enkel linjär regressionsmodell är lämplig eller inte.

51. 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.

52. 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.

- Skriv upp populationsmodellen (F8)

- 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?

53. I en regression mellan 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 2?

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

Kan du förkasta nollhypotesen?

54. Vad är en dummyvariabel och hur används dessa i regressionsanalys? Ge ett exempel.

55. Hur kan man använda dummyvariabler för att skatta effekten av kategoriska variabler med flera nivåer? Ge ett exempel.

56. Beskriv hur man tolkar koefficienter i multipel regression. Vad betyder ”ceteris paribus” i detta sammanhang? (F9; inlämningsuppgift, uppgift 4.5)

57. Beskriv hur justerad förklaringsgrad kan användas för att förbättra en regressionsmodell. Vad är backward elimination?

58. Vad betyder multikollinearitet?

59. Vilka antaganden ligger till grund för multipel linjär regression?

60. Vad är risken med att skatta en sådan modell när modellens antaganden inte alls stämmer?

61. Beskriv fördelar och eventuella nackdelar eller utmaningar med att använda registerdata.

62. Beskriv, utan matematiska formler, olika feltyper vi kan ha i statistiska studier. Om det underlättar, använd en exempelstudie som du hittar på.

63. 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, utifrån statistiska resonemang, ett problem som nedgången har lett till.

64. 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 andra 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.