

Statistik och sannolikhetslära, TNIU66

Exempelsamling, 2019

1. Summasymbolen

1.1. Skriv utan summasymbol:

$$\sum_{k=1}^4 x_k \quad \sum_{k=2}^5 x_{k-1} \quad \sum_{i=1}^3 x_i^2 \quad \sum_{i=1}^5 4$$

Räkna ut summorna om $x_j = 2j - 1$.

1.2. Skriv med summasymbol:

- (a) $x_1 + x_2 + x_3 + x_4$
- (b) $x_1 + x_2 + x_3 + \cdots + x_n$
- (c) $(x_1 + x_2 + x_3 + \cdots + x_n)/n$

1.3. Skriv $x_3 + x_4 + x_5 + x_6$ med summasymbol på två olika sätt, med olika gränser.

1.4. Låt $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ och $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ (dvs de vanliga uttrycken för medelvärde och stickprovsstandardavvikelse). Bestäm s utan summasymbol uttryckt enbart i x_i då $n = 2$. Kommentera resultatet.

2. Deskriptiv statistik

2.1. Rita ett stolpdigram och ett histogram för talen 1, 2, 3, 3, 3, 3, 5, 6, 6, 6. Rita in medelvärde och median i de bägge diagrammen med varsitt streck.

2.2. Rita histogram där datamängden har följande egenskaper:

- (a) medelvärdet är mindre än medianen
- (b) standardavvikelsen är större än medelvärdet
- (c) medelvärdet är större än standardavvikelsen

Totalt skall alltså tre olika histogram ritas.

2.3. Man vill bestämma medellönen för kvinnor respektive män vid ett företag med fyra avdelningar. Man samlar in data per avdelning, och erhåller:

	Antal kvinnor	Medellön	Antal män	Medellön
Avdelning 1	22	26	7	20
Avdelning 2	34	28	14	26
Avdelning 3	14	38	32	32
Avdelning 4	9	46	26	44

Bestäm totala medellönen samt medellönen för kvinnor och män separat. Vilken grupp har högst medellön? Vilken grupp har högst medellön på respektive avdelning? Finns det skäl utifrån denna kartläggning att misstänka lönediskriminering?

- 2.4. Ett sjukhus skall göra ett förnyat inköp av medicin mot högt blodtryck, och låter därför undersöka hur effektiva de två mest populära sorterna har varit under det gångna året. Resultatet framgår av tabellen:

	Antal patienter	Effektiv	Ej effektiv
Medicin A	556	233	323
Medicin B	281	87	194
Summa	837	320	517

Beräkna hur stor andel av patienterna som har blivit bättre efter användande av medicin A respektive av medicin B.

Innan man slutgiltigt bestämmer sig för att köpa in den medicin som störst andel hade nytta av, beslutar man också att se om det är någon skillnad i medicinens verkan på män och kvinnor. Av den orsaken går man tillbaka till samma datamängd, men delar nu upp den även på kön.

För kvinnorna finner man:

	Antal kvinnor	Effektiv	Ej effektiv
Medicin A	185	150	35
Medicin B	18	16	2
Summa	203	166	37

För männen finner man:

	Antal män	Effektiv	Ej effektiv
Medicin A	371	83	288
Medicin B	263	71	192
Summa	634	154	480

Beräkna hur stor andel av kvinnorna som har blivit bättre efter användande av medicin A respektive av medicin B. Beräkna samma sak för de manliga patienterna.

Anser du att sjukhuset bör köpa in medicin A eller medicin B (man har tidigare bestämt sig för att köpa in enbart en av sorterna)? Motivera.

- 2.5.(a) Ge exempel på en mängd med fem tal där medel är större än medianen, och båda är större än noll. Beräkna standardavvikelsen för den mängden.
- (b) Ge exempel på en mängd med fem tal där både medel och standardavvikelse är tio gånger större än i (a)-uppgiften.
- (c) Ge exempel på en mängd med fem tal där medel är tio gånger större än i (a)-uppgiften, men med samma standardavvikelse.
- 2.6. Ge ett eget exempel när medianen är ett mer rättvisande centralmått än medelvärdet. Ge ytterligare ett eget exempel när omvändningen gäller, dvs medelvärdet ger mer användbar information än medianen.
- 2.7. Rania har en stor hög med kulor, och vill ta reda på medelvikten och spridningen i vikt hos kulorna. Hon drar nio kulor slumpmässigt, och finner att dessa har medelvärde 57 gram och stickprovsstandardavvikelse 5,0 gram. För att bli mer säker drar hon ytterligare en kula, som visar sig väga just 57 gram. Vad blir medelvikt och stickprovsstandardavvikelse för de tio kulorna?

- 2.8. En undersökning görs i X-stad och en i Y-stad avseende vilka preferenser folk har för olika former av kollektivtrafik. Av 100 tillfrågade personer i X-stad är det 32 som föredrar spårvagn, 34 föredrar buss och 34 tunnelbana. I Y-stad är det endast 68 personer som svarar, och av dem är 65 för spårvagn, 1 för buss och 2 för tunnelbana. Rita stapeldiagram för de två städerna. Var är spridningen i åsikter störst? Inför, om möjligt, ett sätt att mäta spridningen i åsikter för respektive stad, och beräkna måttet för spridningen inom de två städerna. Motivera annars varför du inte lyckas ordna något sådant mått.

3. Inledande sannolikhetslära

- 3.1. Lisa och Göran äter middag på restaurang *Den Gyllene Moroten*. Restaurangen har på menyn sju förrätter och nio huvudrätter. Då både Lisa och Göran är hungriga beställer de fyra förrätter och fem huvudrätter var.

(a) Beräkna antalet måltidskombinationer som Lisa kan välja mellan då hon äter sina förrätter en i taget¹, och sina huvudrätter en i taget¹. Lisa beställer alltid olika förrätter och huvudrätter under ett besök.

¹Med "en i taget" menas att man tar hänsyn till ordning.

(b) Beräkna antalet måltidskombinationer som Göran kan välja mellan då han äter alla sina förrätter på en gång², och alla sina huvudrätter på en gång². Även Göran beställer alltid olika förrätter och huvudrätter under ett besök.

² Med "alla på en gång" menas att man inte tar hänsyn till ordning.

- 3.2. Den 27 december 1991 inträffade i Gottröra, strax bortanför Arlandas startbana, en osannolikt lyckosam olycka när ett SAS-plan kraschade och bröts av på två ställen.



Otroligt nog överlevde alla personer ombord. Efter olyckan sägs det att en av passagerarna intervjuades, varvid reportern frågade: "Kommer du nu att våga flyga igen?" Svaret blev: "Oja, ännu mer nu. Enligt statistiken händer en sådan här olycka mindre än en gång på miljonen, så när jag nu har varit med om det en gång blir det ännu osannolikare att jag kommer att råka ut för det en gång till. Sannolikheten att råka ut för två olyckor måste ju bli mindre än en på en million millioner."

Avge ett vänligt, men korrekt, svar till passageraren, där du tydligt förklarar vad som är feltänkt i resonemanget.

- 3.3. Betrakta tre helt vanliga, sexsidiga tärningar, kallade A, B och C. Gör nu om dem, så att för tärning A blir sidorna markerade 3, 3, 3, 3, 3 och 6, för tärning B markeras sidorna med 2, 2, 2, 5, 5 och 5, och till sist tärning C får sina sidor markerade med 1, 4, 4, 4, 4 och 4.

Tänk dig ett spel där det gäller att få så högt värde som möjligt. Tre spelare, kallade A, B och C utifrån sin respektive tärning, spelar. De kastar sina tärningar och utfallen jämförs. Visa att:

$$P(A \text{ besegrar } B) > 1/2$$

$$P(B \text{ besegrar } C) > 1/2$$

$$P(C \text{ besegrar } A) > 1/2$$

Finner du resultatet märkligt? Varför (inte)?

- 3.4. Matti skall köra bil från Norrköping till Göteborg en sen höstkväll. Han är orolig för att det skall bli dåligt väder med åtföljande olycksrisk. SMHI har meddelat att sannolikheterna för uppehåll, regn respektive snö är 60%, 39% respektive 1%. Matti har läst att risken för bilolycka på den aktuella vägsträckan är 0,1%, 0,3% respektive 0,9% vid de tre vädertyperna.

- (a) Vilken är risken (=sannolikheten) att han råkar ut för en olycka?
 (b) Vad är sannolikheten, givet att det blir en olycka, att det är snöoväder?

- 3.5. Ett järnverk producerar varje dag 3000 armeringsjärn. En dag är det 38 av järnen som har för dålig hållfasthet (något man självklart inte vet). Ett urval av 100 av de 3000 armeringsjärnen görs slumpmässigt för kvalitetskontroll. Hur sannolikt är det att kontrollurvalet inte innehåller något defekt armeringsjärn?

4. Slumpvariabler

- 4.1. Låt slumpvariabeln X vara antalet driftstörningar nästa vecka i ett nytt datorsystem. Sannolikhetsfördelningen ges av tabellen nedan.

x	0	1	2	3	4
$P(X = x)$	0,45	0,31	0,14	0,07	0,03

Sök sannolikheterna $P(X \leq 2)$ och $P(X \geq 2)$ samt väntevärdet $E(X)$ och standardavvikelsen $SD(X)$.

- 4.2. Låt oss studera kast med två vanliga tärningar. Markera med "U" att *en* tärning får udda antal ögon upp, och med "J" jämt antal ögon. Skriv ned försökets utfallsrum. Låt slumpvariabeln X vara antalet tärningar med udda antal ögon upp. Bestäm värdemängden V_X och sannolikhetsfördelningen för X . Bestäm väntevärde och standardavvikelse för X .
- 4.3. På en tentamen i statistik och sannolikhetslära förekommer en uppgift med sex stycken flervalsfrågor, där var och en skall besvaras med "ja" eller "nej". Varje rätt svar ger en poäng och varje fel svar minus en poäng, dock kan totala poängsumman inte bli mindre än noll. Tentanden Eva har missat att läsa på det avsnittet, och svarar fullständigt slumpmässigt. Tentanden Andreas däremot är väl förberedd, och har sannolikheten 85% att svara rätt på varje enskild fråga. Bägge besvarar samtliga frågor.
- (a) Bestäm sannolikheten att Andreas svarar rätt på exakt fyra deluppgifter.
 (b) Bestäm sannolikheten att Eva får mer än noll poäng på uppgiften.
 (c) Bestäm förväntansvärdet (dvs väntevärdet) för hur många rätt Andreas kommer att ha. Hur många poäng kan han därmed förvänta sig?

4.4. Antalet ambulansutryckningar dagtid i en viss stad antas ha konstant intensitet. Erfarenheten har visat att i den staden är sannolikheten för ingen utryckning under en timme 20%. Dagtid definieras här som att vara mellan kl 06.00 och 18.00.

- (a) Visa att den konstanta intensiteten är cirka 1,6 utryckningar/timme.
- (b) Beräkna det förväntade antalet utryckningar dagtid en vanlig dag.
- (c) Beräkna sannolikheten att det blir som mest två utryckningar dagtid under en dag.
- (d) Beräkna sannolikheten att det mellan kl 12 och 13 blir tre eller flera utryckningar.

4.5. För en likformigt fördelad slumpvariabel med täthetsfunktion

$$f(x) = \begin{cases} 1 & \text{om } 0 \leq x \leq 1 \\ 0 & \text{för övrigt,} \end{cases}$$

bestäm sannolikheten att få en observation inom intervallen:

- (a) $[0, 4, 0, 6]$
- (b) $[0, 9, 1, 1]$
- (c) $[\mu - \sigma/2, \mu + \sigma/2]$
- (d) $[\mu - \sigma, \mu + \sigma]$
- (e) $[\mu - 2\sigma, \mu + 2\sigma]$
- (f) $[\mu - 10\sigma, \mu + 10\sigma]$

Här är $\mu = E(X)$ väntevärdet för X och $\sigma = SD(X)$ standardavvikelsen för X .

4.6. Den kontinuerliga slumpvariabeln X har täthetsfunktionen

$$f(x) = \begin{cases} kx^2 & \text{för } -1 \leq x \leq 2 \\ 0 & \text{för övrigt.} \end{cases}$$

- (a) Bestäm konstanten k så att f blir en täthetsfunktion.
- (b) Bestäm väntevärdet $E(X)$.
- (c) Bestäm standardavvikelsen $SD(X)$.
- (d) Bestäm den betingade sannolikheten $P(X \geq 0 \mid X \leq 1)$.

4.7. Låt X vara en kontinuerlig slumpvariabel med täthetsfunktion

$$f(x) = \begin{cases} 1/2 & \text{om } 0 \leq x \leq 2 \\ 0 & \text{för övrigt.} \end{cases}$$

Bestäm sannolikheterna

- (a) $P(\mu - \sigma \leq X \leq \mu + \sigma)$
- (b) $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$.

Här är $\mu = E(X)$ väntevärdet för X och $\sigma = SD(X)$ standardavvikelsen för X . Vad skiljer den här uppgiften från uppgift 4.5?

4.8. Den kontinuerliga slumpvariabeln X har täthetsfunktionen

$$f(x) = \begin{cases} ke^{-3x} & \text{för } x \geq 0 \\ 0 & \text{för övrigt.} \end{cases}$$

- (a) Bestäm konstanten k så att f blir en täthetsfunktion.
- (b) Bestäm väntevärdet $E(X)$.
- (c) Bestäm standardavvikelsen $SD(X)$.
- (d) Bestäm den betingade sannolikheten $P(X \geq 1 \mid X \leq 2)$.

Ledning/påminnelse: $\int f(x)g(x) dx = [F(x)g(x)] - \int F(x)g'(x) dx$

4.9. Bestäm för en normalfördelad slumpvariabel med väntevärde $\mu = 3$ och standardavvikelse $\sigma = 2$ sannolikheten att en observation faller inom intervallen:

- (a) $[3, 5, 5]$
- (b) $[-10, 3]$
- (c) $[\mu - \sigma, \mu + \sigma]$
- (d) $[\mu - 2\sigma, \mu + 2\sigma]$
- (e) $[\mu - 10\sigma, \mu + 10\sigma]$

Skriv ned täthetsfunktionen till slumpvariabeln och de integraler som beräkningarna i (a) — (e) motsvarar. Jämför med räkningarna i uppgift 4.5.

4.10. Bestäm för en exponentialfördelad¹ slumpvariabel, X , med väntevärde $\mu = 3$, sannolikheten att en av observationerna faller inom intervallet:

- (a) $[3, 5, 5]$
- (b) $[\mu - \sigma/2, \mu + \sigma/2]$
- (c) $[\mu - \sigma, \mu + \sigma]$
- (d) $[\mu - 2\sigma, \mu + 2\sigma]$
- (e) $[\mu - 10\sigma, \mu + 10\sigma]$

Jämför med räkningarna i uppgifterna 4.5 och 4.9.

4.11. Den äldre sortens glödlampor, sådana med glödtråd i sig, har en livslängd som med god approximation kan beskrivas med en exponentialfördelning. I ett hus finns det 50 sådana glödlampor, där var och en av dem har en förväntad livslängd om 1500 timmar. Bestäm sannolikheten att den faktiska medellivslängden för dessa 50 glödlampor överstiger 1540 timmar.

5. Stickprovsteori

5.1. Vid vägning av åtta slumpmässigt utvalda barn i sjuårsåldern registreras följande vikter (i kg): 31, 29, 33, 35, 27, 32, 30, 29. Vi betraktar detta som ett stickprov av alla sjuåringar i Sverige (eftersom mätningen är gjord här). Från tidigare undersökningar vet man att standardavvikelsen är $\sigma = 2,5$ kg. Bestäm sannolikheten att:

- (a) En slumpmässigt vald sjuåring i Sverige väger mer än 35 kg.
- (b) Medelvikten bland sjuåringar i Sverige är mer än 35 kg.

Ange speciellt vilka antaganden du gör om mätdata.

¹Se kompletterande materiel på Lisam och/eller föreläsning B3.

6. Statistik

- 6.1. Vid vägning av åtta slumpmässigt utvalda barn i sjuårsåldern registreras följande vikter (i kg):

31, 29000, 33, 35, 27, 32, 30, 29

Vi betraktar detta som ett stickprov av alla sjuåringar i Sverige (eftersom mätningen är gjord här). Bestäm ett konfidensintervall med konfidensgrad 95% för medelvikten av svenska sjuåringar. Ange speciellt hur du hanterar det extrema värdet 29000 och vilka antaganden du gör om mätdata för övrigt.

- 6.2. En konsultfirma önskar uppskatta maximal tidsåtgång för en viss, väldefinierad typ av projekt. Detta så man kan lämna lägsta möjliga anbud utan att behöva riskera att göra en förlust. Tidigare erfarenhet av liknande projekt har visat att den tid de tagit har varit (i timmar):

10, 12, 11, 8, 13, 12, 10, 13, 12.

Bestäm utifrån dessa mätningar en övre gräns i timmar så att väntevärdet för projekttiden med 95% sannolikhet inte överstiger denna gräns. Ange noggrant alla dina antaganden.

- 6.3. Flygbolaget "Far & Flyg AB" värnar alla sina passagerare lika, och speciellt lika värnar man de som köper biljett till business-class. För att kunna reservera erforderligt antal stolar till dessa resenärer önskar man göra en uppskattning av hur många affärsresenärerna är på varje flygning. Man räknar därför antalet i business-klass varje tisdag för 08.30-avgången, trafikerad med en Boeing 737-900 med plats för 189 passagerare, i tjugo veckor, och finner att man totalt har 900 personer i business-klass. Bestäm ett approximativt 90% konfidensintervall för antalet affärsresenärer på avgången tisdag morgon.

- 6.4. Inför valet hösten 2018 genomfördes en undersökning av vilken fråga väljarkåren uppfattade som viktigast. Man tillfrågade 1048 personer (av de 7 495 936 röstberättigade, enligt Valmyndigheten 2018-09-14), och av dessa angav 20% att invandringen är viktigast, 18% sjukvården, 12% integrationen, 11% lag och ordning, samt 9% vardera på skola respektive miljö.

(a) Bestäm ett 90% konfidensintervall för hur stor andel av väljarna som ansåg att invandringen var den viktigaste frågan.

(b) Det visade sig att opinionsinstitutet som gjort undersökningen hade fuskat, och enbart tillfrågat väljare från Norrköping (med en folkmängd som är 1,4 % av Sveriges, 104 100 röstberättigade). Därmed ansågs inte resultatet kunna spegla hela nationen, men det borde gå att säga något vettigt om väljarkåren i kommunen istället. Bestäm ett 90% konfidensintervall för hur stor andel av väljarna i Norrköping som ansåg att invandringen är den viktigaste frågan.

(c) Är det verkligen så att resultatet inte är giltigt för hela landet? Trots allt var det ett helt slumpmässigt urval, om än med inskränkningen att enbart Norrköpingsbor råkade komma med?

Not: Undersökningen är korrekt återgiven från Demoskop, däremot finns ingen täckning för påståendet i (b).

- 6.5. En kommun i Mellansverige önskar marknadsföra sig som en bra ort för att pendla till Stockholm. Man kontrollerar tågtidtabellen, och noterar att där anges att tågen behöver 58 minuter för turen mellan orten och Stockholm. Man går därför ut i en annonskampanj som hävdar "Mindre än en timme från Stockholms central!".

Politikerna i grannkommunen blir irriterade över detta tilltag, man "vet" ju att tågen alltid är försenade, och låter beställa en undersökning för att se om påståendet om körtid verkligen är sant. Man anlitar en statistik konsult, som kontrollerar 10 tåg för att se verklig körtid. Konsulten finner att genomsnittlig körtid är 62,7 minuter (dvs 62 minuter och 42 sekunder), med en standardavvikelse om 5,0 minuter.

Kan marknadsföringens påstående förkastas utifrån dessa uppgifter på signifikansnivån 5%?

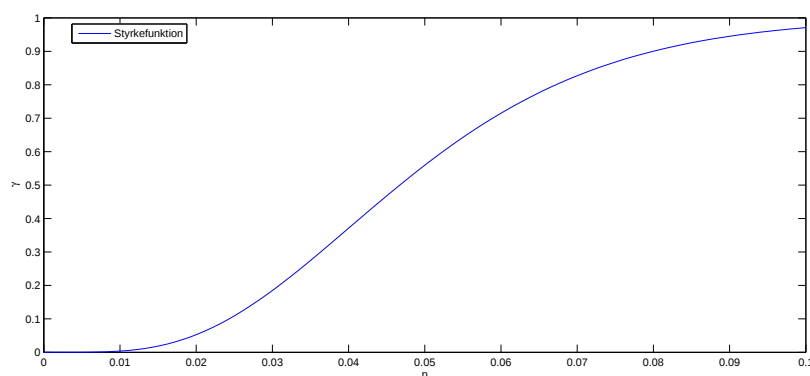
- 6.6. Ett transportföretag garanterar på sin hemsida att minst 98% av deras leveranser sker inom utlovad tid. Två studenter tränade i statistik tycker det låter alltför bra för att vara sant, och avser att undersöka påståendet ordentligt. De sätter därför upp följande hypoteser:

$$H_0 : \quad \text{Andel försenade leveranser } p \leq p_0 = 0,02$$

$$H_a : \quad \text{Andel försenade leveranser } p > p_0 = 0,02$$

Man genomför sedan en hypotesprövning genom att kontakta 100 slumpvist utvalda personer som använt sig av företagets tjänster på senare tid och tillfrågar dem om leveransen kom i tid. Innan man erhöll svaren bestämde man sig för att förkasta företagets påstående, nollhypotesen, om det skulle visa sig att fem eller flera leveranser var försenade (vilket kan tyckas vara generöst, det "borde" ju endast vara två försenade, men för att inte ogrundat anklaga någon för falsk marknadsföring valde man en försiktig linje).

För att undersöka hur bra testet är antar studenterna att företaget har betydligt fler än hundra leveranser, dvs det är verkligen ett stickprov man har tagit. Styrkefunktionen γ beräknas, och blir enligt figuren nedan.



Bestäm följande ur grafen för styrkefunktionen:

- Vilken signifikansnivå, α , har testet?
- Vad är sannolikheten att förkasta en sann nollhypotes, dvs att begå ett fel av typ I (ange även för vilket värde på p ditt svar gäller)?
- Vad är sannolikheten att *inte* förkasta en falsk nollhypotes, dvs att begå ett fel av typ II (ange även för vilket värde på p ditt svar gäller)?
- Skissa hur en ideal styrkefunktion skulle se ut i det här fallet. Hur många observationer behövs i testet för att erhålla en sådan?

- 6.7.(a) Åtta bestämningar av järnhalten i ett malmprov ger följande data (promille):

96	99	103	105	106	110	114	115
----	----	-----	-----	-----	-----	-----	-----

Bilda ett 90% konfidensintervall för väntevärdet av järnhalten. Det är rimligt att anta att varje mätning kommer från slumpvariabler X_i , $i = 1, \dots, 8$, som är normalfördelade $N(\mu, \sigma)$ och oberoende. Väntevärdet μ och standardavvikelsen σ är båda okända.

- (b) Därefter gör man tio nya mätningar från samma malmprov, och finner värdena:

102	99	98	104	97	110	112	98	122	99
-----	----	----	-----	----	-----	-----	----	-----	----

Bestäm ett motsvarande konfidensintervall om 90% som i (a)-uppgiften. Uppenbart är de inte helt lika. Vilket kan man lägga störst tilltro till?

- 6.8. Flygbolaget "Far & Flyg AB" påstår i sin reklam att "mer än 60% av alla resenärer föredrar att resa med oss". I vilken mån har bolaget goda belägg för detta påstående, om det baseras på en undersökning där:

- (a) 7 av 9 (dvs 77,8%) slumpmässigt utvalda resenärer föredrog "Far & Flyg AB"?
(b) 70 av 100 (dvs 70,0%) slumpmässigt utvalda resenärer föredrog "Far & Flyg AB"?

Besvara frågan genom att betrakta nollhypotesen $H_0 : p \leq p_0 = 60\%$, där p är den faktiska andel som föredrar "Far & Flyg AB", och beräkna sannolikheterna

- (a) $P(X \geq 7)$
(b) $P(Y \geq 70)$

under antagandet att H_0 är sann (i gränspunkten $p = p_0$). Här är X och Y de slumpvariabler som ger antalet som föredrar "Far & Flyg AB" i respektive undersökning (innan antalet blev känt).

Tolka resultaten, dvs ange om det är rimligt att förkasta H_0 i något eller båda fallen. Har bolaget i något eller båda fallen gott belägg för alternativhypotesen, den att "mer än 60% av alla resenärer föredrar att resa med oss" är sann?

- 6.9. Vilket eller vilka påståenden är sanna?

- (a) Signifikansnivån, α , vid en hypotesprövning är sannolikheten att godta en felaktig nollhypotes.
(b) Signifikansnivån, α , vid en hypotesprövning är sannolikheten att förkasta en sann nollhypotes.
(c) Sannolikheten för godtagandefel kan beräknas ur styrkefunktionen.
(d) Det så kallade p -värdet anger hur stor andel av nollhypoteserna som inte kan förkastas, givet att hälften av dem är sanna.
(e) Det så kallade p -värdet anger sannolikheten att få ett resultat som är åtminstone lika extremt som det man fick, givet att nollhypotesen är sann.
(f) Det så kallade p -värdet anger sannolikheten för att nollhypotesen falsk.

7. Sambandsanalys

- 7.1. En högskolekurs examineras via två tentamina, istället för en sluttentamen. Ett visst år har studenterna erhållit följande resultat:

Student	Tentamen 1	Tentamen 2
Adam	51	47
Beata	51	50
Ceasar	47	52
Dagny	49	50
Eiman	48	53
Filippa	47	49
Gustaf	49	53
Helga	-1	-1
Igor	46	51
Johanna	54	51
Kimleang	52	54

Examinatorn önskar rationalisera examinationen till nästa år genom att ta bort den ena tentamen; på det sättet skulle mängden undervisning till studenterna kunna ökas och ändå rymmas inom kursens ram. Resultaten är så välkorrelerade att det är onödigt med två skrivningar, tänker examinatorn som har beräknat korrelationskoefficienten. Den som kan den ena delen av kursen kan uppenbart den andra ungefär lika bra.

- Bestäm korrelationskoefficienten mellan de två tentamina för att se om det är sant att resultaten är välkorrelerade. Alla värden skall användas, även de som eventuellt förefaller vara avvikande. Använd gärna din miniräknarens statistikfunktioner för beräkningen.
- En kollega påtalar att Helgas resultat är egendomligt, och en närmare undersökning visar att ”-1” är en kodbeteckning för att hon alls inte deltagit i examinationen (i själva verket hade hon hoppat av kursen, men missat att avregistrera sig). Bortse från hennes resultat, och bestäm vad korrelationskoefficienten nu blir. Använd gärna din miniräknarens statistikfunktioner för beräkningen.
- Är det vettigt utifrån data och resultaten ovan att genomföra den planerade rationaliseringen? Ge examinatorn ett gott råd.

7.2. För att undersöka sambandet mellan materialtätthet och elasticitet för en viss typ av plast bestämdes elasticiteten för nio pressade plaststavar med olika täthet. Resultatet av försöket blev:

Täthet, x_i	1,13	1,15	1,17	1,19	1,21	1,23	1,25	1,27	1,29
Elasticitet, y_i	8,91	10,21	9,64	11,27	11,45	13,61	13,17	14,57	15,04

Dessa data stoppades in i Excel och dess dataanalysverktyg för regression, varvid följande utdata erhöles:

UTDATASAMMANFATTNING						
Regressionsstatistik						
Multipel-R	0,972068857					
R-kvadrat	0,944917862					
Justerad R-kvadrat	0,937048985					
Standardfel	0,553708995					
Observationer	9					
ANOVA						
	fg	KvS	Mkv	F	p-värde för F	
Regression	1	36,81666667	36,81666667	120,082939	1,1668E-05	
Residual	7	2,146155556	0,306593651			
Totalt	8	38,96282222				
	Koefficienter	Standardfel	t-kvot	p-värde	Nedre 95%	Övre 95%
Konstant	-35,40611111	4,328689886	-8,179405788	7,90924E-05	-45,64183619	-25,17038603
X-variabel 1	39,16666667	3,574176192	10,95823613	1,1668E-05	30,71508296	47,61825037

- Rita ett spridningsdiagram för data. Rita in regressionslinjen och gör en första bedömning av rimligheten i en enkel, linjär regressionsmodell.
- Hur stor del av variationen i data förklaras av regressionsmodellen.
- Sök ett 90% konfidensintervall för riktningskoefficienten β_1 baserat på datamaterialet. Förklara med ord vad det här intervallet innebär för det exakta, men okända, värdet på β_1 och vilken sannolikhet det är som är 90%.

7.3. Månadsavgiften för de bostadsrätter som fanns med till salu i lokaltidningen 20 april 2017 var:

En mäklarfirma är intresserad av hur priserna varierar med antal rum och önskar dessutom veta hur pass säkra förutsägelser de kan göra (för att kunna anpassa sin provision). För att åstadkomma detta genomför man en regressionsanalys. Data ovan

Lägenhet	A	B	C	D	E	F	G	H
Rum (antal)	1	1	2	2	3	3	4	4
Månadsavgift (kr)	1800	1160	2130	2650	2530	2850	3100	3780

stoppas in i Excel och dess dataanalysverktyg för regression, med antal rum som bakgrundsvariabel, x , och månadsavgift som responsvariabel, y , varvid följande utdata erhålls:

UTDATASAMMANFATTNING						
Regressionsstatistik						
Multipel-R	0,91450					
R-kvadrat	0,83631					
Justerad R-kvadrat	0,80902					
Standardfel	352,98					
Observationer	8					
ANOVA						
	fg	KvS	Mkv	F	p-värde för F	
Regression	1	3819240	3819240	30,653647	0,001464	
Residual	6	747560	124593,33			
Totalt	7	4566800				
	Koefficienter	Standardfel	t-kvot	p-värde	Nedre 95%	Övre 95%
Konstant	955	305,69	3,12	0,0205	207,01	1702,99
X-variabel 1	618	111,62	5,54	0,0015	344,87	891,13

- Illustrera sambandet mellan avgift och antal rum i ett spridningsdiagram.
- Ange ekvationen för regressionlinjen och rita in den i spridningsdiagrammet.
- Beskriv med ord vad riktningskoefficienten (β_1) betyder i det här sammanhanget.
- Bestäm ett 90%-igt konfidensintervall för vad den förväntade medelavgiften för trerumslägenheter på orten kan vara.
- Beskriv med ord vad detta konfidensintervall innebär, där det framgår vilken sannolikhet det är som är 90%.

7.4. Den lokala fackklubben på ett större företag i Sverige misstänker lönemässig åldersdiskriminering bland personalen, och ber som en pilotstudie därför alla som arbetar som löneadministratörer att rapportera in sin ålder och sin månadslön, den senare uttryckt i tusentals kronor. Resultat blir som följer:

Ålder (år)	32	26	54	21	25	61	33	40
Månadslön (tkr)	21	19	30	18	29	35	24	29500

Dessa data stoppades in i Excel och dess dataanalysverktyg för regression, med ålder som bakgrundsvariabel x och lön som responsvariabel y , varvid nedanstående utdata erhöles.

- Ange ekvationen för regressionlinjen.
- Hur stor del av variationen i data förklaras av regressionsmodellen.
- Verkar det skäligt att ansätta en linjär regressionsmodell?
- Vad borde ha gjorts med data innan analysen utfördes? Gör det, och beräkna nu hur stor andel av variationen som förklaras av regressionslinjen.
- Finns det stöd för misstanken om åldersdiskriminering ur den här undersökningen?

	A	B	C	D	E	F	G
1	UTDATASAMMANFATTNING						
2							
3	Regressionsstatistik						
4	Multipl-R	0,099267116					
5	R-kvadrat	0,00985396					
6	Justerad R-kvadrat	-0,15517038					
7	Standardfel	11200,30391					
8	Observationer	8					
9							
10	ANOVA						
11		<i>fg</i>	<i>KvS</i>	<i>Mkv</i>	<i>F</i>	<i>p-värde för F</i>	
12	Regression	1	7490700,254	7490700,254	0,059712163	0,815093261	
13	Residual	6	752680845,7	125446807,6			
14	Totalt	7	760171546				
15							
16		<i>Koefficienter</i>	<i>Standardfel</i>	<i>t-kvot</i>	<i>p-värde</i>	<i>Nedre 95%</i>	<i>Övre 95%</i>
17	Konstant	1071,471409	11498,97969	0,093179694	0,928794336	-27065,51828	29208,4611
18	X-variabel 1	72,27475593	295,7707559	0,244360724	0,815093261	-651,4502119	795,9997238

7.5. Vilket eller vilka påståenden är sanna?

- (a) Om korrelationskoefficienten $r = -1$ innebär det att alla punkter ligger exakt på en rät linje.
- (b) Om korrelationskoefficienten $r = -1$ innebär det att modellen "förklarar" (dvs beskriver) 100% av variationen i datamängden.
- (c) Om korrelationskoefficienten $r = 0,9$ innebär det att det (med 81% sannolikhet) finns ett orsakssamband mellan x -värdena och y -värdena.
- (d) Regressionslinjens riktningskoefficient β_1 ges av korrelationskoefficienten r i kvadrat.
- (e) Om korrelationskoefficienten $r = 1,4$ innebär det att y -värdena i medeltal växer med ökande x -värden.
- (f) Om korrelationskoefficienten $r = 0$ innebär det att inget samband finns mellan x -värden och y -värden.

7.6. Finns det något samband mellan resultat i skolan och framgångrika studier på universitetet? Forskaren Eva valde ut tio slumpmässiga studenter på ett visst treårigt utbildningsprogram vid ett större universitet i sydöstra Sverige och tog fram studenternas medelbetyg samt meritvärde (dvs medelbetyg plus meritpoäng) från gymnasiet. Dessa uppgifter jämförde hon med hur många högskolepoäng studenterna hade blivit godkända på under första universitetsåret. Resultatet framgår av tabellen.

Medelbetyg	Meritvärde	Poäng på högskolan
14,0	15,0	54
15,5	17,0	36
14,0	14,5	36
15,0	16,0	54
18,5	19,0	60
18,5	18,5	64
12,5	15,0	30
20,0	22,5	60
19,0	21,0	36
12,0	12,0	18

Regressionsanalys utförd med Excel ger för medelbetyg resultat:

	A	B	C	D	E	F	G
1	UTDATASAMMANFATTNING						
2							
3	Regressionsstatistik						
4	Multipl-R	0,679207853					
5	R-kvadrat	0,461323308					
6	Justerad R-kvadrat	0,393988722					
7	Standardfel	12,08673317					
8	Observationer	10					
9							
10	ANOVA						
11		<i>fg</i>	<i>KvS</i>	<i>Mkv</i>	<i>F</i>	<i>p-värde för F</i>	
12	Regression	1	1000,887049	1000,887049	6,85120875	0,030770604	
13	Residual	8	1168,712951	146,0891188			
14	Totalt	9	2169,6				
15							
16		<i>Koefficienter</i>	<i>Standardfel</i>	<i>t-kvot</i>	<i>p-värde</i>	<i>Nedre 95%</i>	<i>Övre 95%</i>
17	Konstant	-13,3230975	22,53227847	-0,5912894	0,57065178	-65,2826248	38,6364299
18	X-variabel 1	3,655540721	1,396587099	2,617481376	0,0307706	0,435005096	6,87607635

Besvara nu följande frågor/uppgifter:

- (a) Vilket resultat (=poängproduktion på högskolan) kan Eva utifrån givna data i medeltal förvänta sig av studenter med i genomsnitt betyget VG från skolan, dvs de som har betygsmedelvärde 15,0? Svara med ett 90 procents konfidsensintervall.
- (b) I regressionsanalysen kan man utläsa att p -värdet för "X-variabel 1" är 0,0307706. Vad betyder det?