Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1912/SF1915 SANNOLIKHETSTEORI OCH STATISTIK, TORSDAG 19 DECEMBER 2019 KL 8.00–13.00.

Examinator SF1912: Camilla Landén, 08-790 6197.
Examinator SF1915: Björn-Olof Skytt, 08-790 8649.

Tillåtna hjälpmedel: Formel- och tabellsamling i Matematisk statistik (utdelas vid tentamen), miniräknare.

Tentamen består av två delar, benämnda del I och del II. Del I består av uppgifterna 1-12. På denna del skall endast svar anges, antingen i form av ett numeriskt värde med tre värdesiffrors noggrannhet eller i form av val av ett av de möjliga svarsalternativen. Studenter som är godkända på kontrollskrivningen behöver ej besvara uppgift 1-3, utan får tillgodoräkna sig dessa tre uppgifter. Studenter som är godkända på datorlaborationen behöver ej besvara uppgift 12, utan får tillgodoräkna sig denna uppgift. Detta gäller vid ordinarie tentamen och vid första omtentamen. Gränsen för godkänt är preliminärt 9 poäng. Möjlighet att komplettera ges för tentander med, preliminärt, 8 poäng.
Del II består av uppgifterna 13-16 och varje korrekt lösning ger 10 poäng. Del II rättas bara för studenter som är godkända på eller får komplettera del I och poäng på del II krävs för högre betyg än E. På denna del skall resonemang och uträkningar skall vara så utförliga och väl motiverade att de är lätta att följa. Införda beteckningar skall förklaras och definieras och numeriska svar skall anges med minst två värdesiffrors noggrannhet. Studenter som är godkända på datorlaborationen får 3 bonuspoäng på del II vid ordinarie tentamenstillfället och det första omtentamenstillfället.

Tentamen kommer att vara rättad inom tre arbetsveckor från skrivningstillfället och kommer att finnas tillgänglig på studentexpeditionen minst sju veckor efter skrivningstillfället.

Del I

Uppgift 1

Antag att den stokastiska variabeln $X \in U(-2, 2)$ och låt $Y = e^X$. Beräkna $P(2 \leq Y < 5)$.

Uppgift 2

Låt A, B och C vara händelser sådana att

$$P(A) = 0.3, \quad P(B) = 0.2, \quad P(A \cap B) = 0.1$$

samt att $P(C|A^* \cap B^*) = 0.2$. Beräkna $P(A \cup B \cup C)$.
Ledning: Börja med att beräkna $P(A \cup B)$. Vad saknas då för att få $P(A \cup B \cup C)$?

A: 0.48

B: 0.60

C: 0.52

D: 0.56

Uppgift 3

Antag att den diskreta stokastiska variabeln X har sannolikhetsfunktionen

$$p_X(1) = \frac{3}{10}, \quad p_X(2) = \frac{3}{10}, \quad p_X(5) = \frac{4}{10}.$$

Bestäm $V(X)$.

A: 1.07

B: 1.15

C: 1.76

D: 3.09

Uppgift 4

Ur ett varuparti om 50 enheter tas ett stickprov på 6 enheter ut på måfå och utan återläggning. Köparen accepterar partiet om högst en enhet i stickprovet klassas som defekt. Vad är sannolikheten att köparen accepterar partiet om felkvoten är 0.10?

A: 0.8970

B: 0.3844

C: 0.8857

D: 0.3543

Uppgift 5

Låt X och Y vara två oberoende normalfördelade stokastiska variabler för vilka det gäller att $E(X) = E(Y) = 4$ och $D(X) = D(Y) = 2$.
Bestäm a så att $P(|X - Y| < a) = 0.99$

A: 3.62

B: 5.54

C: 6.58

D: 7.29

Uppgift 6

Antal meter toalettpapper som förbrukas i ett visst hushåll under en dag kan anses vara en stokastisk variabel med väntevärdet 9 och variansen 5. Antag att hushållet har ett lager på 3200 meter toalettpapper när året börjar. Vad är då den approximativa sannolikheten att toalettpapperet kommer att räcka hela året ut?
Vi antar att året har 365 dagar och att toalettpappersförbrukningen under en dag är oberoende av hur mycket toalettpapper det förbrukas under andra dagar.

Uppgift 7

Antag att X och Y är oberoende stokastiska variabler sådana att $E(X) = E(Y) = \theta$, $V(X) = 3\sigma^2$ och $V(Y) = 2\sigma^2$. Låt x och y vara observationer av X, respektive Y. Två skattningar av θ har föreslagits;

$$\theta^*_{\text{obs}} = \frac{1}{2}(x + y) \quad \text{respektive} \quad \hat{\theta}_{\text{obs}} = \frac{1}{5}(2x + 3y).$$

Vilket av nedanstående påståenden är sant?

A: θ^*_{obs} är den effektivaste skattningen av θ.

B: $\hat{\theta}_{\text{obs}}$ är den effektivaste skattningen av θ.

C: Bägge skattningarna är lika effektiva.

D: Man kan inte avgöra vilken av skattningarna som är effektivast, eftersom minst en av dem inte är väntevärdesriktig.

Uppgift 8

Antag att $X \in Exp(\lambda)$. Vi vill skatta λ med hjälp av Minsta Kvadrat-metoden. Vi gör tre oberoende försök och får stickprovet $x_1 = 18$, $x_2 = 34$, respektive $x_3 = 12$.
Bestäm MK-skattningen av λ utgående från dessa data.

A: 0.0469

B: 0.140

C: 7.11

D: 21.33

Uppgift 9

Antag att $X \in Po(\mu)$ där μ är okänt.
Ange övre gränsen för det ensidigt uppåt begränsade konfidensintervallet för μ som har konfidensgraden 90%. Utgå från följande värden som är framtagna från fyra oberoende försök. $x_1 = 12$, $x_2 = 16$, $x_3 = 19$, $x_4 = 15$

A: 18.02

B: 19.64

C: 20.55

D: 23.80

Uppgift 10

En forskare analyserar $n = 5$ olika prov med två olika mätmetoder A och B och erhåller då n par av värden (x_i, y_i), $i = 1, \ldots, n$. Antag att x_i:na är observationer av $X_i \in N(\mu_i, \sigma_1)$ och y_i:na är observationer av $Y_i \in N(\mu_i + \Delta, \sigma_2)$ och att alla stokastiska variabler är oberoende. Parametern Δ skattas med det numeriska värdet $\bar{y} - \bar{x} = 1.144$. Vad är skattningens medelfel? Observera att parametrarna $\mu_1, \ldots, \mu_5$ samt σ_1 och σ_2 alla är okända.

Prov	1	2	3	4	5
Metod A	1.08	0.75	2.81	4.85	5.90
Metod B	1.68	2.28	4.11	5.99	7.05

A: 0.3425

B: 1.4484

C: 0.1532

D: 2.2901

Uppgift 11

Antag att $X \in \text{ffg}(p)$ och låt $H_0 : p = 0.25$. Vi vill testa H_0 mot alternativet $H_1 : p < 0.25$ och förkastar H_0 om vi observerar ett stort värde x på X. Antag att vi observerat x=5. Bestäm testets P-värde.

A: 0.2500

B: 0.6836

C: 0.3164

D: 0.0791

Uppgift 12

Vid enkel linjär regression gäller följande samband

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

där man observerat paren (x_i, y_i), $i = 1, \ldots, n$ och ε_i betecknar de slumpmässiga felen. Man önskar testa nollhypotesen $H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$ och får P-värdet till 0.042. Vilket av följande påståenden är sant?

A: Vi förkastar H_0 och kan därmed inte förkasta att y_i:na beror av x_i:na på risknivån 1%

B: Vi förkastar H_0 och kan därmed inte förkasta att y_i:na beror av x_i:na på risknivån 5%

C: Vi förkastar H_0 och kan därmed förkasta att y_i:na beror av x_i:na på risknivån 1%

D: Vi förkastar H_0 och kan därmed förkasta att y_i:na beror av x_i:na på risknivån 5%

Del II

Uppgift 13

a) De två händelserna A och C är oberoende. Dessutom är följande givet:

$$P(B) = \frac{1}{4}, \quad P(A \cup C) = \frac{1}{3}, \quad P(C^*|A) = \frac{5}{6}, \quad P(A^* \cap B^*) = \frac{11}{20}.$$

Avgör om A och B är disjunkta händelser. Noggrann motivering krävs. (5 p)

b) Tomten har mycket att göra, så han ger Mary Poppins och Jan Långben uppdraget att sortera julklapparna till alla invånare i Ankeborg. Mary Poppins är oerhört noggrann, så bara 1% av alla julklappar hon sorterar blir felsorterade. Mary Poppins är också mycket snabbare än Jan Långben, så hon sorterar dubbelt så många julklappar som han gör. Jan Långben är utomordentligt slarvig, men trots detta blir 90% av de julklappar som han sorterar rätt sorterade.

Kalle Anka har givetvis oturen att få fel julklapp. Hur stor är sannolikheten att det är Jan Långben som sorterat denna julklapp? (5 p)

Uppgift 14

Observationerna $x_1, \ldots, x_n$ är utfall av oberoende stokastiska variabler $X_1, \ldots, X_n$, alla med täthetsfunktion

$$f_X(x) = (\theta + 1)(1 - x)^\theta, \qquad 0 \leq x \leq 1 \quad \text{och 0 för övrigt.}$$

Här är $\theta \geq 0$ parametern i fördelningen. Tio observationer $x_1, \ldots, x_{10}$

0.5634 0.0888 0.0407 0.2629 0.1652 0.1064 0.1327 0.3357 0.0999 0.0212

sammanfattas av

$$\overline{x} = 0.1817 \qquad \frac{1}{n}\sum_{i=1}^{n} \ln(1 - x_i) = -0.2239 \qquad \frac{1}{n}\sum_{i=1}^{n} \ln(x_i) = -2.0845.$$

Bestäm för dessa observationer Maximum-likelihoodskattningen av θ. (10 p)

Uppgift 15

a) Tabellen visar en del av resultatet av Sifos väljarbarometer från november 2019 i vilken de intervjuade svarade på frågan “Vilket parti skulle du rösta på i riksvalet om det var val idag?”.

	Socialdemokraterna (S)	Sverigedemokraterna (SD)	Totalt antal tillfrågade
Nov 2019	1771	1670	7260

I en kommentar till undersökningen kan man läsa följande påstående “Skillnaden mellan SD och S är precis som i oktober inte signifikant”. Detta är av intresse då Socialdemokratena och Sverigedemorakterna är de två partier som flest säger sig rösta på. Om skillnaden inte är signifikant kan man alltså inte avgöra vilket som är Sveriges största parti utgående från denna undersökning. Visa att påståendet är sant på (approximativa) signifikansnivån 1% om man med skillnaden avser skillnaden i andel som säger sig rösta på partierna och gör det förenklande antagandet att antalet som säger sig rösta på Socialdemokraterna är oberoende av antalet som säger sig rösta på Sverigedemokraterna. (5 p)

b) Tabellen nedan visar hur många personer som i Sifos väljarbarometrar från november 2019 respektive november 2018 sade att de skulle ha röstat på något av januaripartierna, något av partierna i högeroppositionen, respektive något av övriga partier om det vore riksdagsval den dag då de tillfrågades om sina partisympatier.

	Januaripartierna	Högeropposition	Övriga	Totalt
November 2018	3856	3611	738	8205
November 2019	2977	3477	806	7260

Testa på signifikansnivån 1% om den politiska opinionen förblivit oförändrad från november 2018 till november 2019. Den politiska opinionen får här representeras av fördelningen mellan de tre grupperna ovan. (5 p)

Uppgift 16

a) Formulera stora talens lag. För full poäng krävs att precisa förutsättningar och följder angivits. (2 p)

Låt A vara en händelse sådan att $P(A) = p$ där $0 < p < 1$. Låt vidare $U_1, U_2, \ldots$ vara oberoende stokastiska variabler sådana att

$$U_i = \begin{cases} 1 & \text{om } A \text{ inträffar} \\ 0 & \text{om } A \text{ inte inträffar} \end{cases} \qquad i = 1, 2, \ldots$$

b) Tillämpa stora talens lag på de stokastiska variablerna U_i, $i = 1, 2, \ldots$ och ange vad den säger om den relativa frekvensen för händelsen A. Den relativa frekvensen för en händelse A, rf_A, definieras som

$$rf_A = \frac{\text{antalet för } A \text{ gynnsamma utfall}}{\text{antalet slumpförsök}}.$$

(2 p)

c) Tillämpa Tjebysjovs olikhet (finns i formelsamlingen) på

$$X = \frac{1}{n}\sum_{i=1}^{n} U_i$$

för att bestämma ett n sådant att sannolikheten att relativa frekvensen för A avviker mer än 0.01 från p är högst 0.01. (6 p)

Lycka till!

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG TENTAMEN I SF1900/SF1912 SANNOLIKHETSTEORI OCH STATISTIK, TORSDAG 19 DECEMBER 2019 KL 8.00–13.00.

Del I

Rätt rad:

1. 0.229
2. C
3. D
4. A
5. D
6. 0.0233
7. B
8. A
9. A
10. C
11. C
12. B

Kortfattade lösningar:

Uppgift 1

Vi har att

$$P(2 \leq Y < 5) = P(2 \leq e^X < 5) = P(\ln 2 \leq X < \ln 5) = \int_{\ln 2}^{\ln 5} f_X(x)dx$$

Använd nu att $X \in U(-2, 2)$ så fås

$$\int_{\ln 2}^{\ln 5} f_X(x)dx = \int_{\ln 2}^{\ln 5} \frac{1}{4}dx = \frac{1}{4}[x]_{\ln 2}^{\ln 5} = 0.22907$$

Uppgift 2

Med hjälp av additionssatsen för två händelser fås att

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.2 - 0.1 = 0.4.$$

Vidare gäller att

$$P(C|A^* \cap B^*) = \frac{P(C \cap A^* \cap B^*)}{P(A^* \cap B^*)}$$

dvs

$$\begin{aligned} &P(C \cap A^* \cap B^*) = P(C|A^* \cap B^*)P(A^* \cap B^*) = \{\text{De Morgan}\} = \\ &P(C|A^* \cap B^*)[1 - P(A \cup B)] = 0.2 \cdot (1 - 0.4) = 0.12 \end{aligned}$$

Slutligen fås att

$$P(A \cup B \cup C) = P(A \cup B) + P(C \cap A^* \cap B^*) = 0.4 + 0.12 = 0.52.$$

Alternativ lösning:
Med hjälp av additionssatsen för två händelser fås att

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.2 - 0.1 = 0.4.$$

Vidare gäller att

$$1 - P(C|A^* \cap B^*) = P(C^*|A^* \cap B^*) = \frac{P(C^* \cap A^* \cap B^*)}{P(A^* \cap B^*)} = \frac{1 - P(C \cup A \cup B)}{1 - P(A \cup B)}$$

Dvs

$$0.8 = \frac{1 - P(C \cup A \cup B)}{1 - 0.4}$$

Dvs

$$0.8 \cdot 0.6 = 1 - P(C \cup A \cup B)$$

Dvs $\underline{P(A \cup B \cup C) = 0.52}$

Uppgift 3

Vi har att

$$E(X) = \sum_{\text{alla } k} k p_X(k) = 1 \cdot \frac{3}{10} + 2 \cdot \frac{3}{10} + 5 \cdot \frac{4}{10} = \frac{29}{10} = 2.9$$

Vidare gäller att

$$E(X^2) = \sum_{\text{alla } k} k^2 p_X(k) = 1^2 \cdot \frac{3}{10} + 2^2 \cdot \frac{3}{10} + 5^2 \cdot \frac{4}{10} = \frac{115}{10} = 11.5$$

Slutligen fås att

$$V(X) = E(X^2) - E^2(X) = \frac{1150}{100} - \frac{29^2}{100} = \frac{309}{100} = 3.09$$

Uppgift 4

Låt X beteckan antalet felaktiga enheter i stickprovet. Det gäller att $X \in Hyp(50, 6, 0.1)$ och därmed fås att

$$P(X \leq 1) = P(X = 0) + P(X = 1) = \left(\binom{5}{0} \cdot \binom{45}{6} + \binom{5}{1} \cdot \binom{45}{5}\right) / \binom{50}{6} = 0.8970.$$

Uppgift 5

Sätt $Z = X - Y$. Eftersom Z är en linjärkombination av två oberoende normalfördelade stokastiska variabler, så gäller att $Z \in N(\mu_z, \sigma_z)$ där $\mu_z = \mu_x - \mu_y = 4 - 4 = 0$, och $V(Z) = V(X - Y) = V(X) + (-1)^2 \cdot V(Y) = 2^2 + 2^2 = 8 \Rightarrow \sigma_z = \sqrt{8} \Rightarrow Z \in N(0, \sqrt{8})$
Nu fås att

$$P(|Z| < a) = 0.99 \Leftrightarrow P(Z > a) = 0.005$$

Gör om till $N(0, 1)$

$$P(\frac{Z - 0}{\sqrt{8}} > \frac{a - 0}{\sqrt{8}}) = 0.005$$

Dvs.

$$\lambda_{0.005} = \frac{a}{\sqrt{8}} \Rightarrow a = 2.5758 \cdot \sqrt{8} = 7.29$$

Uppgift 6

Låt X_i vara antalet meter som förbrukas dygn i.
X_i:na antas vara likafördelade, oberoende och många.
Detta leder till att Centrala Gränsvärdes-satsen kan tillämpas på summan Y avX_i:na

$$Y = \sum_{i=1}^{365} X_i \sim N(n \cdot \mu_x, \sigma_x \cdot \sqrt{n}) = N(365 \cdot 9, \sqrt{5} \cdot \sqrt{365}) = N(3285, \sqrt{5 \cdot 365})$$

$$\Rightarrow P(Y < 3200) \approx \Phi\left(\frac{3200 - 3285}{\sqrt{5 \cdot 365}}\right) \approx \Phi(-1.99) = 1 - 09767 = 0.0233$$

Uppgift 7

$$\theta^*_{\text{obs}} = \frac{1}{2}(x + y) \quad \text{respektive} \quad \hat{\theta}_{\text{obs}} = \frac{1}{5}(2x + 3y).$$

Det gäller att

$$\begin{aligned}
E\left[\theta^*_{\text{obs}}\right] &= E[\frac{1}{2}(X+Y)] = \frac{1}{2}(E[X]+E[Y]) = \frac{1}{2}(\theta+\theta) = \theta \\
E\left[\hat{\theta}_{\text{obs}}\right] &= E[\frac{1}{5}(2X+3Y)] = \frac{1}{5}(2E[X]+3E[Y]) = \frac{1}{5}(2\theta+3\theta) = \theta \\
\\
V\left(\theta^*_{\text{obs}}\right) &= V(\frac{1}{2}(X+Y)) = \frac{1}{2^2}V(X+Y) = \{\text{oberoende}\} = \frac{1}{4}(V(X)+V(Y)) \\
&= \frac{1}{4}(3\sigma^2+2\sigma^2) = \frac{5\sigma^2}{4} \\
V\left(\hat{\theta}_{\text{obs}}\right) &= V(\frac{1}{5}(2X+3Y) = \frac{1}{5^2}V(2X+3Y) = \{\text{oberoende}\} = \frac{1}{25}(2^2V(X)+3^2V(Y)) \\
&= \frac{1}{25}(4\cdot 3\sigma^2+9\cdot 2\sigma^2) = \frac{6\sigma^2}{5}
\end{aligned}$$

Således är bägge skattningarna väntevärdesriktiga och eftersom $V(\hat{\theta}_{\text{obs}}) < V(\theta^*_{\text{obs}})$ är $\hat{\theta}_{\text{obs}}$ effektivast.

Uppgift 8

Vi bildar Q enligt §9.2.$\Rightarrow$

$$Q(\lambda) = \sum_{i=1}^{3}(x_i - (\mu_i(\lambda))^2 = (18-\frac{1}{\lambda})^2 + (34-\frac{1}{\lambda})^2 + (12-\frac{1}{\lambda})^2$$

$$\frac{dQ}{d\lambda} = 0 \Rightarrow 2\cdot(18-\frac{1}{\lambda})\cdot\frac{1}{(\lambda)^2} + 2\cdot(34-\frac{1}{\lambda})\cdot\frac{1}{(\lambda)^2} + 2\cdot(12-\frac{1}{\lambda})\cdot\frac{1}{(\lambda)^2} = 0$$

$$\Rightarrow 18+34+12 = \frac{3}{\lambda} \Rightarrow \lambda = \frac{3}{64} = 0.0469$$

Uppgift 9

Vi vill använda §12.3 för att kunna bilda ett konfidensintervall med den approximativa konfidensgraden 90% Enligt §12.3 gäller att

$$I_\theta = \theta^*_{obs} \pm D^*_{obs}\cdot\lambda_{\frac{\alpha}{2}}$$

Här är vårt $\theta = \mu$ och $\theta^*_{obs} = \mu^*_{obs} = \bar{x}$. Villkoret för att §12.3 ska kunna användas är att θ^* är approximativt normalfördelat. Detta uppfylls då $\mu^*_{obs} = \bar{x} = 15.5 \geq 15$ vilket är villkoret för att normalapproximation av en poissonfördelning skall vara rimligt.

$$V(\bar{X}) = \frac{\sigma^2}{n} = \{\text{ty Po-fördelning }\} = \frac{\mu}{n} \Rightarrow D^*_{obs} = \sqrt{\frac{\mu^*_{obs}}{n}} = \sqrt{\frac{15.5}{4}}$$

Nu ska vi ha ett uppåt begränsat konfidensintervall, så vi får

$$I_\theta = (-\infty, \theta^*_{obs} + D^*_{obs}\cdot\lambda_\alpha) = (-\infty, \bar{x} + \sqrt{\frac{\bar{x}}{n}}\cdot\lambda_{0.10}) = (-\infty, 15.5 + \sqrt{\frac{15.5}{4}}\cdot 1.2816) = (-\infty, 18.02).$$

Övre gränsen är alltså 18.02.

Uppgift 10

Här har vi stickprov i par och om man bildar differenserna $z_i = y_i - x_i$ så får man

Prov	1	2	3	4	5
Metod A, x_i	1.08	0.75	2.81	4.85	5.90
Metod B, y_i	1.68	2.28	4.11	5.99	7.05
$z_i = y_i - x_i$	0.60	1.53	1.30	1.14	1.15

Då gäller att z_i är observationer av $Z_i \in N(\Delta, \sigma)$, där $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ och vår punktskattning av Δ ges av

$$\Delta^*_{obs} = \bar{x} - \bar{y} = \bar{z}$$

Standardavvikelsen för motsvarande stickprovsvariable ges av

$$D(\Delta^*) = \frac{\sigma}{\sqrt{5}}$$

och medelfelet ges av

$$d(\Delta^*) = \frac{s}{\sqrt{5}} = \frac{\frac{1}{5-1}\sum_{i=1}^{5}(z_i - \bar{z})^2}{\sqrt{5}} = \frac{0.3425}{\sqrt{5}} = 0.1532.$$

Uppgift 11

För P-värdet gäller att det är sannolikheten att få den observation vi fått eller något ännu värre. Här

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - \left[p + p(1-p) + p(1-p)^2 + p(1-p)^3\right] = \{p = 0.25\} = 0.3164.$$

Uppgift 12

Vi kan inte förkasta H_0 när P-värdet är större än risknivån. Detta gör att alternativ A och C faller. Om vi förkastar att $\beta = 0$ så innebär det att vi inte kan förkasta att y_i:na beror av x_i:na. Alltså är B rätt alternativ i detta fall.

Del II

Uppgift 13

a) Från komplementsatsen och det faktum att A och C är oberoende följer att

$$P(C^*|A) = \frac{5}{6} \Rightarrow P(C|A) = \frac{1}{6} \Rightarrow P(C) = \frac{1}{6}.$$

Vidare följer det från additionssatsen och det faktum att A och C är oberoende att

$$P(A \cup C) = P(A) + P(C) - P(A \cap C) = P(A) + P(C) - P(A)P(C).$$

Insättning av kända sannolikheter i denna ekvation ger

$$\frac{1}{3} = P(A) + \frac{1}{6} - P(A) \cdot \frac{1}{6} \Rightarrow P(A) = \frac{1}{5}.$$

Att händelserna A och B är disjunkta är ekvivalent med att $P(A \cap B) = 0$, vilket i sin tur är ekvivalent med att $P(A \cup B) = P(A) + P(B)$. Från de Morgans lag följer att

$$P(A \cup B) = 1 - P(A^* \cap B^*) = 1 - \frac{11}{20} = \frac{9}{20},$$

å ena sidan och

$$P(A) + P(B) = \frac{1}{5} + \frac{1}{4} = \frac{9}{20},$$

å andra sidan. Alltså gäller det att $P(A \cup B) = P(A) + P(B)$.

Svar: <u>Händelserna A och B disjunkta</u>.

b) Låt händelserna vara :
M =Mary Poppins sorterar julklappen.
L =Jan Långben sorterar julklappen.
F =Julklappen blir felsorterad.
Sökt är $P(L|F)$. Bayes sats ger

$$P(L|F) = \frac{P(F|L)P(L)}{P(F|L)P(L) + P(F|M)P(M)} = \frac{\frac{10}{100} \cdot \frac{1}{3}}{\frac{10}{100} \cdot \frac{1}{3} + \frac{1}{100} \cdot \frac{2}{3}} = \frac{10}{12} \approx 0.83$$

Svar: <u>Sannolikheten att Janne Långben har sorterat julklappen är 0.83</u>.

Uppgift 14

Maximum-likelihoodmetodens skattning av θ är det parametervärde som maximerar

$$\begin{aligned}\ln(L(\theta)) &= \ln(f_{X_1,\dots,X_n}(x_1,\dots,x_n)) = \{\text{oberoende}\} = \ln(f_{X_1}(x_1)\cdots f_X(x_n)) \\ &= \sum_{i=1}^{n} \ln(f_{X_i}(x_i)) = \sum_{i=1}^{n} \ln((1+\theta)(1-x_i)^{\theta}) = n\ln(1+\theta) + \theta\sum_{i=1}^{n} \ln(1-x_i).\end{aligned}$$

Derivering med avseende på θ ger

$$\frac{d}{d\theta}\ln(L(\theta)) = \frac{n}{1+\theta} + \sum_{i=1}^{n}\ln(1-x_i),$$

och om vi sätter derivatan lika med noll får vi extremvärdet

$$\frac{n}{1+\theta} + \sum_{i=1}^{n}\ln(1-x_i) = 0 \quad \Rightarrow \quad \theta^*_{ML} = -\frac{n}{\sum_{i=1}^{n}\ln(1-x_i)} - 1.$$

ML-skattningen av θ är alltså

$$\theta^*_{ML} = -\frac{n}{\sum_{i=1}^{n}\ln(1-x_i)} - 1 = -\frac{1}{-0.2239} - 1 = 3.47.$$

Svar: Maximum-likelihoodskattningen av θ är $\theta^*_{ML} = 3.47$.

Uppgift 15

a) Antalet personer som sades sig vilja rösta på S respektive SD i november 2019 kan ses som utfall av $X \in \text{Bin}(7260, p_1)$ respektive $Y \in \text{Bin}(7260, p_2)$, där p_1 och p_2 är dessa partiers verkliga röstandelar. Vi skattar dessa andelar med $(p_1)^*_{obs} = 1771/7260 \approx 0.2439$ respektive $(p_2)^*_{obs} = 1670/7260 \approx 0.2300$ och skattningen av skillnaden i röstandel är

$$(p_1)^*_{obs} - (p_2)^*_{obs} = 0.2439 - 0.2300 = 0.0139.$$

Vi skattar $p_1 - p_2$ med $X/1721 - Y/1649$ som är en stokastisk variabel med varians

$$\frac{V(X)}{7260^2} + \frac{V(Y)}{7260^2} = \frac{p_1(1-p_1)}{7260} + \frac{p_2(1-p_2)}{7260}.$$

Medelfelet får vi om vi ersätter de okända p_1 och p_2 med skattningarna $(p_1)^*_{obs} = 0.2439$ respektive $(p_2)^*_{obs} = 0.2300$. Vi erhåller då medelfelet

$$\begin{aligned} &\sqrt{\frac{(p_1)^*_{obs}(1-(p_1)^*_{obs})}{7260} + \frac{(p_2)^*_{obs}(1-(p_2)^*_{obs})}{1649}} \\ \approx\ &\sqrt{\frac{0.2439\cdot(1-0.2439)}{7260} + \frac{0.2300\cdot(1-0.2300)}{7260}} \approx 0.007057. \end{aligned}$$

$X/7260$ och $Y/7260$ är approximativt normalfördelade $N(p_1, \sqrt{p_1(1-p_1)/7260})$ respektive $N(p_2, \sqrt{p_2(1-p_2)/7260})$, där det är tillåtet att göra normalapproximation ty (t ex) $7260p_1(1-p_1) \approx 7260\cdot 0.2439(1-0.2439) \approx 1339 \geq 10$ med mycket bred marginal. Vi vill beräkna ett (approximativt) 99%-igt konfidensintervall för $p_1 - p_2$ och får med approximativa metoden enligt FS 11.3 intervallet

$$\begin{aligned} &(p_1)^*_{obs} - (p_2)^*_{obs} \pm \lambda_{0.005}\sqrt{\frac{(p_1)^*_{obs}(1-(p_1)^*_{obs})}{7260} + \frac{(p_2)^*_{obs}(1-(p_2)^*_{obs})}{7260}} \\ \approx\ &0.2439 - 0.2300 \pm 2.5758\cdot 0.007057 = 0.0139 \pm 0.01818 = \underline{(-0.004266, 0.03209)}. \end{aligned}$$

Vi testar nollhypotesen $H_0 : p_1 = p_2$ mot mothypotesen $H_1 : p_1 \neq p_2$ på signifikansnivån 1%. Eftersom konfidensintervallet innehåller 0, så kan vi inte förkasta nollhypotesen. Skillnaden mellan Socialdemokrater och Sverigedemokrater är inte signifikant.

b) Vi utför här ett homogenitetstest. Som nollhypotes H_0 väljer vi att röstfördelningen mellan de tre alternativen var densamma i november 2018 som i november 2019, medan mothypotesen H_1 är att röstfördelningen skilde sig mellan de båda tidpunkterna. Vi gör här en tabell med observerade antal enligt

	Januariavtalet	Högeroppositionen	Övriga
November 2018	3856	3611	738
November 2019	2977	3477	806

Låt x_{ij} beteckna antalet personer som faller inom kategori i gällande undersökningens tidpunkt och inom kategori j gällande röstalternativ. Låt vidare n_i beteckna det totala antalet personer som svarade på undersökningen det i:te året och låt m_j beteckna antalet personer som skulle rösta på ett part inom röstalternativ j. Om vi låter N beteckna det totala antalet svar i båda undersökningarna, så blir teststorheten

$$
\begin{aligned}
Q = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(x_{ij} - \frac{n_i m_j}{N})^2}{\frac{n_i m_j}{N}} &= \frac{(3856 - \frac{8205\cdot 6833}{15465})^2}{\frac{8205\cdot 6833}{15465}} + \frac{(3611 - \frac{8205\cdot 7088}{15465})^2}{\frac{8205\cdot 7088}{15465}} + \frac{(738 - \frac{8205\cdot 1544}{15465})^2}{\frac{8205\cdot 1544}{15465}} \\
&\quad + \frac{(2977 - \frac{7260\cdot 6833}{15465})^2}{\frac{7260\cdot 6833}{15465}} + \frac{(3477 - \frac{7260\cdot 7088}{15465})^2}{\frac{7260\cdot 7088}{15465}} + \frac{(806 - \frac{7260\cdot 1544}{15465})^2}{\frac{7260\cdot 1544}{15465}} \\
&= 61.09.
\end{aligned}
$$

Om H_0 är sann så är 61.09 ett utfall av en stokastisk variabel som är approximativt χ^2-fördelad med $(2-1)(3-1) = 2$ frihetsgrader. Approximationen är applicerbar eftersom $n_i m_j/N \geq 7260 \cdot 1544/15465 = 725 > 5$. Eftersom $\chi^2_{0.01}(2) = 9.21 < 61.09$, så kan H_0 förkastas på signifikansnivån 1%. Vi drar följande slutsats.

Svar: På signifikansnivån 1% har röstfördelningen mellan de tre röstalternativen förändrats mellan november 2018 och november 2019.

Uppgift 16

a) Stora talens lag: Låt $X_1, X_2, \dots$ vara oberoende och likafördelade stokastiska variabler som alla har väntevärdet μ och definiera

$$\overline{X}_n = \frac{1}{n}\sum_{i=0}^{n} X_i.$$

Då gäller, för alla $\varepsilon > 0$, att

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0, \quad n \to \infty.$$

b) De stokastiska variablerna $U_1, U_2, \ldots$ är oberoende, likafördelade och har alla väntevärde

$$E(U_i) = 1p + 0(1-p) = p.$$

Låt rf_A beteckna den relativa frekvensen för en händelse A. rf_A ges, per definition av relativ frekvens och av de stokastiska variablerna U_i, av $\overline{U}_n$. Enligt stora talens lag gäller då, för alla $\varepsilon > 0$, att

$$P(|rf_A - p| > \varepsilon) = P(|\overline{U}_n - p| > \varepsilon) \to 0, \quad n \to \infty.$$

Svar: Den relativa frekvensen för en händelse konvergerar (i ovanstående mening) till sannolikheten för händelsen.

c) För de stokastiska variablerna U_i så gäller $E(U_i)^2 = 1^2 p + 0^2(1-p) = p$ och därmed

$$D(U_i) = \sqrt{E(U_i^2) - (E(U_i))^2} = \sqrt{p(1-p)}.$$

Den stokastiska variabeln $\overline{U}_n = \frac{1}{n}\sum_{i=0}^{n} U_i$ är ett aritmetiskt medelvärde av oberoende, likafördelade stokastiska variabler med väntevärde p och standardavvikelse $\sqrt{p(1-p)}$ och uppfyller därmed $E(\overline{U}_n) = p$ och $D(\overline{U}_n) = \sqrt{p(1-p)/n}$. Insättning av $X = \overline{U}_n$, $\mu = p$ och $\sigma = \sqrt{p(1-p)/n}$ i Tjebysjovs olikhet ger nu att

$$P\left(|\overline{U}_n - p| > k\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{k^2}.$$

Eftersom $\overline{U}_n$ är den relativa frekvensen för händelsen A och vi vill att sannolikheten att denna ska avvika mer än 0.01 från p ska vara högst 0.01, så får vi ekvationerna

$$k\sqrt{\frac{p(1-p)}{n}} = \frac{1}{100} \quad \text{och} \quad \frac{1}{k^2} = \frac{1}{100}.$$

Den andra ekvationen ger $k = 10$, vilket efter insättning i den första ekvationen ger

$$\sqrt{n} = 100k\sqrt{p(1-p)} = 1000\sqrt{p(1-p)} \quad \Rightarrow \quad n = 10^6 p(1-p).$$

Eftersom $p(1-p)$ maximalt kan vara $1/4$ för $p \in (0,1)$, så behöver vi välja $n = 250000$.

Svar: För $n = 250000$, så är sannolikheten att den relativa frekvensen för A avviker mer än 0.01 från p högst 0.01.