Sammanfattning av SF1922 Sannolikhetsteori och statistik

Yashar Honarmandi

1 maj 2018

Sammanfattning

Detta är en sammanfattning av viktiga definitioner och satser i kursen SF1922 Sannolikhetsteori och statistik. Den innehåller även information om viktiga sannolikhetsfördelingar.

Den observanta läsaren vill se att jag inte alltid beskriver skillnaden mellan diskreta och kontinuerliga stokastiska variabler. Detta är huvudsakligen ett resultat av latskap. I många fall handlar detta om en summa i det diskreta fallet som blir en integral i det kontinuerliga fallet. Denna generaliseringen anser jag som tillräckligt liten till att jag ej vill spendera energi på att förklara den, och tror att läsaren själv vill kunna se den, isär givet de fall där jag explicit skriver skillnaden. Detta är även en chans för läsaren att bli mer bekväm med det djupa sambandet mellan summor och integraler.

Innehåll

1 Grunläggande koncept inom slump

1.1 Definitioner

Slumpförsök Ett slumpförsök är en experiment där resultatet ej kan avgöras på förhand.

Utfall Ett utfall är resultatet av ett slumpförsök.

Utfallsrum Ett utfallsrum, betecknad Ω, är mängden av alla möjliga utfall för ett givet slumpförsök.

Händelser En händelse är en uppsättning intressanta utfall, alltså en delmängd av utfallsrummet, och betecknas $A, B, C, \ldots$.

Sannolikheter Sannolikheten för en given händelse A uppfyller följande axiom:

- För varje A gäller det att $0 \leq P(A) \leq 1$.
- För hela Ω gäller att $P(\Omega) = 1$.
- Om $A_1, A_2, \ldots$ är en följd av parvis disjunkta händelser så gäller att $P(A_1 \cup A_2 \cup \ldots) = \sum P(A_i)$.

Disjunkta händelser Två händelser A, B är disjunkta, eller parvis oförenliga, om $A \cap B = \emptyset$.

Betingade sannolikheter Sannolikheten $P(B \mid A)$ är sannolikheten för att B händer givet att A har händt, och definieras som

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}.$$

För tre händelser definieras det som

$$P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid (A \cap B))$$

och motsvarande för flere händelser.

Oberoende händelser Två händelser är oberoende om $P(A \cap B) = P(A)P(B)$. Detta generaliseras till tre händelser om

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(A \cap C) &= P(A)P(C), \\ P(B \cap C) &= P(B)P(C), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned}$$

Slumpmässiga fel Ett slumpmässigt fel är en differans mellan ett enkelt mätvärde och ett väntevärde.

Systematiska fel Ett systematiskt fel är en differanse mellan ett väntevärde och ett korrekt värde.

Precision Precision är när många mätningar motsvarar väntevärdet bra.

Noggrannhet Noggrannhet är när många mätningar motsvarar det korrekta värdet bra.

1.2 Satser

de Morgans lagar När man ska hitta komplement till komplicerade mängder, byta alla delmängder med deras komplement och alla unioner ($\cup$) till snitt ($\cap$), och motsatt.

Regler för sannolikhetskalkyl

$$\begin{aligned} P(A^*) &= 1 - P(A), \\ P(B) &= P(B \cap A) + P(B \cap A^*), \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

Bevis Följer från mängdlära.

Lagen om total sannolikhet Låt $H_1, \ldots, H_n$ vara parvis oförenliga och låt $\bigcup_{i=1}^{n} H_i = \Omega$. Då gäller att

$$P(A) = \sum_{i=1}^{n} P(H_i)P(A \mid H_i).$$

Bevis Från mängdlära har man att

$$A = \bigcup_{i=1}^{n} (A \cap H_i).$$

Eftersom alla H_i är parvis oförenliga, följer formeln för sannolikheten direkt.

Bayes' sats Låt $H_1, \ldots, H_n$ vara disjunkta och låt $\bigcup_{i=1}^{n} H_i = \Omega$. Då gäller att

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\sum_j P(H_j)P(A \mid H_j)}.$$

Bevis Direkt konsekvens av lagen om total sannolikhet och definitionen av betingad sannolikhet.

Oberoende händelser där minst en inträffer Låt $A_1, \ldots, A_n$ vara oberoende och $P(A_i) = p_i$. Då ges sannolikheten för att minst en av dessa händer av

$$1 - \prod_{i=1}^{n} (1 - p_i).$$

Bevis Sannolikheten för att inga av de inträffer är

$$\prod_{i=1}^{n} (1 - p_i),$$

och den givna formeln följer direkt.

2 Stokastiska variabler

2.1 Definitioner

Stokastiska variabler En stokastisk variabel är en funktion definierad på ett utfallsrum.

Diskreta stokastiska variabler En stokastisk variabel är diskret om den kan anta ett ändligt eller uppräkneligt oändligt antal värden.

Kontinuerliga stokastiska variabler En stokastisk variabel är kontinuerlig om det finns en funktion f så att

$$P(X \in A) = \int_A f \, \mathrm{d}x \ \ \forall \ A,$$

eller motsvarande i flera variabler.

Sannolikhetsfunktioner Låt X vara en diskret stokastisk variabel. Då definieras sannolikhetsfunktionen som

$$p(k) = P(X = k).$$

Täthetsfunktioner Låt X vara en kontinuerlig stokastisk variabel. Då definieras täthetsfunktionen som en funktion f som uppfyller

$$P(X \in A) = \int_A f \, \mathrm{d}x \ \ \forall\ A,$$
$$f(x) \geq 0 \ \forall\ x,$$
$$\int_\Omega f \, \mathrm{d}x = 1.$$

Sannolikhetsfunktioner i flera variabler Låt (X, Y) vara en diskret stokastisk variabel. Då definieras sannolikhetsfunktionen som

$$p(j, k) = P(X = j, Y = k).$$

Täthetsfunktioner i flera variabler Låt (X, Y) vara en kontinuerlig stokastisk variabel. Då definieras täthetsfunktionen som en funktion f som uppfyller

$$P(X \in A) = \int_A f(x, y) \, \mathrm{d}x \, \mathrm{d}y \ \ \forall\ A,$$
$$f(x, y) \geq 0 \ \forall\ x, y,$$
$$\int_{\mathbb{R}^2} f \, \mathrm{d}x = 1.$$

Fördelningsfunktioner Låt X vara en stokastisk variabel. Funktionen $F : x \to P(X \leq x)$ är fördelningsfunktionen för X.

Fördelningsfunktioner i flera variabler Låt (X, Y) vara en tvådimensionell stokastisk variabel. Funktionen $F_{X,Y} : (x, y) \to P(X \leq x, Y \leq y)$ är den simultana fördelningsfunktionen för (X, Y).

Marginalfördelningar Låt $p_{X,Y}$ vara sannolikhetsfunktionen till den stokastiska variabeln (X, Y). Marginalfördelningnen p_X till X definieras då som

$$p_X(j) = \sum_k p(j, k)$$

i det diskreta fallet och

$$f_X(x) = \int_\mathbb{R} f(x, y) \, \mathrm{d}y$$

i det kontinuerliga fallet. En konsekvens av definitionen i det kontinuerliga fallet är

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x, y).$$

Oberoende stokastiska variabler Variablerna X, Y är oberoende om

$$P(X \in C, Y \in D) = P(X \in C)P(Y \in D)\ \forall\ C, D.$$

Väntevärde Låt X vara en stokastisk variabel med sannolikhetsfunktion p. Då definieras variabelns väntevärde som

$$\mathrm{E}\left(E\right) = \sum kp(k).$$

För en kontinuerlig stokastisk variabel definieras det som

$$\mathrm{E}\left(E\right) = \int_{\mathbb{R}} xf \,\mathrm{d}x\,.$$

Varians Låt X vara en stokastisk variabel med väntevärde μ. Variansen till X, med notationen $\mathrm{V}\left(X\right)$, definieras som

$$\sigma^2 = \mathrm{E}\left((X-\mu)^2\right).$$

Standardavvikelse Låt X vara en stokastisk variabel med varians σ^2. Standardavvikelsen till X, med notationen $\mathrm{D}\left(X\right)$, definieras som

$$\sigma = \sqrt{\sigma^2}.$$

Variationskoefficient Låt X vara en stokastisk variabel med väntevärde μ och standardavvikelse σ^2. Variationskoeficienten till X definieras som

$$R = \frac{\sigma}{\mu}.$$

Kovarians Låt X, Y vara stokastiska variabler med väntevärden μ_X, μ_Y. Då definieras kovariansen mellan dessa som

$$\mathrm{C}\left(X, Y\right) = \mathrm{E}\left((X-\mu_X)(Y-\mu_Y)\right).$$

Okorrellerade variabler X, Y är okorrelerade om $\mathrm{C}\left(X, Y\right) = 0$.

Korrelationskoefficient Låt X, Y vara stokastiska variabler. Då definieras korrelationskoefficienten mellan dessa som

$$\rho(X, Y) = \frac{\mathrm{C}\left(X, Y\right)}{\mathrm{D}\left(X\right)\mathrm{D}\left(Y\right)}.$$

Kvantiler Lösningen till

$$F_X(x) = 1 - \alpha$$

kallas α-kvantilen till X.

Standardiserade stokastiska variabler Låt X vara en stokastisk variabel med väntevärde μ och standardavvikelse σ. Då är $Y = \frac{X-\mu}{\sigma}$ en standardiserad variabel.

2.2 Satser

Fördelningsfunktioners egenskaper Låt F vara en fördelningsfunktion. Då gäller att

- $$F(x) \to \begin{cases} 0, x \to -\infty, \\ 1, x \to \infty. \end{cases}$$
- F är växande (eller icke-avtagande för kontinuerliga stokastiska variabler).
- F är kontinuerlig till höger för varje X.

Omvänt gäller även att alla funktioner som uppfyller dessa egenskaper är fördelningsfunktioner.

Bevis

Fördelningsfunktioner och sannolikheter Låt F vara en fördelningsfunktion för variabeln X. Då gäller att

$$F(b) - F(a) = P(a < X \leq b).$$

Bevis

Fördelningsfunktioner och sannolikhetsfunktioner Låt F och p vara fördelnings- respektiva sannolikhetsfunktionen till en diskret stokastisk variabel X. Då gäller att

$$F(x) = \sum_{j \leq x} p(j),$$

$$p(x) = \begin{cases} F(x), x = 0, \\ F(x) - F(x-1), \text{annars}. \end{cases}$$

En motsvarande relation till första ekvationen gäller även för sannolikhets- och fördelningsfunktioner i flera variabler.

Bevis

Fördelningsfunktioner och täthetsfunktioner Låt F och f vara fördelnings- respektiva täthetsfunktionen till en kontinuerlig stokastisk variabel X och låt f vara kontinuerlig i x. Då gäller att

$$\begin{aligned} F(x) &= \int\limits_{-\infty}^{x} f \, \mathrm{d}u \,, \\ \frac{\mathrm{d}F}{\mathrm{d}x}(x) &= f(x). \end{aligned}$$

Bevis

Fördelningsfunktioner och täthetsfunktioner i flera variabler Låt F och f vara fördelnings- respektiva täthetsfunktionen till en kontinuerlig stokastisk variabel (X, Y) och låt f vara kontinuerlig i (x, y). Då gäller att

$$\begin{aligned} F(x,y) &= \int\limits_{-\infty}^{x} \int\limits_{-\infty}^{y} f(u,v) \, \mathrm{d}u \, \mathrm{d}v \,, \\ \frac{\partial^2 F}{\partial x \partial y}(x,y) &= f(x,y). \end{aligned}$$

Bevis

Normalisering av sannolikhetsfunktioner Låt p vara en sannolikhetsfunktion. Då gäller att

$$\sum p(j) = 1.$$

Bevis

Sannolikhetsfunktioner och sannolikheter Låt p vara en sannolihetsfunktion för den stokastiska variabeln X. Då gäller att

$$P(a \leq X \leq b) = \sum_{i=a}^{b} p(i).$$

Bevis

Funktioner av stokastiska variabler Låt X vara en stokastisk variabel. Då har den stokastiska variabeln $Y = g(X)$ sannolikhetsfuktionen $p_Y(k) = \sum\limits_{g(i)=k} p_X(i)$.

Bevis

Väntevärde för funktioner av stokastiska variabler Låt X vara en stokastisk variabel med sannolikhetsfunktion p_X. Då ges väntevärdet till $g(X)$ av

$$\mathrm{E}\left(g(X)\right) = \sum g(k) p_X(k),$$

med en motsvarande relation i det kontinuerliga fallet och i det flerdimensionella fallet.

Bevis

Förenklad formel för varians Låt X vara en stokastisk variabel med väntevärde μ. Då ges variansen till X av

$$\sigma^2 = \mathrm{E}\left(X^2\right) - \mu^2.$$

Bevis

Förenklad formel för kovarians Låt X, Y vara stokastiska variabler. Då ges kovariansen till dessa av

$$\mathrm{C}\left(X, Y\right) = \mathrm{E}\left(XY\right) - \mathrm{E}\left(X\right) \mathrm{E}\left(Y\right).$$

Bevis

Väntevärde för linjärkombination av variabler

$$\mathrm{E}\left(b + \sum a_i X_i\right) = b + \sum a_i \mathrm{E}\left(X_i\right).$$

Bevis

Varians för linjärkombination av variabler

$$\mathrm{V}\left(b + \sum a_i X_i\right) = \sum a_i^2 \mathrm{V}\left(X_i\right) + \sum_{1 \le j < k} a_j a_k \mathrm{C}\left(X_j, X_k\right).$$

Bevis

Oberoende variabler och funktioner X, Y är oberoende om

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

eller

$$p_{X,Y}(j, k) = p_X(j)p_Y(k)$$

i det diskreta fallet och

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

i det kontinuerliga fallet.

Bevis

Oberoende variabler och väntevärde av produktet Låt X, Y vara oberoende. Då gäller att

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y).$$

Bevis

Oberoende variabler och kovarians Oberoende variabler är okorrelerade.

Bevis

Stora talens lag Låt $X_1, \dots, X_n$ vara likfördelade stokastiska variabler med samma väntevärde μ och inför variabeln $\overline{X} = \frac{1}{n}\sum X_i$. Då gäller att

$$\lim_{n\to\infty} P(\mu - \varepsilon < \overline{X} < \mu + \varepsilon) = 1 \ \forall\ \varepsilon.$$

Bevis

Markovs olikhet Låt Y vara en stokastisk variabel och $a \geq 0, Y \geq 0$. Då gäller att

$$P(Y \geq a)) \leq \frac{\mathrm{E}(Y)}{a}.$$

Bevis

Tjebysjobs olikhet Låt X vara en stokastisk variabel med väntevärde μ och standardavvikelse σ. Då gäller att

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \ \forall\ k > 0.$$

3 Kombinatorik

3.1 Definitioner

Permutationer Permutationerna av k element bland n är antalet sätt man kan "dra" k element från n utan återläggning.

Kombinationer Kombinationerna av k element bland n är antalet sätt man kan "dra" k element från n utan återläggning där ordningen ej spelar någon roll.

3.2 Satser

Multiplikationsprincipen Låt åtgärd 1 kunna utföras på a_1 sätt och åtgärd 2 kunna utföras på a_2 sätt. Då kan båda utföras på $a_1 a_2$ sätt.

Bevis

Dragning med återläggning Dragning av k element ur n med återläggning kan utföras på n^k sätt.

Bevis

Dragning utan återläggning Dragning av k element ur n utan återläggning kan utföras på $n(n-1)\ldots(n-k+1)$ olika sätt.

Bevis

Dragning utan återläggning eller ordning Dragning av k element ur n utan återläggning och där ordning ej spelar någon roll kan utföras på

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

olika sätt.

Bevis

4 Diskreta sannolikhetsfunktioner

Enpunktsfördelningen Enpunktsfördelningen ges av $p(a) = 1$ och $p(x) = 0, x \neq a$.

- Väntevärde: a.
- Varians: 0.

Tvåpunktsfördelningen Tvåpunktsfördelningen ges av $p(a) = p$, $p(b) = 1 - p$ och $p(x) = 0, x \neq a, b$.

- Väntevärde: $b + p(a - b)$.
- Varians: ?.

Likformiga fördelningen Om X antar m olika värden, är $p(x) = \frac{1}{m}$ för dessa värden och 0 annars.

- Väntevärde: ?.
- Varians: ?.

För-första-gången-fördelningen Denna sannolikhetsfördelningen ges av

$$p(k) = (1 - p)^{k-1} p, \ \ k \in \mathbb{N}.$$

Om en stokastisk variabel är fördelat så, skrivs det som $X \in \text{ffg}(p)$.

- Väntevärde: $\frac{1}{p}$.
- Varians: $\frac{1-p}{p^2}$.

Geometrisk fördelning Denna sannolikhetsfördelningen ges av

$$p(k) = (1 - p)^k p.$$

Om en stokastisk variabl är fördelat så, skrivs det som $X \in \text{Ge}(p)$.

- Väntevärde: ?.
- Varians: ?.

Binomisk fördelning Denna sannolikhetsfördelningen ges av

$$p(k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Om en stokastisk variabel är fördelat så, skrivs det som $X \in \text{Bin}(n, p)$.

- Väntevärde: np.
- Varians: $np(1 - p)$.

Hypergeometrisk fördelning Denna sannolikhetsfördelningen ges av

$$p(k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}},$$

med $0 \leq k \leq Np$, $0 \leq n-k \leq N(1-p)$ och $N \geq 2$. Om en stokastisk variabel är fördelat så, skrivs det som $X \in \mathrm{Hyp}(N, n, p)$.

- Väntevärde: np.
- Varians: $\frac{N-n}{N-1}np(1-p)$.

Poissonfördelning Denna sannolikhetsfördelningen ges av

$$p(k) = \frac{\mu^k}{k!}e^{-\mu}.$$

Om en stokastisk variabel är fördelat så, skrivs det som $X \in \mathrm{Po}(\mu)$. Fun fact: Poisson betyder fisk på franska.

- Väntevärde: μ.
- Varians: μ.

4.1 Satser

Två binomiskt fördelade variabler Låt $X \in \mathrm{Bin}(n_1, p), Y \in \mathrm{Bin}(n_2, p)$. Då gäller att $X + Y \in \mathrm{Bin}(n_1 + n_2, p)$.

Bevis

Två Poissonfördelade variabler Låt $X \in \mathrm{Po}(\mu_1), Y \in \mathrm{Po}(\mu_2)$. Då gäller att $X + Y \in \mathrm{Po}(\mu_1 + \mu_2)$.

Bevis

Binomisk approximation av hypergeometriska fördelningen Låt $X \in \mathrm{Hyp}(N, n, p)$. Då är X approximativt $\mathrm{Bin}(n, p)$. Approximationen är typiskt bra om $\frac{n}{N} \leq 0.1$.

Bevis

Poissonapproximation av binomiska fördelningen Låt $X \in \mathrm{Bin}(n, p)$. Då är X approximativt $\mathrm{Po}(np)$. Approximationen är typiskt bra om $p \leq 0.1$.

Bevis

Normalapproximation av binomiska fördelningen Låt $X \in \mathrm{Bin}(n, p)$. Då är X approximativt $\mathrm{N}(np, \sqrt{np(1-p)})$. Approximationen är typiskt bra om $\sqrt{np(1-p)} \geq 10$.

Bevis

Normalapproximation av Poissonfördelningen Låt $X \in \mathrm{Po}(\mu)$. Då är X approximativt $\mathrm{N}(\mu, \sqrt{\mu})$. Approximationen är typiskt bra om $\mu \geq 10$.

Bevis

5 Kontinuerliga sannolikhetsfunktioner

Likformiga fördelningen Denna sannolikhetsfördelning ges av

$$f(x) = \frac{1}{b-a}, \ x \in [a, b].$$

Om en stokastisk variabel är fördelad så, skriver vi $X \in \mathrm{U}(a, b)$.

- Väntevärde: $\frac{b-a}{2}$.
- Varians: $\frac{(b-a)^2}{12}$.

Exponentialfördelningen Denna sannolikhetsfördelning ges av

$$f(x) = \lambda e^{-\lambda x}, \ x > 0.$$

Om en stokastisk variabel är fördelad så, skriver vi $X \in \mathrm{Exp}(\lambda)$.

- Väntevärde: $\frac{1}{\lambda}$.
- Varians: $\frac{1}{\lambda^2}$.

Standardnormalfördelningen En standardiserad normalfördelning har täthetsfunktion

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

och motsvarande fördelningsfunktion Φ. Dessa kommer användas sen utan vidare förklarning. Om en stokastisk variabel är fördelad så, skriver vi $X \in \mathrm{N}(0, 1)$.

- Väntevärde: 0.
- Varians: 1.

Kvantiler i standardnomalfördelningen Vi definierar α-kvantiler för en standardiserad normalfördelat variabel som λ_α så att

$$P(X > \lambda_\alpha) = \alpha.$$

Allmäna normalfördelningen $X \in \mathrm{N}(\mu, \sigma)$ om och endast om $Y = \frac{X-\mu}{\sigma} \in \mathrm{N}(0,1)$. Då gäller:

$$f_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right), F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

- Väntevärde: μ.
- Varians: σ^2.

Asymptotiskt normalfördelade variabler Om Z_n vara en oändlig följd av stokasitska variabler och det finns A_n, B_n så att

$$\lim_{n\to\infty} P\left(a < \frac{Z_n - A_n}{B_n} zb\right) = \Phi(b) - \Phi(a)$$

säjs Z_n vara asymptotiskt normalfördelad. Beteckningen är $Z \in \mathrm{AsN}(A_n, B_n)$.

Kvantiler i t-fördelningen Vi definierar α-kvantiler för en t-fördelat variabel som t_α så att

$$P(X > t_\alpha) = \alpha.$$

5.1 Satser

Standardnormalfördelningens fördelningsfunktion och symmetri Standardormalfördelningens fördelningsfunktion uppfyller

$$\Phi(-x) = 1 - \Phi(x).$$

Bevis

$$\Phi(-x) = \int_{-\infty}^{-x} \phi(x)\,\mathrm{d}x = \int_{\mathbb{R}} \phi(x)\,\mathrm{d}x + \int_{\infty}^{-x} \phi(x)\,\mathrm{d}x$$
$$= 1 + \int_{\infty}^{-x} \phi(x)\,\mathrm{d}x\,.$$

Substituera nu $u = -x$ och få

$$\Phi(-x) = 1 - \int_{-\infty}^{x} \phi(-u)\,\mathrm{d}u$$
$$= 1 - \int_{-\infty}^{x} \phi(u)\,\mathrm{d}u$$
$$= 1 - \Phi(x).$$

Eftersom ϕ är symmetrisk kring 0,

Linjärkombinationer av normalfördelade variabler Låt $X_1, \ldots, X_n$ vara oberoende och normalfördelade med väntevärde μ_i och varians σ_i^2. Då gäller att:

$$\sum a_i X_i + b \in \mathrm{N}\left(\sum a_i \mu_i + b, \sqrt{\sum a_i^2 \sigma_i^2}\right).$$

Bevis

Fördelning av medelvärde Låt $\bar{X} = \frac{1}{n} \sum X_i$ för oberoende och likafördelade X_i med väntevärde μ och standardavvikelse σ. Då gäller att $\bar{X} \in \mathrm{AsN}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Fördelning av kvadrat Låt $\bar{X} = \frac{1}{n} \sum X_i$ för oberoende och likafördelade X_i med väntevärde μ och standardavvikelse σ. Då gäller att $\bar{X}$ och $\sum(X_i - \bar{X})^2$ är oberoende stokastiska variabler och att $\frac{1}{\sigma^2} \sum(X_i - \bar{X})^2 \in \chi^2(n-1)$.

Fördelning av skattad standardisering Låt X_i vara oberoende och likafördelade X_i med väntevärde μ, och skriv $X_i = \mu + \sigma\varepsilon_i$. Då gäller att

$$\frac{\overline{X} - \mu}{\frac{\sqrt{\frac{1}{n-1}\sum(X_i - \overline{X})^2}}{\sqrt{n}}} = \frac{\overline{\varepsilon}}{\frac{\sqrt{\frac{1}{n-1}\sum(\varepsilon_i - \overline{\varepsilon})^2}}{\sqrt{n}}} \in t(n-1),$$

där $t(n-1)$ är t-fördenlingen med $n-1$ frihetsgrader.

Bevis

Centrala gränsvärdesatsen Låt $X_1, \ldots, X_n$ vara oberoende, likafördelade stokastiska variabler med väntevärde μ och standardavvikelse σ. Då uppfyller $Y_n = \sum X_i$

$$\lim_{n\to\infty} P\left(a < \frac{Y_n - n\mu}{\sigma\sqrt{n}} < b\right) = \Phi(b) - \Phi(a).$$

6 Deskriptiv statistik

Definitionerna som dyker upp i denna del kan virka redundanta, men det är underförstått att detta är punktskattningar av parametrar och inte själva parametrarna som definieras här.

6.1 Definitioner

Punktskattningar En punktskattning av en parameter θ är en funktion av utfallen $x_1, \ldots, x_n$ av de stokastiska variablerna $X_1, \ldots, X_n$ vars fördelning beror av θ. Därmed är punktskattningen ett utfall av stickprovsvariabeln θ^*.

Väntevärdesriktighet En punktskattning är väntevärdesriktig om $\mathrm{E}\left(\theta^*\right) = \theta$.

Konsistens Punktskattningen θ^* är konsistent om det för varje θ och $\varepsilon > 0$ gäller att

$$\lim_{n \to \infty} P(|\theta_n^* - \theta| > \varepsilon) = 0.$$

Medelkvadratfel Medelkvadratfelet definieras som $\mathrm{E}\left((\theta^* - \theta)^2\right)$.

Medelfel Medelfelet definieras som en skattning av $\mathrm{D}\left(\theta^*\right)$, och betecknas $\mathrm{d}(\theta^*)$.

Effektivitet Om två skattningar $\theta^*, \hat{\theta}$ uppfyller $\mathrm{V}\left(\theta^*\right) \leq \mathrm{V}\left(\hat{\theta}\right)$ är θ^* effektivare än $\hat{\theta}$.

Medelvärde Medelvärdet definieras som

$$\bar{x} = \frac{1}{n} \sum x_i.$$

Varians Variansen definieras som

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2,$$

med en analog definition av standardavvikelsen s.

Kovarians Kovariansen definieras som

$$c_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}).$$

Korrelationskoefficient Korrelationskoefficienten definieras som

$$r = \frac{c_{xy}}{s_x s_y}.$$

Konfidensintervall Intervallet I_θ som med sannolikhet $1-\alpha$ täcker över den okända parametern θ kallas konfidensintervallet för θ med konfidensgrad $1-\alpha$.

6.2 Satser

Medelvärdets egenskaper Medelvärdet är en konsistent och väntevärdesriktig skattning av en stokastisk variabels väntevärde.

Bevis Väntevärdesriktigheten följer direkt från väntevärdets egenskaper.

Variansens egenskaper Variansen är en konsistent och väntevärdesriktig skattning av en stokastisk variabels varians.

Bevis

Konfidensintervall för väntevärde, känd varians Låt $X_1, \ldots, X_n$ vara normalfördelade med väntevärde μ och varians σ. Då är

$$I_\mu = \left[\bar{x} - \lambda_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + \lambda_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

ett konfidensintervall för väntevärdet med konfidensgrad $1-\alpha$, där $\lambda_{\frac{\alpha}{2}}$ är $\frac{\alpha}{2}$-kvantilen i normalfördelningen.

Bevis Vi har att $\overline{X} \in \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. Detta betyder att

$$P\left(-\lambda_{-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \lambda_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Från detta får vi två olikheter som ger det givna konfidensintervallet.

Konfidensintervall för väntevärde, okänd varians Låt $X_1, \ldots, X_n$ vara normalfördelade med väntevärde μ. Då är

$$I_\mu = \left[\bar{x} - t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}\right]$$

ett konfidensintervall för väntevärdet, där där $t_{\frac{\alpha}{2}}(n+1)$ är $\frac{\alpha}{2}$-kvantilen i t-fördelningen med $n-1$ frihetsgrader.

Bevis Vi har att

$$\frac{\overline{X}-\mu}{\frac{\sqrt{\frac{1}{n-1}\sum(X_i-\overline{X})^2}}{\sqrt{n}}} \in t(n-1).$$

Därmed har vi

$$P\left(-t_{-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sqrt{s^2}}{\sqrt{n}}} \in t(n-1) < t_{\frac{\alpha}{2}}\right) = 1-\alpha,$$

där vi använder beteckningen s^2 för skattningen av standardavvikelsen och använder t-fördelningens symmetri. Detta ger två olikheter som ger det givna konfidensintervallet.

Konfidensintervall för standardavvikelse, okänd medelvärde Låt $X_1, \ldots, X_n$ vara normalfördelade med standardavvikelse σ. Då är

$$I_\mu = \left[\sqrt{\frac{n-1}{\chi^2_{\frac{\alpha}{2}}(n-1)}}s, \sqrt{\frac{n-1}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}}s\right]$$

ett konfidensintervall för väntevärdet, där $\chi^2_{\frac{\alpha}{2}}(n-1)$ är $\frac{\alpha}{2}$-kvantilen i χ^2-fördelningen med $n-1$ frihetsgrader.

För stora n kan man skriva intervallen som

$$\left[1-\frac{\lambda_{\frac{\alpha}{2}}}{\sqrt{2(n-1)}}, 1+\frac{\lambda_{\frac{\alpha}{2}}}{\sqrt{2(n-1)}}\right]$$

Bevis

Konfidensintervall för differans mellan väntevärden för olika objekt Låt $X_1, \ldots, X_{n_1} \in N(\mu_1, \sigma_1)$ och $Y_1, \ldots, Y_{n_2} \in N(\mu_2, \sigma_2)$, där dessa kan betraktas som stickprov från två olika objekt. Då gäller att:

- Om σ_1, σ_2 är kända är

$$\left[\bar{x}-\bar{y}-\lambda_{\frac{\alpha}{2}}D, \bar{x}-\bar{y}+\lambda_{\frac{\alpha}{2}}D\right]$$

  ett konfidensintervall för $\mu_1-\mu_2$ med konfidensgrad $1-\alpha$, där $D = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$.

- Om $\sigma_1 = \sigma_2 = \sigma$ är okända är

$$\left[\bar{x}-\bar{y}-t_{\frac{\alpha}{2}}(f)d, \bar{x}-\bar{y}+t_{\frac{\alpha}{2}}(f)d\right]$$

  ett konfidensintervall för $\mu_1-\mu_2$ med konfidensgrad $1-\alpha$, där $d = s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ och $f = n_1+n_2-2$.

Bevis

Konfidensintervall för differans mellan väntevärden för och efter Låt $X_1, \ldots, X_{n_1} \in N(\mu_1, \sigma_1)$ och $Y_1, \ldots, Y_{n_2} \in N(\mu_2, \sigma_2)$, där samma i motsvarar stickprov från två olika objekt. Då gäller att:

- Om vi definierar $Z_i = X_i - Y_i$, är

  $$I_\mu = \left[\bar{z} - t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}, \bar{z} + t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}}\right]$$

  ett konfidensintervall för $\mu_1 - \mu_2$, där s är skattningen av standardavvikelsen från de olika Z_i.

- Om σ_1, σ_2 är okända är

  $$\left[\bar{x} - \bar{y} - \lambda_{\frac{\alpha}{2}} d, \bar{x} - \bar{y} + \lambda_{\frac{\alpha}{2}} d\right]$$

  ett konfidensintervall för $\mu_1 - \mu_2$ med approximativ konfidensgrad $1-\alpha$, där $d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

Bevis

Allmän skattning av normalfördelade stokastiska variabler Låt skattningen av en parameter θ vara normalfördelad med väntevärde θ och standardavvikelse D. Då beräknas konfidensintervall med approximativ konfidensgrad $1 - \alpha$ som

- $$\left[\theta * - \lambda_{\frac{\alpha}{2}} D, \theta * + \lambda_{\frac{\alpha}{2}} D\right]$$

  om D ej beror av θ.

- $$\left[\theta * - \lambda_{\frac{\alpha}{2}} d, \theta * + \lambda_{\frac{\alpha}{2}} d\right]$$

  om d beror av θ, för något lämpligt val av d.

Bevis

Felförplantning Givet medelfelet till någon skattning av en parameter θ, önskar vi nu att estimera medelfelet och väntevärdet av skattningen av någon funktion av θ. Vi skriver denna som $\psi = g(\theta)$.

Första satsen vi har säjer att om θ^* är en approximativt väntevärdesriktig skattning av θ med medelfel $d(\theta^*)$, är $\psi^* = g(\theta^*)$ en approximativt väntevärdesriktig skattning av $\psi = g(\theta)$, eventuellt med korrektionstermen $\frac{1}{2}d^2(\theta^*)\frac{\mathrm{d}^2 g}{\mathrm{d}\theta^{*2}}(\theta^*)$. Dens medelfel ges av

$$d(\psi^*) \approx \left|\frac{\mathrm{d}g}{\mathrm{d}\theta^*}(\theta^*)\right| d(\theta^*).$$

I fallet där ψ^* beror av två variabler θ^* och η^*, gäller ett motsvarande kriterie. Om kritereiet uppfylls, ges väntevärdet på motsvarande vis och medelfelet ges då av

$$d^2(\psi^*) \approx \left(\frac{\partial^2 g}{\partial\theta^{*2}}(\theta^*, \eta^*)\right)^2 d^2(\theta^*) + \left(\frac{\partial^2 g}{\partial\eta^{*2}}(\theta^*, \eta^*)\right)^2 d^2(\eta^*).$$

Bevis