

# Sammanfattning av SF1922 Sannolikhetssteori och statistik

Yashar Honarmandi

11 maj 2018

## Sammanfattning

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

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

<b>1</b>	<b>Grunläggande koncept inom slump</b>	<b>1</b>
1.1	Definitioner . . . . .	1
1.2	Satser . . . . .	2
<b>2</b>	<b>Stokastiska variabler</b>	<b>3</b>
2.1	Definitioner . . . . .	3
2.2	Satser . . . . .	6
<b>3</b>	<b>Kombinatorik</b>	<b>10</b>
3.1	Definitioner . . . . .	10
3.2	Satser . . . . .	10
<b>4</b>	<b>Diskreta sannolikhetsfunktioner</b>	<b>10</b>
4.1	Satser . . . . .	12
<b>5</b>	<b>Kontinuerliga sannolikhetsfunktioner</b>	<b>13</b>
5.1	Satser . . . . .	14
<b>6</b>	<b>Deskriptiv statistik</b>	<b>16</b>
6.1	Definitioner . . . . .	16
6.2	Satser . . . . .	17
<b>7</b>	<b>Hypotesprövning</b>	<b>20</b>
7.1	$\chi^2$ -tester . . . . .	21

# 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  $\Omega$ , 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  $A, B, C, \dots$ .

**Sannolikheter** Sannolikheten för en given händelse  $A$  uppfyller följande axiom:

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

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

**Betingade sannolikheter** Sannolikheten  $P(B | A)$  är sannolikheten för att  $B$  händer givet att  $A$  har hänt, och definieras som

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

För tre händelser definieras det som

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

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 differans 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  $H_1, \dots, H_n$  vara parvis oförenliga och låt  $\bigcup_{i=1}^n H_i = \Omega$ . Då gäller att

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

**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  $H_1, \dots, H_n$  vara disjunkta och låt  $\bigcup_{i=1}^n H_i = \Omega$ . Då gäller att

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

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

**Oberoende händelser där minst en inträffar** Låt  $A_1, \dots, A_n$  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 inträffar 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 uppräknligt oändligt antal värden.

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

$$P(X \in A) = \int_A f \, dx \quad \forall A,$$

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

$$\begin{aligned} P(X \in A) &= \int_A f \, dx \quad \forall A, \\ f(x) &\geq 0 \quad \forall x, \\ \int_{\Omega} f \, dx &= 1. \end{aligned}$$

**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

$$\begin{aligned} P(X \in A) &= \int_A f(x, y) \, dx \, dy \quad \forall A, \\ f(x, y) &\geq 0 \quad \forall x, y, \\ \int_{\mathbb{R}^2} f \, dx &= 1. \end{aligned}$$

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

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

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

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

i det diskreta fallet och

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

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

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

**Oberoende stokastiska variabler** Variablerna  $X, Y$  är oberoende om

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

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

$$E(X) = \sum k p(k).$$

För en kontinuerlig stokastisk variabel definieras det som

$$E(X) = \int_{\mathbb{R}} x f(x) dx.$$

**Varians** Låt  $X$  vara en stokastisk variabel med väntevärde  $\mu$ . Variansen till  $X$ , med notationen  $V(X)$ , definieras som

$$\sigma^2 = E((X - \mu)^2).$$

**Standardavvikelse** Låt  $X$  vara en stokastisk variabel med varians  $\sigma^2$ . Standardavvikelsen till  $X$ , med notationen  $D(X)$ , definieras som

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

**Variationskoefficient** Låt  $X$  vara en stokastisk variabel med väntevärde  $\mu$  och standardavvikelse  $\sigma$ . Variationskoefficienten till  $X$  definieras som

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

**Kovarians** Låt  $X, Y$  vara stokastiska variabler med väntevärden  $\mu_X, \mu_Y$ . Då definieras kovariansen mellan dessa som

$$C(X, Y) = E((X - \mu_X)(Y - \mu_Y)).$$

**Okorrelerade variabler**  $X, Y$  är okorrelerade om  $C(X, Y) = 0$ .

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

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

**Kvantiler** Lösningen till

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

kallas  $\alpha$ -kvantilen till  $X$ .

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

## 2.2 Satser

**Fördelningsfunktioners egenskaper** Låt  $F$  vara en fördelningsfunktion. Då gäller att

•

$$F(x) \rightarrow \begin{cases} 0, & x \rightarrow -\infty, \\ 1, & x \rightarrow \infty. \end{cases}$$

- $F$  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- respektive 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

$$F(x) = \int_{-\infty}^x f(u) \, du,$$
$$\frac{dF}{dx}(x) = f(x).$$

## 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

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) \, du \, dv,$$
$$\frac{\partial^2 F}{\partial x \partial y}(x, y) = f(x, y).$$

## Bevis

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

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

## Bevis

**Sannolikhetsfunktioner och sannolikheter** Låt  $p$  vara en sannolikhetsfunktion 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  $Y = g(X)$  sannolikhetsfunktionen  $p_Y(k) = \sum_{g(i)=k} p_X(i)$ .

### Bevis

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

$$E(g(X)) = \sum g(k)p_X(k),$$

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  $\mu$ . Då ges variansen till  $X$  av

$$\sigma^2 = E(X^2) - \mu^2.$$

### Bevis

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

$$C(X, Y) = E(XY) - E(X)E(Y).$$

### Bevis

**Väntevärde för linjärkombination av variabler**

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

### Bevis

**Varians för linjärkombination av variabler**

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

### 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

$$E(XY) = E(X)E(Y).$$

### 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  $\mu$  och inför variabeln  $\bar{X} = \frac{1}{n} \sum X_i$ . Då gäller att

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

### Bevis

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

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

### Bevis

**Tjebysjovs olikhet** Låt  $X$  vara en stokastisk variabel med väntevärde  $\mu$  och standardavvikelse  $\sigma$ . Då gäller att

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

## 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  $n(n-1)\dots(n-k+1)$  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

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

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 variabel 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  $X \in \text{Hyp}(N, n, p)$ .

- 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  $X \in \text{Po}(\mu)$ . Fun fact: Poisson betyder fisk på franska.

- Väntevärde:  $\mu$ .
- Varians:  $\mu$ .

#### 4.1 Satser

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

**Bevis**

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

**Bevis**

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

**Bevis**

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

### Bevis

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

### Bevis

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

### Bevis

## 5 Kontinuerliga sannolikhetsfunktioner

**Likformiga fördelningen** Denna sannolikhetsfördelning ges av

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

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

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

**Exponentialfördelningen** Denna sannolikhetsfördelning ges av

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

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

- 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  $\Phi$ . Dessa kommer användas sen utan vidare förklaring. Om en stokastisk variabel är fördelad så, skriver vi  $X \in N(0, 1)$ .

- Väntevärde: 0.
- Varians: 1.

**Kvantiler i standardnormalfördelningen** Vi definierar  $\alpha$ -kvantiler för en standardiserad normalfördelat variabel som  $\lambda_\alpha$  så att

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

**Allmänna normalfördelningen**  $X \in N(\mu, \sigma)$  om och endast om  $Y = \frac{X-\mu}{\sigma} \in N(0, 1)$ . 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:  $\mu$ .
- Varians:  $\sigma^2$ .

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

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

såjs  $Z_n$  vara asymptotiskt normalfördelat. Beteckningen är  $Z \in \text{AsN}(A_n, B_n)$ .

**Kvantiler i  $t$ -fördelningen** Vi definierar  $\alpha$ -kvantiler för en  $t$ -fördelat variabel som  $t_\alpha$  så att

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

## 5.1 Satser

**Standardnormalfördelningens fördelningsfunktion och symmetri** Standardnormalfördelningens fördelningsfunktion uppfyller

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

**Bevis**

$$\begin{aligned} \Phi(-x) &= \int_{-\infty}^{-x} \phi(x) \, dx = \int_{\mathbb{R}} \phi(x) \, dx + \int_{\infty}^{-x} \phi(x) \, dx \\ &= 1 + \int_{\infty}^{-x} \phi(x) \, dx. \end{aligned}$$



Substituera nu  $u = -x$  och få

$$\begin{aligned}\Phi(-x) &= 1 - \int_{-\infty}^x \phi(-u) \, du \\ &= 1 - \int_{-\infty}^x \phi(u) \, du \\ &= 1 - \Phi(x).\end{aligned}$$

Eftersom  $\phi$  är symmetrisk kring 0,

**Linjärkombinationer av normalfördelade variabler** Låt  $X_1, \dots, X_n$  vara oberoende och normalfördelade med väntevärde  $\mu_i$  och varians  $\sigma_i^2$ . Då gäller att:

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

**Bevis**

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

**Fördelning av kvadrat** Låt  $\bar{X} = \frac{1}{n} \sum X_i$  för oberoende och likafördelade  $X_i$  med väntevärde  $\mu$  och standardavvikelse  $\sigma$ . 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  $\mu$ , och skriv  $X_i = \mu + \sigma \varepsilon_i$ . Då gäller att

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

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

**Bevis**

**Centrala gränsvärdesatsen** Låt  $X_1, \dots, X_n$  vara oberoende, likafördelade stokastiska variabler med väntevärde  $\mu$  och standardavvikelse  $\sigma$ . Då uppfyller  $Y_n = \sum X_i$

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

## 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  $\theta$  är en funktion av utfallen  $x_1, \dots, x_n$  av de stokastiska variablerna  $X_1, \dots, X_n$  vars fördelning beror av  $\theta$ . Därmed är punktskattningen ett utfall av stickprovsvariabeln  $\theta^*$ .

**Väntevärdesriktighet** En punktskattning är väntevärdesriktig om  $E(\theta^*) = \theta$ .

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

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

**Medelkvadratfel** Medelkvadratfelet definieras som  $E((\theta^* - \theta)^2)$ .

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

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

**Likelihood-funktionen** Funktionen

$$L(\theta) = \begin{cases} P(X_1 = x_1, \dots, X_n = x_n), & \text{diskret fall,} \\ f_{X_1, \dots, X_n}(x_1, \dots, x_1), & \text{kontinuerligt fall,} \end{cases}$$

givet att  $\theta$  är den sanna parametern, är likelihood-funktionen.

**Maximum likelihood-skattningar** Maximumet till  $L(\theta)$  är maximum likelihood-skattningen av parametern  $\theta$ .

**Minsta kvadrat-skattningar** Låt  $x_1, \dots, x_n$  vara utfall av  $X_1, \dots, X_n$  med väntevärden  $E(X_i) = \mu_i(\theta)$  och definiera kvadratfelet

$$Q = \sum_{i=1}^n (x_i - \mu_i(\theta)).$$

Det värdet av  $\theta$  som minimerar  $Q$  är minsta kvadrat-skattningen av  $\theta$ .

**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

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

**Konfidsensintervall** Intervallet  $I_\theta$  som med sannolikhet  $1 - \alpha$  täcker över den okända parametern  $\theta$  kallas konfidsensintervallet för  $\theta$  med konfidsensgrad  $1 - \alpha$ .

## 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  $X_1, \dots, X_n$  vara normalfördelade med väntevärde  $\mu$  och varians  $\sigma$ . 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  $\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ . Detta betyder att

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

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

**Konfidensintervall för väntevärde, okänd varians** Låt  $X_1, \dots, X_n$  vara normalfördelade med väntevärde  $\mu$ . 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  $t_{\frac{\alpha}{2}}(n-1)$  är  $\frac{\alpha}{2}$ -kvantilen i  $t$ -fördelningen med  $n-1$  frihetsgrader.

**Bevis** Vi har att

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

Därmed har vi

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

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  $X_1, \dots, X_n$  vara normalfördelade med standardavvikelse  $\sigma$ . Då är

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

ett konfidensintervall för väntevärdet, där  $\chi^2_{\frac{\alpha}{2}}(n-1)$  är  $\frac{\alpha}{2}$ -kvantilen i  $\chi^2$ -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  $X_1, \dots, X_{n_1} \in N(\mu_1, \sigma_1)$  och  $Y_1, \dots, Y_{n_2} \in N(\mu_2, \sigma_2)$ , där dessa kan betraktas som stickprov från två olika objekt. Då gäller att:

- Om  $\sigma_1, \sigma_2$  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  $X_1, \dots, X_{n_1} \in N(\mu_1, \sigma_1)$  och  $Y_1, \dots, Y_{n_2} \in N(\mu_2, \sigma_2)$ , 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  $\sigma_1, \sigma_2$  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  $\theta$  vara normalfördelad med väntevärde  $\theta$  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  $\theta$ .

•

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

om  $d$  beror av  $\theta$ , för något lämpligt val av  $d$ .

## Bevis

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

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

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

I fallet där  $\psi^*$  beror av två variabler  $\theta^*$  och  $\eta^*$ , gäller ett motsvarande kriterie. Om kriteriet 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

## 7 Hypotesprövning

Hypotesprövning baseras på stickproc  $X_1, \dots, X_n$  från någon fördelning. Vi önskar pröva någon grundhypotes, eller nollhypotes,  $H_0$  om hur fördelningen ser ut. Nollhypotesen testas gärna mot en alternativ hypotes  $H_1$ .

**Styrkefunktionen** För en given test kan man definiera testens styrkefunktion  $h(\theta)$  som sannolikheten för att  $H_0$  förkastas om  $\theta$  är det rätta värdet på någon parameter. Vi önskar att denna skall vara stor när  $H_1$  är uppfylld.

**Signifikanstester** Från stickproven får man någon teststorhet  $t(X_1, \dots, X_n) = t_{\text{obs}}$ . Man anger sen ett kritiskt område  $C$ , och gör ett signifikanstest:

- Om  $t_{\text{obs}} \in C$  förkastas  $H_0$ .
- Om  $t_{\text{obs}} \notin C$  förkastas ej  $H_0$ .

Man väljer  $C$  på ett sådant sätt att om  $H_0$  är sann, är  $P(t \in C) = \alpha$  för något  $\alpha$ . Detta  $\alpha$  kallas testens signifikansnivå, eller felrisk, och anger sannolikheten för att  $H_0$  förkastas om  $H_0$  är sann. Denna önskas typisk låg.

**$P$ -värden** Man kan även definiera ett  $P$ -värde, eller observerad signifikansnivå. Detta definieras som  $P = P(t \geq t_{\text{obs}})$  under förutsättningen att  $H_0$  är sann. Om  $P \leq \alpha$  förkastar man  $H_0$ .

**Konfidensmetoden** Signifikanstester är ekvivalent med konfidensmetoden, där man hittar ett konfidensintervall med konfidensgrad  $1 - \alpha$ , där  $\alpha$  är testens signifikans, och undersöka om värdet  $\theta_0$ , specificerat i  $H_0$ , ligger i konfidensintervallet.

## 7.1 $\chi^2$ -tester

**Test av förutspådda sannolikheter** Gör  $n$  oberoende försök med möjliga utfall  $A_1, \dots, A_r$ , där alla utfall har frekvens  $x_i$ . Vi vill testa  $H_0$ : alla  $P(A_i) = p_i$  för några  $p_i$ . Vi definierar testvariabeln

$$Q = \sum_{i=1}^r \frac{(x_i - np_i)^2}{np_i}.$$

Om  $H_0$  är sann, är  $Q$  approximativt  $\chi^2(r-1)$ . Det är även så att  $Q$  ökar med avviket av  $p_i$  från det reella värdet. Vi förkastar då  $H_0$  om  $Q_{\text{obs}} > \chi_{\alpha}^2(r-1)$  med signifikans  $\alpha$ . Testen fungerar bra om  $np_i \geq 5 \forall i$ .

**Test av parameter i fördelning** Gör  $n$  oberoende försök med möjliga utfall  $A_1, \dots, A_r$ , där alla utfall har frekvens  $x_i$ . Vi vill testa  $H_0$ : alla  $P(A_i) = p_i(\theta_1, \dots, \theta_k)$  för några  $\theta_i$ . Vi skattar alla  $\theta_i$  och definierar igen testvariabeln

$$Q = \sum_{i=1}^r \frac{(x_i - np_i^*)^2}{np_i^*},$$

där  $p_i^*$  är sannolikheten baserad på skattningarna av  $\theta_i$ . Om  $H_0$  är sann, är  $Q$  approximativt  $\chi^2(r - k - 1)$ . Vi förkastar då  $H_0$  om  $Q_{\text{obs}} > \chi_\alpha^2(r - k - 1)$  med signifikans  $\alpha$ . Testen fungerar bra om  $np_i^* \geq 5 \forall i$ .

**Test av homogenitet i sannolikheter** Gör  $s$  serier med  $n_i$  oberoende försök per serie, med möjliga utfall  $A_1, \dots, A_r$  med frekvens  $x_{sr}$ . Vi vill testa  $H_0$ : alla  $P(A_i)$  är lika från experiment till experiment. Vi definierar testvariabeln

$$Q = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*},$$

där  $p_j^* = \frac{1}{n} \sum_{i=1}^s x_{ij}$ . Om  $H_0$  är sann, är  $Q$  approximativt  $\chi^2((r - 1)(s - 1))$ . Vi förkastar då  $H_0$  om  $Q_{\text{obs}} > \chi_\alpha^2((r - 1)(s - 1))$  med signifikans  $\alpha$ . Testen fungerar bra om  $n_i p_j^* \geq 5 \forall i$ .

**Test av oberoende** Gör  $n$  oberoende försök, med  $sr$  möjliga utfall  $B_1 A_1, \dots, B_s A_r$  med frekvens  $x_{sr}$ . Det gäller att

$$P(A_j) = p_{.j} = \sum_{i=1}^s p_{ij},$$

och motsvarande för  $B_i$ . Vi skattar dessa som

$$p_{.j}^* = \frac{1}{n} \sum_{i=1}^s x_{ij}.$$

Vi vill testa  $H_0$ :  $A$  och  $B$  är oberoende, med andra ord att  $p_{ij} = p_i p_{.j}$ . Vi definierar testvariabeln

$$Q = \sum_{i=1}^s \sum_{j=1}^r \frac{(x_{ij} - n p_i^* p_{.j}^*)^2}{n p_i^* p_{.j}^*}.$$

Om  $H_0$  är sann, är  $Q$  approximativt  $\chi^2((r - 1)(s - 1))$ . Vi förkastar då  $H_0$  om  $Q_{\text{obs}} > \chi_\alpha^2((r - 1)(s - 1))$  med signifikans  $\alpha$ . Testen fungerar bra om  $n p_i^* p_{.j}^* \geq 5 \forall i, j$ .