Verjetnost in statistika - zapiski s predavanj prof. Drnovška

Tomaž Poljanšek

študijsko leto 2022/23

Kazalo

1 Verjetnost

1.1 Neformalni uvod v verjetnost

Začetki verjetnosti (kot vede) so v 17. stoletju, motivacija igre na srečo
17. stol: Fermat, Pascal, Bernoulli
18. in 19 stol: Laplace, Poisson, Čebišev, Markov
20. stol: Kolmogorov (okoli 1930), utemeljitelj sodobnega verjetnostnega računa

Definicija 1.1 (Dogodek)**.** Izvajamo poskus, opazujemo nek pojav, ki se lahko zgodi in ga imenujemo dogodek.

Primer. Met poštene kocke, dogodek je npr. pade šestica, ali npr. pade sodo število pik.

Definicija 1.2 (Frekvenca)**.** Poskus ponovimo n-krat. Opazujemo dogodek A.
Naj bo $K_n(A)$ frekvenca dogodka A, t.j. število tistih ponovitev, pri katerih se je dogodek A zgodil.
Relativna frekvenca je $f_n(A) = \frac{K_n(A)}{n} \in [0,1]$

Dokazati je mogoče, da zaporedje $\{f_n(A)\}$ konvergira, recimo h $p \in [0,1]$.
Statistična definicija verjetnosti: $P(A) := p$.
Pogosto verjetnost lahko določimo vnaprej:
Klasična definicija verjetnosti: $P(A) = \frac{\text{število ugodnih izidov za dogodek } A}{\text{število vseh izidov}}$ pri pogoju, da imajo vsi izidi enake možnosti

Primer. met kocke:
$P(\text{pade šestica}) = \frac{1}{6}$
$P(\text{pade sodo število pik}) = \frac{3}{6} = \frac{1}{2}$

Primer. Kolikšna je verjetnost, da pri metu dveh kock znaša vsota pik 7?
Možne vsote: $2, 3, 4 \cdots 7 \cdots 12$: 11 možnosti
Ali je $P(\text{vsota } 7) = \frac{1}{11}$? Ne, ker te vsote nimajo enakih možnosti:
$2 = 1 + 1, 3 = 2 + 1 = 1 + 2$
Vsi možni izidi so $\{(i,j) : i,j \in \{1,2\cdots 6\}\} = \{1,2\cdots 6\} \times \{1,2\cdots 6\}$

$$\begin{matrix} (1,1) & (1,2) & \cdots & (1,6) \\ (2,1) & (2,2) & \cdots & (2,6) \\ \vdots & \vdots & \ddots & \vdots \\ (6,1) & (6,2) & \cdots & (6,6) \end{matrix}$$

$P(\text{vsota } 7) = \frac{6}{36} = \frac{1}{6}$

Če je vseh izidov neskončno, si lahko pomagamo z geometrijsko definicijo verjetnosti.

Primer. Osebi se dogovorita za sestanek med 10. in 11. uro; čas prihoda je slučajen. Vsak čaka največ 20 minut, najdlje do 11. ure; če v tem času drugega ni, odide. Kolikšna je verjetnost srečanja?
Začnimo čas šteti ob 10. uri. Naj bo x čas prihoda 1. osebe, y pa čas prihoda druge osebe
Možni izidi so kvadrat $[0,1]^2 = [0,1] \times [0,1]$
Ugodni izidi so $|x-y| \leq \frac{1}{3}$

1. $x \geq y : \ x - y \leq \frac{1}{3}$ oz. $x - \frac{1}{3} \leq y$
2. $x \leq y : \ y - x \leq \frac{1}{3}$ oz. $y \leq x + \frac{1}{3}$

$$P(\text{srečanje}) = \frac{\text{ploščina označenega lika}}{\text{pološčina kvadrata}} = \frac{1-(\frac{2}{3})^2}{1} = \frac{5}{9}$$

Primer. Slučajno razporedimo n kroglic v m posod, kjer je $m > n$. Kolikšna je verjetnost, da so vse kroglice v prvih n posodah, v vsaki ena?
Obravnavajmo 3 variante:

1. kroglice razlikujemo
   Vsi izidi: $m \cdot m \cdots m = m^2$ variacije
   Ugodni izidi: $n \cdot (n-1) \cdots 1 = n!$ permutacija
   $\implies P(A) = \frac{n!}{m^n}$
2. kroglic ne razlikujemo
   n kroglic, $m-1$ črtic $\implies (m+n-1)$ mest
   Vsi izidi: $\binom{m+n-1}{n}$ kombinacije s ponavljanjem
   Ugidni izidi: 1
   $P(A) = \frac{1}{\binom{m+n-1}{n}}$
3. kroglic ne razlikujemo, v vsaki posodi je kvečjemu ena
   Število vseh izidov je $\binom{m}{n}$
   Ugoden izid je samo eden
   Torej je $P(A) = \frac{1}{\binom{m}{n}}$

V kvantni mehaniki so kroglice različni delci, posode so energetska stanja.
V primeru (a) imamo Maxwell-Boltzmanovi statistiki, velja za molekule plina.

V primeru (b) imamo Bose-Einsteinovo statistiko, velja za bozone (npr. fotoni).
V primeru (c) imamo Fermi-Diracovo statistiko, velja za fermione (npr elektroni); zanje velja Diracovo izključitveno načelo: v vsakem stanju je največ en delec.

1.2 Aksiomatična definicija verjetnosti

Kolmogorov (okoli 1930)

Definicija 1.3 (Dogodek)**.** Imamo prostor vseh dogodkov Ω (možna oznaka je G). Dogodki so nekatere (ne nujno vse) podmnožice $A \subseteq \Omega$.

Primer. Met kocke: $\Omega = \{1, 2, 3, 4, 5, 6\}$, dogodki so vse podmnožice, $\{6\} \cdots$ dogodek, da pade šestica, $\{2, 4, 6\} \cdots$ dogodek, da pade sodo število pik.

Računanje z dogodki

1. Vsota dogodkov oz. unija dogodkov: $A + B$ oz. $A \cup B$: dogodek, da se zgodi vsaj eden od A in B
2. Produkt dogodkov oz. presek dogodkov: $A \cdot B$ oz. $A \cap B$: dogodek, da se zgodita oda dogodka A in B
3. Nasprotni dogodek oz. komplement dogodka: $\overline{A} = A^C$

Pravila za računanje z dogodki:

- idempotentnost:
$$A \cup A = A = A \cap A$$
- komutativnost:
$$A \cap B = B \cap A$$
$$A \cup B = B \cup A$$
- asociativnost:
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$
- distributivnost:
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$
$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

oziroma

$$(A \cdot B) + C = (A + C) \cdot (B + C)$$
$$(A + B) \cdot C = (A \cdot C) + (B \cdot C)$$

- deMorganova zakona:

$$(A \cup B)^C = A^C \cap B^C$$
$$(A \cap B)^C = A^C \cup B^C$$

še več:

$$(\cup_{i \in I} A_i)^C = \cap_{i \in I} A_i^C$$
$$(\cap_{i \in I} A_i)^C = \cup_{i \in I} A_i^C$$

Definicija 1.4 (σ-algebra)**.** Neprazna družina podmnožic dogodkov F v Ω je σ-algebra, če velja:

1. $A \in F \implies A^C \in F$ (zaprtost za komplement)
2. $A_1, A_2 \dots \in F \implies \cup_{i=1}^{\infty} A_i \in F$ (zaprtost za števne unije)

Če v 2) zahtevamo manj:
$A, B \in F \implies A \cup B \in F$ (šibkejši pogoj) pravimo, da je F algebra.

V algebri imamo zaprtost za končne unije, t.j. $A_1 \cdots A_n \in F \implies \cup_{i=1}^{n} A_j \in F$ (zaradi indukcije). Ker je $\cap_i A_i^C = (\cup_i A_i)^C$ (deMorgan), je algebra zaprta za končne preseke, σ-algebra pa za števne preseke.
Ker je $A \setminus B = A \cap B^C$, je algebra zaprta za razlike dogodkov.
Vsaka algebra vsebuje $\{\emptyset, \Omega\}$: ker je neprazna, obstaja dogodek $A \in F$, potem je $A^C \in F$ in zato je

$$\Omega = A \cup A^C \in F, \emptyset = A \cap A^C \in F$$

Najmanjša (σ-)algebra je $F = \{\emptyset, \Omega\}$, največja ($\sigma$-)algebra je potenčna množica $P(\Omega)$.

Primer. Izberimo $\emptyset \neq A \subsetneq \Omega$. Najmanjša σ-algebra, ki vsebuje $\{1\}, \{2\}, \{3\} \cdots$ je $P(\mathbb{N})$, saj je $A = \cup_{k \in A}\{k\}$ za $\forall k \subseteq \mathbb{N}$ (končna ali števna unija)
Najmanjša algebra F, ki vsebuje $\{1\}, \{2\}, \{3\} \cdots$ je enaka algebri

$$g = \{A \subseteq \mathbb{N} : A \text{ je končna ali } A^C \text{ je končna}\}$$

Dokazujemo g je algebra:

1. zaprtost za komplemente: $A \in g \implies$
   (a) bodisi A je končna množica $\implies A = (A^C)^C \implies A^C \in g$ (zaprtost za komplement)
   (b) bodisi je A^C končna množica $\implies A^C \in g$
2. $A, B \in g \overset{?}{\implies} A \cup B \in g$
   (a) $A \cup B$ je končna $\implies (A \cup B) \in g$ (vse končne množice)
   (b) $A \cup B$ ni končna $\implies$ vsaj ena izmed A in B ni končna, recimo A je neskončna.
   Toda $A \in G \implies A^C$ je končna množica.
   Ker je $(A \cup B)^C \subseteq A^C$ in A^C je končna, je $(A \cup B)^c$ tudi končna množica, torej $A \cup B \in g$
   $A \in g$:
   - A je končna: $A = \cup_{k \in A}\{k\} \in F$ (končna unija)
   - A^C je končna: $A^C = (\cup_{k \in A^C})^C \in F$ (končna unija)

   $\implies g \in F$

Ker je F najmanjša algebra, ki vsebuje $\{1\}, \{2\} \cdots$, je tukaj enačaj, torej $g = F$
Ker npr. množica sodih števil ni v $g \in F$, je $g \notin P(\mathbb{N})$, torej g ni σ-algebra

Definicija 1.5 (Nezdružljivost dogodkov)**.** Dogodka A in B sta nezdružljiva ali disjunktna, če je $A \cap B = \emptyset$

Definicija 1.6 (Popoln sistem dogodkov)**.** Zaporedje $\{A_i\}_i$ (končno ali števno mnogo) je popoln sistem dogodkov, če $\Omega = \cup_i A_i$ in $A_i \cap A_j = \emptyset \forall i \neq j$

Definicija 1.7 (Verjetnost)**.** Naj bo $F \sigma$-algebra na Ω. Verjetnost na (Ω, F) je preslikava $P : F \to \mathbb{R}$ z lastnostmi:

1. $P(A) \geq 0 \forall A \in F$
2. $P(\Omega) = 1$
3. Za poljubne paroma nezdružljive dogodke $A_1, A_2 \cdots$ velja

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

števna aditivnost (verjetnostne preslikave)

Lastnosti preslikave P:

1. $P(\emptyset) = 0$

   Dokaz. v 3) vstavimo $A_1 = A_2 = \cdots = \emptyset$:

   $$P(\emptyset) = P(\emptyset) + P(\emptyset) + \cdots + P(\emptyset) = k \cdot P(\emptyset) \implies$$
   $$\implies (k-1)P(\emptyset) = 0 \implies P(\emptyset) = 0$$

   ■

2. P je končno aditivna, t.j. za poljubne paroma nezdružljive dogodke $A_1 \cdots A_n$ velja

   $$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

   Dokaz. v 3) vzamemo $A_{n+1} = A_{n+2} = \cdots = \emptyset$:

   $$P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{n} P(A_i) (\text{zaradi} P(\emptyset) = 0)$$

   ■

3. P je monotona, t.j. iz $A \subseteq B$ $(A, B \in F)$ sledi $P(A) \subseteq P(B)$, še več: $A \subseteq B \implies P(B \backslash A) = P(B) - P(A)$

   Dokaz. Ker je $B = A \cup (B \backslash A)$ in $A \cap (B \backslash A) = \emptyset$, je po b) $P(B) = P(A) + P(B \backslash A)$ ■

4. $P(A^C) = 1 - P(A)$ za $A \in F$

   Dokaz.

   $$B = \Omega \implies P(A^C) = P(\Omega \backslash A) =$$
   $$\overset{c)}{=} P(\Omega) - P(A) \overset{2)}{=} 1 - P(A)$$

   ■

5. P je zvezna, t.j.

   (a) iz $A_1 \subseteq A_2 \subseteq \cdots A_i \in F$ sledi $P(\cup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$

   (b) iz $B_1 \supseteq B_2 \supseteq \cdots B_i \in F$ sledi $P(\cap_{i=1}^{\infty} B_i) = \lim_{i \to \infty} P(B_i)$

Definiramo

$$C_1 = A_1$$
$$C_i = A_i - A_{i-1} \text{ za } i \geq 2$$

Potem $C_i \cap C_j = \emptyset$ za $i \neq j$, $A_n = C_1 \cup \cdots \cup C_n$ in $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} C_i$
Torej imamo

$$P(\cup_{i=1}^{\infty} A_i) = P(\cup_{i=1}^{\infty} C_i) =$$
$$\stackrel{3)}{=} \sum_{i=1}^{\infty} P(C_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(C_i) =$$
$$\stackrel{b)}{=} \lim_{n\to\infty} P(A_i)$$

Dokaz. Dokazujemo ii): iz $B_1 \supseteq B_2 \supseteq \cdots$ sledi $B_1^C \subseteq B_2^C \subseteq \cdots$ in zato po (i)

$$P(\cup_{i=1}^{\infty} B_i^C) = \lim_{i\to\infty} P(B_i^C) \stackrel{a)}{=} 1 - \lim_{i\to\infty} P(B_i)$$

Toda
$$P(\cup_{i=1}^{\infty} B_i^C) = P((\cup_{i=1}^{\infty} B_i)^C) \stackrel{d)}{=} 1 - P(\cap_{i=1}^{\infty} B_i)$$
Od tod sledi želena enakost

$$P(\cap_{i=1}^{\infty} B_i^C) = \lim_{i\to\infty} P(B_i)$$

■

(Ω, F, P) verjetnostni prostor

Primer. (končni ali števni verjetnostni prostor)
$\Omega = \{w_1, w_2 \cdots\}$ končna ali števna množica, paroma različni,
$F = P(\Omega), A = \cup_{i\in A}\{w_i\}$ končna ali števna unija
$\{w_1\}, \{w_2\}\ldots$ so popoln sistem dogodkov
če je $p_i := P(\{w_i\})$, je $P(A) = \sum_{i:w_i\in A} p_i$ in $\sum_i p_i = 1 = P(\Omega)$
Poseben primer: Ω ima n elementov in $p_i = \frac{1}{n}, P(A)\frac{\text{moč}(A)}{n}$
To je klašična definicija verjetnosti.

(Ω, Φ, P)

Primer. (Neštevni neskončni verjetnostni prostor)
$\Omega = [0,1] \times [0,1]$
$\Phi :=$ najmanjša σ-algebra, ki vsebuje vse odprte pravokotnike $(a,b)\times(c,d)$, $a,b,c,d \in (0,1)$

npr. elipse: $\frac{1}{n}$ radij, za $\forall n$ vzamemo kvadrate v elipsi, izberemo unijo
= najmanjša σ-algebra, ki vsebuje vse zaprte pravokotnike $[a,b]\times[c,d]$, $a,b,c,d \in [0,1]$ - Borelova σ-algebra
Izkaže se, da $\Phi \neq P(\Omega)$

Verjetnost P definiramo na pravokotnikih s $P((a,b)\times(c,d)) = (b-a)(d-c)$
Ni lahko videti, da je to možno razširiti do števno aditivne preslikave na $P(\Omega)$
Verjetnostna preslikava P (na Φ) se imenuje Lebesgueova mera
To je geometrijska definicija verjetnosti:

$$\square = \cap_{n=1}^{\infty}(a - \frac{1}{n}, b + \frac{1}{n}) \times (c - \frac{1}{n}, d + \frac{1}{n})$$

1.3 Pogojna verjetnost

Definicija 1.8 (Pogojna verjetnost)**.** Fiksirajmo dogodek B s $P(B) > 0$. Pogojna verjetnost dogodka A pri pogoju B je

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

Primer. V posodi sta 2 beli in ena črna kroglica. Slučajno izberemo eno kroglico, jo vrnemo v posodo in potem ponovno izberemo kroglico. Kolikšna je verjetnost, da smo v drugo izbrali belo kroglico, če smo v prvo izbrali belo kroglico? $\Omega = \begin{matrix} B_1B_1 & B_1B_2 & B_1\check{C} \\ B_2B_1 & B_2B_2 & B_2\check{C} \\ \check{C}B_1 & \check{C}B_2 & \check{C}\check{C} \end{matrix}$

$$P(\text{prvič bela}) = \frac{6}{9} = \frac{2}{3}$$
$$P(\text{drugič bela} \mid \text{prvič bela}) = \frac{P(\text{prvič in drugič bela})}{P(\text{prvič bela})} = \frac{\frac{4}{9}}{\frac{2}{3}} = \frac{2}{3}$$

Iz definicije sledi $P(A \cap B) = P(B) \cdot P(A \mid B)$
Za poljubne dogodke A, B, C velja

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

oz. “lepše”

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

To posplošimo na n dogodkov $A_1, A_2 \cdots A_n$:

$$P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}) =$$
$$= P(A_1) \cdot \prod_{i=2}^{n} P(A_i \mid \cap_{j=1}^{i-1} A_j)$$

Desna stran:

$$P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdots \frac{P(A_1 \cap \cdots \cap A_n)}{P(A_1 \cap \cdots \cap A_{n-1})}$$

Imejmo poskus v dveh korakih (fazah). V 1. koraku se zgodi natanko en dogodek iz popolnega sistema dogodkov $H_1, H_2 \cdots$ (končno/števno mnogo). V drugem koraku nas zanima dogodek A. Izrazimo $P(A)$ z verjetnostmi $P(H_1), P(H_2 \cdots)$ in $P(A \mid H_1), P(A \mid H_2) \cdots$.
Ker je $A = A \cap \Omega = A \cap (\cup_i H_i) = \cup_i (A \cap H_i)$ in ker so $\{A \cap H_i\}_i$ paroma nezrdužljivi dogodki (zaradi H_i), je

$$P(A) = \sum_i P(A \cap H_i) = \sum_i P(H_i) \cdot P(A \mid H_i)$$

To je formula o popolni verjetnosti

Primer. Na srečolovu je n srečk, od tega je m dobitnih ($m < n$). Ali imamo pred začetkom srečolova večje možnosti za dobitek, če izbiramo prvi ali drugi?
H_1: prvi dobi, H_2: prvi ne dobi, A: drugi zadane

$$P(\text{prvi dobi}) = \frac{m}{n}$$
$$P(\text{drugi dobi}) = P(\text{prvi dobi}) \cdot P(\text{drugi dobi} \mid \text{prvi dobi}) +$$
$$+ P(\text{prvi ne dobi}) \cdot P(\text{drugi dobi} \mid \text{prvi ne dobi}) =$$
$$= \frac{m}{n} \cdot \frac{m-1}{n-1} + \frac{n-m}{n} \cdot \frac{m}{n-1} = \cdots = \frac{m}{n}$$

Pri dvofaznem poskusu nas zanima

$$P(H_k \mid A) = \frac{P(H_k \cap A)}{P(A)} = \frac{P(H_k) \cdot P(A \mid H_k)}{\sum_i P(H_i) \cdot P(A \mid H_i)}$$

- Bayesova formula

Primer. Test s poligrafom (= detektor laži)
Resnicoljub opravi test s poligrafom z verjetnostjo 0.95. Z enako verjetnostjo poligraf prepozna lažnivca. Izmed 1000 oseb, med katerimi je natanko en lažnivec, slučajno izberemo eno osebo, katero poligraf proglasi za lažnivca. Kolikšna je pogojna verjetnost, da je ta oseba res lažnivec?
Naj bo L dogodek, da je oseba lažnivec.
Naj bo L_p dogodek, da poligraf za osebo pravi, da je lažnivec. Potem je

$$\begin{aligned}
&P(L_p \mid L) = 0.95 \text{ in } P(L_P^C \mid L^C) = 0.95 \text{ oz.} \\
&P(L_P \mid L^C) = 0.05 \\
&P(L) = 0.001
\end{aligned}$$

Iščemo verjetnost $P(L \mid L_p)$
$H_1 = L, H_2 = L^C, A = L_p$

$$\begin{aligned}
P(L \mid L_p) &= \frac{P(L) \cdot P(L_p \mid L)}{P(L) \cdot P(L_p \mid L) + P(L^C) \cdot P(L_p^C \mid L^C)} = \\
&= \frac{0.001 \cdot 0.95}{0.001 \cdot 0.95 + 0.999 \cdot 0.05} = \frac{95}{5050} \doteq 0.02 = \frac{1}{50}
\end{aligned}$$

Matematično ekvivalenten problem je presejalni test, npr. program DORA.
(Pogojna) verjetnost, da je oseba bolna, če je test pozitiven, je majhna.
Dogodka A in B sta neodvisna, če je $P(A \cap B) = P(A) \cdot P(B)$
Če je $P(B) > 0$, potem lahko ta pogoj zapišemo kot $P(A) = \frac{P(A \cap B)}{P(B)} = P(A \mid B)$

Definicija 1.9 (Neodvisnost)**.** A in B sta neodvisna, če $P(A \cap B) = P(A) \cdot P(B)$
Dogodki $\{A_i\}_i$ so neodvisni, če za poljuben končen nabor različnih dogodkov $A_{i_1}, A_{i_2} \cdots A_{i_n}$ velja

$$P(A_{i_i} \cap \cdots \cap A_{i_n}) = P(A_{i_1}) \cdot \cdots \cdot P(A_{i_n})$$

Če zahtevamo le za $n = 2$, t.j. $P(A_i \cap A_j) = P(A_i) \cdot P(A_j), i \neq j$, tedaj so dogodki paroma neodvisni

Očitno iz neodvisnosti sledi paroma neodvisnost. Obratno ne velja

Primer.

$$\begin{aligned}
&\Omega = \{1,2,3,4\}, P(\{k\}) = \frac{1}{4} \text{ za } k = 1,2,3,4 \text{ npr. met tetraedra}\\
&A = \{1,2\}, B = \{1,3\}, C = \{1,4\}\\
&P(A) = P(B) = P(C) = \frac{2}{4} = \frac{1}{2}\\
&A \cap B = B \cap C = A \cap C = \{1\}\\
&\implies P(A \cap B) = P(B \cap C) = P(A \cap C) = \frac{1}{4}\\
&\implies A, B, C \text{ so paroma neodvisni}\\
&A \cap B \cap C = \{1\}\\
&P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A) \cdot P(B) \cdot P(C)\\
&\implies \text{ niso neodvisni}
\end{aligned}$$

Trditev 1.10. Naj bosta A in B neodvisna dogodka. Potem sta neodvisna tudi A in B^C. Prav tako tudo A^C in B ter A^C in B^C (komplementiranje ohranja neodvisnost)

Dokaz. Ker je $A \cap B^C = A \backslash (A \cap B)$ je

$$\begin{aligned}
&P(A \cap B^C) = P(A \backslash (A \cap B)) = P(A) - P(A \cap B) =\\
&\overset{A,B \text{ neodvisna}}{=} P(A) - P(A) \cdot P(B) = P(A)(1 - P(B)) = P(A) \cdot P(B^C)
\end{aligned}$$

podobno za ostale kombinacije ■

1.4 Zaporedja neodvisnih ponovitev poskusa

Definicija 1.11. Imejmo zaporedje n neodvisnih ponovitev poskusa, določenega v verjetnostnem prostoru (Ω, Φ, P), v katerem je možen A s $P(A) = p \in (0,1)$. Potem je $q := P(A^C) = 1 - p$

Z $A_n(k)$ označimo dogodek, da se v k ponovitvah poskusa A zgodi natanko n-krat, $k = 0, 1 \cdots n$
Pokažimo, da je njegova verjetnost $P_n(k) := P(A_n(k)) = \binom{n}{k} p^k q^{n-k}$ - Bernoullijeva formula
$A_n(k)$ je disjunktna unija $\binom{n}{k}$ dogodkov, da se A zgodi na predpisanih k mestih, A^C pa na preostalih $(n-k)$ mestih. Verjetnost le teh je produkt p-jev in q-jev: $p^k q^{n-k}$. Od tod sledi Bernoullijeva formula

Primer. Kaljivost semen je 95%. Kolikšna je verjetnost, da izmed 1000 semen vzkali točno 950 semen?
A = seme ne vzkali
$P(A) = p = 0.05, q = 0.95$
$P_{1000}(50) = \binom{1000}{50} 0.05^{50} \cdot 0.95^{950} \doteq 0.05779$

Brez računala je to težko izračunati tudi če uporabimo Stirlingovo formulo na $n!$:

$$n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$$

Tukaj $\sim$ pomeni: $a_n \sim b_n$ če $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$
Torej je $\lim_{n\to\infty} \frac{\sqrt{2\pi n}}{n!}(\frac{n}{e})^n = 1$

1.4.1 Aproksimacijski formuli za $P_n(k)$

1.4.2 Poissonova formula

Če je n velik in k majhen, je $P_n(k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$, kjer je $\lambda = np$

Dokaz.

$$P_n(k) \overset{\text{def}}{=} \binom{n}{k} p^k q^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}(\frac{\lambda}{n})^k(1-\frac{\lambda}{n})^{n-k} =$$
$$= \frac{\lambda}{k!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k} \approx$$
$$\frac{n-i}{n} \to 1, (1-\frac{\lambda}{n})^n \to e^{-\lambda}, (1-\frac{\lambda}{n})^{-k} \to 1$$
$$\approx \frac{\lambda^k}{k!} e^{-\lambda}$$

■

Primer. Kaljivost semen

$$
\begin{aligned}
P_{1000}(50) &\doteq \frac{50^{50}}{50!}e^{-50} = \frac{1}{50!}(\frac{50}{e})^{50} = \\
&\overset{\text{Stirling}}{=} \frac{1}{\sqrt{2\pi 50}} = \frac{1}{10\sqrt{pi}} \doteq 0.05642
\end{aligned}
$$

1.4.3 Laplaceova lokalna formula

Če je n velik, potem je $P_n(k) \approx \frac{1}{\sqrt{2\pi npq}} \cdot e^{-\frac{(k-np)^2}{2npq}}$
Kasneje (2. semester) bomo dokazali splošnejši izrek (centralni limitni izrek)
Narišimo zaporedje $\{P_n(k)\}_{k=0}^{n}$, n fiksen

$$
\begin{aligned}
&P_n(0) = q^n \\
&P_n(1) = npq^{n-1} \\
&P_n(2) = \frac{n(n-1)}{2}p^2q^{n-2}
\end{aligned}
$$

Pomaknjena in raztegnjena funkcija $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

$$
\begin{aligned}
&P_n(k) \leq P_n(k+1)? \\
&\frac{n!}{k!(n-k)!}p^k q^{n-k} \leq \frac{n!}{(k+1)!(n-k-1)!}p^{k+1}q^{n-k-1} \\
&\frac{q}{n-k} \leq \frac{p}{k+1} \iff kq + q \leq np - kp \iff \\
&\iff k(p+q) + q \leq np \iff k + q \leq np
\end{aligned}
$$

Neenakost se obrne pri $k \approx np$

Primer. Kaljivost semen

$$
\begin{aligned}
&p = 0.05, q = 0.95, k = 50 \implies np = 50 \\
&P_{1000}(50) \approx \frac{1}{\sqrt{2\pi \cdot 50 \cdot 0.95}} = \frac{1}{\sqrt{95\pi}} \doteq 0.05788
\end{aligned}
$$

1.4.4 Laplaceova integralska formula

Zanima nas dogodek $B_n(k_1, k_2)$, da se n n ponovitvah poskusa dogodek A zgodi vsaj k_1-krat in manj kot k_2-krat, $0 \leq k_1 < k_2 \leq n + 1$
Ker je

$$B_n(k_1, k_2) = A_n(k_1) \cup A_n(k_1 + 1) \cup \cdots \cup A_n(k_2 - 1)$$

(disjunktna unija), je

$$P_n(k_1, k_2) := P(B_n(k_1, k_2)) = \sum_{k=k_1}^{k_2-1} |A_n(k)| = \sum_{k=k_1}^{k_2-1} P_n(k)$$

Po Laplaceovi lokalni formuli je

$$P_n(k_1, k_2) \approx \frac{1}{\sqrt{2\pi npq}} \sum_{k=k_1}^{k_2-1} e^{-\frac{(k-nq)^2}{2npq}} =$$
$$\doteq \frac{1}{\sqrt{2\pi}} \sum_{k=k_1}^{k_2-1} e^{-\frac{1}{2}x_k^2} \Delta x_k$$

kjer je

$$x_k := \frac{k - np}{\sqrt{npq}}$$
$$\implies \Delta x_k := x_{k-1} - x_k = \frac{k + 1 - np}{\sqrt{npq}} - \frac{k - np}{\sqrt{npq}} = \frac{1}{\sqrt{npq}}$$

To je integralaska (Riemannova) vsota za funkcijo $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
$P_n(k_1, k_2) \approx \sum_{k=k_1}^{k_2-1} f(x_k) \Delta x_k$ na intervalu $a = \frac{k_1-np}{\sqrt{npq}}, b = \frac{k_2-np}{\sqrt{npq}}$
Za velik n torej velja:

$$P_n(k_1, k_2) \approx \int_a^b f(x)dx = \int_{\frac{k_1-np}{\sqrt{npq}}}^{\frac{k_2-np}{\sqrt{npq}}} e^{-\frac{x^2}{2}} dx$$

- Laplaceova integralska formula
$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ - verjetnostni integral

Vpeljimo verjetnostni integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

Φ je liha funkcija, zvezno odvedljiva in strogo naraščajoča
$\Phi(0) = 0$ in $\Phi(x) = f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
Pokažimo, da je $\lim_{x\to\infty} \Phi(x) = \frac{1}{2}$. S pomočjo Γ funkcije imamo

$$\begin{aligned}
&\lim_{x\to\infty} \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-\frac{x^2}{2}} dt = \\
&x = \frac{t^2}{2}, dx = tdt, dt = \frac{dx}{t} = \frac{dx}{\sqrt{2x}} \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-x} \frac{dx}{\sqrt{2x}} = \\
&= \frac{1}{2\sqrt{\pi}} \int_0^\infty x^{-\frac{1}{2}} e^{-x} dx = \\
&\overset{\Gamma(\frac{1}{2})=\sqrt{pi}}{=} \frac{1}{2}
\end{aligned}$$

Laplaceova formula se glasi:

$$P_n(k_1, k_2) = \Phi(\frac{k_2 - np}{\sqrt{npq}}) - \Phi(\frac{k_1 - np}{\sqrt{npq}})$$

Primer. Kaljivost semen
Kolikšna je verjetnost, da vzkali več kot 950 semen v zavojčku s 1000 semeni
A: seme ne vzkali, $p = P(A) = 0.05, q = 0.95, n = 1000 \implies np = 1000$

$$\begin{aligned}
&P_{1000}(0, 50) = \Phi(\frac{50 - 50}{\sqrt{50 \cdot 0.95}}) - \Phi(\frac{0 - 50}{\sqrt{50 \cdot 0.95}}) \doteq \\
&\doteq \Phi(7.36) \approx 0.500
\end{aligned}$$

- verjetnost, da ne vzkali manj kot 50 semen

1.5 slučajne spremenljivke

Danemu poskusu priredimo določeno številsko količino, katere verjetnost je odvisna od slučajna

Primer.

1. Met kocke, število pik
2. Streljanje v tarčo, razdalja zadetka od središča tarče

Definicija 1.12 (Slučajna spremenljivka)**.** Realna slučajna spremenljivka na verjetnostnem prostoru (Ω, Φ, P) je funkcije $X : \Omega \to \mathbb{R}$ z lastnostjo, da je za $\forall x \in \mathbb{R}$ množica $\{\omega \in \Omega : X(\omega) \leq x\}$ v Φ, se pravi dogodek

Oznaka: $\{\omega \in \Omega : X(\omega) \leq x\} \equiv X^{-1}((-\infty, x]) \equiv (X \leq x)$ (ali $\{X \leq x\}$)

Definicija 1.13 (Porazdelitvena funkcija)**.** Porazdelitvena funkcija $F_X : \mathbb{R} \to \mathbb{R}$ je funkcija, definirana s predpisom $F_X(x) = P(X \leq x) \equiv P((X \leq x))$

Dogovor: $P((X \leq x)) \leftrightarrow P(X \leq x)$
Lastnosti porazdelitvene funkcije $F_X \equiv F$:

1. $0 \leq F(X) \leq 1$ za $\forall x \in \mathbb{R}$ (verjetnost)
2. F je naraščajoča funkcija, t.j. iz $x_1 < x_2$ sledi $F(x_1) \leq F(x_2)$

   Dokaz. sledi iz $(X \leq x_1) \subseteq (X \leq x_2) \quad /P()$ ■

3. $\lim_{x\to\infty} F(x) = 1, \lim_{x\to-\infty} F(x) = 0$

   Dokaz. limita $\lim_{x\to\infty} F(x)$ obstaja, ker je F naraščajoča in navzgor omejena z 1.
   Vzemimo strogo naraščajoče zaporedje $\{x_n\} \subseteq \mathbb{R}$, ki je neomejeno. Potem je

$$
\begin{aligned}
&\lim_{x\to\infty} F(x) = \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P(X \leq x_n) = \\
&\cup_{n=1}^{\infty} (X \leq x_n) = \Omega : \\
&\quad (\subseteq) : \text{logično} \\
&\quad (\supseteq) : \omega \in \Omega \implies \exists n \in \mathbb{N} : X(\omega) = x_n \\
&\qquad \implies \omega \in (X \leq x_n) \\
&\overset{P \text{ je zvezna}}{=} P(\cup_{n=1}^{\infty}(X \leq x_n)) = P(\Omega) = 1
\end{aligned}
$$

   Drugo pokažemo podobno (namesto $\cup$ je $\cap$) ■

4. F je zvezna z desne, t.j. $F(X+) = F(X) \forall x \in \mathbb{R}$

   Dokaz. obstoj limite ni problematičen: $F(x+) = \lim_{x\to 0} F(x+h) = \lim_{n\to\infty} F(x_n)$, kjer je $\{x_n\}_n \subseteq \mathbb{R}$ strogo naraščajoče zaporedje z limito v x
   $\{(X \leq x_n)\}_{n\in\mathbb{N}}$ je padajoče zaporedje s presekom

$$\begin{aligned}
&\{(X \leq x_n)\} = \cap_{n=1}^{\infty}\{\omega \in \Omega : X(\omega) \leq x_n\} = \\
&= \{\omega \in \Omega : X(\omega) \leq x\} = (X \leq x) : \\
&\qquad (\supseteq) : \text{ očitno} \\
&\qquad (\subseteq) : \omega \in \Omega \implies \text{za vsak n izpolnjeno} \implies \text{lim obstaja}
\end{aligned}$$

$$\begin{aligned}
F(x+) &= \lim_{n\to\infty} F(x_n) = \lim_{n\to\infty} P(X \leq x_n) = \\
&= P(\cap_{n=1}^{\infty}(X \leq x_n)) = P(X \leq x) = F(x)
\end{aligned}$$

■

5. $F(X-) = P(X < x) \neq F(x)$ v splošnem

$$\begin{aligned}
&P(x_1 < X \leq x_2) = P((X \leq x_2)\backslash(X \leq x_1)) = \\
&= P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1) \\
&P(x_1 < X < x_2) = P(X < x_2) - P(X \leq x_1) = F(x_2-) - F(x_1) \\
&P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1-) \\
&P(x_1 \leq X < x_2) = F(x_2-) - F(x_1-)
\end{aligned}$$

Opomba. V nekaterih učbenikih je porazdelitvena funkcija definirana z $F(x) = P(X < x)$ - zvezna z leve

Najpomembnejša razreda slučajnih spremenljivk sta

1.5.1 Diskretna slučajna spremenljivka

Definicija 1.14 (Diskretna slučajna spremenljivka)**.** Slučajna spremenljivka $X : \Omega \to \mathbb{R}$ je diskretno porazdeljena, če je njena zaloga vrednosti končna ali števna množica. Naj bo $\{x_1, x_2 \cdots\}$ zaloga vrednosti slučajne spremenljivke X.

Vpeljimo verjetnostno funkcijo $p_n := P(X = x_n)$ $n = 1, 2\cdots$. Potem je

$$\sum_n p_n = P(\cup_n(X = x_n)) = P(\Omega) = 1$$

in

$$\begin{aligned}&F_X(x) = P(X \leq x) = P(\cup_{n:x_n \leq x}(X = x_n)) = \\ &\text{paroma nezdružljivi dogodki} \\ &= \sum_{n:x_n\leq x} P(X = x_n) = \sum_{n:x_n\leq x} p_n\end{aligned}$$

npr. naj bodo $x_1 < x_2 < x_3$ v zalogi vrednosti slučajne spremenljivke X
F je odsekoma konstantna

$$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$$

Pomembnejše diskretne porazdelitve:

1.5.1.1 Enakomerna diskretna porazdelitev

na n točkah

$$X : \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

Primer. Met kocke, $X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$

1.5.1.2 Binomska porazdelitev

$Bin(n, p), n \in \mathbb{N}, p \in (0, 1), n-$krat ponovimo poskus, gledamo dogodek A z verjetnostjo $P(A) = p$, X je frekvenca dogodka A v n ponovitvah

$$X : \begin{pmatrix} 0 & 1 & \cdots & n \\ p_0 & p_1 & \cdots & p_n \end{pmatrix}$$

$$p_k = \binom{n}{k} p^k q^{n-k}$$

Primer. n-krat vržemo kocko. X je frekvenca šestice. $X \sim Bin(n, \frac{1}{6})$

1.5.1.3 Poissonova porazdelitev

$Poi(\lambda), \lambda > 0$

$$p_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \; k = 0, 1, 2 \cdots$$
$$\sum_{k=0}^{\infty} p_k = (\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}) e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1$$

$\implies$ to je res porazdelitev ($p_i \geq 0, \sum p_i = 1$)

Primer. Število klicev v telefonskem omrežju v časovni enoti
Binomska, velik n, majhen $p \implies$ Poissonova
Lahko modeliramo z binomsko porazdelitvijo $Bin(n, p)$, kjer je n število naročnikov in p verjetnost, da se posameznik odloči za klic v časovni enoti. Ker je n velik in p majhen, je to približno $Poi(\lambda)$, kjer je $\lambda = np$ (v praksi ni za vse ista)

Primer. Število napačnih črk v knjigi
(Veliko črk v knigi, melo verjetno, da se zmotimo.)
Lahko modeliramo z $Bin(n, p)$, kjer je n število vseh črk v knjigi, p je verjetnost, da si izberemo napačno črko
Ker je n velik, p pa majhen, lahko to aproksimiramo s $Poi(\lambda)$, kjer je $\lambda = np$
Raje vzamemo $Poi(\lambda)$ kot $Bin(n, p)$, ker je preprostejša

1.5.1.4 Geometrijska porazdelitev

$Geo(p), \; p \in (0, 1)$
Ponavljamo poskus, v katerem opazujemo dogodek A s $P(A) = p, q = 1 - p$.
$(X = k)$ je dogodek, da se A zgodi prvič v k-ti ponovitvi

$$p_k = P(A = k) = p \cdot q^{k-1} \; k = 1, 2 \cdots$$

$$\sum_{k=1}^{\infty} p_k = p \cdot \sum_{k=1}^{\infty} q^{k-1} = p \sum_{k=0}^{\infty} q^k = p \frac{1}{1-q} = \frac{p}{p} = 1$$

Primer. Mečemo kocko, X je število metov, da pade šestica prvič. Potem je $X \sim Geo(\frac{1}{6})$

1.5.1.5 Pascalova ali negativna binomska porazdelitev

$Pas(m, p), \ m \in \mathbb{N}, p \in (0, 1)$
Ponavljamo poskus, v katerem nas zanima dogodek A s $P(A) = p$. $(X = k)$ je dogodek, da se A zgodi m-tič v k-ti ponovitvi poskusa. Torej $Pas(1, p) = Geo(p)$

$$p_k = P(X = k) = \binom{k-1}{m-1} p^m q^{k-m} \ k = m, m+1 \cdots$$

(A se zgodi $(m-1)$-krat, $\overline{A}$ pa $(k-m)$-krat)
DN: Enakost $\sum_{k=m}^{\infty} p_k = 1$ analitično preverimo z $(m-1)$-kratnim odvajanjem geometrijske vrste

$$\sum_{k=0}^{\infty} q^{k-1} = \frac{1}{1-q}$$

oz. z direktno uporabo binomske vrste:

$$(1-q)^{-m} = \sum_{j=0}^{\infty} \binom{-m}{j} q^j$$

Primer. Mečemo kocko, X je število potrebnih metov, da pade šestica m-krat. Potem je $X \sim Pas(m, \frac{1}{6})$

1.5.1.6 Hipergeometrijska porazdelitev

$Hip(n; M, N), \ 0 < M < N, n, M, N \in \mathbb{N}, n \leq \min\{M, N-M\}$
V posodi je N kroglic, od tega M belih, ostale črne. Slučajno izberemo n kroglic (brez vračanja). X je število belih kroglic med izbranimi kroglicami. Torej $(X = k)$ je dogodek, da je med izbranimi n kroglicami k belih

$$p_k = P(X=k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \quad k = 0, 1 \cdots n$$

$\binom{m}{k} \cdots k$ belih

$\binom{N-m}{n-k} \cdots$ ostale črne

$\binom{n}{N} \cdots$ izberemo n izmed N

Ker je $\{(X=k)\}^n$ popoln sistem dogodkov, je jasno, da je $\sum_{k=0}^{n} p_k = 1$
Torej velja binomska identiteta

$$\sum_{k=0}^{n} \binom{m}{k}\binom{N-m}{n-k} = \binom{N}{n}$$

- verjetnostni dokaz

Primer. V ribniku je N rib, od tega M krapov. Ulovimo n rib. Naj bo X število ulovljenih krapov. Potem je $X \sim Hip(n; M, N)$

Če je $n << \min\{M, N-M\}$, potem je $Hip(n; M, n) \approx Bin(n, \frac{M}{N})$:

$$p_k = \frac{\frac{M(M-1)\cdots(M_k+1)}{k!} \frac{(N-m)(N-m+1)\cdots(N-m-n+k+1)}{(n-k)!}}{\frac{N(N-1)\cdots(N-n+1)}{n!}} \approx$$

$$\overset{\substack{k\leq m\\ n\leq N}}{\approx} \frac{\frac{M^k}{k!} \frac{(N-m)^{n-k}}{(n-k)!}}{\frac{N^n}{n!}} = \binom{n}{k}(\frac{M}{N})^k(\frac{N-M}{N})^{n-k} = \binom{n}{k}p^k q^{n-k}$$

Intuicija: vzemanje kroglic, $n << \min\{M, N-M\}$
Če je $n << \min\{M, N-M\}$, ne naredimo velike napake, če kroglice vračamo. Tedaj je število belih izvlečenih kroglic binomsko porazdeljeno: $X \sim Bin(n, \frac{M}{N})$

1.5.2 Zvezno porazdeljene slučajne spremenljivke

Definicija 1.15 (Zvezna porazdelitev)**.** Slučajna spremenljivka X je zvezno porazdeljena (zvezna), če obstaja nenegativa integrabilna funkcije p_X, imenovana gostota porazdelitve, da je

$$F_X(x) = \int_{-\infty}^{\infty} p_X(t)dt \text{ za } \forall x \in \mathbb{R}$$

Analogija z diskretnimi porazdelitvami: $F_X(x) = \sum_{n:X_n \leq x} p_k, X : \begin{pmatrix} x_1 & \cdots \\ p_1 & \cdots \end{pmatrix}$
Tedaj je F_X zvezna funkcija. V točkah, kjer je p_X zvezna, je F_X zvezno odvedljiva in velja $F_X'(x) = p_X(x)$
Ker je $\lim_{x\to\infty} F_X(x) = 1$, je $\int_{-\infty}^{\infty} p_X(t)dt = 1$
Za $x_1 < x_2$ velja

$$P(x_1 < X < x_2) = F_X(x_2-) - F_X(x_1+) = \int_{-\infty}^{x_2} p_X(t)dx - \int_{-\infty}^{x_1} p_X(t)dt = \int_{x_1}^{x_2} p_X(t)dt$$

Pomembnejše zvezne porazdelitve:

1.5.2.1 Enakomerna zvezna porazdelitev na $[a, b]$

$$p_X(x) = \begin{cases} \frac{1}{b-a} \text{ če } a < x < b \\ 0 \text{ sicer} \end{cases}$$

$$F_X(x) = \begin{cases} 0 \text{ če } x \leq a \\ \frac{x-a}{b-a} \text{ če } a < x < b \\ 1 \text{ če } x \geq b \end{cases}$$

Primer. Slučajno izberemo X na $[0, 1]$

1.5.2.2 Normalna ali Gaussova porazdelitev

$N(\mu, \sigma),\ \mu \in \mathbb{R}, \sigma > 0$

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

$N(0, 1) : p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ - standardizirana normalna porazdelitev

σ velik:
σ majhen:
Porazdelitvena funkcija:

$$
\begin{aligned}
F(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(\frac{t-\mu}{\sigma})^2} dt = \\
u &= \frac{t-\mu}{\sigma}, du = \frac{dt}{\sigma} \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{1}{2}u^2} du = \\
&= \frac{1}{\sqrt{2\pi}} (\int_{-\infty}^{0} \cdots + \int_{0}^{\frac{x-\mu}{\sigma}} \cdots) = \\
&= \frac{1}{2} + \Phi(\frac{x-\mu}{\sigma})
\end{aligned}
$$

Laplaceova integralaska formula pravi, da je $Bin(n,p) \approx N(np, \sqrt{npq})$ za velik n :

$$
P_n(k) = \frac{1}{\sqrt{2\pi npq}} - \frac{1}{2}(\frac{k-np}{\sqrt{npq}})^2
$$

Primer. Sistolični krvni tlak
Verjetnost, da ima slučajno oseba krvni tlak med 120 in 130 mmHg

1.5.2.3 Eksponentna porazdelitev

$$
\begin{aligned}
&Exp(\lambda), \lambda > 0 \\
&p(x) = \begin{cases} \lambda e^{-\lambda x} \lambda \geq 0 \\ 0 \quad \text{sicer} \end{cases}
\end{aligned}
$$

$$
F(x) = \begin{cases} 1 - e^{-\lambda x} \text{ če } x \geq 0 \\ 0 \quad \text{če } x \leq 0 \end{cases}
$$

Primer. Radioaktivni razpad
$F(x)$ je verjetnost, da se radioaktivni razpad zgodi pred trenutkom $x \in \mathbb{R}^+$

1.5.2.4 Porazdelitev gama

$$
\begin{aligned}
&\Gamma(b,c),\ b,c > 0 \\
&p(x) = \begin{cases} \frac{c^b}{\Gamma(b)} x^{b-1} e^{-cx} x > 0 \\ 0 \quad \text{sicer} \end{cases}
\end{aligned}
$$

Očitno je $Exp(\lambda) = \Gamma(1, \lambda)$

$$\Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$$

$$\begin{aligned}
&\int_{-\infty}^\infty p(x)dx = \frac{c^b}{\Gamma(b)} \int_0^\infty x^{b-1} e^{-cx} dx = \\
&t = cx, dt = cdx \\
&= \frac{c^b}{\Gamma(b)} \int_0^\infty (cx)^{b-1} e^{-cx} cdx = \\
&= \frac{1}{\Gamma(b)} \cdot \Gamma(b) = 1
\end{aligned}$$

- je porazdelitev

1.5.2.5 Porazdelitev $\chi^2(n)$

(hi-kvadrat), $n \in \mathbb{N}$, n je število prostorskih stopenj

$$\begin{aligned}
&\chi^2(n) = \Gamma(\frac{n}{2}, \frac{1}{2}) \\
&p(x) = \begin{cases} \frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} x > 0 \\ 0 \qquad \text{sicer} \end{cases}
\end{aligned}$$

1.5.2.6 Cauchyjeva porazdelitev

$$p(x) = \frac{1}{\pi(1+x^2)} \quad x \in \mathbb{R}$$

$$\begin{aligned}
F(x) &= \frac{1}{\pi} \int_{-\infty}^\infty \frac{dt}{1+t^2} = \frac{1}{\pi} \arctan t|_{-\infty}^x = \\
&= \frac{1}{\pi} \arctan x - \frac{1}{\pi} \cdot \frac{\pi}{2} = \frac{1}{\pi} \arctan x + \frac{1}{2}
\end{aligned}$$

Primer. Slučajna spremenljivka, ki ni niti zvezno niti disktretno porazdeljena Vržemo kovanec, če pade grbc, postavimo $X = 1$, če pade cifra, pa naj bo X

slučajno izbrano stevilo na $[0,2]$
Izračunamo porazdelitveno funkcijo:

$$F(x) = P(X \leq x) \stackrel{x\in[0,2]}{=} P(\text{grb}) \cdot P(X \leq x \mid \text{grb}) + P(\text{cifra}) \cdot P(X \leq x \mid \text{cifra})$$

Če je $0 \leq x \leq 1$, potem je

$$F(x) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{x}{2} = \frac{x}{4}$$

Če je $1 \leq x \leq 2$, potem je

$$F(x) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{x}{2} = \frac{1}{2} + \frac{x}{4}$$

$$F(x) = \begin{cases} 0 \text{ če } x \leq 0 \\ \frac{x}{4} \text{ če } 0 \leq x < 1 \\ \frac{1}{2} + \frac{x}{4} \text{ če } 1 \leq x \leq 2 \\ 1 \text{ če } x \geq 2 \end{cases}$$

Ker F ni zvezna funkcija, X ni zvezno porazdeljena
Ker F ni odsekoma konstantna, X ni diskretno porazdeljena

1.6 Slučajni vektorji

Definicija 1.16 (Slučajni vektor)**.** Naj bo (Ω, Φ, P) verjetnostni prostor. Slučajni vektor je n-terica slučajnih spremenljivk $x = (x_1 \cdots x_n) : \Omega \to \mathbb{R}^n$ z lastnostjo, da je množica

$$(X_1 \leq x_1 \cdots X_n \leq x_n) := \{\omega \in \Omega : X_1(\omega) \leq x_1 \cdots X_n(\omega) \leq x_n\}$$

dogodek za vse n-terice $x = (x_1 \cdots x_n)$, se pravi v Φ za $\forall x = (x_1 \cdots x_n) \in \mathbb{R}^n$

Definicija 1.17 (Porazdelitvena funkcija)**.** Porazdelitvena funkcija slučajnega vektorja $X = (X_1 \cdots X_n)$ je funkcija, definirana z

$$F_X(x) = F_{(X_1 \cdots X_n)}(x_1 \cdots x_n) := P(X_1 \leq x_1 \cdots X_n \leq x_n)$$

Torej $F_X : \mathbb{R}^n \to \mathbb{R}$

F_X ima podobne lastnosti kot v primeru $n = 1$
Očitno je $0 \leq F_X(x) \leq 1$ za $\forall x \in \mathbb{R}^n$, glede na vsako spremenljivko je F_X naraščajoča in z desne zvezna, velja še:

$$\lim_{\substack{x_1 \to \infty \\ \vdots \\ x_n \to \infty}} F_{(X_1 \cdots X_n)}(x_1 \cdots x_n) = 1$$

Definicija 1.18 (Robna porazdelitev)**.** Če pošljemo v ∞ samo nekatere spremenljivke, dobimo porazdelitveno funkcijo slučajnega podvektorja, npr.

$$\lim_{\substack{x_2 \to \infty \\ \vdots \\ x_n \to \infty}} F_{(X_1 \cdots X_n)}(x_1 \cdots x_n) = F_{X_1}(x_1)$$

ali pa

$$\lim_{x_n \to \infty} F_{(X_1 \cdots X_n)}(x_1 \cdots x_n) = F_{X_1 \cdots X_{n-1}}(x_1 \cdots x_{n-1})$$

Takim porazdelitvam rečemo robne (marginalne) porazdelitve

Oglejmo si dvorazsežni primer ($n = 2$):

$$(X, Y) : \Omega \to \mathbb{R}^2$$

za $\forall (x, y) \in \mathbb{R}^2$ je

$$(X \leq x, Y \leq y) := \{\omega \in \Omega : X(\omega) \leq x, Y(\omega) \leq y\}$$

dogodek
Porazdelitvena funkcija $F_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}$ je definirana z

$$\begin{aligned}
&F_{(X,Y)}(x, y) := P(X \leq x, Y \leq y) \\
&\lim_{x \to \infty} F_{(X,Y)}(x, y) = P(Y \leq y) = F_Y(y) \\
&\lim_{y \to \infty} F_{(X,Y)}(x, y) = P(X \leq x) = F_X(x)
\end{aligned}$$

Izrazimo $P(a < X \leq b, c < Y \leq d)$ s porazdelitveno fuknciјo $F(X, Y) = F$. To bo posplošitev formule

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

ki smo jo imeli v primeru $n = 1$

$$\begin{aligned}
&(X, Y) : \Omega \to \mathbb{R}^2 \text{ slučajni vektor} \\
&F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y) = P((x, y) \in (-\infty, x] \times (-\infty, y])
\end{aligned}$$

Izrazimo z $F_{(X,Y)} = F$ verjetnost $P(a < X < b, c < Y < d)$. To bo posplošitev formule $P(a < X < b) = F_X(b) - F_X(a)$
Najprej vzemimo posebni primer:

$$\begin{aligned}
&P(a < X \leq b, Y \leq d) = P((X \leq b, Y \leq d) \backslash (X \leq a, Y \leq d)) = \\
&= P(X \leq b, Y \leq d) - P(X \leq a, Y \leq d) = F(b, d) - F(a, b)
\end{aligned}$$

V splošnem primeru pa imamo

$$\begin{aligned}
&P(a < X \leq b, c < Y \leq d) = P((a < X \leq b, Y \leq d) \backslash (a < X \leq b, Y \leq c)) = \\
&= P(a < X \leq b, Y \leq d) - P(a < X \leq b, Y \leq c) = \\
&\overset{\text{fiks. y}}{=} (F(b, d) - F(a, d)) - (F(b, c) - F(a, c))
\end{aligned}$$

Torej je

$$P(a < X \leq b, c < Y \leq d) = F(b, d) - F(a, d) - F(b, c) + F(a, c)$$

Najpomembnejša razreda večrazsežnih porazdelitev sta

1.6.1 Diskretne porazdelitve

Definicija 1.19. Slučajni vektor $X = (X_1 \cdots X_n) : \Omega \to \mathbb{R}^n$ je diskretno porazdeljen, če je njegova zaloga vrednosti končna/števna množica točk v $\mathbb{R}^n$. Omejimo se na $n = 2 : \Omega \to \mathbb{R}^2$.

Naj bo $\{x_1, x_2 \cdots\}$ zaloga vrednosti slučajne spremenljivke X in $\{y_1, y_2 \cdots\}$ zaloga vrednosti slučajne spremenljivke Y. Potem je zaloga vrednosti vektorja (X, Y) vsebovana v $\{(x_i, y_i) : i = 1, 2 \cdots j = 1, 2 \cdots\}$.
Definiramo verjetnostno funkcijo $p_{ij} := P(X = x_i, Y = y_j) i = 1, 2 \cdots j = 1, 2 \cdots$
Ker je $\{(X = x_i, Y = y_j)\}_{ij}$ popoln sistem dogodkov, je $\sum_i \sum_j p_{ij} = 1$

$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$
$$p_i = P(X = x_i) = P(\cup_j (X = x_i, Y = y_j)) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} \; i = 1, 2 \cdots$$
če je $Y : \begin{pmatrix} y_1 & y_2 & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}$, je
$$q_j = P(Y = y_i) = P(\cup_i (X = x_i, Y = y_j)) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} \; j = 1, 2 \cdots$$

Primer. Met dveh kock: X število pik na 1. kocki, Y na 2.

1.6.2 Zvezne porazdelitve

Definicija 1.20. Slučajni vektor $X = (X_1 \cdots X_n)$ je zvezno porazdeljen, če obstaja integrabilna funkcija $p_X : \mathbb{R}^n \to \mathbb{R}$, imenovana gostota porazdelitve, da je

$$F_X(x) = F_{(X_1 \cdots X_n)}(x_1 \cdots x_n) = \int_{-\infty}^{x_1} dt_1 \int_{-\infty}^{x_2} dt_2 \cdots \int_{-\infty}^{x_n} p_X(t_1 \cdots t_n) dt_n \text{ za } \forall x = (x_1 \cdots x_n) \in \mathbb{R}^n$$

Ker je $\lim_{\substack{x_1 \to \infty \\ \vdots \\ x_n \to \infty}} F_X(x_1 \cdots x_n) = 1$, je

$$\int \cdots_{\mathbb{R}^n} \int p_X(t_1 \cdots t_n) dt_1 \cdots dt_n = 1$$

Za vsako Borelovo množico $A \subseteq \mathbb{R}^n$ (najmanjša σ-algebra z vsemi odprtimi pravokotniki) je

$$P(X \in A) \equiv P((x_1 \cdots x_n) \in A) = \int \cdots_A \int p_X(t_1 \cdots t_n) dt_1 \cdots dt_n$$

Omejimo se na $n = 2 : F_{(X,Y)}(x,y) = \int_{-\infty}^{x} du \int_{-\infty}^{y} p_{(X,Y)}(u,v)dv$
Robni porazdelitvi sta:

$$\begin{aligned} F_X(x) &= \lim_{y\to\infty} F_{(X,Y)}(x,y) = \text{ (brez utemeljevanja)} \\ &= \int_{-\infty}^{x} du \int_{-\infty}^{\infty} p_{(X,Y)}(u,v)dv \end{aligned}$$

ki ima gostoto

$$p_X(x) = \int_{-\infty}^{\infty} p_{(X,Y)}(x,y)dy$$

in

$$\begin{aligned} F_Y(y) &= \lim_{x\to\infty} F_{(X,Y)}(x,y) = \\ &= \int_{-\infty}^{y} dv \int_{-\infty}^{\infty} p_{(X,Y)}(u,v)du \end{aligned}$$

ki ima gostoto

$$p_Y(y) = \int_{-\infty}^{\infty} p_{(X,Y)}(x,y)dx$$

(ekvivalentno vsoti v diskretnem primeru).
Najpomembnejša dvorazsežna zvezna porazdelitev je normalna:

$$\begin{aligned} &N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho), \mu_x, \mu_y \in \mathbb{R}, \sigma_x, \sigma_y > 0, \rho \in (-1,1) \\ &p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}((\frac{x-\mu_x}{\sigma_x})^2 - 2\rho\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + (\frac{y-\mu_y}{\sigma_y})^2)} \\ &(\mu_x, \mu_y) \text{ premik, } (\sigma_x, \sigma_y) \text{ razteg} \\ &N(0,0,1,1,\rho) : p(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)} \end{aligned}$$

Nivojnice, izohipse se: $x^2 - 2\rho xy + y^2 = c$

- $\rho = 0$: krožnica
- $\rho \in (-1, 1)$: elipsa

Robni porazdelitvi sta

$$p_X(x) = \int_{-\infty}^{\infty} p(x,y)dx = \cdots = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2}$$

torej $X \sim N(\mu_x, \sigma_x)$. Podobno $Y \sim N(\mu_y, \sigma_y)$

Primer. Krvni tlak, X je sistolični, Y je diastolični krvni tlak
$\mu_x = 120, \mu_y = 75, \rho \doteq 0.7$

Dvorazsežna normalna porazdelitev je posebni primer večrazsežne normalne porazdelitve $N(\mu, A)$, kjer je $\mu = (\mu_1 \cdots \mu_n)^T$ in A pozitivno definitna matrika.
Gostota v točki $x = (x_1 \cdots x_n)^T$ je

$$p(X) = \sqrt{\frac{detA}{(2\pi)^n}} e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)}$$
$$(x-\mu)^T A(x-\mu) = \langle A(x-\mu), x-\mu \rangle$$

Za dokaz enakosti

$$\int \cdots_{\mathbb{R}^n} \int p(x) dx_1 \cdots dx_n = 1$$

izračunajmo integral

$$\int \cdots_{\mathbb{R}^n} \int e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)} dx_1 \cdots dx_n = \sqrt{\frac{(2\pi)^n}{detA}}$$

$N(\mu, A), \mu \in \mathbb{R}^n, A \in \mathbb{R}^{m\times n}$ pozitivna definitna matrika, t.j. sebi adjungirana matrika, za katero velja

$$x^T Ax = \langle Ax, x \rangle > 0 \ \forall x \in \mathbb{R}^n \backslash \{0^n\}$$

V točki $x = (x_1 \cdots x_n)^T$ je

$$p(x) = \sqrt{\frac{detA}{(2\pi)^n}} \cdot e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)}$$

Izračunajmo integral

$$\begin{aligned}
&\underbrace{\int \cdots \int}_{\mathbb{R}^n} e^{-\frac{1}{2}(x-\mu)^T A(x-\mu)} dx = \\
&y = x - \mu \implies dy = dx \\
&= \underbrace{\int \cdots \int}_{\mathbb{R}^n} e^{-\frac{1}{2}y^T Ay} dy
\end{aligned}$$

Ker je A pozitivna definitna matrika, obstaja ortogonalna matrika U in diagonalna matrika $D = diag(\lambda_1 \cdots \lambda_n)$, da je $A = U^T DU$

$$\begin{aligned}
&= \underbrace{\int \cdots \int}_{\mathbb{R}^n} e^{-\frac{1}{2}y^T U^T DUy} dy = \\
&z = Uy, y = U^T z, dy = |detU^T| dz = dz \\
&= \underbrace{\int \cdots \int}_{\mathbb{R}^n} e^{-\frac{1}{2}z^T Dz} dz = \\
&= \underbrace{\int \cdots \int}_{\mathbb{R}^n} e^{-\frac{1}{2}(\lambda_1 z_1^2 + \cdots + \lambda_n z_n^2)} dz_1 \cdots dz_n = \\
&= \int_{\mathbb{R}} e^{-\frac{1}{2}\lambda_1 z_1^2} dz_1 \cdots \int_{\mathbb{R}} e^{-\frac{1}{2}\lambda_1 z_n^2} dz_n =
\end{aligned}$$

Ker je $\int_{\mathbb{R}} e^{-\frac{1}{2}\lambda z^2} dz = \sqrt{\frac{2\pi}{\lambda}}$ - $z \in \mathbb{R}$ - s pomočjo Γ funkcije, Bronsterin, sledi iz

$$\frac{1}{\sqrt{2\pi}\sigma} = \int_{\mathbb{R}} e^{-\frac{1}{2}(\frac{x}{\sigma})^2} dx = 1$$

Gostota za $N(0, \sigma), \lambda := \frac{1}{\sigma^2}, \sigma = \frac{1}{\sqrt{\lambda}}$

$$= \sqrt{\frac{2\pi}{\lambda_1}} \cdot \cdots \cdot \sqrt{\frac{2\pi}{\lambda_1}} = \sqrt{\frac{(2\pi)^n}{detA}}$$

Torej je $\int \underbrace{\cdots}_{\mathbb{R}^n} \int p(x)dx = 1$
Dvorazšežni primer je posebni primer

$$A = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x \sigma_y} \\ -\frac{\rho}{\sigma_x \sigma_y} & \frac{q}{\sigma_y^2} \end{bmatrix}, \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$
$$detA = \frac{1}{1-\rho^2}\left(\frac{1}{\sigma_x^2 \sigma_y^2} - \frac{\rho^2}{\sigma_x^2 \sigma_y^2}\right) \stackrel{?}{=} \frac{1}{\sigma_x^2 \sigma_y^2}$$

$K = A^{-1} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$ kovariančna matrika (slučajnemu vektorju X, Y)

1.7 Neodvisnost slučajnih spremenljivk

Definicija 1.21 (Neodvisnost)**.** Slučjane spremenljivke $x_1, x_2 \cdots x_n$ v slučjanem vektorju $x = (x_1 \cdots x_n)$ so neodvisne, če je

$$F_X(x_1 \cdots x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n) \text{ za } \forall x \in \mathbb{R}^n$$

oziroma

$$P(X_1 \leq x_1, X_2 \leq x_2 \cdots X_n \leq x_n) = P(X_1 \leq x_1) \cdots P(X_n \leq x_n)$$

oziroma dogodki $(X_1 \leq x_1) \cdots (X_n \leq x_n)$ so neodvisni

Oglejmo si dvorazsežni diskretni primer

Trditev 1.22. Naj bo (X, Y) diskretno porazdeljen vektor:

$$p_{ij} = P(X = x_i, Y = y_j), p_i = P(X = x_i), q_j = P(Y = y_j)$$

Potem sta X in Y neodvisni $\iff p_{ij} = p_i \cdot q_j \ \forall i, j$

Dokaz. $F \equiv F_{(X,Y)}$ porazdelitvena funkcija vektorja (x, y)
$(\Rightarrow)$

$$
\begin{aligned}
&p_{ij} \overset{\text{def}}{=} P(X = x_i, Y = y_j) = \lim_{h \to 0} P(x_i - h < X \leq x_i, y_j - h < Y \leq y_j) = \\
&= \lim_{h \to 0} (F_X(x_i)F_Y(y_j) - F_X(x_i - h)F_Y(y_j) - F_X(x_i)F_Y(y_j - h) - F_X(x_i - h)F_Y(y_j - h)) = \\
&\overset{\text{neodv.}}{=} \lim_{h \to 0} (F_X(x_i) - F_X(x_i - h))(F_Y(y_j) - F_Y(y_j - h)) = \\
&= \lim_{h \to 0} P(x_i - h < X \leq x_i) \cdot P(y_j - h < Y \leq y_j) = \\
&= \lim_{h \to 0} P(x_i - h < X \leq x_i) \cdot \lim_{h \to 0} P(y_j - h < Y \leq y_j) = \\
&= P(X = x_i) \cdot P(Y = y_j) = p_i \cdot q_j
\end{aligned}
$$

($\Leftarrow$)

$$
\begin{aligned}
&F_{(X,Y)}(x, y) = P(X \leq x, Y \leq y) = P(\cup_{i:x_i \leq x} \cup_{j:y_j \leq y} (X = x_i, Y = y_j)) = \\
&\overset{\text{disjunktni}}{=} \sum_{i:x_i \leq x} \sum_{j:y_j \leq y} P(X = x_i, Y = y_j) = \\
&\overset{\text{predpostavka}}{=} \sum_{i:x_i \leq x} \sum_{j:y_j \leq y} p_i q_j = \\
&= (\sum_{i:x_i \leq x} p_i)(\sum_{j:y_j \leq y} q_j) = \\
&= P(X \leq x_i) \cdot P(Y \leq y_j) = F_X(x) \cdot F_Y(y)
\end{aligned}
$$

■

Torej sta X in Y neodvisni slučajni spremenljivki

Trditev 1.23. Naj bo (X, Y) zvezno porazdeljen slučajni vektore z gostoto $p(x, y)$. Potem sta X in Y neodvisni slučajni spremenljivki $\iff$ $p_{(X,Y)}(x, y) = p_X(x) \cdot p_Y(y)$ za (skoraj) vse $x, y \in \mathbb{R}$

Dokaz. (ideja): X in Y sta neodvisni, če $F_{(X,Y)}(x, y) = F_X(x) \cdot F_Y(y) \forall x, y \in \mathbb{R}$. Če parcialno odvajamo po x in po y, dobimo $p_{(X,Y)}(x, y) = p_X(x) \cdot p_Y(y)$. Obratno dobimo z integriranjem po x in po y ■

Primer. $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$. Tedaj je

$$
\begin{aligned}
&X \sim N(\mu_x, \sigma), Y \sim N(\mu_y, \sigma_y)\\
&X \text{ in } Y \text{ sta neodvisni } \iff \rho = 0\\
&p_{(X,Y)}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}((\frac{x-\mu_x}{\sigma_x})^2 + (\frac{y-\mu_y}{\sigma_y})^2)} =\\
&= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{1}{2}(\frac{x-\mu_x}{\sigma_x})^2} + \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2} = p_X(x) \cdot p_Y(y)
\end{aligned}
$$

$$
\begin{aligned}
N(0,0,1,1,\rho) &: x^2 - 2\rho xy + y^2 = c \text{ - ovojnica}\\
\rho = 0 &: x^2 + y^2 = c \text{ - krožnica}
\end{aligned}
$$

Trditev 1.24. Naj bo (X,Y) zvezno porazdeljen slučajni vektor. Potem sta X in Y neodvisni $\iff$ $p_{(X,Y)}(x,y) = f(x) \cdot g(y)$ za neki integrabilni funkciji f in g

Dokaz. ($\Rightarrow$) jasno na zadnji trditvi
($\Leftarrow$)

$$
\begin{aligned}
p_X(x) &= \int_{-\infty}^{\infty} p_{(X,Y)}(x,y)dy \overset{\text{predpostavka}}{=} f(x) \int_{-\infty}^{\infty} g(y)dy \text{ in}\\
p_Y(y) &= \int_{-\infty}^{\infty} p_{(X,Y)}(x,y)dx \overset{\text{predpostavka}}{=} g(y) \int_{-\infty}^{\infty} f(x)dx
\end{aligned}
$$

Ker je $\iint\limits_{\mathbb{R}^2} p_{(X,Y)}(x,y)dxdy = 1$, je

$$\int_{-\infty}^{\infty} f(x)dx \cdot \int_{-\infty}^{\infty} g(y)dy = 1 \text{ predpostavka}$$

Zato je $p_X(x) \cdot p_Y(y) = f(x) \cdot g(y) = p_{(X,Y)}(x,y)$, kar pomeni neodvisnost po prejšnji trditvi ∎

Izrek 1.25. Slučajni spremenljivki X in Y sta neodvisni $\iff$ za vsaki Borelovi množici $A, B \subseteq \mathbb{R}$ velja

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

t.j. dogodka $(X \in A)$ in $(Y \in B)$ sta neodvisna
(Borelova σ-algebra: najmanjša σ-algebra z odprtimi množicami)

Dokaz. $(\Leftarrow)$

$$
\begin{aligned}
&A = (-\infty, x], B = (-\infty, y] \\
&P(X \leq x, Y \leq y) = P(X \in (-\infty, x], Y \in (-\infty, y]) = \\
&= P(x \in (-\infty, x]) \cdot P(Y \in (-\infty, y]) = P(X \leq x)P(Y \leq y) \\
&\implies F_{(X,Y)}(x, y) = F_X(x) \cdot F_Y(y)
\end{aligned}
$$

$(\Rightarrow)$ izpustimo ■

1.8 Funkcije slučajnih spremenljivk in slučajnih vektorjev

Naj bo $X : \Omega \to \mathbb{R}$ slučajna spremenljivka in $f : \mathbb{R} \to \mathbb{R}$ zvezna. Potem je $Y := f \circ X : \omega \to \mathbb{R}$ tudi slučajna spremenljivka.

$f \circ X = f(X)$
saj je množica

$$
\begin{aligned}
&(Y \leq y) \equiv \{\omega \in \Omega : f(X(\omega)) \leq y\} = \\
&= \{\omega \in \Omega : f(X(\omega)) \in (-\infty, y]\} = \\
&= \{\omega \in \Omega : X(\omega) \in f^{-1}((-\infty, y])\} = \\
&= \{X \in f^{-1}((-\infty, y])\}
\end{aligned}
$$

dogodek, ker je $f^{-1}((-\infty, y])$ zaprta množica, torej Borelova.
y je funkcija slučajne spremenljivke X.
Naj bo f strogo naraščajoča funkcija z zalogo vrednosti (a, b), kjer je $-\infty \leq a < b \leq \infty$
Vzemimo $y \in (a, b)$. Potem je

$$
\begin{aligned}
&F_Y(y) \stackrel{\text{def}}{=} P(Y \leq y) = P(f \circ X \leq y) = \\
&\text{f naraščajoča} \to \text{obrnljiva} \\
&= P(X \leq f^{-1}(y)) = F_X(f^{-1}(y))
\end{aligned}
$$

kjer je $f^{-1} : (a, b) \to \mathbb{R}$ inverzna funkcija k funkciji f
če je $y \geq b$ je $F_Y(y) = 1$
če je $y \leq a$ je $F_Y(y) = 0$
Če je še f zvezno odvedljiva in X zvezno porazdeljena slučajna spremenljivka, potem je y tudi zvezno porazdeljena z gostoto Φ

$$\Phi_Y(y) = F_Y^{'}(y) = F_X^{'}(f^{-1}(y)) \cdot (f^{-1}(y))^{'}$$

za $y \in (a, b)$, če je $y \notin (a, b)$, je $p_Y(y) = 0$
Podobno ravnamo v primeru, ko je f strogo padajoča ((a, b) zaloga vrednosti)

$$\begin{aligned} F_Y(y) = P(Y \leq y) = P(f \circ X \leq y) = P(X \geq f^{-1}(y)) = \\ = 1 - P(X \leq f^{-1}(y)) = 1 - F_X(f^{-1}(y)-) \end{aligned}$$

Primer. $X \sim N(0, 1)$, $f(x) = kx + n, k \neq 0, n \in \mathbb{R}$
Vzemimo, da je $k > 0$. Definiramo $Y = f(X)$. Tedaj je

$$p_Y(y) = p_X(\frac{y - n}{k}) \cdot \frac{1}{k}$$

po formuli (prej).

$$y = kx + n \implies x = \frac{y - n}{k} = f^{-1}(y)$$

To je enako

$$p_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{y-n}{k})^2} \frac{1}{k}$$

torej je $Y \sim N(n, k)$
Če je $k < 0$, potem je $p_Y(y) = p_X(\frac{y-n}{k}) \cdot \frac{1}{-k}$, torej za poljuben $k \in \mathbb{R} \backslash \{0\}$ je $Y \sim N(n, |k|)$

Primer. $X \sim N(0, 1), f(x) = x^2$. Tedaj ima $Y = f(X) = X^2$ porazdelitveno funkcijo

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = 0$$

za $y \leq 0$ in

$$F_Y(y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) =$$
$$= \frac{1}{\sqrt{2\pi}\int_{-\sqrt{y}}^{\sqrt{y}}} e^{-\frac{x^2}{2}} dx \overset{e^{-\frac{x^2}{2}} \text{ soda}}{under} = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{x^2}{2}} dx$$

za $y \geq 0$
Gostota za Y pa je

$$p_Y(y) = F_Y'(y) = \frac{2}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\sqrt{y})^2} \cdot \frac{1}{2\sqrt{y}} =$$
$$= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}}$$

kar je $\chi^2(1)$, saj je

$$\chi^2(n) : p_X(x) = \frac{1}{2^{\frac{n}{2}}\gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

za $x > 0$, sicer $p_X(x) = 0$

Trditev 1.26. Če sta X in Y neodvisni slučajni spremenljivki, f in $g : \mathbb{R} \to \mathbb{R}$ zvezni funkciji, potem sta tudi $U = f(X)$ in $V = g(Y)$ neodvisni slučajni spremenljivki

Dokaz.

$F_{(U,V)}(u, v) = P(f(x) \leq u, g(y) \leq v) = P(X \in f^{-1}((-\infty, u]), Y \in g^{-1}((-\infty, v])) =$
$f^{-1}((-\infty, u])$ in $g^{-1}((-\infty, v])$ zaprti $\implies$ Borelovi
$\overset{\text{Borelov izrek}}{=} P(X \in f^{-1}((-\infty, u])) \cdot P(Y \in g^{-1}((-\infty, v])) =$
$= P(f(X) \leq u) \cdot P(g(Y) \leq v) = F_U(u) \cdot F_V(v)\ \forall u, v \in \mathbb{R}$

■

Izrek 1.27. Naj bo $X = (X_1 \cdots X_n) : \Omega \to \mathbb{R}^n$ slučajni vektor in $f = (f_1 \cdots f_m) : \mathbb{R}^n \to \mathbb{R}^m$ zvezna preslikava. Potem je $Y = f \circ X \equiv f(X) : \Omega \to \mathbb{R}^m$ tudi slučajni vektor (brez dokaza).

Y je funkcija slučajnega vektorja X in ima m komponent $Y = (Y_1 \cdots Y_m)$. Porazdelitvena funkcija za $Y_j, (j = 1, 2 \cdots m)$ je

$F_{Y_j}(y) = P(f_i(x_1 \cdots x_n) \leq y) = P((x_1 \cdots x_n) \in f_j^{-1}((-\infty, y]))$ množica v$\mathbb{R}^n$

Če je $X = (X_1 \cdots X_n)$ zvezno porazdeljena, je torej

$$F_{Y_j}(y) = \int \underbrace{\cdots}_{f^{-1}((-\infty,y])} \int p_X(x_1 \cdots x_n) dx_1 \cdots dx_n$$

Primer. $n = 2, m = 1$, $(x, y) : \Omega \to \mathbb{R}^2$ zvezno porazdeljen

$$\begin{aligned} F_Z(z) &= P(Z \leq z) = P(f(x,y) \leq z) = P((X,Y) \in f^{-1}((-\infty, z])) = \\ &= \iint\limits_{x+y \leq z} p_{(X,Y)}(x,y) dx dy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} p_{(X,Y)}(x,y) dy \end{aligned}$$

od tod sledi, da je gostota slučajne spremenljivke Z

$$p_Z(z) = F_Z'(z) = \int_{-\infty}^{\infty} p_{(X,Y)}(x, z-x) dx$$

Če sta še X in Y neodvisni slučajni spremenljivki, potem je

$$p_Z(z) = \int_{-\infty}^{\infty} p_X(x) \cdot p_Y(z-x) dx \text{ - konvolucija funkcij } p_X \text{ in } p_Y$$

Vzemimo posebni primer $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, torej

$$p_X(x) = \frac{1}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \text{ za } x > 0 \text{ in } 0 \text{ sicer}$$

za $p_Y(y)$ podobno.
Po zadnji formuli je $p_Z(z) = \int_{-\infty}^{\infty} p_X(x) \cdot p_Y(z-x) dx = 0$ za $z \leq 0$,
sicer je za $z > 0$

$$p_Z(z) = \frac{1}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})2^{\frac{m}{2}}\Gamma(\frac{n}{2})}e^{-\frac{z}{2}}\int_0^z x^{\frac{m}{2}-1}(z-x)^{\frac{m}{2}-1+\frac{n}{2}-1+1}e^{-\frac{x}{2}}e^{-\frac{z-x}{2}}dx =$$
$$= \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}e^{-\frac{z}{2}}\int_0^z x^{\frac{m}{2}-1}(z-x)^{\frac{n}{2}-1}dx =$$

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$$
$$x = tz \; dx = zdt$$

$$= \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}e^{-\frac{z}{2}}z^{\frac{m}{2}-1+\frac{n}{2}-1+1}\int_0^1 t^{\frac{m}{2}-1}(1-t)^{\frac{n}{2}-1}dt =$$

$$B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$$
$$\to B(\frac{m}{2},\frac{n}{2}) = \frac{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}{\Gamma(\frac{m+n}{2})}$$

$$= \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{m+n}{2})}e^{-\frac{z}{2}}z^{\frac{m+n}{2}-1}$$

Torej $X + Y \sim \chi^n(m+n)$
Dokazali smo

Trditev 1.28. Naj bosta neodvisni slučajni spremenljivki $X \sim \chi^2(m), Y \sim \chi^2(n)$z. Potem je $X + Y \sim \chi^2(m+n)$

Posledica 1.29. Če so $X_1, X_2 \cdots X_n$ neodvisne slučajne spremenljivke, porazdeljene $N(0,1)$, potem je $Y := X_1^2 + \cdots + X_n^2$ porazdeljena po $\chi^2(n)$

Dokaz. Vemo, da je $X_i^2 \sim \chi^2(1)$ in da so $X_1^2 \cdots X_n^2$ neodvisne spremenljivke. Potem je po trditvi + indukciji $Y \sim \chi^2(1 + \cdots + 1) = \chi^2(n)$ ■

Oglejmo si transformacijo $f : \mathbb{R}^2 \to \mathbb{R}, (x, y) \to (u, v)$, ki preslika zvezno porazdeljen slučajni vektor (x, y) v zvezno porazdeljen slučajni vektor (u, v), torej $U = u(x, y), V = v(x, y)$
Označimo še $A_{u,v} = (-\infty, u] \times (-\infty, v]$
Potem je

$$F_{(U,V)}(u, v) = \iint\limits_{A_{u,v}} p_{(U,V)}(s, t) ds dt$$

Pot drugi strani pa je

$$F_{(U,V)}(u, v) = P((U, V) \in A_{u,v}) = P((X, Y) \in f^{-1}(A_{u,v})) = \iint\limits_{f^{-1}(A_{u,v})} p_{(X,Y)}(x, y) dx dy$$

Privzemimo še, da je f zvezno odvedljiva bijekcija. Potem je $f : \mathbb{R}^2 \to \mathbb{R}^2, (u, v) \to (x, y)$ tudi zvezno odvedljiva. Z zamenjavo spremenljivk $x = X(u, v), y = Y(u, v)$ v zadnjem intergalu dobimo

$$F_{(U,V)}(u, v) = \iint\limits_{A_{u,v}} p_{(X,Y)}(x(s, t), y(s, t)) \cdot |J(s, t)| dx ds$$

kjer je

$$J(u, v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix} (u, v)$$

Jacobijeva determinanta.

Zaradi 1.8 imamo torej $p_{(U,V)}(u, v) = p_{(X,Y)}(x(u, v), y(u, v))|J(u, v)|$

Oglejmo si poseben primer

Primer. $U = X, V = v(x, y)$ oz $X = U, Y = y(u, v)$
Tedaj je $p_{(U,V)}(u, v) = p_{(X,Y)}(u, y(u, v))|\frac{\partial y}{\partial v}(u, v)|$
Gostota spremenljivke V je $\int_{-\infty}^{\infty} p_{(X,Y)}(u, y(u, v))|\frac{\partial y}{\partial v}(u, v)| dv = p_V(v)$
Pišimo $Z = V$, torej je $Y = y(x, z)$, saj je $U = X$
Potem prepišemo $p_Z(z) = \int_{-\infty}^{\infty} p_{(X,Y)}(u, y(x, z))|\frac{\partial y}{\partial z}(x, z)| dx$

Primer.

1. $Z = X + Y \implies Y = Z - X$, torej je $y(x,z) = z - x$, $\frac{\partial y}{\partial z}(x,z) = 1$

$$p_{X+Y}(z) = \int_{-\infty}^{\infty} p_{(X,Y)}(x, z-x) \cdot 1 dx$$

2. $Z = X \cdot Y \implies Y = \frac{Z}{X}$, torej je $y(x,z) = \frac{z}{x}$, $\frac{\partial y}{\partial z}(x,z) = \frac{1}{x}$

$$p_{X \cdot Y}(z) = \int_{-\infty}^{\infty} p_{(X,Y)}(x, \frac{z}{x}) \frac{1}{|x|} dx$$

Če sta še X in Y neodvisni slučajni spremenljivki, potem je

$$p_{X \cdot Y}(z) = \int_{-\infty}^{\infty} p_X(x) \cdot p_Y(\frac{z}{x}) \cdot \frac{1}{|x|} dx$$

1.9 Matematično upanje oz. pričakovana vrednost

V primeru $X : \begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$ je matematično upanje oz. pricakovana vrednost vsota $E(X) := \sum_{k=1}^{n} x_k \cdot p_k$

V posebnem primeru $p_1 = \cdots = p_n = \frac{1}{n}$ je $E(X) = \frac{1}{n} \sum_{k=1}^{n} x_k = \frac{x_1 + \cdots + x_n}{n}$ - povprečje števil $x_1 \cdots x_n$

expected value, expectation, mean value

Naj bo X diskretno porazdeljena slučajna spremenljivka z neskončno zalogo vrednosti:

$$X : \begin{pmatrix} x_1 & x_2 & x_3 & \cdots \\ p_1 & p_2 & p_3 & \cdots \end{pmatrix}$$

X ima matematično upanje oz. pričakovano vrednost, če je $\sum_{k=1}^{\infty} |x_k| p_k < \infty$
Tedaj je matematično upanje definirano kot $E(X) = \sum_{k=1}^{\infty} x_k \cdot p_k$
Naj bo sedaj X zvezno porazdeljena slučajna spremenljivka z gostoto p_X. Potem ima X matematično upanje, če je $\int_{-\infty}^{\infty} |x| \cdot p_X(x) dx < \infty$. Tedaje je matematično upanje slučajne spremenljivke X enako $E(X) = \int_{-\infty}^{\infty} x \cdot p_X(x) dx$

Primer.

1. $X \sim Ber(p)$ oz. $X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix} q = 1 - p, E(X) = 0 \cdot q + 1 \cdot p = p$

2. $X \sim Poi(\lambda)$, torej $p_k = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} k = 0, 1 \cdots$

$$E(X) = \sum_{k=0}^{\infty} k \cdot p_k = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \cdot \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda$$

3. Enakomerna porazdelitev na $[a, b]$

$$p(X) = \begin{cases} \frac{1}{b-a} & \text{če } a \leq x \leq b \\ 0 & \text{sicer} \end{cases}$$

$$E(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^2}{2} |_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

4. $X \sim N(\mu, \sigma)$

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{-\infty}{\infty} x \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx =$$

$$U = \frac{x - \mu}{\sigma} \implies du = \frac{dx}{\sigma}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma u + \mu) e^{-\frac{1}{2}u^2} du =$$
$$= \frac{1}{\sqrt{2\pi}} \sigma \int_{-\infty}^{\infty} u \cdot e^{-\frac{1}{2}u^2} du + \mu \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \cdot e^{-\frac{1}{2}u^2} du =$$
$$= \mu$$

Ker je v predzadnjem koraku 1. funkcija (v integralu) liha, 2. pa je gostota porazdelitve $N(0, 1)$

5. Cauchyjeva porazdelitev $p(x) = \frac{1}{\pi(1+x^2)}$
Nima matematičnega upanja, saj je $\int_{-\infty}^{\infty} |x| \cdot \frac{1}{\pi(1+x^2)} dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx = \frac{1}{\pi} ln(1 + x^2)|_0^{\infty} = \infty$

6. $1 - \frac{1}{2} + \frac{1}{3} - \cdots$ je pogojno konvergentna vrsta, t.j. konvergira, a ne

absolutno

$$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}, \sum_{k=1}^{\infty} x_k \cdot p_k = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k}$$

$$x_k \cdot p_k = \frac{(-1)^{k-1}}{k}$$

$$\sum_{k=1}^{\infty} p_k = 1$$

$$p_k := \frac{1}{2^k} \text{ npr. ker je vsota 1}$$

$$x_k := \frac{(-1)^{k-1}}{k} = 2^k$$

Ta porazdelitev nima matematičnega upanja, ker vrsta ne konvergira absolutno.

Trditev 1.30. Naj bo $f : \mathbb{R} \to \mathbb{R}$ zvezna funkcija

(a) Če je $X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$
potem je $E(f \circ X) \equiv E(f(X)) = \sum_{k=1}^{\infty} f(x_k) \cdot p_k$ (če le to matematično upanje obstaja)

(b) Če je X zvezno porazdeljena z gostoto p_X, potem je $E(f \circ X) = \int_{-\infty}^{\infty} f(x) \cdot p_X(x) dx$

Dokaz. (samo (a)):

$$f \circ X : \begin{pmatrix} f(x_1) & f(x_2) & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$$

npr če $f(x_1) = f(x_3) \implies \begin{pmatrix} f(x_1) & f(x_2) & \cdots \\ p_1 + p_3 & p_2 & \cdots \end{pmatrix}$
$(E(f \circ X) = \int_{-\infty}^{\infty} x \cdot p_{f(x)}(x) dx = \cdots = \int_{-\infty}^{\infty} f(x) \cdot p_X(x) dx$ - substitucija $y = f(x)$ v integralu) ∎

Posledica 1.31. Slučajna spremenljivka X ima matematično upanje $\iff$ X ima matematično upanje. Tedaj velja $|E(X)| \leq E(|X|)$

Dokaz. (samo diskreten primer):

$$E(|X|) \overset{\text{trd.} f(x)=|x|}{=} \sum_i |x_i| \cdot p_i \geq |\sum_i x_i \cdot p_i| = |E(X)|$$

∎

Posledica 1.32. Za $\forall a \in \mathbb{R}$ in vsako slučjano spremenljivko X z matematičnim upanjem velja $E(a \cdot X) = a \cdot E(X)$ (homogenost)

Dokaz. $f(x) = a \cdot x$, trditev (od prej) ∎

Podobno kot zadnjo trditev se dokaže

Trditev 1.33. Naj bo $f : \mathbb{R}^2 \to \mathbb{R}$ zvezna funkcija in (X, Y) slučajni vektor

(a) Naj bo (X, Y) diskretno porazdeljen $p_{ij} := P(X = x_i, Y = y_j)$. Potem je $E(f(X, Y)) = \sum_i \sum_i f(x_i, y_j) \cdot p_{ij}$ (če le vrsta (oz. končna vsota) absolutno konvergira)

(b) Naj bo (X, Y) zvezno porazdeljen z gostoto $p(X, Y)$. Potem je $E(f(X, Y)) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} f(x, y)p_{(X,Y)}(x, y)dy$ (če le integral absolutno konvergira)

Posledica 1.34. Če slučajni spremenljivki X in Y imata matamatično upanje, potem ga ima tudi $X+Y$ in velja $E(X+Y) = E(X)+E(Y)$ (aditivnost)

Dokaz. (samo zvezen primer):

$$
\begin{aligned}
E(X, Y) &\overset{f(x,y)=x+y}{=} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} (x + y)p_{(X,Y)}(x, y)dy = \\
&= \int_{-\infty}^{\infty} xdx \int_{-\infty}^{\infty} p_{(X,Y)}(x, y)dy + \int_{-\infty}^{\infty} ydy \int_{-\infty}^{\infty} p_{(X,Y)}(x, y)dx = \\
&= \int_{-\infty}^{\infty} xp_X(x)dx + \int_{-\infty}^{\infty} yp_Y(y)dy = E(X) + E(Y)
\end{aligned}
$$

∎

Posledica 1.35. Za slučajne spremenljivke $X_1 \cdots X_n$, ki imajo matematično upanje, velja $E(a_1X_1+\cdots+a_nX_n) = a_1E(X_1)+\cdots+a_nE(X_n)$ z $\forall a_1 \cdots a_n \in \mathbb{R}$

$$E(X + Y) = \int_{-\infty}^{\infty} x \cdot p_{X+Y}(x)dx \overset{?}{=} E(X) + E(Y) \text{ ni očitno iz tega}$$

Primer. 1. Če ima X matematično upanje, potem $E(X-E(X)) = E(X)-E(E(X)) = E(X) - E(X) = 0$

2. $X_k \sim Ber(p)$, t.j. $X_k \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, q = 1 - p$

$$X = X_1 + \cdots + X_n \implies E(X) = E(X_1) + \cdots + E(X_n) = n \cdot p$$

Posebej to (v 2. zgledu) velja v primeru, ko so $\{X_k\}_{i=1}^n$ neodvisne. To velja tudi za Bernoullijevo zaporedje ponovitev poskusa: opazujemo dogodek A s $P(A) = p$. X je frekvenca dogodka A v n ponovitvah poskusa. Potem je $X \sim Bin(n,p)$ in $X = X_1 + \cdots + X_n$, kjer je $(X_k = 1)$ dogodek, da se A zgodi v k-ti ponovitvi poskusa, sicer je $(X_k = 0)$. Po zgornjem je $E(X) = n \cdot p$. Do tega lahko pridemo tudi direktno:

$$E(X) = \sum_{k=0}^{n} k \cdot p_k = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k q^{n-k} =$$
$$= \sum_{k=1}^{n} k \cdot \frac{n}{k} \binom{n-1}{k-1} p^k q^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} \overset{j=k-1}{=}$$
$$= np(\sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j}) = np(p+q)^{n-1} = np$$

Trditev 1.36 (Cauchy-Schwartzova neenakost)**.** Če obstajata $E(X^2)$ in $E(Y^2)$, potem obstaja tudi $E(X,Y)$ in velja $E(|X \cdot Y|) \leq \sqrt{E(X^2) \cdot E(Y^2)}$. Enačaj velja samo v primeru $|Y| = \sqrt{\frac{E(Y^2)}{E(X^2)}}|X|$ z verjetnostjo 1

Dokaz. Ker za nenegativa realna števila velja neenakost

$$u \cdot v \leq \frac{1}{2}(u^2 + v^2) \iff (u-v)^2 \geq 0$$

za nenegativni slučajni spremenljivki U in V velja neenakost

$$U \cdot V \leq \frac{1}{2}(U^2 + V^2)$$

Enakost velja samo v točkah $\omega \in \Omega$, za katere je $U(\omega) = V(\omega)$
Če vstavimo $U = a \cdot |X|$ in $V = \frac{1}{a}|Y|$ za $a > 0$, dobimo $|X \cdot Y| \leq \frac{1}{2}(a^2 Y^2 + \frac{1}{a^2} Y^2)$ in zato je

$$E(|X \cdot Y|) \leq \frac{1}{2}(a^2 E(X^2) + \frac{1}{a^2} E(Y^2)) \text{ za } \forall a > 0 \tag{1}$$

Če vstavimo $a^2 = \sqrt{\frac{E(Y^2)}{E(X^2)}}$ na desni strani dobimo

$$\frac{1}{2}(\sqrt{E(Y^2) + E(X^2)} + \sqrt{E(X^2 + E(Y^2))}) = \sqrt{E(X^2) + E(Y^2)}$$

Torej je

$$E(|X \cdot Y|) \leq \sqrt{E(X^2) \cdot E(Y^2)}$$

Enakost v neenakosti velja $\iff a|X| = \frac{1}{a}|Y|$, torej $|Y| = a^2|X| = \frac{E(Y^2)}{E(X^2)}|X|$ z verjetnostjo 1 ■

Posledica 1.37. Če obstaja $E(X^2)$, potem obstaja $E(X)$ in velja $(E(X))^2 \leq E(X^2)$

Dokaz. $Y = 1$, t.j. $Y : \begin{pmatrix} 1 \\ 1 \end{pmatrix} \implies$

$$
\begin{aligned}
&E(|X \cdot 1|) \leq \sqrt{E(X^2) \cdot 1} /^2 \\
&(E(|X|))^2 \leq E(X^2)
\end{aligned}
$$

■

Trditev 1.38. Naj bosta X in Y neodvisni slučajni spremenljivki, ki imata matematični upanji. Potem ima matematično upanje tudi $X \cdot Y$ in velja $E(X \cdot Y) = E(X) \cdot E(Y)$

Dokaz. (samo zvezem primer):

$$
\begin{aligned}
E(X \cdot Y) &\overset{\text{trd}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \cdot y \cdot p_{(X,Y)}(x, y) dx dy \\
&\overset{\text{neodvisnost}}{=} \iint_{\mathbb{R}} x \cdot y \cdot p_X(x) \cdot p_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} x p_X(x) dx \cdot \int_{-\infty}^{\infty} x p_Y(y) dy \\
&= E(X) \cdot E(Y)
\end{aligned}
$$

■

Definicija 1.39 (Nekoreliranost)**.** Slučajni spremenljivki X in Y sta nekorelirani, če velja $E(X \cdot Y) = E(X) \cdot E(Y)$, sicer sta korelirani.

Po trditvi iz neodvisnosti sledi nekoreliranost. Obratno pa ne velja:

Primer.

$$U = \begin{pmatrix} 0 & \frac{\pi}{2} & \pi \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$
$$X = cos(U) : \begin{pmatrix} 1 & 0 & -1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$$
$$Y = sin(U) : \begin{pmatrix} 0 & 1 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$
$$E(X) = 0, E(Y) = \frac{1}{3}$$
$$X \cdot Y = sin(U) \cdot cos(U) = 0 \implies E(X \cdot Y) = 0 \implies$$
$$\implies \text{ X in Y sta nekorelirani slučajni spremenljivki}$$

X \Y	0	1	Σ
-1	$\frac{1}{3}$	0	$\frac{1}{3}$
0	0	$\frac{1}{3}$	$\frac{1}{3}$
-1	$\frac{1}{3}$	0	$\frac{1}{3}$
$\sum$	$\frac{2}{3}$	$\frac{1}{3}$	1

$\implies$ nista neodvisni, npr

$\frac{1}{3} = P(X = 1, Y = 0) \neq P(X = 1) \cdot P(Y = 0) = \frac{1}{3} \cdot \frac{2}{3}$

Trditev 1.40. $X : \begin{pmatrix} x_1 & x_2 \\ p_1 & p_2 \end{pmatrix}, Y : \begin{pmatrix} y_1 & y_2 \\ q_1 & q_2 \end{pmatrix}$
Potem sta X in Y neodvisni $\iff$ nekorelirani
$\iff E(X \cdot Y) = E(X) \cdot E(Y)$

1.10 Disperzija, kovarianco in korelacijski koeficient

Definicija 1.41 (Disperzija)**.** Naj obstaja $E(X^2)$. Disperzija oz. varianca slučajne spremenljivke X je $D(X) \equiv var(X) := E((X - E(X))^2)$

Disperzija meri razpršenost slučajne spremenljivke X okoli $E(X)$
Ker je $E((X-E(X))^2) = E(X^2-2E(X)X+(E(X))^2) = E(X^2)-2E(X)E(X)+(E(X))^2 = E(X^2) - (E(X))^2$, je $D(X) = E(X^2) - (E(X))^2$

Lastnosti disperzije:

- $D(X) \geq 0$ in $D(X) = 0 \iff P(X = E(X)) = 1$, t.j. X je izrojena slučajna spremenljivka
- $D(a \cdot X) = a^2 D(X)$ $a \in \mathbb{R}$
- $\forall a \in \mathbb{R}$ velja: $E((X - a)^2) \geq D(X)$. Enakost velja le v primeru $a = E(X)$

Dokaz.

$$\begin{aligned}E((X-a)^2) &= E(X^2-2aX+a^2) = E(X^2)-2E(x)|a|+a^2\\&= (a-E(X))^2+E(X^2)-(E(X))^2\\&= D(X)+(a-E(X))^2 \geq D(X)\end{aligned}$$

Enakost velja samo za $a = E(X)$ ■

Definicija 1.42 (Standardna deviacija)**.** Standardna deviacija ali standardni odklon slučajne spremenljivke X je $\sigma(X) := \sqrt{D(X)}$

Zanjo velja $\sigma(aX) = |a| \cdot \sigma(X)$ za $\forall a \in \mathbb{R}$
Primeri nekaterih $E(X)$ in $D(X)$

1. enakomerna diskretna porazdelitev: $\begin{pmatrix} x_1 & \cdots & x_n \\ \frac{1}{n} & \cdots & \frac{1}{n}\end{pmatrix}$

$$E(X) = \frac{x_1+\cdots+x_n}{n},$$
$$D(X) = E(X^2)-(E(X))^2 = \frac{x_1^2+\cdots+x_n^2}{2} - (\frac{x_1+\cdots+x_n}{2})^2$$

2. Binomska porazdelitev $Bin(n,p), n \in \mathbb{N}, p \in (0,1), q = 1-p$

$$E(X) = n \cdot p, D(X) = npq, \sigma(X) = \sqrt{npq}$$

3. Poissonova porazdelitev $Poi(\lambda), \lambda > 0$

$$E(X) = \lambda, D(X) = \lambda$$

4. Geometrijska porazdelitev $geo(p), p \in (0,1), q = 1-p$

$$E(X) = \frac{1}{p}, D(X) = \frac{q}{p^2}$$

5. Pascalova porazdelitev $Pas(m,p), m \in \mathbb{N}, p \in (0,1)$

$$E(X) = \frac{m}{p}, D(X) = \frac{mq}{p^2}$$

6. Enakomerna zvezna porazdelitev Ed na $[a,b]$

$$E(X) = \frac{a+b}{2}, D(X) = \frac{(b-a)^2}{12}$$

7. Normalna porazdelitev $N(\mu, \sigma)$

$$E(X) = \mu, D(X) = \sigma^2, \sigma(X) = \sigma$$

8. Porazdelitev gama $\gamma(b, c)$

$$E(X) = \frac{b}{c}, D(X) = \frac{b}{c^2}$$

9. Porazdelitev $\chi^2(n) = \gamma(\frac{n}{2}, \frac{1}{2})$

$$E(X) = n, D(X) = 2n$$

10. Eksponentna porazdelitev $Exp(\lambda), \lambda > 0 = \gamma(1, \lambda)$

$$E(X) = \frac{1}{\lambda}, D(X) = \frac{1}{\lambda^2}, \sigma(X) = \frac{1}{\lambda}$$

Preverimo, da je $D(X) = \sigma^2$ za $X \sim N(\mu, \sigma)$

$$D(X) = E((X - E(X))^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x - \mu)^2 \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

$$t = \frac{x - \mu}{\sigma} \implies x - \mu = \sigma t, dx = \sigma dt$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{1}{2}t^2} =$$

$$u = t, dv = t \cdot e^{-\frac{1}{2}t^2}$$
$$du = dt, v = -e^{-\frac{1}{2}t^2}$$

$$\frac{\sigma^2}{\sqrt{2\pi}} (-te^{-\frac{1}{2}t^2}|_{-\infty}^{\infty}) + \int_{-\infty}^{\infty} e^{-\frac{1}{2}t^2} dt = \qquad = \frac{\sigma^2}{\sqrt{2\pi}}(0 + \sqrt{2\pi}) = \sigma^2$$

Definicija 1.43 (Kovarianca)**.** Kovarianca slučajnih spremenljivk $K(X, Y) \equiv Cov(X, Y) := E((X - E(X))(Y - E(Y)))$

Ker je

$$
\begin{aligned}
E((X-E(X))(Y-E(Y))) &= E(XY - E(Y)X - E(X)Y + E(X)E(Y)) \\
&= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y)
\end{aligned}
$$

je $cov(X,Y) = E(XY) - E(X)E(Y)$

Lastnosti:

1. $K(X,X) = D(X)$
2. $K(X,Y) = 0 \iff X$ in Y sta nekorelirani
3. K je simetrična in bilinearna funkcija:
   - $K(X,Y) = K(Y,X)$
   - $K(aX+bY,Z) = aK(X,Z) + bK(Y,Z) \forall a,b \in \mathbb{R}$
4. Če obstajata $D(X)$ in $D(Y)$, potem obstaja tudi $K(X,Y)$. Tedaj velja

   $$|K(X,Y)| \le \sqrt{D(X)\cdot D(Y)} = \sigma(X)\cdot\sigma(Y)$$

   To sledi iz Cauchy-Schwartzove neenakosti ($|E(U \cdot V)| \le \sqrt{E(U^2)\cdot E(V^2)}$) za slučajni spremenljivki $X - E(X)$ in $Y - E(Y)$. Enačaj v neenakosti velja $\iff Y - E(Y) \pm \frac{\sigma(Y)}{\sigma(X)}(X - E(X))$ z verjetnostjo 1
5. Če X in Y imata disperziji, potem jo ima tudi $X+Y$ in valja $D(X+Y) = D(X) + D(Y) + 2K(X,Y)$
   če sta X in Y nekorelirani (posebej neodvisni), potem je $D(X+Y) = D(X) + D(Y)$

   Dokaz. Sledi iz enakosti

   $$
   \begin{aligned}
   (X+Y-E(X+Y))^2 &= ((X-E(X)) + (Y-E(Y)))^2 \\
   &= (X-E(X))^2 + (Y-E(Y))^2 + 2(X-E(X))(Y-E(Y))
   \end{aligned}
   $$

   $$
   \begin{aligned}
   D(X+Y) &= E((X-E(X))^2) + E((Y-E(Y))^2) + E(2(X-E(X))(Y-E(Y))) \\
   &= D(X) + D(Y) + 2K(X,Y)
   \end{aligned}
   $$

   ■

6. Posplošitev: $D(X_1 + \cdots + X_n) = D(X_1) + \cdots + D(X_n) + 2\sum_{i<j} K(X_i, X_j)$
   Če so $X_1 \cdots X_n$ paroma nekorelirani (posebej neodvisni), potem je $D(X_1 + \cdots + X_n) = D(X_1) + \cdots + D(X_n)$

Primer. $Bin(n,p)$ je vsota $X = X_1 + \cdots + X_n$, kjer je $X_i \sim Ber(p)$, t.j. $X_i \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, ki so neodvisne
Zato je $D(X) = D(X_1 + \cdots + X_n) = n \cdot D(X_1) = n \cdot p \cdot q$, saj je $D(X_n) = E(X_n^2) - (E(X_n))^2 = p - p^2 = pq$

Definicija 1.44 (Standardizacija slučajne spremenljivke)**.** Standardizacija skučajne spremenljivke X je slučajna spremenljivka $X_s = \frac{X - E(X)}{\sigma(X)}$

Zanjo velja:

- $E(X_s) = 0$
- $D(X_s) = \frac{1}{\sigma(x)^2} \cdot D(X - E(X)) = \frac{1}{\sigma(X)^2} D(X) = 1$

Primer.

$$X \sim N(\mu, \sigma) \implies X_s = \frac{X - E(X)}{\sigma(X)} = \frac{X - \mu}{\sigma} \sim N(0,1)$$

Definicija 1.45 (Korelacijski koeficient)**.** Korelacijski koeficient slučajnih spremenljivk X in Y je

$$r(X,Y) = \frac{K(X,Y)}{\sigma(X)\sigma(Y)} = \frac{E((X - E(X))(Y - E(Y)))}{\sigma(X)\sigma(Y)} = E(X_s \cdot Y_s)$$

Lastnosti:

1. $r(X,Y) = 0 \iff X$ in Y sta nekorelirani
2. $r(X,Y) \in [-1,1]$, kar sledi iz lastnosti (4) za kovarianco
3. - $r(X,Y) = 1 \iff Y = \frac{\sigma(Y)}{\sigma(X)}(X - E(X)) + E(Y)$ z verjetnostjo 1
   - $r(X,Y) = -1 \iff Y = -\frac{\sigma(Y)}{\sigma(X)}(X - E(X)) + E(Y)$ z verjetnostjo 1

   Tedaj imamo linearno zvezo med X in Y

Primer.

$$(X,Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)\ \mu_x, \mu_y \in \mathbb{R}, \sigma_x, \sigma_y \in [0, \infty], \rho \in [-1,1]$$

Trdimo, da je $r(X,Y)=\rho$

$$
\begin{aligned}
r(X,Y) &= E(X_s \cdot Y_s)\\
&= \frac{1}{2\pi\sqrt{1-p^2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} X\cdot Y\cdot exp(-\frac{1}{2(1-p)^2}\underbrace{(x^2-2\rho xy+y^2)}_{(x-\rho y)^2+(1-\rho^2)y^2})dxdy\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}}dy\cdot\underbrace{\frac{1}{\sqrt{2\pi(1-\rho^2)}}\cdot\int_{-\infty}^{\infty} x\cdot exp(-\frac{1}{2}(\frac{x-\rho y}{\sqrt{1-\rho^2}})^2)dx}_{E(N(\rho y,\sqrt{1-\rho^2}))=\rho y}\\
&= \rho\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y^2e^{-\frac{1}{2}y^2}dy\\
&= \rho
\end{aligned}
$$

Torej sta X in Y nekorelirani $\overset{\text{v splošnem}}{\Longleftrightarrow}$ $\rho=0$ $\overset{\text{ta primer}}{\Longleftrightarrow}$ X,Y neodvisni
Kakšna je gostota, če je ρ blizu 1? $\rho\uparrow 1:\rho\downarrow -1:$
gostota je škoraj skoncentriranaňa neki premici, torej med X in Y obstaja skoraj linearna zveza

1.11 Pogojna porazdelitev in pogojno matematično upanje

Izberimo si dogodek B s $P(B)>0$

Definicija 1.46. Pogojna porazdelitvena funkcija slučajne spremnljivke X glede na B je $F_X(X\mid B):=P(X\le x\mid B)=\frac{P(X\le x\wedge B)}{P(B)}$

Ima enake lastnosti kot porazdelitvena funkcija

A Diskreten primer
Naj bo (X,Y) diskretno porazdeljen slučajni vektor z verjetnostno funkcijo $p_{ij}=P(X=x_i,Y=y_i)i,j=1,2\cdots$
Za pogoj B vzemimo $B=(Y=y_j)$ pri nekem j, torej $q_j=P(Y=Y_j)$
Potem je pogojna porazdelitvena funkcija slučajne spremenljivke X glede $F_X(X\mid Y=y):=\frac{P(X\le x|Y=y_j)}{P(Y=y_j)}=\frac{1}{q_j}\sum_{j:x_j\le x}p_{ij}$
Če vpeljemo pogojno verjetnostno funkcije $P_{i|j}=P(X=x_i\mid Y=y_j)=\frac{p_{ij}}{q_j}$, $F_X(X\mid Y=y_j)=\sum_{i:x_i\le X}p_{i|j}$
Pogojno matematično upanje slučajne spremenljivke X glede na $Y=y_j$ je matematično upanje te porazdelitve:

$$E(X\mid Y=y_j):=\sum_i x_i\cdot p_{i|j}=\frac{1}{q_j}\sum_i x_j\cdot p_{ij}$$

Regresijska funkcija $\ell(y_j) = \sum(X \mid Y = y_j)$, ki je definirana na zalogi vrednoti slučajne spremenljivke Y

Definirajmo novo slučajno spremenljivko $E(X \mid Y) = \ell(y)$, ki ji rečemo pogojno matematično upanje slučajne spremenljivke X glede slučajne spremenljivke Y

Ta ima shemo $E(X \mid Y) = \begin{pmatrix} \ell(y_1) & \ell(y_2) & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix} = \begin{pmatrix} E(X \mid Y = y_1) & \cdots \\ q_1 & \cdots \end{pmatrix}$

Zanjo velja

$$E(X \mid Y) = \sum_j \ell(y_j) \cdot q_j = \sum_j \sum_i x_i \cdot p_{ij} = \sum_i x_i (\sum_j p_{ij}) = \sum_i x_i \cdot p_i = E(X)$$

kjer je $p_i = P(X = x_i)$

Kaj dobimo, če sta X in Y neodvisni slučajni spremenljivki?

Tedaj je $p_{i|j} = \frac{p_{ij}}{q_j} = \frac{p_i \cdot q_j}{q_j} = p_i$ in $\ell(y_j) = E(E(X \mid Y = y_j)) = \sum_i x_i \cdot p_{i|j} = \sum_i x_i \cdot p_i = E(X)$, torej je regresijska funkcija kar konstanta $E(X)$ oz. je $E(X \mid Y)$ izrojena slučajna spremenljivka z vrednostjo $E(X)$

Primer. Kokoš znese N jajc, kjer je $N \sim Poi(\lambda)$ z $\lambda > 0$. Iz vsakega jajca se z verjetnostjo $p \in (0, 1)$ izvali piščanec, neodvisno od drugih jajc. Naj bo K število piščancev Dolocino $E(K \mid N), E(K) in E(N \mid K)$

$$P(N = n) = \frac{\lambda^n}{n!} e^{-\lambda} \; n = 0, 1, 2 \cdots$$

$$P(K = k \mid N = n) = \binom{n}{k} p^k q^{n-k} \; k = 0, 1 \cdots n$$

$$\ell(n) = E(K \mid N = n) = E(Bin(n, p)) = n \cdot p$$

torej je $E(K \mid N) = \ell(n) = p \cdot N$

$$E(K \mid N) = \begin{pmatrix} p \cdot 0 & p \cdot 1 & p \cdot 2 & \cdots \\ P(N = 0) & P(N = 1) & P(N = 2) & \cdots \end{pmatrix}$$

$$E(K) = E(E(K \mid N)) = E(p \cdot N) = p \cdot E(N) = p \cdot \lambda$$

$$P(K = k) = \sum_{n=k}^{\infty} P(K = k \mid N = n) \cdot P(N \leq n) = \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} p^k q^{n-k} \cdot \frac{\lambda^n}{n!} e^{-\lambda} =$$

$$= \frac{1}{k!} e^{-\lambda} p^k \lambda^k \sum_{n=k}^{\infty} \frac{(qk)^{n-k}}{(n-k)!} = \frac{(p\lambda)^k}{k!} e^{-\lambda} e^{q\lambda} = \frac{(p\lambda)^k}{k!} e^{-p\lambda} \; k = 0, 1 \cdots n$$

Torej je $K \sim Poi(p \cdot \lambda)$

$$P(N = n \mid K = k) = \frac{P(N = n, K = k)}{P(K = k)} = \frac{P(K = k \mid N = n) \cdot P(N = n)}{P(K = k)} =$$

$$= \frac{n!p^k q^{n-k}}{k!(n-k)!} \cdot \frac{\lambda^n e^{-\lambda}}{n!} \cdot \frac{pk!e^{p\lambda}}{(p\lambda)^k} = \frac{(q\lambda)^{n-k}}{(n-k)!} \cdot e^{-q\lambda} n = k, k+1 \cdots$$

To je za k premaknjena Poissonova porazdelitev: $k + Poi(q\lambda)$
Potem je $\psi(k) = E(N \mid K = k) = E(k + Poi(qk)) = k + q \cdot \lambda$ in zato je $E(N \mid K) = \psi(k) = k \cdot q + \lambda$
Preizkus: $E(E(N \mid K)) = E(k + q \cdot \lambda) = p\lambda + q\lambda = \lambda = E(N)$ (ok)
Regresijsko premico je vpeljal Golten (1822-1911)

B Zvezni primer
Naj bo (X, Y) zvezno porazdeljen slučajni vektor z gostoto $p_{(X,Y)}(x, y)$.
Vzemimo $B = (y < Y \leq y + k)$ za nek $y \in \mathbb{R}, k > 0$.
Potem je $F_X(X \mid y < Y \leq y + k) = P(x \leq x \mid y < Y \leq y + k) = \frac{P(X \leq x, y < Y \leq y+k)}{P(y < Y \leq y+k)} = \frac{F_{(X,Y)}(x,y+k) - F_{(X,Y)}(x,y)}{F_Y(y+k) - F_Y(y)}$
Pogojna porazdelitvena funkcija slučajne spremenljivke X glede na dogodek $(Y = y)$ je limita, če obstaja:

$$F_X(x \mid Y = y) = \lim_{h \downarrow 0} F_X(x \mid y < Y \leq y+h) = \lim_{h \downarrow 0} \frac{F_{(X,Y)}(x, y+h) - F_{(X,Y)}(x, y)}{F_Y(y+h) - F_Y(y)}$$

Denimo sedaj, da sta $p_{X,Y}$ in p_Y zvezni funkciji. Tedaj je $F_X(X \mid Y = y) = \frac{\frac{\partial}{\partial y} F_{(X,Y)}(x,y)}{F'_Y(y)} = \frac{1}{p_Y(y)} \int_{-\infty}^{x} p_{(X,Y)}(x, v) dv$
Če vpeljemo pogojno pogojno gostoto $p_X(x \mid Y = y) := \frac{p_{(X,Y)}(x,y)}{p_Y(y)}$, je torej

$$F_{(X,Y)}(x \mid Y = y) = \int_{-\infty}^{x} p_X(u \mid y) du$$

Pogojno matematično upanje slučajne spremenljivke X glede na dogodek $(Y = y)$ je

$$E(X \mid Y = y) := \int_{-\infty}^{\infty} x \cdot p_X(x|y) dx = \frac{1}{p_Y(y)} \cdot \int_{-\infty}^{\infty} x p_{(X,Y)}(x, y) dx$$

Vpeljimo regresijsko funkcijo $l(y) := E(X \mid Y = y)$, definirano na zalogi vrednosti slučajne spremenljivke Y. Tako dobimo novo slučajno spremenljivko $E(X \mid Y) := l(y)$: pogojno matematično upanje slučajne spremenljivke X glede na slucajno spremenljivko Y.
Kot v diskretnem primeru se pokaže enakost $E(E(X \mid Y)) = E(X)$

Primer. $(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$
Robna gostota za Y je $N(\mu_y, \sigma_y)$
Zato je pogojna gostota

$$p_X(x \mid y) = \frac{p_{(X,Y)}(x, y)}{p_y(x)} \overset{\text{D.N.}}{=\cdots=} \frac{1}{\sigma_x \sqrt{(2\pi)(1-\rho^2)}} exp(-\frac{1}{2(1-\rho)^2}(\frac{x - \mu_x}{\sigma_x} - \rho \frac{y - \mu_y}{\sigma_y})^2)$$

torej je $N(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x\sqrt{1-\rho^2})$
Eksponent: $\frac{1}{2(1-\rho^2)}\sigma_x^2(x-(\mu_x+\rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)))^2$
$\implies l(y) = E(X \mid Y = y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)$ - 1. parameter
$= \alpha + \beta y : \beta = \rho\frac{\sigma_x}{\sigma_y}, \alpha = \mu_x - \frac{\sigma_x}{\sigma_y} \cdot \mu_y$
Torej je $E(x \mid y) = \alpha + \beta y$

Primer. Meritev onesnaženosti zraka
Slučajna spremenljivka X meri koncentracijo ogljikovih delcev (v $\mu g/m^3$), Y pa koncentracijo ozona (v $\mu l/l = ppm$)
Podatki kažejo, da ima (X,Y) približno dvorazsežno normalno porazdelitev, $\mu_x = 10.7, \sigma_x^2 = 29, \mu_y = 0.1, \sigma_y^2 = 0.02, \rho = 0.72$
Koncentracija ozona je škodljiva zdravju, če je ≥ 0.3
Denimo, da naprava za merjenje ozona odpove, koncentracija škodljivih delcev je $X = 200$

a kolikšna je pričakovana koncentracija ozona?

b kolikšna je verjetnost, da je stopnja ozona zdravju skodljiva

a

$$E(Y \mid X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x) = 0.1 + 0.72\sqrt{\frac{0.02}{29}}(20-10.7) \doteq 0.28$$

b Pogojna porazdelitev $Y \mid X = x$ je $N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_x\sqrt{1-\rho^2}) = N(0.28, 0.1)$

$$P(Y > 0.3 \mid X = 20) = 1 - P(Y \leq 0.3 \mid X = 20) = 1 - F_{N(0,1)}(\frac{0.3-0.28}{0.1}) \doteq 0.42$$

1.12 Višji momenti in vrstilne karakteristike

Definicija 1.47 (Momenti)**.** Naj bo $k \in \mathbb{N}$ in $a \in \mathbb{R}$. Moment reda k glede na točko a je $m_k(a) := E((X-a)^k)$ (če obstaja)

Za a običajno vzamemo

1. $a = 0$: $z_k := m_k(0) = E(X^k)$ začetni moment reda k
2. $a = E(X)$: $m_k := m_k(E(X))$ cenralni moment reda k

Ocitno je $z_1 = E(X), m_2 = D(X)$

Trditev 1.48. Če $\exists m_n(a)$, potem obstajaj tudi moment $m_k(a)$ za vse $k < n$

Dokaz. (V zveznem primeru):

$$\begin{aligned}
E((X-a)^k) &= \int_{-\infty}^{\infty} (x-a)^k p_X(x)dx \\
&= \int a - 1^{a+1}(X-a)^k p_X(x)dx + \int_{(-\infty,a-1)\cup(a+1,\infty)} (x-a)^k p_X(x)dx \\
&\leq \int_{-\infty}^{\infty} p_X(x)dx + \int_{(-\infty,a-1)\cup(a+1,\infty)} (x-a)^k p_X(x)dx \\
&\leq 1 + E((X-a)^k) \\
&< \infty
\end{aligned}$$

■

Trditev 1.49. Če obstaja zacetni moment z_n, potem obstaja $m_n(a)$ glede na poljubno točko $a \in \mathbb{R}$

Dokaz.

$$E((X-a)^n) \leq E((|X|+|a|)^n) = \sum_{k=0}^{n} \binom{n}{k} E(a)^{n-k} \cdot E(|X|^k) < \infty$$

■

Centralne momente lahko izrazimo z začetnimi:

$$m_n(a) = E((X-a)^n) = \sum_{k=0}^{n} \binom{n}{k} (-a)^{n-k} E(X^k)$$

$$a = E(X) \implies m_k = \sum_{k=0}^{n} \binom{n}{k} (-1)^{n-k} z_1^{n-k} z_k$$

Asimetrija slučajne spremenljivke X je $A(X) := E(X_s^3) = E((\frac{X-E(X)}{\sigma_x})^3) = \frac{m_3}{m_2^{\frac{3}{2}}}$ $m_2 = \sigma^2 = D(X)$
$A(N(\mu,\sigma)) = 0$, ker $A(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 e^{-\frac{1}{2}x^2} dx$
Sploščenost (kurtozis) $K(X) := E(X_s^4) = \frac{m_4}{m_2^2}$
$K(N(\mu,\sigma)) = 3$
Ce momenti ne obstajajo (npr. že $E(X)$ ne), potem si lahko pomagamo z vrstilnimi karakteristikami

Definicija 1.50 (Mediana)**.** Mediana slučajne spremenljivke X je vsaka vrednost $x \in \mathbb{R}$, za katero velja $P(X \leq x) \leq \frac{1}{2}$ in $P(Y \geq x) \geq \frac{1}{2}(1 - P(X < x) = 1 - F(x-))$

Če je F porazdelitvena funkcija za X, je to ekvivalentno s pogojem $F(x-) \leq \frac{1}{2} \leq F(x)$
Če je X zvezno porazdeljena slučajna spremenljivka, dobimo $F(X) = \frac{1}{2}$ oz. $\int_{-\infty}^{\infty} p(t)dx = \frac{1}{2}$

Te vrednosti (lahko jih je več) označimo z $X_{\frac{1}{2}}$

Primer.

- $X \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{5} & \frac{4}{5} \end{pmatrix}$
  $x_{\frac{1}{2}} = 1, E(X) = \frac{4}{5}$
- $X : \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{2}{4} \end{pmatrix}$
  Mediane so $[0, 1]$
-
- $X \sim N(0, 1)$
  $x_{\frac{1}{2}} = \mu = E(X)$

Definicija 1.51 (Kvantil)**.** Kvantil reda p ($p \in (0, 1)$) je vsaka vrednost x_p, za katero velja $P(X \leq x_p) \geq p$ in $P(X \geq x_p) \geq 1 - p$
Ekvivalentno je $F(x_p-) \leq p \leq F(x_p)$

Če je X zvezno porazdeljena, je pogoj $F(x_p) = p$ t.j. $\int_{-\infty}^{\infty} p(t)dt = p$

- Kvartili: $X_{\frac{1}{4}}, X_{\frac{2}{4}}, X_{\frac{3}{4}}$
- Percentili: $X_{\frac{1}{100}}, X_{\frac{2}{100}}, \cdots X_{\frac{99}{100}}$

Primer. Telesna višina odraslih moških

Definicija 1.52 ((Semiinter)kvartilni razmik)**.** $s := \frac{1}{2}(x_{\frac{3}{4}} - x_{\frac{1}{4}})$

je nadomestek (analog) za standardno deviacijo

Primer.

- $X \sim N(0, 1)$
  $X_{\frac{1}{2}} = 0$

$\int_{-\infty}^{\frac{1}{4}} p(t)dt = \frac{1}{4} \xrightarrow{\text{tabelca}} x_{\frac{1}{4}} \doteq -0.67$
$\xrightarrow{\text{simetrija}} x_{\frac{3}{4}} \doteq 0.67 \implies s = 0.67, \sigma(x) = 1$

- X naj ima Cauchyjevo porazdelitev
  $p(x) = \frac{1}{\pi(1+x^2)}$
  $x_{\frac{1}{2}} = 0$
  Momenti ne obstajajo

$$\int_{-\infty}^{x_{\frac{1}{4}}} \frac{1}{\pi} \frac{1}{1+x^2} dx = \frac{1}{4}$$
$$\frac{1}{\pi} \arctan x|_{x=-\infty}^{x_{\frac{1}{4}}} = \frac{1}{4}$$
$$\frac{1}{\pi} arctan x_{\frac{1}{4}} + \frac{1}{2} = \frac{1}{4}$$
$$arctan x_{\frac{1}{4}} = \frac{1}{4} \implies x_{\frac{1}{4}} = -1$$
$$\xrightarrow{\text{simetrija}} x_{\frac{3}{4}} = 1, s = 1$$

1.13 Rodovne funkcije

Definicija 1.53. Naj bo X slučajna spremenljivka z vrednostmi v $\mathbb{N} \cup \{0\}$:
$p_k = P(X = k) k = 0, 1, 2 \cdots\ p_k \geq 0, \sum_{k=0}^{\infty} = 1$
Rodovna funkcija skučajne spremenljivke X je

$$G_X(s) = p_0 + p_1 s + p_2 s^2 + \cdots = \sum_{k=0}^{\infty} p_k \cdots s^k$$

za $\forall s \in \mathbb{R}$, za katere vrsta absolutno konvergira.

Očitno je $G_X(0) = p_0, G_X(1) = \sum_{k=0}^{\infty} p_k = 1$
Ker je $s^X : \begin{pmatrix} s^0 & s^1 & s^2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \end{pmatrix}$, je $G_X(s) = E(s^X)$
Za $s \in [-1, 1]$ velja $|p_k \cdot s^k| \leq P_k$ in $\sum_{k=0}^{\infty} p_k = 1$. Zato je vrsta konvergentna, če je $|s| \leq 1$. Torej je konvergenčni radij vrste vsaj 1

Primer.

- $X \sim geo(p)$, $p \in (0,1)$

$$p_k = P(X=k) = p \cdot q^{k-1} \; k = 1,2,3\cdots$$
$$G_X(s) = \sum_{k=1}^{\infty} p \cdot q^{k-1} s^k = ps \sum_{k=0}^{\infty} (qs)^{k-1}$$
$$= ps \frac{1}{1-qs}$$

  konvergira, ko $|qs| < 1 \Leftrightarrow |s| < \frac{1}{|q|} =: R$

- $p_k = P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$

$$G_X(s) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} s^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} =$$
$$= e^{-\lambda} \cdot e^{\lambda s} = e^{\lambda(s-1)}$$

  $R = \infty \; \forall s \in \mathbb{R}$

Iz teorije Taylorjevih vrst sledi

Izrek 1.54 (O eniličnosti)**.** Naj imata X in Y rodovni funkciji G_X in G_Y. Potem je $G_X(s) = G_Y(s)$ za $\forall s \in [-1,1] \leftrightarrow P(X=k) = P(Y=k)$ za vse $k = 0,1,2\cdots$
Tedaj velja $P(X=k) = \frac{1}{k!} G_X^k(0)$

$G_X(s) = \sum_{k=0}^{\infty} p_k s^k$, $p_k = P(X=k)$
Naj ima rodovna funkcija G_X slučajne spremenljivke X konvergenčni radij R > 1. Potem za $\forall s \in (-R,R)$ velja $G_X^{'}(s) = \sum_{k=1}^{\infty} k \cdot p_k s^{k-1}$
Če postavimo $s = 1$, dobimo $G^{'}(1) = \sum_{k=1}^{\infty} k \cdot p_k = E(X)$

Izrek 1.55. Naj ima X rodovno funkcijo $G_X(s)$ in naj bo $n \in \mathbb{N}$. Potem je

$$G_X^n(1-) \equiv \lim_{s \nearrow 1} G_X^n(s) = E(X(X-1)(X-2)\cdots(X-N+1))$$

Dokaz. Za $\forall s \in [0,1)$ je $G_X^n(s) = \sum_{k=n}^{\infty} k(k-1)(k-2)\cdots(k-n+1) p_k s^{k-n+1} =$
$$= E(X(X-1)(X-2)\cdots(X-n+1) \cdot s^{X-n})$$

Ko gre $s \uparrow 1$, z uporabo Abelove leme dobimo

$$\lim_{s \nearrow 1} G_X^n(s) = \lim_{s \nearrow 1} \sum_{k=n}^{\infty} k(k-1) \cdot (k-n+1) =$$

$$\overset{\text{Abelova lema}}{=} \sum_{k=n}^{\infty} lim_{s \nearrow 1} k(k-1) \cdot (k-n+1) = \sum_{k=n}^{\infty} k(k-1) \cdot (k-n+1) p_k = E(X(X-1)\cdots(X-n+1))$$

■

Posledica 1.56.

$$E(X) = G_X^{'}(1-)$$

$$D(X) = E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2 = G_X^{(2)}(1-) + G_X^{(1)}(1) - (G_X^{(1)}(1-))^2$$

Izrek 1.57. Naj bosta X in Y neodvisni slučajni spremenljivki z rodovnima funkcijama G_X in G_Y. Potem je $G_{X+Y}(s) = G_X(s) \cdot G_Y(s)$ za $s \in [-1, 1]$

Dokaz. $G_{X+Y}(s) = E(s^{X+Y}) = E(s^X \cdot s^Y) \overset{\text{izrek}}{=} E(s^X) \cdot E(s^Y) = G_X(s) \cdot G_Y(s)$, saj sta s^X in s^Y neodvisni slučajni spremenljivki ■

Posplo"sitev 1.58. Če so $X_1, X_2 \cdots X_n$ neodvisne slučajne spremenljivke, potem je za vse $s \in [-1,1] G_{X_1+\cdots+X_n}(s) = G_{X_1}(s) \cdot \cdots \cdot G_{X_n}(s)$.
Če so $X_1, X_2 \cdots X_n$ enako porazdeljene in neodvisne, potem je

$$G_{X_1+\cdots+X_n}(s) = (G_X(s))^n$$

Izrek 1.59. Naj bodo za $\forall n \in \mathbb{N}$ slučajne spremenljivke $N, X_1, X_2 \cdots X_n$ neodvisne. Naj ima N rodovno funkcijo G_N, X_n pa rodovno funkcijo G_X. Potem ima slučajna spemenljivka $S := X_1 + X_2 + \cdots + X_n$ rodovno funkcijo enako $G_S = G_N \circ G_X$ oz. $G_S(s) = G_N(G_X(s))$ za $s \in [-1, 1]$

To je posplošitev formule dd: $P(N = n) = 1, G_N(s) = 1 \cdot s^n = s^n$

Dokaz. Zaradi neodvisnosti imamo $P(S = k) = \sum_{n=0}^{\infty} P(S = k, N = n) =$

$$= \sum_{n=0}^{\infty} P(N = n, X_1 + \cdots + X_n = k) \overset{\text{neodvisnost}}{=} \sum_{n=0}^{\infty} P(N = n) \cdot P(X_1 + \cdots + X_n = k)$$

Zato je

$$G_S(s) = \sum_{k=0}^{\infty} P(S = k) \cdot s^k = \sum_{k=0}^{\infty} \sum_{n=1}^{\infty} P(N = n) \cdot P(X_1 + \cdots + X_n = k) \cdot s^k =$$

$$= \sum_{n=1}^{\infty} P(N = n) (\sum_{k=0}^{\infty} P(X_1 + \cdots + X_n = k) \cdot s^k) =$$

$$\overset{G_{X_1+\cdots+X_n}(s)^{\text{neodvisnost}}(G_X(s)^n)}{=} \sum_{n=1}^{\infty} P(N = n) \cdot (G_X(s))^n = G_N(G_X(s))$$

za vse $s \in [-1, 1]$ ■

Posledica 1.60. Pri predpostavkah iz izreka velja Waldova enakost:

$$E(S) = E(N) \cdot E(X)$$

Dokaz.

$$G_S(s) = G_N(G_X(s)) \forall s \in [-1,1] \tag{2}$$

$$E(S) = G_s^{'}(1-) = G_N^{'}(G_X(1-)) \cdot G_X^{'}(1-) = E(N) \cdot E(X) \tag{3}$$

■

Primer. Kokoš, jajca, piščanci
N jajc, $N \sim Poi(\lambda)$
K je število piščancev
Definiramo $X_i = 1$ dogodek, da se iz i-tega jajca izvali piščanec, sicer $X_i = 0$.
Potem je $X_i : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, $q = 1 - p$ in X_i so neodvisne slučajne spremenljivke.
Očitno je $K = X_1 + X_2 + \cdots + X_n$
Ker je $G_N(s) = e^{\lambda(s-1)}$ in $G_X(s) = q \cdot s^0 + p \cdot s = q + ps$, je po izreku $G_K(s) = G_N(G_X(s)) = e^{\lambda(q+ps-1)} = e^{\lambda(ps-p)} = e^{\lambda p(s-1)} \forall s \in [-1,1]$, zato je $K \sim Poi(\lambda p)$

1.14 Momentno rodovna funkcija

Definicija 1.61 (Momentno rodovna funkcija)**.** Momentno rodovna funkcija je $M_X(t) = E(e^{tX})$ za $t \in \mathbb{R}$, za katere obstaja matematično upanje

V primeru zvezne porazdelitve je $M_X(t) = \int_{-\infty}^{\infty} e^{tx} p_X(x) dx$
To je Laplaceova transformacija funkcije p_X
V diskretnem primeru $X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$ je $M_X(t) = \sum_i e^{tx} p_i$

V posebnem primeru, ko ima X nenegative celoštevilske vrednosti, je $M_X(t) = \sum_{i=0}^{\infty} e^{it} p_i =$

$$= \sum_{i=0}^{\infty} p_i (e^t)^i = G_X(e^t) \; (M_X(t) = E((e^t)^X) = G_X(e^t))$$

$$G_X(s) = E(s^X)$$

Očitno je $M_X(0) = E(e^0) = E(1) = 1$

Primer.

$$X \sim N(0,1)$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx =$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}} dx \cdot e^{-\frac{t^2}{2}} =$$

$$= e^{\frac{t^2}{2}} \forall t \in \mathbb{R}$$

ker je $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-t)^2}{2}}$ gostota za $N(0,1)$

Izrek 1.62. Naj bo $M_X(t) < \infty$ (obstaja, $< \infty$ zato, ker je $e^t > 0$) za $\forall t \in (-\delta, \delta)$ pri nekem $\delta > 0$. Potem je porazdelitev za X natanko določena z M_X, vsi začetni momenti obstajajo, $z_k = E(X^k) = M_X^k(0)$ za $\forall k \in \mathbb{N}$ in velja $M_X(t) = \sum_{k=0}^{\infty} \frac{z_k}{k!} t^k$ za $\forall t \in (-\delta, \delta)$

Dokaz. (bistvo)

$$M_X(t) = E(e^{t \cdot X}) = E(\sum_{k=0}^{\infty} t^k \frac{x^k}{k!}) =$$

$$\sum_{k=0}^{\infty} \frac{E(X^k)}{k!} t^k = \sum_{k=0}^{\infty} \frac{z^k}{k!} t^k$$

■

Trditev 1.63. $M_{aX+b}(t) = e^{bt} M_X(at), a \neq 0, b \in R$

Dokaz. $M_{aX+b}(t) = E(e^{t(aX+b)}) = E(e^{(at)X} \cdot e^{bt}) = e^{bt} M_X(at)$ ■

Izrek 1.64. Če sta X in Y neodvisni slučajni spremenljivki, potem je $M_{X+Y}(t) = M_X(t) \cdot M_Y(t)$

Dokaz. $M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX} \cdot e^{tY}) \overset{e^{tX},\, e^{tY} \text{ neodvisni}}{=}$
$= E(e^{tX}) \cdot E(e^{tY}) = M_X(t) \cdot M_Y(t)$ ■

Trditev 1.65. Naj bosta X in Y neodvisni slučajni spremenljivki in $X \sim N(\mu_x, \sigma_x), Y \sim N(\mu_y, \sigma_y)$. Potem je $X + Y \sim N(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2})$

Dokaz. Ker je

$$U := \frac{X - \mu_x}{\sigma_x} = \frac{X - E(X)}{\sigma(X)} \sim N(0,1)$$

(standardizacija), je

$$X = \sigma_x \cdot U + \mu_x$$

in zato je

$$M_X(t) = e^{\mu_x t} \cdot M_U(\sigma_x t)$$

po zadnji trditvi. Potem je

$$M_U(t) = e^{\frac{t^2}{2}}$$

je

$$M_X(t) = e^{\mu_x t} \cdot e^{\frac{\sigma_x^2 t^2}{2}} = e^{\frac{\sigma_x^2 t^2}{2} + \mu_x t} \; \forall t \; in \mathbb{R}$$

za Y velja podobno. Po zadnjem izreku je

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t) = e^{\frac{\sigma_x^2 t^2}{2} + \mu_x t} \cdot e^{\frac{\sigma_y^2 t^2}{2} + \mu_y t} =$$

$$= e^{\frac{(\sigma_x^2 + \sigma_y^2)t^2}{2} + (\mu_x + \mu_y)t}$$

Po izreku je

$$X + Y \sim N(\mu_x + \mu_y, \sqrt{\sigma_x^2 + \sigma_y^2})$$

■

Opomba. Če bi vedeli, da je $X + Y$ porazdeljena normalno, bi "samo" izračunali parametra

Primer.

$$X \sim N(0,1), M_X(t) = e^{\frac{t^2}{2}} = \sum_{k=0}^{\infty} \frac{(\frac{t^2}{2})^k}{k!} = \sum_{k=0}^{\infty} \frac{1}{2^k \cdot k!} t^{2k}$$

Po drugi strani je $M_X(t) = \sum_{j=0}^{\infty} \frac{z_j}{j!} t^j \; \forall t \in \mathbb{R}$
Primerjamo koeficiente:

- lihi koeficienti: $z_{2k-1} = 0 \; k \in \mathbb{N}$
- sodi koeficienti:

$$\frac{z_{2k}}{(2k)!} = \frac{1}{k! 2^k} \implies z_{2k} = \frac{(2k)!}{k! 2^k} =$$

$$= \frac{1 \cdot 2 \cdot 3 \cdot \dots \cdot (2k)}{2 \cdot 4 \cdot 5 \cdot \dots \cdot (2k)} = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2k-1) = (2k-1)!! \; k \in \mathbb{N}$$

1.15 Šibki in krepki zakon velikih števil

Definicija 1.66 (Verjetnostna konvergenca)**.** Zaporedje slučajnih spremenljivk $\{X_n\}_{n \in \mathbb{N}}$ verjetnostno konvergira proti skučajni spremenljivki X, če za $\forall \epsilon > 0$ velja

$$\lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0$$

oziroma

$$\lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1$$

Definicija 1.67 (Skoraj gotova konvergenca)**.** Zaporedje slučajnih spremenljivk $\{X_n\}_{n\in\mathbb{N}}$ skoraj gotovo konvergira proti skučajni spremenljivki X, če velja

$$P(\lim_{n\to\infty} X_n = X) = 1$$

Tukaj je

$$\begin{aligned}(\lim_{n\to\infty} X_n = X) &= \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} \\ &= \{\omega \in \Omega : \forall k(\in \mathbb{N})\exists m \in \mathbb{N}\forall n \geq m : |X_n(\omega) - X(\omega)| < \frac{1}{k}\} \\ &= \{\cap_{k\in\mathbb{N}} \cup_{m\in\mathbb{N}} \cap_{n\geq m}\omega \in \Omega : |X_n(\omega) - X(\omega)| < \frac{1}{k}\}\end{aligned}$$

Opomba. Števne unije in preseki $\implies$ smo v $\sigma-$algebri, torej je to res dogodek

Trditev 1.68. Če $X_n \xrightarrow{n\to\infty} X$ skoraj gotovo, potem za $\forall\epsilon > 0 \lim_{n\to\infty} P(|X_n - X| < \epsilon \text{ za } n \geq m) = 1$

Dokaz. Označimo $c_m := (|X_n - X| < \epsilon \text{ za } n \geq m) = \cap_{n=m}^{\infty}(|x_n - X| < \epsilon)$.
Potem je $c_1 \subseteq c_2 \subseteq \cdots$
je c_m za $\epsilon = \frac{1}{k}$ in $(\lim_{n\to\infty} X_n = X) \subseteq \cup_{n=1}^{\infty} c_m$ (presek)
Torej je $1 = P(\lim_{n\to\infty} X_n = X) \subseteq= (\cup_{m=1}^{\infty} c_m) = \lim_{m\to\infty} P(c_m)$
Od tod sledi $\lim_{m\to\infty} P(c_m) = 1$ ■

Posledica 1.69. Če $X_n \xrightarrow{n\to\infty} X$ skoraj gotovo, potem $X_n \xrightarrow{n\to\infty} X$ verjetnostno konvergira.

Dokaz. Izberemo $\epsilon > 0$. Potem velja

$$P(|X_n - X| < \epsilon \text{ za } \forall n \geq m) \leq P(|X_m - X| < \epsilon)$$

Če uporabimo trditev, dobimo $\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$ (leva stran). ■

Opomba. Obratna implikacija ne velja

Definicija 1.70. Naj bo $X_1, X_2, X_3 \cdots$ zaporedje slučajnih spremenljivk, ki imajo matematično upanje. Definirajmo $Y_n = \frac{S_n - E(S_n)}{n} = \frac{X_1 + \cdots + X_n}{n} - \frac{E(X_1) + \cdots + E(X_n)}{n}$
Potem je $E(Y_n) = 0$
Za $\{Y_n\}_{n\in\mathbb{N}}$ velja šibki zakon velikih števil (ŠZVŠ), kadar $Y_n \overset{n\to\infty}{\to} 0$ verjetnostno, torej za $\forall\epsilon > 0 \lim_{n\to\infty}(|y| < \epsilon) = 1 = \lim_{n\to\infty}(|\frac{S_n - E(S_n)}{n}| < \epsilon)$ Za

$\{Y_n\}_{n\in\mathbb{N}}$ velja krepki zakon velikih števil (KZVŠ), kadar $Y_n \overset{n\to\infty}{\to} 0$ skoraj gotovo, torej $P(\lim_{n\to\infty} \frac{S_n - E(S_n)}{n} = 0) = 1$
Če velja KVZŠ, potem velja ŠVZŠ

Primer. Mečemo kocko, X_k je # pik v k-tem metu. Potem je $E(X_k) = \frac{7}{2}$ in $Y_n = \frac{X_1+\cdots+X_n}{n} - \frac{7}{2}$
Ali konvergira $\frac{X_1+\cdots+X_n}{n} \overset{n\to\infty}{\to} \frac{7}{2}$ skoraj gotovo? (Da)

Izrek 1.71.

a Neenakost Markova: če slučajna spremenljivka X ima matematično upanje, potem je $P(|X| \geq a) \leq \frac{E(|X|)}{a}$ za $\forall a > 0$

b Neenakost Čebiševa: če slučajna spremenljivka X ima disperzijo, potem je $P(|X - E(X)| \geq a \cdot \sigma(x)) \leq \frac{1}{a^2}$ za $\forall a > 0$ (pomembno za $a \geq 1$, ker je verjetnost ≤ 1)
oz. če pišemo $\epsilon = a \cdot \sigma(x) \implies P(|X - E(X)| \geq \epsilon) \leq \frac{D(X)}{\epsilon^2}$ za $\forall \epsilon > 0$

Dokaz. (samo zvezni primer)

a

$$E(X) = \int_{-\infty}^{\infty} |x| p_x(x) dx \geq \int_{\{x:|x|\geq a\}} |x| p_x(x) dx \geq$$
$$|a| \int_{\{x:|x|\geq a\}} p_x(x) dx = a \cdot P(|X| \geq a)$$

b

$$P((X - E(X)) \geq \epsilon) = P((X - E(X))^2 \geq \epsilon^2) \overset{\text{(a) za X-E(X)}}{\leq} \frac{E((X - E(X))^2)}{\epsilon^2} = \frac{D(X)}{\epsilon^2}$$

■

Izrek 1.72 (Markov)**.** Če za zaporedje slučajnih spremenljivk $\{X_n\}_{n\in\mathbb{N}}$ velja $\frac{D(S_n)}{n^2} \overset{n\to\infty}{\to} 0$, potem velja ŠZVŠ. Tukaj je $S_n := X_1 + \cdots + X_n$

Dokaz. V neenakosti Čebiševa vzamemo $X = \frac{S_n}{n}$

$$P(\frac{|S_n - E(S_n)|}{n} \geq \epsilon) \leq \frac{P(S_n)}{n^2 \epsilon^2} \overset{n\to\infty}{\to} 0$$

Če vzamemo $Y_n = \frac{|S_n - E(S_n)|}{n}$, je $P(|Y_n| \geq \epsilon) \overset{n\to\infty}{\to} 0$
oz. $P(|Y_n| < \epsilon) \overset{n\to\infty}{\to} 1$
Zato $Y_n \overset{n\to\infty}{\to} 0$ verjetnostno, torej velja ŠZVŠ za zaporedje $\{X_n\}_{n\in\mathbb{N}}$ ■

Posledica 1.73 (Izrek Čebišev)**.** Če so $X_1, X_2 \cdots X_n$ paroma nekorelirane slučajne spremenljivke in $\sup_{n\in\mathbb{N}} D(X_n) < \infty$, potem za $\{X_n\}_{n\in\infty}$ velja ŠVZŠ

Dokaz. Ker je $D(S_n) = D(X_1)+\cdots+D(X_n) \leq n\cdot c$, je $\frac{D(S_n)}{n^2} \leq \frac{n\cdot c}{n^2} = \frac{c}{n} \overset{n\to\infty}{\to} 0$, zato po izreku Markova velja ŠZVŠ ■

Primer. $X_n : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ neodvisne slučajne spremenljivke, $D(X_n) = pq, E(X_n) = p, E(S_n) = n \cdot p$
Po izreku Čebiševa velja ŠZVŠ: $P(\frac{|S_n - E(S_n)|}{n} \geq \epsilon) \overset{n\to\infty}{\to} 0$

$$\implies P(|\frac{S_n}{n} - p| \geq \epsilon) \overset{n\to\infty}{\to} 0$$

S_n je frekvenca dogodka, $\frac{S_n}{n}$ je relativna frekvenca, $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n} \overset{n\to\infty}{\to} p$ verjetnostno
To je Bernoulijev zakon velikih števil iz 1713

Izrek 1.74 (Kolmogorov)**.** Če za neodvisne slučajne spremenljivke $\{X_n\}_{n\in\mathbb{N}}$ velja $\sum_{n=1}^{\infty} \frac{D_n}{n^2} < \infty$, potem velja KZVŠ, t.j. $P(\lim_{n\to\infty} \frac{S-n-E(S_n)}{n} = 0) = 1$. Posebej je pogoj za vrsto izpolnjen, če je $\sup_n D(X_n) < \infty$

Primer. $X_n : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ neodvisne slučajne spremenljivke, $D(X_n) = pq$
Po izreku Kolmogorova velja KVZŠ, t.j. $\frac{S_n}{n} = \frac{X_1+\cdots+X_n}{n} \overset{n\to\infty}{\to} p$ skoraj gotovo.
To posplošuje Bernoullijev zakon

1.16 Centralni limitni izrek

Definicija 1.75. Naj bo $\{X_n\}_{n\in\mathbb{N}}$ zaporedje slučajnih spremenljivk s končnimi disperzijami. Definiramo $S_n := X_1 + \cdots + X_n$ in standardizirajmo:

$$Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)}$$

torej imamo

$$E(Z_n) = 0, D(Z_n) = 1$$

Za $\{X_n\}_{n\in\mathbb{N}}$ velja centralni limitni izrek, če je

$$F_{Z_n}(x) = P(Z_n \leq x) \overset{n\to\infty}{\to} F_{N(0,1)} \forall x \in \mathbb{R}$$

to je

$$P(\frac{S_n - E(S_n)}{\sigma(S_n)} \leq x) \overset{n\to\infty}{\to} \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dx \text{ za } \forall x \in \mathbb{R}$$

Pravimo, da $\{Z_n\}_{n\in\mathbb{N}}$ po porazdelitvi konvergira proti standardizirani normalni porazdelitvi.

Izrek 1.76 (Centralni limitni izrek (CLI, osnovna verzija)). Naj bodo $X_1, X_2 \cdots$ neodvisne in enako porazdeljene slučajne spremenljivke. Potem zanje velja centralni limitni zakon, t.j

$$P(\frac{S_n - E(S_n)}{\sigma(S_n)} \leq x) \overset{n\to\infty}{\to} \int_{-\infty}^{x} e^{\frac{t^2}{2}} dx \text{ za } \forall x \in \mathbb{R}$$

Dokazal je Ljapunov (1900), s tem je posplošil Laplaceov izrek iz leta 1812. V dokazu bomo uporabili

Izrek 1.77 (O zveznosti rodovne funkcije). Naj za zaporedje $\{Z_n\}_{n\in\mathbb{N}}$ slučajnih spremenljivk velja:
$M_{Z_n}(t) \to M_{N(0,1)}(t) = e^{\frac{t^2}{2}}$ za vse $t \in (-\delta, \delta)$ pri nekem $\delta > 0$
Potem $F_{Z_n}(x) \to F_{N(0,1)}(x)$ za $\forall x \in \mathbb{R}$

Dokaz. CLI v primeru, ko X_n imajo momentno rodovno funkcijo
$M_X(t) = E(e^{tX_n})$ na neki okolici točke 0
Naj bo $E(X_n) = \mu, D(X_n) = \sigma^2$ in $U_n := X_n - \mu = X_n - E(X_n)$. Torej je $E(U_n) = 0$ in $D(U_n) = \sigma^2$ ter $M_U(t) = 1 + tE(U_n) + \frac{t^2}{2!}E(U_n^2) + o(t^2) =$
$= 1 + \frac{t^2}{2}\sigma^2 + o(t^2)$ $(\lim_{n\to\infty} \frac{o(n)}{n} = 0)$
Ker je $D(S_n) \overset{\text{neodvisne}}{=} D(X_1) + \cdots + D(X_n) = n \cdot \sigma^2$ in $E(S_n) = n \cdot \mu = E(X_1) + \cdots + E(X_n)$, je $Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)} =$
$= \frac{1}{\sigma\sqrt{n}} \left(\sum_{n=0}^{n} U_i\right)$
Potem je $M_{Z_n}(t) = E(e^{tZ_n}) = E(e^{\frac{t}{\sigma\sqrt{n}}(U_1+\cdots+U_n)}) = E(e^{\frac{t}{\sigma\sqrt{n}}U_1}) \cdot \cdots \cdot E(e^{\frac{t}{\sigma\sqrt{n}}U_n}) =$
$\overset{\text{enaki}}{=} (M_U(\frac{t}{\sigma\sqrt{n}}))^n = (1 + \frac{t^2}{2n} + o(\frac{1}{n}))^n$
$\overset{n\to\infty \equiv o(\frac{1}{n}\to 0)}{\to} e^{\frac{t^2}{2}}$

Lema 1.78. Če $X_n \to X$, potem $(1 + \frac{X_n}{n})^n \overset{n\to\infty}{\to} e^x$

Po prejšnjem izreku: $F_{Z_n}(x) \overset{n\to\infty}{\to} F_{N(0,1)}(x)$

$$\epsilon > 0 : x - \epsilon \leq x_n \leq x + \epsilon \text{ za dovolj velik n}$$

$$\implies (1+\frac{x-\epsilon}{n})^n \leq (1+\frac{x_n}{n})^n \leq (1+\frac{x+\epsilon}{n})^n$$
$$\implies (1+\frac{x-\epsilon}{n})^n \to e^{x-\epsilon}$$
$$\implies (1+\frac{x_n}{n})^n \to e^x$$
$$\implies (1+\frac{x+\epsilon}{n})^n \to e^{x+\epsilon}$$

■

V splošnem se CLI dokaže s pomočjo karakterističnih funkcij:
naj bo X slučajna spremenljivka, $\ell_X(t) := E(e^{itX}) = E(cos(tX)) + iE(sin(tX)) t \in \mathbb{R}$
za razliko od momentno rodovnih funkcij karakteristične funkcije vedno odstajajo
v zveznem primeru je $\int_{-\infty}^{\infty} e^{itx} p(x) dx$ - Fourierova transformacija funkcije $p_X(x)$
$X_1, X_2 \cdots X_n$ neodvisne, enako porazdeljene

$$\mu := E(X_n), \sigma := \sigma(X_n)$$
$$E(S_n) \overset{\text{neodvisnost}}{=} E(X_1) + \cdots + E(X_n) = n\mu$$
$$D(S_n) \overset{\text{neodvisnost}}{=} D(X_1) + \cdots + D(X_n) = n\sigma^2$$

$X_1, X_2 \cdots X_n$ neodvisne slučajne spremenljivke

$$Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)} = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
$$\overline{Z_n} := \frac{S_n}{n} = \frac{X_1 + \cdots + X_n}{n} \implies Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Po CLI za velike n velja $Z_n \approx N(0,1)$, zato je $\overline{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$ oz. $S_n \approx N(n\mu, \sigma\sqrt{n})$
Če so $X_1, X_2 \cdots$ porazdeljene normalno $N(\mu, \sigma)$, potem je $Z_n \sim N(0,1)$, torej $F_{Z_n}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$

Primer. Laplaceova formula je poseben primer CLI:
$X_n : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, $X_n = 1$ je dogodek, da se dogodek A (s $P(A) = p$) zgodi v n-ti ponovitvi poskusa, sicer je $X_n = 0$
$E(X_n) = p, S_n = X_1 + \cdots + X_n$ frekvenca dogodka A v prvih n ponovitvah
$S_n \sim Bin(n,p), E(S_n) = np, D(S_n) = npq$, ker je $D(X_1) = pq$
$Z_n = \frac{S_n - E(S_n)}{\sigma(S_n)} = \frac{S_n - np}{\sqrt{npq}} \overset{\text{CLI}}{\approx} N(0,1)$, če je n velik

$$\begin{aligned} & P(S_n \leq X) = P(\frac{S_n - np}{\sqrt{npq}} \leq \frac{X - np}{\sqrt{npq}}) \approx \\ & \approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-np}{\sqrt{npq}}} e^{-\frac{t^2}{2}} dt = \\ & = \frac{1}{2} + \Phi(\frac{x - np}{\sqrt{npq}}) \end{aligned}$$

kjer je

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

verjetnostni integral

$$\begin{aligned} & P(\alpha < S_n \leq \beta) = \\ & = P(S_n \leq \beta) - P(S_n \leq \alpha) \approx \\ & \approx \frac{1}{2} + \Phi(\frac{\beta - np}{\sqrt{npq}}) - \frac{1}{2} - \Phi(\frac{\alpha - no}{\sqrt{npq}}) = \\ & = \Phi(\frac{\beta - np}{\sqrt{npq}}) - \Phi(\frac{\alpha - np}{\sqrt{npq}}) \end{aligned}$$

Laplaceova aproksimacijska formula

Primer. Teža vrečke kostanja je porazdeljena približno normalno, saj je vsota tež posameznih kostanjev, ki so neodvisne, enako porazdeljene slučajne spremenljivke
$X_n \cdots$ teža n-tega kostanja, $S_n = X_1 + \cdots + X_n \approx$ normalno - aditiven efekt

Primer.

$$p_{X_n}(x) = \begin{cases} \frac{1}{2}; x \in [-1,1] \\ 0 \text{ sicer} \end{cases}$$

$$E(X_1) = 0, D(X_1) = \frac{(b-a)^2}{12} = \frac{1}{3}$$

$$S_1 = X_1, Z_1 = \frac{X_1 - E(X_1)}{\sigma(X_1)} = \frac{X_1}{\sqrt{\frac{1}{3}}} = x_1\sqrt{3}$$

$$S_2 = X_1 + X_2, Z_2 = \frac{S_2 - E(S_2)}{\sigma(S_2)} = \frac{X_1 + X_2 - E(X_1 + X_2)}{\sigma(X_1 + X_2)}$$

$$S_3 = X_1 + X_2 + X_3, Z_3 = \frac{S_3 - E(S_3)}{\sigma(S_3)}$$

2 Statistika

2.1 Osnovni pojmi

Kot vedo statistiko razdelimo na:

1. opisno statistiko: zbiranje, razvrščanje, prikazovanje podatkov, računanje osnovnih količin
2. analitično statistiko: upraba podatkov pri sklepanju glede zakonitosti danega področja

Definicija 2.1 (Populacija)**.** Populacija je končna ali neskončna množica elementov, pri katerih merimo ali opazujemo neko količino

Primer.

(a) kontrole kvalitete: populacija je množica (serija) izdelka, npr. dnevna proizvodnja, merimo lastnosti izdelkov, npr. življensko dobo

(b) testiranje seb: populacija je množica vseh zaposlenih v državi, merimo npr. starost, višino place $\cdots$

Matematični pogled: na verjetnostnem prostoru $(\Omega, \mathcal{F})$ imamo slučajno spremenljivko X.
Praviloma ne moremo izmeriti cele populacije, ampak meritve opravimo na relativno majhnem delu populacije, na vzorcu. Le-ta mora biti reprezentativen, izbran nepristransko in dovolj velik.
Matematični pogled: vzorec velikosti n je slučajni vektor $(x_1 \cdots x_n)$, kjer so

komponente enako porazdeljene kot slučajna spremenljivka X in med seboj neodvisne.
Vrednost tega slučajnega vektorja pri enem naboru n meritev je realizacija vzorca: $(x_1 \cdots x_n)$: to so konkretni podatki, ki jih analiziramo. Pri opisni statistiki predstavimo in obdelamo te podatke.
Iz teh vzorčnih podatkov želimo oceniti nekatere lastnosti populacije, kot sta:

1. sredina populacije μ, t.i. matematično upanje slučajne spremenljivke X
2. povprečni odklon σ od sredine populacije, t.i. Standardna deviacija slučajne spremenljivke X

Ocene za μ so:

- vzorčno povprečje: $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$
- vzorčni modus: najpogostejša vrednost v vzorcu
- vzorčna mediana: srednja vrednost v vzorcu, urejenem po velikosti

Ocene za σ so:

- vzorčni razmak: razlika med največjo in najmanjšo vrednostjo v vzorcu
- vzorčna disperzija: $s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- popravljena vzorčna disperzija: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1} s_0^2$

2.2 Vzorčne statistike in cenilke

Definicija 2.2 (Vzorčna statistika)**.** Naj bo $(X_1, X_2 \cdots X_n)$ vzorec t.i. slučajni vektor, kjer so $X_1 \cdots X_n$ enako porazdeljene kot slucajna spremenljivka X in med seboj neodvisne.
Vzorčna statistika je simetrična funkcija vzorca $y = g(X_1, X_2 \cdots X_n)$, kjer je g simetricna funkcije n spremenljivk

Praviloma vzorčna statistika ocenjuje vrednost nekega parametra ξ. Tedaj je y cenilka za parameter.
y je odvisna od n, zato pišemo tudi $y_n = g(X_1 \cdots X_n)$.

Definicija 2.3 (Nepristranskost, doslednost)**.** Če je $E(Y) = \xi$, je Y nepristranska cenilka za parameter xi
Cenilka $Y = Y_n$ je dosledna, če $Y_n \overset{n \to \infty}{\to} \xi$ verjetnostno, t.i. $\forall \epsilon > 0$ je $\lim_{n \to \infty} P(|Y_n - \xi| \geq \epsilon) = 0$ oz. $\lim_{n \to \infty} P(|Y_n - \xi| < \epsilon) = 1$

Definicija 2.4 (Standardna napaka)**.** Standardna napaka vzorčne statistike Y je standardna deviacija slučajne spremenljivke Y: $SE(Y) := \sigma(Y)$

Definicija 2.5 (Vzorčno povprecje)**.** Naj bo X slučajna spremenljivka na populaciji, ki ima matematično upanje $E(X) = \mu$ in standardno deviacijo $\sigma(X) = \sigma$. Naj bo $(X_1 \cdots X_n)$ vzorec. Definirajmo vzorčno povprecje

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

ki je vzorčna statistika.

Je cenilka za $\overline{X}$, ki je nepristranska:

$$E(\overline{X}) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \frac{1}{n}n \cdot \mu = \mu$$

Po ŠZVŠ (izreku Čebiševa) je to dosledna cenilka za μ.
Ker je

$$D(\overline{X}) \overset{\text{neodv}}{=} \frac{1}{n^2}\sum_{i=1}^{n} D(X_i) = \frac{1}{n^2}n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

je standardna napaka

$$SE(Y) = \frac{\sigma}{\sqrt{n}}$$

- čim večji n, bolje oceni parameter μ
Po CLI je pri velikem n slučajna spremenljivka $Z_n := \frac{S-n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ porazdeljena približno $N(0,1)$ oz. $\overline{X}$ je porazdeljen približno $N(\mu, \frac{\sigma}{\sqrt{n}})$
Če je X normalno porazdeljena $N(\mu,\sigma)$, potem je $\overline{X}$ porazdeljen $N(\mu, \frac{\sigma}{\sqrt{n}})$ za vsak n

Trditev 2.6. Naj bo Y_n cenilka za ξ. Če je $E(Y_n) \overset{n\to\infty}{\to} \xi$ in $D(Y_n) \overset{n\to\infty}{\to} 0$, potem je $Y = Y_n$ dosledna cenilka za ξ

Dokaz. Fiksirajmo $\epsilon > 0$. Dokazati moramo $\lim_{n\to\infty} P(|Y_n - \xi| \geq \epsilon) = 0$
Ker je $E(Y_n) \overset{n\to\infty}{\xi}$, obstaja $n_0 \in \mathbb{N}$: $|E(Y_n) - \xi| < \frac{\epsilon}{2}$ zato je dogodek

$$(|Y_n - \xi| \geq \epsilon) \subseteq (|Y_n - E(Y_n)| + |E(Y_n) - \xi| \geq \epsilon) \text{ za } \forall n \subseteq$$
$$\overset{n\geq n_0}{\subseteq} (|Y_{n_0} - E(Y_{n_0})| + |E(Y_{n_0}) - \xi| \geq \epsilon)$$

Torej je za $n \geq n_0$

$$P(|Y_n - \xi| \geq \epsilon) \leq P(|Y_n - E(Y_n)| \geq \frac{\epsilon}{2}) \leq \frac{D(Y_n)}{\epsilon^2} \cdot 4 \overset{n\to\infty}{\to} 0 \text{ (doslednost)}$$

Neenakost Čebiševa: $P(|X - E(X)| \geq \epsilon) \leq \frac{D(X)}{\epsilon^2}$
Tako imamo doslednost cenilke: $P(|Y_n - \xi| \geq \epsilon) \overset{n\to\infty}{\to} 0$ ■

Primer. Porazdelitev χ^2, n število prostorskih stopenj

$$p(X) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} x > 0 \\ 0 \text{ sicer} \end{cases}$$

Modus $= n - 2$, $E(X) = n$, $D(X) = 2n$
Mediana $\approx n \cdot (1 - \frac{2}{9n})^3$

Definicija 2.7 (Vzorcna disperzija)**.** Naj bo X slučajna spremenljivka na populaciji. Vzorčna disperzija je definirana s

$$s_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

popravljena vzorčna disperzija pa je

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Kako sta porazdeljeni, če je $X \sim N(\mu, \sigma)$?
Raje vzemimo vzorčno statistiko: $\chi^2 := \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{\sigma^2} s_0^2 = \frac{n-1}{\sigma^2} s^2$
Ni lahko izračunati, da je $\chi^2 \sim \chi^2(n-1)$
Ideja izpeljave je $\chi^2 = Z_1^2 + \cdots + Z_{n-1}^2$ za $Z_i \sim N(0,1)$ in med seboj neodvisne. Potem uporabimo trditev iz verjetnosti: $Z_i^2 \sim \chi^2(1)$, torej $E(\chi^2) = n - 1$, $D(\chi^2) = 2(n-1)$. Od tod sledi

$$E(s_0^2) = E(\frac{\sigma^2}{n} \chi^2) = \frac{\sigma^2}{n} E(\chi^2) = \frac{n-1}{n} \sigma^2$$

torej s_0^2 ni nepristranska za σ^2, je pa asimptotično nepristranska, t.i. $E(s_0^2) \overset{n\to\infty}{\to} \sigma^2$
Podobno je $E(s^2) = \frac{\sigma^2}{n-1} E(\chi^2) = \sigma^2$, torej je s^2 nepristranska cenilka za σ^2
Ker je $D(s_0^2) = \frac{\sigma^4}{n^2} D(\chi^2) = \frac{\sigma^4 2(n-1)}{n^4} \overset{n\to\infty}{\to} 0$ in $D(s^2) = \frac{2\sigma^4}{(n-1)^2} \overset{n\to\infty}{\to} 0$, iz trditve sledi, da sta s_0^2 in s^2 dosledni cenilki za σ^2

Studentova t-porazdelitev

$$p(x) = \frac{1}{\sqrt{n}B(\frac{n}{2},\frac{1}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$$

kjer je $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ Beta funkcija.

$n=1: \quad \frac{1}{\pi}(1+x^2)^{-1} = \frac{1}{\pi(1+x^2)}$Cauchyjeva porazdelitev

ko gre $n \to \infty$, gre $\sqrt{n}B(\frac{n}{2},\frac{1}{n}) \to \sqrt{2\pi}$ in $(1+\frac{x^2}{n})^{-\frac{n-1}{2}} = ((1+\frac{x^2}{n})^n)^{-\frac{n+1}{2n}} \to e^{-\frac{x^2}{2}}$

torej je pri velikih n gostota približno $N(0,1)$

$n=2: \quad \frac{1}{\sqrt{2}B(1,\frac{1}{2})}(1+\frac{x^2}{2})^{-\frac{3}{2}}$

za $n \geq 2$ je $E(X) = 0$

$n=3: \quad c \cdot (1+\frac{x^2}{2})^{-2} \approx \frac{1}{x^4}$ za velike x

za $n \geq 3$ je $D(X) = \frac{n}{n-2} > 1$

Leta 1908 jo je odkril W.S. Gosset, statistik v pivovarni guiness v Dublinu. Student je njegov psevdonim.

Pri normalni porazdelitvi slučajne spremenljivke $X \sim N(\mu,\sigma)$ je vzorčno povprečje $\overline{X}$ porazdeljeno $N(\mu,\frac{\sigma}{\sqrt{n}}), \overline{X} = \frac{X_1+\cdots+X_n}{n}$, torej je $Z := \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}-\mu}{\sigma}\sqrt{n}$ porazdeljena N(0,1). Če poznamo σ, potem bomo znali povedati, kako dobra ocena za μ je $\overline{X}$ ($\to$ intervali zaupanja).

Kako ravnati, če σ ne poznamo?
Lahko jo ocenimo s $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$, tako da potem vzorčna statistika $T = \frac{\overline{X}-\mu}{s}\sqrt{n}$ ni več porazdeljena po $N(0,1)$, niti približno normalna, razen če je n velik in je s potem skoraj konstanta σ.

Kako je porazdeljena vzorčna statistika T?
Ker je $\chi^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{(n-1)S^2}{\sigma^2}$, je $\frac{Z}{T} = \frac{S}{\sigma} = \sqrt{\frac{\chi^2}{n-1}}$, torej je $T = \frac{Z}{\sqrt{\frac{\chi^2}{n-1}}}$

Izkaže se, da sta $Z \sim N(0,1)$ in $\chi^2 \sim \chi^2(n-1)$ neodvisni slučajni spremenljivki. Od tod lahko izračunamo, da ima T Studentovo porazdelitev z $n-1$ prostorskimi stopnjami:

$$p_T(t) = \frac{1}{(n-1)B(\frac{n-1}{2}, \frac{1}{2})} \cdot \frac{1}{(1 + \frac{x^2}{n-1})^{\frac{n}{2}}}$$

2.3 Metode za pridobivanje cenilk

2.3.1 Metoda momentov

Definicija 2.8 (Vzročni moment)**.** Naj bo $(X_1, X_2 \cdots X_n)$ vzorec velikosti n, torej $X_1 \cdots X_n$ neodvisne slučajne spremenljivke, porazdeljene kot slučajna spremenljivka X. Začetni moment reda k je $z_k = E(X^k)$. Definiramo k-ti vzročni moment $z_k := \frac{X_1^k + \cdots + X_n^k}{n}$. Le ta je nepristranska cenilka za z_k : $E(Z_k) = \frac{1}{n}(E(X_1^k) + \cdots + E(X_n^k)) = z_k$. Z_k je tudi dosledna cenilka za z_k.

Naj bo gostota slučajne spremenljivke X odvisna od parametrov $\xi_1 \cdots \xi_n$: $p(X; \xi_1 \cdots \xi_m)$. Naj odstajajo začetni momenti $z_k = E(X^k) = \int_{-\infty}^{\infty} p(x; \xi_1 \cdots \xi_n)dx, k = 1, 2 \cdots m$. Denimo, da iz teh m enačb lahko izrazimo parametre: $\xi_k = \phi_k(z_1, z_2 \cdots z_m), k = 1 \cdots m$ za neko funkcijo ϕ_k. Potem je $c_k := \phi_k(z_1 \cdots z_m)$ cenilka za parameter $\xi_k, k = 1 \cdots n$

Primer. Naj bo $X \sim N(\mu, \sigma)$, kjer sta μ in σ neznana parametra. Potem je $z_1 = E(X) = \mu, z_2 = E(X^2) = E(X^2) - (E(X))^2 + (E(X))^2 = D(X) + (E(X))^2 = \sigma^2 + \mu^2$ (m = 2)
Iz teh dveh enačb izrazimo parametra μ in σ: $\mu = z_1, \sigma^2 = z_2 - \mu^2 = z_2 - z_1^2$. Cenilka za μ je $Z_1 = \overline{X} = \frac{X_1 + \cdots + X_n}{n}$, cenilka za σ^2 je $Z_2 - Z_1^2 = \frac{X_1^2 + \cdots + X_n^2}{n} - \overline{X}^2$. To je enako

$$\begin{aligned} S_0^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \\ &= \frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\overline{X}\overline{X} + \overline{X}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{2}X_i^2 - \overline{X}^2 \end{aligned}$$

Torej bodimo že znani cenilki za parametra μ in σ^2

Primer. Naj bo X porazdeljena enakomerno na $[a, b]$, kjer sta a in b neznana parametra. Iščemo cenilki za a in b. Po metodi momentov moramo izračunati 2 začetna momenta

$$
\begin{aligned}
z_1 &= E(X) = \frac{a+b}{2} \\
z_2 &= E(X^2) = \int_{-\infty}^{\infty} x^2 p(x; a, b) dx = \frac{1}{b-a} \int_a^b x^2 dx = \\
&= \frac{1}{b-a} \frac{x^3}{3} |_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}
\end{aligned}
$$

Iz 1. enačbe dobimo $b = 2z_1 - a$, kar vstavimo v 2. enačbo

$$
\begin{aligned}
& 3z_2^2 = b^2 + ab + a^2 = 4z_1^2 - 4z_1 a + a^2 + 2az_1 - a^2 + a^2 \\
& \implies 3z_2 = 4z_1^2 - 2z_1 a + a^2 \\
& a^2 - 2az_1 + (4z_1^2 - 3z_2) = 0 \\
& \quad D = 4z_1^2 - 4(4z_1^2 - 3z_2) = 12(z_2 - z_1^2) \\
& a_{1,2} = \frac{1}{2}(2z_1 \pm \sqrt{D}) = z_1 \pm \frac{1}{2} 2\sqrt{3}\sqrt{z_2 - z_1^2} = z_1 \pm \sqrt{3}\sqrt{z_2 - z_1^2}
\end{aligned}
$$

Ker je $a < b$, je torej

$$
\begin{aligned}
a &= z_1 - \frac{1}{2} 2\sqrt{3}\sqrt{z_2 - z_1^2} \\
b &= z_1 + \frac{1}{2} 2\sqrt{3}\sqrt{z_2 - z_1^2}
\end{aligned}
$$

Cenilka za a je

$$A := Z_1 \pm \frac{1}{2} 2\sqrt{3}\sqrt{Z_2 - Z_1^2} \quad A := Z_1 \pm \frac{1}{2} 2\sqrt{3}\sqrt{Z_2 - Z_1^2} = Z_1 - S_0\sqrt{3} \text{ po prejšnjem primeru } = \overline{X}$$

Cenilka za b je $B = \overline{X} + S_0\sqrt{3}$ Denimo da imamo konkreten vzorec $-2, 0, 1, 2, 4 (n = 5)$
$\overline{X} = \frac{-2+0+1+2+4}{5} = 1$
$S_0^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{5}((-3)^2 + (-1)^2 + 0^2 + 1^2 + 3^2) = 4$ Vzorčna vrednost za A je $\overline{X} - S_0\sqrt{3} = 1 - 2\sqrt{3} =\doteq -2.46$, vzorčna vrednost z B je $\overline{X} + S_0\sqrt{3} = 1 + 2\sqrt{3} \doteq 4.46$

2.3.2 Metoda maksimalne zanesljivosti (oz. največjega verjetja)

Definicija 2.9 (Funkcija zanesljivosti)**.** Naj bo gostota slučajne spremenljivke X odvisna od parametra ξ, torej $p(x; \xi)$. Funkcija zanesljivosti (likelihood function) je

$$L(x_1 \cdots x_n; \xi) = p(x_1; \xi) \cdot \cdots \cdot p(x_n; \xi)$$

Pri danih $x_1 \cdots x_n$ izberimo tak ξ_{max}, da ima L tam maksimum. Ta vrednost parametra je odvisna od $x_1 \cdots x_n$, torej $\xi_{max} = \phi(x_1, x_2 \cdots x_n)$ za neko funkcijo ϕ. Tako dobimo cenilko $c := \phi(x_1 \cdots x_n)$ za parameter ξ

Primer.

$$p(x; \lambda) := \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x < 0 \end{cases}$$

λ je neznan parameter, ki ga ocenjujemo

$$L(x_1 \cdots x_n; \lambda) = \lambda e^{-\lambda x_1} \cdot \cdots \cdot \lambda e^{-\lambda x_n} = \lambda^n e^{-(x_1 + \cdots + x_n)}$$

Poiskati moramo λ_{max}, pri katerem je dosežen maksimum funkcije L (oz. maksimum funkcije $\ln(L)$)

$$\ln L(x_1 \cdots x_n; \lambda) = n \cdot \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \lambda}(\ln L(x_1 \cdots x_n; \lambda)) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

$$\implies \lambda_{max} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\overline{x}}$$

Ker je $\frac{\partial^2}{\partial \lambda^2} \ln L(x_1 \cdots x_n; \lambda) = -\frac{n}{\lambda^2} < 0$, je v λ_{max} maksimum.
Cenilka za λ je $c := \frac{1}{\overline{X}}$
Isto cenilko dobimo z metodo momentov:

$$z_1 = E(X) = \frac{0}{\infty} x \lambda e^{-\lambda x} dx \overset{\text{D.N.}}{= \cdots} \frac{1}{\lambda} \qquad \implies \lambda = \frac{1}{z_1} = \frac{1}{\overline{x}}$$

cenilka za λ je $c := \frac{1}{\overline{X}}$

Primer. $X \sim N(\mu, \sigma)$, μ, σ neznana parametra, ki ju ocenjujemo

$$L(x_1 \cdots x_n; \mu, \sigma) := \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_1-\mu}{\sigma})^2} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x_n-\mu}{\sigma})^2} =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \cdot \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2+\cdots+(x_n-\mu)^2}$$

$$\ln L = -\frac{n}{2}\ln 2\pi - n \cdot \ln \sigma - \frac{1}{2\sigma^2}((x_1-\mu)^2 + \cdots + (x_n-\mu)^2)$$

$$\frac{\partial}{\partial \mu} \ln L = -\frac{1}{2\sigma^2}(2(x_1-\mu)(-1) + \cdots + 2(x_1-\mu)(-1)) = \frac{1}{\sigma^2}(x_1 - \mu + \cdots + x_n - \mu) = 0$$

$$x_1 + \cdots + x_n - n\mu = 0 \implies \mu = \frac{x_1 + \cdots + x_n}{n} = \overline{x}$$

$$\frac{\partial}{\partial \sigma} \ln L = -\frac{n}{\sigma} + \frac{1}{\sigma^3}((x_1-\mu)^2 + \cdots + (x_n-\mu)^2) = 0$$

$$\implies \sigma^2 = \frac{1}{n}((x_1-\mu)^2 + \cdots + (x_n-\mu)^2) =$$

$$= \frac{1}{n}((x_1-\overline{x})^2 + \cdots + (x_n-\overline{x})^2) = s_0^2$$

Cenilka za μ je $\overline{X}$, cenilka za σ^2 je $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

Primer. $Bin(1,p) = Ber(p), X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$ $q = 1-p, p$ neznan parameter

$$P(X = x) = p^x(1-p)^{1-x} x \in \{0, 1\}$$
$$L(x_1 \cdots x_n; p) = p^{x_1}(1-p)^{1-x_1} \cdot \ldots \cdot p^{x_n}(1-p)^{1-x_n} =$$
$$= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)}$$
$$x := x_1 + \cdots + x_n \implies L(x_1 \cdots x_n; p) = p^x(1-p)^{1-x} x \in \{0, 1 \cdots n\}$$
$$\ln L = x \ln p + (n-x)\ln(1-p)$$
$$\frac{\partial}{\partial p} \ln L = \frac{x}{p} - \frac{n-x}{1-p} = 0$$
$$\implies x(1-p) = (n-x)p \implies x - xp = np - xp \implies p = \frac{x}{n} = \overline{x}$$

Cenilka za p je $P := \overline{X} = \frac{X_1+\cdots+X_n}{n}$
Ker je

$$E(P) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = p$$

je P nepristranska cenilka
Ker je

$$D(P) = \frac{1}{n^2}(D(X_1 + \cdots + D(X_n))) = \frac{1}{n^2} n D(X_1) = \frac{1}{n} D(X_1) \overset{n\to\infty}{\to} 0$$

po trditvi sledi, da je $\overline{X}$ dosledna cenilka za P

2.4 Intervalsko ocenjevanje parametrov

Definicija 2.10 (Interval zaupanja)**.** Naj bo gostota slučajne spremenljivke X odvisna od parametra ξ. Interval $[A, B]$ (odvisen le od $(x_1 \cdots x_n)$ in ne do ξ) je interval zaupanja za parameter ξ, pri stopnji tveganja $\alpha \in (0, 1)$, če je

$$P(\xi \in [A, B]) = 1 - \alpha \text{ oz. } P(\xi \notin [A, B]) = \alpha$$

Za α običajno vzamemo vrednost 0.05 (ali 0.01)
A in B sta vzorčni statistiki, $1 - \alpha$ je stopnja zaupanja

Primer. $X \sim N(\mu, \sigma), \sigma$ poznamo, μ pa je neznan parameter.
Slučajna spremenljivka $Z := \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim N(0, 1)$
Pri dani stopnji tveganja α najdemo $z_{\frac{\alpha}{2}} > 0$, da je $P(-z_{\frac{\alpha}{2}} < Z < z_{\frac{\alpha}{2}}) = 1-\alpha$ oz. $P(|Z| > z_{\frac{\alpha}{2}}) = \alpha$ oz. $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$
Pogoj $|Z| < z_{\frac{\alpha}{2}}$ pomeni: $|\overline{X} - \mu| < z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

$$\begin{aligned} A := \overline{X} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \\ < \overline{X} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} =: B \end{aligned}$$

$[A, B]$ je interval zaupranja za μ pri stopnji tveganja α
Konkreten zgled: imejmo vzorec velikosti $n = 36$, za katerega izračunamo $\overline{x} = 2.6$ in $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.3$. Predpostavimo, da imamo $X \sim N(\mu, \sigma)$ in predpostavimo, da je $\sigma := s = 0.3$ (kar pogosto naredimo, če je n razmeroma velik). Vzemimo $\alpha = 0.05$. Iz tabele razberemo $z_{\frac{\alpha}{2}} = 1.96$, torej $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. Tedaj je vzorčna vrednost za A enaka

$$\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2.6 - 1.96 \frac{0.3}{\sqrt{36}} = 2.5$$

vzorčna vrednost za B je $\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2.7$
Interval zaupanja za μ je $[2.5, 2.7]$, t.j.

$$P(\mu \in [2.5, 2.7]) = 1 - \alpha = 0.95$$

Primer. $X \sim N(\mu, \sigma)$, μ in σ sta neznana.
Iščemo interval zaupanja za μ.

Slučajna spremenljivka $T := \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim Student(n-1)$
Pri danem tveganju α izberemo $t_{\frac{\alpha}{2}} > 0$, da je $P(|T| < t_{\frac{\alpha}{2}}) = 1-\alpha$ oz. $P(T > t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$
Sedaj imamo podobno situacijo kot v primeru 1.
Pogoj $|T| < t_{\frac{\alpha}{2}}$ pomeni

$$A := \overline{X} - t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} =: B$$

Konkreten zgled: življenska doba žarnic v vzorcu je $9.8, 10.2, 10.4, 9.8, 10.0, 10.2, 9.6$ (v dneh), $n = 7$. Predpostavimo normalni model $N(\mu,\sigma)$ z neznanima parametroma μ in σ

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 10.0$$
$$s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2} = 0.283$$

Vzemimo $\alpha = 0.05$, iz tabele za $Student(5)$ razberemo $t_{\frac{\alpha}{2}} = 2.45$
Vzorčna vrednost za A je $a = \overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 9.74$
Vzorčna vrednost za B je $b = \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} = 10.26$
Interval zaupanja za μ je $[9.74, 10.26]$, kar zapišemo kot $\mu = 10.0 \pm 0.26$,
Verjetnost, da je $\mu \in [9.74, 10.26]$ je 0.95

Primer. Pri normalni porazdelitvi $N(\mu,\sigma)$ ocenjujemo parameter σ. Vzorčna statistika

$$\chi^2 := \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{n-1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 S$$

je porazdeljena po $\chi^2(n-1)$
Izberimo c_1 in c_2 da je

$$P(\chi^2 < c_1) = \frac{\alpha}{2} = P(\chi^2 > c_2)$$

oz.

$$P(c_1 < \chi^2 < c_2) = 1-\alpha$$

Pogoj $c_1 < \chi^2 < c_2$ pomeni

$$
\begin{aligned}
&c_1 < \frac{n-1}{\sigma^2} s^2 < c_2 \iff \\
&\iff \frac{1}{c_1} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{c_2} \iff \\
&\iff B := \frac{(n-1)s^2}{c_1} > \sigma^2 > \frac{(n-1)s^2}{c_1} =: A
\end{aligned}
$$

$[A, B]$ je interval zaupanja za σ^2 pri stopnji tveganja α

$$
\begin{aligned}
A &= \frac{1}{c_2} \sum_{i=1}^{n} (x_i - \overline{x})^2, \\
B &= \frac{1}{c_1} \sum_{i=1}^{n} (x_i - \overline{x})^2
\end{aligned}
$$

Primer. Žarnice iz prejšnjega primera:
$n = 7, (n-1)s^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 = 0.481, \alpha = 0.059$

$$
\begin{aligned}
&\chi^2(6) : c_1 = 1.24, c_2 = 14.45 \\
&a = \frac{1}{14.45} 0.481 = 0.033, b = \frac{1}{1.24} 0.481 = 0.388 \\
&\implies P(0.033 < \sigma^2 < 0.388) = 0.95 \\
&P(0.182 < \sigma < 0.623) = 0.95
\end{aligned}
$$

$[0.182, 0.623]$ je interval zaupanja za σ pri stopnji tveganja 0.05

Primer. $X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, q = 1 - p, p$ neznan parameter, ki ga ocenjujemo
$(x_1 \cdots x_n)$ vzorec. Potem je $S_n = X_1 + \cdots + X_n \sim Bin(n,p)$ in $\overline{X} = \frac{S_n}{n}$ je nepristranska in dosledna cenilka za p. Po CLI (Laplaceovi formuli) je pri velikih n $Z := \frac{S_n - np}{\sqrt{npq}} \sim N(0,1)$ oz. $Z = \frac{\overline{X}-p}{\sqrt{pq}} \sqrt{n} \sim N(0,1)$ oz. $\overline{X} \sim N(p, \sqrt{\frac{pq}{n}})$
Pri danem $\alpha > 0$ izberimo $z_{\frac{\alpha}{2}} > 0$, da je $P(|Z| < z_{\frac{\alpha}{2}}) = 1 - \alpha$
P Pogoj $|Z| < z_{\frac{\alpha}{2}}$ pomeni $|S - np| < z_{\frac{\alpha}{2}} \sqrt{npq}$ oz. $|\overline{X} - p| < z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{pq}{n}}$
Če na desni strani naredimo aproksimacijo $\overline{X} \approx p$, dobimo pogoj

$$|\overline{X}-p| < z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$$

od koder dobimo interval zaupanja za p:

$$\begin{aligned}
&A := \overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\\
&B := \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\\
&A < p < B
\end{aligned}$$

Primer. Predsedniške volitve v ZDA leta 2000:
Anketa na 2207 volivcev: $n = 2207, \alpha = 0.05 \implies z_{\frac{\alpha}{2}} = 1.96$
George Bush: 47%, Algore: 44%, Ralph Nader: 2%
Določimo intervale zaupanja

$$p_{Bush} = 0.47 \pm 1.96\sqrt{\frac{0.47(1-0.47)}{2207}} \doteq 0.47 \pm 0.02$$

Interval zaupanja za p_{Bush} je $[0.45, 0.49]$

$$p_{Algore} = 0.44 \pm 1.96\sqrt{\frac{0.44 \cdot 0.56}{2207}} \doteq 0.44 \pm 0.02$$

Interval zaupanja za p_{Algore} je $[0.42, 0.46]$

$$p_{Nader} = 0.02 \pm 1.96\sqrt{\frac{0.02 \cdot 0.98}{2207}} \doteq 0.02 \pm 0.006$$

Odstopanje:

$$\begin{gathered}
z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} < 2\sqrt{\frac{\frac{1}{4}}{n}} = \frac{1}{\sqrt{n}}\\
x(1-x) \le \frac{1}{4} \iff x - x^2 \le \frac{1}{4} \iff\\
\iff 0 \le x^2 - x + \frac{1}{4} = (x - \frac{1}{2})^2
\end{gathered}$$

2.5 Preizkušanje statističnih hipotez

Definicija 2.11 (Statistična hipoteza)**.** Statistična hipoteza je vsaka domneva o porazdelitvi slučajne spremenljivke X na populaciji

Definicija 2.12 (Enostavnost hipoteze)**.** Hipoteza je enostavna, če natanko določa porazdelitev, sicer je sestavljena

Primer. $X \sim N(\mu, \sigma), \sigma$ poznamo, μ je neznan parameter
$H(\mu = 0)$ je primer enostavne hipoteze. Če σ ne poznamo, je to sestavljena hipoteza

Vedno preizkušamo eno ničelno hipotezo H_0 nasproti alternativni hipotezi H_1

Primer. $X \sim N(\mu, \sigma), \sigma$ poznamo
$H_0(\mu = 0) : H_1(\mu \neq 0)$

Za H_0 običajno vzamemo enostavno hipotezo, za katero upamo, da jo bomo zavrnili
Hipoteza je lahko pravilna ali nepravilna. Ideal je sprejeti pravilno in zavrniti nepravilno. Odločiti se moramo na osnovi vzorca. Če vzorčni podatki preveč odstopajo od hipoteze, potem niso konsistentni z njo oz. so razlike značilne (signifikantne); tedaj hipotezo zavrnemo
Vnaprej določimo stopnjo značilnosti $\alpha \in [0, 1]$, to je verjetnost, da zavrnemo pravilo hipotezo. Običajno je $\alpha = 0.05$ ali $\alpha = 0.01$. Take teste imenujemo testi značilnosti

Primeri testov znacilnosti

2.5.1 test Z

$X \sim N(\mu, \sigma), \sigma$ znan parameter
Ničelna domneva je $H = (\mu = \mu_0)$, kjer je μ_0 damo realno število.
Pri predpostavki $H_0(\mu = \mu_0)$ je $Z := \frac{\overline{X} - \mu}{\sigma}\sqrt{n}$ porazdeljena $N(0, 1)$, saj je $\overline{X} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$
Vzemimo $H_1(\mu \neq \mu_0)$. Tedaj H_0 zavrnemo, če vzorčna vrednost za Z leži na kritičnem območju

$$K_\alpha = (-\infty, -z_{\frac{\alpha}{2}}] \cup [z_{\frac{\alpha}{2}}, \infty)$$

kjer je α stopnja značilnosti in $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

$Z \ldots$ testna statistika
Pri stopnji značilnosti α določimo $z_{\frac{\alpha}{2}} > 0$, da je

$$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

K_α kritično območje
Če je vzorčna vrednost za Z na K_α, hipotezo H_0 zavrnemo

Primer. Izdelovalec vrvic trdi, da je povprečna sila, pri kateri se vrvica strga 150N s standardno deviacijo 5N. Na vzorcu 50 vrvic ($n = 50$) je bila povprečna sila 148N. Privzamemo normalno porazdelitev $N(\mu, 5)$. Pri stopnji značilnosti $\alpha = 0.01$ testiramo hipotezo

$$H_0(\mu = 150) : H_1(\mu \neq 150)$$

Iz tabel razberemo $z_{\frac{\alpha}{2}} = 2.58$
Kritično območje $K_\alpha = (-\infty, -2.58] \cup [2.58, \infty)$
Testna statistika $Z = \frac{\overline{X}-150}{5}\sqrt{50} = (\overline{X} - 150)\sqrt{2}$, njena vzorčna vrednost je

$$z = (148 - 150)\sqrt{2} = -2\sqrt{2} = -2.82$$

Ker je $z = -2.82 \in K_\alpha, H_0$ zavrnemo; razlike so značilne (signifikantne)
Ker gledamo odstopanja v obe smeri, je to dvostranski test Z.
Smiselen bi bil enostranski test Z: $H_0(\mu = 150) : H_1(\mu < 150)$
$K_\alpha = (-\infty, -z_\alpha)$, pri $\alpha = 0.01$ je $z_\alpha = 2.33$
Ker $z = -2.82 \in K_\alpha, H_0$ zavrnemo tudi sedaj

2.5.2 test T: $X \sim N(\mu, \sigma)$, μ neznan

$H_0(\mu = \mu_0) : H_1(\mu \neq \mu_0), \mu_0$ je dano število
Testna statistika

$$T = \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

S je vzorčna deviacija
Pri predpostavki H_0 je porazdeljena po Student(n-1)
$K_\alpha = (-\infty, -t_{\frac{\alpha}{2}}] \cup [t_{\frac{\alpha}{2}}, \infty)$
Če vzorčna vrednost za T leži na K_α, hipotezo zavrnemo

Primer. Nadaljevanje prejšnjega primera
Privzamemo deviacijo 5N izračunano iz vzorca, torej

$$S = 5N, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Kot prej je $\overline{x} = 148N$
Iz tabel razberemo, da je $t_{\frac{\alpha}{2}} = 2.68$ (pri 49 prostorskih stopnjah in $\alpha = 0.01$)
$K_\alpha = (-\infty, -2.68] \cup [2.68, \infty)$, vzorčna vrednost za T je (tako kot prej) $t = -2.82$
Ker $t \in K_\alpha, H_0$ zavrnemo

Primer. V kliničnem poskusu testirajo zdravila za zniževanje krvnega tlaka so 10 bolnikom izmerili sistolično krvni tlak pred in po zdravljenju. Razlike ̈pred-poš̈o $-8, 0, 2, 4, 9, 14, 19, 22, 32, 35 mmHg$. Predpostavimo normalni model $N(\mu, \sigma)$ z neznanima μ in σ. Testiramo hipotezo $H_0(\mu = 0) : H_1(\mu > 0)$. Naredimo test značilnosti (parametrični test značilnosti)

1. enostranski test T $H_0(\mu = 0) : H_1(\mu > 0), H_0 :=$ zdravilo ne učinkuje
2. stopnja značilnosti $\alpha = 0.05$
3. testna statistika $T = \frac{\overline{X}}{S}\sqrt{10}$, $n = 10$
4. kritično območje $K_\alpha = [t_\alpha, \infty)$
   $Student(9) \implies t_\alpha = 1.83$
5. vzorčna vrednost za T je

$$t = \frac{12.9}{14.1}\sqrt{10} = 2.89$$

Sklep: ker $t = 2.89 \in K_\alpha$, hipotezo H_0 zavrnemo

Definicija 2.13 (P-vrednost)**.** P-vrednost je najmanjša stopnja značilnosti, pri kateri še lahko zavrnemo hipotezo (pro danih vzorčnih podatkih)

V našem primeru je $P = 0.89\% = 0.0089$

2.5.3 Studentov primerjalni test

Imejmo 2 neodvisna vzorca velikosti m in n. Prvi je vzet iz populacije, na kateri ime slučajna spremenljivka $X \sim N(\mu_x, \sigma)$, druga pa iz populacije, na kateri imamo $Y \sim N(\mu_x, \sigma)$. Predpostavljamo torej enakost disperzij. Če sta s_x^2 in s_y^2 povprečni vzorčni disperziji

$$S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$
$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

potem definiramo skupno vzorčno varianco

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} = \frac{1}{m+n-1}(\sum_{i=1}^{m}(x_i - \overline{x})^2 + \sum_{i=1}^{n}(y_i - \overline{y})^2)$$

Testiramo hipotezo $H_0(\mu_x = \mu_y) : H_1(\mu_x \neq \mu_y)$. Testna statistika

$$T = \frac{\overline{X} - \overline{Y}}{S}\sqrt{\frac{mn}{m+n}}$$

Potem je

$$\overline{X} - \overline{Y} \sim N(0, \sqrt{(\frac{\sigma}{\sqrt{m}})^2 + (\frac{\sigma}{\sqrt{n}})^2}) =$$
$$= N(0, \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}) = N(0, \sigma\sqrt{\frac{m+n}{mn}})$$

Zato ima spremenljivka

$$Z = \frac{\overline{X} - \overline{Y}}{S}\sqrt{\frac{mn}{m+n}} \sim N(0,1)$$

Ker je spremenljivka

$$U = \frac{(m+n-2)S^2}{\sigma^2} \sim \chi^2(m+n-2)$$

je

$$T = \frac{Z}{\sqrt{\frac{U}{m+n-2}}} \sim Student(m+n-2)$$

Primer. 2 zdravili proti nespečnosti preizkusijo na vzorcih velikosti $m = n = 10$.
Dodatno število ur pri prvem zdravilu: $1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4$
Dodatno število ur pri drugem zdravilu pa $0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0.0, 2.0$

1. dvostranski standardni primerjalni test $H_0(\mu_x = \mu_y) : H_1(\mu_x \neq \mu_y)$
2. stopnja značilnosti $\alpha = 0.05$
3. testna statistika $T = \frac{\overline{X}-\overline{Y}}{S}\sqrt{\frac{mn}{m+n}} \sim Student(m+n-2) = Student(18)$, če velja hipoteza H_0
4. kritično območje $K_\alpha = (-\infty, -t_{\frac{\alpha}{2}}] \cup [t_{\frac{\alpha}{2}}, \infty), t_{\frac{\alpha}{2}} = 2.10$ (iz tabele, 18 prostorskih stopenj)
5. vzorčna vrednost za T je

$$t = \frac{\overline{x} - \overline{y}}{s}\sqrt{5} = \frac{2.33 - 0.75}{\sqrt{3.70}}\sqrt{5} = 1.84$$

Sklep: ker $1.84 \notin K_\alpha$, hipoteze H_0 ne moremo zavrniti pri stopnji značilnosti 5%
P-vrednost pride 0.079

Poleg populacijskega povprečja μ lahko testiramo tudi druge količine:

- standardna deviacija $\sigma : H_0(\sigma = \sigma_0), \sigma_0$ je dano število
- tip porazdelitvenega zakona: $H_0(F = F_0)$
- neodvisnost dveh spremenljivk
- korelacijski koeficient

2.5.4 Test hi-kvadrat

(Pearson)
Preizkus domneve o tipu porazdelitvenega zakona, torej $H_0(F = F_0) : H_1(F \neq F_0)$, kjer je F_0 dana porazdelitvena funkcija. Zalogo vrednosti slučajne spremenljivke X razdelimo na r razredov (disjunktno) $S_1, S_2 \cdots S_r$, da je

$$p_k = P(X \in S_k \mid H_k) > 0 \; \forall k = 1, 2 \cdots r$$

Potem je $\sum_{k=1}^r p_k = 1, \sum_{k=1}^r N_k = n$ ter $N_k \sim Bin(n, p_k)$ in $E(N_k) = p \cdot n_k$ kar je pričakovana vrednost za k-ti razred.

Pri velikem n ima testna statistika $\chi^2 = \sum_{k=1}^{r} \frac{(N_k - np_k)^2}{np_k}$ približno porazdelitev $\chi^2(r-1)$
Če χ^2 zavzame preveliko vrednost, hipotezo H_0 zavrnemo
$K_\alpha = [c_\alpha, \infty), P(\chi^2 > c_\alpha) = \alpha$

Primer. Preizkušamo "poštenost" kocke
v $n = 120$ dobimo za število pik

število pik	1	2	3	4	5	6	
opazovane frekvence	20	22	17	18	19	24	N_k
pričakovane frekvence	20	20	20	20	20	20	np_k

$r = 6, p_k = \frac{1}{6}$

$$\chi^2 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \cdots + \frac{(24-20)^2}{20} = \frac{34}{20} = 1.7$$

Pri $\alpha = 0.05$ in 5 prostorskih stopnjah je $c_\alpha = 11.1$
Ker $1.7 \notin K_\alpha$, hipoetzo H_0, da gre za enakomerno porazdelitev na 6 točkah ne moremo zavrniti

Primer. Podatki o število močnih potresov (vsaj 8 stopnje po Richterjevi lestvici) v obdobju 1969-2001, $n = 33$

število potresov	0	1	2	3	4	5	$\cdots$
število let s toliko potresiv	15	13	4	1	0	0	$\cdots$

Hipoteza H_0 : podatki so porazdeljeni po $Poisson(1.5)$: $p_k = e^{-1.5}\frac{(1.5)^k}{k!} k = 0, 1, 2 \cdots$
Pričakovane frekvence so

število potresov	0	1	2	3	4	$\cdots$
število let s toliko potresiv	7.4	11.0	8.3	4.1	1.6	$\cdots$

Vsota za od (vključno z) 3 naprej je 6.3
Teorija priporoča, da so pričakovane frekvence vsaj 5, zato vpeljemo razred "≥ 3"

razred	0	1	2	3
opazovane frekvence	15	13	4	1
pričakovane frekvence	7.4	11.0	8.3	6.3

$r = 4$

$$\chi^2 = \frac{(15-7.4)^2}{7.4} + \frac{(13-11)^2}{11} + \frac{(4-8.3)^2}{8.3} + \frac{(1-6.3)^2}{6.3} = 14.9$$

Pri $\chi^2(3)$ je $c_\alpha = 7.82$
Ker $14.9 \in K_\alpha = [7.82, \infty)$, hipotezo H_0 zavrnemo
P-vrednost je 0.002

Opomba. Če so v testu χ^2 frekvence p_k odvedljivo odvisne od parametra θ, torej $p_k(\theta)$, potem ima statistika $\chi^2 = \sum_{k=1}^r \frac{(N_k - np_k(\hat{\theta}))^2}{np_k(\hat{\theta})}$ približno porazdelitev $\chi^2(r-2)$, kjer je $\hat{\theta}$ cenilka za parameter θ po metodi maksimalne zanesljivosti

Primer. Potresi (od prej), H_0: podatki imajo Poissonovo porazdelitev
Cenilka za λ je

$$\hat{\lambda} = \overline{X} = \frac{0 \cdot 15 + 1 \cdot 13 + 2 \cdot 4 + 3 \cdot 1}{33} = \frac{8}{11} = 0.73$$

(ista cenilka tako po metodi momentov kot po moetodi največjega verjetja)
Torej

$$p_k(\hat{\lambda}) = e^{-0.73} \frac{(0.73)^k}{k!} \quad k = 0, 1 \cdots$$

razred	0	1	≥ 2	k
opažene frekvence	15	13	5	N_k
pričakovane frekvence	15.9	11.6	5.5	$n \cdot p_k(\hat{\lambda})$

$r = 3$

$$\chi^2 = \frac{(15 - 15.9)^2}{15.9} + \frac{(13 - 11.6)^2}{11.6} + \frac{(5 - 5.5)^2}{5.5} = 0.27$$

Za $\chi^2(r-2) = \chi^2(1)$ in $\alpha = 0.05$ je $c_\alpha = 3.84, K_\alpha = [3.84, \infty)$
Ker $0.27 \notin K_\alpha$, hipoteze H_0 ne moremo zavrniti

Opomba. Računi bi bili drugačni če bi imeli $\lambda = 0.73$ podan na začetku
($\chi^2(r-1)$ vs. $\chi^2(r-2)$)

2.6 Linearna regresija

Definicija 2.14 (Linearni regresijski model)**.** Linearni regresijski model: $Y = a + bx + U$
Pri fiksnem $x \in \mathbb{R}$ predpostavljamo, da je $y = a + bx + U$, kjer sta a in b konstanti ter $U \sim N(0, \sigma)$ za nek pozitiven σ oz. $Y \sim N(a + bx, \sigma)$

Za različne vrednosti $x_1, x_2 \cdots x_n$ dobimo slučajni vektor $(y_1, y_2 \cdots y_n)$, kjer je $y_k \sim N(a + bx_k, \sigma)$
$y = a + bx$ je regresijska premica
y_k je vrednost za $Y_k, k = 1, 2 \cdots n$
Radi bi ocenili a in b
Z metodo maksimalne zanesljivosti se dobi cenilki

$$\hat{b} = \frac{\sum_{k=1}^{n}(x_k - \overline{X})(x_k - \overline{Y})}{\sum_{k=0}^{n}(x_k - \overline{X})^2}$$

in

$$\hat{a} = \overline{Y} - \hat{b} \cdot \overline{X}$$

Če vpeljemo naslednje vsote

$$S_x := \sum_{k=1}^{n} x_k$$
$$S_Y := \sum_{k=1}^{n} y_k$$
$$S_{xx} := \sum_{k=1}^{n} x_k^2$$
$$S_{xY} := \sum_{k=1}^{n} x_k y_k$$

je števec

$$\begin{aligned}
&\sum_{k=1}^{n}(x_k - \overline{X})(y_k - \overline{Y}) = \sum_{k=1}^{n}(x_k y_k - \overline{X} y_k - x_k \overline{Y} + \overline{XY}) = \\
&= S_{xx} - \frac{1}{n}\overline{X} S_y - \overline{Y} S_x + n\overline{XY} = \\
&= S_{xY} - \frac{1}{n} S_x S_Y - \frac{1}{n} S_Y S_x + \frac{1}{n} S_y = \\
&= S_{xY} - \frac{1}{n} S_x S_Y
\end{aligned}$$

in

$$\begin{aligned}
&\sum_{k=1}^{n}(x_k - \overline{X})^2 = \sum_{k=1}^{n}(x_k^2 2x_n\overline{X})^2 = \\
&= S_{xx} - 2\overline{X}S_x + n\overline{X}^2 = \\
&= S_{xx} - 2\frac{1}{n}S_x^2 + \frac{1}{n}S_x^2 = \\
&= S_{xx} - \frac{1}{n}S_x^2
\end{aligned}$$

Torej je

$$\hat{b} = \frac{nS_{xY} - S_x S_y}{nS_{xx} - S_x^2}$$

in

$$\hat{a} = \frac{1}{n}S_Y + \hat{b}\frac{1}{n}S_x$$

Primer. Astronom Hubble je leta 1929 ugotavljal, da je hitrost oddaljevanja galaksije od zemlje linearno odvisna od oddaljenosti galaksije
$n = 24$

oddaljenost [mega Parsek]	0.032	0.034	0.214	$\cdots$	2.0	2.0
hitrost oddaljevanja [km/s]	170	290	-130	$\cdots$	800	1090

$$S_x = 21.9, S_Y = 8065, S_{xx} = 29.5, S_{xY} = 12.519$$

$$\hat{b} \to 456,\ \hat{a} \to -34$$

Regresijska permica ja $y = -34 + 456x$

Regresijska premica se zelo približa izhodišču, kar se ujema s teorijo o velikem poku.

Do cenilk $\hat{a}$ in $\hat{b}$ lahko pridemo po metodi najmanjših kvadratov: minimiziramo funkcijo

$$f(a,b) := \sum_{k=1}^{n}(y_k - (a - bx_k))^2$$

Porebna pogoja za minimum sta:

$$
\begin{aligned}
0 = \frac{\partial f}{\partial a} &= -2\sum_{k=1}^{n}(y_k - a - bx_k) = (-2)(S_Y - na - bS_x) \\
0 = \frac{\partial f}{\partial b} &= -2\sum_{k=1}^{n}x_k(y_k - a - bx_k) = (-2)(S_{xY} - aS_x - bS_{xx})
\end{aligned}
$$

Torej

$$
\begin{aligned}
S_Y &= na + bS_x \quad / \cdot S_x \\
S_{xY} &= aS_x + bS_{xx} \quad \cdot n
\end{aligned}
$$

Enačbi odštejemo

$$
\begin{aligned}
S_xS_Y - nS_{xY} &= b(S_x^2 - nS_{xx}) \\
\implies b &= \frac{nS_{xx} - S_xS_y}{nS_{xx} - S_x^2} \\
a &= \frac{1}{n}(S_Y - bS_x) = \frac{1}{n}S_Y - b\frac{1}{n}S_x
\end{aligned}
$$

$$(X, Y) \sim N(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$$

$$E(Y \mid X = x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x) = \alpha + \beta x$$

kjer je $\beta = \rho\frac{\sigma_y}{\sigma_x}, \alpha = \mu_y - \beta\mu_x$

$$\beta = \rho\frac{\sigma_y}{\sigma_x} = \frac{K}{\sigma_x\sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \frac{K}{\sigma_x^2}$$

$$\hat{b} = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{X})(y_k - \overline{Y})}{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{X})^2}$$

števec: vzorčna kovarianca
imenovalec: vzorčna disperzija za x

$$\hat{a} = \overline{Y} - \beta\overline{X}$$

2.7 Testiranje zanesljivosti

Poseben primer testa χ^2, imenuje se prilagoditveni test (Goodness-of-Fit Test).
Ničelna hipoteza H_0: dogodka A in B sta neodvisna
Če je $p = P(A)$ in $q = P(B)$, imamo 4 razrede (kategorije):

kategorija	$A \cap B$	$A \cap B^C$	$A^C \cap B$	$A^C \cap B^C$
verjetnost	pq	$p(1-q)$	$(1-p)q$	$(1-p)(1-q)$

Če sta p in q znana parametra (običajno nista), uporabimo test χ^2 za $r = 4$:

$$\chi^2 = \frac{(N_{A\cap B} - npq)^2}{npq} + \cdots + \frac{(N_{A^C \cap B^C} - n(1-p)(1-q))^2}{n(1-p)(1-q)}$$

Kjer je $N_{A\cap B}$ opažena frekvenca dogodka $A \cap B$ in n velikost vzorca
Če H_0 velja, ima $\chi^2 \sim \chi^2(3)$ pri velikem n

$$K_\alpha = [c_\alpha, \infty)$$

"

Če je vzorčna vrednost za χ^2 na K_α, hipotezo H_0 zavrnemo
Običajno p in q nista znana parametra, zato ju ocenimo iz podatkov

Kontngenčna matrika:

	B	B^C
A	X_{11}	X_{12}
A^C	X_{21}	X_{22}

$$X_{11} + X_{12} + X_{21} + X_{22} = n$$

kjer je n velikost vzorca
Cenilki za p in q sta $\hat{p} := \frac{X_{11}+X_{12}}{n}$, $\hat{q} := \frac{X_{11}+X_{21}}{n}$
Statistika χ^2 je potem

$$\chi^2 = \frac{(X_{11} - n\hat{p}\hat{q})^2}{n\hat{p}\hat{q}} + \cdots + \frac{(X_{22} - n(1-\hat{p})(1-\hat{q}))^2}{n(1-\hat{p})(1-\hat{q})}$$

Izkaže se, da je $\chi^2 \sim \chi^2(1)$ za velik n

Primer. Na univerzo Barkley prijavljene študente razvrstimo glede na izbiro področja v skupini "lažje"in "težje"(glede na to ali se je lahko ali težko vpisati na področje)

lažji	1385	133	1518
težji	1306	1702	3008
	2691	1835	$4526 = n$

Ali je izbira področja neodvisna os spola?
Testirajno pri stopnji značilnosti $\alpha = 0.05$

$$n = 4526, X_{11} = 1385 \cdots$$

$$\hat{p} \to \frac{1516}{4526} = 0.34, \ \hat{q} \to \frac{2691}{4526} = 0.59$$

Pričakovane frekvence:

lažji	903 ($n\hat{p}\hat{q}$)	615	1518
težji	1788	1220	3008
	2691	1835	4526

$$n\hat{p}\hat{q} = (X_{11} + X_{12})(X_{11} + X_{21}) \cdot \frac{1}{n}$$

$$\chi^2 = \frac{(1385 - 903)^2}{903} + \cdots + \frac{(1702 - 1220)^2}{1220} = 957.1$$

Pri $\alpha = 0.05$ je $c_\alpha = 3.84$

Hipotezo H_0 zavrnemo
P vrednost $4 \cdot 10^{-210}$

Opisani test lahko posplošimo na večje kontingenčne tabele:
Denimo, da 1. karakteristika določa r kategorij $A_1, A_2 \cdots A_r$, 2. pa s kategorij $B_1, B_2 \cdots B_s$.
Naj bo $p_i = P(A_i) i = 1, 2 \cdots r$ in $q_j = P(B_j) j = 1, 2 \cdots s$ $\quad (\sum_{i=1}^{r} p_i = 1, \sum_{j=1}^{s} q_j = 1)$
Ničelna hipoteza H_0: A_i in B_j sta neodvisna za vsak i in vsak j
Vzorec velikosti n
Opažene frekvence:

	B_1	B_2	$\cdots$	B_r	
A_1	X_{11}	X_{12}	$\cdots$	X_{1s}	$n \cdot \hat{p}_1$
A_2	X_{21}	X_{22}	$\cdots$	X_{2s}	$n \cdot \hat{p}_2$
$\vdots$	$\vdots$				$\vdots$
A_r	X_{r1}	X_{r2}	$\cdots$	X_{rs}	$n \cdot \hat{p}_r$
	$n \cdot \hat{q}_1$	$n \cdot \hat{q}_2$	$\cdots$	$n \cdot \hat{q}_s$	n

$$\hat{p}_i := \frac{1}{n} \sum_{j=1}^{s} x_{ij}$$

$$\hat{q}_j := \frac{1}{n} \sum_{i=1}^{r} x_{ij}$$

Cenilke za p_i in q_j:
Definiramo

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(X_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}$$

Izkaže se, da je pri velikih n χ^2 pribli"no porazdeljena $\chi^2((r-1)(s-1))$

Če je $n\hat{p}_i\hat{q}_j < 5$ za kakšna i in j, potem se priporoča, da se združi nekatere razrede

Primer. $n = 1031$ darovalcev krvi glede na krvno skupino (A, B, AB, 0) in RH faktor (+, -)
Opažene frekvence:

	A	B	AB	0	
RH+	320	96	40	412	868
RH-	66	23	9	65	163
	386	119	49	477	1031

Pričakovane frekvence:

	A	B	AB	0	
RH+	325.0	100.2	41.3	401.5	868
RH-	61.0	18.8	7.7	75.5	163
	386	119	49	477	1031

$$\chi^2 = \frac{(320-325)^2}{325} + \cdots + \frac{(65-75.5)^2}{75.5} = 3.54$$

$\chi^2(3), \alpha = 0.05$
$c_\alpha = 7.8$

Ker $3.54 \notin K_\alpha$, hipoteze H_0 ne moremo zavrniti

2.7.1 Teoretične osnove testa χ^2

Oglejmo si primer, ko je $r = 2$

	S_1	S_2
opažene frekvence	N	$n - N$
pričakovane frekvence	np	$n(1-p)$

$p = P(\text{prvi razred})$
$N \cdots$ število vrednosti vzorca, ki padejo v 1. razred S_1
Potem je

$$\begin{aligned}
\chi^2 &= \frac{(N-np)^2}{np} + \frac{(n-N-n(1-p))^2}{n(1-p)} = \\
&= \frac{(N-np)^2}{np} + \frac{(N-np)^2}{n(1-p)} = \frac{(N-np)^2}{np(1-p)}((1-p)+p) = \\
&= (\frac{N-np}{\sqrt{np(1-p)}})^2
\end{aligned}$$

Ker je N porazdeljena binomsko $Bin(n,p)$, je pri velikem n slučajna spremenljivka

$$\frac{N - E(N)}{\sigma(N)} = \frac{N - np}{\sqrt{np(1-p)}}$$

porazdeljena po $N(0,1)$ (Laplaceova formula oz. CLI)
Iz verjetnostnega dela vemo, da je kvadrat porazdelitve $N(0,1)$ porazdeljen po $\chi^2(1)$. Torej je χ^2 porazdeljena po $\chi^2(1)$ pri velikem n.
V splošnem primeru (pri poljubnem $r \in \mathbb{N}$) se χ^2 zapiše kotvsota $(r-1)$ kvadratov slučajnih spremenljivk, ki so porazdeljene po $N(0,1)$ in neodvisne. Ker za porazdelitev χ^2 velja $\chi^2(m) + \chi^2(n) \sim \chi^2(m+n)$ (za neodvisne slučajne spremenljivke), je potem χ^2 porazdeljena po $\chi^2(1+1+\cdots+1) = \chi^2(r-1)$

2.8 Test za neznan delež

Naj bo $X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, kjer je p neznan parameter, ki bi ga radi testirali. Testiramo $H_0(p = p_0) : H_1(p \neq p_0)$, kjer je $p \in (0,1)$ dano število. Vemo, da je $\overline{X}$ nepristranska cenilka za p. Po Laplaceovi formuli je

$$\overline{X} \approx N(p_0, \sqrt{\frac{p_0 q_0}{n}})$$

za velike n, če velja hipoteza H_0. Torej je

$$Z := \frac{\overline{X} - p_0}{\sqrt{p_0 q_0 / n}} \approx N(0,1)$$

za velike n.
Pri dani stopnji značilnosti $\alpha > 0$ določimo $z_{\frac{\alpha}{2}}$, da je

$$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$K_\alpha = (-\infty, -z_{\frac{\alpha}{2}}] \cup [z_{\frac{\alpha}{2}}, \infty)$
Če vzorčna vrednost za Z leži na kritičnem območju K_α, potem hipotezo zavrnemo

Primer. Proizvodnjo izdelkov je treba prilagoditi, če delež defektnih izdelkov preseže 10%. Izmed 500 slučajno izbranih izdelkov je 55 defektnih. Pri stopnji značilnosti 5% testiramo hipotezo, ali je potrebno prilagoditi proizvodnjo
$H_0(p_0 = 0.1) : H_1(p_0 > 0.1)$ - enostranski test

$$\begin{aligned} Z &= \frac{\overline{X} - 0.1}{\sqrt{0.1 \cdot 0.9/500}} = \\ &= \frac{\frac{55}{100} - 0.1}{\sqrt{0.1 \cdot 0.9/500}} = \\ &= \frac{0.01}{\sqrt{0.1 \cdot 0.9/500}} = 0.75 \end{aligned}$$

$K_\alpha = [z_\alpha, \infty)$ (ker delamo enostranski test)
Iz tabel razberemo $z_\alpha = 1.54$ ($z_\alpha > 0$ določimo s pogojem $P(Z > z_\alpha) = \alpha$)
Ker $0.75 \notin K_\alpha$, hipoteze ne moremo zavrniti

Za smiselnost testa je pomembno hipoteze formulirati pred analizo podatkov

Primer. Pri neki loteriji izmed števil $0, 1 \cdots 9$ žrebajo po eno število vsak dan. Pri pregledu podatkov za daljše časovno odbobje opaimo, da se število 7 pojavi velikokrat ob sredah: izmed 150 števil se 7 pojavi 22 krat. Testirajmo pri stopnji značilnosti $\alpha = 0.05$ hipotezo, da se 7 pojavi prevečkrat ob sredah
$H_0(p = 0.1) : H_1(p > 0.1)$ enostranski test
$\overline{X} = \frac{22}{150} = 0.147, z_\alpha = 1.64$ kot v prejšnjem primeru

$K_\alpha = [z_\alpha, \infty), n = 150$
$Z = \frac{\overline{X} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.147 - 0.1}{\sqrt{0.1 \cdot 0.9/150}} = 1.91$
Ker $1.91 \in K_\alpha$, hipotezo H_0 zavrnemo
To je primer vohljanja podatkov
P-vrednost $P(Z \geq 1.91) = 0.028$, t.j. verjetnost, da se 7 pojavi ob sredah vsaj 22 krat
Verjetnost, da se 7 pojavi 22 krat v nekem dnevu je $1 - (1 - 0.028)^7 = 0.18$, kar ni več zelo majhno število
Verjetnost, da se neko število pojavi vsaj 22 krat v nekem dnevu pa je še večja

2.9 Neparametrični testi

Doslej smo preizkušali hipoteze o neznanih parametrih v danih porazdelitvah (običajno smo privzeli normalno porazdelitev). To so parametrični testi.
Če na porazdelitev slučajne spremenljivke X ne moremo nič privzeti, potem lahko uporabimo neparametrične teste

2.9.1 Test z znaki

To je analog testa T
Na populaciji imamo 2 slučajni spremenljivki: X s porazdelitveno funkcijo F_X in Y s porazdelitveno funkcijo F_Y.
Obravnavamo 2 slučajna vektorja $(X_1, X_2 \cdots X_m)$ in $(Y_1, Y_2 \cdots Y_n)$, kjer je X_i in Y_i dobimo na istem elementu v populaciji.

Testiramo hipotezo $H_0(F_X = F_Y)$
Definiramo razlike

$$D_i := X_i - Y_i \; i = 1, 2 \cdots n$$

Tukaj smo privzeli, da so vrednosti slučajnega vektorja $(D_1, D_2 \cdots D_n)$ različne od 0, sicer jih izpustimo in zmanjšamo n. Če velja $H_0(F_X = F_Y)$, potem je $P(D_i > 0) = \frac{1}{2} = P(D_i < 0)$ za vsak $i = 1, 2 \cdots n$
Naj bo S^+ število pozitivnih D_i-jev, S^- pa negativnih. Seveda je $S^+ + S^- = n$
Tedaj je $S^+ \sim Bin(n, \frac{1}{2})$, torej je

$$p_k = P(S^+ = k) = \binom{n}{k} 2^{-n} \; k = 0, 1 \cdots n$$

Pri dani stopnji značilnosti $\alpha > 0$ je kritično območje

$$H_\alpha = \{k : k \leq k_\alpha \text{ ali } k \geq n - k_\alpha\}$$

kjer je k_α določen z zahtevama

$$\sum_{k=0}^{k_\alpha} p_k = P(S^+ \leq k_\alpha) \leq \frac{\alpha}{2}$$

in

$$\sum_{k=0}^{k_\alpha + 1} p_k = P(S^+ \leq k_\alpha + 1) > \frac{\alpha}{2}$$

Pri velikem n je S^+ približno normalno porazdeljen $N(\frac{n}{2}, \frac{\sqrt{n}}{2})$, $(\sqrt{npq} = \sqrt{n\frac{1}{2}\frac{1}{2}})$ torej je slučajna spremenljivka

$$Z := \frac{S^+ - \frac{n}{2}}{\frac{\sqrt{n}}{2}} = \frac{2S^+ - n}{\sqrt{n}}$$

približno normalno $N(0, 1)$

Primer. Zdravilo za zniževanje krvnega tlaka pri $n = 10$ bolnikih

X_i	sistolični krvni tlak pred zdravljenjem									
Y_i	sistolični krvni tlak po zdravljenju									
D_i	-8	0	2	4	9	14	19	22	32	35

$H_0(F_X = F_Y)$ zdravilo ne učinkuje
0 prečrtamo in n za 1 zmanjšamo: $n = 9$
Vzorčna vrednost za S^+ je 8
Določimo kritično območje K_α pri stopnji zmačilnosti $\alpha = 0.05$

$$P(S^+ = 0) = \binom{9}{0} 2^{-9} < \frac{\alpha}{2}$$
$$P(S^+ \leq 1) = \binom{9}{0} 2^{-9} < \frac{\alpha}{2} + \binom{9}{1} 2^{-9} = 10 \cdot 2^{-9} = 0.020 \leq \frac{\alpha}{2} = 0.025$$
$$P(S^+ \leq 2) = P(S^+ \leq 1) + P(S^+ = 2) = 10 \cdot 2^{-9} + \binom{9}{2} 2^{-9} =$$
$$= 46 \cdot 2^{-9} \doteq 0.090 > \frac{\alpha}{2} = 0.025$$

Vidimo, da je $k_\alpha = 1$, saj je $P(S^+ \leq 1) < \frac{\alpha}{2}$ in $P(S^+ \leq 2) > \frac{\alpha}{2}$
Kritično območje $K_\alpha = \{0, 1, 8, 9\}$
Ker vzorčna vrednost za S^+, ki je 8, leži na K_α, hipotezo H_0 zavrnemo

Slabost tega testa je, da gledamo samo predznak razlike in ne velikost

$$Z = \frac{2S^+ - n}{\sqrt{n}},\ n = 9$$

Čeprav n ni velik, izračunajmo velikost za Z

$$Z = \frac{2 \cdot 9 - 8}{\sqrt{9}} = \frac{7}{3} = 2.33$$

2.9.2 Inverzijski test

Wilcoxon-Mann-Whitney, 1945
X, Y naj imata porazdelitveni funkciji F_X in F_Y. Vzorca $(X_1, X_2 \cdots X_m)$ in $(Y_1, Y_2 \cdots Y_n)$ sta neodvisna in $m \leq n$ (če to ni res zamenjamo vlogi X in Y). Testirajmo hipotezo $H_0(F_X = F_Y)$. Vzorčne vrednosti vzorcev

$X_1, X_2 \cdots X_m, Y_1, Y_2 \cdots Y_n$ razvrstimo po velikosti $z_1 \leq z_2 \leq \cdots \leq z_{m+n}$ (zapomnimo si ali je iz X ali Y). Pripišimo mesta (Range)

$$R_i = rang(x_i) = h, \text{ če je } z_k = x_i$$

Slučajna spremenljivka $r = R_1 + \cdots + R_m$ ima vrednosti med $\frac{m(m+1)}{2}$ in $mn + \frac{m(m+1)}{2}$
Vrednost $\frac{m(m+1)}{2}$ dobimo, če so X_i na začetku zaporedja $\{z_m\} : 1 + 2 + \cdots + m = \frac{m(m+1)}{2}$
Največjo vrednost pa dobimo, če so X_i na koncu zaporedja: $Y_1, Y_2 \cdots Y_n, X_1, X_2 \cdots X_m$:

$$(n+1) + (n+2) + \cdots + (n+m) = n \cdot m + \frac{m(m+1)}{2}$$

Če velja $H_0(F_X = F_Y)$ in $m + n \geq 0, m, n \geq 4$, potem smemo privzeti, da je V približno normalno porazdeljena

$$V \sim N(\frac{(m+n+1) \cdot m}{2}, \sqrt{\frac{mn(m+n+1)}{12}})$$

oz.

$$Z := \sqrt{\frac{3}{mn(m+n+1)}}(2V - m(m+n+1)) \sim N(0,1)$$

Definicija 2.15 (Inverzija)**.** Inverzija med x_i in y_j se pojavi, če ima y_j manjši rang kot x_i

Zaporedje $x_1, x_2 \cdots x_m, y_1 \cdots y_n$ nima inverzije
Zaporedje $x_1 \cdots x_{m-1}, y_1, x_m, y_2 \cdots y_n$ ima eno inverzijo
Naj bo U število vseh inverzij
Vsaka inverzija, ki jo naredimo, poveča število rangov V za 1. Ker je pri

$$U = 0, V = \frac{m(m+1)}{2}$$

je

$$V = U + \frac{m(m+1)}{2}$$

Primer. 2 zdravili proti nespečnosti. Vzorčne vrednosti razvrstimo po velikosti $m = n = 1$

z_i	zdravilo	rang
-1.6	y	1
-1.2	y	2
-0.2	y	3
-0.1	y	4.5
-0.1	x	4.5
0.1	y	6
0.1	x	7
0.7	y	8
0.8	x	9.5
0.8	y	9.5
1.1	x	11
1.6	x	12
1.9	x	13
2.0	y	14
3.4	x	15.5
3.4	y	15.5
3.7	y	17
4.4	x	18
4.9	x	19
5.5	x	20

Vsota rangov V je $4.5 + 7 + 9.5 + 11 + 12 + 13 + 15.5 + 18 + 19 + 20$

$$Z = \sqrt{\frac{3}{10 \cdot 10 \cdot 21}}(2V - 10 \cdot 21) = \frac{1}{5\sqrt{7}}(V - 105)$$

$K_\alpha = (-\infty, -z_{\frac{\alpha}{2}}] \cup [z_{\frac{\alpha}{2}}, \infty)$
Pri $\alpha = 0.05$ je $z_{\frac{\alpha}{2}} = 1.96$
Vzorčna vrednost za Z je 1.85
Hipoteze $H_0(F_X = F_Y) =$ "zdravili sta enako učinkoviti" ne moremo zavrniti

Inverzijski test je neparametrični analog primerjalnega Studentovaga testa
Inverzijski test "gleda" samo urejenost in ne velikost podatkov