# Zápočtová písemka 23.11.2020 – ukázkové řešení

**1.** Nechť $X_1, \ldots, X_n$ jsou n.n.v s rozdělením $U(0,1)$. Označme $M = \max(X_1, \ldots, X_n)$.

(a) Určete distribuční funkci $M$.

(b) Určete hustotní funkci $M$.

(c) Určete $\mathbb{E}(M)$.

**Řešení:** (a) Napřed si uvědomme jak určit $M$ z $X_1$, …, $X_n$. Podle definice pro každý bod pravděpodobnostního prostoru $\omega \in \Omega$ platí $M(\omega) = \max(X_1(\omega), \ldots, X_n(\omega))$, tj. je to maximum z $n$ hodnot. Pro náš výpočet se ale víc hodí začít s distribuční funkcí $F_M$. Připomeňme si, že to je funkce $\mathbb{R} \to \mathbb{R}$. Její hodnota v $t$ je rovna (podle definice distribuční funkce) $P(M \leq t)$. Teď použijeme definici $M$: maximum z několika čísel je nejvýše $t$ právě tehdy, když je každé z nich nejvýše $t$. Formulkou:

$$F_M(t) = P(M \leq t) = P(X_1 \leq t, X_2 \leq t, \ldots, X_n \leq t)$$

Dále využijeme nezávislosti náhodných veličin $X_1$, …, $X_n$, díky cemuž je poslední výraz roven součinu $\prod_{i=1}^{n} P(X_i \leq t) = F(t)^n$. Tady jsme označili $F$ distribuční funkci proměnných $X_1$, …, $X_n$ (všechny mají stejnou distribuční funkci, protože všechny jsou uniformně rozdělené na $[0,1]$). Jak víme, $F(t) = t$ (pro $t \in [0,1]$), $F(t) = 0$ (pro $t < 0$) a $F(t) = 1$ (pro $t > 1$). Odsud tedy

$$F_M(t) = \begin{cases} 0 & \text{pro } t < 0 \\ t^n & \text{pro } t \in [0,1] \\ 1 & \text{pro } t > 1. \end{cases}$$

(b) Jak víme, hustotu dostaneme derivací na každém intervalu, kde to jde, tj.

$$f_M(t) = \begin{cases} 0 & \text{pro } t < 0 \\ nt^{n-1} & \text{pro } t \in [0,1] \\ 0 & \text{pro } t > 1. \end{cases}$$

(c) Použijeme vzorec pro střední hodnotu spojité náhodné veličiny a už odvozený vzorec pro hustotu.

$$\mathbb{E}(M) = \int_{-\infty}^{\infty} t f_M(t) dt = \int_0^1 t \cdot nt^{n-1} dt = n \int_0^1 t^n dt = n \left[ \frac{t^{n+1}}{n+1} \right]_0^1 = \frac{n}{n+1}$$

**Poznámky k řešením:** Distribuční funkce i hustota jsou funkce – tj. je potřeba spočítat $F_M(t)$ jako funkci čísla $t$. Tato funkce by asi měla záviset na $t$ a na $n$ – pokud některá z těchto proměnných ve vaší odpovědi chybí, je to přinejmenším důvod ke kontrole. Rychlá kontrola je také dosazení extrémních hodnot $n$. Výpočet má platit pro všechna $n$, tj. např. pro $n = 1$, kdy je $M = X_1$, takže získané vzorce by po dosazení $n = 1$ měly odpovídat rozdělení $U(0,1)$. Naopak pokud bude $n$ hodně velké (resp. limitně $n \to \infty$), bude asi $M$ čím dál vyšší – hustota bude „soustředěná více vpravo“, distribuční funkce „vzroste z 0 na 1 později“, a střední hodnota poroste. Nalezené vzorce tohle všechno splňují (rozmyslete).

**2.** Pro soutěž ve sprintu potřebujeme změřit 100 m. Naše měření ale není přesné, rozdíl oproti skutečné hodnotě má normální rozdělení. Předpokládáme, že tento rozdíl má střední hodnotu 0 a směrodatnou odchylku 0.2 m.

(a) Jaká je pravděpodobnost, že naměřená vzdálenost bude kratší než 99.8 m?

(b) Jaká je pravděpodobnost, že naměřená vzdálenost nebude mezi 99.9 a 100.1 m?

Možná se vám bude hodit nabulka distribuční funkce standardního normálního rozdělení $N(0,1)$.

| $x$ | $-2$ | $-1.5$ | $-1$ | $-0.5$ | $0$ | $0.5$ | $1$ | $1.5$ | $2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi(x)$ | 0.0228 | 0.0668 | 0.1587 | 0.3085 | 0.5000 | 0.6915 | 0.8413 | 0.9332 | 0.9772 |

**Řešení:** Podle zadání naměříme hodnotu $X = 100 + R$, kde $R \sim N(\mu, \sigma^2)$, kde $\mu = 0$ a $\sigma = 0.2$ m. Podle znalostí z přednášky má $Z = (R - \mu)/\sigma$ standardní normální rozdělení, tj. $Z \sim N(0, 1)$. Budeme tedy používat $\Phi$, distribuční funkci $Z$.

(a)

$$P(X < 99.8) = P(R < -0.2) = P(Z < -1) = \Phi(-1) \doteq 0.1587.$$

Ekvivalentně, pro $X \sim N(\mu, \sigma^2)$ platí $P(R < \mu + (-1) \cdot \sigma) = \Phi(-1)$.

(b)

$$\begin{aligned} P(X < 99.9 \text{ nebo } X > 100.1) &= P(R < -0.1 \text{ nebo } R > 0.1) \\ &= P(Z < -0.5 \text{ nebo } Z > 0.5) \\ &= P(Z < -0.5) + P(Z > 0.5) \\ &= \Phi(-0.5) + 1 - \Phi(0.5) \\ &= 2\Phi(-0.5) \doteq 2 \cdot 0.3085 = 0.617 \end{aligned}$$

**Poznámky k řešením:** Někteří si popletli rozdělení $R$ a $X$, a pak dosazovali do $\Phi$ hodnoty okolo 100. To by mělo být podezřelé, takové hodnoty jsou zanedbatelně malé, zatímco hledané pravděpodobnosti od oka moc malé nejsou. (Samozřejmě by mohla být diskuze o tom, jaká je reálná hodnota $\sigma$ – záleželo by jak na použité technice, tak i na tom, kdo měří.)

Několik lidí převedlo úlohu na integrál z funkce $e^{-t^2/2}$ (nebo podobné). To je věcně správně, a s vhodným numerickým software by to umožnilo i získat správnou odpověď. Na druhou stranu pokud ten software nebude jen numerický, ale bude to například Wolfram Alpha, tak pozná známý integrál a podívá se do tabulky. A když už ta tabulka byla v zadání ...

**3.** Nechť $X, Y$ jsou nezávislé náhodné veličiny s rozdělením $Geom(p)$.

(a) Napište sdruženou pravděpodobnostní funkci $p_{X,Y}$.

(b) Spočtěte $P(X + Y = n)$.

(c) Spočtěte $P(X = k \mid X + Y = n)$.

**Řešení:** (a) Připomeňme si, že geometrické rozdělení jsme definovali předpisem $p_X(k) = (1-p)^{k-1}p$ pro $k \geq 1$ (a 0 jinak). Stejnou hodnotu má samozřejmě i $p_Y(k)$. Sdružená pravděpodobnostní funkce je dána vzorcem (pro $k, \ell \geq 1$)

$$\begin{aligned} p_{X,Y}(k,\ell) &= P(X = k, Y = \ell) && \text{podle definice} \\ &= P(X = k)P(Y = \ell) && \text{protože } X, Y \text{ jsou nezávislé} \\ &= (1-p)^{k-1}p(1-p)^{\ell-1}p && \text{podle vzorce pro geometrické rozdělení} \\ &= (1-p)^{k+\ell-2}p^2. \end{aligned}$$

(b) Potřebujeme konvoluční vzorec (nebo ho znovu dokážeme):

$$\begin{aligned} P(X + Y = n) &= \sum_{k \in Im X} P(X = k, Y = n - k) && \text{rozbor všech možností, jak může být } X + Y = n \\ &= \sum_{k \in Im X} p_{X,Y}(k, n-k) && \text{definice } p_{X,Y} \\ &= \sum_{k \in Im X} p_X(k)p_Y(n-k) && \text{nezávislost – tohle je konvoluční vzorec} \\ &= \sum_{k=1}^{n-1} (1-p)^{k+(n-k)-2}p^2 && \text{podle části (a)} \\ &= (n-1)(1-p)^{n-2}p^2 && \text{všechny sčítance jsou stejné} \end{aligned}$$

(c)

$$\begin{aligned}
P(X=k \mid X+Y=n) &= \frac{P(X=k, X+Y=n)}{P(X+Y=n)} && \text{definice podmíněné pravděpodobnosti!}\\
&= \frac{P(X=k, Y=n-k)}{P(X+Y=n)} && \text{čítatel napíšeme jinak, ale ekvivalentně}\\
&= \frac{(1-p)^{k+(n-k)-2}p^2}{(n-1)(1-p)^{n-2}p^2} && \text{čítatel podle (a), jmenovatel podle (b)}\\
&= \frac{1}{n-1}.
\end{aligned}$$

Všimněte si, že na druhém řádku jsme mohli nahradit jev $X+Y=n$ jevem $Y=n-k$ – protože se pohybujeme uvnitř jevu $X=k$, tj. víme, čemu se $X$ rovná.

**Poznámky k řešením:** Hodně lidí zapomínalo, že sdružená pravděpodobnostní funkce je funkce více proměnných – pro každou možnost hodnoty $X$ a $Y$ potřebuji zjistit pravděpodobnost té kombinace. Několik lidí v části (b) správně napsali sumu, ale nevšimli si, že všechny členy jsou stejné, takže sečíst ji je triviální. V části (c) přiliš mnoho lidí tápalo v definici podmíněné pravděpodobnosti!!

Interpretace výsledku úlohy: Házíme kostkou, dokud nepadne druhá šestka. Pokud je to v $n$-tém hodu, tak ta první šestka mohla padnout při prvním, druhém, ..., $(k-1)$-ním hodu (to je jasné) a to ve všech se stejnou pravděpodobností $\frac{1}{n-1}$. To už tolik jasné není, i když to asi není tak nečekané. Rozdělení náhodné veličiny $N=X+Y$ se nazývá *Pascalovo řádu* 2. Obecně Pascalovo rozdělení řádu $r$ popisuje, jak dlouho musíme čekat na $r$-tý úspěch (po $r$-té nám padne šestka). Rozdělení o $r$ menší (počítáme jen neúspěchy) se nazývá *negativní binomické*.

**4.** Hugo pro přijímání emailů používá vlastní server. Ten má každý den poruchu s pravděpodobností $p$ (a pak ten den nedoručí žádný email). Pokud nemá poruchu, doručí server všechny emaily – a můžeme předpokládat, že jejich počet má Poissonovo rozdělení s parametrem $\lambda$. (To znamená, že pokud Hugo nedostal žádný email, tak je to buď kvůli poruše serveru, nebo proto, že mu nikdo nenapsal, což Poissonovo rozdělení umožňuje.) Předpokládejme také, že porucha serveru je nezávislá na počtu emailů. Označme $X$ počet emailů, které Hugo za jeden den dostal.

(a) Napište pravděpodobnostní funkci náhodné veličiny $X$.
(b) Vyjádřete $X$ jako součin Bernoulliovské a Poissonovské náhodné veličiny.
(c) Spočtěte $\mathbb{E}(X)$.
(d) Spočtěte $var(X)$.

V částech (c), (d) můžete použít část (b), nebo vyjít přímo z definice a použít (a).

**Řešení:** (a) Pro $k>0$ máme jen jeden způsob, jak $X=k$ (server nemá poruchu s pravděpodobností $p$ a Poisson vygeneruje $k$). Pro $k=0$ máme dvě možnosti, jak je popsáno v zadání. Z toho plyne vzorec

$$p_X(k) = \begin{cases} (1-p)\frac{\lambda^k}{k!}e^{-\lambda} & \text{pro } k>0 \\ (1-p)e^{-\lambda}+p & \text{pro } k=0 \end{cases}$$

(b) Označme jako $U$ indikátorovou veličinu jevu „server nemá poruchu“, tj. $U=1$ pokud server nemá poruchu, jinak $U=0$. Je tedy $U \sim Bern(1-p)$. Dále buď $M$ počet emailů, které měly být doručeny; podle zadání $M \sim Pois(\lambda)$. Rozborem možností (server má/nemá poruchu) ověříme, že $X=UM$.

(c) Veličiny $U$, $M$ jsou podle zadání nezávislé. Proto platí

$$\mathbb{E}(X) = \mathbb{E}(UM) = \mathbb{E}(U)\mathbb{E}(M) = (1-p)\lambda.$$

Alternativně bychom mohli postupovat rozborem možností. Označíme $D$ jev „server má poruchu“ a počítáme

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X \mid D)P(D) + \mathbb{E}(X \mid D^c)P(D^c)\\
&= 0 \cdot p + \lambda \cdot (1-p).
\end{aligned}$$

Ještě jiný postup (asi nejpracnější) spočívá v postupu dle definice:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{\infty} k p_X(k) && \text{definice střední hodnoty} && (1)\\
&= \sum_{k=1}^{\infty} k p_X(k) && \text{člen s } k=0 \text{ je nulový} && (2)\\
&= \sum_{k=1}^{\infty} k(1-p)\frac{\lambda^k}{k!}e^{-\lambda} && \text{dosadíme podle (a)} && (3)\\
&= (1-p)\sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} && \text{vytkneme před závorku} && (4)\\
&= (1-p)\lambda\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}e^{-\lambda} && \text{ještě jednou} && (5)\\
&= (1-p)\lambda e^{\lambda}e^{-\lambda} && \text{Taylorova řada pro } e^{\lambda} && (6)\\
&= (1-p)\lambda && && (7)
\end{aligned}$$

Taky jsme mohli na řádku (4) zjistit, že je to stejná řada jako při výpočtu střední hodnoty Poissonova rozdělení, a tudíž je rovna $\lambda$.

(d) Použijeme rozklad z části (b) a vzorec $var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Protože $\mathbb{E}(X)$ už známe, stačí spočítat jen

$$\begin{aligned}
\mathbb{E}(X^2) &= \mathbb{E}(U^2M^2) && \\
&= \mathbb{E}(U^2)\cdot\mathbb{E}(M^2) && \text{nezávislost}\\
&= \mathbb{E}(U)\cdot\mathbb{E}(M^2) && \text{protože } U^2 = U\\
&= \mathbb{E}(U)\cdot(var(M)+\mathbb{E}(M)^2) && \text{vzorec pro rozptyl použitý obráceně}\\
&= (1-p)\cdot(\lambda+\lambda^2) && \text{abychom mohli použít, co si pamatujeme}
\end{aligned}$$

Dohromady tedy dostáváme

$$var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = (1-p)\cdot(\lambda+\lambda^2) - p^2(1-\lambda)^2 = \lambda(1-p)(1+\lambda p).$$

**Poznámky k řešením:** Kupodivu hodně lidí tápalo už v části (a) – a ne proto, že by nevěděli, jaký je vzorec pro Poissonovo rozdělení, ale nevěděli, jak ho zkombinovat s pravděpodobností poruchy serveru. To je přitom jenom základní výpočet s pravděpodobnostmi (nezávislé jevy – pravděpodobnosti se násobí). Ale možná to bylo tím, že je to poslední úloha, navíc s dlouhým zadáním.

K interpretaci výsledku: jak jsme spočítali, je zde $var(X) \geq \mathbb{E}(X)$ (s rovností pro $p = 0$, kdy se dostáváme zpět do případu Poissonova rozdělení). Je to tedy dobré rozdělení pro případ, kdy data mají větší rozptyl než střední hodnotu (tedy zejména, pokud je ten rozptyl způsobený hodně častou nulou ...).