Matematika IV – 10. přednáška
Transformace a číselné charakteristiky náhodných veličin

Michal Bulant

Masarykova univerzita
Fakulta informatiky

21. 4. 2008

Obsah přednášky

Doporučené zdroje

- Martin Panák, Jan Slovák, **Drsná matematika**, *e*-text.
- Karel Zvára, Josef Štěpán, **Pravděpodobnost a matematická statistika**, Matfyzpress, 4. vydání, 2006, 230 stran, ISBN 80-867-3271-1.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Teorie pravděpodobnosti a matematická statistika (sbírka příkladů)**, Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, **Popisná statistika**, Masarykova univerzita, 3. vydání, 2002, 48 stran, ISBN 80-210-1831-3.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, **Základní statistické metody**, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.

Příklad (rozdělení $\chi^2(1)$)

Nechť Z má normované normální rozdělení. Určete hustotu transforomované náhodné veličiny $X = Z^2$.

Řešení

Zřejmě je pro $x \leq 0$ distribuční funkce nulová, pro $x > 0$ dostáváme: $F_X(x) = P[Z^2 < x] = P[-\sqrt{x} < Z < \sqrt{x}] =$

$$= \int_{-\sqrt{x}}^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}} \, \mathrm{d}z = \int_0^{\sqrt{x}} \frac{1}{\sqrt{2\pi}} t^{-\frac{1}{2}} \mathrm{e}^{-\frac{t}{2}} \, \mathrm{d}t$$

a derivací podle x dostaneme hustotu

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} \mathrm{e}^{-\frac{x}{2}} \, .$$

Rozdělení náhodné veličiny s touto hustotou se nazývá (Pearsonovo) χ^2 rozdělení s jedním stupněm volnosti a značí se $X \sim \chi^2(1)$.

Transformace náhodných veličin

Místo náhodné veličiny X, např. „roční plat zaměstnance”, budeme vyčíslovat jinou závislou hodnotu $\psi(X)$, např. „roční čistý příjem zaměstnance po zdanění a včetně sociálních dávek”. V systému se značnou sociální solidaritou je první veličina hodně variabilní, zatímco druhá může být skoro konstantní. Statisticky se proto budou značně odlišovat.

Připomeňme si přechod od binomického k Poissonovu rozdělení z minulé přednášky:

Věta (Poissonova)

Je-li $X_n \sim \mathrm{Bi}(n, p_n)$ *taková, že* $\lim_{n\to\infty} np_n = \lambda$ *a* $X \sim \mathrm{Po}(\lambda)$, *pak*

$$\lim_{n\to\infty} P[X_n = k] = P[X = k]$$

pro $k = 0, 1, \ldots$.

Binomické rozdělení $\mathrm{Bi}(n, p)$ odpovídá n–krát nezávisle opakovanému pokusu popsanému alternativním rozdělením, přičemž naše náhodná veličina měří počet zdarů. Je tedy

$$f_X(t) = \begin{cases} \binom{n}{t} p^t (1-p)^{1-t} & t \in \{0, 1, \dots, n\} \\ 0 & \text{jinak} \end{cases} .$$

$$\begin{aligned} \lim_{n\to\infty} P(X_n = k) &= \lim_{n\to\infty} \binom{r_n}{k} \frac{(n-1)^{r_n - k}}{n^{r_n}} \\ &= \lim_{n\to\infty} \frac{r_n(r_n-1)\dots(r_n-k+1)}{(n-1)^k} \frac{1}{k!} \left(1 - \frac{1}{n}\right)^{r_n} \\ &= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1 + \frac{-\frac{r_n}{n}}{r_n}\right)^{r_n} = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \end{aligned}$$

Poissonovo rozdělení popisuje náhodné veličiny s pravděpodobnostní funkcí $f_X(t) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}$ pro $t \in \mathbb{N}$

Nejjednodušší funkcí, po konstantách, je afinní závislost $\psi(x) = a + bx$.
V případě afinní závislosti $x = \frac{1}{b}(y - a)$ je proto pravděpodobnostní funkce nenulová právě v bodech $y_i = ax_i + b$. V případě rozdělení X_n typu $\mathrm{Bi}(n, p)$ převádí transformace $x = y\sqrt{np(1-p)} + np$ náhodnou veličinu X_n na rozdělení Y_n s distribuční funkcí blízkou distribuční funkci spojitého rozdělení $N(0, 1)$.
Dříve uvedená Poissonova věta popisuje asymptotické chování binomického rozdělení pří $n \to \infty$ a $p \to 0$, následující věta pak chování v případě konstantní pravděpodobnosti zdaru p.

Věta (de Moivre-Laplaceova)

Pro náhodné veličiny X_n s rozdělením $\mathrm{Bi}(n, p)$ platí

$$\lim_{n\to\infty} P\left[a < \frac{X_n - np}{\sqrt{np(1-p)}} < b\right] = \Phi(b) - \Phi(a),$$

kde Φ je distribuční funkce normovaného normálního rozdělení.

Příklad

Hodíme kostkou celkem 12 000 krát. Určete pravděpodobnost toho, že počet hozených šestek je mezi 1 800 a 2 100.

Řešení

Přesná pravděpodobnost je dána výrazem
$\sum_{k=1800}^{2100} \binom{12000}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{12000-k}$, což je obtížně vyčíslitelné.
Využijeme tvrzení Moivre-Laplaceovy věty, přepsaného do tvaru

$$P\left[A < X_n < B\right] - \left(\Phi\left(\frac{B - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{A - np}{\sqrt{np(1-p)}}\right) \right) \to 0$$

pro $n \to 0$.

Řešení (pokr.)

Volbou $p = 1/6, A = 1800, B = 2100, n = 12000$ dostáváme odhad

$$P \approx \Phi\left(\frac{2100 - 2000}{\sqrt{12000 \cdot \frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{1800 - 2000}{\sqrt{12000 \cdot \frac{1}{6}\frac{5}{6}}}\right) =$$

$$= \Phi(\sqrt{6}) - \Phi(-2\sqrt{6}) \approx 0,992.$$

Poznámka

Statistické tabulky – viz např. `https://is.muni.cz/auth/el/1433/jaro2008/MB104/um/StatTab.pdf` nebo sbírka příkladů [BMO]. Osecký.

Příklad

Pravděpodobnost narození chlapce je $0,515$. Jaká je pravděpodobnost, že mezi tisíci novorozenci bude alespoň tolik děvčat jako chlapců?

Příklad

Nezávisle opakujeme pokus s výsledky 1 a 0, které mají **neznámé** pravděpodobnosti p a $1-p$. Parametr p chceme odhadnout pomocí *relativních četností* X_n/n (X_n je počet jedniček při n pokusech). Víme, že je $X_n \sim \mathrm{Bi}(n,p)$, proto nám Moivre-Laplaceova věta umožní určit počet pokusů n potřebný k zajištění požadované přesnosti odhadu δ se spolehlivostí $1-\beta$.

Řešení

Využijeme Moivre-Laplaceovu větu zapsanou ve tvaru

$$0 = \lim_{n\to\infty}\left| P\left[\left|\frac{X_n}{n} - p\right| < \delta\right] - \left(\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{np(1-p)}}\right)\right)\right|$$

Řešení

Hledáme nejmenší n, splňující nerovnost $P[|X_n/n - p| < \delta] \geq 1 - \beta$, kterou můžeme podle věty aproximovat nerovností

$$\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - \Phi\left(-\frac{n\delta}{\sqrt{np(1-p)}}\right) =$$
$$= 2\Phi\left(\frac{n\delta}{\sqrt{np(1-p)}}\right) - 1 \geq 1 - \beta.$$

Ta je ekvivalentní s podmínkou $n\delta/\sqrt{np(1-p)} \geq z(\beta/2)$, kde $z(p)$ je řešení rovnice $\Phi(z(p)) = 1 - p$ (tzv. *kritická hodnota* normovaného normálního rozdělení). Pro $\delta = 0,05$ a $1 - \beta = 0,9$ máme z tabulek $z(\beta/2) \approx 1,645$ a s využitím zřejmého odhadu $p(1-p) \leq 1/4$ dostáváme $n \geq (z(\beta/2)/2\delta)^2 \approx 270,6$.

Transformace normálně rozložené veličiny

Podobně zkusme opačnou transformaci provést na veličinu Y s normálním rozdělením $N(0,1)$. Pro pevně zvolená čísla $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$ spočtěme rozdělení náhodné veličiny $Z = \mu + \sigma Y$. Dostáváme distribuční funkci

$$\begin{aligned} F_Z(z) &= P(Z < z) = P(\mu + \sigma Y < z) \\ &= F_Y\left(\frac{z-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{z-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\, dt \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x, \end{aligned}$$

kde poslední úprava vychází ze substituce $x = \mu + \sigma t$. Hustota naší nové náhodné veličiny Z je proto

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

a takovému rozdělení se říká normální typu $\mathrm{N}(\mu, \sigma)$.

Střední hodnota

Při statistickém zkoumání hodnot náhodných veličin (např. zpracování výsledků nějakého měření) hledáme výpovědi o náhodné veličině pomocí různých z ní odvozených čísel.
Jako nejjednodušší příklad může sloužit **střední hodnota**[1] $E(X)$ náhodné veličiny X, která je definována

$$E(X) = \begin{cases} \sum_i x_i \cdot f_X(x_i) & \text{pro diskrétní veličinu} \\ \int_{-\infty}^{\infty} x \cdot f_X(x)\,\mathrm{d}x & \text{pro spojitou veličinu.} \end{cases}$$

Obecně střední hodnota náhodných veličin nemusí existovat, protože příslušné sumy či integrály nemusí konvergovat.

[1] Často se místo $E(X)$ píše EX.

Střední hodnota transforomované náhodné veličiny

Střední hodnotu můžeme přímo vyjádřit také pro funkce $Y = \psi(X)$ náhodné veličiny X. V diskrétním případě můžeme přímo spočíst

$$
\begin{aligned}
E(Y) &= \sum_j y_j P(Y = y_j) \\
&= \sum_j y_j \sum_{\psi(x_i)=y_j} P(X = x_i) \\
&= \sum_i \psi(x_i) P(X = x_i).
\end{aligned}
$$

Je tedy $E(\psi(X))$ přímo spočítatelná pomocí pravděpodobnostní funkce f_X.
Podobně vyjadřujeme střední hodnotu funkce ze spojité náhodné veličiny:

$$E(\psi(X)) = \int_{-\infty}^{\infty} \psi(x) f_X(x) \, \mathrm{d}x,$$

pokud tento integrál absolutně konverguje.

Příklad

Spočtěme střední hodnotu binomického rozdělení.

Řešení

Pro $X \sim \mathrm{Bi}(n, p)$ je

$$
\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!(k-1)!} p^{k-1} (1-p)^{n-k} = \\
&= np \sum_{j=0}^{n-1} \frac{(n-1)!}{(n-1-j)!j!} p^{j} (1-p)^{n-1-j} = \\
&= np.
\end{aligned}
$$

Základní vlastnosti střední hodnoty

Věta

Nechť $a, b \in \mathbb{R}$ a X, Y jsou náhodné veličina s existující střední hodnotou. Pak

- $E(a) = a$,
- $E(a + bX) = a + bE(X)$,
- $E(X + Y) = E(X) + E(Y)$,
- *jsou-li X a Y* **nezávislé**, *pak $E(XY) = E(X) \cdot E(Y)$.*

Důkazy těchto tvrzení jsou přímočaré, zkuste si je udělat!
Analogická tvrzení platí i pro náhodné vektory.

Příklad

Spočtěme ještě jednou střední hodnotu binomického rozdělení, tentokrát s využitím vlastností střední hodnoty.

Řešení

Vyjádříme počet zdarů v n pokusech jako počet zdarů v jednotlivých pokusech

$$X = \sum_{k=1}^{n} Y_i,$$

přičemž náhodné veličiny Y_k mají všechny alternativní rozdělení $A(p)$. Snadno spočítáme $E(Y_k) = 1 \cdot p + 0 \cdot (1 - p) = p$. Dále víme, že střední hodnota součtu je součtem středních hodnot, proto

$$E(X) = \sum_{k=1}^{n} E(Y_k) = np.$$

Kvantily

Dalšími užitečnými charakteristikami jsou tzv. **kvantily**. Pro ryze monotóní distribuční funkci F_X (tj. spojitou náhodnou veličinu X s všude nenulovou hustotou, jako je tomu např. u normálního rozdělení) jde prostě o inverzní funkci $F_X^{-1} : (0,1) \to \mathbb{R}$. To znamená, že hodnota $y = F^{-1}(\alpha)$ je taková, že $P(X < y) = \alpha$. Obecněji, je-li $F_X(x)$ distribuční funkce náhodné veličiny X, pak definujeme **kvantilovou funkci**[2]

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R};\ F(x) \geq \alpha\},\ \ \alpha \in (0,1).$$

Zřejmě jde o zobecnění předchozí definice.
Nejčastěji jsou používané kvantily s $\alpha = 0.5$, tzv. **medián**, s $\alpha = 0.25$, tzv. **první kvartil**, $\alpha = 0.75$, tzv. **třetí kvartil**, a podobně pro **decily** a **percentily** (kdy je α rovno násobkům desetin a setin). K těmto hodnotám se vrátíme v popisné statistice později.

[2] Uvědomte si, že jsme se již s kvantily setkali, jen jsme jím tak zatím neříkali.

Rozptyl a směrodatná odchylka

Tyto číselné charakteristiky rozdělení náhodné veličiny nepopisují nějakou střední či typickou hodnotu (jako střední hodnota či medián), ale míru ,,kolísání" náhodné veličiny kolem střední hodnoty.

Rozptylem (variancí) náhodné veličiny X, která má konečnou střední hodnotu, nazýváme číslo

$$D(X) = \operatorname{var} X = E([X - E(X)]^2),$$

odmocnina z rozptylu $\sqrt{D(x)}$ se pak nazývá **směrodatná odchylka**.

Základní vlastnosti rozptylu

Věta

Pro náhodnou veličinu X a reálná čísla a, b platí:

1. $D(X) = E(X^2) - E(X)^2$,
2. $D(a + bX) = b^2 D(X)$,
3. $\sqrt{D(a + bX)} = |b|\sqrt{D(X)}$.

Důkaz.

Důkaz je přímočarý – nejprve se dokáže 2. tvrzení, pak se z něj tvrzení první. Poznamenejme, že tvrzení 1 se často používá k výpočtům $D(X)$. $\square$

Kovariance

O závislosti dvou náhodných veličin do jisté míry vypovídá tzv. **kovariance**, definovaná předpisem

$$C(X,Y) = \text{cov}(X,Y) = E([X - E(X)][Y - E(Y)]).$$

Veličinám X, Y, pro něž je $C(X,Y) = 0$, říkáme **nekorelované** .

Věta

Pro náhodné veličiny s existujícími rozptyly platí:

1. $(X,Y) = C(Y,X)$,
2. $C(X,X) = D(X)$,
3. $C(X,Y) = E(XY) - E(X)E(Y)$,
4. $C(a + bX, c + dY) = bdC(X,Y)$,
5. $D(X+Y) = D(X) + D(Y) + 2C(X,Y)$, *speciálně, jsou-li* X, Y *nezávislé, je* $D(X+Y) = D(X) + D(Y)$, *tj.* $C(X,Y) = 0$ *a* X, Y *jsou nekorelované.*

Koeficient korelace

Koeficient korelace je jen speciální název pro kovarianci dvou normovaných náhodných veličin:

$$R(X, Y) = \rho_{X,Y} = C\left(\frac{X - E(X)}{\sqrt{D(X)}}, \frac{Y - E(Y)}{\sqrt{D(Y)}}\right).$$

Věta

1. $R(X, X) = 1$,
2. $R(a + bX, c + dY) = \operatorname{sgn}(bd) R(X, Y)$,
3. *jsou-li X, Y nezávislé, je* $R(X, Y) = 0$,
4. $|R(X, Y)| \leq 1$.

Příklad

Spočtěme rozptyl binomického rozdělení.

Řešení

Stejně jako dříve lze psát $X = \sum_{k=1}^{n} Y_k$, kde $Y_1, \ldots, Y_n$ jsou nezávislé náhodné veličiny vyjadřující úspěch v k-tém pokusu. Snadno vypočteme $E(Y_k^2) = 1^2 \cdot p + 0^2 \cdot (1-p) = p$, proto $D(Y_k) = E(Y_k^2) - E(Y_k)^2 = p - p^2 = p(1-p)$. Protože pro **nezávislé** Y_k platí $D(\sum Y_k) = \sum D(Y_k)$, je

$$D(X) = np(1-p).$$

Všimněme si, že výraz $X_n - np/\sqrt{np(1-p)}$ vystupující v Moivre-Laplaceově větě je totéž, co $X_n - E(X_n)/\sqrt{D(x)}$ a jde tedy o tzv. normovanou náhodnou veličinu (tj. veličinu lineárně transforomovanou tak, aby měla střední hodnotu 0 a rozptyl 1). Moivre-Laplaceova věta pak říká, že pro $n \to \infty$ se rozložení této náhodné veličiny blíží normovanému normálnímu rozdělení $N(0,1)$.

Další momenty

Někdy je užitečné studovat řadu dalších charakteristik rozdělení náhodných veličin. Za rozumných předpokladů jsou definovány *k*-**té obecné momenty**

$$\mu'_k = E(X^k)$$

a *k*-**té centrální momenty**

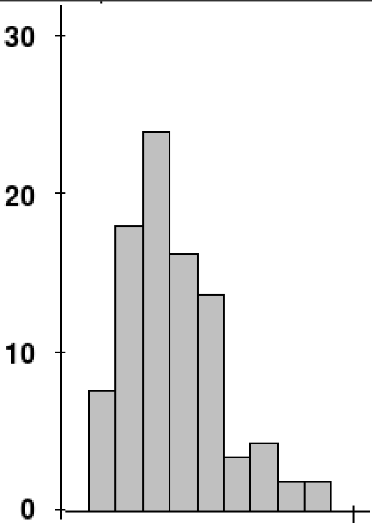
$$\mu_k = E([X - E(X)]^k).$$

Pomocí momentů pak definujeme např. **šikmost** (asymetrii) náhodné veličiny X jako

$$\frac{\mu_3}{\sqrt{D(x)}^3}$$

nebo **špičatost** (exces) jako

$$\frac{\mu_4}{D(x)^2} - 3.$$



Kladná šikmost distribuce (více vysokých kladných hodnot než odpovídá normálnímu rozdělení s nulovou šikmostí).

Momentová vytvořující funkce

Definice

Reálnou funkci proměnné $t \in \mathbb{R}$ $M_X(t) = E(e^{tX})$ nazveme **momentovou vytvořující funkcí** náhodné veličiny X.

Poznámka

Je-li X např. spojitá, platí

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x = \\ &= \int_{-\infty}^{\infty} \left(1 + tx + \frac{t^2x^2}{2!} + \cdots\right) f(x)\,\mathrm{d}x = \\ &= 1 + t\mu_1' + \frac{t^2\mu_2'}{2!} + \cdots \end{aligned}$$

a jde vlastně o *exponenciální vytvořující funkci* posloupnosti k-tých obecných momentů μ_k'.

Věta

Pro momentovou vytvořující funkci platí:

- $\mu'_k = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)\,|_{t=0}$.
- *Platí-li* $M_X(t) = M_Y(t)$ *pro všechna* $t \in \langle -b, b \rangle$, *mají náhodné veličiny stejné rozdělení, tj.* $F_X(x) = F_Y(x)$.
- $M_{a+bX}(t) = e^{at} M_X(bt)$.
- *Jsou-li* X, Y *nezávislé, je* $M_{X+Y}(t) = M_X(t) M_Y(t)$.

Příklad

Určete momentovou vytvořující funkci binomického rozdělení.

Řešení

$$M(t) = E(e^{tX}) = \sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} =$$
$$= \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} =$$
$$= (pe^t + (1-p))^n = (p(e^t - 1) + 1)^n.$$

Snáze jsme mohli funkci určit s využitím předchozích vět a momentové vytvořující funkce alternativního rozdělení, neboť $E(e^{tX}) = e^{t\cdot 1} \cdot p + e^{t \cdot 0}(1-p) = p(e^t - 1) + 1.$

Příklad

Naposled spočtěme střední hodnotu a rozptyl binomického rozdělení, tentokrát s využitím vytvořující funkce.

Řešení

$M(t) = (p(e^t - 1) + 1)^n$, proto je

$$\frac{\mathrm{d}}{\mathrm{d}t} M(t) = n(p(e^t - 1) + 1)^{n-1} e^t p,$$

což pro $t = 0$ dá $E(X) = \mu_1' = np$.
Podobně spočítáme i $D(x) = \mu_2' - (\mu_1')^2$.

Momenty normálního rozdělení

Přímý výpočet střední hodnoty a rozptylu normovaného normálního rozdělení není triviální. S využitím momentové vytvořující funkce je ale poměrně jednoduchý.

Nechť $Z \sim N(0,1)$. Pak

$$
\begin{aligned}
M_Z(t) &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z = \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2 - 2tz + t^2 - t^2}{2}\right) \mathrm{d}z = \\
&= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2}\right) \mathrm{d}z = \exp\left(\frac{t^2}{2}\right).
\end{aligned}
$$

Poslední integrál je roven 1 díky tomu, že na místě integrované funkce je funkce s vlastnostmi hustoty.

Střední hodnota a rozptyl normálního rozdělení

S využitím předchozího výpočtu $M_Z(t) = \exp\left(\frac{t^2}{2}\right)$ snadno spočítáme, že

$$M'_Z(t) = t\exp\left(\frac{t^2}{2}\right),$$
$$M''_Z(t) = t^2\exp\left(\frac{t^2}{2}\right) + \exp\left(\frac{t^2}{2}\right).$$

Dosazením $t = 0$ pak dostaneme

$$E(Z) = 0, D(Z) = 1.$$

Pro transformovanou náhodnou veličinu $Y = \mu + \sigma Z \sim N(\mu, \sigma^2)$ pak snadno odvodíme z vlastností střední hodnoty, resp. rozptylu, že $E(Y) = \mu, D(Y) = \sigma^2$ (což zpětně zdůvodňuje zápis $N(\mu, \sigma^2)$). Momentová vytvořující funkce má tvar

$$M_Y(t) = \exp\left(\mu t + \sigma^2\frac{t^2}{2}\right).$$