Pravděpodobnost a matematická statistika

Doc. RNDr. Gejza Dohnal, CSc.

dohnal@nipax.cz

Pravděpodobnost a matematická statistika
2010

1.týden (20.09.-24.09.) Data, typy dat, variabilita, frekvenční analýza (histogramy, četnosti absolutní, relativní, prosté, kumulativní), základní statistické charakteristiky (průměr, výběr.rozptyl, minimum, maximum, medián, kvartily, boxplot), sešikmenná rozdělení (vzájemná poloha mediánu a střední hodnoty), chvosty, kvantily

2. týden (27.09.-01.10.) Princip statistické indukce, výběr, vlastnosti výběru, experiment. Náhodná veličina, rozdělení pravděpodobnosti a jeho souvislost s histogramem. Pravděpodobnost, pravidla pro počítání s pravděpodobností, podmíněná pravděpodobnost, závislost náhodných veličin.

3.týden (04.10.-08.10.) Využití závislosti při stanovení pravděpodobnosti - věta o úplné pravděpodobnosti a Bayesova věta

4.týden (11.10.-15.10.) Rozdělení chyb měření - normální rozdělení a počítání s ním. Odhady parametrů normálního rozdělení. Intervaly spolehlivosti pro normální data. Jednovýběrové testy o střední hodnotě

5.týden (18.10.-24.10.) Výběrový poměr jako odhad pravděpodobnosti sledovaného jevu. Alternativní rozdělení, binomické rozdělení. Intervalový odhad výběrového poměru. Výběry s vracením a bez vracení (binomické a hypergeometrické rozdělení)

6.týden (25.10.-29.10.) odpadá

7.týden (01.11.-05.11.) Poruchy v čase (Poissonův proces). Poissonovo rozdělení, exponenciální rozdělení, jeho výhody a nevýhody, modelování doby do poruchy pomocí Weibullova rozdělení, lognormálního rozdělení, případně useknuté normální rozdělení.

8.týden (08.11.-12.11.) Testy dobré shody, Q-Q graf (pouze vysvětlení), testy normality. Některé neparametrické testy

9.týden (15.11.-19.11.) Dvě náhodné veličiny - srovnání dvou výběrů (dvouvýběrové testy)

10. týden (22.11.-26.11.) Dvě náhodné veličiny. Dvourozměrné četnosti jako odhad dvourozměrného rozdělení, frekvenční tabulka. Marginální rozdělení (vše pouze diskrétně s tabulkou)

11. týden (29.11.-03.12.) Závislost náhodných veličin, míry závislosti (kovariance, korelace), test významnosti korelačního koeficientu

12. týden (06.12.-10.12.) Regrese, lineární regresní model (přímková, kvadratická, polynomická regrese), analýza reziduí, pásy spolehlivosti

13. týden (13.12.-17.12.) Více výběrů, jednoduché třídění, ANOVA.

14. týden (20.12.-22.12.) Rezerva, opakování, testy normality (náhrada za 28.10.)

Výběrový poměr

Úloha: Jaká je pravděpodobnost, že balíček kávy, který si koupí náhodný zákazník, bude mít hmotnost menší, než je dolní hranice intervalu spolehlivosti pro průměr?

24.52586	24.17119	24.54486	24.44240	23.93455	24.20389	24.19974	24.34851
23.94024	24.21022	24.87474	25.06155	25.48924	25.32572	23.71721	24.61622
25.06676	24.90055	24.36213	24.98580	24.80591	24.20853	24.72623	24.64437
24.70405	23.97645	25.29837	24.46910	24.99453	25.42994	24.66147	24.75773
25.03970	24.44901	25.13285	24.40205	24.78721	23.83656	24.17186	23.65390
24.48244	24.68550	24.22988	23.83956	24.09777	24.52098	24.89240	24.25332
24.14259	25.12906						

$$\langle 24.868, 25.132 \rangle$$

Výběrový poměr

Úloha: Jaká je pravděpodobnost, že balíček kávy, který si koupí náhodný zákazník, bude mít hmotnost menší, než je dolní hranice intervalu spolehlivosti pro průměr?

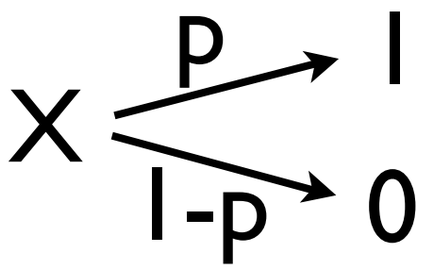
24.52586	24.17119	24.54486	24.44240	23.93455	24.20389	24.19974	24.34851
23.94024	24.21022	24.87474	25.06155	25.48924	25.32572	23.71721	24.61622
25.06676	24.90055	24.36213	24.98580	24.80591	24.20853	24.72623	24.64437
24.70405	23.97645	25.29837	24.46910	24.99453	25.42994	24.66147	24.75773
25.03970	24.44901	25.13285	24.40205	24.78721	23.83656	24.17186	23.65390
24.48244	24.68550	24.22988	23.83956	24.09777	24.52098	24.89240	24.25332
24.14259	25.12906						

$$\langle 24.868, 25.132 \rangle$$

pod hranicí: 36 36/50 = 0.72
v mezích: 14 14/50 = 0.28
celkem: 50

Výběrový poměr = statistický bodový odhad pravděpodobnosti sledovaného jevu

Alternativní rozdělení



přibližně v 100.p% případů nastane výsledek 1

přibližně v 100.(1-p)% případů nastane výsledek 0

střední hodnota X: $E(X) = p.1 + (1-p).0 = p$

rozptyl X: $Var(X) = E\left(X^2\right) - \left(E(X)\right)^2 = p.1 + (1-p).0 - p^2 = p(1-p)$

absolutní četnost kladných výsledků = součet pozorování $Y = \sum_{i=1}^{n} X_i$

$E(Y) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np \qquad Var(Y) = E(Y - np)^2 = np(1-p)$

relativní četnost kladných výsledků = aritmetický průměr pozorování $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}np = p$

$Var(\bar{X}) = E(\bar{X} - p)^2 = \frac{p(1-p)}{n}$

Intervalový odhad výběrového poměru

$$U = \frac{Y - np}{\sqrt{np(1-p)}} \sim N(0,1) \qquad U = \frac{\bar{X} - p}{\sqrt{p(1-p)}}\sqrt{n} \sim N(0,1)$$

Intervalový odhad pro výběrový poměr = Intervalový odhad pravděpodobnosti sledovaného jevu

$$\left\langle \bar{X} - \sqrt{\frac{p(1-p)}{n}}u_\alpha, \bar{X} + \sqrt{\frac{p(1-p)}{n}}u_\alpha \right\rangle$$

Test hypotézy o výběrovém poměru:

$H_0 : p = p_0$

$H_A : p \neq p_0$

$$T = \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \qquad T = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$$

Nulovou hypotézu zamítneme, když $|T| \geq u_\alpha$ pro námi stanovené α

Výběr bez vracení

Sportka: 49 čísel, ze kterých 6 vyhrává (jsou vytaženy).
Jaká je pravděpodobnost, že při výběru 6ti čísel vybereme 4 z tažených?

Kontrola jakosti: 1000 výrobků, mezi nimi jsou 3% vadných.
Jaká je pravděpodobnost, že při výběru 10 výrobků vybereme alespoň 1 zmetek?

Výběr uchazečů o práci: z 15ti uchazečů o zaměstnání, mezi kterými je 10 žen, vybíráme anonymně podle výsledku testu 5 osob.
Jaká je pravděpodobnost, že to budou samé ženy?

Obecně: N prvků, mezi nimiž je M s určitou sledovanou vlastností.
Jaká je pravděpodobnost, že při výběru n prvků bez vracení vybereme k prvků se sledovanou vlastností?

$$P(k; n, N, M) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

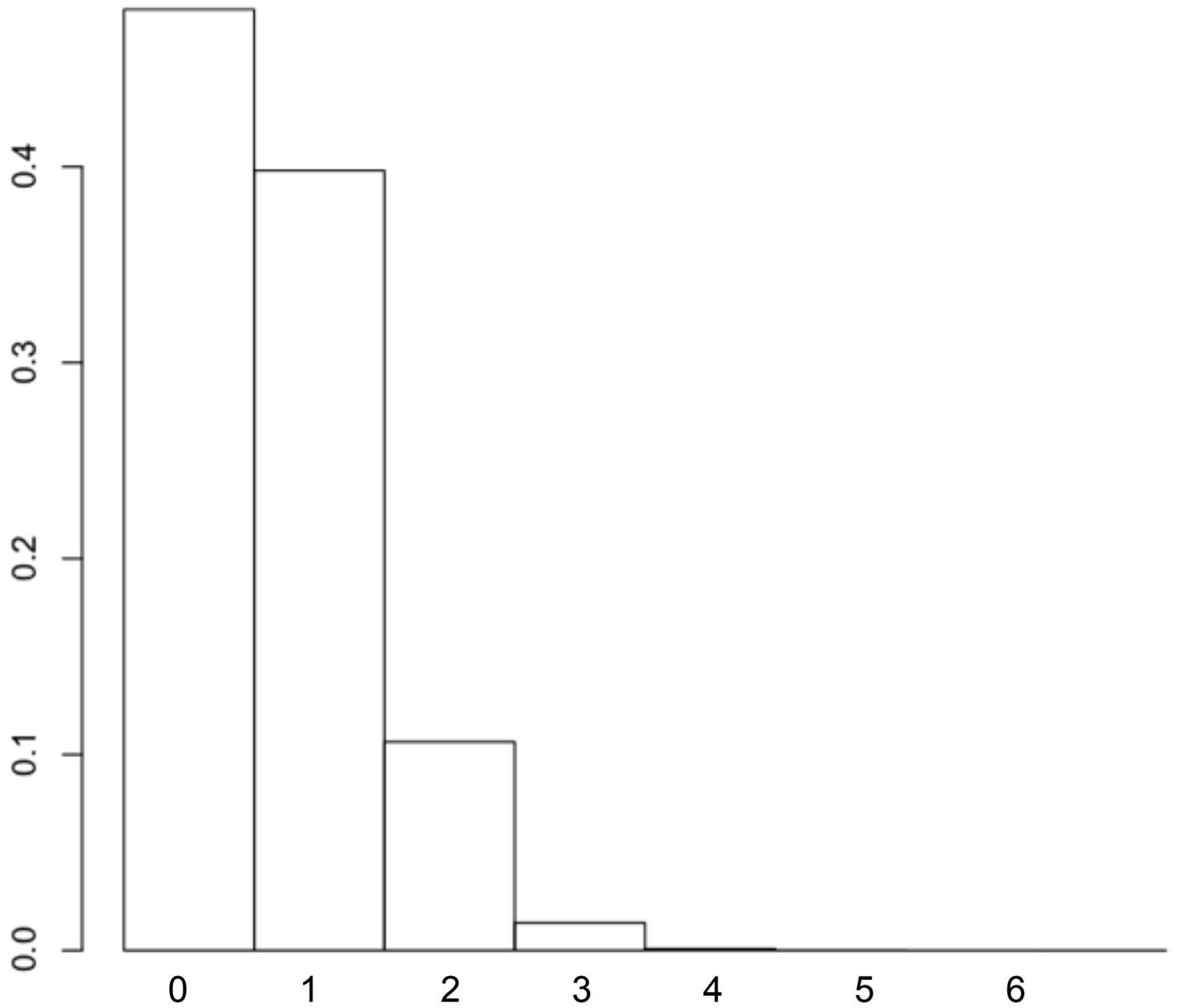
počet k-tic v M prvcích

počet zbylých (n-k)-tic z ostatních (N-M) prvků

počet všech možností = počet n-tic z N prvků

Výběr bez vracení

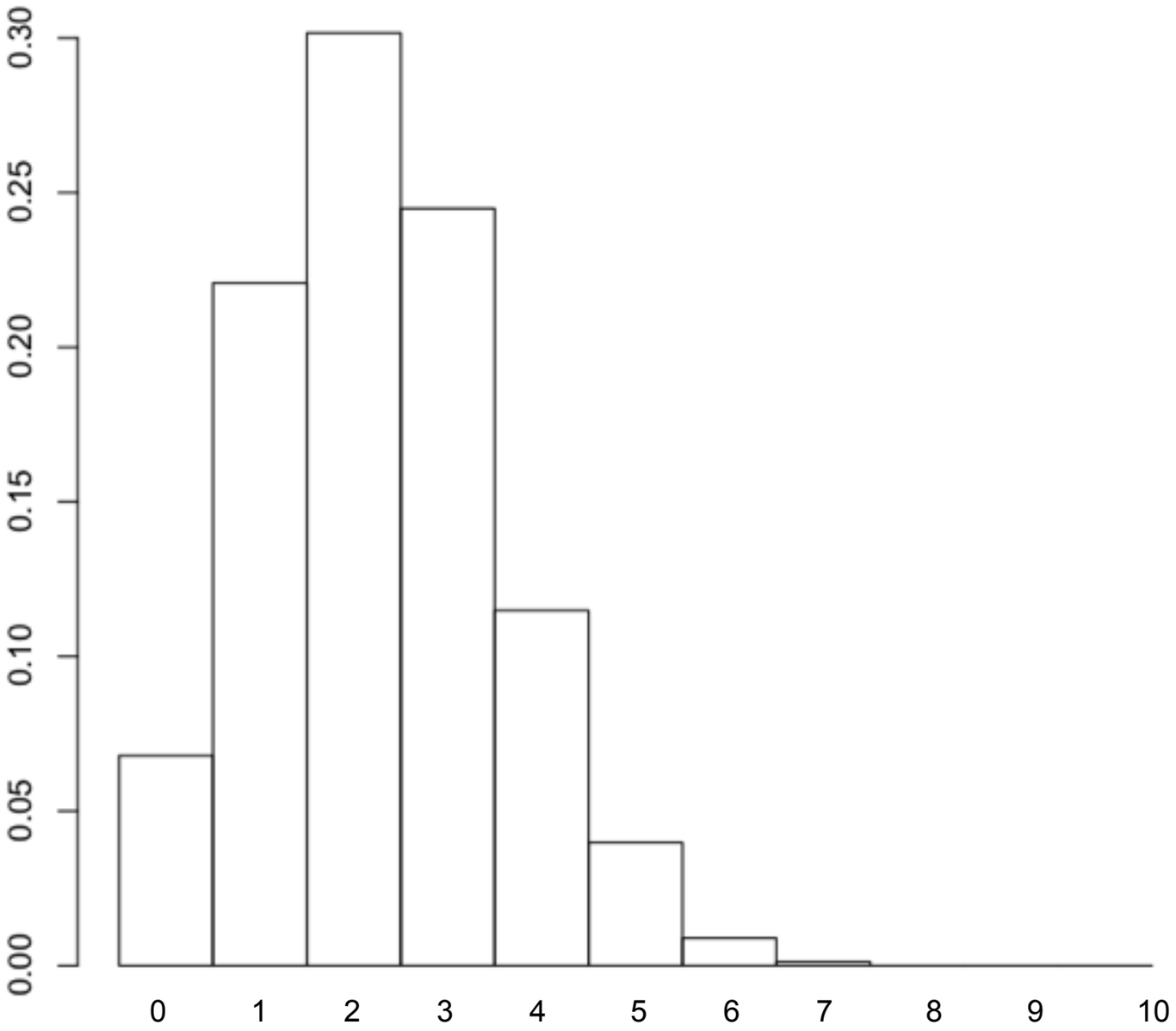
Sportka: 49 čísel, ze kterých 6 vyhrává (jsou vytaženy).
Jaká je pravděpodobnost, že při výběru 6ti čísel vybereme 4 z tažených?



Výběr bez vracení

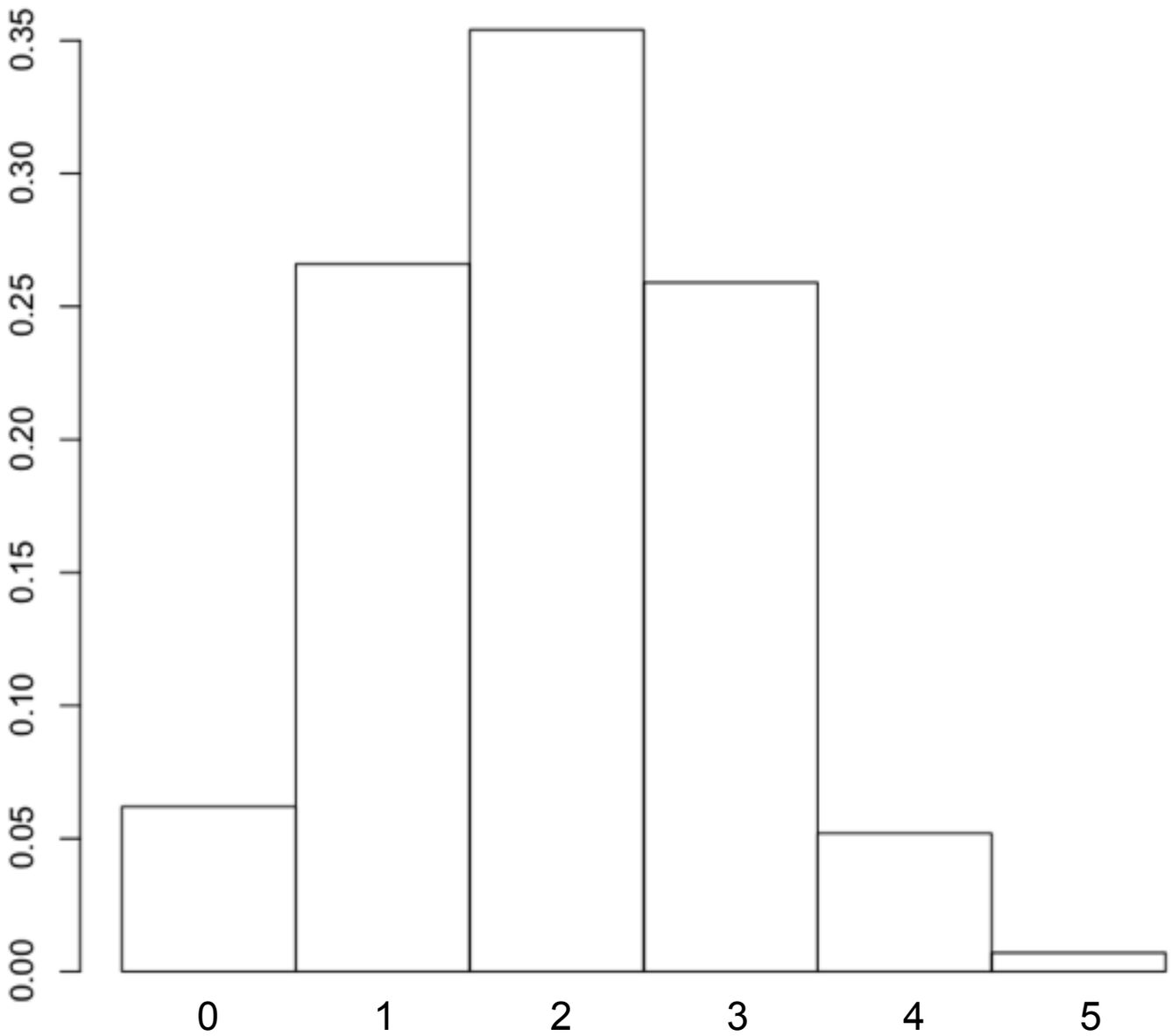
Kontrola jakosti: 1000 výrobků, mezi nimi jsou 3% vadných.

Jaká je pravděpodobnost, že při výběru 10 výrobků vybereme alespoň 1 zmetek?



Výběr bez vracení

Výběr uchazečů o práci: z 15ti uchazečů o zaměstnání, mezi kterými je 10 žen, vybíráme anonymně podle výsledku testu 5 osob.
Jaká je pravděpodobnost, že to budou samé ženy?



Hypergeometrické rozdělení

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

$$N = 1, 2, \ldots, \quad M \leq N, \quad n \leq N, \quad max(0, n + M - N) \leq k \leq min(n, M)$$

$$E(X) = n\frac{M}{N}$$

$$Var(X) = \frac{nM(N-n)(N-M)}{N^2(N-1)}$$

Výběr s vracením

Házení kostkou: házíme třemi hracími kostkami současně (nebo jednou třikrát po sobě).
Jaká je pravděpodobnost, že padnou alespoň dvě šestky?

Kontrola jakosti: Z výrobní linky odebíráme nezávisle na sobě 10 výrobků. Víme, že v produkci jsou 3% vadných. Jaká je pravděpodobnost, že při výběru vybereme alespoň 1 zmetek?

Losování zaměstnance: každý den v týdnu losujeme jednoho z 15ti zaměstnanců, který provede odpolední úklid. Mezi zaměstnanci 10 žen.
Jaká je pravděpodobnost, že v týdnu vybereme samé ženy?

Obecně: N prvků, mezi nimiž je M s určitou sledovanou vlastností.
Jaká je pravděpodobnost, že při výběru n prvků s vracením vybereme k prvků se sledovanou vlastností?

počet k-tic v n prvcích

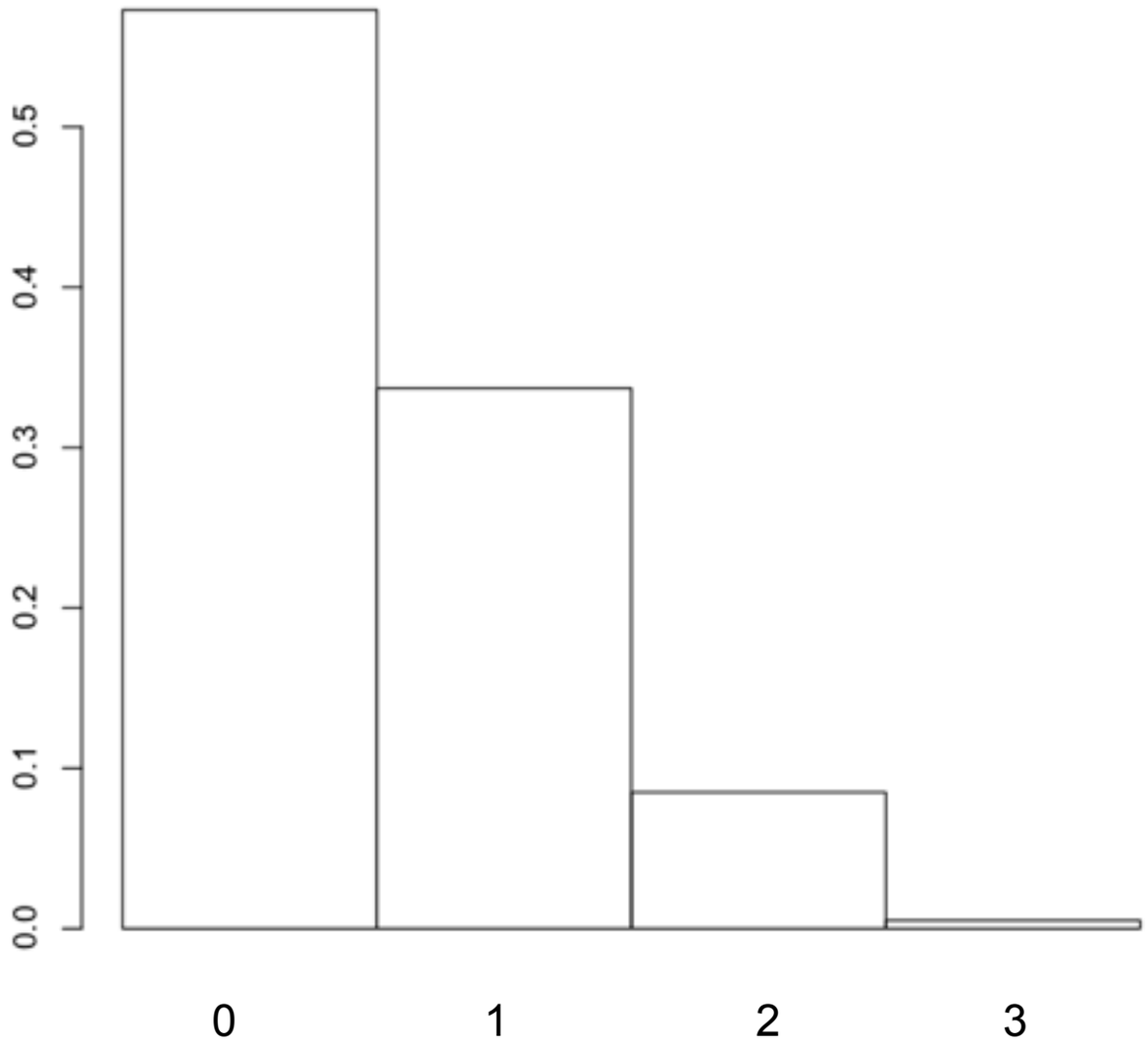
$(n\text{-}k)$-krát vybereme prvek s pravděpodobností $(1\text{-}\frac{M}{N})$

$$P(k;n,N,M) = \binom{n}{k}\left(\frac{M}{N}\right)^{k}\left(\frac{N-M}{N}\right)^{n-k}$$

k-krát vybereme prvek s pravděpodobností $\frac{M}{N}$

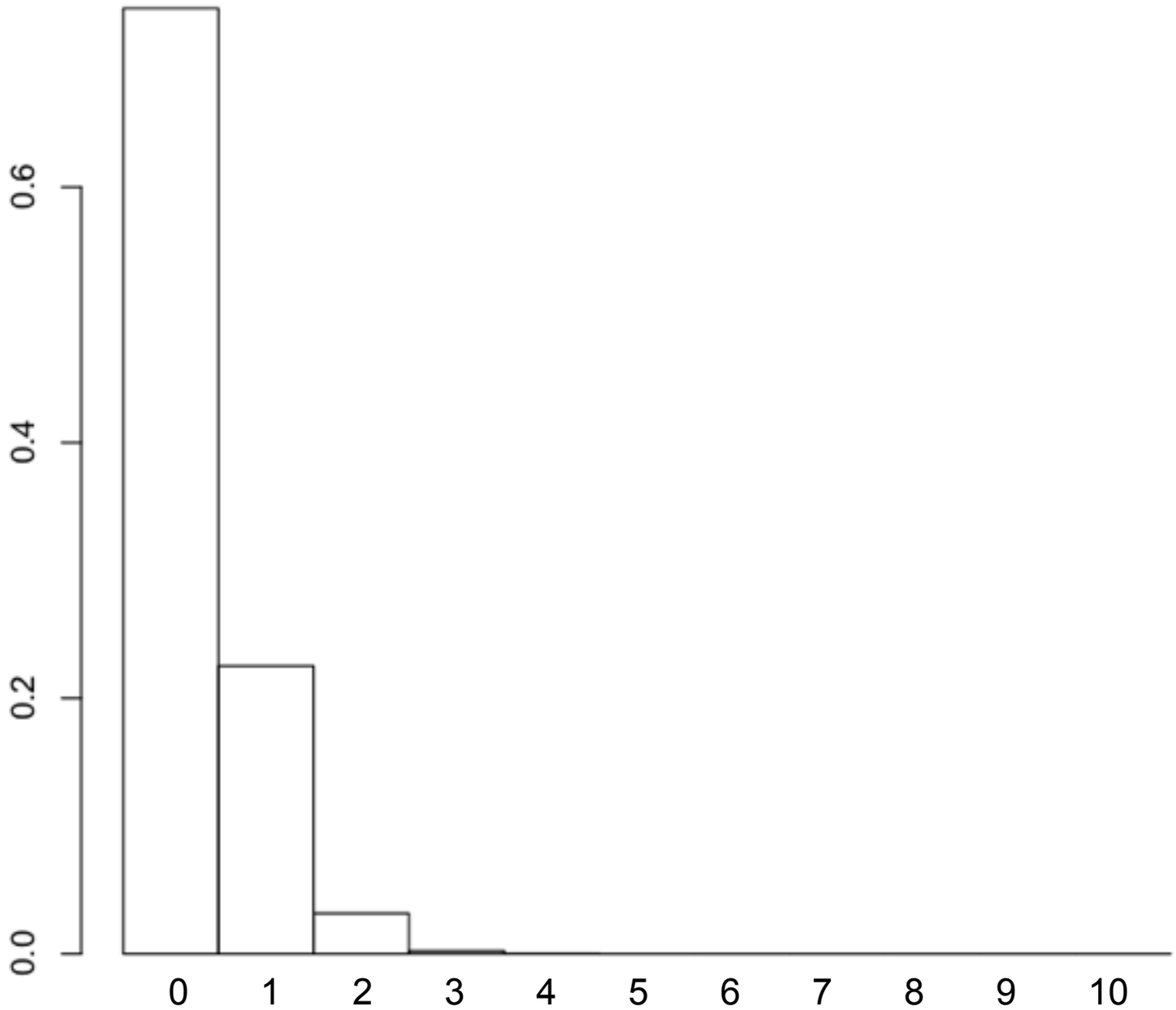
Výběr s vracením

Házení kostkou: házíme třemi hracími kostkami současně (nebo jednou třikrát po sobě).
Jaká je pravděpodobnost, že padnou alespoň dvě šestky?



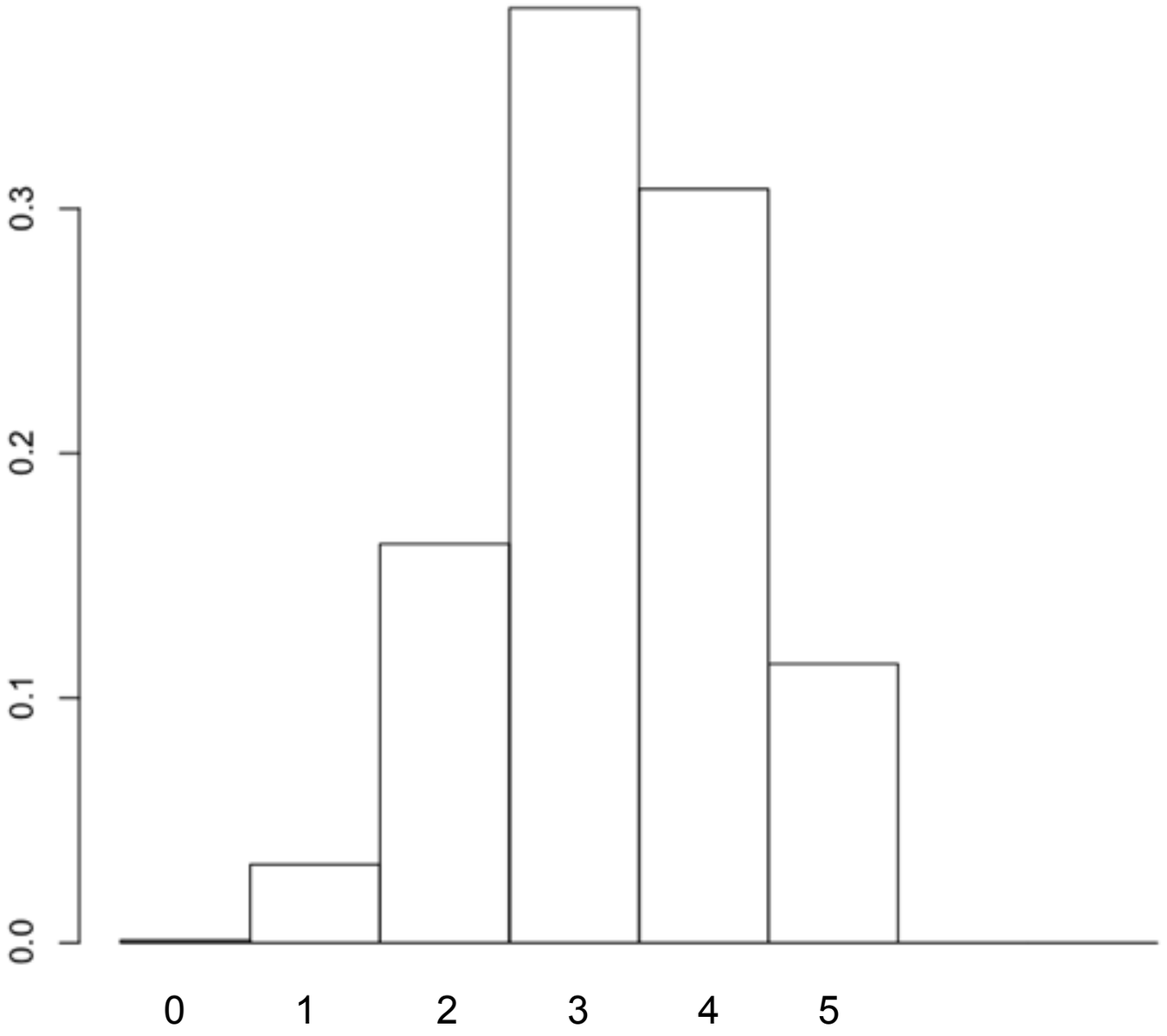
Výběr s vracením

Kontrola jakosti: Z výrobní linky odebíráme nezávisle na sobě 10 výrobků. Víme, že v produkci jsou 3% vadných. Jaká je pravděpodobnost, že při výběru vybereme alespoň 1 zmetek?



Výběr s vracením

Losování zaměstnance: každý den v týdnu losujeme jednoho z 15ti zaměstnanců, který provede odpolední úklid. Mezi zaměstnanci 10 žen.
Jaká je pravděpodobnost, že v týdnu vybereme samé ženy?



Binomické rozdělení

$$P(k; n, N, M) = \binom{n}{k} \left(\frac{M}{N}\right)^k \left(\frac{N-M}{N}\right)^{n-k}$$

$$N = 1, 2, \ldots, \quad M \leq N, \quad n \leq N, \quad max(0, n + M - N) \leq k \leq min(n, M)$$

Obvyklejší je tvar

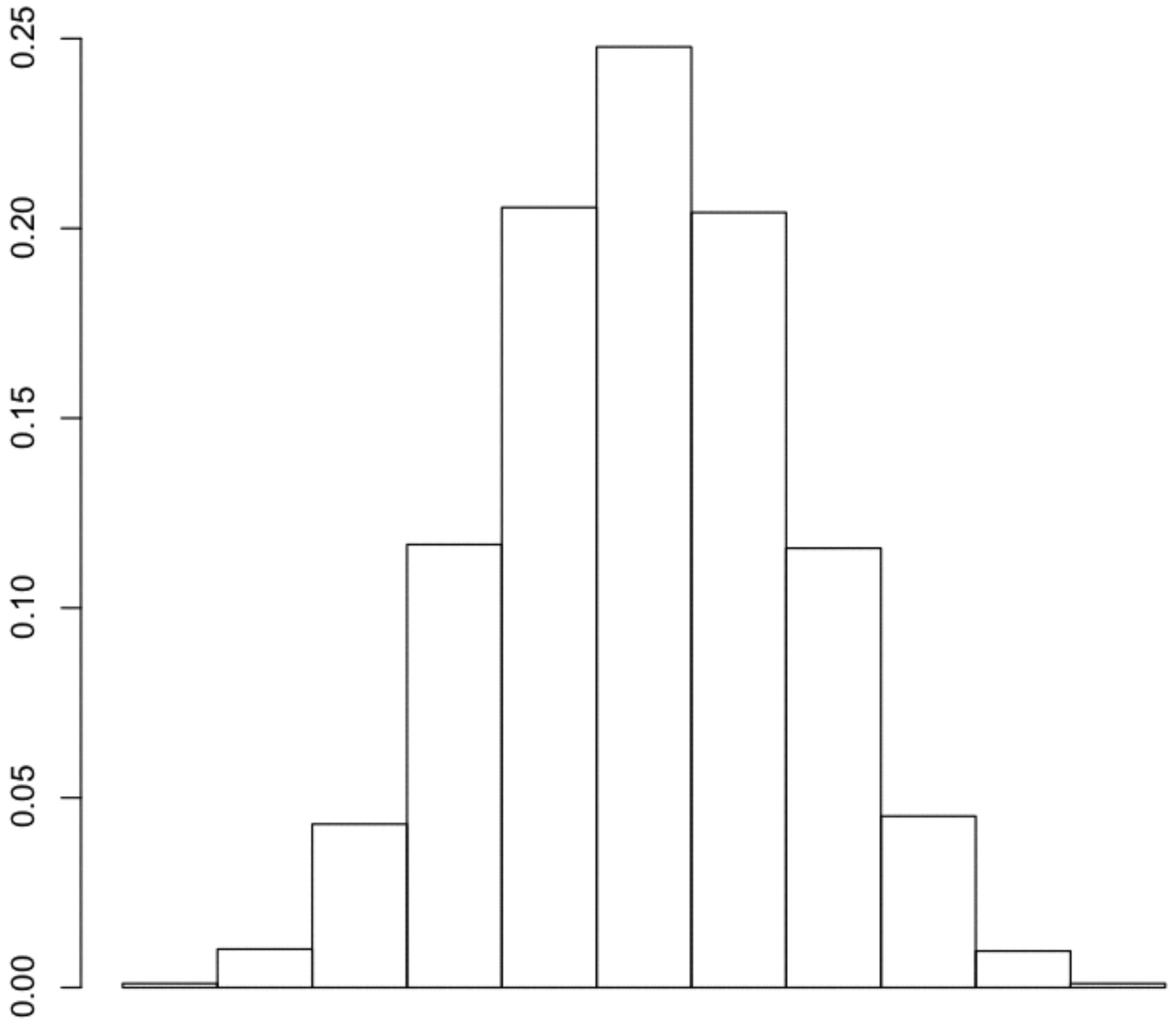
$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$n = 1, 2, \ldots, \quad p \in (0, 1), \quad k = 0, 1, \ldots, n$$

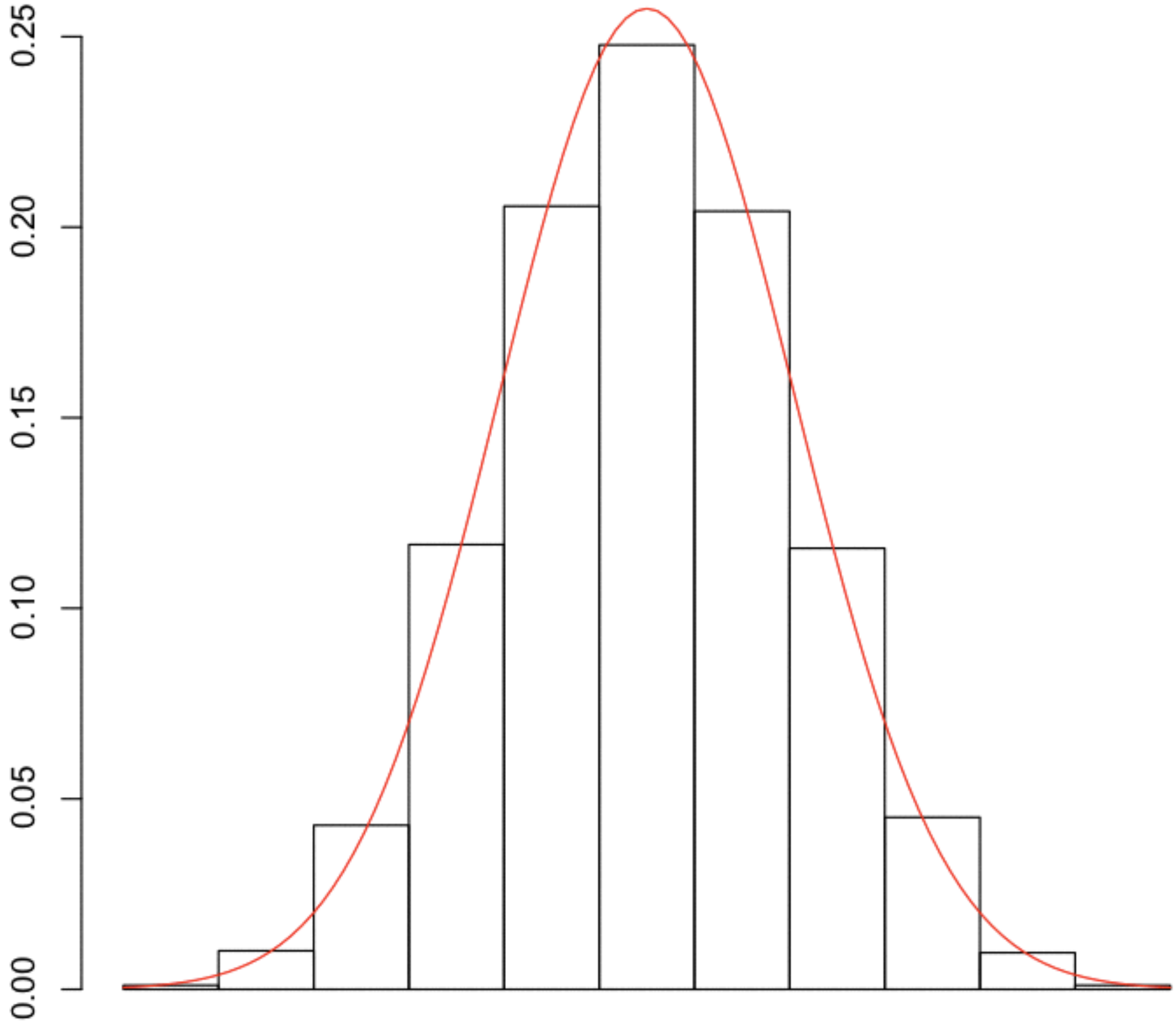
Náhodná veličina s binomickým rozdělením popisuje počet úspěchů při n nezávislých opakováních bernoulliovských pokusů s pravděpodobností úspěchu p.

$$E(X) = np \qquad Var(X) = np(1-p)$$

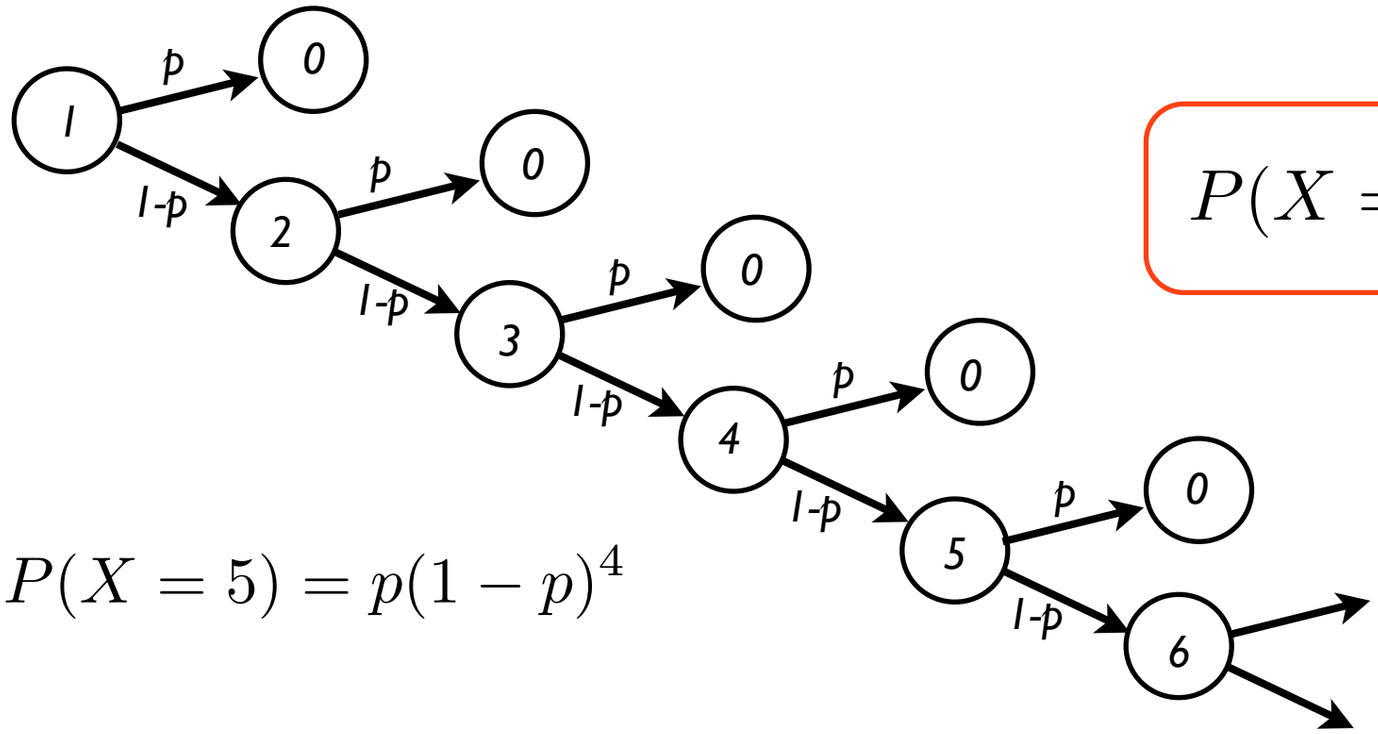
Binomické rozdělení



Binomické rozdělení



Geometrické rozdělení



$$P(X = k) = p(1-p)^{k-1}$$

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

$$P(X = 5) = p(1-p)^4$$

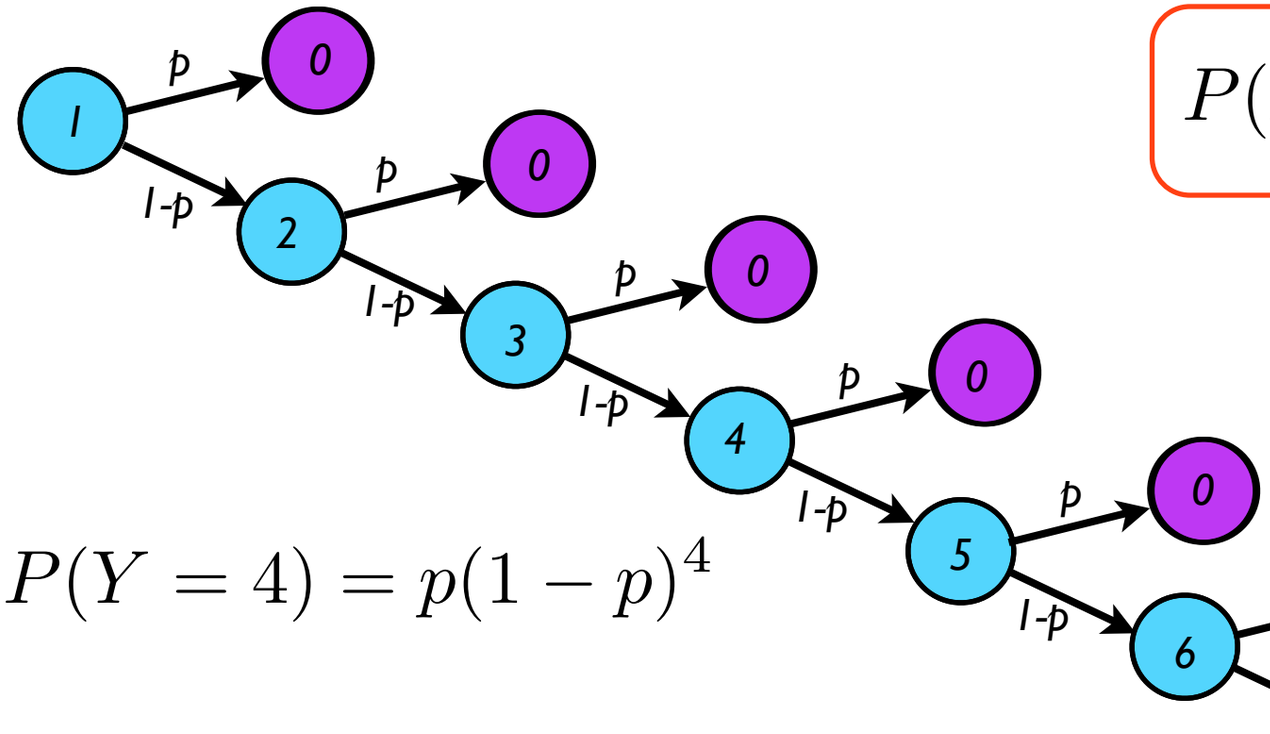
X je počet kroků, které je třeba učinit, aby nastal první výskyt sledovaného jevu

$$E(X) = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$

Geometrické rozdělení

X je počet kroků, které je třeba učinit, aby nastal první výskyt sledovaného jevu



$$P(X = k) = p(1-p)^{k-1}$$

$$k = 1, 2, \ldots$$

$$P(X = 5) = p(1-p)^4$$

$$P(Y = 4) = p(1-p)^4$$

Y je počet kroků, které předcházejí prvnímu výskytu sledovaného jevu

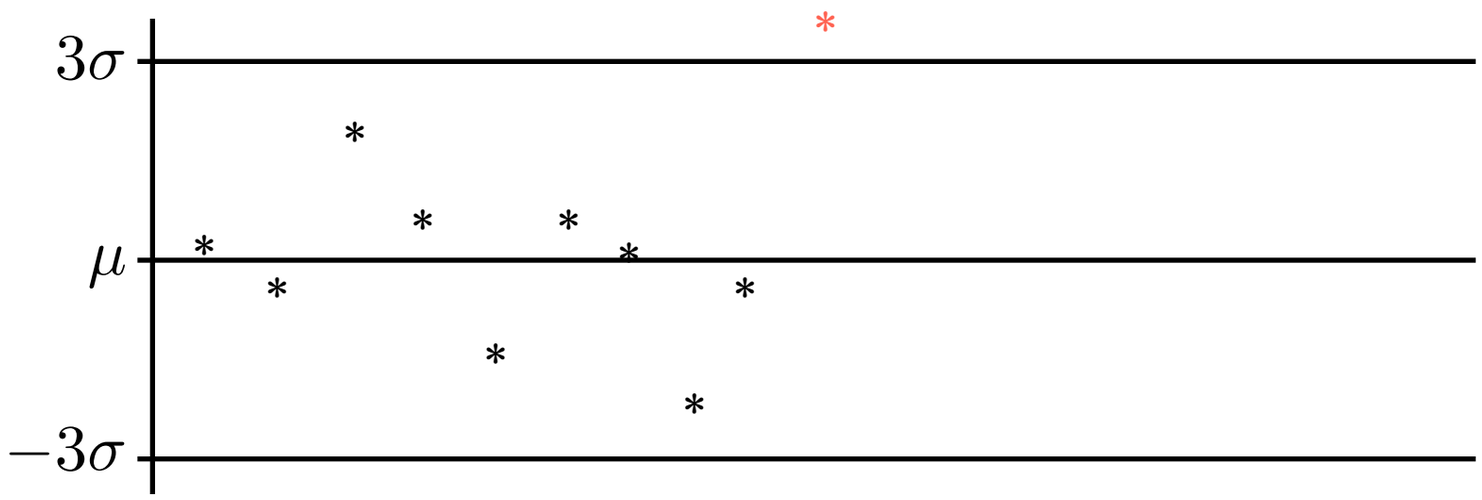
$$P(Y = k) = p(1-p)^k$$

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

$$E(Y) = \frac{1-p}{p} \qquad Var(Y) = \frac{1-p}{p^2}$$

Je-li sledovaný jev porucha, potom se Y nazývá “diskrétní doba života”

Geometrické rozdělení



$$P(|X-\mu| \leq 3\sigma) = \Phi(3) - \Phi(-3) = 2\Phi(3) - 1 = 0,9973$$

$$P(|X-\mu| \geq 3\sigma) = 0,0027$$

N = počet inspekcí před signálem

$$p = 0,0027 \qquad E(N) = \frac{1}{p} = \frac{1}{0,0027} = 370$$

Počet inspekcí před prvním falešným signálem (ARL = Average Run Length)