Normaaljaotus

Normaalne juhuslik suurus leiab teiste juhuslike suurustega võrreldes kõige enam kasutamist, sest suur osa loodust ja ühiskonda kujutavatest juhuslikest suurustest allub normaaljaotusele. Normaaljaotusega on näiteks:

- inimeste jaotus intelligentsi koefitsiendi IQ alusel;
- inimeste pikkus ja kaal;
- ühe kuu keskmised temperatuurid ühes kohas pikema aja (sajandi) jooksul;
- keskmine teraviljasaak;
- juhuslikud mõõtmisvead;
- detailide kulumine teatud mehhanismides.

Normaaljaotus tekib järgmiste tingimuste korral:

- tunnuse väärtustel on olemas mingi fikseeritud keskmine tase;
- tunnuse väärtus kujuneb paljude üksteisest sõltumatute nõrgalt mõjuvate faktorite toimel;
- tunnuse väärtuste suurenemine üle keskmise taseme ja vähenemine alla keskmist taset on võrdvõimalikud.

Esimesena võttis normaalse juhusliku suuruse mõiste kasutusele saksa matemaatik Carl Friedrich Gauss (1777 - 1855). Ta uuris astronoomiliste mõõtmiste mõõtmisvigu ja leidis, et need alluvad kindlale seaduspärasusele, mis nüüdseks kannab normaaljaotuse nime. Saksakeelses kirjanduses kutsutakse normaaljaotust Gaussi auks Gauss'i jaotuseks. Samal ajal prantsuskeelses kirjanduses normaaljaotus kannab Laplace'i jaotuse nime. Asi on nimelt selles, et matemaatiline aparatuur (valemid), mida Gauss kasutas, oli varem välja töötatud Laplace'i poolt kuigi hoopis teisel eesmärgil. Räägitakse, et selleks, et mitte häirida sakslasi ja prantslasi tegid inglise bioloogid ettepaneku nimetada antud juhuslikku suurust neutraalselt normaalseks juhuslikuks suuruseks, kuna ta väga sageli esineb meid ümbritsevas elus (looduses).

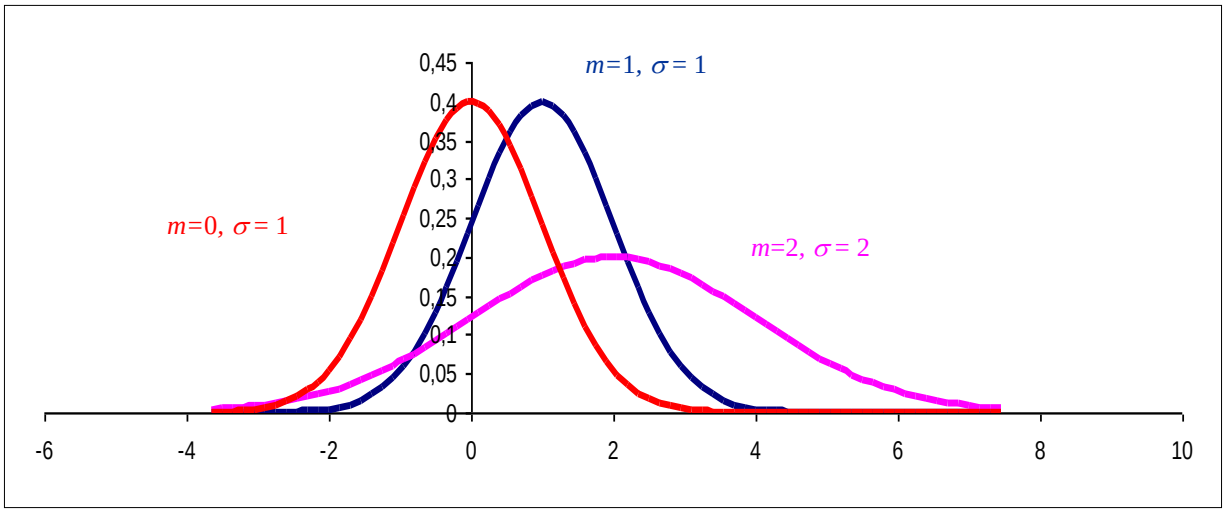
Kui pideva juhusliku suuruse tihedusfunktsiooniks on funktsioon

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

siis öeldakse, et see suurus **on normaaljaotusega** e. **Gaussi jaotusega**.

Asjaolu, et juhuslik suurus on normaaljaotusega parameetritega m ja σ, tähistatakse sümboolselt $\boldsymbol{X \sim N(m,\sigma)}$.

Üldise normaaljaotuse kõrval kasutatakse väga sageli ka **normeeritud normaaljaotust** e. standardset normaaljaotust, mis saadakse üldisest normaaljaotusest kui $m = 0$ ja $\sigma = 1$. Järgneval joonisel on esitatud normaaljaotuse tihedusfunktsiooni kuju mõningate m ja σ väärtuste korral.



Normaaljaotuse parameetrid

Normaaljaotuse keskväärtus: $EX = m$

Normaaljaotuse dispersioon: $DX = \sigma^2$

Normaaljaotuse standardhälve: $\sigma(X) = \sigma$

Järgnevalt veendume, et kehtib $EX = m$. Võrduse $DX = \sigma^2$ kontrollimine on analoogne.

Normaaljaotuse keskväärtus

Normaaljaotuse keskväärtuse avaldamiseks kasutame pideva juhusliku suuruse keskväärtuse valemit:

$$EX = \int_{-\infty}^{\infty} xp(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} x dx .$$

Teeme viimases integraalis muutujate vahetuse $t = \frac{x-m}{\sigma\sqrt{2}}$, siis

$$x - m = \sqrt{2}\sigma t \text{ ja } dx = \sqrt{2}\sigma\, dt .$$

Seega

$$EX = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} (\sqrt{2}\sigma t + m)\sqrt{2}\sigma\, dt = \frac{\sqrt{2}\sigma}{\sqrt{\pi}} \int_{-\infty}^{\infty} te^{-t^2} dt + \frac{m}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt .$$

Saadud summas on esimene integraal võrdne nulliga kuna integreeritav funktsioon on paaritu ja rajad on nullpunkti suhtes sümmeetrilised. Viimane integraal on aga nn. Poissoni integraal, mille kohta on kõrgemast matemaatikast teada valem $\int_{-\infty}^{\infty} e^{-t^2} dt = \sqrt{\pi}$.

Kokkuvõttes oleme saanud, et

$$EX = 0 + \frac{m}{\sqrt{\pi}}\sqrt{\pi} = m .$$

Normaaljaotusega juhusliku suuruse antud vahemikku sattumise tõenäosus

Teame, et tõenäosus, et juhusliku suuruse võimalikud väärtused satuvad vahemikku (α, β), avaldub tihedusfunktsiooni kaudu järgnevalt:

$$P(\alpha < X < \beta) = \int_{\alpha}^{\beta} p(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{(x-m)^2}{2\sigma^2}} dx .$$

Teeme viimases integraalis muutujate vahetuse $t = \frac{x-m}{\sigma}$, siis

$$x = m + \sigma t \text{ ja } dx = \sigma\, dt ,$$

lisaks arvutame vastavalt tehtud muutujate vahetusele ka uued integreerimisrajad

$$t_{\text{alumine}} = \frac{x-m}{\sigma} = \frac{\alpha-m}{\sigma}$$

$$t_{\text{ülemine}} = \frac{x-m}{\sigma} = \frac{\beta-m}{\sigma} .$$

Seega

$$P(\alpha < X < \beta) = \frac{1}{\sqrt{2\pi}} \int_{(\alpha-m)/\sigma}^{(\beta-m)/\sigma} e^{-\frac{t^2}{2}} dt = -\frac{1}{\sqrt{2\pi}} \int_{0}^{(\alpha-m)/\sigma} e^{-\frac{t^2}{2}} dt + \frac{1}{\sqrt{2\pi}} \int_{0}^{(\beta-m)/\sigma} e^{-\frac{t^2}{2}} dt .$$

Et integraal $\int e^{-\frac{t^2}{2}} dt$ ei avaldu elementaarfunktsioonides, siis tekkinud integraalide arvutamine taandatakse ühele tabuleeritud erifunktsioonile – Laplace'i funktsioonile.

Funktsiooni $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt$ nimetatakse **Laplace'i funktsiooniks** e. **tõenäosuse integraaliks**.

Selle funktsiooni kaudu:

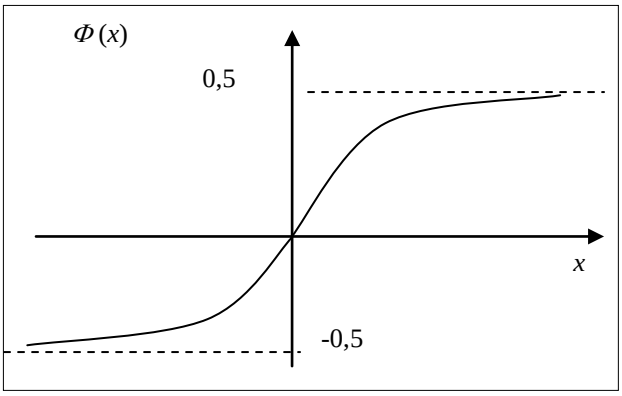
$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta-m}{\sigma}\right) - \Phi\left(\frac{\alpha-m}{\sigma}\right) .$$

(Laplace'i funktsiooni tabelid jagatakse praktikumides.)

Laplace'i funktsiooni omadusi

- $\Phi(0) = 0$ (sest kui $x = 0$, siis integraali rajad on võrdsed)
- $\Phi(\infty) = 0{,}5$ (kusjuures koondumine toimub küllalt kiiresti, juba $\Phi(3) = 0{,}49865$)
- $\Phi(-x) = -\Phi(x)$ (ehk Laplace'i funktsioon on paaritu funktsioon)

Laplace'i funktsiooni põhimõtteline kuju on esitatud järgneval joonisel.



Näide. Kui tõenäone on, et normaaljaotusega juhusliku suuruse $X \sim N$ (100, 5) väärtused langevad vahemikku (90, 105)?

Lahendus. Kasutame valemit $P(\alpha < X < \beta) = \Phi\left(\frac{\beta - m}{\sigma}\right) - \Phi\left(\frac{\alpha - m}{\sigma}\right)$ ja Laplace'i funktsiooni tabelit:

$$P(90 < X < 105) = \Phi\left(\frac{105-100}{5}\right) - \Phi\left(\frac{90-100}{5}\right) =$$
$$= \Phi(1) - \Phi(-2) = \Phi(1) + \Phi(2) = 0{,}34134 + 0{,}47725 = 0{,}81859.$$

Näide. Kui tõenäone on, et normaaljaotusega juhusliku suuruse $X \sim N$ (m, σ) tsentreeritud hälbe absoluutväärtus $|X - m|$ ei ületa standardhälbe k-kordset (k = 1, 2, 3)?

Lahendus.

$$P(|X - m| < k\sigma) = P(-k\sigma < X - m < k\sigma) = P(-k\sigma + m < X < k\sigma + m) =$$
$$= \Phi\left(\frac{-k\sigma + m - m}{\sigma}\right) - \Phi\left(\frac{k\sigma + m - m}{\sigma}\right) = \Phi(-k) - \Phi(-k) = 2\Phi(k).$$

Arvutame tulemused k = 1, 2, 3 korral:

$k = 1$: $P(-\sigma + m < X < \sigma + m) = 2\Phi(1) \approx 0{,}6826$

Seega umbes 2/3 normaaljaotusega juhusliku suuruse väärtustest ei hälbi keskväärtusest rohkem kui ühekordse standardhälbe võrra.

$k = 2$: $P(-2\sigma + m < X < 2\sigma + m) = 2\Phi(2) \approx 0{,}9544$

$k = 3$: $P(-3\sigma + m < X < 3\sigma + m) = 2\Phi(3) \approx 0{,}9974$

Seega praktiliselt võib kindel olla (tõenäosusega 0,9974), et väärtuste hälbed keskväärtusest ei ületa kolmekordset standardhälvet. Viimast tulemust tuntakse nn. **"3σ"-reegli** nime all.

Normaaljaotuse jaotusfunktsioon

Normaaljaotuse jaotusfunktsioon avaldub Laplace'i funktsiooni kaudu järgmiselt:

$$F(x) = \int_{-\infty}^{x} p(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \ldots = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right)$$

Tõenäosusteooria piirteoreemid

Teoreeme, mis seovad massiliselt esinevate juhuslike nähtuste teoreetilisi ja eksperimentaalseid karakteristikuid, tuntakse **tõenäosusteooria piirteoreemidena**. Piirteoreeme on kahte liiki:

1. Suurte arvude seadus

Nende teoreemide põhijäreldus seisneb selles, et katsete arvu suurenedes juhuslikkuse mõju väheneb ning katsete arvu lähenedes lõpmatusele juhuslikkuse mõju kaob hoopiski. Seega küllalt suure katsete arvu korral juhuslike sündmuste karakteristikud muutuvad peaaegu mittejuhuslikeks. Nii näiteks stabiliseerub suure katsete arvu korral sündmuse sagedus, sama kehtib ka juhusliku suuruse keskväärtuse kohta.

Seoses suurte arvude seadusega vaatleme kahte piirteoreemi:

- Bernoulli teoreem
- Tšebõševi teoreem

2. Tsentraalne piirteoreem e. Ljapunovi teoreem

See teoreem annab tingimused juhuslike suuruste jada koondumiseks normaaljaotuseks.

Kõigepealt tutvume veel ühe mõistega, mida järgnevas kasutame. Öeldakse, et juhuslike suuruste jada X_1, X_2, ... koondub **tõenäosuse järgi** juhuslikuks suuruseks X, kui iga $\varepsilon > 0$ korral

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

Tõenäosuse järgi koondumine tähendab seda, et tõenäosus juhuslike suuruste X_n ja X oluliseks erinemiseks läheneb nullile ehk seda esineb väga harva.

Bernoulli suurte arvude seadus

Bernoulli teoreem on (ajalooliselt) esimene teoreemidest, mis nüüd kannavad suurte arvude seaduse nime. J. Bernoulli uuris põhjalikult sõltumatuid katseid ning nende uurimiste tulemusena tõestas ta 1713. a. teoreemi, mis nüüd kannab tema nime.

Korduvatel sõltumatutel katsetel koondub sündmuse suhteline sagedus katsete arvu n kasvamisel tõenäosuse järgi sündmuse tõenäosuseks:

$$\lim_{n\to\infty} P(\left|\frac{m}{n} - p\right| < \varepsilon) = 1,$$

kus m on sündmus toimumiskordade arv katseseerias ja ε kuitahes väike positiivne arv.

Teisiti öeldes: sündmuse suhteline sagedus koondub tõenäosuse järgi selle sündmuse tõenäosuseks üksikkatsel.

Bernoulli suurte arvude seadus kinnitab veelkord, et suhteline sagedus küllalt pikas katseseerias ehk nn. statistiline tõenäosus sobib sündmuse tõenäosuse hinnanguks.

Bernoulli suurte arvude seaduse kehtivust on püütud korduvalt katseliselt kontrollida. Buffon viskas 1777. a. münti 4040 korda. Kull saadi 2048 korral, vastav suhteline sagedus oleks seega $w = \frac{2048}{4040} \approx 0{,}507$ (kulli saamise tõenäosus üksikviskel on teatavasti 0,5). Quetelet registreeris 1846. a. 20 musta ja 20 valget kuuli sisaldavast urnist kuulide võtmise tulemusi. Kuuli võeti 4096 korda, sealjuures valge kuul saadi 2066 korral, vastav suhteline sagedus oleks seega $w = \frac{2066}{4096} \approx 0{,}504$.

Tšebõševi suurte arvude seadus

Suurte arvude seaduse põhiteoreemiks on Tšebõševi teoreem. Selle teoreemi tõestas 1867. aastal vene matemaatik P. Tšebõšev. Oluline on, et Tšebõševi teoreem kehtib küllaltki üldistel tingimustel, mis praktiliselt alati on täidetud.

Küllalt suure arvu võrdsete keskväärtuste EX ja võrdsete dispersioonidega sõltumatute juhuslike suuruste X_1, X_2, ... , X_n puhul läheneb nende suuruste aritmeetiline keskmine tõenäosuse järgi keskväärtusele: ehk iga $\varepsilon > 0$ korral

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - EX\right| < \varepsilon\right) = 1.$$

Selles teoreemis ei väideta absoluutse kindlusega, et $\frac{1}{n}\sum_{i=1}^{n} X_i \to EX$ vaid $\frac{1}{n}\sum_{i=1}^{n} X_i \to EX$ tõenäosuse järgi.

Tšebõševi teoreemist järeldub:

- Ka suure hulga juhuslike suuruste aritmeetiline keskmine võib oluliselt hälbida keskväärtusest, kuid seda juhtub väga harva, s.t. suurte hälvete tõenäosus läheneb nullile.
- Tšebõševi suurte arvude seadus on kinnituseks sellele, et katseliselt määratava suuruse tõelise väärtuse parimaks lähendiks on katsetulemuste aritmeetiline keskmine ja viimane on seda usaldusväärsem, mida pikema katseseeria põhjal see on leitud.

Ljapunovi teoreem(tsentraalne piirteoreem)

Ljapunovi teoreem annab seletuse, miks rakendustes kasutatakse kõige sagedamini normaaljaotusega või sellele lähedase jaotusega suurusi. Nii näiteks allub juhuslik mõõtmisviga normaaljaotusele. Mõõtmisviga moodustub paljude juhuslike põhjuste koosmõjul, kus iga üksikpõhjus tingib summaarse vea seisukohalt tühise vea. Vea komponentide põhjustajate hulk on tavaliselt suur.

Olgu juhuslik suurus X küllalt suure hulga n sõltumatu juhusliku suuruse X_i (i = 1, ... , n) summa: $X = \sum_{i=1}^{n} X_i$. Osasuuruste X_i jaotusseadused võivad olla väga erinevad, kuid iga osasuuruse osatähtsus summaarse juhusliku suuruse X moodustamisel olgu väike. Siis läheneb summaarse juhusliku suuruse X jaotus normaaljaotusele liidetavate arvu n kasvamisel.

Moivre – Laplace'i integraalne piirteoreem

Eespool rääkisime, et kui sündmuse esinemise tõenäosus on väike, siis saab binoomjaotust lähendada Poissoni jaotusega. Osutub, et teatud tingimustel saab binoomjaotust lähendada ka normaaljaotusega. Praktikas arvestatakse, et binoomjaotuse lähendamisel normaaljaotusega saadakse piisav täpsus juhul kui on täidetud tingimused $np > 5$ ja $nq > 5$.

Kui m on sündmuse A toimumiste arv n sõltumatu katse korral ja sündmuse A tõenäosus p on jääv, siis kehtib valem

$$\lim_{n\to\infty} P(\alpha \le \frac{m-np}{\sqrt{npq}} < \beta) = \Phi(\beta) - \Phi(\alpha),$$

kus $q = 1 - p$ ja $\Phi(x)$ on Laplace'i funktsioon.

Toodud valem on teisendatav praktikas sagedasemat kasutamist leidvale kujule:

$$\lim_{n\to\infty} P(k_1 \le m < k_2) = \Phi(\frac{k_2 - np}{\sqrt{npq}}) - \Phi(\frac{k_1 - np}{\sqrt{npq}}).$$

See valem võimaldab küllalt suure katsete arvu korral kasutada diskreetse binoomjaotuse asemel ligikaudse lähendina pidevat normaaljaotust parameetritega $EX = np$ ja $\sigma(X) = \sqrt{npq}$.

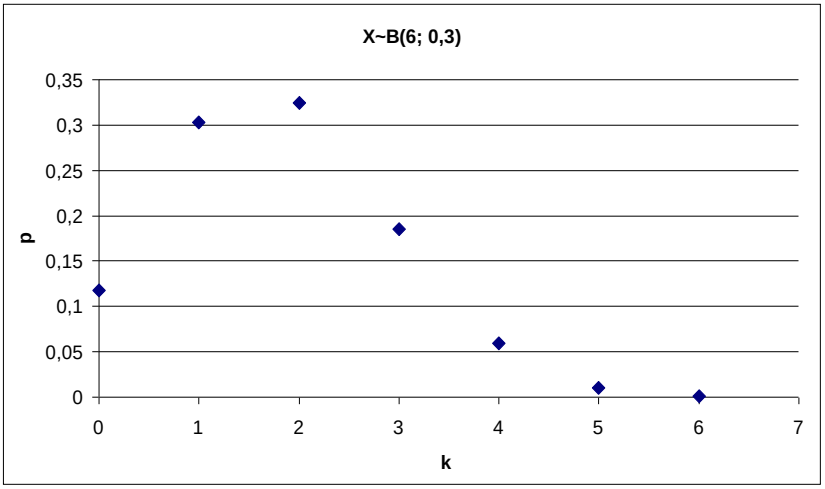
Moivre – Laplace'i lokaalne piirteoreem

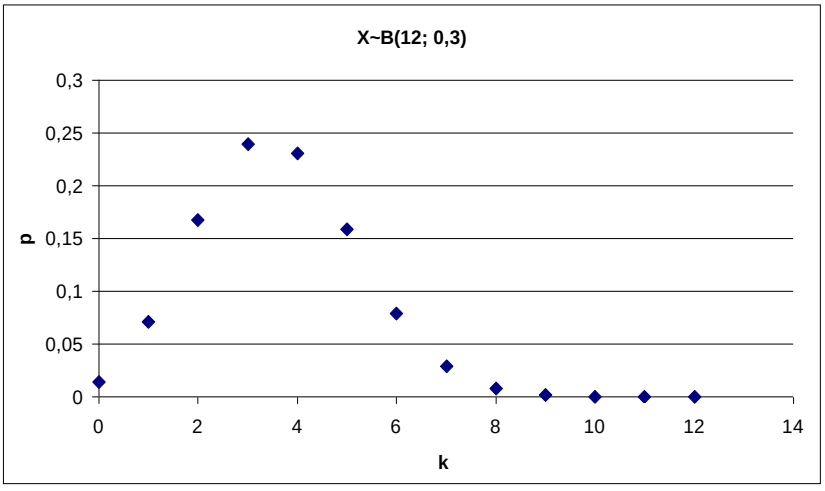
Kuna katsete arvu n tõkestamatul kasvamisel koondub binoomjaotus normaaljaotuseks parameetritega $EX = np$ ja $\sigma(X) = \sqrt{npq}$, siis küllalt suure katsete arvu korral võime sündmuse sageduse m tõenäosuse arvutamiseks kasutada Bernoulli valemi asemel ligikaudset valemit

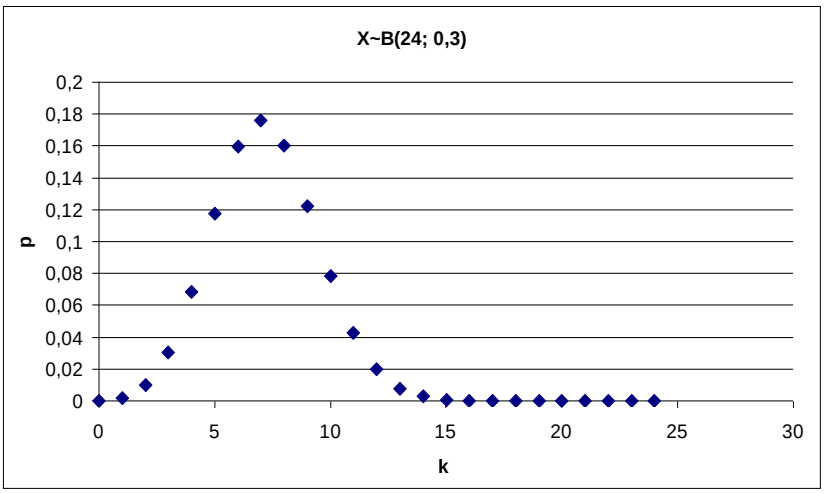
$$P_{m,n} \approx \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \text{kus} \quad x = \frac{m - np}{\sqrt{npq}}.$$

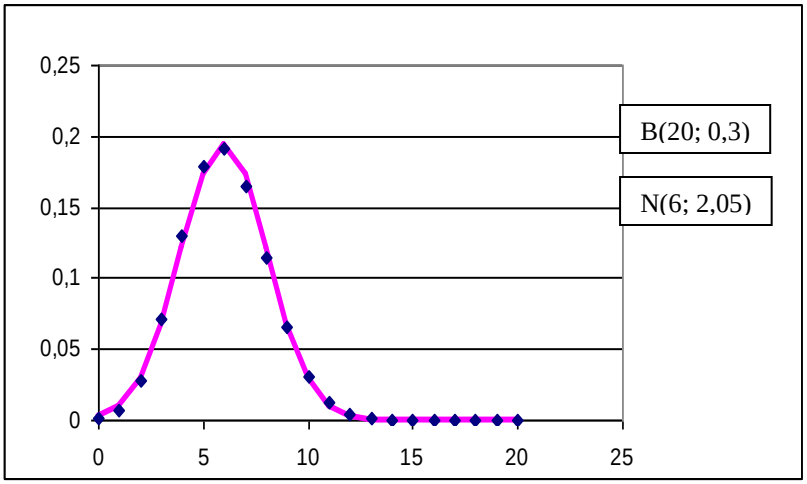
Binoomjaotuse koondumine normaaljaotuseks

Järgnevatel joonistel on kujutatud suuruse $P_{m,n}$ jaotus vastavalt Bernoulli valemile. On näha, et katsete arvu n kasvades paigutuvad vastavad punktid järjest korrapärasemalt, lähenedes normaaljaotuse tihedusfunktsiooni kujutavale kõverale.









Näide. Münti visatakse 50 korda. Kui tõenäone on kulli esinemine 20 korda? Kulli esinemissageduse asumine 20 ja 28 vahel?

Lahendus. See on tüüpiline binoomjaotuse ülesanne, mida on kasulik lahendada tuginedes Moivre – Laplace'i teoreemidele. Kulli saamise tõenäosus üksikkatsel on $p = 0{,}5$ ja ka $q = 0{,}5$. Et $np = 50 \cdot 0{,}5 = 25 > 5$ ja $nq = 50 \cdot 0{,}5 = 25 > 5$, siis võib binoomjaotuse asemel ligikaudse lähendina kasutada normaaljaotust parameetritega $EX = np = 25$ ja $\sigma(X) = \sqrt{npq} \approx 3{,}536$. Kasutades Moivre – Laplace'i lokaalset piirteoreemi, saame

$$x = \frac{m - np}{\sqrt{npq}} = \frac{20 - 25}{3{,}536} \approx -1{,}42.$$

$$P_{50,20} \approx \frac{1}{\sqrt{npq}} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = \frac{1}{3{,}536} \cdot \frac{1}{\sqrt{2\pi}} e^{-(-1{,}42)^2/2} \approx 0{,}041.$$

Vastavalt Moivre – Laplace'i integraalsele piirteoreemile leiame tõenäosuse, et kulli esinemissagedus on 20 ja 28 vahel

$$P(20 < k < 28) \approx \Phi\left(\frac{28 - 25}{3{,}536}\right) - \Phi\left(\frac{20 - 25}{3{,}536}\right) \approx 0{,}725.$$