

Metody matematyczne astronomii 2

Wykład 3

Mariusz Tarnopolski

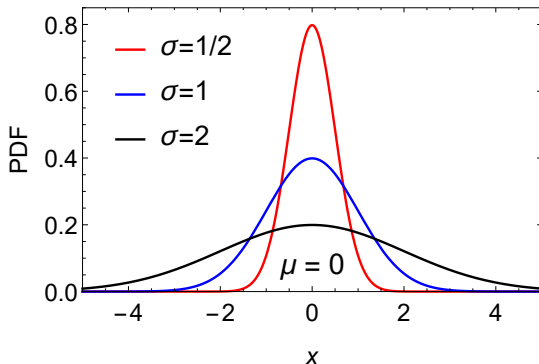
Instytut Astronomii UMK

MMA 2 ©2023



Rozkład Gaussa (normalny), $\mathcal{N}(\mu, \sigma)$

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$



Rozkład Gaussa

- $f(x) = f(-x)$ oraz $f(x) > 0$ dla $-\infty < x < +\infty$.
- Można *ustandaryzować*: $x \rightarrow \frac{x-\mu}{\sigma}$; mamy wtedy¹ $\mathcal{N}(0, 1)$, którego PDF, $f(x; 0, 1)$, ozn. jako $\varphi(x)$ — to jego wartości są stabilizowane.
- $\varphi(x)$ unormowany, zatem $f(x; \mu, \sigma) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$ — też unormowany:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$$

- Dystrybuanta

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right]$$

gdzie $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ to tzw. funkcja błędu.

¹Por. wykład 2, slajd 6.

Rozkład Gaussa

- $f(x) = f(-x)$ oraz $f(x) > 0$ dla $-\infty < x < +\infty$.
- Można *ustandaryzować*: $x \rightarrow \frac{x-\mu}{\sigma}$; mamy wtedy¹ $\mathcal{N}(0, 1)$, którego PDF, $f(x; 0, 1)$, ozn. jako $\varphi(x)$ — to jego wartości są tablicowane.
- $\varphi(x)$ unormowany, zatem $f(x; \mu, \sigma) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$ — też unormowany:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$$

- Dystrybuanta

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right]$$

gdzie $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ to tzw. funkcja błędu.

¹Por. wykład 2, slajd 6.

Rozkład Gaussa

- $f(x) = f(-x)$ oraz $f(x) > 0$ dla $-\infty < x < +\infty$.
- Można *ustandaryzować*: $x \rightarrow \frac{x-\mu}{\sigma}$; mamy wtedy¹ $\mathcal{N}(0, 1)$, którego PDF, $f(x; 0, 1)$, ozn. jako $\varphi(x)$ — to jego wartości są tablicowane.
- $\varphi(x)$ unormowany, zatem $f(x; \mu, \sigma) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$ — też unormowany:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$$

- Dystrybuanta

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right]$$

gdzie $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ to tzw. funkcja błędu.

¹Por. wykład 2, slajd 6.

Rozkład Gaussa

- $f(x) = f(-x)$ oraz $f(x) > 0$ dla $-\infty < x < +\infty$.
- Można *ustandaryzować*: $x \rightarrow \frac{x-\mu}{\sigma}$; mamy wtedy¹ $\mathcal{N}(0, 1)$, którego PDF, $f(x; 0, 1)$, ozn. jako $\varphi(x)$ — to jego wartości są stabilizowane.
- $\varphi(x)$ unormowany, zatem $f(x; \mu, \sigma) = \frac{1}{\sigma} \varphi\left(\frac{x-\mu}{\sigma}\right)$ — też unormowany:

$$\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1$$

- Dystrybuanta

$$\Phi(x) = \int_{-\infty}^x \varphi(t) dt = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) \right]$$

gdzie $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ to tzw. funkcja błędu.

¹Por. wykład 2, slajd 6.

Przykład: niech wzrost mężczyzny² ma rozkład $\mathcal{N}(180, 8)$. Wtedy prawdopodobieństwo

$$P(X > 190) = P\left[\left(Y = \frac{X - 180}{8}\right) > \frac{190 - 180}{8} = 1.25\right]$$

Z tablic dystrybuanty standardowego rozkładu normalnego odczytujemy (historycznie) lub liczymy numerycznie na komputerze (obecnie), że

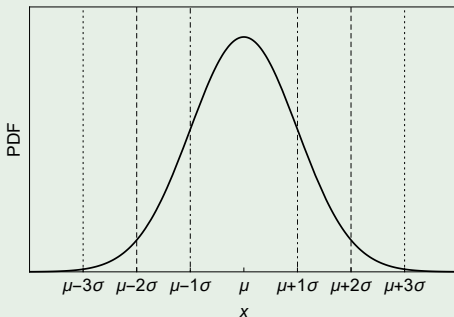
$$P(Y > 1.25) = 1 - \Phi(1.25) = 0.105$$

czyli jest 10.5% szans by znaleźć mężczyznę o wzroście powyżej 190 cm.

²Wzrost kobiet i mężczyzn osobno jest dobrze opisany rozkładami Gaussa o różnych parametrach. Wzrost człowieka będzie zatem tzw. rozkładem mieszanym (*mixture distribution*), tj. unormowaną sumą tych dwóch rozkładów normalnych (zob. slajd 16).

Zadanie 5

Przecałkować numerycznie PDF (równoważnie: skorzystać z odpowiadających wartości dystrybucyj; wskazówka: `scipy.stats.norm`) żeby wyznaczyć „poziomy sigma”, tj. jakie jest prawdopodobieństwo $P(|X - \mu| \leq k\sigma)$ znalezienia zmiennej normalnej (ustandaryzowanej) w przedziałach $[\mu - k\sigma, \mu + k\sigma]$ dla $k \in \{1, 2, 3, 4, 5\}$. Jakim poziomom sigma odpowiadają prawdopodobieństwa 90%, 95% i 99%?



Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem*, CLT)

Średnia n niezależnych zmiennych losowych X_i (dla uproszczenia niech najpierw X_i będą iid, tj. $\mu_i \equiv \mu < \infty$ oraz $\sigma_i^2 \equiv \sigma^2 < \infty$)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

zbiega do rozkładu normalnego gdy n jest dostatecznie duże.

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem, CLT*)

Średnia n niezależnych zmiennych losowych X_i (dla uproszczenia niech najpierw X_i będą iid, tj. $\mu_i \equiv \mu < \infty$ oraz $\sigma_i^2 \equiv \sigma^2 < \infty$)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

zbiega do rozkładu normalnego gdy n jest dostatecznie duże.
Dokładniej: zmienna ustandaryzowana

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{gdy} \quad n \rightarrow \infty$$

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem, CLT*)

Średnia n niezależnych zmiennych losowych X_i (dla uproszczenia niech najpierw X_i będą iid, tj. $\mu_i \equiv \mu < \infty$ oraz $\sigma_i^2 \equiv \sigma^2 < \infty$)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

zbiega do rozkładu normalnego gdy n jest dostatecznie duże.
Dokładniej: zmienna ustandaryzowana

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{gdy} \quad n \rightarrow \infty$$

czyli

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{gdy} \quad n \rightarrow \infty$$

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem, CLT*)

Średnia n niezależnych zmiennych losowych X_i (dla uproszczenia niech najpierw X_i będą iid, tj. $\mu_i \equiv \mu < \infty$ oraz $\sigma_i^2 \equiv \sigma^2 < \infty$)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

zbiega do rozkładu normalnego gdy n jest dostatecznie duże.
Dokładniej: zmienna ustandaryzowana

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \quad \text{gdy} \quad n \rightarrow \infty$$

czyli

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{gdy} \quad n \rightarrow \infty$$

zatem $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$.

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem*, CLT)

Jest to niezwykle użyteczne np. przy wyznaczaniu niepewności (przedziałów ufności — zob. wykł. 4) dla średniej wartości pewnej wielkości na podstawie próbki obserwacji pewnej populacji — bez względu na to jaki rozkład ma populacja (jednostajny, wielomodalny, jakikolwiek inny) średnia będzie miała rozkład Gaussa, zaś jej niepewność to $\Delta \bar{X} = \frac{\sigma}{\sqrt{n}}$.

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem, CLT*)

Jest to niezwykle użyteczne np. przy wyznaczaniu niepewności (przedziałów ufności — zob. wykł. 4) dla średniej wartości pewnej wielkości na podstawie próbki obserwacji pewnej populacji — bez względu na to jaki rozkład ma populacja (jednostajny, wielomodalny, jakiegokolwiek inny) średnia będzie miała rozkład Gaussa, zaś jej niepewność to $\Delta \bar{X} = \frac{\sigma}{\sqrt{n}}$.

Z prawa propagacji błędów:

$$\Delta \bar{X} = \left[\sum_{i=1}^n \left(\frac{\partial \bar{X}}{\partial X_i} \right)^2 \Delta X_i^2 \right]^{1/2} = \left[\sum_{i=1}^n \left(\frac{1}{n} \cdot 1 \right)^2 \sigma^2 \right]^{1/2} = \left(n \frac{\sigma^2}{n^2} \right)^{1/2} = \frac{\sigma}{\sqrt{n}}$$

Centralne twierdzenie graniczne (Lindeberga-Lévy'ego)

(*Central limit theorem, CLT*)

Jest to niezwykle użyteczne np. przy wyznaczaniu niepewności (przedziałów ufności — zob. wykł. 4) dla średniej wartości pewnej wielkości na podstawie próbki obserwacji pewnej populacji — bez względu na to jaki rozkład ma populacja (jednostajny, wielomodalny, jakiegokolwiek inny) średnia będzie miała rozkład Gaussa, zaś jej niepewność to $\Delta \bar{X} = \frac{\sigma}{\sqrt{n}}$.

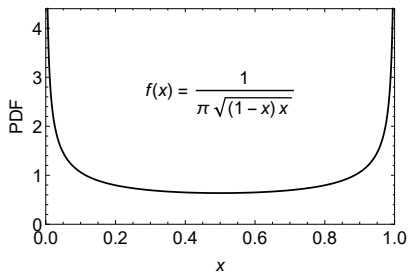
Z prawa propagacji błędów:

$$\Delta \bar{X} = \left[\sum_{i=1}^n \left(\frac{\partial \bar{X}}{\partial X_i} \right)^2 \Delta X_i^2 \right]^{1/2} = \left[\sum_{i=1}^n \left(\frac{1}{n} \cdot 1 \right)^2 \sigma^2 \right]^{1/2} = \left(n \frac{\sigma^2}{n^2} \right)^{1/2} = \frac{\sigma}{\sqrt{n}}$$

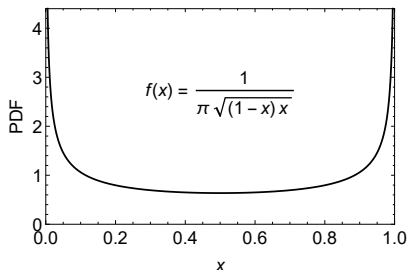
Z własności wariancji:

$$V(\bar{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

CLT — przykład



CLT — przykład



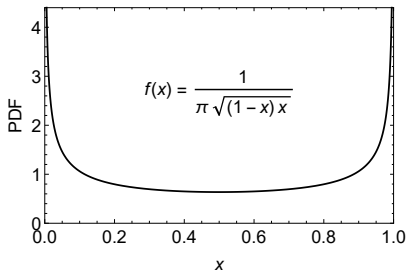
Losujemy 10 wartości:

$\{0.12, 0.95, 0.77, 0.97, 0.03, 0.32, 0.6, 0.87, 0.84, 0.06\}$.

Ich średnia to 0.553.

Powtarzamy 1000 razy, otrzymując rozkład średnich:

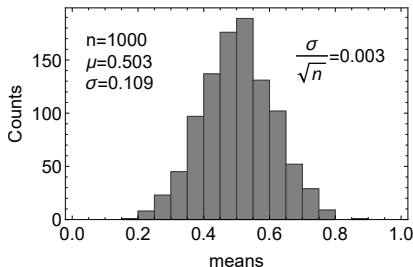
CLT — przykład



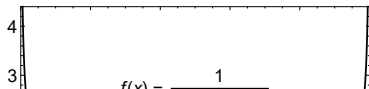
Losujemy 10 wartości:
 $\{0.12, 0.95, 0.77, 0.97, 0.03,$
 $0.32, 0.6, 0.87, 0.84, 0.06\}.$

Ich średnia to 0.553.

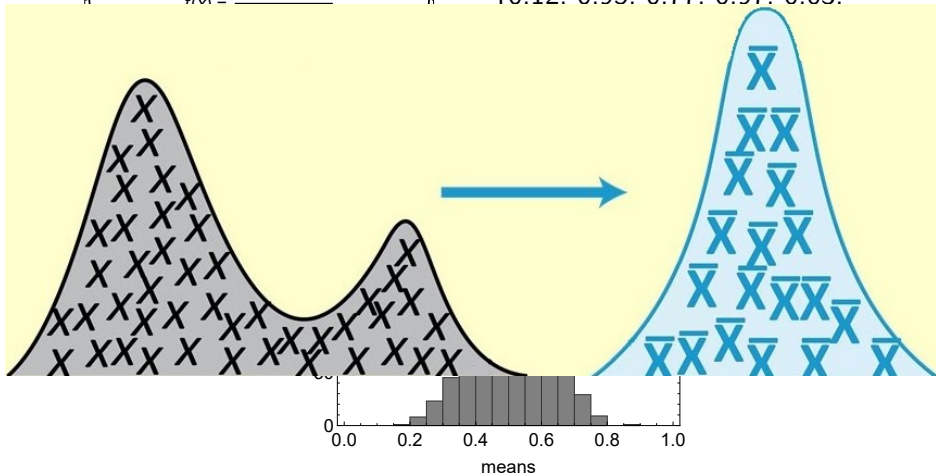
Powtarzamy 1000 razy, otrzymując
rozkład średnich:



CLT — przykład



Losujemy 10 wartości:
 $\{0.12, 0.95, 0.77, 0.97, 0.03, \dots\}$



Twierdzenie Lindeberga-Lévy'ego ma swoje ograniczenia — w szczególności, mówi *explicite* o **iid zmiennych losowych**. W wielu praktycznych sytuacjach mamy do czynienia z sumą niezależnych zmiennych losowych o **różnych rozkładach**. Wtedy zachodzi:

Twierdzenie Lindeberga-Fellera. Niech będzie ciąg $X_1, X_2, X_3, \dots$ niezależnych zmiennych losowych takich, że

$$S_n = \sum_{i=1}^n X_i, \quad \sigma_n^2 = \sum_{i=1}^n V(X_i)$$

Ustandaryzujemy zmienną S_n :

$$Y_n = \frac{S_n - \mathbb{E}(S_n)}{\sigma_n}$$

CLT Lindeberga-Fellera

Jeśli spełniony jest tzw. warunek Lindeberga:

$$\forall \varepsilon > 0 : \quad \frac{1}{\sigma_n^2} \sum_{i=1}^n \mathbb{E} \left[(X_i - \mathbb{E}(X_i))^2 \mathbb{1}_{|X_i - \mathbb{E}(X_i)| > \varepsilon \sigma_n} \right] \xrightarrow{n \rightarrow \infty} 0$$

to Y_n ma rozkład $\mathcal{N}(0, 1)$.

Ten warunek sprawdza, czy żadna pojedyncza zmienna losowa nie fluktuuje na tyle silnie by dominować w sumie będącej przedmiotem zainteresowania CLT (tj. czy rzadkie duże odchylenia nie są **zbyt** duże w porównaniu z typową skalą fluktuacji).

Funkcja $\mathbb{1}_\omega$ to tzw. funkcja wskaźnikowa:

$$\mathbb{1}_\omega = \begin{cases} 1 & \text{jeśli warunek } \omega \text{ jest spełniony} \\ 0 & \text{jeśli warunek } \omega \text{ nie jest spełniony} \end{cases}$$

Np. $\mathbb{1}_{1 < 2} = 1$, ale $\mathbb{1}_{2 > 3} = 0$.

W astronomii problemem jest brak powtarzalności: jeśli zaobserwujemy 10 supernowych, to nie da się „na zawołanie” zaobserwować kolejnych 9990.

W astronomii problemem jest brak powtarzalności: jeśli zaobserwujemy 10 supernowych, to nie da się „na zawołanie” zaobserwować kolejnych 9990. Rozwiązanie — bootstrap (jedna z metod Monte Carlo):

- 1 mamy próbkę N obserwacji. Dokonujemy **losowania z powtórzeniem** N wartości spośród N obserwacji — niektóre mogą się powtarzać, inne nie zostaną wylosowane w danej próbie. Każde takie losowanie tworzy nową próbkę. Powtarzamy 1000 razy i postępujemy jak na slajdzie 8.

W astronomii problemem jest brak powtarzalności: jeśli zaobserwujemy 10 supernowych, to nie da się „na zawołanie” zaobserwować kolejnych 9990. Rozwiązanie — bootstrap (jedna z metod Monte Carlo):

- 1 mamy próbkę N obserwacji. Dokonujemy **losowania z powtórzeniem** N wartości spośród N obserwacji — niektóre mogą się powtarzać, inne nie zostaną wylosowane w danej próbie. Każde takie losowanie tworzy nową próbkę. Powtarzamy 1000 razy i postępujemy jak na slajdzie 8.
- 2 Jeśli mamy N obserwacji postaci $x \pm \Delta x$, gdzie Δx jest błędem standardowym pomiaru x (tj. 68% przedział ufności dla danego pomiaru x to $[x - \Delta x, x + \Delta x]$) — losujemy dla każdego pomiaru nową realizację z rozkładu $\mathcal{N}(x, \Delta x)$, tworząc nową próbkę N obserwacji. Dalej postępujemy jak wyżej.

Zadanie 6

Moduły odległości do galaktyki M87, wyznaczone różnymi metodami przez różnych autorów, wynoszą [de Grijs & Bono, ApJS 246:3 (2020)]:

$\mu_0 = \{31.01, 31.44, 31.01, 31.03, 30.8, 30.86, 31.35, 30.9, 30.87, 31.06, 31.06, 31.18, 31.08, 30.86, 31.17, 31.14, 31.02, 31.18, 31.14, 31.02, 31.19, 30.89, 30.89, 31.09\}$ mag

oraz odpowiadające im niepewności:

$\Delta\mu_0 = \{0.08, 0.21, 0.03, 0.06, 0.04, 0.17, 0.13, 0.12, 0.1, 0.05, 0.14, 0.04, 0.06, 0.09, 0.1, 0.03, 0.16, 0.12, 0.03, 0.16, 0.06, 0.1, 0.05, 0.14\}$ mag

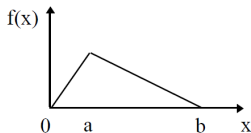
Obliczyć średnią $\bar{\mu}_0$ oraz jej niepewność. Wykonać bootstrap obiema metodami (tj. wykorzystując (i) wyłącznie wartości μ_0 oraz (ii) biorąc pod uwagę niepewności $\Delta\mu_0$) oraz porównać wyniki.

Przeliczyć $\bar{\mu}_0 \pm \Delta\bar{\mu}_0$ na odległości fizyczne D wg wzoru $\log_{10} D = \frac{\mu_0}{5} + 1$ (skorzystać z prawa propagacji błędów). Wyniki podać w Mpc.

Różnorodne rozkłady

- Rozkład trójkątny:

$$f(x) = \begin{cases} \frac{2x}{ab} & \text{dla } 0 \leq x \leq a, \\ \frac{2x}{ab-b^2} + \frac{2}{b-a} & \text{dla } a \leq x \leq b, \\ 0 & \text{dla } x < 0 \text{ i } x > b. \end{cases}$$



- Rozkład t Studenta ($\nu > 0$, ale zazwyczaj $\nu \in \mathbb{N}$; dla dużych ν staje się normalny; dla $\nu = 1$ daje rozkład Cauchy'ego):

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

- Rozkład Poissona (dyskretny) — statystyka zliczeń fotonów:

$$p(k) = \frac{\lambda^k}{k!} \exp(-\lambda) \text{ dla } k \in \mathbb{N}, \lambda \in \mathbb{R}_+. \text{ Średnia i wariancja są równe } \lambda.$$

Dla dużych λ staje się normalny.

Zadanie 7

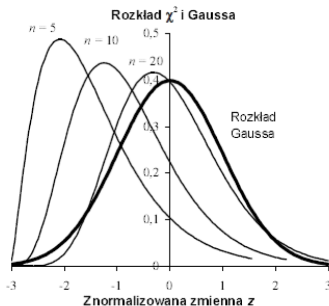
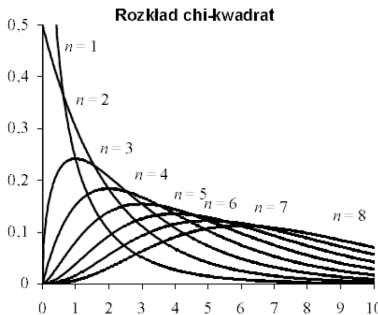
- 1 Detektor zebrał 1400 fotonów z obszaru nieba z kierunku na potencjalne źródło, oraz 1100 fotonów z obserwacji tła tego samego obszaru (te same czasy ekspozycji). Jaka jest istotność detekcji źródła (tj. stosunek sygnału do szumu, *signal-to-noise ratio*, S/N)?
- 2 Kolejne, niezależne obserwacje tła z tego samego obszaru nieba zebrały 1150 fotonów. Czy istotność sygnału wobec dwóch niezależnych poziomów tła wzrosła czy zmalała?

Różnorodne rozkłady

- Rozkład chi-kwadrat o n stopniach swobody — rozkład sumy kwadratów n zmiennych losowych normalnych $X_i \sim \mathcal{N}(0, 1)$, tj.

$\sum_{i=1}^n X_i^2$ (gdy $n \rightarrow \infty$ to rozkład χ^2 staje się normalny):

$$f(x) = \begin{cases} \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2} & \text{dla } x > 0, \\ 0 & \text{dla } x \leq 0. \end{cases}$$



- Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

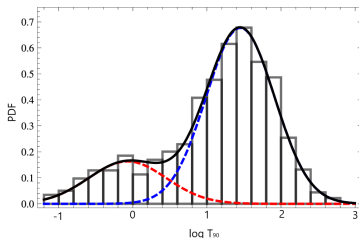
czyli rozkład zmiennej $\ln X$ jest normalny.

Różnorodne rozkłady

- Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

czyli rozkład zmiennej $\ln X$ jest normalny. Np. czasy trwania rozbłysków gamma:



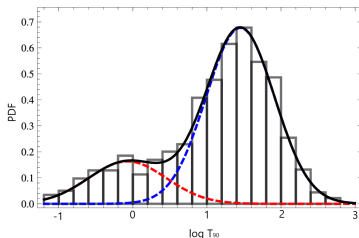
Różnorodne rozkłady

- Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

czyli rozkład zmiennej $\ln X$ jest normalny. Np. czasy trwania rozbłyków gamma:

Pojawia się przy modelowaniu iloczynu zmiennych losowych.



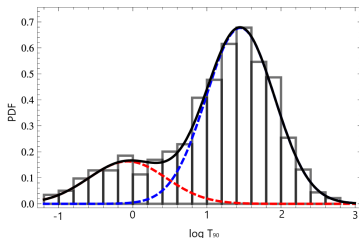
Różnorodne rozkłady

- Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

czyli rozkład zmiennej $\ln X$ jest normalny. Np. czasy trwania rozbłyków gamma:

Pojawia się przy modelowaniu iloczynu zmiennych losowych.



Zadanie 8

Wyjaśnić (poprzez obliczenia) pojawienie się multiplikatywnego czynnika $\frac{1}{x}$ w rozkładzie log-normalnym.

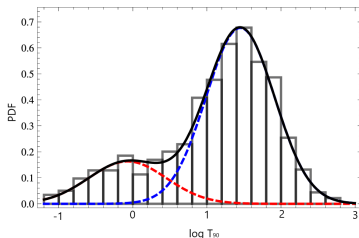
Różnorodne rozkłady

- Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

czyli rozkład zmiennej $\ln X$ jest normalny. Np. czasy trwania rozbłyków gamma:

Pojawia się przy modelowaniu iloczynu zmiennych losowych.



Zadanie 8

Wyjaśnić (poprzez obliczenia) pojawienie się multiplikatywnego czynnika $\frac{1}{x}$ w rozkładzie log-normalnym.

- Rozkład mieszany k rozkładów $f_i(x)$, tj. $f(x) = \sum_{i=1}^k p_i f_i(x)$, gdzie $\sum_{i=1}^k p_i = 1$ oraz $p_i > 0$.

Różnorodne rozkłady

Rozkład log-normalny:

$$f(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log x)^2}{2\sigma^2}\right)$$

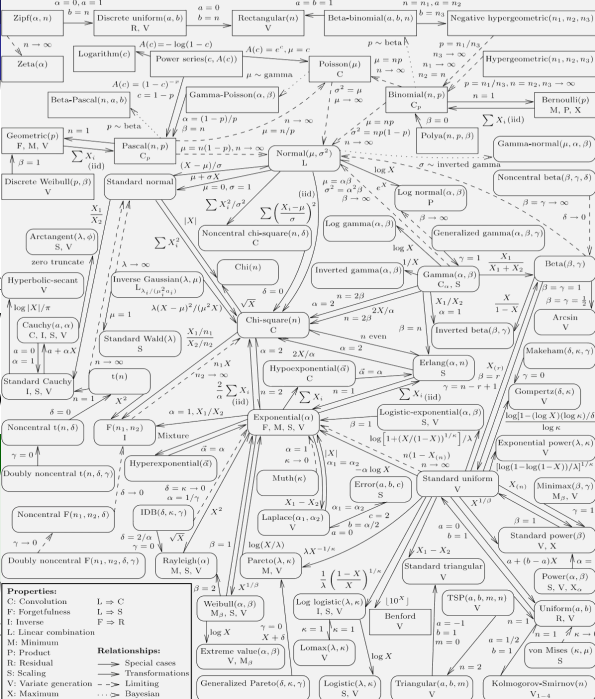
czyli rozkład zmiennej $\ln X$ normalny. Np. czasy trw rozbłysków gamma:

Pojawia się przy modelowaniu

Zadanie 8

Wyjaśnić (poprzez obliczenia w rozkładzie log-normalnym).

- Rozkład mieszany k rozkładów gdzie $\sum_{i=1}^k p_i = 1$ oraz



Funkcje zmiennych losowych

Rozważmy zmienną losową ciągłą X o rozkładzie prawdopodobieństwa $f_X(x)$ oraz zmienną losową Y związaną z X przekształceniem funkcyjnym $Y = g(X)$. Jeśli g jest monotoniczna³ to istnieje funkcja odwrotna $h = g^{-1}$, zatem

$$Y = g(X), \quad y = g(x), \quad X = h(Y), \quad x = h(y)$$

³Jeśli nie jest to można osobno rozważać przedziały, w których jest monotoniczna.

Funkcje zmiennych losowych

Rozważmy zmienną losową ciągłą X o rozkładzie prawdopodobieństwa $f_X(x)$ oraz zmienną losową Y związaną z X przekształceniem funkcyjnym $Y = g(X)$. Jeśli g jest monotoniczna³ to istnieje funkcja odwrotna $h = g^{-1}$, zatem

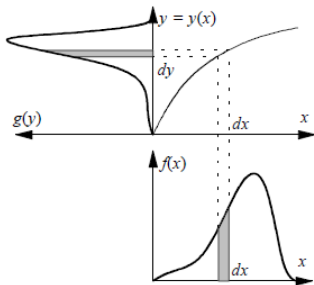
$$Y = g(X), \quad y = g(x), \quad X = h(Y), \quad x = h(y)$$

Z „zachowania” prawdopodobieństw (por. też zmienne gaussowskie ze slajdu 3) mamy

$$|f_Y(y)dy| = |f_X(x)dx|$$

skąd ostatecznie

$$f_Y(y) = f_X[h(y)]|h'(y)|$$



³ Jeśli nie jest to można osobno rozważać przedziały, w których jest monotoniczna.

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Para (X, Y) zmiennych losowych X i Y tworzy dwuwymiarową zmienną losową.

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Para (X, Y) zmiennych losowych X i Y tworzy dwuwymiarową zmienną losową.

Dystrybuanta zmiennej (X, Y) to funkcja $F(x, y)$, która każdej parze $(x, y) \in \mathbb{R}^2$ przypisuje prawdopodobieństwo zdarzenia takiego, że $X < x \wedge Y < y$:

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Para (X, Y) zmiennych losowych X i Y tworzy dwuwymiarową zmienną losową.

Dystrybuanta zmiennej (X, Y) to funkcja $F(x, y)$, która każdej parze $(x, y) \in \mathbb{R}^2$ przypisuje prawdopodobieństwo zdarzenia takiego, że $X < x \wedge Y < y$:

$$F(x, y) = P(X < x, Y < y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Para (X, Y) zmiennych losowych X i Y tworzy dwuwymiarową zmienną losową.

Dystrybuenta zmiennej (X, Y) to funkcja $F(x, y)$, która każdej parze $(x, y) \in \mathbb{R}^2$ przypisuje prawdopodobieństwo zdarzenia takiego, że $X < x \wedge Y < y$:

$$F(x, y) = P(X < x, Y < y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

gdzie PDF spełnia

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

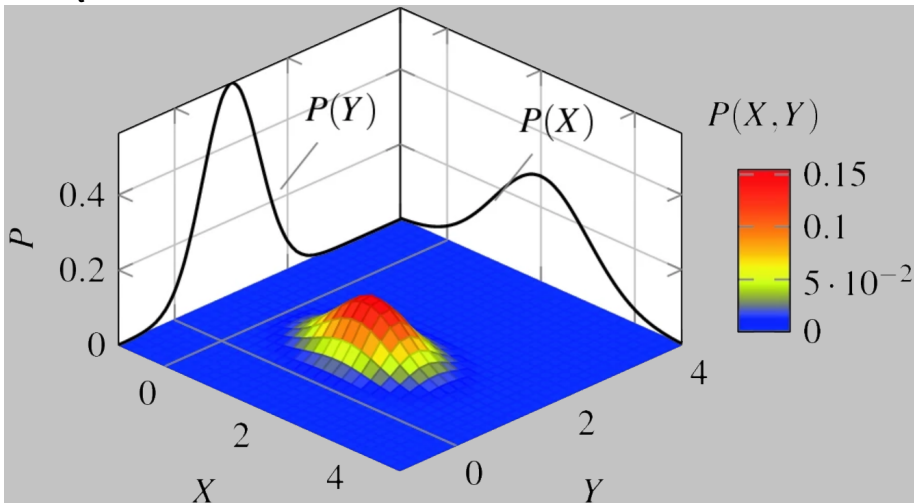
Prawdopodobieństwo zajścia zdarzenia $A \subset \mathbb{R}^2$ obliczamy jako

$$P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Para (X, Y) zmiennych losowych X i Y tworzy dwuwymiarową zmienną losową.



Funkcje zmiennych losowych

Zmienne wielowymiarowe

Gęstości brzegowe zmiennych X i Y w rozkładzie zmiennej dwuwymiarowej:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

$$f_2(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

Powiązane dystrybuanty brzegowe:

$$F_1(x) \equiv F(x, +\infty) = \int_{-\infty}^x f_1(u) du$$

$$F_2(y) \equiv F(+\infty, y) = \int_{-\infty}^y f_2(v) dv$$

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Gęstości brzegowe zmiennych X i Y w rozkładzie zmiennej dwuwymiarowej:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

$$f_2(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

Powiązane dystrybuanty brzegowe:

$$F_1(x) \equiv F(x, +\infty) = \int_{-\infty}^x f_1(u) du$$

$$F_2(y) \equiv F(+\infty, y) = \int_{-\infty}^y f_2(v) dv$$

Zmienne losowe X i Y są niezależne wtedy i tylko wtedy, gdy $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ lub równoważnie: $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Niech dwuwymiarowa zmienna losowa (X, Y) będzie przekształcona w dwuwymiarową zmienną (U_1, U_2) za pomocą wzajemnie jednoznacznych przekształceń ciągłych:

$$U_1 = g_1(X, Y), \quad U_2 = g_2(X, Y)$$

o ciągłych pochodnych. Oznaczmy przez $h_1 = g_1^{-1}$ i $h_2 = g_2^{-1}$ przekształcenia odwrotne.

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Niech dwuwymiarowa zmienna losowa (X, Y) będzie przekształcona w dwuwymiarową zmienną (U_1, U_2) za pomocą wzajemnie jednoznacznych przekształceń ciągłych:

$$U_1 = g_1(X, Y), \quad U_2 = g_2(X, Y)$$

o ciągłych pochodnych. Oznaczmy przez $h_1 = g_1^{-1}$ i $h_2 = g_2^{-1}$ przekształcenia odwrotne. Jakobian przekształcenia odwrotnego:

$$J = \begin{pmatrix} \frac{\partial x}{\partial u_1} & \frac{\partial x}{\partial u_2} \\ \frac{\partial y}{\partial u_1} & \frac{\partial y}{\partial u_2} \end{pmatrix}$$

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Niech dwuwymiarowa zmienna losowa (X, Y) będzie przekształcona w dwuwymiarową zmienną (U_1, U_2) za pomocą wzajemnie jednoznacznych przekształceń ciągłych:

$$U_1 = g_1(X, Y), \quad U_2 = g_2(X, Y)$$

o ciągłych pochodnych. Oznaczmy przez $h_1 = g_1^{-1}$ i $h_2 = g_2^{-1}$ przekształcenia odwrotne. Jakobian przekształcenia odwrotnego:

$$J = \begin{pmatrix} \frac{\partial x}{\partial u_1} & \frac{\partial x}{\partial u_2} \\ \frac{\partial y}{\partial u_1} & \frac{\partial y}{\partial u_2} \end{pmatrix}$$

Wtedy dwuwymiarowy PDF zmiennych U_1 i U_2 to

$$f_{U_1, U_2}(u_1, u_2) = f_{X, Y}[h_1(u_1, u_2), h_2(u_1, u_2)]|J|$$

Funkcje zmiennych losowych — przykład

Z: Rozważmy dwie niezależne zmienne losowe X, Y o rozkładzie jednostajnym $\mathcal{U}(0, 1)$. Jaki rozkład ma ich suma $X + Y$?

Funkcje zmiennych losowych — przykład

Z: Rozważmy dwie niezależne zmienne losowe X, Y o rozkładzie jednostajnym $\mathcal{U}(0, 1)$. Jaki rozkład ma ich suma $X + Y$?

R: Niech $U_1 = X + Y$, $U_2 = Y$. Wtedy $X = U_1 - U_2$, $Y = U_2$. Dla takiej transformacji wyznacznik jakobianu wynosi $|J| = 1$.

Funkcje zmiennych losowych — przykład

Z: Rozważmy dwie niezależne zmienne losowe X, Y o rozkładzie jednostajnym $\mathcal{U}(0, 1)$. Jaki rozkład ma ich suma $X + Y$?

R: Niech $U_1 = X + Y$, $U_2 = Y$. Wtedy $X = U_1 - U_2$, $Y = U_2$. Dla takiej transformacji wyznacznik jacobianu wynosi $|J| = 1$. Rozkład wspólny zmiennych (U_1, U_2) wynosi

$$f_{U_1, U_2}(u_1, u_2) = f_{X, Y}[x(u_1, u_2), y(u_1, u_2)] \cdot |J| = f_X(u_1 - u_2)f_Y(u_2)$$

Funkcje zmiennych losowych — przykład

Z: Rozważmy dwie niezależne zmienne losowe X, Y o rozkładzie jednostajnym $\mathcal{U}(0, 1)$. Jaki rozkład ma ich suma $X + Y$?

R: Niech $U_1 = X + Y$, $U_2 = Y$. Wtedy $X = U_1 - U_2$, $Y = U_2$. Dla takiej transformacji wyznacznik jacobianu wynosi $|J| = 1$. Rozkład wspólny zmiennych (U_1, U_2) wynosi

$$f_{U_1, U_2}(u_1, u_2) = f_{X, Y}[x(u_1, u_2), y(u_1, u_2)] \cdot |J| = f_X(u_1 - u_2) f_Y(u_2)$$

Rozkład zmiennej U_1 jest rozkładem brzegowym rozkładu wspólnego (U_1, U_2) :

$$f_{U_1=X+Y}(u_1) = \int_{-\infty}^{\infty} f_X(u_1 - u_2) f_Y(u_2) du_2 = \int_0^1 f_X(u_1 - u_2) du_2$$

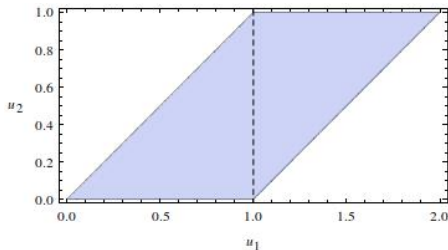
gdzie skorzystaliśmy z faktu, że $f_Y \neq 0$ dla $0 \leq u_2 \leq 1$. Funkcja podcałkowa jest niezerowa dla $0 \leq u_1 - u_2 \leq 1$, czyli wewnątrz równoległoboku $u_2 \leq u_1$, $u_2 \geq u_1 - 1$, w którym $f_X(u_1, u_2) = 1$.

Funkcje zmiennych losowych — przykład

Rozbija nam to całkę na dwie:

$$\int_0^{u_1} du_2 = u_1 \text{ dla } u_1 \in [0, 1]$$

$$\int_{u_1-1}^1 du_2 = 2 - u_1 \text{ dla } u_1 \in (1, 2]$$



Funkcje zmiennych losowych — przykład

Rozbija nam to całkę na dwie:

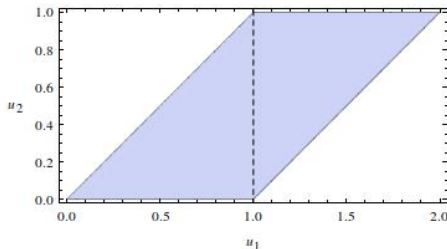
$$\int_0^{u_1} du_2 = u_1 \text{ dla } u_1 \in [0, 1]$$

$$\int_{u_1-1}^1 du_2 = 2 - u_1 \text{ dla } u_1 \in (1, 2]$$

Rozkład sumy

$$f_{U_1=X+Y}(u_1) = \begin{cases} u_1 & \text{dla } u_1 \in [0, 1] \\ 2 - u_1 & \text{dla } u_1 \in (1, 2] \end{cases}$$

jest zatem rozkładem trójkątnym:

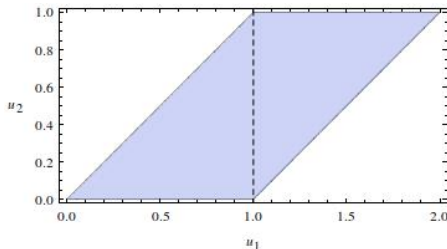


Funkcje zmiennych losowych — przykład

Rozbija nam to całkę na dwie:

$$\int_0^{u_1} du_2 = u_1 \text{ dla } u_1 \in [0, 1]$$

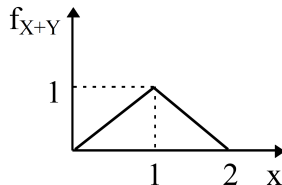
$$\int_{u_1-1}^1 du_2 = 2 - u_1 \text{ dla } u_1 \in (1, 2]$$



Rozkład sumy

$$f_{U_1=X+Y}(u_1) = \begin{cases} u_1 & \text{dla } u_1 \in [0, 1] \\ 2 - u_1 & \text{dla } u_1 \in (1, 2] \end{cases}$$

jest zatem rozkładem trójkątnym:



Funkcje zmiennych losowych

Podobnie można obliczyć rozkłady różnicy $X - Y$, iloczynu XY , pierwiastka $\sqrt{X}$, kwadratu X^2 , różnicy kwadratów $(X - Y)^2$ itd.

Funkcje zmiennych losowych

Podobnie można obliczyć rozkłady różnicy $X - Y$, iloczynu XY , pierwiastka $\sqrt{X}$, kwadratu X^2 , różnicy kwadratów $(X - Y)^2$ itd. Staje się to jednak żmudne i jest ograniczone istnieniem całek elementarnych.

Funkcje zmiennych losowych

Podobnie można obliczyć rozkłady różnicy $X - Y$, iloczynu XY , pierwiastka $\sqrt{X}$, kwadratu X^2 , różnicy kwadratów $(X - Y)^2$ itd. Staje się to jednak żmudne i jest ograniczone istnieniem całek elementarnych.

Z pomocą przychodzą metody Monte Carlo: symulujemy numerycznie wiele realizacji zmiennych $X, Y, \dots$, z których następnie konstruujemy interesującą nas zmienną losową jako funkcję tych zmiennych $X, Y, \dots$.

Funkcje zmiennych losowych

Podobnie można obliczyć rozkłady różnicy $X - Y$, iloczynu XY , pierwiastka $\sqrt{X}$, kwadratu X^2 , różnicy kwadratów $(X - Y)^2$ itd. Staje się to jednak żmudne i jest ograniczone istnieniem całek elementarnych.

Z pomocą przychodzą metody Monte Carlo: symulujemy numerycznie wiele realizacji zmiennych $X, Y, \dots$, z których następnie konstruujemy interesującą nas zmienną losową jako funkcję tych zmiennych $X, Y, \dots$.

Na przykład:

Zadanie 9

Rozważyć macierze postaci $\begin{pmatrix} a & b \\ b & d \end{pmatrix}$. Obliczyć różnicę między wartościami własnymi $\Delta\lambda = \lambda_2 - \lambda_1$, gdzie $\lambda_2 \geq \lambda_1$. Wykorzystać powyższą procedurę Monte Carlo by narysować rozkłady (histogramy) zmiennej losowej $\Delta\lambda$ gdy

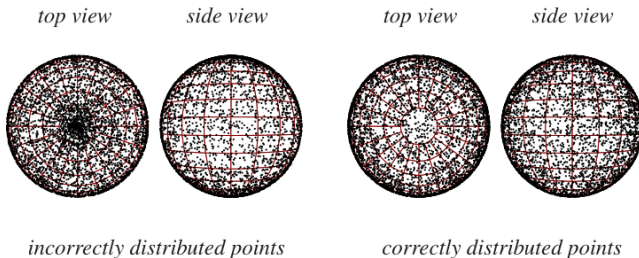
- 1 $a, b, d \sim \mathcal{U}(-1, 1)$ i są niezależne,
- 2 $a, b, d \sim \mathcal{N}(0, 1)$ i są niezależne.

Izotropowy rozkład na sferze niebieskiej

Weźmy np. współrzędne galaktyczne. Jeśli będą miały rozkłady jednostajne, $b \sim \mathcal{U}(-\pi/2, \pi/2)$, $l \sim \mathcal{U}(0, 2\pi)$,

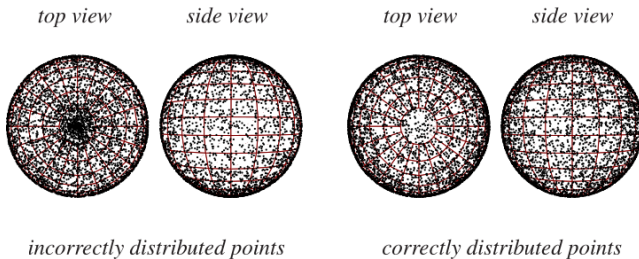
Izotropowy rozkład na sferze niebieskiej

Weźmy np. współrzędne galaktyczne. Jeśli będą miały rozkłady jednostajne, $b \sim \mathcal{U}(-\pi/2, \pi/2)$, $l \sim \mathcal{U}(0, 2\pi)$, to losowo rozmieszczone punkty będą się grupować wokół biegunów:



Izotropowy rozkład na sferze niebieskiej

Weźmy np. współrzędne galaktyczne. Jeśli będą miały rozkłady jednostajne, $b \sim \mathcal{U}(-\pi/2, \pi/2)$, $l \sim \mathcal{U}(0, 2\pi)$, to losowo rozmieszczone punkty będą się grupować wokół biegunów:



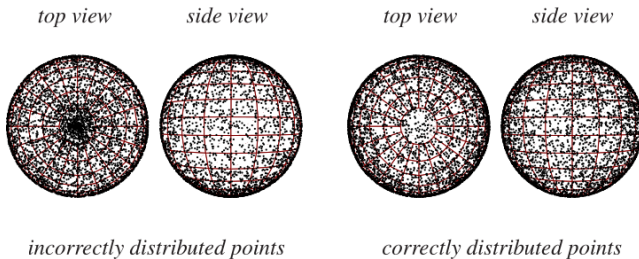
Żeby uzyskać rozkład izotropowy trzeba wygenerować współrzędne (b, l) wg

$$b = \arcsin(2u - 1), \quad l = 2\pi v$$

gdzie $u, v \sim \mathcal{U}(0, 1)$ są niezależne.

Izotropowy rozkład na sferze niebieskiej

Weźmy np. współrzędne galaktyczne. Jeśli będą miały rozkłady jednostajne, $b \sim \mathcal{U}(-\pi/2, \pi/2)$, $l \sim \mathcal{U}(0, 2\pi)$, to losowo rozmieszczone punkty będą się grupować wokół biegunów:



Żeby uzyskać rozkład izotropowy trzeba wygenerować współrzędne (b, l) wg

$$b = \arcsin(2u - 1), \quad l = 2\pi v$$

gdzie $u, v \sim \mathcal{U}(0, 1)$ są niezależne. Poprawny rozkład jest konieczny przy badaniu (an)izotropii wszelakich źródeł [[MT](#), [MNRAS 472\(4\), 4819 \(2017\)](#)].

Funkcje zmiennych losowych

Zmienne wielowymiarowe

Moment zwykły zmiennej dwuwymiarowej to

$$m_{kl} = \mathbb{E}[X^k Y^l]$$

Moment centralny to

$$\mu_{kl} = \mathbb{E}[(X - m_{10})^k (Y - m_{01})^l]$$

Z tych definicji:

$$m_{10} = \mathbb{E}(X), \quad m_{01} = \mathbb{E}(Y), \quad \mu_{20} = \text{Var}(X), \quad \mu_{02} = \text{Var}(Y)$$

Moment mieszany to kowariancja: $\mu_{11} = \text{Cov}(X, Y)$.

Zachodzi $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \equiv m_{11} - m_{10}m_{01}$.

Macierz kowariancji:

$$C = \begin{pmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \text{Var}(Y) \end{pmatrix}$$