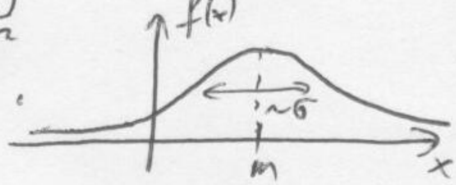
Normális eloszlás, centrális határeloszlás tétel

Minden folytonos eloszlások legfontosabbika: ! $m \in \mathbb{R}, \sigma > 0$.

Def: Az X val. változó normális avagy Gauss eloszlású (m, σ^2) paraméterekkel, ha sűrűségfüggvénye

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



Jelölés: $X \sim N(m, \sigma^2)$

Spec. eset, X standard normális eloszlású, ha $m = 0$, $\sigma^2 = 1$.

Jelölés: $X \sim N(0, 1)$. Sűrűségfv-e $\boxed{\varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-x^2/2}}$

Tény (nem trivi számolás): Ez az f_{m,σ^2} tényleg sűrűségfüggvény, vagyis $\int_{-\infty}^{\infty} f_{m,\sigma^2}(x)\,dx = 1$.

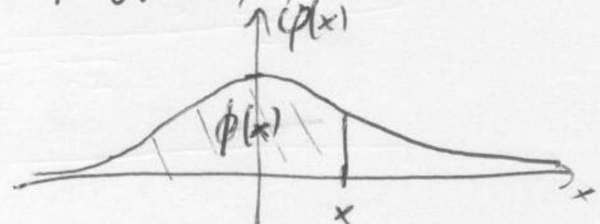
Nehézség (ami miatt nem trivi): az $\int f_{m,\sigma^2}(x)\,dx$ határozatlan integrál nem írható fel elemi függvények segítségével. Ez kényelmetlenné teszi a normális eloszlással való (papíron, ceruzával) számolást, de ettől még az eloszlás nem lesz kevésbé fontos.

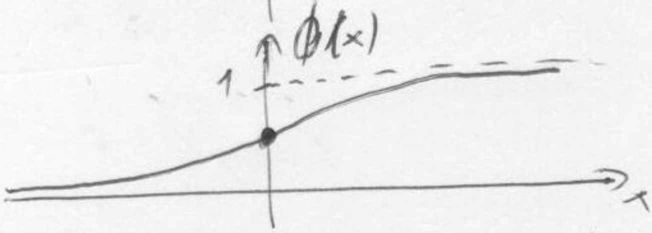
<u>Köv.:</u> $N(m,\sigma^2)$ <u>eloszlásfüggvényére</u> nincs „szép" képlet.

<u>Pontosabban:</u>

Def./jelölés: Legyen $\Phi : \mathbb{R} \to [0,1]$ a standard normális eloszlás eloszlásfüggvénye, vagyis

$$\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$$





<u>Szimmetria:</u> $\varphi(-x) = \varphi(x)$

$\Phi(-x) = 1 - \Phi(x)$

tengelyes szimmetria az y tengelyre

középpontos szimmetria a $(0, \frac{1}{2})$ pontra

<u>Tétel</u> (könnyű számolás): Ha $X \sim N(m,\sigma^2)$, akkor

$\mathbb{E}X = m$, $\mathrm{Var}\,X = \sigma^2$ — így persze $DX = \sigma$.

<u>Megj.:</u> Az irodalom nem egységes a jelölésben: nálam $N(m,\sigma^2)$-ben a 2. paraméter a szórás<u>négyzet</u>, de van, aki $N(m,\sigma)$-t használ, ahol σ a szórás.

Nagyen szerencsés tulajdonság:

Ha X normális elosztlású ~~…~~, $a, b \in \mathbb{R}$, ~~…~~ és $Y = aX+b$, akkor Y is normális elosztlású,

és persze $\mathbb{E}Y = a\mathbb{E}X + b$

$$\operatorname{Var} Y = a^2 \operatorname{Var} X$$

$$DY = |a| DX.$$

Avagy: Ha $X \sim N(m, \sigma^2)$,

akkor $aX+b \sim N(am+b, a^2\sigma^2)$

Spec:

- Ha $X \sim N(0,1)$, akkor $\sigma X + m \sim N(m, \sigma^2)$
- $\boxed{\text{Ha } X \sim N(m, \sigma^2), \text{ akkor } \frac{X-m}{\sigma} \sim N(0,1)}$

Köv: Egy $X \sim N(m, \sigma^2)$ ~~…~~ val. változó elosztlásfüggvénye

$$\boxed{F_{m,\sigma^2}(x) = \phi\left(\frac{x-m}{\sigma}\right)}$$

(ahol ϕ még mindig a standard normális elosztlásfüggvény.)

Biz.: $\mathbb{P}(X<x) = \mathbb{P}\left(\frac{X-m}{\sigma} < \frac{x-m}{\sigma}\right) \overset{\frac{X-m}{\sigma} \sim N(0,1)}{=\!=\!=} \phi\left(\frac{x-m}{\sigma}\right)$ ∎

Megj: Ez a képlet mégiscsak „szép". Egyetlen hátránya, hogy ϕ-t nem lehet beütni a butább számológépekbe, mint az e^x-et vagy a $\tan x$-et. Ez a számítógépek korában nem probléma.

Megj: Mivel $\phi(t)$ a zsebszámológépek nem ~~ismerik~~ (mind) ismerik, szokás (volt) táblázatban szedni. Ezek a táblázatok máig ~~fenn~~ fennmaradtak — a kurzus weblapján is van egy — de kizárólag egyetemisták használják, kizárólag ZH-íráskor.

Alkalmazás: Ha $a<b \in \mathbb{R}$, ~~akkor~~ és ~~$X \sim N(m, \sigma^2)$, akkor~~ $X \sim N(m, \sigma^2)$, akkor

$\cancel{\mathbb{P}(X<b) = \phi(b) \quad \mathbb{P}(X>a) = 1-}$

$$\mathbb{P}(X<b) = \phi\left(\frac{b-m}{\sigma}\right) ; \quad \mathbb{P}(X>a) = 1-\phi\left(\frac{a-m}{\sigma}\right),$$

$$\mathbb{P}(a<X<b) = \phi\left(\frac{b-m}{\sigma}\right) - \phi\left(\frac{a-m}{\sigma}\right)$$

Gyakorlati célokra ennyit kell tudni a normális eloszlásról (pl. a definíciót nem.)

A normális eloszlás jelentőségét az alábbi tétel adja:

Tétel (Centrális határeloszlás tétel = CHT):

Legyen $X_1, X_2, X_3 \ldots$ $\boxed{\text{független}}$ és $\boxed{\text{azonos eloszlású}}$ és $\boxed{\text{véges szórású}}$, $\mathbb{E}X_i = m$, $\operatorname{Var} X_i = \sigma^2$ (minden i-re), és legyen $S_n = X_1 + X_2 + \ldots + X_n$.

Ekkor, $\boxed{\text{ha } n \text{ nagy}}$, akkor S_n jól közelíthető normális eloszlással.

Pontosabban: $\dfrac{S_n - nm}{\sqrt{n}\,\sigma}$ jól közelíthető $N(0,1)$ eloszlással.

[Hát persze: $\mathbb{E}S_n = nm$, $\operatorname{Var} S_n = n\sigma^2$, $DS_n = \sqrt{n}\,\sigma$]

Még pontosabban: minden $x \in \mathbb{R}$-re

$$\mathbb{P}\left(\frac{S_n - nm}{\sqrt{n}\,\sigma} < x\right) \xrightarrow{n \to \infty} \Phi(x).$$

centrális határ eloszlás

Fontos: $\mathbb{P}\left(\dfrac{S_n - nm}{\sqrt{n}\,\sigma} < x\right) = \mathbb{P}\big(S_n < \underbrace{nm + \sqrt{n}\,\sigma x}_{K}\big) \approx$

$$\approx \Phi(x) = \Phi\left(\frac{K - nm}{\sqrt{n}\,\sigma}\right), \quad \text{de}$$

x-től (avagy K-tól) függ, hogy

- mennyire kell az n-nek nagynak lenni, és
- mennyire jó a közelítés.

Erről szól a Berry-Esseen tétel és (részben) a nagy eltérés tételek.

Tanulság röviden: Ha x túl nagy (n-hez képest), avagy $\underbrace{K - nm}_{\text{eltérés a várható értéktől}} \gg \sqrt{n}$, akkor a CHT közelítés rossz.

<u>Megj.</u>: A CHT fenti (legegyszerűbb) formája messzemenően általánosítható, a feltételek gyengíthetők:

- nem kell szigorúan vett függetlenség (de persze kell, hogy a függés gyenge legyen)
- ~~nem~~ nem kell csupa azonos eloszlású összeadandó (de persze kell, hogy ne ~~legyen~~ dominálja semelyik a többit)
- nem kell véges szórás (de azért „nagyon végtelen" nem lehet),

És még úgy is kijön a normális eloszlás, mint jó közelítés.

<u>Ezért</u> van, hogy a normális eloszlás a természetben lépten-nyomon szembejön.

A CHT történetéből:

1733: de Moivre kiszámolja a leges-legegyszerűbb esetre, de nem nevezi nevén a gyereket

1809: Gauss felfedezi a normális eloszlást

1810: Laplace bizonyítja az első ~~CHT-t~~ ilyen állítást

⋮

1920: Pólya György adja neki a „centrális határeloszlás tétel" nevet.

Pólyánál a „centrális" a vál.számban betöltött szerepre utal.

1922: Lindeberg bizonyítja először úgy, ahogy én kimondtam. [És annál általánosabban is.]

⋮

2000-es évek: minden évben több száz cikk a további általánosításokról

REKLÁM $\boxed{\text{CHT}}$ és egyéb határeloszlás tételek

Miért szeretjük őket?

Mire jó a $\mathbb{P}(X_1+X_2+\dots+X_n<K)$ valószínűséget CHT-vel közelíteni?

Miért nem számoljuk ki inkább pontosan?

3. számú ok: A modellünk úgyis csak durva közelítése a valóságnak: kár a pontos számolással szenvedni.

2. számú ok: Ha n nagy, a pontos számolás ~~fe~~ számítógéppel is reménytelenül nehéz (értsd: időigényes), számítógéppel is.

$\boxed{\text{1. számú ok:}}$ A pontos számoláshoz sokszor $\boxed{\text{nincs elég}}$ $\boxed{\text{információ}}$: az X_i-k eloszlását nem ismerjük.

A CHT alkalmazásához sokkal kevesebb infó is elég: csak m és σ, semmi más.