

SF1901: Sannolikhetslära och statistik

Föreläsning 7.

Normalfördelning, Linjära kombinationer av normalfördelade variabler, Centrala gränsvärdessatsen, Approximationer

Jan Grandell & Timo Koski

07.02.2016



From a Data Set to a Probability Distribution

Adolphe Quetelet (1796–1874) was an influential statistician and a tireless promoter of statistical data collection¹, who wanted to mold social science to an exact science like physics, he talked about 'social physics'. He presented a theory of *the average man* invoking the normal (Gaussian) distribution. He extended his statistical analysis of the demographic and anthropometric characteristics of man to behaviour, mind and soul.

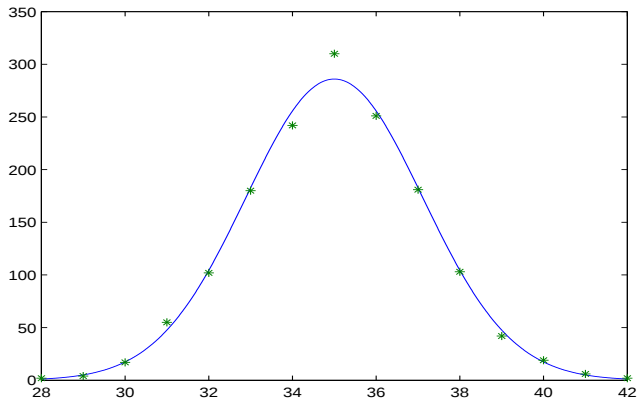
¹or is it a hunter of factoids

A data set due to Quetelet according to a textbook by J.W. Lindeberg, 1927 : chest breadths of 1516 Belgian soldiers

Class i	Absolute frequency f_i	The values of $g(x) = 286e^{-0.112(x-35)^2}$ $g(x_i)$
28	2	1.2
29	4	5.1
30	17	17.4
31	55	47.7
32	102	104.4
33	180	182.7
34	242	255.7
35	310	286.0
36	251	255.7
37	181	182.7
38	103	104.4
39	42	47.7
40	19	17.4
41	6	5.1

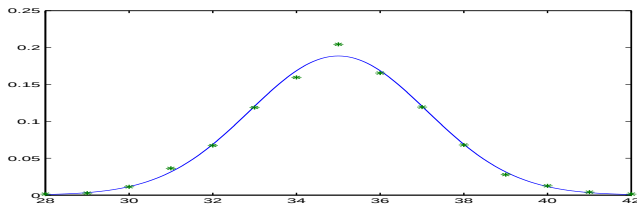
From a Data Set to a Probability Distribution

A plot of the frequency data and the function $286e^{-0.112(x-35)^2}$



'stickprov ur en fördelning' !

Plotten av de relativa frekvenserna av Quetelets data och funktionen $\frac{286}{1516} e^{-0.112(x-35)^2}$.

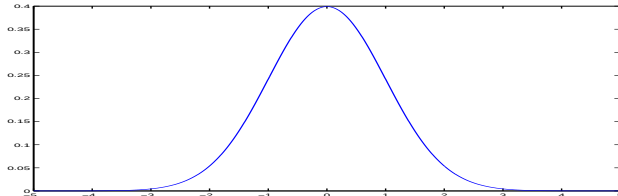


Bell-Shaped Curve

$\sigma > 0,$

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

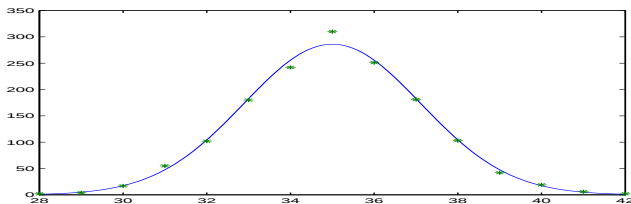
I figuren $\mu = 1, \sigma = 1$



Funktionen $286e^{-0.112(x-35)^2}$ svarar mot

$$g(x) = \frac{1516}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

med $\mu = 35, \sigma^2 = 4.46$ (1516 är summan av klassfrekvenserna)

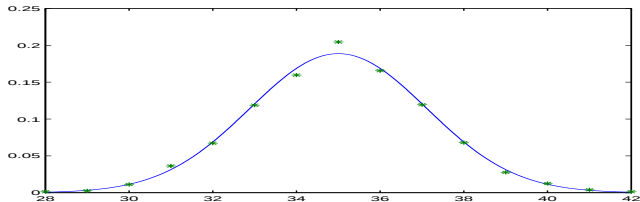


Normal (Probability) Density Curve

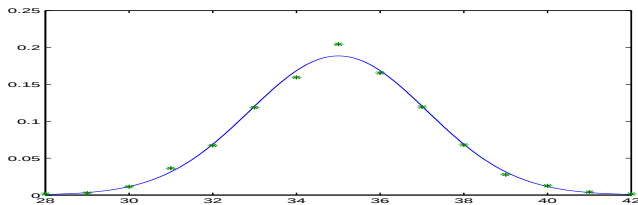
De relativa frekvenserna i Quetelets data och

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

med $\mu = 35, \sigma^2 = 4.46$



SLUT PÅ INLEDNINGEN MED QUETELET-



Standardiserad normalfördelning, $N(0, 1)$

Definition

En s.v. Z säges vara standardiserad normalfördelad om den är $N(0, 1)$ -fördelad, dvs. om den har täthetsfunktionen

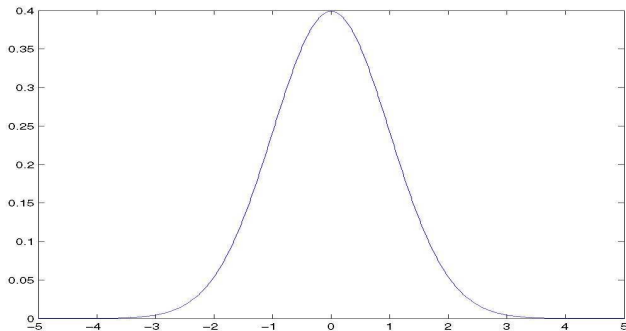
$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Dess fördelningsfunktion betecknas med $\Phi(z)$, dvs.

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Standardiserad normalfördelning, $N(0, 1)$

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$



En bilaga i fin stil (kan överhoppas)

Hur vet vi att $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ är en sannolikhetstäthet ? Dvs. varför gäller det att

$$\int_{-\infty}^{\infty} \varphi(z) dz = 1.$$

Svaret ges t.ex. i Eike Petermann: *Analytiska metoder II*, Studentlitteratur 2002, sid. 235, Ex. 9.14, Anmärkning 9.6 eller bilagan nedan



Bilaga (ur Eike Petermann: *Analytiska metoder II*) :

$$\int_{-\infty}^{\infty} \varphi(z) dz = 1.$$

Exempel 9.14:

$$\begin{aligned} \iint_{\mathbb{R}^2} e^{-(x^2+y^2)} dx dy &= \left[\begin{array}{l} x = r \cos v, \quad 0 \leq r \\ y = r \sin v, \quad 0 \leq v < 2\pi, \quad \det \left(\frac{d(x,y)}{d(r,v)} \right) = r \end{array} \right] = \\ &= \iint_{0 \leq r} e^{-r^2} r dr dv = \int_0^{2\pi} \left[\int_0^{\infty} e^{-r^2} r dr \right] dv = \int_0^{\infty} e^{-r^2} r dr \cdot \int_0^{2\pi} dv = \\ &= \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty} \cdot \left[v \right]_0^{2\pi} = \frac{1}{2} \cdot 2\pi = \pi. \quad \blacksquare \end{aligned}$$

$$\int_{-\infty}^{\infty} \varphi(z) dz = 1.$$

9.6

Anmärkning 9.6:

Använder man sats 9.11 utan föregående substitution på integralen i exemplet ovan, så får man:

$$\begin{aligned} \iint_{\mathbb{R}^2} e^{-(x^2+y^2)} dx dy &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} e^{-x^2-y^2} dx \right] dy = \int_{-\infty}^{\infty} e^{-x^2} dx \cdot \int_{-\infty}^{\infty} e^{-y^2} dy = \\ &= \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^2 \end{aligned}$$

Eftersom dubbelintegralens värde beräknades till π , får man det anmärkningsvärda resultatet:

$$\boxed{\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}}$$

Trots att e^{-x^2} saknar elementär primitiv funktion²⁰ så kan alltså den speciella integralen $\int_{-\infty}^{\infty} e^{-x^2} dx$ beräknas exakt!

Sista sidan i bilagan :

$$\int_{-\infty}^{\infty} \varphi(z) dz = 1.$$

Sambandet är av intresse bland annat för normalfördelningen i sannolikhetsläran (jämför exempel 9.12): Konstanten k i frekvensfunktionen

$$f(x) = k \cdot e^{-x^2/2\sigma^2}$$

måste väljas så att
$$\int_{-\infty}^{\infty} k \cdot e^{-x^2/2\sigma^2} dx = 1$$

Efter variabelsubstitutionen $x = \sigma\sqrt{2}t$ får man

$$\frac{1}{k} = \sigma\sqrt{2} \int_{-\infty}^{\infty} e^{-t^2} dt = \sigma\sqrt{2\pi}$$

dvs.

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



KTH Matematik

Standardiserad normalfördelning, $N(0, 1)$

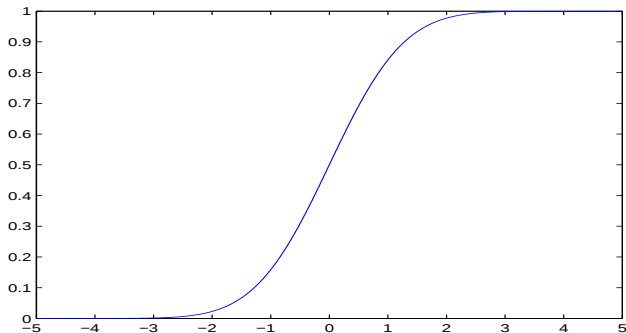
Fördelningsfunktionen betecknas med $\Phi(z)$, dvs.

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

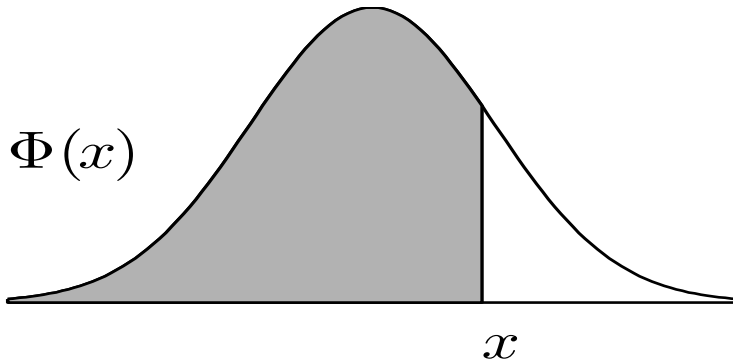
Av symmetriskäl gäller klart att

$$\Phi(0) = \frac{1}{2}.$$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$



$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$



Beräkning av $\Phi(z)$: MATLAB

Ett problem är att fördelningsfunktionen inte kan ges på en sluten analytisk form. Det är dock lätt att numeriskt beräkna fördelningsfunktionen och vi använder programvara för beräkning av $\Phi(x)$. I MATLAB Statistics Toolbox beräknas $\Phi(1)$ av :

```
>> normcdf(1,0,1)
```

```
ans =
```

```
0.84134474607
```



Beräkning av $\Phi(z)$: TI-82 STATS

Ta fram menyn DISTR med tangentsekvensen 2nd DISTR. Gå ned till **normalcdf**(. Då beräknas $\Phi(1)$ av²:

normcdf(-1E99, 1, 0, 1)

.8413447404

Man kan även skriva (d.v.s. 0 och 1 är defaultvärden)

normcdf(-1E99, 1)

.8413447404

²E tas fram med tangentsekvensen 2ND E

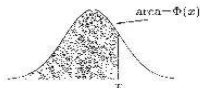
Tabellen för $\Phi(x)$ ur kursens formelsamling

Vi slår upp $\Phi(1)$ som .8413

Tabell 1. Standard normalfördelning.

$\Phi(x) = P(X \leq x)$, där $X \in N(0, 1)$.

För negativa x , utnyttja att $\Phi(-x) = 1 - \Phi(x)$



x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

$$\Phi(-z) = 1 - \Phi(z).$$

Vi observerar att $\varphi(-z) = \varphi(z)$. $\Phi(z)$ är tabulerad i kursens formelsamling endast för $z \geq 0$. Vi har dock

$$\begin{aligned}\Phi(-z) &= \int_{-\infty}^{-z} \varphi(x) dx = [y = -x] = - \int_{\infty}^z \varphi(-y) dy \\ &= \int_z^{\infty} \varphi(y) dy = 1 - \Phi(z).\end{aligned}$$

Sats

$$\Phi(-z) = 1 - \Phi(z).$$

$$Z \in N(0, 1), E(Z), V(Z)$$

Om Z är $N(0, 1)$ -fördelad, så kan man visa att

$$\begin{aligned} E(Z) &= 0 \quad (\text{ty } \varphi(-z) = \varphi(z)) \\ V(Z) &= 1. \end{aligned}$$



När vi kommer till statistikdelen av kursen behöver vi ofta lösa ekvationer av följande slag:

Bestäm z så att vi för givet α har

- $P(Z \leq z) = 1 - \alpha;$
- $P(Z > z) = 1 - \alpha;$
- $P(-z < Z \leq z) = 1 - \alpha.$

För att lösa sådana ekvationer inför vi α -kvantilen λ_α definierad av $P(Z > \lambda_\alpha) = \alpha$ eller

$$\alpha = 1 - \Phi(\lambda_\alpha).$$

Kvantiler i TI-82 STATS invNorm(

Vi vill ta fram $\lambda_{0.05}$ definierad av ekvationen $P(Z > \lambda_{0.05}) = 0.05$ eller

$$0.05 = 1 - \Phi(\lambda_{0.05}) \Leftrightarrow \Phi(\lambda_{0.05}) = 0.95$$

Vi knappar in i räknären (meny DISTR)

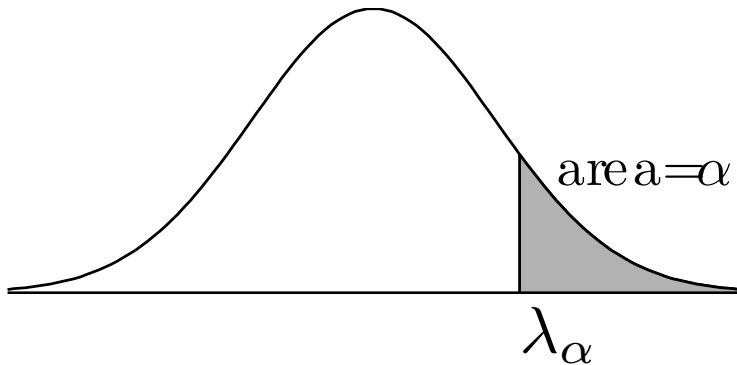
$$\text{invNorm}(0.95, 0, 1)$$

som ger lösningen

$$1.644853626$$

(≈ 1.6449).





Tabell 2 i kursens formelsamling ger $\lambda_{0.05} = 1.6449$.

Normalfördelningens kvantiler $P(X > \lambda_\alpha) = \alpha$

där $X \in N(0, 1)$

$$\alpha = 1 - \Phi(\lambda_\alpha).$$

Det är då bra att observera att

$$1 - \alpha = 1 - \Phi(\lambda_{1-\alpha})$$

$$\Leftrightarrow$$

$$\alpha = \Phi(\lambda_{1-\alpha})$$

$$\Leftrightarrow$$

$$\alpha = 1 - \Phi(-\lambda_{1-\alpha}),$$

vilket ger

$$\lambda_{1-\alpha} = -\lambda_\alpha.$$

Definition

En s.v. X säges vara $N(\mu, \sigma)$ -fördelad, där μ reell och $\sigma > 0$, om

$$Z = \frac{X - \mu}{\sigma} \text{ är } N(0, 1)\text{-fördelad.}$$

Sats

Låt X vara $N(\mu, \sigma)$ -fördelad. Då gäller

$$f_X(x) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

och

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Bevis. Vi har

$$\begin{aligned} F_X(x) &= P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right). \end{aligned}$$

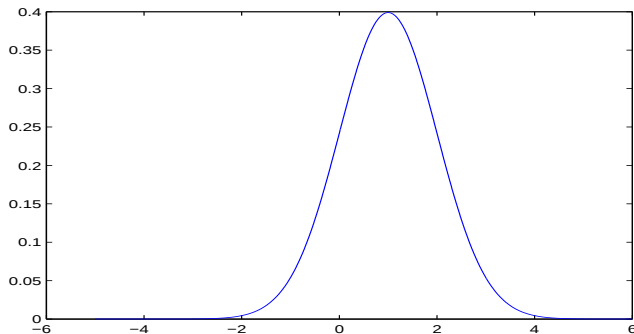
Derivation ger $f_X(x) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}\right)$.



$X \in N(\mu, \sigma)$ med

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

där μ godtycklig konstant och $\sigma > 0$. I figuren för $f_X(x)$ har vi $\mu = 1, \sigma = 1$



Normalfördelning (även känd som Gaussfördelning efter C.F. Gauss, 1777-1855)



$X \in N(0, \sigma)$ med

$$f_X(x) = \sqrt{\frac{m}{2\pi k_B T}} e^{-mx^2/2k_B T}$$

d.v.s. $\sigma = \sqrt{\frac{k_B T}{m}}$, k_B = Boltzmanns konstant, T = temperatur, m = partikelns massa. "the fraction of particles in with velocities in the x -direction within $x, x + dx = f_X(x) dx$ ". .

En viktig regel

Beviset ovan innehåller en viktig räkneregel. Om X är $N(\mu, \sigma)$ -fördelad, så gäller det att

$$F_X(x) = P(X \leq x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

Man kan m.a.o. använda tabellen för $\Phi(x)$ även för att beräkna $F_X(x)$ för $X \in N(\mu, \sigma)$.

Smidigare än med regeln ovan och tabellen kan vi beräkna, om $X \in N(\mu, \sigma)$, sannolikheten $F_X(x) = P(X \leq x)$ som

$$\text{normcdf}(-1E99, x, \mu, \sigma)$$

Till exempel, om $X \in N(2, 2)$ och beräknas $P(0 \leq X \leq 3.5)$ med TI-82 STATS som

$$\begin{aligned} &\text{normcdf}(0, 3.5, 2, 2) \\ &.614717461 \end{aligned}$$

Sats

Om X är $N(\mu, \sigma)$ -fördelad så gäller

$$E(X) = \mu \quad \text{och} \quad V(X) = \sigma^2.$$

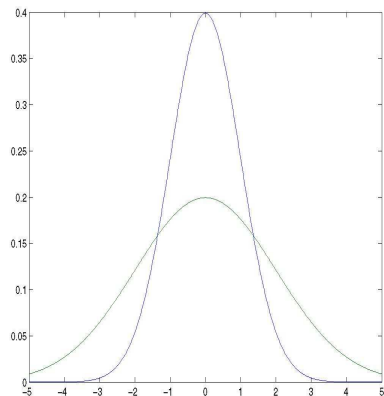
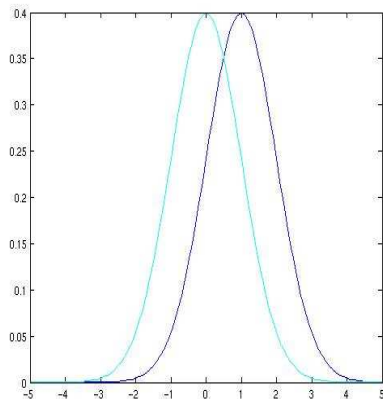
Bevis. Vi ska nu se hur listig definitionen är!

$$X = \sigma Z + \mu$$

$$E(X) = \sigma E(Z) + \mu = 0 + \mu = \mu$$

$$V(X) = \sigma^2 V(Z) + 0 = \sigma^2.$$

Täthetsfunktionerna för $N(0, 1)$ och $N(1, 1)$ och $N(0, 1)$ och $N(0, 2)$ (från vänster till höger)



Sats

Låt X vara $N(\mu, \sigma)$ -fördelad och sätt $Y = aX + b$. Då gäller det att

Y är $N(a\mu + b, |a|\sigma)$ -fördelad.

Bevis. Från definitionen följer att $X = \mu + \sigma Z$ där Z är $N(0, 1)$ -fördelad. Detta ger

$$Y = aX + b = a(\mu + \sigma Z) + b = a\mu + b + a\sigma Z$$

$$\frac{Y - (a\mu + b)}{a\sigma} = Z.$$

Om $a > 0$ följer satsen. Om $a < 0$ utnyttjar vi att Z och $-Z$ har samma fördelning.



Sats

Om X är $N(\mu_X, \sigma_X)$ -fördelad, Y är $N(\mu_Y, \sigma_Y)$ -fördelad och X och Y är oberoende så gäller att

$$X + Y \text{ är } N\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right)\text{-fördelad}$$

och

$$X - Y \text{ är } N\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right)\text{-fördelad.}$$

Repetition om väntevärden

Återkalla i minnet att Låt X och Y vara två oberoende (okorrelerade räcker) s.v. Då gäller

$$E(X + Y) = E(X) + E(Y) \quad V(X + Y) = V(X) + V(Y)$$

$$E(X - Y) = E(X) - E(Y) \quad V(X - Y) = V(X) + V(Y).$$

Det nya är att vi kan ge fördelningen för summan av oberoende normalfördelade variabler



Sats

Låt $X_1, \dots, X_n$ vara oberoende och $N(\mu_1, \sigma_1), \dots, N(\mu_n, \sigma_n)$. Då gäller att

$$\sum_{k=1}^n c_k X_k \text{ är } N\left(\sum_{k=1}^n c_k \mu_k, \sqrt{\sum_{k=1}^n c_k^2 \sigma_k^2}\right)\text{-fördelad.}$$

Allmän regel: Linjärkombinationer av oberoende normalfördelade stokastiska variabler är normalfördelade med rätt väntevärde och rätt standardavvikelse.

Sats

Låt $X_1, \dots, X_n$ vara oberoende (okorrelerade ränder) s.v. och sätt

$$Y = c_1 X_1 + \dots + c_n X_n.$$

Då gäller

$$E(Y) = c_1 E(X_1) + \dots + c_n E(X_n)$$

och

$$V(Y) = c_1^2 V(X_1) + \dots + c_n^2 V(X_n)$$

Det nya är att vi för linjärkombinationer av oberoende normalfördelade stokastiska variabler kan ge hela fördelningen.

Följdsats

Låt $X_1, X_2, \dots, X_n$ vara oberoende och $N(\mu, \sigma)$ -fördelade s.v. Då gäller att

$$\bar{X} \text{ är } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)\text{-fördelad.}$$

Återkalla i minnet:

Sats

Låt $X_1, X_2, \dots, X_n$ vara oberoende och likafördelade s.v. med väntevärde μ och standardavvikelse σ . Då gäller att

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n} \quad \text{och} \quad D(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Centrala gränsvärdessatsen

Vi har sett några exempel på att normalfördelningen har trevliga statistiska egenskaper. Detta skulle vi inte ha så stor glädje av, om normalfördelningen inte dessutom var vanligt förekommande. Centrala gränsvärdessatsen CGS, som är den huvudsakliga motiveringen för normalfördelningen, kan utan vidare sägas vara ett av sannolikhetsteorins och statistikens allra viktigaste resultat.

Sats

(CGS) Låt $X_1, X_2, \dots$ vara oberoende och lika fördelade s.v. med väntevärde μ och standardavvikelse σ . Då gäller att

$$P\left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \rightarrow \Phi(x) \quad \text{då } n \rightarrow \infty.$$

Ofta uttrycker man slutsatsen i CGS som att

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \text{ är approximativt } N(0, 1)\text{-fördelad}$$

eller att

$$\sum_{i=1}^n X_i \text{ är approximativt } N(n\mu, \sigma\sqrt{n})\text{-fördelad.}$$

En, för statistiken mycket vanlig användning av CGS är följande:

Följsats

Låt $X_1, X_2, \dots$ vara oberoende och lika fördelade s.v. med väntevärde μ och standardavvikelse σ . Då gäller att

$$P(a < \bar{X} \leq b) \approx \Phi\left(\frac{b - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right)$$

om n är tillräckligt stort.

Det är tyvärr inte möjligt att ge några generella och enkla tumregler om hur stort n måste vara för att normalapproximationen ska vara användbar. Detta beror på hur "normalliknande" de enskilda variablerna X_k är. Om X_k na är normalfördelade så "gäller" ju CGS för alla n . En tumregel är att om X_k na är någorlunda symmetriskt fördelade så räcker ganska små n , säg något tiotal. Om X_k na är påtagligt skevt fördelade så behöver n var något eller i värsta fall några hundratal.

Det är svårt att formulera strikt, men det räcker i CGS att X_k na är någorlunda oberoende och någorlunda lika fördelade. Med "någorlunda lika fördelade" menas framförallt att det inte finns vissa X_k som är mycket dominerande. Detta innebär att mätfel i valgjorda försök kan anses vara approximativt normalfördelade. I mindre valgjorda försök kan det däremot mycket väl finnas någon dominerande felkälla som inte alls behöver vara approximativt normalfördelad.