Sannolikhet och statistik

Normalfördelningen

VT 2009

Uwe.Menzel@math.uu.se

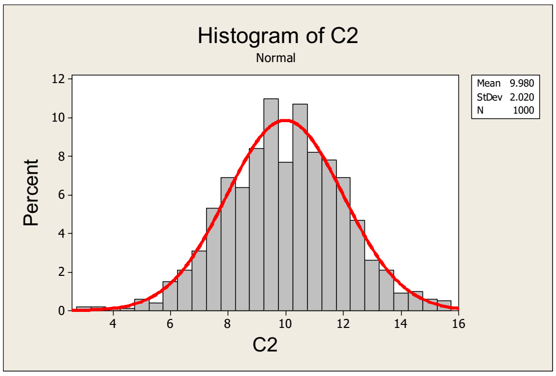
http://www.math.uu.se/~uwe/

Data

85, 87, 150, 100, 100, 90, 70, 72, 75, 70, 85, 143, 100, 121, 92, 66, 70, 69, 75, 80, 140, 92, 130, 83, 70, 68, 67, 75, 83, 149, 95, 130, 80, 68, 85, 75, 73, 78, 140, 90, 124, 86, 69, 70, 75, 77, 110, 165, 110, 150, 110, 115, 80, 75, 75, 98, 172, 110, 145, 110, 95, 52, 80, 96, 110, 168, 110, 145, 110, 80, 80, 75, 89, 95, 170, 110, 145, 120, 89, 72, 79, 75, 95, 220, 100, 149, 100, 110, 80, 85, 80, 90, 165, 103, 135, 95, 77, 76, 85, 80, 88, 155, 103, 120, 85, 79, 78, 82, 75, 85, 150, 103, 135, 90, 75, 85, 78, 75, 88, 150, 95, 130, 90, 70, 76, 89, 82, 95, 145, 100, 133, 90, 77, 89, 79, 80, 90, 165, 103, 135, 95, 77, 86, 80, 85, 100, 160, 120, 140, 100, 90, 79, 92, 70, 100, 165, 120, 140, 100, 120, 86, 71, 95, 100, 155, 120, 139, 100, 89, 86, 78, 78, 110, 158, 122, 145, 108, 95, 95, 78

- Delar in i intervaller
- Hur många värden finns i varje intervall?

Normalfördelningen är "viktigast"



Moivre 1733 (myntkast)
Laplace 1812
Ljapunov

Approximation av en diskret fördelning

Normalfördelningen

Philosophically speaking, the Normal distribution represents one of the <u>empirically verified</u> elementary "truths about the general nature of reality," and its status can be compared to the one of fundamental laws of natural sciences.

http://www.statsoft.com/textbook/esc.html

Täthetsfunktion, väntevärde och varians

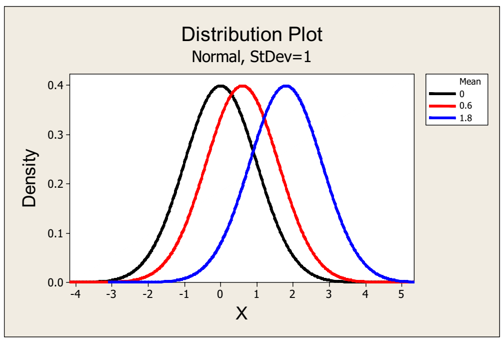
$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty}$$

$$E(X) = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \mu$$

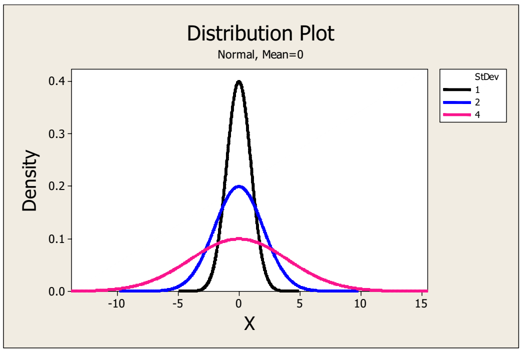
$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}\, dx = \sigma^2$$

$$D(X) = \sigma$$

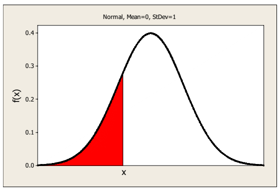
Väntevärdet μ



Standardavvikelsen σ



Fördelningsfunktionen (cdf)



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

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

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

Det finns inget analytiskt uttryck för $F_X(x)$

Låt X vara normalfördelad: $X \in N(\mu, \sigma)$.

Låt $Y = (X-\mu)/\sigma$. Vilken fördelningsfunktion har Y ?

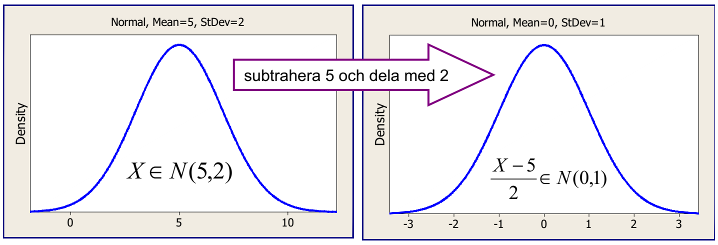
$$\begin{aligned}
F_Y(y) &= P(Y \le y) = P\left(\frac{X-\mu}{\sigma} \le y\right) \\
&= P(X \le \mu + \sigma y) = F_X(\mu + \sigma y) \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\mu+\sigma y} e^{-(t-\mu)^2/2\sigma^2}\,dt \quad \text{subst. } u = (t-\mu)/\sigma \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-(u)^2/2}\,du \quad = \Phi(y)
\end{aligned}$$

Det är fördelningsfunktionen för $N(0,1)$, alltså $\boxed{Y \in N(0,1)}$

Standardiserad normalvariabel

$$X \in N(\mu,\sigma) \quad \Rightarrow \quad \frac{X-\mu}{\sigma} \in N(0,1)$$

Om X är normalfördelad, så är (X-μ)/σ standard-normalfördelad.



Fördelningsfunktion för allmän normalfördelning

$$X \in N(\mu,\sigma) \quad \Rightarrow$$

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

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

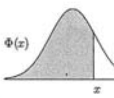
Med hjälp av denna formel och en tabell över Φ kan man beräkna fördelningsfunktionen för alla värden av x, för alla μ och σ (med en viss precision).

Standard normalfördelning: Φ(x)

μ = 0
σ= 1

Blom s. 397

Tabell 1. Standard normalfördelning

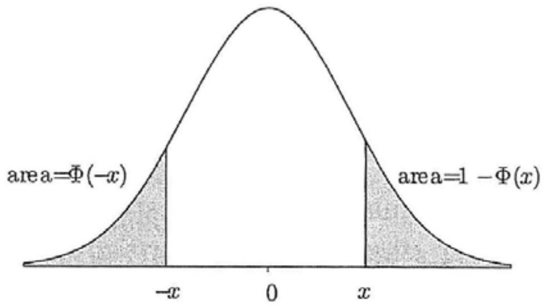


$\Phi(x) = P(X \le x)$, där $X \in N(0,1)$.
För negativa x, utnyttja att $\Phi(-x) = 1 - \Phi(x)$

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
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

tabellerad från 0 .. 4.0 i 0.01-steg

$\Phi(x)$ för negativa argument



Fördelningsfunktion för allmän normalfördelning

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

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

Exempel 1 och 2

Exempel 1: $X \in N(4,2)$ Vad är $P(X \le 3)$??

$$\begin{aligned} P(X \le 3) &= F_X(3) = \Phi\left(\frac{3-4}{2}\right) = \Phi(-0.5) \\ &= 1 - \Phi(0.5) = 1 - 0.6915 = \mathbf{0.3085} \end{aligned}$$

Exempel 2: $X \in N(4,2)$ Vad är $P(X \ge 5)$??

$$\begin{aligned} P(X \ge 5) &= 1 - P(X < 5) = 1 - P(X \le 5) \\ &= 1 - F_X(5) = 1 - \Phi\left(\frac{5-4}{2}\right) \\ &= 1 - \Phi(0.5) = 1 - 0.6915 = \mathbf{0.3085} \end{aligned}$$

Sannolikhet att en normalfördelad s.v. hamnar mellan värden a och b

$$P(a < X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

gäller allmänt

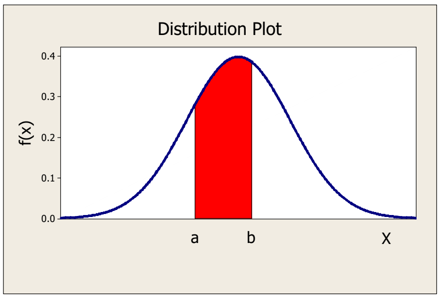
gäller om X är normalfördelad N(μ, σ)

s.v.= slumpvariabel

Sannolikhet att s.v. X hamnar mellan a och b

s.v.= slumpvariabel

$$P(a < X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$



Exempel 3

$$\boxed{P(a < X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

Exempel: $X \in N(4,2)$ Vad är $P(3 \le X \le 5)$??

$$\begin{aligned} P(3 \le X \le 5) &= F_X(5) - F_X(3) \\ &= \Phi\left(\frac{5-4}{2}\right) - \Phi\left(\frac{3-4}{2}\right) \\ &= \Phi(0.5) - \Phi(-0.5) \\ &= \Phi(0.5) - [1 - \Phi(0.5)] \\ &= 2 \cdot \Phi(0.5) - 1 \\ &= 2 \cdot 0.6915 - 1 = 0.383 \end{aligned}$$

Sannolikheten inom σ-intervaller

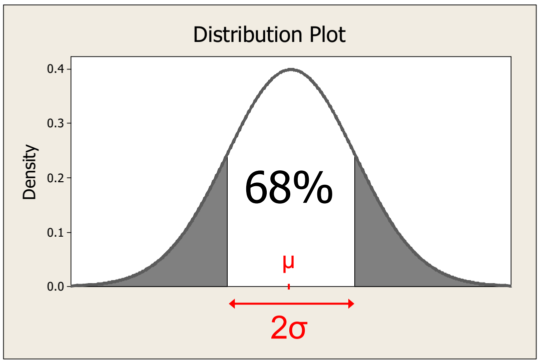
$$P(a < X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$$\begin{aligned} P(\mu - 2\cdot\sigma < X \le \mu + 2\cdot\sigma) &= F_X(\mu + 2\cdot\sigma) - F_X(\mu - 2\cdot\sigma) \\ &= \Phi\left(\frac{\mu + 2\cdot\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\cdot\sigma - \mu}{\sigma}\right) \\ &= \Phi(2) - \Phi(-2) = 0.954 \end{aligned}$$

Samma med 1 och 3:

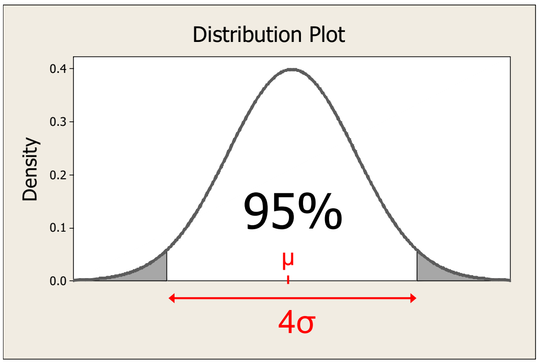
$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= 0.682 \\ P(\mu - 2\sigma \le X \le \mu + 2\sigma) &= 0.954 \\ P(\mu - 3\sigma \le X \le \mu + 3\sigma) &= 0.997 \end{aligned}$$

$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= 0.682 \\ P(\mu - 2\sigma \le X \le \mu + 2\sigma) &= 0.954 \\ P(\mu - 3\sigma \le X \le \mu + 3\sigma) &= 0.997 \end{aligned}$$

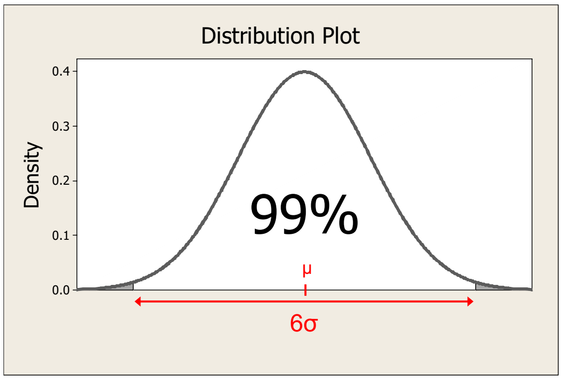


nsigma.MPJ

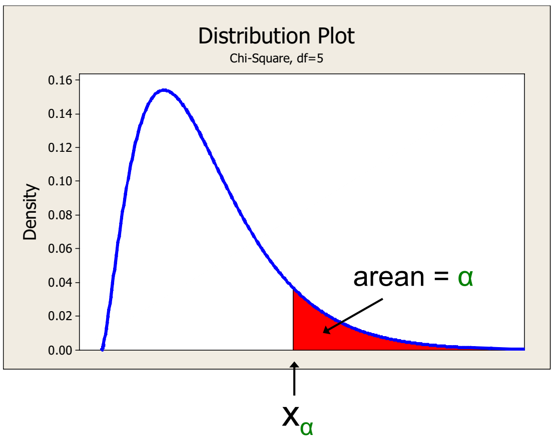
$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= 0.682 \\ P(\mu - 2\sigma \le X \le \mu + 2\sigma) &= 0.954 \\ P(\mu - 3\sigma \le X \le \mu + 3\sigma) &= 0.997 \end{aligned}$$



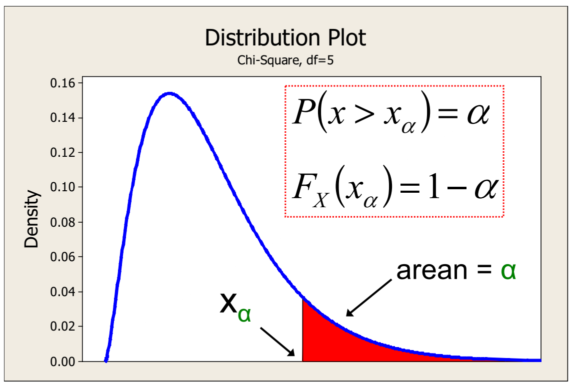
$$\begin{aligned} P(\mu - \sigma \le X \le \mu + \sigma) &= 0.682 \\ P(\mu - 2\sigma \le X \le \mu + 2\sigma) &= 0.954 \\ P(\mu - 3\sigma \le X \le \mu + 3\sigma) &= 0.997 \end{aligned}$$



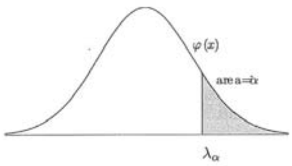
Definition av kvantilerna



Kvantiler



Kvantilerna för N(0,1) kallas λ_α



Figur: Kvantiler för $N(0,1)$

α	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
λ_α	3.29	3.09	2.58	2.33	1.96	1.64	1.28

Sannolikhet inom $\mu \pm \lambda_{\alpha/2}\sigma$ för allmän normalfördelning

$$\boxed{P(a < X \le b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

$$\begin{aligned} P(\alpha) &= P\left(\mu - \lambda_{\alpha/2}\sigma < X \le \mu + \lambda_{\alpha/2}\sigma\right) \\ &= F_X\left(\mu + \lambda_{\alpha/2}\sigma\right) - F_X\left(\mu - \lambda_{\alpha/2}\sigma\right) \\ &= \Phi\left(\frac{\mu + \lambda_{\alpha/2}\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \lambda_{\alpha/2}\sigma - \mu}{\sigma}\right) \\ &= \Phi\left(\lambda_{\alpha/2}\right) - \Phi\left(-\lambda_{\alpha/2}\right) \\ &= \Phi\left(\lambda_{\alpha/2}\right) - \left[1 - \Phi\left(\lambda_{\alpha/2}\right)\right] \\ &= 2 \cdot \Phi\left(\lambda_{\alpha/2}\right) - 1 \\ &= 2\left[1 - \alpha/2\right] - 1 \\ &= 1 - \alpha \end{aligned}$$

Sannolikhet inom $\mu \pm \lambda_{\alpha/2}\sigma$ för allmän normalfördelning

$$\boxed{P\left(\mu - \lambda_{\alpha/2}\sigma < X \le \mu + \lambda_{\alpha/2}\sigma\right) = 1 - \alpha}$$

Sätt $\alpha = 0.05,\ 0.001,\ 0.0001$

$$\begin{aligned} P(\mu - 1.96\sigma < X \le \mu + 1.96\sigma) &= 0.95 \\ P(\mu - 2.58\sigma < X \le \mu + 2.58\sigma) &= 0.99 \\ P(\mu - 3.29\sigma < X \le \mu + 3.29\sigma) &= 0.999 \end{aligned}$$

Ex: $\alpha = 0.05,\ \alpha/2 = 0.025,\ \lambda_{\alpha/2} = 1.96,\ 1 - \alpha = 0.95$

Låt $X \in N(3,1) \implies P(1 < X \le 5) \approx 95\%$

Summa och differens av oberoende normalfördelade s.v. med olika μ och σ

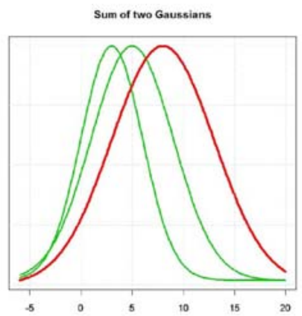
Om $X \in N(\mu_X, \sigma_X)$ och $Y \in N(\mu_Y, \sigma_Y)$

$$\Longrightarrow$$

$$\begin{aligned} X + Y &\in N\left(\mu_X + \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right) \\ X - Y &\in N\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2 + \sigma_Y^2}\right) \end{aligned}$$

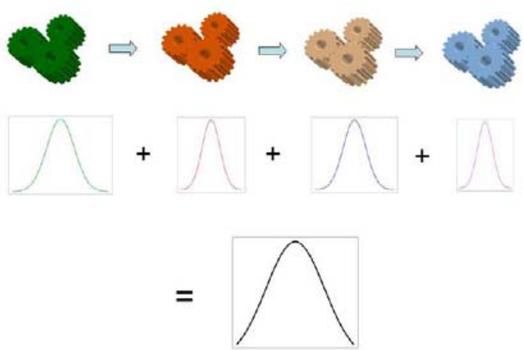
Summor (differenser) av oberoende N är också N, även om ingående komponenter har olika väntevärden och varianser.

Summa och differens av oberoende normalfördelade s.v. med olika μ och σ



Figur: $\mu_1 = 3;\ \mu_2 = 5;\ \mu = 8;\ \sigma_1 = 3;\ \sigma_2 = 4;\ \sigma = 5$

Summa och differens av flera oberoende normalfördelade s.v.



Summa och differens av flera oberoende normalfördelade s.v.

$$X_i \in N(\mu_i, \sigma_i)$$

$$\Longrightarrow$$

$$\sum_1^n a_i X_i + b \in N\left(\sum_1^n a_i\mu_i + b, \sqrt{\sum_1^n a_i^2\sigma_i^2}\right)$$

Medelvärdet är en slumpvariabel !

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\cdot(X_1 + X_2 + \ldots + X_n)$$

Slumpvariabel

Slumpvariabler

Försök	X_i	medelvärde
1	10, 12, 11, 10, 9, 11, 11, 13, 9, 10	10,6
2	10, 13, 11, 9, 9, 11, 10, 13, 9, 14	10,9
3	12, 12, 13, 9, 10, 11, 12, 10, 11, 10	11,0
...	...	mindre spridning

$s_1 = 1.26 \quad s_2 = 1.85 \quad s_3 = 1.25 \qquad s_{\bar{x}} = 0.21$

Summa och medelvärde av flera oberoende normalfördelade s.v. med samma μ och σ

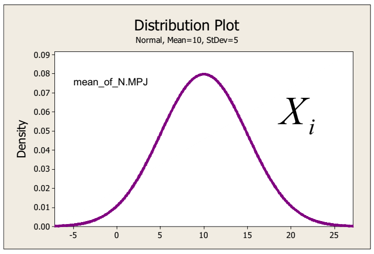
Låt $E(X_i) = \mu$ och $D(X_i) = \sigma$.

Om $Y = \sum_{i=1}^{n} X_i$ så gäller $Y \in N(n\mu, \sigma\sqrt{n})$

Om $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ så gäller $\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

OBS!!: Vi förutsätter här att komponenterna $X_1, X_2, \ldots X_n$ är normalfördelade (satsen gäller exakt).

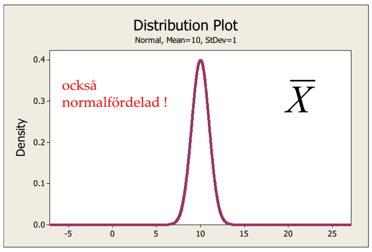
Flera N-fördelade variabler – fördelning av deras medelvärde



$$X_i \in N(\mu, \sigma)$$

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

$$\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$



Medelvärdet har en mindre spridning: det är just därför vi tar medelvärdet när vi ska mäta någonting!

Exempel: Fördelning för X_i

Exempel: Ett mätfel X är fördelad $N(0, 10)$.

Hur stor är slh. att felet är begränsad till $(-3, 3)$??

$$\begin{aligned} P(-3 < X < 3) &= F_X(3) - F_X(-3) \\ &= \Phi\left(\frac{3-0}{10}\right) - \Phi\left(\frac{-3-0}{10}\right) \\ &= \Phi(0.3) - \Phi(-0.3) = 2\cdot\Phi(0.3) - 1 \approx 0.236 \end{aligned}$$

Sannolikheten är 23% att felet hållas i gränserna mellan -3 och 3

Exempel: Fördelning för X

Exempel: Ett mätfel X är fördelad $N(0, 10)$.

Hur stor är slh. att felets medelvärde (över 25 mätningar) är begränsad till $(-3, 3)$??

$$X \in N(0, 10) \;\Rightarrow\; \bar{X} \in N\left(0, \frac{10}{\sqrt{25}}\right) = N(0, 2)$$

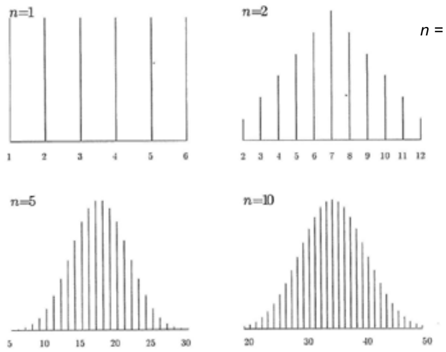
$$\begin{aligned} P(-3 < \bar{X} < 3) &= F_{\bar{X}}(3) - F_{\bar{X}}(-3) \\ &= \Phi\left(\frac{3-0}{2}\right) - \Phi\left(\frac{-3-0}{2}\right) \\ &= \Phi(1.5) - \Phi(-1.5) = 2\Phi(1.5) - 1 \approx 0.866 \end{aligned}$$

Sannolikheten är 87% att felets medelvärde hållas i gränserna mellan -3 och 3. Mycket bättre!

Centrala gränsvärdessatsen

- Vi har sett att summor av normalfördelade variabler är också normalfördelade
- **CGS:** även summor av *godtyckligt* fördelade slumpvariabler är (ungefär) normalfördelade, *bara de är*
  - många
  - oberoende
  - likafördelade

Nytt experiment: kasta flera tärningar och beräkna summan av ögontalet



Sammanfattning

Komponenterna måste vara oberoende och likafördelade (samma μ och σ).

	normalfördelad	godtycklig fördelad
$Y = \sum_{i=1}^{n} X_i$	$Y \in N\left(n\mu, \sigma\sqrt{n}\right)$ $\forall n$	$Y \in N\left(n\mu, \sigma\sqrt{n}\right)$ $n \to \infty$
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ $\forall n$	$\boxed{\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)\ n \to \infty}$
	exakt	asymptotiskt