Sannolikhet och statistik

Normalfördelningen

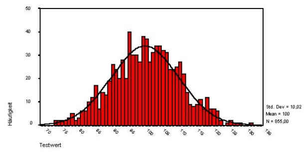
HT 2008

Uwe.Menzel@math.uu.se

http://www.math.uu.se/∼uwe/

Normalfördelningens betydelse

Empiriskt(se figur): många storheter approximativt normalfördelade
Summan av många (ungefär) oberoende och (ungefär) likafördelade s.v. är approximativt normalfördelad (CGS)
Exempel: mätfel = summa av många oberoende felkällor som är av samma storleksordning



Figur: Approximativt normalfördelad

Figur: 10 D-Mark

Täthetsfunktion, väntevärde och varians för $N(\mu, \sigma)$

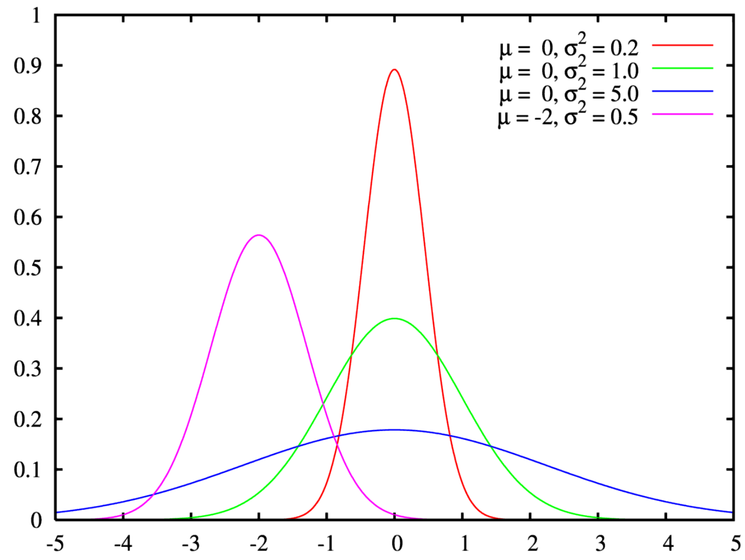
$$\boxed{f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \qquad -\infty < x < \infty}$$

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

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

$$D(X) = \sigma$$

Täthetsfunktion för $N(\mu, \sigma)$ för olika μ och σ



Figur: Normalfördelningen

Fördelningsfunktion för $N(\mu, \sigma)$

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

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

Tyvärr ingen analytisk lösning

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

$$\Longleftrightarrow$$

$$(X-\mu)/\sigma \in N(0,1)$$

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

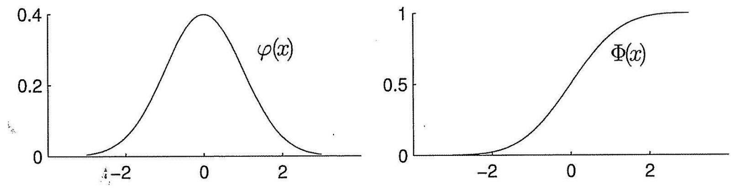
Låt $Y = (X-\mu)/\sigma$. Vilken fördelningsfunktion har Y ?

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

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

Fördelningsfunktionen för $N(0,1)$ kallas $\Phi(x)$

$$\begin{aligned} \Phi_X(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t)^2/2}\, dt \\ \varphi_X(x) &= \frac{1}{\sqrt{2\pi}} e^{-(x)^2/2} \end{aligned}$$



Figur: Standardiserad normalfördelning

Beräkning av $F_X(x)$ för godtyckliga x, μ och σ:

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

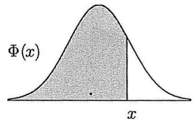
$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \\ &= P\left(Y \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Vi behöver Φ helst för hela $\mathbb{R}$

Standardiserad normalfördelning: Tabell över $\Phi(x)$

Blom s. 397

Tabell 1. Standard normalfördelning

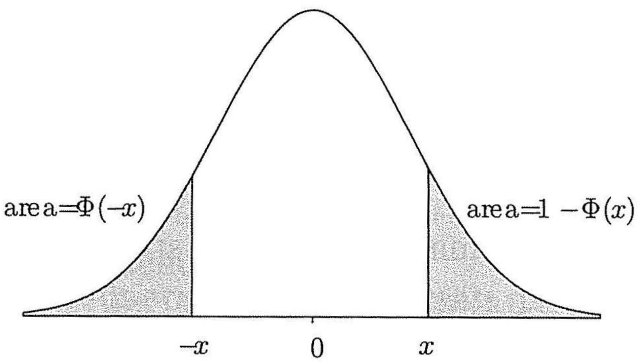


$\Phi(x) = P(X \leq x)$, där $X \in N(0,1)$.
För negativa x, utnyttja att $\Phi(-x) = 1 - \Phi(x)$

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
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
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

$\Phi(x)$ för negativa argument

$$\begin{aligned} \varphi(-x) &= \varphi(x) \\ \Phi(-x) &= 1 - \Phi(x) \end{aligned}$$



Figur: $N(0,1)$

Beräkning av slh. för allmän normalfördelning

$$\boxed{X \in N(\mu, \sigma) \Rightarrow F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)}$$

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

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

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

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

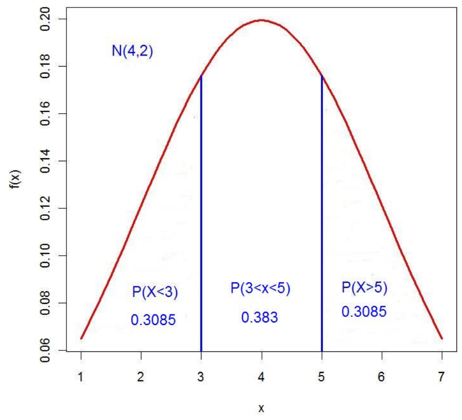
Sannolikheten att X ligger mellan två värden

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

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

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

Sammanfattning av alla 3 exempel



Figur: $P(X < 3) + P(3 \leq X < 5) + P(X \geq 5) = 1$

Sannolikhet inom $\mu \pm \sigma, \mu \pm 2\sigma, \mu \pm 3\sigma$

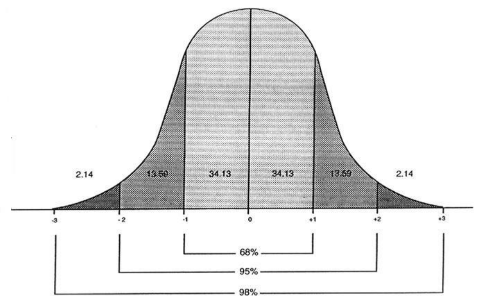
$$P(a < X \leq b) = F_X(b) - F_X(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

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

Samma med 1 och 3:

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

Sannolikhet inom $\mu \pm \sigma, \mu \pm 2\sigma, \mu \pm 3\sigma$

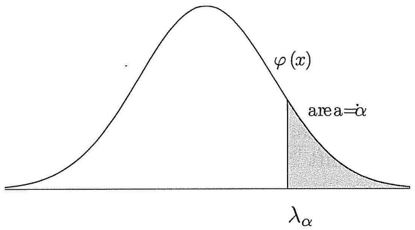


Figur: $k \cdot \sigma$ intervaller

Kvantiler för den standardiserade normalfördelningen

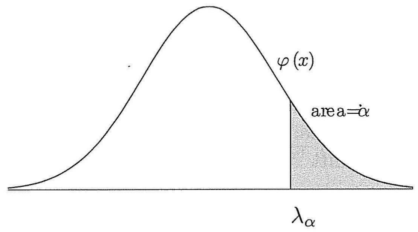
Definition kvantil λ_α:

$$P(X > \lambda_\alpha) = \alpha \quad \text{eller} \quad \Phi(\lambda_\alpha) = 1 - \alpha$$



Figur: Kvantiler för $N(0,1)$

Kvantiler för den standardiserade normalfördelningen



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

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

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

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

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

Sätt $\alpha = 0.05,\ 0.001,\ 0.0001$

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

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

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

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

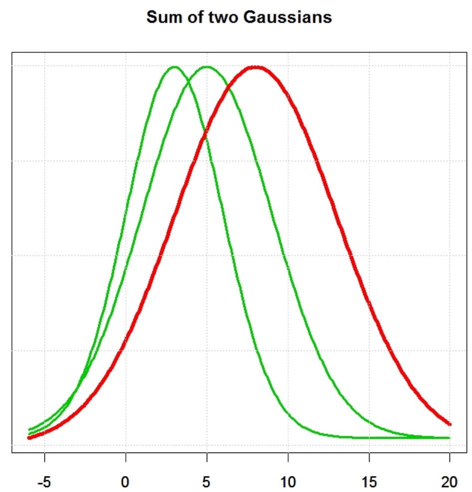
$$\text{Om} \quad X \in N(\mu_X, \sigma_X) \quad \text{och} \quad Y \in N(\mu_Y, \sigma_Y)$$

$$\Longrightarrow$$

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

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

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



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

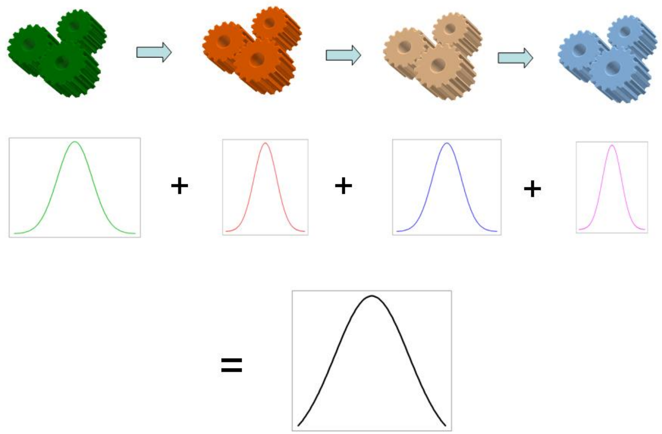
Summa av flera normalfördelade s.v. med olika μ och σ

X_i oberoende

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

$$\Longrightarrow$$

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



Figur: $t = t_1 + t_2 + t_3 + t_4$

Fördelning för summan och medelvärdet av flera normalfördelade s.v. med samma μ och σ

Summan av oberoende, normalfördelade s.v. med samma väntevärde och standardavvikelse:

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

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

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

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

Exempel: Fördelning för $\bar{X}$

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 absolutbeloppet av felet är ≤ 3.

Exempel: Fortsättning

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

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

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

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

Sannolikheten är 87% att absolutbeloppet av felets medelvärde blir mindre än 3. Mycket bättre!

Differens av två medelvärden från N med olika μ, σ, n

Vi har sett hur medelvärdet av N's är fördelad, och vi har sett hur summan (differensen) av N's är fördelad.

Om man kombinerar dessa formler, ser man t. ex. hur en **differens av medelvärden** är fördelad:

Låt $\bar{X}$ och $\bar{Y}$ vara två sådana medelvärden:

$$\bar{X} \in N\left(\mu_X, \frac{\sigma_X}{\sqrt{n_X}}\right) \quad \text{och} \quad \bar{Y} \in N\left(\mu_Y, \frac{\sigma_Y}{\sqrt{n_Y}}\right)$$

$$\Longrightarrow$$

$$\bar{X} - \bar{Y} \in N\left(\mu_X - \mu_Y, \sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}\right)$$

Centrala gränsvärdessatsen

Vi har sett att summor av oberoende normalfördelade s.v. är också normalfördelade.

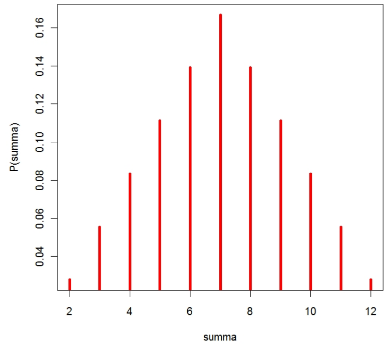
Även summor av godtycklig fördelade, oberoende och likafördelade slumpvariabler är (i regel) *ungefär* normalfördelad, bara antalet komponenter i summan är tillräckligt stort.

$$\Longrightarrow$$

Normalfördelningen har ett stort användningsområde

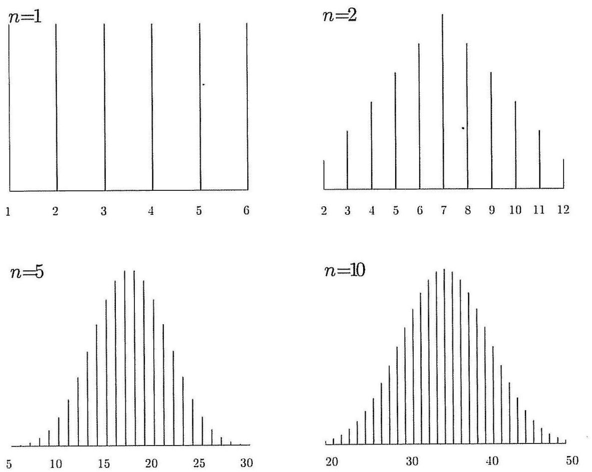
Centrala gränsvärdessatsen

Kom ihåg vart experiment där vi kastade 2 tärningar:



Figur: Poängsumma av tärningskast, n = 2

Nytt experiment: Vi kastar n tärningar



Figur: Poängsumma av tärningskast, n = antalet tärningar

Centrala gränsvärdessatsen

Summan av oberoende, likafördelade s.v. är approximativt normalfördelad om antalet komponenter är tillräckligt stort.

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

$$\text{Om} \quad Y = \sum_{i=1}^{n} X_i \quad \text{så gäller} \quad Y \in AsN\left(n\mu, \sigma\sqrt{n}\right)$$

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

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

Centrala gränsvärdessatsen

$$\boxed{Y = \sum_{i=1}^{n} X_i \implies Y \in AsN\left(n\mu, \sigma\sqrt{n}\right)}$$

$$\frac{Y - n\mu}{\sigma\sqrt{n}} \in N\left(0, 1\right)$$

Centrala gränsvärdessatsen

Om $X_1, X_2, \ldots$ är en oändlig följd av oberoende likafördelade s.v. med väntevärdet μ och standardavvikelsen σ, så gäller för $Y_n = X_1 + X_2 \ldots X_n$ att

$$P\left(a < \frac{Y_n - n\mu}{\sigma\sqrt{n}} \leq b\right) \rightarrow \Phi\left(b\right) - \Phi\left(a\right) \quad \text{om} \quad n \rightarrow \infty$$

I ex. *"mätfel"* var vi alltså inte tvungna att förutsätta $X_i \in N$

Sammanfattning summa och medelvärde

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

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