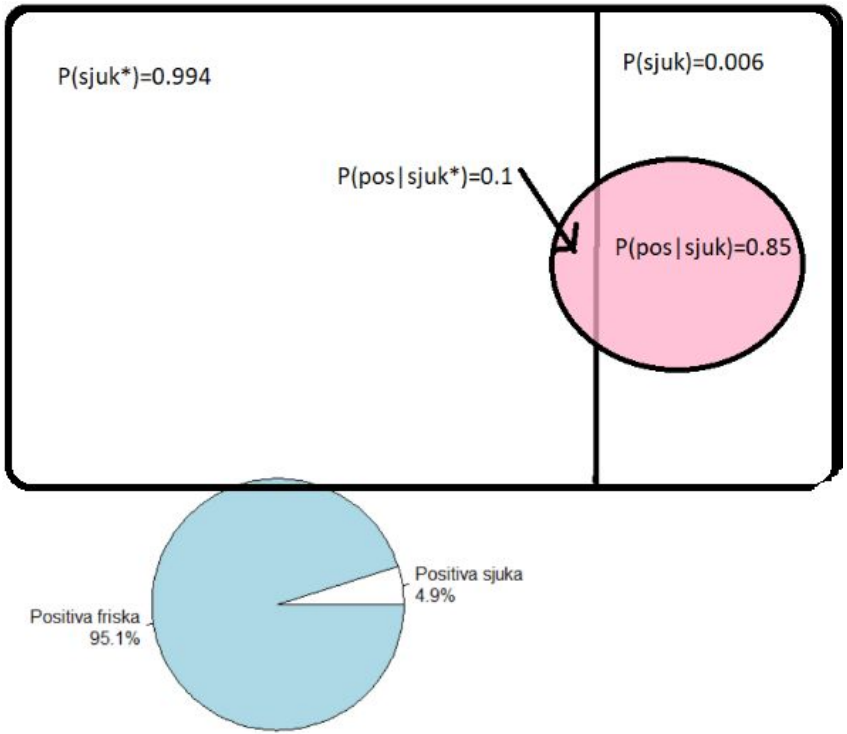
Summor och linjärkombinationer, Centrala gränsvärdessatsen och sannolikhetsmodeller

Snabb repetition

Satsen om total sannolikhet och Bayes' sats (2.39)



P(pos)=P(pos|sjuk)*P(sjuk)+P(pos|sjuk*)*P(sjuk*)=

0.85*0.006+0.1*0.994=0.0051+0.0994=0.1045

P(sjuk|pos)=P(pos|sjuk)/P(pos)*P(sjuk)=

(0.85/0.1045)*0.006=8.134*0.006=0.0488

P(sjuk*|pos*)=P(pos*|sjuk*)/P(pos*)*P(sjuk*)=

1.005025*0.994=0.998995

$V(X)=E(X^2)-[E(X)]^2$; $V(X)=E(X^2)-\mu^2$

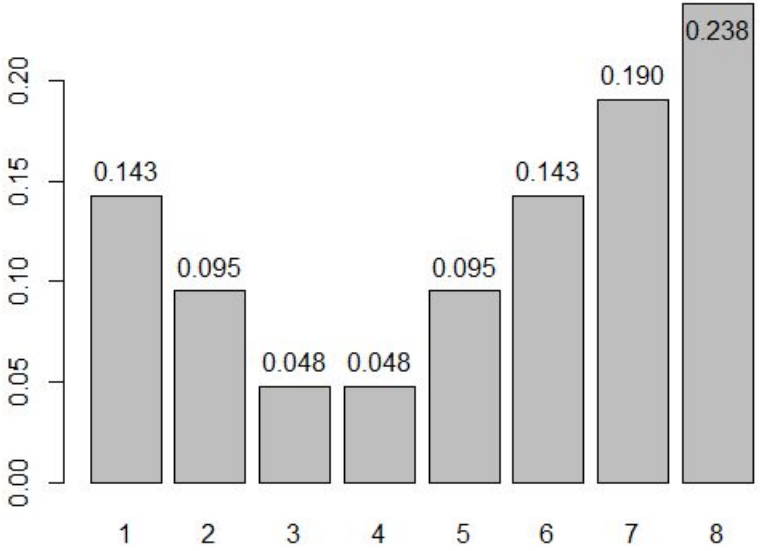
$E(X) = \mu =$

0.143*1+0.095*2+0.048*3+0.048*4+0.095*5+0.143*6+0.190*7+0.238*8=5.236

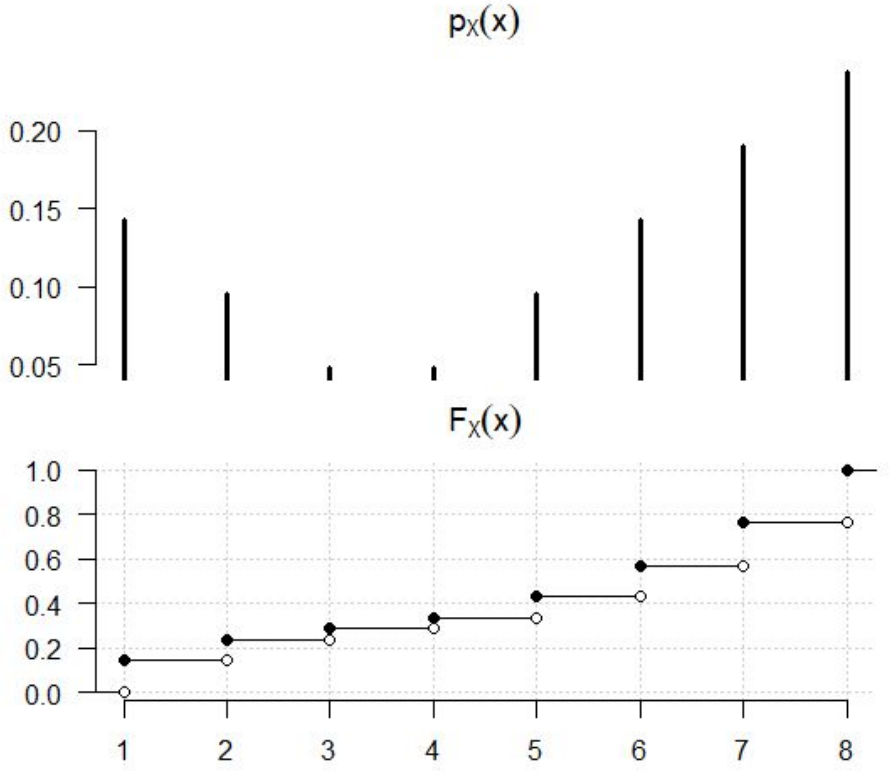
$E(X^2)$=0.143*1^2+0.095*2^2+0.048*3^2+0.048*4^2+0.095*5^2+0.143*6^2+0.190*7^2+0.238*8^2=33.788

$V(X)=E(X^2)-[E(X)]^2$=33.788-5.236^2=6.372304

D(X)=6.372304^(½)=2.52



Fördelningsfunktion



Normalfördelning

500 mm långa skruvar

Standardavvikelse

Eine Maschine produziert 500mm lange Schrauben mit einer Standardabweichung von 10mm. Die Länge der Schrauben kann als normalverteilt angesehen werden.

kortare

a) Berechne die Wahrscheinlichkeit dafür, dass eine Schraube kürzer ist als 485 mm.

Normalverteilung und Tafelwerk der Stochastik

Die Länge der Schrauben ist normalverteilt mit Erwartungswert $\mu = 500$ und Standardabweichung $\sigma = 10$. Gesucht ist die Wahrscheinlichkeit dafür, dass eine Schraube kürzer ist als 485 mm, also $P(X \leq 485)$.

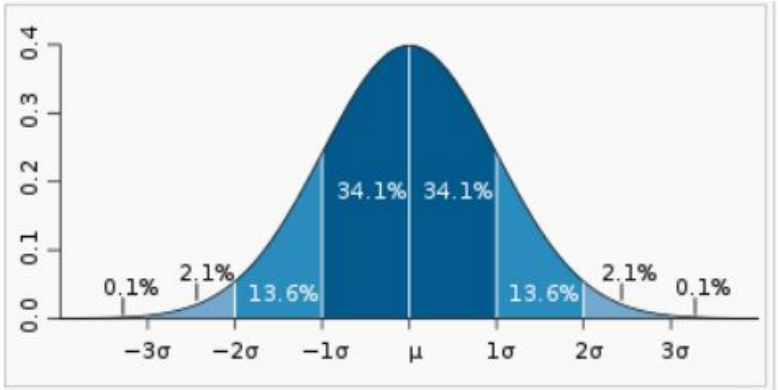
$P(X \leq k) \approx \Phi\left(\frac{k-\mu}{\sigma}\right)$ Setz die Werte ein.

$P(X \leq 485) \approx \Phi\left(\frac{485-500}{10}\right)$ Vereinfache.

$= \Phi(-1,5) = 1 - \Phi(1,5)$ Lies den Wert im Tafelwerk der Stochastik ab.

$\approx 1 - 0,93319 = 0,06681$

Die Wahrscheinlichkeit dafür, dass eine Schraube kürzer als 4,85cm ist, beträgt also etwa 6,7%.



Représentation graphique d'une loi normale. Chaque bande colorée a la largeur d'un écart type.

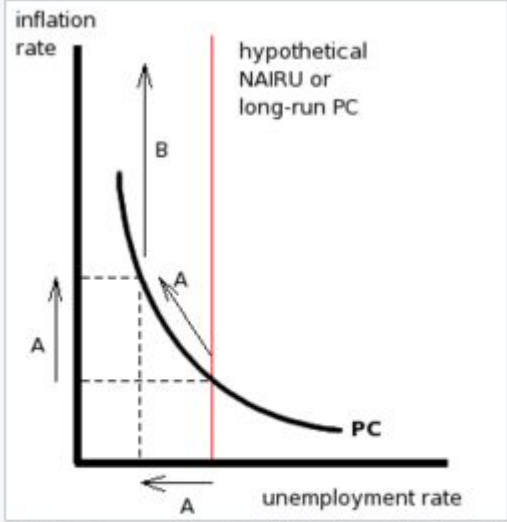
Normalfördelning

- $X \in N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \in N(0, 1)$
- $F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ där $\Phi(\cdot)$ ges av tabell
- $X_1, \ldots, X_n$ oberoende och $N(\mu_1, \sigma_1^2), \ldots, N(\mu_n, \sigma_n^2) \Rightarrow \sum_{i=1}^{n} a_i X_i \in N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$

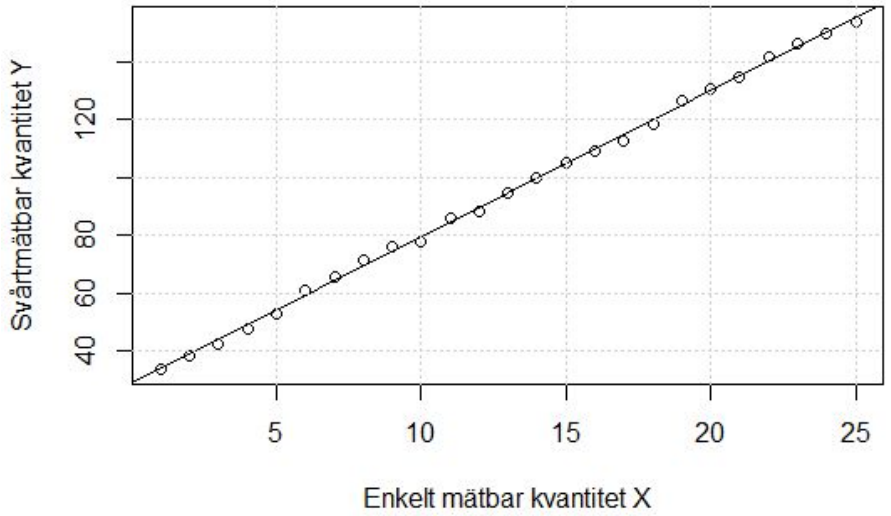
https://de.serlo.org/mathe/stochastik/wichtige-modelle-verteilungen/normalverteilung/aufgaben-normalverteilung

Dagens första tema är flera slumpvariabler

Man är naturligtvis inte bara intresserad av mätningar utan av samband mellan mätningar.



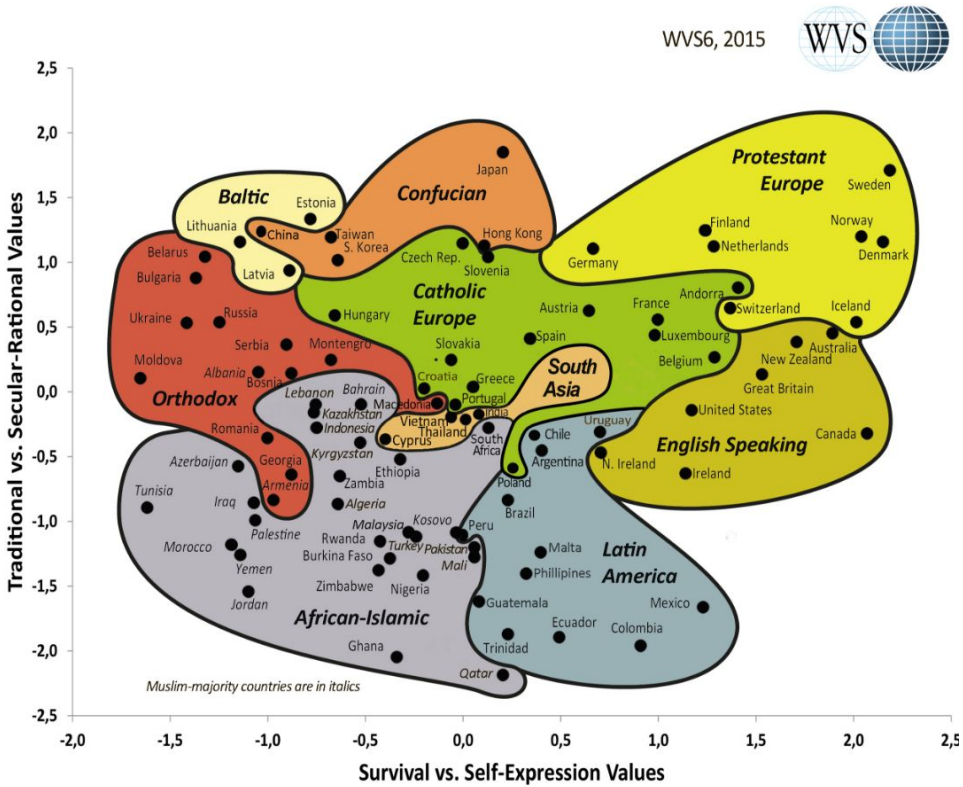
Phillipskurvan i sin klassiska form. Det råder ett negativt samband mellan inflation och arbetslöshet.



World value survey

Egentligen har ”World values survey” många olika dimensioner, som kokas ihop till två med hjälp av lite linjär algebra.

Riktigt så långt kommer vi inte här.



Intressanta frågor

1) Hur samvarierar X och Y (och Z,W,U,V,...)?
2) Hur beror Y av X (och Z,W,U,V,...)?
3) Vad händer när vi kombinerar ihop flera slumpvariabler exempel BMI = V/L^2, där V är vikten i kilo och L längden i meter.

1) Detta är kommer vi att tala om denna timme.
2) Detta är ämne för kommande föreläsningar om linjär regression och faktorförsök.
3) Fallet då kombinationen är en summa behandlas här. Mer komplicerade samband är föremålet för nästa föreläsning.

Samvariation och addition av flera slumpvariabler

Avsnittets mål

- Korrelation är ett mått på beroende, men inte allt beroende
- Regler för väntevärde och varians av linjärkombinationer:

$E(a+bX+cY)=a+bE(X)+c(E(Y)$

$V(bX+cY)=b^2E(X)+c^2(E(Y)$

- Standardavvikelser adderar som Pytagoras' sats

$$E(S_n) = n\mu \; ; E(\overline{X}) = \mu$$

$$V(S_n) = n\mu \; ; V(\bar{X}) = \frac{\sigma}{n}$$

$$D(S_n) = \sqrt{n}\sigma \; ; D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

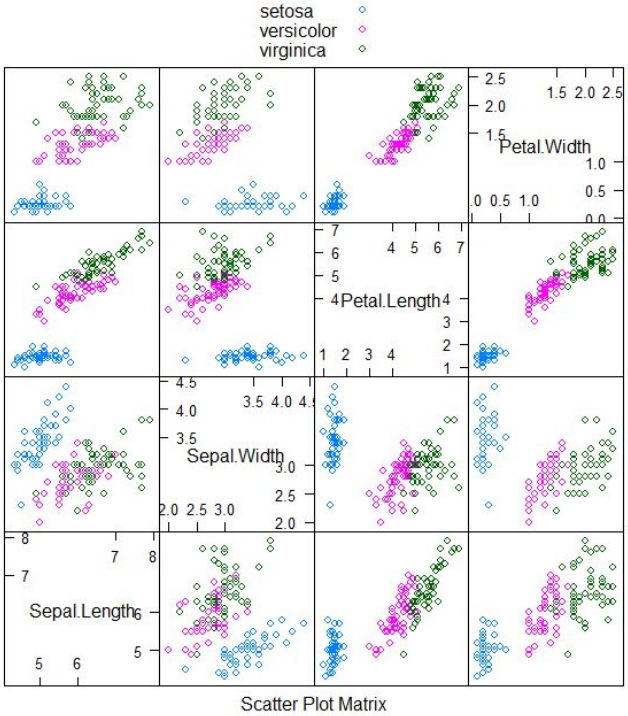
Om de ingående variablerna är oberoende och likafördelade (iid).

Tuviris (iris setosa)

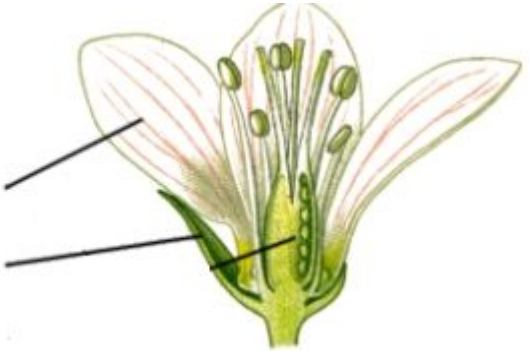
Brokiris (isis versicolor)

Iris virginica

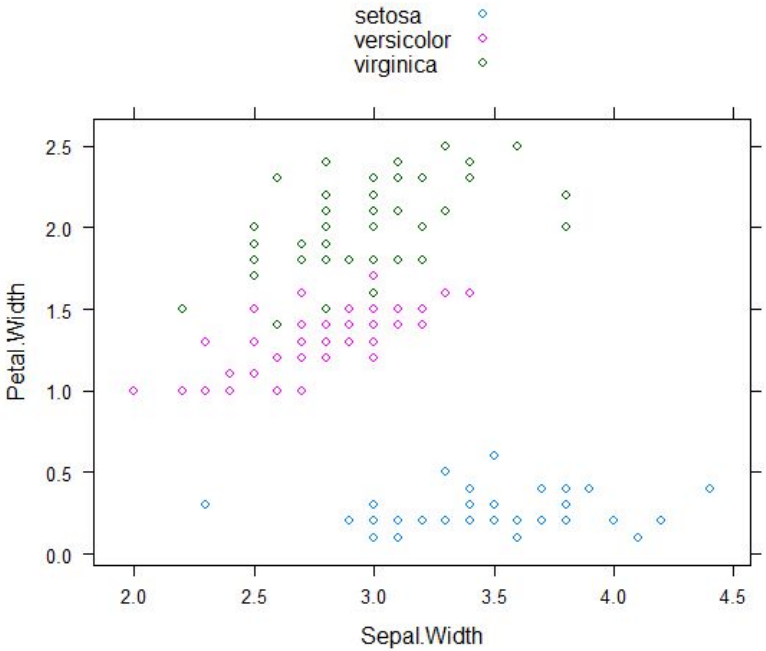
Fyra mått: bredd och längd av kron- och foderplad



"petal"
kronblad
foderblad
"sepal"



Jämför! Korrelationskoeficient: $-1 \leq \rho \leq 1$.

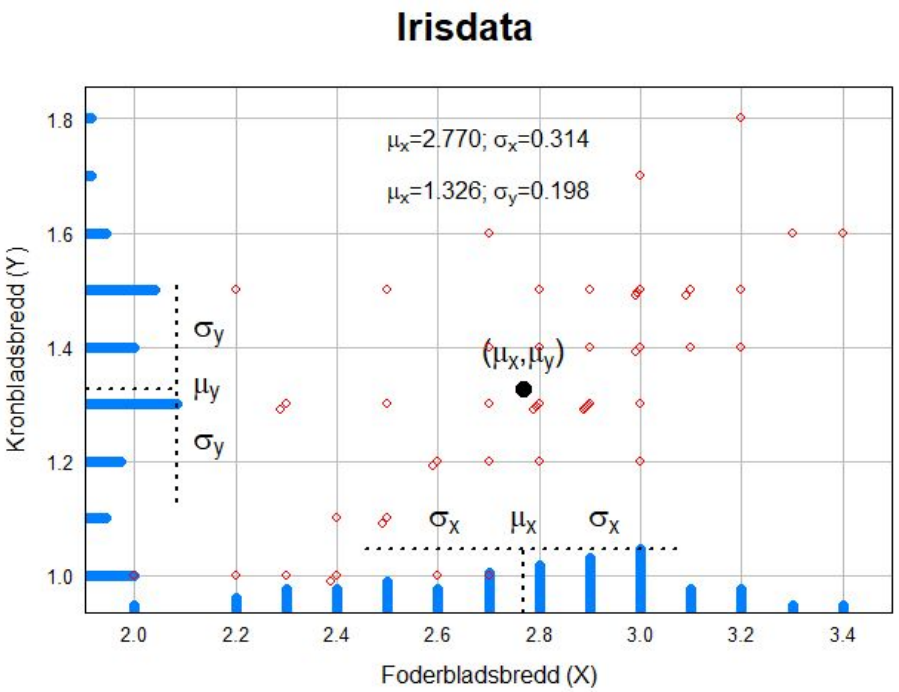


Functionen `cor` i R ger

```
$setosa

[1] 0.232752

$versicolor

[1] 0.6639987

$virginica

[1] 0.537728
```

För att se vad som menas ser vi på brokiris

(För resonemangets skull låtsas jag att detta är en fördelning och inte ett stickprov.)



X och Y har egna, ”marginella”, fördelningar.

$p_X(2)=0.02$

$p_X(2.2)=0.04$

$p_X(2.3)=0.06$

$p_X(2.4)=0.06$

$p_X(2.5)=0.08$

$p_X(2.6)=0.06$

$p_X(2.7)=0.10$

$p_X(2.8)=0.12$

$p_X(2.9)=0.14$

$p_X(3)=0.16$

$p_X(3.1)=0.06$

$p_X(3.2)=0.06$

$p_X(3.3)=0.02$

$p_X(3.4)=0.02$

$p_Y(1)=0.14$

$p_Y(1.1)=0.06$

$p_Y(1.2)=0.10$

$p_Y(1.3)=0.26$

$p_Y(1.4)=0.14$

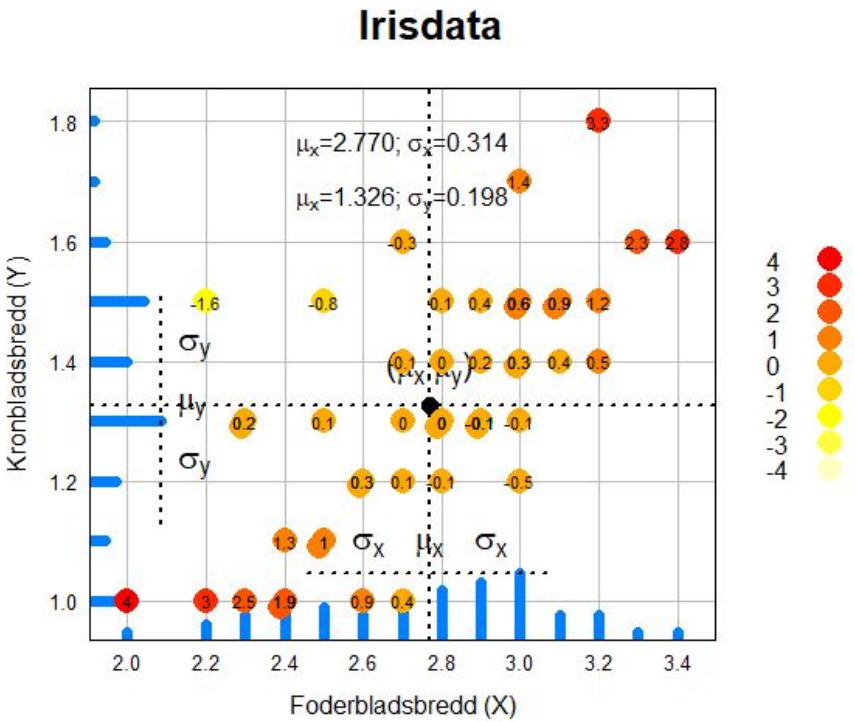
$p_Y(1.5)=0.20$

$p_Y(1.6)=0.06$

$p_Y(1.7)=0.02$

$p_Y(1.8)=0.02$

Korrelation

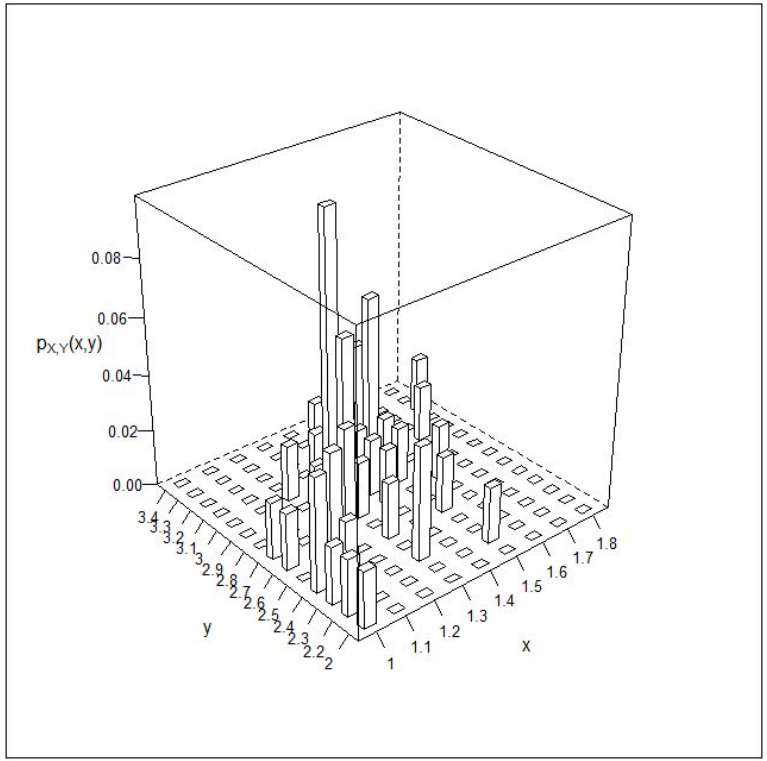


Ju större termen $\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}$ är, desto mer tyder det på en positiv samvariation mellan x och y. Negativa värden antyder ett negativt samband. Ett mått på det hur stort samband det är totalt är genomsnittet:

$$\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sigma_y} = \frac{1}{\sigma_x\sigma_y}\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y).$$

När man skattar ur data delar man med n-1, inte med n, för att kompensera för osäkerheten i uppskattningarna.

Korrelationen är ett mått på samvariation



”Simultan sannolikhetsfunktion” $p_{XY}(x,y)$

$$C(X,Y) = \sum_{x,y} p_{XY}(x,y)(x-\mu_X)(y-\mu_Y)$$

Kovariansen: C(X,Y) = E[(X-μ_x)(Y-μ_y)]

p(1,2)*(1-μ_x)(2-μ_y)+p(1,2.2)*(1-μ_x)(2.2-μ_y)+...+p(1.8,3.2)*(1.8-μ_x)(3.2-μ_y)= 0.0412

Den simultana sannolikhetsfunktionen

```
> threeD
           Sepal.Width
Petal.Width    2  2.2  2.3  2.4  2.5  2.6  2.7  2.8  2.9    3  3.1  3.2  3.3  3.4
        1   0.02 0.02 0.02 0.04 0.00 0.02 0.02 0.00 0.00 0.00 0.00 0.00 0.00 0.00
        1.1 0.00 0.00 0.00 0.02 0.04 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
        1.2 0.00 0.00 0.00 0.00 0.00 0.04 0.02 0.02 0.00 0.02 0.00 0.00 0.00 0.00
        1.3 0.00 0.00 0.04 0.00 0.02 0.00 0.02 0.06 0.10 0.02 0.00 0.00 0.00 0.00
        1.4 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.02 0.02 0.04 0.02 0.02 0.00 0.00
        1.5 0.00 0.02 0.00 0.00 0.02 0.00 0.00 0.02 0.02 0.06 0.04 0.02 0.00 0.00
        1.6 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.00 0.02 0.02
        1.7 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00
        1.8 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.00
```

ϱ=C(X,Y)/D(X)/D(Y)= 0.0412/.314/.198=

0.664

Enkel ”momentalgebra”

3.72 Beräkna vv och varians för olika komb av X1, X2 och X3

En teknisk manick byggs av en del som varierar i längd enligt $X \sim N(4.5, 1^2)$ och en som som oberoende av denna varierar som $Y \sim N(15, 2^2)$. Vad är väntevärde och standardavvikelse för den totala längden $Z = X+Y$?

Lösning: $E(Z)=E(X+Y)=E(X)+E(Y)=4.5+15=19.5$

$V(Z)=V(X+Y)=[\text{oberoende}]=1+2^2=5$. $D(Z)=5^{1/2}\sim 2.24$.

Lägg märke till att dubbel standardavvikelse gör att bidrag Ys helt dominerar.

$D(X+Y)^2=D(X)^2+D(Y)^2$

Väntevärden är väldigt lätta:

$E(aX+bY)=aE(X)+bE(Y)$

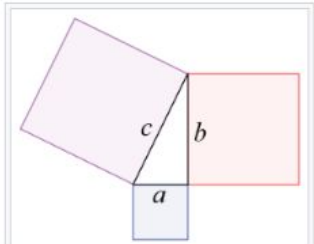
Standardavvikelser är lite knepigare, men Pythagoras' sats ger en minnesregel.

$V(X+Y)=V(X)+V(Y)$.

“”Räkna med varianser; tolka standardavvikelser!”

Gauss's Pythagorean right triangle proposal is an idea attributed to Carl Friedrich Gauss for a method to signal extraterrestrial beings by constructing an immense right triangle and three squares on the surface of the Earth. The e Pythagorean theorem,

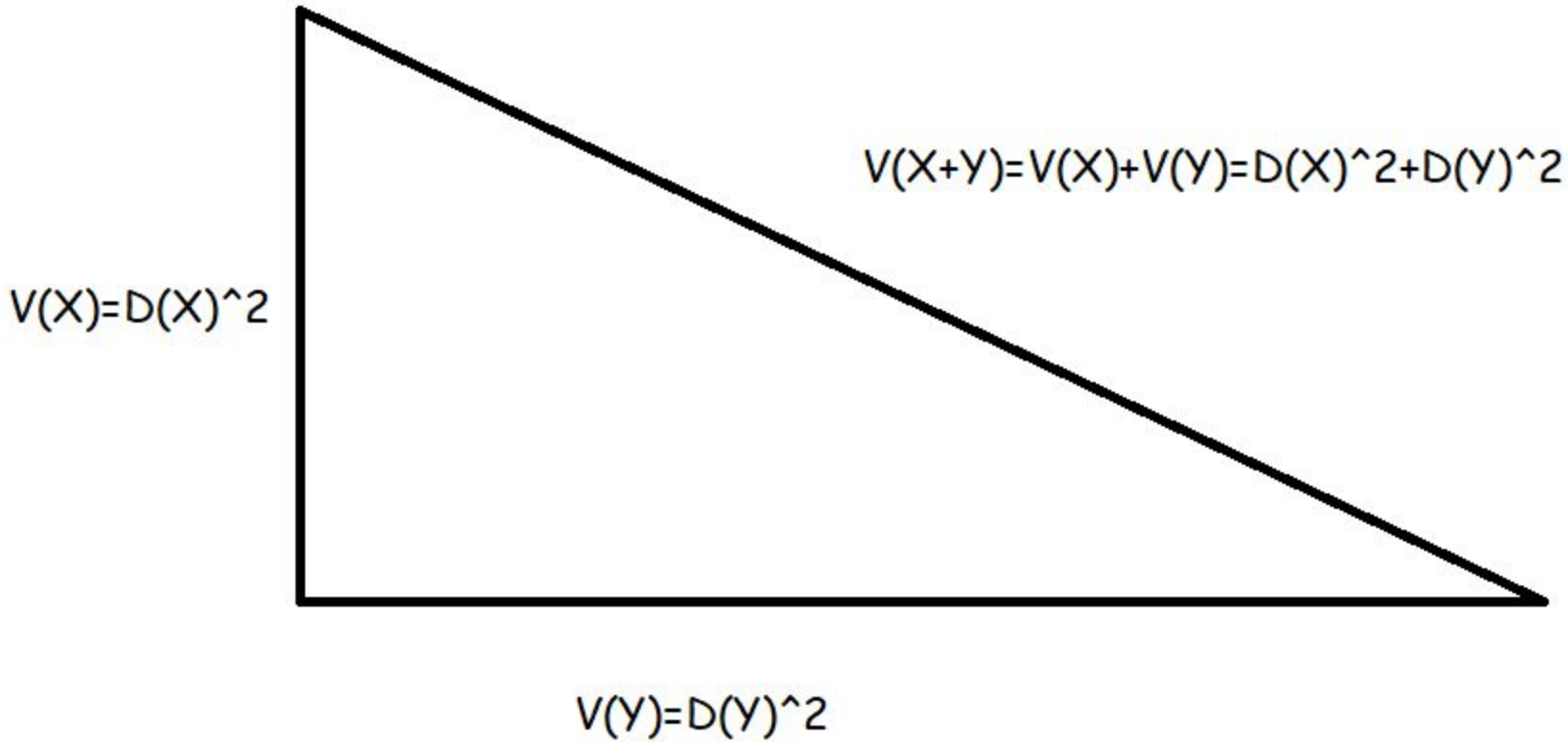
ing with Gauss, with exact ail, and even whether Gauss . Many of the earliest itor, instead crediting a escriptors, and in some s of the proposal also rtheless, Gauss's writings ontact extraterrestrial life, d light using a heliotrope, his s of the Moon.



Visual representation of the Pythagorean theorem. Under the proposal the shape seen here would be drawn at vast size on the Siberian tundra using pine trees and fields of wheat.

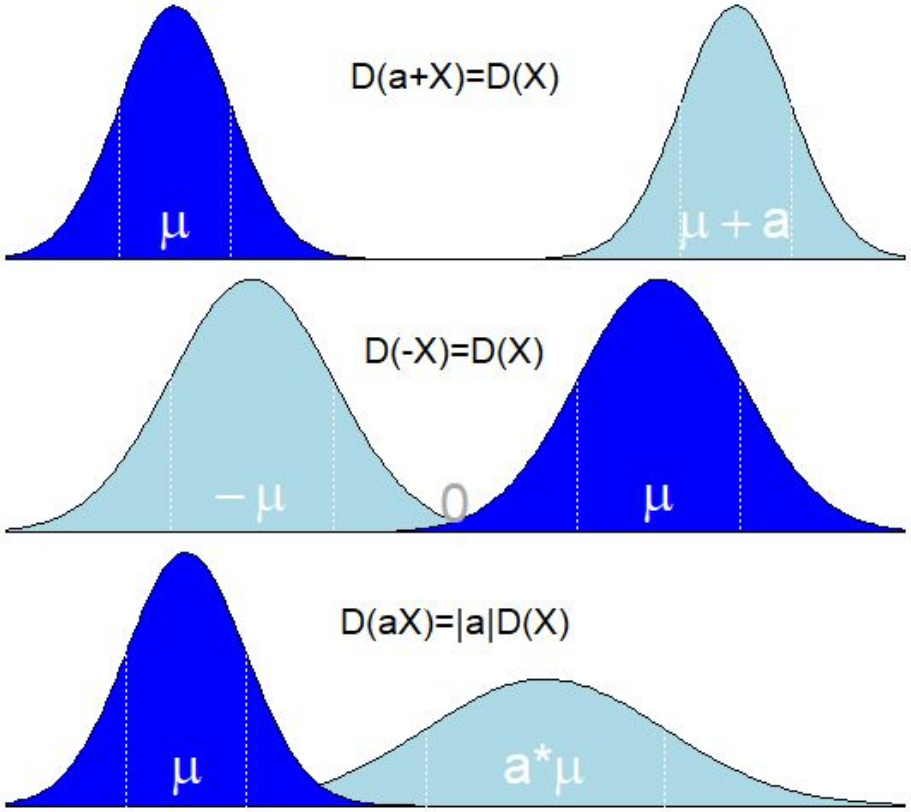
Proposal [edit]

Carl Friedrich Gauss is credited with an 1820 proposal[1] for a method to signal extraterrestrial beings in the form of drawing an immense right triangle and three squares on the surface of the Earth, intended as a symbolical representation of the Pythagorean theorem, large enough to be seen from the Moon or Mars. Details vary between sources, but typically the "drawing" was to be constructed on the Siberian tundra, and made up of vast strips of pine forest forming the right triangle's borders, with the interior of the drawing and exterior squares composed of fields of wheat.[2] Gauss is said to have been convinced that Mars harbored intelligent life and that this geometric figure, invoking the Pythagorean theorem through the squares on the outside borders[3] (sometimes called a "windmill diagram", as originated by Euclid),[4] would demonstrate to such alien observers the reciprocal existence of intelligent life on Earth and its grounding in mathematics.[5] Wheat was said to be chosen by Gauss for contrast with the pine tree borders "because of its uniform color".[6]


V(X+Y)=V(X)+V(Y)=D(X)^2+D(Y)^2
V(X)=D(X)^2
V(Y)=D(Y)^2

Regler vid addition och multiplikation med konstant

- D(a+X)=D(X). Ingenting händer med spridningen när en konstant läggs till.
- D(-X)=D(X). Spegling i noll påverkar inte spridningen.
- D(aX)=|a|D(X). När alla tal observationer multipliceras med a, hänger spridningen med.
- V(a+X)=V(X)
- $V(aX)=a^2V(X)$
- $V(a+bX+cY)=b^2V(X)+c^2V(Y)$



Exempel

Exempel: Antag att E(X)=3, D(X)=2;E(Y)=-2, D(Y)=4

Beräkna väntevärde och standardavvikelse för 2X+1.5Y+5!

E(2X+1.5Y+5)=2E(X)+1.5E(Y)+5=2*3+1.5*(-2)+5=8

V(2X+1.5Y+5)=4*V(X)+2.25*V(Y)=4*4+2.25*16=52

D((2X+1.5Y+5)=$44^{1/2}$= 7.21

Lite justerad version av uppgift 3.74

En nioåring och en tolvåring, båda slumpvis valda, ska bära en tiokilos låda över en bro som tål en vikt på 85 kg. Nioåringars vikt kan antas vara normalfördelad med väntevärde 30 kilo och standardavvikelse 3 kilo. Tolvåringars vikt antas normalfördelad med väntevärde 40 och standardavvikelse 4. Vad är sannolikheten att bron brister?

Lösning: Låt $X \sim N(30,3)$ vara nioåringens vikt och $Y \sim N(40,4)$ vara tolvåringens. Bilda $Z=10+X+Y$.

Vi är intresserade av $P(Z>85)$.

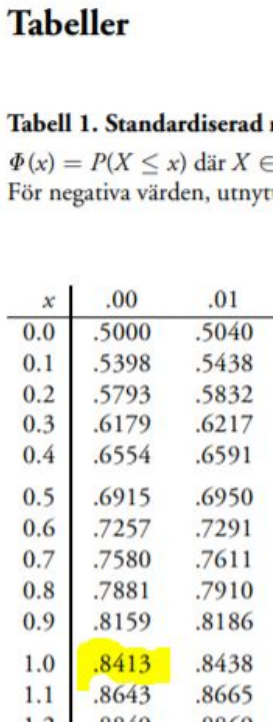
X+Y är normalfördelad, om X och Y är oberoende och normalfördelade

$E(Z)=E(10+X+Y) = 10+E(X)+E(Y)=10+30+40=80$

$V(Z)=V(10+X+Y)=V(X)+V(Y)=3^2+4^2=25$

$D(Z)=25^{1/2} =5.$

$P(Z>85)=1-P(Z<85)=1-\Phi((85-80)/5)=1-\Phi(1)=1-0.8413=0.1587 = 15.9\ \%.$



Tabeller

Tabell 1. Standardiserad

$\Phi(x) = P(X \leq x)$ där $X \in$

För negativa värden, utnyt

x	.00	.01
0.0	.5000	.5040
0.1	.5398	.5438
0.2	.5793	.5832
0.3	.6179	.6217
0.4	.6554	.6591
0.5	.6915	.6950
0.6	.7257	.7291
0.7	.7580	.7611
0.8	.7881	.7910
0.9	.8159	.8186
1.0	.8413	.8438
1.1	.8643	.8665

Variation på ett tema från Hans-Uno Bengtsson (som i sin tur snodde det från något annat ställe)

- En svårartat berusad man vaknar och vinglar sträckan X_1 längs en gata. Han somnar igen.
- Så vaknar han igen och rör sig sträckan X_2 och somnar om.
- …
- Varje förflytttning kan ske högerut och vänsterut så att $E(X_i)=0$. Alla variabler är oberoende med standardavvikelse σ.
- Hur långt ifrån utgångspunkten befinner sig fylleristen efter *n* steg?

”Fysiklösning”

$S_n=X_1+...+X_n$;
I genomsnitt befinner han sig på mitten (0), men genomsnittet av S_n^2 är intressantare:
$S_n^2=X_1^2+...+X_n^2+2X_1X_2+....+2X_{n-1}X_n$.
$\langle S_n^2\rangle=\langle X_1^2\rangle+...+\langle X_n^2\rangle+2\langle X_1X_2\rangle+....+2\langle X_{n-1}X_n\rangle=\boldsymbol{\sigma}^2+...+\boldsymbol{\sigma}^2+0+0+...+0=n\boldsymbol{\sigma}^2$.

$\langle S_n^2\rangle^{1/2}=n^{1/2}\boldsymbol{\sigma}$. Ett mer statistiskt sätt att skriva samma sak: Om $X_1,...,X_n$ är oberoende och normalfördelade, så gäller:

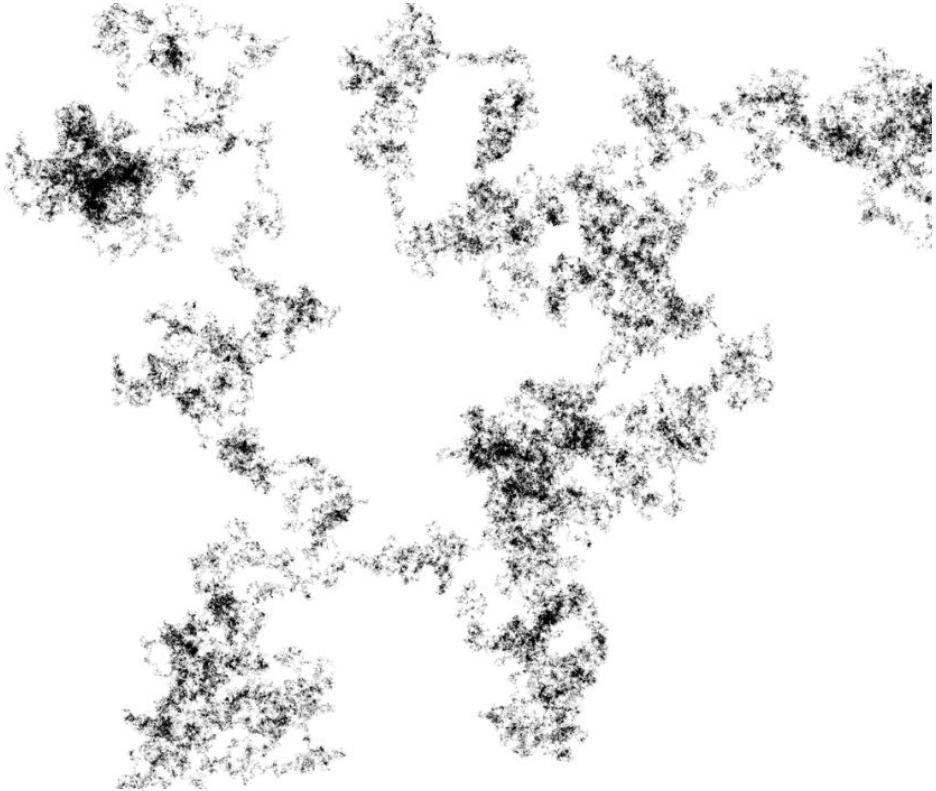
$$D(\sum_{i=1}^{n} X_i) = \sqrt{n}\sigma$$

Brownsk rörelse

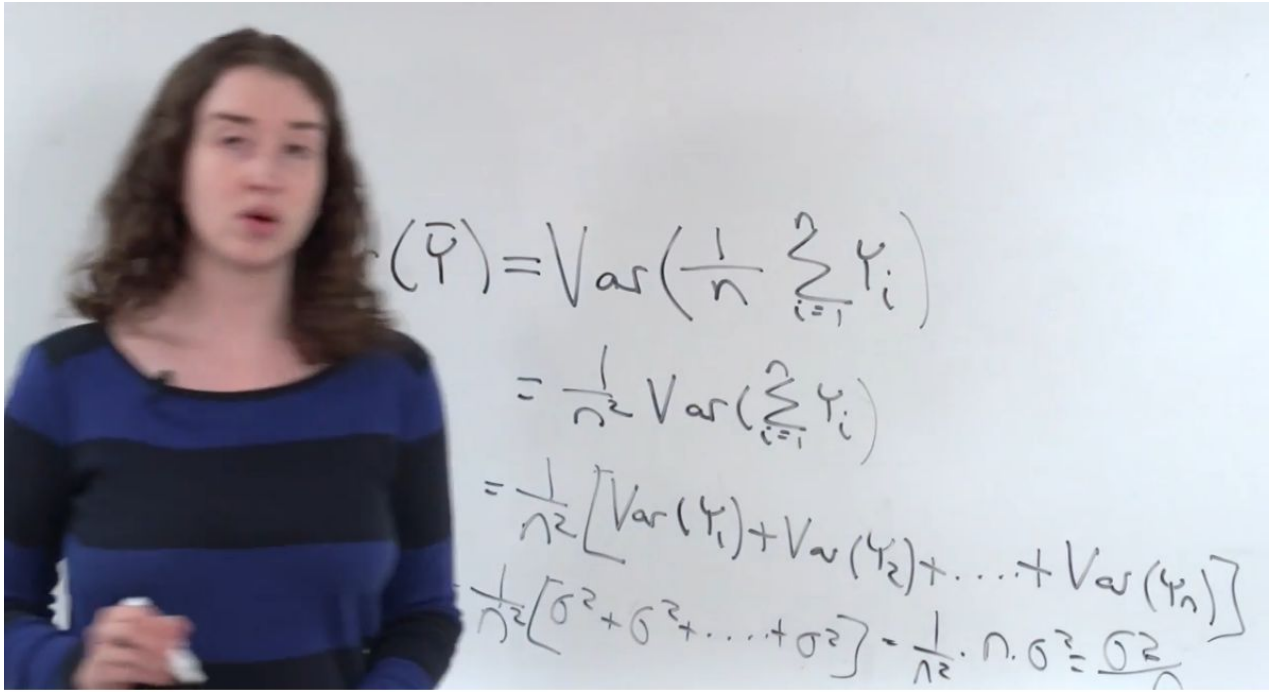
Om detta utförs i två eller tre dimensioner.

Enkelt uttryckt: När du öppnar surströmmingsburken, har dess luktmolekyler efter *n* tidsenheter rört sig så att de är inom ett avstånd som är proportionellt mot roten ur n.

Detta har uppenbara tillämpningar inom kemi och biologi. Mindre uppenbart är att det används för att modellera optionspriser.



Samma beräkningar för medelvärdet



https://www.youtube.com/watch?v=RwS_iA-xb0w

Allmänna regler för summa och medelvärde

Låt $S_n = \sum_{i=1}^{n} X_i$ och $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ för oberoende slumpvariabler med $E(X_i) = \mu$ och $V(X_i) = \sigma^2$. Då gäller

$$E(S_n) = n\mu \; ; E(\overline{X}) = \mu$$

$$V(S_n) = n\mu \; ; V(\bar{X}) = \frac{\sigma}{n}$$

$$D(S_n) = \sqrt{n}\sigma \; ; D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Avsnittets mål

- Korrelation är ett mått på beroende, men inte allt beroende
- Regler för väntevärde och varians av linjärkombinationer:

$E(a+bX+cY)=a+bE(X)+c(E(Y)$

$V(bX+cY)=b^2E(X)+c^2(E(Y)$

- Standardavvikelser adderar som Pytagoras' sats

$$E(S_n) = n\mu \; ; E(\overline{X}) = \mu$$

$$V(S_n) = n\mu \; ; V(\bar{X}) = \frac{\sigma}{n}$$

$$D(S_n) = \sqrt{n}\sigma \; ; D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Om de ingående variablerna är oberoende och likafördelade (iid).

Centrala gränsvärdessatsen

Avsnittets mål

- Centrala gränsvärdessatsen säger något som vi egentligen kände till, enligt formeln plus/minus standardavvikelsen genom roten ur antalet observationer.
- Även rätt exotiska fördelningar konvergerar, så länge som variansden existerar.

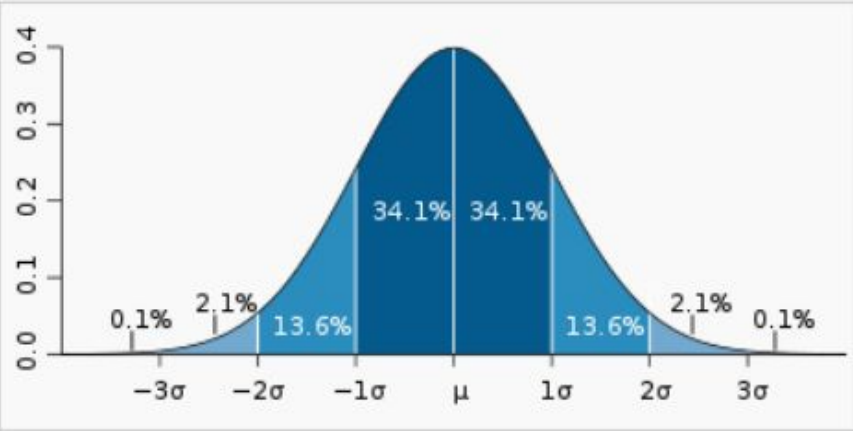
CGS (CLT) kan ni redan ha använt

$$\bar{x} \pm 1.96 \frac{SD}{\sqrt{n}},$$

där SD är standardavvikelsen.

$$D(\bar{x}) \frac{\sigma}{\sqrt{n}},$$

där σ är den teoretiska standardavvikelsen (okänd, men uppskattas av SD) och n är antalet värden.



En normalfördelad variabel med väntevärde μ kommer med sannolikhet 95 % att ligga inom plus/minus 1.96 standardavvikelser från sitt väntevärde.

Nu vet vi att standardavvikelsen av genomsnittet är $\sigma/n^{1/2}$ och väntevärdet är μ.

Centrala gränsvärdessatsen ger den saknade länken: Jo, genomsnittet är ungefär normalfördelat, varför vi kan anta att μ med 95 procents sannolikhet täcks av intervallet ovan.

Formell form

Lindeberg–Lévy CLT. Suppose $\{X_1, X_2, \ldots\}$ is a sequence of i.i.d. random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $N(0,\sigma^2)$:[3]

$$\sqrt{n}\left(S_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right).$$

Mindre formellt: Om du lägger ihop ett någorlunda stort antal någorlunda väluppfostrade, oberoende, slumpvariabler, så kan är resultatet ungefär normalfördelat.

Medelvärden tenderar att centreras

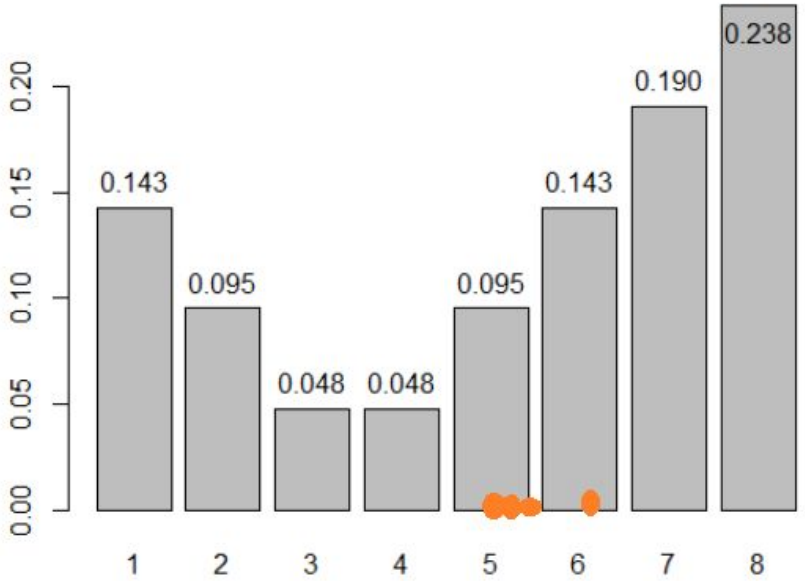
7 1 7 7 5 8 8 6 6 1 6 1 7 1 1 7 6 6 8 7 8 8 2 6 6 8
6 8 3 7 8 2 5 7 4 7 6 6 7 2 ...

(7+1+7+7+5+8+8+6+6+1)/10= 5.6

(6+1+7+1+1+7+6+6+8+7)/10 =5.0

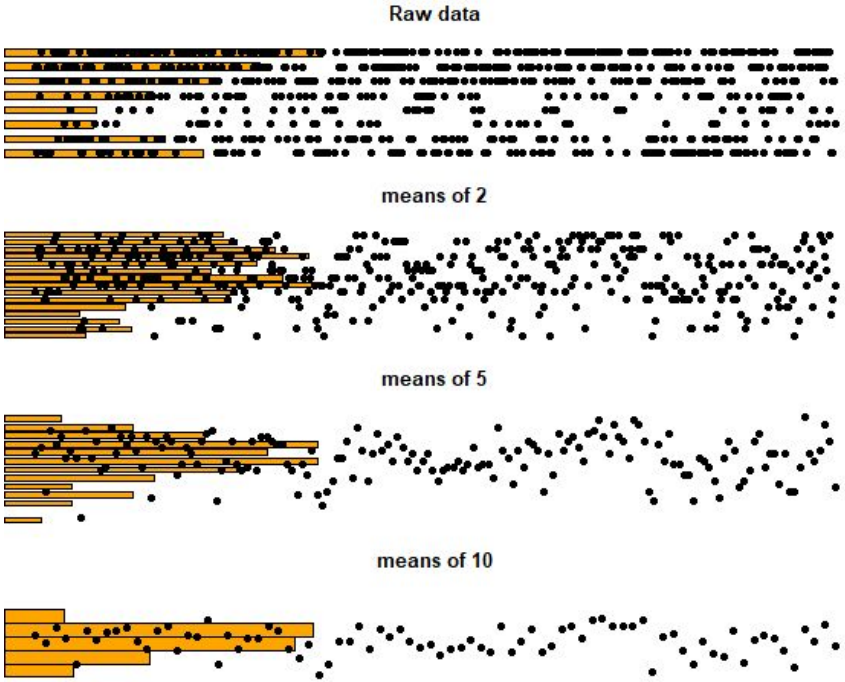
(8+8+2+6+6+8+6+8+3+7)/10=6.2

(8+2+5+7+4+7+6+6+7+2)/10= 5.4



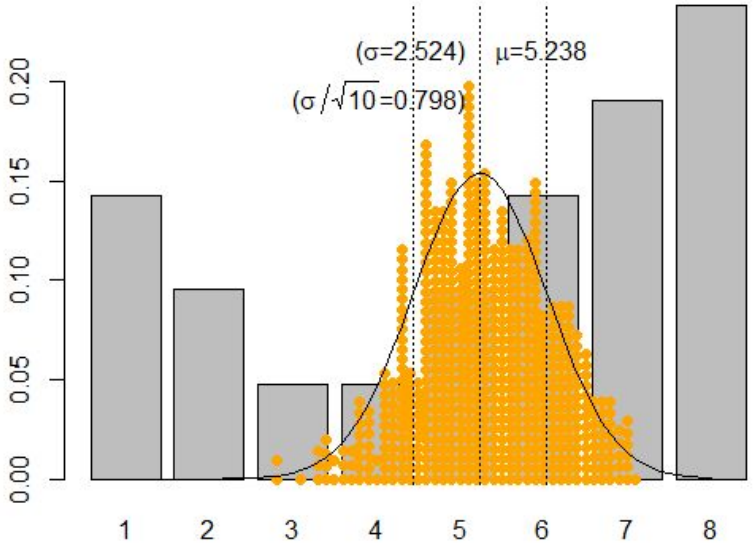
Medelvärdet av tillräckligt många är normalfördelat

- Frekvenstolkningen säger att andelen observationer kommer att fördela sig som staplarna i den första raden.
- Centrala gränsvärdessatsen säger att om vi tar medelvärdet av flera variabler, så kommer resultatet att bli ungefär normalfördelat. Ju fler, dess mer normalfördelat.



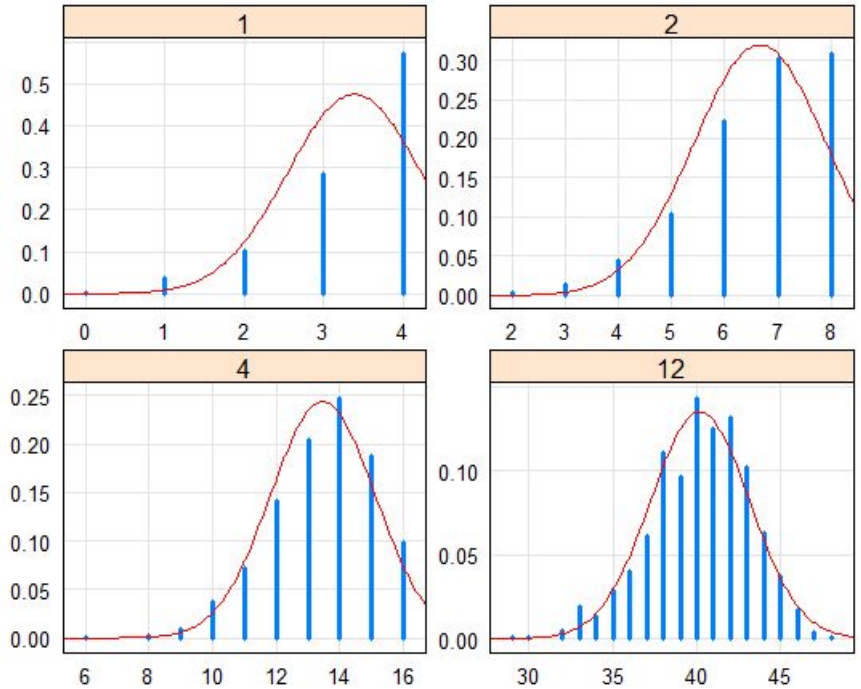
640 medelvärden av 10

Medelvärdet av n observationer koncentreras kring μ och liknar en N(μ,σ/n$^{1/2}$)-fördelning. Ju större *n*, desto mer likt.



Lägg ihop oberoende och likafördelade variabler

3.125 CGS tillsammans med tidigare kunskaper om diskreta fördelningar



Exempel

Viktfördelning är inte normalfördelad.

- Kvinnor är 15 kilo lättare än män (i genomsnitt)
- Extrema värden uppåt förekommer på ett helt annat sätt än extrema värden nedåt.

Längd, vikt och BMI. Medelvärden samt felmarginal (95-procentigt konfidensintervall)

Redovisning efter ålder och kön. 16 år och äldre

Källa: SCB, Undersökningarna av levnadsförhållanden (ULF/SILC)

	Medellängd						Medelvikt					
Definitioner	1988-89		2008-09		2010-11		1988-89		2008-09		2010-11	
	Andel	Fel-marginal	Andel	Fel-marginal	Andel	Fel-marginal	Andel	Fel-marginal	Andel	Fel-marginal	Andel	Fel-marginal
Samtliga 16+ år	171,3	± 0,2	172,6	± 0,1	172,5	± 0,1	70,3	± 0,2	74,8	± 0,2	75,2	± 0,2
Män 16+ år	178,2	± 0,2	179,5	± 0,2	179,4	± 0,2	77,5	± 0,3	82,6	± 0,3	82,9	± 0,4
Kvinnor 16+ år	164,7	± 0,1	165,8	± 0,1	165,7	± 0,2	63,3	± 0,2	67,1	± 0,3	67,4	± 0,3

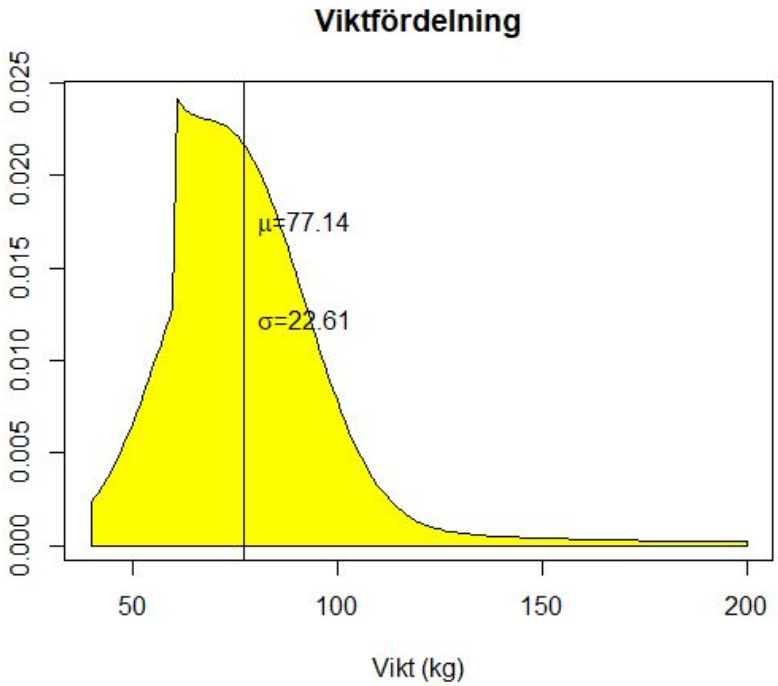
Högst åtta personer får åka i hissen

$S_n=X_1+X_2+X_3+X_4+X_5+X_6+X_7+X_8$.

$E(S_n)=8*\boldsymbol{\mu}=8*77.14=617$

$D(S_n)=8^{1/2}*\boldsymbol{\sigma}= 2.83*22.61=64.0$.

Sannolikheten att åtta personer väger mer än 800 kilo: $P(S_8>800)=$ $1-P(S_8<800)\sim1-\boldsymbol{\Phi}((800-617)/64)= 1-\boldsymbol{\Phi}(2.859375)=$[använd pnorm eller normcdf]$=1-0.9978=0.0022$.



Avsnittets mål

- Centrala gränsvärdessatsen säger något som vi egentligen kände till, enligt formeln plus/minus standardavvikelsen genom roten ur antalet observationer.
- Även rätt exotiska fördelningar konvergerar, så länge som variansden existerar.

Kända diskreta fördelningar

Avsnittets mål

- Geometrisk fördelning är att vänta på något.
- Binomialfördelning är att välja med återläggning.
- Poissonfördelning är osannolika experiment utförda många gånger.
- Hypergeometrisk fördelning är att välja utan återläggning
- Om man kan känna igen respektive situation, har man mycket på fötterna.

Geometrisk (ffg-fördelning)

Det finns forskning som tyder på att föräldrar i Norge och Sverige hellre har flickor än pojkar, till den grad att vissa skaffar barn efter barn tills de får en flicka och sedan slutar. (Det är tveksamt om effekten är så stark, men här är en länk: https://www.viforaldrar.se/gravid/onskas-dotter-varfor-ar-det-sa/)

Vi antar att ett barn vid varje graviditet är en flicka med sannolikhet 0.485, oberoende av tidigare resultat, en siffra man kan komma fram till genom att titta i human mortality database, som, något förvånande, har antal barn. (https://www.mortality.org). Antag att föräldrar skaffar barn efter barn tills det blir en flicka.

Låt N vara antalet barn de får

P(N=1)=0.485

P(N=2)=0.515*0.485

...

P(N=n)=0.515*...*0.515*0.485=0.515^(n-1)*0.485

N: antal gånger man måste göra ett försök som lyckas oberoende av varandra med slh *p*

$p_N(n)=q^{n-1}*p$, där q=1-p

$$E(N) = p\sum_{n=1}^{\infty} nq^{n-1} = p\frac{d}{dq}\sum_{n=0}^{\infty} q^n = p\frac{d}{dq}\frac{1}{1-q}$$

$$p\frac{d}{dq}\frac{1}{1-q} = p\frac{1}{(1-q)^2} = \frac{1}{p}$$

$$V(X) = \frac{q}{p^2}$$

Fördelningsfunktion

- Explicit uttryck:

$F(x)=1-q^x$, där $q=1-p$.

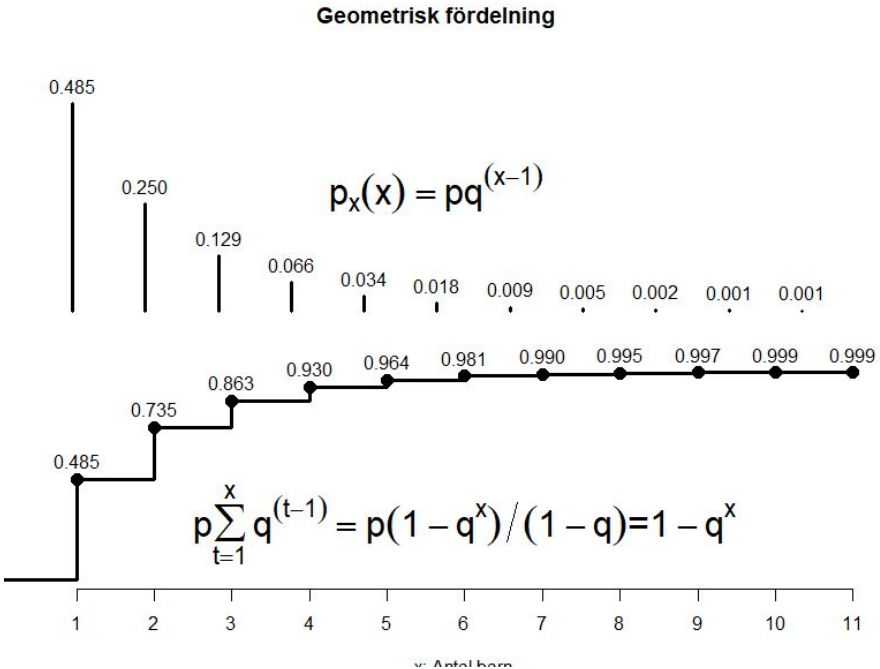
- Notera att kvartilen är lätt att beräkna, om det definieras som det minsta x så att

$1-F(x)<\alpha$

$q^x<\alpha; x*\log(q)<\log(\alpha); x>\log(\alpha)/\log(q)$

α=0.05 och q=0.515 ger x> 4.51, dvs

$\lambda_{0.05}$=5. Längre än 5 % sannolikhet att få fem barn eller mer.



Geometrisk fördelning

Y: antal gånger tills man slår sexa $p=\frac{1}{6}$ med en tärning. Man får E(Y)=1/p=6 betyder att man i genomsnitt behöver vänta sex gånger. V(Y)=q/p^2=$\frac{5}{6}$ /$(\frac{1}{6})^2$=30

D(Y)=$30^{1/2}$~5.48

E(Antal pojkar)=E(Antal barn)-1=1/0.485-1=1.061. Paradoxalt nog leder flicksatsningen inte till fler flickor.

Den geometriska fördelningen är en diskret version av den kontinuerliga exponentialfördelningen, som studeras i nästa föreläsning.

Preludium om diskret matematik. Dragning utan återläggning

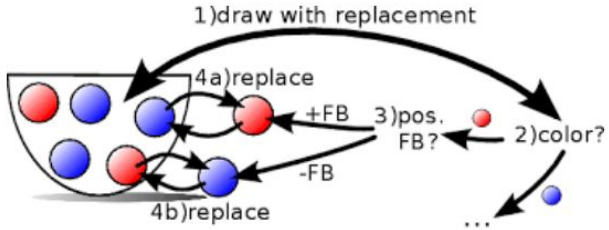
På hur många sätt kan man dra *k* bollar från *n*?

Den första kan dras på n sätt, den andra på n-1 och nummer k på n-k+1 sätt.

Det finns alltså n*(n-1)*…(n-k+1) sätt att dra bollarna, men då har vi tagit hänsyn till ordningsföljden. Det finns för varje dragning k*(k-1)*…*2*1 = k! olika ordningsföljder. Därför finns

$$\binom{n}{k} = \frac{n * (n-1) * \ldots (n-k+1)}{k!}$$

Sätt att välja k bollar ur en urna med n bollar.



Binomialfördelning (”dragning med återläggning”)

Gör N försök som oberoende av varandra lyckas med sannolikhet *p*. Låt *X* vara antalet lyckade försök. Minns hur vi tidigare hade sannolikheter att en maskin fungerar.

P(X=0) =⅔ *¾ *½ =¼ = 25 %

P(maskin 1 fungerar) = ⅓;P(maskin 2 fungerar) = ¼ ;

P(maskin 3 fungerar) =½ ;

P(X=1)= ⅓ *¾ *½+⅔*¼*½ +⅔*¾ *½=11/24 = 45.8 %

p(x)

Situationen är liknande, men här fungerar varje maskin med samma sannolikhet p. Antag igen att q=1-p. Då behöver vi summera ihop alla n termer
$p*q*q*p*...*p=p^kq^{n-k}$, där k av faktorerna är p och n-k q. Platserna för p kan väljas på $\binom{n}{k}$ sätt. Totalt blir sannolikheten.

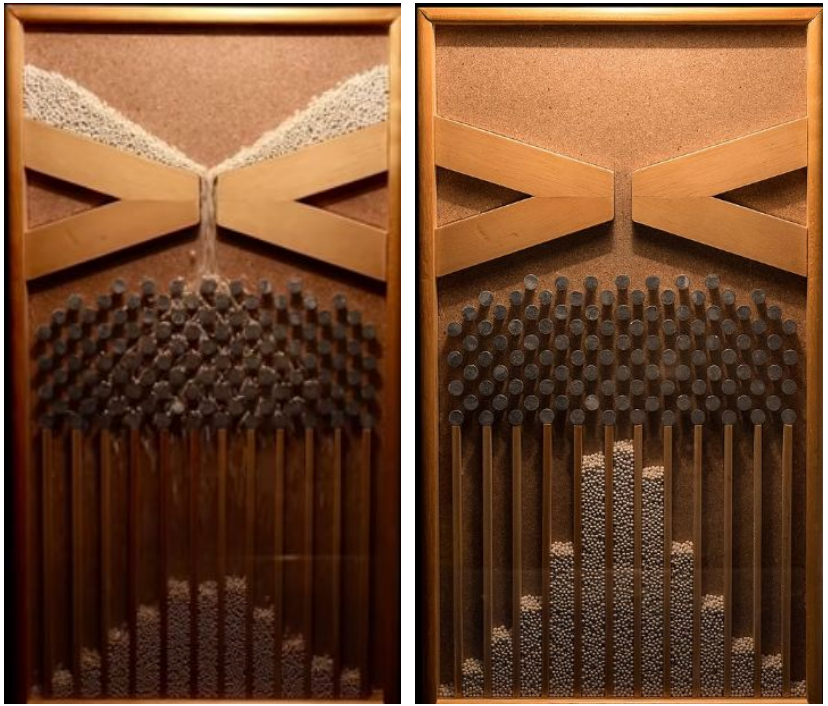
$$p_N(k) = \binom{n}{k} p^k q^{n-k}$$

Quincunx (https://en.wikipedia.org/wiki/Bean_machine)

Varje liten kula studsar mot tio pinnar och går med sannolikhet ½ åt vänster och ½ åt höger. (Oberoendeantagandena kan infrågasättas, men det skippar vi.) Om vi numrera vänster till höger med 0,...,10 får vi:

$$p_N(k) = \binom{n}{k} p^k q^{n-k} \text{ med } p = \frac{1}{2}$$

(Hoppas principen är klar. Denna quincunx är inte helt ortodox.)



Fördelningsfunktion

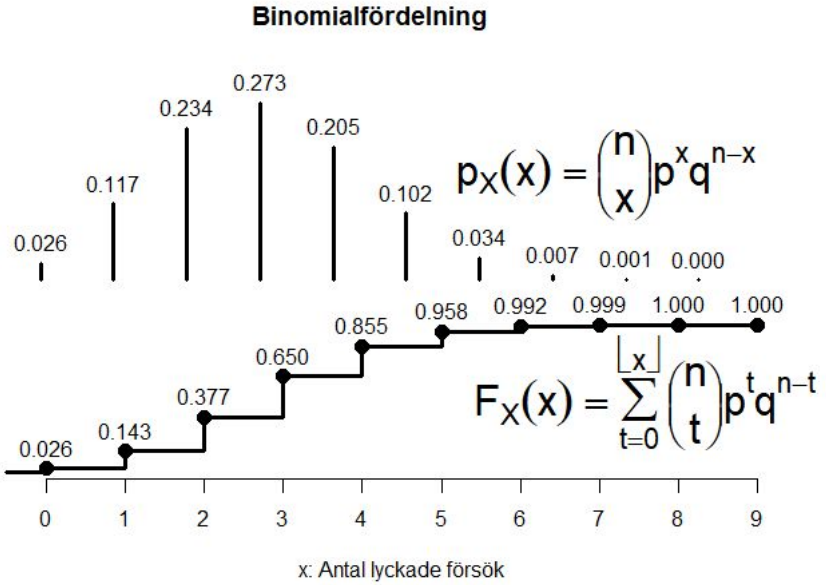
Fallet p = ⅓ syns i bilden.

Summan kan inte förenklas ytterligare. För p = 0.05,.10,0.15,...,0.5 och n = 2,3,...19 finns tabeller.

Notera att fallet p>0.5 återförs på p<0.5 genom att byta roll på p och q.

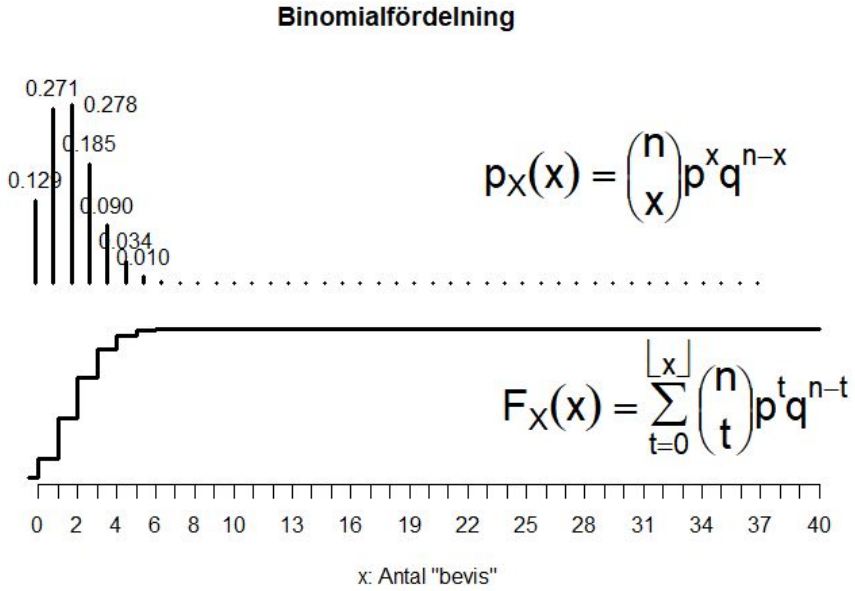
E(X)=np

V(X)=npq.



Metaanalys

- Det är idag populärt att försöka bevisa att statin är farligt.
- Ett test konstrueras så att sannolikheten att det visar att statin är farligt är 0.05 om det egentligen är ofarligt.
- Antag att 40 forskargrupper oberoende av varandra utför sådana försök.
- Låt oss se på fördelningen för N, antalet test som (felaktigt) visar att statin är farligt.
- E(N)=40*0.05=8
- V(N)=40*0.05*0.95=1.9
- D(N)=1.39



Jämför det inledande exemplet när man testar för en ovanlig sjukdom! Detta är ett reellt problem både i sceening för sjukdomar och i rapportering av forskningsresultat.

Många försök med liten sannolikhet: Poissonfördelningen

Låt oss hålla väntevärdet n*p=**λ** konstant, men låt n gå mot oändligheten!

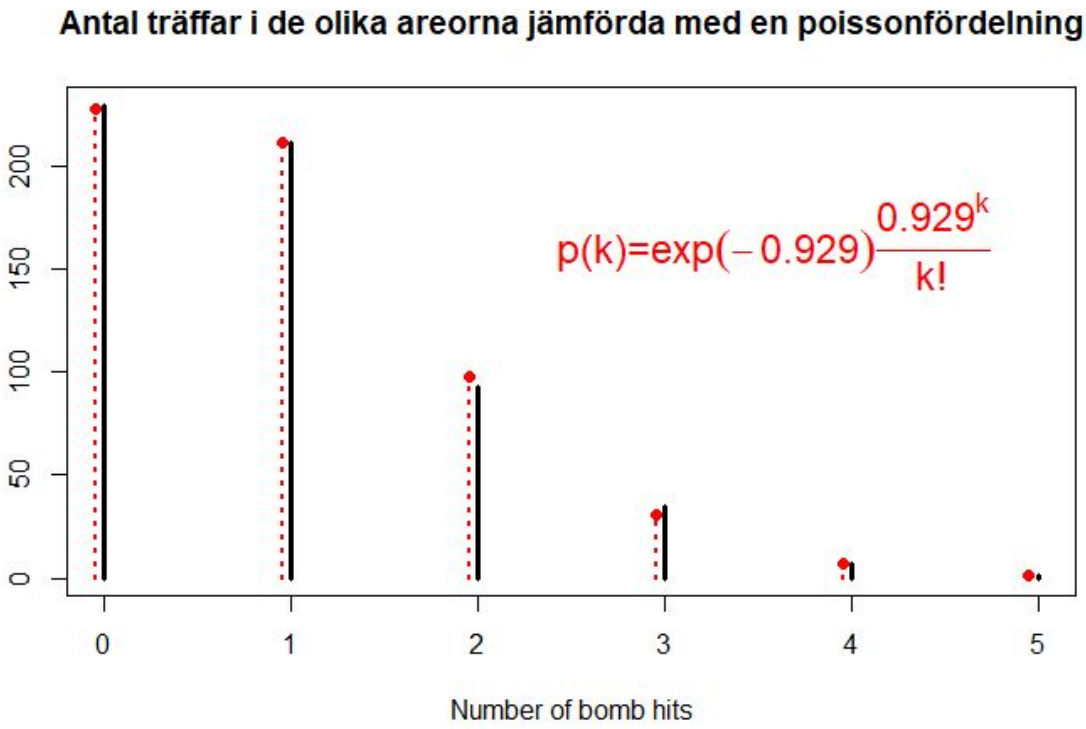
Tolkning: När man gör ett försök många gånger som lyckas med lyckas med liten sannolikhet, behöver vi inte veta det exakta antalet försök, bara det väntevärdet. Den resulterande fördelningen kallas Poissonfördelning.

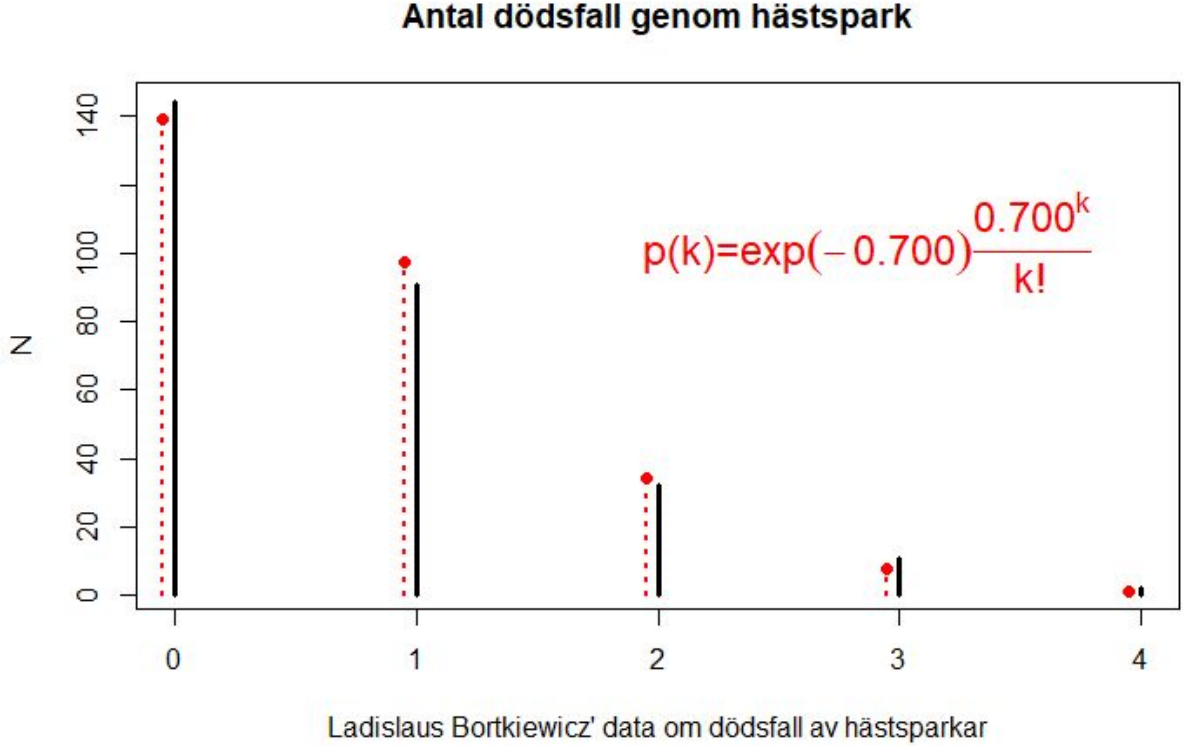
$$\frac{n!}{(n-k)!\,k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}=\frac{n!}{(n-k)!}\left(\frac{\frac{1}{n}}{1-\frac{\lambda}{n}}\right)^k\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n=$$

$$\left(\frac{1}{1-\frac{\lambda}{n}}\right)^k\left(\frac{n-1}{n}*\ldots*\frac{n-k+1}{n}\right)\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\to\frac{\lambda^k}{k!}e^{-\lambda}$$

$$\text{då } n\to\infty$$

Poissonfördelning - bombexemplet



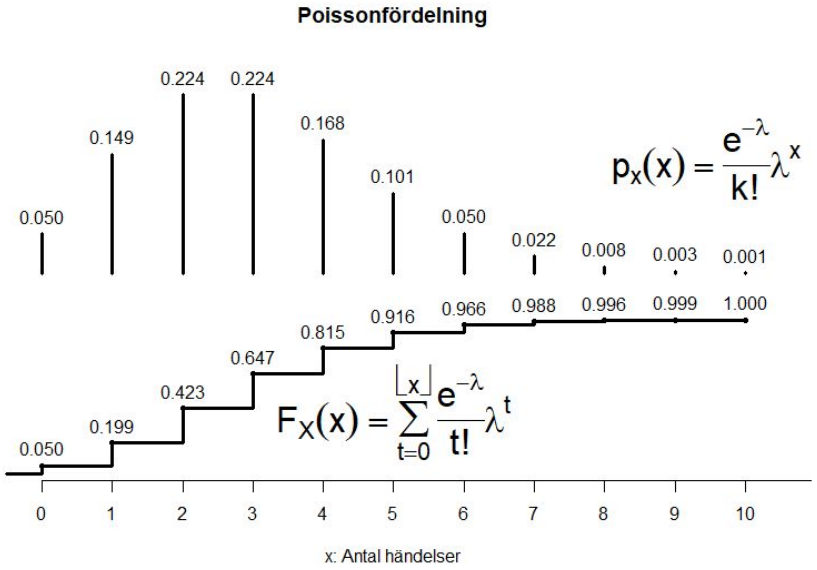

Antal dödsfall genom hästspark
N
0
20
40
60
80
100
140
0
1
2
3
4
p(k)=exp(−0.700)0.700^k / k!
Ladislaus Bortkiewicz' data om dödsfall av hästsparkar

Fördelningsfunktion

Tabell 5. Poissonfördelningen

$P(X \leq x)$ där $X \in Po(\mu)$

x \ μ	0.1	0.2	0.3	0.4	0.5
0	0.90484	0.81873	0.74082	0.67032	0.60653
1	0.99532	0.98248	0.96306	0.93845	0.90980
2	0.99985	0.99885	0.99640	0.99207	0.98561
3	1.00000	0.99994	0.99973	0.99922	0.99825
4	1.00000	1.00000	0.99998	0.99994	0.99983
5	1.00000	1.00000	1.00000	1.00000	0.99999
6	1.00000	1.00000	1.00000	1.00000	1.00000
7	1.00000	1.00000	1.00000	1.00000	1.00000



Exempel

Utspädningsförsök. Antag att det finns N bakterier i en lösning. Blanda ordentligt, späd ut och odla ett antal prover om med samma mängd vätska. Låt X vara det antal av proverna där bakterierna växer till.

Sannolikheten att ett prov växer till är precis $P(X>0)=1-P(X=0)=1-e^{-\lambda}$, λ där är det förväntade antalet bakterier per prov. Alltså är antalet prover med bakterietillväxt binomialfördelat med $p= 1-e^{-\lambda}$. Detta kan användas för att skatta λ, som i sin tur kan ge en uppskattning av bakteriekoncentrationen.

Hypergeometrisk fördelning

- Tant Agda är väldigt noga med att man ska hälla i te först och mjölk sedan.
- Tant Edit ifrågasätter Agda och sätter henne på ett test. Agda får åtta numrerade koppar med te. Edit, men inte Agda, vet vilka fyra som hon hällt te i först i. I resten har hon hällt mjölk först.
- Agda kan egentligen inte alls känna skillnaden. Vad är sannolikheten att hon klarar Edits test?

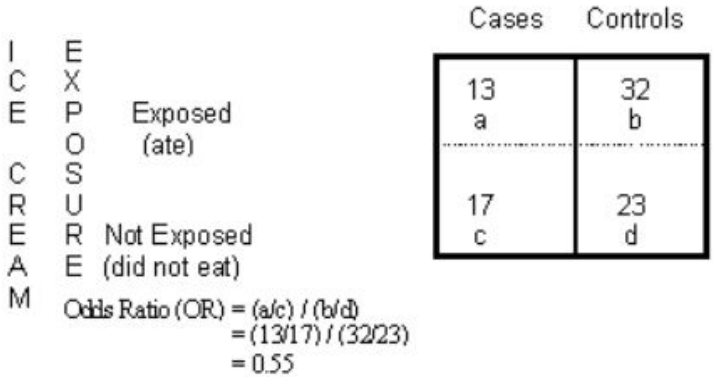
4 röda och 4 blå kulor.

Det finns bara ett sätt att välja bara blå kulor.

Det finns $\binom{8}{4} = \frac{8*7*6*5}{4*3*2*1} =$ 70 sätt att välja fyra kulor.

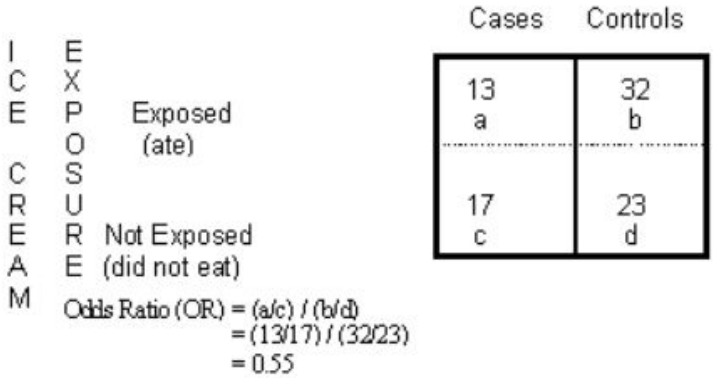
Sannolikheten är alltså $\frac{1}{70}$ att Agda inte kommer att avslöjas som den okunniga snobb hon är.

Case-controlstudier



- En läkare misstänker att fall (case) av magsjuka är kopplat till att äta en viss sorts glass.
- Han frågar därför alla patienter han har med magsjuka om de ätit det aktuella sorten.
- Så letar han upp patienter utan magsjuka (controls) om de ätit glassorten. Resultat som i tabellen.
- Fråga: Kan en så stor skillnad uppstå av en slump?

Fishers exakta test



- *Om* det inte finns någon koppling, kan man likna situationen vid en urna med 45 röda bollar (ätit) och 40 som blå (inte ätit).
- Vid dragning av 55 bollar, fick man 32 röda. Sannolikheten att få 32 eller fler kan fås genom den så kallade *hypergeometriska* fördelningen och är 0.14, dvs. Ungefär en gång på sju.
- Detta är inte tillräckligt för att vi ska dra slutsatsen att det finns en koppling.
- Detta test kallas *Fishers exakta test* och är mycket vanligt.

Sannolikhetsfunktion, hellre dator än tabell

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

$$X \in Hyp(N, n, m)$$

Den hypergeometriska fördelningen har tre parametrar:

N:totalt antal kulor

m antal blå kulor

n: antal man drar

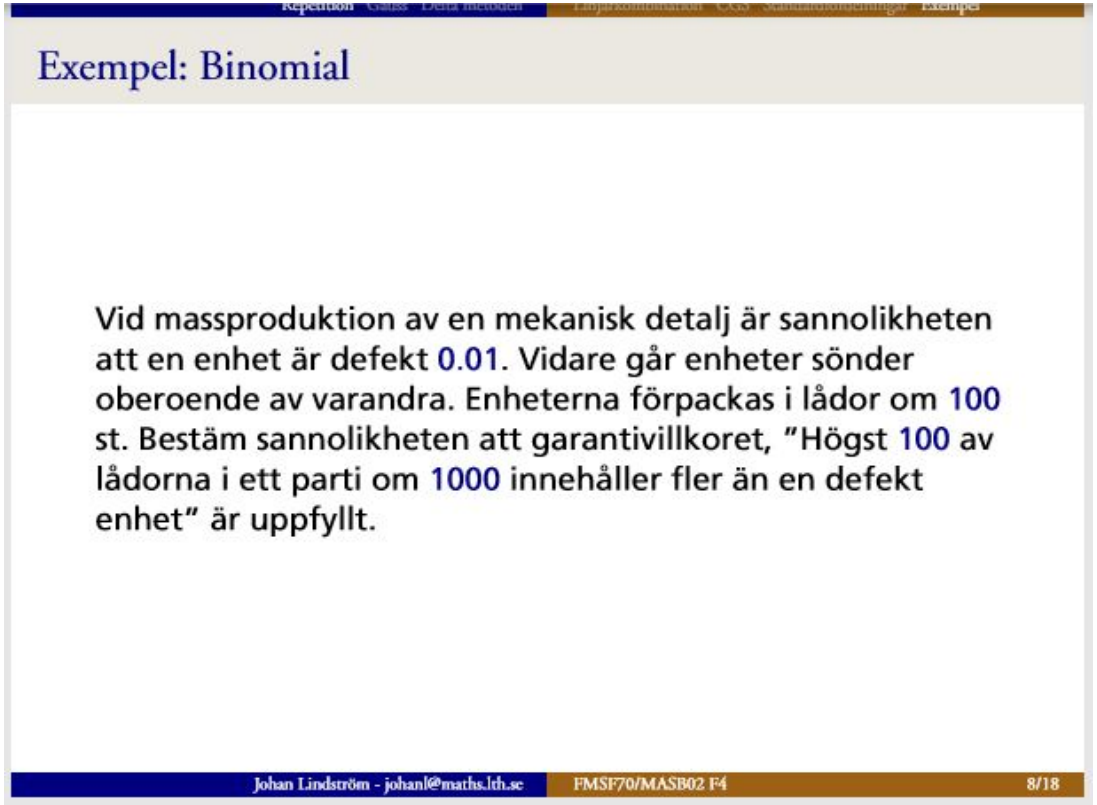
```
> fisher.test(matrix(c(13,32,17,23),2,2),alternative = "less")

        Fisher's Exact Test for Count Data

data:  matrix(c(13, 32, 17, 23), 2, 2)
p-value = 0.1394
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 1.285951
sample estimates:
odds ratio
 0.5535703

> 1-phyper(31,45,40,55)
[1] 0.1393503
```

Quiz - vilken fördelning svarar på vilken fråga?

Geometrisk fördelning	$p_N(n)=q^{n-1}*p$, där $q=1-p$
Binomialfördelning, $Bin(n, p)$	$p(k) = \binom{n}{k} p^k(1-p)^{n-k}$
Poissonfördelning, $Po(\mu)$	$p(k) = e^{-\mu}\frac{\mu^k}{k!}$
Hypergeometrisk fördelning $Hyp(N, n, m)$	$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$

1) Stina är så charmig att hon får alla jobb hon söker med sannolikhet 75 %. Vad är sannolikheten att hon behöver få på fler än tre intervjuer?
2) Sockerbagaren lägger russin i en jättedeg som räcker till tre i varje bulle. Hur stor andel av bullarna saknar russin?
3) Till fredagens TP delas korridorens medlemmar slumpmässigt in i fyra lag. Johanna vill gärna ha med någon sportnörd, som det finns tre av. Hur troligt är det att det blir så?
4) Gunnel har tre barn. Hon har en genetisk sjukdom, som bara beror av en dominant gen, som inte sitter på könskromosomen. Eftersom sjukdomen debuterar i vuxen ålder undrar hon hur många av hennes barn som har genen.

Från förra året. Är binomial det enda svaret?



Avsnittets mål

- Geometrisk fördelning är att vänta på något.
- Binomialfördelning är att välja med återläggning.
- Poissonfördelning är osannolika experiment utförda många gånger.
- Hypergeometrisk fördelning är att välja utan återläggning
- Om man kan känna igen respektive situation, har man mycket på fötterna.