Statistik och Dataanalys I

Föreläsning 14 - Slumpvariabler

Mattias Villani

Statistiska institutionen
Stockholms universitet

mattiasvillani.com @matvil @matvil mattiasvillani

Översikt

- **Slumpvariabler** och **sannolikhetsfördelningar**
- Sammanfatta sannolikhetsfördelningar - **väntevärde** och **varians**
- **Kontinuerliga slumpvariabler** - en första titt på **normalfördelningen**.
- **Räkna med slumpvariabler** - skift, skalning, linjärkombination och summor
- **Beroende slumpvariabler** - **korrelation** och **kovarians**

Slumpvariabler

- **Slumpvariabel** mäter ett numeriskt värde från slumpmässigt försök. T ex antal prickar vid kast med tärning, eller

$$X = \begin{cases} 0 & \text{om minusgrader} \\ 1 & \text{om plusgrader} \end{cases}$$

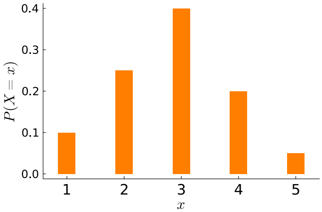
- Vi skriver **slumpvariabler** med **stora bokstäver** X och deras numeriska **utfall** med **små bokstäver** x.
- Slumpvariabeln "antal prickar" X fick utfallet $x = 3$. 
- En slumpvariabel kan vara:
  - **diskret** (utfallen går att räkna, även $0, 1, 2, \ldots$ till oändligt)
  - **kontinuerlig** (utfallen går inte att räkna, många decimaler)
- Exempel
  - Diskret: $X =$ antal prickar på tärning
  - Kontinuerlig: $X =$ temperatur (med decimaler)

Sannolikhetsfördelning

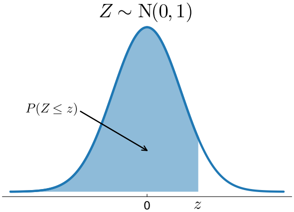
- Varje värde x som slumpvariabeln X kan anta har en **sannolikhet** $P(X = x)$ (eller bara $P(x)$).
- **Sannolikhetsfördelningen** för X är sannolikheterna för alla möjliga utfall. $\sum P(x) = 1$.

x	1	2	3	4	5	Σ
P(x)	0.10	0.25	0.40	0.20	0.05	1

Diskret slumpvariabel



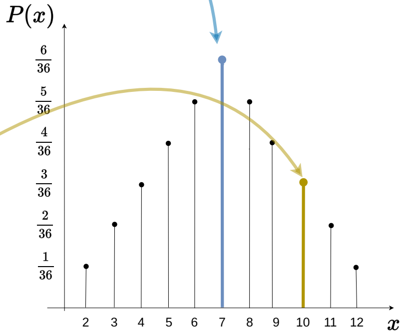
Kontinuerlig slumpvariabel



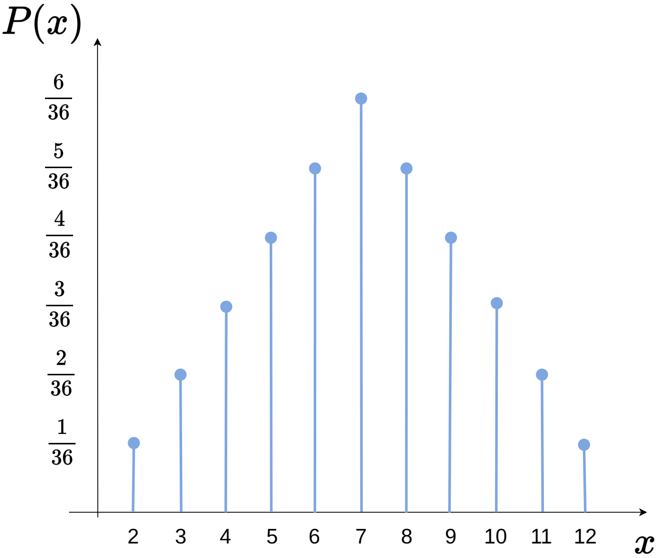
Kasta tärning - fördelning för slumpvariabel

- **Slumpvariabel**: Händelser $\Longrightarrow$ numeriska värden.

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12



Kasta tärning - fördelning för slumpvariabel



Väntevärde - fördelningens centrum

- **Medelvärdet** för ett **stickprov**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \ldots + \frac{1}{n}x_n$$

- X är en slumpvariabel med sannolikhetsfördelning $P(X = x)$.
- **Väntevärdet** för **slumpvariabeln** X är (**expected value**)

$$E(X) = \sum_{\text{alla } x} x \cdot P(x)$$

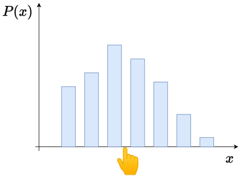
- Vi använder ofta grekiska bokstaven μ för $E(X)$.
  Grekiska bokstaven för m, m som i **mean**. "lilla my".
- Mer utförligt: om X kan anta värdena $\{x_1, x_2, \ldots, x_m\}$ så är

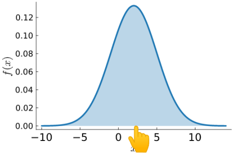
$$E(X) = \sum_{i=1}^{m} x_i \cdot P(x_i)$$

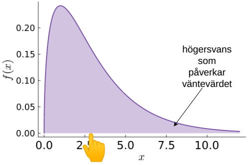
Väntevärde - mått fördelningens centrum (läge)

- Väntevärde - sannolikhetsfördelningens **centrum**.
- Väntevärdet - punkt där sannolikhetsfördelning **’balanserar’**.
- Medelvärdet $\bar{x}$ påverkas mycket av extrema värden.
- Väntevärdet påverkas mycket av fördelningens ’svansar’.

Diskret slumpvariabel	Kontinuerlig symmetrisk slumpvariabel	Kontinuerlig skev slumpvariabel





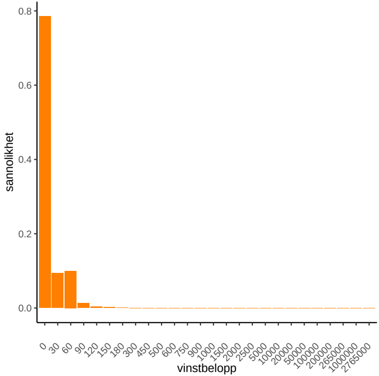


Förväntad vinst - Trisslott

sda1 triss

vinst	antal	probs
0	4713000	0.7855000000
30	565563	0.0942605000
60	601056	0.1001760000
90	78000	0.0130000000
120	21600	0.0036000000
150	11280	0.0018800000
180	3600	0.0006000000
300	2790	0.0004650000
450	375	0.0000625000
500	600	0.0001000000
600	600	0.0001000000
750	150	0.0000250000
900	180	0.0000300000
1000	480	0.0000800000
1500	240	0.0000400000
2000	150	0.0000250000
2500	45	0.0000075000
5000	90	0.0000150000
10000	132	0.0000220000
20000	21	0.0000035000
50000	9	0.0000015000
100000	6	0.0000010000
200000	3	0.0000005000
265000	26	0.0000043333
1000000	1	0.0000001667
2765000	3	0.0000005000
summa:	6000000	1

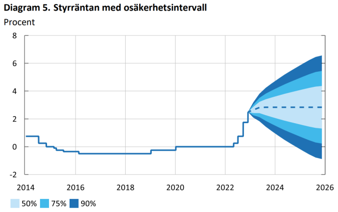
$$\begin{aligned}\mathrm{E}(\text{vinst}) = {} & 0 \cdot 0.7855 + 30 \cdot 0.0942605 + \\ & 60 \cdot 0.100176 + \cdots + 2765000 \cdot 0.0000005 \\ = {} & 14.7 \text{ kr}\end{aligned}$$

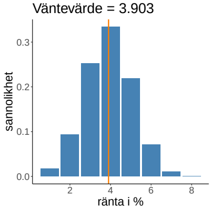


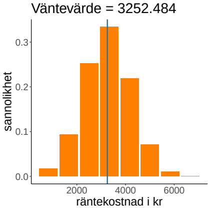
Källa: Svenska spel - `https://www.svenskaspel.se/triss/spelguide/triss-30`

Vilken räntekostnad för bolån i slutet av 2023?

- Antag: lån på 1 miljon. 1% högre ränta än styrräntan.







bankränta i %	sannolikhet	månadskostnad
1	0.017	833
2	0.094	1667
3	0.252	2500
4	0.334	3333
5	0.219	4167
6	0.071	5000
7	0.011	5833
8	0.001	6667

$$E(\text{ränta}) = 1 \cdot 0.017 + 2 \cdot 0.094 + \ldots + 8 \cdot 0.001 \approx 3.9\%$$

$$E(\text{kostnad}) = 833 \cdot 0.017 + 1667 \cdot 0.094 + \ldots + 6667 \cdot 0.001 \approx 3252 \text{ kr}$$

Diagram 5 från Penningpolitisk rapport, Nov 2022, Sveriges Riksbank, `https://www.riksbank.se`

Varians - fördelningens spridning

- Väntevärdet μ är bara en slags bästa gissning.
- Ofta viktigt att veta fördelningens **spridning**. Osäkerhet.
- Medelavvikelse från μ som spridning?
  - Avvikelser från centrum $x - \mu$.
  - Problem: Negativa och positiva avvikelser kan ta ut varandra.
  - Lösning: kvadrera avvikelserna $(x-\mu)^2$ först.
- **Variansen** för en slumpvariabel

$$Var(X) = \sum_{\text{alla } x} (x-\mu)^2 P(x)$$

- Variansen skrivs ofta med symbolen σ^2.
- Exempel: X = räntekostnad. $\mu = E(X) = 3252$.

$$Var(X) = (833-3252)^2 \cdot 0.017 + (1667-3252)^2 \cdot 0.094 + \ldots + (6667-3252)^2 \cdot 0.001 \approx 965553.1 \text{ kr}^2$$

Standardavvikelse - ett mått på medelspridning

- **Variansen** för en slumpvariabel

$$Var(X) = \sum_{\text{alla } x} (x - \mu)^2 P(x)$$

  Variansen har **enheter i kvadrat**. Ingen trevlig tolkning.

- **Standardavvikelsen** har samma enheter som slumpvariabeln

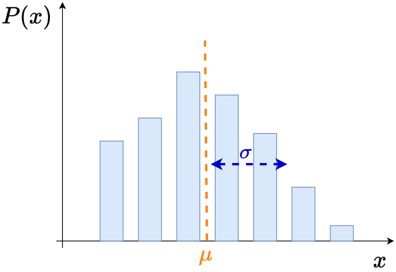
$$\sigma = SD(X) = \sqrt{Var(X)}$$

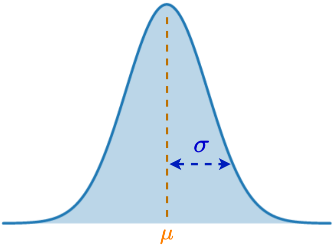
- Exempel: $X =$ räntekostnad.

$$\sigma = \sqrt{965553.1} \approx 982.63 \ \texttt{kr}$$

- Vår "bästa gissning" av räntekostnad: $\mu = 3252$ kr
- Men en "typisk avvikelse" från denna gissning är cirka 983 kr.

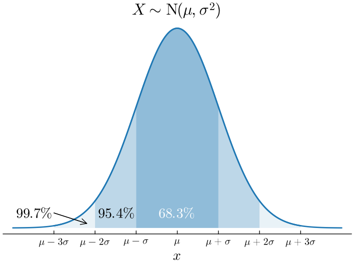
Väntevärde och standardavvikelse

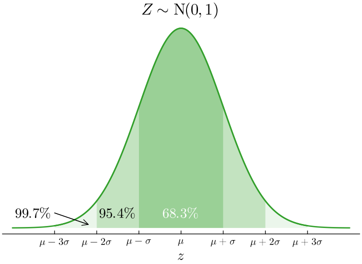




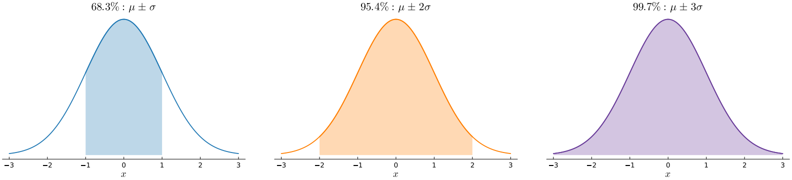
Normalfördelning - 68-95-99.7% regeln

- **Normalfördelning**, $X \sim N(\mu, \sigma^2)$
  - Väntevärde $E(X) = \mu$
  - Varians $Var(X) = \sigma^2$
- **Parametrarna** μ och σ^2 är just väntevärdet och variansen!
- **68-95-99.7% regeln**





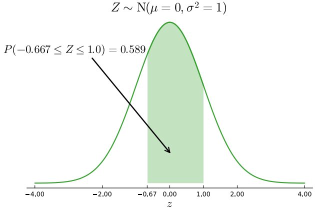
68-95-99.7% regeln



Kontinuerliga slumpvariabler och täthetsfunktionen

- **Kontinuerlig slumpvariabel** antar alla värden, men $P(X = x) = 0$ för alla x! 🤯
- **Täthetsfunktion**: $f(x)$.
- Positiv $f(x) > 0$ för alla x.
- Täthetsfunktion ger **inte** sannolikheter. OK om $f(x) > 1$.
- **Täthetsfunktionen** används för att **beräkna sannolikheter**:

$$P(a \leq X \leq b) = \text{arean under } f(x) \text{ mellan } a \text{ och } b$$



- **SDAIII**: räkna arean under funktion med **integration**.

Normalfördelning - interaktivt

Normalfördelningen

μ : 0

σ : 1

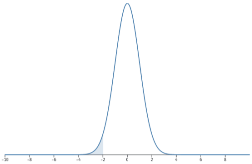
Kvantil: -1.96

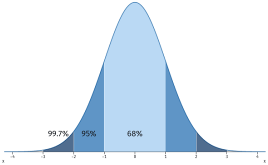
Om $X \sim \mathrm{N}(0, 1)$ så gäller att

$E(X) = \mu = 0.00$

$Var(X) = \sigma^2 = 1.00$

$P(X \leq -1.96) = 0.02500$





Mattias Villani Normalfördelningen

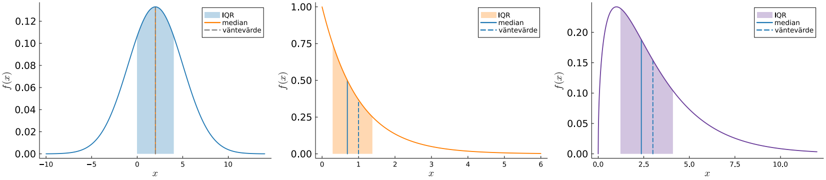
Observable

Median och interkvartilavstånd

- **Median**, m: värde med 50% av sannolikhetsmassan till vänster.

$$P(X \leq m) = 0.5$$

- 10%-**kvantil**: 10% av sannolikhetsmassan till vänster.
- **Kvartiler**: 25%, 50%, 75%.
- **Interkvartilavstånd** (**IQR**): avstånd mellan 25%-kvartil och 75%-kvartil.



Skifta slumpvariabler

- Exempel: X ränta i procent på mitt banklån. $E(X) = 3.9\%$.
- Sämre förhandlare: bankräntan 2% högre än styrräntan.
- Din ränta: $Y = X + 2$. **Skiftar/förskjuter** slumpvariabeln.
- Måste vi göra om alla beräkningar för dig? Nope.

$$E(Y) = E(X) + 2 = 3.9 + 2 = 5.9\%$$

> **Väntevärde - skiftade slumpvariabler.**
>
> $E(X \pm c) = E(X) \pm c$ för godtycklig konstant c

- **Variansen ändras inte** av ett skift:

> **Varians - skiftade slumpvariabler.**
>
> $Var(X \pm c) = Var(X)$ för godtycklig konstant c

Skala slumpvariabler

- Exempel: får dra av 30% på skatten för räntekostnad.
- Räntekostnad efter skatt: $Y = 0.7 \cdot X$. **Skalar** slumpvariabeln.

Väntevärde - skalning.

$$E(aX) = a \cdot E(X) \quad \text{för godtycklig konstant } a$$

Varians - skalning.

$$Var(aX) = a^2 Var(X) \quad \text{för godtycklig konstant } a$$

Standardavvikelse - skalning.

$$SD(aX) = |a| \cdot Var(X) \quad \text{för godtycklig konstant } a$$

- $E(0.7 \cdot X) = 0.7 \cdot 3252 = 2276.4$ kr
- $SD(0.7 \cdot X) = 0.7 \cdot 982.63 \approx 687.84$ kr

Linjärkombinationer av slumpvariabler 😍

- **Linjärkombination** av slumpvariabel = **skift och skalning**.

$$Y = c + aX$$

Väntevärde - linjärkombination.

$$E(c \pm aX) = c \pm aE(X) \qquad \text{för konstanter } a \text{ och } c$$

Varians - linjärkombination.

$$Var(c \pm aX) = a^2 Var(X) \qquad \text{för konstanter } a \text{ och } c$$

- Exempel företags produktionskostnader:
  - X antal efterfrågade enheter (slumpvariabel).
  - Fast produktionskostnad c
  - Rörlig produktionskostnad per enhet a
  - Produktionskostnad: $Y = c + aX$

Standardisering

- Om $X \sim N(\mu, \sigma^2)$ så gäller att

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

- **Standardisering**: från allmän normalfördelning till standard normal genom **skift** och **skalning**.
- Beräkna sannolikheter för $X \sim N(\mu, \sigma^2)$ från standard normal

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

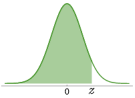
- Exempel: $X \sim N(2, 3^2)$, vad är sannolikheten att $X \leq 5$?

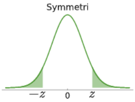
$$P(X \leq 5) = P\left(\frac{X - 2}{3} \leq \frac{5 - 2}{3}\right) = P\left(Z \leq 1\right) = 0.8413$$

Normalfördelning - Z-tabell

Normalfördelning

Tabellen ger sannolikheten $\Phi(z) = P(Z \le z)$ för olika z där Z är standardnormal, $Z \sim N(0, 1)$.
Sannolikheter i den vänstra svansen fås genom symmetri: $P(Z \le -z) = 1 - P(Z \le z)$.





Andra decimalen i z

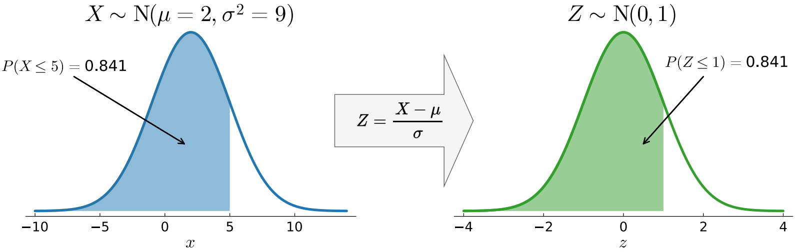
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817

Normalfördelning i R

- $X \sim N(\mu, \sigma^2)$.

Beräkning	R kommando
$f(2)$	`dnorm(x = 2, mean = 1, sd = 1.5)`
$P(X \leq 2)$	`pnorm(q = 2, mean = 1, sd = 1.5)`
Kvantil	`qnorm(p = 0.5, mean = 1, sd = 1.5)`
10 slumptal	`rnorm(n = 10, mean = 1, sd = 1.5)`

Standardisering



Väntevärde - summa av slumpvariabler

- X och Y är två olika slumpvaribler
  - X antal prickar på 1:a tärningen
  - Y antal prickar på 2:a tärningen
  - $X + Y =$ totalt antal prickar på båda tärningarna.

Väntevärde - summa av slumpvariabler.

$$E(X + Y) = E(X) + E(Y)$$

Varians - summa av oberoende slumpvariabler

- För **variansen** måste vi vara försiktiga med eventuella **beroenden mellan variabler**.
- Vadslagning:
  - X är din vinst/förlust i ett vad.
  - Y är din motståndares vinst/förlust.
  - $X + Y = 0$, dvs har ingen varians alls! Perfekt beroende.
- Aktieportfölj:
  - X är avkastning aktie.
  - Y är avkastning på annan aktie.
  - Total avkastning: $X + Y$. Varians?
- Om vi antar att X och Y är oberoende blir variansen enkel:

Väntevärde - summa av oberoende slumpvariabler.

$$Var(X + Y) = Var(X) + Var(Y)$$

Väntevärde och varians - många oberoende variabler

- Låt X_1, X_2 och X_3 vara tre oberoende slumpvariabler.

$$E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3)$$

$$Var(X_1 + X_2 + X_3) = Var(X_1) + Var(X_2) + Var(X_3)$$

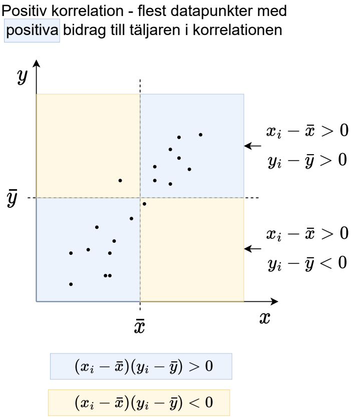
Väntevärde - summa av slumpvariabler.

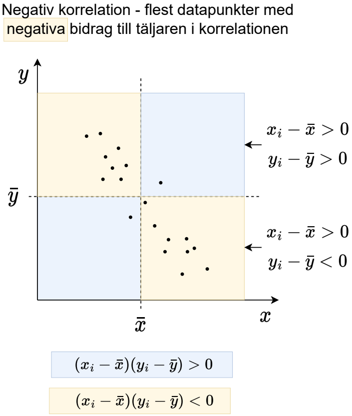
$$E(X_1 + X_2 + \ldots, X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$$

Varians - summa av oberoende slumpvariabler.

$$V(X_1 + X_2 + \ldots, X_n) = Var(X_1) + Var(X_2) + \ldots + Var(X_n)$$

- Korrelation: linjärt beroende mellan variabler.

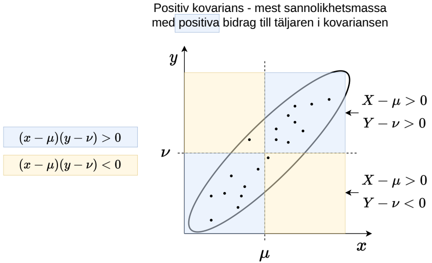




Beroende variabler - Kovarians och Korrelation

- Låt X ha väntevärde μ och Y väntevärde ν.
- **Kovarians**: **linjärt beroende** mellan slumpvariabler.

$$Cov(X, Y) = E\left((X-\mu)(Y-\nu)\right)$$



- **Korrelation** $(-1 \leq Corr(X, Y) \leq 1)$ 😍

$$Corr(X, Y) = \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

- Så kovariansen kan också uttryckas

$$Cov(X, Y) = \rho_{XY} \cdot \sigma_X \cdot \sigma_Y$$

Variansen av en summan av beroende variabler

Varians - summa av beroende slumpvariabler.

$$V(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

- **Positiv kovarians** - **variansen** för summan **större** än vid oberoende.
- **Negativ kovarians** - **variansen** för summan **mindre** än vid oberoende.
- Säker aktieportfölj: välj aktier var priser tenderar att röra sig i olika riktningar. Even Steven.