

---

---

# Statistiek in Data Analyse

— Lydia Brenner —  
@Fysica\_Interact

---

---

# PIZZA!



# Schafen





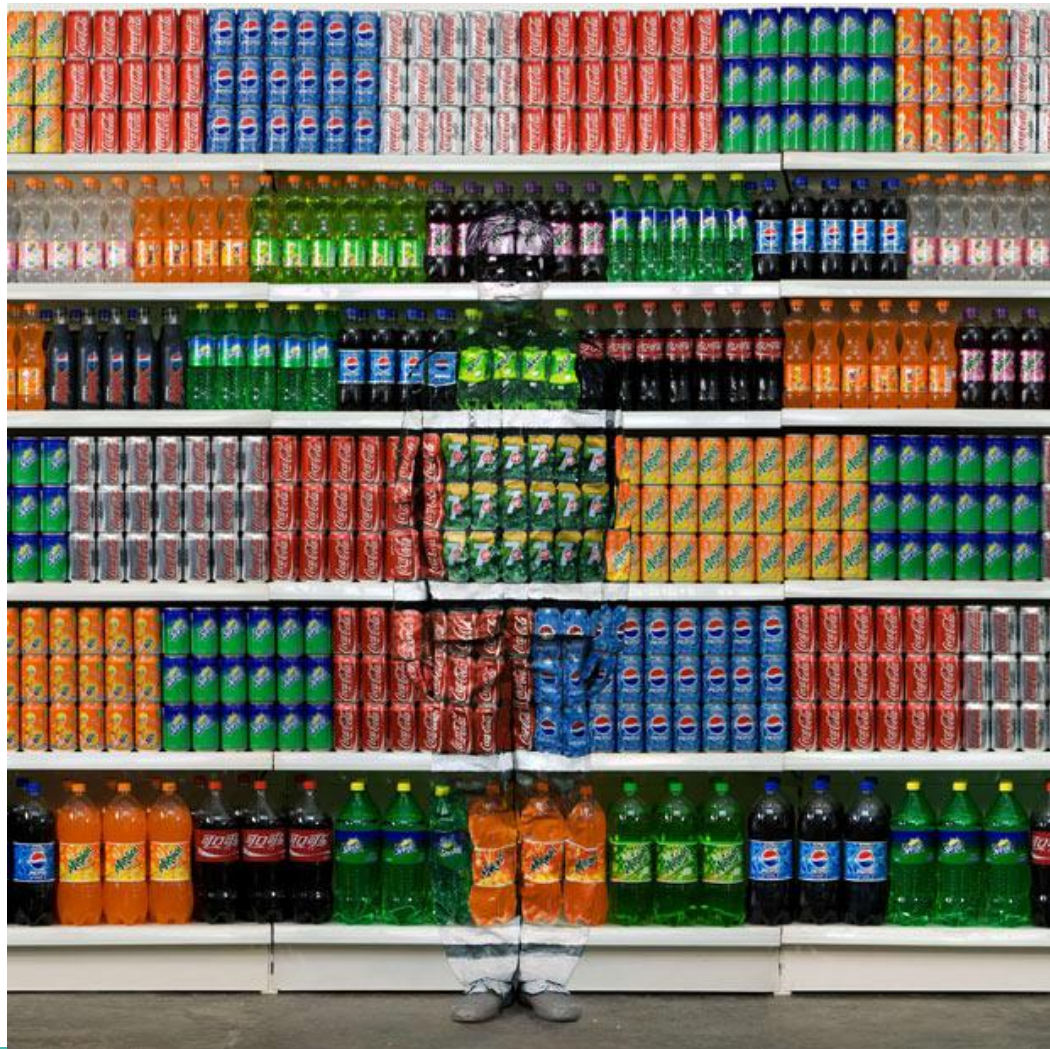
# Liu Bolin

Zie jij hem?



# Liu Bolin

Zie jij hem?





# Liu Bolin

Zie jij hem?

Een andere  
hoek



# Wat is de kans?

Wat wordt er bedoeld met de kans  $P_A$  van een event  $A$ ?

1. Een getal met bepaalde wiskunde regels
2. Een eigenschap van  $A$  die bepaald hoe vaak  $A$  gebeurt
3. Uit  $N$  keer gebeurt  $A$   $N_A$  keer, dan is  $P_A$  het limit of  $N_A/N$  voor grote  $N$
4.  $P_A$  is hoeveel je in  $A$  gelooft, te meten door welke kansen je accepteert in een weddenschap

# Wat is de kans?

Wat wordt er bedoeld met de kans  $P_A$  van een event  $A$ ?

Frequentist (meest gebruikte) interpretatie;

3. Uit  $N$  keer gebeurt  $A$   $N_A$  keer, dan is  $P_A$  het limit of  $N_A/N$  voor grote  $N$

$P_A$  is niet een eigenschap van  $A$ , maar van de combinatie van  $A$  en een dataset met  $N$  datapunten

$$P_A = \lim_{N \rightarrow \infty} \frac{N_A}{N}$$



# Wat is de kans?

Wat wordt er bedoeld met de kans  $P_A$  van een event  $A$ ?

Baysian interpretatie;

4.  $P_A$  is hoeveel je in  $A$  gelooft, te meten door welke kansen je accepteert in een weddenschap

$$P(A|B) = \frac{P(B|A)}{P(B)} \times P(A)$$

# Wat willen we eigenlijk weten?

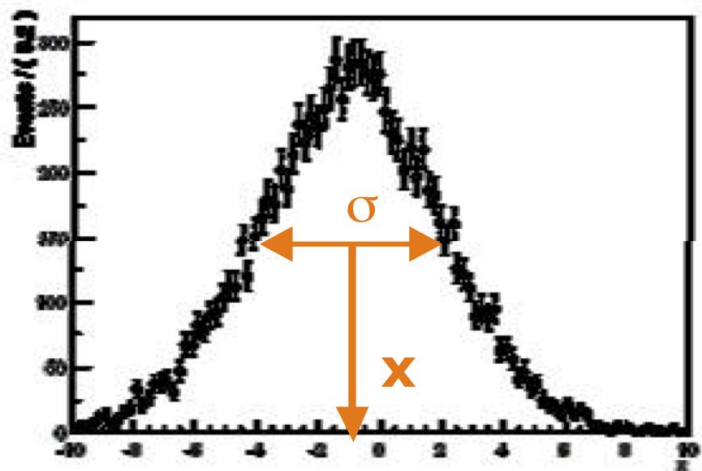
Wat is de waarschijnlijkheidsverdeling?


$$P(\textit{Theory}|\textit{Data}) = \frac{P(\textit{Data}|\textit{Theory})}{P(\textit{Data})} \times P(\textit{Theory})$$

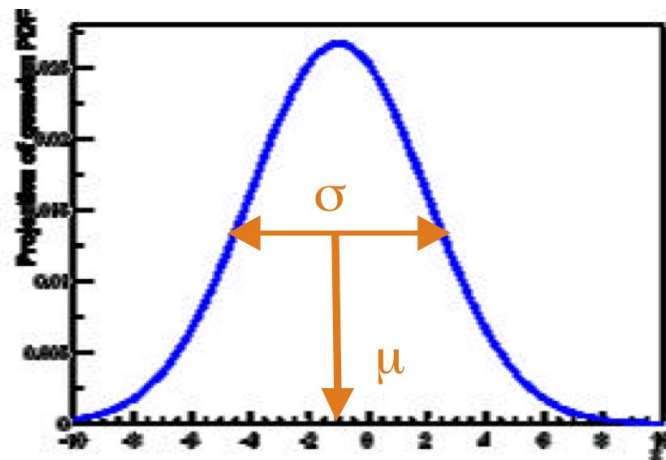
# Gemiddelde en Standaardafwijking

Let op! Standaardafwijking heeft dezelfde letter

## Data Sample



## Parent Distribution (from which data sample was drawn)



$\bar{x}$  - mean of our sample

$\mu$  - mean of our parent dist

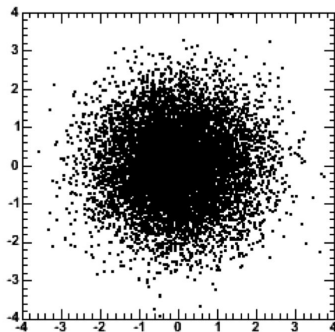


# Standaardafwijking in meer dimensies

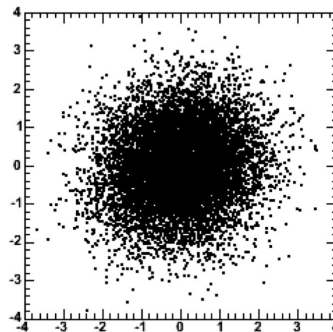
Wat is de spreiding in elke richting?

Wat is de correlatie tussen de variabelen?

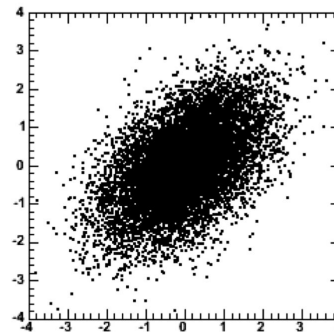
$r = 0$



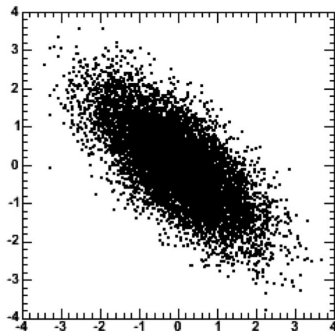
$r = 0.1$



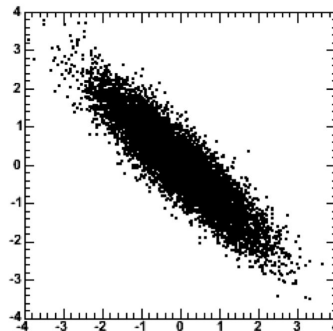
$r = 0.5$



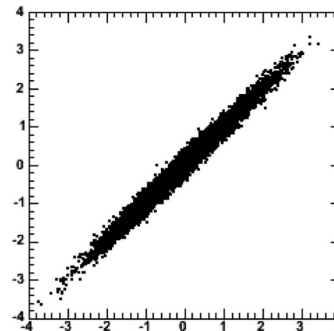
$r = -0.7$



$r = -0.9$



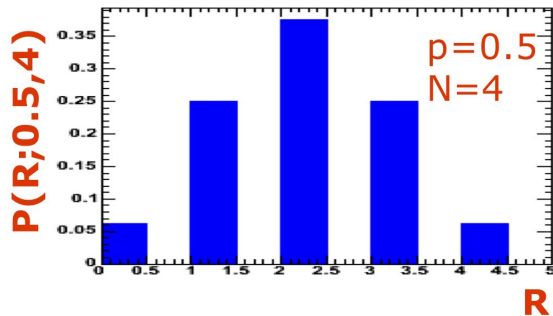
$r = 0.99$



# Binomiale distributie

Voorbeeld: Rode en witte knikkers in een vaas

- Een fractie  $p$  van de knikkers is rood
- Je trekt  $N$  knikkers na elkaar uit de vaas (en legt elke keer de knikker terug)
- $R$  is het aantal rode knikkers dat je trekt

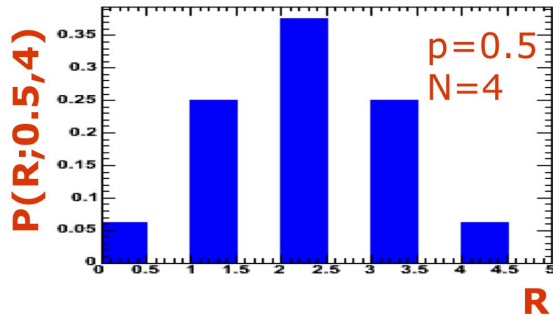


$$P(R; p, N) = p^R (1 - p)^{N-R} \frac{N!}{R!(N - R)!}$$

# Binomiale distributie

Voorbeeld: Rode en witte knikkers in een vaas

- Gemiddelde  $\langle R \rangle = n \cdot p$
- Standaardafwijking  $\sigma = \sqrt{np(1-p)}$



Aantal permutaties van een specifieke uitkomst;  
RWR=RRW=WRR



$$P(R; p, N) = p^R (1-p)^{N-R} \frac{N!}{R!(N-R)!}$$



Kans op een specifieke uitkomst;  
RWWRRR



# Poisson distributie

Some weet je het aantal trekkingen  $N$  niet

- Voorbeeld: Geiger teller
- Tijd-afhankelijke gebeurtenissen

Hoeveel deeltjes verwachten we na een bepaalde tijd?

- Deel de tijd  $\lambda$  op in  $n$  blokjes
- Gebruik de binomiale formule met als kans  $p=\lambda/n$  en laat  $n$  naar oneindig gaan

# Poisson distributie

Some weet je het aantal trekkingen  $N$  niet

$$\langle r \rangle = \lambda$$

$$\sigma = \sqrt{\lambda}$$

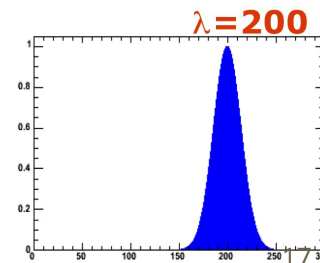
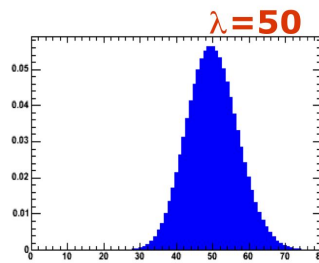
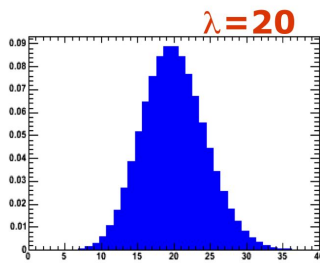
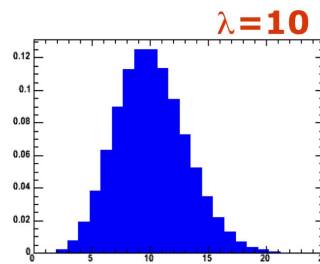
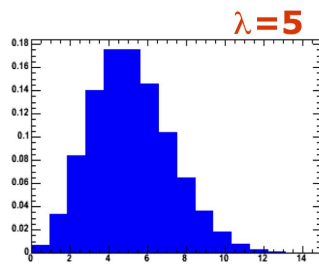
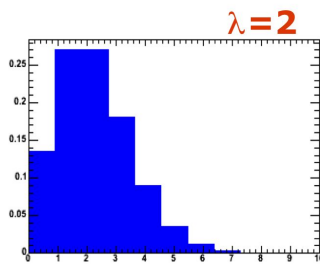
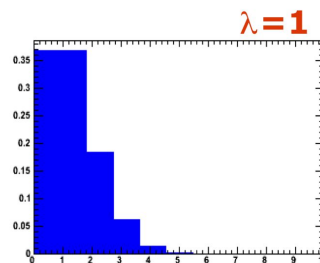
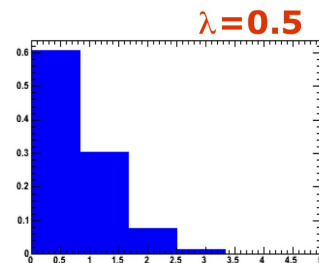
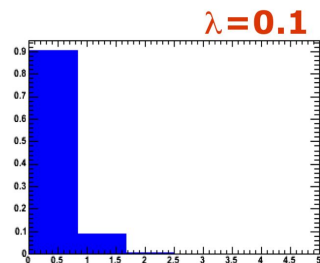
$$P(r; \lambda / n, n) = \frac{\lambda^r}{n^r} \left(1 - \frac{\lambda}{n}\right)^{n-r} \frac{n!}{r!(n-r)!}$$

$$P(r; \lambda) = \frac{e^{-\lambda} \lambda^r}{r!}$$

$$\lim_{n \rightarrow \infty} \frac{n!}{r!(n-r)!} = n^r,$$

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^{n-r} = e^{-\lambda}$$

# Poisson distributie



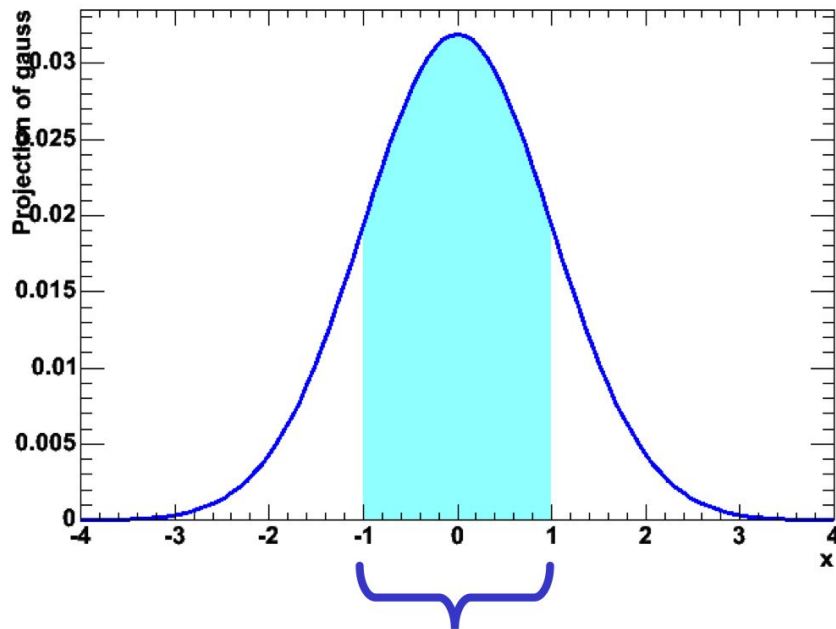


# Gaussiaanse distributie

Poisson distributie in het limiet voor grote  $N$

Integraal van de Gauss

|   |                                |
|---|--------------------------------|
| <b>68.27% within <math>1\sigma</math></b> | 90% $\rightarrow 1.645\sigma$  |
| 95.43% within $2\sigma$                   | 95% $\rightarrow 1.96\sigma$   |
| 99.73% within $3\sigma$                   | 99% $\rightarrow 2.58\sigma$   |
|   | 99.9% $\rightarrow 3.29\sigma$ |



# Onzekerheden (errors)

Experiment is het doen van metingen

- Meting is niet perfect; in-perfectie is beschreven in de resolutie op de error

Errors zijn meestal Gaussiaans voor veel onafhankelijke metingen

- 68% kans dat de echte waarde binnen de errors valt

# Central-limit-theorem

Willekeurige nummers optellen

$N=1; X = x_1$

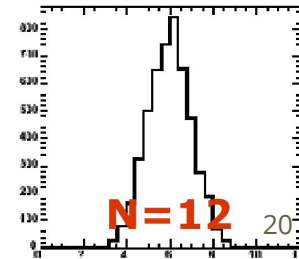
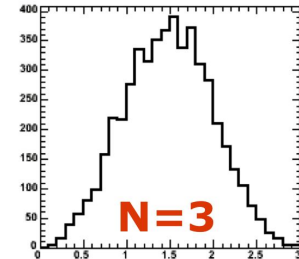
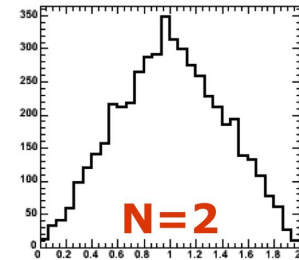
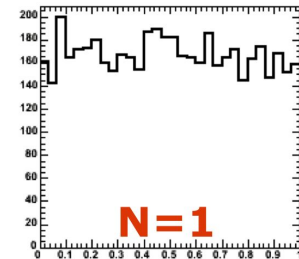
$N=2; X = x_1 + x_2$

$N=3; X = x_1 + x_2 + x_3$

etc.

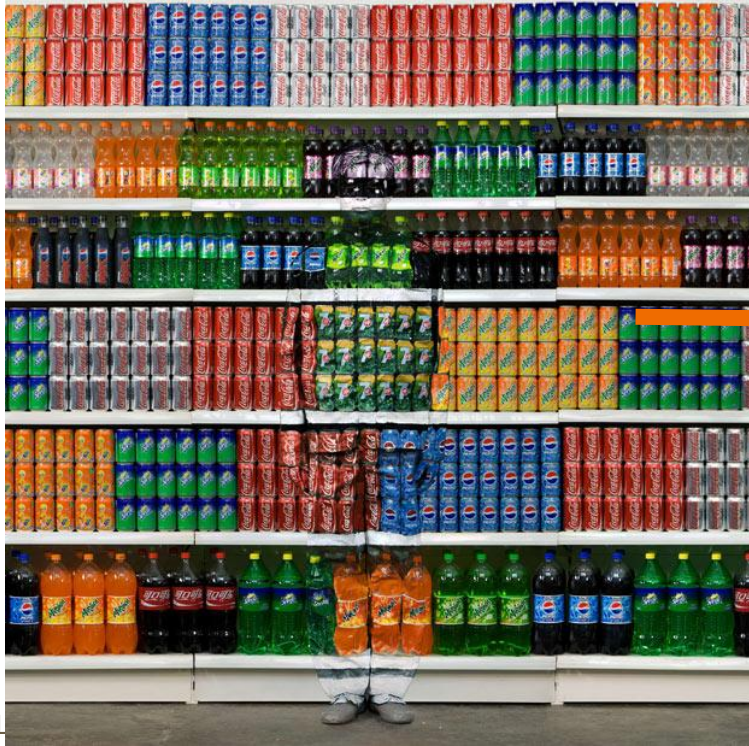
- Meeste errors komen uit veel verschillende onzekerheden waardoor een Gauss een goede aanname is

Gauss 





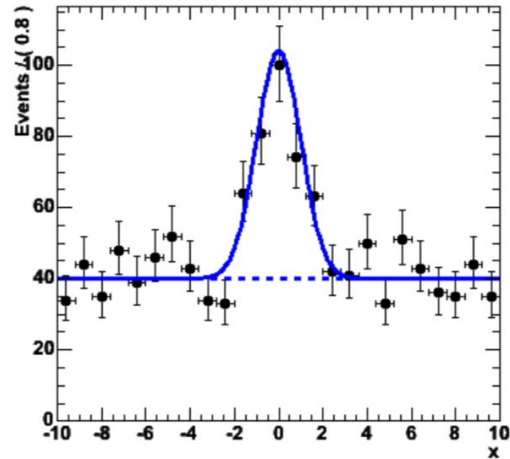
# Intermezzo: Errors klein houden



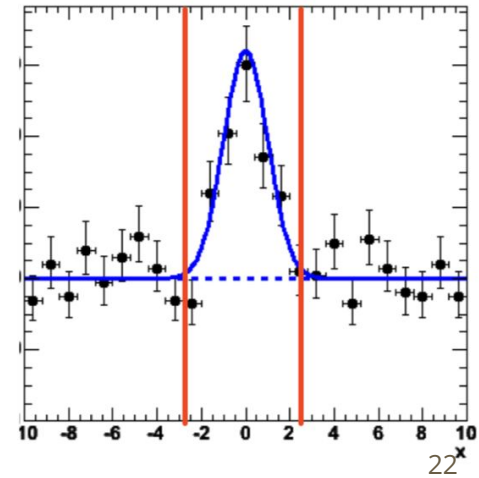
# Intermezzo: Achtergronden verwijderen



Full Sample

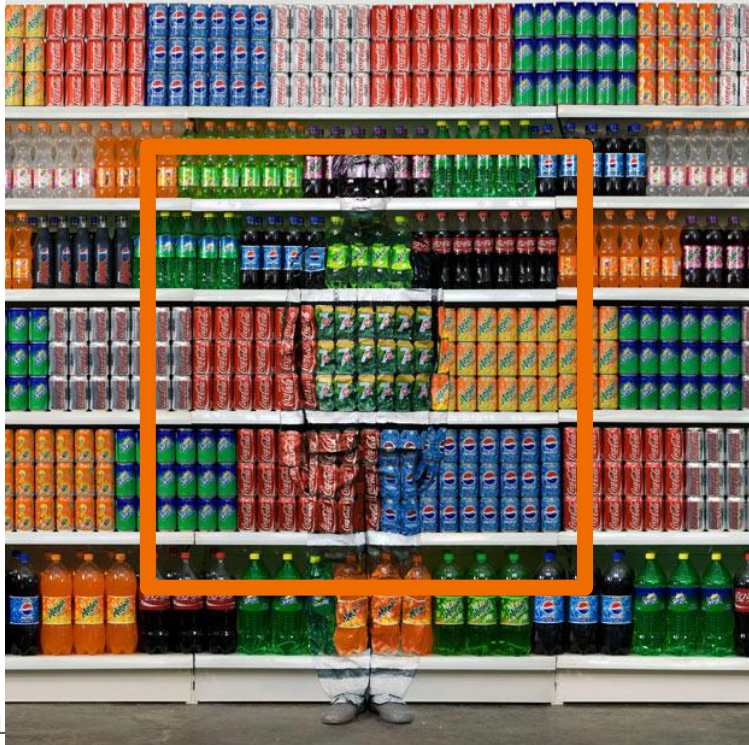


Signal Enriched Sample

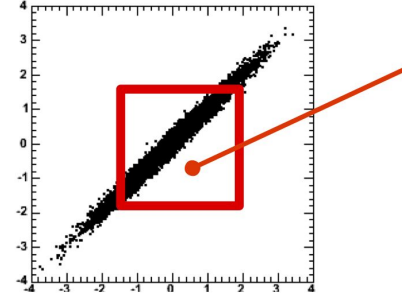
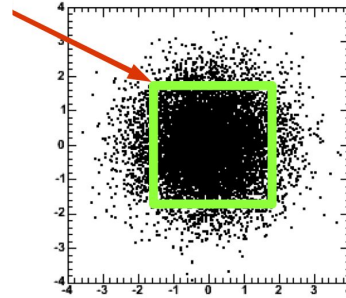




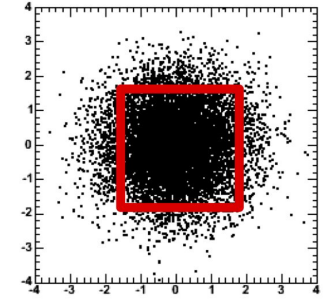
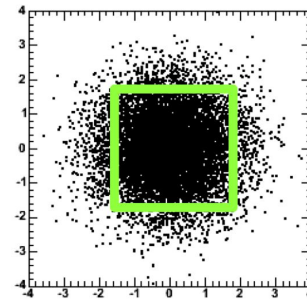
# Intermezzo: Achtergronden verwijderen



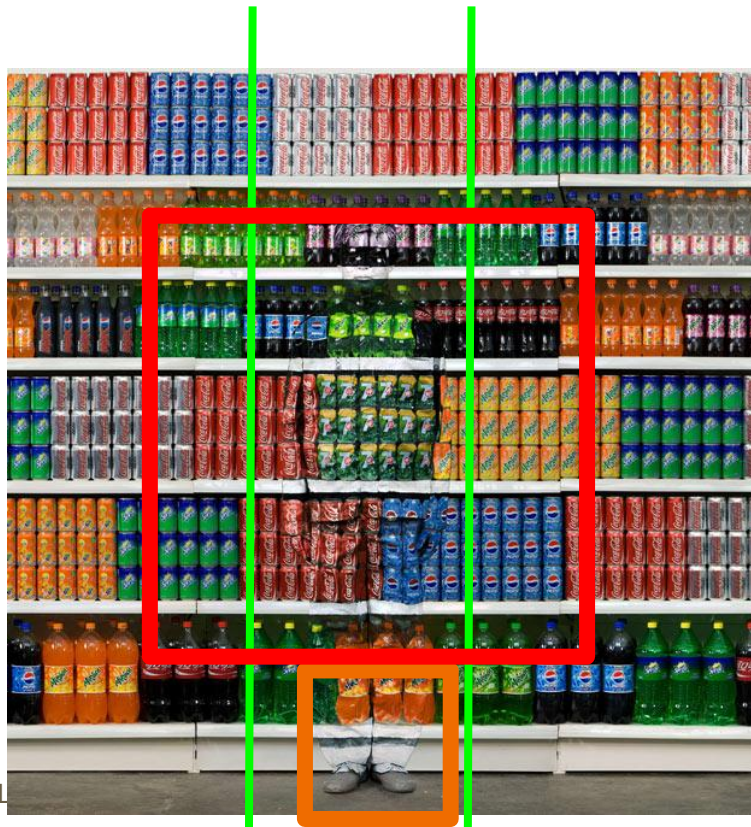
Signal



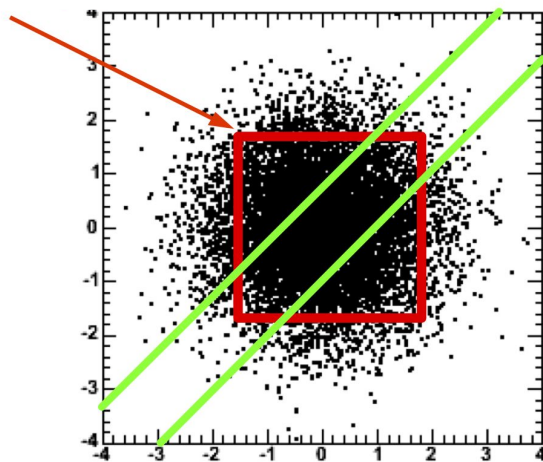
Background



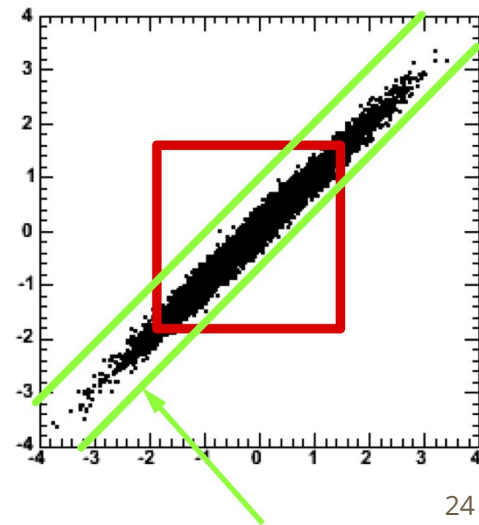
# Intermezzo: Achtergronden verwijderen - opties!



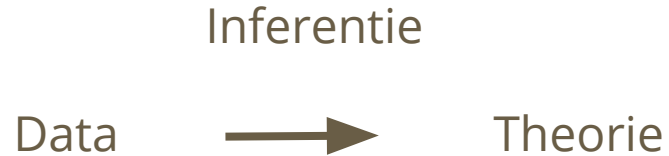
Background



Signal



# Schatten



Methode nodig om de parameters van de theorie distributie te schatten

- Fitten!



# Schatten

De ideale schatter is

- Consistent; in het limiet van oneindig metingen veranderd de geschatte waarde niet
- Onpartijdig; met weinig statistiek krijg je gemiddeld het correcte antwoord
- Efficiënt; De standaardafwijking is klein

De ideale schatter bestaat niet!

- Meest gebruikte schatters zijn de Chi-squared en Likelihood schatters

# Schatten



Waarschijnlijkheid waar (van links naar rechts) Liu Bolin staat

Geschatte plek

# Kanttekening; Betrouwbaarheidsinterval



Aan 1 of 2 kanten snijden?

Wat is nu de kans dat Liu Bolin volledig in het interval valt?

# Kanttekening 2; Systematische errors

Systematische errors zijn **NIET** systematische fouten

- Deze errors komen door bijvoorbeeld het niet goed kunnen modelleren van de detector
- Deze errors komen **niet** doordat mensen herhaaldelijk fouten maken
- Alles wat niet door statistische fluctuaties komt noemen we systematisch

# Hypothese testen

Wat is hypothese testen? Het maken van keuzes

- Is dit deeltje type A of type B?
- Is dit het deeltje wat ik zoek of achtergrond?
- Gaat de kwaliteit van de detector achteruit?
- Komt de data overeen met het standaard model of niet?



# De nulhypothese

Om te laten zien dat een effect bestaat

- Door het eten van wortels krijg je beter nachtzicht
- Adverteren op Facebook verhoogd je verkoop
- Een nieuw medicijn zorgt voor hogere overlevingskansen
- De data bevat nieuwe deeltjes

Moet je je best doen om het omgekeerde (de nulhypothese) te bewijzen, en dat moet mislukken

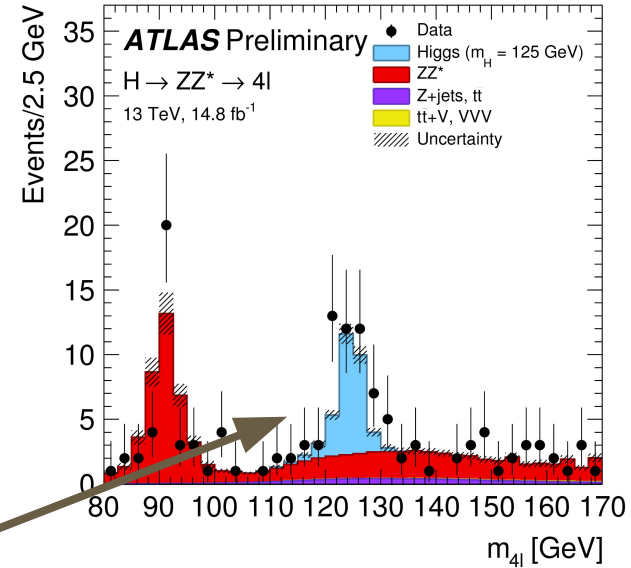
- Mensen die wel en geen wortels eten hebben hetzelfde nachtzicht
- Verkoop is onafhankelijk van Facebook advertenties
- De overlevingskansen met en zonder het nieuwe medicijn zijn hetzelfde
- Het standaard model beschrijft de data goed

# De nulhypothese

Als de nulhypothese niet bij je data past, dan bestaat je effect

- De significantie is de kans dat de nulhypothese wordt afgedankt en toch correct is - je claimt een ontdekking die niet bestaat

Oftewel; Als je data niet overeenkomt met het standaard model, dan is er een nieuw deeltje



De ontdekking van het Higgs boson!

# Experiment

**Telefoon gebruik:** Heb je tijdens dit praatje je telefoon gebruikt?

Wat is de kans dat iemand op zijn/haar telefoon heeft gekeken?

# Experiment

Telefoon gebruik: Heb je tijdens dit praatje je telefoon gebruikt?

- Ik gok dat iedereen ongeveer 1 keer per uur op hun telefoon kijkt, maar heel beleefd is, dus we kiezen  $p=0.5$
- We hebben  $N$  mensen in de zaal in dit praatje
- $R$  is het aantal mensen dat op hun telefoon heeft gekeken

Wat is de kans dat iemand op zijn/haar telefoon heeft gekeken?

# Experiment

Telefoon gebruik: Heb je tijdens dit praatje je telefoon gebruikt?

- Ik gok dat iedereen ongeveer 1 keer per uur op hun telefoon kijkt, maar heel beleefd is, dus we kiezen  $p=0.5$
- We hebben  $N$  mensen in de zaal in dit praatje
- $R$  is het aantal mensen dat op hun telefoon heeft gekeken

$$P(R; p, N) = p^R (1 - p)^{N-R} \frac{N!}{R!(N - R)!}$$

Wat is de kans dat iemand op zijn/haar telefoon heeft gekeken?



# Experiment

Telefoon gebruik: Heb je tijdens dit praatje je telefoon gebruikt?

Wat gebeurt er met de kans als we een andere  $p$  kiezen?

Wat gebeurt er met de kans als we heel veel mensen vragen  
( $N$  wordt oneindig groot)

$$P(R; p, N) = p^R (1 - p)^{N-R} \frac{N!}{R!(N - R)!}$$

# Vragen?