Onderzoeksmethoden: Statistiek 2

Peter de Waal
(gebaseerd op slides Peter de Waal, Marjan van den Akker)

Departement Informatica
Beta-faculteit, Universiteit Utrecht

Vandaag

- Theoretische kansverdelingen
- Z-scores
- Normale verdeling
- Steekproefverdeling
- Centrale limietstelling

Theoretische kansverdelingen

- Worden bepaald door een wiskundige funktie
- Geven theoretische basis of theoretisch model voor een verdeling in een grote of oneindige populatie
- Worden gebruikt om hypotheses te testen
- Worden gebruikt om te modelleren

Basisbegrippen

- **Experiment**: proces waarvan de uitkomst onzeker is,
- **Stochastische variabele**: (numerieke) representatie van de uitkomst van een experiment,
- Stochastische variabele X kan zijn:
  - Discreet
  - Continu

Discrete stochastische variabele X

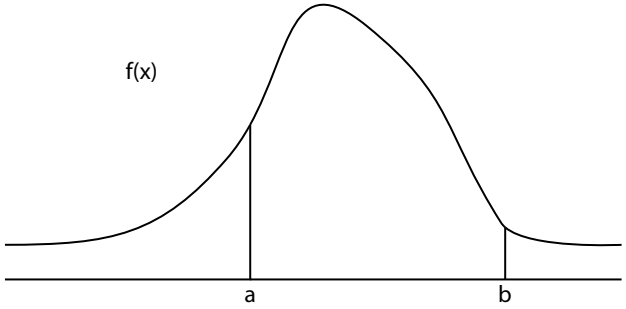
- Mogelijke uitkomsten: $x_1,\ x_2,\ \ldots,\ x_n$.
- **Kansmassafunctie** : beschrijving kansverdeling:
  - $P(X = x_i)$
  - $0 \leq P(X = x_i) \leq 1$
  - $\sum_i P(X = x_i) = 1$

Voorbeeld

- Zuivere munt
- Gooi 4 keer met munt, X is het aantal keren kop.

Continue stochastische variabele X

- Kan elke waarde in een interval aannemen
- **Kansdichtheid** f: (Eng: **probability density function**)
- Totale oppervlakte onder de grafiek is 1
- $P(a \leq X \leq b) = \int_a^b f(x)dx$



Modus, mediaan

Modus:

- Discreet: waarde met maximale kansmassa
- Continu: waarde met maximale kansdichtheid

Mediaan:

- Een mediaan is een waarde m waarvoor geldt:

$$P(X \leq m) \geq 0.5 \quad \text{én} \quad P(X \geq m) \geq 0.5$$

Verwachtingswaarde ("gemiddelde")

Definities

- $\mu = E(X) = \sum_i x_i \cdot P(X = x_i)$
- $\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)dx$

Engels:

- *Mean* of *Expectation*

Algemener:

- $E(g(X)) = \sum_i g(x_i)P(X = x_i)$
- $E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

Eigenschappen:

- $E(cX) = cE(X)$
- $E(X + Y) = E(X) + E(Y)$

Variantie, standaarddeviatie

Variantie

- $\sigma_X^2 = \text{var}(X) = E\Big((X-\mu)^2\Big) = E(X^2) - \mu^2$
- $\text{var}(aX+b) = a^2\text{var}(X)$
- $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$

Standaarddeviatie, standaardafwijking

- $\sigma_X = \sqrt{\text{var}(X)}$

Covariantie, correlatie

Gegeven 2 stochastische variabelen X en Y met

- *Simultane* verdeling: $P(X = x_i \wedge Y = y_j) = p_{i,j}$

Covariantie:

- $\text{cov}(X,Y) = E\Big((X-\mu_X)(Y-\mu_Y)\Big) = \sum_{i,j} p_{i,j}(x_i-\mu_X)(y_j-\mu_Y)$

Correlatie:

- $r_{X,Y} = \dfrac{\text{cov}(X,Y)}{\sigma_x\sigma_Y}$

Onafhankelijkheid

Twee stochastische variabelen X en Y zijn *onafhankelijk* als:

- Continu:

  $P(X \in A \wedge Y \in B) = P(X \in A)P(Y \in B)$ voor *alle* verzamelingen A en B

- Discreet:

  $$P(X = a \wedge Y = b) = P(X = a)P(Y = b) \text{ voor alle } a \text{ en } b$$

Als X en Y onafhankelijke stochastische variabelen zijn, dan:

- $E(XY) = E(X)E(Y)$,
- $\text{cov}(X,Y) = 0$,
- $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$

Theoretische verdelingen: voorbeeld 1

Lars speelt computer game

- Op level 5 is de succes kans per poging 60%
- Bij succes door naar volgende ronde, anders nieuwe poging
- Wat is de kans dat Lars na 4 pogingen doorgaat van level 5 naar level 6?

Geometrische verdeling

- Aantal failures tot het eerste succes bij succes kans p
- $P(X = k) = p(1-p)^k$
- $E(X) = \dfrac{1-p}{p}$

Theoretische verdelingen: voorbeeld 2

- Computerspel van 8 ronden per level
- Elke ronde 40% kans op 1 gouden munt
- Speler mag na 8 ronden altijd door naar volgende level
  - Wat is de kans op 8 munten in een level?
  - Wat is de kans op precies 1 munt?
  - Wat is de kans op precies 5 munten?

Binomiale verdeling

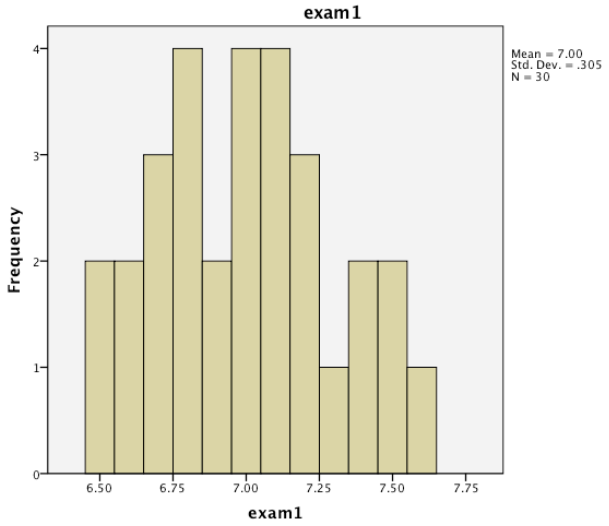
Aantal successen uit n onafhankelijke pogingen met elk succeskans p

- $\binom{n}{k}$ = aantal deelverzamelingen met k elementen uit een verzameling met n elementen.
- $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}$
- $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $E(X) = np$.

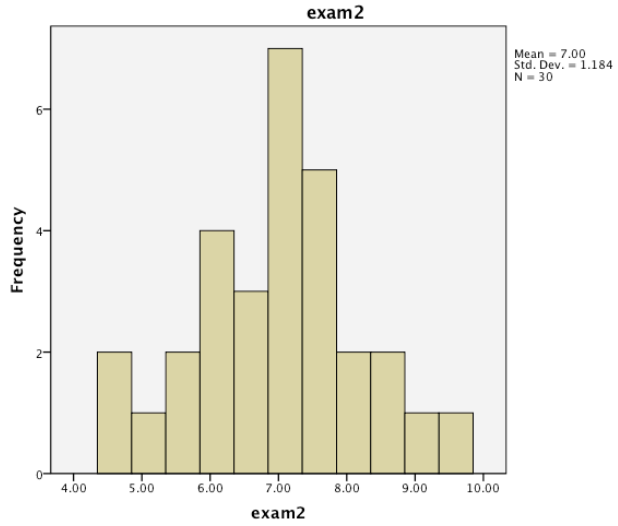
Terug naar de Statistiek: Z-scores

- De beste drie eindcijfers krijgen gratis smartphone.
- Je eindcijfer is 7.6. Wat zegt dat?
- Het gemiddelde is 7.0 Wat nu?
- Het gemiddelde is 8.0 ...
- Het gemiddelde is 7.0 en standaarddeviatie is 0.3
- Het gemiddelde is 7.0 en standaarddeviatie is 1.1

$\overline{X} = 7.0, s = 0.3$



$\overline{X} = 7.0, s = 1.1$



Z-score

- Standaardscore die de afstand tot het gemiddelde uitdrukt in termen van standaardafwijking
  oftewel
- Het aantal standaarddeviaties s dat een bepaalde score X van het gemiddelde $\overline{X}$ af ligt
- $Z = \dfrac{X - \overline{X}}{s}$

  met
- $s = \sqrt{\dfrac{\sum(X - \overline{X})^2}{n-1}}$ is standaarddeviatie van de steekproef.
- $SS = \sum(X - \overline{X})^2$ wordt vaak aangeduid als de *Sum of Squares*.

Z-score: voorbeelden

Data set met gemeten lengtes:

- Gemiddelde 180
- Standaarddeviatie 12

Vragen:

- Wat is Z-score van iemand met lengte $X = 204$? ($Z = 2$)
- Wat is Z-score van iemand met lengte $X = 162$? ($Z = -1.5$)
- Wat is de lengte van iemand met Z-score $Z = 0$? ($X = 180$)

Kenmerken van Z-scores

- Gestandaardiseerde variabelen
- Lineaire transformatie van ruwe scores
- Dimensieloos
- $\overline{Z} = 0$,
- $\sigma_Z^2 = \sigma_Z = 1$.

Steekproef, populatie en nog een stap verder...

- Soms kennen we de kansverdeling (theoretisch model), dan kun je er goed mee rekenen.
- Soms kom je hier moeilijk achter, bijvoorbeeld:
  - Hoe zijn de resultaten van datastructuren verdeeld?
- Vaak kun je aannames maken...

De normale verdeling

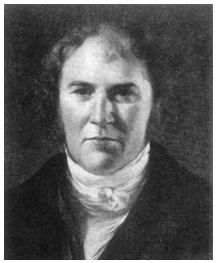
Normale verdeling

De **normale** of **Gaussische** verdeling heeft twee parameters, n.l. μ (gemiddelde), en σ (standaarddeviatie), en heeft dichtheid:

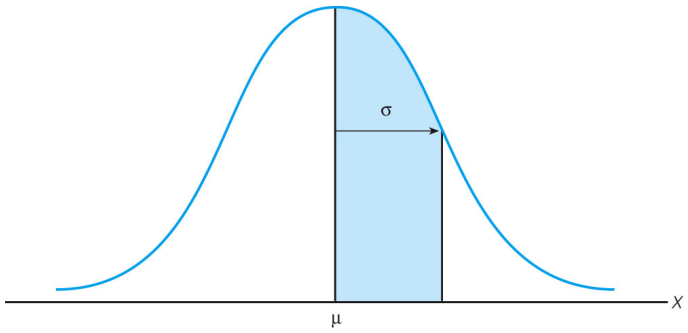
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

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

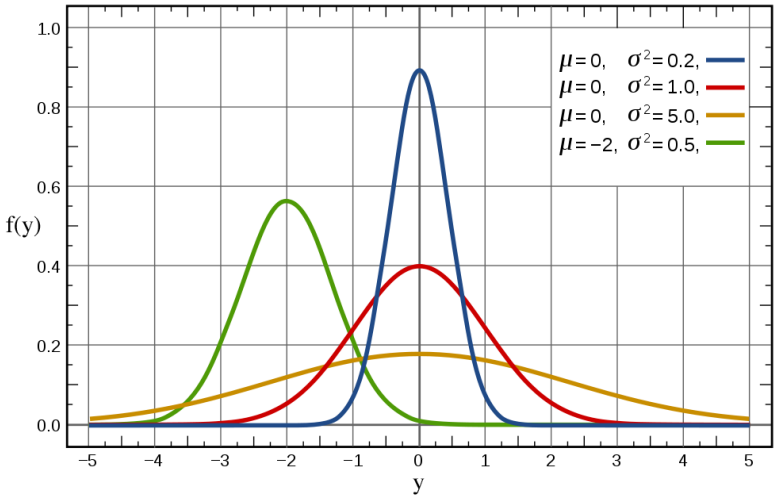
Genoemd naar Carl Friedrich Gauss (1777–1855).
Tegelijkertijd en onafhankelijk ontwikkeld door Gauss (1808) en Adrain (1809)



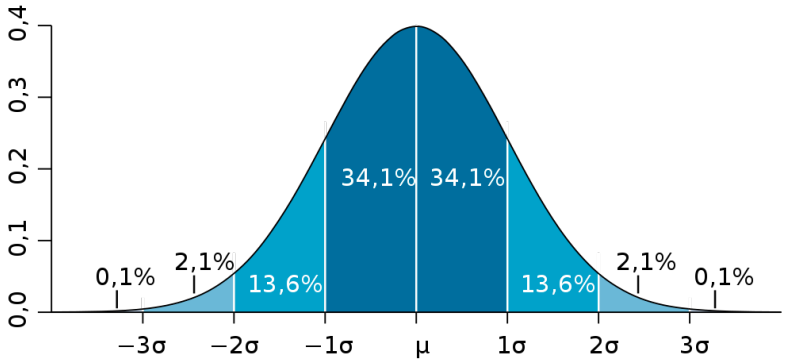
De normale verdeling



De normale verdeling, voorbeelden



Kansen in normale verdelingen



Eigenschappen

- Symmetrisch
- Continue variabele
- Mediaan = Modus = Gemiddelde
- Van veel natuurlijke eigenschappen wordt aangenomen dat zij bij benadering een normale verdeling volgen.
- Voorbeelden: lengte van Nederlanders, cholesterolgehalte in het bloed, IQ, inhoud van pak rijst...
- Afwijkingen, fouten, volgen normale verdeling, 'witte ruis'

De standaardnormale verdeling

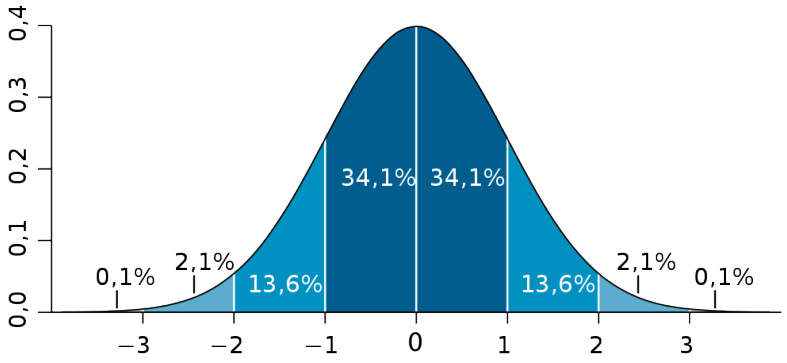
- De *scores van een normale verdeling* omzetten in standaardscores (Z) levert de *standaardnormaalverdeling*
- Deze heeft altijd een gemiddelde van 0, en een standaardafwijking van 1
- Dus als X normaal verdeeld is met gemiddelde μ en standaardafwijking σ (Notatie: $X \sim N(\mu, \sigma^2)$) dan is

$$Z = \frac{X - \mu}{\sigma}$$

standaardnormaal verdeeld (dus $Z \sim N(0, 1)$!).

- Terugtransformatie: $X = \mu + \sigma Z$.

Kansen voor de standaardnormale verdeling



Deze kansen (en meer) vind je in de tabellen op pag. 290 en 291.

Kansen berekening voor normale verdelingen

Toepassingen van de normale verdeling als norm: Kansuitspraken in twee "richtingen"

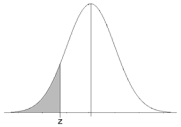
Stel: de lengte van mannen in een populatie is normaal verdeeld met $\mu = 180$ en $\sigma = 12$?

1. Wat is het 90^e percentiel ?
2. Hoeveel procent van de mannen is groter dan 186?
3. Is het waarschijnlijk dat een gevonden score van 190 of meer afkomstig is uit een normaal verdeelde populatie met $\mu = 180$ en $\sigma = 12$?

De tweede richting speelt een cruciale rol bij toetsende (=verklarende) statistiek

De standaardnormaal tabel

Standard Normal Cumulative Probability Table



Cumulative probabilities for NEGATIVE z-values are shown in the following table:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985

De standaardnormaal tabel (vervolg)

Standard Normal Cumulative Probability Table

Cumulative probabilities for POSITIVE z-values are shown in the following table:

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
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890

Voorbeeld

Stel: de lengte van mannen in een populatie is normaal verdeeld met $\mu = 180$ en $\sigma = 12$?

1. Wat is het 90^e percentiel ?

In Z-score:

Antwoord A (Tabel p.290): $z = 1.28$, want de oppervlakte links van -1.28 is ongeveer 0.10.

Antwoord B (Tabel p.291): $z = 1.28$, want de oppervlakte links van 1.28 is ongeveer 0.90.

Voorbeeld (vervolg 1)

We weten dus $P(Z \geq 1.28) = 0.10$ (en ook $P(Z \leq 1.28) = 0.90$)

Combineer met $\mu = 180$ en $\sigma = 12$:

$$\begin{aligned} 0.10 &= P(Z \geq 1.28) \\ &\quad (\text{transformeer } Z\text{-score } = 1.28 \rightarrow X: \ 180 + 1.28 \times 12 = 195.36) \\ &= P(X \geq 195.36) \end{aligned}$$

Dus $P_{90} = 195.36$

Voorbeeld (vervolg 2)

Stel: de lengte van mannen in een populatie is normaal verdeeld met $\mu = 180$ en $\sigma = 12$?

1. Hoeveel procent van de mannen is groter dan 186 cm?

$$\begin{aligned} ?? &= P(X \geq 186) \\ &\quad (\text{transformeer } X \rightarrow Z\text{-score: } 186 \rightarrow \tfrac{186-180}{12} = 0.50) \\ &= P(Z \geq 0.50) \\ &= 0.3085 \qquad (\text{Tabel p.290}) \end{aligned}$$

Antwoord: 30.85%

Check!

Q: Een Z-score heeft altijd een standaardnormale verdeling?

A: Nee, alleen als de (niet-genormaliseerde) variable X normaal verdeeld is.

Reality check

Q: Wat als scores uit de "echte wereld" helemaal niet lijken op een theoretische verdeling?

Oplossing:

- Kijk niet naar scores van individuen, maar naar scores van steekproeven
- Verbind deze geobserveerde scores (van bijv. gemiddelden van steekproeven) met de theoretische norm m.b.v. de centrale limietstelling

Ook:

- We weten vaak niet de populatieparameters (als μ, σ).
- Probeer deze te achterhalen met steekproef.
- Hoe verhouden observaties uit steekproef zich tot de (theoretische) populatieverdeling?

Oplossing

Steekproefgemiddelde

$$\overline{X}(n) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Het steekproefgemiddelde, zoals in bovenstaande formule is een stochastische variabele en heeft dus een verdeling, verwachting, variantie, ...

Q: Hoe bepalen we die?

Centrale limietstelling

- Steekproef $X_1, \ldots, X_n$: onafhankelijk, en identiek *willekeurig* verdeeld, gemiddelde μ en variantie σ^2.
- $\overline{X}(n) = (X_1 + X_2 + \cdots + X_n)/n$

Stelling

Als $n \to \infty$, dan is $\overline{X}(n)$ bij benadering normaal verdeeld met gemiddelde μ en variantie σ^2/n.

- $\frac{\sigma}{\sqrt{n}}$ heet de **standaardfout** (van het gemiddelde)
- In de praktijk is $n \geq 30$ voldoende.
- $\overline{X}(n) \sim N(\mu, \sigma^2)$, dus $Z(n) = \dfrac{\overline{X}(n) - \mu}{\left(\sigma/\sqrt{n}\right)} \sim N(0,1)$ bij benadering.

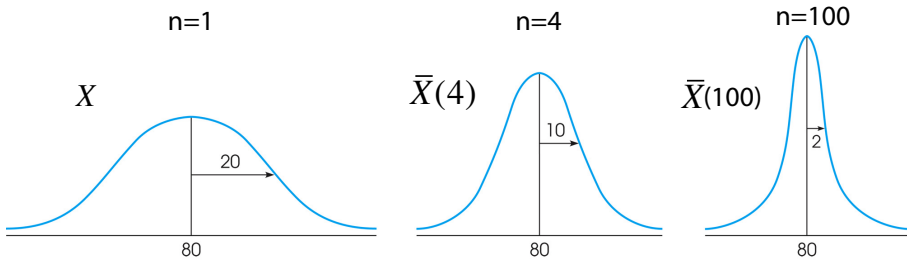
Centrale limietstelling: wat betekent dit?

- Theoretische of steekproefverdeling:
  - Meet individuen op variabele X
  - Herhaal dit n keer (b.v. $n = 1000$) en noteer individuele waarden
  - Histogram representeert **steekproefverdeling** (Eng: **sample distribution**)
- Verdelingen van steekproeven
  - Neem een steekproef van n individuen en meet en bepaal steekproefgemiddelde $\overline{X}(n)$.
  - Herhaal dit m keer (b.v. $m = 1000$) en noteer de afzonderlijke steekproefgemiddelden
  - Histogram representeert de **verdeling van steekproeven** (Eng: **sampling distribution**)

Voorbeeld 1

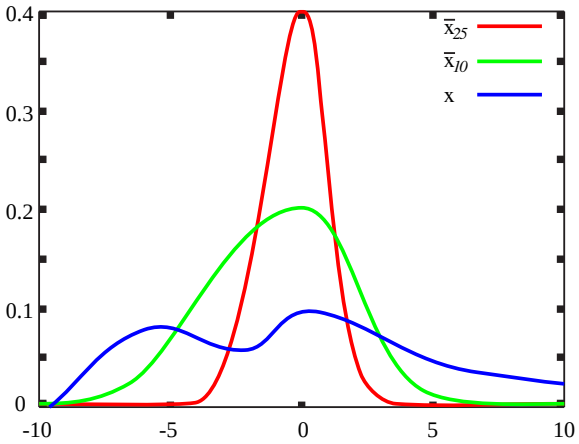
Voorbeeld

- Gemeten variabele: gewicht X
- Normale verdeling ($\mu = 80$ kg, $\sigma = 20$ kg).



Voorbeeld 2

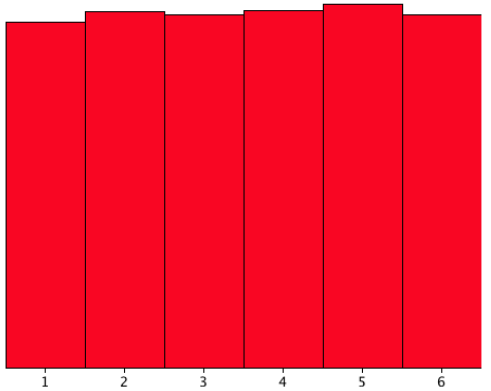
- Stochastische variabele X met gemiddelde $\mu = 0$ en $\sigma = 5$:



Voorbeeld 3

Voorbeeld: worp van één dobbelsteen:

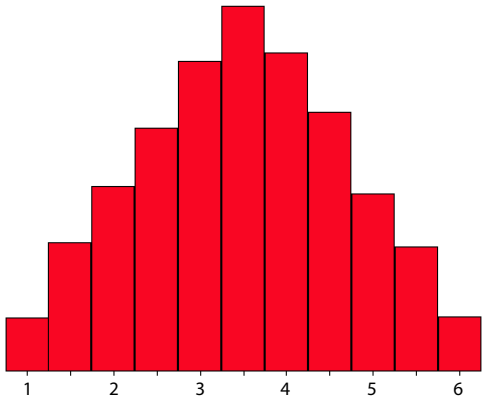
- Elke steekproef = 1 worp:
  histogram: tel aantal ogen (sample distribution)



Voorbeeld 3

Voorbeeld: worp van twee dobbelstenen:

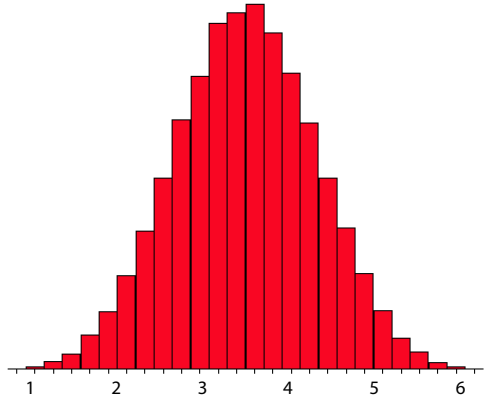
- Elke steekproef = 2 worpen:
  histogram: tel gemiddeld aantal ogen (steekproef verdeling bij steekproefgrootte $n = 2$)



Voorbeeld 3

Voorbeeld: worp van vijf dobbelstenen:

- Elke steekproef = 5 worpen:
  histogram: tel gemiddeld aantal ogen (steekproef verdeling voor $n = 5$)



Tot zover...

- Theoretische verdelingen
  - Geometrische verdeling
  - Binomiaalverdeling
  - Normale verdeling
- Z-scores
- Centrale Limietstelling

Volgende keer:

- t-verdeling
- Toetsen van hypotheses.