Statistiek voor A.I.

College 9

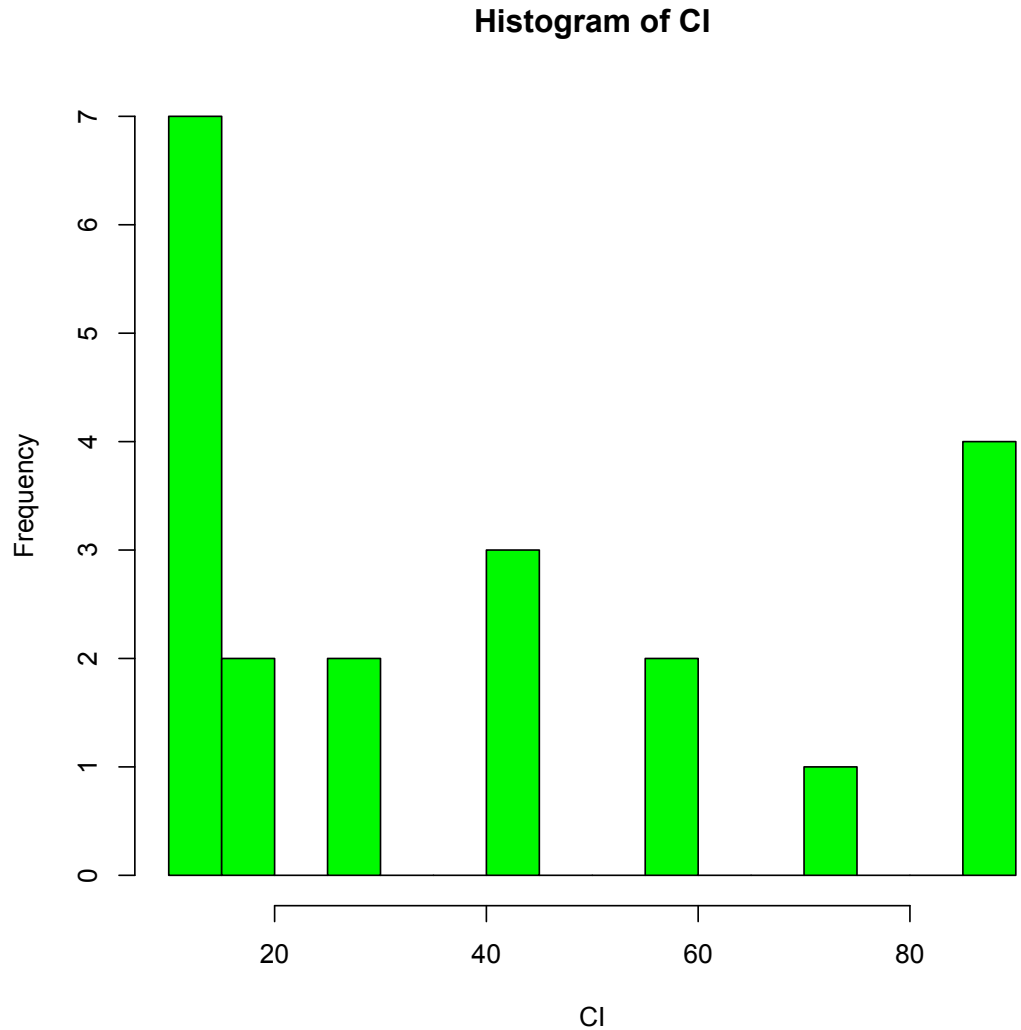
Donderdag 11 Oktober

2 Deductieve statistiek

Bayesiaanse statistiek

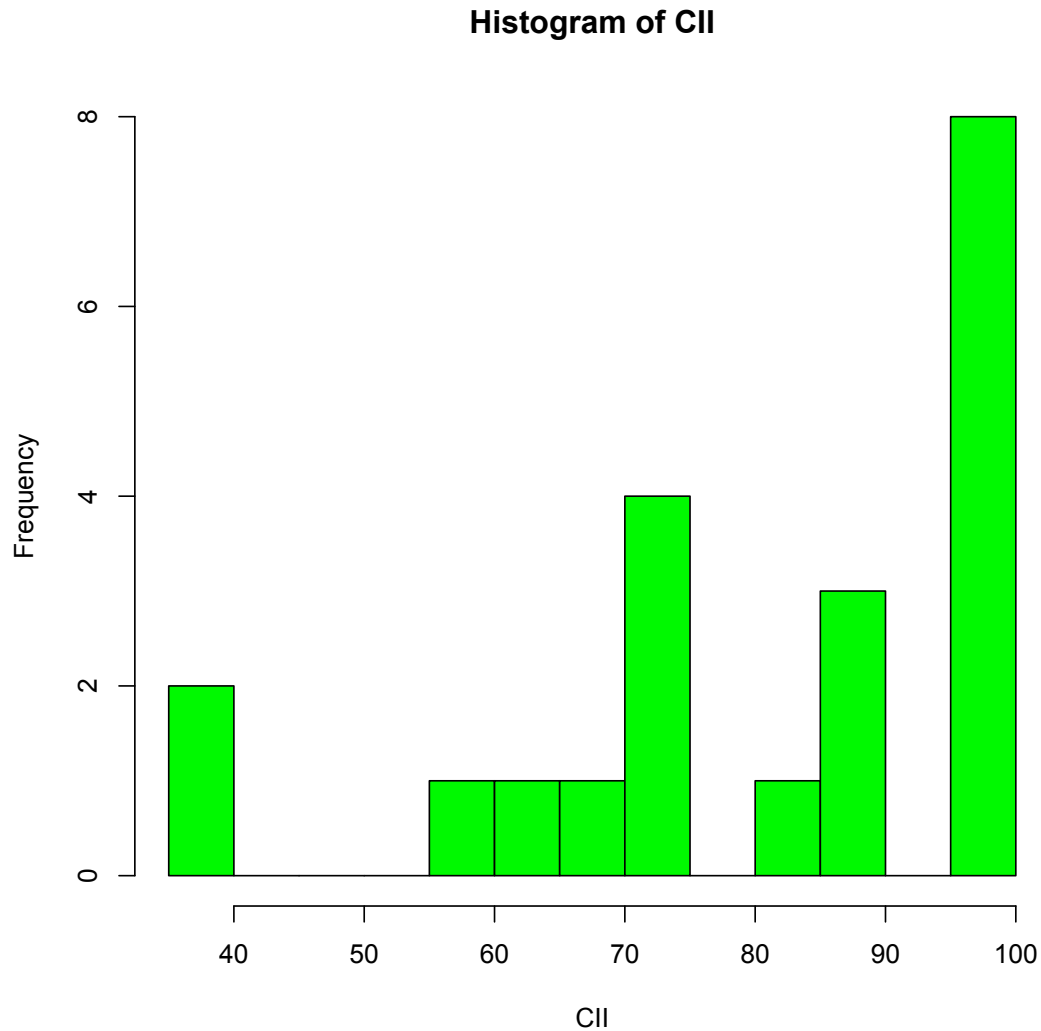
Jullie - onderzoek Tim

Reistijd naar college (minuten).



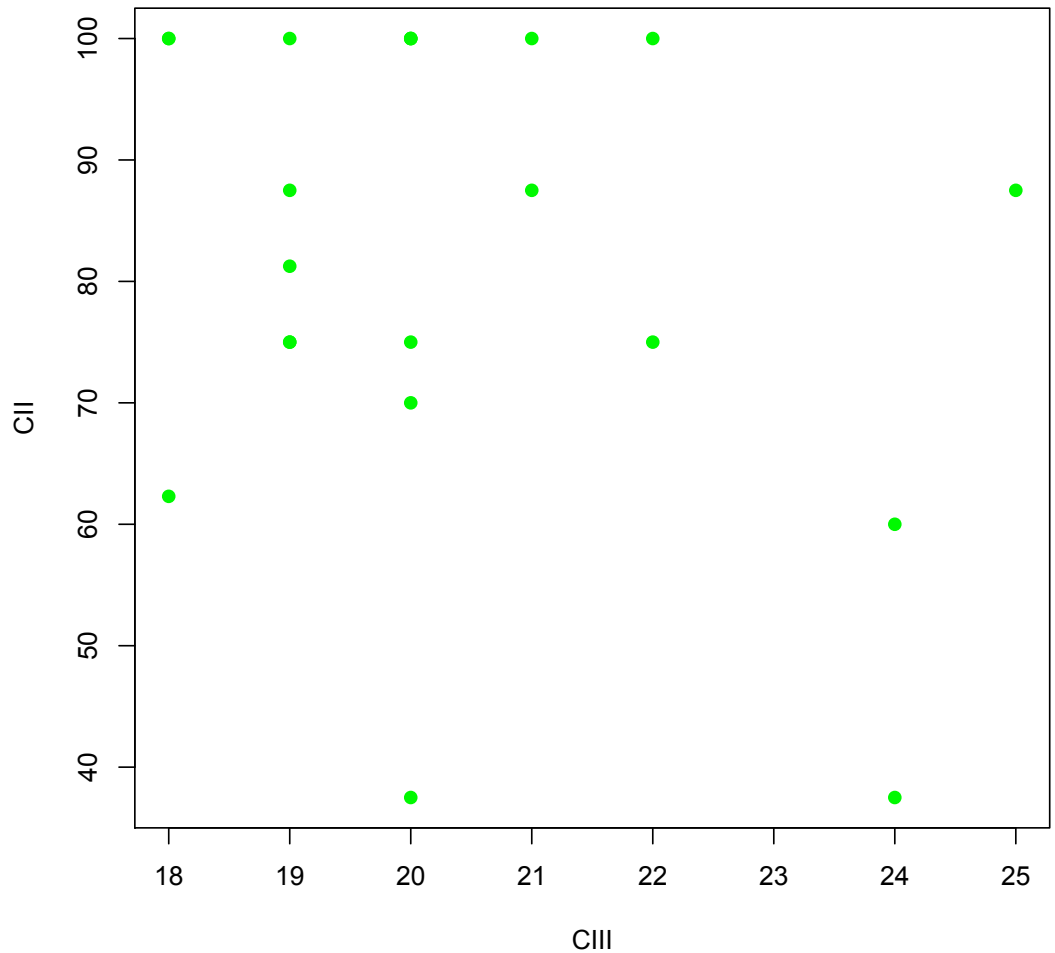
Jullie - onderzoek Tim

Percentage behaalde vakken.



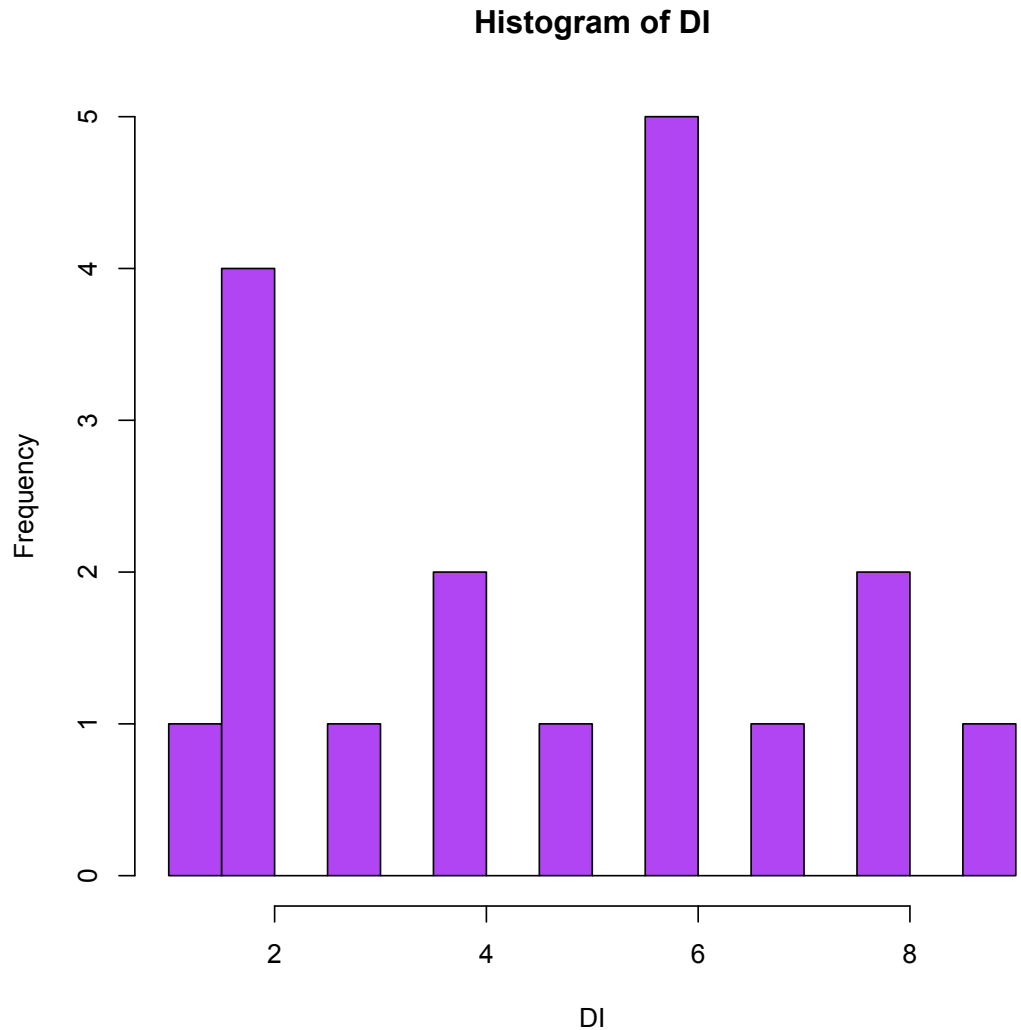
Jullie - onderzoek Tim

Horizontaal: leeftijd.
Verticaal: percentage behaalde vakken.

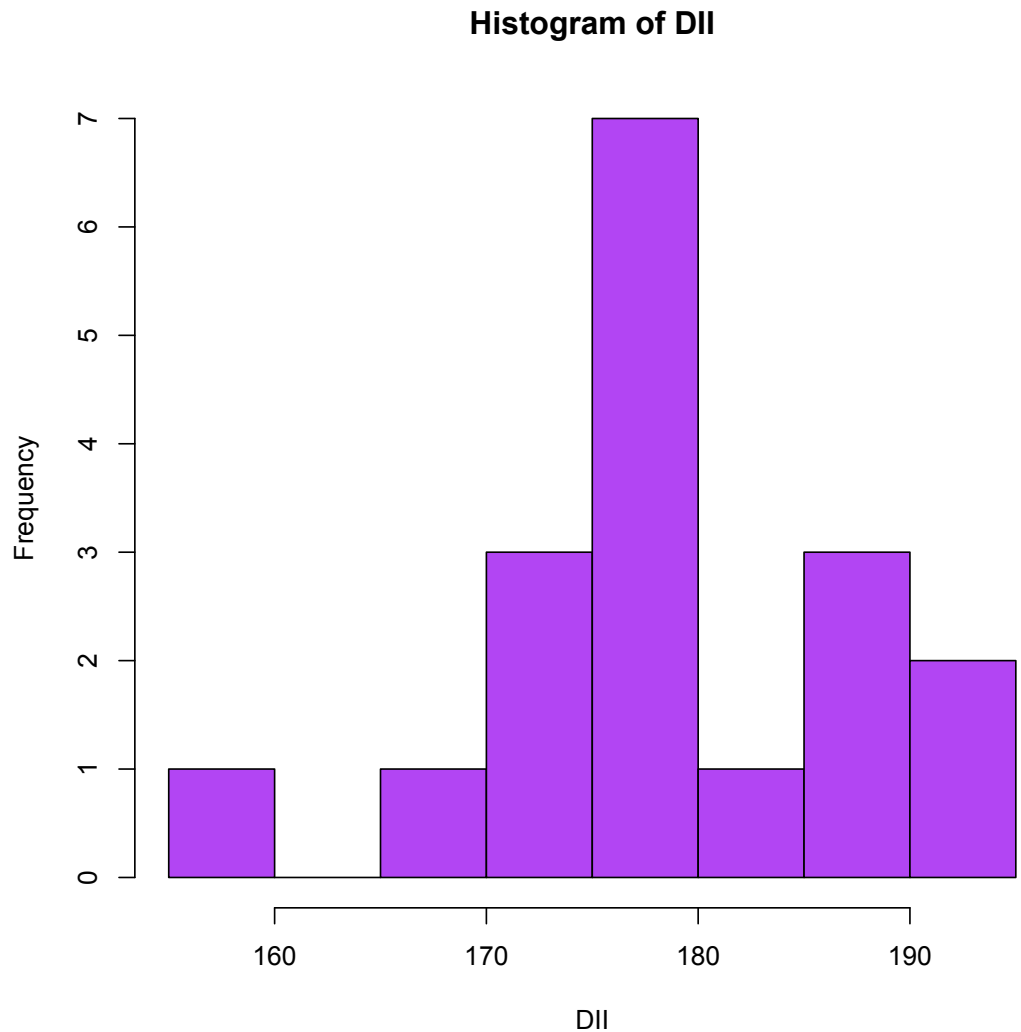


Correlatiecoëfficiënt: -0.29.

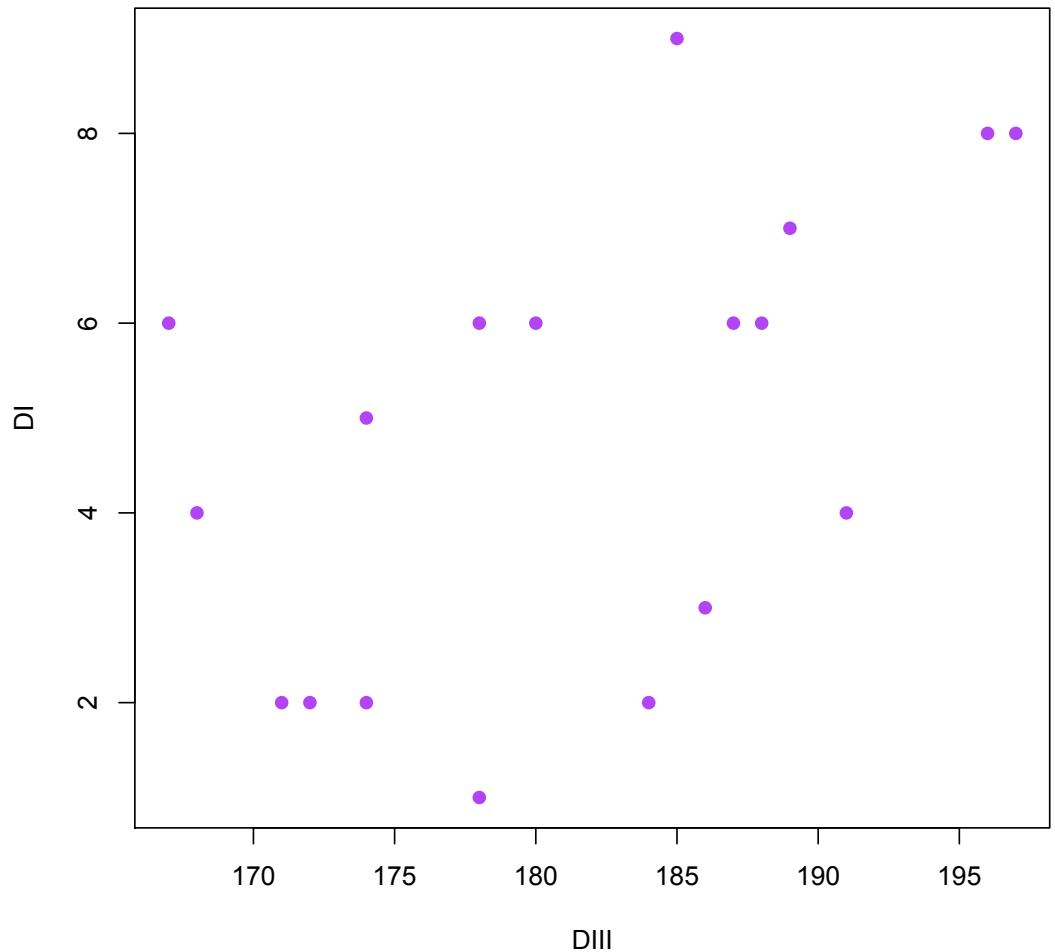
Aantal boterhammen per dag.



Lengte in cm.



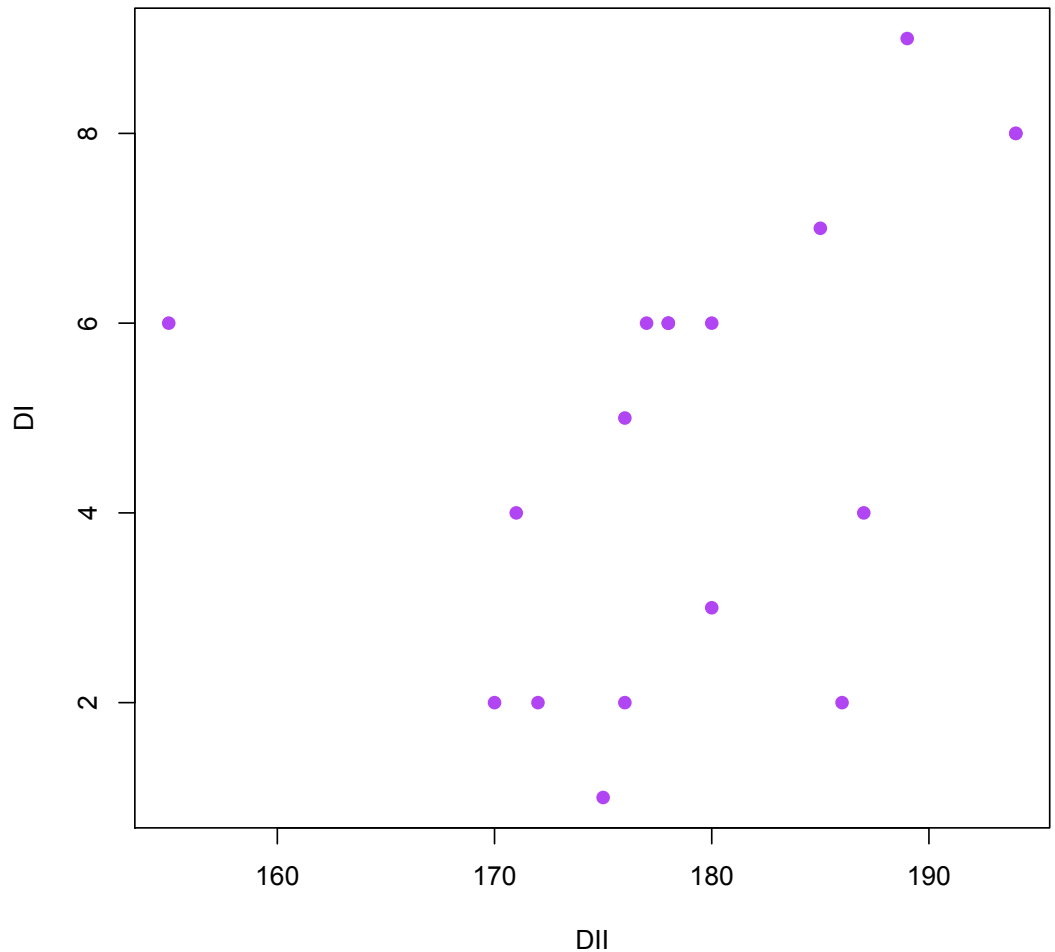
Horizontaal: lengte.
Verticaal: aantal boterhammen per dag.



Correlatiecoëfficiënt: 0.42.

Jullie - onderzoek Geert-Jan, Joris, Brechje

Horizontaal: lengte tussen topjes middelvingers met gestrekte armen.
Verticaal: aantal boterhammen per dag.

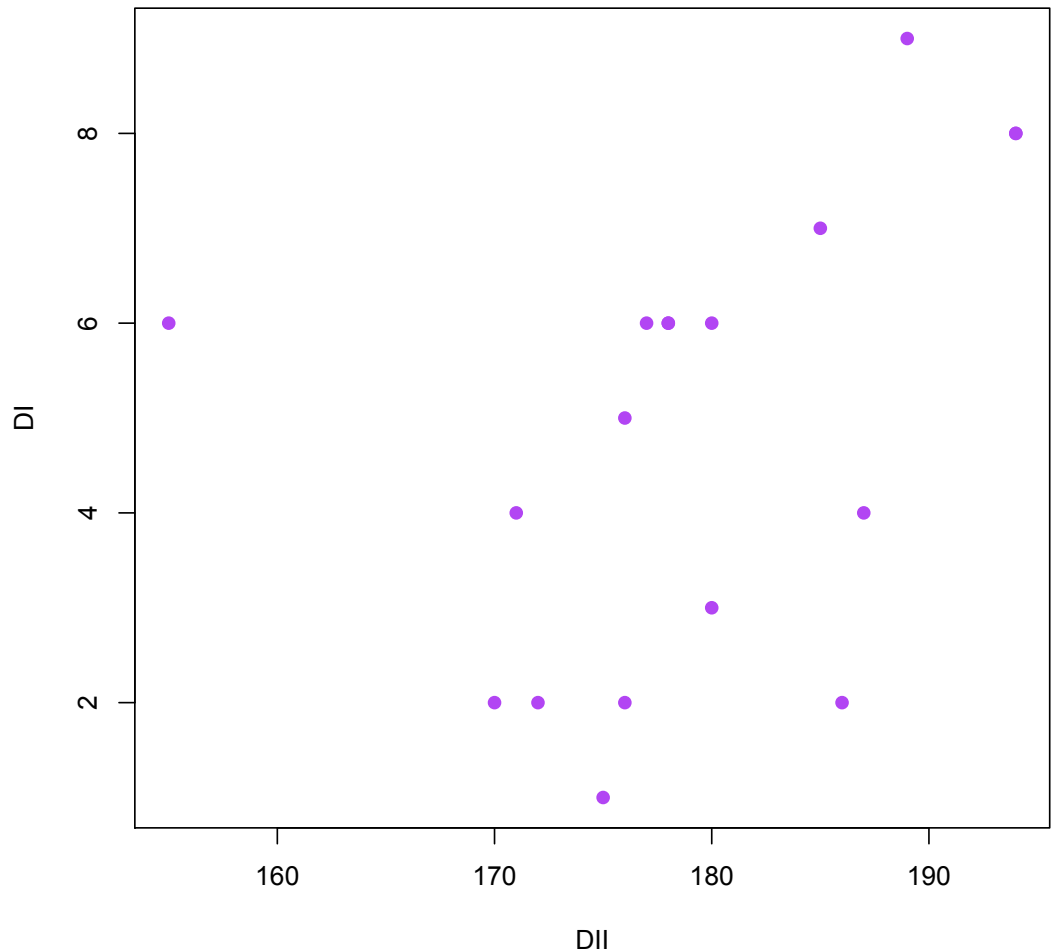


Correlatiecoëfficiënt: 0.51.

Jullie - onderzoek Geert-Jan, Joris, Brechje

Horizontaal: lengte

Verticaal: lengte tussen topjes middelvingers met gestrekte armen.



Correlatiecoëfficiënt: 0.89

Vandaag :

- Bayesiaans leren
- Belangrijke verdelingen
- Normaliseren

Bayesiaans leren

Definitie Bij een *classificatie probleem* moeten data die uit rijtjes eigenschappen bestaan geclassificeerd worden. De uitkomstenruimte bestaat uit de hypothesen, die alle mogelijke volledige classificaties vormen.

Als de data uit rijtjes van m eigenschappen bestaan, en er zijn n mogelijke classificaties (klassen) $k_1, \ldots, k_n$, dan moet elke hypothese dus elk van de 2^m mogelijke combinaties van eigenschappen classificeren. Er zijn dus n^{2^m} hypothesen, die bestaan uit rijtjes ter lengte $m+1$ (m eigenschappen en de classificatie).

Er wordt geleerd op grond van geclassificeerde rijtjes. Bij data "de combinatie van eigenschappen $r_1, \ldots, r_m$ wordt in k_i geclassificeerd" (d.w.z. informatie $r = (r_1, \ldots, r_m, k_i)$) wordt een Bayesiaanse update uitgevoerd, waarbij r wordt geïdentificeerd met de gebeurtenis $E_r = \{H \mid r \in H\}$. Dat geeft likelihood

$$P(r \mid H) = \begin{cases} 1 & \text{als } r \in H \\ 0 & \text{als } r \notin H. \end{cases}$$

De a-posteriori waarschijnlijkheden zijn:

$$P(H_i \mid r) = \frac{P(r \mid H_i)P(H_i)}{\sum_{j=1}^{n^{2^m}} P(r \mid H_j)P(H_j)} = \begin{cases} \frac{P(H_i)}{\sum_{\{j \mid r \in H_j\}} P(H_j)} & \text{als } r \in H_i \\ 0 & \text{als } r \notin H_i. \end{cases}$$

Als bij een classificatie probleem de hypothesen aanvankelijk even waarschijnlijk worden geacht dan heeft het Bayesiaans leren de volgende simpele vorm.

Op grond van data d zijn de a-posteriori kansen:

$$P(H \mid d) = \begin{cases} \frac{1}{\text{het aantal hypothesen dat } d \text{ bevat}} & \text{als } d \in H \\ 0 & \text{als } d \notin H. \end{cases}$$

Voorbeeld Een diagnostisch algoritme (DA) moet een zekere ziekte Z leren diagnostiseren op grond van drie symptomen S_1, S_2, S_3. Als leerdata krijgt het algoritme rijtjes ($r = r_1 r_2 r_3 r_4$), waarbij r_4 staat voor niet ziek (0) of ziek (1) en de andere r_i voor het wel (0) of niet (1) hebben van symptoom S_i.

Een hypothese bestaat uit een maximaal grote consistente verzameling van rijtjes. De uitkomstenruimte bestaat uit de $2^{2^3} = 256$ hypothesen, die aanvankelijk dezelfde waarschijnlijkheid hebben. De likelihoods zijn dus:

$$P(r \mid H) = \begin{cases} 1 & \text{als } r \in H \\ 0 & \text{anders.} \end{cases}$$

De eerste data bestaan uit een zieke patiënt met symptomen S_2 en S_3. Daardoor vallen alle hypothesen die 0111 niet bevatten af, en de overige zijn allen even waarschijnlijk:

$$P_1(H) = P_0(H \mid 0111) = \begin{cases} \frac{1}{128} & \text{als } 0111 \in H \\ 0 & \text{anders.} \end{cases}$$

De volgende data bestaan uit een gezond persoon met alle symptomen: 1110. Dat geeft:

$$P_2(H) = P_1(H \mid 1110) = \begin{cases} \frac{1}{64} & \text{als } 0111, 1110 \in H \\ 0 & \text{anders.} \end{cases}$$

Etc.

Voorbeeld Een automatische vertaler moet het woord *bank* in een tekst op grond van de voorkomens van de woorden *stoel*, *kamer*, *geld* en *kantoor* vertalen/classificeren als *coach* of als *bank*.

De mogelijke voorkomens van de woorden zijn rijtjes $w_1 \ldots w_4$, waar $w_i = 1$ als het i-de woord van (*stoel*,*kamer*,*geld*,*kantoor*) in de tekst voorkomt en 0 anders. Een hypothese classificeert elk rijtje als 1 (vertaal als *coach*) of 0 (vertaal als *bank*). Er zijn dus $2^{2^4} = 65536$ hypothesen.

Hypothesen die 11000 of 00111 bevatten kijgen a-priori waarschijnlijkheid 0. De overige 16384 hypothesen hebben a-priori waarschijnlijkheid $\frac{1}{8192} = 0.00012$ als ze geen rijtje $11xy0$ bevatten en $\frac{1}{24576} = 0.00004$ anders. $\mathcal{H}$ staat voor de verzameling hypothesen van de eerste soort, dus $|\mathcal{H}| = 2^{12} = 4096$.

Op grond van data 01011 zijn de a-posteriori waarschijnlijkheden:

$$P(H_i \mid 01011) = \begin{cases} \frac{P(H_i)}{\sum_{\{j \mid 01011 \in H_j\}} P(H_j)} & \text{als } 01011 \in H_i \\ 0 & \text{anders.} \end{cases}$$

$$= \begin{cases} \frac{\frac{1}{8192}}{\frac{12288}{2 \cdot 24576} + \frac{4096}{2 \cdot 8192}} = 0.00024 & \text{als } 01011 \in H_i \in \mathcal{H} \\ \frac{\frac{1}{24576}}{\frac{12288}{2 \cdot 24576} + \frac{4096}{2 \cdot 8192}} = 0.00008 & \text{als } 01011 \in H_i \notin \mathcal{H} \\ 0 & \text{anders.} \end{cases}$$

Definitie Beschouw een classificatie probleem waarbij op grond van m eigenschappen een classificatie in één van de n klassen $k_1, \ldots, k_n$ gegeven moet worden. Gegeven een rijtje $r = (r_1, \ldots, r_m)$, is rk_i een afkorting voor $(r_1, \ldots, r_m, k_i)$. Twee bekende classificeerders zijn:

De *MAP-classificeerder* ν_{MAP}:

$$\nu_{\mathrm{MAP}}(r) = j \text{ precies dan als } rk_j \in h_{\mathrm{MAP}}.$$

De *optimale Bayesiaanse classificeerder* ν_{OB}:

$$\nu_{\mathrm{OB}}(r) = \mathrm{argmax}_j \sum_{i=1}^{n} P(rk_j \mid H_i) P(H_i).$$

Belangrijke verdelingen

Definitie Een continue stochast X heeft een *exponentiële verdeling* met *parameter* λ als de kansdichtheid $f : [0, \infty)$ deze vorm heeft:

$$f(x) = \lambda e^{-\lambda x}.$$

De cumulatieve distributiefunctie:

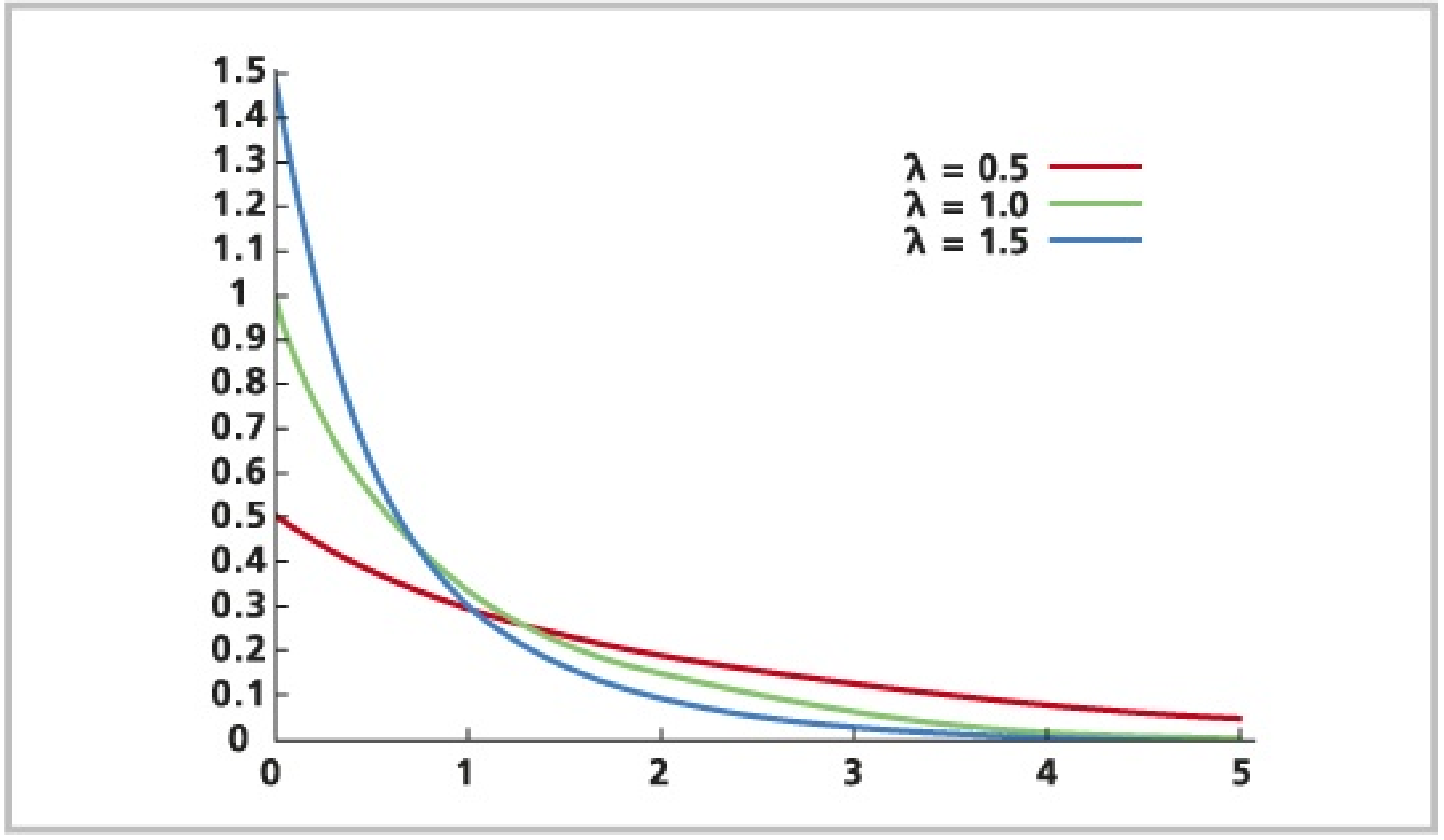
$$F(t) = P(X \leq t) = \int_0^t \lambda e^{-\lambda x} dx = -e^{-\lambda x}|_0^t = 1 - e^{-\lambda t}.$$

Typische stochasten met een (verdeling die benaderd wordt door een) exponentiële verdeling: de tijd die verstrijkt tot een zekere gebeurtenis plaatsvindt (een telefoontje, een emmissie, een bezoeker, een vallende ster).

Eigenschap van geheugenloosheid:

$$P(X \geq t + s \,|\, X \geq t) = \frac{\int_{t+s}^{\infty} \lambda e^{-\lambda x} dx}{\int_t^{\infty} \lambda e^{-\lambda x} dx} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X \geq s).$$

Voorbeeld



Definitie Een continue stochast X heeft een *normale verdeling* met *gemiddelde* μ en *standaardafwijking* σ als de kansdichtheid deze vorm heeft:

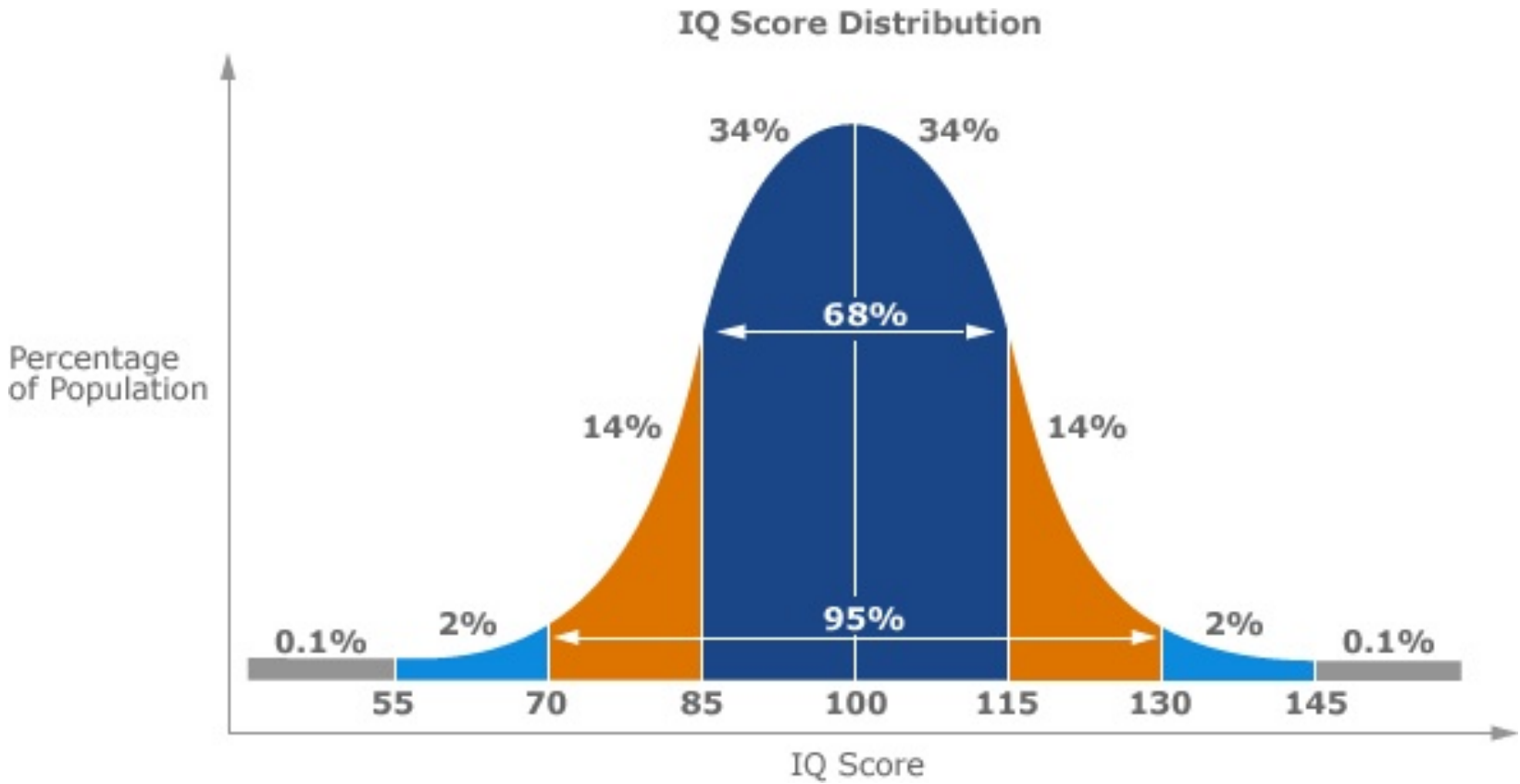
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

De *standaard normale verdeling* is de normale verdeling met gemiddelde $\mu = 0$ en standaardafwijking $\sigma = 1$. De kansdichtheid is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

De verdeling wordt vaak aangeduid met P_s.

Voorbeeld De verdeling van het IQ:



Voorbeeld

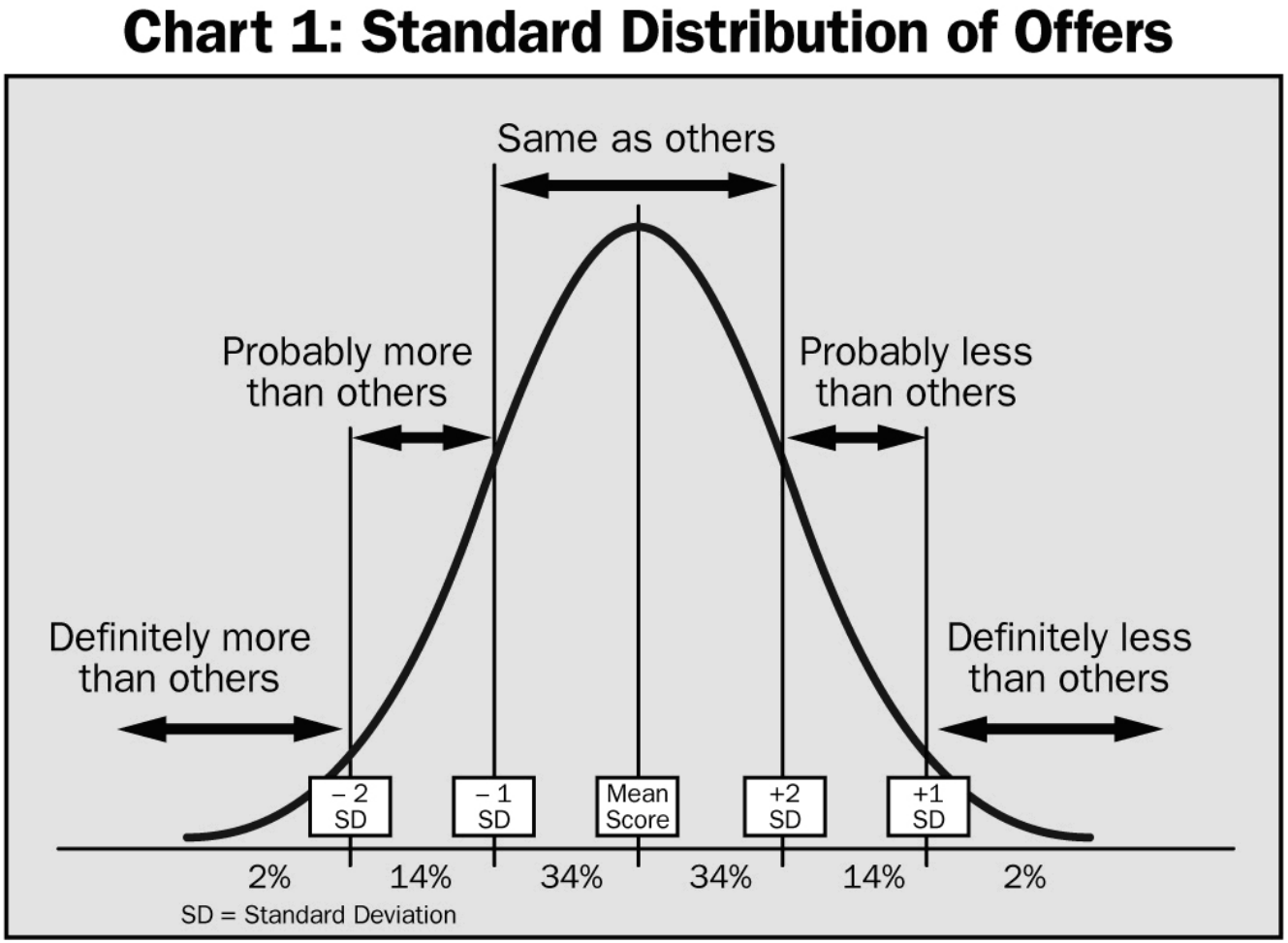


Chart 1: Standard Distribution of Offers
Same as others
Probably more than others
Probably less than others
Definitely more than others
Definitely less than others
– 2 SD
– 1 SD
Mean Score
+2 SD
+1 SD
2%
14%
34%
34%
14%
2%
SD = Standard Deviation

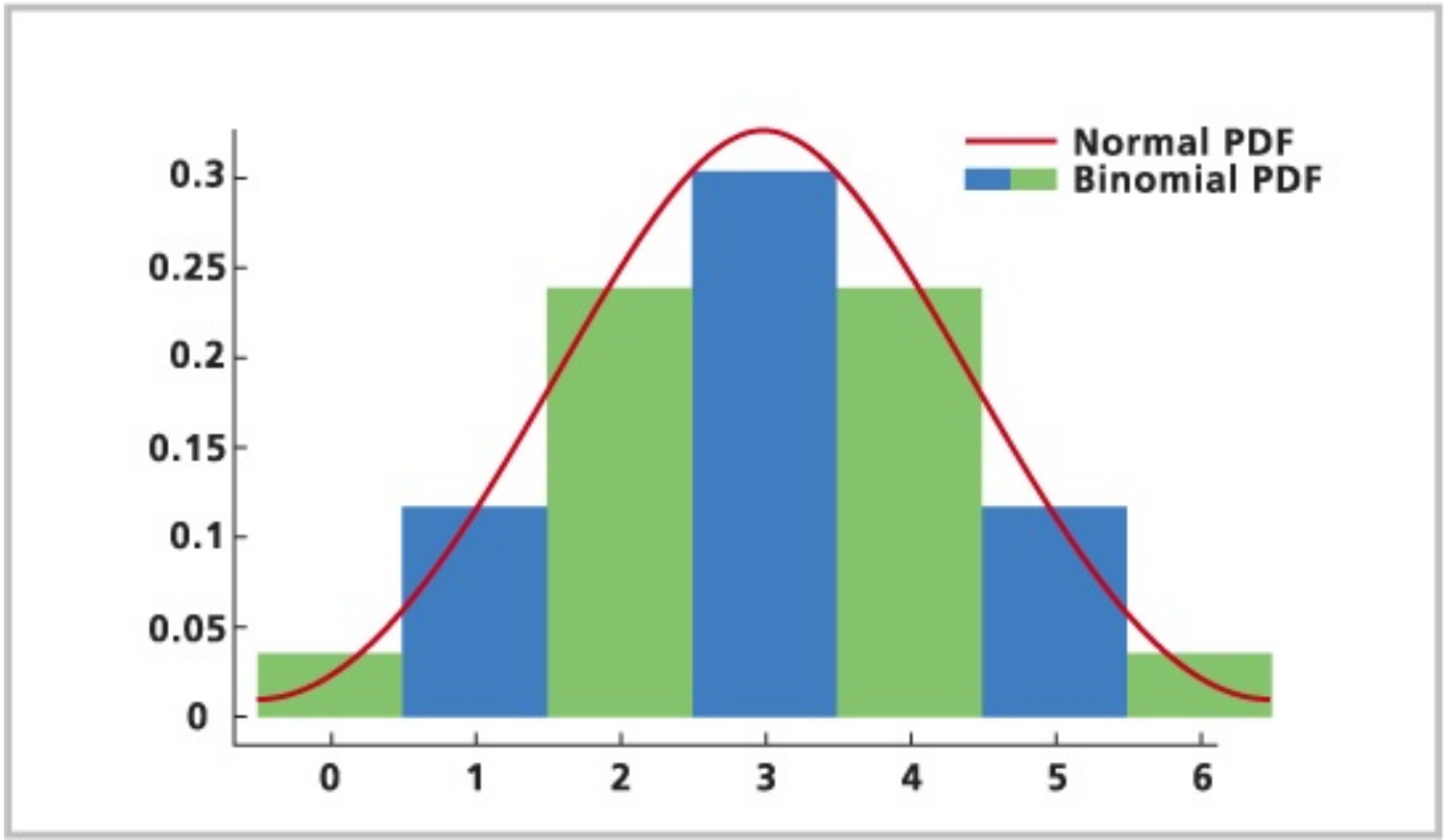
Stelling Voor elke een normaal verdeelde stochast X met *gemiddelde* μ en *standaardafwijking* σ geldt:

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68.$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95.$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.99.$$

Voorbeeld



Stelling Bij toenemende n gaat de binomiale verdeling

$$\binom{n}{k} p^k (1-p)^{n-k}$$

lijken op de normale verdeling met gemiddelde p en variantie $p(1-p)$.

Dat wil zeggen:

Zij X een binomiaal verdeelde stochast met verdeling $\binom{n}{k} p^k (1-p)^{n-k}$. Dan geldt voor grote n dat *de kans dat het aantal successen tussen k_1 en k_2 ligt*, benaderd kan worden door de normale verdeling:

$$P(k_1 \leq X \leq k_2) \approx \int_{k_1}^{k_2} \frac{1}{\sqrt{p(1-p)}\sqrt{2\pi}} e^{-\frac{(x-p)^2}{2p(1-p)}} dx.$$

Feit: Veel verdelingen lijken op de normale verdeling (Centrale Limietstelling).

Belangrijke verdelingen: normale verdeling

Voorbeeld

Normaliseren

Stelling Als X een normale verdeling heeft met gemiddelde μ en standaardafwijking σ, dan heeft de stochast

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

de standaard normale verdeling. Dat wil zeggen dat de standaardscores van X de standaard normale verdeling hebben.

Omdat $\frac{X-\mu}{\sigma}$ de standaardscore van X is, geldt dat

$$P(X \geq a) = P(\frac{X-\mu}{\sigma} \geq \frac{a-\mu}{\sigma}) = P_s(\frac{X-\mu}{\sigma} \geq \frac{a-\mu}{\sigma}).$$

Met tabel C.1 kan de X waarvoor $P_s(\frac{X-\mu}{\sigma} \geq \frac{a-\mu}{\sigma})$ geldt bepaald worden.

Voorbeeld Zij X een normaal verdeelde stochast met gemiddelde 7 en standaardafwijking 3.

De kans dat $X \geq 8$ is

$$P(X \geq 8) = P(\frac{X-7}{3} \geq \frac{8-7}{3}) = P_s(\frac{X-7}{3} \geq \frac{1}{3}).$$

Uit tabel C.1 blijkt dat ($\frac{1}{3}$ afgerond op 2 decimalen)

$$P_s(z \geq \frac{1}{3}) = 0.3707.$$

Dus $P(X \geq 8) = 0.3707$.

Voorbeeld Zij X een normaal verdeelde stochast met gemiddelde 7 en standaardafwijking 3.

De kans dat $X \geq 5$ is

$$P(X \geq 5) = P(\frac{X-7}{3} \geq \frac{5-7}{3}) = P_s(\frac{X-7}{3} \geq -\frac{2}{3}).$$

Uit tabel C.1 blijkt dat $P_s(z \geq \frac{2}{3}) = 0.2546$. Dus $P_s(z \leq -\frac{2}{3}) = 0.2546$, en daarmee

$$P_s(z \geq -\frac{2}{3}) = 1 - 0.2546 = 0.7454.$$

Dus $P(X \geq 5) = 0.7454$.

Voorbeeld Zij X een normaal verdeelde stochast met gemiddelde -3 en standaardafwijking 2.

Het 40^e percentiel is de waarde x waarvoor $P(X \leq x) = 0.4$. Dat wil zeggen, de x waarvoor $P_s(\frac{X+3}{2} \leq \frac{x+3}{2}) = 0.4$.

Uit tabel C.1 blijkt dat $P_s(z \geq 0.26) = 0.4$, dus $P_s(z \leq -0.26) = 0.4$.

Daarmee $P_s(\frac{X+3}{2} \leq -0.26) = 0.4$.

Het 40^e percentiel is dus $x = 2 \cdot (-0.26) - 3 = -3.52$.

Finis