

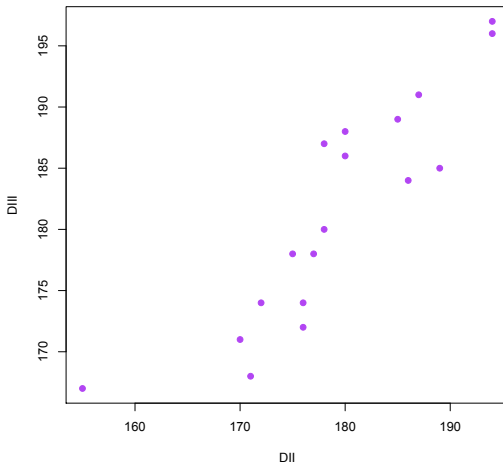
Statistiek voor A.I.

College 10

Dinsdag 16 Oktober

Horizontaal: lengte

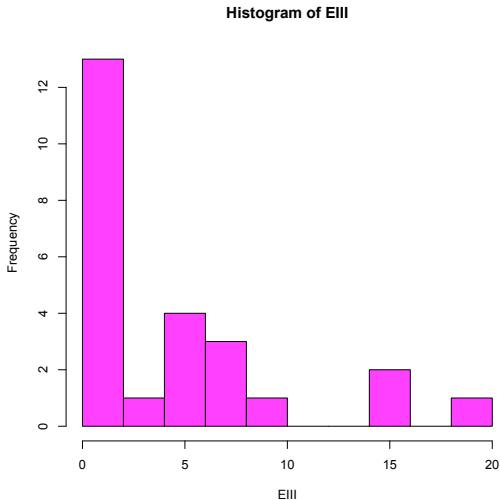
Verticaal: lengte tussen topjes middelvingers met gestrekte armen.



Correlatiecoëfficiënt: 0.89

Jullie - onderzoek Bjorn, Alex, Rianne

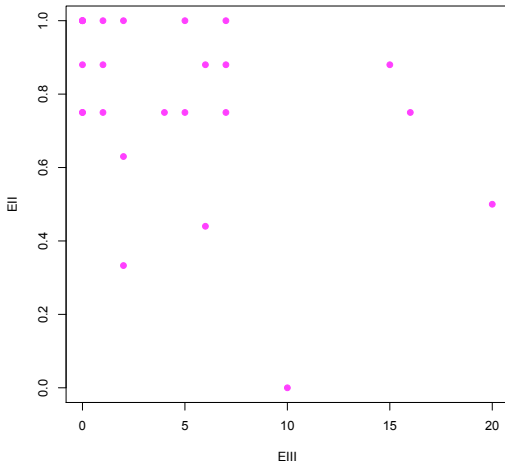
Gemiddeld aantal uren per week doorgebracht met gamen tijdens collegeweken vorig jaar.



Julie - onderzoek Bjorn, Alex, Rianne

Horizontaal: gemiddeld aantal uren per week doorgebracht met gamen

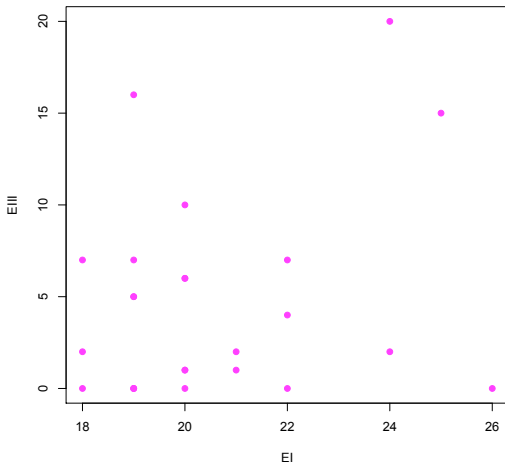
Verticaal: aantal vakken gehaald vorig jaar/aantal vakken gevolgd vorig jaar.



Correlatiecoëfficiënt: -0.36.

Horizontaal: leeftijd

Verticaal: gemiddeld aantal uren per week doorgebracht met gamen



Correlatiecoëfficiënt: 0.24.

Vandaag :

- Filosofie
- Experimenten herhalen
- Wet van de Grote Getallen

Filosofie van de kansrekening

Vraag:

Wat is een kans?

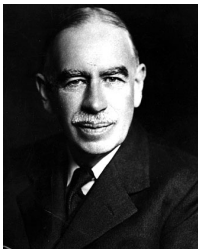
Wat betekent de uitspraak: de kans op A is p ?

Wat is de rechtvaardiging voor de axioma's van de kansrekening?

Filosofie: drie van de vele theorieën

- Logische theorie
- Subjectieve theorie
- Frequentietheorie

John Maynard Keynes (1883-1946)



Henry Moore: *Principia Ethica*. Regels voor ethisch handelen.

Bertrand Russell en Alfred Whitehead: *Principia Mathematica*. Deductie (logica).

Keynes: *Treatise on Probability*. Inductie (waarschijnlijkheid/kansrekening) .

Waarschijnlijkheid is de mate van *rationeel* geloof. Het is objectief (in de Platonistische zin van het woord).

Het is mogelijk dat er aan de waarschijnlijkheid van zekere gebeurtenissen geen numerieke waarde toe te kennen is.

Principe van Onvoldoende Reden (Principle of Indifference)

Als er geen reden bekend is om een mogelijkheid over een andere te verkiezen, dan is, gegeven deze kennis, de waarschijnlijkheid van elk van de mogelijkheden gelijk.

Filosofie: paradoxen door het Principe van Onvoldoende Reden

Hoewel het Principe van Onvoldoende Reden redelijk lijkt, leidt het in sommige gevallen tot “paradoxen”.

Betrands Paradox: (eerder college).

Water-Wijn Paradox:

Stel dat over een glas dat wijn en water bevat slechts bekend is dat het van het een ten hoogste 3 maal zoveel bevat als van het ander: $\frac{1}{3} \leq \frac{\text{wijn}}{\text{water}} \leq 3$.

Met het Principe van Onvoldoende Reden is de kansdichtheid van de verhouding wijn/water uniform op het interval $[\frac{1}{3}, 3]$. Daarom geldt

$$P\left(\frac{\text{wijn}}{\text{water}} \leq 2\right) = \frac{2 - \frac{1}{3}}{3 - \frac{1}{3}} = \frac{5}{8}.$$

Omdat tevens $\frac{1}{3} \leq \frac{\text{water}}{\text{wijn}} \leq 3$ is ook de kansdichtheid van de verhouding water/wijn uniform op het interval $[\frac{1}{3}, 3]$. Daarom geldt

$$P\left(\frac{\text{water}}{\text{wijn}} \geq \frac{1}{2}\right) = \frac{3 - \frac{1}{2}}{3 - \frac{1}{3}} = \frac{15}{16}.$$

Maar

$$\frac{\text{wijn}}{\text{water}} \leq 2 = \frac{\text{water}}{\text{wijn}} \geq \frac{1}{2}.$$

Als daarmee ook de kansen gelijk moeten zijn geeft dat een tegenspraak.

Filosofie: paradoxen door het Principe van Onvoldoende Reden

Water-Wijn Paradadox

In plaats van de verhoudingen $\frac{\text{wijn}}{\text{water}}$ en $\frac{\text{water}}{\text{wijn}}$ kan ook het percentage water of wijn in het glas als uniform verdeeld beschouwd worden. Dat geeft:

Het percentage wijn in het glas ligt tussen 25% ($\frac{\text{wijn}}{\text{water}} = \frac{1}{3}$) en 75% ($\frac{\text{wijn}}{\text{water}} = 3$), en is uniform verdeeld over het interval $[25, 75]$. Idem voor water.

Er geldt

$$P\left(\frac{\text{wijn}}{\text{water}} \leq 2\right) = P(\text{percentage wijn} \leq \frac{2}{3} \cdot 100\%)$$

$$P\left(\frac{\text{water}}{\text{wijn}} \geq \frac{1}{2}\right) = P(\text{percentage water} \geq \frac{1}{3} \cdot 100\%)$$

$$P\left(\frac{\text{wijn}}{\text{water}} \leq 2\right) = \frac{\frac{2}{3} \cdot 100 - 25}{75 - 25} = \frac{5}{6}$$

$$P\left(\frac{\text{water}}{\text{wijn}} \geq \frac{1}{2}\right) = \frac{75 - \frac{1}{3} \cdot 100}{75 - 25} = \frac{5}{6}.$$

Dus ontstaat onder deze opvatting van de situatie geen tegenspraak:

$$P\left(\frac{\text{wijn}}{\text{water}} \leq 2\right) = P\left(\frac{\text{water}}{\text{wijn}} \geq \frac{1}{2}\right).$$

Frank Plumpton Ramsey (1903-1930)



Bruno de Finetti (1906-1985)

Waarschijnlijkheid is de mate van *persoonlijk* geloof. Het is subjectief.

De kans die een persoon aan een gebeurtenis toekent wordt vastgesteld door middel van een denkbeeldige weddenschap (volgende slide).

Definitie De kans die een persoon aan een gebeurtenis A toekent is dat wat de persoon als *weddenschapsquotiënt* kiest in de volgende zetting:

Persoon I kiest een getal q , de weddenschaapsquotiënt, en een andere persoon, II , kiest vervolgens de inzet z . I betaalt qz aan II , en als A plaatsvindt betaalt II z aan I .

De weddenschaapsquotiënt q wordt beschouwd als de kans die persoon I aan gebeurtenis A toekent:

$$P_I(A) \equiv_{\text{def}} q.$$

Definitie De weddenschaapsquotiënten van I voor de gebeurtenissen $A_1, \dots, A_m$ zijn *coherent* als persoon II geen inzetten kan kiezen die hem altijd winst opleveren.

Probleem: Hoe komt men tot overeenstemming aangaande de “objectieve” kansen? Bijvoorbeeld de kans op K bij een (on)zuivere munt?

Stelling Onder aanname van coherentie zijn in de subjectieve interpretatie de axioma's van de kansrekening waar.

Bewijs Alleen het bewijs van het axioma $0 \leq P(A) \leq 1$ wordt hier gegeven.

Als de weddenschapsquotiënt $P_I(A) > 1$, dan zal // door $z > 0$ te kiezen altijd winst maken.

En als $P_I(A) < 0$, dan zal // door $z < 0$ te kiezen altijd winst maken.

Omdat aangenomen is dat de weddenschapsquotiënten coherent zijn volgt hieruit dat noch $P_I(A) < 0$ noch $P_I(A) > 1$. Dus geldt $0 \leq P(A) \leq 1$.

Richard Edler von Mises (1883-1953)



Waarschijnlijkheidsrekening is een empirische wetenschap, zoals de natuurkunde, en houdt zich bezig met *problemen waarbij ofwel dezelfde gebeurtenis zich herhaalt en herhaalt, ofwel een groot aantal uniforme elementen betrokken zijn op hetzelfde moment.*

Axioma van Convergentie:

Voor elk willekeurig attribuut A van een collectief C bestaat $\lim_{n \rightarrow \infty} \frac{\#A}{n}$, en dat getal is $P(A | C)$.

Merk op: Een kans is altijd relatief een collectief.

Probleem: Er kunnen alleen kansen toegekend worden aan fenomenen die herhaald kunnen worden. Hoe moet het collectief gekozen worden?

Filosofie: frequentietheorie versus subjectieve theorie

Subjectieve theorie



Bayesiaans leren, maximum likelihood schatters.

Frequentietheorie



Hypothese toetsen, zuivere schatters, betrouwbaarheidsintervallen.

Experimenten herhalen

Definitie Stel dat X een stochast is bij een experiment dat als uitkomsten getallen heeft. Als dat experiment n maal wordt herhaald, dan staat X_i voor de uitkomst van de i -de keer. $\bar{X}_n$ is de gemiddelde waarde van de n uitkomsten:

$$\frac{\sum_{i=1}^n X_i}{n}.$$

$\bar{X}_n$ is ook een stochast.

Wanneer de populatie groot is t.o.v. de steekproef kan een steekproef ter grootte n gezien worden als het n maal herhalen van een experiment.

Daarom zijn beide notaties $\bar{X}$ (gemiddelde steekproef) en $\bar{X}_n$ (gemiddelde waarde bij het n maal herhalen van het experiment behorende bij de stochast X) met elkaar in overeenstemming.

Experimenten herhalen: binomiale verdelingen

Stelling Zij X een stochast met een Bernoulli verdeling: $X = 1$ met kans p en $X = 0$ met kans $(1 - p)$. Laat bij het n maal herhalen van het experiment X_i de uitkomst van het i^e experiment zijn. Dan geldt

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

De mogelijke waarden van $\bar{X}_n$ zijn: $0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1$.

$\bar{X}_n$ heeft een binomiale verdeling:

$$P(\bar{X}_n = \frac{k}{n}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

De verwachtingswaarden van X en $\bar{X}_n$ zijn gelijk, de variantie neemt af:

$$E(X) = p \quad \text{Var}(X) = p(1-p).$$

$$E(\bar{X}_n) = p \quad \text{Var}(\bar{X}_n) = \frac{p(1-p)}{n}.$$

Experimenten herhalen

Voorbeeld Stel de populatie bestaat uit vier mensen met respectievelijke lengtes (in m.) $m_1 = 1.73$, $m_2 = 1.60$, $m_3 = 1.81$ en $m_4 = 1.74$. $\mu = 1.72$.

Stel dat er telkens een steekproef ter grootte 3 uit de populatie genomen wordt, wat is de kansverdeling van $\bar{X}_3$?

Er zijn vier mogelijke steekproeven en dus vier mogelijke waarden voor $\bar{X}_3$:

steekproef	steekproefgemiddelde $\bar{X}_3$
$\{m_1, m_2, m_3\}$	1.71
$\{m_1, m_2, m_4\}$	1.69
$\{m_1, m_3, m_4\}$	1.76
$\{m_2, m_3, m_4\}$	1.72

Merk op: Het gemiddelde van de steekproefgemiddelden is $1.72 = \bar{X}$.

De variantie van de populatie is:

$$\frac{(1.73 - 1.72)^2 + (1.60 - 1.72)^2 + (1.81 - 1.72)^2 + (1.74 - 1.72)^2}{4} = 0.00575.$$

De variantie van het steekproefgemiddelde is

$$\frac{(1.71 - 1.72)^2 + (1.69 - 1.72)^2 + (1.76 - 1.72)^2 + (1.72 - 1.72)^2}{4} = 0.00065.$$

Stelling Als X een stochast is met $E(X) = \mu$ en $\text{Var}(X) = \sigma^2$ en het bijbehorende experiment wordt n maal herhaald, dan geldt

$$E(\bar{X}_n) = \mu \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Stelling Wanneer uit een populatie met gemiddelde μ en variantie σ^2 steekproeven ter grootte n worden genomen, dan is het steekproefgemiddelde $\bar{X}_n$ een stochast, met als gemiddelde en variantie respectievelijk

$$E(\bar{X}_n) = \mu \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Wet van de Grote Getallen

Stelling (Markov Ongelijkheid)

Zij X een stochast waarvan alle waardes positief zijn. Dan geldt voor $t \geq 0$:

$$P(X \geq t) \leq \frac{E(X)}{t}.$$

Bewijs Alleen het geval dat X een discrete stochast met waarden $X_1, \dots, X_n$ is wordt behandeld. Er geldt

$$E(X) = \sum_{i=1}^n X_i P(X = X_i) = \sum_{X_i < t} X_i P(X = X_i) + \sum_{X_i \geq t} X_i P(X = X_i).$$

Omdat alle waarden van X positief zijn volgt hieruit

$$E(X) \geq \sum_{X_i \geq t} X_i P(X = X_i) \geq \sum_{X_i \geq t} t P(X = X_i) = t P(X \geq t).$$

Stelling (Chebyshev Ongelijkheid)

Voor elk getal $t > 0$ geldt: $P(|X - E(X)| \geq t) \leq \frac{\text{Var}(X)}{t^2}$.

Bewijs Laat Y de stochast $(X - E(X))^2$ zijn. Dan zijn alle waarden van Y positief en $E(Y) = \text{Var}(X)$. Uit de Markov ongelijkheid volgt:

$$P(|X - E(X)| \geq t) = P(Y \geq t^2) \leq \frac{E(Y)}{t^2} = \frac{\text{Var}(X)}{t^2}.$$

Stelling Voor elk getal $t > 0$ geldt: $P(|X - E(X)| < t) \geq 1 - \frac{\text{Var}(X)}{t^2}$.

Finis