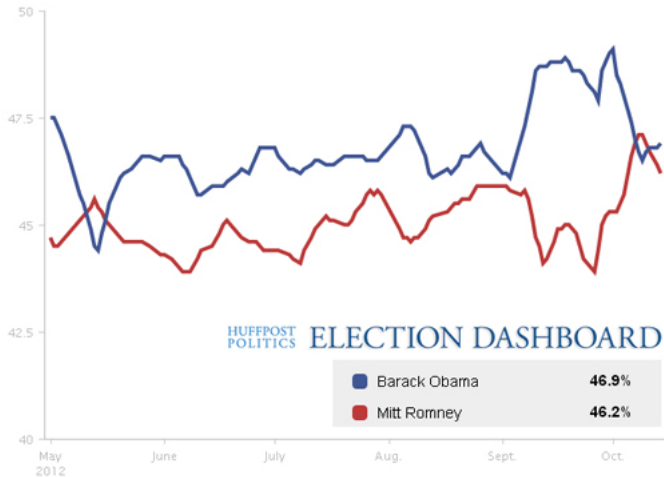


# Statistiek voor A.I.

College 10

Donderdag 18 Oktober

## Huffington Post poll verkiezingen VS - 12 Oktober 2012



### *Likely Voter Preferences in the 2012 Presidential Election*

Suppose the presidential election were held today, and it included Barack Obama and Joe Biden as the Democratic Party's candidates and Mitt Romney and Paul Ryan as the Republican Party's candidates. Who would you vote for [Barack Obama and Joe Biden, the Democrats (or) Mitt Romney and Paul Ryan, the Republicans]? If unsure: As of today do you lean more toward Obama and Biden, the Democrats (or) Romney and Ryan, the Republicans]?



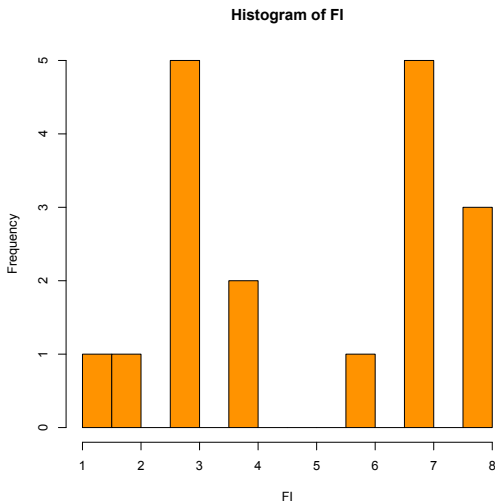
Based on continuous seven-day rolling averages from Oct. 1-7, 2012, through Oct. 9-15, 2012.

GALLUP®

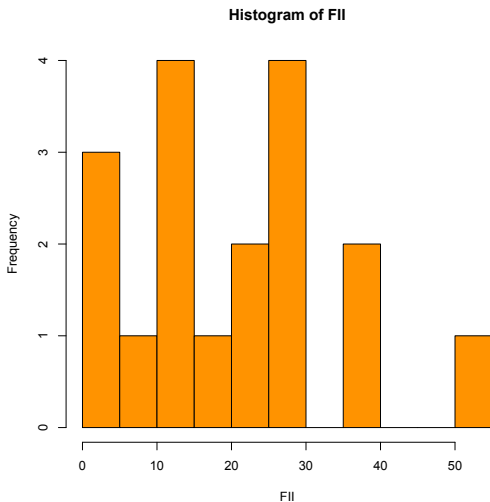
## Jullie - onderzoek Kimberly, Josca en Niels

Hoe mannelijk/vrouwelijk schat je jezelf? (1 = zeer mannelijk, 10 = zeer vrouwelijk)

Steekproef bevat 9 vrouwen en 9 mannen.

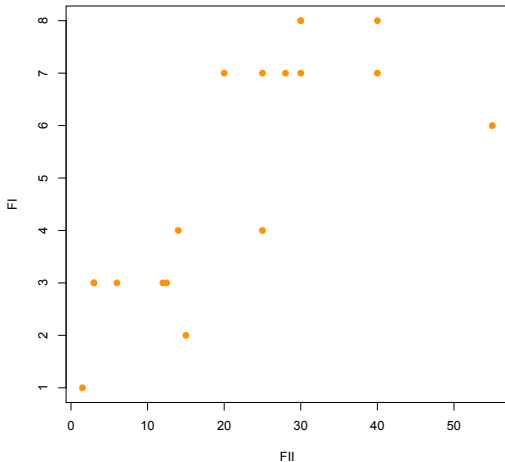


Hoe lang is je haar (in cm)?



Horizontaal: Hoe lang is je haar (in cm)?

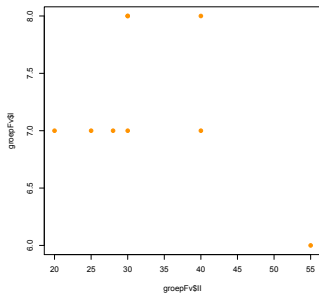
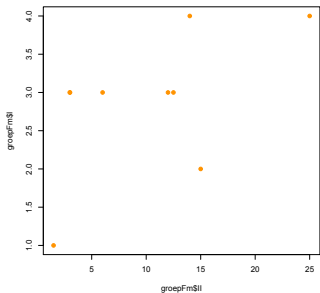
Verticaal: Hoe mannelijk/vrouwelijk schat je jezelf?



Correlatiecoëfficiënt: 0.77

Horizontaal: Hoe lang is je haar (in cm)?

Verticaal: Hoe mannelijk/vrouwelijk schat je jezelf?



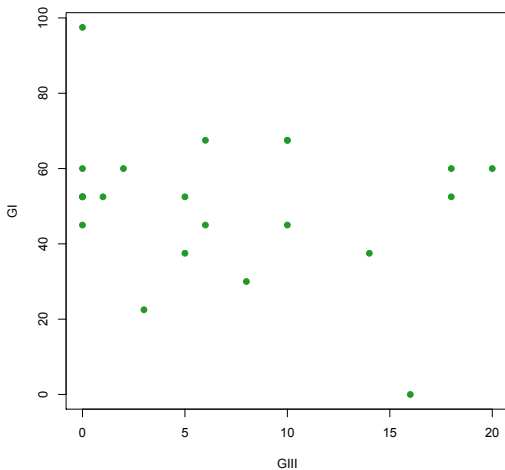
Links: mannen, correlatiecoëfficiënt: 0.55.

Rechts: vrouwen, correlatiecoëfficiënt: -0.38.

## Jullie - onderzoek Peter, Huub en Rens

Horizontaal: Gemiddeld aantal glazen bier per week.

Verticaal: Aantal ECTS behaald vorig jaar.



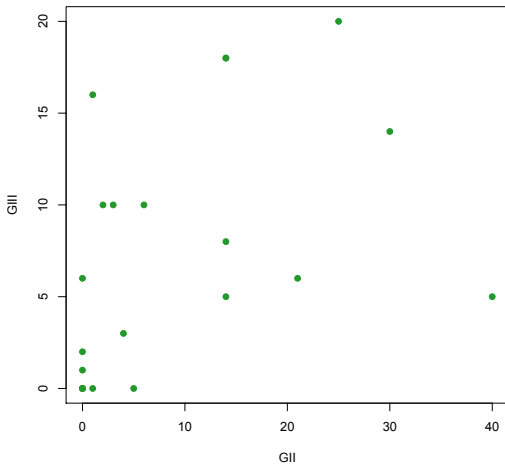
Correlatiecoëfficiënt: -0.21.



## Julie - onderzoek Peter, Huub en Rens

Horizontaal: Gemiddeld aantal koppen koffie per week.

Verticaal: Gemiddeld aantal glazen bier per week.



Correlatiecoëfficiënt: 0.45.

Vandaag :

- Wet van de Grote Getallen
- Centrale Limietstelling
- Hypothese toetsen

### 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:  $P(|X - E(X)| < t) \geq 1 - \frac{\text{Var}(X)}{t^2}$ .

## Wet van de Grote Getallen: gevolg van Chebyshev ongelijkheid

**Stelling** Zij  $X$  een stochast met waarden in  $\mathbb{R}$  en gemiddelde  $\mu$  en variantie  $\sigma^2$ . Dan geldt voor elke  $\epsilon > 0$ :

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \quad P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

**Bewijs** Eerder werd aangetoond:

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

Dus volgt uit Chebyshev:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Omdat

$$P(|\bar{X}_n - \mu| < \epsilon) = 1 - P(|\bar{X}_n - \mu| \geq \epsilon),$$

volgt ook

$$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$



**Merk op:** in de gevallen  $n\epsilon^2 \leq \sigma^2$  geeft de stelling geen informatie.

## Wet van de Grote Getallen

**Stelling** (Wet van de Grote Getallen) Voor een stochast  $X$  met verwachtingswaarde  $\mu$  geldt voor elke  $\epsilon > 0$ :

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1.$$

Bij toenemende  $n$  neemt de waarschijnlijkheid dat  $\bar{X}_n$  dicht bij  $\mu$  ligt toe.

**Bewijs** Dit volgt uit de laatste stelling op de vorige stelling en het feit dat  $\lim_{n \rightarrow \infty} \frac{\sigma^2}{n\epsilon^2} = 0$ .



**Voorbeeld** Kan door vaak een munt te gooien de waarschijnlijkheid dat het gemiddelde aantal keren dat kop gegooit wordt meer dan 0.3 afwijkt van 0.5, kleiner dan 0.01 worden?

$\bar{X}_n$  is het steekproefgemiddelde bij het  $n$  maal gooien van de munt. Gevraagd wordt naar  $n$  waarvoor  $P(|\bar{X}_n - 0.5| \geq 0.3) \leq 0.01$ .

Met Chebyshev:

$$P(|\bar{X}_n - 0.5| \geq 0.3) \leq \frac{0.25}{n(0.3)^2}.$$

Dus  $P(|\bar{X}_n - 0.5| \geq 0.3) \leq 0.01$  voor  $n \geq \frac{0.25}{0.01(0.3)^2} = \frac{2500}{9}$ .

Voor een duizendmaal zo grote nauwkeurigheid moet  $n$  duizendmaal zo groot zijn:  $P(|\bar{X}_n - 0.5| \geq 0.3) \leq 0.00001$  voor  $n \geq \frac{0.25}{0.00001(0.3)^2} = \frac{2500000}{9}$ .

**Voorbeeld** Een wijnboer verkoopt wijn per doos en wil dat de kans dat het percentage bedorven flessen in een doos meer dan 4% afwijkt van gemiddelde  $\mu$  hoogstens 0.2 is. Hij weet  $\mu$  niet en neemt voor de variantie 0.005.

Hoeveel flessen wijn moeten de dozen minimaal bevatten?

Met Chebyshev:

$$P(|\bar{X}_n - \mu| \geq 0.04) \leq \frac{0.005}{n(0.04)^2}.$$

Dus  $P(|\bar{X}_n - \mu| \geq 0.04) \leq 0.2$  voor  $n \geq \frac{0.005}{0.2(0.04)^2} = 15.625$ .

De dozen moeten minimaal 16 flessen bevatten.

**Stelling** Voor elke  $\epsilon > 0$  en  $1 > \delta > 0$  kan door het kiezen van de juiste steekproefgrootte (door het experiment een voldoende aantal maal te herhalen) de waarschijnlijkheid dat het steekproefgemiddelde  $\bar{X}_n$  minder dan  $\epsilon$  afwijkt van het ware gemiddelde  $\mu$ , groter dan  $\delta$  worden.

**Bewijs** Gezocht wordt een  $n$  waarvoor  $P(|\bar{X}_n - \mu| < \epsilon) \geq \delta$ .

Met Chebyshev:

$$P(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

Daarmee  $P(|\bar{X}_n - \mu| < \epsilon) \geq \delta$  voor  $n$  waarvoor  $1 - \frac{\sigma^2}{n\epsilon^2} \geq \delta$ .

Dat wil zeggen, voor

$$n \geq \frac{\sigma^2}{\epsilon^2(1 - \delta)}.$$



**Merk op:** voor een  $x$  maal zo grote nauwkeurigheid ( $\epsilon$   $x$  maal zo klein), moet  $n$   $x^2$  maal zo groot worden.



## Wet van de Grote Getallen: Obama versus Romney

**Voorbeeld** Gallup poll 15 Oktober 2012: Obama 46%. Hoe groot moet de steekproef minimaal zijn om met tenminste 0.9 waarschijnlijkheid tot op 1% nauwkeurigheid vast te stellen welk percentage van de bevolking op Obama stemt?  $\mu$  is het werkelijke percentage Obama-stemmers.

Zij  $X$  de stochast bij het wel of niet Obama stemmen van een persoon.  $X$  heeft een Bernoulli verdeling met variantie  $\sigma^2 = \mu(1 - \mu)$ , dus  $0 \leq \sigma^2 \leq 0.25$ .

Gevraagd wordt naar de minimale  $n$  waarvoor

$$P(|\bar{X}_n - \mu| < 0.01) \geq 0.9.$$

De stelling op de vorige slide geeft

$$n \geq \frac{\sigma^2}{\epsilon^2(1 - \delta)} = \frac{\sigma^2}{(0.01)^2(0.1)} \geq \frac{(0.25)^2}{(0.01)^2(0.1)} = 6250.$$

Bij het werkelijke aantal  $n = 2700$ : minstens 0.9 waarschijnlijkheid is slechts te bereiken bij lagere nauwkeurigheid:

$$P(|\bar{X}_{2700} - \mu| < \epsilon) \geq 0.9 \Rightarrow \epsilon^2 \geq \frac{\sigma^2}{2700(1 - 0.9)} \Rightarrow \epsilon \geq 0.015,$$

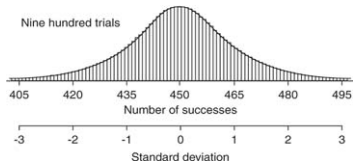
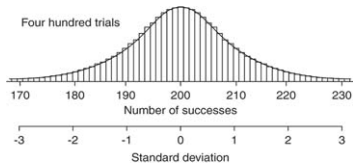
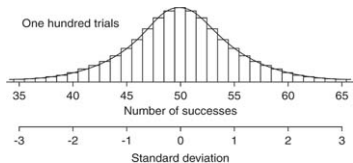
en 1%-nauwkeurigheid is slechts te bereiken bij lagere waarschijnlijkheid:

$$P(|\bar{X}_{2700} - \mu| < 0.01) \geq \delta \Rightarrow \delta \leq 1 - \frac{\sigma^2}{2700(0.01)^2} \leq 0.769.$$

## Centrale Limietstelling

## Centrale Limietstelling

Voorbeeld Benadering van een binomiale verdeling door de normale verdeling.



**Stelling** (Centrale Limietstelling) Voor een stochast  $X$  met verwachtingswaarde  $\mu$  en standaardafwijking  $\sigma$  geldt voor elke  $a$ :

$$\lim_{n \rightarrow \infty} P\left(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq a\right) = P_s(z \leq a).$$

“Bij toenemende  $n$  benadert  $\bar{X}_n$  de normale verdeling met gemiddelde  $\mu$  en standaardafwijking  $\frac{\sigma}{\sqrt{n}}$ .”

De Centrale Limietstelling is een versterking en precisering van de Wet van de Grote Getallen.

**Definitie**  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$  is de *standaardfout van het gemiddelde*.

## Centrale Limietstelling

**Voorbeeld** Zij  $X$  een stochast met  $E(X) = 3$  en  $\text{Var}(X) = 4$  waarvoor de verdeling onbekend is of moeilijk te berekenen. Er wordt gevraagd naar de kans dat het steekproef gemiddelde van een willekeurige steekproef ter grootte  $n = 10.000$  kleiner is dan 2.95.

Aangenomen wordt dat  $n$  voldoende groot is om de verdeling van  $\bar{X} = \bar{X}_{10.000}$  als een normale verdeling met verwachtingswaarde  $\mu = 3$  en standaardafwijking  $\frac{\sqrt{4}}{\sqrt{n}} = \frac{2}{100} = 0.02$  te beschouwen.

Onder die aanname geldt:

$$P(\bar{X} \leq 2.95) = P_s\left(\frac{\bar{X} - 3}{0.02} \leq \frac{2.95 - 3}{0.02}\right) = P_s\left(\frac{\bar{X} - 3}{0.02} \leq \frac{-0.05}{0.02}\right) = P_s\left(\frac{\bar{X} - 3}{0.02} \leq -2.25\right).$$

Wegens de symmetrie van de standaard normale verdeling rond 0 geldt

$$P_s\left(\frac{\bar{X} - 3}{0.02} \leq -2.25\right) = P_s\left(\frac{\bar{X} - 3}{0.02} \geq 2.25\right).$$

Uit tabel C.1 blijkt dat  $P_s(z \geq 2.25) = 0.0122$ . Dus  $P(\bar{X} \leq 2.95) = 0.0122$ .

## Hypothese toetsen

I CAN'T BELIEVE SCHOOLS  
ARE STILL TEACHING KIDS  
ABOUT THE NULL HYPOTHESIS.

I  
I REMEMBER READING A BIG  
STUDY THAT CONCLUSIVELY  
DISPROVED IT YEARS AGO.



### Voorbeeld

- Een medicijn wordt aan een groep patiënten toegediend en aan een controlegroep niet. Er moet besloten worden of het medicijn werkt of niet.
- Een fabriek laat een gedeelte van een lading geproduceerde auto's testen op een defect. Er moet besloten worden of de lading goed genoeg is om bij de leverancier af te leveren.
- De schedelgrootte van enkele Egyptenaren uit de oudheid is bekend. Op grond van deze data wil men vaststellen of de gemiddelde schedelgrootte van mensen in het oude Egypte gelijk is aan die van de huidige mens.



## Hypothese toetsen: de hypothesen

**Definitie** Bij *hypothese toetsen* wordt een hypothese  $H_0$  getest tegen een hypothese  $H_a$ .

$H_0$  is de *nulhypothese* en  $H_a$  is de *alternatieve hypothese*.  $H_a$  wordt ook wel  $H_1$  genoemd.

De alternatieve hypothese is meestal de *onderzoekshypothese*, die alleen aangenomen (ondersteund) wordt door de nulhypothese te verwerpen.

Vaak is de nulhypothese de hypothese van *geen verschil* of *geen onderscheid* of *geen relatie*, en de test dient om aan te tonen dat er wel een verschil, onderscheid of relatie is.

## Hypothese toetsen: de zijde

**Definitie** De hypotheses die we gaan beschouwen zijn meestal van de vorm:

$$\begin{array}{ll} \text{tweezijdig} & H_0 : \mu = a \\ & H_a : \mu \neq a \end{array}$$

$$\begin{array}{ll} \text{linkszijdig} & H_0 : \mu \geq a \\ & H_a : \mu < a \end{array}$$

$$\begin{array}{ll} \text{rechtszijdig} & H_0 : \mu \leq a \\ & H_a : \mu > a \end{array}$$

## Hypothese toetsen: criterium voor verwerpen $H_0$

**Definitie**  $H_0$  en  $H_a$  worden *verworpen* of *aangenomen* (niet verworpen).

Als  $H_0$  aangenomen wordt, wordt  $H_a$  verworpen.

Als  $H_0$  verworpen wordt, wordt  $H_a$  aangenomen.

### Definitie

- $H_0$  is waar en wordt aangenomen: correcte beslissing.
- $H_0$  is waar en wordt verworpen: fout van de 1<sup>e</sup> soort.
- $H_a$  is waar en wordt aangenomen: correcte beslissing.
- $H_a$  is waar en wordt verworpen: fout van de 2<sup>e</sup> soort.

Criterium voor het verwerpen van de nulhypothese:

Voor een zekere constante  $x_k$  (of  $x_k^r$  en  $x_k^l$ ) wordt op grond van een steekproef met gemiddelde  $\bar{X}$

$$\begin{array}{ll} H_0 \text{ verworpen als } \bar{X} \geq x_k^r \text{ of } \bar{X} \leq x_k^l & \text{(tweezijdige toets)} \\ H_0 \text{ verworpen als } \bar{X} \geq x_k & \text{(rechtszijdige toets)} \\ H_0 \text{ verworpen als } \bar{X} \leq x_k & \text{(linkszijdige toets).} \end{array}$$

Finis