Onderzoeksmethoden: Statistiek 4

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

Departement Informatica
Beta-faculteit, Universiteit Utrecht

Vandaag

- Recap:
  - Hypothese toetsen
  - t-toets met één steekproef
  - Betrouwbaarheidsinterval
- Meer t-toetsen:
  - t-toets met gepaarde metingen
  - t-toets met onafhankelijke metingen
- Chi-kwadraat toetsen:
  - Goodness-of-fit toets
  - Toets van onafhankelijkheid

Recap: Hypothese toetsen

Procedure:

- Formuleer hypotheses
- Kies teststatistiek en leg criterium vast
- Bereken teststatistiek uit steekproef
- Neem beslissing

Recap: One-sample t-toets

- Vergelijking van één steekproefgemiddelde met een 'norm' (een van te voren bepaald gemiddelde, zeg μ_0).
- σ uit populatie is niet bekend en wordt geschat met behulp van s.
- Het steekproefaantal is klein ($n < 120$).
- Meetwaarden onafhankelijk en identiek normaal verdeeld (met zelfde gemiddelde en variantie).
- Oplossing: t-verdelingen

$$t_{obs} = \frac{\overline{X}(n) - \mu_0}{(s/\sqrt{n})}$$

- Onder H_0: t_{obs} heeft t-verdeling met $df = n - 1$ vrijheidsgraden
- t-table: zie boek, online, Excel, calculator...

Hypotheses toetsen en omgekeerd

Bij t-toets:

- Hypothese $H_0: \ \mu = \mu_0$.
- Geobserveerd steekproefgemiddelde $\overline{X}$.
- Q: Wanneer is $\overline{X}$ in "overeenstemming" met H_0?
- A: Als $\overline{X}$ niet meer afwijkt van μ_0 dan $\frac{s}{\sqrt{n}} \cdot t_{crit}$.

Andersom:

- Geobserveerd steekproefgemiddelde $\overline{X}$.
- Q: Met welke hypothesewaarden μ_0 is $\overline{X}$ in "overeenstemming"?
- A: Als μ_0 niet meer afwijkt van $\overline{X}$ dan $\frac{s}{\sqrt{n}} \cdot t_{crit}$.

Betrouwbaarheidsinterval

- Stel: ik meet steekproefgemiddelde $\overline{X}(n) = 23.4$
- Kan ik nu met 95% betrouwbaarheid zeggen in welk gebied het onbekende populatiegemiddelde μ ligt?
- 95% betrouwbaarheidsinterval:

$$\left[\overline{X}(n) - \frac{s}{\sqrt{n}} \times t_{crit} \quad ; \quad \overline{X}(n) + \frac{s}{\sqrt{n}} \times t_{crit}\right]$$

waarbij t_{crit} de kritieke waarde is voor $\alpha = 0.05$ bij $df = n - 1$.

t-toets: 3 soorten onderzoeksvraagstellingen

1. t-toets met één steekproef:
   (one-sample t-test)
2. t-toets met gepaarde metingen:
   (dependent-samples t-test)
   (matched-subjects t-test)
   (between-subjects t-test)
3. t-toets voor twee onafhankelijke steekproeven
   (independent-samples t-test)

t-toets met gepaarde metingen (Paired-samples t-test)

Voorbeeld 1:

- 30 zware rokers worden aan een trainingsprogramma onderworpen om van het roken af te komen,
- Vóór de training rookten zij gemiddeld 36 sigaretten per dag;
- Eén maand na de training rookten dezelfde rokers gemiddeld 28 sigaretten per dag.
- Is dit verschil groot genoeg om te mogen zeggen dat het trainingsprogramma effect heeft?

Voorbeeld 2:

- Convergeert het force-directed graph algoritme bij parameters (s^A, u^A, r^A) langzamer of sneller dan bij (s^B, u^B, r^B)?

t-toets met gepaarde metingen: voorbeeldduitwerking

- Paren van observaties:
  - B.v. 40 testgrafen met 100 punten
  - Run op elk van de testgrafen het force-directed graph algoritme bij parameterkeuze resp. u^A en u^B.
  - Voor elke graaf:
    X_A = het aantal iteraties bij keuze u^A en
    X_B = het aantal iteraties bij keuze u^B
  - Neem aan: X_A en X_B 'ongeveer' normaal verdeeld
  - Voor ieder paar waarnemingen bereken je de verschilscore $D = X_A - X_B$
- Verder werken met D als bij de one-sample t-toets met hypothese $H_0: \ \mu_D = 0$.

t-toets met gepaarde metingen: voorbeelduitw. (2)

- Toetsingsgrootheid: $t = \dfrac{\overline{D} - \mu_D}{(s_D/\sqrt{n})}$.
- Significantieniveau: $\alpha = .05$.
- We observeren: $\overline{D}_{obs} = -23$ en $s_D = 51$:

  (Parameterkeuze u^B leverde gemiddeld 23 méér iteraties op dan keuze u^A)
- $\dfrac{s_D}{\sqrt{n}} = \dfrac{51}{\sqrt{40}} = 8.064$
- $t = \dfrac{-23 - 0}{8.064} = -2.852$

t-toets met gepaarde metingen: voorbeelduitw. (3)

- Beslissingsregel:
  - Verwerp H_0 indien $t_{obs} \leq -t_{crit}$ of $t_{crit} \geq t_{obs}$.
  - Verwerp H_0 niet indien $-t_{crit} < t_{obs} < t_{crit}$.
- Criterium: Tweezijdige toetsing met $\alpha = .05$:
  $t_{crit}(df = 39) \approx 2.023$
- Toetsingsgrootheid: $t = -2.852$.
- Conclusie?
  Verwerp H_0: er is een significant verschil tussen de twee parameterinstellingen.

t-toets met gepaarde metingen: formules

- $D = X_A - X_B$,
- $t = \dfrac{\overline{D} - \mu_D}{s_{\overline{D}}} = \dfrac{\overline{D} - \mu_D}{(s_D/\sqrt{n})}, \quad (df = n - 1)$.

Onder aanname van (meest gebruikelijke) $H_0: \ \mu_D = 0$ is

- $t = \dfrac{\overline{D}}{s_{\overline{D}}} = \dfrac{\overline{D}}{(s_D/\sqrt{n})}, (df = n - 1)$.
- en het $(1 - \alpha)100\%$ betrouwbaarheidsinterval voor μ_D:

$$\left[\overline{D} - s_{\overline{D}} \cdot t_{crit}(df) \ ; \ \overline{D} + s_{\overline{D}} \cdot t_{crit}(df)\right]$$

Consistentie van het effect

- Verschil is dus significant:
- Maar nader onderzoek levert:
  - 11 grafen gaven bij parameterkeuze u^B meer iteraties dan bij u^A
  - 20 grafen gaven bij u^B gelijk # iteraties als bij u^A
  - 9 grafen gaven bij u^B minder iteraties dan u^A
- Bij significantie wil je eigenlijk meer:
  - Minstens de helft?
  - Of nog meer (toevalseffect)?
- Waargenomen verschillen geven consistent effect als:

  Minimale aantal $= \lceil 0.5(n+1) + 0.823\sqrt{n} \rceil$ bij $\alpha = 0.05$

  Minimale aantal $= \lceil 0.5(n+1) + 1.163\sqrt{n} \rceil$ bij $\alpha = 0.01$

Kwaliteit van het effect: voorbeelduitwerking

We zien

parameter u^B levert méér iteraties op dan bij u^A

$n = 40$

Minimale aantal $= \lceil 0.5 \cdot 41 + 1.163\sqrt{40} \rceil = \lceil 25.7 \rceil = 26$ bij $\alpha = 0.05$

Als tenminste 26 van de testgrafen meer iteraties nodig heeft bij u^B dan bij u^A dan mogen we concluderen dat in het algemeen dit bij tenminste de helft van de grafen het geval zal zijn.

t-toetsen

- *t*-toets met één steekproef
  (one-sample t-test)
- *t*-toets met gepaarde metingen
  (dependent-samples t-test)
  (matched-subjects t-test)
  (between-subjects t-test)
- *t*-toets voor twee onafhankelijke steekproeven
  (independent-samples t-test)

t-toets voor twee onafhankelijke metingen

Eng: Independent samples *t*-test.

- Bestaat er een (significant) verschil tussen bachelor studenten en masterstudenten wat betreft leeftijd?
- Is er een (significant) verschil tussen uitwonende en thuiswonende studenten Informatica wat betreft het geld dat zij te besteden hebben?

Rekenvoorbeeld:

- Is er een significant verschil in looptijd van het force-directed graph algoritme tussen bipartiete grafen en niet-bipartiete grafen (met gelijk aantal 100 knopen en 200 kanten)?

t-toets voor twee onafhankelijke metingen

- We meten één afhankelijke interval of ratio variabele.
- Twee onafhankelijke steekproeven:
  - Steekproef 1: Grootte n_1, met metingen $X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$
  - Steekproef 2: Grootte n_2, met metingen $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$
- Aannames:
  - Afhankelijke variabele (meetresultaten) zijn normaal verdeeld
  - Gelijke variantie (niet strict: klopt meestal wel met even grote steekproeven)
- $H_0: \ \mu_1 = \mu_2$ (ook wel $\mu_1 - \mu_2 = 0$)
- Vrijheidsgraden: $df = n_1 + n_2 - 2$.
- In Excel: t-Test: Two-Sample Assuming Equal Variances

t-toets voor twee onafhankelijke metingen

Onder aanname $H_o: \ \mu_1 = \mu_2$ heeft toetsingsgrootheid:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{\overline{X}_1 - \overline{X}_2}} = \frac{(\overline{X}_1 - \overline{X}_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

een t-verdeling met $df = n_1 + n_2 - 2$ vrijheidsgraden,

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s^2_X}{n_1} + \frac{s^2_X}{n_2}}$$

Met "*pooled variance*" $s^2_X = \frac{s^2_{X_1}(n_1 - 1) + s^2_{X_2}(n_2 - 1)}{n_1 + n_2 - 2}$ en $s^2_{X_1}, s^2_{X_2}$ steekproefvarianties van de twee respectievelijke steekproeven.

Let op: formule in boek veronderstelt $s_{X_1} = s_{X_2}$ en wijkt dus af.

t-toets voor twee onafhankelijke metingen: voorbeeld

1. Formuleer hypothese

- $H_0: \ \mu_1 = \mu_2$
- $H_1: \ \mu_1 \neq \mu_2$.

2. Kies test-statistiek en leg criterium vast:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{s_{\overline{X}_1 - \overline{X}_2}}$$

- Significantieniveau: $\alpha = .05$
- $n_1 = 46$ (bipartiet) en $n_2 = 56$ (niet bipartiet)
- Onder H_0 heeft t een t-verdeling met $df = 100$
- Kritieke waarde $t_{crit} = 1.99$ (tweezijdig)

t-toets voor twee onafhankelijke metingen: voorbeeld

3. Bereken teststatistiek uit steekproef:

$$\overline{X}_1 = 3.8698, \ S_{X_1} = 1.6714, \ n_1 = 46,$$

$$\overline{X}_2 = 4.5819, \ S_{X_2} = 1.5216, \ n_2 = 56,$$

$$s^2_X = \frac{(1.6714)^2 \cdot 45 + (1.5216)^2 \cdot 55}{100} = 2.531$$

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{2.531}{46} + \frac{2.531}{56}} = 0.3165$$

$$t = \frac{3.8698 - 4.5819}{0.3165} = -2.25$$

4. Neem beslissing: ??$t < -t_{crit}$, dus H_0 verwerpen.

t-toets voor twee onafhankelijke metingen

$(100 - \alpha)\%$ **Betrouwbaarheidsinterval** voor verschil:

$$\left[(\overline{X}_1 - \overline{X}_2) - s_{\overline{X}_1 - \overline{X}_2} \cdot t(df)_{\alpha/2} \;\;,\;\; (\overline{X}_1 - \overline{X}_2) + s_{\overline{X}_1 - \overline{X}_2} \cdot t(df)_{\alpha/2}\right]$$

Effectgrootte

- Bij onafhankelijke metingen en significant verschil: percentage verklaarde variantie
- Hoeveel van de verschillen in de scores op de afhankelijke variabele wordt verklaard doordat ze uit een verschillende groep afkomstig zijn?
- Percentage verklaarde variantie is $\omega^2 \cdot 100\%$ met

  $$\omega^2 = \frac{t_{obs}^2}{t_{obs}^2 + df}$$

  waarbij (vuistregel):
  - 0 – 5% is een *zwak* effect
  - 5 – 20% is een *matig* effect
  - $>$ 20% is een *sterk* effect
- In dit geval: $\omega^2 = 0.0481$ dus percentage verklaarde variantie is 4.81% (zwak effect)

Wat als waarnemingen niet normaal verdeeld zijn?

- T-toets is een zgn. parametrische toets, i.e.
  - Hypothese gaat over een parameter van de verdeling.
  - Gebaseerd op aanname dat steekproefwaarnemingen bepaalde verdeling hebben (vaak Normale verdeling).

Q: Wat te doen als niet aan aannames is voldaan?

1. Neem het gemiddelde van een aantal waarnemingen. Dit is normaal verdeeld volgens Centrale Limietstelling.
   - Bijv. de gemiddelde looptijd van 20 restarts van Simulated Annealing
2. Wilcoxon signed-rank test (Ref. bijv. Wikipedia)
   Dit is een niet parametrische test (geen aanname op de verdeling van de steekproefwaarnemingen).

χ^2 (Chi-kwadraat) toets

- Met een Chi2 (χ^2) toets ga je na hoe waarschijnlijk het is dat verhoudingsmaten aan bepaalde verwachtingen of voorwaarden voldoen
- Twee soorten:
  - Goodness-of-fit
  - Onafhankelijkheid
- Geen aanname vooraf op bepaalde verdeling!

χ^2-toets type 1: goodness of fit

Verwachtingen over een variabele in een distributie.

Voorbeeld:

- Zijn verschillende typen internet-aansluitingen gelijk verdeeld?
- Zijn de verschillende soorten smartphones in deze klas gelijk verdeeld als in de rest van Nederland?

Smartphonegebruik in NL 2012 (Q4) onder mobile phone users:

Source: Telecompaper

Smartphone Android	Smartphone iOS	Smartphone Anders	Geen Smartphone
47%	13%	10%	30%

Voorbeeld

Steekproef van $n = 60$ proefpersonen met mobiel:

Hypothese H_0: resultaten komen overeen met de verwachte aantallen volgens de gegeven verdeling:

Goodness-of-fit	Android	iOS	Anders	geen
Verwacht %	47%	13%	10%	30%
Verwachte frequentie E_i	28.2	7.8	6.0	18
Geobserveerde freq. O_i	25	12	5	18

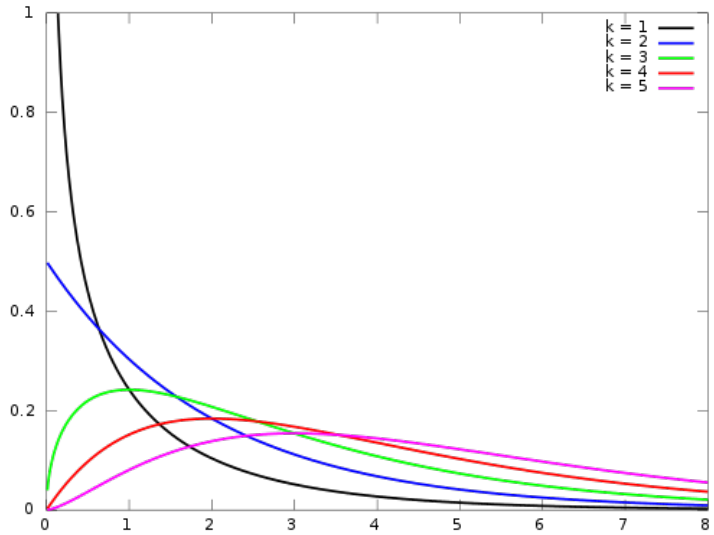
Toetsingsgrootheid: $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$

met K gelijk aan het aantal mogelijke uitkomstwaarden

(Dit geval $K = 4$)

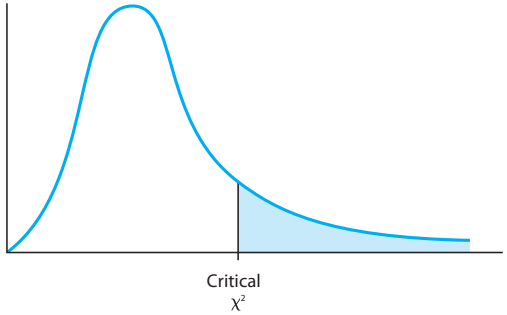
χ^2 volgt Chi-kwadraat (χ^2) verdeling

Familie van verdelingen met vrijheidsgraad:



χ^2 voor Goodness-of-Fit test

χ^2 volgt een chi-kwadraat verdeling met $df = K - 1$ vrijheidsgraden.



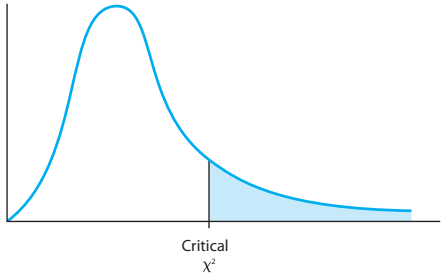
Voorbeeld:

- $\alpha = .05$
- $\chi^2 = 2.791$
- $\chi^2_\alpha(df) = \chi^2_\alpha(3) = 7.815$ (Tabel in boek of in Excel CHISQ.INV(0.95;3))
- Verwerp H_0 als $\chi^2 \geq \chi^2_\alpha(df)$.

Beslissing?

H_0 niet verwerpen.

χ^2 tabel



	Proportion in Critical Region				
df	0.10	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.64	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28

Voorbeeld

Steekproef van $n = 57$ studenten (2013-3) met mobiel:

Goodness-of-fit	Android	iOS	Anders	geen
Verwacht %	47%	13%	10%	30%
Verwachte frequentie E_i	26.79	7.41	5.7	17.1
Geobserveerde freq. O_i	36	7	6	8

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} =$$
$$\frac{(36-26.79)^2}{26.79} + \frac{(7-7.41)^2}{7.41} + \frac{(6-5.7)^2}{5.7} + \frac{(8-17.1)^2}{17.1} = 8.047$$
met $K = 4$.

$df = 3$, $\chi^2_\alpha(3) = 7.815$, dus significant verschillend van verdeling over NL bevolking.

Voorwaarden voor een χ^2-toets

- De steekproefelementen zijn onafhankelijk van elkaar en willekeurig getrokken
- Iedere observatie kan in precies één cel van de tabel worden geklassificeerd
- De verwachte celfrequenties zijn voldoende groot, d.w.z.
  - minder dan 20% van de cellen heeft $E_i < 5$
  - geen enkele cel heeft $E_i < 1$

χ^2-toets type 2: onafhankelijkheid

Hangen twee nominale variabelen samen?

- Hangt gezinssamenstelling samen met type internetaansluiting
- Hangt keuze smartphone OS samen met afstudeerrichting?
- Hangt keuze spelcomputer samen met geslacht+leeftijd?

Kruistabel voorbeeld voorkeur Spelcomputer

Observed	Man	Vrouw	Kind	Totaal
Xbox	4	10	16	30
PlayStation	4	9	17	30
Wii	2	11	27	40
Totaal	10	30	60	100

Kruistabel voorbeeld voorkeur Spelcomputer

Observed	Man	Vrouw	Kind	Totaal
Xbox	4	10	16	30
PlayStation	4	9	17	30
Wii	2	11	27	40
Totaal	10	30	60	100

Expected	Man	Vrouw	Kind	Totaal
Xbox	??	??	??	30
PlayStation	??	??	??	30
Wii	??	??	??	40
Totaal	10	30	60	100

Kruistabel voorbeeld voorkeur Spelcomputer

Observed	Man	Vrouw	Kind	Totaal
Xbox	4	10	16	30
PlayStation	4	9	17	30
Wii	2	11	27	40
Totaal	10	30	60	100

Expected	Man	Vrouw	Kind	Totaal
Xbox	3	9	18	30
PlayStation	3	9	18	30
Wii	4	12	24	40
Totaal	10	30	60	100

χ^2-toets voor onafhankelijkheid kruistabel

- Hypotheses:
  - H_0: Elke groep (man, vrouw, kind) heeft dezelfde verdeling van voorkeuren over de verschillende spelcomputers, ofwel H_0: voorkeur is onafhankelijk van de groep.
  - H_1: bij minstens één van de groepen is de verdeling anders. (H_1: keuze spelcomputer is afhankelijk van groep)
- R = aantal rijen en C = aantal kolommen
- 'Expected' celaantallen $E_{ij} = \dfrac{(O_{i\cdot} \times O_{\cdot j})}{N} = \dfrac{\text{rijsom} \times \text{kolomsom}}{\text{totaal}}$
- Toetsstatistiek: $\chi^2 = \sum_{i,j} \dfrac{(O_{ij} - E_{ij})^2}{E_{ij}}$, met $df = (R-1)(C-1)$

Kruistabel voorbeeld voorkeur Spelcomputer

Obs./Exp.	Man	Vrouw	Kind	Totaal
Xbox	4 / 3	10 / 9	16 / 18	30 / 30
PlayStation	4 / 3	9 / 9	17 / 18	30 / 30
Wii	2 / 4	11 / 12	27 / 24	40 / 30
Totaal	10 / 10	30 / 30	60 / 60	100 / 100

$(O_{ij} - E_{ij})^2/E_{ij}$	Man	Vrouw	Kind	
Xbox	0.333	0.111	0.222	
PlayStation	0.333	0	0.056	
Wii	1.000	0.083	0.375	
			Sum =	2.513

Rekenvoorbeeld Voorkeur vs. Spelcomputer

- $\chi^2 = 2.513$
- $df = (R-1)\cdot(C-1) = 2\cdot 2 = 4$
- Opzoeken in Tabel: $\chi^2_\alpha(df = 4) = 9.488$, voor $\alpha = .05$.
- Dus? H_0 niet verwerpen.

Smartphone vs. Afstudeerrichting

Obs./Exp.	Android	iOS	Anders	Geen	Totaal
Informatica Classic	14 / 13.9	2 / 2.7	2 / 2.3	4 / 3.1	22 / 22.0
Informatica Gametech	22 / 22.1	5 / 4.3	4 / 3.7	4 / 4.9	35 / 35.0
Totaal	36 / 36.0	7 / 7.0	6 / 6.0	8 / 8.0	57 / 57.0

- $df = (R-1)\cdot(C-1) = 1\cdot 3 = 3$,
- $\chi^2(df = 3) = 7.815$,
- $\chi^2 = \frac{(14-13.9)^2}{13.9} + \frac{(2-2.7)^2}{2.7} + \ldots + \frac{(4-4.9)^2}{4.9} = 0.787$

Dus H_0 niet verwerpen: er is geen aanleiding te concluderen dat smartphone keuze afhangt van studierichting.

χ^2-test Type 2: Onafhankelijkheid

Oplossing voor lege cellen:

- Fischer Exact test (zie Wikipedia)
- Cellen samenvoegen (zorg voor logische samenvoegingen)

Tot zover

- Morgen: Werkcollege
- Volgende week: Toets
  - Datum: Dinsdag 11 oktober
  - Tijdstip: 13:30 – 15:30
  - Plaats: EDUC-GAMMA
  - Meenemen:
    - Calculator
    - A4 (tweezijdig) handgeschreven/bedrukt met formules en aantekeningen
  - Papier en kopieën van tabellen worden verstrekt
  - Let op: je moet zelf kunnen bepalen welke toets je moet gebruiken!