Les 5 Vergelijken van verdelingen

In de vorige les hebben we naar toetsen voor hypothesen gekeken, waarbij de hypothese een uitspraak over een parameter van een kansverdeling was, bijvoorbeeld over het gemiddelde of een relatieve frequentie. Maar als we bijvoorbeeld willen toetsen, of een dobbelsteen eerlijk is, zullen we na 120 worpen niet alleen maar het gemiddelde en de variantie bepalen, maar kijken of de getallen 1 t/m 6 alle ongeveer 20 keer gevallen zijn. Op deze manier zouden we natuurlijk onmiddellijk zien, dat de stochast X met

$$P(X=1)=\frac{5}{24},\quad P(X=2)=\frac{1}{6},\quad P(X=3)=\frac{1}{12},$$
$$P(X=4)=0,\quad P(X=5)=\frac{13}{24},\quad P(X=6)=0$$

geen eerlijke dobbelsteen beschrijft, terwijl $E[X]=3\frac{1}{2}$ en $Var(X)=\frac{35}{12}$, net zo als bij een eerlijke dobbelsteen (ga dit na). We zouden dus met een toets op het gemiddelde of de variantie niet aan het licht kunnen brengen dat de dobbelsteen oneerlijk is, maar natuurlijk zouden we dit ook niet op zo'n stomme manier proberen.

De dobbelsteen is een voorbeeld van een verdeling, waar we niet alleen maar een parameter van de kansverdeling willen toetsen, maar de volledige verdeling willen bekijken. De nulhypothese, die we in dit geval zouden toetsen is

$$H_0 : P(X=1)=\frac{1}{6},\ldots,P(X=6)=\frac{1}{6}$$

en de alternatieve hypothese luidt, dat niet alle van deze kansen gelijk aan $\frac{1}{6}$ zijn. Natuurlijk kunnen we niet verwachten, dat we bij een steekproef precies de kansen van de nulhypothese vinden, maar naarmate de steekproef groter wordt, zouden we steeds kleinere afwijkingen verwachten. Het vergelijken van de onder de nulhypothese verwachte aantallen en de daadwerkelijk waargenomen aantallen geeft aanleiding tot een belangrijke klasse van toetsen voor hypothesen over kansverdelingen, namelijk de χ^2-toetsen.

5.1 De χ^2-aanpassingstoets

De situatie die we nu bekijken is als volgt: Gegeven is een stochast X met een zekere kansverdeling, bijvoorbeeld de uniforme verdeling voor een eerlijke dobbelsteen. De nulhypothese is, dat een steekproef door de stochast X is voortgebracht en we willen toetsen of deze hypothese plausibel is.

De algemene aanpak is, de mogelijke uitkomsten van de stochast X in een aantal klassen in te delen. Voor een stochast met een discrete kansverdeling zijn de klassen vaak de verschillende mogelijke uitkomsten, maar soms is het handig verschillende uitkomsten in één klasse samen te vatten.

Voor continue kansverdelingen kiest men als klassen meestal intervallen, deze zijn vaak van dezelfde breedte, maar dit is niet noodzakelijk zo.

Voorbeeld: Voor een stochast $X \in \mathcal{N}(\mu,\sigma^2)$ waarvoor men een normale verdeling met verwachtingswaarde μ en variantie σ^2 verondersteld, worden de

intervalgrenzen vaak op veelvouden van de standaardafwijking σ gelegd. Men krijgt zo bijvoorbeeld de klassen

$$\begin{array}{ll} K_1 : -\infty < X < \mu - 3\sigma, & K_2 : \mu - 3\sigma \leq X < \mu - 2\sigma, \\ K_3 : \mu - 2\sigma \leq X < \mu - \sigma, & K_4 : \mu - \sigma \leq X < \mu, \\ K_5 : \mu < X \leq \mu + \sigma, & K_6 : \mu + \sigma \leq X < \mu + 2\sigma, \\ K_7 : \mu + 2\sigma \leq X < \mu + 3\sigma, & K_8 : \mu + 3\sigma \leq X < \infty \end{array}$$

Als de mogelijke uitkomsten van X in k klassen ingedeeld zijn, wordt voor elke van de klassen de kans p_i bepaalt, dat X een uitkomst in de i-de klasse produceert. Bij een steekproef van n stuks zullen we dan np_i waarden in de i-de klasse verwachten.

In het voorbeeld van de normale verdeling met 8 klassen kunnen we uit de standaard-normale verdeling de volgende kansen afleiden:

i	1	2	3	4	5	6	7	8
p_i	0.0013	0.0214	0.1359	0.3413	0.3413	0.1359	0.0214	0.0013

We beschrijven nu met een stochast X_i het aantal uitkomsten in een steekproef van n stuks, die in de i-de klasse vallen. Uit de verschillen van X_i en np_i moeten we nu een toets afleiden, die aangeeft of het plausibel is dat de steekproef volgens de veronderstelde kansverdeling is voortgebracht.

Voor het speciaal geval van slechts 2 klassen hebben we dit probleem al eerder bekeken, in dit geval vallen de uitkomsten met kans p in de eerste klasse en met kans $q = 1-p$ in de tweede klasse. Maar dit betekent, dat X de stochast van een Bernoulli-experiment met kans p is en de stochast X_1 is binomiaal verdeeld met parameters n en p. De relatieve frequentie p van een binomiale verdeling hadden we in de vorige les getoetst, door X_1 op een (bij benadering) standaard-normale verdeling te transformeren, namelijk door

$$Z := \frac{X_1 - np}{\sqrt{np(1-p)}}.$$

Als Z standaard-normaal verdeeld is, heeft Z^2 een χ^2-verdeling met 1 vrijheidsgraad en we kunnen Z^2 als volgt herschrijven:

$$\begin{aligned} Z^2 &= \frac{(X_1 - np)^2}{np(1-p)} = (1-p)\,\frac{(X_1 - np)^2}{np(1-p)} + p\,\frac{(X_1 - np)^2}{np(1-p)} \\ &= \frac{(X_1 - np)^2}{np} + \frac{((n - X_1) - n(1-p))^2}{n(1-p)} \\ &= \frac{(X_1 - np)^2}{np} + \frac{(X_2 - nq)^2}{nq}. \end{aligned}$$

We zien dus dat we Z^2 kunnen beschrijven als som van de kwadratische afwijkingen tussen waargenomen aantallen en verwachte aantallen, genormeerd op de verwachte aantallen.

In plaats van de waarde van Z met de z-waarden van de standaard-normale verdeling te vergelijken, kunnen we de waarde van Z^2 tegen de waarden χ^2_α van een χ^2-verdeling met 1 vrijheidsgraad toetsen die gedefinieerd zijn door

$$P(Z^2 > \chi^2_\alpha) = \alpha$$

want er geldt $P(Z^2 > \chi^2_\alpha) = P(Z > z_\alpha) = \alpha$.

De veralgemening van 2 tot k klassen is nu enigszins voor de hand liggend: De gekwadrateerde afwijkingen van de waargenomen aantallen van de verwachte aantallen worden door de verwachte aantallen gedeeld en deze hoeveelheden worden voor de verschillende klassen bij elkaar opgeteld. Het idee achter de normering op het aantal verwachte uitkomsten in een klasse is dat bij een verwacht aantal van 100 uitkomsten een afwijking van 3 minder sterk weegt dan bij een verwacht aantal van 10 uitkomsten. Men definieert dus de stochast χ^2 door

$$\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \frac{(X_1 - np_1)^2}{np_1} + \ldots + \frac{(X_k - np_k)^2}{np_k}.$$

Er laat zich aantonen dat χ^2 voor $n \to \infty$ een χ^2-verdeling met $k-1$ vrijheidsgraden heeft. Voor het geval $k = 2$ hebben we dit boven ingezien, want we hebben aangetoond dat

$$\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2$$

en het laatste heeft voor $n \to \infty$ inderdaad een χ^2-verdeling met 1 vrijheidsgraad. Het bewijs voor algemene k vergt behoorlijk meer moeite en wordt hier onderdrukt.

We geven wel een iets handigere manier aan om χ^2 uit te rekenen. Uit $(X_i - np_i)^2 = X_i^2 - 2X_i np_i + n^2 p_i^2$ volgt dat $\frac{(X_i - np_i)^2}{np_i} = \frac{X_i^2}{np_i} - 2X_i + np_i$. We hebben $\sum_{i=1}^k p_i = 1$ en omdat de som van de X_i het totaal aantal n van uitkomsten aangeeft, geldt $\sum_{i=1}^k X_i = n$. Hiermee krijgen we

$$\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{X_i^2}{np_i} - \sum_{i=1}^{k} 2X_i + \sum_{i=1}^{k} np_i = \left(\sum_{i=1}^{k} \frac{X_i^2}{np_i} \right) - n.$$

De kansverdeling die de verdeling van n uitkomsten over k klassen beschrijft, waarbij een uitkomst met kans p_i in de i-de klasse valt, heet de *multinomiale verdeling* met parameters $p_1, \ldots, p_k$ (die aan $p_1 + \ldots + p_k = 1$ moeten voldoen). Er geldt

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!}\, p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

waarbij $n_1 + \ldots + n_k = n$ is. De multinomiale verdeling voor het speciaal geval $k = 2$ is natuurlijk juist de binomiale verdeling.

Het idee van een toets, de zogeheten χ^2-*aanpassingstoets* of kort χ^2-*toets*, is nu hetzelfde als bij de toetsen die we in de vorige les hebben gezien. Voor de verschillende aantallen ν van vrijheidsgraden en de verschillende levels α van onbetrouwbaarheid worden waarden $\chi^2_{\nu,\alpha}$ bepaald zo dat

$$P(\chi^2 > \chi^2_{\nu,\alpha}) = \alpha.$$

Onder de aanname van de nulhypothese geeft een steekproef dus (slechts) met kans α een χ^2-waarde die zo groot of groter is dan χ^2 en de nulhypothese wordt verworpen als een waarde van χ^2 wordt gevonden die groter is dan $\chi^2_{\nu,\alpha}$ voor de gekozen level α. Vaak wordt ook hier de P-waarde van χ^2 bepaald, dus de kans waarmee de stochast X van de nulhypothese een steekproef oplevert die een χ^2-waarde oplevert die zo groot of groter is dan de gevonden χ^2.

Merk op: Een belangrijke voorwaarde voor de toepasbaarheid van de χ^2-toets is, dat voor iedere klasse de verwachte aantallen $np_i \geq 5$ zijn, want anders wordt de verdeling van χ^2 niet nauwkeurig genoeg door een χ^2-verdeling benaderd. Dit eist soms dat men klassen samenvoegt die anders te weinig waarnemingen laten verwachten.

In het voorbeeld van de normale verdeling heeft de klasse K_1 de verwachte relatieve frequentie $p_1 = 0.0013$: Om hier op $np_1 \geq 5$ te komen, moeten we een steekproef van grootte $n \geq 3847$ hebben. Als dit niet haalbaar is, kunnen we bijvoorbeeld de klassen K_1 en K_2 samenvoegen, de gecombineerde kans voor deze twee klassen is $p'_1 = 0.02275$ en om nu aan de voorwaarde $np'_1 \geq 5$ te voldoen is een steekproef van grootte $n \geq 220$ voldoende.

Voorbeeld: We nemen aan dat we voor onze oneerlijke dobbelsteen met kansen $(\frac{5}{24}, \frac{1}{6}, \frac{1}{12}, 0, \frac{13}{24}, 0)$ bij een steekproef met $n = 120$ worpen precies de juiste aantallen vinden, dus $(25, 20, 10, 0, 65, 0)$. Bij een eerlijke dobbelsteen is $p_1 = \ldots = p_6 = \frac{1}{6}$ en we zouden dus voor elke klasse 20 uitkomsten verwachten. De waarde voor χ^2 is in dit geval

$$\begin{aligned}\chi^2 &= \frac{(25-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(0-20)^2}{20} + \frac{(65-20)^2}{20} + \frac{(0-20)^2}{20} \\ &= \frac{1}{20}(25 + 0 + 100 + 400 + 2025 + 400) = 147.5\end{aligned}$$

en voor $\alpha = 0.01$ vind men in de tabellen voor een χ^2-verdeling met 5 vrijheidsgraden de waarde $\chi^2_{5,0.01} = 15.1$ en zelfs voor $\alpha = 0.001$ is $\chi^2_{5,0.001} = 20.5$ veel kleiner dan de gevonden waarde voor χ^2. De P-waarde voor $\chi^2 = 147.5$ is in feite $4.5 \cdot 10^{-30}$ dus is het nagenoeg uitgesloten dat een resultaat met zo'n grote waarde voor χ^2 toevallig door een eerlijke dobbelsteen opgeleverd zou worden.

Voorbeeld: Van een bepaalde plantensoort komen volgens de wetten van Mendel vier variaties voor in de verhouding $9 : 3 : 3 : 1$. De verwachte relatieve frequenties zijn dus $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$ en $p_4 = \frac{1}{16}$. In een steekproef van 160 exemplaren vindt men de volgende aantallen n_i, de met de verwachte aantallen np_i vergeleken worden:

	variatie				
	1	2	3	4	totaal
n_i	88	35	24	13	160
np_i	90	30	30	10	160

Omdat de verdeling 4 klassen bevat, hebben we de kritieke waarden van de χ^2-verdeling met 3 vrijheidsgraden nodig. Voor $\alpha = 0.1$ is $\chi^2_{3,0.1} = 6.25$ en voor $\alpha = 0.05$ is $\chi^2_{3,0.05} = 7.81$. Als waarde voor χ^2 krijgen we

$$\chi^2 = \frac{(88-90)^2}{90} + \frac{(35-30)^2}{30} + \frac{(24-30)^2}{30} + \frac{(13-10)^2}{10} \approx 2.98$$

dus geeft dit experiment niet eens op een onbetrouwbaarheidslevel van 10% evidentie tegen de wetten van Mendel. De P-waarde van $\chi^2 = 2.98$ is 0.395, dit betekent dat 39.5% van de steekproeven minstens een χ^2-waarde van 2.98 zou opleveren, dus is onze steekproef zeker geen atypisch resultaat.

Meestal wordt de χ^2-aanpassingstoets als rechtséénzijdige toets toegepast, die aangeeft wat de kans is dat een steekproef in het geval van de nulhypothese een zo grote χ^2-waarde geeft. Er zijn echter ook gevallen waarbij een tweezijdige χ^2-toets uitgevoerd wordt, omdat men steekproeven ook verdacht vindt, als ze *te goed* bij de nulhypothese passen.

Een voorbeeld hiervoor is het toetsen van *toevalsgetallen*. Voor toevalsgetallen tussen 0 en 1 kan men bijvoorbeeld als klassen bijvoorbeeld de deelintervallen van lengte 0.1 kiezen. Als een toevalsgenerator nu 10000 toevalsgetallen produceert, zou men ongeveer 1000 getallen in ieder deelinterval verwachten en men berekent hiervoor de waarde van χ^2. Natuurlijk mag χ^2 in dit geval niet te groot zijn, omdat dit evidentie tegen de nulhypothese geeft dat de toevalsgenerator onbevooroordeeld is. Maar omgekeerd geeft een te kleine χ^2-waarde aanleiding tot de aanname dat er te veel regelmaat in de toevalsgetallen zit en de rij toevalsgetallen voorspelbaar is. Dit is evidentie tegen de nulhypothese dat de toevalsgenerator de getallen onafhankelijk van elkaar produceert. Men zou in dit geval de toevalsgenerator als ongeschikt verwerpen als de χ^2-waarde niet tussen $\chi^2_{0.05}$ en $\chi^2_{0.95}$ ligt.

> Een van de grondleggers van de statistiek, R.A. Fisher, heeft de χ^2-toets op de experimenten van Gregor Mendel met erwten toegepast, waardoor deze tot de ontdekking van de genen werd geleid (zonder ze zo te noemen). Fisher kwam tot het resultaat dat χ^2 een P-waarde van 0.99996 had, dus slechts 4 in 100000 steekproeven zouden een zo kleine χ^2-waarde opleveren. Het lijkt erop dat Mendel's tuin assistent precies wist, welke uitslag Mendel bij zijn experimenten verwachte en hier een handje bij heeft geholpen.

De waarden $\chi^2_{\nu,\alpha}$

De $\chi^2_{\nu,\alpha}$-waarden zijn net zo als de z-waarden en t-waarden voor verschillende parameters ν en α in tabellen opgeslagen of worden door software pakketten berekend. Voor grotere aantallen van vrijheidsgraden zijn er zekere benaderingen die op het verband van de χ^2-verdeling met de normale verdeling berusten.

(1) Voor een stochast χ^2 met een χ^2-verdeling met ν vrijheidsgraden is

$$Z := \sqrt{2\chi^2} - \sqrt{2\nu - 1}$$

bij benadering standaard-normaal verdeeld, waarbij deze benadering zeker voor $\nu > 100$ toegepast mag worden. Door dit naar χ^2 op te lossen, volgt dat men $\chi^2_{\nu,\alpha}$ kan benaderen door

$$\chi^2_{\nu,\alpha} \approx \frac{1}{2}\left(z_\alpha + \sqrt{2\nu - 1}\right)^2.$$

(2) Een betere benadering krijgt men uit het feit dat ook

$$Z := \frac{\sqrt[3]{\frac{\chi^2}{\nu}} - (1 - \frac{2}{9\nu})}{\frac{2}{9\nu}}$$

bij benadering standaard-normaal verdeeld is. Oplossen hiervan naar χ^2 geeft de benadering

$$\chi^2_{\nu,\alpha} \approx \nu \cdot \left(1 - \frac{2}{9\nu} + z_\alpha\sqrt{\frac{2}{9\nu}}\right)^3.$$

Er wordt soms aangegeven de benadering (1) voor $\nu > 100$ en de betere benadering (2) voor $\nu > 30$ toe te passen, maar met deze voorwaarden zit men zeker aan de veilige kant.

Voor $\nu = 50$ en $\alpha = 0.05$ is bijvoorbeeld de juiste waarde $\chi_{50,0.05} = 67.5048$, benadering (1) geeft $\chi_{50,0.05} \approx 67.2189$ en benadering (2) $\chi_{50,0.05} \approx 67.5006$. Zelfs voor $\nu = 10$ en $\alpha = 0.05$ is de fout van de twee benaderingen nog klein, de juiste waarde is hier $\chi_{10,0.05} = 18.3070$, benadering (1) geeft $\chi_{10,0.05} \approx 18.0225$ en benadering (2) $\chi_{10,0.05} \approx 18.2918$.

Verschillende kritieke waarden $\chi^2_{\nu,\alpha}$ zijn in Tabel 3 te vinden. Voor aantallen van vrijheidsgraden die niet in de tabel genoteerd zijn, kan men (voor voldoende grote ν) de boven aangegeven benaderingen toepassen, of een waarde voor een hoger aantal vrijheidsgraden kiezen, die wel genoteerd is. Op deze manier wordt in ieder geval de kans op een type I fout niet vergroot.

Onbekende parameters

In veel gevallen wil men toetsen of een steekproef door een stochast met een zeker *type* van kansverdeling geproduceerd is, bijvoorbeeld met een binomiale verdeling of een normale verdeling. In dit geval hangt de verdeling voor de nulhypothese van onbekende parameters af die uit de steekproef geschat moeten worden. Bij een schatter voor het gemiddelde van een kansverdeling hebben we gezien dat door het vervangen van de variantie door een schatting de verdeling breder wordt, omdat er meer onzekerheid in de schatting zit. We moesten daarom de normale verdeling door de Student-t verdeling vervangen.

Iets soortgelijks gebeurt ook bij de χ^2-toetsen. Als we de parameters van de verdeling waarmee we de verwachte kansen p_i berekenen door schattingen

$\nu\backslash\alpha$	0.95	0.1	0.05	0.01	0.001
1	.0039	2.71	3.84	6.63	10.8
2	.103	4.61	5.99	9.21	13.8
3	.352	6.25	7.81	11.3	16.3
4	.711	7.78	9.49	13.3	18.5
5	1.15	9.24	11.1	15.1	20.5
6	1.64	10.6	12.6	16.8	22.5
7	2.17	12.0	14.1	18.5	24.3
8	2.73	13.4	15.5	20.1	26.1
9	3.33	14.7	16.9	21.7	27.9
10	3.94	16.0	18.3	23.2	29.6
12	5.23	18.5	21.0	26.2	32.9
15	7.26	22.3	25.0	30.6	37.7
20	10.9	28.4	31.4	37.6	45.3
25	14.6	34.4	37.7	44.3	52.6
30	18.5	40.3	43.8	50.9	59.7
40	26.5	51.8	55.8	63.7	73.4
50	34.8	63.2	67.5	76.2	86.7
70	51.7	85.5	90.5	100	112
100	77.9	118	124	136	149

Tabel 3: Kritieke waarden $\chi_{\nu,\alpha}$ voor de χ^2-verdelingen met ν vrijheidsgraden.

vervangen, passen we de waarden p_i al aan de steekproef aan, daarom wordt in dit geval de onzekerheid kleiner tegenover het geval van bekende parameters. Op een gegeven onbetrouwbaarheidslevel α moeten de kritieke waarden dus kleiner worden. Gelukkig laat zich bewijzen dat dit op een overzichtelijke manier gebeurt, er moet namelijk voor elke parameter die we uit de steekproef schatten één vrijheidsgraad afgetrokken worden. Er geldt:

Stelling: Als voor het berekenen van de verwachte kansen p_i voor een uitkomst in de i-de klasse r parameters voor de kansverdeling van X met een maximum likelihood schatting worden bepaald, dan heeft $\chi^2 := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$ voor $n \to \infty$ een χ^2-verdeling met $k - 1 - r$ vrijheidsgraden.

Merk op: Voor het gemiddelde μ van een verdeling is de maximum likelihood schatting gewoon het steekproefgemiddelde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ en voor de parameter p van een binomiale verdeling is $\overline{p} = \frac{k}{n}$ de maximum likelihood schatting, waarbij k het aantal successen bij n pogingen is. Aan de andere kant geldt dat de maximum likelihood schatting voor de variantie niet de steekproefvariantie $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is, maar $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n-1}{n}\, s^2$. Maar omdat de verdeling van χ^2 toch alleen maar voor $n \to \infty$ een χ^2-verdeling geeft, maakt het niet zo veel uit of we de variantie σ^2 door de (zuivere) schatting s^2 of de asymptotisch zuivere maximum likelihood schatting $\frac{n-1}{n}\, s^2$ vervangen. Vaak wordt daarom in de literatuur ook alleen maar aangegeven, dat een parameter door *een* schatting wordt vervangen, maar niet of door de maximum likelihood schatting of door een andere schatting.

Voorbeeld: Om het uur worden uit een productieproces steekproeven genomen van 5 stuks en het aantal defecte stukken wordt genoteerd. In 200 zulke steekproeven zijn de volgende resultaten gevonden:

aantal defecte stukken	0	1	2	3	4	5
aantal steekproeven	104	58	26	8	4	0

We willen toetsen of het aantal defecte stukken een binomiale verdeling heeft omdat dit het geval zou zijn als de kans op defecte stukken over de tijd constant gebleven is. Omdat de parameter p van de binomiale verdeling niet bekend is, moeten we deze uit de steekproeven schatten. We krijgen hiervoor

$$\overline{p} = \frac{1}{1000}(104 \cdot 0 + 58 \cdot 1 + 26 \cdot 2 + 8 \cdot 3 + 4 \cdot 4 + 0 \cdot 5) = \frac{150}{1000} = 0.15.$$

Als indeling van de steekproeven in klassen kiezen we de aantallen defecte stukken in een steekproef (van 5 stuks). De verwachte relatieve frequentie p_i voor de i-de klasse (met i defecte stukken) is dan volgens de binomiale verdeling met parameters $m = 5$ en $p = \overline{p} = 0.15$ gegeven door

$$p_i = \binom{m}{i}\overline{p}^i \cdot (1-\overline{p})^{m-i} = \binom{5}{i}0.15^i \cdot 0.85^{5-i}$$

en voor de $n = 200$ steekproeven krijgen we als verwachte aantallen voor de klassen

defect	0	1	2	3	4	5
p_i	0.444	0.392	0.138	0.024	0.002	0.0001
np_i	88.74	78.30	27.64	4.88	0.43	0.02

Omdat de verwachte aantallen voor de klassen met 3, 4 en 5 defecte stukken te klein zijn, voegen we deze samen tot één klasse met ≥ 3 defecte stukken. We krijgen zo de volgende statistiek waarvoor we de χ^2-waarde moeten bepalen:

defect	0	1	2	≥ 3
n_i	104	58	26	12
np_i	88.74	78.30	27.64	5.32

Omdat we de parameter $\overline{p}$ van de binomiale verdeling uit de steekproeven hebben geschat, heeft de χ^2-verdeling $4 - 1 - 1 = 2$ vrijheidsgraden. Op de levels $\alpha = 0.05$ en $\alpha = 0.01$ hebben we de kritieke waarden $\chi^2_{2,0.05} = 5.99$ en $\chi^2_{2,0.01} = 9.21$. Er geldt nu

$$\chi^2 = \frac{(104-88.74)^2}{88.74} + \frac{(58-78.30)^2}{78.30} + \frac{(26-27.64)^2}{27.64} + \frac{(12-5.32)^2}{5.32} \approx 16.37$$

dus kunnen we de nulhypothese van een binomiale verdeling zelfs op de onbetrouwbaarheidslevel $\alpha = 0.01$ veilig verwerpen. De P-waarde van $\chi^2 = 16.37$ is in feite 0.0003, een veel te lage waarde voor de aanname dat de afwijking van de binomiale verdeling toevallig is. We zouden dus concluderen, dat de kans p op defecte stukken in het productieproces over de tijd niet constant was.

5.2 χ^2-toets voor contingentietabellen

We hebben met de χ^2-aanpassingstoets getoetst of een steekproef bij een zekere kansverdeling past. Vaak komt men echter een iets andere vraag tegen, namelijk of twee of meer steekproeven bij een gemeenschappelijke kansverdeling horen, waarbij het niet nodig is deze gemeenschappelijke verdeling nader te bepalen. Dit probleem wordt meestal met een variatie van de χ^2-toets uit de vorige sectie aangepakt, waarbij men de verwachte aantallen uit de steekproeven bepaald. Hierbij gebruikt men een *contingentietabel*.

Stel we hebben r steekproeven met omvangen $n_1, \ldots, n_r$. Ieder van de steekproeven wordt op k klassen verdeeld, dit geeft de aantallen n_{ij} van elementen in de i-de steekproef, die in de j-de klasse vallen. We krijgen zo een $r \times k$-matrix met als elementen de hoeveelheden van elementen in de doorsnede van een steekproef en een klasse en dit noemen we een *contingentietabel*.

Met $n := \sum_{i=1}^{r} n_i = n_1 + \ldots + n_r$ noteren we de gemeenschappelijke omvang van alle steekproeven. We definiëren nu

$$p_j := \frac{n_{1j} + \ldots + n_{rj}}{n}$$

als kans voor een uitkomst in de j-de klasse, dit is juist de relatieve frequentie van uitkomsten in de j-de klasse in alle steekproeven. Met de kansen p_j krijgen we als verwachte waarde op positie (i, j) in de contingentietabel de waarde $n_i \cdot p_j$, want dit is het aantal uitkomsten in de j-de klasse die we bij een steekproef van omvang n_i zouden verwachten. We vatten nu de cellen van de contingentietabel als nieuwe klassen op en berekenen voor deze klassen de χ^2-waarde, dus

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - n_i \cdot p_j)^2}{n_i \cdot p_j}.$$

Er laat zich ook in dit geval aantonen, dat χ^2 voor $n \to \infty$ een χ^2-verdeling heeft, en het aantal vrijheidsgraden is $\nu = (r-1)(k-1)$. Dit kunnen we als volgt inzien: Als de p_j bekend waren, hadden we voor iedere steekproef $k-1$ vrijheidsgraden, dus in het geheel $r(k-1)$ vrijheidsgraden. Maar omdat we de p_j uit de steekproeven schatten, moeten we hiervan $k-1$ aftrekken (niet k, want p_k laat zich door $p_k = 1 - p_1 - \ldots - p_{k-1}$ uit de andere schattingen berekenen). Dit geeft dus $\nu = r(k-1) - (k-1) = (r-1)(k-1)$ vrijheidsgraden.

Voorbeeld: Bij een enquête in drie steden A, B en C werd een contingentietabel met de volgende resultaten gevonden:

stad	voor	tegen	neutraal	geen antwoord	totaal
A	105	61	87	167	420
B	118	60	130	145	453
C	88	58	62	101	309
totaal	311	179	279	413	1182

We hebben dus

$$n_1 = 420, \quad n_2 = 453, \quad n_3 = 309, \quad n = 1182,$$

$p_1 = \frac{311}{1182} \approx 0.263$, $p_2 = \frac{179}{1182} \approx 0.151$, $p_3 = \frac{279}{1182} \approx 0.236$, $p_4 = \frac{413}{1182} \approx 0.349$ en dit geeft als tabel met de verwachte aantallen $n_i \cdot p_j$:

stad	voor	tegen	neutraal	geen antwoord
A	110.5	63.6	99.1	146.8
B	119.2	68.6	106.9	158.3
C	81.3	46.8	72.9	108.0

Als we nu de waarde van χ^2 berekenen, zijn de cellen van de tabellen de nieuwe klassen en we krijgen

$$\chi^2 = \frac{(105 - 110.5)^2}{110.5} + \frac{(61 - 63.6)^2}{63.6} + \ldots + \frac{(101 - 108.0)^2}{108.0} \approx 17.2.$$

Dit moeten we vergelijken met de kritieke waarden van de χ^2-verdeling met $(3-1)\cdot(4-1) = 6$ vrijheidsgraden. We hebben $\chi^2_{6,0.05} = 12.6$ en $\chi^2_{6,0.01} = 16.8$, dus zijn de resultaten van de drie steden op de level $\alpha = 0.01$ significant verschillend.

In het geval van $r = 2$ steekproeven hebben we natuurlijk al eerder toetsen op verschillen van de verdelingen gezien, bijvoorbeeld toetsen op hetzelfde gemiddelde. Het hangt vaak van de vraagstukken af, of een χ^2-toets hier beter geschikt zou zijn. In het algemeen is de χ^2-toets minder scherp dan een toets op verschillen van de gemiddelden, aan de andere kant kan deze ook nog verschillen detecteren als de gemiddelden wel overeenkomen. In het bijzonder is de χ^2-toets ook toepasbaar, als de veronderstelling van een normaal verdeelde schatter niet meer houdbaar is.

Voorbeeld: Bij een niet nader toegelicht experiment met mogelijke uitslagen $1, \ldots, 10$ worden met twee verschillende methoden I en II de volgende aantallen uitslagen bereikt:

methode	1	2	3	4	5	6	7	8	9	10	totaal
I	6	16	22	38	44	30	18	12	8	6	200
II	2	6	12	22	29	30	21	16	8	4	150
totaal	8	22	34	60	73	60	39	28	16	10	350

Als geschatte kansen p_j voor de uitkomsten krijgen we

j	1	2	3	4	5	6	7	8	9	10
p_j	0.023	0.063	0.097	0.171	0.209	0.171	0.111	0.080	0.046	0.029

en als we hiermee de χ^2-waarde berekenen, krijgen we $\chi^2 \approx 11.12$. Voor een χ^2-verdeling met $(2-1)\cdot(10-1) = 9$ vrijheidsgraden hebben we $\chi_{9,0.1} = 14.7$, dus geeft de χ^2-toets met onbetrouwbaarheid $\alpha = 0.1$ geen evidentie voor een verschil van de twee methoden. De P-waarde van $\chi^2 = 11.12$ is 0.268.

Maar we kunnen met onze kennis uit de vorige les natuurlijk ook toetsen, of de twee methoden hetzelfde gemiddelde hebben. Hiervoor kijken we naar de

steekproefgemiddelden $\overline{x_I}$ en $\overline{x_{II}}$ en de steekproefvarianties s_I^2 en s_{II}^2 voor de twee steekproeven met omvangen $n_I = 200$ en $n_{II} = 150$. We hebben

$$\overline{x_I} = \frac{1}{200}(6 \cdot 1 + \ldots + 6 \cdot 10) = 5.05, \quad \overline{x_{II}} = \frac{1}{150}(2 \cdot 1 + \ldots + 4 \cdot 10) = 5.67$$

$$s_I^2 = 4.29, \quad s_{II}^2 = 3.86$$

en hieruit krijgen we voor de gepoolde variantie s^2 en standaardafwijking s:

$$s^2 = \frac{(n_I - 1)s_I^2 + (n_{II} - 1)s_{II}^2}{n_I + n_{II} - 2} = \frac{199 \cdot s_I^2 + 149 \cdot s_{II}^2}{348} = 4.11, \quad s = 2.03.$$

Als t-waarde die we met de kritieke waarden van de Student-t verdeling met 348 vrijheidsgraden moeten toetsen, hebben we

$$t = \frac{\overline{x_{II}} - \overline{x_I}}{s\sqrt{\frac{1}{n_I} + \frac{1}{n_{II}}}} \approx 2.82.$$

De verdeling van t is nagenoeg een standaard-normale verdeling en als P-waarde voor $t = 2.82$ vinden we 0.0024, dus vinden we met deze toets een significant verschil voor de gemiddelden van de twee methoden.

Toets op onafhankelijkheid van kenmerken

Een variatie op het vergelijken van r steekproeven geeft een toets op onafhankelijkheid van twee kenmerken in een steekproef. Bijvoorbeeld wil men weten, of het interesse in verschillende studievakken onafhankelijk is van het geslacht van de student. Men interpreteert nu de studenten van de verschillende studievakken als verschillende steekproeven en de indeling vrouw/man als indeling in klassen. De nulhypothese is, dat de kenmerken onafhankelijk zijn, in dit geval zou de kansverdeling voor iedere steekproef hetzelfde zijn en we zijn terug bij de situatie van de vorige sectie.

Voor het gemak nemen we aan dat het eerste kenmerk de waarden $\{1, \ldots, r\}$ kan hebben en het tweede kenmerk de waarden $\{1, \ldots, k\}$. Als n elementen in de steekproef zitten, noteren we met n_{ij} het aantal elementen met waarde i voor het eerste kenmerk en waarde j voor het tweede kenmerk. Als schatting p_{i*} voor de relatieve frequentie van elementen met waarde i voor het eerste kenmerk krijgen we

$$p_{i*} := \frac{n_{i1} + \ldots + n_{ik}}{n}$$

en als schatting p_{*j} voor de relatieve frequentie van elementen met waarde j voor het tweede kenmerk krijgen we

$$p_{*j} := \frac{n_{1j} + \ldots + n_{rj}}{n}.$$

De kansen p_{i*} en p_{*j} heten ook *marginale kansen*, omdat ze met de totale aantallen corresponderen die we aan de rand van de contingentietabel schrijven.

Onder de aanname van de nulhypothese zijn de twee kenmerken onafhankelijk, dus is de kans op een uitkomst in de cel (i, j) van de contingentietabel

$p_{i*} \cdot p_{*j}$ en het verwachte aantal uitkomsten voor deze cel is dus $n \cdot p_{i*} \cdot p_{*j}$. We kijken dus in dit geval naar de χ^2-waarde

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(n_{ij} - np_{i*}p_{*j})^2}{np_{i*}p_{*j}}$$

en er laat zich weer aantonen dat dit voor $n \to \infty$ een χ^2-verdeling heeft. Omdat het schatten van de p_{i*} uit de steekproef $r-1$ vrijheidsgraden wegneemt en het schatten van de p_{*j} het aantal vrijheidsgraden om $k-1$ reduceert, hebben we $\nu = rk - 1 - (r-1) - (k-1) = (r-1)(k-1)$ vrijheidsgraden.

Voorbeeld: In een onderzoek werd getoetst of de prestaties van leerlingen in de vakken Engels en Wiskunde onafhankelijk van elkaar zijn. Men deelt de resultaten in 3 klassen, cijfers 6 en lager, cijfers 7 en 8 en cijfer 9 en 10.

	Wiskunde			
Engels	≤ 6	$7,8$	$9,10$	totaal
≤ 6	85	42	14	141
$7,8$	38	163	47	248
$9,10$	12	71	56	139
totaal	135	276	117	528

Hieruit krijgen we voor de marginale kansen:

$$p_{1*} = \frac{141}{528} = 0.267, \quad p_{2*} = \frac{248}{528} = 0.470, \quad p_{3*} = \frac{139}{528} = 0.263$$

$$p_{*1} = \frac{135}{528} = 0.256, \quad p_{*2} = \frac{276}{528} = 0.523, \quad p_{*3} = \frac{117}{528} = 0.222$$

Onder de aanname van de nulhypothese dat de twee kenmerken onafhankelijk zijn, zouden we voor de combinatie (i,j) van de kenmerken $n \cdot p_{i*} \cdot p_{*j}$ leerlingen in de steekproef verwachten. Dit geeft de verwachte waarden in de volgende tabel:

	Wiskunde		
Engels	≤ 6	$7,8$	$9,10$
≤ 6	36.1	73.7	31.2
$7,8$	63.4	129.6	55.0
$9,10$	35.5	72.7	30.8

We zien al dat dit behoorlijk afwijkt van de gevonden waarden. Als we hiervoor de χ^2-waarde berekenen, krijgen we

$$\chi^2 = \frac{(85-36.1)^2}{36.1} + \frac{(42-73.7)^2}{73.7} + \ldots + \frac{(56-30.8)^2}{30.8} \approx 145.8$$

terwijl we voor een χ^2-verdeling met $(3-1)\cdot(3-1) = 4$ vrijheidsgraden op significantie level $\alpha = 0.001$ de waarde $\chi^2_{4,0.001} = 18.5$ vinden. Het is dus duidelijk dat de resultaten in de twee vakken niet onafhankelijk van elkaar zijn.

Yates-correctie

In het speciaal geval van een 2×2 contingentietabel wordt vaak de *Yates-correctie* toegepast, die rekening ermee houdt, dat de data discreet is, maar de χ^2-verdeling een continue kansverdeling. In het algemeen wordt de χ^2-waarde met Yates-correctie bij l klassen met kansen $p_1, \ldots, p_l$ berekend door

$$\chi^2 := \sum_{i=1}^{l} \frac{(|X_i - np_i| - \frac{1}{2})^2}{np_i}$$

maar dit wordt eigenlijk alleen maar in het geval van 1 vrijheidsgraad toegepast, en dit is juist het geval voor $r = 2$ en $k = 2$.

De Yates-correctie heeft het effect dat de χ^2-waarde die berekent wordt iets lager is dan zonder de correctie. Dit leidt ertoe dat de nulhypothese met Yates-correctie minder snel verworpen wordt dan zonder Yates-correctie.

Voor grote steekproeven maakt de Yates-correctie bijna geen verschil en inmiddels wordt soms aanbevolen, de Yates-correctie *niet* toe te passen. Als alle gevonden aantallen van de cellen kleine zijn (bijvoorbeeld tussen 5 en 10 liggen) is het verstandig om de χ^2-waarde met en zonder Yates-correctie te bepalen. Als de twee manieren tot verschillende conclusies leiden (verwerpen van de nulhypothese bij de ene, niet verwerpen bij de andere), zou men de steekproef moeten vergroten om tot een duidelijke beslissing te kunnen komen.

Voorbeeld: In een proef wordt aan een groep van mensen met een bepaalde ziekte een nieuw medicijn gegeven, terwijl een tweede groep met dezelfde ziekte een placebo krijgt. Er wordt nu gekeken hoe veel van de mensen binnen een bepaalde periode gezond zijn geworden.

	gezond	ziek	totaal
medicijn	75	25	100
placebo	65	35	100
totaal	140	60	200

Als marginale kansen krijgen we hieruit

$$p_{1*} = p_{2*} = \frac{100}{200} = 0.5 \quad \text{en} \quad p_{*1} = \frac{140}{200} = 0.7, \quad p_{*2} = \frac{60}{200} = 0.3.$$

De aanname van onafhankelijkheid betekent in dit geval dat de nieuwe medicijn hetzelfde effect heeft als het placebo. Omdat de groepen even groot zijn, zouden we onder de aanname van onafhankelijkheid verwachten dat in beide groepen $200 \cdot 0.5 \cdot 0.7 = 70$ mensen gezond worden en dat $200 \cdot 0.5 \cdot 0.3 = 30$ ziek blijven.

Zonder Yates-correctie krijgen we hieruit de χ^2-waarde

$$\chi^2 = \frac{(75-70)^2}{70} + \frac{(25-30)^2}{30} + \frac{(65-70)^2}{70} + \frac{(35-30)^2}{30} \approx 2.38$$

en met Yates-correctie

$$\chi^2 = \frac{(|75-70|-0.5)^2}{70} + \frac{(|25-30|-0.5)^2}{30} + \frac{(|65-70|-0.5)^2}{70} + \frac{(|35-30|-0.5)^2}{30} \approx 1.93.$$

In beide gevallen kunnen we de nulhypothese op onafhankelijkheid op een level van $\alpha = 0.1$ niet verwerpen, want voor een χ^2-verdeling met 1 vrijheidsgraad vinden we $\chi^2_{1,0.1} = 2.71$. De P-waarde zonder Yates-correctie is 0.123 en de P-waarde met Yates-correctie is 0.165 en dit zijn allebij geen afzonderlijk kleine waarden. Om aan te tonen dat de nieuwe medicijn wel een effect heeft, zijn dus verdere experimenten nodig.

2×2-tabellen

In het voorbeeld hier boven hebben we kunnen zien, dat bij een 2×2-contingentietabel de tellers in de som voor χ^2 alle hetzelfde zijn (in het voorbeeld 5^2). Dit is geen toeval, maar in feite altijd het geval voor 2×2-tabellen en heeft tot gevolg dat we voor dit belangrijke speciaal geval de χ^2-waarde op een veel makkelijkere manier kunnen uitrekenen.

> Het zal geen verrassing zijn, dat een 2×2-tabel een speciaal geval is, want hier gaan we toetsen of twee relatieve frequenties hetzelfde zijn. In de vorige les hebben we gezien, dat we dit voor twee relatieve frequenties p_1 en p_2 kunnen doen, door de z-waarde
>
> $$z := \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$
>
> met $p_0 := \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$ te berekenen, die onder de aanname van de nulhypothese $p_1 = p_2$ standaard-normaal verdeeld is. De waarde χ^2 voor de χ^2-toets die we nu gaan berekenen is in dit speciaal geval juist het kwadraat van z.

We noteren de 2-contingentietabel als volgt:

	A	B	totaal
1	a	b	n_1
2	c	d	n_2
totaal	n_A	n_B	n

Hiervoor berekenen we de χ^2-waarde door

$$\begin{aligned}
\chi^2 &= \frac{(a - \frac{n_1 n_A}{n})^2}{\frac{n_1 n_A}{n}} + \frac{(b - \frac{n_1 n_B}{n})^2}{\frac{n_1 n_B}{n}} + \frac{(c - \frac{n_2 n_A}{n})^2}{\frac{n_2 n_A}{n}} + \frac{(d - \frac{n_2 n_B}{n})^2}{\frac{n_2 n_B}{n}} \\
&= \frac{n}{n_1 n_2 n_A n_B}\Bigg(n_2 n_B (a - \frac{n_1 n_A}{n})^2 + n_2 n_A (b - \frac{n_1 n_B}{n})^2 \\
&+ n_1 n_B (c - \frac{n_2 n_A}{n})^2 + n_1 n_A (d - \frac{n_2 n_B}{n})^2 \Bigg) \\
&= \frac{n}{n_1 n_2 n_A n_B}\Bigg(\frac{n_2 n_B}{n^2}(na - n_1 n_A)^2 + \frac{n_2 n_A}{n^2}(nb - n_1 n_B)^2 \\
&+ \frac{n_1 n_B}{n^2}(nc - n_2 n_A)^2 + \frac{n_1 n_A}{n^2}(nd - n_2 n_B)^2 \Bigg).
\end{aligned}$$

Dit ziet nog niet naar een verbetering uit, maar nu vullen we in dat $n = a+b+c+d$, $n_1 = a+b$, $n_2 = c+d$, $n_A = a+c$ en $n_B = b+d$. Dit geeft

$$\begin{aligned} na - n_1 n_A &= (a+b+c+d)a - (a+b)(a+c) \\ &= a^2 + ab + ac + ad - a^2 - ab - ac - bc = ad - bc =: \Delta. \end{aligned}$$

Op een soortgelijke manier zien we in, dat ook

$$nb - n_1 n_B = \Delta, \quad nc - n_2 n_A = \Delta, \quad nd - n_2 n_B = \Delta.$$

Dit is in feite het bewijs, dat we in de tellers van de termen voor χ^2 altijd hetzelfde getal vinden, namelijk $(\frac{\Delta}{n})^2$.

Als we nu nog invullen dat $n_1 + n_2 = n$ en $n_A + n_B = n$, zien we dat $n_2 n_B + n_2 n_A + n_1 n_B + n_1 n_A = n_2(n_B + n_A) + n_1(n_B + n_A) = (n_2 + n_1)n = n^2$ en daarom geldt

$$\frac{n_2 n_B}{n^2}\Delta^2 + \frac{n_2 n_A}{n^2}\Delta^2 + \frac{n_1 n_B}{n^2}\Delta^2 + \frac{n_1 n_A}{n^2}\Delta^2 = \Delta^2 = (ad-bc)^2.$$

Alles bij elkaar genomen, hebben we dus aangetoond dat

$$\chi^2 = \frac{n}{n_1 n_2 n_A n_B}(ad-bc)^2$$

en dit is voor 2×2-contingentietabllen inderdaad veel handiger dan de algemene formule van boven.

5.3 Variantie-analyse

Met de χ^2-toetsen zijn we nagegaan of verschillende steekproeven bij dezelfde verdeling horen. Vaak komt men echter ook de vraag tegen of meerdere verdelingen hetzelfde gemiddelde hebben, bijvoorbeeld als het om verschillende behandelingen van een zekere soort groente gaat. Voor twee steekproeven hebben we hier al naar gekeken, dit konden we met een toets op het verschil van de twee gemiddelden oplossen. Hiervoor hadden we onder de veronderstelling dat de twee steekproeven uit verdelingen met dezelfde variantie komen, gekeken naar de verdeling van de schatter

$$T := \frac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

waarbij $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ de gepoolde variantie van de steekproeven was.

Net zo als we met de χ^2-toets een veralgemening van het vergelijken van 2 relatieve frequenties op relatieve frequenties voor k klassen hebben gevonden, gaan we nu de toets op gelijkheid van gemiddelden op meer dan 2 steekproeven uitbreiden.

Het idee hierbij is, de totale variantie van de steekproeven te analyseren en deze te verdelen in de variantie binnen de enkele steekproeven en de variantie tussen de steekproeven. Daarom heet deze methode dan ook *variantie-analyse* of kort *ANOVA* (voor **AN**alysis **O**f **VA**riance).

We veronderstellen, dat we k steekproeven hebben die afkomstig zijn van normale verdelingen met dezelfde (onbekende) variantie σ^2 en met (onbekende) verwachtingswaarden $\mu_1, \ldots, \mu_k$. De i-de steekproef heeft omvang n_i en wordt met $x_{i1}, \ldots, x_{in_i}$ genoteerd. De totale omvang van alle steekproeven is $n := n_1 + \ldots + n_k$. De nulhypothese is

$$H_0 : \mu_1 = \ldots = \mu_k.$$

We berekenen de steekproefgemiddelden $\overline{x_i}$ en het gemiddelde $\overline{x}$ *en gros* (d.w.z. het gemiddelde over alle steekproeven), dus

$$\overline{x_i} := \frac{1}{n_i} \sum_j x_{ij} \quad \text{en} \quad \overline{x} := \frac{1}{n} \sum_{i,j} x_{ij} = \sum_i \frac{n_i}{n} \overline{x_i}.$$

De totale kwadratische afwijking

$$v := \sum_{i,j} (x_{ij} - \overline{x})^2$$

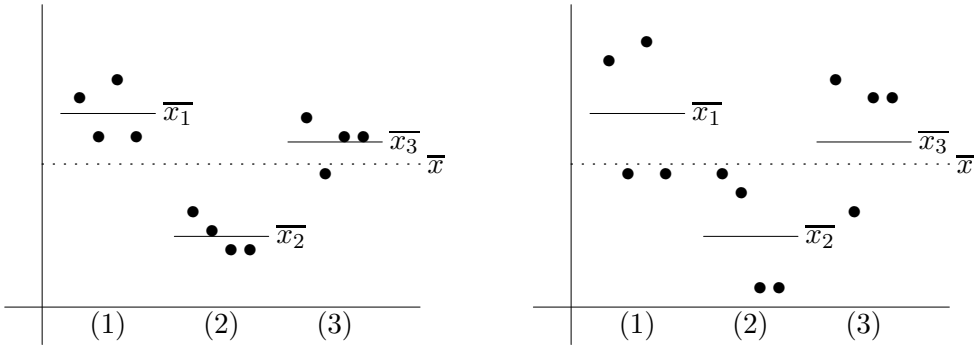
heeft nu twee bronnen, namelijk de kwadratische afwijkingen

$$v_i := \sum_j (x_{ij} - \overline{x_i})^2$$

binnen de steekproeven en de kwadratische afwijking

$$\sum_i (\overline{x_i} - \overline{x})^2.$$

tussen de steekproeven. Het idee achter deze opsplitsing van de kwadratische afwijkingen is in het volgende plaatje te zien:



In beide plaatjes zien we 3 steekproeven met telkens 4 waarden en de steekproefgemiddelden $\overline{x_i}$ zijn in beide gevallen hetzelfde.

In het linkerplaatje liggen de elementen van de steekproeven dicht bij de steekproefgemiddelden, daarom is de bijdrage van de kwadratische afwijkingen binnen de steekproeven in dit geval klein en de totale kwadratische afwijking wordt vooral veroorzaakt door de afwijkingen tussen de steekproefgemiddelden. Dit is sterke evidentie tegen de nulhypothese dat de gemiddelden van de verdelingen gelijk zijn.

In het rechterplaatje zijn de kwadratische afwijkingen binnen de steekproeven veel groter terwijl de kwadratische afwijkingen tussen de steekproefgemiddelden nog steeds hetzelfde zijn. Omdat in dit geval de kwadratische afwijkingen binnen de steekproeven relatief een groter deel bijdragen aan de totale kwadratische afwijking, zou men de nulhypothese moeilijker kunnen verwerpen, want de grote spreiding binnen de steekproeven maakt het plausibel, dat alle steekproeven door een verdeling met hetzelfde gemiddelde voortgebracht zijn.

Om het opsplitsen van de totale kwadratische afwijking binnen en tussen de steekproeven precies te analyseren, maken we weer gebruik van onze succesvolle aanpak, de elementen x_{ij} van de steekproeven als realisaties van onafhankelijke stochasten X_{ij} te zien. Ons uitgangspunt is hierbij, dat $X_{ij} \in \mathcal{N}(\mu_i, \sigma^2)$ is, dus normaal verdeeld met gemiddelde μ_i en variantie σ^2. De schatters $\overline{X_i}$ voor de gemiddelden van de steekproeven en $\overline{X}$ voor het gemiddelde over alle steekproeven zijn dan gegeven door

$$\overline{X_i} := \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad \text{en} \quad \overline{X} := \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k}\frac{n_i}{n}\overline{X_i}.$$

Er geldt nu

$$\begin{aligned}\sum_{i,j}(X_{ij}-\overline{X})^2 &= \sum_{i,j}((X_{ij}-\overline{X_i})+(\overline{X_i}-\overline{X}))^2\\ &= \sum_{i,j}(X_{ij}-\overline{X_i})^2 + \sum_{i,j}(\overline{X_i}-\overline{X})^2 + 2\sum_{i,j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X})\\ &= \sum_{i,j}(X_{ij}-\overline{X_i})^2 + \sum_{i} n_i(\overline{X_i}-\overline{X})^2 + 2\sum_{i,j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X}).\end{aligned}$$

Maar de laatste som geeft 0, omdat

$$\begin{aligned}\sum_{j}(X_{ij}-\overline{X_i})(\overline{X_i}-\overline{X}) &= (\overline{X_i}-\overline{X})(\sum_{j}(X_{ij}-\overline{X_i}))\\ &= (\overline{X_i}-\overline{X})((\sum_{j}X_{ij}) - n_i\overline{X_i}) = (\overline{X_i}-\overline{X})(n_i\overline{X_i}-n_i\overline{X_i}) = 0.\end{aligned}$$

Dus hebben we aangetoond dat

$$\sum_{i,j}(X_{ij}-\overline{X})^2 = \underbrace{\sum_{i,j}(X_{ij}-\overline{X_i})^2}_{V_b} + \underbrace{\sum_{i} n_i(\overline{X_i}-\overline{X})^2}_{V_t}.$$

We gaan nu de twee stochasten V_b (b voor **b**innen) en V_t (t voor **t**ussen) die gedefinieerd zijn door

$$V_b := \sum_{i,j}(X_{ij}-\overline{X_i})^2 \quad \text{en} \quad V_t := \sum_{i} n_i(\overline{X_i}-\overline{X})^2$$

apart onderzoeken.

We weten dat $S_i^2 = \frac{1}{n_i-1}\sum_j (X_{ij} - \overline{X_i})^2$ een zuivere schatter voor σ^2 is, daarom is $\sum_j (X_{ij} - \overline{X_i})^2$ een zuivere schatter voor $(n_i - 1)\sigma^2$. De som V_b over de kwadratische afwijkingen *binnen* de steekproeven is dus een zuivere schatter voor $\sum_i (n_i - 1)\sigma^2 = (n-k)\sigma^2$ en dus geldt:

$$S_b^2 := \frac{V_b}{n-k} \text{ is een zuivere schatter voor } \sigma^2.$$

Om de variantie tussen de steekproeven te analyseren, schrijven we de stochasten X_{ij} voor de uitkomsten in de steekproeven als $X_{ij} = \mu_i + E_{ij}$, waarbij E_{ij} de afwijking van de verwachtingswaarde μ_i van X_{ij} aangeeft. In het bijzonder is E_{ij} normaal verdeeld met verwachtingswaarde 0 en variantie σ^2.

Omdat de schatters $\overline{X_i}$ verwachtingswaarde μ_i hebben, heeft $\overline{X}$ de verwachtingswaarde

$$\mu := \frac{1}{n}\sum_i n_i \mu_i.$$

We schrijven nu $\mu_i = \mu + \alpha_i$, dan zijn de α_i juist de afwijkingen tussen de gemiddelden van de enkele verdelingen en het gemiddelde over alle verdelingen. In het bijzonder volgt uit $\mu = \frac{1}{n}\sum_i n_i\mu_i$ dat

$$\sum_i n_i\alpha_i = \sum_i n_i(\mu_i - \mu) = (\sum_i n_i\mu_i) - n\mu = 0.$$

Voor de stochast V_t geldt nu:

$$\begin{aligned}
V_t &= \sum_i n_i(\overline{X_i} - \overline{X})^2 = \sum_i n_i((\overline{X_i} - \mu_i) + (\mu - \overline{X}) + (\mu_i - \mu))^2 \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 + \sum_i n_i(\mu - \overline{X})^2 + \sum_i n_i(\mu_i - \mu)^2 \\
&+2\sum_i n_i(\overline{X_i} - \mu_i)(\mu - \overline{X}) +2\sum_i n_i(\overline{X_i} - \mu_i)(\mu_i - \mu) +2\sum_i n_i(\mu - \overline{X})(\mu_i - \mu) \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 + n(\mu - \overline{X})^2 + \sum_i n_i\alpha_i^2 \\
&+ 2(\mu - \overline{X})\underbrace{\sum_i n_i(\overline{X_i} - \mu_i)}_{n(\overline{X}-\mu)} +2\sum_i n_i(\overline{X_i} - \mu_i)\alpha_i + 2(\mu - \overline{X})\sum_i n_i\alpha_i \\
&= \sum_i n_i(\overline{X_i} - \mu_i)^2 - n(\mu - \overline{X})^2 + \sum_i n_i\alpha_i^2 + 2\sum_i n_i(\overline{X_i} - \mu_i)\alpha_i
\end{aligned}$$

We kijken nu naar de verwachtingswaarde van V_t: Omdat $E[\overline{X_i}] = \mu_i$ geldt, is $E[(\overline{X_i} - \mu_i)^2] = \frac{\sigma^2}{n_i}$ en omdat $E[\overline{X}] = \mu$ is $E[(\overline{X} - \mu)^2] = \frac{\sigma^2}{n}$. Verder hebben we natuurlijk $E[\overline{X_i} - \mu_i] = 0$, daarom geldt

$$\begin{aligned}
E[V_t] &= \sum_i n_i E[(\overline{X_i} - \mu_i)^2] - nE[(\mu - \overline{X})^2] + \sum_i n_i\alpha_i^2 + 2\sum_i n_i\alpha_i E[(\overline{X_i} - \mu_i)] \\
&= \sum_i n_i\frac{\sigma^2}{n_i} - n\frac{\sigma^2}{n} + \sum_i n_i\alpha_i^2 = (k-1)\sigma^2 + \sum_i n_i\alpha_i^2.
\end{aligned}$$

De nulhypothese luidt dat alle μ_i hetzelfde zijn, dus dat alle $\alpha_i = 0$ zijn, de alternatieve hypothese is, dat minstens een $\alpha_i \neq 0$ is. Hieruit volgt:

(1) Onder de aanname van de nulhypothese $\alpha_i = 0$ voor alle i is

$$S_t^2 := \frac{V_t}{k-1} \text{ is een zuivere schatter voor } \sigma^2.$$

(2) Onder de aanname van de alternatieve hypothese $\alpha_i \neq 0$ voor een i is

$$S_t^2 := \frac{V_t}{k-1} \text{ is een zuivere schatter voor } \sigma^2 + \frac{1}{k-1}\sum_i n_i\alpha_i^2 > \sigma^2.$$

Voor gegeven steekproeven berekenen we nu de concrete realisaties s_b^2 en s_t^2 van de schatters S_b^2 en S_t^2 voor σ^2, dus

$$s_b^2 := \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x_i})^2 \quad \text{en} \quad s_t^2 := \frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{x_i} - \overline{x})^2.$$

Omdat onder de aanname van de nulhypothese S_b^2 en S_t^2 beide zuivere schatters voor σ^2 zijn, kunnen we in dit geval verwachten dat $s_b^2 \approx s_t^2$. Andersom geeft een waarde $s_t^2 \gg s_b^2$ evidentie tegen de nulhypothese. Men kijkt daarom naar de verdeling van de stochast

$$F := \frac{S_t^2}{S_b^2}$$

waarvoor men in het geval van de nulhypothese een waarde rond 1 verwacht. Analoog met de andere toetsen bepaalt men nu weer f-waarden f_α, zo dat onder de aanname van de nulhypothese steekproeven met een waarde van f_α of hoger voor F met kans α optreden, dus

$$P(F > f_\alpha) = \alpha.$$

Omdat men bij de nulhypothese een waarde van F rond 1 verwacht, zullen de $f_\alpha > 1$ zijn. Bij de F-toets met onbetrouwbaarheid α verwerpt men nu de nulhypothese als $\frac{s_t^2}{s_b^2} > f_\alpha$ is.

> De naam *variantie-analyse* voor de F-toets zou inmiddels duidelijk zijn. Men analyseert hoe veel van de totale kwadratische afwijking door de afwijkingen binnen de steekproeven veroorzaakt wordt en hoeveel door de afwijkingen tussen de steekproeven. Als het laatste relatief gezien te veel wordt, geeft dit evidentie tegen de nulhypothese dat de verdelingen van de steekproeven alle hetzelfde gemiddelde hebben.

De verdeling van F heet de *Fisher-verdeling* of *F-verdeling* en wordt afgeleid uit de χ^2-verdelingen.

De F-verdeling van Fisher

We weten dat $\frac{k-1}{\sigma}\,S_t^2$ een χ^2-verdeling χ^2_{k-1} met $k-1$ vrijheidsgraden heeft en $\frac{n-k}{\sigma}\,S_b^2$ een χ^2-verdeling χ^2_{n-k} met $n-k$ vrijheidsgraden. Hieruit volgt dat de F-verdeling gegeven is door

$$F = \frac{S_t^2}{S_b^2} = \frac{\frac{\chi^2_{k-1}}{k-1}}{\frac{\chi^2_{n-k}}{n-k}}$$

dus is F (tot op constanten na) een quotiënt van χ^2-verdeelde stochasten met $k-1$ en $n-k$ vrijheidsgraden. Deze twee aantallen van vrijheidsgraden karakteriseren de F-verdeling en we noteren de F-verdeling met $k-1$ en $n-k$ vrijheidsgraden met $F_{k-1,n-k}$.

Voor de geïnteresseerde lezer vermelden we hier de expliciete dichtheidsfunctie $f_{m,n}$ voor de F-verdeling $F_{m,n}$. Het zal geen verrassing zijn, dat deze op een quotiënt van de dichtheidsfuncties van χ^2-verdelingen lijkt:

$$f_{m,n}(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\,\Gamma(\frac{n}{2})} m^{\frac{m}{2}} n^{\frac{n}{2}} x^{\frac{m}{2}-1} (n+mx)^{-\frac{m+n}{2}}$$

De verwachtingswaarde en variantie van $F_{m,n}$ zijn

$$E[F_{m,n}] = \frac{n}{n-2} \quad \text{en} \quad Var(F_{m,n}) = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)}.$$

In het speciaal geval met $k=2$ steekproeven laat zich aantonen dat de verdeling $F_{1,n}$ juist de verdeling van het kwadraat T^2 van een stochast T met Student-t verdeling met n vrijheidsgraden is.

Verder geldt dat voor $n \to \infty$ de verdeling $F_{m,n}$ tegen de verdeling van $\frac{\chi^2_m}{m}$ convergeert en voor $m \to \infty$ gaat $F_{m,n}$ tegen $\frac{n}{\chi^2_n}$.

Variantie-analyse tabellen

De resultaten van een variantie-analyse worden meestal in een bepaalde soort tabellen aangegeven, die er typisch als volgt uit zien:

bron	vrijheids-graden	kwadratische afwijkingen	schattingen voor σ^2	F-waarde	P-waarde
tussen	$k-1$	$\sum_i n_i(\overline{x_i}-\overline{x})^2$	s_t^2	$f = \frac{s_t^2}{s_b^2}$	$P(F_{k-1,n-k} > f)$
binnen	$n-k$	$\sum_{i,j}(x_{ij}-\overline{x_i})^2$	s_b^2		
totaal	$n-1$	$\sum_{i,j}(x_{ij}-\overline{x})^2$			

Voorbeeld: Bij vier leveranciers van een zekere stof worden steekproeven genomen en de zuiverheid van de stof bepaald (die in procent aangegeven wordt). De vraag is, of er evidentie tegen de nulhypothese is, dat de vier leveranciers even zuiver produceren. De steekproeven en hun gemiddelden zijn in de volgende tabel aangegeven:

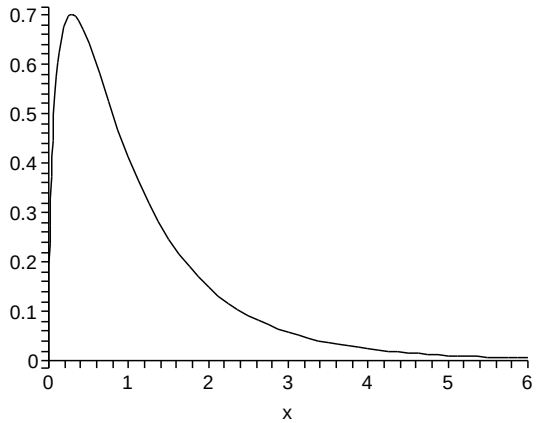
leverancier	steekproeven					n_i	$\overline{x_i}$
1	99.3	99.4	98.8	99.4		4	99.225
2	99.8	97.4	98.9	99.0	98.6	5	98.740
3	98.2	97.2	96.4	98.3		4	97.525
4	98.7	99.6	99.2			3	99.167
totaal						16	98.6375

We hebben $k = 4$ leveranciers en $n = 16$ steekproeven, daarom hebben we de F-verdeling met 3 en 12 vrijheidsgraden nodig. Uit deze gegevens berekent men de volgende variantie-analyse tabel:

bron	vrijheids-graden	kwadratische afwijkingen	schattingen voor σ^2	F-waarde	P-waarde
tussen	3	7.224	2.408	4.726	0.021
binnen	12	6.114	0.509		
totaal	15	13.337			

Afhankelijk van de gebruikte software wordt de P-waarde niet berekend, in dit geval vindt men in de tabellen voor $\alpha = 0.05$ de kritieke waarde $f_{3,12,0.05} = 3.49$ en voor $\alpha = 0.01$ de kritieke waarde $f_{3,12,0.01} = 5.95$. Men zou dus op een onbetrouwbaarheidslevel van 5% de nulhypothese wel kunnen verwerpen, maar op een onbetrouwbaarheidslevel van 1% niet meer. De P-waarde van 0.021 zegt juist, dat onder de aanname van de nulhypothese slechts 2.1% van de steekproeven een F-waarde van 4.726 of groter zouden opleveren.

We zien ook in Figuur 17 dat de gevonden waarde 4.726 van F al redelijk ver in de staart van de F-verdeling ligt, dus zou men in dit geval in ieder geval twijfels hebben of de leveranciers even zuivere stof produceren.



Figuur 17: F-verdeling met 3 en 12 vrijheidsgraden.

Belangrijke begrippen in deze les

- χ^2-aanpassingstoets

- kritieke waarden $\chi^2_{\nu,\alpha}$
- χ^2-toets bij onbekende parameters
- contingentietabel
- χ^2-toets op onafhankelijkheid
- Yates-correcte
- variantie-analyse (ANOVA)
- afwijkingen binnen en tussen steekproeven
- F-verdeling van Fisher
- F-toets

Opgaven

28. Er wordt 120 keer met een dobbelsteen geworpen. De aantallen voor de verschillende uitkomsten zijn:

$$1:12, \quad 2:21, \quad 3:27, \quad 4:22, \quad 5:20, \quad 6:18.$$

Is dit een zuivere dobbelsteen?

29. Bij een reukproef werd aan 50 willekeurig gekozen vrouwen gevraagd of zij parfum A lekkerder vonden dan B of omgekeerd. Aan A gaven 37 vrouwen de voorkeur, de overige vonden B lekkerder. Toets op de significantie level $\alpha = 0.1$ de nulhypothese dat er geen voorkeur voor één van de twee merken bestaat. Voer de toets zonder en met Yates-correctie uit.

30. In een weverij zijn in het verleden gemiddeld 2 weeffouten per $100m^2$ geweven doek opgetreden. Een recente steekproef op 100 stukken doek van $100m^2$ heeft het volgende resultaat opgeleverd:

fouten	0	1	2	3	4	5	6	7	8	9	10
aantal doeken	16	22	28	15	8	3	3	1	2	1	1

(i) Toets op een significantie level van $\alpha = 0.05$ de nulhypothese dat het aantal fouten Poisson-verdeeld met parameter $\lambda = 2$ is.

(ii) Toets op een significantie level van $\alpha = 0.05$ de nulhypothese dat het aantal fouten überhaupt Poisson-verdeeld is.

31. Van 1000 aselect gekozen personen is nagegaan of ze kleurenblind zijn. Van de 480 mannen bleken dit er 38 te zijn, bij de vrouwen was het aantal 6.

(i) Toets op de level $\alpha = 0.1$ of kleurenblindheid onafhankelijk is van het geslacht.

(ii) Wat is het minimale aantal vrouwen dat kleurenblind mag zijn, waarvoor de nulhypothese op level $\alpha = 0.1$ niet verworpen wordt (waarbij we nog steeds van 38 kleurenblinde mannen uit gaan)?

32. Twee groepen A en B van elk 100 patiënten hebben een bepaalde ziekte. Groep A wordt behandeld met een zeker serum, groep B met een ander serum. Na een bepaalde tijd zijn 75 patiënten van groep A en 65 patiënten van groep B genezen. Toets met onbetrouwbaarheid $\alpha = 0.05$ of beide sera evenveel effect hebben.

33. Bij een computerbedrijf wordt in 3 ploegen (ochtend, middag, nacht) op vier verschillende types van computers (A, B, C, D) gewerkt. De manager vraagt zich af of er bij het aantal reboots van computers een samenhang tussen de ploeg en de type computer bestaat. Hij heeft de volgende contingentietabel voor reboots gemaakt:

	type computer			
	A	B	C	D
ochtend	5	3	2	7
middag	7	12	9	16
nacht	1	2	4	2

Wat kan hij op een onbetrouwbaarheidslevel van $\alpha = 0.05$ zeggen?

34. Bij een crash-test met telkens 6 auto's van 3 verschillende merken wordt gekeken, wat de herstelling van de auto's kost. Er worden de volgende resultaten verkregen:

	kosten
A	200€ 50€ 150€ 75€ 100€ 250€
B	75€ 470€ 20€ 140€ 220€ 210€
C	120€ 570€ 600€ 450€ 700€ 350€

Kan op grond van deze waarden de nulhypothese dat de gemiddelde kosten bij iedere merk hetzelfde zijn op een onbetrouwbaarheidslevel van $\alpha = 0.05$ verworpen worden? Hoe zit het met $\alpha = 0.01$? De relevante kritieke waarden voor de F-verdeling zijn $f_{2,15,0.05} = 3.68$ en $f_{2,15,0.01} = 6.36$.