Les 4 Toetsen van hypothesen

We hebben tot nu toe enigszins algemeen naar grootheden van populaties gekeken en bediscussieerd hoe we deze grootheden uit steekproeven kunnen schatten. Vaak hebben we echter redelijk concrete voorstellingen over de waarde van een zeker parameter. In dit geval kan het resultaat van een steekproef onze idee over de parameter steunen of aanduiden dat we ons waarschijnlijk vergissen.

Vaak wordt deze situatie door het opstellen van een *hypothese* gerealiseerd en een steekproef kan wel of niet evidentie voor het verwerpen van de hypothese geven. We zullen zien dat het toetsen van een hypothese min of meer een herformulering van de ideeën achter intervalschatters en in het bijzonder betrouwbaarheidsintervallen zijn.

4.1 Hypothesen

In een hypothese maken we een uitspraak over een eigenschap van een stochast, bijvoorbeeld over de verwachtingswaarde. Hiervoor geven we aan dat een parameter θ waarvan de kansverdeling van de stochast afhangt een zekere waarde heeft. Vervolgens proberen we aan de hand van een steekproef voor de stochast evidentie voor of tegen de hypothese te vinden. Als we bijvoorbeeld de hypothese hebben dat de gemiddelde Nederlander $180cm$ groot is, dan geeft een (aselecte) steekproef van 1000 Nederlanders met een steekproefgemiddelde van $190cm$ hier sterke evidentie tegen, terwijl een steekproefgemiddelde van $181cm$ dit niet doet.

Hypothesen worden altijd in paren bekeken:

(i) De *nulhypothese* H_0 zegt dat een parameter θ een zekere waarde θ_0 heeft.

(ii) De *alternatieve hypothese* H_1 of H_a zegt dat de parameter θ van θ_0 afwijkt.

In het eenvoudigste geval zien de hypothesen er dus als volgt uit:

$$H_0 : \theta = \theta_0 \qquad H_1 : \theta \neq \theta_0.$$

In het voorbeeld van de gemiddelde grootte houdt de alternatieve hypothese de mogelijkheden in, dat de gemiddelde Nederlander (duidelijk) groter of kleiner is dan $180cm$. Dit geval leidt tot een *tweezijdige toets*.

Vaak is men echter alleen maar geïnteresseerd of een parameter in een zekere richting van de nulhypothese afwijkt. Bijvoorbeeld wil een sporter weten of hij door een nieuwe training methode (of door een nieuw dopingmiddel) harder kan lopen dan eerder. In dit geval zijn de hypothesen

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_0$$

en dit geeft aanleiding tot een *rechtséénzijdige* toets, want met de alternatieve hypothese gaan we na of de parameter θ naar *rechts* van de nulhypothese afwijkt.

Analoog test men met een *linkséénzijdige* toets of de parameter θ naar *links* van de nulhypothese afwijkt, in dit geval zijn de hypothesen

$$H_0 : \theta \geq \theta_0 \qquad H_1 : \theta < \theta_0.$$

Definitie: Een *toets* is een procedure die op grond van een steekproef beslist of de nulhypothese verworpen wordt of niet.

Bij een toets kunnen er twee soorten van fouten gemaakt worden omdat het gemiddelde van een steekproef (met een geringe kans) sterk van het gemiddelde van de volledige populatie kan afwijken:

I: De nulhypothese wordt verworpen terwijl hij juist is.

Dit heet een *type I fout* of een *fout van de eerste soort*. De kans α op een type I fout heet de *onbetrouwbaarheid* (of *onbetrouwbaarheidsdrempel*) van de toets.

II: De nulhypothese wordt niet verworpen terwijl hij onjuist is.

Dit heet een *type II fout* of een *fout van de tweede soort*. De kans β op een type II fout levert het *onderscheidingsvermogen* (power) $1-\beta$ van de toets.

We kunnen deze terminologie in het volgende schema weergeven:

	H_0 is juist	H_0 is onjuist
H_0 niet verwerpen	juiste beslissing kans $1-\alpha$	type II fout kans β
H_0 verwerpen	type I fout kans α	juiste beslissing kans $1-\beta$

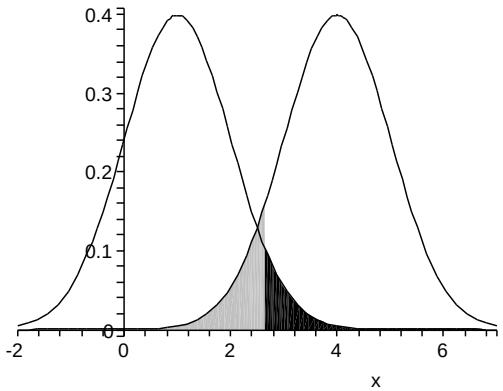
Het is natuurlijk heel eenvoudig, de kans op een type I fout te minimaliseren door de nulhypothese bijna nooit te verwerpen. Maar dit betekent dat veel resultaten van steekproeven als niet strijdig met H_0 geaccepteerd worden die eigenlijk evidentie voor de alternatieve hypothese geven. In dit geval is dus de kans op een type II fout hoog en het onderscheidingsvermogen van de toets slecht.

Merk op dat het onderscheidingsvermogen $1-\beta$ van een toets alleen bepaald kan worden als de alternatieve hypothese $H_1 : \theta \neq \theta_0$ vervangen wordt door een concrete alternatieve hypothese

$$H_1 : \theta = \theta_1.$$

Vaak worden toetsen vergeleken, door bij een vaste onbetrouwbaarheid α naar het onderscheidingsvermogen te kijken. De betere toets heeft dan het hogere onderscheidingsvermogen. Men kan ook het onderscheidingsvermogen $1-\beta$ als functie van de onbetrouwbaarheid opvatten, dit geeft de zogeheten *operating characteristic*. (Let wel: Er zijn ongeveer zo veel definities van operating characteristic als er auteurs zijn, maar de achterliggende gedachten zijn hetzelfde.) Een ideale toets zou al voor zeer kleine waarden van α naar een onderscheidingsvermogen $1-\beta$ dicht bij 1 stijgen.

In Figuur 16 is het concept van type I en type II fouten geïllustreerd. We kijken hierbij naar de nulhypothese $H_0 : \theta = 1$ en kiezen een onbetrouwbaarheid α van $\alpha = 0.05$. Het zwarte gebied onder de linker normale verdeling heeft juist de oppervlakte 0.05, dus leiden steekproefwaarden $\overline{\theta}$ die in dit gebied vallen tot verwerpen van de nulhypothese. Als we als alternatieve hypothese $H_1 : \theta = 4$ nemen, dan is de kans op een type II fout de oppervlakte onder de rechter normale verdeling, waar we de nulhypothese niet verwerpen, dus het grijze gebied. In het voorbeeld is deze oppervlakte ongeveer 0.0877, dus is het onderscheidingsvermogen van deze toets ongeveer 92.2%.



Figuur 16: Gebieden voor type I (zwart) en type II fouten (grijs).

4.2 Toetsen en betrouwbaarheidsintervallen

Aan de hand van het begrip van een type I fout kunnen we nu een verband leggen tussen toetsen en betrouwbaarheidsintervallen. We hadden een betrouwbaarheidsinterval op level γ rond een schatting $\overline{\theta}$ van een parameter zo gekozen, dat over alle mogelijke steekproeven gezien het interval de juiste waarde van θ met kans γ bevat. Dit was equivalent met de uitspraak, dat de schatting $\overline{\theta}$ met kans γ binnen het interval rond θ met dezelfde lengte als het betrouwbaarheidsinterval valt, omdat dit interval juist de kansmassa γ bevat.

Deze aanpak kunnen we nu omdraaien om een toets met onbetrouwbaarheid $\alpha = 1 - \gamma$ te krijgen: Voor de nulhypothese $H_0 : \theta = \theta_0$ kiezen we een interval $[\theta_-, \theta_+]$ rond θ_0 zo dat onder de aanname dat H_0 juist is de kans op een steekproefwaarde $\overline{\theta}$ buiten dit interval hoogstens α is, dus

$$P(\theta_- \leq \overline{\theta} \leq \theta_+) = 1 - \alpha = \gamma.$$

Als de schatting $\overline{\theta}$ buiten het interval $[\theta_-, \theta_+]$ ligt, wordt dit als evidentie *tegen* de nulhypothese H_0 beschouwd omdat dit slechts met de (kleine) kans α gebeurt en in dit geval wordt de nulhypothese verworpen.

Bij éénzijdige toetsen is het interval $[\theta_-, \theta_+]$ aan een kant open, omdat we de nulhypothese alleen maar bij afwijking in één richting verwerpen:

- Bij een *rechtséénzijdige toets* wordt H_0 verworpen, als de schatting $\overline{\theta}$ buiten het interval $[-\infty, \theta_+]$ ligt, dus als $\overline{\theta}$ te sterk naar *rechts* van de nulhypothese afwijkt.
- Bij een *linkséénzijdige toets* wordt H_0 verworpen, als de schatting $\overline{\theta}$ buiten het interval $[\theta_-, \infty]$ ligt, dus als $\overline{\theta}$ te sterk naar *links* van de nulhypothese afwijkt.

Merk op: Het lijkt op het eerste gezicht verwarrend, dat bij een rechtséénzijdige toets het interval $[-\infty, \theta_+]$ waarvoor we de nulhypothese niet verwerpen naar *links* open is, terwijl het rechtséénzijdige betrouwbaarheidsinterval voor een schatting naar *rechts* open is. Maar dit schijnbare paradox maakt juist het verband tussen toetsen en betrouwbaarheidsintervallen duidelijk:

Stelling: Het betrouwbaarheidsinterval op level $\gamma = 1 - \alpha$ rond een schatting $\overline{\theta}$ bevat precies de waarden θ_0 waarvoor $\overline{\theta}$ bij een toets met onbetrouwbaarheid α geen aanleiding geeft om de nulhypothese $\theta = \theta_0$ te verwerpen.

Andersom: Een toets met onbetrouwbaarheid α verwerpt de nulhypothese $H_0 : \theta = \theta_0$ op grond van de schatting $\overline{\theta}$ dan en slechts dan als θ_0 buiten het betrouwbaarheidsinterval van level $\gamma = 1 - \alpha$ rond $\overline{\theta}$ valt.

Toetsen voor gemiddelden

In de meeste situaties zal onder de voorwaarde dat de nulhypothese juist is de schatter T voor de schattingen $\overline{\theta}$ een normale verdeling met gemiddelde θ_0 en variantie $\frac{\sigma^2}{n}$ hebben. Dit is in het bijzonder het geval als T de schatter voor het gemiddelde van een normale verdeling is, maar bij benadering ook voor de schatter van het gemiddelde van niet-normale verdelingen (als n niet te klein is). In dit geval weten we dat de stochast

$$Z := \frac{T - \theta_0}{\frac{\sigma}{\sqrt{n}}} = \frac{(T - \theta_0)\sqrt{n}}{\sigma}$$

standaard-normaal verdeeld is en we kunnen daarom net zo als bij de betrouwbaarheidsintervallen met behulp van de z-waarden makkelijk een interval aangeven, dat een *tweezijdige toets* met onbetrouwbaarheid α oplevert, want er geldt

$$P\left(\theta_0 - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq T \leq \theta_0 + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

We zullen bij deze toets de nulhypothese dus verwerpen als de schatting $\overline{\theta}$ meer dan $z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ van θ_0 afwijkt, dus als

$$|\overline{\theta} - \theta_0| > z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}.$$

Dit zou namelijk onder de aanname van H_0 slechts met kans α gebeuren en omdat de kans α laag is, geeft dit evidentie tegen H_0. De kans α dat de beslissing om H_0 te verwerpen onjuist is, is juist de kans op een type I fout.

Merk op: De foutmarge rond θ_0 die we toelaten zonder H_0 te verwerpen is precies hetzelfde als de foutmarge die we voor het betrouwbaarheidsinterval rond $\overline{\theta}$ hebben gekozen. Dit is geen toeval, omdat de definitie van een toets met onbetrouwbaarheid α in principe alleen maar een herformulering van de definitie van een betrouwbaarheidsinterval van level $1-\alpha$ is.

Als we een *rechtséénzijdige toets* met onbetrouwbaarheid α willen hebben, moeten we een interval $[-\infty, \theta_+]$ vinden zo dat $P(T > \theta_+) = \alpha$. Maar omdat

$$P\left(T \leq \theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

is $[-\infty, \theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}}]$ zo'n interval en we verwerpen $H_0 : \theta \leq \theta_0$ als

$$\overline{\theta} > \theta_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Analoog krijgen we een *linkséénzijdige toets* met onbetrouwbaarheid α door H_0 te verwerpen als

$$\overline{\theta} < \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}},$$

want $P(T < \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}) = \alpha$, of te wel

$$P\left(T \geq \theta_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Voorbeeld: Een eierhandelaar koopt een grote partij eieren van een kippenfokker. We mogen aannemen dat het gewicht X van de eieren in een homogene partij normaal verdeeld is en dat de standaardafwijking van de gewichten $6g$ is. De fokker garandeert dat het gemiddelde van de eieren in deze partij boven de $60g$ ligt. De handelaar neemt nu een steekproef van 5 eieren en constateert dat deze samen $275g$ wegen. Hij wil de levering alleen maar reclameren als hij de nulhypothese $H_0 : \mu = 60$ op een onbetrouwbaarheidslevel van $\alpha = 0.05$ kan verwerpen. Omdat hij natuurlijk alleen maar bij te lichte eieren gaat reclameren, past hij een linkséénzijdige toets toe. Er geldt $z_{0.05} = 1.6449$ en dus zal hij de nulhypothese verwerpen, als zijn schatting $\overline{\mu}$ voldoet aan $\overline{\mu} < 60 - z_{0.05}\frac{6}{\sqrt{5}} \approx 55.6$. Zijn steekproef geeft $\overline{\mu} = \frac{275}{5} = 55$, dus zal hij inderdaad reclameren.

Aanpassingen bij kleine steekproeven

We zijn er tot nu toe van uit gegaan dat de schatter T voor de schattingen $\overline{\theta}$ de variantie $\frac{\sigma^2}{n}$ heeft. Vaak is de hiervoor benodigde variantie σ^2 van de onderliggende kansverdeling echter onbekend, in dit geval wordt de variantie $\frac{\sigma^2}{n}$ vervangen door de schatting $\frac{s^2}{n}$, waarbij s^2 de steekproefvariantie is. Maar het vervangen van σ^2 door de schatting s^2 leidt ertoe dat de getransformeerde stochast

$$\frac{(T - \theta_0)\sqrt{n}}{s}$$

geen normale verdeling maar een Student-t verdeling met $n-1$ vrijheidsgraden heeft. We moeten dus de z-waarden in de boven aangegeven intervallen voor de verschillende toetsen vervangen door de t-waarden van de Student-t verdeling, net zo als bij de betrouwbaarheidsintervallen. We krijgen dus een tweezijdige toets met onbetrouwbaarheid α door de nulhypothese H_0 te verwerpen als

$$|\overline{\theta} - \theta_0| > t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Bij de rechts- en linkséénzijdige toetsen zijn de criteria voor het verwerpen van de nulhypothese analoog

$$\overline{\theta} > \theta_0 + t_{n-1,\alpha} \frac{s}{\sqrt{n}} \quad \text{en} \quad \overline{\theta} < \theta_0 - t_{n-1,\alpha} \frac{s}{\sqrt{n}}.$$

Als n groot is (meestal wordt hier $n \geq 50$ als vuistregel gehanteerd), ligt de Student-t verdeling met $n-1$ vrijheidsgraden zo dicht bij de standaard-normale verdeling, dat deze correctie verwaarloosd kan worden omdat dan $z_\alpha \approx t_{n-1,\alpha}$ is. Maar bij onbekende variantie σ^2 *en* kleine steekproeven moeten de toetsen inderdaad zo als aangegeven aangepast worden.

Toetsen voor relatieve frequenties

Stel we willen de hypothese toetsen dat defecte stukken bij een productie met kans p_0 optreden, dus dat de parameter p van een binomiale verdeling gelijk is aan p_0. Hiervoor tellen we met de stochast X het aantal k van successen bij n pogingen en krijgen hiermee de schatting $\overline{p} = \frac{k}{n}$ voor p. We weten dat bij een niet te kleine steekproef ($np_0 \geq 5$, $n(1-p_0) \geq 5$) de stochast

$$Z := \frac{X - np_0}{\sqrt{np_0(1-p_0)}}$$

bij benadering standaard-normaal verdeeld is. Voor de de standaard-normale verdeling geldt (zie boven) dat $P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1-\alpha$, dus is

$$P\left(np_0 - z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)} \leq X \leq np_0 + z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)}\right) = 1-\alpha,$$

dus zullen we bij een tweezijdige toets met onbetrouwbaarheid α de nulhypothese $H_0 : p = p_0$ verwerpen als bij een steekproef met k successen in n pogingen geldt dat

$$|k - np_0| > z_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)}.$$

Als we beide zijden door n delen, kunnen we dit ook rechtstreeks als criterium voor de relatieve frequenties formuleren, we verwerpen de nulhypothese als

$$|\overline{p} - p_0| > z_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}.$$

De rechts- en linkséénzijdige toetsen kunnen we inmiddels zonder na te denken afleiden, we verwerpen bij de relatieve frequenties de nulhypothese H_0 als

$$\overline{p} > p_0 + z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \text{ (rechts) of } \overline{p} < p_0 - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}} \text{ (links)}.$$

Voorbeeld: Een handelaar verkoopt een grote partij goederen en deelt de koper mee dat er hoogstens 5% ondeugdelijke exemplaren in zitten. Om dit te verifiëren neemt de koper een steekproef van 150 stuks. Hij zal reclameren als hij op een onbetrouwbaarheidslevel van $\alpha = 0.05$ de bewering van de handelaar kan verwerpen. Omdat $0.05 \cdot 150 = 7.5 > 5$, kunnen we de normale benadering van de binomiale verdeling toepassen. Te koper zal natuurlijk alleen maar bij een te hoog aantal ondeugdelijke exemplaren reclameren, daarom moeten we een rechtséénzijdige toets toepassen. Er geldt $z_{0.05} = 1.6449$, $n = 150$ en $p_0 = 0.05$, dus is $z_\alpha \sqrt{np_0(1-p_0)} \approx 4.39$, de koper zal dus vanaf $7.5 + 4.39$, dus vanaf 12 ondeugdelijke stukken reclameren.

Als een steekproef te klein is om de normale benadering toe te passen, is het meestal mogelijk de kans op een steekproef met k of meer successen expliciet met de binomiale verdeling te berekenen, namelijk door

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}.$$

Bij een rechtséénzijdige toets wordt H_0 verworpen als $P(X \geq k) < \alpha$. Analoog berekent men met

$$P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

de kans op een steekproef met hoogstens k successen en verwerpt bij een linksséénzijdige toets de nulhypothese als $P(X \leq k) < \alpha$.

Bij een tweezijdige toets hangt het criterium ervan af of $k > np_0$ of $k < np_0$. Als kans dat een steekproef zo sterk van p_0 afwijkt als $\overline{p}$ krijgt men in dit geval $2 \cdot \min(P(X \geq k), P(X \leq k))$ omdat ook met de afwijking in de andere richting rekening gehouden moet worden. Als criterium voor het verwerpen van de nulhypothese krijgt men zo

$$\min(P(X \geq k), P(X \leq k)) < \frac{\alpha}{2}.$$

Significantie en P-waarden

Als we een toets zo opzetten dat we de nulhypothese verwerpen als de schatting voor een parameter θ buiten het betrouwbaarheidsinterval van level $\gamma = 1 - \alpha$ rond de nulhypothese θ_0 ligt, dan noemen we α ook de *significantie level* van de toets. De significantie is dus gelijk aan de kans op een type I fout onder de aanname dat de nulhypothese juist is.

We noemen een resultaat dus *significant op level* α als de kans dat dit resultaat optreedt terwijl de nulhypothese geldt, hoogstens α is.

Het woord *significant* (van het Latijnse *signum* = teken) is gekozen om aan te duiden, dat het gevonden resultaat iets *betekent* en niet meer als *toevallige afwijking* beschouwd kan worden.

Soms ligt een schatting $\overline{\theta}$ veel verder af van de nulhypothese dan het betrouwbaarheidsinterval op de gekozen level α aangeeft. De schatting geeft dus zelfs op een hogere level nog evidentie tegen de nulhypothese. In dit geval kijkt men vaak naar de hoogste mogelijke waarde van α, zo dat de schatting nog net tot verwerpen van de nulhypothese zou leiden en noemt dit de *P-waarde* van de schatting:

Definitie: De P-waarde p van een schatting $\overline{\theta}$ geeft aan dat onder de aanname van de nulhypothese $H_0 : \theta = \theta_0$ steekproeven die verder dan $\overline{\theta}$ van θ_0 afwijken slechts met kans p voorkomen.

De P-waarde van een schatting maakt dus een kwantitatieve uitspraak over de evidentie tegen de nulhypothese, terwijl een gewone toets met significantie level α alleen maar aangeeft of de evidentie sterker dan een gekozen level is of niet.

Soms wordt de mate van significantie met zekere intervallen van P-waarden verbonden, men leest bijvoorbeeld aanduidingen zo als

$P < 0.001$:	zeer sterk significant
$0.001 < P < 0.01$:	sterk significant
$0.01 < P < 0.05$:	zwak significant

maar er bestaan geen conventies die enigszins uniform gehandhaafd worden.

4.3 Toetsen op verschillen tussen twee verdelingen

We hebben tot nu toe naar de situatie gekeken dat we een hypothese over een parameter van een kansverdeling hebben en deze hypothese met een steekproef willen toetsen. In de praktijk is echter vaak een iets andere vraag van belang, namelijk of een parameter bij twee verdelingen dezelfde waarde heeft, dus bijvoorbeeld of twee verdelingen hetzelfde gemiddelde hebben. In dit geval is het niet zo interessant wat de waarden van de gemiddelden zijn, maar alleen maar of hun verschil 0 is of niet.

In plaats van een enkele steekproef moeten we hier voor ieder van de twee verdelingen een aparte steekproef nemen, en de verdelingen van de schattingen met behulp van deze steekproeven worden door twee onafhankelijke schatters T_1 en T_2 beschreven.

We gaan ervan uit dat T_1 een zuivere schatter voor de parameter θ_1 van de eerste verdeling en T_2 een zuivere schatter voor de parameter θ_2 van de tweede verdeling is. Verder veronderstellen we dat de varianties σ_1^2 en σ_2^2 van de twee verdelingen bekend zijn en we steekproeven van grootte n_1 en n_2 nemen. In dit geval geldt

$$E[T_1 - T_2] = \theta_1 - \theta_2 \quad \text{en} \quad Var(T_1 - T_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

De nulhypothese is dat de parameters θ_1 en θ_2 gelijk zijn, dus

$$H_0 : \theta_1 = \theta_2 \quad \text{of} \quad \theta_1 - \theta_2 = 0.$$

Als we weer veronderstellen dat T_1 en T_2 bij benadering normaal verdeeld zijn dan is

$$Z := \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

(bij benadering) een standaard-normale verdeling en we kunnen weer de z-waarden gebruiken om een toets te formuleren:

Als de steekproef voor de eerste verdeling de schatting $\overline{\theta_1}$ en de steekproef voor de tweede verdeling de schatting $\overline{\theta_2}$ oplevert, dan wordt op significantie level α de nulhypothese $\theta_1 = \theta_2$ verworpen als

$$|\overline{\theta_1} - \overline{\theta_2}| > z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Voorbeeld: Stel de normaal verdeelde stochast X heeft variantie $\sigma_X^2 = 0.09$ en de normaal verdeelde stochast Y heeft variantie $\sigma_Y^2 = 0.16$. Een steekproef van 9 stuk geeft een gemiddelde van $\overline{x} = 21.7$ voor X en een steekproef van 4 stuk geeft een gemiddelde van $\overline{y} = 21.2$ voor Y. Kunnen we op een onbetrouwbaarheidslevel van $\alpha = 0.05$ de nulhypothese verwerpen dat X en Y hetzelfde gemiddelde hebben?

Er geldt $z_{0.025} = 1.96$ en $\sqrt{\frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}} = \sqrt{0.05}$, dus zullen we de nulhypothese inderdaad verwerpen omdat $|\overline{x} - \overline{y}| = 0.5 > 1.96 \cdot \sqrt{0.05} \approx 0.44$.

Ook éénzijdige toetsen spelen hier weer een belangrijke rol, bijvoorbeeld wil men aantonen dat een nieuwe medicijn beter is dan een oude. Als de parameter θ_1 de oude en de parameter θ_2 de nieuwe medicijn beschrijft, is de nulhypothese $H_0 : \theta_2 \leq \theta_1$ en men probeert met een rechtséénzijdige toets evidenties ervoor te vinden om deze hypothese te verwerpen, dus $\theta_2 > \theta_1$ te ondersteunen. Met dezelfde redeneringen die we eerder hebben toegepast, geeft dit op significantie level α het criterium

$$\overline{\theta_2} - \overline{\theta_1} > z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

om de nulhypothese te verwerpen. De schatting voor het verschil tussen de nieuwe en oude medicijn moet dus een zekere marge overschrijden om met hoge kans een toevallig effect uit te kunnen sluiten.

Aanpassingen bij kleine steekproeven

We zijn weer ervan uitgegaan dat de varianties σ_1^2 en σ_2^2 van de twee onderliggende verdelingen bekend zijn. Als dit niet het geval is, moeten we net als bij de toetsen voor een enkele verdeling de varianties door de geschatte steekproefvarianties s_1^2 en s_2^2 vervangen. Het probleem is, dat de verdeling van

$$T := \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

geen Student-t verdeling meer is en we dus niet zonder meer met de t-waarden kunnen werken. Maar gelukkig laat zich de verdeling van T wel door een Student-t verdeling benaderen, alleen moet men hiervoor nog een geschikt aantal ν van vrijheidsgraden bepalen.

Men kan inzien, dat het aantal vrijheidsgraden groter dan het minimum van $n_1 - 1$ en $n_2 - 1$ moet zijn, omdat dit de vrijheidsgraden voor de aparte stochasten T_1 en T_2 zijn. Aan de andere kant kan het aantal vrijheidsgraden ook niet groter dan $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$ zijn, want dit zou men bij samenvoegen van de twee steekproeven krijgen.

Als men aan de conservatieve kant zit en de nulhypothese niet te snel wil verwerpen, is $\nu := \min(n_1 - 1, n_2 - 1)$ een mogelijke keuze voor het aantal vrijheidsgraden. Maar meestal wordt het aantal vrijheidsgraden uit de grootten van de steekproeven en de steekproefvarianties berekend, bijvoorbeeld door

$$\nu := \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}\frac{s_1^2}{n_1} + \frac{1}{n_2-1}\frac{s_2^2}{n_2}}$$

De situatie is iets eenvoudiger en overzichtelijker als bekend is dat de twee verdelingen dezelfde (onbekende) variantie hebben. In dit geval noemt men het gewogen gemiddelde

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

van de steekproefvarianties de *gepoolde variantie* van de twee steekproeven.

Het idee achter de gepoolde variantie is, de twee steekproeven samen te vatten en uit de verzamelde waarden een schatting voor de variantie te maken. Stel X en Y zijn stochasten met dezelfde variantie σ^2. Voor een steekproef van grootte n_1 is $S_1^2 := \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2$ een zuivere schatter voor σ^2 en net zo is $S_2^2 := \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_i - \overline{Y})^2$ een zuivere schatter voor σ^2. Hieruit volgt, dat $(n_1-1)S_1^2 + (n_2-1)S_2^2$ een zuivere schatter voor $(n_1 + n_2 - 2)\sigma^2$ is, en dus is

$$S^2 := \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad \text{een zuivere schatter voor } \sigma^2.$$

De gepoolde variantie is dus juist de realisatie van deze zuivere schatter voor σ^2 op twee concrete steekproeven.

Het voordeel van de gepoolde variantie is, dat men hiermee weer naar een Student-t verdeling met een bekend aantal vrijheidsgraden komt, er geldt namelijk dat

$$T := \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = \frac{(T_1 - T_2) - (\theta_1 - \theta_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

een Student-t verdeling met $n_1 + n_2 - 2$ vrijheidsgraden is.

Een tweezijdige toets zou in deze situatie de nulhypothese $H_0 : \theta_1 = \theta_2$ verwerpen als

$$|\overline{\theta_1} - \overline{\theta_2}| > t_{n_1+n_2-2,\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

De vraag of de aanname dat twee steekproeven uit verdelingen met dezelfde variantie $\sigma_1^2 = \sigma_2^2 = \sigma^2$ komen juist is, kan zijnerzijds ook weer met een toets onderzocht worden. Hiervoor kijkt men naar het quotiënt $\frac{\sigma_1^2}{\sigma_2^2}$, waarvoor $\frac{s_1^2}{s_2^2}$ een schatting is en de verdeling van deze schattingen heet de F-verdeling. De nulhypothese is $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ en de zogeheten F-toets geeft aan, wanneer H_0 op een zekere onbetrouwbaarheidslevel moet worden verworpen. In dit college gaan we de F-toets echter alleen maar in verband met de variantie-analyse behandelen.

Verschillen tussen relatieve frequenties

De ideeën die we net hebben bediscussieerd, kunnen we ook toepassen op de vraag of twee relatieve frequenties significant verschillen. Als P_1 een zuivere schatter voor de relatieve frequentie p_1 is en P_2 een zuivere schatter voor de relatieve frequentie p_2, dan is $P_1 - P_2$ een schatter met verwachtingswaarde $E[P_1 - P_2] = p_1 - p_2$ en met variantie $Var(P_1 - P_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$, waarbij n_1 en n_2 de grootten van de steekproeven zijn.

Als we willen laten zien, dat de twee relatieve frequenties verschillend zijn, is de nulhypothese natuurlijk dat p_1 en p_2 gelijk zijn, dus

$$H_0 : p_1 = p_2.$$

Onder de aanname dat de nulhypothese juist is, is dus

$$Var(P_1 - P_2) = p_1(1-p_1)(\frac{1}{n_1} + \frac{1}{n_2}) = p_2(1-p_2)(\frac{1}{n_1} + \frac{1}{n_2}).$$

Omdat we niet ervan kunnen uitgaan dat p_1 of p_2 bekend is, moeten we hier weer een schatting invullen, en hiervoor nemen we de schatting p_0 die we uit de combinatie van de twee steekproeven krijgen, dus

$$p_0 := \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}.$$

Als de steekproeven niet te klein zijn (dus weer $n_1 p_1 \geq 5$ en $n_2 p_2 \geq 5$, d.w.z. in ieder steekproef hebben we minsten 5 successen) is de stochast

$$Z := \frac{P_1 - P_2}{\sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

bij benadering standaard-normaal verdeeld en we kunnen hiermee weer met behulp van de z-waarden tweezijdige en éénzijdige toetsen formuleren.

Als we de schattingen $\overline{p_1}$ en $\overline{p_2}$ voor de relatieve frequenties in de twee steekproeven vinden, zullen we bij een tweezijdige toets de nulhypothese H_0 : $p_1 = p_2$ verwerpen als

$$|\overline{p_1} - \overline{p_2}| > z_{\frac{\alpha}{2}} \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}.$$

Bij een éénzijdige toets krijgen we analoog, dat we de nulhypothese moeten verwerpen als

$$\overline{p_2} - \overline{p_1} > z_\alpha \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})} \quad \text{of} \quad \overline{p_1} - \overline{p_2} < z_\alpha \sqrt{p_0(1-p_0)(\frac{1}{n_1} + \frac{1}{n_2})}$$

afhankelijk ervan of we willen aantonen dat p_2 groter of kleiner is dan p_1.

Belangrijke begrippen in deze les

- nulhypothese, alternatieve hypothese
- toets (tweezijdig, éénzijdig)
- onbetrouwbaarheid van een toets
- onderscheidingsvermogen van een toets
- type I fout, type II fout
- significantie
- P-waarde
- aanpassingen bij kleine steekproeven
- gepoolde variantie

Opgaven

21. Men past op elk van twee (aselecte, onafhankelijke) steekproeven een toets met onbetrouwbaarheid α toe. Hoe groot moet α worden gekozen zo dat de kans dat minstens één van de nulhypothesen ten onrechte wordt verworpen hoogstens 10% is?

22. Het gewicht van sinaasappels was tot nu toe normaal verdeeld met gemiddelde $\mu_0 = 50g$ en standaardafwijking $\sigma = 2g$. Van een nieuwe goedkopere behandeling van de sinaasappelbomen wordt beweerd dat ze minstens even zware vruchten oplevert. Een kweker wil deze bewering toetsen tegen het alternatief dat $\mu < 50g$ (waarbij de standaardafwijking onveranderd blijft). De sinaasappels in een steekproef van 100 stuks hebben een gemiddeld gewicht van $49.65g$. Heeft de kweker op een onbetrouwbaarheidslevel van $\alpha = 0.05$ reden om de nieuwe methode niet toe te passen?

23. Zij X een normaal verdeelde stochast met standaardafwijking $\sigma = 10$ en onbekende gemiddelde μ. Op grond van een steekproef willen we de hypothese $H_0 : \mu = 50$ rechtséénzijdig toetsen met onbetrouwbaarheid $\alpha = 0.05$. We eisen daarbij dat het onderscheidingsvermogen bij de alternatieve hypothese $H_1 : \mu = 52$ gelijk aan 90% moet zijn.

    (i) Hoe groot moet de steekproef minstens zijn?

    (ii) Hoe groot is bij de steekproefgrootte uit (i) het onderscheidingsvermogen bij de alternatieve hypothese $\mu = 51$?

24. In een fabriek staan 2 vulmachines, A en B, waarmee flessen worden gevuld. Bij een juiste instelling van de machines is de inhoud van de flessen normaal verdeeld met een gemiddelde van $250g$. De standaardafwijking is onafhankelijk van de instelling steeds $2.5g$. Om na te gaan of de machines goed zijn ingesteld wordt voor elke machine de inhoud van 4 net gevulde flessen nauwkeurig bepaald. De gemiddelde inhoud voor flessen van machine A bedraagt $251.68g$, terwijl hij $252.68g$ voor flessen van machine B is.

    (i) Toets met onbetrouwbaarheid $\alpha = 0.05$ of de machines A en B op het juiste vulgewicht van $250g$ ingesteld zijn.

    (ii) Toets met onbetrouwbaarheid $\alpha = 0.05$ of de instellingen van de machines A en B onderling verschillen.

25. Een examen bestaat uit 20 vragen met telkens 4 mogelijke antwoorden. De kandidaten zijn geslaagd als op minstens 10 vragen het juist antwoord is gekozen. Beschouw het tentamen als een statistische toets.

    (i) Formuleer een nulhypothese H_0 en een alternatieve hypothese H_1.

    (ii) Definieer de grootheid die voor de toets uit de steekproef bepaald wordt en bepaal de kansverdeling van deze grootheid onder de aanname van H_0.

    (iii) Bereken de onbetrouwbaarheid van de toets.

    (iv) Bereken het onderscheidingsvermogen van de toets als de kans op het geven van het juiste antwoord door een kandidaat per vraag $\frac{1}{2}$ is.

26. Uit een baal katoen werd een aselecte steekproef genomen van 4000 draden om de vezellengte te bepalen. De gemiddelde lengte was $2.33cm$ en de standaardafwijking $0.48cm$. Uit dezelfde baal werd een andere steekproef genomen van 200 draden volgens een andere methode dan de eerste. Van deze tweede steekproef was de gemiddelde vezellengte $2.54cm$. Aangenomen mag worden dat de vezellengte normaal verdeeld is. Toets met onbetrouwbaarheid $\alpha = 0.05$ of er verschil is tussen de twee steekproefmethoden.

27. Een fabrikant betrekt al jaren transistoren van A, die hem gemiddeld 8% kapotte levert. Van een vertegenwoordiger van B koopt hij 75 stuks die wat duurder zijn, maar waarvan beweerd wordt dat er minder kapot zijn. Bij controle blijken 5 van deze 75 transistoren ondeugdelijk te zijn. Zijn de percentages kapotte exemplaren in de producten van A en B op een significantie level van $\alpha = 0.05$ verschillend?

28. Een medicus beweert dat de kans op een jongengeboorte groter is dan die op de geboorte van een meisje. Hij komt tot deze conclusie omdat 51% van de pasgeboren baby's uit zijn praktijk jongens zijn. Hoeveel geboorten moeten dat zijn om deze conclusie of een onbetrouwbaarheidslevel van $\alpha = 0.05$ te rechtvaardigen?