To P or not to P?

Zowel in wetenschappelijke artikelen als in de actuariële praktijk wordt veelvuldig gebruik gemaakt van nulhypotheses en p-waarden. Toch verschijnen er al jarenlang vanuit de statistische wetenschap publicaties die waarschuwen voor verkeerd gebruik en verkeerde interpretatie van p-waarden. In dit artikel gaat Richard Plat in op deze ogenschijnlijke tegenstelling, de bezwaren vanuit de statistische wetenschap, en de mogelijke alternatieven.

GEBRUIK P-WAARDEN

In de actuariële praktijk wordt veelvuldig gebruik gemaakt van nulhypotheses en bijbehorende p-waarden. Een veel voorkomende situatie waarbij hiervan gebruik van wordt gemaakt is bij het selecteren van verklarende variabelen bij regressie. Een voorbeeld is lineaire regressie:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \varepsilon_i$$

waarbij x_i de verklarende variabelen zijn (met $i = 1,\ldots,N$), y_i de afhankelijke variabele, ε_i de storingsterm en β_K de te schatten coëfficiënten (met $k = 1,\ldots,K$). Bij het selecteren van de verklarende variabelen komt het vaak voor dat naar een grote selectie potentiële variabelen gekeken wordt, en dat vervolgens per variabele de volgende nulhypothese H_0 wordt getest voor de bijbehorende coëfficiënten:

$$H_0 : \beta_k = 0$$

Als deze nulhypothese waar is, dan heeft de statistiek $t_k = b_k / \sigma(b_k)$ een t verdeling met $N - K$ vrijheidsgraden, waarbij b_k de schatter is van coëfficiënt β_k en $\sigma(b_k)$ de standaardfout van die schatter. De p-waarde p_k (voor een tweezijdige test) wordt als volgt bepaald:

$$p_k = 2 \times P_{t_{N-K}}\{ t \geq | t_k | \}$$

Als de p-waarde onder een bepaald significantieniveau α (meestal gelijk gesteld aan 5%) ligt, dan wordt de coëfficiënt als statistisch significant aangemerkt. Het lijkt dan voor de hand te liggen om alleen de statistisch significante variabelen op te nemen in het uiteindelijk regressiemodel.

Dr. R. Plat AAG RBA is partner van Risk at Work.

Bovenstaande werkwijze wordt veel gehanteerd in zowel wetenschappelijke literatuur als in de actuariële praktijk. Hiervoor zijn verschillende redenen:
- De p-waarde is een relatief eenvoudig te communiceren maatstaf;
- Onderbouwing van een model op basis van p-waarden is een transparant en te automatiseren proces, en behoeft weinig 'expert judgment';
- Als gevolg daarvan is het eenvoudiger om het model goedgekeurd te krijgen door reviewers (bijvoorbeeld model validatie, auditors, accountants, DNB).

BEZWAREN VANUIT STATISTISCHE WETENSCHAP

Ondanks het veelvuldig gebruik van de hierboven beschreven methode, is er in het vorige decennium een reeks wetenschappelijke artikelen gepubliceerd met bezwaren over het gebruik en de interpretatie van p-waarden, zie bijvoorbeeld Cumming (2014), Nuzzo (2014) en Greenland et al (2016). Dit heeft de American Statistical Association (ASA) ertoe bewogen om voor het eerst in hun bestaan een standpunt te publiceren over een specifieke statistische praktijk, namelijk het gebruik van p-waarden en de notie van statistische significantie. Dit is vastgelegd in 6 principes in Wasserstein & Lazar (2016):

1 **P-waarden geven een indicatie hoe incompatibel de data is met een gespecificeerd statistisch model.** Eventuele incompatibiliteit (lage p-waarde) kan worden geïnterpreteerd als bewijs tegen de nulhypothese.

2 **P-waarden representeren niet de kans dat de nulhypothese waar is, of de kans dat de data alleen een gevolg zijn van willekeurig toeval.** Het representeert hoe de data zich verhoudt tot een specifieke hypothese, $P\{data|H_0\}$, niet de kans of de hypothese waar is gegeven de data, $P\{H_0|data\}$.

3 **Wetenschappelijke conclusies en business- of beleidsbeslissingen zouden niet gebaseerd moeten zijn op of een p-waarde hoger of lager is dan een bepaalde grens.** Hoewel praktische overwegingen veelal binaire beslissingen vereisen, zouden meerdere factoren in ogenschouw genomen moeten worden, zoals het ontwerp van de analyse, de datakwaliteit, eventueel beschikbaar extern bewijs, en de geldigheid van assumpties zoals gebruikt in de data-analyse.

4 **Een juiste statistische analyse vereist volledige rapportage en transparantie.** Inclusief het aantal hypotheses die beschouwd zijn, alle beslissingen omtrent gebruikte data, alle uitgevoerde statistische analyses, en alle berekende p-waarden.

5 **Een p-waarde, of statistische significantie, meet niet de grootte van een effect of de importantie van een resultaat.** Ieder effect, hoe klein ook, kan een lage p-waarde opleveren als het aantal waarnemingen hoog genoeg is of de standaardfout laag genoeg.

6 **Op zichzelf is de p-waarde geen goede maatstaf voor bewijs tegen of voor een model of hypothese.** Een p-waarde zonder context of ander bewijs geeft beperkte informatie. Een hoge p-waarde betekent bijvoorbeeld niet noodzakelijk dat de nulhypothese waar is: er zouden vele andere nulhypothesen even of meer consistent kunnen zijn met de data.

Samenvattend stelt de ASA dat de p-waarde weliswaar nog steeds een rol kan spelen in statistische analyses (volgens principe 1), maar als onderdeel van een bredere analyse en zeker niet als doorslaggevende factor (in combinatie met een bepaalde grens). Meer detail en analyses zijn gegeven in Greenland et al (2016).

P-WAARDE: EEN STOCHASTISCHE VARIABELE

Naast de genoemde bezwaren, is het ook van belang zich te realiseren dat de p-waarde ook een stochastische maatstaf is. Afhankelijk van de specifieke trekkingen uit de daadwerkelijke verdeling kan de p-waarde relatief hoog of laag uitvallen. Om dit te illustreren is een voorbeeld uitgewerkt. Stel er zijn twee variabelen A_t en B_t. Variabele A_t is standaardnormaal verdeeld en voor variabele B_t geldt het volgende proces:

$$B_t = 0.4 \times A_t + \varepsilon_t$$

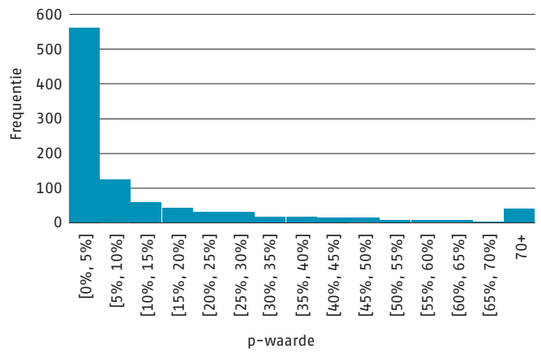
waarbij ε_t eveneens standaardnormaal verdeeld is. Met andere woorden, de relatie tussen A_t en B_t is bekend en de betreffende coëfficiënt van 0.4 is duidelijk verschillend van 0. Vervolgens zijn de volgende stappen doorlopen:

1 Er is een trekking gedaan van 30 observaties van A_t en B_t.

2 Gegeven de trekking in stap 1), is een lineaire regressie ($B_t = \beta \times A_t + \varepsilon_t$) uitgevoerd en is de p-waarde bepaald van de coëfficiënt β.

3 Stap 1) en 2) zijn vervolgens 1.000 maal herhaald.

De verdeling van de resulterende 1.000 p-waarden is gegeven in onderstaande figuur.



Figuur 1: verdeling p-waarden van coëfficiënt B_t

De figuur laat zien dat in slechts ~55% van de scenario's de p-waarde < 5% is. De overige ~45% laat een grote variabiliteit zien, met zelfs p-waardes boven de 70%. Dit voorbeeld toont aan dat de onzekerheid van de gemeten p-waarde hoog kan zijn (afhankelijk van onder andere het aantal waarnemingen).

ALTERNATIEVEN

Eén van de redenen voor het veelvuldig gebruik van p-waarden is omdat er in de literatuur nog geen consensus is over wat het beste alternatief is. In de meeste in de literatuur genoemde alternatieven worden schattingsresultaten en/of het effect van de verklarende variabelen geprefereerd boven hypothese testen. Halsey (2019) noemt vier alternatieven / aanvullingen:

1 Vul analyses op basis van p-waarden aan met betrouwbaarheids-intervallen van die p-waarden, of de kans dat een statistisch significante hypothese een vals positieve observatie is.

2 Focus op het effect van de variabele (op basis van de geschatte coëfficiënt) in combinatie met de betrouwbaarheid daarvan. Hoewel deze begrippen op hetzelfde onderliggende statistische raamwerk gebaseerd zijn, is een focus op het effect van een variabele wel fundamenteel anders dan of een nulhypothese al dan niet verworpen wordt.

3 Gebruik de Bayes factor: deze representeert het relatieve bewijs voor de nulhypothese versus de alternatieve hypothese. Een Bayes factor van bijvoorbeeld 5 geeft aan dat het bewijs voor de alternatieve hypothese 5 maal hoger is dan voor de nulhypothese.

4 Laat maatstaven voor de fit van mogelijke modellen leidend zijn, bijvoorbeeld het Akaike Information Criterion (AIC) of het Bayesian Information Criterion (BIC). Beiden zijn maatstaven waarbij een penalty op het aantal parameters wordt toegepast, waardoor een balans gevonden wordt tussen fit kwaliteit en de complexiteit van het model.

De maatstaven genoemd onder 2) en 4) zijn over het algemeen al onderdeel van de standaard output van statistische pakketten, dus kunnen eenvoudig toegepast worden.

CONCLUSIE

In dit artikel zijn bezwaren vanuit de statistische wetenschap beschreven met betrekking tot verkeerd gebruik en verkeerde interpretatie van p-waarden bij hypothese testen. Ook is middels een voorbeeld aangetoond hoe hoog de onzekerheid van een gemeten p-waarde kan zijn. Alternatieven zijn beschikbaar die eenvoudig toegepast kunnen worden. ■

Referenties

Cumming (2014), The New Statistics: Why and How

Greenland et al (2016), Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations

Halsey (2019), The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?

Nuzzo (2014), Scientific Method: Statistical Errors

Wasserstein & Lazar (2016), The ASA Statement on p-Values: Context, Process, and Purpose,