STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Verklarende statistiek

9. Toetsen van hypothesen

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Toetsen van hypothesen

Deze tekst gaat ervan uit dat je goed weet hoe kansmodellen werken en dat je bovendien de teksten over betrouwbaarheidsintervallen (voor proporties en gemiddelden) grondig hebt bestudeerd.

DEEL 1. Basisideeën

1. Hoe extreem mag je zijn?

Een belangrijk begrip bij het toetsen van hypothesen gaat over uitkomsten "die je niet verwacht" of "die weinig kans hebben om op te treden" of "die extreem groot of klein zijn". Dit begrip gaan we even vooraf bekijken.

1.1. Zeldzame gebeurtenissen

Veronderstel dat ik je zeg dat ik een nieuw computerprogramma geschreven heb dat lukraak nullen en enen genereert. Ik stel voor dat we met mijn programma een spel spelen. Bij een één verlies ik en geef ik jou een euro. Als het een nul is dan verlies jij en geef jij mij een euro.
De eerste keer is het een nul en dus geef jij mij een euro. De tweede keer is het ook een nul en krijg ik van jou een euro. De derde keer is het terug een nul en weer moet jij mij een euro geven. De vierde keer verschijnt er terug een nul zodat jij weer een euro moet betalen. We spelen verder en ook de vijfde keer is het een nul. Deze keer wil je niet meer betalen. Je hebt een sterk vermoeden dat mijn programma niet eerlijk is.

Soortgelijke proeven werden bij heel wat mensen gedaan. Na 4 of 5 mislukkingen op een rij hadden de meesten een sterk vermoeden dat er iets niet klopte. Dat was een reactie gewoon op het gevoel. De kans dat zoiets gebeurt kan je eenvoudig uitrekenen. Als je een spel speelt dat eerlijk is dan moet de kans op een één even groot zijn als de kans op een nul. Beide uitkomsten moeten optreden met kans 1/2. Bij de eerste trekking heb je kans 1/2 om een nul te hebben. Voor het vervolg maak je gebruik van onafhankelijke gebeurtenissen (de uitkomst van een volgende trekking heeft niets te maken met wat de vorige trekkingen opleverden) zodat (productregel)

$$P(\textit{eerste} = 0 \textbf{ en } \textit{tweede} = 0) \;=\; P(\textit{eerste} = 0) \cdot P(\textit{tweede} = 0) \;=\; \frac{1}{2} \cdot \frac{1}{2} \;=\; \frac{1}{4} \;=\; 25\,\%.$$

Op analoge manier is de kans van 3 mislukkingen op een rij gelijk aan $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8} = 0.125 = 12.5\,\%$.
Voor 4 mislukkingen op een rij is de kans 6.25 % en 5 mislukkingen op een rij gebeurt met kans 3.125 %. De meeste mensen beschouwen een gebeurtenis als "onverwacht" wanneer ze een kleine kans heeft om op te treden. Dat onverwachte kadert natuurlijk binnen de context van het verhaal. Hier gelooft men dat het spel "eerlijk" is. Binnen die context is een start met 4 à 5 mislukkingen op een rij "uitzonderlijk". De kans dat zoiets gebeurt ligt in de buurt van 5 %.

In de statistiek gebruikt men al lang een 5 % criterium om over "onverwachte" gebeurtenissen te spreken. Je kan dat ook bekijken "in the long run". Als je tegen 100 mensen dit spel speelt en als het een eerlijk spel is, dan gebeurt het slechts bij (ongeveer) 5 mensen dat ze starten met 4 à 5 mislukkingen op een rij terwijl dat bij die andere 95 mensen niet zo is (en zij binnen de eerste 4 à 5 pogingen toch al minstens één keer – of meerdere keren – winnen en een euro krijgen).

> Bij het toetsen van hypothesen gebruikt men in de meeste studies het 5 % criterium. Een uitkomst is "onverwacht", "uitzonderlijk", "zeldzaam", "extreem", … als de kans hoogstens 5 % is dat je, binnen de context van de studie, zo'n uitkomst ziet.

1.2. Waarden of gebieden?

Een bedrijf gebruikt een grote ingewikkelde machine in haar productieproces. Die machine wordt elke maand gecontroleerd. Men meet dan gedurende meerdere dagen en op verschillende tijdstippen de kwaliteit van de output. Dat gebeurt volgens een vastgelegd schema waarbij al die meetresultaten uiteindelijk tot één eindresultaat worden samengebracht. Men weet dat die manier van meten een hoge eindscore oplevert als de machine dringend toe is aan een onderhoud. Als dit dan niet gebeurt gaat ze stuk.

Volgens de constructeur levert een goed werkende machine niet altijd dezelfde eindscore. De mogelijke eindscores en de kansen dat je die scores opmeet bij een goed werkende machine zien er als volgt uit:

Mogelijke eindscore	≤ 20	24	25	30	36
Kans op die score bij een goed werkende machine	$\frac{30}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Een machine die dringend aan onderhoud toe is levert meestal een hoge score (zoals 25, 30 of 36).

Om de machine een onderhoudsbeurt te geven moet je het productieproces een dag stilleggen. Zoiets doe je liever niet als de machine nog goed werkt en eigenlijk geen onderhoud nodig heeft. Maar als de machine stuk gaat dan duurt het weken vooraleer ze kan hersteld worden. Dat zou pas een echte ramp zijn voor het bedrijf. Daarom ben je zeer achterdochtig van zodra je een hoge eindscore ziet.

Als je nu de klassieke 5 % grens gebruikt om over “een extreme gebeurtenis” te spreken, welke score moet je dan vinden om te beslissen dat de machine een onderhoudsbeurt nodig heeft? Je weet dat je moet kijken naar hoge scores.

Als de machine nog goed werkt dan is:

- de kans op een score van 36 gelijk aan $\frac{1}{36} \cong 0.028 = 2.8\,\%$
- de kans op een score van 30 gelijk aan $\frac{2}{36} \cong 0.056 = 5.6\,\%$
- de kans op een score van 25 gelijk aan $\frac{1}{36} \cong 0.028 = 2.8\,\%$

Als je een score van 36 vindt, dan beslis je dat de machine een onderhoud nodig heeft want een goed werkende machine levert de score 36 slechts met kans 2.8 % (en dat is minder dan de 5 % grens zodat je hier kan spreken van een “onverwacht groot” resultaat).

Als je een score van 25 vindt, dan moet je, om consequent te zijn, ook beslissen dat de machine een onderhoud nodig heeft, want een score van 25 krijg je met kans 2.8 % en dat is minder dan 5 %.

Maar nu komt de tegenspraak. Als jij moet beslissen of de machine een onderhoud nodig heeft en je zegt “ja” bij een score van 25 omdat je die score te groot vindt, wat zeg je dan als je een score van 30 ziet? Dat zou toch nog meer moeten wijzen op “dringend onderhoud”! Maar een score van 30 krijg je met een goede machine met kans 5.6 %, wat meer dan 5 % is.

Besluit.

Als een grote uitkomst aanleiding geeft tot de beslissing dat je met “een onverwachte gebeurtenis” te maken hebt, dan moet je bij een nog grotere uitkomst zeker ook beslissen dat de gebeurtenis “onverwacht” is. Maar meer extreme uitkomsten gaan niet noodzakelijk samen met kleinere kansen. Kijk maar naar de waarde 25 die, op basis van haar kans, meer extreem zou zijn dan de waarde 30.

Die moeilijkheid heb je niet bij gebieden. Als je bijvoorbeeld kijkt naar rechterstaarten die meer en meer extreem worden dan gebeurt het nooit dat hun kansen vergroten. Zo heb je:

- de kans op minstens 25 = $P(X \geq 25) = P(X = 25) + P(X = 30) + P(X = 36) = \frac{1}{36} + \frac{2}{36} + \frac{1}{36} = \frac{4}{36}$

- de kans op minstens 30 = $P(X \geq 30) = P(X = 30) + P(X = 36) = \frac{2}{36} + \frac{1}{36} = \frac{3}{36}$

- de kans op minstens 36 = $P(X \geq 36) = P(X = 36) = \frac{1}{36}$

Bij het toetsen van hypothesen kijk je niet naar extreme uitkomsten (extreme getalwaarden) maar naar extreme gebieden. Die gebieden zijn rechterstaarten (waarin extreem grote uitkomsten terechtkomen) of linkerstaarten (waarin extreem kleine uitkomsten terechtkomen). De kansen die je berekent zijn altijd kansen om in gebieden (in staarten) terecht te komen.

<u>Nota</u> (facultatief).

Als je een concreet voorbeeld wil hebben van een kansmodel zoals hierboven beschreven voor de score van een goed werkende machine, dan kan je kijken naar het product van twee eerlijke dobbelstenen. Die kansverdeling kan je eenvoudig zelf opstellen.

In de volgende tabel zie je alle producten staan. Als de dobbelstenen eerlijk zijn dan heeft elke cel een kans van $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

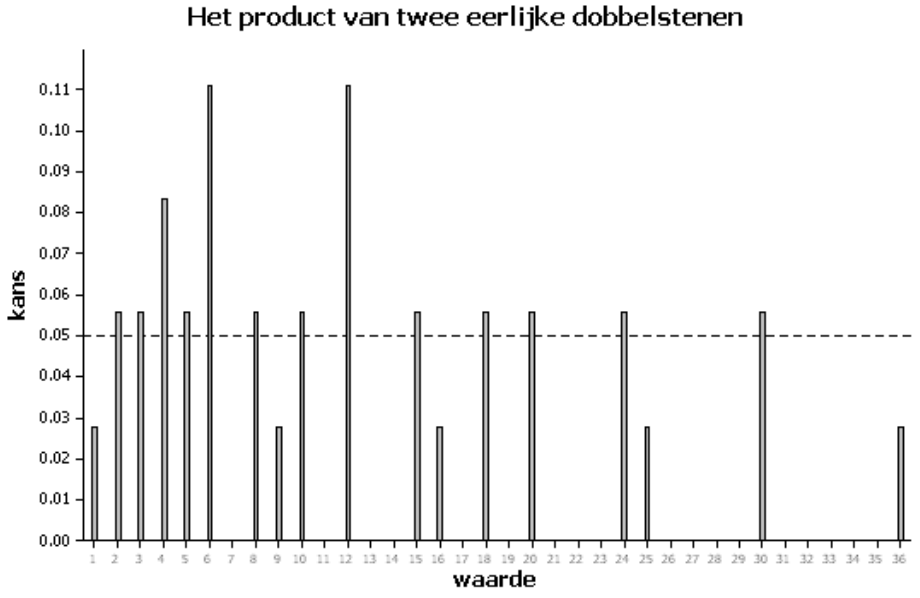
	1	2	3	4	5	6
	2	4	6	8	10	12
	3	6	9	12	15	18
	4	8	12	16	20	24
	5	10	15	20	25	30
	6	12	18	24	30	36

In de notatie van kansmodellen, met X = het product van 2 eerlijke dobbelstenen, heb je hier:

x	1	2	3	4	5	6	8	9	10
$P(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$

x	12	15	16	18	20	24	25	30	36
$P(X=x)$	$\frac{4}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Je kan deze kansverdeling ook voorstellen door een staafdiagram. Zo zie je grafisch alle mogelijke waarden samen met hun kansen. Je ziet ook dat meer extreme (grotere) waarden niet noodzakelijk samengaan met kleinere kansen.



2. Hypothesen

2.1. Een hypothese is…

Een hypothese is **een bewering over een populatie**.
Meestal gaat zo'n bewering over een *populatieparameter*.

Bij een 0–1 populatie (zoals het geslacht van baby's, met 0 = meisje en 1 = jongen) kan je een bewering formuleren over de proportie jongens die in 1995 in Vlaanderen werden geboren. Je kan er bijvoorbeeld van overtuigd zijn dat er evenveel meisjes als jongens geboren worden. In dat geval zeg je dat volgens jou $\pi = \frac{1}{2}$. Hierbij is π de notatie voor **de populatieproportie** die in dit voorbeeld de proportie jongens is bij alle baby's die in 1995 in Vlaanderen werden geboren.

Bij een continue populatie (zoals de lengte van Vlaamse meisjes van 17) formuleer je een bewering over het gemiddelde. Jij hebt ergens gelezen dat die meisjes gemiddeld 166 cm groot zijn. Zelf denk je dat meisjes van 17 gemiddeld groter zijn. Jouw opinie kan je formuleren als $\mu > 166$. Hierbij is μ de notatie voor **het populatiegemiddelde** dat in dit voorbeeld de gemiddelde lengte is van alle 17-jarige Vlaamse meisjes.

Een hypothese is een bewering over een populatie.
Een hypothese formuleer je vooraleer je de steekproef trekt.

2.2. Nulhypothese en alternatieve hypothese

Een **nulhypothese** is een bewering dat een populatieparameter **een welbepaalde waarde** aanneemt. De notatie voor een nulhypothese is een hoofdletter H met index nul: H_0.

In de meeste gevallen reflecteert de nulhypothese "de gangbare mening" of "de klassieke standaard". Als men vroeger dacht dat er evenveel jongens als meisjes worden geboren dan was dat "de gangbare mening" die men kon opschrijven als: $H_0 : \pi = \frac{1}{2}$ met π de proportie jongens bij alle baby's die geboren worden.
Bij studies naar bepaalde effecten (het geneesmiddel doet patiënten langer leven, vrouwen worden gediscrimineerd,…) stelt men als nulhypothese "dat er geen effect is" (het geneesmiddel helpt niet, er is geen discriminatie,…).

Een **alternatieve hypothese** is een bewering over dezelfde populatieparameter waarbij je iets anders zegt dan wat er in de nulhypothese staat. De notatie voor een alternatieve hypothese is een hoofdletter H met index één: H_1. In de meeste studies fixeert men zich hier niet op een welbepaalde waarde maar zegt de alternatieve hypothese dat de populatieparameter **in een bepaald gebied** ligt.

Als je leest dat Vlaamse meisjes van 17 gemiddeld 166 cm groot zijn dan is dat "de huidige standaard". Maar jij hoeft dat niet te geloven en je kan er vast van overtuigd zijn dat ze groter zijn. In deze context is de nulhypothese $H_0 : \mu = 166$ en de alternatieve hypothese is $H_1 : \mu > 166$. Als je integendeel denkt dat die meisjes gemiddeld kleiner zijn dan is je alternatieve hypothese $H_1 : \mu < 166$. Je kan ook zeggen dat 166 cm ongeloofwaardig overkomt maar dat je zelf niet weet of ze nu groter of kleiner zijn. In zo'n situatie werk je met $H_1 : \mu \neq 166$. In de drie gevallen heb je als alternatieve hypothese een gebied aangegeven waarin jij denkt dat de populatieparameter μ ligt.

Een hypothese van de vorm $H_1 : \mu > 166$ of $H_1 : \mu < 166$ noemt men een **éénzijdige** alternatieve hypothese. Zij wijst in één richting naar alleen maar grotere waarden of naar alleen maar kleinere waarden.
Een hypothese van de vorm $H_1 : \mu \neq 166$ noemt men een **tweezijdige** alternatieve hypothese. Zij wijst gelijktijdig naar waarden die groter of kleiner zijn.

Opdracht 1

In 2000 haalden de leerlingen van de derde graad ASO in Vlaanderen een gemiddelde score van 57 op 100 op het eindexamen wiskunde. Jij twijfelt eraan of dat vorig schooljaar ook het geval was. Dat wil je nu, met behulp van toetsen van hypothesen, onderzoeken.

1. Wat is hier de populatie en welke populatieparameter (met juiste notatie) bestudeer je?
2. Formuleer de nulhypothese (in woorden en met de juiste notatie).
3. Formuleer de alternatieve hypothese (in woorden en met de juiste notatie). Is zij één– of tweezijdig?

Opdracht 2

Astrologen geloven dat de positie van de maan en de planeten op het ogenblik van je geboorte je persoonlijkheidskenmerken bepalen. Zij gebruiken tijdstip en plaats van geboorte om een horoscoop en een astrologisch profiel op te stellen. Maar werkt dat echt? Om dat te onderzoeken werd in de Verenigde Staten aan astrologen de volgende opdracht gegeven. Van 3 personen werd een uitgebreid rapport over hun persoonlijkheidskenmerken gemaakt. Dat gebeurde op een gestandaardiseerde manier, op basis van wetenschappelijk geteste vragenlijsten. Dan werd aan de astroloog tijdstip en plaats van de geboorte van één van die drie personen gegeven met de vraag om daarbij de juiste persoon te identificeren (op basis van de 3 rapporten met persoonlijkheidskenmerken).

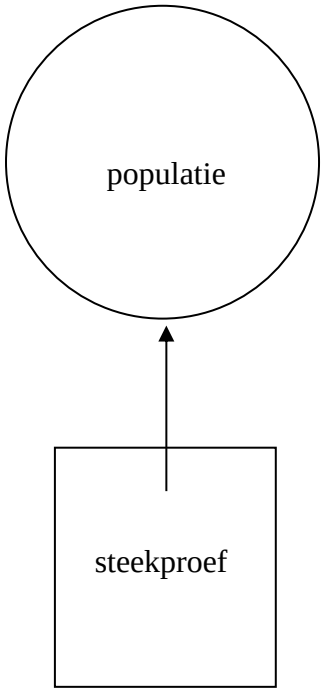
1. Wat is hier de populatie en welke populatieparameter (met juiste notatie) bestudeert men?
2. Formuleer de nulhypothese en gebruik de juiste notatie.
3. Formuleer de alternatieve hypothese en gebruik de juiste notatie. Is deze hypothese één– of tweezijdig?

Opdracht 3

In een advertentie lees je dat je een dieet met weinig koolhydraten moet volgen als je gewicht wil verliezen. Eet weinig of geen brood, pasta, zoetigheid… en trek je voor de rest niets aan van het aantal calorieën op je bord. Helpt dat echt? Je wil dit onderzoeken door, over een periode van 3 maanden, het gewichtsverlies te noteren van personen die zo'n dieet volgen.

1. Wat is hier de populatie en welke populatieparameter (met juiste notatie) bestudeer je?
2. Formuleer de nulhypothese en gebruik de juiste notatie.
3. Formuleer de alternatieve hypothese en gebruik de juiste notatie.

3. Toetsen



In de verklarende statistiek doe je, op basis van steekproefresultaten, een uitspraak over een populatieparameter.

Bij betrouwbaarheidsintervallen gebruik je een steekproef om een betrouwbaarheidsinterval op te stellen voor een ongekende populatieparameter. Bij toetsen van hypothesen worden beweringen gemaakt over een populatieparameter. Nu gebruik je steekproefresultaten om die beweringen te beoordelen.

3.1. Wat je (niet) bewijst

Bij toetsen van hypothesen start je altijd met de nulhypothese H_0. Dat wil zeggen dat jij gelooft dat H_0 waar is. Zo heb je te maken met een populatieparameter die je kent en die de waarde heeft die in H_0 wordt vermeld. Dat blijf je nu geloven, tenzij je in je steekproef resultaten vindt die de bewering van de nulhypothese **zeer sterk ongeloofwaardig** maken. Pas dan mag je H_0 verwerpen en overstappen op de alternatieve hypothese H_1.

Ann heeft in een doos 500 kaartjes gelegd die allemaal identiek zijn behalve de kleur (wit en zwart). Met die doos wil zij spelen tegen Els die blindelings een kaartje uit de doos moet trekken. Als het kaartje wit is wint Ann.
Els denkt dat Ann veel meer witte dan zwarte kaartjes in die doos heeft gestopt maar Ann zegt dat het er van elke kleur evenveel zijn. Toch vertrouwt Els het zaakje niet. Zij wil vooraf een controle doen door eerst 20 keer (telkens met terugleggen) een kaartje uit die doos te trekken. Op basis van die uitkomsten zal zij beslissen of zij met die doos wil spelen.

Als je dit even naar een algemeen kader vertaalt dan heb je hier te maken met een 0–1 populatie. Vanuit het standpunt van Els is 0 = wit (mislukking) en 1 = zwart (succes). Els denkt dat haar succesproportie (= de proportie zwarte kaartjes in de doos = de populatieproportie) kleiner dan 1/2 is maar Ann beweert dat het een eerlijke doos is. Als nulhypothese moet Els starten met Ann te geloven zodat $H_0 : \pi = 0.5$. Maar zelf denkt zij dat ze benadeeld is zodat haar alternatieve hypothese eruit ziet als $H_1 : \pi < 0.5$. En nu trekt Els haar steekproef van grootte $n = 20$.

De verdere redenering doe je in het kader van de nulhypothese.

- Als H_0 waar is, is het dan uitzonderlijk dat Els hoogstens 9 zwarte kaartjes vindt? De kans dat er tussen die 20 kaartjes hoogstens 9 zwarte zitten is 41 %. Als je dus 100 keer 20 kaartjes trekt dan zal het ongeveer 41 keer gebeuren dat daar niet meer dan 9 zwarte bij zijn. Zo'n uitkomst is helemaal niet "uitzonderlijk" of "onverwacht" wanneer je met een eerlijke doos te maken hebt.
- Als H_0 waar is, is het dan uitzonderlijk dat Els hoogstens 8 zwarte kaartjes vindt? De kans dat je zoiets vindt is 25 %. Dat is niet uitzonderlijk.
- Je kan zo verder gaan om uiteindelijk te komen tot de kans om uit een eerlijke doos 20 kaartjes te trekken die allemaal wit zijn. De kans dat zoiets gebeurt is $\left(\frac{1}{2}\right)^{20} = \frac{1}{1\,048576} \cong 0.000001$. Als Els dat zou tegenkomen en als ze dan nog zou geloven dat Ann haar een eerlijke doos heeft voorgeschoteld, dan is ze toch wel op haar hoofd gevallen.

Waar het hier om gaat, is dat je, vanaf een bepaald ogenblik, dingen begint te zien die je helemaal niet had verwacht als de nulhypothese waar zou zijn. En dan besluit je dat je de nulhypothese verwerpt en voor de alternatieve hypothese kiest. Bemerk dat je dit pas doet als je terechtkomt in "extreme gebieden" waarin je slechts met 5 % kans valt als de nulhypothese waar is. In alle andere gevallen blijf je bij de nulhypothese. Toetsen van hypothesen is **een asymmetrische procedure** waarbij je heel lang aan de nulhypothese vasthoudt. Pas bij extreme uitkomsten stap je over op de alternatieve hypothese. Je hebt dan zolang gewacht dat je mag zeggen dat er nu een statistisch bewijs is dat de alternatieve hypothese waar is.

Over de nulhypothese kan je zo'n uitspraak niet doen. Het is niet omdat "ongeveer" 10 van de 20 getrokken kaartjes zwart zijn dat je daarom te maken hebt met een eerlijke doos. Zo'n resultaat krijg je ook gemakkelijk met een doos waarin 251 witte en 249 zwarte kaartjes zitten. De procedure die je gebruikt bij het toetsen van hypothesen levert geen statistisch bewijs dat de nulhypothese waar is. Je start met een nulhypothese en je blijft er mee zitten zolang je ze niet kan verwerpen. Maar of ze juist is, dat vertellen de steekproefresultaten je niet.

Wat je wil bewijzen plaats je in de alternatieve hypothese H_1 want toetsen van hypothesen levert alleen een bewijs voor de bewering die in H_1 staat.

Als je de nulhypothese H_0 kan verwerpen, dan heb je een statistisch bewijs dat H_1 waar is.

Als je de nulhypothese H_0 niet kan verwerpen, dan bewijst dat niet dat H_0 waar is.

Opdracht 4

Je bent verdacht van een misdaad en wordt aangehouden. Ofwel ben je schuldig en dan wacht je de gevangenis. Ofwel ben je onschuldig en dan word je vrijgelaten.

1. Wat zijn de gevolgen als men start met de nulhypothese dat je onschuldig bent? Wat is de taak van de aanklager en wat is jouw taak? Als je wordt vrijgelaten, ben je dan onschuldig?
2. Wat zijn de gevolgen als men start met de nulhypothese dat je schuldig bent? Wat is de taak van de aanklager en wat is jouw taak? Als je in de gevangenis blijft, ben je dan schuldig?

3.2. Soorten fouten

De volgende tabel geeft een samenvatting van alles wat er kan gebeuren:
- wat **de populatie** betreft, gedraagt zij zich zoals beweerd wordt in de nulhypothese of niet
- wat **jouw beslissing** betreft, verwerp je de nulhypothese of niet.

		Voor het gedrag van de populatie geldt:	
		H_0 is waar	H_0 is niet waar
Op basis van de steekproef beslis je:	verwerp H_0	**type I fout**	**correct**
	verwerp H_0 niet	**correct**	**type II fout**

Ofwel is de nulhypothese waar. Voor de doos met 500 kaartjes was die $H_0 : \pi = 0.5$ zodat je echt met een eerlijke doos te maken hebt. Maar ook dan kan het gebeuren dat je uit zo'n doos een extreme steekproef trekt. Dat is de enige informatie die je hebt en je besluit dus dat die doos niet eerlijk is. In dat geval maak je een fout. Zo'n fout wordt een type I fout genoemd.

> Een type I fout maak je als je de nulhypothese H_0 verwerpt terwijl ze in feite waar is.

Het kan ook zijn dat de doos meer witte dan zwarte kaartjes bevat zodat de nulhypothese $H_0 : \pi = 0.5$ niet waar is. Als je uit zo'n doos een steekproef trekt dan kan het gebeuren dat je een resultaat vindt dat helemaal niet onverwacht is wanneer dat uit een eerlijke doos zou komen. En dus zie je geen reden om $H_0 : \pi = 0.5$ te verwerpen. De fout die je nu maakt is een type II fout.

> Een type II fout maak je als je de nulhypothese H_0 niet verwerpt terwijl ze in feite niet waar is.

Opdracht 5

In onze rechtspraak start men met de nulhypothese dat je onschuldig bent. Wanneer wordt hier een type I fout gemaakt? Wanneer een type II fout? Wat vind je het ergste?

3.3. Significantieniveau

Bij toetsen van hypothesen zorgt men ervoor dat de kans om een type I fout te maken klein is. Gewoonlijk neemt men 5 % zodat $P(type\ I\ \ fout) = 5\,\%$. Deze kans wordt **het significantieniveau** genoemd, genoteerd als α.

α = significantieniveau
$= P(type\ I\ \ fout)$
$= P(verwerp\ H_0\ terwijl\ H_0\ waar\ is)$

Opdracht 6

Een type I fout maken kan zeer erge gevolgen hebben. Dat zag je in vorige opdracht. Waarom werkt men eigenlijk met een procedure waarbij je kans hebt om een fout te maken zoals $P(type\ I\ \ fout) = 5\,\%$? Waarom stelt men niet $P(type\ I\ \ fout) = 0$? Wat zou dat betekenen in de rechtspraak?

4. Proporties

4.1. Een toets opstellen

We starten met begrippen die je vroeger geleerd hebt bij het opstellen van een betrouwbaarheidsinterval voor een populatieproportie.

In een studie waarbij de populatie behandeld wordt als categorisch met slechts 2 verschillende uitkomsten stel je die uitkomsten gelijk aan 0 = mislukking en 1 = succes. Je hebt dan te maken met een $0-1$ populatie. In zo'n populatie wordt de kans op succes genoteerd als π. Dat is ook de succesproportie in die populatie of kortweg de populatieproportie.

Als je uit die populatie een steekproef trekt dan heb je een aantal mislukkingen (nullen) en een aantal successen (enen). Je kan daarmee de succesproportie in je steekproef (de waarde van jouw steekproefproportie) berekenen. Die noteer je als $\hat{p}$.

Je weet dat een andere steekproef uit diezelfde populatie je een andere waarde van de steekproefproportie zal opleveren. Daarom kijk je naar het onderliggende kansmodel dat al die verschillende steekproefproporties naar jou stuurt. Dat kansmodel noteer je met een hoofdletter: $\hat{P}$.
Als de steekproef groot genoeg is zodat gelijktijdig voldaan is aan $\begin{cases} n\,\pi \geq 15 \\ n\,(1-\pi) \geq 15 \end{cases}$ dan kan je de kansverdeling van de steekproefproportie $\hat{P}$ benaderen door de normale verdeling.
Je hebt dan dat $\frac{\hat{P}-E(\hat{P})}{se(\hat{P})} \sim N(0;1)$ wat hetzelfde is als $\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}/\sqrt{n}} \sim N(0;1)$ want $E(\hat{P})=\pi$ en $se(\hat{P})=\frac{\sqrt{\pi(1-\pi)}}{\sqrt{n}}$.

Nu komt het verschil met betrouwbaarheidsintervallen.

Het model $\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}/\sqrt{n}}$ bevat de populatieproportie π.

Bij betrouwbaarheidsintervallen ga je ervan uit dat je π niet kent.
Bij toetsen van hypothesen start je met een veronderstelling over π. Je veronderstelt dat de populatieproportie π gelijk is aan wat er beweerd wordt in de nulhypothese. Die waarde noteer je als π_0 en je spreekt dan over "*de populatieproportie onder de nulhypothese"*.

Om te zien hoe dat werkt, bekijk je volgend voorbeeld.

Je vriendin beweert dat bij bevallingen van een (eerste, tweede, derde, …) kind er ongeveer evenveel moeders jonger zijn dan 30 jaar als moeders van minstens 30 jaar. Jij denkt dat er meer jonge moeders zijn. Je geraakt akkoord om dat eens uit te testen voor de geboorten in 1995 in Vlaanderen waarbij je met een steekproef van grootte $n=70$ zal werken. Leeftijd is een continue veranderlijke maar de gestelde onderzoeksvraag maakt er een dichotomie van: minstens 30 jaar noem je

mislukking (en aan die moeders geef je een code 0) en jonger dan 30 jaar noem je succes (en aan die moeders geef je een code 1). Op die manier heb je een $0-1$ populatie gemaakt. De proportie moeders die bij een bevalling in 1995 jonger dan 30 waren, is de populatieproportie, genoteerd als π . Over deze proportie zijn er beweringen gemaakt:

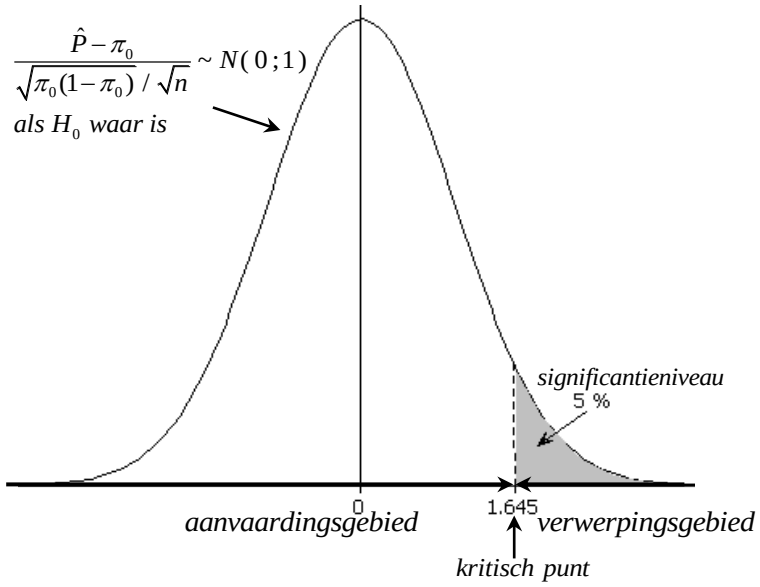
- je vriendin –> nulhypothese $H_0 : \pi = \frac{1}{2}$
- jij –> alternatieve hypothese $H_1 : \pi > \frac{1}{2}$

Je start nu met de veronderstelling dat de nulhypothese waar is. In dat geval zijn er in de populatie evenveel moeders jonger dan 30 jaar als er moeders zijn van minstens 30 jaar. In het model $\frac{\hat{P}-\pi}{\sqrt{\pi(1-\pi)}/\sqrt{n}}$ is π niet langer onbekend. Onder de nulhypothese is $\pi = \frac{1}{2}$, waarbij je de waarde onder de nulhypothese (dat is hier $\frac{1}{2}$) algemeen noteert als π_0.

Er is nu voldaan aan $n\,\pi_0 = 70\,\frac{1}{2} = 35 \geq 15$ en $n\,(1-\pi_0) = 70\,(1-\frac{1}{2}) = 35 \geq 15$ zodat

$\frac{\hat{P}-\pi_0}{\sqrt{\pi_0(1-\pi_0)}/\sqrt{n}} \sim N(0;1)$ of $\frac{\hat{P}-0.5}{\sqrt{0.5\,(1-0.5)}/\sqrt{70}} \sim N(0;1)$ of $\frac{\hat{P}-0.5}{0.06} \sim N(0;1)$.

Je hebt hier een kansmodel dat observeerbaar is want er staat geen enkele onbekende parameter meer in. Bovendien gedraagt dat model zich zoals de standaard normale, wat een verdeling is die je goed kent. Het enige wat je nu nog moet doen is een “extreem gebied” bepalen. Om te zien of je met een linkerstaart of met een rechterstaart of met beide staarten moet werken kijk je naar de alternatieve hypothese. Die is hier éénzijdig en wijst naar grotere waarden. Daarom ga je hier **éénzijdig rechts** toetsen, wat betekent dat je een rechterstaart als “extreem gebied” afbakent. Je werkt hierbij met het klassieke 5 % criterium, zodat je slechts met 5 % kans in die rechterstaart terechtkomt. Het kritische punt is 1.645 want $P(Z \geq 1.645) = 0.05$ als $Z \sim N(0;1)$. Hierbij is Z (hoofdletter) de gangbare notatie voor een kansmodel dat zich gedraagt zoals de standaard normale.



Je hebt nu 2 gebieden bepaald. Het verwerpingsgebied is [1.645 ; + ∞ [. Als je daarin terechtkomt dan verwerp je de nulhypothese. Als je in het aanvaardingsgebied] – ∞ ; 1.645 [terechtkomt dan kan je de nulhypothese niet verwerpen. Dit betekent niet dat je aanvaardt dat H_0 juist is. Eigenlijk zou je "aanvaardingsgebied" moeten vervangen door "niet-verwerpingsgebied" maar dat is niet gebruikelijk in teksten over statistiek.

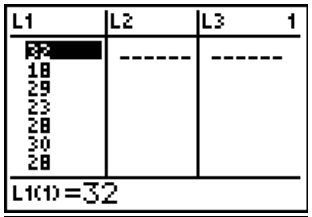
Een toets opstellen komt neer op het bepalen van een regel die zegt: "Als je in dat gebied terechtkomt dan moet je H_0 verwerpen en anders niet". Die regel stel je op door alleen maar gebruik te maken van eigenschappen van het kansmodel. Dat kansmodel is voor iedere leerling in je klas hetzelfde en dus heeft iedereen dezelfde regel opgesteld. Pas daarna trekt iedereen een steekproef. En nu zullen je medeleerlingen uiteraard niet allemaal hetzelfde vinden.

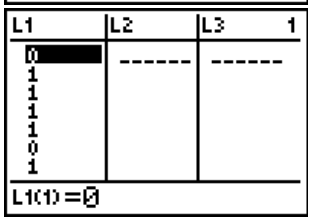
4.2. Een toets uitvoeren

Wanneer je een toets hebt opgesteld dan is het niet moeilijk om ze ook uit te voeren. Je moet er wel altijd op letten dat je op een goede manier data verzamelt. Trek nu uit de databank een steekproef van grootte $n = 70$ en let erop dat je alleen maar geboorten van het jaar 1995 neemt. De leeftijd van de moeder breng je over naar de lijst [L1] van je GRM. Die leeftijd moet je nu een $0 - 1$ code geven. Dat gaat als volgt: druk achtereenvolgens [2nd] [L1] en [2nd] [TEST] en kies 5:< . Tik dan het getal 30 gevolgd door [STO▶] [2nd] [L1] en [ENTER].

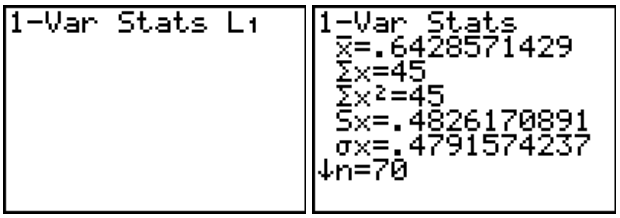
Hiernaast zie je een voorbeeld van zo'n steekproef. Voor de rest van dit verhaal is het de bedoeling dat je werkt met de steekproef die je zelf getrokken hebt.

Volgnr.	Duur	Gew	Sex	Lft_m	Gebjaar
1	39	3670	1	32	1995
2	42	2760	1	18	1995
3	39	3080	1	29	1995
4	40	3700	1	23	1995
5	40	3000	0	28	1995
6	37	2760	1	30	1995
7	38	3300	0	28	1995
8	38	3140	1	32	1995
9	39	2590	0	27	1995
10	39	2970	0	22	1995
11	40	3500	1	36	1995
12	41	4390	0	31	1995

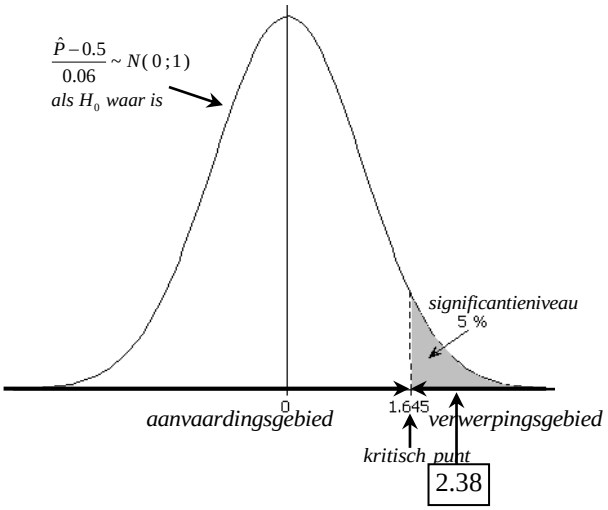




Onder de nulhypothese werk je met **het model** $\frac{\hat{P} - 0.5}{0.06}$. Voor de steekproef die nu getrokken is bereken je **de waarde** (kleine letter) $\frac{\hat{p} - 0.5}{0.06}$. Daarvoor moet je de succesproportie $\hat{p}$ in je steekproef kennen. Die is niets anders dan $\frac{\textit{aantal successen}}{\textit{totaal aantal}} = \frac{\textit{aantal enen}}{\textit{totaal aantal}}$. Het aantal enen vind je door de som te maken van je uitkomsten want die zijn niets anders dan nullen en enen.



Druk [STAT] loop naar CALC en druk 1:1–Var Stat. Vervolledig het commando met [2nd] [L1] en [ENTER]. In dit voorbeeld is $\sum x = 45$ zodat $\hat{p} = \frac{45}{70} \cong 0.643$.



Je bent op $\frac{\hat{p}-0.5}{0.06} = \frac{0.643-0.5}{0.06} \cong 2.38$ terechtgekomen, wat een punt is in het verwerpingsgebied.

Als je in het verwerpingsgebied terechtkomt dan zeg je dat de geziene afwijking (ten opzichte van de bewering van de nulhypothese) geen "toevallig verschil" meer is, te wijten aan schommelingen van steekproeven. Neen, dan zeg je dat je een **significant verschil** gevonden hebt en dan stap je over van de nulhypothese op de alternatieve hypothese.

Besluit.
Verwerp de nulhypothese op het 5 % significantieniveau en aanvaard dat, bij de geboorten in 1995 in Vlaanderen, de proportie moeders jonger dan 30 jaar groter is dan 1/2.

Bemerkingen.

1. Bij toetsen van hypothesen formuleer je een besluit waarbij je zegt "welk criterium voor extreem" (welk significantieniveau) je gebruikt hebt.

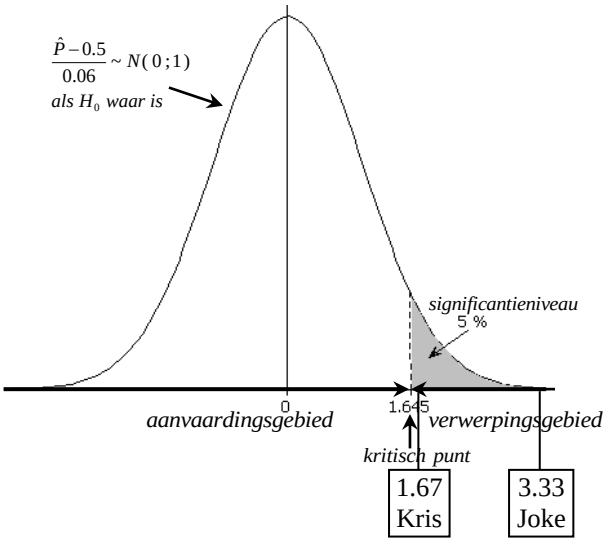
2. Zoals bij betrouwbaarheidsintervallen bestaat ook toetsen van hypothesen uit twee grote stappen: een stap "**vooraf**" en een stap "**nadien**".

   Vooraf stel je een procedure op waarbij je zelf bepaalt wat de kans op een type I fout is. Als je met een significantieniveau van 5 % werkt, dan is de kans om de nulhypothese te verwerpen, terwijl ze toch waar is, gelijk aan 5 %. Je kan dat "in the long run" bekijken. Als je, terwijl de nulhypothese waar is, 100 keer zo'n toets uitvoert, dan zal je (ongeveer) 5 keer die nulhypothese verwerpen (en een foutieve beslissing nemen).

   Nadien trek je de steekproef. Dan sta je daar met één welbepaalde waarde van je model. Dat is de basis waarop je een beslissing neemt (H_0 verwerpen of niet). Die beslissing is juist of fout, daar hoort geen kansuitspraak meer bij. Dat is ook zo bij betrouwbaarheidsintervallen, waar je achteraf een interval vindt dat juist of fout is.

4.3. De p–waarde

Kris en Joke hebben ook een steekproef van grootte $n = 70$ getrokken. Kris vond 42 moeders jonger dan 30 jaar zodat voor haar steekproef $\hat{p} = \frac{42}{70} = 0.6$. Zij komt dus op de waarde $\frac{\hat{p} - 0.5}{0.06} = \frac{0.6 - 0.5}{0.06} = 1.67$ terecht. Bij Joke waren er 49 moeders jonger dan 30 jaar. Zij vindt $\hat{p} = \frac{49}{70} = 0.7$ en komt op de waarde $\frac{\hat{p} - 0.5}{0.06} = \frac{0.7 - 0.5}{0.06} = 3.33$ terecht. Grafisch kan je hun resultaten voorstellen zoals hiernaast. Zij zijn beiden in het verwerpingsgebied terecht gekomen zodat hun besluit identiek is: "verwerp de nulhypothese op het 5 % significantieniveau".



Op deze manier een besluit formuleren is correct maar het vertelt niet alles. Als de nulhypothese waar is en de kans op een moeder jonger dan 30 gelijk is aan 50 %, dan is het resultaat van Joke toch nog veel meer onverwacht dan dat van Kris. In de steekproef van Kris waren er 60 % jonge moeders, maar bij Joke waren er dat 70 %.

Om deze extra informatie weer te geven wil je graag zeggen hoe extreem jouw steekproef wel was. Je weet dat je dit moet aangeven met "kansen om in extreme gebieden terecht te komen". Dat doet de p–waarde (= *p–value* = *probability value*). Het is deze waarde die je in de output van de meeste statistische software vindt.

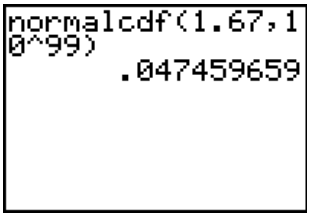
De p–waarde is de kans dat het kansmodel, als de nulhypothese waar is, even extreme of nog extremere waarden aanneemt, vergeleken met de waarde waarop jij met jouw steekproef bent terechtgekomen.

Voor een éénzijdig rechtse toets wijzen extreme waarden naar "groter". De p–waarde kan je hier opschrijven als $P\left(\frac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}/\sqrt{n}} \geq \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}/\sqrt{n}}\right)$. Bekijk deze kansuitspraak goed. Het gaat over de kans dat het kansmodel onder de nulhypothese, dus $\frac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$, een waarde aanneemt die minstens zo groot is als wat jij gevonden hebt vanuit je steekproef, namelijk $\frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$.

Als je voor het kansmodel (benaderend) werkt met de standaard normale verdeling dan noteert men de gevonden waarde dikwijls met een (kleine) letter z want het standaard normaal kansmodel wordt met een (hoofdletter) Z genoteerd.

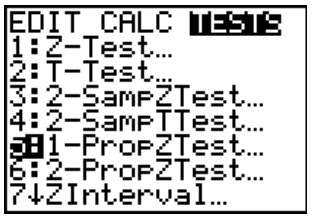
Voor Kris is $z = \dfrac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}} = 1.67$ zodat haar p–waarde gelijk is aan $P\left(\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}/\sqrt{n}} \geq 1.67 \right)$ of aan $P\left(\dfrac{\hat{P} - 0.5}{0.06} \geq 1.67 \right)$.

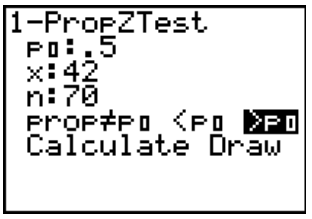
Aangezien $\dfrac{\hat{P} - 0.5}{0.06} \sim N(0;1)$ en $P(Z \geq 1.67) = P(1.67 \leq Z < +\infty) \cong 0.047$ is de p-waarde van Kris gelijk aan 4.7 %. Met de GRM kan je dit als volgt vinden: druk [2nd] [DISTR], kies 2:normalcdf(, vul verder in zoals aangegeven en druk [ENTER].

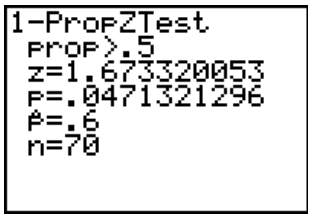


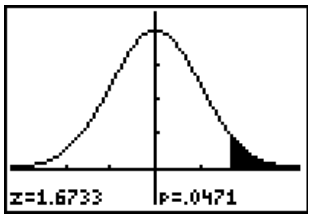
Het extreme gebied dat door het resultaat van Kris wordt afgebakend is $[\,1.67\,;+\infty\,[$. Dat is meer extreem dan $[\,1.645\,;+\infty\,[$ en dus kom je daar met een kleinere kans in terecht. Daarom is de p-waarde van Kris kleiner dan 5 %.

Zo'n p–waarde kan je ook anders vinden. Druk [STAT], loop naar TESTS en kies 5:1–PropZTest…









Je GRM schrijft p_0 waar er eigenlijk π_0 moet staan. Om te beginnen tik je de populatieproportie onder de nulhypothese in. In deze studie is $\pi_0 = 0.5$. Met de letter x vraagt je GRM naar het aantal successen (dat is hier 42) en n staat voor de steekproefgrootte (dat is hier 70). Voor de alternatieve hypothese moet je kiezen uit $H_1 : \pi \neq \pi_0$ of $H_1 : \pi < \pi_0$ of $H_1 : \pi > \pi_0$. In deze studie veronderstelt de alternatieve hypothese dat de populatieproportie groter is dan π_0. Ga dus naar die plaats ($> \pi_0$) en druk [ENTER]. Ga dan naar Calculate en druk [ENTER]. Je krijgt nu een scherm dat zegt dat je rechts éénzijdig toetst (prop>.5 betekent $\pi > 0.5$), dat je terechtgekomen bent op de waarde 1.67 (z = 1.67) en dat de p–waarde 4.7 % is (p=.047). Verder staat er dat de succesproportie in de steekproef 60 % was ($\hat{p} = .6$) bij een steekproef van grootte $n = 70$.

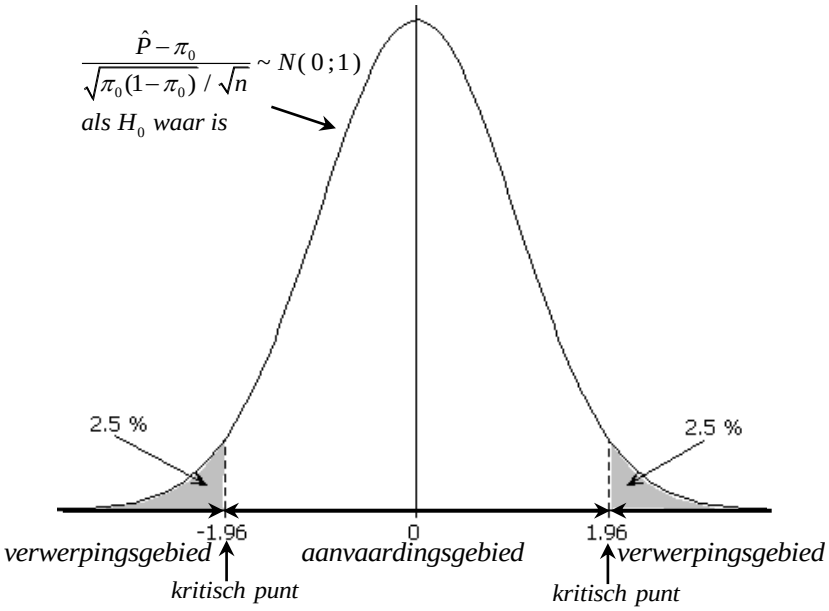
Als je bij hetzelfde commando niet Calculate gebruikt maar Draw en dan [ENTER] drukt, dan krijg je een figuur van de standaard normale dichtheidsfunctie met de waarde waar jij bent terechtgekomen (z=1.67) en een gearceerde oppervlakte boven de rechterstaart $[\,1.67\,;+\infty\,[$. Het maatgetal van die oppervlakte is de kans dat een standaard normale in het gebied $[\,1.67\,;+\infty\,[$ terechtkomt. Die kans is 4.7 %, aangegeven door p=.047. Dat is de p–waarde.

Opdracht 7

1. Bereken de p–waarde van Joke. Schrijf eerst (met de juiste notatie) de kansuitspraak op die hierbij hoort en formuleer ze ook in woorden. Gebruik daarna je GRM om de p-waarde te berekenen. Doe dit met het commando voor het toetsen van 1 proportie.
2. *Schrap wat niet past*: “Een grotere p-waarde duidt op een meer extreme uitkomst: JA / NEEN”. Motiveer je antwoord.
3. *Schrap wat niet past*: “Als je een verwerpingsgebied hebt opgesteld op basis van een 5 % significantieniveau en je vindt een p–waarde die kleiner is dan 5 % dan ben je WEL / NIET in dat verwerpingsgebied terechtgekomen”. Geef uitleg bij deze uitspraak en gebruik daarbij de resultaten van Kris en Joke ter illustratie.

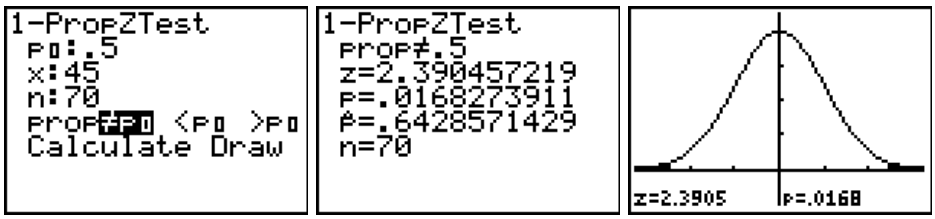
4.4. Tweezijdig toetsen

Als je vriendin beweert dat er bij de bevallingen in 1995 in Vlaanderen evenveel vrouwen waren jonger dan 30 als vrouwen van minstens 30, dan zou je dat gewoon in twijfel kunnen trekken zonder zelf een specifiek tegenvoorstel te doen. Volgens jou is de proportie π van vrouwen jonger dan 30 niet gelijk aan 1/2. Voor deze studie werk je dan met de nulhypothese $H_0 : \pi = 0.5$ tegenover de alternatieve hypothese $H_1 : \pi \neq 0.5$. Een "extreme" uitkomst die je niet had verwacht als de nulhypothese waar zou zijn is nu een uitkomst die "te groot" of "te klein" is. Het verwerpingsgebied bestaat dan uit twee staarten (extreem links en extreem rechts). Als je met een significantieniveau $\alpha = 5\,\%$ werkt, dan zorg je ervoor dat het kansmodel in het verwerpingsgebied terechtkomt met kans 5 %. Wanneer de voorwaarden voldaan zijn om de normale benadering te gebruiken is $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}} \sim N(0;1)$ en bestaat het verwerpingsgebied uit $]-\infty ; -1.96]$ samen met $[1.96 ; +\infty[$. Dat zie je ook op de figuur.



Om p–waarden te berekenen gebruik je een analoge redenering. In de getrokken steekproef die we als voorbeeld gebruikt hebben waren er 45 successen zodat $\hat{p} = \dfrac{45}{70} \cong 0.643$. Hiermee kom je terecht op de z–waarde $z = \dfrac{\hat{p} - 0.5}{0.06} \cong \dfrac{0.643 - 0.5}{0.06} \cong 2.38$. Hier zitten wat afgeronde getallen in de berekening. Je GRM geeft aan dat je terechtkomt op $z = 2.39$.

Wat is nu de kans dat het model $\dfrac{\hat{P} - 0.5}{0.06}$ terechtkomt in een gebied dat even ver of nog verder van het centrum verwijderd is dan jouw z–waarde $z = 2.39$? Dat is de kans dat de standaard normale valt in $]-\infty ; -2.39]$ of in $[2.39 ; +\infty[$. Die kans is de p–waarde en die is hier gelijk aan 0.017 of 1.7 %. Hoe je die met je GRM berekent zie je hiernaast.



Opdracht 8

1. De steekproef van het voorbeeld hierboven leverde een p-waarde van 1.7 %. Welke conclusie trek je hieruit? Formuleer die conclusie nauwkeurig en volledig.
2. Bereken (met je GRM) de p–waarde van de steekproef die je zelf hebt getrokken en veronderstel daarbij dat het gaat over een studie van $H_0 : \pi = 0.5$ tegenover $H_1 : \pi \neq 0.5$. Kan je uit de vorm van de alternatieve hypothese afleiden welk type toets je moet doen? Hoe heet zo'n toets?

4.5. Een robuuste procedure

In de vorige paragrafen heb je gezien dat je start met de steekproefproportie $\hat{P}$ als kansmodel bij het toetsen van een populatieproportie π. Dat model standaardiseer je waarbij je π vervangt door π_0 omdat je veronderstelt dat de nulhypothese waar is. Zo krijg je het model $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$. Om te weten met welke kans dit model in bepaalde gebieden valt moet je het gedrag van $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$ kennen.

Dit gedrag kan je benaderen met de standaard normale verdeling zodat je kan werken met

$$\frac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}} \sim N(0;1)$$

- als de steekproef "voldoende groot" is zodat $n\pi_0 \geq 15$ en $n(1-\pi_0) \geq 15$

Maar ook wanneer de steekproefgrootte niet aan deze eisen voldoet, dan nog is de normale benadering dikwijls goed genoeg. Je spreekt dan over een "robuuste" methode die toch nog redelijk goed blijft werken zelfs als de voorwaarden niet helemaal voldaan zijn (zonder in extremen te vervallen natuurlijk). Daarom mag je ook de standaard normale verdeling blijven gebruiken:

- bij kleinere steekproeven waarbij je tweezijdig toetst
- bij kleinere steekproeven waarbij je éénzijdig toetst en waarbij de nulhypothese gelijk is aan $H_0 : \pi = \dfrac{1}{2}$.

De situatie van een kleine steekproef bij een éénzijdige toets waarbij de nulhypothese niet gelijk is aan $H_0 : \pi = \dfrac{1}{2}$ behandelen we niet op het niveau van het secundair onderwijs.

4.6. Samenvatting

Toetsen van hypothesen voor een populatieproportie.

1. De onderzoeksvraag zorgt ervoor dat de populatie kan behandeld worden als een 0 – 1 populatie waarbij π de proportie successen (of de kans op succes) in de populatie is.

2. Over de populatieproportie wordt een nulhypothese en een alternatieve hypothese geformuleerd van de vorm:
   $H_0 : \pi = \pi_0$
   $H_1 : \pi \neq \pi_0$ of $H_1 : \pi > \pi_0$ of $H_1 : \pi < \pi_0$
   Deze hypothesen worden geformuleerd zonder naar de resultaten van de steekproef te kijken.

3. Als kansmodel werk je met de gestandaardiseerde steekproefproportie $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$ waarbij je veronderstelt dat de nulhypothese waar is zodat je π kan vervangen door π_0.

4. Dan trek je een "goede" steekproef (EAS) of anders argumenteer je waarom de data die je ter beschikking hebt, representatief zijn voor de populatie, zoals een EAS.

5. Je gaat na (onder meer op basis van de steekproefgrootte) of je de standaard normale verdeling mag gebruiken zodat $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}} \sim N(0;1)$.

6. Dan voer je, met de GRM, de berekeningen uit. In je verslag noteer je de steekproefgrootte n, de gevonden steekproefproportie $\hat{p}$, de z-waarde $z = \dfrac{\hat{p} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}}$ en de p-waarde.

7. Geef ook aan wat de p–waarde betekent in de context van de uitgevoerde studie.
   Zeg welk significantieniveau (bv. $\alpha = 5\,\%$) je gebruikt en trek dan volgend besluit:
   - als de p-waarde niet groter is dan het significantieniveau dan is je resultaat **statistisch significant** zodat je **de nulhypothese verwerpt op dat significantieniveau** en kiest voor de alternatieve hypothese
   - in het andere geval zeg je dat het gevonden resultaat je niet toelaat om de nulhypothese te verwerpen. Tot nader order is π_0 *een aannemelijke waarde* voor de populatieproportie.

Opdracht 9

In de Verenigde Staten werd bij een goed getrokken steekproef van 116 astrologen de studie uitgevoerd die in opdracht 2 staat beschreven. Er waren 40 astrologen die een goed antwoord gaven. Bevestigt dit dat astrologie (bij Amerikaanse astrologen) werkt? Toets op het 5 % significantieniveau. Gebruik als leidraad de stappen die in de samenvatting zijn aangegeven. Een deel van de antwoorden van opdracht 2 komen ook hier van pas. Je kan die kort herhalen.

Opdracht 10

Men schat dat 32 % van de Vlamingen dagelijks rookt. De overheid wil een gerichte antirookcampagne voeren en zoekt eerst naar gemeenten waar het percentage rokers beduidend hoger ligt. In die gemeenten zal een extra campagne gevoerd worden. Men gaat daarbij als volgt te werk. In elk van de 308 Vlaamse gemeenten wordt een goede en voldoend grote steekproef getrokken. Men gebruikt de code 0 = geen dagelijkse roker en 1 = wel een dagelijkse roker. Dan voert men voor elke gemeente een hypothesetoets uit op het 5 % significantieniveau.

1. Formuleer de nulhypothese en de alternatieve hypothese en zeg over welke populatieparameter het gaat.
2. In die studie vindt men dat in jouw gemeente het percentage rokers significant groter is dan 32 %. Is dat een goede reden om daar (met belastingsgeld) een extra campagne te voeren? Welke bedenkingen kan je hierbij maken?

5. Gemiddelden

Wanneer je te maken hebt met populatiekarakteristieken die je als continu kan behandelen (zoals leeftijd, gewicht, lengte, …) dan kan je een hypothese toetsen over het populatiegemiddelde μ. Dat is dan de populatieparameter waarin je geïnteresseerd bent.

We maken nu gebruik van wat je al weet over
- toetsen van hypothesen voor proporties
- betrouwbaarheidsintervallen voor gemiddelden.

Als je dit goed bestudeerd hebt dan is het niet moeilijk om te begrijpen hoe je hypothesen over gemiddelden toetst.

5.1. Een toets voor μ

Het populatiegemiddelde μ bestudeer je met behulp van het steekproefgemiddelde $\bar{X}$. Dat is het kansmodel waarmee je start.

Je weet dat $E(\bar{X}) = \mu$ en $se(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ zodat $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ een gestandaardiseerd kansmodel is. In geen enkele realistische studie ken je de standaardafwijking σ van de populatie. Die grootheid vervang je door de steekproefstandaardafwijking $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. Uiteindelijk kom je zo terecht op het kansmodel $\frac{\bar{X} - \mu}{S / \sqrt{n}}$. Omdat je σ vervangen hebt door S werk je niet meer met de standaard normale verdeling maar met een t–verdeling.

Om te zien hoe een toets voor μ werkt, start je met een voorbeeld.

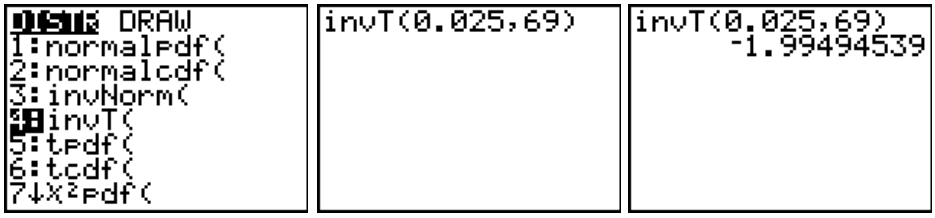
Je vriend beweert dat het gemiddelde geboortegewicht van meisjes 3 kg is. Jij hebt geen idee of dat waar is, maar dat het 3 kg is geloof je niet. Je komt overeen om dit eens na te gaan voor de geboorten van meisjes in 2000 in Vlaanderen. Je beslist om met een steekproef van grootte $n = 70$ te werken.

Geboortegewicht is een continue veranderlijke en de onderzoeksvraag gaat hier over het gemiddelde geboortegewicht μ van alle meisjes die in 2000 in Vlaanderen geboren zijn.
Over dit populatiegemiddelde heb je de volgende hypothesen (gewichten uitgedrukt in gram):
- je vriend –> nulhypothese $H_0 : \mu = 3000$
- jij –> alternatieve hypothese $H_1 : \mu \neq 3000$

Je start met de veronderstelling dat de nulhypothese waar is. In dat geval is μ niet langer onbekend in het model $\frac{\bar{X} - \mu}{S / \sqrt{n}}$. Algemeen schrijf je μ_0 voor de waarde van μ onder de nulhypothese. Je werkt dan met $\frac{\bar{X} - \mu_0}{S / \sqrt{n}}$. Hier is $\mu_0 = 3000$ en dus werk je vanaf nu met $\frac{\bar{X} - 3000}{S / \sqrt{n}}$. Een steekproef met $n = 70$ is voldoende groot om te mogen veronderstellen dat $\frac{\bar{X} - 3000}{S / \sqrt{n}}$ zich gedraagt als een t-verdeling met $(n-1) = 69$ vrijheidsgraden.

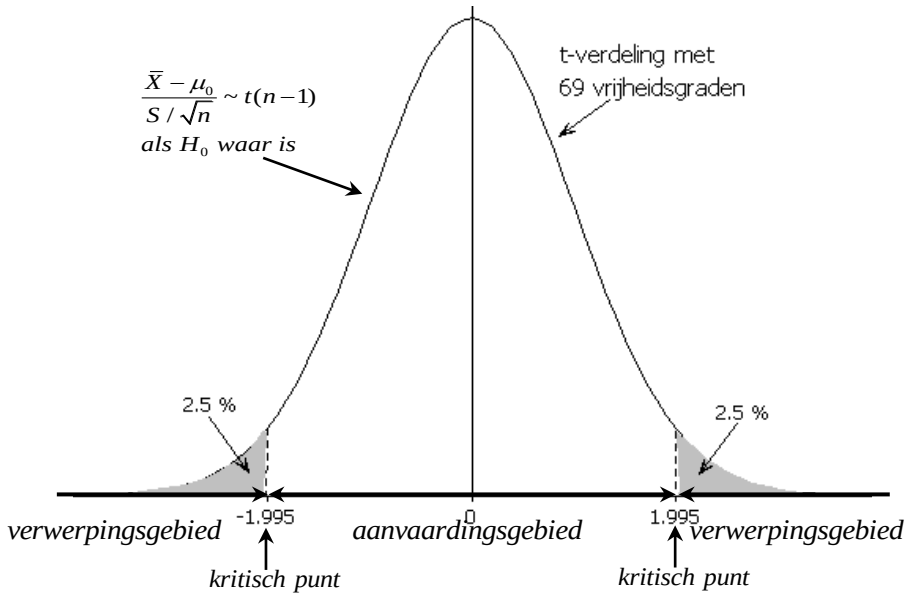
Een model dat zich gedraagt als een t–verdeling wordt dikwijls genoteerd met een (hoofdletter) T. De notatie $T \sim t(n-1)$ betekent dat het kansmodel T een t–verdeling met (n–1) vrijheidsgraden heeft. De waarde die zo'n model aanneemt noteer je als t (kleine letter). Je spreekt dan over een gevonden t–waarde.

De alternatieve hypothese $H_1 : \mu \neq 3000$ zegt dat zowel "te grote" als "te kleine" uitkomsten als "onverwacht" of "extreem" beschouwd worden wanneer de nulhypothese $\mu = 3000$ waar zou zijn. Je moet dus gelijktijdig een linker– en rechterstaart afbakenen waar je onder de nulhypothese met 5 % kans in terechtkomt als je werkt met een significantieniveau $\alpha = 5\,\%$. Wegens de symmetrie van de t-verdeling bepaal je eerst een linkerstaart waar je met kans 2.5 % in terechtkomt. Je kan dat met je GRM. Druk [2nd] [DISTR] en kies 4:invT(. Vervolledig het commando zoals aangegeven en druk [ENTER].



Je vindt hier als kritisch punt –1.995.

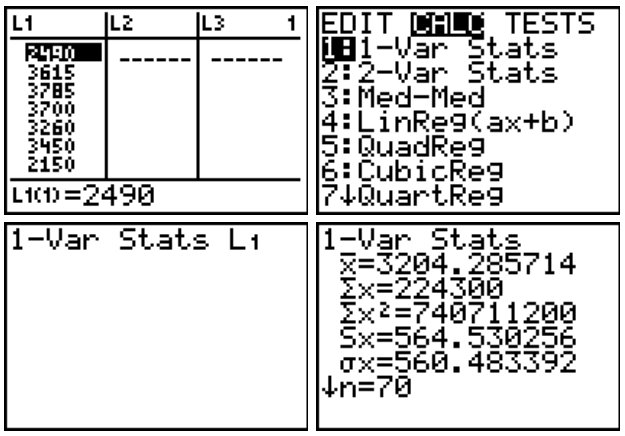
De t–verdeling met 69 vrijheidsgraden valt in de linkerstaart $]-\infty\,;-1.995]$ met kans 2.5 %. Wegens symmetrie is er ook 2.5 % kans om in de rechterstaart $[1.995\,;+\infty[$ terecht te komen. Als je in die staarten terechtkomt dan verwerp je de nulhypothese op het 5 % significantieniveau. De opgestelde toets kan je ook goed grafisch voorstellen. Dat zie je op onderstaande figuur.



Bemerk dat je nog altijd geen steekproef getrokken hebt

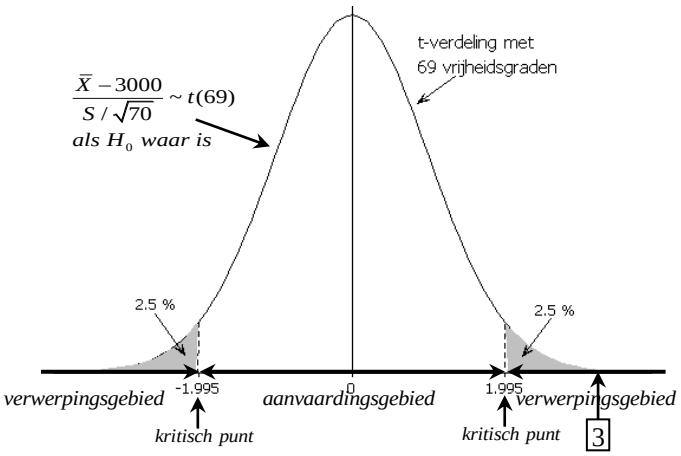
Trek nu uit de databank een steekproef van grootte $n = 70$ en let erop dat je alleen maar meisjes van het jaar 2000 neemt. Breng de geboortegewichten over naar de lijst [L1] van je GRM.
Hiernaast zie je een voorbeeld van zo'n steekproef. Voor de rest van dit verhaal is het de bedoeling dat je werkt met de steekproef die je zelf getrokken hebt.

Volgnr.	Duur	Gew	Sex	Lft_m	Gebjaar
1	35	2490	0	27	2000
2	40	3615	0	24	2000
3	41	3785	0	27	2000
4	40	3700	0	25	2000
5	38	3260	0	27	2000
6	40	3450	0	29	2000
7	37	2150	0	23	2000
8	41	3490	0	32	2000
9	40	2995	0	29	2000
10	36	2540	0	18	2000
11	40	4120	0	27	2000
12	37	2830	0	26	2000



Onder de nulhypothese werk je met **het model** $\dfrac{\bar{X} - 3000}{S/\sqrt{70}}$. Voor de steekproef die nu getrokken is bereken je **de t–waarde** (kleine letters) $\dfrac{\bar{x} - 3000}{s/\sqrt{70}}$. Met je GRM vind je dat $\bar{x} = 3204.3$ en $s = 564.53$ zodat $t = \dfrac{\bar{x} - 3000}{s/\sqrt{70}} = \dfrac{3204.3 - 3000}{564.53/\sqrt{70}} \cong 3$. Je bent met je t–waarde in het verwerpingsgebied terechtgekomen.

Als je in het verwerpingsgebied terechtkomt dan zeg je dat het gevonden verschil niet meer te wijten is aan toevallige schommelingen van steekproeven. Je spreekt dan over een **significant verschil** en je stapt over van de nulhypothese op de alternatieve hypothese.

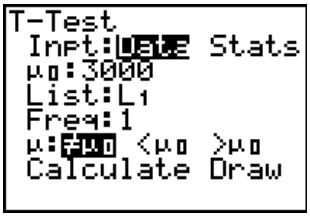


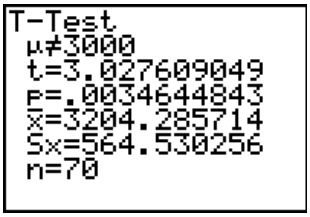
Besluit.
Verwerp de nulhypothese op het 5 % significantieniveau en aanvaard dat, in het jaar 2000 in Vlaanderen, het gemiddelde geboortegewicht van meisjes niet gelijk was aan 3 kg. Het gevonden resultaat suggereert dat het groter was.

Je weet dat je beter kan dan alleen maar besluiten of je een nulhypothese al dan niet verwerpt. Je kan aangeven hoe sterk je steekproefresultaten de nulhypothese ongeloofwaardig maken. Dat doe je met de p-waarde.

Bij de tweezijdige toets die je hier uitvoert kijk je naar staarten (links en rechts) met waarden die even extreem zijn of nog extremer dan je gevonden t–waarde. In dit voorbeeld kijk je dus naar zowel $]-\infty;-3]$ als naar $[3;+\infty[$. De kans om met een t-verdeling met 69 vrijheidsgraden in die staarten terecht te komen bereken je met de GRM.

Druk [STAT], loop naar TESTS en kies 2:T-Test… Vul in zoals aangegeven, loop naar Calculate en druk [ENTER].





Besluit.
De getrokken steekproef van grootte $n = 70$ levert een gemiddeld geboortegewicht dat gelijk is aan $\bar{x} = 3204.3$.
De gevonden t–waarde is $t = 3$ en de p–waarde is $p = 0.003 = 0.3\,\%$.
Aangezien 0.3 % < 5 % kan je de nulhypothese verwerpen op het 5 % significantieniveau. Je aanvaardt dat, in het jaar 2000 in Vlaanderen, het gemiddelde geboortegewicht van meisjes niet gelijk was aan 3 kg. Het gevonden resultaat suggereert dat het groter was.

5.2. Een robuuste procedure

Het kansmodel $\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}}$ dat je gebruikt bij het toetsen van een populatiegemiddelde μ heeft een t-verdeling wanneer de onderliggende populatie een normale verdeling heeft.

Juist zoals bij een proportie is ook de methode die je gebruikt bij een gemiddelde "robuust".
Je mag, als benadering, werken met $\dfrac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t(n-1)$

- als de steekproef "voldoende groot" is zodat n ongeveer 30 is of meer
- bij kleinere steekproeven waarbij je tweezijdig toetst
- bij kleinere steekproeven waarbij je éénzijdig toetst en waarbij de onderliggende populatie niet te scheef is.

5.3. Samenvatting

Toetsen van hypothesen voor een populatiegemiddelde.

1. De onderzoeksvraag gaat over het gemiddelde μ van een populatie die als een continue veranderlijke kan behandeld worden.

2. Over het populatiegemiddelde wordt een nulhypothese en een alternatieve hypothese geformuleerd van de vorm:
   $H_0 : \mu = \mu_0$
   $H_1 : \mu \neq \mu_0$ of $H_1 : \mu > \mu_0$ of $H_1 : \mu < \mu_0$
   Deze hypothesen worden geformuleerd zonder naar de resultaten van de steekproef te kijken.

3. Als kansmodel werk je met het gestandaardiseerde steekproefgemiddelde $\frac{\bar{X} - \mu_0}{S / \sqrt{n}}$ waarbij je σ vervangt door S en veronderstelt dat de nulhypothese waar is zodat je μ kan vervangen door μ_0.

4. Dan trek je een "goede" steekproef (EAS) of anders argumenteer je waarom de data die je ter beschikking hebt, representatief zijn voor de populatie, zoals een EAS.

5. Je gaat na (onder meer op basis van de steekproefgrootte) of je de t-verdeling kan gebruiken zodat $\frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim t\,(n-1)$.

6. Dan voer je, met de GRM, de berekeningen uit. In je verslag noteer je de steekproefgrootte n, het gemiddelde $\bar{x}$ en de standaardafwijking s van de steekproef, de t-waarde $t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$ en de p-waarde.

7. Geef ook aan wat de p–waarde betekent in de context van de uitgevoerde studie.
   Zeg welk significantieniveau (bv. $\alpha = 5\,\%$) je gebruikt en trek dan volgend besluit:
   - als de p-waarde niet groter is dan het significantieniveau dan is je resultaat **statistisch significant** zodat je **de nulhypothese verwerpt op dat significantieniveau** en kiest voor de alternatieve hypothese
   - in het andere geval zeg je dat het gevonden resultaat je niet toelaat om de nulhypothese te verwerpen. Tot nader order is μ_0 *een aannemelijke waarde* voor het populatiegemiddelde.

Opdracht 11

Om het onderzoek beschreven in opdracht 1 uit te voeren, heb jij een steekproef getrokken van 105 leerlingen die vorig jaar in Vlaanderen in de derde graad ASO zaten. Van die leerlingen heb je de score op het eindexamen wiskunde genoteerd (op een maximum van 100). Dat leverde een gemiddelde score van 55 en een standaardafwijking van 11.
Gebruik deze informatie om het onderzoek van opdracht 1 uit te voeren. Toets op het 5 % significantieniveau. Gebruik als leidraad de stappen die in de samenvatting zijn aangegeven. De antwoorden van opdracht 1 die hier van pas komen, kan je kort herhalen.

Opdracht 12

Hypothesen moet je formuleren vooraleer je naar de data kijkt. Veronderstel eens dat je in de vorige opdracht vooraf zou gekeken hebben naar de scores die je vond in je steekproef. In je steekproef is het gemiddelde gelijk aan 55 wat minder is dan de gemiddelde score van 57 die in 2000 werd behaald. Dat brengt je op het idee om links éénzijdig te toetsen: $H_0 : \mu = 57$ tegenover $H_1 : \mu < 57$. Om deze toets uit te voeren gebruik je daarna dezelfde data die je op dat idee hebben gebracht. Dat is een voorbeeld van "data snooping" (wat niet mag in de statistiek).

1. Gebruik je GRM om voor deze éénzijdige toets de p–waarde te berekenen.
2. Welk besluit zou je nu trekken op het 5 % significantieniveau?

5.4. Is significant belangrijk?

Een resultaat dat statistisch significant is wijst niet noodzakelijk op een verschil dat in de praktijk belangrijk is. Dat zijn twee verschillende dingen. Als de steekproef zeer groot is dan vind je een uitkomst die bijna altijd *significant* verschilt van de nulhypothese, ook als die uitkomst dicht tegen de nulhypothese ligt. Dat ontdek je in de volgende opdracht.

Opdracht 13

Tussen 1993 en 2008 was in Vlaanderen het jaarlijkse gemiddelde geboortegewicht μ ongeveer gelijk aan 3305 g. Dat jaargemiddelde was zeer stabiel over de jaren heen, met afwijkingen die nooit groter dan 15 g waren. Een verschil van meer dan 20 g zou in de medische wereld misschien even de wenkbrauwen doen fronsen maar kleinere afwijkingen vindt men niet alarmerend.
Jij trekt een steekproef bij de baby's die vorig jaar geboren werden en vindt een gemiddeld gewicht van 3314 g met een standaardafwijking van 350 g. Je voert een toets uit van $H_0 : \mu = 3305$ tegenover $H_1 : \mu \neq 3305$ op het 5 % significantieniveau.

1. Bereken de p–waarde en trek een besluit als de steekproef een grootte had van $n = 400$.
2. Bereken de p–waarde en trek een besluit als de steekproef een grootte had van $n = 6000$.

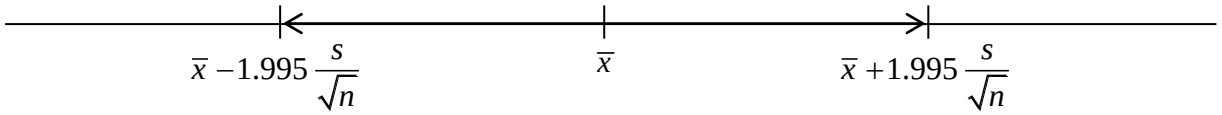
5.5. Toetsen en betrouwbaarheidsintervallen

Als je een populatiegemiddelde μ bestudeert dan ontdek je dat ***tweezijdige*** toetsen en betrouwbaarheidsintervallen nauw verwant zijn (je moet dan uiteraard een interval met 95 % betrouwbaarheid nemen wanneer je toetst op het 5 % significantieniveau).

De reden is heel eenvoudig en misschien heb je die al ontdekt: je start vanuit eenzelfde kansmodel $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t\,(n-1)$.

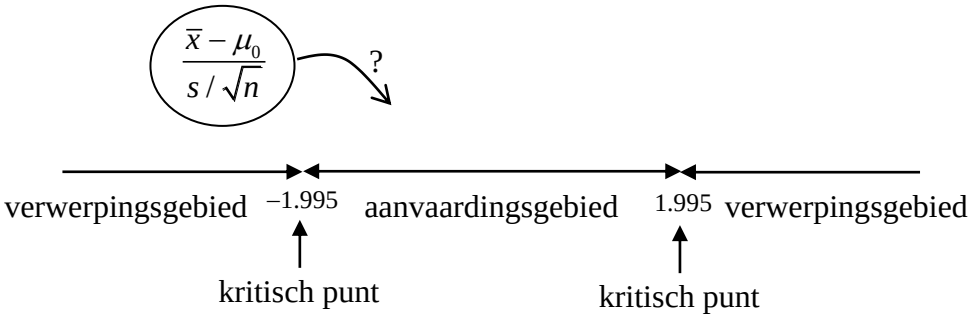
Voor betrouwbaarheidsintervallen leidt $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t\,(n-1)$ tot $\bar{X} \pm (t-waarde)\,\frac{S}{\sqrt{n}}$. Dit is het kansmodel dat betrouwbaarheidsintervallen genereert. Hierin is $(t-waarde)$ het kritisch punt van de t-verdeling. Als je even veronderstelt dat je werkt met een steekproef van grootte $n=70$ dan vind je 1.995 als kritisch punt bij een t–verdeling met $(n-1)=69$ vrijheidsgraden.

Je trekt nu de steekproef en berekent het gemiddelde $\bar{x}$ en de standaardafwijking s.
Je 95 % betrouwbaarheidsinterval ziet er dan als volgt uit:



De waarden in het interval [$\bar{x}-1.995\frac{s}{\sqrt{n}}$; $\bar{x}+1.995\frac{s}{\sqrt{n}}$] zijn *aannemelijke* waarden voor het populatiegemiddelde μ.

Als je de hypothese $H_0:\mu=\mu_0$ tegenover $H_1:\mu\neq\mu_0$ toetst dan werk je met $\frac{\bar{X}-\mu_0}{S/\sqrt{n}} \sim t\,(n-1)$ om de toets op te stellen.
Na het trekken van de steekproef kijk je in welk gebied de gevonden $\frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ terechtkomt.



Als je in het aanvaardingsgebied valt, dan kan je de nulhypothese $H_0:\mu=\mu_0$ niet verwerpen. In dat geval is μ_0 een *aannemelijke* waarde voor het populatiegemiddelde μ.

Je komt in het aanvaardingsgebied terecht enkel en alleen als $-1.995 < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < 1.995$. Je kan dit herschrijven en dan krijg je $\bar{x} - 1.995\, s/\sqrt{n} < \mu_0 < \bar{x} + 1.995\, s/\sqrt{n}$. Wanneer je dus als nulhypothese een waarde μ_0 neemt die in het betrouwbaarheidsinterval ligt dan zal je (met dezelfde gegevens) die nulhypothese niet kunnen verwerpen. Het betrouwbaarheidsinterval valt samen met de verzameling van alle nulhypothesen die niet kunnen verworpen worden. Die verzameling is strikt genomen het open interval $\left] \bar{x} - 1.995 \frac{s}{\sqrt{n}} \; ; \; \bar{x} + 1.995 \frac{s}{\sqrt{n}} \right[$.

Opdracht 14

In de tekst over betrouwbaarheidsintervallen voor gemiddelden werd een steekproef van grootte $n = 70$ getrokken van geboortegewichten in het jaar 2003 in Vlaanderen. Voor die steekproef was het gemiddelde $\bar{x} = 3247.7$ en de standaardafwijking $s = 612.03$. Er werd toen een 95 % betrouwbaarheidsinterval voor het gemiddelde geboortegewicht μ van alle baby's van 2003 opgesteld. Dat interval was [3101.8 ; 3393.6].
Voer nu, met dezelfde steekproef, een tweezijdige toets uit op het 5 % significantieniveau (gebruik je GRM) en

1. kies als nulhypothese een willekeurig getal in] 3101.8 ; 3393.6 [en voorspel of je die nulhypothese zal kunnen verwerpen. Bereken dan de p-waarde en trek je besluit.
2. kies als nulhypothese een willekeurig getal dat niet in] 3101.8 ; 3393.6 [ligt en voorspel of je die nulhypothese zal kunnen verwerpen. Bereken dan de p-waarde en trek je besluit.

Toetsen van hypothesen levert minder informatie dan betrouwbaarheidsintervallen.

- Als je de nulhypothese niet kan verwerpen dan weet je dat μ_0 een *aannemelijke* waarde is voor het populatiegemiddelde. Maar dan zal je μ_0 niet aanvaarden als "de puntwaarde" van μ. Er is immers een hele verzameling *aannemelijke* waarden van μ. Die wordt gegeven door het betrouwbaarheidsinterval.

- Als je de nulhypothese wel kan verwerpen dan weet je dat μ_0 *geen aannemelijke* waarde is voor μ. Maar je hebt dan geen idee of μ_0 ver weg ligt van de *aannemelijke waarden* of er heel dicht bij. Die informatie zou nochtans kunnen helpen om te weten of het gevonden significante verschil ook een belangrijk verschil is in de praktijk.

Opdracht 15

In een vorige opdracht heb je $H_0 : \mu = 3305$ tegenover $H_1 : \mu \neq 3305$ getoetst voor het gemiddelde geboortegewicht μ van vorig jaar. Je hebt daar ondermeer gewerkt met $n = 6000$, $\bar{x} = 3314$ en $s = 350$. Voor die gegevens vond je een p-waarde van 4.6 % zodat je de nulhypothese kon verwerpen op het 5 % significantieniveau. Dat betekent dat 3305 geen *aannemelijke* waarde is voor μ. Maar wijst dit op een belangrijk verschil? Zoek de verzameling van *alle aannemelijke* waarden in deze context en trek je besluit.

DEEL 2. Verdere begrippen (*facultatief*)

Bij een verdere studie van toetsen van hypothesen ontmoet je begrippen die belangrijk zijn als je in de praktijk een statistische studie moet opzetten. In de appendix voor de leerkracht staan voorbeelden waar deze begrippen (en hun onderlinge wisselwerking) worden geïllustreerd.

6. Het onderscheidingsvermogen

Tot nu toe heb je bij het opstellen van een toets bijna uitsluitend de nulhypothese H_0 gebruikt. Je veronderstelt dat de nulhypothese waar is om het gepaste kansmodel te kiezen. En dan wapen je je tegen de fout om H_0 te verwerpen wanneer H_0 waar is. Dat doe je door het significantieniveau klein (meestal 5 %) te houden. Dat is immers de kans op een type I fout (H_0 verwerpen als H_0 waar is). De alternatieve hypothese gebruik je om te weten of je éénzijdig (rechts of links) of tweezijdig moet toetsen.

6.1. Werken onder de alternatieve

Wat er gebeurt als de alternatieve hypothese H_1 waar is kan je ook bestuderen. Ook daar kan je een fout maken. Dat is de type II fout (H_0 niet verwerpen als H_0 niet waar is). De kans om zo'n fout te maken noteer je als β. De kans om zo'n fout <u>niet</u> te maken is **het onderscheidingsvermogen** (= *power* in het Engels).

$$
\begin{aligned}
1-\beta \text{ = onderscheidingsvermogen} \\
&= 1-P(\textit{type II fout}) \\
&= 1-P(\textit{verwerp } H_0 \textit{ niet als } H_1 \textit{ waar is}) \\
&= P(\textit{verwerp } H_0 \textit{ als } H_1 \textit{ waar is}) \\
&= P(\textit{aanvaard } H_1 \textit{ als } H_1 \textit{ waar is})
\end{aligned}
$$

Vanaf nu moet je goed opletten als je een kansuitspraak doet. Soms veronderstel je dat de nulhypothese waar is en in andere gevallen veronderstel je dat de alternatieve hypothese waar is. Om het onderscheid duidelijk te maken gebruik je in het vervolg een extra notatie:

- $P_{H_0}(\textit{gebeurtenis})$ is de kans van de gebeurtenis in de veronderstelling dat H_0 waar is
- $P_{H_1}(\textit{gebeurtenis})$ is de kans van de gebeurtenis in de veronderstelling dat H_1 waar is.

Een eenvoudig voorbeeld (zelfs al is het niet erg realistisch) helpt soms om de begrippen "significantieniveau" en "onderscheidingsvermogen" beter te begrijpen.

Veronderstel dat jij een nieuw product maakt, bijvoorbeeld een nieuw soort speculaaspasta. Als de nieuwe pasta niet beter is dan wat er momenteel op de markt is dan krijg je die niet verkocht. Zo'n pasta ligt immers al in de grote winkelketens en daar kan jij niet tegen concurreren. Maar jij bent ervan overtuigd dat je nieuwe pasta veel beter is en dat de mensen die daarom zeker zullen kopen. Dat wil je bewijzen.

In het kader van toetsen van hypothesen start je met de nulhypothese H_0 dat je pasta alleen maar even goed is als al die andere pasta's. In de alternatieve hypothese H_1 zet je wat je wil bewijzen: jouw pasta is beter.

Een type I fout bega je nu als volgt. In feite is je pasta niet beter (de nulhypothese is waar). Maar jij hebt ervan geproefd, je bent ervan overtuigd dat hij wel beter is (je verwerpt de nulhypothese) en je start de productie. Resultaat: niemand koopt en jij bent failliet. De **kans** om zoiets tegen te komen wil je klein houden, bv. 5 %. Dat is het **significantieniveau**. Daarom zet je vooraf een uitgebreide studie op. Je laat meerdere mensen proeven en oordelen. En je wacht tot er heel veel mensen zeggen dat die pasta beter is vooraleer je aan de productie durft beginnen.

Maar hoelang moet je wachten? Als je nieuwe pasta nu eens echt beter is (de alternatieve hypothese is waar) dan wil je zo snel mogelijk beginnen produceren. De studie die je opzet moet dus ook liefst met grote kans aantonen dat je pasta beter is (de nulhypothese verwerpen) wanneer hij dat echt is. Dat is het **onderscheidingsvermogen**. Het is de **kans** dat je kan aantonen dat je product beter is als het echt beter is.

6.2. Paranormale gaven ontdekken

Je hebt een vriendin die beweert dat zij paranormaal begaafd is: zij kan gedachten lezen. Om dat te toetsen doe je de volgende test. Je neemt een doos met 10 rode en 10 witte kaarten. Je trekt lukraak een kaart uit de doos en denkt aan de getrokken kleur. Dan vraag je aan je vriendin aan welke kleur je denkt. Zij staat achter een gordijn en ziet je niet. Je noteert of haar antwoord juist is (code 1) of fout (code 0). Dan leg je de kaart terug in de doos, je schudt eens goed en trekt terug een kaart. Dat herhaal je 50 keer.

Je hebt hier een 0–1 populatie waarbij je de kans op succes (kans op een juist antwoord) voorstelt als π. De nulhypothese veronderstelt dat er geen effect is en dat de antwoorden van je vriendin eruit zien als "lukraak raden". De kans op succes is dan gelijk aan 1/2 zodat $H_0 : \pi = 0.5$. Als je vriendin echt paranormaal begaafd is dan zullen er van haar antwoorden, in the long run, meer dan de helft juist zijn. De alternatieve hypothese is hier éénzijdig rechts: $H_1 : \pi > 0.5$.

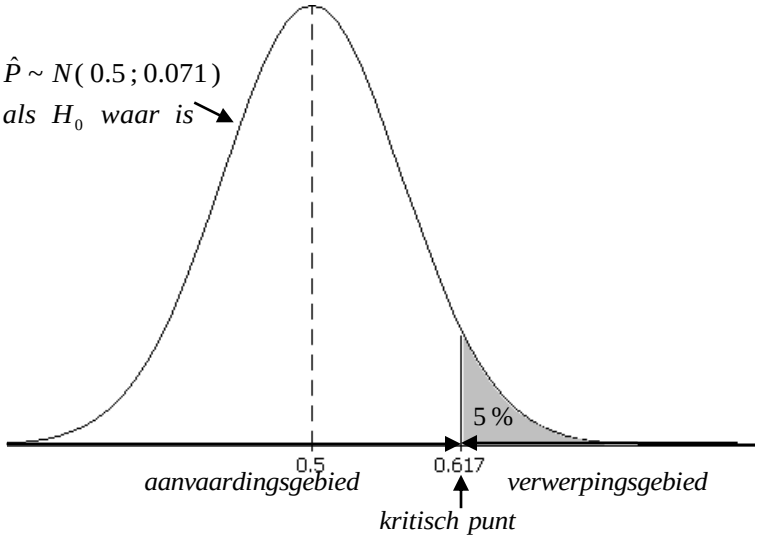
Zoals altijd begin je met de veronderstelling dat de nulhypothese waar is. Het basis-kansmodel om de populatieproportie π te bestuderen is de steekproefproportie $\hat{P}$. Onder de nulhypothese is (in algemene notatie) $E(\hat{P}) = \pi_0$ en $se(\hat{P}) = \sqrt{\pi_0(1-\pi_0)} / \sqrt{n}$. Standaardiseren leidt dan tot $\dfrac{\hat{P} - \pi_0}{\sqrt{\pi_0(1-\pi_0)} / \sqrt{n}} \sim N(0;1)$ wanneer de steekproef groot genoeg is.

Om de begrippen mooi grafisch te kunnen illustreren werk je hier even zonder te standaardiseren. Je hebt dat $n\pi_0 = (50)(0.5) = 25 \geq 15$ en $n(1-\pi_0) = (50)(1-0.5) = 25 \geq 15$ zodat je mag veronderstellen dat $\hat{P}$ normaal verdeeld is. Bovendien is $E(\hat{P}) = 0.5$ en $se(\hat{P}) = \sqrt{0.5(1-0.5)} / \sqrt{50} \cong 0.071$. Dat betekent dat $\hat{P} \sim N(0.5; 0.071)$.

Nu gebruik je dit (ongestandaardiseerde) kansmodel om een toets op te stellen. Dit betekent dat je een aanvaardings– en verwerpingsgebied bepaalt. In dit voorbeeld toets je éénzijdig rechts zodat je op zoek gaat naar een rechterstaart waarin $\hat{P}$, onder H_0, terechtkomt met kans 5 %. Als je later met je gevonden steekproefproportie $\hat{p}$ in die staart terechtkomt, dan zal je de nulhypothese verwerpen terwijl ze waar is en een type I fout maken.

Het kritische punt dat de gebieden afbakent vind je met de GRM. Druk 2nd [DISTR], kies 3:invNorm(en vul in zoals aangegeven. Het kritische punt is gelijk aan 0.617.

```
invNorm(0.95,0.5
,0.071)
      .6167846074
```



Op dit ogenblik heb je een toets opgesteld voor $H_0 : \pi = 0.5$ tegenover $H_1 : \pi > 0.5$. Dat betekent dat je vanaf nu de volgende regel hebt vastgelegd:

> Trek een steekproef van grootte $n = 50$ en bereken de succesproportie $\hat{p}$ in je steekproef. Als $\hat{p} \in [\,0.617\,;1\,]$ dan verwerp je de nulhypothese en als $\hat{p} \in [\,0\,;0.617\,[$ dan verwerp je ze niet. Dat alles gebeurt op het 5 % significantieniveau.

Als je deze regel gebruikt, hoe groot is dan je kans om te ontdekken dat je vriendin paranormaal begaafd is als ze dat echt is? Die kans is het onderscheidingsvermogen.

Om die kans te berekenen moet je wat specifieker zijn. Zeggen dat $\pi > 0.5$ geeft nog veel alternatieve mogelijkheden voor π en daarbij hoort telkens een ander onderscheidingsvermogen. Als voorbeeld neem je $H_1 : \pi = 0.7$. Dat betekent dat je vriendin, als zij zo'n test miljoenen keren zou herhalen, ongeveer 70 % juiste antwoorden zou geven. Dat is veel beter dan raden.

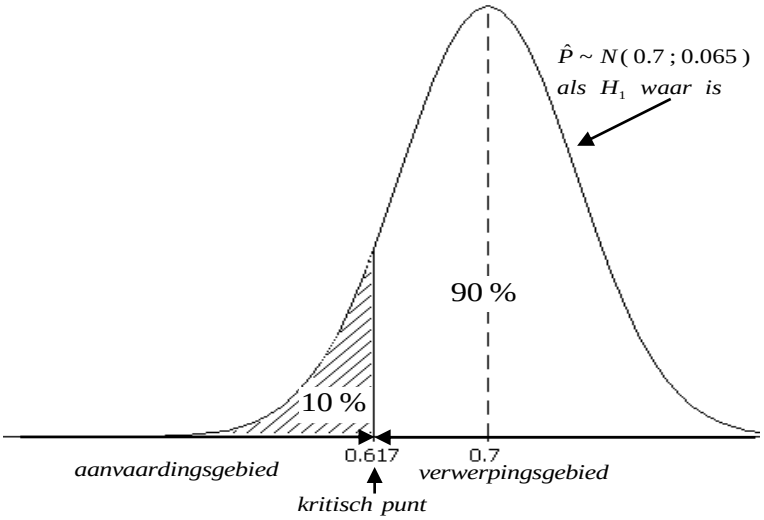
Nota.
Onder de alternatieve hypothese veronderstel je hier (en ook verder in de tekst) dat je te maken hebt met een 0–1 populatie (met $\pi > 0.5$) waaruit je een EAS trekt. Dat is zoals trekken met terugleggen uit een doos met vaste succeskans π. Of paranormale begaafdheid echt zo werkt weet je eigenlijk niet. Misschien hangt een volgend antwoord af van een vorig of is er per antwoord geen vaste succeskans. In dat geval kan je de kansmodellen waarmee je hier werkt niet gebruiken. Zo'n situaties worden in deze tekst niet behandeld.

Je kijkt nu wat er gebeurt als $H_1 : \pi = 0.7$ waar is. Ook in deze situatie mag je veronderstellen dat $\hat{P}$ normaal verdeeld is want $n\pi = (50)(0.7) = 35 \geq 15$ en $n(1-\pi) = (50)(1-0.7) = 15$. Maar nu is $E(\hat{P}) = 0.7$ en $se(\hat{P}) = \sqrt{0.7(1-0.7)} / \sqrt{50} \cong 0.065$. Dat betekent dat $\hat{P} \sim N(\,0.7\,;0.065\,)$. Als je met dit model een steekproef trekt, $\hat{p}$ berekent en in $[\,0.617\,;1\,]$ terechtkomt dan zal je, volgens de vastgelegde regel, de nulhypothese verwerpen. Dat betekent dat je de alternatieve hypothese aanvaardt terwijl ze juist is, of dat jij een statistisch bewijs hebt dat je vriendin paranormaal begaafd is als ze dat echt ook is.

De kans om in $[0.617;1]$ terecht te komen bereken je met de GRM. Druk 2nd [DISTR], kies 2:normalcdf(en vul in zoals aangegeven.

```
normalcdf(0.617,
10^99,0.7,0.065)
      .8991852314
normalcdf(0.617,
1,0.7,0.065)
      .8991832675
```

Het onderscheidingsvermogen is hier gelijk aan 90 %. De kansuitspraak die daarbij hoort ziet er als volgt uit: $P_{H_1}(\hat{P} \geq 0.617) = 90\,\%$.



Nota. Een proportie kan nooit groter dan 1 zijn en daarom kijk je of de gevonden $\hat{p}$ in $[0.617;1]$ terechtkomt. Anderzijds gebruik je de normale als benadering voor kansen waarmee $\hat{P}$ in gebieden terechtkomt. Een rechterstaart loopt dan tot $+\infty$ zodat je naar het gebied $[0.617;+\infty[$ moet kijken. Zoals je bij de GRM output merkt, is er essentieel geen verschil voor de berekende kans.

7. Vier basisgrootheden

Bij het toetsen van hypothesen zijn 4 basisgrootheden met elkaar verbonden:

- het significantieniveau: de kans om de nulhypothese H_0 te verwerpen als H_0 waar is.
- het onderscheidingsvermogen: de kans om de alternatieve hypothese H_1 te aanvaarden als H_1 waar is.
- de effectgrootte: het verschil dat je wil ontdekken (hoeveel wijkt de alternatieve hypothese af van de nulhypothese?)
- de steekproefgrootte: hoeveel elementen (dossiers, patiënten, respondenten,..) heb je nodig in je studie?

Bij je paranormaal begaafde vriendin heb je een procedure opgesteld met significantieniveau $\alpha = 5\,\%$. Je werkte met een steekproefgrootte $n = 50$. Als paranormale begaafdheid de succeskans opdrijft naar 70 % dan wil jij zo'n effect van $\pi_1 - \pi_0 = 0.7 - 0.5 = 0.2 = 20\,\%$ met grote kans kunnen ontdekken. Die kans is het onderscheidingsvermogen en die is in deze studie gelijk aan 90 %.

Bemerk dat je vooraf kijkt naar het samenspel tussen de 4 basisgrootheden van een toets. Je kan daar bepaalde eisen aan opleggen en dan een toets opstellen die aan die eisen voldoet. Dat doe je allemaal op basis van onderliggende kansmodellen. ***Je hebt nog altijd geen steekproef getrokken***.

8. Enkele TI-84 Plus commando's

In het kader van betrouwbaarheidsintervallen en toetsen van hypothesen kan het handig zijn dat je de betekenis van sommige afkortingen kent. Dat helpt om commando's gemakkelijker te onthouden.

In onze teksten hebben wij ervoor gekozen om de begrippen aan te brengen aan de hand van één kansmodel voor proporties en één kansmodel voor gemiddelden. Op die manier kan je de toegemeten tijd optimaal besteden aan "de onderliggende ideeën". Als leerlingen later een stap naar speciale gevallen of uitbreidingen moeten zetten, zal dat vanuit hun basiskennis weinig problemen opleveren. De commando's op de TI-84 Plus behandelen veel meer dan wat wij besproken hebben.

Een eerste groep van 6 commando's gaat over toetsen van hypothesen voor gemiddelden en proporties.

1:Z-Test… is een toets voor het gemiddelde μ van één populatie waarbij je veronderstelt dat de standaardafwijking σ van de populatie gekend is. Het kansmodel waarmee je dan werkt is (onder de gepaste voorwaarden) normaal verdeeld (met Z als klassieke notatie).

2:T-Test… is een toets voor het gemiddelde μ van één populatie waarbij de standaardafwijking σ van de populatie niet gekend is. Het kansmodel waarmee je dan werkt volgt (onder de gepaste voorwaarden) een t-verdeling (met T als notatie voor zo'n kansmodel).

3:2–SampZTest… is een toets om de gemiddelden μ_1 en μ_2 van twee populaties te vergelijken waarbij je veronderstelt dat de standaardafwijkingen σ_1 en σ_2 van de populaties gekend zijn. Het kansmodel waarmee je werkt is (onder de gepaste voorwaarden) normaal verdeeld. Voor deze test heb je 2 steekproeven nodig (2-Samp = two samples) en Z verwijst naar de normale verdeling.

4:2–SampTTest… is een toets om de gemiddelden μ_1 en μ_2 van twee populaties te vergelijken waarbij de standaardafwijkingen σ_1 en σ_2 van de populaties niet gekend zijn. Het kansmodel waarmee je werkt volgt (onder de gepaste voorwaarden) een t-verdeling.

5:1–PropZTest… is een toets voor de proportie π van één populatie. Het kansmodel waarmee je werkt is (onder de gepaste voorwaarden) normaal verdeeld.

6:2–PropZTest… is een toets om de proporties π_1 en π_2 van twee populaties te vergelijken. Het kansmodel waarmee je werkt is (onder de gepaste voorwaarden) normaal verdeeld.

De notatie voor betrouwbaarheidsintervallen is volledig analoog.