Statistiek
Hoorcollege 4, 30-11-2015

In dit college:
- Hypothese formuleren en schematisch toetsen dmv statistische technieken en daaruit de juiste conclusies strekken
- Samenhang tussen twee variabelen beschrijven en interpreteren alsmede kunnen bepalen wanneer dit een causaal verband betreft

(DEELTOETS!)

Tentamenvraag leerdoel 4
Een steekproef onder 53 Nederlanders over de gemiddelde lengte van een Nederlander (uitgedrukt in meters) heeft een gemiddelde van 1,75 en een variantie van 0.0025. Ga er vanuit dat de Nederlandse populatie normaal verdeeld is. Waardoor wordt een 95%-betrouwbaarheidsinterval voor μ gegeven?
Antwoord: 174 < μ < 176
- Gemiddelde lengte populatie: 1,75
- Normaal verdeeld: Z-waarde
- Variantie: 0.0025
- Standaardafwijking: $\sqrt{0.0025}$ = 0,05
- 95% betrouwbaarheidsinterval invullen:
  $P(\bar{x} - 1.96\frac{Standaardafwijking}{}$ SLIDE

Zie slide voor de berekening van een betrouwbaarheidsinterval van het populatiegemiddelde.

Bound on the error of estimation: Formule zie slide 11
Sample size to estimate a mean: Formule zie slide 11

Testen van hypotheses: Centrale ideeën

Hypotheses: statements waar je vanuit gaat dat deze waar is. Als er genoeg bewijs is wordt deze aangenomen, anders verworpen
Nul hypothese: H_0 = … (testprocedure begint met de aanname dat deze hypothese klopt)
Alternatieve hypothese: H_1 = … (bij genoeg bewijs wordt deze hypothese aangenomen)

Er zijn twee mogelijkheden
- Er is genoeg bewijs voor de alternatieve hypothese
- Er is niet genoeg bewijs voor de alternatieve hypothese

Er zijn twee mogelijkheden foute:
- Type 1: Verwerpen als H_0 klopt, of
- Type 2: niet verwerpen als H_0 niet klopt.

Voor voorbeeld zie slide 16.

Rejection region: kritiek gebied, range aan variabelen, als de test statistic hier in valt verwerpen we H_0 hypothese.
Two-tailed test (tweezijdige test): Hierbij kan de nul hypothese bij kleiner en bij groter verworpen worden
One-tailed test (eenzijdige test): H_1 zegt dat het gemiddelde groter is dan het gemiddelde van de nul hypothese, of H_1 zegt dat het gemiddelde kleiner is dan het gemiddelde van de nul hypothese.

6 stappen bij het gebruik van de rejection region:
1. Schrijf de 0 en alternatieve hypothese op
2. Kies het niveau van siginicantie α (meestal 5%)

3. Bereken de test statistiek (T waarde) en geef de distributie van de test statistiek aan wanneer de 0 hypothese waar is.
4. Bereken de kritieke waarden van de rejection region
5. Bekijk of de teststatistiek in de kritieke waarde valt. Zo ja, verwerp de nul hypothese
6. Vertel dit in woorde

Voor voorbeeld: zie slide 24 en 25.

Z-waarde hou je uit de tabel 1, hier ga je op zoek naar 95%

Gebruik deze formule: $z = \frac{\bar{X} - \mu}{\frac{Standaardafwijking}{\sqrt{n}}}$

De p-waarde methode

De p-waarde van een test is de kans dat je een teststatistiek vindt die zo extreem is, is gelijk aan jou p-waarde. Een p-waarde <1% is een groot bewijs voor de alternatieve hypothese. Een p-waarde tussen 1 en 5% is een sterk bewijs, een p-waarde van 5 en 10% is zwak bewijs voor de alternatieve hypothese, en een p-waarde>10% onderbouwd de alternatieve hypothese helemaal niet.

6 stappen bij het berekenen van de p-waarde:

1. Formuleer4 de nul en alternatieve hypothese
2. Kies het significantieniveau α
3. Bereken de test statistiek (T waarde) en geef de distributie van de test statistiek aan wanneer de 0 hypothese waar is.
4. Bereken de p-waarde
5. Bekijk of de p-waarde kleiner is dan α
6. Formuleer de conclusie in woorden

Betrouwbaarheidsinterval bij tweezijdige toets:
Als het buiten betrouwbaarheidsinterval valt, verwerp je de nul hypothese ook.

6 stappen bij een hypothesis test

1. Formuleer de nul en alternatieve methode
2. Kies het significantieniveau
3. Breken de test statistiek en geef de distributie van de test staistiek aan waar de nul hypothese waar is
4. Bereken de kritieke waardes, de rejection region, de p-waard OF de confidence interval (alleen bij een tweezijdige toets)
5. Bekijk of de test statistiek valt in de rejection region of bekijk of de p-waarde kleiner is dan a, of buiten de Confidence interval ligt. Zo ja, verwerp de nul hypothese.
6. Formuleer de conclusie in woorden.

Voor voorbeeld zie slide 33 en 34.

3 benaderingen voor het testen van een hypothese:

1. Rejection region approach: bereken de rejection region en bepaal of de test statistiek in deze regio ligt
2. P-value approach: bereken de waarde van de test statistiek, vind de kans dat de test statistiek ten minste zo extreem is als die waarde, en vergelijk de p-waarde met a.
3. Confidence interval approach: Bereken de 1-a confidence interval en bepaal of het gemiddelde vna de nul hypothese in het interval valt. Dit geldt alleen bij een tweezijdige toets.

Je kunt de standaardafwijking schatten door de sample standard deviation

$$t_{n-1} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Dit is een Students's t distributie met v = n-1 mate van vrijheid. Er zijn twee onzekere objecten, waardoor er meer informatie nodig is om een hypothese te verwerpen. Kijk in tabel 8 voor de kritieke waarden voor $t_{n-1,a/2}$

Als de sample standaardafwijking bekend is in plaats van de populatiestandaardafwijking, kun je de t-test gebruiken. De test statistic is dan: T = $\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$.

Simple linear regressie analyse
In veel gevallen zijn we geïnteresseerd in de vraag of een specifieke random variabele (x) invloed heeft op een andere random variabele (Y). Dit kun je leiden uit een regressie analyse.

Elk van deze relaties kan uitgedrukt worden in een lineair model. Probabilistic model: een (kans)model waar de willekeurigheid dat deel is van een levensecht proces in berekend is.

First-Order Lineair Model

De simple lineair population regression model is een model met een rechte lijn met één onafhankelijke variabele
Het wordt gegeven door: y = SLIDE
Waar:

- Y = afhankelijke variabele (interval variabele)
- X = onafhankelijke variabele (interval variabele)
- β_0 = y-intercept
- β_1 = slope of the line
- E = error variabele (random variabele)

De restfactoren zijn normaal verdeeld met een gemiddelde van 0 en een variantie.

Schatten van de coefficienten
Het doel is om de relatie tussen de onafhankelijke en afhankelijke variabele X en Y te analyseren. Hiervoor moeten we β_0 en β_1 schatten. Voor de formules hiervan zie slide 46.