Statistiek
Hoorcollege 6, 14-12-2015

Regression analyse using Excel
Excel: options →add-ins → Excel add-ins → Analysis ToolPak (om statistische analyses uit te voeren)
Regressie analysis: Input X range, input Y range, confidence Level bijvoorbeeld 90%.
Residuals: extra opties
Zie slide 5 voor een samenvatting van de uitvoer van de regressie analyse.

Inference on about a population variance
In plaats van het kijken naar een centrale locatie, kunnen we ook kijken naar de variabiliteit van een populatie. Toepassingen:

- Mate van risico
- Consistentie van een productie proces

De populatie variantie is σ^2.
S^2 = steekproef variantie (daarom moet je delen door n-1!)
De vraag is wat je kunt zeggen over de populatie wanneer de variantie bekend is.

Hypothesis testing regarding the population variance
Om de hypotheses over de populatie variantie te testen gebruiken we de **chi-squared test statistic**
De random variabele $x^2 = \frac{(n-1)s^2}{\sigma^2}$ is chi-squared gedistrubeerd met v=n-1 graden vrijheid, zolang de steekproefpopulatie normaal is.
Dit verschilt van de z-statistiek en de t-statistiek door dat de uitkomst alleen positief kan zijn en ook is de chi-squared distributie niet symmetrisch. Tabel 7 geeft de waarden van de chi-square voor verschillende maten van vrijheid en significantie niveaus.
Voorbeeld: zie slide 9-12.

Kritieke waarde $x^2_{\sigma;n-1}$
Belangrijk: net zoals bij z-testen en t-testen gebruiken we de waarde van de relevante parameter gegeven in H_0

Observational and experimental data
We onderscheiden twee soorten data: observational data en experimental data
Observational data: De data is niet onder controle van de onderzoeker (geen causaal verband)
Experimental data: De data is wel onder controle van de onderzoeker, de onderzoeker draait aan de knoppen (causaal verband)

Voorbeeld: effecten van wijn, zie slide 19 en 20.

Experimental data
Je verdeelt de personen in twee groepen:

- **Treatment Group** (deze mensen laat je bijv. 2 glazen wijn per dag drinken)
- **Control Group** (deze mensen mogen niet drinken)

Je zorgt dat beide groepen gelijk zijn in alle andere aspecten, hiervoor controleer je andere eigenschappen zoals eetgewoontes of levensstijl. Als je een significant verschil tussen gemiddelde ziet, is het aannemelijk dat het verschil kan toegeschreven worden aan een bepaalde oorzaak (in het voorbeeld dus wijn), en hierdoor kun je stellen dat wijn een direct effect heeft op de bloeddruk.

Matched pairs experiments

Een **matched pairs experiment** is een type van een experimenteel design: verschillende testpersonen aan elkaar koppelen op basis van observeerbare karakteristieken.
Je neemt twee onafhankelijke steekproeven. Categoriseer de onderwerpen onder sommige andere dimensies. Je matcht een onderwerp in één steekproef met een onderwerp in een andere steekproef, volgens jou categorie. Deze benadering verminderd variatie in de data.

<u>Voorbeeld matched pairs experiment: zie slide25-27</u>

De parameter van interest is het gemiddelde van de populatie van verschil: μ_D. Om hier over hypotheses te kunnen testen gebruiken we de volgende test statistiek: t = $\frac{\bar{X}_D - \mu_D}{S_D/\sqrt{n}}$.
Deze test statistiek is t-gedistributeerd met n-1 graden van vrijheid.

Sampling distribution of the difference between two means

Het verschil $\bar{X}_1 - \bar{X}2$ is normaal gedistributeerd als beide populations normaal en ongeveer normaal als de populaties niet-normaal is en de steekproefgroottes groot zijn.
De verwachte waarde hiervan is: E($\bar{X}_1 - \bar{X}2$) = $\mu_1 - \mu_2$
De variantie van $\bar{X}_1 - \bar{X}2$ is (bij onafhankelijke steekproeven): V($\bar{X}_1 - \bar{X}2$) =$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
De standaardafwijking hiervan is: $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Equal variances

De test statstic voor $\mu_1 - \mu_2$ wanneer $\sigma_1 \neq \sigma_2$ is gegeven door de formule:
t = $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$ waarin $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$ en v = $n_1 + n_2 - 2$ graden vrijheid.
s_p^2 wordt de **pooled variance estimator** genoemd. Dit is het gewogen gemiddelde van de twee steekproef varianties.

Unequal variances

De test statstic voor $\mu_1 - \mu_2$ wanneer $\sigma_1 \neq \sigma_2$ is gegeven door de formule:
t = $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$ waarin v = **<u>Zie slide 30 voor formule.</u>**

Confidence intervals

Voor formules: **<u>zie slide31</u>**

The F-distribution

Voor formules: **<u>zie slide33</u>**