

Projekt 8.5 Hypotesetest med anvendelse af t -Test

(Dette materiale har været anvendt som forberedelsesmateriale til den skriftlige prøve 2012 for netforsøget)

Indhold

Indledning	1
χ^2 -test	2
Numeriske data og hypotesetest	2
Middelværdi, varians og spredning	3
Stokastiske variable og empiriske størrelser	4
Normalfordelte observationer	5
Hypotesetest med et datasæt (en variabel)	7
Anvendelse af t -test for et datasæt (en variabel) i <i>Muffins</i>	9
Fordeling af t -teststørrelsen	10
Hypotesetest med to datasæt (to variable)	12
Anvendelse af t -test for to datasæt (to variable) i <i>Muffins</i>	14
Kilder	16

Indledning

I statistik opstilles hypoteser for variable i en stikprøve. Dette forberedelsesmateriale omhandler to tests, som kan bruges, når du skal vurdere om en hypotese skal forkastes eller accepteres. Det ene test er χ^2 -testet, som du allerede kender, og det andet test er t -testet. Materialet indeholder to dele. I første del genopfrisker vi centrale dele ved χ^2 -testet, mens vi i anden del introducerer t -testet og anvender dette til at besvare spørgsmål, som kan være interessante at stille, når vi betragter et datamateriale.

Datamaterialet som bruges i denne sammenhæng kaldes *Muffins*¹. Det indeholder en stor mængde rådata, som stammer fra en spørgeskemaundersøgelse blandt 539 unge tyskere. Vi anvender et større udklip herfra, som findes i den vedlagte fil: [MuffinsUdklip.xls](#).

Når man skal håndtere store datasæt, er det vigtigt at kunne sortere datamaterialet, således at de sammenhænge, man ønsker at undersøge, træder tydeligt frem, og data bliver tilgængeligt i den rigtige form. Fx kunne det i nogle tilfælde være interessant at sortere datamaterialet efter køn, således at pigernes svar er adskilt fra drengenes svar. En sådan sortering vil være aktuel, når man skal undersøge sammenhængen mellem køn og en anden *kategorisk* variabel. Man optæller så blot pigernes og drengenes svar i relation til den anden kategoriske variabel. Denne optælling kan foretages automatisk i et værktøjsprogram. Tilsvarende vil det være hensigtsmæssigt i behandling af datamaterialet at kunne opstille en krydstabel automatisk i et værktøjsprogram.

Datamaterialet samt din bearbejdning af dette skal medbringes til eksamen. Du skal under prøven kunne importere data fra et regneark af samme slags som [MuffinsUdklip.xls](#) og bearbejde disse i dit værktøjsprogram.

Opgaver indenfor χ^2 -test hører under kernestoffet og er derfor en naturlig del af den skriftlige prøves 2. delprøve. Derudover vil den skriftlige prøves 2. delprøve indeholde opgaver, der handler om anvendelse af t -test.

¹ *Medien- und Freizeitgestaltung für interessanten Stochastikunterricht* af Rolf Biehler, Klaus Kombrink, Stefan Schweynoch, 2003. Se eventuelt <http://www.mathematik.uni-kassel.de/~fathom/projekt.htm>.

χ^2 -test

I forbindelse med χ^2 -test har du arbejdet med datasæt og krydstabeller fx fra en spørgeskemaundersøgelse med to eller flere kategoriske variable. Vi genopfrisker i følgende eksempel sammensat af en række øvelser anvendelsen af χ^2 -uafhængighedstestet, hvor vi samtidigt får et indblik i data fra *Muffins*.

Åben [MuffinsUdklip.xls](#), og åben fanebladet *Udklip1* heri. I *Udklip1* er der to kategoriske variable: *Køn* og *Hvor ofte man går i kirke*.

Øvelse 1: Sortering

Hvor mange piger og drenge er der i udklippet? Sorter data efter køn og tæl op automatisk i dit værktøjsprogram.

Øvelse 2: Krydstabel

Nedenfor ses en krydstabel for de to nævnte variable fra Øvelse 1 (her vist i Excel):

Antal af	Hvor ofte man går i kirke	Kolonnenavne					
Rækkenavne		aldrig	en_eller_to_pr_Måned	en_gang_pr_uge	mere_end_1_pr_uge	sjældent	Hovedtotal
dreng		125	8	20	15	45	213
pige		122	13	22	18	74	249
Hovedtotal		247	21	42	33	119	462

Opstil nu selv med udgangspunkt i *Udklip1* krydstabellen ovenfor vha. dit værktøjsprogram.

Øvelse 3: Stikprøve & population

Gør rede for, hvilken af betegnelserne *stikprøve* og *population* der kan anvendes om datasættet i *Udklip1*.

Øvelse 4: Nulhypotese

Vi ønsker at undersøge, om der er sammenhæng mellem køn og hvor ofte man går i kirke (kaldes for overskuelighedens skyld nu *kirkegang*).

Opstil en nulhypotese, som kan anvendes til at undersøge, om der er sammenhæng mellem køn og kirkegang.

Øvelse 5: Forventede værdier

Bestem vha. dit værktøjsprogram de forventede værdier under forudsætning af at din nulhypotese er sand, og opstil dem i en krydstabel svarende til krydstabellen for de observerede værdier.

Øvelse 6: Kritisk værdi og p -værdi i χ^2 -testet

Beregn χ^2 -teststørrelsen, og benyt dit værktøjsprogram til at bestemme den kritiske værdi svarende til et 5% signifikansniveau.

Gør rede for, om nulhypotesen skal forkastes eller accepteres på baggrund af χ^2 -teststørrelsen og den kritiske værdi.

Gør rede for, at p -værdien sammenlignet med signifikansniveauet giver samme resultat.

Bemærk: Nogle værktøjsprogrammer kan udføre dette χ^2 -test automatisk ud fra de observerede værdier i krydstabellen.

Numeriske data og hypotesetest

Et datasæt kan også indeholde *numeriske data*, og disse skal behandles på en anden måde end de kategoriske. Hver søjle i et numerisk datasæt (tabellen med alle observationer) repræsenterer i den sammenhæng *målinger* af en variabel og i datasættet *Muffins* kan dette for eksempel være respondenternes højde.

Kategoriske og numeriske variable

Når vi analyserer et datasæt bestående af *kategoriske data*, sammenligner vi *antallet* af observationer i hver *kategori* på tværs af de variable.

Når vi analyserer et datasæt bestående af *numeriske data*, tager vi udgangspunkt i de

såkalde *empiriske størrelser*: *middelværdi* og *varians*, og sammenligner hver af disse på tværs af de variable.

For at illustrere forskellene på numeriske og kategoriske data betragter vi to forskellige datasæt. I afsnittet ovenfor om χ^2 -testet, så vi på kategoriske data, hvor vi *optalte* antal piger og drenge efter hvor ofte de går i kirke. I Tabel 1 nedenfor har vi *målt* højden af en række piger og drenge – og altså ikke optalt, antallet af drenge og piger med en bestemt højde.

	A	B		A	B
1	Køn	Højde	214	dreng	1.97
2	dreng	1.7	215	pige	1.64
3	dreng	1.65	216	pige	1.58
4	dreng	1.71	217	pige	1.6
5	dreng	1.86	218	pige	1.6
6	dreng	1.7	219	pige	1.62
7	dreng	1.7	220	pige	1.6
8	dreng	1.7	221	pige	1.61
9	dreng	1.76	222	pige	1.64
10	dreng	1.7	223	pige	1.6
11	dreng	1.75	224	pige	1.61
12	dreng	1.75	225	pige	1.63
13	dreng	1.84	226	pige	1.68

Tabel 1: Køn og Højde fra *Udklip2*, som består af datasættene for de to variable fra *Udklip1*.

I både afsnittet ovenfor og i Tabel 1 er der imidlertid tale om stikprøver. Vi har udtaget et – forhåbentligt – repræsentativt udsnit af den population, som vi ønsker at vide noget om, og ud fra behandling af disse udtaler vi os om bestemte egenskaber ved hele populationen.

Øvelse 7: Diagrammer skaber overblik

Konstruer relevante diagrammer for data fra både eksemplet med kategoriske data og fra Tabel 1 med numeriske data, og diskuter, hvad diagrammerne viser om de variable, der indgår. Data svarende til Tabel 1 finder du i [MuffinsUdklip.xls](#) under fanebladet *Udklip2*. Du kan vælge boksplot, søjlediagrammer, prikplot, histogrammer eller andre relevante figurer evt. opdelt efter køn.

Middelværdi, varians og spredning

I Tabel 1 kan vi ikke sammenligne kønnene med de redskaber, vi har fra χ^2 -testet, da søjlerne indeholder *numeriske data*. Vi får brug for en anden metode, og her anvender vi som nævnt størrelserne *middelværdi* og *varians*. Eller rettere sagt: De *bedste bud* vi ud fra Tabel 1 kan give på middelværdi og varians for højdemålingerne i stikprøven. Disse *bedste bud* kaldes de *empiriske størrelser* for *middelværdi* og *varians*, og de betegnes med de to matematiske symboler: \bar{x} (empirisk middelværdi) og s^2 (empirisk varians).

Empirisk middelværdi, varians og spredning

De empiriske størrelser \bar{x} og s^2 kan beregnes med formlerne:

$$\bar{x} = \frac{1}{n} \sum_i x_i = \frac{x_1 + x_2 + x_3 + x_4 + \dots + x_n}{n}$$

og

$$s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2,$$

hvor $x_1, x_2, x_3, x_4, \dots, x_n$ betegner observationerne, dvs. de værdier som variabelen antager. Ud fra variansen kan vi også bestemme *spredningen*, som er kvadratroden af variansen:

$$s = \sqrt{s^2}.$$

Vi har indført sumtegnet \sum her for at forenkle opskrivningen, men de fleste værktøjsprogrammer kan bestemme de empiriske størrelser direkte ud fra observationerne.

I Tabel 1 svarer $x_1, x_2, x_3, x_4, \dots, x_n$ til de observerede højdemålinger, hvor indiceringen angiver målingens nummer i datasættet. n angiver det samlede antal målinger, som er henholdsvis 213 for drengene, og 249 for pigerne.

Øvelse 8: Middelværdi, varians og spredning

Benyt dit værktøjsprogram til at bestemme de empiriske størrelser for hhv. drengenes og pigernes højde i Tabel 1. Data finder du som tidligere nævnt i *Udklip2*.

Stokastiske variable og empiriske størrelser

Forskellen på den *empirisk middelværdi* og den værdi, som vi i det følgende vil kalde den *sande middelværdi*, kan illustreres ved at betragte datasættet i Tabel 2 nedenfor, som viser højdemålingerne for 5 drenge fra *Udklip2*.

dreng	1.87
dreng	1.88
dreng	1.9
dreng	1.96
dreng	1.8

Tabel 2: Uddrag af højdemålinger for 5 drenge i *Muffins*.

Betragter vi for eksempel højdemålingen 1,9 meter, kan man gøre sig følgende overvejelse:

Hvor præcist blev denne højde målt?

Forestiller vi os, at fem personer – en af gangen – måler denne drengs højde, vil de fem personers målinger efter al sandsynlighed variere en smule. En vil måske måle højden til 1,895 m, en anden 1,900 m osv. Pointen her er, at denne variation ikke giver os mulighed for at udtale os om, hvad drengens *sande* højde er. Vi accepterer dog, at vi har et rimeligt godt bud på den *sande* højde. Når vi beregner middelværdi og spredning som ovenfor, så giver vi et *empirisk* bud på middelværdien for højden af henholdsvis piger og drenge i *Udklip2* – som jo ikke nødvendigvis er fuldstændigt entydigt korrekt. Derfor skelner vi mellem begreberne *empiriske* og *sande* størrelser, hvor de empiriske er knyttet til de faktiske observationer, mens de *sande* er teoretiske værdier.

Stokastisk variabel

I statistikkens verden kaldes de empiriske størrelser *repræsentationer* for de enkelte højdemålinger. Man indfører begrebet *stokastisk variabel*, som repræsentant for observationerne.

Den stokastiske variabel betegnes X .

I relation til Tabel 1 og 2 taler man om den stokastiske variabel X_i , som repræsentant for højdemålingen x_i . Der forekommer som nævnt en (tilfældig) variation i det konkrete udfald x_i , men med den stokastiske variabel kan vi beskrive disse variationer.

De værdier, som vi ovenfor har kaldt de *sande værdier*, betegnes matematisk som følger:

Den sande middelværdi: μ . Den sande varians: σ^2 . Den sande spredning: $\sigma = \sqrt{\sigma^2}$.

Normalfordelte observationer

For at vi kan komme nærmere en behandling og vurdering af middelværdien for højdemålingerne bag udsnittet i Tabel 2, kræves det, at højdemålinger følger en kendt fordeling. Dette krav indebærer, at der er en bestemt systematik i observationerne og at den stokastiske variabel, der ligger bag målingerne, kan beskrives systematisk med en *matematisk model* – her *normalfordelingen*.

Når man undersøger om et datasæt er *normalfordelt* gør man det grafisk med *et fraktildiagram*.

I et fraktildiagram plotter man datasættets målinger på førsteaksen og de forventede værdier, under forudsætning af at datasættet er normalfordelt, på andenaksen. Hvis datasættet rent faktisk er normalfordelt, så vil alle målingerne derfor ligge præcist på en ret linje.

Disse forventede værdier beregnes ved at anvende den omvendte funktion til *standardnormalfordelingens* tæthedsfunktion på $\frac{i}{n}$, hvor i betegner observationens nummer på den sorterede liste (her sorteret efter stigende højde), og n betegner det samlede antal observationer i datasættet.

Øvelse 9: Standardnormalfordelingen

Hvad kendetegner standardnormalfordelingen? Søg fx oplysningerne på internettet.

Øvelse 10: Fraktildiagram

Sorter drengenes højde efter størrelse (stigende) i *Udklip2* ved hjælp af dit værktøjsprogram. Opret to nye kolonner, som skal indeholde henholdsvis observationsnummer og den forventede værdi for hver af observationerne. De forventede værdier for højdemålingerne beregnes med værktøjsprogrammets indbyggede facilitet dertil². Her er det vist i Excel, hvor $i = 1$ og $n = 213$, og hvor 0 og 1 er henholdsvis middelværdi og spredning i standardnormalfordelingen.

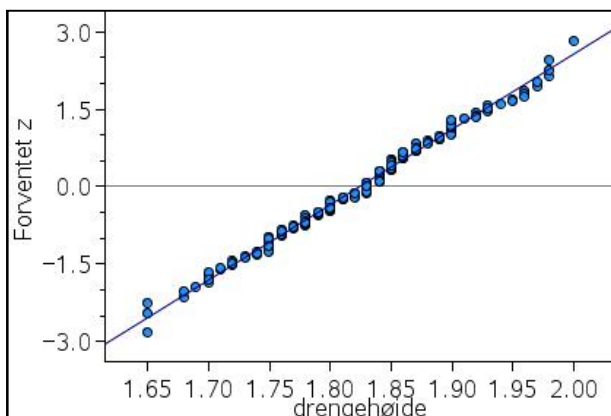
=NORMINV(1/213;0;1)	Resultat:
NORMINV(sandsynlighed; middelværdi; standardafv)	-2,59753

Fortsættes beregningerne for $i = 2$, $i = 3$ osv. dvs. for $\frac{i}{n} = \frac{2}{213}$, $\frac{i}{n} = \frac{3}{213}$ osv., så bliver de første 6 punkter i fraktildiagrammet som følger:

$$(1,65; -2,60), (1,65; -2,35), (1,65; -2,19), (1,68; -2,08), (1,68; -1,99), (1,69; -1,91).$$

Plot nu fraktildiagrammet med højdemålingerne på førsteaksen og de forventede værdier på andenaksen. Plot også den bedste rette linje (regressionslinjen) gennem punkterne. Hvis punkterne ligger tilnærmelsesvist på en ret linje, så kan vi opfatte højdemålingerne som normalfordelte. Hvis punkterne derimod afviger systematisk fra en ret linje, så må man lede efter en anden model for fordelingen. Typisk gælder det, at jo flere målinger, des mindre variation omkring en ret linje vil man forvente, og punkterne i et fraktildiagram må gerne sno sig omkring linjen. For drengene ser det ud som på Figur 1 nedenfor.

² I mange værktøjsprogrammer kaldes denne omvendte normalfordelingstæthedsfunktion for norminv(...) eller invnorm(...).



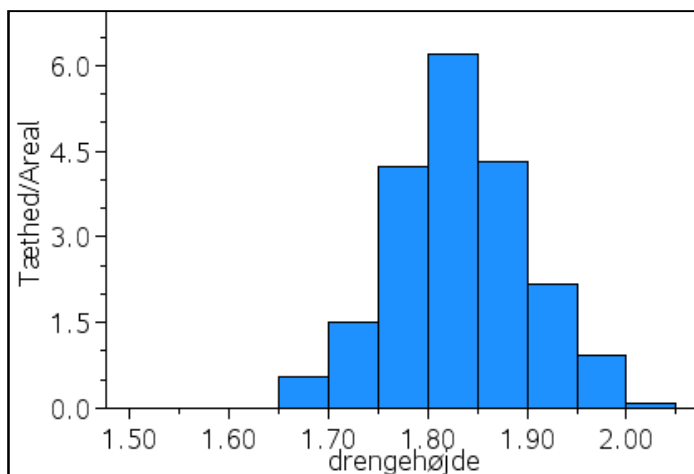
Figur 1: Fraktildiagram for fordelingen af drengenes højdemålinger, som angivet i Tabel 1.

De fleste værktøjsprogrammer kan dog lave fraktildiagrammer direkte, uden at du behøver at udregne de forventede værdier.

Fraktildiagram
 Plot af punkter (*observeret værdi, forventet værdi*) for et datasæt, hvor den forventede værdi er baseret på standardnormalfordelingen.
 Hvis punkterne tilnærmelsesvist ligger på en ret linje, kan datasættet betragtes som værende normalfordelt.

En mindre præcis måde at vurdere på om et datasæt er normalfordelt, er ved at gruppere data og konstruere et søjlediagram. Er datasættet normalfordelt, vil man forvente at observationerne fordeler sig symmetrisk omkring den empiriske middelværdi, og at der er flest observationer tæt på middelværdien, mens der er færrest længst væk fra middelværdien. Fordelingen af observationerne siges at være *klokkeformet*.

For alle højdemålingerne for drengene, som ligger bag udklipet i Tabel 2, ser den grupperede grafiske fremstilling (med intervalbredde 5 cm) ud som i Figur 2:



Figur 2. Søjlediagram for fordelingen af alle højdemålingerne for drengene angivet i Tabel 1.

Fraktildiagrammet og søjlediagrammet i Figur 1 og 2, giver god grund til tro, at drengenes højder er normalfordelte, og vi kan derfor gå ud fra, at fordelingen af højdemålingerne kan beskrives ved en normalfordeling. Vi siger også, at den stokastiske variabel X , der ligger bag data er normalfordelt, og vi skriver med matematisk notation: $X \sim N(\mu, \sigma^2)$. Med denne notationen fortæller vi samtidigt, at fordelingen er bestemt ud fra de to *parametre*, μ og σ^2 . Det er jo disse parametre, vi *estimerer*, og vi vil i næste afsnit undersøge dem nærmere.

Øvelse 11: Fraktildiagram og søjlediagram

Sortér pigernes højdemålinger i *Udklip2* efter størrelse, og undersøg vha. et fraktildiagram og et søjlediagram, om pigernes højder kan antages at være normalfordelte.

Der vil til den skriftlige prøve ikke blive stillet krav om fremstilling af fraktildiagrammer.

Hypotesetest med et datasæt (en variabel)

Hvis vi ønsker at undersøge, om de empiriske størrelser knyttet til et datasæt stemmer overens med andre undersøgelers bestemmelse af samme (oplyst værdi), kan vi gøre dette med et t -test. Som i χ^2 -test opstilles og undersøges en nulhypotese, som typisk formuleres som følger: *Middelværdi i datasæt og den oplyste værdi er ens*. Antag, at danske drenge i alderen 16-18 år i gennemsnit er 1,82 m høje. Hvis vi fx ønsker at undersøge, om gennemsnitshøjden for de tyske drenge i *Muffins* er den samme som de danske drenges gennemsnitshøjde, opstiller vi nulhypotesen:

De tyske drenges gennemsnitshøjde er 1,82 m

Intuitivt kan man undersøge nulhypotesen ved at betragte *forskellen* mellem den empiriske middelværdi, $\bar{x} = 1,824$ (som du fandt i Øvelse 8), og de danske drenges højde $\mu_0 = 1,82$ m. Men vi må også tage datasættets spredning i betragtning. Hvis spredningen i datasættet er stor, så vil man acceptere en større margin på forskellen mellem de to middelværdier \bar{x} og μ_0 . Er spredningen derimod lille, så vil kun en mindre margin blive accepteret.

Man kan vise at den empiriske middelværdi, \bar{x} , er normalfordelt med spredning $\sqrt{\frac{s^2}{n}}$, og det er den værdi, vi skal dividere forskellen mellem middelværdierne med, således at spredningen tages i betragtning:

t -teststørrelse for et datasæt

Den generelle formel til beregning af t -teststørrelsen er:

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Da \bar{x} er normalfordelt, kan man vise at t -teststørrelsen er:

t -fordelt med $f = n - 1$ frihedsgrader

og vi skriver:

$$t \sim t_{n-1}$$

Nulhypotesen kan nu formuleres som:

H_0 : *Middelværdien for de tyske drenges højde er 1,82 m, dvs. $\mu_0 = 1,82$.*

t -teststørrelsen kan på baggrund af udregningerne fra Øvelse 8 og bemærkningen ovenfor beregnes til:

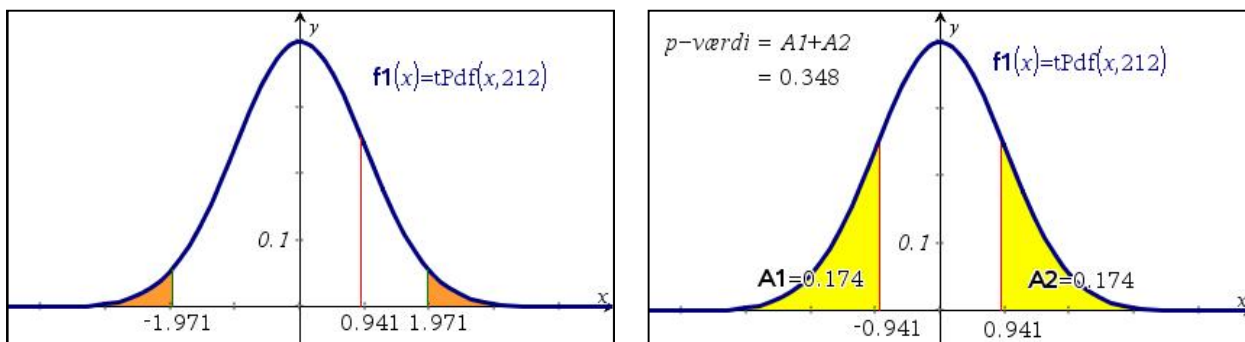
$$t \cong 0,941$$

Vi har nu brug for at vurdere om teststørrelsen 0,941 er *kritisk* i forhold til vores nulhypotese. t -fordelingen har et meget nært slægtskab med normalfordelingen. Vi får brug at kunne vurdere teststørrelsen i forhold til en kritisk værdi, der kan fortælle, om vi skal forkaste nulhypotesen på et givet signifikansniveau. Som ved χ^2 -testet bruger vi den funktion, som teststørrelsen er knyttet til, her t -fordelingen med $n - 1$ frihedsgrader, til at bestemme den sandsynlighed, der svarer til teststørrelsen. Den sandsynlighed er ligesom i χ^2 -testet netop et udtryk for, hvor svært/let det er, at frembringe en teststørrelse, der er mindst lige så stor som den observerede – altså p -værdien! Grafisk set er p -værdien ligesom ved χ^2 -testet med stor tilnærmelse givet ved arealet af det område, som grafen for fordelings tæthedsfunktion afgrænser sammen med førsteaksen i intervallet afgrænset af teststørrelsen. t -testet er i modsætning til χ^2 -testet dog to-sidet, dvs. området er to-delt – en del til højre for den positive værdi af teststørrelsen og en del til venstre for den negative værdi af teststørrelsen (se Figur 3 nedenfor).

Vi kan også vælge at sammenligne teststørrelsen direkte med den kritiske værdi, ligesom i forbindelse med χ^2 -testet. De kritiske værdier kan beregnes med et værktøjsprogram.

På Figur 3 nedenfor ses grafen for tæthedsfunktionen for t -fordeling med $f = 213 - 1 = 212$ frihedsgrader.

På figuren til venstre ses også plot af teststørrelsen 0,941 og den kritiske værdi 1,97 svarende til et signifikansniveau på 5% (dvs. der afskæres 2,5% i hver side). På figuren til højre ser vi, at arealet af det skraverede (gule) område afgrænset af grafen og førsteaksen hhv. til venstre for den negative værdi og til højre for den positive værdi af teststørrelsen er 34,8%, hvilket netop svarer til p -værdien.



Figur 3: t -fordelingen, med 212 frihedsgrader. Til venstre ses teststørrelsen $t = 0,941$ og den kritiske værdi 1,97. Til højre ses den grafiske beregning af p -værdien.

Det giver os, under forudsætning af at nulhypotesen er sand, anledning til, *ikke at forkaste nulhypotesen* på signifikansniveau 5%, da p -værdien er 0,348, dvs. sandsynligheden for at finde en teststørrelse, der er mindst lige så stor som den observerede er 34,8%.

Vi kan naturligvis også beregne p -værdien (her vist i Excel):

<code>=TFORDELING(0,941;212;2)</code>	Resultat:
<code>TFORDELING(x; frihedsgrader; haler)</code>	0,347776

Vi kan også beregne den kritiske værdi for t -testet på et 5% signifikansniveau med 212 frihedsgrader (her vist i Excel):

<code>=tinv(0,05;212)</code>	Resultat:
<code>TINV(sandsynlighed; frihedsgrader)</code>	1,971217

dvs. en kritisk værdi på 1,971, altså er vores teststørrelse på 0,941 *mindre end* den kritiske værdi, og vi kan således konkludere, at vi ikke kan forkaste nulhypotesen om, at gennemsnittet af de tyske drenges højde er 1,82 m.

p -værdi og kritisk værdi

Hvis p -værdien er mindre end signifikansniveauet, forkastes nulhypotesen.

Hvis t -teststørrelsen er positiv og er større end den kritiske værdi ved det valgte signifikansniveau, så forkastes nulhypotesen.

Hvis t -teststørrelsen er negativ og er mindre end den negative kritiske værdi ved det valgte signifikansniveau, så forkastes nulhypotesen.

Anvendelse af *t*-test for et datasæt (en variabel) i *Muffins*

I et udklippet *Udklip1* finder vi også andre variable, som vist i Tabel 3 nedenfor.

	A	B	C	D	E
1	Køn	Hvor ofte man går i kirke	Alder	Højde	Vægt
2	pige	aldrig	16	1.64	40
3	pige	sjældent	16	1.58	42
4	pige	aldrig	17	1.6	44
5	pige	aldrig	16	1.6	45
6	pige	aldrig	17	1.62	45
7	pige	sjældent	17	1.6	46
8	pige	aldrig	17	1.61	47
9	pige	en_gang_pr_uge	17	1.64	47
10	pige	aldrig	16	1.6	48
11	pige	en_eller_to_pr_Måned	16	1.61	48
12	pige	aldrig	16	1.63	48

Tabel 3. Uddrag af observationer for flere variable fra *Udklip1*.

Vi vil nu undersøge, om gennemsnitshøjden for pigerne i stikprøven er 1,685 m, som oplyst på folkesundhed.dk.

Vores nulhypotese bliver da:

$$H_0 : \text{Middelværdien for de tyske pigers højde er 1,685 m, dvs. } \mu_{\text{højdepiger}} = 1,685$$

Vi bruger datasættet *Udklip2* og bestemmer først deskriptorerne for datasættet – her vist i Excel ved hjælp af *Dataanalyse*³:

The screenshot shows the 'Dataanalyse' dialog box with 'Beskrivende statistik' selected. To the right, a summary table for 'Kolonne1' is displayed with the following values:

Kolonne1	
Middelværdi	1,693494
Standardfejl	0,003567
Median	1,7
Tilstand	1,7
Standardafvigelse	0,056285
Stikprøvevarians	0,003168
Kurtosis	-0,04746
Skævhed	0,107263
Område	0,32
Minimum	1,52
Maksimum	1,84
Sum	421,68
Antal	249
Konfidensniveau	0,007025

Herfra får vi den empiriske middelværdi, empiriske varians samt antal (alle indrammet ovenfor). Vi har dermed de størrelser, vi skal bruge til at teste nulhypotesen, og vi får teststørrelsen:

$$t \cong 2,381.$$

Den kritiske værdi for *t*-fordelingen med signifikansniveau 5% og med 248 frihedsgrader kan bestemmes til:

<code>=tinv(0,05;248)</code> TINV(sandsynlighed; frihedsgrader)	Resultat: 1,969576
--	------------------------------

Dvs. vores teststørrelse er større end den kritiske værdi, og vi må derfor forkaste vores nulhypotese på et 5% signifikansniveau, under forudsætning af at nulhypotesen er sand, dvs. gennemsnitshøjden for pigerne i

³ Installation af *Dataanalyse* i Excel sker under *Indstillinger > Tilføjelsesprogrammer*. En vejledning hertil kan findes flere steder på internettet.

stikprøven er ikke 1,685 m. p -værdien kan beregnes til 0,017 dvs. der er kun 1,7% sandsynlighed for at få en lige så stor værdi, som den fra folkesundhed.dk, hvilket jo giver samme konklusion.

Bemærk: Nogle værktøjsprogrammer kan udføre dette t -test automatisk ud fra fx de observerede værdier i tabellen.

Øvelse 12: t -test gennemsnitshøjde for drenge

Undersøg med udgangspunkt i stikprøven *Udklip1*, om der er belæg for, at drengenes gennemsnitshøjde er 1,810 m. Du vælger selv signifikansniveau.

Øvelse 13: t -test gennemsnitsvægt

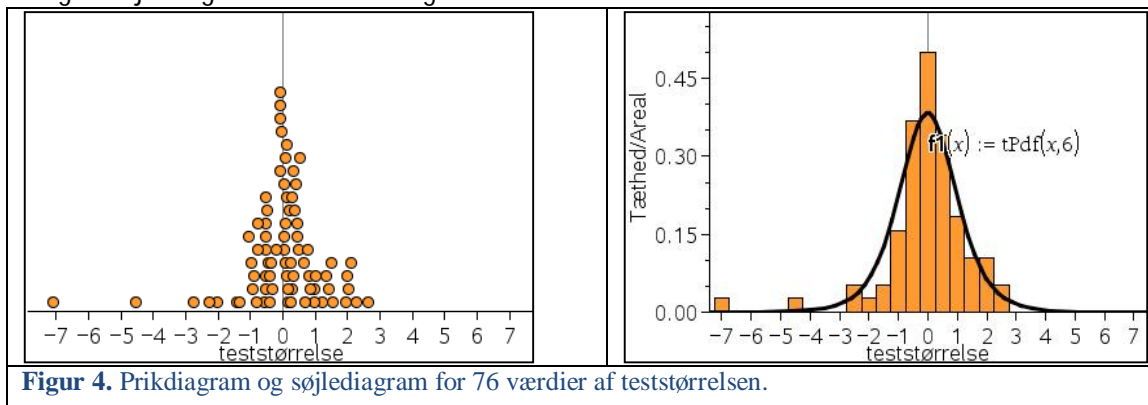
Undersøg med udgangspunkt i stikprøven *Udklip1*, om disse respondenter har en gennemsnitsvægt på 60 kg. Du vælger selv signifikansniveau.

Fordeling af t -teststørrelsen

Vi vil i det følgende undersøge t -teststørrelsen vha. simulering. Vi opfatter nu vores stikprøve med højderne for de 462 piger og drenge som vores population. Fra populationen udtager vi stikprøver af forskellig størrelse rigtig mange gange for at se, hvordan de mange teststørrelser for hver af disse stikprøver fordeler sig.

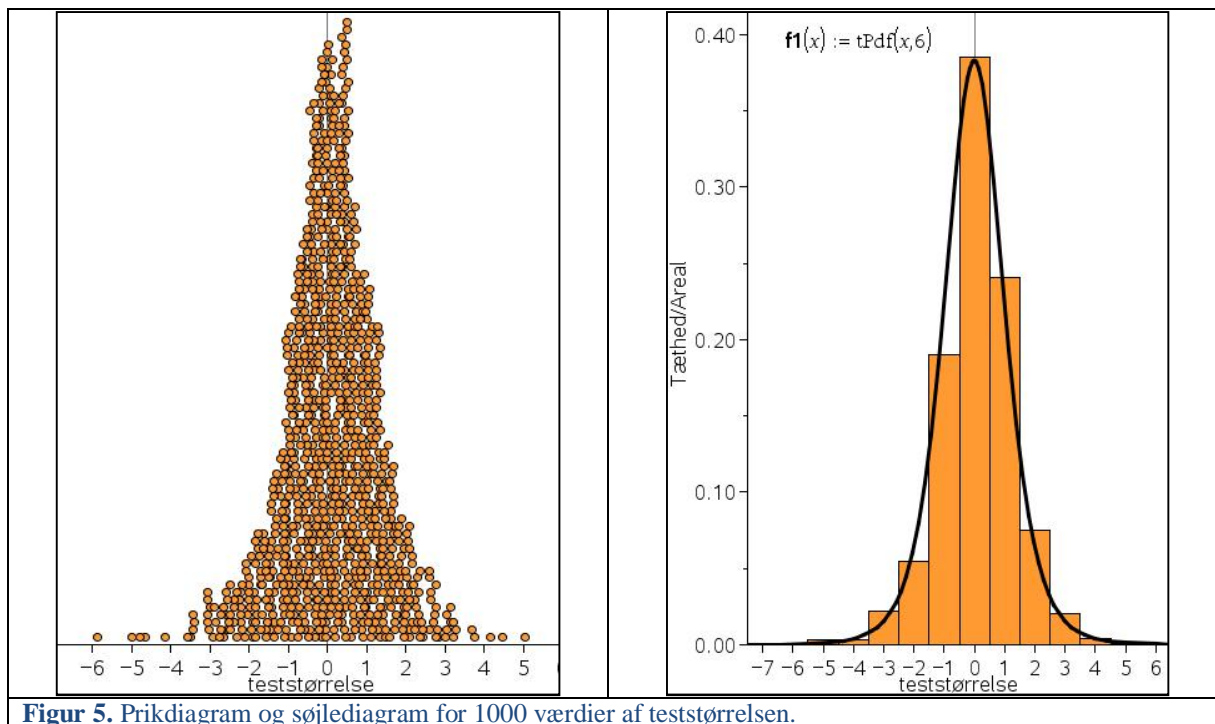
Stikprøve på 7 personer

Vi vælger at tage 76 stikprøver på hver 7 tilfældige personer blandt de 462. På figuren nedenfor ses et prikplot og et søjlediagram over fordelingen af de 76 teststørrelser:



Figur 4. Prikdiagram og søjlediagram for 76 værdier af teststørrelsen.

Prikplottet viser de 76 værdier af teststørrelsen, og vi ser, at teststørrelserne er både positive og negative. Mange af teststørrelserne samler sig omkring 0 og færre længere væk fra 0, og vi ser også en vis grad af symmetri omkring 0. Histogrammet viser, at de 76 teststørrelser fordeler sig med tilnærmelse til grafen for tæthedsfunktionen for en t -fordeling med 6 frihedsgrader, og vi skriver $t_{\text{teststørrelse}} \sim t_6$. Udvider vi fra 76 til 1000 stikprøver og tegner et prikplot med de 1000 teststørrelser, så får vi:



Figur 5. Prikdiagram og søjlediagram for 1000 værdier af teststørrelsen.

Konstruerer vi et histogram over fordelingen af de 1000 teststørrelser, finder vi, at histogrammet i endnu højere grad end med de 76 stikprøver følger grafen for tæthedsfunktionen for en t -fordeling med 6 frihedsgrader.

Øvelse 14: Frihedsgrader

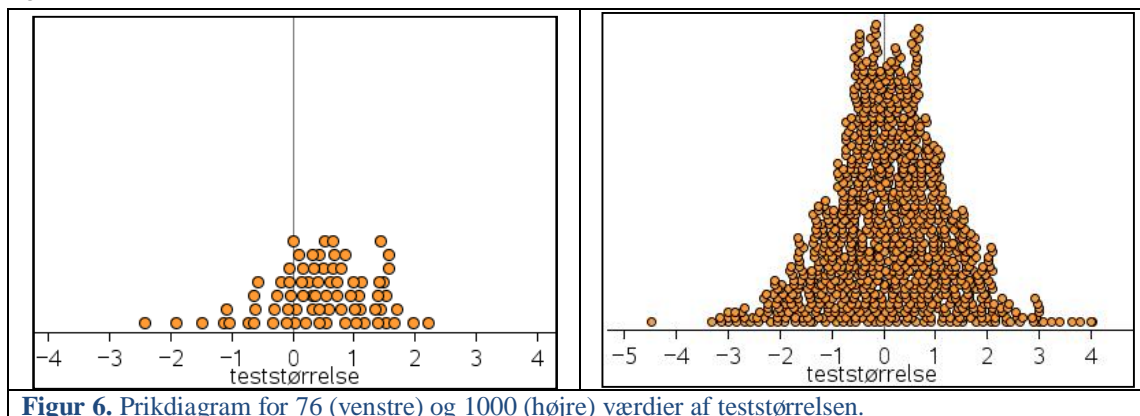
Tegn grafen for tæthedsfunktionen for t -fordelingen (benyt den indbyggede tæthedsfunktion i dit værktøjsprogram) med 1, 2, 3, 100 og 1000 frihedsgrader (brug evt. et skyderelement). Hvad minder grafens form dig om? På hvilken måde ændrer grafen udseende, når antal frihedsgrader øges?

Øvelse 15: Kritiske værdier

Bestem og indtegn de kritiske værdier på signifikans niveau 5% sammen med graferne i Øvelse 15 – husk, at t -testet er to-sidet, dvs. p -værdien svarende til den positive kritiske værdi er 2,5%.

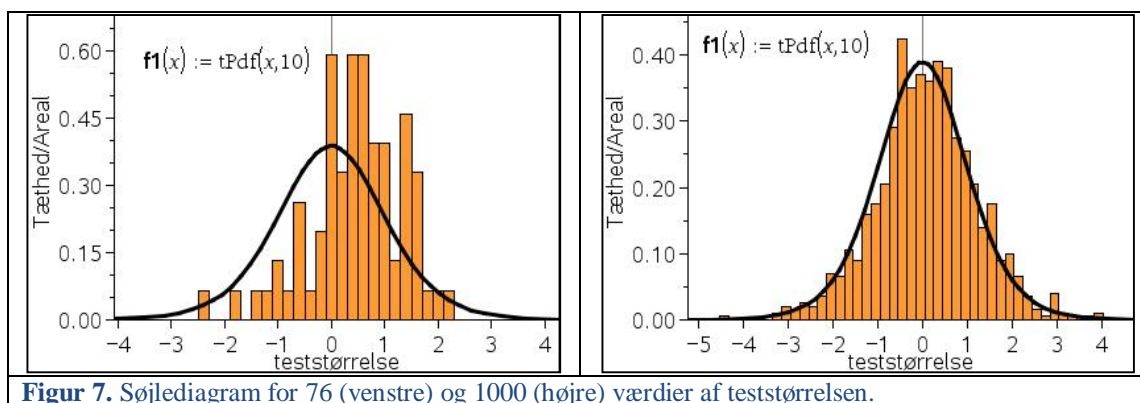
Stikprøve på 11 personer

Vi udvider nu stikprøvestørrelsen til 11 personer. På samme vis som før, så tager vi tilfældigt henholdsvis 76 og 1000 stikprøver med hver 11 personer fra de 462 personer. Prikplottene ser nu ud som vist på Figur 6 nedenfor.



Figur 6. Prikdiagram for 76 (venstre) og 1000 (højre) værdier af teststørrelsen.

På Figur 7 ses søjlediagrammer for fordelingen af teststørrelserne både i tilfældet med 76 stikprøver og 1000 stikprøver.



Igen ser vi, at søjlediagrammet følger grafen for tæthedsfunktionen for t -fordelingen med 10 frihedsgrader bedre des flere stikprøver, vi udtager.

Der gælder følgende sammenhæng mellem stikprøvens størrelse og antallet af frihedsgrader i den aktuelle t -fordeling:

Antal observationer og frihedsgrader
Antallet af frihedsgrader er én mindre end antallet af observationer ved et t -test.

Hypotesetest med to datasæt (to variable)

Arbejdet med t -testet i forrige afsnit efterlader et billede af, at noget lignende kan gøres for *to datasæt* eksempelvis med data for drengene og pigerne i Tabel 1. Her har vi målinger for to variable, som hver kan betragtes som udfald af *stokastiske variable*, hhv. X_{i1} og X_{i2} .

Øvelse 16: Boksplot til sammenligning

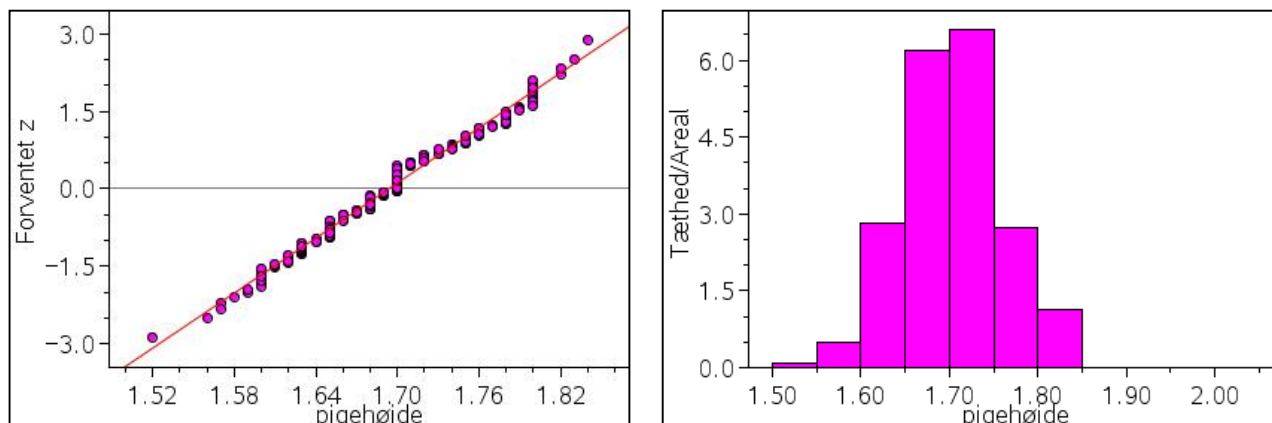
Bestem kvartilsæt samt middelværdi og konstruer et boksplot for hver af de to variable højde og vægt for hhv. alle piger og alle drenge i *Udklip1*. Sammenlign herefter med udgangspunkt i boksplot opdelt efter køn pigernes og drengenes vægt. Sammenlign på samme måde pigernes og drengenes højde.

Hvis vi vil *sammenligne* datasættene, er det oplagt at undersøge, om drengenes og pigernes højde er ens. I stedet for at sammenligne et datasæt med en tabelværdi, som vi gjorde før, vil vi her sammenligne middelværdierne for de to køns højdemålinger. Som ved testet på ét datasæt, tager vi udgangspunkt i de *empiriske størrelser* og undersøger nulhypotesen:

Pigernes gennemsnitshøjde er den samme som drengenes gennemsnitshøjde

Først må vi dog undersøge, om de to datasæt hver for sig er normalfordelte.

Vi ved allerede at drengenes højdemålinger er normalfordelte. I Figur 8 nedenfor ser vi fraktildiagrammet samt søjlediagrammet for fordelingen af pigernes højde (jf. Øvelse 11).



Figur 8: Fraktildiagram og søjlediagram for fordelingen af pigernes højdemålinger.

Fraktildiagrammet giver os god grund til at tro på, at pigernes højder er normalfordelte. Dvs. vi kan gå ud fra, at fordelingen af højdemålingerne for både piger og drenge kan beskrives ved en normalfordeling. Vi siger også, at den stokastiske variabel X_{ij} , der ligger bag de kønsopdelte data er normalfordelt, og vi skriver med matematisk notation: $X_{ij} \sim N(\mu_j, \sigma_j^2)$, hvor j 'et betyder, at de to datasæt har forskellige *parametre*. Vi skal så nu undersøge om disse parametre kan antages at være ens, dvs. om pigerne og drengene er gennemsnitligt lige høje.

For at kunne sammenligne middelværdierne for hhv. pigernes og drengenes højde, dvs. vurdere forskellen mellem de empiriske middelværdier, må vi også her tage hvert af datasættenes spredning i betragtning.

Derfor får vi brug for en størrelse, der beskriver den fælles varians, s_f^2 :

Fælles varians for de to datasæt

Den fælles varians for de to datasæt kan beregnes ved

$$s_f^2 = \frac{f_d \cdot s_d^2 + f_p \cdot s_p^2}{f_d + f_p}$$

hvor f_d betegner frihedsgrader, og s_d betegner spredningen i relation til drengenes højde – tilsvarende for f_p og s_p for pigernes højde.

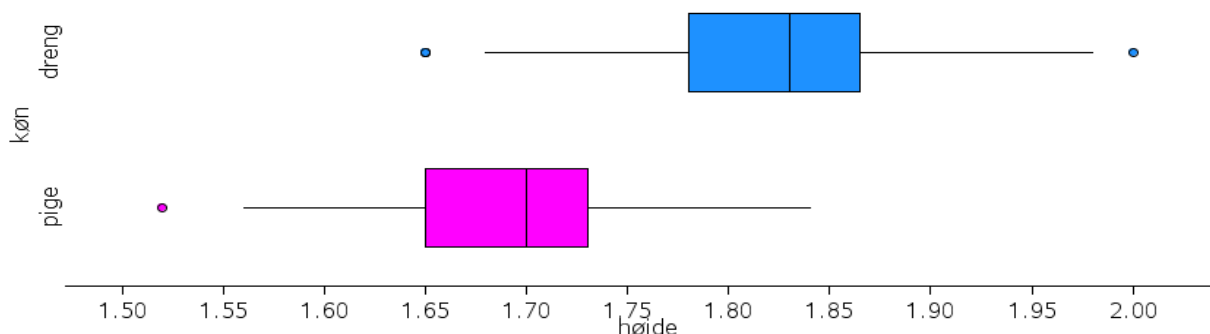
Øvelse 17: Fælles varians

Benyt dit værktøjsprogram til at bestemme s_f^2 for de to datasæt i Tabel 1.

Vi vil ikke diskutere s_f^2 dybere, men blot konstatere at s_f^2 indeholder bidrag fra begge datasæt, og at der er et krav til variationen *internt* i begge disse, for at vi kan bruge s_f^2 :

Der må ikke være signifikant forskel på de to variationer.

Det vil sige, at pigernes højdemålinger skal variere ligeså meget (eller lidt) som drengenes højdemålinger, for at vi kan sammenligne de to middelværdier. Hvordan man vurderer om dette er tilfældet, er ikke helt ligetil, men fx kan et boksplot give os et godt billede af om variationerne følger samme mønster:



Figur 9. Boksplot for højdemålingerne for drengene og pigerne fra Tabel 1.

Af Figur 9 fremgår det, at variationsbredden og kvartilbredden i hvert af de to boksplot er nogenlunde ens, og vi har derfor grund til at tro, at der ikke er signifikant forskel på variationen i de to variable⁴. Bemærk, at det kun er variansen, vi forholder os til her – og ikke middelværdien.

⁴ Det er også muligt at teste om forskellen mellem de to varianser er signifikant, i dette tilfælde bruges et *F*-test.

På baggrund af overvejelserne ovenfor opstiller vi nulhypotesen og udregner teststørrelsen for sammenligningen af de to datasæts middelværdier:

H_0 : Middelværdien for pigernes højde og drengenes højde er ens, dvs. $\mu_d = \mu_p$

Generelt gælder der om t -teststørrelsen i denne sammenhæng:

t -test for to datasæt med ens varians

Den generelle formel til beregning af t -teststørrelsen for to datasæt (her dog eksemplificeret ved datasæt for piger og drenge) er:

$$t = \frac{\bar{x}_d - \bar{x}_p}{\sqrt{s_f^2 \cdot \left(\frac{1}{n_d} + \frac{1}{n_p} \right)}}$$

hvor \bar{x}_d og \bar{x}_p betegner de empiriske middelværdier for hhv. drenge og piger, s_f betegner den fælles varians for piger og drenge, mens n_d og n_p betegner antallet af observationer for hhv. drengenes højde og pigernes højde.

Teststørrelsen er t -fordelt med $f = n_d + n_p - 2$ frihedsgrader, og vi skriver: $t \sim t_{n_d+n_p-2}$

Ved indsættelse af de kendte værdier i formelen bliver t -teststørrelsen således her:

$$t \cong 22,55$$

Vi undersøger om denne teststørrelse er kritisk i forhold til nulhypotesen på et 5% signifikansniveau. Da begge datasæt er normalfordelte, finder man, at teststørrelsen er t -fordelt med

$$f = n_d + n_p - 2 = 213 + 249 - 2 = 460 \text{ frihedsgrader.}$$

Den kritiske værdi bestemmes (her beregnet i Excel):

<code>=TINV(0,05;460)</code> TINV(sandsynlighed; frihedsgrader)	Resultat: 1,965134
--	------------------------------

Dvs. vores teststørrelse er *større end* den kritiske værdi, og vi må således forkaste nulhypotesen på et 5% signifikansniveau. Dvs. højderne er ikke ens for de to køn.

Vi kan tilsvarende udregne, at p -værdien bliver meget lille, nemlig $4,3 \cdot 10^{-78}$, og at konklusionen ud fra p -værdien derfor som ventet er den samme.

Bemærk: I det følgende og tilsvarende i de opgaver, der stilles til den skriftlige prøve, vil vi forudsætte, at variansen for de aktuelle variable er ens.

Anvendelse af t -test for to datasæt (to variable) i *Muffins*

I *Udklip3* i [MuffinsUdklip.xls](#) har vi følgende datasæt:

	A	B	C	D	E	F
1	Køn	Fritid	Fjernsynstid	Læsetid		
2	dreng	74.25	0	4		Tallene er per uge.
3	dreng	61.75	4	3		Læsetid er udover lektielæsning.
4	pige	71	2	5		
5	dreng	73.25	15	1		
6	dreng	71.25	10	1		
7	dreng	78	2	2		
8	dreng	70	20	2		
9	pige	79.25	10	1		
10	pige	71	10	2		
11	dreng	69	2	3		
12	pige	73	10	5		
13	pige	65.5	4	3		
14	pige	72	12	5		

Tabel 4. Udklip3 med både kategoriske og numeriske variable.

Øvelse 18: Numeriske og kategoriske variable

Hvilke variable indgår i udklipet? Hvilke af disse variable er numeriske henholdsvis kategoriske?

Øvelse 19: Boksplot til sammenligning

Bestem kvartilsæt samt middelværdi, og konstruer boksplot for de tre variable *Fritid*, *Fjernsynstid* og *Læsetid* for hhv. alle piger og alle drenge. Kommenter fordelingen af data for de tre variable.

Med udgangspunkt i denne stikprøve kan vi undersøge forskellige hypoteser, hvor de fire variable indgår. Vi vil undersøge om gennemsnittet af fritidstimer på en uge for piger er den samme som gennemsnittet af fritidstimer på en uge for drenge. Vi antager, at variansen for piger og drenges fritidstimer er ens. Vores nulhypotese bliver da:

$$H_0 : \mu_{\text{fritidpiger}} = \mu_{\text{fritiddrenge}}$$

Vi sorterer først data, således at vi først har pigernes data og derefter drengenes data. På den måde får vi opdelt vores stikprøve i to mindre stikprøver.

Vi tester nulhypotesen med et *t*-test og får (her vist i Excel):

The image shows two screenshots from Excel. The left screenshot shows the 'Dataanalyse' (Data Analysis) task pane with 't-test: To stikprøver med ens varians' selected. The right screenshot shows the 't-test: To stikprøver med ens varians' dialog box. The dialog box has the following fields and options:

- Input:**
 - Område for variabel 1: [Empty]
 - Område for variabel 2: [Empty]
 - Hypotese for forskel i middelværdi: [Empty]
 - Etiketter
 - Alpha: 0,05
- Outputindstillinger:**
 - Outputområde: [Empty]
 - Ny regnearksfane: [Empty]
 - Ny projektmappe

t-test: To stikprøver med ens varians

Input

Område for variabel 1:

Område for variabel 2:

Hypotese for forskel i middelværdi:

Etiket

Alpha:

Outputindstillinger

Outputområde:

Ny regnearksfane:

Ny projektmappe

OK
Annuller
Hjælp

	Variabel 1	Variabel 2
Middelværdi	69.53185	72.05227
Varians	31.32015	30.61826
Observationer	281	220
Puljevarians	31.01211	
Hypotese for forsk	0	
fg	499	
t-stat	-5.02751	
P(T<=t) en-halet	3.47E-07	
t-kritisk en-halet	1.647913	
P(T<=t) to-halet	6.94E-07	
t-kritisk to-halet	1.964729	

Da p -værdien er meget lille, kan vi klart afvise nulhypotesen på et 5% signifikansniveau under forudsætning af, at der er ens varians. Dvs. gennemsnittet af fritidstimer på en uge er forskellig for piger og drenge.

Øvelse 20: Køn og fjernsyn

I *Udklip3* finder du observationer fra *Muffins* for de to variable *Køn* og *Fjernsynstid*, hvor variabelen *Fjernsynstid* indeholder respondenternes svar på, hvor mange timer de ser tv pr. uge.

Bestem kvartilsættene, og tegn for hvert køn et boksplot, der viser fordelingen af fjernsynstid for begge køn.

Det oplyses, at variansen af pigernes og drengenes fjernsynstid kan antages at være ens. Overvej, hvordan dette kommer til udtryk i boksplottene.

Opstil en nulhypotese, der kan anvendes til at undersøge om middelværdien for pigers og drenges fjernsynstid er ens, og undersøg om nulhypotesen kan forkastes på et 5% signifikansniveau.

Øvelse 21: Køn og vægt

I *Udklip1* finder du observationer fra *Muffins* for de to variable *Køn* og *Vægt*.

Bestem kvartilsættene, og tegn for hvert køn et boksplot, der viser vægtfordelingen for begge køn.

Det oplyses, at variansen af pigernes og drengenes vægt kan antages at være ens. Overvej, hvordan dette kommer til udtryk i boksplottene.

Opstil en nulhypotese, der kan anvendes til at undersøge om middelværdien for pigers og drenges vægt er ens, og undersøg om nulhypotesen kan forkastes på et 5% signifikansniveau.

Kilder

1. *Muffins* databasen:
<http://www.mathematik.uni-kassel.de/didaktik/muffins/>
2. Statistik, Thomas Bendsen, VIA University College Bioanalytikeruddannelsen, 2009-2012:
<http://statnoter.biolyt.dk/>
3. *Statistik i løb*, Preben Blæsild og Lars Bo Kristensen, LMFK, 2005, kan købes her:
<http://www.lmfk.dk/> under bogsalg.
4. *Idrætsstatistik*, Preben Blæsild og Jørgen Granfeldt, Institut for Matematiske Fag AU, 2001:
<http://www.uvmat.dk/stat/index.htm>.
5. Danskernes højde ifølge Folkesundhed.dk:
<http://www.si-folkesundhed.dk/Ugens%20tal%20for%20folkesundhed/Ugens%20tal/Danskernes%20h%C3%B8jde.aspx>

a)