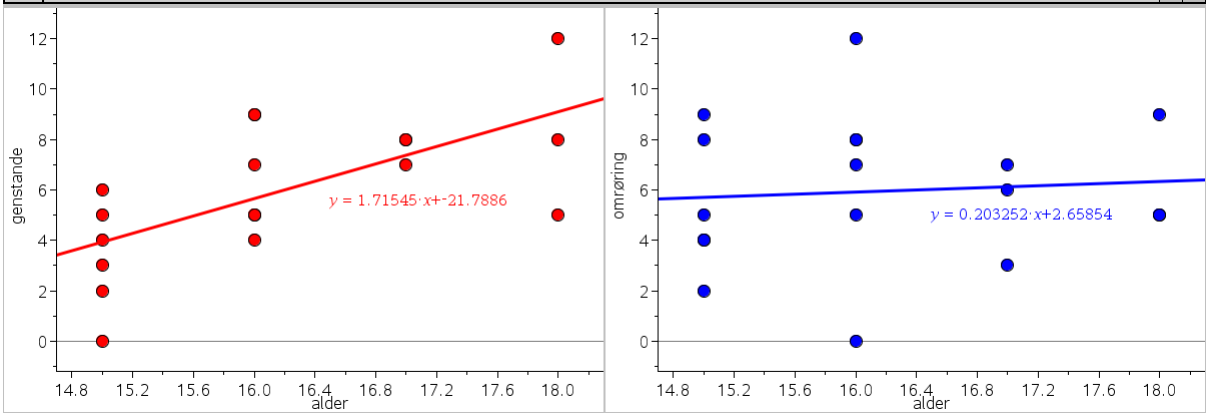
	A køn	B alder	C genstande	D omrøring	E	F	G	H	I	J	K	L	M	N
◆				=randsamp(gens		=LinRegMx('alc			=LinRegMx('alde					
1	Dreng	15	0	4	Titel	Lineær regres...	Restvariation=	Titel	Lineær regress...					
2	Dreng	15	4	4	RegEqn	m*x+b	80.6179	RegEq...	m*x+b					
3	Dreng	15	3	5	m	1.71545	Totalvariation=	m	0.203252					
4	Dreng	16	7	8	b	-21.7886	140.944	b	2.65854					
5	Dreng	16	9	5	r^2	0.428017	Forklaringsgrad=	r^2	0.006009					
6	Dreng	16	5	12	r	0.65423	0.428017	r	0.077515					
7	Dreng	16	4	0	Resid	{-3.94308943...		Resid	{-1.707317073...					
8	Dreng	17	8	3										
9	Dreng	17	7	7										
10	Dreng	17	8	6										
11	Dreng	18	8	5										
12	Dreng	18	12	9										



Uafhængighed – et eksempel på en rød tråd i statistikken

Statistiknoter til TI-Nspire CAS version 3.1

Bjørn Felsager
Revideret November 2011

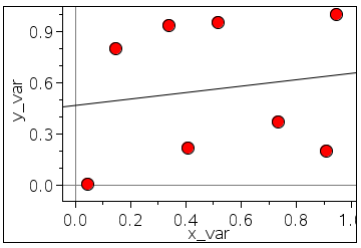
Uafhængighed – et eksempel på en rød tråd i statistikken

Chi-i-anden-testen er nu obligatorisk stof på matematik B og A, og ligeledes er tolkning af p-værdier obligatorisk stof på samfundsfag A. Både matematik og samfundsfag skal altså fremover kunne samarbejde omkring hypotesetest. Samtidigt anvendes statistiske tests hyppigt i biologiundervisningen og også her er det derfor oplagt med et samarbejde mellem matematik og biologi omkring hypotesetest. Det kan derfor være nyttigt at sætte chi-i-anden-testen ind i en lidt større ramme.

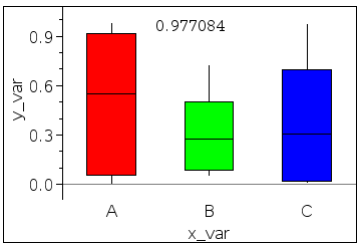
Der findes som bekendt to typer af chi-i-anden test: **Uafhængighedstesten** og **Goodness-of-fit-testen**. Begge typer rækker ud til andre nærtbeslægtede tests. I denne note vil vi fokusere på uafhængighedstesten, der især rækker ud til **t-testen**.

Vi skal altså se på den generelle problemstilling: **Hvordan afgør vi om to stokastiske variable er uafhængige**?

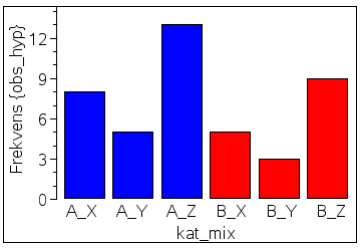
Der findes nu tre forskellige typer af stokastiske variable: De *numeriske/kvantitative* variable og de *kategoriske/kvalitative* variable som kan være uordnede (nominale) eller ordnede (ordinale). Det giver anledning til følgende tre nye kombinationer af uafhængighedstest:



Type I:To numeriske variable



Type II: Én kategorisk og én numerisk variabel



Type III: To ordinale variable

Den første type bygger på **lineær regression** og giver derfor også mulighed for at uddybe forståelsen af den lineære regression, der er grundlæggende for et andet af de nye obligatoriske samarbejdsemner mellem matematik og samfundsfag: regressionsmodeller. Det giver endnu en god begrundelse for at dykke ned i uafhængighedssporet. Denne type fører til **t-testen** for hældningen af regressionslinjen.

Den anden type bygger på **sammenligning af fordelinger**, herunder specielt sammenligninger af niveauerne, fx hvilken type pølse indeholder mest nitrit osv. Det er ikke helt nemt at håndtere denne type statistisk, så ofte vil man transformere den over i en af de to andre typer. Men hvis man kigger på en kategorisk variabel med kun *to værdier*, dvs. kun vil sammenligne to fordelinger, fører den til **t-testen** for forskellen mellem to middelværdier, som er nært beslægtet med den foregående t-test.

Den tredje type fører direkte til **krydstabeller** og **gamma-testen**. Denne test er velegnet som supplement til chi-i-anden testen, idet den inddrager rangfølgen af de ordnede kategoriske variable og dermed giver flere oplysninger om en eventuel sammenhæng end chi-i-anden testen, ligesom den ofte giver en stærkere test.

Den grundlæggede testteori for de tre typer af uafhængighedstest er den samme. Specielt er det den samme måde man simulerer uafhængighed på i de tre tilfælde, idet man sikrer uafhængighed af de to involverede variable ved at **omrøre** værdierne for den ene. Men den teoretiske udformning er derimod forskellig i de tre tilfælde, som vi især skal se nærmere på i det følgende.

Det fører også til følgende karakterisering af de eksperimentelle metoder versus de teoretiske metoder i de tre tilfælde: Det er forholdsvis kompliceret at gennemføre en eksperimentel test i de tre tilfælde, men det er den samme overordnede metode i alle tre tilfælde og den er nem at forstå. Omvendt er det forholdsvis enkelt at gennemføre en teoretisk test i alle tre tilfælde. Men til gengæld er der tale om tre væsensforskellige metoder, og det er ikke helt nemt at forstå de bagvedliggende formler. Det er derfor en klar fordel at starte med den eksperimentelle metode for at bygge forståelse op for testteorien og derved skaffe sig et brugbart kørekort til at kunne anvende de teoretiske metoder i praksis.

Men først skal vi altså have en ramme for testteorien. Her benytter vi os af **Fishers testteori**. Det grundlæggende spørgsmål i testteorien kan formuleres således:

Hvordan ser man forskel på en tilfældig variation og en systematisk variation.

Når man anvender testteori handler det om at kunne træffe et rationelt valg på basis af nogle observationer. Disse observationer kan forklares ud fra en helt tilfældig variation af de indgående stokastiske variable eller ud fra en systematisk variation af de indgående stokastiske variable (hvor der typisk vil være overlejret en tilfældig variation). Der er tradition for at kalde antagelsen om en tilfældig variation for **nulhypotesen H_0** og tilsvarende antagelsen om en (delvis) systematisk variation for **den alternative hypotese Ha**. Begge antagelser er principielt mulige, men vi skal have truffet et valg og vi skal derfor på en eller anden måde fremskaffe en vurdering af de to antagelsers troværdighed.

Det gøres på baggrund af nulhypotesen om uafhængighed. Man skal altså forsøge at karakterisere den tilfældige variation frembragt af to uafhængige stokastiske variable, så man bliver i stand til at genkende en typisk tilfældig variation knyttet til to uafhængige variable.

Det kan gøres **eksperimentelt** ved at simulere nulhypotesen. En sådan simulering er nem at udføre i et dynamisk statistikprogram som TI-Nspire CAS. Og det giver megen forståelse for fordelingen af teststørrelse, når man selv har prøvet at frembringe dem. Tilsvarende giver det bedre forståelse for brugen af den universelle teoretiske fordeling, når man helt konkret har kunnet sammenligne den med en simuleret fordeling. Derimod er det uden for rækkevidde af almindelig gymnasieundervisning i matematik at udlede formlerne for tæthedsfunktionerne for de universelle fordelinger (z-normalfordelinger, t-fordelinger, χ^2-fordelinger osv.) endsige bevise de centrale grænseværdisætninger for disse universelle fordelinger. Alt hvad vi kan håbe på i den almindelige undervisning er derfor at give eleverne en god fornemmelse for brugen af disse universelle fordelinger uden at begrunde i matematiske detaljer, hvor de egentlig kommer fra.

I den foreliggende note arbejdes der derfor meget konkret med simulering af uafhængighed som grundlag for et uafhængighedstest, og det vises samtidigt, hvordan man i praksis kan gennemføre en teoretisk test ved hjælp af de indbyggede rutiner. Endelig vises det hvordan man kan udvide simuleringen, så den også dækker den teoretiske udformning af testen, men det er mere kompliceret, så disse afsnit er markeret med en * - de kan roligt overspringes i en første gennemlæsning.

For at kunne arbejde med simuleringer har vi brug for et **konkret datamateriale**. Det kan man få forskellige steder fra. Fx kan man tage udgangspunkt i **en historisk case**. Her ville fx Challenger-ulykken egne sig fortrinligt til en illustration af de forskellige typer uafhængighedstest idet nøglespørgsmålet i opklaringen af ulykken jo er en påvisning af at uheldene med de såkaldte O-ringe afhang signifikant af temperaturen på opsendelsestidspunktet for rumfærgen. Her vil vi i stedet tage udgangspunkt i **et elevfrembragt datamateriale**. I dette tilfælde drejer sig om en spørgeskemaundersøgelse om unges alkoholforbrug, som nemt kan udbygges hjemme på skolen. Eksemplet er lånt fra den glimrende metodebog 'Fold dig ud – i samfundsfaglige metoder' af Per Henriksen og Torben Steiner Nielsen, udgivet på forlaget Columbus. Eksemplet diskuteres dels i kapitel 6: Tal, der taler, side 69-72, dels i kapitel 7, Kvantitative sammenhænge og forskelle, side 85-88. Den foreliggende note kan ses som et forsøg på at udfolde de samme emner fra en mere matematisk synsvinkel, og dermed forhåbentligt give anledning til en frugtbar udveksling af samfundsfaglige og matematiske metoder i arbejdet med et sådant projekt.

Referencer:

Ønsker man at læse noget baggrundsmateriale om statistik for bedre at kunne forstå emnet kan man dels orientere sig historisk, dels teoretisk. Hvad angår den historiske udvikling af statistikken er det stærkt anbefalelsesværdigt at læse den charmerede anekdotiske beretning om den moderne statistiks historie:

David Salsburg: *The Lady Tasting Tea*: *How Statistics Revolutionized Science in the Twentieth Century*, Freeman, 2001.

Ønsker man en mere akademisk diskussion af statistikkens historie er det værd at læse:

Lorraine Daston et al: *The Empire of Chance*: *How probability changed science and everyday life*, Cambridge Press, 1989

Ønsker man at fordybe sig i den mere matematisk/teoretiske baggrund for formler osv. kan man på gymnasieniveau fordybe sig i den følgende glimrende introduktion til statistik fra en matematisk synsvinkel:

Axel Bertelsen: *Statistik med matematik*, Systime, 2005.

På universitetsniveau kan man fx læse den følgende kompakte detaljerede fremstilling af moderne statistik:

Jørgen Larsen: *Basisstatistik*, Imfufatekst nr. 435 fra 2004.

Ønsker man at vide mere om hvordan statistik anvendes i praksis i samfundsfag kan man fx gribe til den lettilgængelige:

Per Vejrup-Hansen: *Statistik med Excel*, Samfundslitteratur, 2001

Endelig kan man hente nyttige informationer om hvordan man bruger regneark i samfundsfag i den overskuelige indføring

Christina Black Hansen, Per Henriksen: *Excel i samfundsfag*, Forlaget Columbus, 2011.

Noten er forfattet/revideret af Bjørn Felsager, november 2011.

1. Unges alkoholforbrug som funktion af alder: Lineær regression t-test

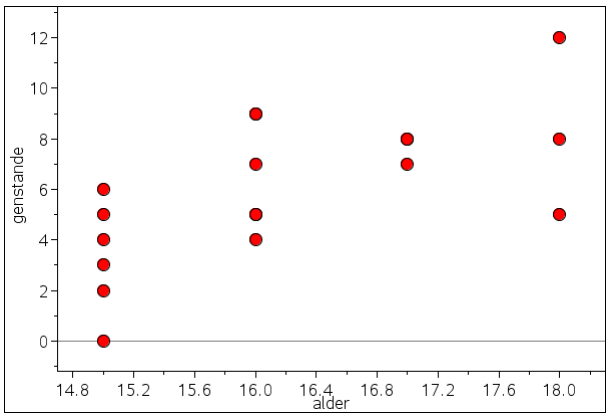
1.1 Lineær regression t-test: Den beskrivende statistik

Lad os som et gennemgående eksempel se på et spørgeskema der handler om *unges alkoholforbrug*: Eksemplet stammer fra bogen Fold dig ud, der er en metodeteori for samfundsfag. Eksemplet omhandler alkoholforbruget hos 18 elever i en klasse. Eksemplet er tænkt som et illustrativt eksempel med forholdsvis få data. I praksis skal man bruge en væsentlig større stikprøve, men for overskuelighedens skyld holder vi os til eksemplet med de 18 elever. Idéen er nu at undersøge om alkoholforbruget hos de unge kunne tænkes at afhænge af deres køn henholdsvis deres alder. **Alkoholforbruget** (antal genstande pr uge) er altså *responsvariablen*, mens **køn** og **alder** er de *forklarende variable*. Her er kønnet en kategorisk variable, mens alder er en numerisk variabel. Vi har derfor de to første typer af kombinationer repræsenteret. Vi starter med at undersøge sammenhængen mellem alder og alkoholforbrug.

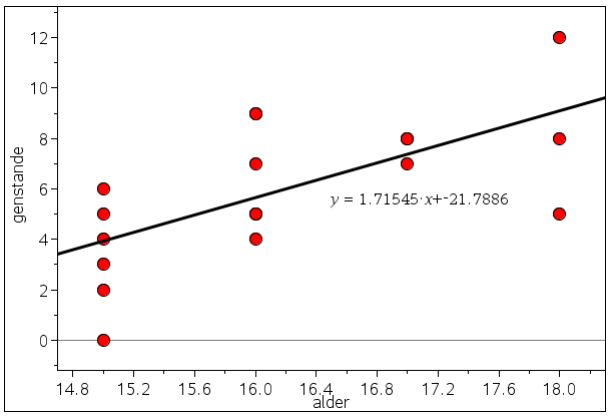
Som sædvanlig lægger vi ud med en **beskrivende** undersøgelse af datasættet (deskriptiv statistik).

	A køn	B alder	C genstande
◆			
1	Dreng	15	0
2	Dreng	15	4
3	Dreng	15	3
4	Dreng	16	7
5	Dreng	16	9
6	Dreng	16	5
7	Dreng	16	4
8	Dreng	17	8
9	Dreng	17	7
10	Dreng	17	8
11	Dreng	18	8
12	Dreng	18	12
13	Pige	15	5
14	Pige	15	6
15	Pige	15	2
16	Pige	16	9
17	Pige	16	5
18	Pige	18	5

Datasættet er sorteret primært efter **køn** og sekundært efter **alder**, dvs. først har vi listet alle drengene efter stigende alder og dernæst alle pigerne efter stigende alder. Det sker ved at markere de tre søjler, højre klikke og vælge menupunktet **sorter**, idet vi *sorterer baglæns*, dvs. i første omgang efter **alder** og i anden omgang efter **køn**. For at undersøge om der kan være tale om en sammenhæng mellem **alder** og **genstande** afbilder vi nu de to variable i et **punktplot**. Vi tilføjer derfor et **Diagrammeer og statistik**-værksted og afsætter **alder** vandret og **genstande** lodret:

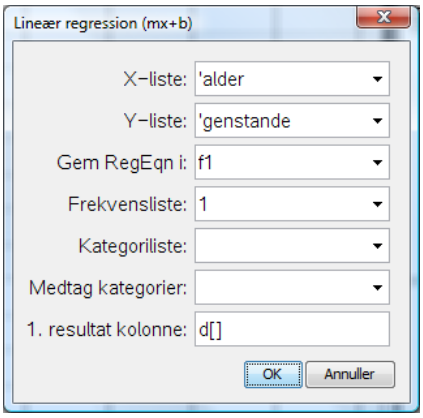


Det kunne godt se ud som om der var en svagt stigende tendens, og vi tilføjer derfor en regressionslinje ved at gå ind i **Undersøg data**-menuen:



Vi ser af regressionsligningen, at hældningen for regressionslinjen er 1.7, dvs. alkoholforbruget i denne klasse stiger med 1.7 genstande pr. uge, for hver gang alderen stiger med 1 år. Skæringen med aksen er ikke relevant, da det ikke giver mening at tillægge babyer et alkoholforbrug. Den viser sig da også i dette tilfælde at være negativ. Men selv om det ser ud som om alkoholforbruget udviser en svagt stigende tendens med alderen, så er der også tydelige individuelle variationer. Det kan tolkes som resultatet af tilfældige variationer eller som resultatet af andre forklarende variable. Vi ønsker derfor et mål for *kvaliteten af modellen*: Hvor god er den lineære model til at forklare den observerede variation i alkoholforbruget? Det er her forklaringsgraden r^2 kommer ind i billedet. Hvis vi vil beregne forklaringsgraden er det nemmest at gå tilbage til regnearket og udføre den lineære regression dér (vælg **statistik**-menuen).

Vi skal da som vist udfylde en dialogboks, der først og fremmest handler om hvilke to variable, der er på spil, men også giver mulighed for at inkludere hyppighedslister (hvilket ikke er relevant i dette tilfælde, hvor vi har adgang til de rå data), ligesom vi kan splitte beregningen efter en tredje kategorisk variabel.



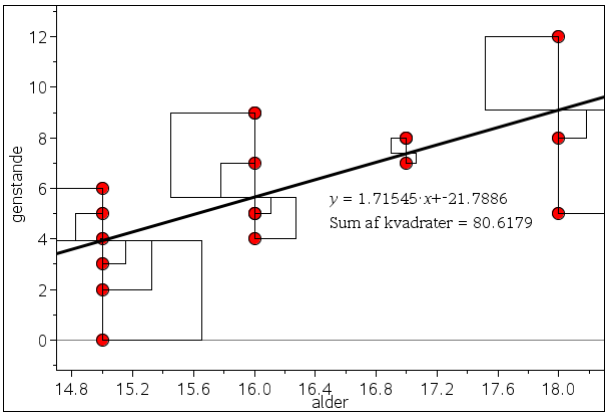
i dette tilfælde kunne vi altså inkludere **køn** som en kategoriliste og sige at vi kun ville se resultatet for drengene eller at vi kun ville se resultatet for pigerne[1]. Men her ser vi altså på det samlede resultat:

D	E
	=LinRegMx('alder,'genstande,
Titel	Lineær regression (mx+b)
RegEqn	m*x+b
m	1.71545
b	-21.7886
r^2	0.428017
r	0.65423
Resid	{-3.9430894308935,0.05691...

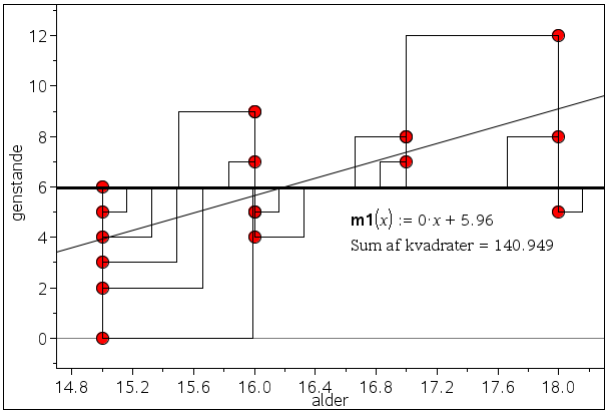
Vi får da ligesom i grafrummet oplyst ligningen for regressionsligningen i form af koefficienterne m og b, men denne gang får vi også oplyst såvel korrelationskoefficienten r som forklaringsgraden r^2. I matematisk modellering vil vi være mest interesseret i forklaringsgraden, fordi den - i modsætning til korrelationskoefficienten - giver mening for alle slags modeller. I dette tilfælde er forklaringsgraden altså 42.8%, dvs. den lineære model kan forklare 42.8% af den observerede variation, mens den resterende variation må forklares af anden vej, fx som et resultat af tilfældige variationer i alkoholforbruget.

Spørgsmålet er så hvad vi mener mere præcist med at den lineære model kan forklare 42.8 % af variationen. Her kan vi med fordel se lidt nærmere på mindste kvadraters metode. Vælger vi **Residualer** i **Undersøg data**-menuen kan vi få vist de *residuelle kvadrater* for den lineære regressionsmodel. Vi ser da, at summen af kvadraterne, dvs. summen af de kvadratiske afvigelser hørende til den lineære model, er givet ved 80.6179. Det er den såkaldte *rest-variation*, dvs. den variation, som vi *ikke* kan forklare ved hjælp af den lineære model. Hvis den er lille er modellen god, hvis den er stor, er modellen dårlig. Problemet er blot at et tal i sig selv ikke kan være lille eller stort. Det kan det kun i sammenligning med en standard.

[1] Dette ville kræve at vi først oprettede særskilte lister for delkategorierne {Dreng} og {Pige}



Her er det vi med fordel kan inddrage *nulhypotesen* som en standard. Ifølge nulhypotesen kan vi også prøve at forklare den observerede variation som et resultat af en tilfældig variation. Ifølge nulhypotesen er der nu slet ingen sammenhæng mellem alkoholforbruget og alderen. Men det betyder jo at alkoholforbruget ifølge nulhypotesen blot svinger tilfældigt rundt om sin middelværdi. Vi kan derfor modellere nulhypotesen med en konstant funktion, der er lig med middelværdien[2]. Vi tegner derfor den konstante funktion givet ved middelværdien ved at vælge menupunktet **Plot funktion** i **Undersøg graf**-menuen:



Desværre kan vi ikke få vist de *residuelle kvadrater* for en funktion. Men i stedet kan vi få et tilnærmet billede af situationen ved at vælge en **flytbar linje** i **Undersøg data**-menuen og så lægge den oven på grafen for den konstante funktion. Vi får da den tilnærmede værdi 140.49 for summen af kvadraterne ifølge nulhypotesen. Denne sum kaldes for *totalvariationen*. Den udgør nu den standard, som vi kan måle restvariationen i forhold til. Da restvariationen på 80.62 udgør 57.2% af totalvariationen på 140.9 viser det netop at det ud fra den lineære model er lykkedes os at forklare 100% - 57.2% = 42.8% af variationen. Dette er den præcise betydning af *forklaringsgraden*. Vi kan nemt formalisere udregningen i regnearket:

[2] Man kan forholdsvis simpelt vise at middelværdien netop er den bedste konstante funktion ifølge mindste kvadraters metode.

	D	E	F
◆		=LinRegMx('alder,'genstande,	
1	Titel	Lineær regression (mx+b)	Restvariation=
2	RegEqn	m*x+b	80.6179
3	m	1.71545	Totalvariation=
4	b	-21.7886	140.944
5	r²	0.428017	Forklaringsgrad=
6	r	0.65423	0.428017
7	Resid	{-3.9430894308935,0.05691...	

De indtastede formler i cellerne ser således ud:

$F2$	$=\text{sum}\left(\left(f1(\textbf{alder})-\textbf{genstande}\right)^2\right)$
$F4$	$=\text{sum}\left(\left(\text{mean}(\textbf{genstande})-\textbf{genstande}\right)^2\right)\cdot 1.$
$F6$	$=1-\frac{f2}{f4}$

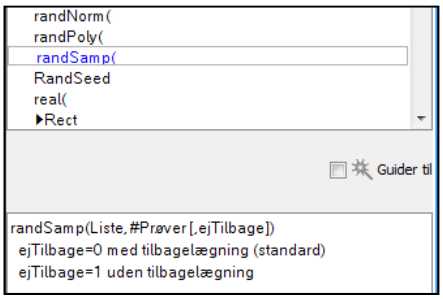
Dermed har vi styr på forklaringsgraden og vi er færdige med den beskrivende statistik, dvs. vi har gjort rede for strukturen af observationerne.

1.2 Lineær regression t-test: Den bekræftende statistik

Vi skifter så til den **bekræftende statistik**, dvs. vi stiller nu spørgsmålet, hvorvidt det observerede mønster kan udvides til en større population, hvor den pågældende klasse kan betragtes som en *repræsentativ stikprøve* for denne større population, fx populationen af elever i gymnasiet generelt på det pågældende klassetrin. Læg godt mærke til repræsentativiteten. De pågældende elever er jo næppe udvalgt tilfældigt blandt alle landets elever – så repræsentativiteten er noget man må argumentere for ad anden vej.

I den bekræftende statistik skal vi nu afgøre hvorvidt den observerede variation ligner en typisk tilfældig variation eller ej. Hvis den ikke ligner en tilfældig variation, må vi forkaste nulhypotesen og vælge at tro på den systematiske variation, dvs. vi må vælge den alternative hypotese. Men det kræver altså, at vi kan genkende en typisk tilfældig variation. For at finde ud af hvordan den tilfældige variation opfører sig kan vi nu enten udnytte et rent teoretisk kendskab til den tilfældige variation eller vi kan simulere nulhypotesen og dermed rent eksperimentelt afgøre, hvordan den tilfældige variation opfører sig, herunder hvad der er den typiske forventede opførsel. Her vil vi se på begge angrebsvinkler.

Først den *eksperimentelle undersøgelse*. For at *simulere nulhypotesen* skal vi kunne simulere *uafhængighed af to stokastiske variable*, da vi jo netop tester for uafhængighed. Men det foregår på præcis den samme måde uafhængigt af om de stokastiske variable er kategoriske eller numeriske. Man foretager simpelthen en *omrøring*, dvs. bytter tilfældigt om på rækkefølgen af værdierne for den ene stokastiske variable, og bryder derved enhver sammenhæng den måtte have haft med den anden stokastiske variabel. I vores tilfælde laver vi derfor en omrøring på den afhængige variabel **genstande**. Det gøres ved hjælp af kommandoen for en dynamisk stikprøve, der hedder **randsamp()**. Vi slår kommandoen op i kataloget i venstrepanelet, og ser at vi skal oplyse hvilken liste vi vil udtage stikprøven fra, hvor stor den skal være og om den skal være med eller uden tilbagelægning. I vores tilfælde skal det være en tilfældig permutation af elementerne i listen for **genstande**, dvs. det skal netop være en stikprøveudtagning *uden tilbagelægning*, og vi skal netop trække det samme antal elementer i stikprøven, som i den oprindelige liste.



Vi indsætter derfor en ny søjle ved siden af søjlen for genstande og udfører kommandoen

D	**omrøring**:=randsamp(**genstande**,dim(**genstande**),1)

	A køn	B alder	C genstande	D omrøring
◆				=randsamp(gens
1	Dreng	15	0	8
2	Dreng	15	4	2
3	Dreng	15	3	9
4	Dreng	16	7	9
5	Dreng	16	9	7
6	Dreng	16	5	4
7	Dreng	16	4	0
8	Dreng	17	8	8
9	Dreng	17	7	12
10	Dreng	17	8	6
11	Dreng	18	8	3
12	Dreng	18	12	4
13	Pige	15	5	5
14	Pige	15	6	5
15	Pige	15	2	8
16	Pige	16	9	5
17	Pige	16	5	5
18	Pige	18	5	7

Derved får vi mulighed for at udføre en lineær regression på de *simulerede data*, dvs. med **alder** ud af *x*-aksen og **omrøring** op af *y*-aksen. Da der er tale om den samme uafhængige variabel, **alder**, kunne vi have afsat simuleringen i det samme koordinatsystem ved at højreklikke og vælge **Tilføj Y-**

variabel, men for tydeligere at kunne skelne mellem det *observerede datasæt* og det *simulerede datasæt* har vi afbildet dem i hvert deres grafrum. Yderligere har vi farvelagt dem, så de data, der hører til de oprindelige observationer fremstår rødt, mens de data, der hører til simuleringen fremstår med blåt. Når vi befinder os i regnearket kan vi nu gentage simuleringen ved at taste **CTRL-R** (for recalculate). De typiske hældninger for simuleringen ses da at ligge meget tættere på nul, som er den forventede værdi for hældningen, hvis de to variable virkeligt var uafhængige.

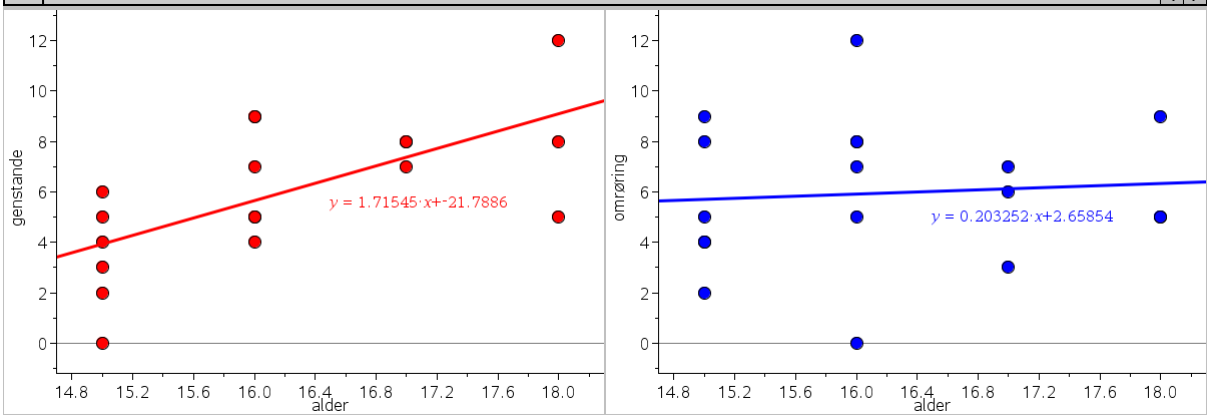
J
obs_hæld=
1.71545
sim_hæld=
0.203252

J2	obs_hæld:=f3
J4	sim_hæld:=i3

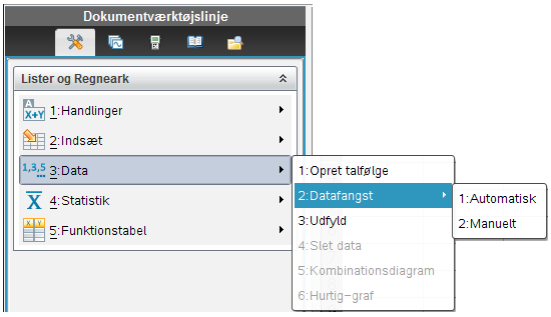
Det er *den simulerede hældning*, vi skal have opsamlet. I den næste søjle indfører vi derfor en variabel **test**, og vælger **Data**-menuen **Automatisk Datafangst**:

	A køn	B alder	C genstande	D omrøring	E	F	G	H	I	J	K	L	M	N
◆				=randsamp(gens		=LinRegMx('alc			=LinRegMx('alde					
1	Dreng	15	0	4	Titel	Lineær regres...	Restvariation=	Titel	Lineær regress...					
2	Dreng	15	4	4	RegEqn	m*x+b	80.6179	RegEq...	m*x+b					
3	Dreng	15	3	5	m	1.71545	Totalvariation=	m	0.203252					
4	Dreng	16	7	8	b	-21.7886	140.944	b	2.65854					
5	Dreng	16	9	5	r²	0.428017	Forklaringsgrad=	r²	0.006009					
6	Dreng	16	5	12	r	0.65423	0.428017	r	0.077515					
7	Dreng	16	4	0	Resid	{-3.94308943...		Resid	{-1.707317073...					
8	Dreng	17	8	3										
9	Dreng	17	7	7										
10	Dreng	17	8	6										
11	Dreng	18	8	5										
12	Dreng	18	12	9										

I9



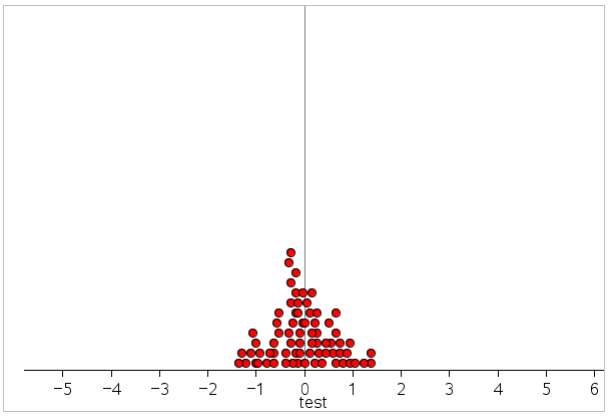
For nu at få en tydelige fornemmelse for hvor typisk den observerede variation er, vil vi *måle* den simulerede hældning og foretage en opsamling af målingerne ved hjælp af datafangst, så vi kan se, hvordan de simulerede hældninger fordeler sig, når vi gentager simuleringen rigtigt mange gange. Vi bruger altså *hældningen for regressionslinjen* som *teststørrelsen*. Vi skal da først have gemt hældningerne som variable. Det sker i celler, som ligger *uden for* søjlerne med de lineære regressioner, da vi ellers lukker for de lineære regressioner, så de ikke længere opdateres. Vi vælger derfor den første ledige søjle og kopierer cellerne med *den observerede hældning* henholdsvis *den simulerede hældning* over i to ledige celler. Derefter lægger vi de to værdier på lager ved at højreklikke i cellerne og vælge menupunktet **Gem var**. Som navne for de gemte variable bruger vi da **obs_hæld** og **sim_hæld**:



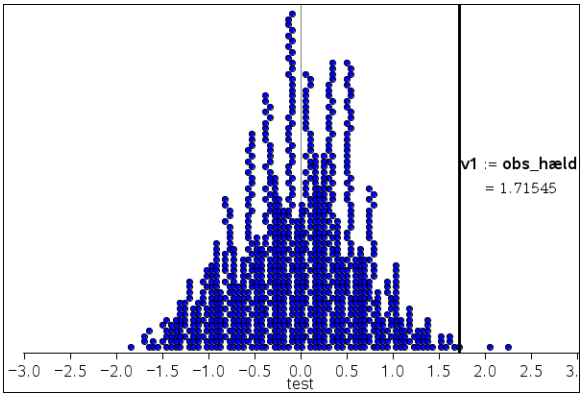
K
test
=capture('si
0.203252

K	test:=capture('sim_hæld,1)

Derefter er vi klar til at se nærmere på fordelingen af teststørrelsen **test**. Vi opretter derfor et nyt **Diagrammer og statistik**-værksted, afbilder variablen **test** på x-aksen, vender tilbage til regnearket og trykker **CTRL-R** gentagne gange. Vi ser da fordelingen langsomt dukke op:



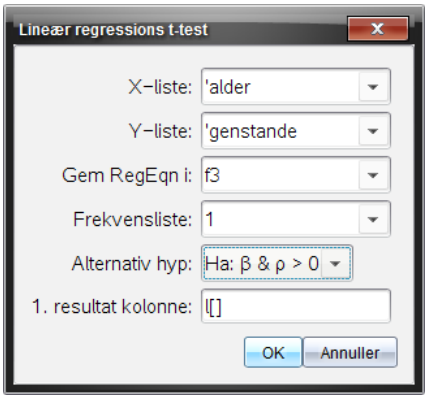
Vi kan nu tilpasse grafvinduet, skifte farve til blå for simuleringen og sætte en *tæller* på, så vi ved, hvor mange målinger, vi har foretaget af teststørrelsen - det sker ved hjælp af kommandoen dim(**test**). Vi tilføjer også med **Plot værdi** den *observerede hældning*. Efter 1000 målinger ser det således ud:



Det er tydeligvis kun meget få simuleringer, i dette tilfælde 3 simuleringer, der har en hældning, der er mindst lige så stor som den observerede. Vi finder altså en *p*-værdi på 3‰ eller 0.3%. *Den observerede hældning er derfor ikke typisk* for en tilfældig variation og vi kan roligt forkaste nulhypotesen om uafhængighed mellem de to variable.

Vi kan også foretage en **teoretisk test**. Det går naturligvis meget hurtigere end når man udfører en eksperimentel simulering af nulhypotesen – men til gengæld kan den være lidt sværere at forstå.

Vi vender derfor tilbage til regnearket og vælger en **lineær regressions t-test** fra **Statistik**-menuen:



Som den alternative hypotese til uafhængigheden, dvs. Ha, vælger vi en *voksende sammenhæng*. De græske bogstaver β og ρ er standardbetegnelser i teoretisk statistik. Ligningen for regressionsligningen skrives $y = a + b \cdot x$ (også i den danske faglitteratur!), så hældningen hedder *b*, men derved kommer den sande værdi for hældningen til at hedde β (beta) som er den græske variant af det latinske bogstav *b*. Korrelationskoefficienten hedder *r*, men derved kommer den sande værdi for korrelationskoefficienten til at hedde ρ (rho) som er den græske variant af det latinske bogstav *r*. Den alternative hypotese

Ha: β & ρ > 0

er altså blot en kompakt skrivemåde for at vi tester nulhypotesen om uafhængighed mod en positiv hældning eller - hvad der er det samme – mod en positiv korrelation.

L	M
	=LinRegtTest(
Titel	Lineær Reg ...
Alternativ...	β & ρ > 0
RegEqn	a+b*x
t	3.46018
PVal	0.001612
df	16.
a	-21.7886
b	1.71545
s	2.24469
SESlope	0.495768
r²	0.428017
r	0.65423
Resid	{-3.9430894...

Vi får da de viste oplysninger fra den lineære regressionstest. En del af oplysningerne er velkendte: Koefficienterne til regressionsligningen a og b, forklaringsgraden r^2, korrelationskoefficienten r samt residualerne. Men vi får også en række nye specifikt statistiske oplysninger:

t-værdien:	3.46
p-værdien:	0.16%
df-værdien:	16
s-værdien:	2.24
SESlope:	0.50

1) **t-værdien** er navnet på teststørrelsen i en t-test. I vores tilfælde er den tæt knyttet til hældningen for regressionslinjen, men af tekniske grunde ønsker man at normere teststørrelsen, så man umiddelbart kan sammenligne den med en universel standardfordeling. Den korte version er derfor, at man finder t-værdien ved at dividere hældningen med usikkerheden på hældningen, dvs. SESlope (Standard Error Slope):

$$t = \frac{b}{SESlope}$$

Det er nemt at kontrollere i regnearket:

L	M	N
	=LinRegtTest(	
Titel	Lineær Reg ...	
Alternativ...	β & ρ > 0	
RegEqn	a+b*x	
t	3.46018	3.46018
PVal	0.001612	
df	16.	
a	-21.7886	
b	1.71545	
s	2.24469	
SESlope	0.495768	

N4	$=\frac{m8}{m10}$

Vi uddyber formlen om lidt!

2) **p-værdien** er sandsynligheden for at finde en test-værdi (dvs. i realiteten en simuleret hældning), der er mindst lige så stor som den observerede. Den ligger rimeligt tæt på vores estimerede p-værdi på 0.3%. Selv med et signifikansniveau på 1% fører den til forkastelse af nulhypotesen.

3) **df-værdien** eller Degrees of Freedom er *antallet af frihedsgrader* i t-testen. Da der er tale om en lineær regressionstest er antallet af frihedsgrader 18-2 = 16, idet vi mister 2 frihedsgrader, fordi vi er nødt til at estimere såvel hældningen som begyndelsesværdien for den rette linje. Frihedsgrader i et t-test fungerer altså på mange måder ligesom frihedsgraderne i et chi2-test.

4) **s-værdien** er *residualspredningen*, dvs. usikkerheden på y-værdien, når denne regnes ud fra mindste kvadraters linje. Den hedder også RMS eller Root Mean Square og udregnes af mange naturvidenskabelige programmer. Den er ikke så svær at stille en formel op for. Blot skal man huske på, at når man midler over de kvadratiske afvigelser skal man dividere med n-2 (dvs. med antallet af frihedsgrader). Formlen for residualspredningen er altså givet ved

$$Residualspredning = RMS = \sqrt{\frac{Restvariationen}{n-1}}$$

Igen er det nemt at kontrollere i regnearket:

	L	M	N
◆	re('si	=LinRegtTest(	
1	3252	Titel	Lineær Reg ...	
2	1626	Alternativ...	β & ρ > 0	
3	6748	RegEqn	a+b*x	
4	2276	t	3.46018	3.46018
5	8618	PVal	0.001612	
6	3496	df	16.	
7	4959	a	-21.7886	
8	8699	b	1.71545	
9	8211	s	2.24469	2.24469

N9	$=\sqrt{\frac{\text{sum}\left(\left(f1(\textbf{alder})-\textbf{genstande}\right)^2\right)}{m6}}$

Læg mærke til at *residualspredningen ikke* afhænger af *totalvariationen*. I modsætning til forklaringsgraden er *residualspredningen* altså ikke et egentligt kvalitetsmål for modellen. Men den fortæller noget om den typiske usikkerhed på *y*-værdierne, og det kan også i mange sammenhænge være en nyttig information. I praksis kan man imidlertid få præcis den samme information fra *residualplottet*, hvor man uden videre kan aflæse størrelsesordenen for den typiske usikkerhed. Residualplottet giver så samtidigt mulighed for at vurdere om modellen er *udtømmende* eller om der er *en systematisk restvariation*, vi ikke har fået gjort rede for i modellen. Men hvis man er vant til at bruge Root Mean Square i forbindelse med lineær regression, så har man den altså til rådighed!

5) **SESlope** = StandardErrorSlope er usikkerheden eller spredningen på estimatet for hældningen. Den kan udregnes ved en forholdsvis simpel formel, der imidlertid ikke er helt nem at begrunde. Løst sagt så afhænger usikkerheden eller spredningen på hældningen af forholdet mellem spredningerne i lodret retning og vandret retning. Ydermere skal man dividere med kvadratroden af stikprøvens størrelse: Jo større stikprøven er, jo mere præcis bliver hældningen fastlagt. Det viser sig alt i alt at føre til formlen:

$$SESlope = \frac{Residualspredningen}{\sqrt{sum\left(\left(alder - mean(alder)\right)^2\right)}}$$

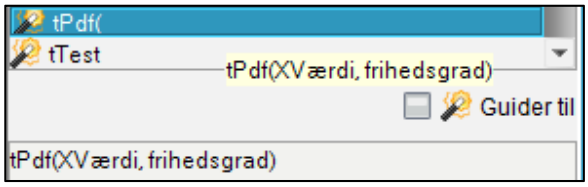
Igen kan man kontrollere formlen i regnearket:

L	M	N
	=LinRegtTest(	
Titel	Lineær Reg ...	
Alternativ...	β & ρ > 0	
RegEqn	a+b*x	
t	3.46018	3.46018
PVal	0.001612	
df	16.	
a	-21.7886	
b	1.71545	
s	2.24469	2.24469
SESlope	0.495768	0.495768

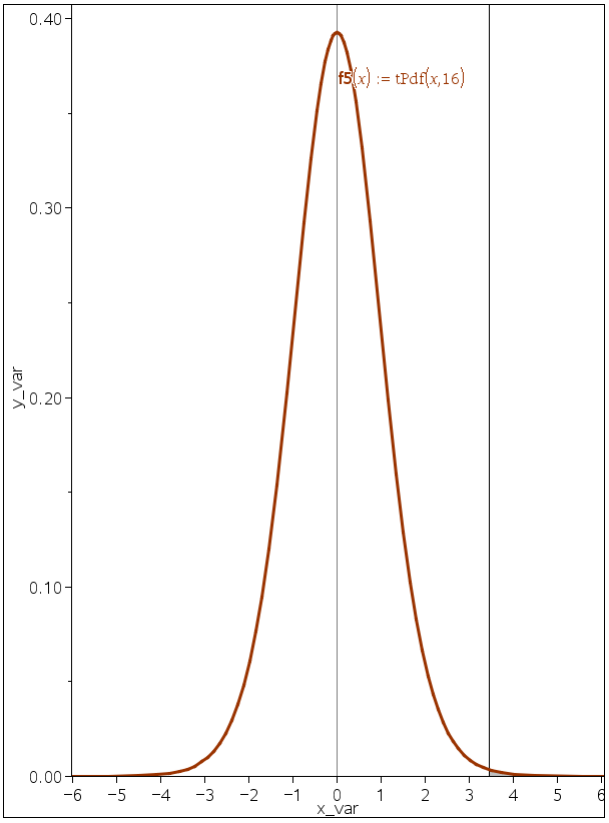
N10	$=\frac{n9}{\sqrt{\text{sum}\left(\left(\textbf{alder}-\text{mean}(\textbf{alder})\right)^2\right)}}$

Vi uddyber nu **p-værdien** og **t-teststørrelsen**.

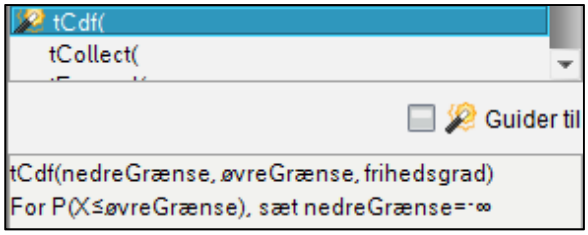
Først **p-værdien**. Hertil skal vi bruge den teoretiske fordeling af t-teststørrelsen. Vi vil illustrere p-værdien grafisk, så vi åbner et nyt diagrammer og statistik-værksted med et tomt koordinatsystem på basis af to tomme variable: **x_var**:={ } og **y_var**:={ } . Herefter vælger vi **Plot funktion** i **Undersøg data**-menuen. Vi finder t-fordelingen i kataloget. Vi skal bruge tæthedsfunktionen **tPdf()**, dvs. tPointDistributionFunction. Vi skal da oplyse navnet på den uafhængige variabel, her x, og antallet af frihedsgrader, her 16.



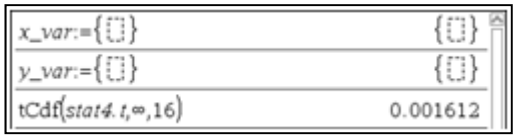
Derefter justerer vi grænserne for grafrummet og finder følgende graf for fordelingsfunktionen. Samtidigt har vi tilføjet teststørrelsen ved at plotte værdien for stat4.t. Vi kan så højreklikke på grafen, vælge **Skraver under funktion**, og udpege grænserne som den lodrette linje ved den observerede t-værdi og boksen med +∞. Vi finder da netop arealet under grafen svarende til de *x*-værdier, der ligger længere ude end den observerede (dvs. hvor den simulerede hældning er større end den observerede). Dette areal er givet ved 0.0016. Dermed har vi fundet en grafisk illustration af p-værdien.



Hvis vi bare vil beregne p-værdien sker det ved brug af den kumulerede fordelingsfunktion. Igen kan vi benytte kataloget.



Som standard foreslås det at sætte den nedre grænse til -∞, hvilket giver den traditionelle kumulerede fordeling, men vi skal bruge intervalfordelingen, så vi sætter den nedre grænse til stat4.t og sætter den øvre grænse til ∞, mens frihedsgraderne som før sættes til 16:



Resultatet er selvfølgelig det samme som før!

Vi vender os derefter mod **t-teststørrelsen**. Hvordan skal den forstås? Her kan vi ikke gå ind i alle de teoretiske detaljer, men må nøjes med at antyde: t-teststørrelsen blev indført af kemikeren Gosset i 1908 i forbindelse med estimation af middelværdier for små stikprøver, hvor spredningen var ukendt. Hvis bare spredningen var kendt på forhånd kunne man teste en kendt middelværdi μ_0 ved at sammenligne den observerede middelværdi $\overline{X}$ med den forventede værdi μ_0. Det er da nærliggende at benytte teststørrelsen

$$z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

hvor σ er spredningen for den stokastiske variabel X og $\sigma / \sqrt{n}$ dermed spredningen for middelværdien $\overline{X}$ den såkaldte standardfejl (SE = Standard Error). Under almene forudsætninger kan man så med god tilnærmelse regne med at teststørrelsen z er standardnormalfordelt med middelværdien 0 og spredningen 1. Det gør det nemt at vurdere troværdigheden af nul-hypotesen: Mellem -2 og 2, dvs. indenfor to standardafvigelser (normalområdet), er den oplagt troværdig, og udenfor intervallet fra -3 til 3, dvs. udenfor tre standardafvigelser (det exceptionelle område) er den oplagt utroværdig. Men ingen vidste hvad man skulle gøre, hvis man ikke kendte spredningen på forhånd? Gosset foreslog nu i 1908, at man i stedet skulle erstatte den ukendte spredning σ med et estimat s for spredningen. Derved fik man en ny teststørrelse

$$t = \frac{\overline{X} - \mu_0}{s / \sqrt{n}}$$

der ikke længere var standardnormalfordelt og hvor fordelingen oven i købet afhang af antallet af frihedsgrader *n*. Men der var ligesom med χ^2-fordelingen stadigvæk tale om en universel fordeling, der kunne tabellægges og bruges til statistiske test. I vores tilfælde er det hældningen, vi ønsker at vurdere. Den har den forventede værdi 0, så vi skal ikke trække noget fra i tælleren, men vi skal selvfølgelig stadigvæk dividere med standardfejlen for hældningen (der i sig selv er et vægtet gennemsnit af de enkelte hældninger målt ud fra tyngdepunktet).

1.3 Simulering af t-fordelingen for en lineær regressionstest *

Sammenhængen mellem det eksperimentelle test, der bygger på den simulerede hældning, og det teoretiske test, der bygger på t-værdien, kan nu også uddybes, om end det ikke er helt så simpelt. Vi indfører en ny teststørrelse i simuleringen svarende til t-værdien, dvs. vi erstatter den simulerede hældning **sim_hæld** med den simulerede t-værdi **sim_t** beregnet ud fra formlen **sim_hæld/sim_SEslope**. Her må vi altså både beregne den simulerede hældning såvel som den simulerede standardfejl for hældningen ud fra den omrørte lineære regression, der er gemt i funktionen **f2**. Det kræver, at vi først udregner værdien af den simulerede standardfejl og gemmer værdien af den simulerede t-værdi i en variabel, så vi kan referere til den indenfor regnearket. Dernæst indfører vi en ny søjlevariabel med navnet **t_test**:

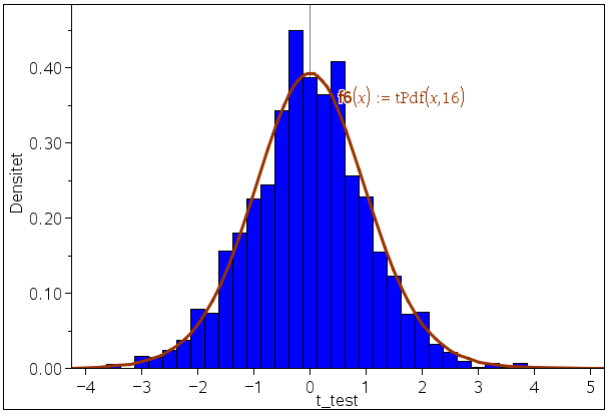
J
sim_hæld=
0.788618
sim_SEslope=
0.625171
sim_t=
1.26144
Antal mål.=
2500

L	t_test
	=capture('sim_t,1)
	-0.436665
	-1.87689
	0.386265
	-0.136506
	-1.23293
	-0.902721
	0.692276
	-0.744305
	-0.361134
	-1.23293

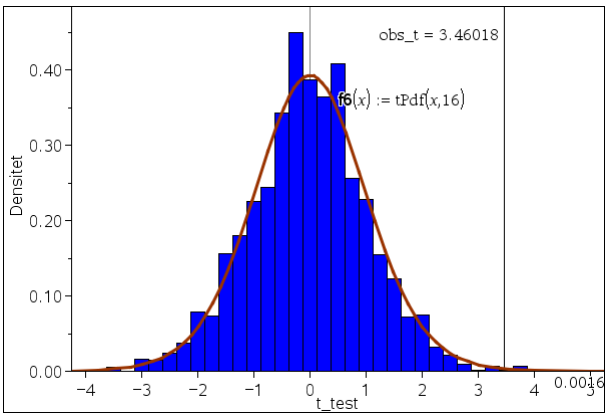
$$J6 \quad \text{sim_seslope} := \sqrt{\frac{\text{sum}\left(\frac{\left(f2(\text{alder})-\text{omrøring}\right)^2}{16}\right)}{\text{sum}\left(\left(\text{alder}-\text{mean}(\text{alder})\right)^2\right)}}$$

$$J8 \quad \text{sim_t} := \frac{\text{sim_hæld}}{\text{sim_seslope}}$$

Derefter afbilder vi **t_test** i et **Diagrammerog statistik**-værksted, skifter til et histogram og sætter skalaen til densitet (ved at højreklikke på histogrammet):



Vi få så nogenlunde frembragt en symmetrisk fordeling omkring 0. Det er denne fordeling vi vil sammenligne med den teoretiske tæthedsfordeling. Vi tilføjer derfor grafen for t-fordelingens tæthedsfunktion **tPdf**(x,16) og ser at de følges så nogenlunde ad. Herefter kan vi præcis som før afbilde den observerede t-værdi og udregne arealet for den del af grafen, der ligger til højre for den observerede t-værdi.

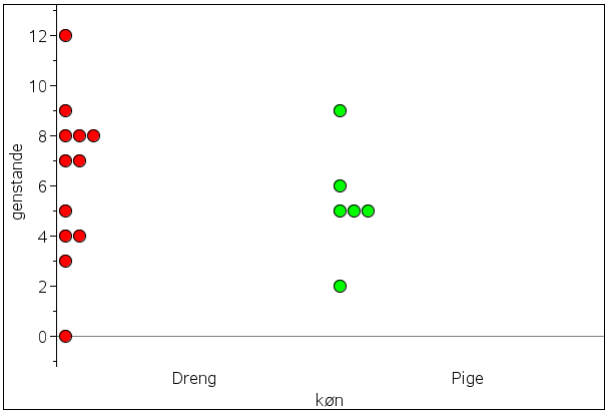


Som det se, er det dog svært at få øje på det skraverede areal med værdien 0.0016 i det nederste højre hjørne, da det fylder så lidt.

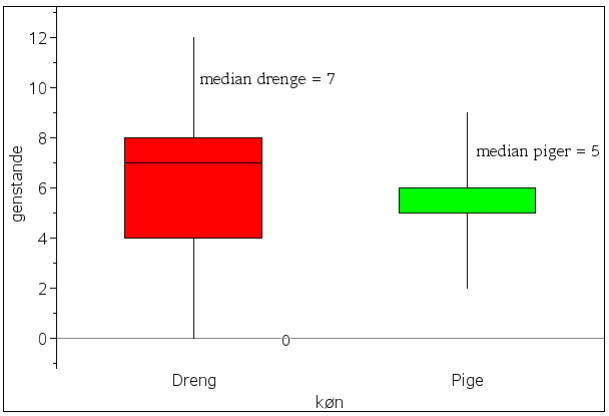
2. Unges alkoholforbrug som funktion af køn: t-test for to middelværdier

2.1 t-test for to middelværdier: Den beskrivende statistik

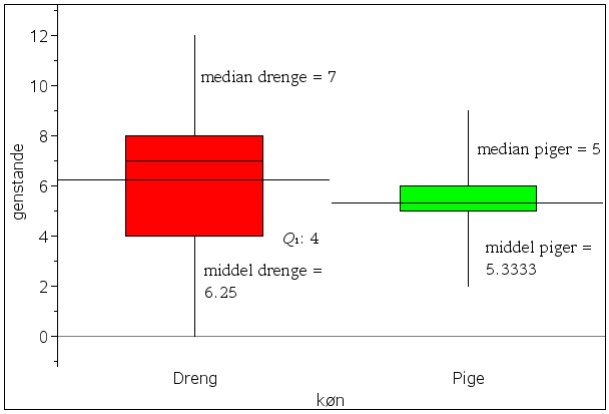
Vi skifter nu spor til en undersøgelse af sammenhængen mellem alkoholforbruget og kønnet. Dvs. denne gang undersøger vi sammenhængen mellem en kategorisk og en numerisk variabel. Igen lægger vi ud med en *beskrivende* undersøgelse af datasættet (deskriptiv statistik). For at undersøge om der kan være tale om en sammenhæng mellem **køn** og **genstande** afbilder vi nu de to variable i et **prikplot**. Vi tilføjer derfor et **Diagrammer og statistik**-værksted og afsætter **køn** vandret og **genstande** lodret:



Vi ser da de to fordelinger af alkoholforbruget for henholdsvis drenge og piger. Allerede af prikplottet synes det som om niveauet for drengenes alkoholforbrug ligger højere end for pigernes, ligesom det ser ud som om drengenes alkoholforbrug har en større spredning. Vi kan omsætte prikplottet til et **boksplot** og skifte til et sammenhængende boksplot ved at højreklikke et par gange:



Vi har tilføjet et par tekstbokse for at fastfryse informationen om drengenes median på 7 og pigernes median på 5. Vi kan også afbilde middelværdien ved at vælge **Plot værdi** i **Undersøg data**-menuen. Vi ser da at drengenes niveau også ligger højere end pigernes når vi måler på middelværdien, om end forskellen er skrumpet lidt ind.



Konklusionen på den beskrivende statistik er altså at i den pågældende klasse er drengenes niveau for alkoholforbruget højere end pigernes ligesom drengenes variation i alkoholforbruget er større end variationen hos pigerne.

Hvis vi ønsker at understøtte den beskrivende statistik med beregninger i regnearket løber vi ind i det problem at enkeltvariabel-statistikken skal bruge lister med de kategorier vi ønsker at splitte efter. Vi tilføjer derfor to nye kategoriske variable **dreng** og **pige** som indeholder kategorierne "Dreng" og "Pige":

	A køn	B alder	C genstande	D dreng	E pige
◆					
1	Dreng	15	0	Dreng	Pige
2	Dreng	15	4		
3	Dreng	15	3		
4	Dreng	16	7		
5	Dreng	16	9		
6	Dreng	16	5		
7	Dreng	16	4		
8	Dreng	17	8		
9	Dreng	17	7		
10	Dreng	17	8		
11	Dreng	18	8		
12	Dreng	18	12		
13	Pige	15	5		
14	Pige	15	6		
15	Pige	15	2		
16	Pige	16	9		
17	Pige	16	5		
18	Pige	18	5		

Så snart vi har tilføjet disse splitvariable kan vi udføre enkeltvariabelstatistik med variablene **dreng** og **pige** som filtre. De fleste oplysninger giver sig selv. Vi får blandt andet oplyst middelværdierne $\bar{x}$, medianerne og variationerne SSX for de to variable.

F	G	H
	=OneVar('genstande,1,'	=OneVar('genstande,1,'
Titel	Statistik med én varia...	Statistik med én varia...
$\bar{x}$	6.25	5.33333
Σx	75.	32.
Σx^2	581.	196.
$sx := s_{n-1}x$	3.19446	2.25093
$\sigma x := \sigma_n x$	3.05846	2.0548
n	12.	6.
MinX	0.	2.
Q_1X	4.	5.
MedianX	7.	5.
Q_3X	8.	6.
MaxX	12.	9.
$SSX := \Sigma(x-\bar{x})^2$	112.25	25.3333

Læg mærke til, at der oplyses to spredningsmål:

Populationsspredningen σx, der udregnes ved at midle med *n*, dvs. ud fra formlen $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$.

Stikprøvespredningen sx, der udregnes ved at midle med *n*−1, dvs. ud fra formlen $s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$.

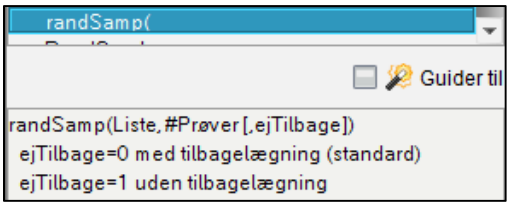
2.2 t-test for to middelværdier: Den bekræftende statistik

Vi går derefter over til den **bekræftende statistik**, dvs. vi stiller nu spørgsmålet hvorvidt det observerede mønster kan udvides til en større population, hvor den pågældende klasse kan betragtes som en repræsentativ stikprøve for en større population, fx populationen af elever i gymnasiet generelt på det pågældende klassetrin. Læg godt mærke til repræsentativiteten. De pågældende elever er jo næppe udvalgt tilfældigt blandt alle landets elever – så repræsentativiteten er noget man må argumentere for ad anden vej.

I den bekræftende statistik skal vi nu afgøre hvorvidt den observerede variation ligner en typisk tilfældig variation eller ej. Hvis den ikke ligner en tilfældig variation må vi forkaste nulhypotesen og vælge at tro på den systematiske variation, dvs. vi må vælge den alternative hypotese. Men det kræver altså, at vi kan genkende en typisk tilfældig variation. For at finde ud af hvordan den tilfældige variation opfører sig kan vi enten udnytte et rent teoretisk kendskab til den tilfældige variation eller vi kan simulere nulhypotesen og dermed rent eksperimentelt afgøre, hvordan den tilfældige variation opfører sig, herunder hvad der er den typiske forventede opførsel. Her vil vi se på begge angrebsvinkler:

Først den eksperimentelle undersøgelse: For at simulere nulhypotesen skal vi kunne simulere uafhængighed af to stokastiske variable, da vi jo netop tester for uafhængighed. Men det foregår på den samme måde uafhængigt af om de stokastiske variable er kategoriske eller numeriske. Man foretager en omrøring, dvs. bytter tilfældigt om på rækkefølgen af værdierne for den ene stokastiske variable, og bryder derved enhver sammenhæng den måtte have haft med den anden stokastiske variabel.

I vores tilfælde laver vi derfor en omrøring på variablen **genstande**. Det gøres ved hjælp af kommandoen for en dynamisk stikprøve, der hedder **randsamp**(). Slår vi den op i kataloget i venstrepanelet ser vi derfor, at vi skal altså oplyse hvilken liste vi vil udtage stikprøven fra, hvor stor den skal være og om den skal være med eller uden tilbagelægning.

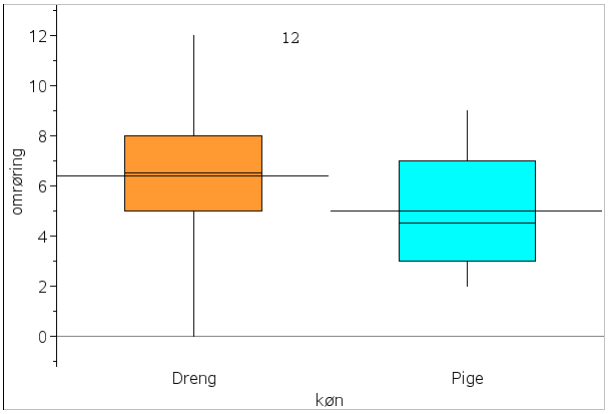


I vores tilfælde skal det være en tilfældig permutation af elementerne, dvs. det skal være uden tilbagelægning, og vi skal netop trække det samme antal elementer i stikprøven, som i den oprindelige liste. Vi indsætter derfor en ny søjle ved siden af søjlen for genstande og udfører kommandoen

D	**omrøring**:=randsamp(**genstande**,dim(**genstande**),1)

	A køn	B alder	C genstande	D omrøring
◆				=randsamp(gens
1	Dreng	15	0	8
2	Dreng	15	4	2
3	Dreng	15	3	9
4	Dreng	16	7	9
5	Dreng	16	9	7
6	Dreng	16	5	4
7	Dreng	16	4	0
8	Dreng	17	8	8
9	Dreng	17	7	12
10	Dreng	17	8	6
11	Dreng	18	8	3
12	Dreng	18	12	4
13	Pige	15	5	5
14	Pige	15	6	5
15	Pige	15	2	8
16	Pige	16	9	5
17	Pige	16	5	5
18	Pige	18	5	7

Vi kan så oprette et nyt **Diagrammer og statistik**-værksted baseret på de *omrørte data*.



Man skal ikke simulere ret mange gange for at se at det er nemt at opnå en lige så stor forskel i niveau ud fra nulhypotesen. Den observerede forskel ligner altså rent faktisk en typisk tilfældig variation, og denne gang kan vi derfor *ikke* forkaste nulhypotesen.

Men derfor kan vi jo godt gennemføre en eksperimentel test af uafhængighed alligevel. For nu at få en tydeligere fornemmelse for hvor typisk den observerede variation er, vil vi måle den simulerede forskel på niveauerne (målt på middelværdierne) og foretage en opsamling af målingerne med datafangst, så vi kan se hvordan de simulerede niveauforskelle fordeler sig, når vi gentager simuleringen rigtigt mange gange. Vi bruger altså såvel de observerede som de simulerede niveauforskelle i form af forskellen mellem de to middelværdier. De udregnes nemmest ved som vist at trække data for drenge og piger ud som særskilte lister. Forskellene lægges på lager ved fx at højreklikke i cellerne og vælge **Gem var**. Som navne bruger vi da **obs_dif** og **sim_dif**:

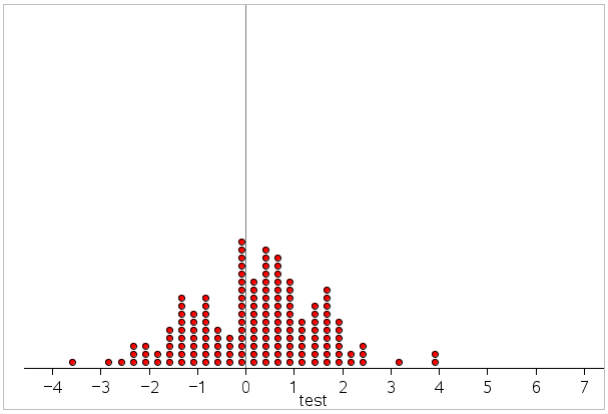
	E
1	obs_dif=
2	0.916667
3	sim_dif=
4	1.41667

E2	**obs_dif**:=(mean(**d_genstande**)−mean(**p_genstande**))·1.
E4	**sim_dif**:=(mean('**d_omrørt**)−mean('**p_omrørt**))·1.

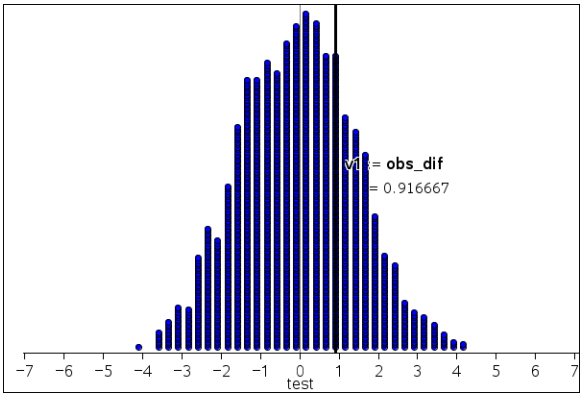
F d_genstande	G p_genstande	H d_omrørt	I p_omrørt
=c1:c12	=c13:c18	=d1:d12	=d13:d18
0	5	8	7
4	6	6	9
3	2	5	2
7	9	9	4
9	5	4	3
5	5	8	5
4		5	
8		0	
7		5	
8		7	
8		12	
12		8	

Igen har vi farvekodet tabellen, så den er i overensstemmelse med graferne. Vi kan så udføre en automatisk datafangst på de simulerede forskelle ved at indsætte en ekstra søjle og kalde den **test** og vælge **Automatisk datafangst** i **Data**-menuen.

F test
=capture('sim_dif,1)
1.41667



Vi kan nu tilpasse grafvinduet, skifte farve til blå for simuleringen og sætte en tæller på, så vi ved, hvor mange målinger vi har foretaget af teststørrelsen (det sker ved hjælp af kommandoen dim(**test**)). Vi tilføjer også med plot værdi den observerede hældning.



Efter 2000 målinger ser vi da, at en betydelig del af målingerne har en større niveauforskel end den faktisk observerede, så vi kan ikke på baggrund af observationerne forkaste nulhypotesen. Vi kan altså godt på det foreliggende grundlag acceptere at alkoholforbruget er uafhængigt at kønnet – men vi kan også se, at med så lille en stikprøve kan det være meget svært at forkaste nulhypotesen – også selvom den observerede forskel rent faktisk skulle vise sig at være reel. Det er ikke sikkert testen er stærkt nok til at kunne se forskel på nulhypotesen og den alternative hypotese. Det skal vi vende tilbage til.

Vi kan optælle de skæve simuleringer, der er mindst lige så store som den observerede. Derved kan vi estimere p-værdien ud fra vores simulering. Vi ser da, at p-værdien estimeres til 28.7%, så nulhypotesen har rimelig høj troværdighed!

E
obs_dif=
0.916667
sim_dif=
-1.83333
Antal mål.=
2000
Antal skæve=
574
p-værdi i %=
28.7

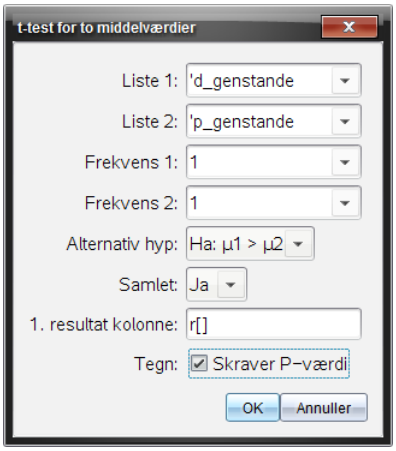
E6	=dim(**test**)

E8	=countif(**test**,?≥**obs_dif**)

E10	$=\frac{e8}{e6}\cdot 100.$

Vi kan også foretage en **teoretisk test**. Det er naturligvis endnu hurtigere, men til gengæld kan det være lidt sværere at forstå. Vi vender derfor tilbage til regnearket og vælger denne gang en t-test for to middelværdier:

Først skal vi vælge om vi vil teste på de rå data eller om vi vil oplyse de statistiske nøgletal direkte. Her vil vi vælge de rå data (det andet er nyttigt i opgavesituationer, hvor man kun får oplyst de statistiske nøgletal, dvs. middelværdier, stikprøvespredninger og størrelsen af stikprøverne). Vi skal så indtaste de relevante oplysninger:



Først skal vi oplyse navnene på de to lister, som vi ønsker at sammenligne. Igen er det altså nødvendigt at splitte den oprindelige liste **genstande** i to dellister, én delliste **d_genstande** hørende til **drengenes forbrug** og én delliste **p_genstande** hørende til **pigernes forbrug**. Så kan vi om nødvendigt tilføje hyppighedslister. I dette tilfælde lader vi den alternative hypotese være Ha: $\mu 1 > \mu 2$, idet vi har en formodning/fordom om at drengenes niveau godt kunne være højere end pigernes. Så skal vi svare på et lidt mindre gennemskueligt spørgsmål: Samlet? Det hentyder til at vi kan håndtere datasættet på to forskellige måder: Når vi under antagelse af nulhypotesen skal estimere spredningen for alkoholforbruget om uafhængighed kan man enten gøre det ud fra den samlede liste **genstande**, eller man kan gøre det ved at kombinere estimaterne fra de to dellister **d_genstande** og **p_genstande**. Langt det enkleste er at svare **Ja** til spørgsmålet Samlet?, dvs. sammenholde de to dellister med den forventede opførsel ud fra den samlede liste i stedet for at holde dem strengt adskilt. Endelig får vi mulighed for at illustrere resultatet grafisk ved at sætte hak i feltet **Skraver P-værdi**.

Resultatet er de følgende søjler i regnearket.

	Q	R
◆		=tTest_2Samp('d_ger
1	Titel	t-test med 2 prøver...
2	Alternativ Hyp	μ1 > μ2
3	t	0.625199
4	PVal	0.270328
5	df	16.
6	x̄1	6.25
7	x̄2	5.33333
8	sx1	3.19446
9	sx2	2.25093
10	sp	2.9324
11	n1	12.
12	n2	6.

Igen er de fleste felter rimeligt oplagte: Middelværdierne $\bar{x}_1, \bar{x}_2$, stikprøvespredningerne sx1, sx2 og størrelsen af stikprøverne n1 og n2. Resten vil vi nu kommentere:

Først er der **t-værdien**. Den er nærtbeslægtet med forskellen på niveauerne, dvs. differensen mellem middelværdierne. Men ligesom i det forrige t-test skal vi normalisere differensen ved at dividere med den forventede usikkerhed. Det viser sig at føre til følgende formel

$$t = \frac{\bar{x}_1 - \bar{x}_2}{sp \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

hvor sp er et mål for modellens residualspredning. Denne formel kan vi nemt kontrollere i regnearket

	Q	R	S
◆		=tTest_2Samp('d_ger	
1	Titel	t-test med 2 prøver...	
2	Alternativ Hyp	μ1 > μ2	
3	t	0.625199	0.625199

S3	$=\frac{r6-r7}{r10\cdot\sqrt{\frac{1}{r11}+\frac{1}{r12}}}$

Så er der **p-værdien**, dvs. sandsynligheden for at få en t-værdi, der er mindst lige så stor, forudsat de to variable er uafhængige. Den udregnes her til 27.0%, mens vi ved simuleringen fandt 28.7%.

Dernæst er der **antallet af frihedsgrader** df (= Degrees of Fredom), som er n-2, idet vi har mistet to frihedsgrader undervejs, da vi estimerede de to middelværdier (dvs. niveauerne for drenges henholdsvis pigers alkoholforbrug). I dette tilfælde er antallet af frihedsgrader derfor 16. Hvis man har

glemt at svare Ja til **Samlet** fås i stedet en skæv værdi for antallet af frihedsgrader!
Endelig er der det fælles samlede estimat sp for resdiualspredningen. Det udregnes ved en kombination af de to dellisters residualspredninger (ifølge sætningen om at en varians for en sum af to uafhængige stokastiske variable er det samme som summen af de to enkelte varianser). Men finder derved den følgende formel:

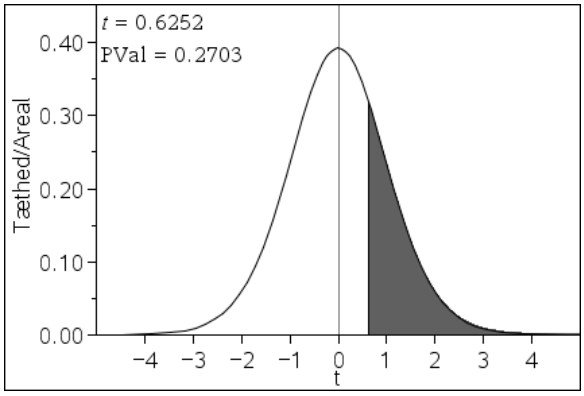
$$sp = \sqrt{\frac{(n_1 - 1)\cdot sx_1^2 + (n_2 - 1)\cdot sx_2^2}{n-2}}$$

Den kan igen kontrolleres i regnearket:

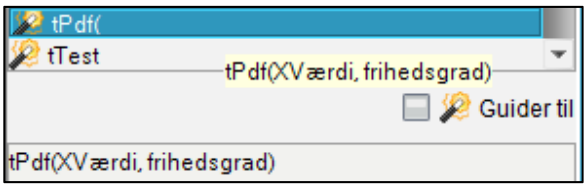
	Q	R	S
◆		=tTest_2Samp('d_ger	
1	Titel	t-test med 2 prøver...	
2	Alternativ Hyp	μ1 > μ2	
3	t	0.625199	0.625199
4	PVal	0.270328	
5	df	16.	
6	x̄1	6.25	
7	x̄2	5.33333	
8	sx1	3.19446	
9	sx2	2.25093	
10	sp	2.9324	2.9324
11	n1	12.	
12	n2	6.	

$$S10 \quad = \sqrt{\frac{(r11-1)\cdot r8^2+(r12-1)\cdot r9^2}{r11+r12-2}}$$

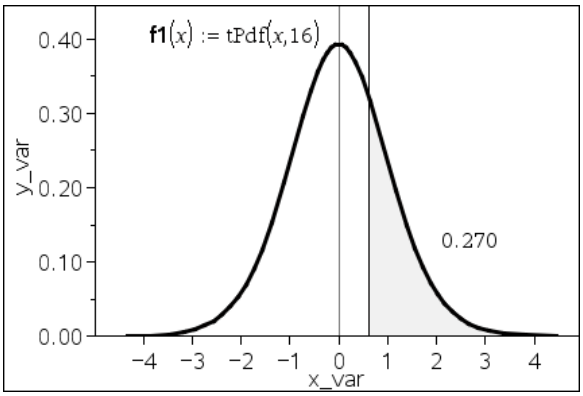
Vi får også foræret en grafisk illustration af testen:



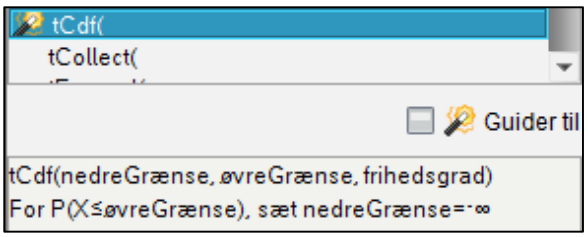
Vi uddyber nu **p-værdien** og **t-teststørrelsen**. Først **p-værdien**. Hertil skal vi bruge den teoretiske fordeling af t-teststørrelsen. Vi vil illustrere p-værdien *grafisk*, så vi åbner et nyt **Diagrammer og statistik**-værksted med et tomt koordinatsystem (på basis af to tomme variable: **x_var** := { } og **y_var** := { }). Herefter vælger vi **Plot funktion** i **Undersøg data**-menuen. Vi finder t-fordelingen i kataloget. Vi skal bruge tæthedsfunktionen **tPdf()**, dvs. tPointDistributionFunction:



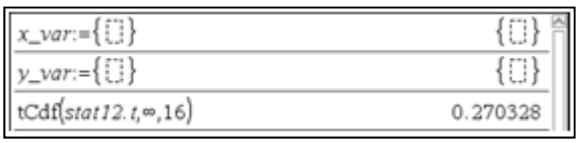
Vi indtaster navnet på den uafhængige variabel, dvs. x, og antallet af frihedsgrader, dvs. 16. Herefter justere vi grænserne for grafrummet og finder følgende graf for fordelingsfunktionen. Samtidigt har vi tilføjet teststørrelsen ved at vælge menupunktet **Plot værdi** i **Undersøg data**-menuen og indtaste navnet for t-værdien, dvs. her **stat12.t**. Vi kan så højreklikke på grafen, vælge **Skraver under funktion**, og udpege grænserne som den lodrette linje ved den observerede t-værdi og boksen med +∞. Vi finder da netop arealet under grafen svarende til de *x*-værdier, der ligger længere ude end den observerede (dvs. hvor den simulerede hældning er større end den observerede). Arealet er givet ved 0.2703, dvs. p-værdien er 27.03%. Dermed har vi selv fundet en grafisk illustration af p-værdien.



Hvis vi bare vil beregne den sker det ved brug af den kumulerede fordelingsfunktion, tCf(), som vi igen henter i kataloget.



Som standard foreslås det at sætte den nedre grænse til -∞, hvilket giver den traditionelle kumulerede fordeling. Men vi skal bruge intervalfordelingen, så vi ændrer den nedre grænse til **stat12.t** og sætter den øvre grænse til ∞, mens frihedsgraderne som før sættes til 16:



Resultatet er selvfølgelig det samme som før!

2.3 Simulering af t-fordelingen for en t-test med to middel værdier *

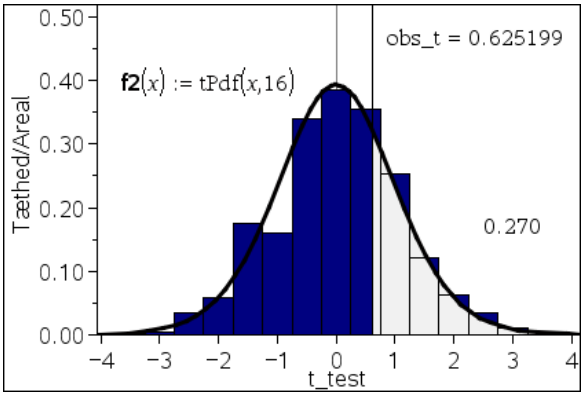
Sammenhængen mellem det eksperimentelle test, der bygger på den simulerede niveauforskel, og det teoretiske test, der bygger på t-værdien, kan nu også uddybes, om end det ikke er helt så simpelt. Vi indfører en ny teststørrelse **sim_t** i simuleringen svarende til t-værdien, dvs. vi erstatter den simulerede niveauforskel **sim_dif** med den simulerede t-værdi beregnet ud fra formlen **sim_dif/sim_SE**, hvor vi både må beregne den simulerede niveauforskel såvel som den simulerede standardfejl for niveauforskellen ud fra den omrørte stikprøve. Det kræver at vi først udregner værdien af den simulerede standardfejl og gemmer værdien af den simulerede t-værdi i en variabel, så vi kan referere til den indenfor regnearket. Dernæst indfører vi en ny søjlevariabel med navnet **t_test**.

	E
◆	
11	sim_se=
12	**1.47461**
13	sim_t=
14	**0.452096**
15	Antal t−mål.=
16	2500

	G t_test
◆	=capture(sim_t,1)
1	-0.224973
2	-0.394994
3	-0.05616
4	-0.742478
5	-0.567162
6	1.78293
7	-0.05616
8	-0.567162
9	3.94208
10	-1.93349

E14	$\text{sim_t} := \frac{\text{sim_dif}}{\text{sim_se}}$

Derefter afbilder vi **t_test** i et **Diagrammer og statistik**-værksted, skifter til et histogram og sætter skalaen til tæthed/Areal (ved at højreklikke på histogrammet).



Vi får så nogenlunde frembragt en symmetrisk fordeling omkring 0. Det er denne fordeling vi vil sammenligne med den teoretiske tæthedsfordeling. Vi tilføjer derfor grafen for tæthedsfunktionen med forskriften **tPdf**(x,16) og ser at de følges så nogenlunde ad. Herefter kan vi som vist afbilde den observerede t-værdi og udregne arealet for den del af grafen, der ligger til højre for den observerede t-værdi.

E12	$\text{sim_se} := \sqrt{\frac{(\dim(\text{d_omrørt})-1)\cdot(\text{stdevsamp}(\text{d_omrørt}))^2+(\dim(\text{p_omrørt})-1)\cdot(\text{stdevsamp}(\text{p_omrørt}))^2}{\dim(\text{d_omrørt})+\dim(\text{p_omrørt})-2}}\cdot\sqrt{\frac{1}{\dim(\text{d_omrørt})}+\frac{1}{\dim(\text{p_omrørt})}}\cdot 1.$

2.4 Hvor stor skal stikprøven være, hvis man skal kunne skelne nulhypotesen fra den alternative hypotese? *

Men hvad nu hvis den observerede forskel mellem piger og drenge rent faktisk var signifikant, men at t-testen bare var for svag til at se det, fordi stikprøven var for lille? Hvis vi nu ønskede at gentage spørgeskemaundersøgelsen, hvor mange skal vi så spørge for at have en rimelig chance for at kunne se forskellen som værende signifikant – hvis den altså rent faktisk er signifikant, hvilket vi jo ikke kan vide på forhånd?

Det kan man undersøge ved at skalere stikprøven op med udgangspunkt i den observerede fordeling af **genstande** og **køn**. Man tilføjer derfor hyppighedslister **d_genstande** og **p_genstande** med en **konstant hyppighed**. Derved holder man fordelingen konstant, men ændrer alene på stikprøvens størrelse. Et t-test med to prøver viser da, at en skalering med faktoren 7 gør forskellen signifikant på 5%-niveau.

Det svarer til 126 adspurgte. Det vil altså være en god ide at inddrage ca. 150 elever i den næste spørgeskemaundersøgelse, hvis vi skal gøre os håb om at se en niveauforskel på ca. 1 som værende signifikant!

	T d_hyp	U p_hyp	V	W
◆	=seq(skala	=seq(skal		=tTest_2Samp('d_gen
1	7	7	Titel	t−test med 2 prøver
2	7	7	Alternativ Hyp	μ1 > μ2
3	7	7	t	1.74048
4	7	7	PVal	0.042128
5	7	7	df	124.
6	7	7	x̄1	6.25
7	7		x̄2	5.33333
8	7		sx1	3.07683
9	7		sx2	2.07971
10	7		sp	2.7869
11	7		n1	84.
12	7		n2	42.

2.5 Dummy-test: Omdan t-test med to middel værdier til en lineær regressionstest

Vi har set at der er tre forskellige typer uafhængighedstests for to stokastiske variable, alt eftersom vi kombinerer to numeriske variable, en kategorisk og en numerisk eller to kategoriske variable. I praksis vil man ofte konvertere den midterste type til en af de to andre: Det kan ske ved enten at omdanne en kategorisk variabel til en numerisk variable, eller ved at gruppere den numeriske variabel, dvs. omdanne den til en kategorisk variabel.

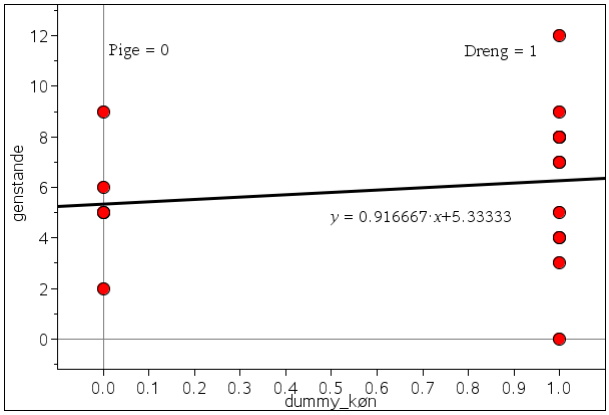
Eksempel 1: I Susanne Christensens noter lægger hun ud med et eksempel, hvor man skal undersøge hvorvidt de unges tøjforbrug er afhængigt af deres køn. Som udgangspunkt skal man altså undersøge sammenhængen mellem en kategorisk variabel, **køn**, og en numerisk variabel, **tøjforbrug** (målt som det gennemsnitlige antal kroner pr måned, den unge bruger på tøj). For at håndtere det som en χ^2-test, der jo handler om sammenhængen mellem to kategoriske variable, må hun derfor først konvertere den numeriske variabel til en kategorisk variabel. Det gør hun ved at gruppere **tøjforbrug**, idet hun indlægger et snit på 1500 kr. pr måned og derved splitter **tøjforbrug** i **højt forbrug**, hvis det overstiger 1500 kr. pr måned, og ellers **lavt forbrug**.

Eksempel 2: Der er imidlertid også den anden mulighed, at man i stedet konverterer den kategoriske variabel, **køn**, til en numerisk variabel. Det vil vi illustrere ved hjælp af eksemplet med de unges alkoholforbrug. Vi omkoder da det ene køn, kvinde, til et 0-tal, og det andet køn, mand, til et 1-tal. Den omkodede variabel kaldes en **dummyvariabel**. I praksis sker det ved hjælp af den viste celleformel, som vi efterfølgende trækker ned gennem listen:

	A køn	B dummy_køn
◆		
1	Dreng	1
2	Dreng	1
3	Dreng	1
4	Dreng	1
5	Dreng	1
6	Dreng	1
7	Dreng	1
8	Dreng	1
9	Dreng	1
10	Dreng	1
11	Dreng	1
12	Dreng	1
13	Pige	0
14	Pige	0
15	Pige	0
16	Pige	0
17	Pige	0
18	Pige	0

$B1$	$=\begin{cases}0, a1=\text{"Pige"}\\1, a1=\text{"Dreng"}\end{cases}$

Da vi nu har to numeriske variable i spil kan vi afbilde dem i et punktplot.



Man kan nu eftervise ved simple symbolske udregninger at den bedste rette linje netop går gennem tyngdepunktet for pigerne, henholdsvis tyngdepunktet for drengene (ligesom den går igennem det samlede tyngdepunkt). Hældningen for dummy-regressionsligningen er derfor netop det samme som niveauforskellen, dvs. forskellen på middelværdierne for de to køn. Spørgsmålet er så om denne hældning er statistisk signifikant forskellig fra 0. Vi kan derfor nu udføre en lineær regressionstest, som er en t-test fuldstændigt ækvivalent med t-testen for 2 prøver:

R	S	T	U
	=tTest_2Samp('d_ge		=LinRegtTest('dumm
Titel	t-test med 2 prøver	Titel	Lineær Reg t-test
Alternativ Hyp	μ1 > μ2	Alternativ Hyp	β & ρ > 0
t	0.625199	RegEqn	a+b*x
PVal	0.270328	t	0.625199
df	16.	PVal	0.270328
x̄1	6.25	df	16.
x̄2	5.33333	a	5.33333
sx1	3.19446	b	0.916667
sx2	2.25093	s	2.9324
sp	2.9324	SESlope	1.4662
n1	12.	r²	0.023847
n2	6.	r	0.154425

Vi ser da også at **t-værdien**, **p-værdien** og **antallet af frihedsgrader** er nøjagtigt de samme. I t-testen for de to prøver får vi derefter oplyst de to middelværdier $\bar{x}_1$ og $\bar{x}_2$, mens vi i dummy-regressionstesten får oplyst skæringen med x-aksen a (som er den første middelværdi) og hældningen b (som er differensen mellem de to middelværdier, dvs. $\bar{x}_1 - \bar{x}_2$. De to stikprøvespredninger sx1 og sx2, optræder kun i t-testen for to prøver, mens standardfejlen for hældningen kun optræder i dummy-regressionstesten. Til gengæld optræder residualspredningen i begge modeller, idet den dog kaldes sp i t-testen for to prøver henholdsvis s i dummytesten. Endelig får vi oplyst antallet af elementer i de to prøver i t-testen for to prøver, mens vi får oplyst korrelationskoefficienten og forklaringsgraden i dummy-regressionstesten.

3. Unges holdning til skolen versus glæden for skolen: Gamma-testen

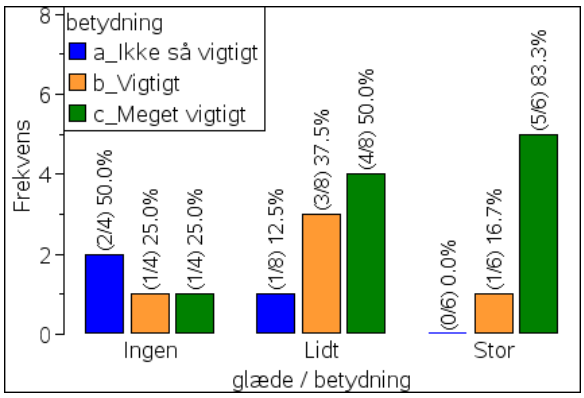
3.1 Gamma-testen: Den beskrivende statistik

Vi tager igen udgangspunkt i 18 elever, men denne gang har vi bedt dem udfylde et spørgeskema om deres holdning til dette og hint i forbindelse med deres skolegang. Vi ser her på to af spørgsmålene: "Har du glæde af at gå i skole?" med svarmulighederne "Ingen", "Lidt", "Stor" og "Hvor vigtigt er det at man klarer sig godt i skolen?" med svarmulighederne "Ikke så vigtigt", "Vigtigt" og "Meget vigtigt". Det er netop typiske eksempler på ordinale variable, hvor de enkelte svar kan rangordnes.

Svarene fordeler sig således på de 18 elever:

	A betydning	B glæde
1	a_Ikke så vigtig...	Ingen
2	a_Ikke så vigtig...	Ingen
3	b_Vigtigt	Ingen
4	c_Meget vigtigt	Ingen
5	a_Ikke så vigtig...	Lidt
6	b_Vigtigt	Lidt
7	b_Vigtigt	Lidt
8	b_Vigtigt	Lidt
9	c_Meget vigtigt	Lidt
10	c_Meget vigtigt	Lidt
11	c_Meget vigtigt	Lidt
12	c_Meget vigtigt	Lidt
13	b_Vigtigt	Stor
14	c_Meget vigtigt	Stor
15	c_Meget vigtigt	Stor
16	c_Meget vigtigt	Stor
17	c_Meget vigtigt	Stor
18	c_Meget vigtigt	Stor

Her har vi indføjet rangordningen i navnene på betydningskategorierne for bedre at kunne styre rangordningen i diagrammerne: "a_ikke så vigtigt", "b_Vigtigt" og "c_Meget vigtigt".



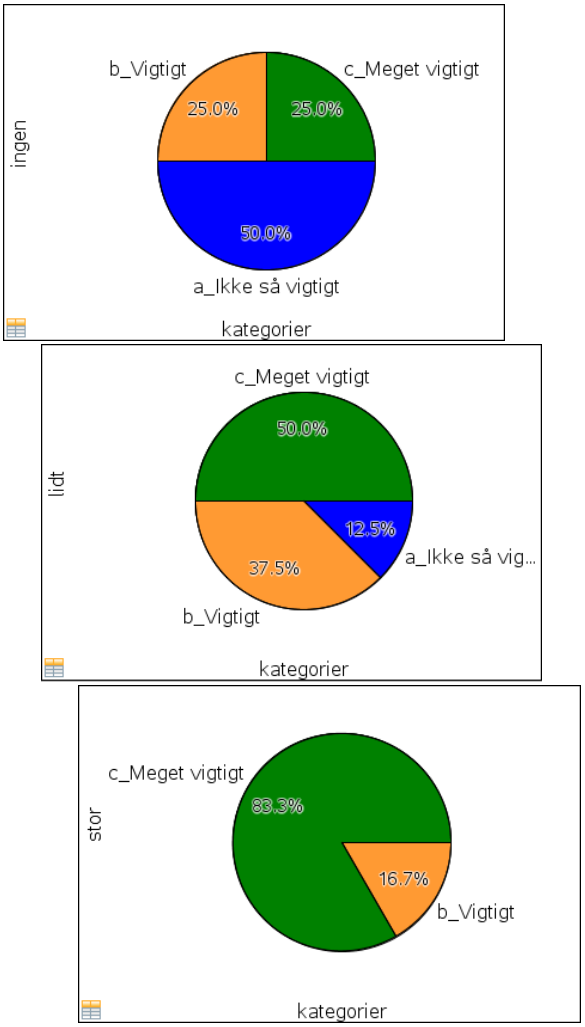
Et grupperet søjlediagram viser da at der synes at være en sammenhæng mellem de to kategoriske variable: Jo større glæde man har af at gå i skole jo vigtigere er det at man klarer sig godt i skolen. Hvis man omdanner de rå data til en krydstabel kan man også få vist de procentuelle fordelinger i form af tre adskilte cirkeldiagrammer.

For at kunne undersøge sammenhængen nærmere omdannes listerne derfor til en krydstabel. Vi skal da have opstillet passende kategorier samt en passende kommando til at styre optællingen (tallene kan også aflæses fra det grupperede søjlediagram, men vi får også brug for at kunne frembringe dem dynamisk):

	C	D	E	F
1		Ingen	Lidt	Stor
2	a_Ikke så vigti...	2	1	0
3	b_Vigtigt	1	3	1
4	c_Meget vigtig...	1	4	5

D2 =sum(iffn(**betydning**=$c2 and **glæde**=d$1,1,0))

Trækkes krydstabellen op over stregen kan vi frembringe cirkeldiagrammerne:



Vi ser den samme tendens som før: Jo mere glæde man har af at gå i skole, jo vigtigere er det at klare sig godt!

G kategorier	H ingen	I lidt	J stor
=c2:c4	=d2:d4	=e2:e4	=f2:f4
a_Ikke så vigtigt	2	1	0
b_Vigtigt	1	3	1
c_Meget vigtigt	1	4	5

Men så snart vi har tallene klar i krydstabellen kan vi begynde at sætte tal på tendensen. Vi ser da på et vilkårligt par af elever. Et sådant par svarer til to celler i tabellen (der eventuelt kan falde sammen). Vi undersøger da hvorvidt de har samme holdning eller forskellige holdninger til de stillede spørgsmål. De har samme holdning til betydningsspørgsmålet, hvis de ligger i samme række. De har samme holdning til glædesspørgsmålet, hvis de ligger i samme søjle. Et sådant par kaldes på engelsk for en 'tie', på dansk et **uafgjort par** (de står lige). De uafgjorte par viser ikke noget om tendensen, så dem vil vi ignorere i det følgende. Vi ser i stedet på de **afgjorte par**, hvor cellerne både skifter række og søjle. Der findes nu to typer:

	G kategorier	H ingen	I lidt	J stor
◆	=c2:c4	=d2:d4	=e2:e4	=f2:f4
1	a_Ikke så vigtigt	2	1	0
2	b_Vigtigt	1	3	1
3	c_Meget vigtigt	1	4	5

Ser vi fx på cellen i midten, så kan den kombineres med cellerne 2 og 5 langs hoveddiagonalen skråt nedad eller med cellerne 1 og 0 langs sidediagonalen skråt opad. I den første kombination er det sådan at når glæden ved at gå i skole vokser gælder det samme om betydningen af at klare sig godt i skolen.

	G kategorier	H ingen	I lidt	J stor
◆	=c2:c4	=d2:d4	=e2:e4	=f2:f4
1	a_Ikke så vigtigt	2	1	0
2	b_Vigtigt	1	3	1
3	c_Meget vigtigt	1	4	5

Et sådant par kaldes på engelsk 'concordant', dvs. de to variables opførsel er i samklang med hinanden. PÅ dansk kan vi kalde et sådant par et **voksende par** (når den ene vokser, gør den anden det også, dvs. de opfører sig på samme måde). I den anden kombination er det sådan at når glæden ved at gå i skole vokser aftager betydningen af at klare sig godt i skolen. Et sådant par kaldes på engelsk 'discordant', dvs. de to variables opførsel er i uoverensstemmelse med hinanden. På dansk kan vi kalde et sådant par et **aftagende par** (når den ene vokser, aftager den anden, dvs. de opfører sig modsat af hinanden). Vi ser da at cellen i midten giver anledning til $3\cdot(2+5)=21$ voksende par, men kun $3\cdot(1+0)=3$ aftagende par. Den overordnede tendens synes altså at være voksende.

Som foreslået af Kruskal og Goodmann giver det nu anledning til det følgende mål for overensstemmelsen/korrelationen mellem de to variables opførsel: Vi finder først antallet af voksende par V og tilsvarende antallet af aftagende par A:

$$V = 2\cdot(3+1+4+5)+1\cdot(1+5)+1\cdot(4+5)+3\cdot 5$$

$$= 2\cdot 13+1\cdot 6+1\cdot 9+3\cdot 5 = 26+6+9+15 = 56$$

$$A = 1\cdot(1+0+3+1)+4\cdot(0+1)+1\cdot(1+0)+3\cdot 0$$

$$= 1\cdot 5+4\cdot 1+1\cdot 1+3\cdot 0 = 5+4+1+0 = 10$$

Der er altså er klart overskud af voksende par! Vi måler dette med **gammagraden**, der defineres som **den relative forskel mellem de voksende og aftagende par**, dvs.

$$\gamma = \frac{V-A}{V+A} = \frac{56-10}{56+10} = \frac{46}{66} \approx 0.69697$$

Vi siger derfor at gammagraden er 69.7%. Vi lægger også mærke til at tælleren altid er mindre end nævneren og at brøken kun kan være 1, hvis A = 0 og tilsvarende kun kan være -1 hvis V = 0. Der gælder derfor:

- Gammagraden ligger mellem -1 og 1.
- Den er positiv for voksende sammenhænge, negativ for aftagende sammenhænge (og 0 for sammenhænge, der hverken er voksende eller aftagende).
- Den er 1 for rent voksende sammenhænge, hvor der slet ikke findes aftagende par.
- Den er -1 for rent aftagende sammenhænge, hvor der slet ikke findes voksende par.

Gammagraden har også en simpel sandsynlighedsteoretisk fortolkning: Når vi trækker et tilfældigt par vil vi ignorere det, hvis der er tale om et uafgjort par. Hvis der derimod er tale om et afgjort par vil vi lade p_+ betegne sandsynligheden for at parret er voksende og tilsvarende lade p_- betegne sandsynligheden for at parret er aftagende. De to betingede sandsynligheder p_+ og p_- er da givet ved

$$p_+ = \frac{V}{V+A} \quad \text{og} \quad p_- = \frac{A}{V+A}$$

Gammagraden er derfor også givet ved $p_+ - p_-$.
I det ovenstående tilfælde fås derfor

$$\begin{aligned} p_+ - p_- &= 69.7\% \\ p_+ + p_- &= 100\% \end{aligned} \quad \text{hvorfor} \quad \begin{aligned} p_+ &= \frac{100\% + 69.7\%}{2} = 84.85\% \\ p_- &= \frac{100\% - 69.7\%}{2} = 15.15\% \end{aligned}$$

Sandsynligheden for at trække et voksende par er altså næsten seks gange så stor som sandsynligheden for at trække et aftagende par.

Vi har altså nu på mange forskellige måder argumenteret for at den fundne sammenhæng peger mod en tydelig voksende sammenhæng mellem elevernes glæde ved at gå i skole og den betydning de tillægger det at man klarer sig godt i skolen. Det er så langt vi kan komme med den beskrivende statistik.

3.2 Gamma-testen: Den bekræftende statistik*

Lad os til indledning udføre en χ^2-test[3] på matricen/krydstabellen {**ingen**, **lidt**, **stor**}:

K	L
	=χ²2way({ingen,lidt,stor}
Titel	χ²−uafhængighedstest...
χ²	5.85
PVal	0.210634
df	4.
ExpMatrix	[[0.66666666666667,1...
CompMatrix	[[2.6666666666667,0....

Vi finder da en p-værdi på 21.1% og det giver *ikke* anledning til at forkaste nulhypotesen: χ^2-testen kan altså ikke se sammenhængen som signifikant.

Vi prøver derfor i stedet at udføre en gammatest. Vi skal da som sædvanligt omrøre den ene af de to stokastiske variable. I dette tilfælde vælger vi at omrøre **glæde:**

	A betydning	B glæde	C omrøring	D
◆			=randsamp(glæde	
1	a_Ikke så vigtig...	Ingen	Lidt	
2	a_Ikke så vigtig...	Ingen	Stor	a_
3	b_Vigtigt	Ingen	Stor	b_
4	c_Meget vigtigt	Ingen	Lidt	c_
5	a_Ikke så vigtig...	Lidt	Ingen	
6	b_Vigtigt	Lidt	Ingen	V
7	b_Vigtigt	Lidt	Ingen	A
8	b_Vigtigt	Lidt	Lidt	G
9	c_Meget vigtigt	Lidt	Lidt	
10	c_Meget vigtigt	Lidt	Stor	
11	c_Meget vigtigt	Lidt	Ingen	
12	c_Meget vigtigt	Lidt	Lidt	
13	b_Vigtigt	Stor	Stor	
14	c_Meget vigtigt	Stor	Stor	
15	c_Meget vigtigt	Stor	Lidt	
16	c_Meget vigtigt	Stor	Lidt	
17	c_Meget vigtigt	Stor	Stor	
18	c_Meget vigtigt	Stor	Lidt	

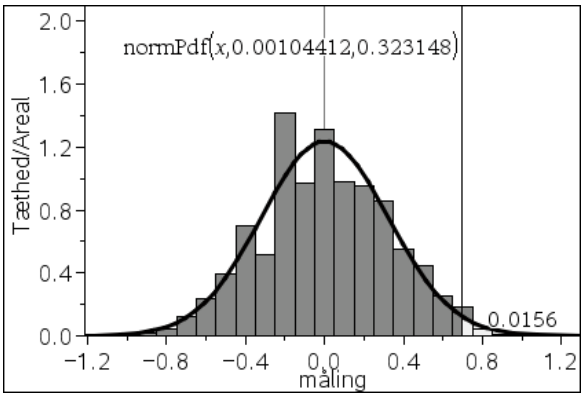
C **omrøring**:=randsamp(**glæde**,dim(**glæde**),1

Det giver mulighed for at opbygge en krydstabel med tilhørende gammagrad for de omrørte data: **betydning** og **omrøring**, der som sædvanlig er defineret ud fra en randsamp()-kommando (samme størrelse som klassen, men **uden tilbagelægning**, hvorfor slutparameteren skal være 1).

	D	E	F	G	H
◆					=d2:
1	Observationer	Ingen	Lidt	Stor	a_Ikl
2	a_Ikke så vigtigt	2	1	0	b_Vi
3	b_Vigtigt	1	3	1	c_M
4	c_Meget vigtigt	1	4	5	
5					
6	Voksende par	56			
7	Aftagende par	10			
8	Gamma−graden	0.69697			
9					
10	Simulering	Ingen	Lidt	Stor	
11	a_Ikke så vigtigt	1	2	0	
12	b_Vigtigt	1	2	2	
13	c_Meget vigtigt	2	4	4	
14					
15	Voksende par	40			
16	Aftagende par	22			
17	Gamma−graden	0.290323			
18					

E11 =sum(iffn(**betydning**=$*d11* and **omrøring**=*e*$1,1,0))

Den simulerede gammagrad vil da somme tider være positiv og somme tider negativ. Da der ingen sammenhæng er mellem de simulerede variable vil der nemlig være samme chance for at finde en positiv sammenhæng som en negativ sammenhæng, dvs. i middel er gammagraden 0. Vi gemmer såvel den observerede gammagrad som den simulerede gammagrad ud fra nulhypotesen i variablene **obs_gamma** og **sim_gamma**. Derefter kan vi fange den simulerede gammagrad med en automatisk datafangst i variablen **måling**:



Her ses resultatet af 1000 målinger afsat som histogram. Der fremkommer en klokkeformet fordeling, som typisk tilnærmes med en normalfordeling (brug menupunktet **Vis normal PDF** i **Undersøg data**-menuen). Vi kan også højreklikke på normalfordelingsgrafen og vælge **Skraver under funktion**. Arealet under grafen til højre for den observerede gammagrad er da 1.56%. Ved tosidet test af uafhængighed fås derfor en p-værdi på det dobbelte, altså 3.1%.

[3] Du kan finde en udførlig diskussion af χ^2-testen i hæftet **Statistik med TI-Nspire CAS ver. 3.1**, som fx kan findes på TI's hjemmeside.

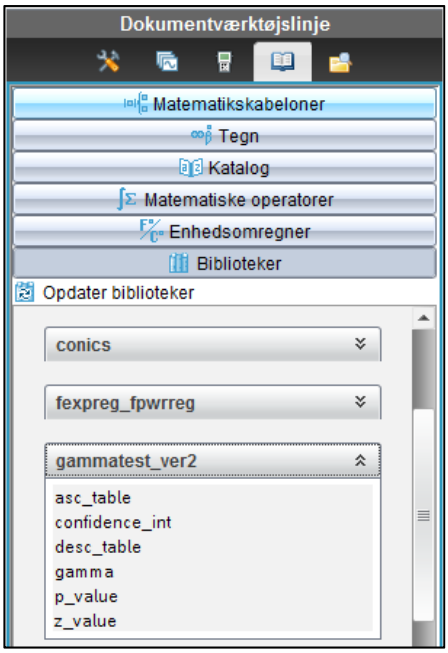
Det samme bekræftes af en optælling af antal skæve udfald på sædvanlig vis:

	D	E	F	G	H måling
◆					=capture(s
5					-0.245902
6	Voksende par	56			-0.52381
7	Aftagende par	10			0.6
8	Gamma-graden	**0.69697**			0.
9					-0.354839
10	Simulering	Ingen	Lidt	Stor	-0.245902
11	a_Ikke så vigtigt	0	2	1	0.290323
12	b_Vigtigt	2	2	1	-0.354839
13	c_Meget vigtigt	2	4	4	0.363636
14					-0.16129
15	Voksende par	34			-0.09375
16	Aftagende par	28			0.164179
17	Gamma-graden	**0.096774**			0.552239
18	Antal målinger	1000			0.363636
19	Antal skæve	32			0.164179
20	p-værdi	3.2			-0.396825
21					0.353846
22					0.193548

E19 =countif(**måling**,?≥**obs_gamma**)

Vi finder da en eksperimentel p-værdi på 3.2%. I begge tilfælde konkluderer vi altså at den fundne sammenhæng mellem de to ordinale variable er statistisk signifikant på signifikansniveauet 5%. Det understreger netop at gammatesten som lovet er stærkere end χ^2-testen for ordinale variable.

Vi kan også udføre en teoretisk gammatest. Da gammatesten ikke er indbygget som kanonisk test i TI-Nspire CAS kræver det at man tilføjer en biblioteksrutine, der kan udføre testen![4]

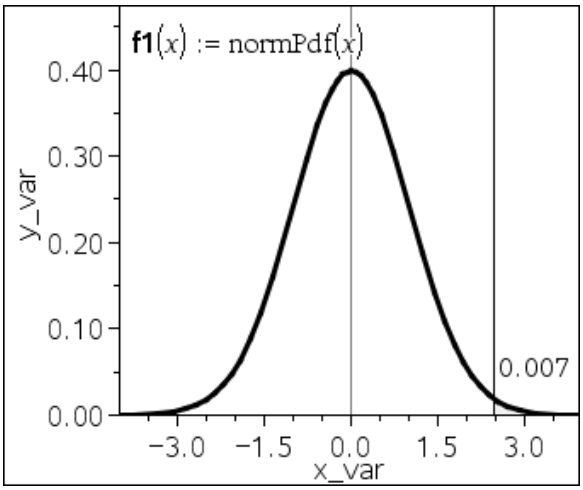


Det er især kommandoerne gamma, p_value og z_value, der er interessante. De skal som argument bruge matricen/krydstabellen, der kan indtastes som under χ^2-testen:

17	Gamma-graden	**0.096774**			0.
18	Antal målinger	1000			0.
19	Antal skæve	32			0.
20	p-værdi	3.2			-0.
21	Gammatest				0.
22	Gammagrad	0.69697			0.
23	p-værdi	0.01337			-0.
24	z-værdi	2.47377			0.
25					0.

E23 =**gammatest_ver2\p_value({ingen,lidt,stor})**·1.

Vi ser da at p-værdien er 1.4% hvorfor sammenhængen er statistisk signifikant. Vi ser også at den kanoniske teststørrelse z har værdien 2.47, dvs. den ligger ca. 2.5 standardafvigelser ude i standardnormalfordelingen. Den teoretiske test er tosidig. Hvis vi selv vil illustrere den skal vi tegne standardnormalfordelingen og afskære 2.47 standardafvigelser i begge sider:



Afskærer vi til højre finder vi arealet 0.007. Da vi skal afskære i begge sider fås derfor den kanoniske p-værdi 1.4%.

Hvis vi vil forstå z-værdien er vi nødt til at kunne estimere standardspredningen for gammagraden. Det er en rimelig indviklet formel, som man godt kan stave sig igennem i regnearket, men den kræver en tilbundsgående diskussion af krydstabellen, som vi udelader i dette hæfte.[5] Men når først standardspredningen s_0 er på plads er sammenhængen den sædvanlige:

$$z = \frac{\gamma}{s_0}.$$

[4] Gammatesten ligger i filen **gammatest_ver2**, der fx kan findes på TI's hjemmeside. Filen skal gemmes i mappen **MyLib** i TI-Nspire mappen under **Dokumenter**.

[5] Detaljerne kan fx findes på hjemmesiden (åbn Lecture 1 om ordinal variables): http://staff.pubhealth.ku.dk/~skm/fsvpage/Analysis%20of%20multivariate%20categorical%20data/CatData%20course.htm

4. En teknisk bemærkning om opsamling af teststørrelsen *

I det foregående har vi anvendt **automatisk data-opsamling** af teststørrelsen. Det er strengt taget ikke helt korrekt. Den automatiske opsamling registrerer nemlig kun teststørrelsen, når der sker *ændringer* i teststørrelsen. Men der er en lille sandsynlighed for at teststørrelsen ved et tilfælde rent faktisk ikke skifter værdi, hvilket fx vil være tilfældet når permutationerne af y-værdierne rent faktisk sker indenfor de samme x-værdier. I eksemplet med de unges alkoholforbrug kan man derfor komme ud for at permutationen kun ombytter de første 12 værdier indbyrdes og tilsvarende med de sidste 6 indbyrdes værdier. Det vil ikke ændre teststørrelsen.

Man kan kontrollere hvor alvorligt problemet er ved at man samtidigt med at man laver datafangst på teststørrelsen også laver datafangst på en dummy-variabel, som med sikkerhed skifter værdi hver gang. Her kan man fx benytte dummyvariablen **rand()**, der hele tiden producerer nye tilfældige tal i intervallet fra 0 til 1. I de foregående eksempler vil man da se, at gentagelserne forekommer i ca. 2% af tilfældene. Det giver en lille skævhed i fordelingen af teststørrelsen, men den er ikke stor nok til at den kan rykke ved konklusionerne. Man kan komme uden om problemet ved at anvende **manuel datafangst** (hvor man skal taste CTRL ., dvs. CTRL PUNKTUM, hver gang man vi samle data op). Men det er mere besværligt, fordi man i så fald skal skiftevis skal taste CTRL R og CTRL . rigtigt mange gange for at opsamle målingerne – og man må ikke overspringe CTRL R en eneste gang, for så får man jo bare det modsatte problem, hvor nogle af målingerne optræder flere gange uden at man har omrørt. Det er derfor en klar fordel at kunne anvende automatisk dataopsamling.

Hvis man vil gøre det helt korrekt, skal man derfor udføre den automatiske datafangst lidt mere omhyggeligt. I praksis kan man da lave **samtidig datafangst** på såvel **summen af teststørrelsen og dummy-variablen** som på **dummy-variablen**. Derved undgår man gentagelser, hvilket man kan kontrollere ved fx at se, at man hele tiden fanger det samme antal målinger. Efterfølgende kan man så trække målingerne af dummy-variablen fra målingerne af summen. Derved rekonstrueres netop den fulde liste af teststørrelser svarende til de gentagne simuleringer. For at undgå at komplicere begrebstilegnelsen har jeg ikke vist denne teknik i forbindelse med eksemplerne. Men de er også afprøvet med denne sikre teknik, og det giver ikke anledning til væsentlige ændringer af de forskellige estimater, der ligger til grund for testene.

Vi viser her resultatet af en sådan sikker måling for den lineære regressions test. Først opretter vi dummyvariablen **dummy** og sum-variablen **sumtest**.

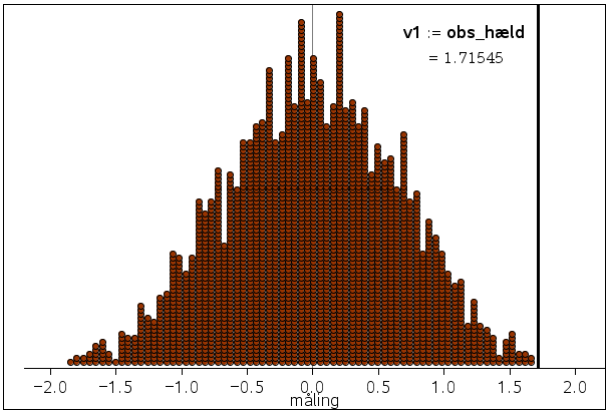
	I	J
◆	=LinRegMx(	
1	Lineær reg...	obs_hæld
2	m*x+b	**1.71545**
3	0.300813	sim_hæld
4	1.0813	**0.300813**
5	0.013161	**0.408449**
6	0.114723	**0.709262**
7	{-0.593495...	Antal test
8		2500
9		Antal dummy
10		2500

J4	**sim_hæld**:=*i3*
J5	**dummy**:=rand()
J6	**sumtest**:=*j4+j5*

Derefter gennemfører vi den simultane måling af sum-variablen og dummy-variablen.

	K test	L safetest	M måling
◆	=capture('sumtest,1)	=capture('dummy,1)	=test−safetest
1	0.74355	0.540298	0.203252
2	0.541919	0.533789	0.00813
3	0.830216	0.773305	0.056911
4	0.957086	0.509932	0.447154
5	1.00922	0.903533	0.105691
6	0.824682	0.426308	0.398374
7	-0.973917	0.627709	-1.60163
8	1.80086	0.573216	1.22764
9	1.00784	0.072878	0.934959
10	-0.349093	0.179362	-0.528455

Endelig afbilder vi de rekonstruerede målinger af den simulerede hældning:



Indholdsfortegnelse:

Forord **side 1**

1. Unges alkoholforbrug som funktion af alder: Lineær regression t-test **side 3**

1.1 Lineær regression t-test: Den beskrivende statistik side 3
1.2 Lineær regression t-test: Den bekræftende statistik side 5
1.3 Simulering af t-fordelingen for en lineær regressionstest* side 10

2. Unges alkoholforbrug som funktion af køn: t-test for to middelværdier **side 11**

2.1 t-test for to middelværdier: Den beskrivende statistik side 11
2.2 t-test for to middelværdier: Den bekræftende statistik side 12
2.3 Simulering af t-fordelingen for en t-test med to prøver* side 16
2.4 Hvor stor skal stikprøven være, hvis man skal kunne skelne nulhypotesen fra den alternative hypotese?* side 16
2.5 Omdan t-test med to middelværdier til lineær regressionstest side 17

3. Unges holdning til skolen versus glæden ved skolen: Gamma-testen: **side 18**
3.1 Gammatesten: Den beskrivende statistik side 18
3.2 Gammatesten: Den beskrivende statistik side 20

4. En teknisk bemærkning om opsamling af teststørrelsen* **side 22**