Excel-øvelser i sannsynlighetsregning

av

Peer Andersen

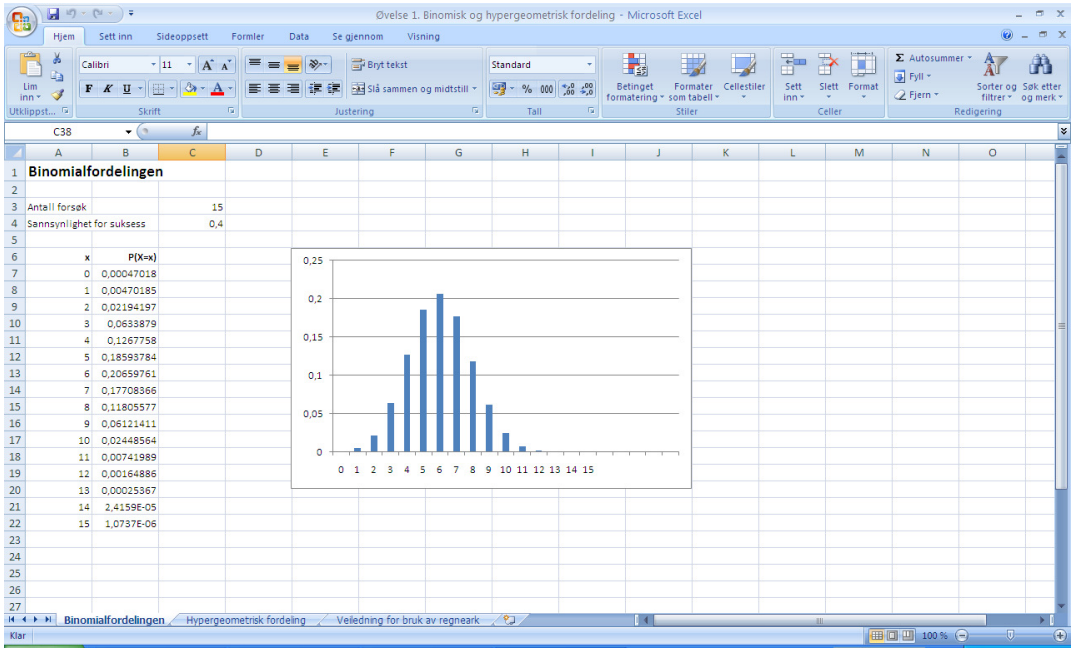
Innhold

Innledning

Dette heftet inneholder 7 øvelser i sannsynlighetsregning. Øvelsene er primært rettet mot fordypningskurs i allmennlærerutdanningen. Øvelsene forutsetter at en har grunnleggende kjennskap til Excel og kjennskap til de fagområdene som øvelsene berører. I boken Excel-øvelser i matematikk av Peer Andersen er det beskrevet mer grunnleggende øvelser med hvordan Excel kan brukes i matematikken. Her finnes det også beskrivelser av hvordan Excel kan brukes i blant annet ulike simuleringsforsøk. Dette heftet går et steg videre. I øvelse 1 ser vi først på hvordan en kan jobbe med binomiske og hypergeometriske sannsynligheter. I øvelse 2 ser vi på hvordan vi kan generere 1000 stikkprøver, hver på 10 elementer og deretter bruke dette til å lage konfidensintervall. Øvelse 3 går på å gjennomføre hypotesetest om gjennomsnittet når standardavviket er kjent. I øvelse 4 gjennomfører vi tilsvarende hypotese når standardavviket er ukjent. I øvelse 5 ser vi på hypoteser i en binomisk situasjon. Sammenligning mellom to populasjoner er tema for øvelse 6. Til slutt i øvelse 7 ser vi på hvordan Excel kan brukes i arbeidet med regresjon.

Øvelse 1. Binomiske og hypergeometriske sannsynligheter

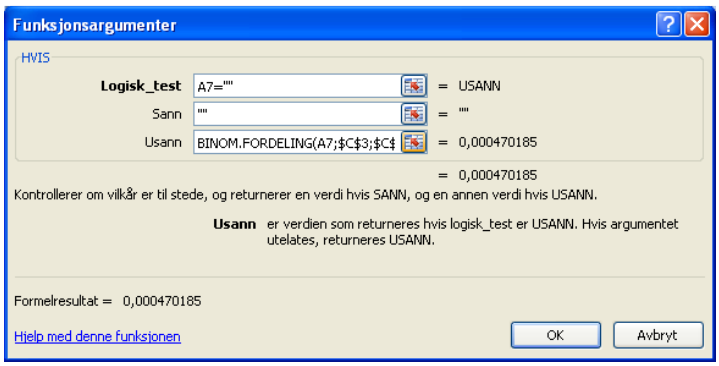
I denne øvelsen skal vi se på hvordan Excel kan benyttes til å beregne binomiske og hypergeometriske sannsynligheter. Vi ser først på hvordan vi kan bruke Excel til å beregne binomiske sannsynligheter. Vi skal konstruere regnearket slik at vi oppgir antall forsøk og sannsynligheten for suksess i hvert enkelt forsøk. Regnearket skal deretter regne ut hele fordelingen og også fremstille det i et diagram. Vi skal konstruere regnearket slik at vi tar høyde for inntil 20 forsøk. Regnearket kan se ut som vist under.



Du kan starte med å skrive inn teksten og verdiene som står i cellene fra A1 til C6. Vi skal deretter ta fatt på beregningen av sannsynlighetene. Vi må først lage oss en tellekolonnene som starter på 0 og går opp til så mange forsøk vi skal gjøre. Her kommer vår første lille utfordring. Problemet er hvordan vi skal få stoppet x på akkurat 15. Løsningen på dette er å bruke HVIS funksjonen som ligger inne i Excel. Først kan du skrive inn 0 i celle A7. I celle A8 skal vi bruke HVIS setningen. Du kan åpne funksjonsveiviseren og finne HVIS funksjonen. Den fyller du ut som vist under

Det denne funksjonen gjør er at den tester om verdien i foregående celle er mindre enn antall forsøk. Dersom det er tilfelle tar den verdien fra foregående celle og plusser på 1. Dersom den ikke er

mindre enn antall forsøk lar den cellen stå blank. For å få frem en blank celle i Excel skriver en inn "". Du kan nå kopiere cellen ned til og med celle A27. I B kolonnen skal vi beregne sannsynligheten for de enkelte utfallene. Vi ser først på celle B7. I Excel er det en funksjon som regner ut binomialfordelingen som vi skal bruke. Imidlertid er det slik at hvis den tilhørende verdien i A kolonnen er blank, ønsker vi at tilsvarende verdi i B kolonnen også skal være blank. Vi må derfor bruke en HVIS setning slik vi gjorde for verdiene i A kolonnen. I celle B7 kan du fylle ut HVIS setningen som vist under



I skjermbilde har litt av teksten etter USANN ikke kommet med. Det som skal stå i feltet etter USANN er

BINOM.FORDELING(A7;C3;C4;USANN)

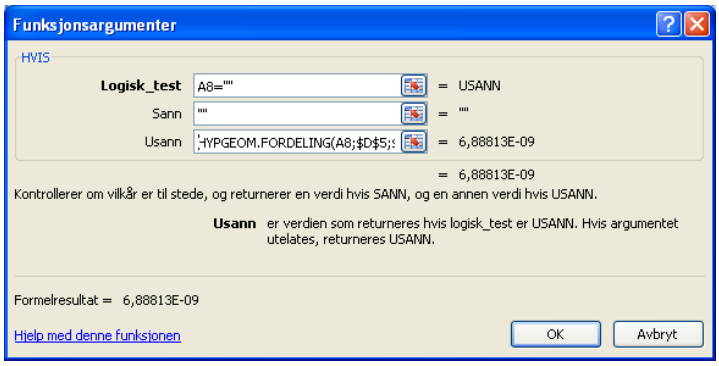
Det funksjonen gjør er at den først tester om celle A7 er blank eller ikke. Dersom den er blank lar den celle B7 også være blank. Hvis den ikke er blank skal vi regne ut sannsynligheten for 0 suksesser. Formelen som er vist over gjør denne beregningen. Verdien A7 indikerer antall suksesser, C3 er antall forsøk og C4 er sannsynligheten for suksess på et enkelt forsøk. USANN har vi skrevet for å angi at det er punktsannsynligheten vi skal beregne. Vi har brukt dollartegn rundt C3 og C4 slik at disse ikke skal forandre seg når vi kopierer formelen. Formelen du har skrevet inn i celle B7 kan du nå kopiere ned til og med celle B27.

Til slutt skal vi lage et diagram over sannsynlighetsfordelingen. Du kan starte med å merke celle B7 til og med celle B27. Klikk deretter på Sett inn og velg Stolpe. Velg deretter et passende diagram. Du vil se at enhetene på x-aksen ikke stemmer. I figuren starten den på 1, mens den burde startet på 0. Her er det Excel som legger inn verdiene på x-aksen, slik den tror det skal være. Men som vi ser så stemmer ikke dette helt. Vi kan imidlertid endre dette slik at det blir riktig. Høyreklikk med musen en plass i diagrammet og velg Merk data. Velg Rediger under der hvor det står Vannrette akseetiketter. Merk området fra A7 til A27 og trykk ok.

Regnearket skal da være klart til bruk. Test det ut på noen kjente problemstillinger og se at det virker som det skal.

Regnearket for hypergeometrisk fordeling skal vi lage på samme måte. På neste side er det vist hvordan det kan se ut. I celle C3 angir vi størrelsen på populasjonen. I celle C4 angir vi antall spesielle i populasjonen og til slutt i celle C5 skriver vi inn hvor stort utvalget skal være. Tellekolonnen lager vi på akkurat samme måte som vi gjorde i det binomiske tilfelle. Vi skal også ta høyde for at vi kan ha et utvalg på inntil 20. I celle B8 skal vi beregne sannsynligheten. I Excel finnes det en funksjon som

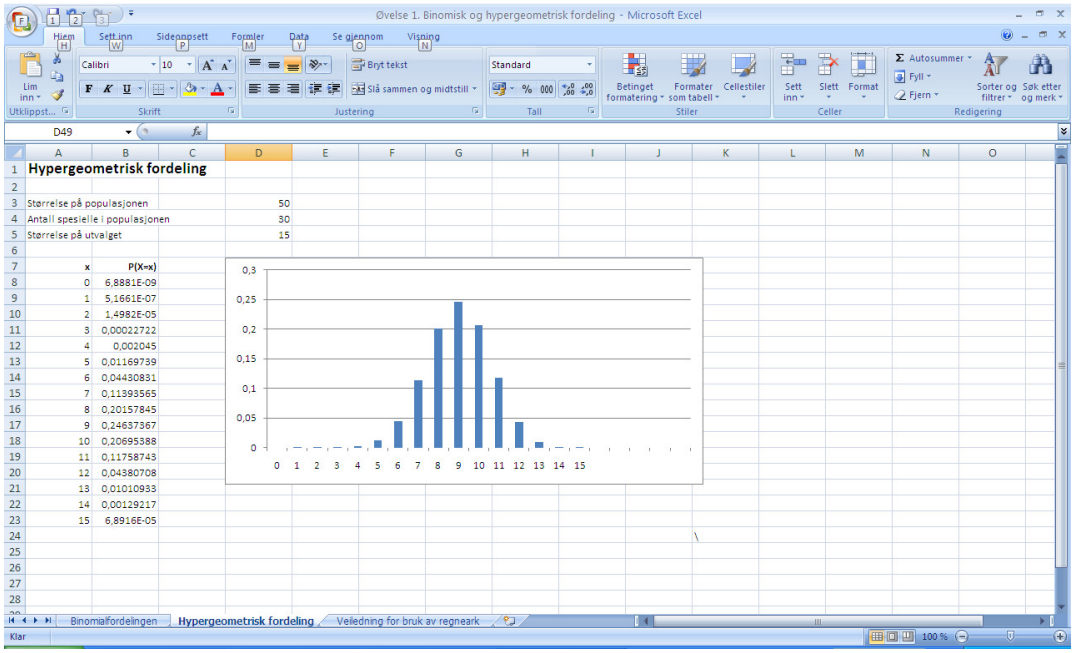
regner ut hypergeometriske sannsynligheter som vi skal bruke. Som i det binomiske tilfelle ønsker vi at cellene i B kolonnen skal være blanke om tilsvarende celler i A kolonnen er blank. Vi bruker derfor HVIS funksjonen. Du kan fylle den ut som vist under



Teksten etter USANN er ufullstendig i skjermbilde, men det som skal stå der er

HYPGEOM.FORDELING(A8;D5;D4;D3)

Du kan nå kopiere funksjonen ned til celle B28. Diagrammet lager du på akkurat samme måte som i det binomiske tilfelle.



Øvelse 2. Konfidensintervall om gjennomsnittet

La oss anta at vi har en populasjon som er normalfordelt med gjennomsnitt μ og standardavvik σ. Gjennomsnittet μ er ukjent. Ved hjelp av statistiske metoder kan en lage et intervall der en med en viss prosent sikkerhet kan si at den ukjente μ vil ligge. Denne prosenten er vanligvis 90, 95 eller 99. Et slikt intervall kalles gjerne for et konfidensintervall.

Hvis vi antar at vi har n uavhengige observasjoner $X_1, X_2, \ldots, X_n$ fra populasjonen kan en vise at et 95 % konfidensintervall for gjennomsnittet vil være angitt ved

$$\left[\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right]$$

Det vil si at den ukjente μ med 95% sikkerhet vil ligge innenfor dette intervallet. Dersom vi skal konstruere et 90% eller 99% konfidensintervall erstatter vi 1,96 med henholdsvis 1,645 og 2,58.

La oss se på et lite eksempel. Vi kjøper inn 10 pakker med kjøttdeig fra Gilde og måler fettinnholdet i hver av dem. La oss anta at fettinnholdet er:

14,3	14,6	13,5	13,9	14,6
14,3	15,1	14,3	12,8	13,6

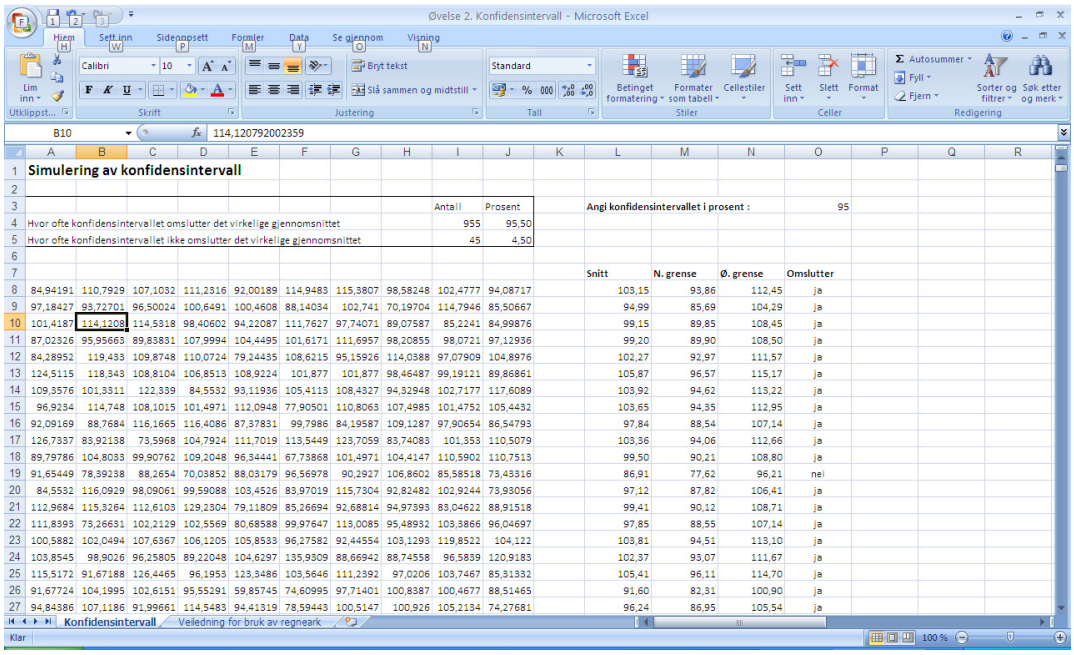
På dette grunnlaget ønsker vi å finne et anslag for hva det gjennomsnittlige fettinnholdet er i alle kjøttdeiger som Gilde produserer. Vi antar at fettinnholdet er normalfordelt og at standardavviket er lik 1. Gjennomsnittet i stikkprøven er 14,1. Ved å bruke formelen over finner vi at konfidensintervallet blir

[13,48 , 14,72]

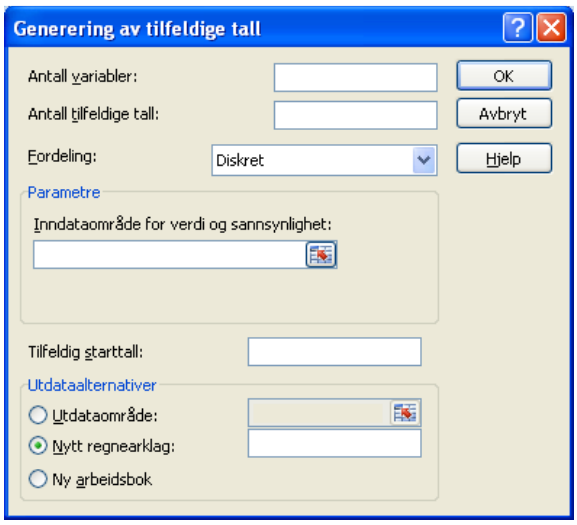
Det forteller oss at det gjennomsnittlige fettinnholdet for alle kjøttdeigene til Gilde med 95% sikkerhet vil ligge mellom 13,48 og 14,72. En annen måte å tolke dette på er å tenke seg at vi trekker ut en rekke ulike stikkprøver, og for hver stikkprøve beregner vi gjennomsnittlig fettinnhold. Da vil ca. 95 % av de konstruerte intervallene omslutte den ukjente fettprosenten.

Konstruksjon av regnearket

Excel er et glimrende verktøy for å lage simuleringer av konfidensintervall. Vi skal her se på hvordan vi kan konstruere et 95% konfidensintervall for gjennomsnittet μ når vi tar en stikkprøve på 10 observasjoner fra et normalfordelt materiale der $\mu = 100$ og standardavviket $\sigma = 15$. Vi skal totalt ta 1000 stikkprøver og for hver og en av disse stikkprøvene skal vi beregne gjennomsnittet og tilhørende konfidensintervall. Vi skal også summere opp hvor mange ganger den virkelige μ ligger innenfor konfidensintervallet og hvor mang ganger den ligger utenfor. Regnearket kan se ut omtrent som vist på neste side



Vi skal nå se på hvordan vi kan konstruere regnearket. Det første vi skal gjøre er å generere de 1000 stikkprøvene. Vi velger å generere stikkprøvene ut i fra en normalfordeling med gjennomsnitt på 100 og standardavvik på 15. For å gjøre dette må vi bruke en funksjon som heter dataanalyse. Hvis den ikke er installert så må du gjøre det først. For å installere den klikker du på Office-knappen oppe i venstre hjørne. Deretter velger du Alternativer for Excel. I menyen du da får opp velger du Tillegg. I listen du får opp så klikker du på Analyseverktøy og deretter på start. I vinduet du da får opp velger du Analyseverktøy og trykker ok. Analyseverktøyet vil da bli installert. Hvis du velger Data fra menyen på øverste linje, skal Dataanalyse ligge helt til høyre. Ved å klikke på Dataanalyse starter du opp dataanalyseverktøyet. Velg der Generering av tilfeldige tall. Følgende vindu vil da dukke opp.



.

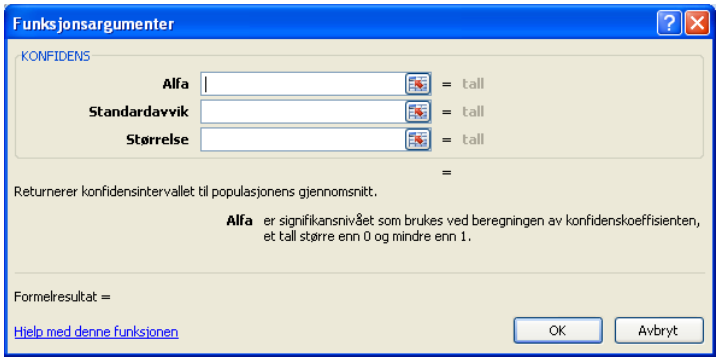
I ruten Antall_variabler skriver du inn hvor mange observasjoner hver stikkprøve skal bestå av. Vi velger 10, men det er selvsagt ikke noe i veien for å velge et annet antall observasjoner. Antall tilfeldige tall vil si hvor mange stikkprøver vi skal ha. Vi velger her 1000. På Fordeling velger du normalfordeling og setter gjennomsnitt til 100 og standardavvik til 15. På Utdataområde skriver du

A8. Stikkprøvene vil da komme fra A8 og nedover. Vi velger å la stikkprøvene gå fra rad 8 og nedover slik at vi kan bruke feltene over til å beregne hvor mange ganger konfidensintervallet omslutter μ. Du kan gjerne merke kolonne A til K og sette kolonnebredden til 8. Det gjør du ved å velge Format og så Kolonnebredde.

Det neste vi skal gjøre er å beregne gjennomsnittet og øvre og nedre grense for konfidensintervallet for hver av stikkprøvene. Gjennomsnittet skal vi beregne i L kolonnen. Flytt markøren til rute L8 og skriv inn formelen =GJENNOMSNITT(A8:J8). Du kan alternativt bruke funksjonsveiviseren. I rute L7 kan du skrive inn en liten overskrift, f. eks Snitt. I rute M8 skal vi beregne nedre grense for konfidensintervallet. Den nedre grensen finner vi ved å regne ut følgende uttrykk.

Nedre grense = $\bar{X} - 1{,}96\frac{\sigma}{\sqrt{n}}$

Når vi skal beregne dette i Excel er det en funksjon som heter Konfidens som vi kan bruke. Den hjelper oss med å beregne siste leddet i uttrykket over. Når vi skal beregne nedre grensen flytter du først musen til rute M8. Start med å skrive inn =L8-. Klikk deretter på funksjonsveiviseren og hent frem funksjonen Konfidens. Du får da opp følgende vindu



Alfa er signifikansnivået og siden det er et 95% konfidensintervall vi skal beregne settes den til 0,05. Standardavviket er i vårt tilfelle 15 og størrelsen på utvalget er 10. Trykk deretter på ok. Den øvre grensen beregnes på tilsvarende måte. Du har nå fått beregnet konfidensintervallet for den første stikkprøven. Du kan nå kopiere cellene L8 til N8 nedover til og med rad 1007, slik at vi får beregnet konfidensintervallet for alle 1000 stikkprøvene. Du kan gjerne sette inn en overskrift i rute M7 og N7 som f. eks N. grense og Ø. grense.

Det neste vi skal gjøre er at vi O kolonnen skal lage en funksjon som kartlegger om konfidensintervallet omslutter det virkelige gjennomsnittet som vi har satt til å være 100. Til det bruker vi HVIS funksjonen som vi skal kombinere med en OG funksjon. Start med å flytte musen til rute O8. Deretter åpner du funksjonsveiviseren og finner frem HVIS funksjonen. Den fyller du ut som vist på neste side

La oss se litt nærmere på første linjen. Det vi sjekker der er om nedre grense (M8) er mindre enn 100 og om øvre grense (N8) er større enn 100. Hvis begge disse er oppfylt vil konfidensintervallet omsluttet gjennomsnittet og det blir skrevet ja i rute O8. I motsatt fall blir det skrevet nei i rute O8. Til slutt kopierer du formelen ned til linje 1007. Det som nå gjenstår er å telle opp hvor mange ganger konfidensintervallet omslutter gjennomsnittet. For å gjøre det skal vi bruke ANTALL.HVIS funksjonen. Du kan flytte musen til rute I4 og åpne ANTALL.HVIS funksjonen med funksjonsveiviseren. Du får da opp følgende bilde

I Området angir du det området vi skal søke i. Det er O8:O1007. I linjen med Vilkår angir du det vi skal se etter, i vårt tilfelle ”ja”. Trykk ok når du har gjort dette. I rute I5 skal vi kartlegge hvor mange ganger vi får nei. Det kan gjøres ved å ta antall forsøk minus antall ”ja”. Det vil si du kan skrive inn =1000-I4 i rute I5. I rute J4 og J5 kan du beregne hva dette blir i prosent.

Til slutt kan du skrive inn en passende overskrift i rute A1 og en passende tekst i rute A4 og A5.

Andre typer konfidensintervall

En kan lett modifisere regnearket slik at det kan beregne et vilkårlig konfidensintervall. Det gjør vi ved at vi angir hvilke konfidensintervall vi skal finne. Vi lar rute O3 være cellen der vi angir hvilke konfidensintervall vi skal beregne. Hvis det er et 95% konfidensintervall skriver vi 95. Hvis det f. eks er 90 skriver vi inn 90 i ruten. I rute L3 kan du skrive inn en liten tekst som f. eks Angi konfidensintervallet i prosent:

Vi må også modifisere formlene våre i rute M8 og N8. I formelen konfidens har vi angitt signifikansnivået til å være 0,05. Når vi henter størrelsen på konfidensintervallet fra rute O3 må vi erstatte 0,05 med (100-O3)/100 for å finne signifikansnivået.

Spørsmål til ettertanke

Se på regnearket ditt, hvor mange ganger omslutter konfidensintervallet det virkelige gjennomsnittet? Ligger det i nærheten av 95% ? Når du skal kjøre en ny simulering går du til Dataanalyse og fyller ut samme vindu som i sted. Du vil da få 1000 nye stikkprøver. Prøv dette noen ganger og se hvilke resultater du får. Prøv også med andre typer konfidensintervall enn 95%, f. eks 90% og 99%. Hvordan blir resultatene i disse tilfellene?

La oss se på noen av stikkprøvene som ikke omslutter det virkelige gjennomsnittet, f. eks en stikkprøve der den nedre grensen er over 100. Hvis vi ser på stikkprøven vil vi oppdage at de fleste verdiene ligger godt over 100, mens på stikkprøver som omslutter det virkelige gjennomsnittet så er det mer jevnt fordelt på begge sider av 100. Når vi genererer tilfeldige tall fra et normalfordelt materiale, vil de fleste observasjonene være i nærheten av gjennomsnittet som i vårt tilfelle er 100. Når vi lager konfidensintervall vil i 95% av tilfelle intervallet omslutte det virkelige gjennomsnittet. Men av og til vil vi få stikkprøver der hovedtyngden av observasjonene ligger på ene siden av gjennomsnittet, og i rundt 5% av tilfellene vil observasjonene ligger så skjevt at konfidensintervallet ikke vil omslutte det virkelige gjennomsnittet.

Øvelse 3. Hypoteser om gjennomsnittet når standardavviket er kjent

Vi skal i denne øvelsen se på hvordan vi kan bruke Excel til å utføre hypotesetesting om gjennomsnittet. I denne øvelsen skal vi ta for oss situasjonen der standardavviket er kjent. I neste øvelse ser på situasjonen der standardavviket ikke er kjent.

En utførlig beskrivelse av hvordan en utfører hypoteseprøving om gjennomsnittet finnes i flere bøker. Et godt alternativ er Sannsynlighetsregning og statistisk metodelære av Knut Ole Lysø. Før vi ser på hvordan vi utfører dette i Excel, så tar vi et lite eksempel som viser gangen i hypoteseprøvingen.

Fettinnholdet i kjøttdeigen til et firma skal i gjennomsnitt ligge på 14%. Vi antar at fettinnholdet er normalfordelt og at standardavviket er 1. Du har lenge hatt mistanke om at fettinnholdet er høyere enn 14%, og bestemmer deg for å sjekke dette. Det gjør du ved å plukke ut 10 forskjellige pakker med kjøttdeig for deretter å registrere fettinnholdet er i hver enkelt pakke. Vi setter opp følgende hypoteser

$$H_0 : \mu = 14 \quad \text{mot} \quad H_1 : \mu > 14$$

Ved hjelp av sentralgrenseteoremet kan en vise at $\bar{X}$ (stikkprøvegjennomsnittet) er normalfordelt med gjennomsnitt på 14 og standardavvik på $\frac{1}{\sqrt{10}}$. Det vi søker nå er en grense slik at

$$P(\bar{X} \geq grense \mid \mu = 14) = 0{,}05$$

En kan da vise at

$$grense = 14 + 1{,}645 \cdot \frac{1}{\sqrt{10}} = 14{,}52$$

Forutsatt at fettinnholdet i snitt faktisk er på 14% så betyr dette at sjansen for at gjennomsnittet på en tilfeldig valgt stikkprøve skal ligge under 14,52% er 95%. Hvis vår stikkprøve har et fettinnhold som er på over 14,52% forkaster vi H_0 , og konkluderer med at fettinnholdet er for høyt.

Konstruksjon av regnearket

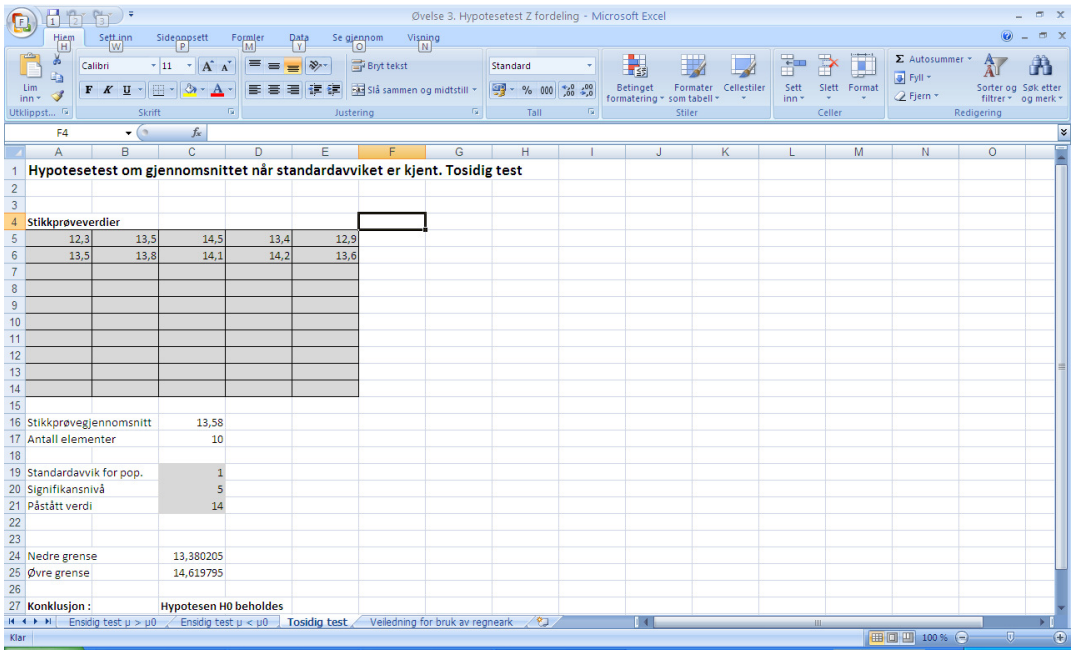
Vi skal nå konstruere et regneark som vi kan bruke til å gjennomføre en hypotesetest. Vi skal ta utgangspunkt i eksempelet med fettinnholdet til kjøttdeig, men vi skal lage det generelt slik at det kan brukes på andre hypoteser også. Vi skal lage tre regneark, et for tosidig test, et for de to typene med ensidige tester. Vi skal først ta for oss hypotesen

$$H_0 : \mu = \mu_0 \quad \text{mot} \quad H_1 : \mu > \mu_0$$

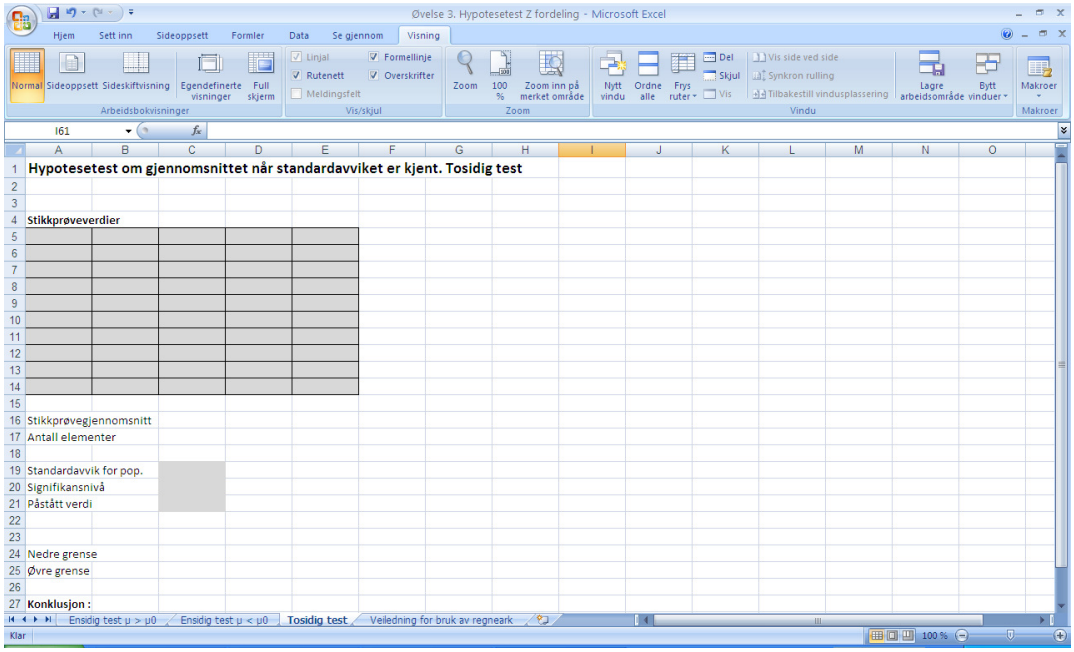
Vi tenker oss at vi tar en stikkprøve på 10 kjøttdeiger med følgende resultat

15,2	14,6	13,5	14,4	14,6
14,3	15,1	14,3	14,6	13,6

Regnearket vi skal konstruere skal se ut omtrent som vist under.

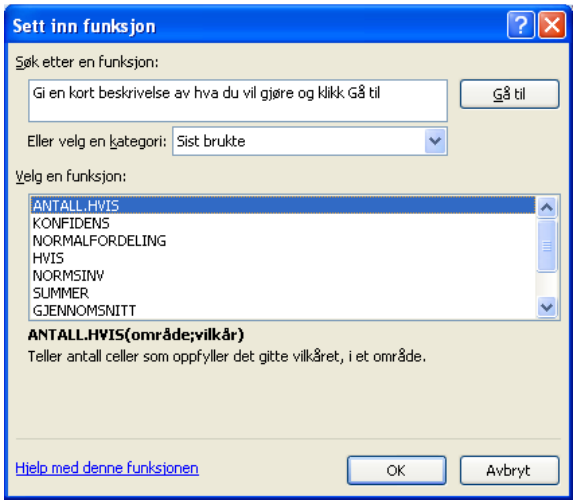


De grå feltene er felter vi skal fylle ut, mens de hvite feltene er felter som maskinen beregner for oss. Vi starter med å lage grunnstrukturen i regnearket. Det vi si at du lager et regneark som vist under.



Når du skal skrive symbolet μ klikker du på Sett inn og velger deretter symbol. Du vil der finne symbol for μ. Vi er nå klare til å starte konstruksjonen av selve regnearket. Først fyller du inn verdiene på stikkprøven i eksempelet i det grå feltet. Regnearket skal konstrueres slik at det kan håndtere stikkprøver med inntil 50 elementer. Når stikkprøveverdiene er fylt inn fyller vi inn standardavvik, konfidensnivå og hva påstanden H_0 er i det nederste grå feltet. I vårt tilfelle har vi satt standardavviket til 1. Vi kan sette signifikansnivået til 5 % og påstanden er at $\mu = 14$. I rute C19 skriver vi inn 1, i rute C20 skriver vi inn 5 og i rute C21 skriver vi inn 14. I rute C16 skal vi regne ut

gjennomsnittet. Klikk først på Formler i menyen og klikk deretter på knappen f_x helt til venstre i menyen. Du får da opp følgende vindu.



Velg Alle istedenfor Sist brukte og blad deg nedover til du finner funksjonen Gjennomsnitt. Dobbeltklikk på denne. Merk deretter området vi skal finne gjennomsnittet av, det vil si cellene A5 til og med celle E14 og klikk på ok. Selv om vi merker alle 50 cellene ignorerer Excel de blanke cellene når den regner ut gjennomsnittet. Det neste vi skal gjøre er å telle opp hvor mange elementer stikkprøven inneholder. Flytt først markøren til rute C17. Vi bruker deretter funksjonsveiviseren slik vi gjorde for gjennomsnitt og leter oss frem til en funksjon som heter Antall. Dobbeltklikk på denne og merk det samme området som i sted og trykk ok.

Det som nå gjenstår er å beregne grensen. I vårt tilfelle vil grensen bli

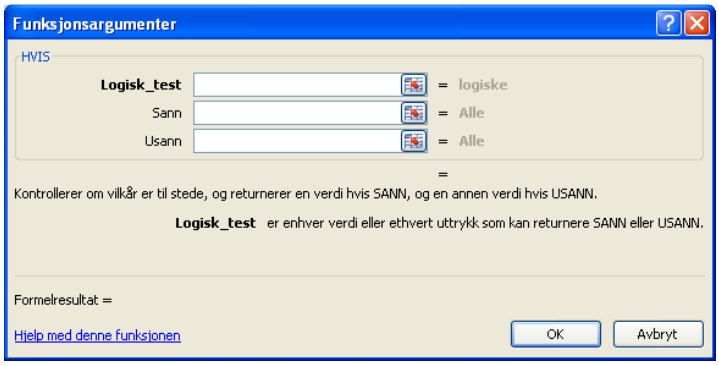
$$grense = 14 + 1{,}645 \cdot \frac{1}{\sqrt{10}} = 14{,}52$$

I regnearket skal vi gjøre dette generelt ved at vi bruker verdien i rute C21 i stedet for 14 og i stedet for 1,645 bruker vi en formel til å beregne hva z_α skal bli på bakgrunn av signifikansnivået vi har satt. Du kan nå flytte musen til rute C24 og starte med å skrive inn formelen

=C21+NORMSINV((100-C20)/100)*C19/ROT(C17)

NORMSINV beregner z_α verdien på bakgrunn av signifikansnivået som vi har skrevet inn i rute C20. Formelen NORMSINV finner z_α på ved at vi bruker 1-signifikansnivået som argument. Siden vi har oppgitt signifikansnivået som heltall i regnearket må vi først ta 100 minus signifikansnivået og deretter dele det på 100. Grensen i vårt eksempel skal da bli 14,52.

Vi ser at gjennomsnittet til stikkprøven vår ligger under grensen vi har funnet. Det betyr at vi beholder hypotesen H_0. Vi har ikke på dette nivået grunnlag for å beskylde kjøttdeigprodusenten for å ha et for høyt fettinnhold i kjøttdeigene. I rute C27 skal vi skrive inn hva konklusjonen blir. Hvis gjennomsnittet på vår stikkprøve ligger over grensen skal vi forkaste H_0. Hvis gjennomsnittet derimot ligger under grensen beholder vi H_0. For å avgjøre dette bruker vi HVIS funksjonen. Du finner HVIS funksjonene ved å bruke funksjonsveiviseren. Når du åpner HVIS funksjonen får du opp følgende vindu.



I feltet etter Logisk_test skriver du inn hva vi skal teste. I vårt tilfelle blir det om C16>C24. Hvis det er sant skal hypotesen forkastes og vi skriver i feltet etter Sann "Hypotesen H0 forkastes". I feltet etter Usann skriver vi "Hypotesen H0 beholdes". Trykk deretter på ok.

Regnearket ditt skal nå se ut slik det er vist innledningsvis i denne øvelsen. Sett nå signifikansnivået til 1%. Hva blir konklusjonen på testen? Prøv deretter med 10%. Hva blir resultatet nå?

Hypotesetest når $\mu < \mu_0$

Vi skal nå se på situasjonen når $\mu < \mu_0$. Vi kan også her bruke eksempelet med fettinnholdet i kjøttdeig. Vi kan tenke oss at lederen for bedriften har mistanke om at fettinnholdet er for lavt i forhold 14% som det skal være. Hvis fettinnholdet blir for lavt vil det påføre bedriften større kostnader siden kjøtt er dyrere enn fett. La oss anta at vi tar en stikkprøver på 10 kjøttdeiger og får følgende verdier.

12,3	13,5	14,5	13,4	12,9
13,5	13,8	14,1	14,2	13,6

Vi ønsker nå å teste

$$H_0 : \mu = 14 \quad \text{mot} \quad H_1 : \mu < 14$$

Grensen vil være gitt ved

$$grense = 14 - 1{,}645 \cdot \frac{1}{\sqrt{10}} = 13{,}48$$

forutsatt at signifikansnivået er 5 %.

Når vi skal konstruere regnearket kan vi bruke det vi gjorde i sted. Kopier hele arket som du laget i sted og kopier det inn i ark2. Du kan gjerne endre navnet på ark2 til noe som er mer beskrivende som for eksempel Ensidig test $\mu < \mu_0$. Stikkprøveverdien endrer du slik at de stemmer med tabellen over. Vi må endre litt på formelen for grense. I stedet for + etter C21 skriver du – slik at formelen blir

=C21-NORMSINV((100-C20)/100)*C19/ROT(C17)

Vi må også endre litt på rute C27 der vi har konklusjonen. I stedet for C16>C24 skal vi skrive C16<C24. Ellers er formelen lik.

Prøv også dette regnearket for ulike signifikansnivåer og se hva du får.

Tosidig test

Det siste vi skal se på i denne øvelsen er tosidig test. I en tosidig test søker vi både en nedre grense og en øvre grense for hva vi kan akseptere. Hypotesen vil teste vil i dette tilfelle være

$$H_0 : \mu = 14 \quad \text{mot} \quad H_1 : \mu \neq 14$$

Den nedre grensen vil da være gitt ved

$$nedre\ grense = 14 - 1{,}96 \cdot \frac{1}{\sqrt{10}} = 13{,}38$$

mens den øvre grensen vil være

$$øvre\ grense = 14 + 1{,}96 \cdot \frac{1}{\sqrt{10}} = 14{,}62$$

forutsatt at signifikansnivået er 5%.

Ved konstruksjon av dette regnearket kan vi bruke det vi allerede har gjort. Det enkleste er å kopiere det forrige arket du laget over til ark3. (Ark3 endrer du navn på til tosidig test.) Vi må gjøre noen små modifikasjoner på arket for å tilpasset det til en tosidig test. Rute A24 hvor det står grense endrer du til nedre grense. I rute A25 kan du skrive øvre grense. Formelen for grense må vi modifisere litt og det er det som står i NORMSINV vi må justere. Det som =NORMSINV(0,95) beregner er z_α i uttrykket

$$P(Z < z_\alpha) = 0{,}95$$

I dette tilfelle vil $z_\alpha = 1{,}645$. I en tosidigtest søker vi imidlertid en z_α som oppfyller kriteriet

$$P(Z < z_\alpha) = 0{,}975$$

Noe som gir $z_\alpha = 1{,}96$. Generelt kan vi modifisere formelen i rute C24 ved å erstatte (100-C20)/100 med (100-C20/2)/100. Formelen som skal stå i rute C24 blir da

=C21-NORMSINV((100-C20/2)/100)*C19/ROT(C17)

Formelen i rute C25 blir helt tilsvarende bare at vi har setter + etter C21 istedenfor -. Også rute C27 der vi avgjør om hypotesen beholdes eller forkastes må vi endre på. Det er feltet etter Logisk_test som må endres. I en tosidig test skal H_0 forkastes hvis gjennomsnittet er under nedre grense eller over den øvre grensen. Det betyr at vi i feltet etter Logisk_test må skrive ELLER(C16<C24;C16>C25) istedenfor C16<C24.

Når dette er gjort skal regnearkene være klare til bruk. Test ut regnearkene med noen oppgaver fra læreboken du bruker og se hvordan de fungerer.

Øvelse 4. Hypoteser om gjennomsnittet når standardavviket er ukjent

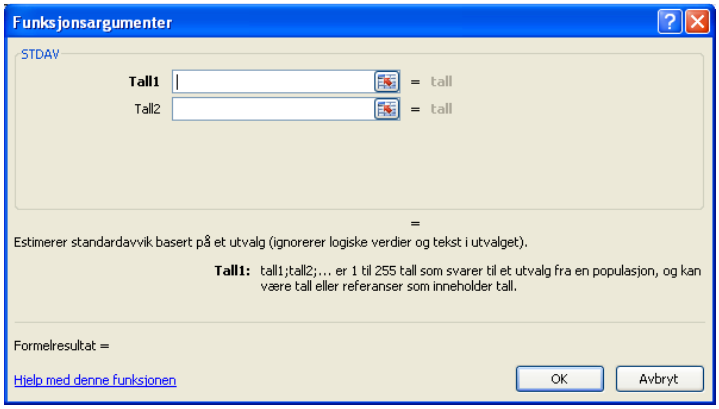
I denne øvelsen skal vi se på hypotesetest om gjennomsnittet når standardavviket er ukjent. Denne øvelsen forutsetter at Øvelse 3 er gjort først, da vi i denne øvelsen kopierer store deler av dette regnearket. Når standardavviket σ til populasjonen er ukjent kan vi estimere standardavviket på bakgrunn av en stikkprøve. Det estimerte standardavviket betegnes gjerne med $\hat{\sigma}$. En kan nå vise at $\bar{X}$ i dette tilfelle følger en såkalt student t-fordeling. Når standardavviket var kjent brukte vi en faktor på 1,645 for en ensidig test på 5% nivå. Når $\bar{X}$ er t-fordelt må vi bruke en annen faktor. Denne faktoren er avhengig av hvor stor stikkprøven er. Generelt snakker vi om antall frihetsgrader og antall frihetsgrader er størrelsen på stikkprøven minus 1. Med andre ord, hvis vi har en stikkprøve på 10 observasjoner vil vi ha 9 frihetsgrader. Det er utarbeidet tabeller som gir den kritiske verdien når signifikansnivået og antall frihetsgrader er kjent.

Konstruksjon av regnearket

Selve regnearket er svært likt det vi allerede har konstruert for tilfelle der standardavviket er kjent. Det enkeleste er nok å lagre dette regnearket som et nytt regneark som vi kan kalle for hypotesetest t-fordeling. Når det er gjort skal vi modifisere regnearket slik at det blir tilpasset situasjonen med at standardavviket er ukjent. Vi starter med regnearket der vi ser på testen

$$H_0 : \mu = \mu_0 \quad \text{mot} \quad H_1 : \mu > \mu_0$$

Vi tar utgangspunkt i det samme eksempelet med fettinnhold i kjøttdeigpakker som vi brukte i forrige oppgave. På forrige regnearket var imidlertid standardavviket kjent, men her er det ukjent. I rute A19 skal vi beregne det estimerte standardavviket. Det gjør vi ved først å flytte markøren til ruta A19 og deretter åpne funksjonsveiviseren. Der leter du deg frem til funksjonen STDAV. Du får da opp følgende vindu



La musen stå i hvite feltet etter Tall1 og merk deretter det grå feltet hvor stikkprøveverdiene står. Husk å merke hele feltet slik at regnearket vil fungere for stikkprøver på opptil 50 observasjoner.

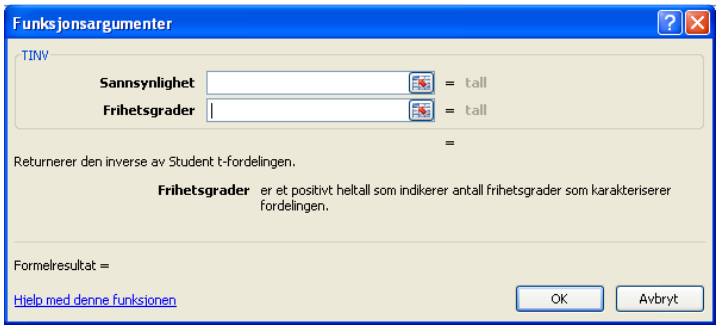
Det neste formelen vi må modifisere er formelen som gir oss grensen. Vi kan vise at grensen er gitt ved

$$grense = \mu_0 + t_\alpha \cdot \frac{\hat{\sigma}}{\sqrt{n}}$$

For å finne faktoren t_α kan vi bruke en formel som heter TINV. Flytt først musen til rute C20. Der skriver vi inn

=C21+

Deretter klikker du på funksjonsveiviseren og åpner funksjonen TINV. Du får da opp følgende vindu



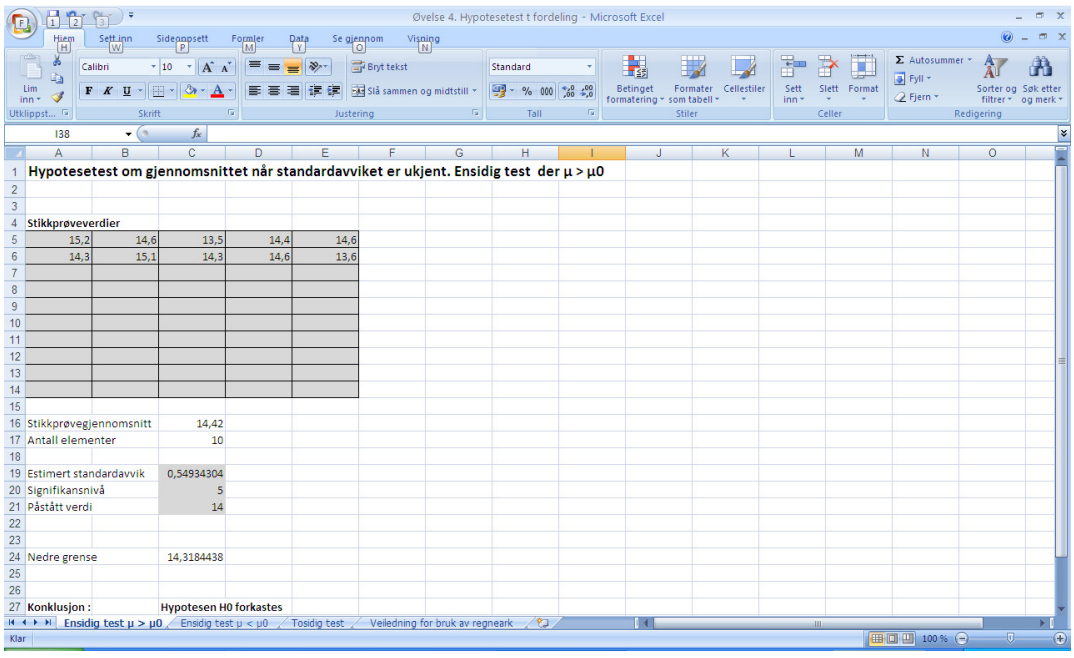
På sannsynlighet skal vi skrive inn signifikansnivået. Nå er denne funksjonen basert på en tosidig test. Siden vi har en ensidig test må vi gange verdien med 2. Det betyr at der hvor det står Sannsynlighet skriver du inn 2*C20/100 og på Frihetsgrader skriver du inn C17-1. Trykk deretter på ok. Du har nå fått formelen

=C21+TINV(2*C20/100;C17-1)

Det siste vi skal gjøre er å multiplisere TINV med uttrykket $\frac{\hat{\sigma}}{\sqrt{n}}$, det vil si med C19/ROT(C17). Vi får da formelen

=C21+TINV(2*C20/100;C17-1)*C19/ROT(C17)

Regnearket skal nå se ut som vist under.



Regnearkene for situasjonene for den andre ensidige testen og for den tosidige testen modifiseres på helt tilsvarende måte.

Øvelse 5. Hypoteser i en binomisk situasjon

I denne øvelsen skal vi se på hvordan vi kan bruke Excel til å utføre hypotesetester om en binomisk p. I boken Sannsynlighetsregning og statistisk metodelære av Knut Ole Lysø er det beskrevet et utmerket eksempel der en tester om en terning gir for mange seksere. Dette vil være en ensidig test og vi kan sette opp følgende hypoteser for denne situasjonen.

$$H_0 : p = \frac{1}{6} \quad \text{mot} \quad H_1 : p > \frac{1}{6}$$

Det vi søker er en grense som oppfyller kravet

$$P\left(X \geq grense \mid p = \frac{1}{6}\right) \leq 0{,}05$$

Vi kan bruke binomialfordeling og regne dette eksakt. Imidlertid vil ofte normalfordelingen bli brukt i praksis, og i de fleste tilfeller gir den akseptable resultater. Når vi bruker normalfordelingen kan vi vise at hvis vi har hypotesen

$$H_0 : p = p_0 \quad \text{mot} \quad H_1 : p > p_0$$

vil grensen være gitt ved

$$grense = np_0 + 0{,}5 + z_\propto \sigma_0$$

Uttrykket np_0 og σ_0 er forventningen og standardavviket forutsatt at H_0 er riktig. Tallet 0,5 er yates korreksjon.

Hvis vi i vårt eksempel med terningen tenker oss at vi kaster en terning 100 ganger og setter signifikansnivået til 5% vil grensen bli

$$grense = 100 \cdot \frac{1}{6} + 0{,}5 + 1{,}645 \cdot \sqrt{100 \cdot \frac{1}{6} \cdot \frac{5}{6}} = 23{,}3$$

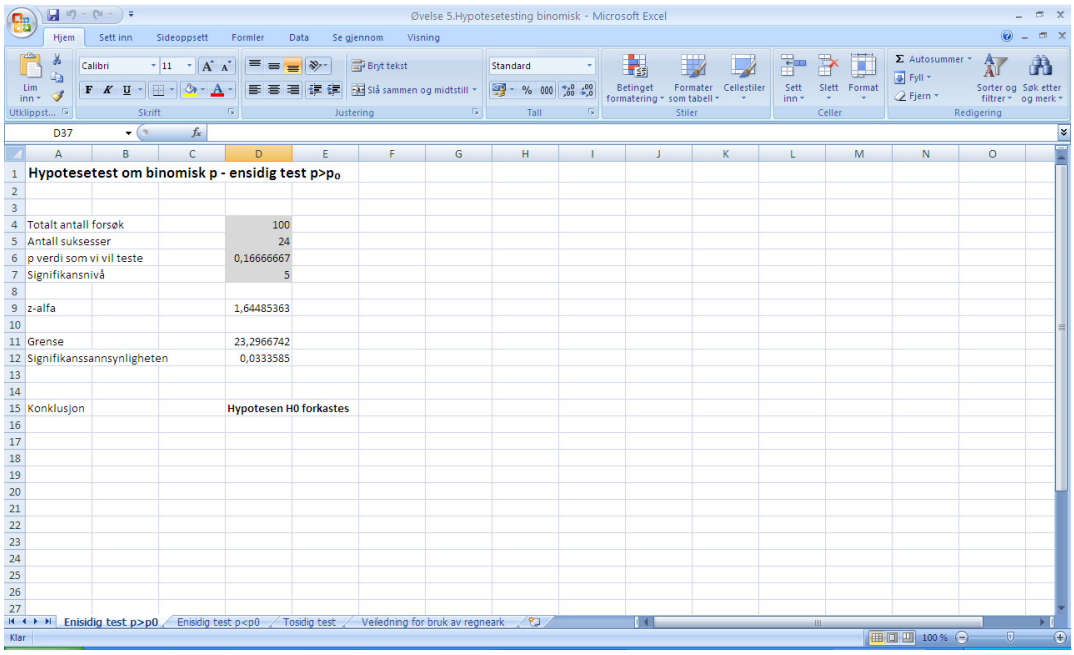
Dette betyr at det er 5% sjanse for å 23,3 eller flere seksere på 100 kast. Siden vi ikke kan få 23,3 seksere er det mest korrekt å runde oppover til 24. Hvis vi i vårt forsøk får 23 eller færre seksere beholder vi H_0 hvilket betyr at ikke har grunnlag for å påstå at det er noe galt med terningen. Hvis vi derimot får 24 eller flere seksere forkaster vi H_0 og trekker konklusjonen at her er det noe galt med terningen. Sjansen for at vi feilaktig trekker denne konklusjonen vil være under 5%.

<u>Konstruksjon av regneark</u>

Vi skal nå se på hvordan vi kan konstruere et regneark som utfører denne hypotesetesten. Vi starter også her med et regneark for den ensidige hypotesen

$$H_0 : p = p_0 \quad \text{mot} \quad H_1 : p > p_0$$

Regnearket kan se ut omtrent som vist på neste side



I regnearket vi skal konstruere skal vi skrive inn antall forsøk, hvor mange suksesser vi har, hva vi vil teste og signifikansnivået. På grunnlag av dette skal regnearket beregne grensen for oss og trekke en konklusjon på testen. Du kan starte med å skrive inn teksten og fylle inn verdiene i de grå rutene. Vi bruker samme eksempel som tidligere med terningkast og vi ser for oss at vi gjør 100 forsøk med 24 suksesser. Antall forsøk kan være vilkårlig, men helst ikke for få forsøk siden vi bruker normaltilnærmelsen. I vårt eksempel setter vi p-verdien vi skal teste til $\frac{1}{6}$. (Husk at du må skrive =1/6 i rute D6). Vi velger i førsteomgang å sette signifikansnivået til 5%.

For å kunne beregne grensen må vi først beregne $z_\propto$. Vi skal beregne $z_\propto$ i rute D9. For å beregne $z_\propto$ bruker vi NORMSINV funksjonen. Dette gjøres på tilsvarende måte som i øvelse 3. I vårt tilfelle blir formelen

=NORMSINV((100-D7)/100)

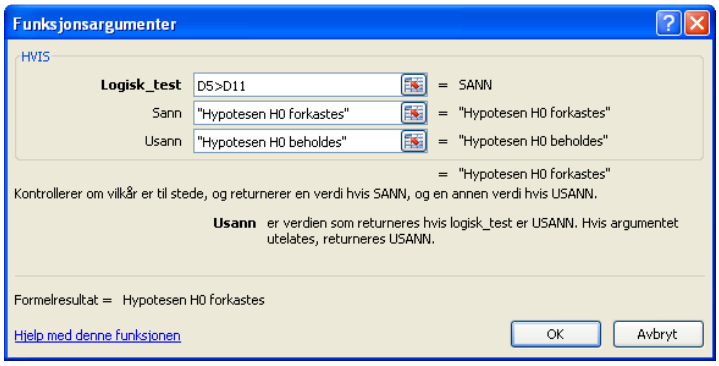
Det neste vi skal gjøre er å beregne grensen i rute D11. Fra i sted så vi at formelen

$$grense = np_0 + 0{,}5 + z_\propto \sigma_0$$

gir oss grenseverdien, der σ_0 er standardavviket som er gitt ved $\sqrt{np_0(1-p_0)}$ Formelen som vi skal skrive inn i rute D11 blir

=D4*D6+0,5+D9*ROT(D4*D6*(1-D6))

Det siste vi skal gjøre er å avgjøre om hypotesen skal forkastes eller ikke. Dette skal vi gjøre i rute D15. Til det må vi bruke HVIS funksjonen for å sjekke om antall suksesser vi har fått ligger over eller under grensen. Vi åpner HVIS funksjonen med funksjonsveiviseren og fyller den ut som vist under.



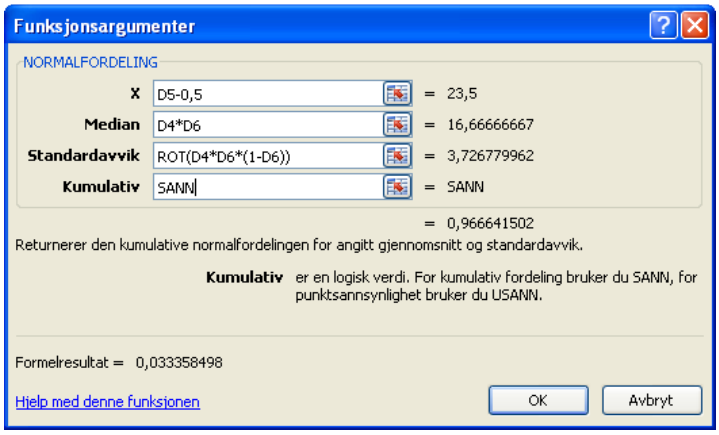
Vi kan gjerne regne ut signifikanssannsynligheten også. Det kan vi gjøre i rute D13. Når vi regner ut signifikanssannsynligheten er det uttrykket

$$P\left(X \geq S \mid p = \frac{1}{6}\right)$$

der S er antall suksesser. Hvis vi bruker normaltilnærmelsen kan vi vise at dette uttrykket kan omformes til

$$1 - P\left(Z \leq \frac{S - 0{,}5 - np_0}{\sigma_0}\right)$$

Dette utrykket kan vi legge inn i rute D13. Start med å skrive =1- i rute D13. Deretter åpner du funksjonsveiviseren og åpner normalfordelingen. Du får da opp et vindu som fyller ut som vist under



Regnearkene for den andre ensidige testen og for den tosidige testen kan konstrueres på helt tilsvarende måte.

Øvelse 6. Test på forskjell i populasjonsgjennomsnitt

I denne øvelsen skal vi teste om to populasjonsgjennomsnitt kan påstås å være like eller ikke. Datamaterialet henter vi fra en tilfeldig stikkprøve fra hver av populasjonene. Vi antar at stikkprøvene er uavhengig av hverandre. Vi antar også at populasjonene er normalfordelte. Det vi ønsker å teste ut er om gjennomsnittene μ_1 og μ_2 i de to populasjonene er like eller ikke. Dette kan vi formulere som en hypotese

$$H_0 : \mu_1 = \mu_2 \quad \text{mot} \quad H_1 : \mu_1 \neq \mu_2$$

Ofte velger vi heller å se på differansen $D = \mu_1 - \mu_2$ og formulere hypotesen slik

$$H_0 : D = 0 \quad \text{mot} \quad H_1 : D \neq 0$$

La oss se på et lite eksempel. Vi tenker oss at en bedrift har to maskiner som begge produserer samme type vare. Varen skal i prinsippet veie 100 gram uavhengig av hvilken maskin som varen produseres av, men det vil likevel være noe variasjon. Vi kan imidlertid anta at vekten er normalfordelt. Vi vil nå undersøke om det er signifikant forskjell i gjennomsnittet til disse to populasjonene. Vi antar nå at vi tar en stikkprøve fra produksjonen til hver av maskinene. Resultatet er vist i tabellen under.

Maskin 1	102	101	99	98	100	97	103	102	100
Maskin 2	101	99	98	97	101	100	101		

Dersom standardavviket til populasjonene er kjent kan vi vise at $\hat{D} = \bar{X} - \bar{Y}$ er normalfordelt med forventning D og standardavvik $\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$ der σ_1, σ_2 og n, m er standardavvik og antall observasjoner for henholdsvis stikkprøve 1 og 2. I praksis er som oftest standardavviket ukjent og vi må da estimere standardavviket og vi bør bruke t-fordelingen istedenfor. Den nedre grensen for hva vi kan akseptere vil derfor være

$$nedre\ grense = -t_{\propto}\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}$$

Den øvre grensen vil tilsvarende være

$$Øvre\ grense = t_{\propto}\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}$$

Faktoren $t_{\propto}$ vil avhenge av antall observasjoner i stikkprøven og signifikansnivået. Dessverre er det ikke en enkel regel for å finne ut hvor mange frihetsgrader vi har. Men det finnes en metode som heter Welsh' metode. Vi må først beregne forholdet mellom stikkprøvevariansene dvs.

$$W = \frac{\hat{\sigma}_1^2/n}{\hat{\sigma}_2^2/m}$$

Antall frihetsgrader beregnes deretter ved hjelp av formelen

$$v = \frac{(1+W)^2}{\frac{W^2}{n-1} + \frac{1}{m-1}}$$

La oss nå se hva vi får i vårt eksempel. Vi finner at

$\bar{X} = 100{,}22$ og $\bar{Y} = 99{,}57$ og

$\hat{\sigma}_1^2 = 3{,}94$ og $\hat{\sigma}_2^2 = 2{,}62$

Dette gir at

$$\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}} = \sqrt{\frac{3{,}94^2}{9} + \frac{2{,}62^2}{7}} = 0{,}90$$

Vi finner videre at

$$W = \frac{\hat{\sigma}_1^2/n}{\hat{\sigma}_2^2/m} = \frac{3{,}94/9}{2{,}62/7} = 1{,}17$$

Antall frihetsgrader blir da

$$v = \frac{(1+1{,}17)^2}{\frac{1{,}17^2}{9-1} + \frac{1}{7-1}} = 13{,}94$$

Som vi runder ned til 13 frihetsgrader. Med signifikansnivå på 95% vil derfor $t_\propto = 2{,}16$ slik at

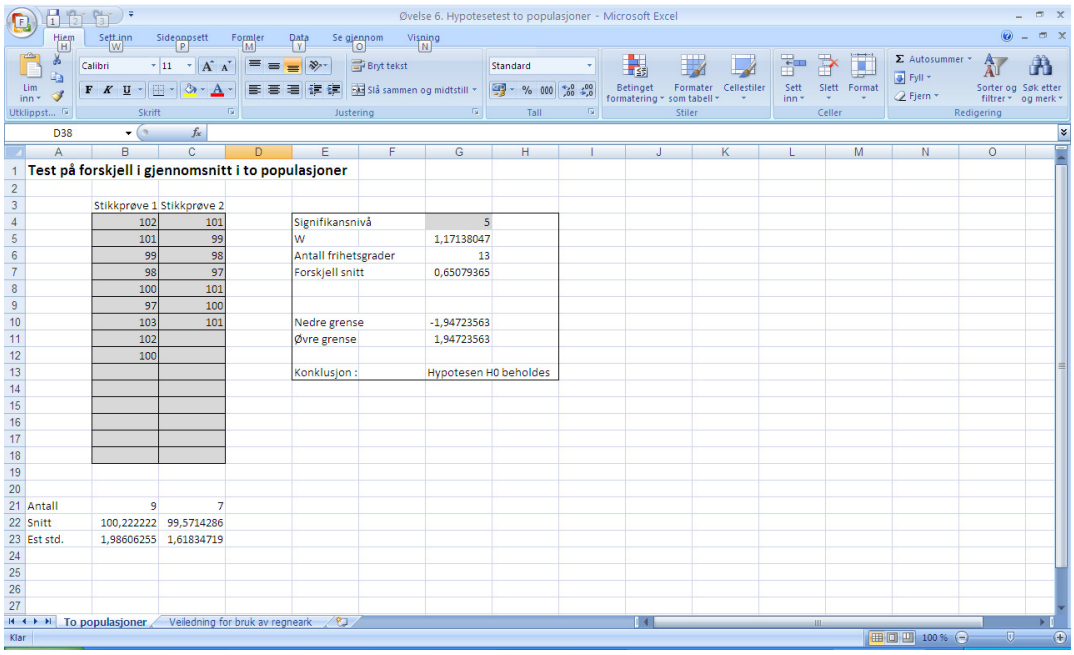
$$nedre\ grense = -2{,}16 \cdot 0{,}90 = -1{,}95$$

$$øvre\ grense = 2{,}16 \cdot 0{,}90 = 1{,}95$$

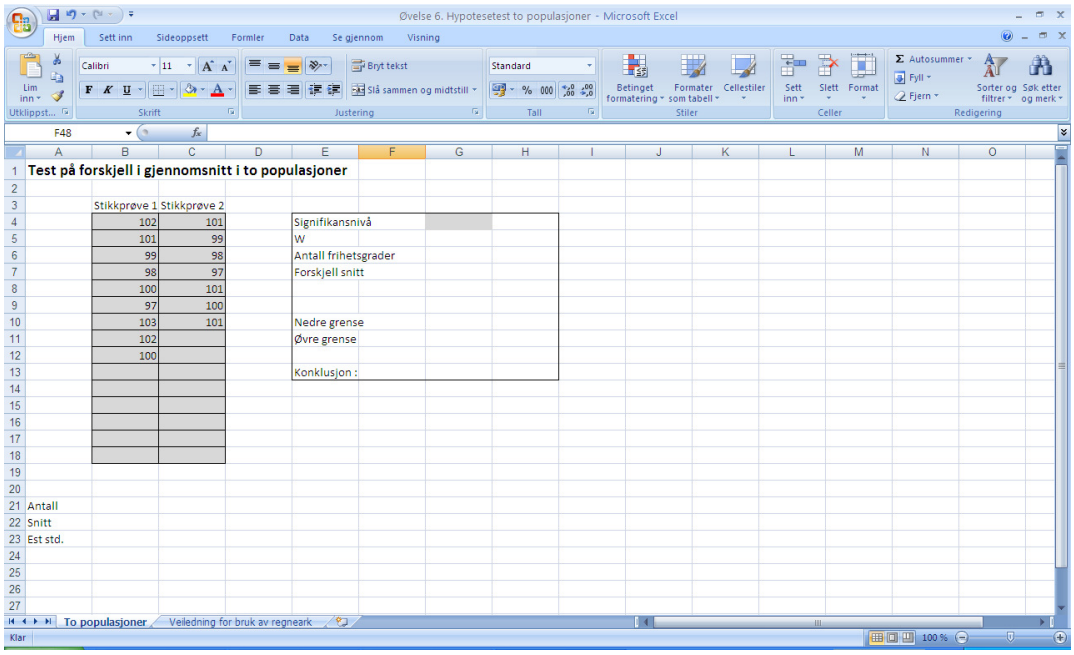
Siden forskjellen på våre stikkprøver bare var 0,65 beholder vi H_0 .

Konstruksjon av regnearket

Vi skal nå se hvordan vi kan konstruere et regneark som gjennomfører denne hypotesetesten. Vi konsentrerer oss om å lage et regneark for situasjonen der standardavviket er ukjent og der vi bruker t-fordelingen. Et regneark basert på Z-fordelingen kan lages på helt tilsvarende måte. Målet vårt er å konstruere et regneark som vist på neste side.



Vi skal konstruere et regneark som håndterer stikkprøver på inntil 15 observasjoner fra hver stikkprøve. (Det er selvsagt ingenting i veien for å lage regneark som tar høyde for stikkprøver med flere observasjoner). Vi skal også her starte med å lage selve strukturen før vi fyller inn formlene. Start med å åpne et nytt regneark og skriv inn teksten som vist i regnearket under.



Feltene som er markert med grått er felter der vi skal fylle inn verdier, mens de øvrige er felter der vi skal bruke formler for å beregne verdiene. Det første vi skal gjøre er å fylle inn stikkprøvene i det store grå feltet og signifikansnivået i det lille grå feltet. Signifikansnivået kan du sette til 5 %. (Skriv inn tallet 5 i rute G4) Det neste vi skal gjøre er å beregne antall observasjoner, gjennomsnitt og estimert standardavvik for stikkprøvene. I rute B21 og C21 skal vi beregne hvor mange observasjoner

det er i hver av stikkprøvene. Det gjør vi ved å bruke ANTALL funksjonen. I rute B22 og C22 skal vi beregne gjennomsnittet ved hjelp av GJENNOMSNITT funksjonen. I rute B23 og C23 skal vi estimere standardavviket for hver av populasjonene. Det finnes en funksjon i Excel som gjør dette og det er funksjonen STDAV.

Neste skritt blir å beregne antall frihetsgrader. Som vi så i sted krever det noe regnearbeid. Vi skal først beregne

$$W = \frac{\hat{\sigma}_1^2/n}{\hat{\sigma}_2^2/m}$$

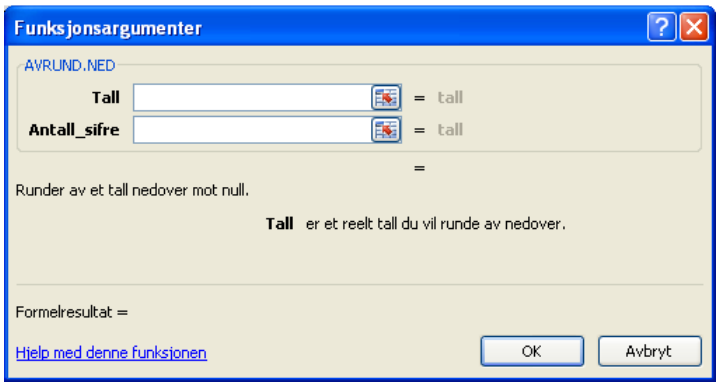
Det skal vi gjøre i celle G5. For å beregne W bruker du følgende formel

=(B23^2/B21)/(C23^2/C21)

Vi henter det estimerte standardavviket fra rute B23 og C23 og antall observasjoner fra rute B21 og C21. Det neste vi skal gjøre er å beregne antall frihetsgrader. Vi husker fra i sted at formelen

$$v = \frac{(1+W)^2}{\frac{W^2}{n-1} + \frac{1}{m-1}}$$

gir oss antall frihetsgrader. Formelen er ikke spesielt pen, og det blir heller ikke formelen i Excel. Vi ønsker også å runde svaret ned til nærmeste heltall. Vi starter med å åpne funksjonen for å avrunde tall. Bruk funksjonsveiviseren til å finne funksjonen AVRUND.NED. Du får da opp følgende vindu.



Etter Tall kan du fylle ut følgende uttrykk som beregner brøken over.

(1+G5)^2/(G5^2/(B21-1)+1/(C21-1))

Etter Antall_sifre kan du fylle ut 0 som angir at vi skal ha 0 desimaler i tallet vi skal runde ned. Du skal da ha fått følgende formel i rute G6:

=AVRUND.NED((1+G5)^2/(G5^2/(B21-1)+1/(C21-1));0)

I rute G7 skal vi beregne forskjellen mellom gjennomsnittene til de to populasjonene. Dersom vi ønsker at differensen skal være positiv bruker vi ABS funksjonen i Excel. Formelen som skal stå i rute G7 blir derfor

=ABS(B22-C22)

Det som gjenstår da er å finne nedre og øvre grense. Den nedre grensen er gitt ved formelen

$$nedre\ grense = -t_{\propto}\sqrt{\frac{\hat{\sigma}_1^2}{n}+\frac{\hat{\sigma}_2^2}{m}}$$

Verdien $t_{\propto}$ finner vi ved å bruke TINV funksjonen. Ved å skrive inn signifikansnivået og antall frihetsgrader beregner Excel TINV for oss. Formelen for nedre grense som vi skriver i rute G10 blir derfor

=-TINV(G4/100;G6)*ROT(B23^2/B21+C23^2/C21)

Tilsvarende blir formelen for øvre grense

=TINV(G4/100;G6)*ROT(B23^2/B21+C23^2/C21)

Til slutt skal vi i rute G13 avgjøre om vi skal beholde hypotesen eller om vi skal forkaste den. Formelen

=HVIS(G7<G11;"Hypotesen H0 beholdes";"Hypotesen H0 forkastes")

Avgjøre dette spørsmålet. Siden vi rute G7 har brukt absoluttverdi er det tilstrekkelig å sjekke verdien i G7 mot den øvre grensen. Dersom G7 er mindre enn G11 beholder vi H_0 . I motsatt fall forkaster vi den.

Øvelse 7. Regresjon

I denne øvelsen skal vi se på hvordan Excel kan brukes i arbeidet med regresjon. Vi skal først se hvordan vi kan fremstille dataene i et diagram og også hvordan vi kan beregne regresjonslinjen. Vi skal deretter se på hvordan vi kan gjennomføre hypotesetesten

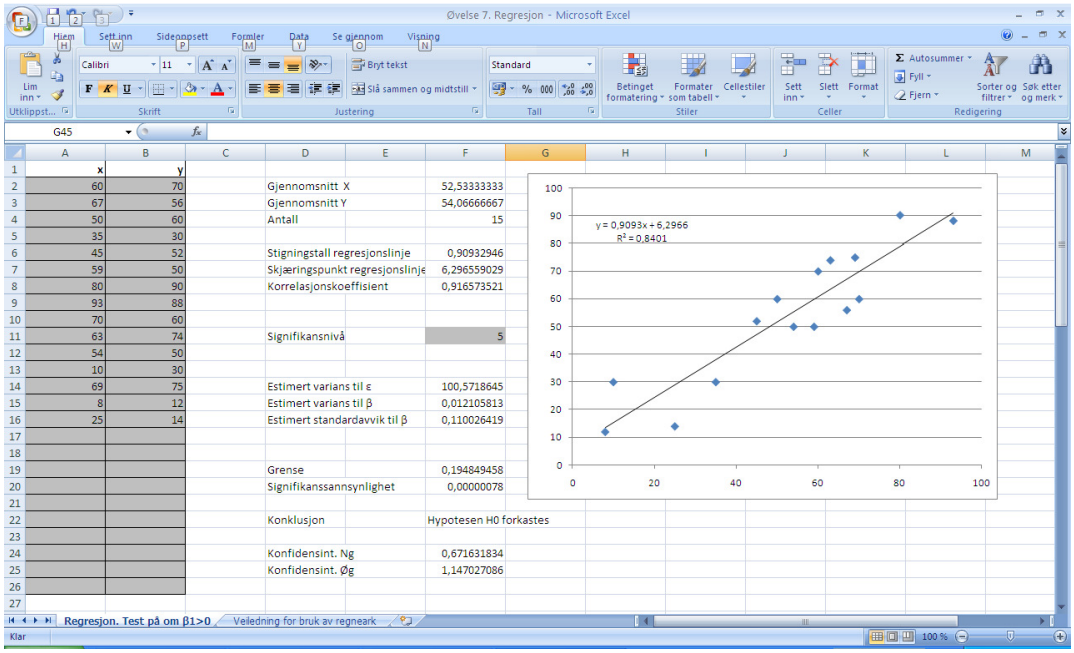
$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 > 0$$

der regresjonslinjen

$$Y = \beta_0 + \beta_1 X$$

beskriver en trend i datamaterialet.

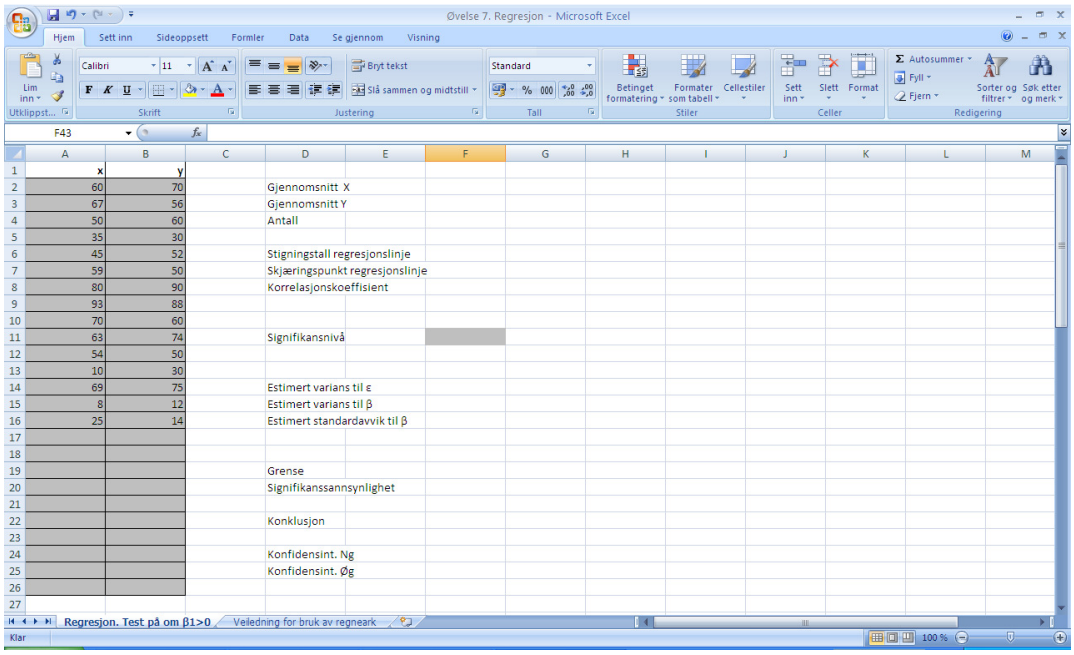
Målet vårt er å konstruere et regneark som vist under.



X	60	67	50	35	45	59	80	93	70	63	54	10	69	8	25
Y	70	56	60	30	52	50	90	88	60	74	50	30	75	12	14

Jeg velger å bruke et konkret eksempel for å vise hva vi skal gjøre. Vi tenker oss at i et matematikkurs gis det to deleksamener, en til jul og en til sommeren. I tabellen under er poengsummen til 15 tilfeldige studenter gitt.

Vi lar X symbolisere første deleksamen og Y den andre deleksamen. Vi skal nå legge dataene inn i Excel og behandle dem. Du kan starte med å skrive inn teksten og deretter verdiene til de to prøvene i det grå feltet. Vi lager regnearket slik at det kan ta hånd om inntil 25 observasjoner, selv om vi i vårt eksempel kun har 15. Regnearket vil da se omtrent slik ut



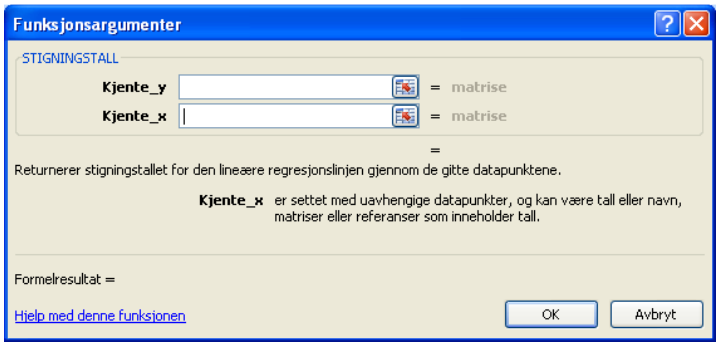
Vi skal nå plotte verdiene i et diagram og tegne opp regresjonslinjen. Start med å merke verdiene i det grå feltet og trykk deretter på Sett inn på menyen. Der velger du punktdiagrammet. Velg diagrammet øverst til venstre. Flytt diagrammet litt til høyre på skjermen slik at det kommer bort fra F kolonnen. Fjern også feltet der hvor det står Serie 1. Når vi skal legge inn regresjonslinjen høyreklikker du med musen på et av punktene og velger Legg til trendlinje. Velg Lineær og hak ut foran Vis formel i diagrammet og hak også ut Vis R kvadrat. Vi har nå fått tegnet opp regresjonslinjen, beregnet likningen for linjen og vi har også fått beregnet korrelasjonskoeffisienten opphøyd i annen. Metoden som Excel bruker for å beregne regresjonslinjen er minste kvadraters metode.

Selve regresjonslinjen kan en også beregne ved hjelp av formlene som finnes i Excel. Før vi gjør det skal vi beregne gjennomsnittet til X og Y i cellene F2 og F3. For å bergene gjennomsnittet bruker vi funksjonen GJENNOMSNITT. I rute F4 beregner vi antall observasjoner ved hjelp av ANATLL funksjonen. Vi trenger ikke disse tallene for å beregne regresjonslinjen, men vi har bruk for dem når vi skal gjennomføre hypotesetesten litt senere.

Vi skal nå se på hvordan regresjonslinjen kan beregnes. Det er ofte vanlig å skrive opp regresjonslikningen på denne måten

$$Y = a + bX$$

I rute F6 skal vi beregne b. Dette er forholdsvis tidkrevende om en skal gjøre det med kalkulator. I Excel finnes det en funksjon som gjør dette for oss. Den funksjonen heter STIGNINGSTALL. Bruk funksjonsveiviseren til å finne denne funksjonen. Du får da opp følgende vindu



I feltet etter Kjente_y merker du av Y verdiene i det grå feltet. Det vil si celle B2 til B16. Etter Kjente_x merker du av tilsvarende X verdier, det vil si celle A2 til A16. Trykk deretter ok. For å beregne a bruker vi funksjonen SKJÆRINGSPUNKT. Den er bygget opp helt likt med funksjonen STIGNINGSTALL. Til slutt i denne delen skal vi beregne korrelasjonskoeffisienten. Funksjonen KORRELASJON hjelper oss med denne beregningen. Finn KORRELASJON i funksjonsveiviseren og merk av X verdiene i feltet Matrise 1 og Y verdiene i feltet Matrise 2. Trykk deretter på ok. Hvis du har gjort dette riktig skal regresjonslinjen vi har funnet samsvare med den som er fremkommet i diagrammet. Uttrykket R^2 som står i diagrammet under regresjonslikningen er korrelasjonskoeffisienten opphøyd i annen.

<u>Hypotesetest</u>

Vi tenker oss at vi har et datamateriale der linjen $Y = \beta_0 + \beta_1 X$ kan oppfattes som trenden i datamaterialet. Dette er linje som populasjonen av punkter vil være samlet omkring. Vi kan vanligvis ikke trekke opp denne linjen fordi β_0 og β_1 er ukjente størrelser. Den generelle regresjonslikningen kan skrives som

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

der en vilkårlig enkeltobservasjon Y_i er uttrykt ved den tilhørende X_i og der ε_i representerer enkeltobservasjoners avvik fra trenden. Som et estimat på trenden i datamaterialet brukes gjerne regresjonslinjen som er gitt ved

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X.$$

Størrelsen som knytter Y til X er β_1. Dersom denne er lik 0 innebærer det at det ikke er noe sammenheng mellom X og Y. Vi ønsker derfor ofte å teste om β_1 er lik 0 eller ikke. Vi skal nå ta for oss hypotesen

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 > 0$$

Vi kan vise at $\hat{\beta}_1$ er normalfordelt med forventningsverdi β_1. Vi betegner standardavviket til β_1 med σ_1. Standardavviket er vanligvis ukjent og en størrelse vi må estimere. Vi kan videre vise at brøken

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_1}$$

er t fordelt med $n - 2$ frihetsgrader. Selve hypotesetesten gjennomføres på tilsvarende måte som hypoteser om gjennomsnittet. Det som gjenstår før vi kan gjøre det, er å beregne det estimerte

standardavviket til $\hat{\beta}_1$. Dette medfører litt utregning da vi først må estimere variansen til ε_i. Den estimerte variansen til ε_i er gitt ved

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum(y_i - \bar{Y})^2 - \hat{\beta}_1^2 \sum(x_i - \bar{X})^2}{n-2}$$

Variansen til $\hat{\beta}_1$ er gitt ved

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum(x_i - \bar{X})^2}$$

Som vi har tidligere har gjort skal vi i regnearket beregne en grense som er slik at

$$P\left(\hat{\beta}_1 \geq grense \,\middle|\, \beta_1 = 0\right) = 0{,}05$$

Denne grensen er gitt ved

$$grense = t_\propto \cdot \hat{\sigma}_1$$

Vi skal nå legge disse verdiene inn i regnearket vårt. Vi starter med å spesifisere signifikansnivået i rute F11. Du kan i første omgang sette det til 5. I rute F14 skal vi beregne estimerte variansen til ε_i det vil si $\hat{\sigma}_\varepsilon^2$. I Excel er det en formel som beregner uttrykk som $\sum(y_i - \bar{Y})^2$ og den heter AVVIK.KVADRER. Den fungerer slik hvis en skal regne ut kvadratavvikene mellom den enkelte y_i og gjennomsnittet $\bar{Y}$, bruker en formelen AVVIK.KVADRERT(B2:B26). Hele uttrykket for $\hat{\sigma}_\varepsilon^2$ blir derfor

=(AVVIK.KVADRERT(B2:B26)-F6*F6*AVVIK.KVADRERT(A2:A26))/(F4-2)

I rute F15 skal vi beregne $\hat{\sigma}_1^2$. Vi bruker også her funksjonen AVVIK.KVADRERT og formelen som skal stå i rute F15 blir da

=F14/AVVIK.KVADRERT(A2:A26)

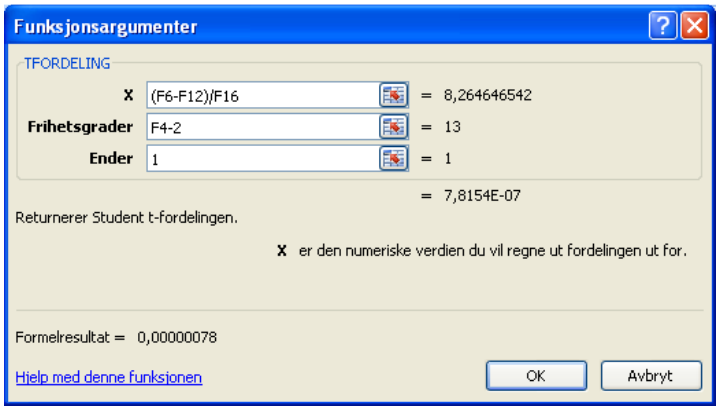
Til sist skal vi i rute F16 beregne standardavviket til $\hat{\beta}_1$. Dette er roten av variansen som vi beregnet i rute F15. Formelen som skal stå i rute F16 blir derfor

=ROT(F15)

Vi er nå klare for å beregne grensen. Som tidligere hjelper funksjonen TINV oss med det slik at formelen i rute F19 blir

=F12+TINV(F11/100*2;F4-2)*F16

Merk her at siden det er en ensidigtest så må vi gange signifikansnivået med 2 for at det skal bli riktig. Vi kan også beregne signifikanssannsynligheten. Funksjonen TFORDELING kan brukes til dette. Flytt musen til celle F20 og åpne når du åpner denne funksjonen i funksjonsveiviseren får du et vindu som du kan fylle ut som vist på neste side.



I feltet etter X er det $\frac{\hat{\beta}_1-\beta_1}{\hat{\sigma}_1}$ vi har fylt inn. I neste feltet skriver du inn antall frihetsgrader. Siden dette er en ensidig test skriver du inn 1 i feltet etter Ender. Det neste vi skal gjøre er å avgjøre om hypotese skal forkastes eller ikke. Hvis F6 > F19 skal vi forkaste H_0 , hvis ikke skal vi beholde den. Vi bruker en HVIS funksjon slik vi tidligere har gjort. Etter at du har funnet den i funksjonsveiviseren fyller du den ut som vist under.

Vi skal også beregne konfidensintervallet for β_1. Det vil være gitt ved

$\hat{\beta}_1 \pm t_\propto \cdot \hat{\sigma}_1$.

Den nedre grensen som vi skriver i rute F24 vil derfor være gitt ved

=F6-TINV(F11/100;F4-2)*F16

Den øvre grensen til konfidensintervallet beregnes ved tilsvarende formel bare at en erstatter – tegnet med + tegn.