Mulig brukte forkortelser:

Forkortelse	Representerer
ΣXi	Sum av verdiene av alle observasjoner av X
$\bar{x}$	Gjennomsnitt av verdiene av alle observasjoner av x
M	Median
C	Typetall (mode)
$\tilde{x}$	Geometrisk gjennomsnitt av verdiene av alle observasjoner av x
SD	Standardavvik
s	Standardavviket i et gitt utvalg
σ	Standardavviket i en gitt populasjon
SE	Standard-"feil"
mdev	Gjennomsnittsavvik
Pr	Sannsynlighet
S	Utfallsrom
P(OI)	Populasjonen (som er relevant)
O	Utfall "hentet" fra utfallsrommet
$\widetilde{\rho_{rn}}$	Observert sannsynlighet etter n "forsøk"
c(X)	Kumulativ sannsynlighet for å observere verdi opp til X
E(X)	Vektet gjennomsnittsforventning for en "tilfeldig" variabel
μ	E(X) på populasjonsnivå
$\bar{x}$	E(X) i et utvalg av populasjonen
E($\bar{x}$)	(Hypotetisk) gjennomsnitt av utvalgs-gjennomsnitt fra "mange" utvalg
V(X)	Variansen av en "tilfeldig" variabel
σ^2	V(X) på populasjonsnivå
v^2	V(X) i et utvalg av populasjonen (/n)
s^2	Samme som v^2, men dividert på (n-1)
$\sigma_{\bar{x}}^2$	(Hypotetisk) varians av $\bar{x}$-verdier fra "mange" utvalg
$\sqrt{v(\bar{x})}$	Standard-"feil"
p	Observert andel av "utfall 1"/"suksess" for en tilfeldig Bernoulli-variabel (i et utvalg)
$\bar{p}$	Sannsynlighet for "utfall 1"/"suksess" for en tilfeldig Bernoulli-variabel (i et utvalg)
π	Sannsynlighet for "utfall 1"/"suksess" for en tilfeldig Bernoulli-variabel (på populasjonsnivå)
EMV	Forventet gjennomsnittlig pengeverdi av hendelse
N	Normal-distribuert variabel
Z	N~(0,1) (Standard normal-distribusjon)
PrDF	Probability density function
n	Utvalgs-størrelse, eller generelt antall observasjoner/"forsøk"
$\bar{N}$	Populasjons-størrelse

t	t-/"student"-distribusjon
df eller v	Grader av frihet
x^2 eller chi	"chi-square"-distribusjon
F	F-distribusjon
CLT	Central limit-teoremet
Q	Estimator
θ	Estimat av populasjonsverdi basert på Q
SSP	Ønsket kvalitet hos Q for å estimere θ, uavhengig as utvalgs-størrelse
LSP	Ønsket kvalitet hos Q for å estimere θ, som fremtrer i store utvalg (asymptotic property)
BLU(E)	Best linear unbiasedness/best linear unbiased estimator
M(S)E	Mean (squared) error (for en estimator)
$\oint$	Vektlegging av lav varians (satt opp mot lav gjennomsnittlig skjevhet)
CL	Confidence level
CI	Confidence interval
PE	Point estimate
IE	Interval estimate
CV	Critical value
H0	Null-hypotese (simple hypothesis)
HA	Alternativ hypotese (composite hypothesis)
Hypo	Hypotese
TS	Test statistic/test observator
α eller LOS	Level of significance (=Pr(type 1 error))
β	Pr(type 2 error)
RR	Rejection region
BR	Basic result
μ^*	Potensielt/mulig populasjonsgjennomsnitt
TC	Test criteria
OCC	Operating characteristics curve
SOS	Sum of squares
p(X,Y)	Joint Pr distribution for combinations of values of the variables X and Y
f(X)	Marginal Pr distribution for values of variable X
g(Y)	Marginal Pr distribution for values of variable Y
Cov	Ko-varians mellom to variable i en populasjon
ρ (ikke p)	Korrelasjons-koeffisient for to variabler i en populasjon
v	Ko-varians mellom to variable i et utvalg
R	Korrelasjons-koeffisient for to variabler i et utvalg

1A)

Variable	Obs	Mean	Std. Dev.	Min	Max
days	146	16.4589	16.25322	0	81
aborigin	146	.4726027	.5009674	0	1
male	146	.4520548	.4994092	0	1
fast_learner	146	.5684932	.4969914	0	1
veryhigh_a~t	146	.2534247	.4364693	0	1

Data-settet inneholder fem variabler, som alle har et registrert resultat for alle de 146 individene i utvalget. «Days» viser til antall dager man har vært borte fra skolen i løpet av en periode (et skoleår?), og vi ser at både gjennomsnittet og standardavviket er omtrent 16 og en halv dag. Det betyr at mange har vært borte veldig få/ingen dager (siden standardavviket er nesten like høyt som gjennomsnittet), men også at mange har vært borte i langt over både 20 og 30 dager (minst en har vært borte i 81 dager, som vi ser av «Max»), siden gjennomsnittet ligger langt over 0.

De fire andre variablene er binomiale (binære), der «veryhigh_a-t» kategoriserer «days» til «ja»/1 eller «nei»/0, avhengig av om «days» er større eller mindre enn 22,5 (23 gir 1, 22 gir 0). Denne variabelen viser et gjennomsnitt på ca. ¼, som betyr at 25 % av de 146 elevene var borte i minst 23 dager. Standardavviket på binomiale variabler sier kanskje ikke så mye, ettersom verdiene alltid er enten 0 eller 1 uansett, slik at standardavviket blir nokså likt for alle binære variabler, men her er det i hvert fall 0,44. (Merk at standardavvik ≈ gjennomsnittlig avvik, de er ikke nødvendigvis helt like.)

«Male» viser rett og slett til kjønnet på hver person i utvalget, og 1 betyr «ja» (altså at det er en mann/gutt), mens «nei» betyr 0 (altså at det er en dame/jente, gitt at vi forutsetter at alle (i dette utvalget) er enten jente eller gutt). Gjennomsnittet på 0,45 kan nok tyde på en viss skjevhet i utvalget (at utvalget ikke representerer alle australske skoleelever perfekt, når kjønn er en faktor, grunnet «sampling variability» (som alltid vil finnes hvis ikke n nærmer seg P)), hvis vi antar et gjennomsnitt nærmere 0,5 i populasjonen. Men, det er jo mulig at dette ikke er snakk om obligatorisk skolegang, og da kan det godt hende at 0,45 stemmer bedre overens med populasjonen. Standardavviket er 0,5 (alle avviker enten 0,45 eller 0,55 fra gjennomsnittet).

«Aborigin» og «fast_learner» er de to siste variablene, også de binomiale. Dette er to «ja» - eller «nei»-verdier, der førstnevnte er relativt enkelt å definere (objektivt), mens sistnevnte må ha skjedd gjennom en form for vurdering av hvert enkelt individ i utvalget. Det kan virke som om «kravet» for å bli ansett som «fast_learner» er satt i forkant av testen, fordi det ellers kanskje ville vært naturlig å sette grensa ved median-eleven, slik at gjennomsnittet ble 0,5 (når gjennomsnittet av en binær variabel er 0,5, vet vi at også medianen må være det, ettersom det innebærer at halvparten av observasjonene er 1, og resten 0). Dette er dog ikke noe vi vet sikkert. Uansett er gjennomsnittene henholdsvis 0,47 og 0,57, og standardavvikene 0,5 for begge (fordi alle avviker enten til 0 eller til 1, slik at standardavviket alltid vil være i nærheten av 0,5, særlig når gjennomsnittet også er det). At 47 % er aboriginere, tyder på at testen er gjort i et område med relativt større andel urbefolkning enn i Australia i sin helhet.

1B)

Is a fast learner	Freq.	Percent	Cum.
Slow learner	23	34.85	34.85
Fast learner	43	65.15	100.00
Total	66	100.00	

43 av 66 gutter i utvalget, som vi av tabellen ser at tilsvarer 65 %, anses som «fast learners».

Is a fast learner	Freq.	Percent	Cum.
Slow learner	40	50.00	50.00
Fast learner	40	50.00	100.00
Total	80	100.00	

For jentene ser vi at 40 av 80, altså halvparten, anses som «fast learners».

Dette viser en relativt tydelig forskjell i utvalget, så mye tyder på at guttene også i befolkningen er «fast_learners» i større grad enn jentene (altså at større andel regnes som å være i den kategorien, hvordan kjønnene fordeler seg innad i de to nokså generelle kategoriene er ukjent ut ifra dette datasettet).

Dette forutsetter dog at de som har bestemt kriteriene for «fast_learners», og deretter vurdert utvalgsmedlemmene opp mot det, har gjort en god og objektiv jobb, og ettersom vi ikke har gjort noen statistisk test, må vi ta et forbehold om at det relativt lille utvalget kan være lite representativt for befolkningen, selv om det ble tilfeldig plukket.

1C)

Is a fast learner	Large n. of days absent		Total
	No very h	Very high	
Slow learner	47	16	63
Fast learner	62	21	83
Total	109	37	146

Begynner med marginal sannsynlighet, slik at neste steg blir enklere. Bruker f(x) og g(y), og får følgende resultater:

f(slow)=Pr(slow)=63/146=0,43

f(fast)=Pr(fast)=83/146=0,57 (kunne også brukt 1-Pr(slow))

g(not very high)=Pr(not very high)=109/146=0,75

g(very high)=Pr(very high)=37/146=0,25 (eller 1-0,75)

Kan dermed bruke disse sannsynlighetene for å regne ut såkalt «joint» (kombinert) sannsynlighet for de fire mulige kombinasjonene, og symboliserer da med p (for Pr):

p(slow, not very high)=0,43*0,75=0,32

p(slow, very high)=0,43*0,25=0,11

p(fast, not very high)=0,57*0,75=0,43

p(fast, very high)=0,57*0,25=0,14

Kunne også satt inn tall rett fra tabellen, og for eksempel fått p(fast, very high)=21/146=0,14.

1D)

”Because the covariance between fast learner and very high absent is zero, the two variables are independent and therefore days of absence from school have no effect on school performance. As a consequences, during the COVID pandemic countries around the world should not be concerned about closing down schools”

Sjekker først om påstanden som brukes som begrunnelse for argumentasjonen, stemmer.

	fast_l~r	veryhi~t
fast_learner	.247	
veryhigh_a~t	-.000236	.190505

Ser at ko-variansen er praktisk talt null (-0,0002) (verdien av ko-variansen gir i seg selv ikke noen informasjon om hvor sterk lineær relasjon de to variablene har, bare om den er positiv eller negativ. Styrken måles egentlig ved å bruke korrelasjons-koeffisienten, men velger å forutsette at et såpass marginalt tall også ville gitt en koeffisient tilnærmet lik 0). Dermed er det riktig at det ikke er noen lineær sammenheng mellom variablene.

Studenten tar imidlertid ikke forbehold om at det kan være en ikke-lineær relasjon, som vil være ensbetydende med en sammenheng mellom variablene. Intuitivt vil det kanskje også kunne anses som noe spesielt å påstå at fravær ikke har noen

betydning for skoleresultater, da det ville implisert at skolegang ikke har noen netto positiv læringseffekt i det hele tatt, sett på gruppenivå.

Vi kan sjekke påstanden om «independence» ved hjelp av følgende regel: X og Y er uavhengige dersom p(X,Y)=f(X)*g(Y) for alle verdier av X og Y.

Hvis vi setter «fast»=X og «very high»=Y, kan vi sette inn «sant»/1 og «usant»/0 og sjekke om likninga holder:

p(0,0)=f(0)*g(0)

0,32=0,43*0,75

0,32=0,32

p(1,0)=f(1)*g(0)

0,43=0,57*0,75

0,43=0,43

p(0,1)=f(0)*g(1)

0,11=0,43*0,25

0,11=0,11

p(1,1)=f(1)*g(1)

0,14=0,57*0,25

0,14=0,14

Ser at likninga holder for alle fire kombinasjonene, nettopp fordi vi antok at disse hendelsene er «uavhengige», slik at vi kunne bruke f(X) og g(Y) for å finne p(X,Y) i første omgang. Kan derfor se at studenten tilsynelatende har rett i at antall dager

borte fra skolen (eller i hvert fall det faktum om man har vært borte i mer enn 22,5 dager eller ikke) ikke påvirker om man er «fast_learner» eller ikke, skjønt begrunnelsen ikke er gyldig. Alternativt kan man kanskje anse «fast_learner» som i hvert fall delvis medfødt, sånn at det vi egentlig ser er at like stor andel av de som lærer raskt har høyt fravær som blant de som ikke anses å gjøre det.

Dette kan f.eks. implisere at alt fravær er medisinsk begrunnet, for om det var det, ville man nok antatt at hvor raskt man lærer ikke har (nevneverdig) påvirkning på risikoen for å bli syk, slik at begge gruppene bør ha like høyt fravær. Det kan også bety at grad av «skulk» er lik i de to gruppene i utvalget, men kanskje skulker de i så fall av ulike grunner?

At rask læring og fravær er uavhengige, er imidlertid langt ifra det samme som at tilstedeværelse ikke er positivt for skoleresultater, som er noe helt annet (gitt at «fast_learners» måler nettopp om man lærer raskt, og ikke skoleresultater i seg selv).

Konklusjonen i påstanden ville helt klart vært logisk hvis den var godt begrunnet/riktig. Det kan vi heller ikke utelukke basert på informasjonen i påstanden, men heller ikke bekrefte, og hvis man skal anta noe, vil det nok være mer naturlig å anta en sammenheng, enn å ikke gjøre det.

1E)

Vi skal finne ut hvor stor sikkerhet utvalget gir oss i å anta gjennomsnittlig fravær i befolkningen, ved å bruke 99 % CL. Vi har at $n=146>>30$, slik at vi kan benytte CLT, som sikrer at «et hypotetisk stort antall utvalg» ville normalfordelt gjennomsnittene sine, uavhengig av befolkningens fordeling (som vi kanskje kan anta at er «skewed to the right»). Vi får følgende CI:

SØK1004, kandidat 10029, 9.12.20

$\bar{X} \sim N(\mu, \sigma^2/n)$

$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) = Z$ score

CL = 0,99 innebærer

$Pr(-X < Z < X) = 0,99$

$Pr(-2,575 < Z < 2,575) = 0,99$

$X = 2,575 = CV$

$Pr\left(-2,575 < \frac{16,46-\mu}{16,25/\sqrt{146}} < 2,575\right) = 0,99$

$Pr\left(16,46 - 2,575\frac{16,25}{12,08} < \mu < 16,46 + 2,575\frac{16,25}{12,08}\right)$

$Pr(13 < \mu < 19,92) = 0,99$

Vi ser at vi med 99 % sikkerhet kan fastslå at en gjennomsnittselev i populasjonen («of interest») er borte mellom ca. 13 og 20 dager, basert på at gjennomsnittet i utvalget var 16,46 dager. Dette ved hjelp av at vi vet at utvalgsvariansen er en «unbiased» estimator for populasjonsvariansen (ettersom Stata dividerer på n-1), og derfor benytter oss av den.

1F)

Vi skal finne et 90% CI for andelen studenter i befolkningen som er borte i minst 23 dager. Vi kan approksimere til («hypotetsisk») normalfordeling dersom vi forventer i gjennomsnitt minst fem «suksesser» og minst fem «fails» fra ethvert gitt hypotetisk utvalg. Vi bekrefter at dette er oppfylt, og tester så på følgende måte:

SØK1004, kandidat 10029, 9.12.20

$n \cdot p = 146 \cdot 0{,}25 = 36{,}5 > 5$,

$n \cdot (1-p) = 146 \cdot 0{,}75 = 109{,}5 > 5$

$p \sim N(\mu, \sigma^2)$

$E(\mu) = p, \quad \mu = \pi$

$\sigma^2 = E(X^2) - \mu^2$

$E(X^2) = 1^2 \cdot \pi + 0^2 \cdot (1-\pi) = \pi$

$\mu^2 = \pi^2$

$\sigma^2 = \pi - \pi^2 = \pi(1-\pi)$

$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

SE for $PE = p$ blir

$\sqrt{\frac{\pi(1-\pi)}{n}}$, slik at,

når π estimeres av sin effektive estimator p, ender vi med at 90% CI gir $\pi =$

$p \pm 1{,}645 \cdot \sqrt{\frac{0{,}25(0{,}75)}{146}} = p \pm 0{,}059$,

$Pr(0{,}191 < \pi < 0{,}309) = 0{,}9$

CV-en 1,645 er funnet ved å se på Z-score=0,45 (som *2 gir CL) som befinner seg midt mellom Z-value 1,64 og 1,65.

1G)

Det ble gjort et forsøk på å øke andelen «fast learners» fra et antatt utgangspunkt på 1/5. Man kan selvsagt tenke seg til mange andre grunner en et «program» som ville kunne påvirke andelen i positiv eller negativ retning, men det mest naturlige uten tilleggsinformasjon er nok å anta at andelen ville holdt seg på 20 % om det ikke var for programmet.

1H)

SØK1004, Kandidat 10029, 9.12.20

$$p(83) = \underbrace{\frac{146!}{(146-83)!\,83!}}_{\text{Antall utfall med "riktig" resultat}} \underbrace{0{,}2^{83} \cdot 0{,}8^{63}}_{\text{Pr. par hver av dem}} \approx 0$$

Formelen viser at sannsynligheten for at 83 av 146 utvalgsdeltakere er «fast learners» hvis populasjonsandelen er 0,2, praktisk talt er 0.

I utvalget er det nettopp dette som er resultatet. Det innebærer at vi kan være så sikre vi kan bli på at utvalget for det første har blitt testet etter (eller underveis i) programmet, og at programmet har hatt en (trolig ganske stor) positiv effekt.

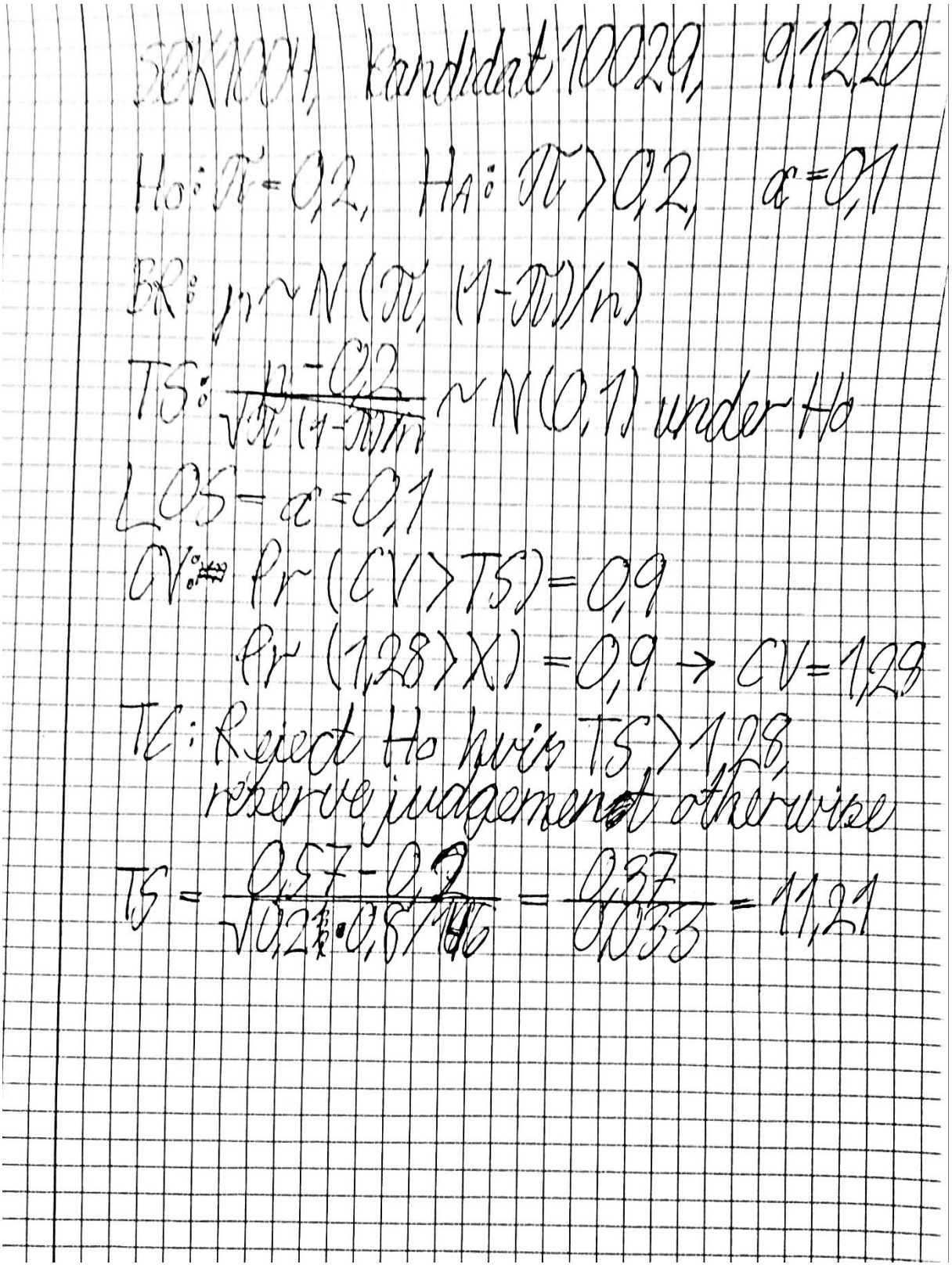
1I)

Forutsetningene for at betraktningene/resultatet fra 1H) skal være gyldig(e), er i hovedsak de samme som gjelder generelt, som for eksempel at utvalget er tilfeldig valgt og relativt stort, og at all testing blir gjort på en god og objektiv måte.

Det vi kan legge til i dette tilfellet er at andelen på 0,2 som ble antatt på forhånd er korrekt, og at utvalget potensielt kan ha blitt (positivt) påvirket nettopp av å være med i utvalget, slik at dette ga lærere og/eller elever ekstra motivasjon som man ikke ville hatt i en normalsituasjon der resultatene ikke nødvendigvis ble registrert på samme måte.

1J)

Vi vet at vi kan forvente over fem 1-ere og over fem 0-ere i ethvert gitt utvalg, og tester derfor nullhypotesen for å se om den kan forkastes:



Ettersom 11,21>>1,28, kan vi trygt forkaste H0, og dermed si tilnærmet sikkert at programmet har hatt en positiv effekt.

1K)

Hvor stort utvalg trengs for å være 95 % sikker på at andelen «fast learners» er innenfor et 3 prosentpoeng stort intervall basert på utvalgsandelen?

Utvalgsandelen er 57 %, så en normalfordeling vil tilsi at intervallet da må bli +/-1,5 %-poeng, som gir følgende forutsetning: $Pr(p-0,015<\pi<p+0,015)=0,95$.

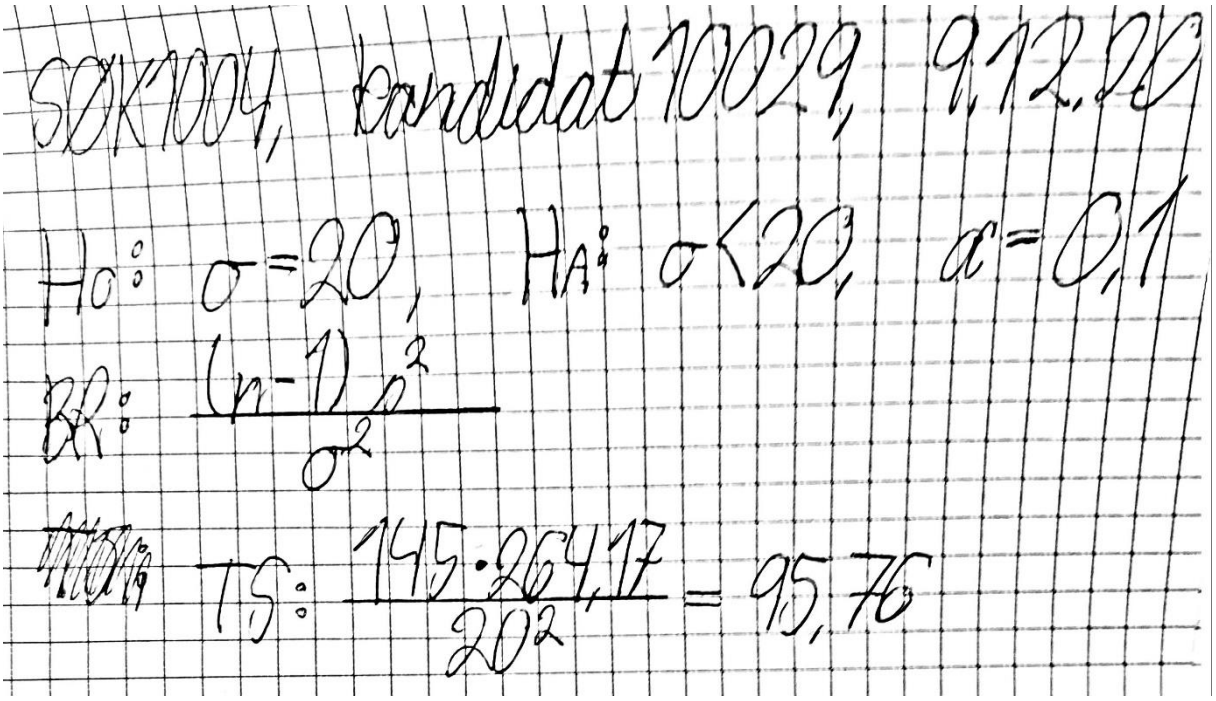
Dette gir en Z-verdi på 1,96, fordi vi da får 0,025=2,5% av hypotetiske utvalg utenfor betingelsen.

Dette innebærer at kvadratroten av n er lik 0,57*1,96/0,015=74,48, som er 5547,27 kvadrert.

Dette betyr at vi med tilnærmet 95 % sikkerhet kunne anslått populasjonsandelen «fast learners» til å være mellom 0,555 og 0,855 hvis et utvalg på n=5547 hadde gitt oss samme utvalgsandel som det vi faktisk har, og at vi ville passert 95%-grensen ved n=5548.

1L)

Kommandoen gitt i oppgaveteksten ga feilmelding, og tabellene viser ikke den etterspurte CV-en. Beregner derfor TS først, og deretter CV v.h.a. en kalkulator jeg fant på internett (https://www.socscistatistics.com/pvalues/chidistribution.aspx):



Fikk at «P-verdien er 0,999451», hvilket jeg antar at betyr at vi ikke kan utelukke at H0 fortsatt gjelder.

Det vil si at reduksjonen fra 20 til 16,46 anses som «statistisk ikke signifikant», som innebærer at det potensielt kan være et resultat av «sampling variability». Vi vet imidlertid ikke, og kan ikke bekrefte H0, så konklusjonen blir da «reserve judgement on H0» som igjen innebærer at vi også reserverer oss fra å konkludere om HA.

2A-2D)

SØK1004, kandidat 10029, 9.12.20

a) $Pr(-1<X<2)$

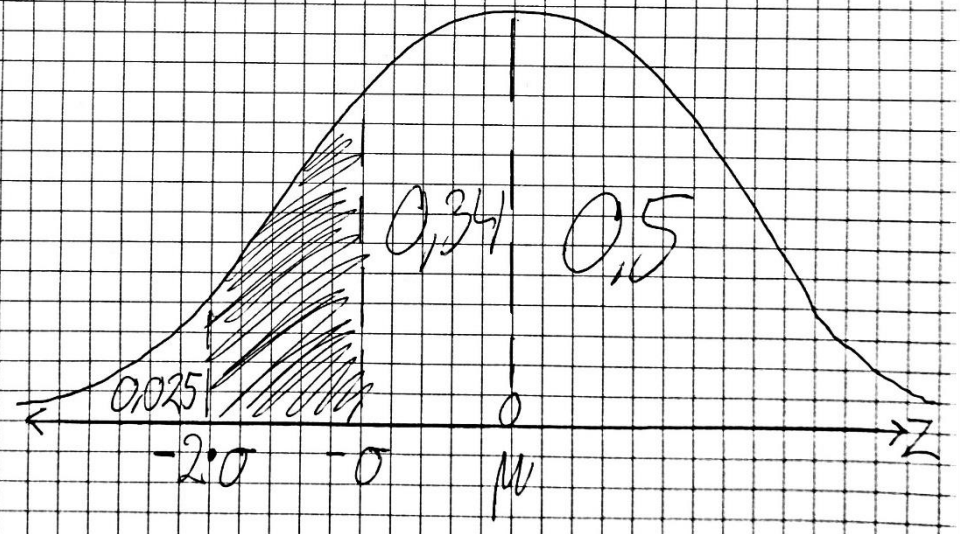
$X \sim N(5,9)$

$Z \sim N(0,1)$

$Z = (X-5)/3$

$Pr\left(\frac{-1-5}{3} < \frac{X-5}{3} < \frac{2-5}{3}\right) =$

$Pr(-2<Z<-1) =$

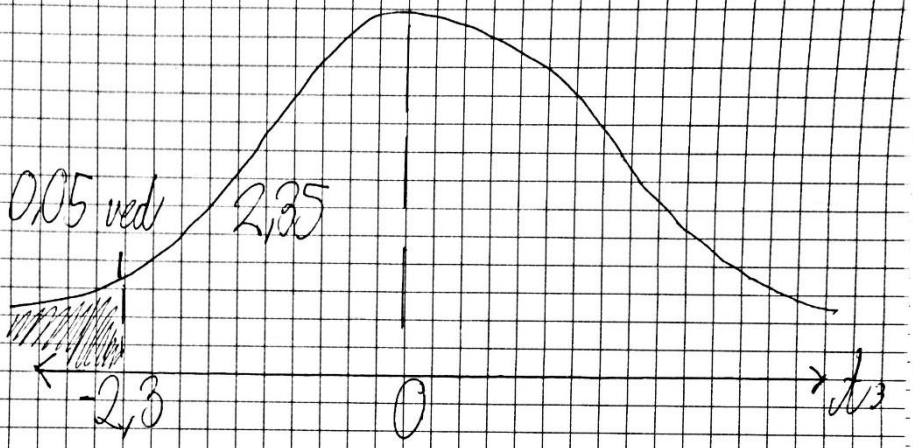


$1-0,025-0,34-0,5=0,135$

SØK1004, kandidat 10029, 9.12.20

b $Pr(X < -2,3) =$ like over 5%

$X \sim t_3$



c $Pr(0,67 < X < 3,32)$

$X \sim \chi^2_9$

0,67 ligger under $Pr = 0,005$ gitt for $(X < 1,735)$, så benytter $Pr(X < 3,32)$, og finner 3,325 ved 0,95, altså er $Pr(X > 3,325) = 0,95$, kan derfor skrive

$Pr(0,67 < X < 3,32) = 0 + (1 - 0,95) = 0,05$

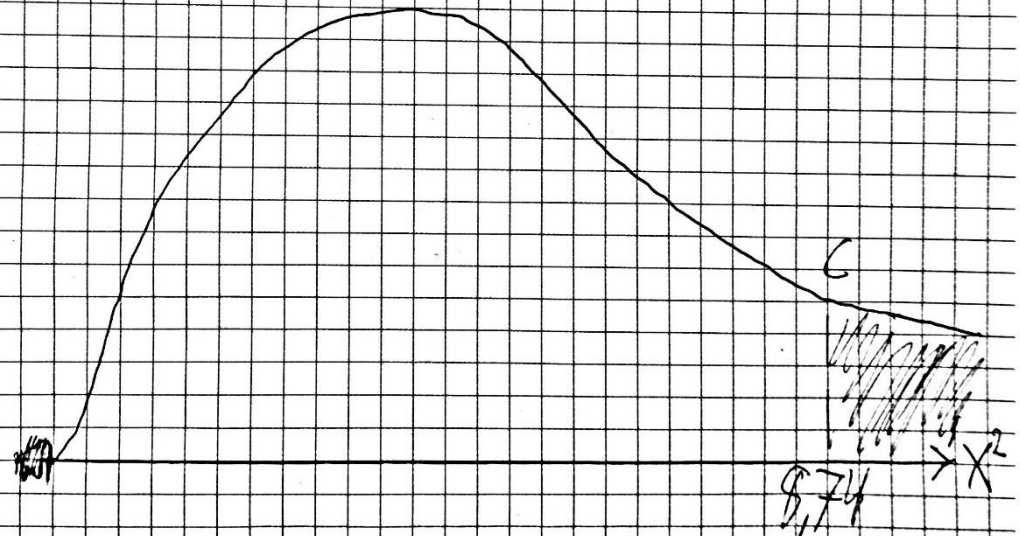
SØK1004, kandidat 10029, 9.12.20

d. $Pr(X > c) = 0,05$

$X \sim F(12,3)$

Slik at $Pr(X > 8,7446) = 0,05$

$c = 8,7446$



3A)

En «unbiased» (forventningsrett) estimator er en utvalgsverdi som ikke har en systematisk tendens til å avvike fra populasjonsverdien den er ment å estimere. Dette betyr at den i et «gjennomsnitts-utvalg» vil representere populasjonsverdien perfekt. Vi kan skrive dette formelt som at $E(Q)=\theta$, som betyr at estimatoren Q i gjennomsnitt vil være lik populasjonsverdien den skal estimere, her representert ved det generelle symbolet θ, som f.eks. kan stå for populasjonsvarians.

Sampling variability gjør at estimatet og realiteten likevel sjelden-aldri vil være like, men det er ingen systematiske avvik; «downward bias» oppveies nøyaktig av «upward bias», sett over en hypotetisk situasjon med mange utvalg.

3B)

I virkeligheten kan man sjelden gjennomføre mange utvalg (i så fall er det etter hvert ikke lenger snakk om «statistical inference», men «descriptive statistics» ved at man har en folketelling/«census»).

Dette kan man imidlertid simulere i f.eks. et dataprogram som Stata. Ved å sjekke gjennomsnittet av gjennomsnittene fra veldig mange utvalg, kan vi se at det (gitt at estimatoren faktisk er forventningsrett) konvergerer mot et mer eller mindre selvvalgt populasjonsgjennomsnitt etter hvert som antall utvalg øker, og når vi når mange utvalg, er de praktisk talt like; vi kan se at det gjennomsnittlige gjennomsnittet korresponderer med det faktisk populasjonsgjennomsnittet, og vet derfor at vi for ethvert utvalg der vi bruker den gitte estimatoren, vil kunne forvente som et gjennomsnitt at utvalget viser til populasjonen helt representativt.

I en eventuell slik empirisk test, vil vi se at «nærheten» mellom estimat og realitet er en funksjon av både antall utvalg, men også antall pr. utvalg, slik at større utvalg gir lavere «variability», eller feilmargin om du vil.

Do-file:

```
cd "C:\Users\norma\OneDrive\Skrivebord\Skole\SØK1004\Eksamen"
use "C:\Users\norma\OneDrive\Skrivebord\Skole\SØK1004\Eksamen\eksamen.dta"
browse
summarize
tabulate fast_learner if male ==1
tabulate fast_learner if male ==0
tabulate fast_learner veryhigh_absent
correlate fast_learner veryhigh_absent, covariance
summarize days
summarize veryhigh_absent
tabulate fast_learner
summarize days, detail
```