DOI: 10.18261/9788215028163-2017-08

Kapittel 8
Ein statistikkdiskusjon med utgangspunkt i normavvik i studenttekstar

TERJE MYKLEBUST OG JAN OLAV FRETLAND

SAMANDRAG Vi diskuterer nokre grunnleggande tilnærmingar til eit datasett om normavvik («rettskrivingsfeil») i studenttekstar. Vi innfører modellar for å analysere tekstane, og vi ser på kva for type spørsmål datasettet kan svare på og kva det ikkje kan svare på. Då er det viktig å klargjere føresetnadane. Vi diskuterer òg forskjellen mellom å analysere eit datasett med tanke på å «bevise» samanhengar statistisk, og det å bruke datasettet for å fremje hypotesar. Statistiske metodar kan brukast i begge tilfella. Analysane her kan sjåast på som ei vurdering av kvaliteten i datasettet for ytterlegare analysar.

NØKKELORD utforskande statistikk, normavvik i nynorsktekstar, modellar for teljedata

ABSTRACT We have a dataset about misspellings in Norwegian texts written by students. We introduce statistical models to analyze the dataset, and we discuss what questions can possibly be answered by the dataset. Moreover, we clearly point out the assumptions necessary for drawing conclusions. We also discuss different approaches to a dataset, like exploratory and confirmatory approaches. The analysis in this paper can be viewed as an evaluation of the quality of the data for further analysis.

INNLEIING

Kunnskap og metodar for kunnskapsutvikling er ein del av vår immaterielle kapital. Statistiske metodar og statistiske vurderingar høyrer til blant slike metodar etter som dei er viktige for å erverve ny kunnskap frå data. Denne artikkelen tar utgangspunkt i eit datasett med talet på rettskrivings- og bøyingsavvik i tekstar på

nynorsk skrivne av studentar frå fire ulike studium ved Høgskulen i Sogn og Fjordane. Registreringsarbeidet er gjort av Jan Olav Fretland og er grunnlaget for Fretland (2007, 2015). Formålet med artikkelen er likevel ikkje å finne ny kunnskap om avvik i nynorsktekstar. Vi går derimot inn på nokre grunnleggande spørsmål som vi møter i mange datasett, og vi diskuterer ulike tilnærmingar til eit datasett generelt og til dette settet spesielt. Datasettet er eit godt utgangspunkt for å diskutere ein ikkje heilt uvanleg situasjon, nemleg ein situasjon der vi står overfor data utan at vi har vel formulerte spørsmål som vi ønskjer å svare på.

Eit av hovudpoenga i artikkelen er å diskutere korleis ein kan modellere teljedata som her er førekomstane av normfeil i nynorsktekstar. Dette er vurderingar som statistikkpakkane ikkje gjer for oss, og kanskje nettopp derfor vert dette viktige grunnlagsabeidet ofte oversett. Ved hjelp av dei statistiske modellane ser vi på kva spørsmål datasettet kan svare på, og kanskje spesielt kva spørsmål det ikkje kan svare på. Vi peikar blant anna på nokre typiske slutningar som datasettet inviterer til, men som det ikkje nødvendigvis er grunnlag for. I alle fall ønskjer vi å synleggjere nødvendige føresetnadar for eventuelle konklusjonar. Indirekte er vi inne på forsøksdesign, det vil seie korleis data må samlast inn for å svare på bestemte spørsmål.

Vi forsøkjer også å vurdere kor mykje informasjon det er i datasettet. Som forskarar må vi vere opne for at ikkje alle datasett inneheld så mykje interessant informasjon som ein skulle ønskje. Det viktigaste er trass alt at vi ikkje forsøkjer å trekke meir ut av eit datasett enn det er grunnlag for. Blant anna kan det vise seg at mykje av det vi ser i datasettet, viser seg å vere innanfor heilt naturlege variasjonar. Eit sentralt spørsmålet i denne samanhengen, er om det er forskjell på feilmengda i tekstane. Og sidan vi har ulike grupper, bør vi i første omgang undersøkje kvar gruppe for seg. Oppfølgingsspørsmålet, som også er inspirert av Fretland (2015), er om det er forskjell mellom gruppene. Her peikar vi også på spørsmål som ein ikkje bør stille i eit datasett.

Vi føreset at lesaren har ein viss kjennskap til nokre grunnleggande sentrale omgrep i statistikk som til dømes forventning, varians, signifikans og p-verdi.

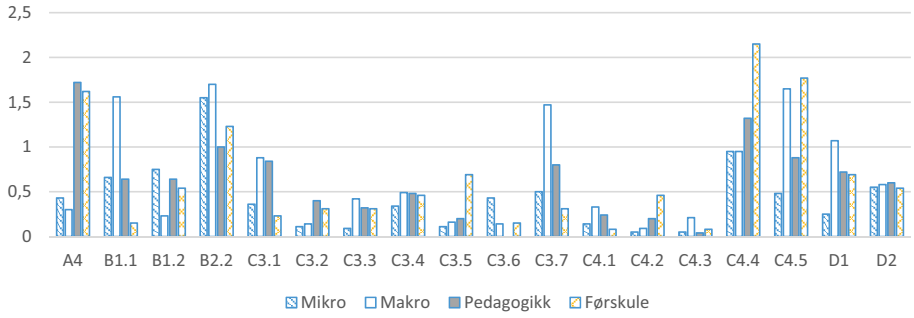
OM DATASETTET

Datasettet består av talet på normavvik i studenttekstar. Feiltypane er delt inn i atten ulike kategoriar. Det er høvesvis 44, 43, 25 og 13 tekstar skrivne av studentar i makroøkonomi, mikroøkonomi, pedagogikk og førskulelærarutdanninga ved Høgskulen i Sogn og Fjordane. Det er éin tekst per person. Registreringa er bygd på dei om lag fire første sidene med normal handskrift i kvar tekst, slik at vi går

ut frå at tekstane er omtrent like lange. Feilkategoriane er merkte som i figur 8.1. Grovt sett står bokstaven A for allmenn ordlegging, B for bøyingsfeil, C for feil med konsonantisme og D for morfemfeil. Desse er så delt inn i fleire underkategoriar:

A4 Enkelt bestemt substantiv: Eg viser til ditt brev for …brevet ditt
B1.1 Feil med hankjønn/hokjønn fleirtal: priser for prisar, skatter/skattar
B1.2 Inkjekjønn fleirtal: møter for møte
B3 Presens/preteritum av verb: auker for aukar, styrka for styrkte
C3 Feil i vokalisme:
C3.1 a/e: døme uten for utan, C3.2 a/o: holda/halda, C3.3 e/i: virke/verke, men og posetivt
C3.4 e/æ: væra for vera, C3.5 o/å: åpen/open, C3.6 o/u/ø: bud for bod osv.
C3.7 e/ei, ø/au: øke for auke, lede for leie osv.
C4 Feil i konsonantisme:
C4.1 g/k/0: lege for lækje, valg/val, C4.2 m/v: jamn/jevn, C4.3 d/t: bedre for betre, sedlar
C4.4 enkel/dobbel konsonant: greit, tall
C4.5 +/– konsonant (seldt for selt, brendt/brent, vigtig/viktig)
D1 Feil kjønn på verbalsubstantiv på -heit og -ing: innstillingen for innstillinga
D2 Andre feil med kjønn på substantiv: ein form for ei form

Her ser vi først og fremst på datasettet i seg sjølv, og vi refererer derfor berre til Fretland, (2015, s. 180–81) for ei nærare forklaring på desse kategoriane.



Figur 8.1: Stolpane viser gjennomsnittleg tal feil per tekst i dei ulike kategoriane for kvar av dei fire gruppene. Den gjennomsnittlege totale feilmengda per tekst er 7,80, 12,37, 11,04 og 11,77 i høvesvis mikro, makro, pedagogikk og førskule.

TRE ULIKE TILNÆRMINGAR TIL EIT DATASETT

Det er grovt sett tre ulike måtar å nærme seg eit datasett på. Nokon gonger er vi interesserte i datasettet i seg sjølv. Vi ser på det som eit komplett bilde av det vi ønskjer å studere, og vi er ikkje interessert i å trekke slutningar som gjeld utover det konkrete datasettet. Om vi til dømes har statistikk over nedbørsmengda på ein bestemt stad i Noreg det siste året, kan det vere interessant å beskrive nedbørsmengda slik ho faktisk var gjennom dette året. Det kan gjerast med tabellar, diagram og ymse mål som gjennomsnitt og standardavvik, men omgrep som signifikans gir inga meining. Vi er då ikkje interesserte i å seie noko om nedbørsmengda utover det vi har observert. Denne tilnærminga kjem vi ikkje nærare inn på i denne artikkelen.

I vitskaplege tilnærmingar ønskjer vi typisk å trekke generelle slutningar som gjeld utover det konkrete datasettet. Då er vi eigentleg ikkje interesserte i datasettet i seg sjølv. Nedbørsstatistikken siste år er då berre interessant i den grad han kan fortelje noko om korleis vi til dømes kan forvente at nedbørsmengde på denne staden er i framtida, eventuelt korleis ho har vore i tidlegare år. I slike situasjonar er vi typisk interesserte i å estimere ymse storleikar og å teste hypotesar, og såkalla signifikante slutningar står sentralt. Det stiller strenge krav både til datainnsamling og analysemetodar.

Den tredje tilnærminga vert ofte omtala som ei utforskande tilnærming. Då er vi eigentleg interesserte i å finne ut noko om noko ved hjelp av data. Vi ønskjer å finne ut noko som gjeld utover det konkrete datasettet, men vi veit ikkje heilt kva vi ser etter. Derimot håpar vi at datasettet skal avsløre eit eller anna om verkelegheita som vi kanskje ikkje hadde tenkt på i utgangspunktet. Også i slike situasjonar kan vi bruke statistiske metodar, men vi kan då ikkje hevde å ha signifikante bevis sjølv om vi får låge p-verdiar.

KVA MÅ TIL FOR Å TREKKE SIGNIFIKANTE SLUTNINGAR?

Den andre tilnærmingsmåten til eit datasett, som vi nemnde i førre avsnitt, dreiar seg om å trekke generelle slutningar som gjeld utover datasettet sjølv. Då står såkalla signifikante slutningar sentralt. Når vi skal trekke signifikante slutningar frå data, må vi ha klart for oss kva spørsmål datasettet skal svare på. Dette kan kanskje synast opplagt, men det er likevel ikkje heilt uvanleg at data vert samla inn utan ei klar formeining om kva spørsmål dei skal svare på. Likevel vert resultata gjerne presenterte som signifikante. I vår konkrete situasjon kan ein tenkje at datasettet skal fortelje noko om førekomsten av ortografiske feil. Men i utgangspunktet er det ei altfor uklar problemstilling om ein vil trekke såkalla signifikante slutningar. Kanskje er ein interessert i om det er samanhengar mellom feiltypane. Er det til

dømes slik at dei som gjer mange feil i ein kategori, også gjer mange andre feil? Men heller ikkje dette er eit særleg presist utgangspunkt for å trekkje signifikante slutningar. Derimot kan vi spørje om det er samanheng mellom talet på feil til dømes i den første kategorien og talet på feil i den andre kategorien. Eller vi kan spørje om alle feiltypar er like hyppige. Om dette er interessante spørsmål, er sjølvsagt eit anna spørsmål. Poenget er at hypotesen må vere formulert svært presist om vi skal kunne trekkje signifikante slutningar. Hypotesen bør faktisk vere formulert før ein startar datainnsamlinga. Det er viktig at forsøket og datainnsamlinga vert designa med tanke på å svare på den eller dei heilt konkrete hypotesane vi har.

I undersøkingar av denne typen er vi typisk interesserte i å finne ut noko allmenngyldig. Vi vil finne ut noko som ikkje berre gjeld for dette konkrete datasettet, men som er gyldig for ein større populasjon. Poenget med forsøket kunne til dømes vere å finne ut noko om frekvensane til desse feiltypane, eventuelt å finne ut noko om samanhengane mellom dei. Og vi håper rimelegvis at det vi ser i dette datasettet, seier oss noko om feilmønster som er gyldige utover desse konkrete tekstane og desse konkrete personane.

Gyldigheita av generaliseringar utover vårt konkrete datasett er blant anna avhengig av kor representative personane er for den populasjonen vi ønskjer å seie noko om. I tillegg er det eit spørsmål om kor representative tekstane er. Om vi kan trekke generelle slutningar utover dei konkrete tekstane, er i tillegg avhengig av kor mange personar og tekstar vi har, i tillegg er lengda på tekstane viktig. Kor vidt utvalet er for lite eller ikkje, kan statistiske metodar gi oss informasjon om. Det vil seie at dei statistiske metodane kan fortelje oss noko om kor sikre konklusjonane våre er med det utvalet vi har, gitt at utvalet er eit tilfeldig (representativt) utval av populasjonen. Statistiske metodar kan også hjelpe oss med å plukke ut representative utval.

Når det gjeld representativitet, er det langt i frå alltid at vi har høve til å plukke ut observasjonane slik vi skulle ønskje. Vi må ofte ta det datasettet som er tilgjengeleg. Nokon gonger har vi til dømes berre ein skuleklasse til rådigheit. I slike tilfelle er kanskje ikkje resultata representative for heile landet. Men om forsøket er gjort ordentleg, er resultata kanskje representative for klassen og kanskje litt utover klassen, og i mange tilfelle kan slike resultat tene som gode hypotesar for stoda i landet generelt. Desse hypotesane kan så testast ved eit seinare høve.

Dei nemnde problemstillingane er reelle nok, men nokon gonger kan vi oppleve at dei vert blanda saman med uklare og direkte feiltolkingar av datasettet. Det hender at det vert trekt konklusjonar frå datasettet som ikkje ein gong er gyldige for den populasjonen datasettet er representativt for. Ofte ser ein at artiklar vert avslutta med ein setning om at ein ikkje kan garantere at gruppa er representativ, og at det er for få data til å konkludere noko sikkert. Slike kommentarar kan ofte

skyggje for andre alvorlege feiltolkingar og uheldige datainnsamlingar. Det hender nemleg at det vert trekt konklusjonar som ikkje ein gong er gyldige for det aktuelle utvalet. Dessutan kan det faktisk vere lett for å omtale utvalet som noko anna enn det det faktisk er. I vårt datasett er det til dømes lett for å sette likskapsteikn mellom tekstane og tekstforfattarane, men vi er ikkje ein gong garanterte at den konkrete teksten er representativ for forfattaren.

I dette arbeidet tar vi som utgangspunkt at personane og tekstane er representative for den gruppa og dei tekstane dei måtte vere representative for, noko anna er trass alt nesten meiningslaust. Vi vil med andre ord ikkje diskutere representativiteten i datasettet. Vi må likevel passe på at vi har grunnlag for dei konklusjonane vi trekkjer om desse tekstane og denne gruppa.

Å UTFORSKE EIT DATASETT?

Det er ikkje alltid at ein har konkrete hypotesar som ein ønskjer å teste. Nokon gonger vil vi berre forsøkje å danne oss eit bilde av verda ved å undersøkje data. Vi er no altså komen til den tredje tilnærmingsmåten som er nemnt i avsnittet «Tre ulike tilnærmingar til eit datasett». Spesielt på nye område der ein har lite eller ingen kunnskap, kan det vere naturleg å forsøke å danne seg eit bilde av stoda ved å utforske eit datasett. Det er sjølvsagt legitimt, men då er vi i ein heilt annan situasjon enn når vi skal teste hypotesar. I ei utforskande tilnærming er det ikkje snakk om å påvise noko, men det er snakk om å sjå etter moglege samanhengar. Utkomet av analysen er då i beste fall ein eller fleire hypotesar om det aktuelle temaet. Desse hypotesane må så testast på eit seinare tidspunkt, om dei skal få status som noko anna enn hypotesar. I slike situasjonar seier vi gjerne at vi utforskar datasettet. Statistiske metodar kan brukast også i slike utforskingar, men konklusjonane har altså ein annan status. Sjølv om vi observerer små p-verdiar, kan vi likevel ikkje hevde at vi har signifikante resultat. I ei utforskande tilnærming er små p-verdiar berre ein indikasjon på at vi kanskje har ein god hypotese. Sjå også (Tukey, 1977; Tukey, 1980).

Det ligg ikkje føre teoriar med hypotesar som vi ønskjer å teste. Det er med andre ord ei utforskande tilnærming vi har grunnlag for å gjere i dette datasettet.

EIN MODELL FOR TALET PÅ NORMFEIL I EIN TEKST

Datasettet har talet på feil i 18 ulike feilkategoriar, nummererte frå 1 til 18. La k stå for kategorinummeret, det vil seie at k er eit heil tal blant tala 1, … , 18. Talet på feil i kategori nummer k kallar vi for Y_k.

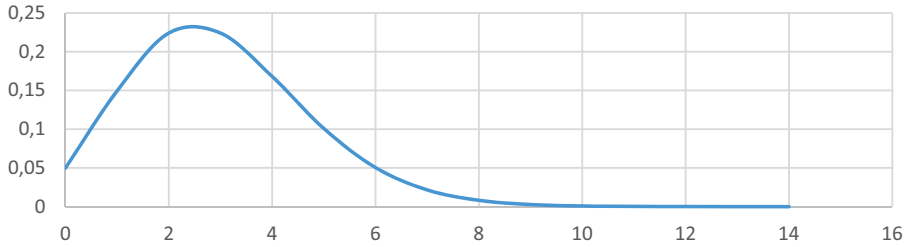
Vi kan tenkje på skriving som ei følgje med ord. For kvart nytt ord som vert skrive, er det ein viss sjanse for at det er feil av kategori k. Ein slik prosess vert ofte modellert med ein såkalla Poisson-prosess. I denne prosessen tel vi førekomstar i kategori k. Vi lar Y_k stå for antal førekomstar i kategori k, og vi seier då at Y_k er Poisson-fordelt. Dei ulike feila opptrer med ulik intensitet. Svært forenkla er intensiteten talet på førekomstar i kategorien, men på grunn av ymse slumpetreff varierer dette talet frå tekst til tekst så vel som i ulike deler av ein tekst. Vi kan sjå på intensiteten som gjennomsnittleg tal feil når vi har mange slike tekstar. Han seier altså ikkje nøyaktig kor mange feil den enkelte har gjort. Meir formelt er intensiteten ein teoretisk verdi (parameter) som bestemmer Poisson-fordelinga, jamfør figur 8.2.

Intensiteten til feilkategori k skriv vi som λ_k (den greske bokstaven lambda). Dess større intensiteten λ_k er, dess større sjanse er det for å få ein feil i kategorien k. I ei Poisson-fordeling er det slik at forventninga til Y_k er lik intensiteten, det vil seie at

$$E(Y_k) = \lambda_k,$$

og variansen er også lik intensiteten

$$Var(Y_k) = \lambda_k.$$



Figur 8.2: Poisson-fordeling med intensitet 3. Kurva antyder at det er rimeleg stor sjanse for å finne alt frå null til seks feil i ein kategori med feilintensitet 3.

Når vi kjenner intensiteten, kjenner vi også Poisson-fordelinga. Men i vår situasjon kjenner vi ikkje intensiteten, vi veit med andre ord ikkje kor mange feil vi kan forvente å finne i kategori k. Spørsmålet er om vi kan estimere intensiteten ved hjelp av datasettet vårt. Faktisk er det slik at antal førekomstar som vi finn i kategorien, er det beste estimatet vi kan finne for intensiteten λ_k. Merk at vi her brukar intensiteten som intensitet per fire sider tekst. Det fungerer her, sidan alle tekstane er fire sider lange. Dersom tekstane hadde hatt ulik lengde, ville det vere naturleg å snakka om intensitet per ord, eventuelt per side.

KAN VI TREKKJE SLUTNINGAR OM FEILSJANSANE TIL INDIVIDA UT FRÅ DETTE DATASETTET?

Datasettet gir oss talet på førekomstar i dei ulike kategoriane. Men at to tekstar har ulike feilintensitetar i ein kategori, skuldast ikkje nødvendigvis at forfattarane er ulike (med omsyn til feilsjanse), det kan like gjerne skuldast at tekstane er ulike. For at vi skal få feil i ein kategori som har å gjere med bøying av substantiv, må vi nødvendigvis ha å gjere med eit substantiv, i tillegg må forfattaren bøye substantivet feil, først då har vi ein førekomst i denne kategorien.

Vi må altså skilje mellom at vi finn ein feil i ein kategori, og det at feilkategorien faktisk kan opptre i den aktuelle sekvensen av teksten. Om vi litt forenkla ser på kvart ord i teksten, så kan ikkje alle ord ha feil i alle mulege kategoriar, i det minste har dei truleg ulik tilbøyelegheit for å ha feil i dei ulike kategoriane. Om vi ser på eit ord i teksten, så kan vi litt forenkla tenkje oss tre mulegheiter for kategori *k*: 1) Ordet er feil av kategori *k,* 2) ordet er rett, men det kunne potensielt hatt ein feil av kategori *k,* og 3) ordet kan ikkje ha denne typen feil. På tilsvarande måte som vi lar Y_k telje talet på feil i kategori *k*, lar vi X_k telje talet på mulege feil i kategori *k*. Då kan vi modellere X_k med ei Poisson-fordeling med intensitet μ_k. Det vil seie at

$$E(X_k) = \mu_k \text{ og } Var(X_k) = \mu_k.$$

Her står E for expectation, altså forventning, og Var står for varians. Vi kan tenke at kvar gong det kjem eit ord som potensielt kan hamne i kategori k, så er det ein viss sjanse for at det faktisk har ein slik type feil. Denne sjansen kan vi kalle feilsannsynet, som vi skriv p_k. Vi lar, som nemnt, Y_k vere antal feil i kategori k. Samanhengen mellom feilintensiteten λ_k (tilhøyrande Y_k) og feilintensiteten μ_k er då

$$\lambda_k = p_k\mu_k.$$

Faktisk kan det visast at om X_k er Poisson-fordelt med intensitet μ_k og vi har feilsannsynet p_k, så er Y_k Poisson-fordelt (slik vi har føresett ovanfor) med intensitet $\lambda_k = p_k\mu_k$. Følgjande døme kan klargjere forholdet mellom variablane X_k og Y_k.

> Døme 1
> Vi har følgjande tekst: «Vi slår saman dei to kategoriane vi registrerte og rekner deretter ut middelverdien.» Her er det tre verb og derfor berre tre mulege verbbøyingsfeil, B2.2, altså er X_3 observert til 3, medan det berre er gjort éin slik type feil, altså er Y_3 observert til å vere 1.

Variabelen X_k er nok litt idealisert. Blant anna vil det i mange høve vere uklart og diskutabelt kva verdi denne variabelen eigentleg har i ein tekst. Men det har ikkje nødvendigvis stor innverknad i denne samanhengen. Det viktigaste er at vi er merksame på at talet på feil i ein kategori kan vere tekstavhengig, og vi treng ein modell som tar omsyn til det. For ei ytterlegare modellvurdering sjå til dømes (Church & Gale, 1995; Katz, 1996).

For å unngå omstendelege setningar vil vi i det etterfølgjande litt upresist seie at kategorien opptrer i teksten. Med det meiner vi at vi har eit ord eller sekvens av teksten der vi potensielt kan finne denne feiltypen.

Vi er vel blant anna interesserte i å finne ut noko om frekvensane til dei ulike feiltypane, og om det er samanhengar mellom dei ulike feilkategoriane. Med andre ord er vi interesserte i feilraten til kategori k, det vil seie p_k, og vi er interesserte i om det er samanhengar mellom feilratane i dei ulike kategoriane. Vi er derimot ikkje interesserte i kor ofte kategorien opptrer i teksten. I utgangspunktet er vi altså ikkje interesserte i å seie noko om intensiteten μ_k. Spørsmålet er om datasettet kan fortelje oss noko om sannsynet for å gjere feil i dei ulike kategoriane, altså om vi kan seie noko om p_k.

Dessverre er det slik at vårt datasett ikkje gir grunnlag for å seie noko om feilratane, altså om sjansane p_k til forfattaren. Bortsett frå at vi sjølvsagt kan konkludere med at sjansen ikkje er null dersom kategorien inneheld minst ein feil. Vi kan nemleg ikkje vite om høg feilintensitet, λ_k, i kategori k skuldast at kategorien har høg intensitet μ_k, eller om det skuldast at feilsannsynet p_k er stort. Det hjelper ikkje om vi får aldri så mange og lange tekstar, eller om vi undersøkjer tekstane til aldri så mange personar. Datasettet kan ikkje svare på dette spørsmålet. Undersøkinga har med andre ord ikkje eit design som gjer at vi kan seie noko om sjansane for å gjere feil i ein kategori. Denne påstanden byggjer riktig nok på modellane som vi legg til grunn, og at vi ikkje gjer ytterlegare føresetjingar om kategoriintensitetane μ_k. Kva type føresetjingar det her eventuelt er snakk om, kjem vi tilbake til i neste avsnitt.

ER DET SKILNADAR INNANFOR KVAR AV GRUPPENE?

Det er rimeleg å spørje om det er forskjell på individa innanfor gruppene. Nok ein gong er det altså feilratane til personane vi er interesserte i, og ikkje intensitetane til dei ulike feilkategoriane. Kan datasettet svare på slike spørsmål? Viss vi er villege til å gjere visse føresetjingar, er svaret i prinsippet ja, viss ikkje er svaret nei.

La oss sjå på om det er forskjell mellom personane innanfor kvar gruppe i kategori nummer k. Vi kan føresetje at kategoriintensiteten μ_{ik} til tekst i og kategori k er lik for alle tekstane innanfor gruppa. Det vil seie at intensitetane er like i alle

tekstane innanfor gruppa. Vi føreset altså at intensiteten er personuavhengig og tekstuavhengig. Vi kan kalle denne felles intensiteten for μ_k. Altså føreset vi at

$$\mu_{1k} = \mu_{2k} = \ldots = \mu_{nk} = \mu_k.$$

Det betyr at tekst nummer i har feilintensiteten

$$\lambda_{ik} = p_{ik}\mu_k,$$

der p_{ik} er sannsynet for at person i gjer feil gitt at ordet er i kategori k.

Realismen i denne føresetnaden kan vi ikkje seie noko om ut frå datasettet. Poenget er at om vi skal seie noko om feilratane til desse personane ut frå tekstane, så må vi legge denne eller liknande føresetnadar til grunn. Med andre ord kan vi vel seie at datasettet ikkje er optimalt for å seie noko om feilratane p_{ik} til dei ulike personane i = 1, …, n. I denne artikkelen peikar vi berre på kva føresetnadar som må gjerast for å teste det vi ønskjer å teste. Vi forsøkjer ikkje å gjere vurderingar utover det som datasettet gir grunnlag for. Men utan ytterlegare kunnskapar om tekstar må vi nok dessverre seie at føresetnaden høyrer til i gruppa av litt tvilsame føresetnadar. Dersom føresetnadane er urealistiske, er det desto viktigare at dei kjem fram i lyset. Dessverre ser ein ofte at føresetnadane ikkje vert klargjorde, dermed får ein gjerne eit inntrykk av at konklusjonane er sterkare enn det er grunnlag for.

Merk likevel at å føresetje at alle har den same kategoriintensiteten μ_k, ikkje betyr at vi føreset at alle tekstane har like mange ord i denne kategorien. Vi føreset berre at denne kategorien har den same Poisson-fordelinga i alle tekstane. Følgjande analogi kan vere klargjerande. Vi føreset at kvar gong vi kastar ein mynt, så har vi femti prosents sjanse for å få kronesida opp. Det er likevel slik at om to personar kastar mynten ti gonger kvar, så vil dei sannsynlegvis ikkje få like mange kronesider i løpet av dei ti kasta.

Med den nemnde føresetjinga kan vi sette opp ei såkalla nullhypotese, H_0, og ei alternativ hypotese H_1.

H_0: Alle personane i gruppa har same sjanse for å gjere feil i kategori k, det vil seie at

$$p_{1k} = p_{2k} = \ldots = p_{nk} = p_k.$$

H_1: Ikkje alle personane i gruppa har same sjanse for å gjere feil i kategori k.

Vi nyttar den så kalla Wald-observatoren, som for den interesserte lesar er

$$\sum_{i=1}^{n} \frac{\left(Y_i - \hat{\lambda}_{ik}\right)^2}{\hat{\lambda}_{ik}},$$

der n er talet på tekstar og

$$\hat{\lambda}_k = \widehat{p_k \mu_k} = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}.$$

Wald-observatoren er tilnærma kji-kvadrat fordelt med n-1 fridomsgrader dersom n er stor.

TABELL 8.1

Nullhypotesen: Den totale feilintensiteten er lik for alle tekstane innanfor gruppa.					
	Antal tekstar i gruppa		**Dei tekstane som skil seg mest ut, er fjerna**		
	N	**p-verdi**	**Antal fjerna**	**N***	**p-verdi**
Mikrostudentar	44	0,025	3	41	0,179
Makrostudentar	43	$1{,}5 \cdot 10^{-28}$			
Pedagogikk	25	$3{,}1 \cdot 10^{-4}$	4	21	0,176
Førskule	13	0,011	2	11	0,255

Andre kolonne gir tal tekstar i gruppa. Tredje kolonne gir p-verdien for å forkaste null-hypotesen at den totale feilintensiteten er lik i alle tekstane innanfor gruppa. I høgre del av tabellen ser vi korleis p-verdien endrar seg dersom vi fjernar dei mest ekstreme tekstane, høvesvis 3, 4 og 2 tekstar i mikro-, pedagogikk, og førskulegruppa. I makrogruppa er p-verdien rimeleg uforandra sjølv om vi fjernar dei meste ekstreme tekstane, og det er derfor ikkje rapportert.

Av tredje kolonne i tabellen ser vi at nullhypotesen vert forkasta på 5 % signifikansnivå for alle gruppene. Dette er ikkje overraskande, det er trass alt ulike teksttypar skrivne av ulike studentar. P-verdien for makrostudentane er svært låg, noko som tyder på at den totale feilintensiteten varierer frå tekst til tekst. Dette er med andre ord ei variert gruppe tekstar med omsyn til den totale feilintensiteten.

I dei andre gruppene er derimot p-verdiane overraskande høge, jamvel om dei er mindre enn 5 % som er eit vanleg nivå for å forkaste nullhypotesen. I mikrogruppa er det trass alt 44 studentar, og då er det i utgangspunktet overraskande at p-verdien er så høg som 0,025, vi har trass alt ulike forfattarar og ulike tekstar. Det tyder på at feilintensitetane (for den totale feilmengda) trass alt ikkje varierer vel-

dig mykje i mikrogruppa, sjølv om vi altså kan forkaste nullhypotesen at dei er like. Dersom vi fjernar dei tre tekstane som skil seg mest ut (i mikrogruppa), får vi ein p-verdi på heile 0,179. Vi må med andre ord konkludere med at desse tekstane er forholdsvis einsarta med omsyn til den totale feilmengda. Det er i utgangspunktet overraskande at 44 tekstar, som er skrivne av ulik personar, skal vere så like. Det vert peika på at dette kan kome av at eksamen i mikroøkonomi er relativt teknisk med mange standardformuleringar som kanskje gir mindre rom for språkfeil. Om det er tilfelle, er det eit godt døme på at talet på feil er tekstavhengig så vel som personavhengig, som vi diskuterte ovanfor.

Også tekstane i førskulegruppa er forholdsvis einsarta. Dersom vi i denne gruppa fjernar dei to tekstane som skil seg mest ut, får vi ein p-verdi på høge 0,255. Med andre ord er dette også ei overraskande homogen gruppe. I pedagogikkgruppa er forskjellane mellom dei totale feilintensitetane meir synlege, men heller ikkje her er p-verdien veldig låg, og vi skal ikkje fjerne meir enn dei fire mest ekstreme tekstane før p-verdien aukar til heile 0,176.

I makrogruppa er dei totale feilintensitetane heilt klart ulike. Her har vi med andre ord ei variert gruppe tekstar med omsyn til dei totale feilintensitetane.

Desse vurderingane er nok ikkje heilt uproblematiske, og dei er kanskje først og fremst aktuelle i ei utforskande tilnærming.

Men kvifor skulle vi vere interesserte i om det ser ut som at tekstane er skrivne av ulike personar eller ikkje? Med andre ord, kva er poenget med å gjere desse testane? At alle ikkje er like gode i rettskriving, er vel relativt velkjent. Til det er det i alle fall tre ting å seie; vi veit trass alt ikkje om det gjeld i vår gruppe. Men viktigare, desse testane seier oss noko om kvaliteten på datasettet vårt. Det er nok rimeleg å tru at personane er ulike med omsyn til rettskriving, men dersom testane ikkje klarer å avsløre det, fortel det oss at tekstane kanskje er for korte. Vi må då vere forsiktige med å trekkje slutningar ut frå dette datasettet. For det tredje har dette spørsmålet relevans for om Poisson-fordelinga kan brukast. Vi kjem tilbake til det seinare.

Når feilintensitetane, til den totale feilmengda, innanfor mikro-, pedagogikk- og førskulegruppa er så vidt like, er det eit signal om at tekstane ikkje er lange nok. Det er eit signal om at forskjellane i talet på observerte feil er innanfor heilt naturlege svingingar som vi kan rekne med å finne innanfor ein og same tekst som er skriven av ein og same person. Det må likevel påpeikast at vi her berre har undersøkt den totale feilmengda i kvar tekst. Vi har ikkje sjekka kvar enkelt feiltype.

ER DET SKILNAD PÅ DEI FIRE GRUPPENE?

Fretland (2015) stiller spørsmål om studentgruppene er ulike. Vi har alt delvis svart på dette spørsmålet. Vurderingane i førre kapittel peika mot at det er større

variasjon innanfor makrogruppa enn det er innanfor dei andre gruppene. Det er spesielt lite variasjon mellom tekstane i mikrogruppa. Det er likevel eit spørsmål om den forventa feilintensiteten er lik i alle gruppene til tross for at variasjonen innanfor gruppene er ulik. Merk at dette spørsmålet impliserer andre fordelingar enn Poisson-fordelinga der variansen er lik forventninga.

I førre kapittel tok vi som utgangspunkt at feila i tekstane er Poisson-fordelte. Vi testa om tekstane innanfor kvar av gruppene hadde like feilintensitetar. Konklusjonen var, som venta, at tekstane innanfor kvar av gruppene har ulike feilintensitetar. Ein konsekvens av dette er at vi ikkje lenger kan sjå på talet på feil i dei ulike tekstane som ulike trekkingar frå den same Poisson-fordelinga. Det kan vi ikkje gjere sjølv innanfor same gruppe.

Som tidlegare er det rimeleg å anta at talet på feil i ein gitt tekst er Poisson-fordelt med ein bestemt intensitet. Men som vi fann ut i førre kapittel, ser det ut til at dei ulike tekstane innanfor kvar gruppe har ulike feilintensitetar. Og det er nok rimeleg, sidan dei er skrivne av ulike personar. Dersom vi plukkar ut ein tilfeldig tekst (innanfor ei gruppe), er det rimeleg å tenkje at talet på feil, Y, er Poisson-fordelt med intensitet Λ, der Λ er ein stokastisk variabel slik at forventninga og variansen er høvesvis

$$E(\Lambda) = \lambda \text{ og } Var(\Lambda) = \sigma^2.$$

For å forenkle har vi utelatt indeksen for kategorinummeret. At Λ er ein stokastisk variabel, betyr at intensiteten kan variere frå person til person. At forventninga til Λ er λ, betyr litt uformelt at i snitt er intensiteten lik λ. Ut frå regelen om dobbel forventning (Høyland, 1988, s. 107) kan vi då vise at forventa tal feil er

$$E(Y) = \lambda \text{ og } Var(Y) = \lambda + \sigma^2.$$

Vi ser at forventninga er i tråd med Poisson-fordelinga, men at variasjonen aukar i forhold til ei Poisson-fordeling, vi får eit ekstra ledd σ^2. I denne modellen kan altså ulike tekstar ha same forventa feilintensitet til tross for at variansen er ulik, det går ikkje i Poisson-fordelinga.

I slike situasjonar er det ofte vanleg å føresette at feilintensiteten Λ har ei spesiell Gamma-fordeling $\Gamma(\nu,\nu)$. Då kan det visast (Cameron & Trivedi, 1998, s. 675) at talet på feil Y har ei såkalla negativ binomialfordeling der

$$E(Y) = \lambda \text{ og } Var(Y) = \lambda + \alpha\lambda^2 = \lambda(1 + \alpha\lambda).$$

Parameteren α vert gjerne kalla dispersjonsparameteren.

Vi kan sjå på dispersjonsparameteren α som eit uttrykk for kor mykje dei ulike tekstane varierer med omsyn til feilintensitet. Dersom alle tekstane hadde hatt lik feilintensitet, ville $\alpha = 0$, då ville feila i ein tilfeldig valt tekst vore Poisson-fordelt med feilintensitet λ. Dess større α er, dess større variasjon er det mellom feilintensitetane i dei ulike tekstane.

Vi kan spørje om det er forskjell mellom dei fire studentgruppene med omsyn til forventa feil i dei ulike kategoriane. I utgangspunktet er datasettet heller ikkje eigna til å svare på dette spørsmålet. Igjen kjem det av at vi ikkje veit noko om kategoriintensitetane, μ_k, i dei ulike tekstane. Vi ser derfor berre på om tekstane innanfor dei ulike studentgruppene er like med omsyn til feilintensitet. Vi vurderer altså ikkje om studentane har lik feilintensitet.

La $\lambda_k^{(Mi)}, \lambda_k^{(Ma)}, \lambda_k^{(Pe)}$ og $\lambda_k^{(F)}$ vere dei forventa feilintensitetane til kategori k for høvesvis tekstar i gruppa mikroøkonomi, makroøkonomi, pedagogikk og førskule. Vi set opp følgjande nullhypotese:

$H_0: \lambda_k^{(Mi)} = \lambda_k^{(Ma)} = \lambda_k^{(Pe)} = \lambda_k^{(F)}$, det vil seie at den forventa feilintensiteten er lik for alle gruppene.

Den alternative hypotesen er at dei ikkje er like.

For ordens skuld tar vi med observatoren og ei kort vurdering av han, men dette er ikkje spesielt viktig for tolkingsdiskusjonen som følgjer etter tabell 8.2. For kvar kategori k kan vi bruke observatoren

$$\sum_{j \in \{Ma, Mi, Pe, F\}} \frac{\left(\bar{Y}_k^j - \bar{Y}_k\right)^2}{\mathrm{Var}\left(\bar{Y}_k^j\right)},$$

der til dømes $\bar{Y}_k^{Ma}$ er gjennomsnittet i makrogruppa, og $\bar{Y}_k$ er gjennomsnittet i heile datasettet for kategori *k*. Merk at vi har estimert $\mathrm{Var}\left(\bar{Y}_k^j\right)$ ved hjelp av heile datasettet, det vil blant anna medføre at vi brukar same varians i alle gruppene, noko som er rett under nullhypotesen, men det gir ein konservativ observator (låg teststyrke). I staden burde vi kanskje estimert $\mathrm{Var}\left(\bar{Y}_k^j\right)$ for kvar gruppe, det ville gitt ein mindre konservativ observator, og vi kunne forventa lågare p-verdiar. Vi vel ei konservativ holdning blant anna fordi at kji-kvadratfordelinga til observatoren byggjer på ei normaltilnærming som neppe er oppfylt når vi har få feil per kategori per tekst og dessutan få observasjonar i førskulegruppa. I tabell 8.3 viser vi likevel p-verdiane for begge alternativa.

Merk også at observatoren tar omsyn til at tekstane innanfor kvar gruppe har ulike feilintensitetar. Dersom det ikkje var tilfelle, ville summen av feila innanfor

gruppa vore Poisson-fordelt, og variansen ville vore mindre. Observatoren ville då gitt langt lågare p-verdiar.

TABELL 8.2

Kategori	Estimert intensitet, $\hat{\lambda}$	Estimert varians	p-verdi
A4	0,77	1,55	0,000
B1.1	0,91	1,98	0,002
B1.2	0,53	0,63	0,020
B2.2	1,46	1,94	0,214
C3.1	0,62	1,25	0,064
C3.2	0,20	0,21	0,050
C3.3	0,27	0,31	0,049
C3.4	0,43	0,37	0,677
C3.5	0,21	0,21	0,001
C3.6	0,22	0,20	0,001
C3.7	0,87	1,23	0,000
C4.1	0,22	0,23	0,208
C4.2	0,14	0,15	0,005
C4.3	0,10	0,14	0,158
C4.4	1,15	1,76	0,021
C4.5	1,10	1,43	0,000
D1	0,67	1,25	0,008
D2	0,57	0,82	0,994
Alle feil	10,43	38,53	0,005

Andre og tredje kolonne viser høvesvis estimert forventa tal feil og estimert varians i dei ulike kategoriane under nullhypotesen om lik forventning og varians i alle dei fire gruppene. Observatoren er tilnærma kji-kvadratfordelt med tre fridomsgrader under nullhypotesen.

P-verdiane i tabell 8.2 er gjennomgåande låge. Det vil seie at det er grunn til å tru at både den totale feilintensiteten og feilintensitetane for dei ulike kategoriane varierer mellom dei fire gruppene.

Dersom vi fjernar mikrostudentane og samanliknar makro-gruppa, pedagogikk- og førskulegruppa, finn vi derimot gjennomgåande høge p-verdiar, sjå tabell 8.3.

Berre kategoriane A4, B1.1, C3.5 og C3.7 har p-verdiar mindre enn 0,05. Med heile 18 testar er det å forvente nokre fåe signifikante utslag sjølv om det eigentleg ikkje er forskjellar. Med andre ord kan vi litt røft seie at det verkar som om vi kan forvente omtrent like mange feil i desse tre gruppene. Det vil seie at dei tre gruppene makro, pedagogikk og førskule er rimeleg like. Dette er kanskje litt overraskande om vi studerer figur 8.1, der vi nok kan få inntrykk av at gruppene er ulike i mange av kategoriane.

Konklusjonen er at mikrostudentane skil seg ut som den gruppa med færrast feil, medan dei andre gruppene er nokså like.

I utgangspunktet kan det sjå ut som om gruppene er nokså tilfeldig utvalde. Ein statistikar vil då gjerne spørje om dei fire gruppene er valt ut fordi ein ønskjer å studere akkurat desse gruppene, eller om dei berre er tilfeldige grupper. Kvifor skulle det vere interessant å finne ut om tilfeldig utvalde grupper er ulike? Om vi finn at gruppene er ulike, kva er i så tilfelle interessant ved at tilfeldige grupper er ulike? Poenget vårt er at vi som forskarar må ha eit reflektert forhold til dei spørsmåla vi stiller oss. Det har lite hensikt å svare på uinteressante spørsmål. Når det er sagt, så kan slike spørsmål seie noko om kvaliteten i datasettet som kan vere viktig for å svare på andre spørsmål. Dersom tilfeldige grupper som i utgangspunktet burde vere like, viser seg å vere svært ulike, bør dei ikkje behandlast som ein populasjon.

Frå ein didaktisk synsvinkel synes vi det er spørsmålet om samanhengar mellom feilkategoriane som er det interessante spørsmålet å stille i dette datasettet. Negativ-binomialfordelinga, som vi kjem fram til her, saman med generaliserte lineære modellar, er grunnlaget for ei slik analyse (Cameron & Trivedi, 1998; Hilbe, 2011; Zeileis, Kleiber, & Jackman, 2008). For ein utfyllande analyse viser vi til (Myklebust, u.d.). I ein slik analyse kan det dessutan vere viktig å vite om studentane og gruppene er tilnærma like eller ikkje. Det kjem av at det kan innverke på korleis vi bør handsame datasettet når vi ser etter samanhengar mellom kategoriane. Dessutan såg vi at det er viktig for å vurdere kva for ei fordeling vi bør legge til grunn. Vi argumenterte for at ei negativ-binomialfordeling passar betre enn ei Poisson-fordeling, nettopp fordi at tekstane ser ut til å ha ulike feilintensitetar. Spørsmålet om kva fordeling vi skal bruke, er viktig for den vidare analysen. Dessutan kan det tenkast at gruppene er for ulike til at dei kan vurderast samla. Vår konklusjon er at ein bør vurdere om mikrogruppa bør utelatast dersom ein skal sjå etter samanhengar mellom feilkategoriane. Med andre ord, sjølv om spørsmålet om gruppene er ulike kanskje ikkje er interessant i seg sjølv, så er det likevel viktig som eit ledd i andre deler av analysen.

TABELL 8.3

Er feilintensiteten lik hjå makro, pedagogikk og førskule?			
Kategori	**$\hat{\lambda}$, estimert intensitet (forventa feil per tekst)**	**p-verdi. Når variansen er antatt lik i alle gruppene**	**p-verdi. Når variansen er estimert innanfor kvar gruppe**
A4	0,95	0,000	0,000
B1.1	1,05	0,018	0,000
B1.2	0,41	0,146	0,125
B2.2	1,41	0,261	0,130
C3.1	0,77	0,388	0,012
C3.2	0,25	0,230	0,126
C3.3	0,37	0,905	0,887
C3.4	0,48	0,999	1,000
C3.5	0,26	0,012	0,060
C3.6	0,10	0,268	0,000
C3.7	1,07	0,005	0,000
C4.1	0,26	0,499	0,083
C4.2	0,19	0,078	0,078
C4.3	0,14	0,391	0,058
C4.4	1,26	0,077	0,120
C4.5	1,43	0,084	0,011
D1	0,90	0,665	0,583
D2	0,58	0,998	0,995
Alle feil	11,86	0,749	0,335

Test av nullhypotesen at makro-, pedagogikk- og førskulegruppene har lik feilintensitet. Andre kolonne viser estimert forventa tal feil under nullhypotesen om lik forventning og varians i makro-, pedagogikk- og førskulegruppene. Tredje kolonne viser p-verdiane der variansen er antatt lik i alle gruppene og estimert utfrå heile datasettet. Fjerde kolonne viser p-verdiane når variansen er estimert individuelt for kvar gruppe. I begge tilfeller er observatoren tilnærma kji-kvadratfordelt med to fridomsgrader under nullhypotesen.

OPPSUMMERING

Vi har tre ulike tilnærmingar til eit datasett: 1) Datasettet gir oss all informasjon, og vi ønskjer å beskrive datasettet. 2) Vi ønskjer å trekke konklusjonar som har gyldigheit utover datasettet, typisk har vi då konkrete hypotesar som vi testar. 3) Vi utforskar datasettet for å sjå om det kan ymte om ny kunnskap.

Dette datasettet er ikkje samla inn for å teste veldefinerte hypotesar, og det er heller ikkje interessant i seg sjølv. Med andre ord er det den tredje tilnærminga ovanfor som er aktuell. Vi studerer datasettet for å forsøke å kartlegge og å forstå noko om rettskrivings- og bøyingsfeil. Statistiske metodar kan også brukast til det, men resultatet av denne typen utforsking er, i beste fall, ein eller fleire hypotesar om ortografiske feil og feilmønster. Eventuelle låge p-verdiar kan ikkje tolkast som statistisk signifikante resultat, men dei indikerer at vi har ein god hypotese.

Vi modellerer både potensielle avvik og faktiske avvik. Poisson-fordelinga og negativ-binomialfordeling er aktuelle fordelingar. Negativ-binomialfordelinga er meir fleksibel og opnar for meir strukturell variasjon mellom tekstane/forfattarane. Utan ytterlegare informasjon om tekstane, og tekstar generelt, kan vi i lita grad seie noko om spørsmål som til dømes kva for kategoriar folk har størst problem med. Det kjem av at vi ikkje kan seie om høg frekvens i ein kategori skuldast at forfattaren har vanskar med denne kategorien, eller om det kjem av at kategorien har høg frekvens i teksten. Mange av konklusjonane må derfor gjelde tekstar, og vi kan berre i liten grad overføre desse konklusjonane til forfattarane. Det siste ville nok vere det mest interessante.

Som forventa, ser det ut til at intensiteten til den totale feilmengda varierer mellom tekstane, men i mikrogruppa er forskjellane, etter ei totalvurdering, trass alt små. Kanskje skuldast det at tekstane i gruppa er prega av mange standardiserte formuleringar. I alle fall er mikro-gruppa nokså homogen med omsyn til den totale feilmengda. Det er likevel liten grunn til å tru at feilintensitetane i 44 tekstar som er skrivne av ulike studentar, skal vere omtrent like. Konklusjonen er derfor at tekstane i mikrogruppa ikkje er lange nok til at vi ser forskjellane på dei, og i mange samanhengar vil det derfor vere lite informasjon å hente frå denne gruppa. Det er derfor eit spørsmål om denne gruppa bør vere med i ein eventuell vidare analyse av datasettet. Som grupper er dei andre gruppene rimeleg like, og det vil i mange samanhengar vere eit kvalitetsteikn med datasettet. Det indikerer nemleg at vi kan sjå på tekstane i desse tre gruppene som observasjonar frå ein populasjon, noko som er ein fordel.

Eit interessant norskdidaktisk spørsmål for vidare analyse er om det er samanhengar mellom dei ulike feilkategoriane, sjå (Myklebust, u.d.). Vår vurdering er at mikrogruppa neppe bør takast med i den analysen, men at dei tre andre grup-

pene i utgangspunktet kan sjåast på som ein populasjon, og slik sett dannar dei eit godt grunnlag for ein slik analyse.

LITTERATUR

Cameron, A.C., & Trivedi, P.K. (1998). *Regression analysis of count data.* New York: Cambridge University Press.

Church, K.W., & Gale, W.A. (1995). Poisson mixture. *Natural language enginering., 1*, ss. 163–190.

Fretland, J.O. (2007). Du skriv feil, lærar. I A.S. Norddal, *Betre nynorskundervisning.* (ss. 71–82). Skriftserien Nasjonalt senter for nynorsk i opplæringa.

Fretland, J.O. (2011). Nynorsk ordval i studentoppgåver: Om «pengeetterspørselens inntektsfølsomheit» og andre nøtter for nynorsk i opplæringa. I Jenstad, & Vikør, *Leksikalsk forsking i norske målføre og nynorsk skriftspråk.* DKNVS Skrifter.

Fretland, J.O. (2015). «Vi analyserar økningen i isokvanter.» Ein analyse av nynorskfeil i studentarbeid. I H. Eiksund, & J.O. Fretland, *Nye røyster i nynorskforskinga.*

Hilbe, J.M. (2011). *Negative binomial regression.* New York: Cambridge University Press.

Høyland, A. (1988). *Sannsynlighetsregning og statistisk metodelære. 1 sannsynlighetsregning.* (5. utg.). Trondheim: Tapir.

Katz, S.M. (1996). Distribution of content words and phrases in text and language modelling. *Natural language enginering, 2*, ss. 15–59.

Myklebust, T. (u.d.). *Ein analyse av studenttekstar med statistiske metodar.* Sogndal: Under arbeid. HISF.

Tukey, J.W. (1977). *Exploratory data analysis.* Pearson.

Tukey, J.W. (1980). We need both exploratory and confirmatory. *The american statistician, 34*(1), ss. 23–25.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software, 27*, ss. 1–25.