Høgskolen i Østfold

Statistikk og statistisk programmering, våren 2022

Obligatorisk oppgave 5

Innleveringsfrist: Fredag 25. mars 2022, kl. 23.59

Leveres på Canvas.

De Python-baserte oppgavene må gjøres for å få godkjent innleveringen.

Du må levere selve Python-fila, og ikke bare bilde av koden. I tillegg skal du levere en pdf-fil med svar på resterende oppgaver. Du kan også skrive Python-koden i Word så vi selv kan kopiere det over i egen fil. Om du skriver oppgavene i Word, eller om du skriver for hånd og limer inn bilde av dette i Word-fila, spiller ingen rolle. Det som er viktig er at det er oversiktlig og lesbart for oss! Outputen du får når du kjører fila tar du et bilde av og limer inn i pdf-en.

Oppgave 1

Denne oppgaven skal løses uten bruk av Python.

Vanlig kjøttdeig skal ikke inneholde mer enn 14 % fett. Vi har mistanke om at kjøttdeig inneholder for mye fett, og bestemmer oss for å utføre en hypotesetest. La X være prosentandelen fett i en tilfeldig 400 g pakke kjøttdeig. Vi antar at X er normalfordelt med forventningsverdi μ og standardavvik $\sigma = 3$.

a) Formuler hypotesene.

b) Utfør testen med signifikansnivå 0.05 når vi har følgende målinger:
14, 15, 18, 12, 17, 13, 15, 19, 16

c) Beregn testens *p*-verdi.

d) Beregn testens styrkefunksjon for verdiene 13, 14, 15, 16 og 17 og skisser den.

e) Vi krever at sannsynligheten for type II feil skal være mindre enn 0.10 når $\mu = 16$. Hvor mange målinger må vi foreta for å oppnå denne styrken?

Oppgave 2

Denne oppgaven skal løses uten bruk av Python (med unntak av spørsmål c).

Denne oppgaven er til forveksling lik den foregående, men nå er standardavviket ukjent. Dessuten skal vi ha et annet signifikansnivå og litt andre målinger. Videre dropper vi spørsmålene om styrkefunksjon (fordi det blir litt komplisert i dette tilfellet).

Vanlig kjøttdeig skal ikke inneholde mer enn 14 % fett. Vi har mistanke om at kjøttdeig inneholder for mye fett, og bestemmer oss for å utføre en hypotesetest. La X være prosentandelen fett i en tilfeldig 400 g pakke kjøttdeig. Vi antar at X er normalfordelt med forventningsverdi μ og ukjent standardavvik σ.

a) Formuler hypotesene.

b) Utfør testen med signifikansnivå 0.01 når vi har følgende målinger:
15, 17, 18, 14, 19, 16, 17, 15, 19, 13

c) Bruk en tabell til å finne et anslag for testens p-verdi (vi kan ikke finne den eksakte verdien fra våre tabeller). Bruk også Python til å finne den eksakte verdien ved hjelp av `scipy.stats.t`-funksjonen.

Oppgave 3

Denne oppgaven skal løses uten bruk av Python.

Et legemiddelfirma har utviklet et nytt medikament til behandling av pasienter med Alzheimers sykdom. Eksisterende medikamenter har en nytteeffekt hos 55 % av pasientene. Legemiddelfirmaet påstår at det nye medikamentet er bedre enn de eksisterende. For å teste hypotesen blir 45 pasienter behandlet med det nye medikamentet. Totalt X av disse vil oppleve nytte av medisinen. Vi kan anta at variabelen X er binomisk fordelt med ukjent parameter p.

Testen av det nye medikamentet viste at 35 av de 45 pasientene hadde nytte av behandlingen. Utfør en hypotesetest med signifikansnivå 0.05.

Oppgave 4

Lag et Python-program som utfører en hypotesetest basert på følgende scenario.
(Tips: du kan bruke funksjonen `scipy.stats.ttest_1samp` i denne oppgaven.)

På den greske øya Kripos kryr det av villkatter. Man antar at mange turister mater kattene når de sitter og spiser. Man ønsker derfor å undersøke om vekten til kattene er høyere på slutten av turistsesongen enn ellers i året.

Anta at vekten til voksne katter på Kripos før turistsesongen har begynt, er normalfordelt med forventningsverdi $\mu = 2.75$ kg.

For å undersøke om vekten har økt i løpet av turistsesongen, veier man en del katter på slutten av turistsesongen. Resultatet av målingene finner du i følgende fil:
http://www.it.hiof.no/statok/katters_vekt.csv

De fire feltene i filen er: løpenummer, kjønn, kattens vekt i kg og hjertets vekt i gram.

a) Angi i besvarelsen hypotesene du bruker.

b) Utfør en hypotesetest ved å beregne testobservatoren t og sammenligne den med kvantilet i t-fordelingen for å finne ut om undersøkelsene gir belegg for å hevde at forventningsverdien til kattenes vekt har økt i løpet av turistsesongen.

Programmet skal be om signifikansnivå «alfa» og forventningsverdien «my_0». «my_0» er altså den forventningsverdien til vekten som vi skal sjekke mot, altså 2.75 i vårt tilfelle: Er forventningsverdien til kattenes vekt høyere enn 2.75 på slutten av turistsesongen?

c) Finn testens p-verdi og skriv den ut.

d) Kjør programmet flere ganger og finn hvilket signifikansnivå som gjør at vi går fra å beholde H_0 til å forkaste H_0. Sammenlign med p-verdien.

e) Undersøk ved å kjøre programmet med ulike verdier av my_0 hvilken my_0 som er grensa mellom å forkaste H_0 og beholde H_0 når vi gjør hypotesetest på signifikansnivå $\alpha = 0.05$.

Oppgave 5

Følgende oppgave skal løses uten bruk av Python.

I tabellen nedenfor har jeg plukket ut data for noen få av kattene fra oppgave 4.

Kattens vekt (kg)	2	2.4	2.6	3.2	3.6
Hjertets vekt (g)	6.5	7.9	8.3	11.9	13.3

Beregn den empiriske korrelasjonskoeffisienten r mellom disse kattenes vekt og vekten av deres hjerter.

Oppgave 6

Lag et Python-program som beregner og skriver ut den empiriske korrelasjonskoeffisienten r mellom kattenes vekt og vekten av deres hjerter, basert på hele datasettet http://www.it.hiof.no/statok/katters_vekt.csv.

Funksjonen `scipy.stats.pearsonr` kan brukes til dette.

Angi i besvarelsen hvilken verdi for r du finner.

(Korrelasjonskoeffisienten r kalles ofte Pearsons korrelasjonskoeffisient, oppkalt etter den engelske matematikeren Karl Pearson. Det er derfor scipy-funksjonen for å beregne den heter «`pearsonr`». Korrelasjonskoeffisienten r er selvsagt oppkalt etter Pearson siden det var Bravais og Galton som først definerte denne …).

Oppgave 7

I denne oppgaven skal du bruke Python kun i spørsmål e.

I forbindelse med Vinmonopolets lansering av årets viner fra Burgund for noen år siden, sto det en artikkel på nettavisen «Børsen». (Du kan finne den her hvis du er interessert: https://www.borsen.no/livsstil/skremt-etter-vinavsloring---skandale/72102130.)

I artikkelen fortelles det at det ble åpnet 74 vinflasker, og at 11 av disse var «korket», altså at vinen hadde blitt helt eller delvis ødelagt på grunn av lekkasje i korken. Dette var en overraskende stor feilprosent. Ifølge artikkelen er det vanlige at 5 % av flaskene har slike feil.

a) Anta at man før starten av denne prøvesmakingen trekker ut en tilfeldig flaske av de 74 flaskene. Hva er sannsynligheten for at denne flasken har korkfeil?

b) Hvis vi i utgangspunktet trekker ut 10 av de 74 flaskene, hva er sannsynligheten for at nøyaktig 5 av disse har korkfeil?

Den ansvarlige i Vinmonopolet kommer i artikkelen med følgende uttalelse:
«Jeg tror det var ganske tilfeldig at det dukket opp så vidt mye korkfeil.»
Vi skal undersøke dette ved hjelp av en hypotesetest.

c) For at vi skal kunne bruke den standardnormalfordelte testobservatoren Z til hypotesetesten, må det stilles et krav til variansen (se regel 6.21, s. 273 i Løvås). Undersøk om dette kravet er oppfylt i dette tilfellet.

d) Uansett om du i spørsmål c fant at kravet var oppfylt eller ikke, skal du gjennomføre en hypotesetest for å vurdere om uttalelsen fra den ansvarlige i Vinmonopolet er en rimelig uttalelse. Benytt signifikansnivå 0.1. Det du skal undersøke er altså om det er rimelig å anta at de 74 flaskene som ble åpnet faktisk kan være et tilfeldig utvalg fra en populasjon med maksimum 5 % feil.

e) Et alternativ til å gjennomføre en hypotesetest basert på Z, er å benytte den eksakte p-verdien utregnet på bakgrunn av at dette er en binomisk fordeling. Bruk Python og finn den eksakte p-verdien, og gjør en hypotesetest basert på denne.

Oppgave 8

I denne oppgaven skal du bruke Python i alle delspørsmål unntatt c.

Det er gjort målinger av den del katters vekt sammen med vekten av deres hjerter. Resultatet av målingene finner du i følgende fil:
http://www.it.hiof.no/statok/katters_vekt.csv

De fire feltene i filen er: løpenummer, kjønn, kattens vekt i kg og hjertets vekt i gram.

Det kan virke som om vekten til kattenes hjerter øker med deres kroppsvekt.

a) Lag et spredeplott som viser vekten til kattenes vekt på førsteaksen og kattenes hjertevekt på annenaksen. Her kan følgende funksjon brukes:

```
matplotlib.pyplot.scatter(x,y)
```

b) Lag et Python-program som finner regresjonslinjen som er best tilpasset vekten til kattens hjerte (Y) som funksjon av kattens vekt (x).

Programmet skal også beregne og skrive ut standardfeilen til estimatoren for β.

I løsningen skal du angi den regresjonslinjen du finner, og også standardfeilen til estimatoren for β.

Her kan følgende funksjon brukes:

```
scipy.stats.linregress(x,y)
```

c) I dette delspørsmålet skal du **ikke** bruke Python, men du kan bruke resultatene fra spørsmål b.

Gjennomfør en hypotesetest av om det er en sammenheng mellom hjertets vekt og kattens vekt. Benytt signifikansnivå $\alpha = 0.05$. Benytt en høyresidig test, altså sjekk om $\beta > 0$ fordi det er dette som er interessant, altså om våre data virkelig tyder på at en større katt som hovedregel har et større hjerte enn en mindre katt. (En $\beta < 0$ betyr at store katter har mindre hjerter enn små katter, noe som åpenbart må være feil, og derfor bruker vi en høyresidig test fremfor en tosidig test).

Antall frihetsgrader her er $v = n - 2 = 97 - 2 = 95$. Dette finnes ikke i tabellen, og du kan derfor benytte $v = 100$. Feilen en gjør ved dette, er ubetydelig.

En slik hypotesetest er forklart i avsnitt 7.3.5 i boka.

d) Bruk Python til å finne og skrive ut et 95 % konfidensintervall for forventningsverdien til hjertets vekt $E(Y)$ for ulike verdier av kattens vekt. Kattens vekt skal brukeren taste inn.

Dette er beskrevet i avsnitt 7.3.6.

I besvarelsen skal du angi resultatet for tre ulike vekter av en katt, og avrundet til én desimal.

e) Bruk Python til å finne et 95 % prediksjonsintervall for hjertets vekt for ulike verdier av kattens vekt. Kattens vekt skal brukeren taste inn.

Dette er beskrevet i avsnitt 7.3.7.

Angi resultatet avrundet til én desimal for de samme tre vektene av kattene som i oppgave d.