EKSAMEN I HSTAT1101, 3. APRIL 2019: LØSNINGSFORSLAG

VERSJON MENT FOR EKSAMENSKANDIDATENE

Knut R. Wangen,

15.04.2019

k.r.wangen@medisin.uio.no

Innledning

Eksamen ble arrangert digitalt på plattformen Inspera og besto av 9 oppgaver. Tre av oppgavene var flervalgsoppgaver der kandidatene måtte velge ett svar blant flere alternativer. De øvrige oppgavene var langsvarsoppgaver som måtte besvares med tekst.

Svarene på flervalgsoppgavene ble vurdert automatisk mens svarene på langsvarsoppgavene ble vurdert skjønnsmessig. Hver av oppgavene ga maksimalt 10 poeng, slik at samlet poengsum var maksimalt 90 poeng.

Følgende sammenheng mellom karakterene (A-F) og poengene [0-90] ble brukt: A: [80, 90], B: [70, 79], C: [50, 69], D: [40, 49], E: [30, 39], F: [0, 29].

Begrunnelser for karakter. Etter at karakterene er kunngjort får kandidatene tilgang til herværende løsningsforslag. I tillegg vil kandidatene ha tilgang til sine oppnådde poeng på de enkelte deloppgavene, ved å logge seg på uio.inspera.no (under flagget "Arkiv"). En kandidats detaljerte oversikt over oppnådde poeng og løsningsforslaget anses som en formell begrunnelse for karakteren, og det vil derfor ikke bli gitt noen annen formell begrunnelse.

Klage på karakter. De formelle reglene for klage på karakter finnes på: http://www.uio.no/studier/emner/medisin/helseadm/HSTAT1101/

Uformelle tilbakemeldinger. Kandidater som ønsker ytterligere tilbakemeldinger på sin besvarelse oppfordres til å sende undertegnede en epost for å avtale en uformell samtale. I eposten må kandidaten foreslå 2–3 alternative tidspunkter for samtalen, som kan skje per telefon/Skype/WhatsApp eller i møte. Tidspunktene må være i vanlig arbeidstid (9:00–17:00), foreslås med minst 48 timers varsel, og være innen 31. mai 2019. Eposten må også oppgi kandidatens kandidatnummer.

Løsningsforslag til oppgavene

Oppgave 1

a) Tobakksforbruk er en kategorisk, nominell variabel (svaralternativene har ingen naturlig rekkefølge). Utdanningsnivå en er kategorisk, ordinal variabel (svaralternativene har en naturlig rekkefølge)

b) Fordelingen til f.eks. utdanningsnivå kan beskrives numerisk ved å beregne andelen personer (i utvalget eller populasjonen) som tilhører hver utdanningskategori. Dette kan fremstilles grafisk f.eks. i stolpediagram eller kakediagram.

I et stolpediagram for en ordinal katetorisk variabel vil man ofte sortere kategoriene i en naturlig rekkefølge (enten fra lav til høy utdanning, eller fra høy til lav utdanning). For en nominelle kategoriske variabel kan rekkefølgen velges mer vilkårlig.

Oppgave 2

$$\begin{aligned} P(6 < X < 18) &= P(X = 7) + P(X = 8) + \ldots + P(X = 17) \\ &= P(X \leq 17) - P(X \leq 6) = 0{,}9444 - 0{,}0512 \approx \underline{\underline{0{,}893}}. \end{aligned}$$

Oppgave 3

$P(|Y| > 1{,}356) = 2 \cdot P(Y > 1{,}356) = 2 \cdot 0{,}10 = \underline{\underline{0{,}20}}.$

Oppgave 4

1. Vi setter $\widehat{p} = 0{,}5$ og finner

$$n = \frac{\widehat{p}(1-\widehat{p})(Z_{\alpha/2})^2}{\mathrm{ME}^2} = \frac{0{,}25 \cdot 1{,}96^2}{0{,}02^2} = \underline{\underline{2401}}.$$

2. Vi setter $\widehat{p} = 0{,}15$ og finner

$$n = \frac{\widehat{p}(1-\widehat{p})(Z_{\alpha/2})^2}{\mathrm{ME}^2} = \frac{0{,}15 \cdot 0{,}85 \cdot 1{,}96^2}{0{,}02^2} = 216{,}09,$$

det vil si $\underline{\underline{217}}$ etter avrunding opp til nærmeste heltall.

Oppgave 5

$P(Z > |-1{,}48|) = 2 \cdot P(Z > 1{,}48) = 2 \cdot 0{,}0694 = \underline{\underline{0{,}139}}$.

Oppgave 6

Vi bruker formelen $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ og setter inn de oppgitte verdiene. Dette gir $2{,}057 \pm 1{,}96\frac{1{,}440773}{\sqrt{1000}}$, som er lik $\underline{\underline{(1{,}97; 2{,}15)}}$.

Det er forutsatt at observasjonene i utvalget (pasientene) er trukket uavhengig av hverandre. Formelen for konfidensintervallet gjelder tilnærmet, under forutsetningen av at utvalget er stort nok til at utvalgsgjennomsnittet blir tilnærmet normalfordelt (sentralgrenseteoremet). Formelen gjelder eksakt når observasjonene trekkes fra en normalfordelt populasjon med kjent varians, men dette virker ikke realistisk fordi fordelingen ser ut til å være skjev mot høyre.

Oppgave 7

La $D_0 = \mu_x - \mu_y$ være differansen mellom forventet søvnkvalitet etter forsøket (μ_x) og forventet søvnkvalitet før forsøket (μ_y). La nullhypotesen være $H_0 : D_0 = \mu_x - \mu_y = 0$, og alternativhypotesen være $H_1 : D_0 \neq 0$. La valgt signifikansnivå være $\alpha = 5\%$.

Testobservatoren blir ved innsetting

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} = \frac{0{,}70}{1{,}97/\sqrt{37}} = 2{,}161.$$

Kritisk verdi ($t_{0{,}025}$) ved 37 frihetsgrader ligger mellom den kritiske verdien for 30 frihetsgrader (2,042) og for 40 frihetsgrader (2,021). Den observerte verdien 2,161 må derfor være større enn den kritiske verdien og vi kan forkaste H_0 på 5% signifikansnivå. Det ser altså ut til at forsøket påvirket søvnkvaliteten.

Oppgave 8

For å undersøke om forskjellene i gjennomsnittene er statistisk signifikant bør det gjøres en formell hypotesetest der alle dataene inngår.[1] La nullhypotesen være $H_0 : \mu_x - \mu_y = 0$, og alternativhypotesen være $H_1 : \mu_x - \mu_y \neq 0$. Siden oppgaven oppgir 95% konfidensintervall, er det naturlig å velge et 5% signifikansnivå.

[1] Et alternativ er å beregne et konfidensintervall for differansen av populasjonsforventningene.

Anta at utvalgene er trukket fra normalfordelte populasjoner med lik varians. Estimatet for variansen blir

$$s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}$$

$$= \frac{(65-1)1{,}34^2 + (65-1)1{,}69^2}{65+65-2} = 2{,}32585,$$

mens testobservatoren blir

$$t = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{s_p^2}{n_x}+\frac{s_p^2}{n_y}}} = \frac{7{,}04-6{,}40}{\sqrt{\frac{2{,}32585}{65}+\frac{2{,}32585}{65}}} = 2{,}392.$$

Den kritiske verdien ($t_{0{,}025}$) for $65+65-2 = 128$ frihetsgrader må ligge mellom verdien for 100 frihetsgrader (1,984) og verdien for uendelig mange frihetsgrader (1,960). Den observerte verdien 2,392 må derfor være større enn den kritiske verdien og vi kan forkaste H_0 på 5% signifikansnivå. Gjennomsnittene er derfor signifikant forskjellige.

Oppgave 9

Hvis en deltager helt tilfeldig velger en plastbolle, så vil sannsynligheten for å velge plastbollen som skjuler mobiltelefonen være $p = 1/4$. Nullhypotesen kan dermed uttrykkes $H_0 : p = 1/4$, mens alternativhypotesen blir $H_1 : p > 1/4$.

Vi kan betrakte hver deltager som et Bernoulli-forsøk, der sannsynligheten for suksess er 1/4. Antall suksesser for alle $n = 64$ deltagere vil da ha en binomisk fordeling. Fra utvalget kan vi beregne utvalgsandelen, $\widehat{p}$. Gitt H_0, vil utvalgsandelen være tilnærmet normalfordelt fordi $n \cdot p(1-p) = 64 \cdot 0.25 \cdot 0.75 = 12 > 5$.

La signifikansnivået være 5% significance level. Da vil testobservatoren

$$Z = \frac{\widehat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\widehat{p}-0{,}25}{\sqrt{\frac{0{,}25(1-0{,}25)}{64}}}$$

ha 1,645 som kritisk veri. Dette betyr at vi vil forkaste H_0 hvis

$$\widehat{p} > 0{,}25 + 1{,}645\sqrt{\frac{0{,}25(1-0{,}25)}{64}} = 0{,}3561.$$

Ved å gange med utvalgsstørrelsen får vi $0{,}3561 \cdot 64 = 21{,}7$, og vi vil derfor forkaste H_0 hvis 22 eller flere deltagere identifiserer riktig plastbolle.