

CA03 - STAT100 - V20

Fábio Rodrigues Pereira

Oppgave 0

Se på løsningsforslaget til forrige ukes kollokvieoppgaver, diskuter en ting hver som dere ikke fikk til helt eller forstod, og som dere nå skjønner litt bedre.

Skriv ned hvilke temaer dere synes var vanskelige, og hva som gjorde at dere klarte å forstå det litt bedre. (For eksempel kollokviediskusjoner, jobbe med øvingsoppgaver, lese i boka, se filmer etc.)

$$\text{Oppgave 2f.) } \text{KI } 95\% = \bar{X} \pm Z_{0,025} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow (91,93; 99,21)$$

$\underbrace{Z_{0,025}}_{1,96}$ $\underbrace{\frac{\sigma}{\sqrt{n}}}_{29}$

$$\text{Oppgave 2g.) } \text{KI } 99\% = \bar{X} \pm Z_{0,005} \cdot \frac{\sigma}{\sqrt{n}} = (90,79; 100,36)$$

$\underbrace{Z_{0,005}}_{2,575}$ $\underbrace{\frac{\sigma}{\sqrt{n}}}_{29}$

Oppgave 1

A) i) Ta utgangspunkt i ditt eget (studie)liv og kom med minst et eksempel (ett eksempel hver i kollokviegruppene) på noe som kan antas å være binomisk fordelt.

ii) Sett opp de tre antakelsene som må være oppfylt for at det dere foreslår skal være binomisk fordelt, og diskutér det i gruppa.

iii) $X \sim Bin(n, p)$ kalles en sannsynlighetsmodell for eksemplet ditt.

Uten at du har undersøkt det, hva vil du gjette at verdien på p er?

iv) Hvordan kunne du samlet data for å få et estimat for p i eksemplet ditt?

i) One example would be a coin flipping where the probability to have one of the both possible independent outcomes (head or tail) are equal to p and $1-p$ respectively;

ii) Then X is a binomial independent variable which contains the sum of the results of n attempts, such that:

$$X = \sum_{i=1}^n x_i \text{ where } x_i = \begin{cases} 1 & \text{if head} \\ 0 & \text{if tail} \end{cases}$$

iii) In this case, the probability p should be 50% or 0,5.

iv) A estimator for $\hat{p} = \frac{\bar{X}}{n}$, then we could sum the results of the attempts (n) and take the average of its results (\bar{X}), then calculate the estimate \hat{p} .

B) i) Ta utgangspunkt i ditt eget (studie)liv og kom med minst et eksempel (ett eksempel hver i kollokviegruppene) på noe som kan tenkes å være normalfordelt. Begrunn hvorfor du/dere mener at normalfordelingen er en rimelig antakelse i eksemplene dere har valgt.

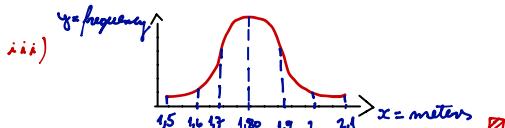
ii) $X \sim N(\mu, \sigma)$ kalles en sannsynlighetsmodell for eksemplet ditt.

Uten at du har undersøkt det, hva vil du gjette at verdien på μ og σ er?

iii) Bruk dette til å tegne en skisse av fordelingen(e) i eksemplet/ene du/dere har valgt.

i) One example would be the height of a population which the frequency should be a curve of normal distribution, because most of the people in a specific population are of average height and the number of taller and/or smaller people are less than the average. Also, a very short amount of people are very tall or small.

ii) For men, $\hat{\mu}$ would be $\bar{x} \approx 1,80$ and $\hat{\sigma} = S = \sqrt{\frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2} \approx 10$



- C) « $X_i \sim N(\mu, \sigma)$, X_i -ene er uavhengige» er måten vi bruker statistisk notasjon for å formulere en sannsynlighetsmodell for den situasjonen at du gjør n uavhengige målinger/observasjoner fra den opprinnelige fordelingen din.

Forklar hvorfor formuleringen

$$X_i \sim N(\mu, \sigma), \text{ og } X_i\text{-ene er uavhengige}$$

er det samme som å skrive

$$X_i = \mu + \varepsilon_i, \text{ der} \\ \varepsilon_i \sim N(0, \sigma), \text{ og } \varepsilon_i\text{-ene er uavhengige}$$

Dette er en øvelse som vil komme til nytte i resten av kurset.

↳ The independent variables (X_i) are approximately chosen according to the distribution $N(\mu, \sigma)$ values, this means that the normal distribution values are roughly equal to the real values when randomly chosen.

↳ $X_i \sim \text{norm}(\mu, \sigma) = X_i = \varepsilon_i + \mu$ where $\varepsilon_i \sim \text{norm}(0, \sigma)$. □

Oppgave 2

En forsker på et rehabiliteringssykehus er opptatt av at pasienter som trener seg opp etter ulykker og sykdommer (for eksempel unge folk som har pådratt seg ryggmargsskade i stupeulykker, og slagpasienter), skal få tilbake normal gangfunksjon. Det betyr at de skal bli i stand til å gå normalt igjen. Men hva betyr egentlig det, og hva er «normal gangfunksjon»?

En mulig måte å måle gangfunksjonen på er å la den som skal undersøkes gå i 6 minutter, og så måle hvor langt vedkommende har gått. Forskeren kaller dette «6 minutters gangtest», eller «6 minutes walking distance (6MWD)».

- A) Hva slags type data vil man få fra en 6MWD?

Hvor langt er det rimelig å anta at en frisk person vil gå på 6 minutter? Bruk det dere kan fra gymtimer, idrett og andre erfaringer til å diskuter i gruppa og foreslå en forventningsverdi og et standardavvik for slike målinger.

↳ The data from an 6MWD is the walking distance of each participant.

↳ I would guess that the walking average would be 500 meters every 6 minutes with standard deviation of 50 meters. □

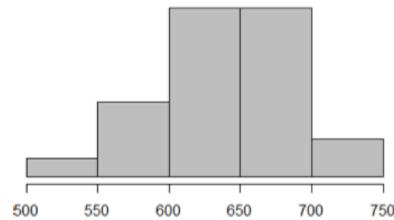
- B) For å få et sammenligningsgrunnlag for pasientene sine, vil forskeren først gjøre et prøveprosjekt, en såkalt pilotstudie. Han vurderer da å enten spørre sine nærmeste 25 kolleger (menn, 20-60 år, de fleste fysioterapeuter) eller å rekruttere 25 tilfeldige personer (menn, 20-60 år) fra nabologatet, og be dem om å gjennomføre en 25 minutters gangtest. Hvilken plan er best? Hvorfor?

↳ The second plan is the best because with a sample of people it is possible to have more accurate estimations than from opinions. □

- C) Uansett hva du/dere har svart i forrige oppgave, velger forskeren de 25 nabolagspersonene.

Etter at han har samlet data, får han følgende histogram:

Stemmer disse observasjonene med det dere ble enige om i A)? *Nope, but almost!*



Forskeren velger å oppsummere utvalget sitt med gjennomsnittet og standardavviket:

$$\bar{x} = 636, SD = 46$$

i) Er dette fornuftige oppsummeringstall? Hvorfor/hvorfor ikke?

ii) Hva betyr disse tallene i praksis? (Hint: Hvis du ser på variasjonen i hvor langt enkeltpersoner vil gå på 6 minutter, kan du finne et intervall som dekker ca 95% av de enkelte gangdistansene?)

- i) Yes, because the graph is quite symmetric, then $\bar{x}=636$ is in the top with $3 \times SD$ for both sides which covers the entire observations. This looks like a normal distribution with \bar{x}, SD ;
- ii) \bar{x} := the average walking distance of the sample outcomes;
 SD := the amount of dispersion of the sample outcomes;
95% coverage will be between $[636 - 2 \times 46 = 544; 636 + 2 \times 46 = 728]$ \square

- D) Forskeren er ikke bare interessert i å oppsummere de 25 personene i utvalget, men også å bruke dem til å si noe om populasjonen, nærmere bestemt forventet verdi for 6MWD i populasjonen.

Sett opp en sannsynlighetsmodell for 25 målinger fra en pilotstudie. Forklar hvilke parametere modellen inneholder, og hvilken parameter du er interessert i å estimere.

Gi et punktestimat for denne parameteren.

- \hookrightarrow $X_i \sim T_{m-1}(\bar{x}, SD)$ where μ and σ are unknown; $i \in \{1, \dots, m\}$ and $m = 25$;
 \hookrightarrow We want to estimate μ of the population with 95% of security;
 \hookrightarrow Unbiased estimator := $\bar{x} = E[\bar{x}] = 636$ and $SD = 46 \Rightarrow$ punktestimat of $\mu = \bar{x}$ \square
 \hookrightarrow $\bar{x} \sim T_{m-1}(\mu, SD/\sqrt{m})$
 \hookrightarrow $KI\ 95 = \left[\bar{x} \pm T_{24, 0.025} \cdot \frac{SD}{\sqrt{m}} \right] \rightarrow P\left(-T_{24, 0.025} \leq \frac{\bar{x} - \mu}{SD/\sqrt{m}} \leq T_{24, 0.025}\right) \rightarrow [617, 011; 654, 99] \square$

- E) Du kan anta at σ er kjent, og at $\sigma = 50$ m.

Beregn et 95% konfidensintervall for μ , basert på den sannsynlighetsmodellen du har formulert i D).

Hvordan tolker du dette intervallet, og hvordan er det forskjellig fra intervallet du fikk i C)?

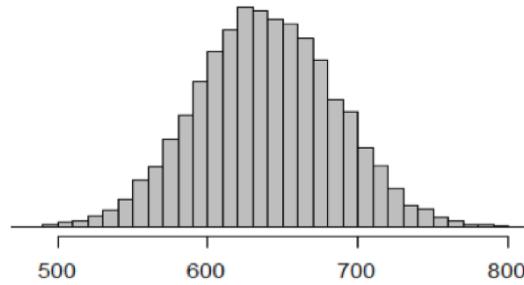
- \hookrightarrow because σ is known!
 \hookrightarrow $X \sim norm(\mu, \sigma^2/m)$ where μ is unknown and σ known; $m = 25$;
 \hookrightarrow $KI\ 95\% \text{ of } \mu = \left[636 - 2,97 \cdot \frac{50}{\sqrt{25}} = 606,3; 636 + 2,97 \cdot \frac{50}{\sqrt{25}} = 665,7 \right]$
 \hookrightarrow The μ of the population has 95% of security to be between $[606,3; 665,7]$;
 \hookrightarrow The 95% interval in 2c means where the random variables $X_i \sim norm(\bar{x}, SD)$ could be 95% of the time frequently. In other hand, The 95% interval here means the security of the estimate μ of the population could be. \square

- F) Forskningsrådet synes studien til forskeren er lovende, og bevilger penger til en større ommer omfattende studie, der forskeren kan rekruttere 10.000 personer til å gjennomføre 6MWD.

Forskeren får nå følgende histogram:

Han velger nok en gang å oppsummere utvalget sitt med gjennomsnittet og standardavviket, som nå er:

$$\bar{x} = 639, SD = 47$$



- i) Er dette fornuftige oppsummeringstall? Hvorfor/hvorfor ikke?
- ii) Hva betyr disse tallene i praksis? (Hint: Hvis du ser på variasjonen i hvor langt enkeltpersoner vil gå på 6 minutter, kan du nå finne et intervall som dekker ca 95% av de enkelte gangdistansene?)
- iii) Kommenter forskjellen/likheten til histogrammet i C) og histogrammet i F)
- iv) Kommenter forskjellen/liheten til oppsummeringstallene i C) og i F), og de intervallene du har regnet ut i C) og i D).
- v) Forklar hvorfor SD ikke blir mindre i F enn i C), selv om n er mye større her.
- i) Yes, because this graph shows a normal distribution (symmetrical), the \bar{x} is in the middle top and $3 \times SD$ covers 99,7% of the sample outcomes. //
- ii) \bar{x} := the average walking distance of the sample outcomes;
 SD := the amount of dispersion of the sample outcomes;
95% coverage will be between $[639 - 2 \times 47 = 545; 639 + 2 \times 47 = 733]$
- iii) In 2C we have a sample size with only $n=25$ observations/outcomes, then the histogram is not that symmetrical than in 2D where $n=10.000$. Samples with bigger n is better to have more accurate estimations and inferences. //
- iv) \bar{x} is smaller in C than in F. SD is smaller in C than in F. In F we have more n, then the parameters are more precise. //
- v) This happened because in F, the range of the outcomes is larger than in C. Then the quantification of the dispersion can be bigger. //

G) Gjenta D) og E) for denne situasjonen.

- D) \rightarrow because n is very large
 $\hookrightarrow X_i \sim \text{norm}(\bar{x}, SD)$ where μ and σ are unknown; $i \in \{1, \dots, n\}$ and $n = 10.000$;
 \hookrightarrow We want to estimate μ of the population with 95% of security;
 \hookrightarrow Unbiased estimator := $\bar{x} = E[\bar{x}] = 639$ and $SD = 47 \therefore$ pointestimate of $\mu = \bar{x}$ //
- $\hookrightarrow \bar{x} \sim \text{norm}(\mu, SD/\sqrt{n})$, because n is large;
 $\hookrightarrow KI 95\% = \left[\bar{x} \pm z_{0,025} \cdot \frac{SD}{\sqrt{n}} \right] \rightarrow P(-z_{0,025} \leq \frac{\bar{x} - \mu}{SD/\sqrt{n}} \leq z_{0,025})$
 $KI 95\% = [637,60; 640,3959] \blacksquare$

- E) $\sigma = 50 \therefore \bar{X} \sim \text{norm}(\mu, \sigma/\sqrt{n})$, $n = 10.000$;
 $\hookrightarrow KI 95\% \text{ of } \mu = \left[639 - 2,97 \cdot \frac{50}{\sqrt{10.000}} = 637,515; 639 + 2,97 \cdot \frac{50}{\sqrt{10.000}} = 640,485 \right]$
 \hookrightarrow The μ of the population has 95% of security. To be between $[637,51; 640,48]$;

H) i) Forklar hvorfor punktestimatet du fikk i D) og G) er ganske like.

ii) Forklar hvorfor konfidensintervallene du regnet ut i E) og G) er veldig forskjellige.

i) Because in both it was used the unbiased estimator of μ which is the average of the samples. As both samples look like a normal distribution, then the estimates are very near. //

ii) because the first sample has a very small size $n=25$, then less accurate than the second sample with size $n=1000$. //

I) Forklar forskjellen på intervallet du regnet ut i F) og konfidensintervallet du regnet ut i G).

\hookrightarrow The 95% interval in 2F means where the random variables $X_i \sim \text{norm}(\bar{x}, s_d)$, $i \in \{1, \dots, n=1000\}$ could be 95% of the time frequently. In other hand, the 95% interval in 2G means the uncertainty range of the estimate μ of the population could be. //

J) Anta så at σ ikke er kjent. Gjenta oppgave D), E) og G), altså beregn punktestimat og 95% konfidensintervall for μ , basert på pilotstudien, og punktestimat og 95% konfidensintervall for μ , basert på den store studien.

Sammenlign de fire estimatene og konfidensintervallene: Hva har størst påvirkning på bredden av konfidensintervallet her? Kjent eller ukjent σ , eller liten eller stor n ?

$$\hat{\bar{x}} = 636; s_d = 46; n = 25; \sigma \text{ unknown}$$

$$\hookrightarrow X_i \sim T_{n-1}(\mu, s_d) \text{ where } i \in \{1, \dots, n=25\}$$

$$\hookrightarrow \text{KI 95} = \left[\hat{\bar{x}} \pm T_{24; 0,025} \cdot \frac{s_d}{\sqrt{n}} \right] \rightarrow P\left(-T_{24; 0,025} \leq \frac{\hat{\bar{x}} - \mu}{s_d/\sqrt{n}} \leq T_{24; 0,025}\right) \rightarrow [617,011; 654,99] //$$

$$\hat{\bar{x}} = 639; s_d = 47; n = 1000; \sigma \text{ unknown}$$

$$\hookrightarrow X_i \sim T_{n-1}(\mu, s_d) \sim \text{norm}(\mu, s_d/\sqrt{n}) \text{ where } i \in \{1, \dots, n=1000\}$$

$$\hookrightarrow \text{KI 95} = \left[\hat{\bar{x}} \pm T_{999; 0,025} \cdot \frac{s_d}{\sqrt{n}} \right] \rightarrow P\left(-T_{999; 0,025} \leq \frac{\hat{\bar{x}} - \mu}{s_d/\sqrt{n}} \leq T_{999; 0,025}\right) \rightarrow [638,07; 639,92]$$

$$* T_{999; 0,025} \approx Z_{0,025} = 1,96 \text{ because } n \text{ is very large!}$$

* The size of the sample size is what influences most the length of the confidence interval. The parameter σ influences the length of the KI, but it is adjusted when we change from normal distribution to T-distribution. //

K) Hvor mange personer kunne forskeren nøyd seg med å undersøke hvis han ønsket et 95% konfidensintervall med en bredde på maks 20 meter?

$\hookrightarrow 4 \times 25 \text{ people is enough for a KI 95\% with length of 20m, such that:}$

$$\text{f} \rightarrow n \text{ long} \\ X_i \sim \text{norm}(\mu, s_d/\sqrt{n}), n = 100, \hat{\bar{x}} = 639; s_d = 47; \sigma \text{ unknown};$$

$$\hookrightarrow \text{KI 95} = 639 \pm 1,96 \cdot \frac{47}{\sqrt{100}} = [629,788; 648,21] //$$

Oppgave 3 (Hvis dårlig tid: Løs oppgave 5 først)

Sett opp en sannsynlighetsmodell for antallet personer i et tilfeldig utvalg på n personer, som sier de ville stemt MDG. Husk antakelsene.

$$X \sim \text{bin}(n, p) \text{ where } \begin{cases} n \text{ is the sample size;} \\ p \text{ is the estimate probability;} \\ X \text{ is the sum of successes answers (total of people who said that note for MDG)} \end{cases}$$

- i) Finn en av de nyligste meningsmålingene på nettet, og bruk observasjonene i denne til å gi et punktestimat for p , og et 95% konfidensintervall for p .
- ii) Hvor mange personer måtte vært med i meningsmålingen for at bredden på intervallet skulle vært maksimalt 2 prosentpoeng (1 prosentpoeng i hver retning fra punktestimatet), altså en maksimal bredde på 0.02?

$$\hat{p} = \frac{\bar{X}}{m} = \frac{41}{952} = 0,043 \quad \text{where } 41 \text{ people said that note for MDG; } 952 \text{ total of people interviewed.}$$

$$\text{KI } 95\% \text{ for } \hat{p} = \hat{p} \pm z_{0,025} \cdot \sqrt{\frac{p(1-p)}{m}} = 0,043 \pm 1,96 \cdot \sqrt{\frac{0,043 \cdot (1-0,043)}{952}} \approx [0,0301; 0,0558] //$$

$$\text{i)} 0,0558 = (0,043 + 0,01) + 1,96 \cdot \sqrt{\frac{0,043 \cdot (1-0,043)}{m}} \rightarrow 0,0558 = \frac{0,053}{\sqrt{m}} \cdot \sqrt{m} + \frac{0,3976}{\sqrt{m}}$$

$$0,0558 \cdot \sqrt{m} = 0,053 \cdot \sqrt{m} + 0,3976$$

$$0,0028 \cdot \sqrt{m} = 0,3976$$

$$(\sqrt{m})^2 = \left(\frac{0,3976}{0,0028}\right)^2 \Rightarrow m = 20164 \text{ people}$$

Oppgave 4

I landet Utopia var det 100% valgdeltakelse, og det seirende partiet «Merpartiet» fikk 52.3% oppslutning. Lederen i partiet feiret grundig, og ble sett sjanglende hjem sent på morgenkvisten, mens han kysset en mann som ikke var hans kone. En tabloidavis gnir seg i hendene og planlegger umiddelbart en meningsmåling for å undersøke om dette kan ha skadet omdømmet til Merpartiet.

- i) Hvilket spørsmål bør meningsmålingen stille?
- ii) Hvilken sannsynlighetsmodell bør du bruke her?
- iii) Sett opp en nullhypotese og en alternativ hypotese for å undersøke om oppslutningen til regjeringspartiet er redusert, sammenlignet med selve valgdagen.
- iv) Hvilket resultat må meningsmålingen gi for at du skal konkludere med «Behold H_0 », og hvilket resultat må den gi for at du skal konkludere med «Forkast H_0 »?

- i) The question would be if his behaviour will result in a reduction of his supporters;
- ii) $X \sim \text{bin}(n, p)$ where X is the total of supporters;
- iii) $H_0 := \text{not have reduction} \text{ vs } H_1 := \text{One-sided } p < 0,523;$
- iv) I would set a significant level $\alpha = 0,05$ and check if the probability to observe at least the same # of extreme observations as it has been before, given H_0 is true, is less than α . If $P(X \geq x | H_0) \leq \alpha$, then I would not discard H_0 , otherwise I would discard H_0 .

Oppgave 5

Anta at forskeren i oppgave 2 har undersøkt samtlige menn 20-60 år i Norge, så han vet at $\mu_{Norge} = 632m$.

En svensk kollega lurer på om forventet gangfunksjon for svenske menn i samme aldersgruppe, $\mu_{Sverige}$, er som i Norge.

Hvilken sannsynlighetsmodell bør du bruke her? ~~Normal distribution because μ parameter is interest.~~

Sett opp en nullhypotese og en alternativ hypotese for å undersøke om det er ulik gangfunksjon blant voksne menn i Norge og Sverige. ✓

Hvis det svenske konfidensintervallet inneholder verdien $\mu_{Sverige} = 632$, vil du da beholde eller forkaste H_0 ? (Viktig spørsmål!)

$H_0 :=$ The walking distance overall is not different between Norway and Sweden. ✗ $H_1 :=$ The walking distance overall is different between Norway and Sweden.

* Both hypothesis testing regards only men between 20-60 years.

L I would keep H_0 because it looks like that both μ are the same, but only using this information is not sufficient to have an accurate decision because we don't know anything about the extreme values of the μ s. □