



# NTNU

Det skapende universitet

## **Pensumsdiskusjonsforelesning** **Utvalgsfordelinger og estimatorer**

Geir-Arne Fuglstad

Uke 10

# Oppgave 1

Du jobber på et kasino og din jobb er å sjekke at det ikke foregår juks ved å observere spillene. I et spill har en terning landet på «6 øyne» i 200 av 1000 kast. Du mistenker at terningen har blitt modifisert og bestemmer deg for å bruke statistikk for å undersøke.

# Oppgave 1

Du jobber på et kasino og din jobb er å sjekke at det ikke foregår juks ved å observere spillene. I et spill har en terning landet på «6 øyne» i 200 av 1000 kast. Du mistenker at terningen har blitt modifisert og bestemmer deg for å bruke statistikk for å undersøke.

- a) Forklar hvorfor kast av en rettferdig terning med utfall 1 («6 øyne») og 0 («ikke 6 øyne») kan sees som å trekke fra en bernoullifordeling med suksessansynlighet  $p = 1/6$ .

# Oppgave 1

Du jobber på et kasino og din jobb er å sjekke at det ikke foregår juks ved å observere spillene. I et spill har en terning landet på «6 øyne» i 200 av 1000 kast. Du mistenker at terningen har blitt modifisert og bestemmer deg for å bruke statistikk for å undersøke.

- b) Anta terningen er rettferdig. Da er sannsynligheten for 200 eller flere kast med «6 øyne» av 1000 kast lik 0.0029. Svekker eller styrker dette dine mistanker om juks?

# Oppgave 1

Du jobber på et kasino og din jobb er å sjekke at det ikke foregår juks ved å observere spillene. I et spill har en terning landet på «6 øyne» i 200 av 1000 kast. Du mistenker at terningen har blitt modifisert og bestemmer deg for å bruke statistikk for å undersøke.

- c) Du anser  $p = P(\text{«6 øyne»})$  som ukjent. La  $X_i$  være utfall (1/0) av kast  $i$ . Du velger estimatoren  $\hat{P} = \frac{1}{1000} \sum_{i=1}^{1000} X_i$  for  $p$ . Finn forventningsverdi, varians, utvalgsfordeling og estimat for  $\hat{P}$ .

## Oppgave 2

To epidemiologer undersøker andel personer i Norge som har en spesifikk sykdom. Forsker 1 har basert på et tilfeldig utvalg kommet fram til en estimator  $\hat{P}_1$ , og forsker 2 har basert på et annet tilfeldig utvalg laget estimatoren  $\hat{P}_2$ . La  $p$  betegne virkelig andel, og anta at  $\hat{P}_1$  og  $\hat{P}_2$  er forventningsrette for  $p$ , og at  $\text{Var}[\hat{P}_1] = p(1 - p)/10$  og  $\text{Var}[\hat{P}_2] = p(1 - p)/2$ .

## Oppgave 2

To epidemiologer undersøker andel personer i Norge som har en spesifikk sykdom. Forsker 1 har basert på et tilfeldig utvalg kommet fram til en estimator  $\hat{P}_1$ , og forsker 2 har basert på et annet tilfeldig utvalg laget estimatoren  $\hat{P}_2$ . La  $p$  betegne virkelig andel, og anta at  $\hat{P}_1$  og  $\hat{P}_2$  er forventningsrette for  $p$ , og at  $\text{Var}[\hat{P}_1] = p(1 - p)/10$  og  $\text{Var}[\hat{P}_2] = p(1 - p)/2$ .

- a) Du ønsker å gjøre en metaanalyse der du kombinerer de to studiene for å oppnå en bedre estimator. Du velger  $\hat{P} = (\hat{P}_1 + \hat{P}_2)/2$ . Er  $\hat{P}$  bedre enn  $\hat{P}_1$  og  $\hat{P}_2$ ?

## Oppgave 2

To epidemiologer undersøker andel personer i Norge som har en spesifikk sykdom. Forsker 1 har basert på et tilfeldig utvalg kommet fram til en estimator  $\hat{P}_1$ , og forsker 2 har basert på et annet tilfeldig utvalg laget estimatoren  $\hat{P}_2$ . La  $p$  betegne virkelig andel, og anta at  $\hat{P}_1$  og  $\hat{P}_2$  er forventningsrette for  $p$ , og at  $\text{Var}[\hat{P}_1] = p(1 - p)/10$  og  $\text{Var}[\hat{P}_2] = p(1 - p)/2$ .

- Du ønsker å gjøre en metaanalyse der du kombinerer de to studiene for å oppnå en bedre estimator. Du velger  $\hat{P} = (\hat{P}_1 + \hat{P}_2)/2$ . Er  $\hat{P}$  bedre enn  $\hat{P}_1$  og  $\hat{P}_2$ ?
- Forklar hvorfor  $\hat{P}$  blir en dårligere estimator enn  $\hat{P}_1$  og hvordan du kan gå frem for å lage en bedre estimator enn  $\hat{P}_1$ .



## Oppgave 2

To epidemiologer undersøker andel personer i Norge som har en spesifikk sykdom. Forsker 1 har basert på et tilfeldig utvalg kommet fram til en estimator  $\hat{P}_1$ , og forsker 2 har basert på et annet tilfeldig utvalg laget estimatoren  $\hat{P}_2$ . La  $p$  betegne virkelig andel, og anta at  $\hat{P}_1$  og  $\hat{P}_2$  er forventningsrette for  $p$ , og at  $\text{Var}[\hat{P}_1] = p(1 - p)/10$  og  $\text{Var}[\hat{P}_2] = p(1 - p)/2$ .

c) Finn den beste estimatoren  $\hat{P} = a\hat{P}_1 + b\hat{P}_2$  der  $a, b \in \mathbb{R}$ .

## Oppgave 2

To epidemiologer undersøker andel personer i Norge som har en spesifikk sykdom. Forsker 1 har basert på et tilfeldig utvalg kommet fram til en estimator  $\hat{P}_1$ , og forsker 2 har basert på et annet tilfeldig utvalg laget estimatoren  $\hat{P}_2$ . La  $p$  betegne virkelig andel, og anta at  $\hat{P}_1$  og  $\hat{P}_2$  er forventningsrette for  $p$ , og at  $\text{Var}[\hat{P}_1] = p(1 - p)/10$  og  $\text{Var}[\hat{P}_2] = p(1 - p)/2$ .

**d)** Det viser seg at  $\hat{P}_1$  og  $\hat{P}_2$  er korrelerte fordi forskerene brukte delvis samme datakilde. Finn beste estimator  $\hat{P} = a\hat{P}_1 + (1 - a)\hat{P}_2$  for  $a \in \mathbb{R}$  når

$$\text{Var}[\hat{P}_1] = 0.009$$

$$\text{Var}[\hat{P}_2] = 0.045$$

$$\text{Cov}[\hat{P}_1, \hat{P}_2] = 0.01.$$

## Oppgave 3

En maskin produserer 200 g sjokoladeplater. På grunn av tilfeldig variasjon varierer vekten mellom sjokoladeplater. Anta vektene er uavhengige. Vi ønsker å estimere «gjennomsnittlig» vekt. Vi observerer vekter 199, 200.1, 205, 198 og 199.5. Anta at variansen til vektene er 1.

## Oppgave 3

En maskin produserer 200 g sjokoladeplater. På grunn av tilfeldig variasjon varierer vekten mellom sjokoladeplater. Anta vektene er uavhengige. Vi ønsker å estimere «gjennomsnittlig» vekt. Vi observerer vekter 199, 200.1, 205, 198 og 199.5. Anta at variansen til vektene er 1.

- Formuler problemet på en statistisk presis måte.

## Oppgave 3

En maskin produserer 200 g sjokoladeplater. På grunn av tilfeldig variasjon varierer vekten mellom sjokoladeplater. Anta vektene er uavhengige. Vi ønsker å estimere «gjennomsnittlig» vekt. Vi observerer vekter 199, 200.1, 205, 198 og 199.5. Anta at variansen til vektene er 1.

- b) Foreslå en estimator, vis at den er forventningsrett, og regn ut dens varians. Hva blir estimatet?

## Oppgave 3

En maskin produserer 200 g sjokoladeplater. På grunn av tilfeldig variasjon varierer vekten mellom sjokoladeplater. Anta vektene er uavhengige. Vi ønsker å estimere «gjennomsnittlig» vekt. Vi observerer vekter 199, 200.1, 205, 198 og 199.5. Anta at variansen til vektene er 1.

- c) Vi måler 30 plater og finner  $\bar{x} = 201$ . Anta at sann forventningsverdi er 200 og regn ut sannsynligheten for at  $\bar{X} > 201$ .

## Oppgave 3

En maskin produserer 200 g sjokoladeplater. På grunn av tilfeldig variasjon varierer vekten mellom sjokoladeplater. Anta vektene er uavhengige. Vi ønsker å estimere «gjennomsnittlig» vekt. Vi observerer vekter 199, 200.1, 205, 198 og 199.5. Anta at variansen til vektene er 1.

- d) Variansen er ukjent. Du observerer for 30 plater,  $\bar{x} = 201$  og  $\sum_{i=1}^{30} x_i^2 = 1212058$ . Velg en forventningsrett estimator for variansen og regn ut estimert varians.

## Oppgave 4

Anta  $X_1, X_2, \dots, X_n$  er uavhengige og identiske normalfordelte med forventningsverdi  $\mu$  og varians  $\sigma^2$ . Vi har to forventningsrette estimatorer for  $\sigma^2$ ,

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Hvilken estimator er best for  $\sigma^2$ ? Hvilken av de er mest nyttige i praksis?



# Noen viktige poeng

- Hvorfor var selve tallverdien til estimatet en så liten del av oppgavene vi gjorde?

# Noen viktige poeng

- Hvorfor var selve tallverdien til estimatet en så liten del av oppgavene vi gjorde?
- Vi vet aldri sannheten så all tillitt til estimatene oppstår gjennom sannynlighetsfordelingene til estimatorene.

# Noen viktige poeng

- Hvorfor var selve tallverdien til estimatet en så liten del av oppgavene vi gjorde?
- Vi vet aldri sannheten så all tillitt til estimatene oppstår gjennom sannynlighetsfordelingene til estimatorene.
- Hvis vi ikke kjenner estimatoren eller dens fordeling (for eksempel «black box»-metoder) er det veldig utfordrende å evaluere egenskapene til metoden.