STK1100 våren 2023

Konfidensintervaller

Svarer til avsnitt 8.1 i læreboka

Matematisk institutt
Universitetet i Oslo

Eksempel 1

En kjemiker er interessert i å bestemme konsentrasjonen μ av et stoff i en løsning

Hun måler konsentrasjonen fem ganger med et bestemt apparat og får resultatene (i mg per liter):

2.52 2.49 2.62 2.45 2.56

Vi antar målingene hun får med apparatet er normalfordelte med forventningsverdi lik den sanne konsentrasjonen μ (så det er ingen systematiske målefeil) og standardavvik σ = 0.10

Et punktestimat for μ er $\bar{x} = 2.53$

Kan vi gi et intervall som med stor sikkerhet inneholder den virkelige konsentrasjonen?

Generelt har vi følgende situasjon:

Observasjonene $x_1, x_2, \ldots, x_n$ er observerte verdier av stokastiske variabler $X_1, X_2, \ldots, X_n$ som er uavhengige og $N(\mu, \sigma^2)$-fordelte, der σ er kjent

Vi vil bestemme et intervall som med stor sikkerhet inneholder μ

Vi har at $\bar{X} \sim N(\mu, \sigma^2 / n)$

Det følger at

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

Vi har at

$$P(-1.96 \le Z \le 1.96) = 0.95$$

dvs

$$P\left(-1.96 \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) = 0.95$$

Ved å omforme ulikhetene gir dette at

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Så det stokastiske intervallet

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\quad \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

inneholder μ med 95% sannsynlighet

Når vi setter inn de observerte verdiene $x_1, x_2, \ldots\ldots, x_n$ av de stokastiske variablene $X_1, X_2, \ldots\ldots, X_n$ får vi intervallet

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \ , \quad \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

Dette er et 95% konfidensintervall for μ

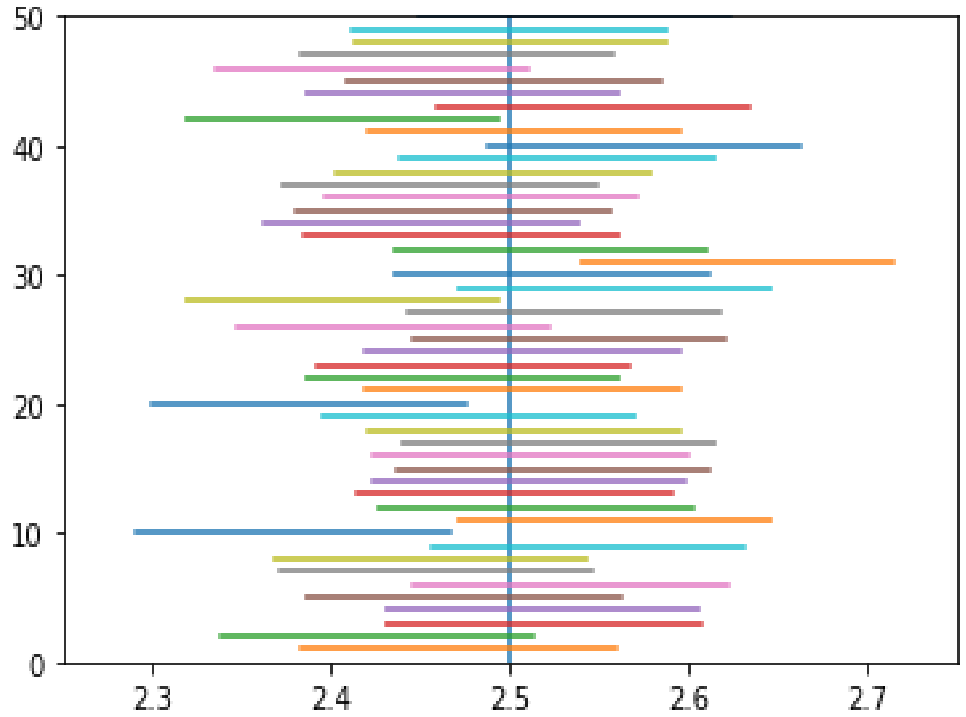
I eksempel 1 får vi intervallet

$$2.53 \pm 1.96 \frac{0.10}{\sqrt{5}} \qquad \text{dvs} \qquad 2.53 \pm 0.09$$

Vi kan si at 0.09 er en feilmargin for estimatet

Simulering av 50 konfidensintervaller

```
import numpy as np
import scipy.stats as stats
import math
import matplotlib.pyplot as plt
m=2.50
s=0.10
n=5
plt.xlim([m-0.25,m+0.25])
plt.ylim([0,50])
plt.plot([m,m],[0,50])
ind=np.arange(1,51)
for k in ind:
    x=stats.norm.rvs(size=n,loc=m,scale=s)
    low=np.mean(x)-1.96*s/np.sqrt(n)
    up=np.mean(x)+1.96*s/np.sqrt(n)
    plt.plot([low,up],[k,k])
```



Antall intervaller som inneholder den sanne verdien av μ er binomisk fordelt med $p = 0.95$

Bredden av konfidensintervallet

Konfidensintervallet til kjemikeren var $2.53 \pm 1.96 \frac{0.10}{\sqrt{5}}$

Bredden av intervallet er $2 \cdot 1.96 \frac{0.10}{\sqrt{5}} = 0.175$

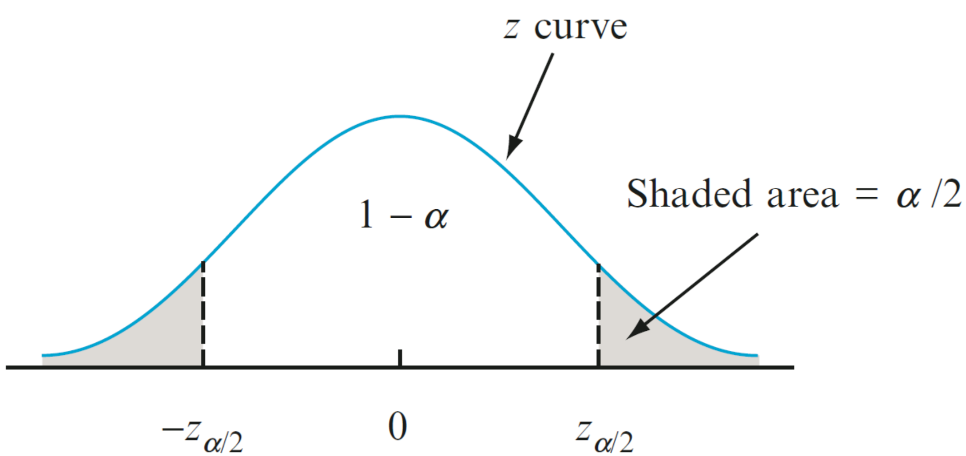
Hvor mange målinger må kjemikeren gjøre hvis hun ønsker at konfidensintervallet skal ha bredde 0.12?

Med *n* observasjoner blir bredden $2 \cdot 1.96 \frac{0.10}{\sqrt{n}}$

Altså $2 \cdot 1.96 \frac{0.10}{\sqrt{n}} \leq 0.12$

Det gir $n \geq \left(\frac{2 \cdot 1.96 \cdot 0.10}{0.12}\right)^2 = 10.67$, dvs 11 målinger

Det er vanlig å se på konfidensintervall med konfidenskoeffisient 90%, 95% eller 99%



Generelt er et $100(1-\alpha)\%$ konfidensintervall for μ gitt ved

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \;,\; \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right)$$

eller kortversjonen $\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$

Eksempel 2

En annen kjemiker er interessert i å bestemme presisjonen til et apparat

Han bruker en standardløsning med kjent konsentrasjon $\mu = 4.00$ og måler konsentrasjonen av denne seks ganger med apparatet og får resultatene

3.86 4.17 3.77 3.99 4.04 4.10

Vi antar som før at måleresultatene er normalfordelte, men nå er det σ som er ukjent mens μ er kjent

Kan vi gi et estimat for σ?

Kan vi gi et intervall som med stor sikkerhet inneholder den virkelige presisjonen σ?

Generelt har vi følgende situasjon:

Observasjonene $x_1, x_2, ..., x_n$ er observerte verdier av stokastiske variabler $X_1, X_2, ..., X_n$ som er uavhengige og $N(\mu, \sigma^2)$-fordelte, der μ er kjent

Vi har at $E\{(X_i - \mu)^2\} = V(X) = \sigma^2$

En forventningsrett estimator for σ^2 er

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

Estimator for σ:

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

I eksempel 2 får vi estimatet $\hat{\sigma} = 0.137$

For å finne et konfidensintervall, merker vi oss først at

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$$

Det følger at $\quad Z_i^2 = \frac{\left(X_i - \mu\right)^2}{\sigma^2}$ er ki-kvadratfordelt

med 1 frihetsgrad, dvs. gamma-fordelt med $\alpha = 1/2$ og $\beta = 2$ (se side 261)

Hvis *U* og *V* er uavhengige og

$$U \sim \text{gamma}(\alpha_1, \beta) \qquad V \sim \text{gamma}(\alpha_2, \beta)$$

så er

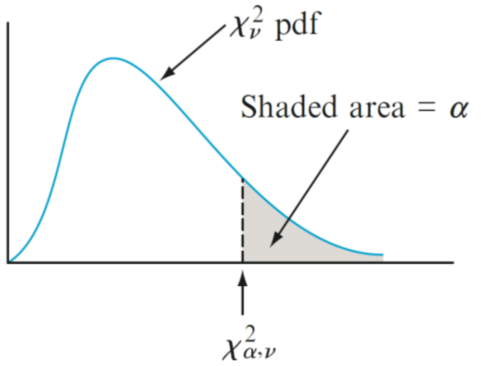
$$U + V \sim \text{gamma}(\alpha_1 + \alpha_2, \beta)$$

Det følger at

$$\sum_{i=1}^{n} Z_i^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \text{gamma}(n/2, 2)$$

dvs. ki-kvadratfordelt med n frihetsgrader

La $\chi^2_{\alpha,\nu}$ være $100(1-\alpha)\%$-persentilen i kji-kvadratfordelingen med ν frihetsgrader



Da har vi at

$$P\left(\chi^2_{0.975,n} \le \frac{n\hat{\sigma}^2}{\sigma^2} \le \chi^2_{0.025,n}\right) = 0.95$$

Det gir

$$P\left(\frac{n\hat{\sigma}^2}{\chi^2_{0.025,n}} \le \sigma^2 \le \frac{n\hat{\sigma}^2}{\chi^2_{0.975,n}} \right) = 0.95$$

$$P\left(\sqrt{\frac{n}{\chi^2_{0.025,n}}}\;\hat{\sigma} \le \sigma \le \sqrt{\frac{n}{\chi^2_{0.975,n}}}\;\hat{\sigma} \right) = 0.95$$

95% konfidensinterval

$$\left[\sqrt{\frac{n}{\chi^2_{0.025,n}}}\;\hat{\sigma} \;,\; \sqrt{\frac{n}{\chi^2_{0.975,n}}}\;\hat{\sigma} \right]$$

I eksempel 2 får vi:

$$\left[\sqrt{\frac{6}{14.45}}\; 0.137 \;,\; \sqrt{\frac{6}{1.237}}\; 0.137 \right] = [\, 0.088 \,,\; 0.302 \,]$$

Generell framgangsmåte for konfidensintervall

Vi har stokastiske variabler $X_1, X_2, \ldots\ldots, X_n$
Vi vil bestemme et 100(1-α)% konfidensintervall
for en parameter θ

Anta at vi har en observator $h(X_1, X_2, \ldots\ldots, X_n, \theta)$
og at fordelingen til observatoren ikke avhenger
av θ eller noen andre parametere

Hvis vi lar *a* og *b* være henholdsvis
$100 \cdot (\alpha/2)\%$-persentilen og
$100 \cdot (1-\alpha/2)\%$-persentilen i fordelingen til
$h(X_1, X_2, \ldots\ldots, X_n, \theta)$, så har vi at

$$P\left(a \leq h(X_1, X_2, \ldots\ldots, X_n, \theta) \leq b\right) = 1 - \alpha$$

Anta nå at vi kan omforme ulikhetene:

$$a \le h(X_1, X_2, \dots\dots, X_n, \theta) \le b$$

$$\Updownarrow$$

$$L(X_1, X_2, \dots\dots, X_n) \le \theta \le U(X_1, X_2, \dots\dots, X_n)$$

Da har vi at:

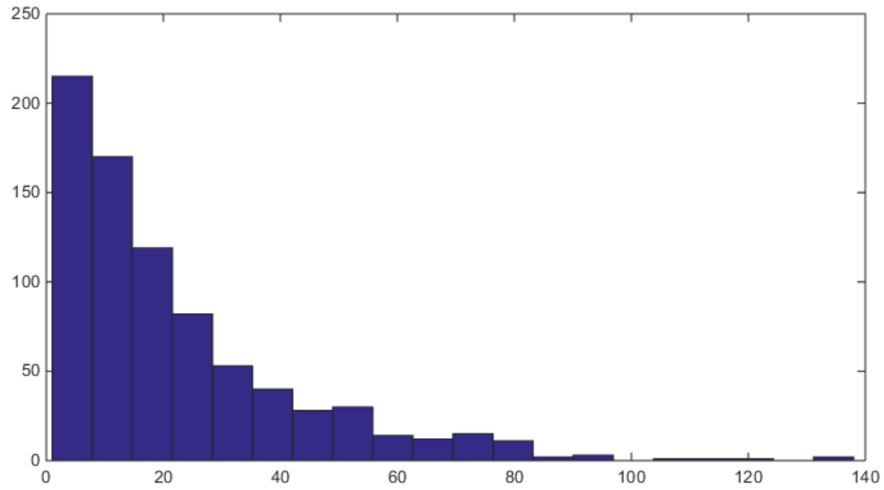
$$P\Big(L(X_1, X_2, \dots\dots, X_n) \le \theta \le U(X_1, X_2, \dots\dots, X_n)\Big) = 1 - \alpha$$

Når vi setter inn de observerte verdiene $x_1, x_2, \dots\dots, x_n$
av de stokastiske variablene $X_1, X_2, \dots\dots, X_n$,
får vi et 100(1-α)% konfidensintervall for θ:

$$\Big[L(x_1, x_2, \dots\dots, x_n)\ ,\ U(x_1, x_2, \dots\dots, x_n)\Big]$$

Eksempel:

En har målt tiden mellom 800 impulser i en nervefiber (http://www.statsci.org/data/general/nerve.html)
Vi kan modellere nerveimpulsene som en Poissonprosess (jf. sidene 160-161). Tiden mellom to impulser er da ekspontialfordelt (jf. Side 235)



$\bar{x} = 21.9$ (enhet 1/100 sekund)

Vi vil finne et 95% konfidensintervall for forventet tid mellom to nerveimpulser

Generelt har vi følgende situasjon:

Observasjonene $x_1, x_2, \ldots\ldots, x_n$ er observerte verdier av stokastiske variabler $X_1, X_2, \ldots\ldots, X_n$ som er uavhengige og eksponentialfordelt, dvs de har tettheten (jf. side 198)

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

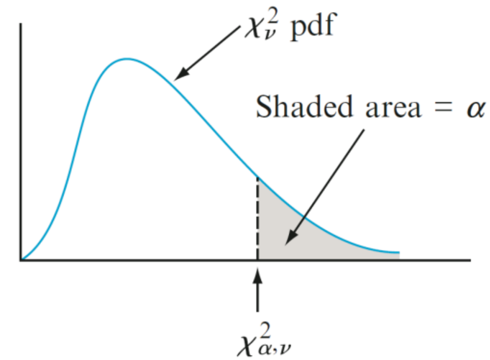
Vi vil finne et konfidensintervall for forventningsverdien $\mu = 1/\lambda$

Vi finner først et konfidensintervall for λ

Vi kan vise at $Y_i = 2\lambda X_i$ er kji-kvadratfordelt med 2 frihetsgrader, dvs gamma-fordelt med $\alpha = 2/2 = 1$ og $\beta = 2$ (detaljer på forelesningen)

Det følger at $$\sum_{i=1}^{n} Y_i = 2\lambda \sum_{i=1}^{n} X_i = 2\lambda n\bar{X}$$
er kji-kvadratfordelt med $2n$ frihetsgrader, dvs. gamma-fordelt med $\alpha = 2n/2 = n$ og $\beta = 2$

La $\chi^2_{\alpha,\nu}$ være $100(1-\alpha)\%$-persentilen i ki-kvadratfordelingen med ν frihetsgrader



Da har vi at

$$P\left(\chi^2_{0.975,2n} \le 2\lambda n\bar{X} \le \chi^2_{0.025,2n}\right) = 0.95$$

Det gir

$$P\left(\frac{\chi^2_{0.975,2n}}{2n\bar{X}} \le \lambda \le \frac{\chi^2_{0.025,2n}}{2n\bar{X}}\right) = 0.95$$

Av dette får vi et 95% konfidensintervall for λ

Forventningsverdien er gitt ved $\mu = 1/\lambda$

Derfor er $\lambda = 1/\mu$ og vi har at

$$P\left(\frac{\chi^2_{0.975,2n}}{2n\bar{X}} \le \frac{1}{\mu} \le \frac{\chi^2_{0.025,2n}}{2n\bar{X}}\right) = 0.95$$

Dette gir

$$P\left(\frac{2n\bar{X}}{\chi^2_{0.025,2n}} \le \mu \le \frac{2n\bar{X}}{\chi^2_{0.975,2n}}\right) = 0.95$$

Et 95% konfidensintervall for forventningsverdien er

$$\left[\frac{2n\bar{x}}{\chi^2_{0.025,2n}}\ ,\ \frac{2n\bar{x}}{\chi^2_{0.975,2n}}\right]$$

For nerveimpulsene har vi at $n = 799$ og $\bar{x} = 21.9$

Fra Python finner vi at

$c^2_{0.975,1598} = 1489.1$ kommando: stats.chi2.ppf(0.025,1598)

$c^2_{0.025,1598} = 1710.7$ kommando: stats.chi2.ppf(0.975,1598)

Et 95% konfidensintervall for forventet tid mellom to nerveimpulser er (målt i 1/100 sekunder)

$$\left[\frac{2 \cdot 799 \cdot 21.9}{1710.7} \; , \; \frac{2 \cdot 799 \cdot 21.9}{1489.1} \right]$$

dvs

$$\left[20.5 \; , \; 23.5 \right]$$