STK1100 våren 2023

Observatorer og deres fordeling
Sentralgrensesetningen
Lineærkombinasjoner

Svarer til avsnittene
6.1, 6.2 og 5.3 i læreboka

Matematisk institutt
Universitetet i Oslo

Illustrasjon (eksempel 6.1 i boken)

Vi måler bruddstyrken $x_1, x_2, ..., x_{10}$ til 10 prøver

Hvis vi måler 10 nye prøver, vil vi ikke få de samme verdiene av $x_1, x_2, ..., x_n$

En modell for forsøket er at $x_1, x_2, ..., x_{10}$ er observerte verdier av uavhengige stokastiske variabler $X_1, X_2, ..., X_n$ som er Weibull-fordelt med formparameter $\alpha = 2$ og skalaparameter $\beta = 5$

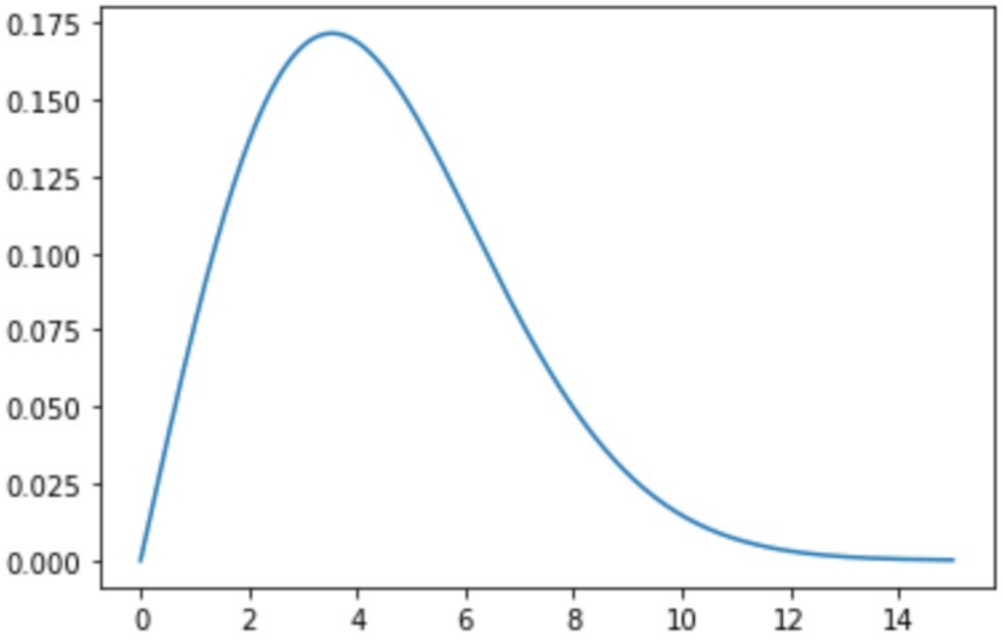
$$f(x;\alpha,\beta) = \frac{\alpha}{\beta^{\alpha}}\, x^{\alpha-1} e^{-(x/\beta)^{\alpha}}$$

Det kan vises (avsnitt 4.5, ikke pensum!)

$$\mu = E(X) = \beta\,\Gamma(1+1/\alpha) = 4.43$$

$$\tilde{\mu} = \eta(0.50) = \beta\,(\ln 2)^{1/\alpha} = 4.16$$

Weibull-fordeling med $\alpha = 2$ $\beta = 5$




```
(4.4311346272637895, 4.162773055788488)
```

Se full kode i Python i Jupyter notebook kapittel_6.ipynb

Vi gjentar forsøket seks ganger og regner ut gjennomsnitt og empirisk median for hvert forsøk

2.2633	6.7963	3.2480	2.9499	4.5383	11.5193
1.5728	0.8638	9.1273	9.2834	4.9079	5.2395
7.1827	1.0462	2.0220	5.6657	2.5846	0.3700
1.5051	4.2510	1.3066	8.7684	2.3935	6.4579
3.3849	2.3600	3.1126	7.6349	6.4756	2.6502
7.6281	6.9870	2.6335	2.2037	4.2244	2.7666
5.6532	4.6457	2.7243	3.0170	4.4955	3.5467
3.8843	1.4835	4.8371	5.3585	3.3033	1.4917
1.0419	2.4132	3.2496	1.1298	2.9300	3.4159
0.9453	1.0167	6.6427	9.1765	2.6526	4.4309

meanX= mean(X)

3.5062	3.1863	3.8904	5.5188	3.8506	4.1888

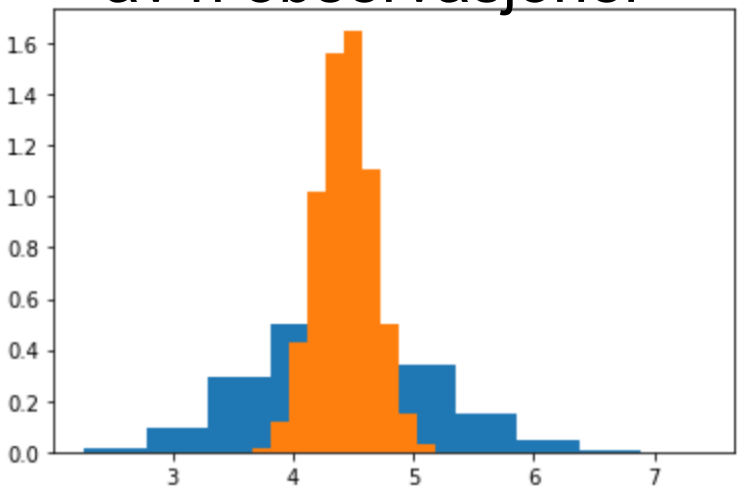
medianX=median(X)

2.8241	2.3866	3.1803	5.5121	3.7639	3.4813

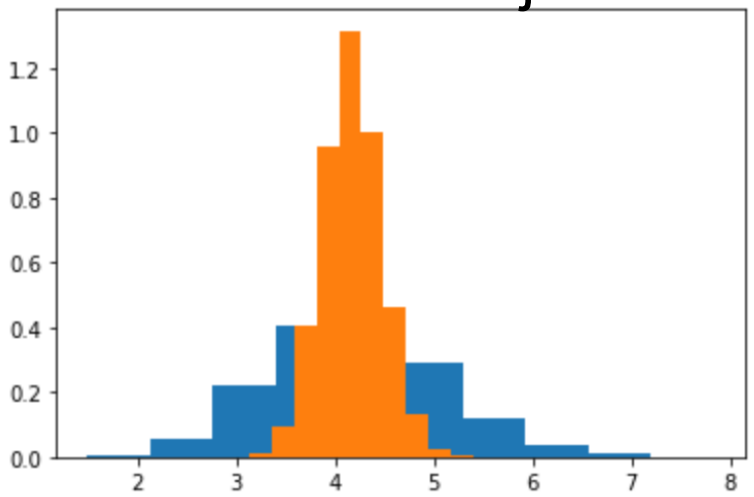
Gjennomsnitt og empirisk median vil variere fra forsøk, de er også stokastiske variabler

Vi kan studere fordelingene deres ved å gjenta forsøket veldig mange ganger (f.eks. 10 000 ganger), for hvert forsøk bestemme gjennomsnitt og empirisk median, og så tegne et histogram av de observerte verdiene

10000 gjennomsnitt av n observasjoner



10000 medianer av n observasjoner



Blå: n=10 Orange: n=100

Tilfeldig utvalg og observatorer

De stokastiske variablene $X_1, X_2, ..., X_n$ utgjør et tilfeldig utvalg hvis

$X_1, X_2, ..., X_n$ er uavhengige

alle X_i-ene har samme fordeling

En observator Y (engelsk: statistic) er en funksjon av utvalget, dvs $Y = h(X_1, X_2, ..., X_n)$

Fordelingen til Y kalles samplingsfordelingen til Y

Fordelingen til gjennomsnittet

Hvis $X_1, X_2, ..., X_n$ er et tilfeldig utvalg fra $N(\mu, \sigma^2)$-fordelingen, så er

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$$

$N(\mu, \sigma^2 / n)$-fordelt

Vi kan vise resultatet ved å bruke momentgenererende funksjoner (jf. forelesningen)

Husk at $M_{X_i}(t) = \exp\left\{\mu t + \sigma^2 t^2 / 2\right\}$

Hvis vi standardiserer, har vi at

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

er standardnormalfordelt

Vi har altså at

$$P\left(\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z \right) = P(Z \leq z) = \Phi(z)$$

der

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \, du$$

Sentralgrensesetningen

Sentralgrensesetningen sier at resultatet på forrige slide holder som en tilnærming for alle fordelinger (så sant variansen eksisterer):

Hvis $X_1, X_2, ..., X_n$ er et tilfeldig utvalg fra en fordeling med forventning μ og standardavvik σ, så har vi at

$$\lim_{n\to\infty} P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z\right) = P(Z \le z) = \Phi(z)$$

Hvis n er tilstrekkelig stor (ofte er $n > 30$ tilstrekkelig), er $\bar{X}$ tilnærmet $N(\mu, \sigma^2/n)$-fordelt

Illustrasjon ved simulering

Vil simulere $X_1, X_2, ..., X_n$ fra ulike fordelinger og sammenligne den empiriske kumulative fordelingen til $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ med den kumulative standardnormalfordelingen

- Uniform(0,1)

  $\mu = 1/2 \quad \sigma = 1/\sqrt{12}$

- Eksp(1)

  $\mu = 1 \quad \sigma = 1$

- Bernoulli med $p = 0.25$

  $\mu = 0.25 \quad \sigma = \sqrt{0.25 \cdot 0.75}$

Sjekk kapittel_6.ipynb for kode!

Forventning og varians til lineær-kombinasjoner av uavhengige variabler

Anta at $X_1, X_2, ..., X_n$ er uavhengige stokastiske variable. Da er

- $E(a_1X_1 + ... + a_nX_n) = a_1E(X_1) + ... + a_nE(X_n)$
- $V(a_1X_1 + ... + a_nX_n) = a_1^2V(X_1) + ... + a_n^2V(X_n)$

Hvis spesielt $X_1, X_2, ..., X_n$ er et tilfeldig utvalg fra en fordeling med forventning μ og standardavvik σ, så har vi at

- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \sigma^2 / n$

Chebyshevs ulikhet og store talls lov

For en stokastisk variable Y med forventning μ og standardavvik σ har vi at

$$P(|Y-\mu| \geq k\sigma) \leq 1/k^2$$

Hvis $X_1, X_2, ..., X_n$ er et tilfeldig utvalg fra en fordeling med forventning μ og standardavvik σ, så har vi for $\varepsilon > 0$

$$P(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Det følger at (store talls lov)

$$P(|\bar{X}-\mu| < \varepsilon) \to 1 \quad \text{når } n \to \infty$$

Forventning og varians til lineær-kombinasjoner av stokastiske variabler

La $X_1, X_2, ..., X_n$ være stokastiske variabler

Vi ser på lineærkombinasjonen $a_1X_1 + ...+ a_nX_n$

Vi har generelt at (forelesning + s. 303-305 bok)

- $E(a_1X_1 + ... + a_nX_n) = a_1E(X_1) + ... + a_nE(X_n)$
- $$V(a_1X_1 + ... + a_nX_n) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(X_i, X_j)$$
$$= \sum_{i=1}^{n} a_i^2 V(X_i) + 2\sum\sum_{i<j} a_i a_j Cov(X_i, X_j)$$

Hvis $X_1, X_2, ..., X_n$ er uavhengige har vi at

- $V(a_1X_1 + ... + a_nX_n) = a_1^2V(X_1) + ... + a_n^2V(X_n)$

Fordeling til lineærkombinasjoner og summer av stokastiske variabler

I noen viktige situasjoner kan vi bruke momentgenererende funksjoner til å finne fordelingen til lineærkombinasjoner og summer av stokastiske variabler (detaljer på forelesningen)

Normalfordelte variabler

Anta at $X_1, X_2, ..., X_n$ er uavhengige og $X_i \sim N(\mu_i, \sigma_i^2)$

Ved å bruke at $M_{X_i}(t) = \exp\left\{\mu_i t + \sigma_i^2 t^2 / 2\right\}$ får vi at

$$\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i \ , \ \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)$$

Poissonfordelte variabler

Anta at $X_1, ..., X_n$ er uavhengige og $X_i \sim \text{Poisson}(\lambda_i)$

Ved å bruke at $M_{X_i}(t) = \exp\left\{\lambda_i (e^t - 1)\right\}$ får vi at

$$\sum_{i=1}^{n} X_i \sim \text{Poisson}\left(\sum_{i=1}^{n} \lambda_i \right)$$

Gammafordelte variabler med samme skalaparameter

Anta at $X_1, ..., X_n$ er uavhengige og $X_i \sim \text{gamma}(\alpha_i, \beta)$

Ved å bruke at $M_{X_i}(t) = 1/(1-\beta t)^{\alpha_i}$ får vi at

$$\sum_{i=1}^{n} X_i \sim \text{gamma}\left(\sum_{i=1}^{n} \alpha_i \ , \ \beta \right)$$