Myntkast og binomialfordelingen

Halvor Aarnes, UiO, 2014

Innhold

Myntkast

Femtiøringen er gått ut av bruk, men er praktisk til eksperimenter med myntkast. Mynt: Revers(bakside) fabeldyr fra Urnes stavkirke. Kron: Advers (forside)

Utfallsrommet for et myntkast er

$$S = \{k,\ \ m\} \qquad k = kron, \quad m = mynt$$

Utfallsrommet ved kast av to mynter er:

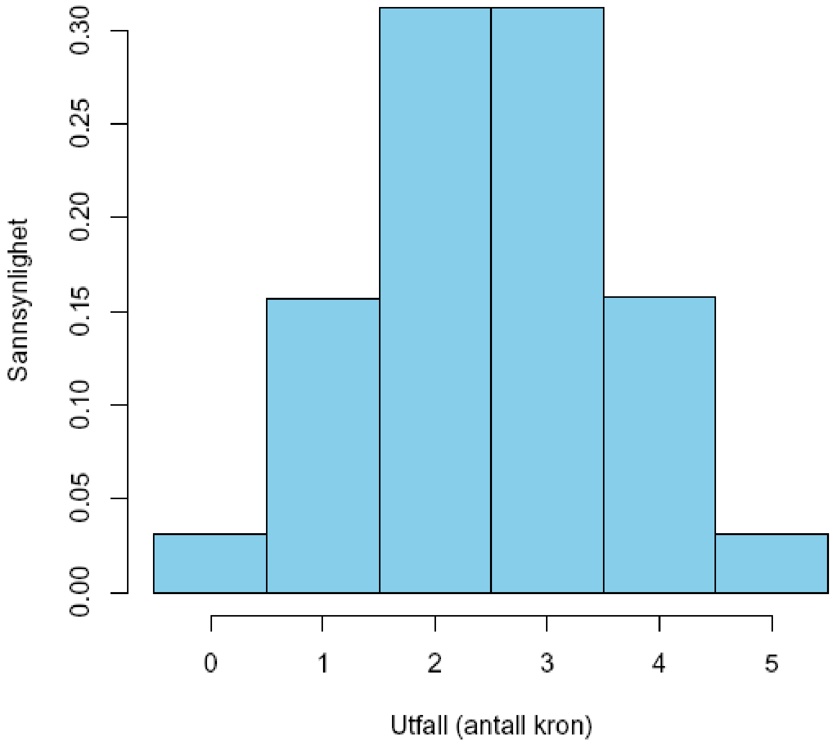
$$S = \ \{kk,\ \ km,\ \ mk,\ \ mm\}$$

Ved kast av en rettferdig mynt tilsier erfaring, fornuft og intuisjon at hvert kast er uavhengig. Utfallet av et kast påvirker ikke utfallet av neste kast. Sannsynligheten for å få kron er 50% (1/2), men dette gjelder bare hvis man foretar mange kast (De store talls lov). Vi kan ikke si noe om utfallet av variabelen som sådan, men vi kan si noe om sannsynligheten. Vi lar utfallet kron være 1 og utfallet mynt lik 0

$$P(X = 1) = P(X = 0) = \frac{1}{2}$$

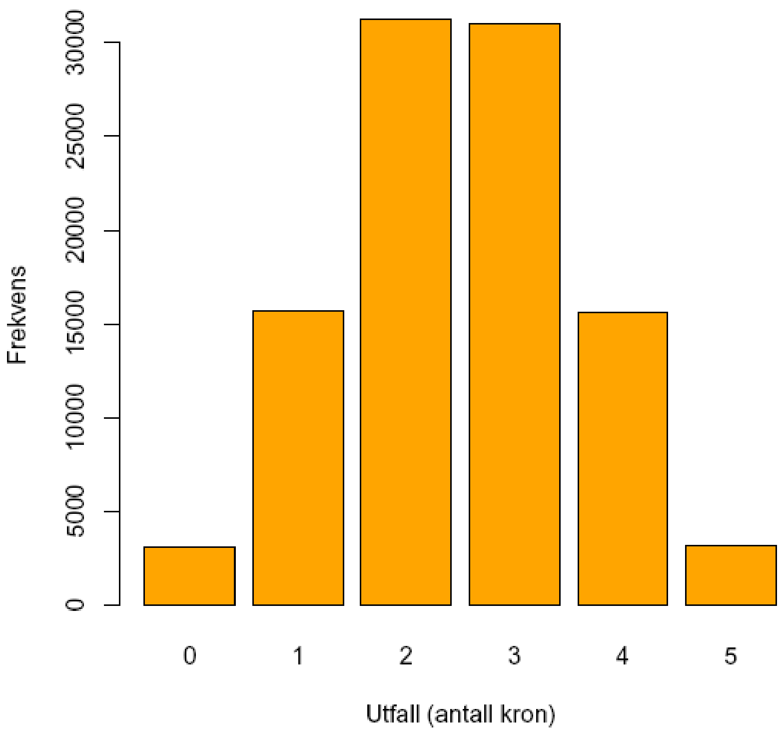
Myntkast er uavhengige hendelser, men et eksempel på avhengige hender er avhengige hendelser hvor neste trinn er påvirket av det forrige er klassifisering av vær (variabel V, hvor V_t er ufallet ved tidspunkt t) som sol (1) og skyet (0. Vær er et eksempel på korrelerte data, været i dag er omtrent som det var i går, har en værtype kommet har den en tendens til å vare ved noen dager. Et annet eksempel er aksjekurser, kursen i dag er omtrent lik den i går.

For eksempel sannsynligheten for skifte fra soldag til overskyet dag, og omvendtsoldag kommer etter en overskyet



Figur 1. Simulering av fem myntkast og telling av antall kron, n=10000.

Alle mulige **kombinasjoner** av de 5 myntkastene er:
1 5 10 10 5 1
Og de tilsvarende sannsynlighetene er
0.03125 0.15625 0.31250 0.31250 0.15625 0.03125
Det betyr e.g. at sannsynligheten for 3 kron og 2 mynt er lik 31.25%
Vi kan se at summen av sannsynlighetene blir lik 1

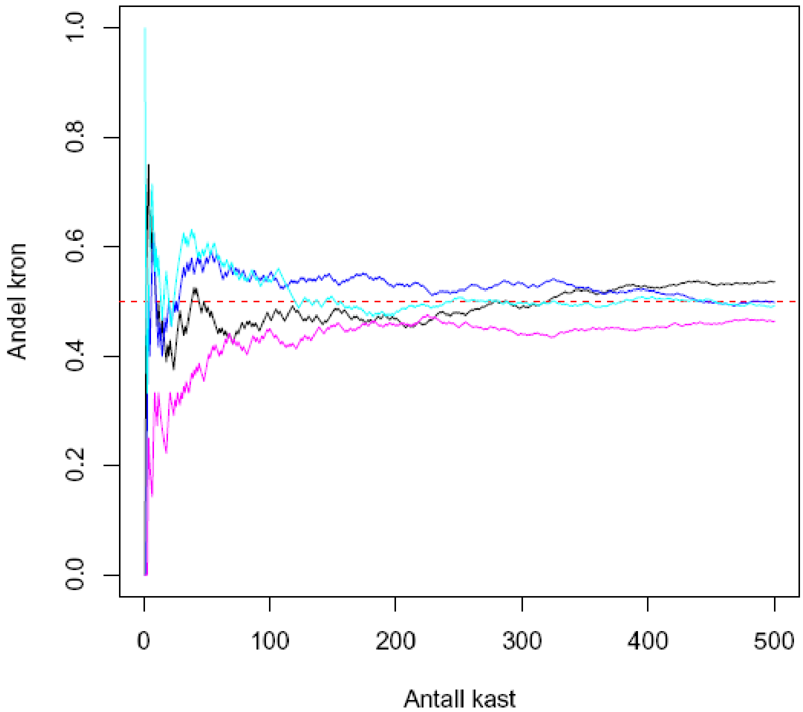


Figur 2. Simulering av kast av fem mynter 10000 ganger og opptelling av antall kron. Stolpediagrammet, som er symmetrisk, viser antall kron.

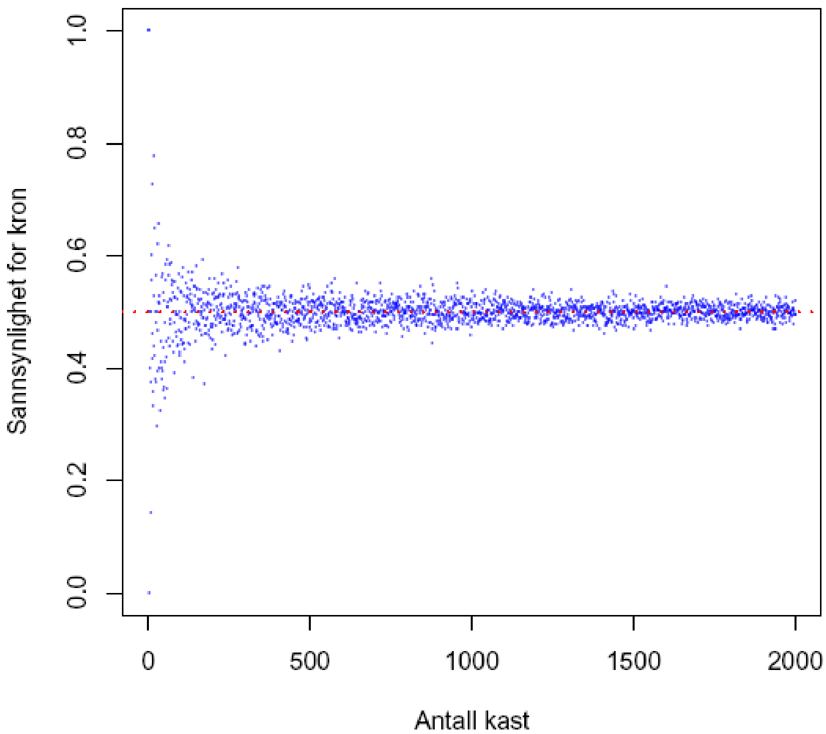
Vi kan regne ut forventet verdi, det vil si forventet antall kron (suksess) i 5 kast: E(X) = np = 5·0.5 = 2.5.
Vi kan også sjekke dette ved å summere produktet av k og punktsannsynlighetene

$$E(X) = \sum_{k=0}^{n} kP(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = n \cdot p$$

Variansen Var(X) = npq = 5·0.5·0.5 = 1.25



Figur 3. Simulering av 500 myntkast og andelen kron. Andelen av kron vil ved mange kast konvergere mot 0.5.



Figur 4. Forholdet mellom antall kron og antall kast nærmer seg 0.5 når antall forsøk øker. Er mynten helt perfekt symmetrisk (rettferdig mynt) forventer man at i det lange løp skal det bli like mange kron som mynt.

1. Hva er sannsynligheten for å få minst en kron ved å kaste en mynt to ganger ?
Alle utkomme er like sannsynlig og punktsannsynligheten for hvert av tilfellene *x* i *S* blir *P(x)*=1/4.
Hendelsen ”minst en kron” betyr undermengden:

$$A = \{kk,\ \ km,\ \ mk\}$$

Sannsynligheten for denne hendelsen er lik summen av punktsannsynlighetene for hvert element,   altså 0.75 (75% sannsynlig).

$$P(A) = \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = \frac{3}{4}$$

Fakultetsfunksjonen og binomialkoeffisienter

Fakultetsfunksjonen n! (n fakultet) er gitt ved:

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

0!=1,   det er bare en måte å ordne null objekter.

Fakultetsfunksjonen kan også uttrykkes som produktnotasjon:

$$n! = \prod_{k=1}^{n} k$$

Fakultetsfunksjonen gjelder bare for positive heltall.
Det er bare når n=0 og n=1 at n! blir et oddetall,   ellers blir det alltid et liketall,   og den øker meget raskt med n:

Vi kan regne ut **binomialkoeffisientene**:

$$\binom{n}{k} = \frac{n!}{k!\ \cdot\ (n-k)!} \qquad\qquad k \le n$$

Binomialkoeffisientene kan settes opp i **Pascals trekant**,   hvor tallene i neste horisontale linje er summen av de ovenfor og havner midt mellom dem:

1

$1\ \ 1$

$$1 \quad 2 \quad 1$$
$$1 \quad 3 \quad 3 \quad 1$$
$$1 \quad 4 \quad 6 \quad 4 \quad 1$$

Trekanten er symmetrisk om midtlinjen og

$$\binom{n}{x} = \binom{n}{n-x}$$

De tilsvarende tallene satt opp som binomialkoeffisienter:

$$\binom{0}{0}$$
$$\binom{1}{0}\binom{1}{1}$$
$$\binom{2}{0}\binom{2}{1}\binom{2}{2}$$
$$\binom{3}{0}\binom{3}{1}\binom{3}{2}\binom{3}{3}$$
$$\binom{4}{0}\binom{4}{1}\binom{4}{2}\binom{4}{3}\binom{4}{4}$$

og slik kan man fortsette nedover. Trekker man skrå linjer i Pascals trekant finner man igjen Fibonaccitallene.

Summen av to koeffisienter som står ved siden av hverandre i en horisontal rekke havner midt mellom dem i raden under:

$$\binom{n}{x} + \binom{n}{x+1} = \binom{n+1}{x+1}$$

Vi ser også at summerer vi tallene som står i de horisontale linjene får vi tallene 1, 2, 4, 16, 64 osv.
Det vil si:

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n} = 2^n$$

Loven om Pascals trekant:

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$$

Koeffisientene i Pascals trekant er dem man finner når man multipliserer ut uttrykkene:

$(x+1)^1$ $\quad (x+1)^2 = x^2 + 2x + 1$ $\quad (x+1)^3 = x^3 + 3x^2 + 3x + 1$ $\quad (x+1)^4 = x^4 + 4x^3 + 6x^2 + 4x + 1$

osv.

Eller mer generelt **binomialteoremet**:

$$(x+y)^n = \binom{n}{0} x^n y^0 + \binom{n}{1} x^{n-1} y^1 + \binom{n}{2} x^{n-2} y^2 + \binom{n}{3} x^{n-3} y^3 + \cdots + \binom{n}{n} x^0 y^n$$

som også kan uttrykkes som:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Jfr. Hardy-Weinberg og fordeling av alleler når n=2.

$(x+y)^0=1$

Sannsynligheten for 3 kron i 5 kast er for eksempel ca. 31%

$$\frac{\binom{5}{3}}{2^5}$$

Bernoullifordeling

Benoulli-eksperimenter er en sekvens av eksperimenter, gjentatte eksperimenter under samme betingelser.

Tetthetsfunksjonen for Bernoullifordelingen er:

$$p(x) = p^x \cdot (1-p)^{1-x}$$

Variansen for en Bernoullifordeling er lik *pq=p(1-p)*. Bernoulli-fordelingen har fått navn etter matematikeren og astronomen Jacob Bernoulli (1654-1705) kjent for *Ars conjectandi* (1713).

Binomialfordelingen og Bernoulli-eksperimenter

Den statistiske fordeling som beskriver myntkast er den **binomial fordeling** som består av en serie med uavhengige forsøk med to mulige utkomme (**Bernoulli-eksperimenter**). Et myntkast er et Benoulli-eksperiment, en sekvens av eksperimenter er gjentatte eksperimenter under samme betingelser.
Sannsynlihetstetthetsfunksjonen for Bernoullifordelingen er:

$$P(x) = p^x \cdot (1-p)^{1-x}$$

Bernoullifordelingen har bare to utkomme: suksess-ikkesuksess, kron-mynt, galt-riktig, tilstede-fraværende, reproduksjon-ikke reproduksjon osv.
De to sannsynlighetene blir *p* for *x(1)* suksess og sannsynlighet q = *1-p* for *x(0)* ikke-suksess. hvor $0 \leq p \leq 1$.
For myntkast har vi P(k) = p, og P(m) = 1-p = q, hvor k er kron og m er mynt
Tallet n angir et fast antall forsøk. Hvert forsøk på være uavhengig av det neste. Utfallet må bare være to kategorier.

For en binomial fordeling er sannsynligheten for suksess *k* ganger i *n* uavhengige forsøk *P(X = k)* lik:

$$P(X=k) = \frac{n!}{(n-k)!\,k!} p^k q^{n-k} \qquad k = 0, 1, 2, 3, \dots, n$$

q er sannsynligheten for ikke-suksess.

Summen av alle punktsannsynlighetene, det vil si fordelingsfunksjonen F(x), er den kumulative sannsynligheten:

$$F(x) = \sum_{k=0}^{n} P(X = k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1$$

For eksempel er *P(a ≤ x ≤ b)* gitt ved

$$F(x) = \sum_{k=a}^{b} \binom{n}{k} p^k q^{n-k}$$

Det vil si fordelingsfunksjonen, den kumulative sannsynligheten, blir:

$$F(x) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$$

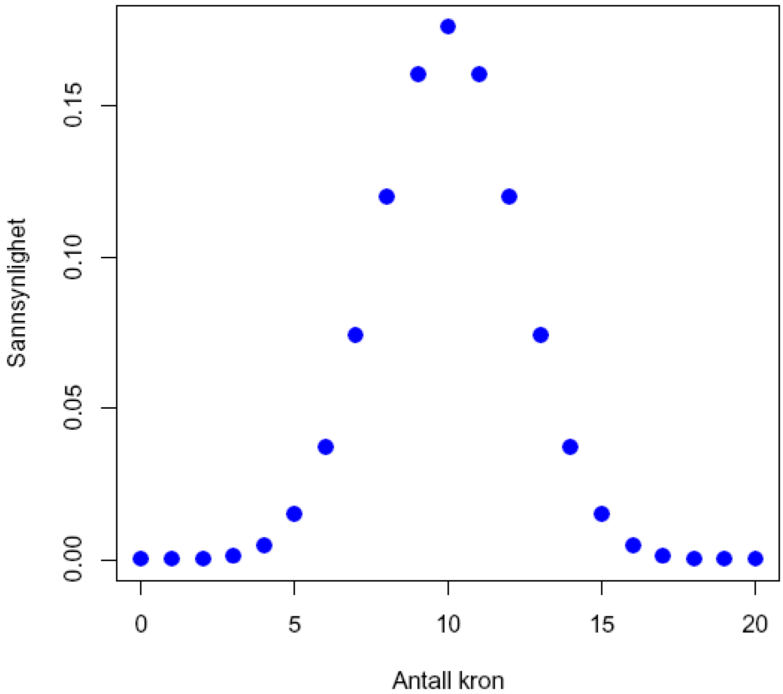
1. *Antall kron i 20 myntkast* ?
Sannsynligheten for å få en sekser i ett terningkast er 1/6, sannsynligheten for å få to seksere i to terningkast er $1/6^2=1/36$, sannsynligheten for å få tre seksere i tre terningkast er $1/6^3=1/216$ osv. Analogt med myntkast. Sannsynligheten for å få en kron i ett myntkast er 1/2, sannsynligheten for to kron i to myntkast er $1/2^2=1/4$, sannsynligheten for tre kron i tre myntkast er $1/2^3=1/8$ osv., og for 20 myntkast (=n) blir sannsynligheten for 20 kron $1/2^{20}=1/1048576$. Eller antall mulige utfall av kron er $2^{20}=1048576$
Antall måter vi kan velge ut antall kron (=k) blir:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{20!}{k!\,(20-k!)} \qquad k = 0,1,2,3,\dots,20$$

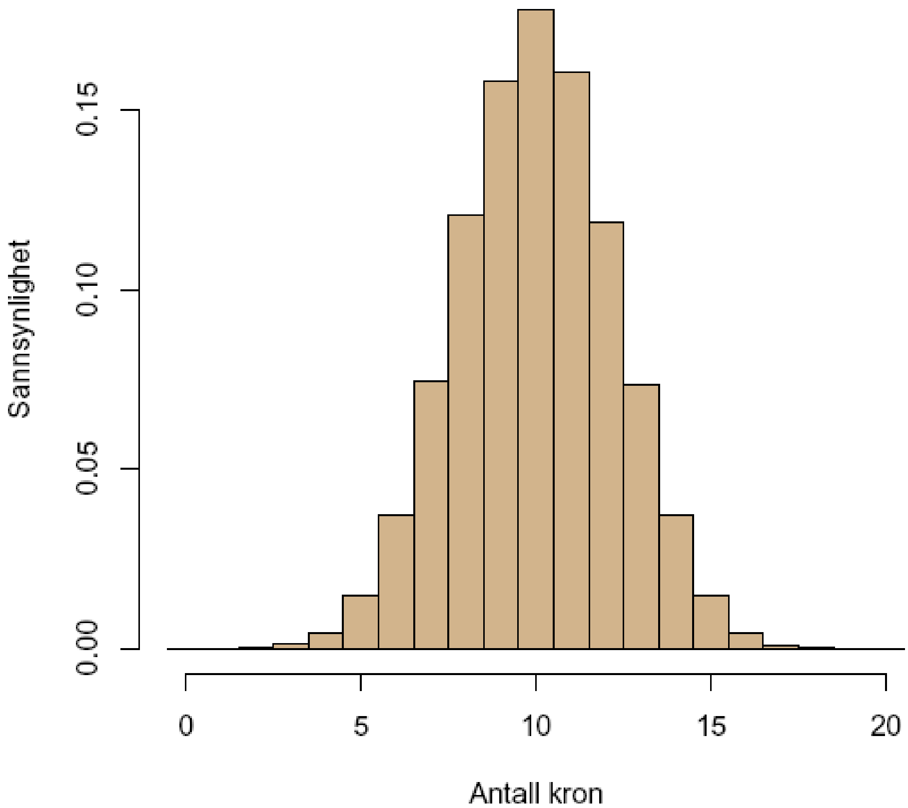
Sannsynligheten for X blir derved

$$P(X = k) = \frac{\binom{20}{k}}{2^{20}}$$

Det er størst sannsynlighet for å få 10 kron i 20 myntkast.



Figur 4. Punktsannsynlighetene for antall kron (”suksess”) i 20 myntkast.



Figur 5. Simulering av utfall av 100000 myntkast.

Oppgave
Kast fem femtiøringer, eller andre typer mynter. Tell antall utfall i kron og mynt. Sett streker i tabellen, gjenta minst 30 ganger og summer utfallene. Lag en grafisk framstilling. Regn ut frekvensene.

Kron	Strek for hvert utfall	Antall (n)
0		
1		
2		
3		
4		
5		

Myntkast som generalisert lineær modell

Myntkast kan også betraktes som en binomisk modell**,** og bli analysert med en generalisert lineær modell med linkfunksjonen logit. Vi kan la *p* være sannsynligheten for kron (suksess), mynt er "ikke-suksess" med sannsynlighet *1-p*.

$$logit(p) = ln\frac{p}{1-p} = \ln(odds) = \beta_0 + \beta_1 x = y$$

$$odds = \frac{p}{1-p} = \frac{p}{q} = \frac{P(suksess)}{P(ikke\ suksess)}$$

Når vi angir family=binomial så er det *logit(p)* vi finner og vi må beregne oss tilbake til sannsynligheten *p* ved å ta eksponentialfunksjonen på begge sider av likhetstegnet i den øverste ligningen.

Vi får da:

$$p = \frac{e^y}{1+e^y} \qquad hvor\ y = \beta_0 + \beta_1 x$$

$$p = \frac{odds}{1+odds}$$

Vi lager en nullmodell hvor vi bestemmer intercept (=β_0) i modellen. For mynt får vi ikke noe β_1-ledd siden x = 0, og β_0 angir sannsynligheten *p* for suksess (kron). Hvis β_0 er lik null, så er p=0.5, og stigningstallet β_1 bestemmer hvor raskt kurven stiger til p=1. Logit-transformasjon omdanner en S-formet logistisk linje til en rett linje.

Tsjebysjevs ulikhetsteorem

Tsjebysjevs ulikhetsteorem (Tjebysjev, P.L. (1821-1894) som gjelder for alle typer fordelinger, sier at for en endimensjonal stokastisk variabel X med forventningsverdi E(X) og varianse Var(X) så vil for et positivt tall c

$$P[|X - E(X) > c|] \leq \frac{Var(X)}{c^2}$$

Hvis vi setter c lik antall ganger standardavviket Nσ

$$P[|X - E(X) > N\sigma|] \leq \frac{\sigma^2}{(N\sigma)^2} = \frac{1}{N^2}$$

Det vil si at sannsynligheten for at X avviker fra E(X) mer enn N standardavvik er mindre enn $1/N^2$. Hvis N=3 standardavvik
så vil sannsynligheten for forskjellen mellom X og E(X) være mindre enn 1/9=0.111. Tsjebysjev studerte også primtallsfordeling og de store talls lov.

Vi kan teste om en mynt er helt jevnt laget ved å kaste den n=10000 og telle antall kron i en binomial fordeling p=1/2.
Forventet verdi er E(X) = np =5000,
variansen Var(X)=npq=2500 og standardavviket σ=50.
±3.29 ganger standardavviket tilsvarer 99.9% av arealet under normalfordelingskurven.
Det vil si at blir gjennomsnittlig antall kron mellom 4836 og 5165 er det 99.9% sannsynlig (p=0.01) at myntene er helt jevnt laget.
Tsjebysjevs teorem sier p=0.09.

Myntkast og maksimum likelihood estimering (MLE)

Likelihoodfunksjonen for en binomial fordeling er:

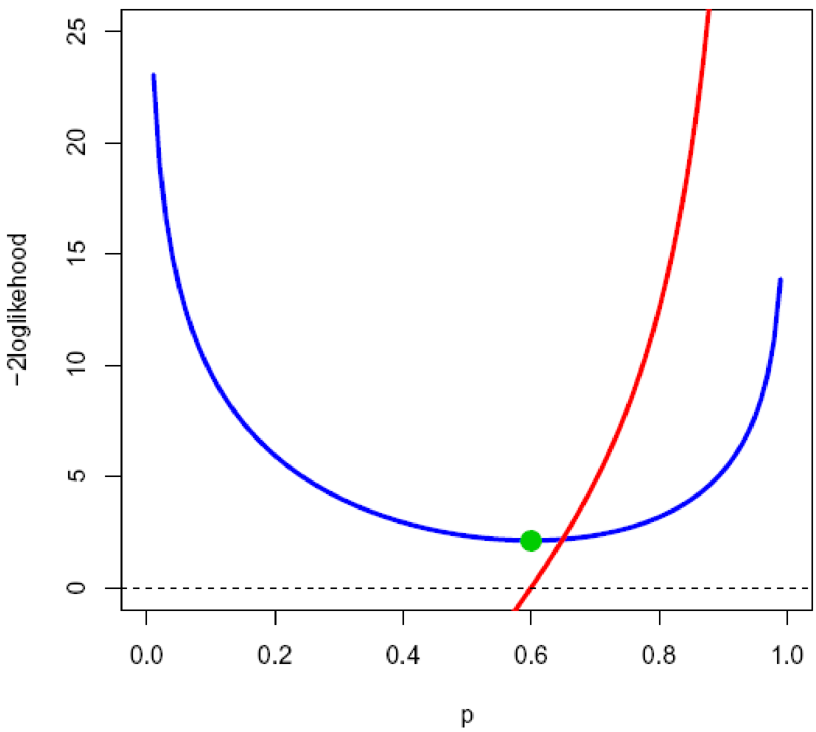
$$L(p|Y) = \binom{n}{Y} p^Y (1-p)^{n-Y}$$

Det er bare en ukjent parameter, p. Vi ønsker ut fra gitte data Y å bestemme parameterverdi p.

-2 loglikehoodfunksjon en blir:

$$-2logL(p|Y) = -2\left(ln\binom{n}{Y} + Ylnp + (n-Y)\ln(1-p)\right)$$

Hvis vi får 3 kron i 5 myntkast, hvor vi bestemmer den partiellderiverte av loglikelihoodfunksjonen og setter denne lik 0, og løsningen av denne gir $\hat{p}$, MLE, den verdien som maksimerer log-likelihoodfunksjoen eller minimerer minus loglikelihood.



Figur 6. Funksjonen for -2logliehood for utfall 3 kron i 5 kast (blå). Minimumspunktet er størst sannsynlighet er 0.6, men funksjonen er relativt bred og antyder at sannsynligheten må være i området 0.4-0.7.

Den rød kurven viser den førstederiverte, og den deriverte er lik null ved minimumspunktet.
Estimatet for p blir $\hat{p} = 0.6$. Hvis vi hadde fått 30 kron i 50 kast minsker dette konfidensintervallet.
Imidlertid, en rettferdig symmetrisk mynt ville gi ca. 25 kron i 50 kast.
95% konfidensintervall kan vi bestemme fra tabellverdien for χ^2-fordelingen

$$\hat{L} + \frac{\chi_1^2(0.95)}{2} \qquad \frac{\chi_1^2(1-\alpha = 0.95)}{2} = 1.92$$

Den mest sannsynlige parameterestimat er gitt ved:

$$\hat{p} = \frac{Y}{n}$$

i vårt eksempel 25/50=0.5
og MLE for variansen er:

$$Var(p) = \frac{\hat{p}(1-\hat{p})}{n}$$

Variansen minsker når n øker, og variansen er største når $\hat{p} = 0.5$
Legg merke til at i Akaikes informasjonskriterum (AIC) inngår -2logLL:

$$AIC = -2logLL + 2k$$

hvor k er antall parametere i modellen.

Tegnestiftkast

For myntkast er utfallsrommet S={kron,mynt} eller S={K,M}
Sannsynligheten for å få en kron i et myntkast er lik 1/2, forutsatt at mynten er helt symmetrisk. Vi forutsetter at mynten ikke blir stående på høykant. Kaster man to mynter er utfallsrommet S={KK,KM,MK,MM). Legg merke til at vi har begge mulighetene KM og MK. Sannsynlighet for to kron er lik 1/4, og altså ikke lik 1/3, en feilslutning d'Alembert (1717-1783) gjorde.

Kaster man en tegnestift ned på en flate blir det ikke samme sannsynligheten for de to utfallene spissen ned og spissen opp. Utfallet blir også påvirket av underlaget, for eksempel om det er en blank glatt flate eller om det er et ullunderlag. Sannsynligheten for antall "spiss opp":

$$P(spiss\ opp) = \frac{antall\ spiss\ opp}{n\ \ kast}$$

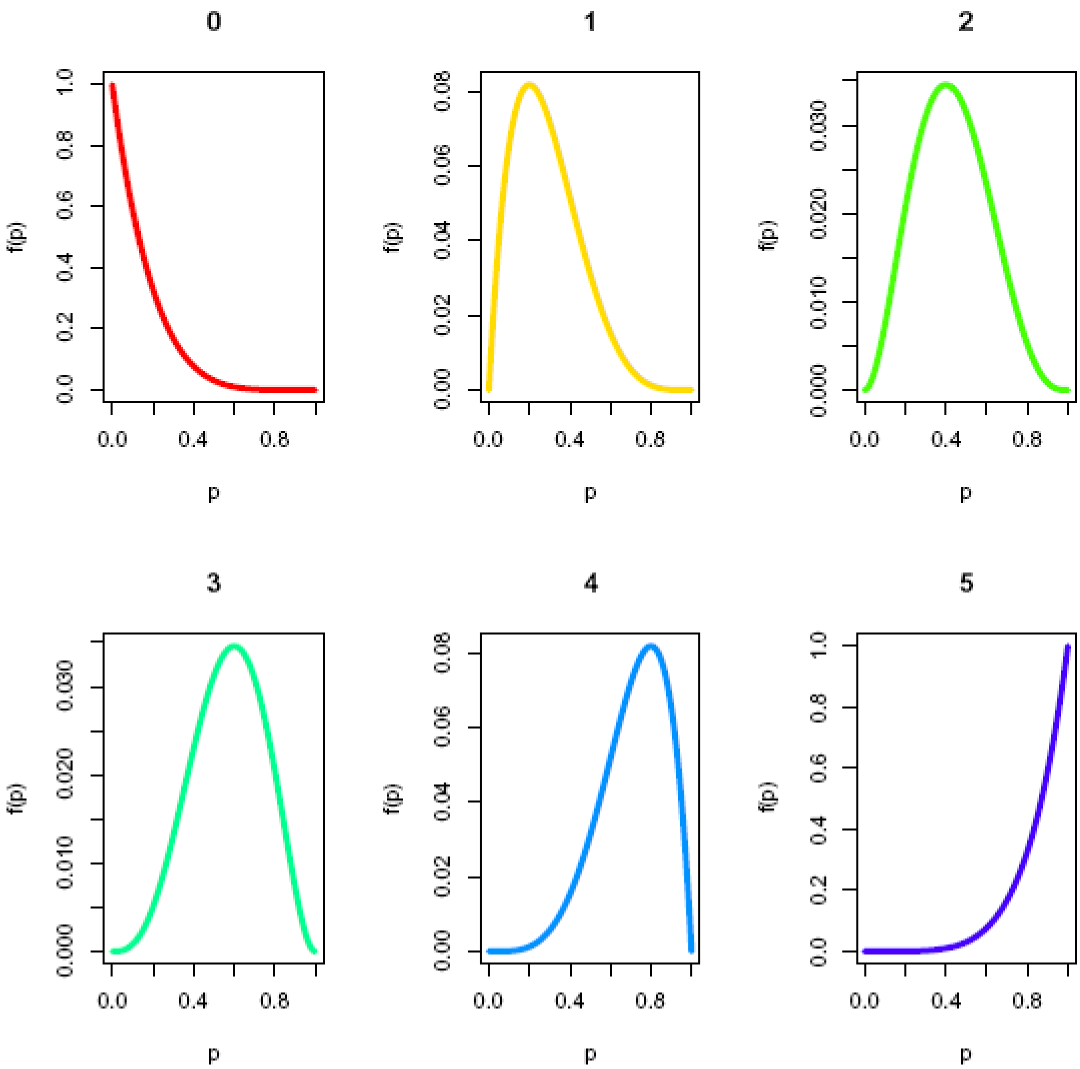
Kast av tegnestifter atskiller seg fra myntkast. Vi har ingen empirisk kunnskap om hva sannsynligheten for ”spiss opp” er. Først må man gjøre et eksperiment som gir data, og ut fra datasettet kan man med maksimum likelihood estimering bestemme sannsynligheten.

Oppgave

Kast 5 tegnestifter, noter antall ”spiss opp” , og gjenta minst 30 ganger. Sammenlign et eksperiment med slipp av tegnestiftene på en glatt bordplate og et stykke tøy.

Sammenlign med et tilsvarende forsøk hvor man bruker mynter.
Spiss opp = kron, spiss ned = mynt

Spissopp=Kron	Strek for hvert utfall	Antall (n)
0		
1		
2		
3		
4		
5		



Figur 7. Sannsynlighet gitt ved maksimumspunktet av kurvene for 0-5 ”tegnestift opp” i fem kast.

Litteratur

Wikipedia

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

(PS. På min gamle “hjemme-PC” har jeg ikke fått somlet meg til en oppdatering).