SF1901: Sannolikhetslära och statistik
Föreläsning 8.
Approximationer av sannolikhetsfördelningar

Jan Grandell & Timo Koski

11.02.2016

KTH Matematik

Centrala gränsvärdessatsen

Vi har sett några exempel på att normalfördelningen har trevliga statistiska egenskaper. Detta skulle vi inte ha så stor glädje av, om normalfördelningen inte dessutom var vanligt förekommande. Centrala gränsvärdessatsen CGS, som är den huvudsakliga motiveringen för normalfördelningen, kan utan vidare sägas vara ett av sannolikhetsteorins och statistikens allra viktigaste resultat.

Sats

(CGS) *Låt $X_1, X_2, \ldots$ vara oberoende och lika fördelade s.v. med väntevärde μ och standardavvikelse σ. Då gäller att*

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq x\right) \to \Phi(x) \quad \text{då } n \to \infty.$$

KTH VETENSKAP OCH KONST
KTH Matematik

CGS

Ofta uttrycker man slutsatsen i CGS som att

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \text{ är approximativt } N(0,1)\text{-fördelad}$$

eller att

$$\sum_{i=1}^{n} X_i \text{ är approximativt } N\left(n\mu, \sigma\sqrt{n}\right)\text{-fördelad.}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Diskreta st. v:er: Binomialfördelningen

Definition

En diskret s.v. X säges vara binomialfördelad med parametrarna n och p, $\mathrm{Bin}(n,p)$*-fördelad, om*

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ för } k = 0, 1, \ldots, n.$$

Vi skriver detta med $X \in \mathrm{Bin}(n,p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Diskreta st. v:er: Binomialfördelningen

De generella villkoren för detta:

- *n oberoende upprepningar av ett försök.*
- *varje försök har två utfall, 0 och 1.*
- *sannolikheten för lyckat försök (=1) är densamma = p vid varje försök.*

X = antalet lyckade försök.

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ för } k = 0, 1, \ldots, n.$$

Vi skriver detta med $X \in \text{Bin}(n, p)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Binomialfördelningen

Antag att vi gör ett försök där en händelse A, med sannolikheten $p = P(A)$, kan inträffa. Vi upprepar försöken n gånger, där försöken är oberoende. Sätt

$X =$ antalet gånger som A inträffar i de n försöken.

Vi säger då att X är binomialfördelad med parametrarna n och p, eller kortare att X är $\mathrm{Bin}(n, p)$-fördelad.

Vi har

$$p_X(k) = \binom{n}{k} p^k q^{n-k}, \text{ för } k = 0, \ldots, n,$$

där $q = 1 - p$.

KTH Matematik

$X \in \mathrm{Bin}(n, p)$, $X = U_1 + \ldots + U_n$

Låt $U_1, \ldots, U_n$ vara s.v. definierade av

$$U_i = \begin{cases} 0 & \text{om } A^* \text{ inträffar i försök nummer } i, \\ 1 & \text{om } A \text{ inträffar i försök nummer } i. \end{cases}$$

Lite eftertanke ger att $U_1, \ldots, U_n$ är oberoende och att

$$X = U_1 + \ldots + U_n.$$

KTH VETENSKAP OCH KONST
KTH Matematik

$X \in \text{Bin}(n, p)$, $X = U_1 + \ldots + U_n$

Då

$$E(U_i) = 0 \cdot (1 - p) + 1 \cdot p = p$$

och

$$V(U_1) = E(U_i^2) - E(U_i)^2 = E(U_i) - E(U_i)^2 = p - p^2 = p(1 - p)$$

så följer

$$E(X) = nE(U_i) = np \quad \text{och} \quad V(X) = nV(U_i) = npq.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Bin(n,p) approximativt $N(np, \sqrt{npq})$-fördelad

Av Xs representation som en summa följer att CGS kan tillämpas.

Sats

Om X är Bin(n,p)*-fördelad med $npq \geq 10$ så är X approximativt $N(np, \sqrt{npq})$-fördelad.*

Detta innebär att

$$\left.\begin{array}{l} P(X \leq k) \\ P(X < k) \end{array}\right\} \approx \Phi\left(\frac{k-np}{\sqrt{npq}}\right).$$

Med *halvkorrektion* menas att vi använder följande approximation:

$$P(X \leq k) \approx \Phi\left(\frac{k+\frac{1}{2}-np}{\sqrt{npq}}\right),$$

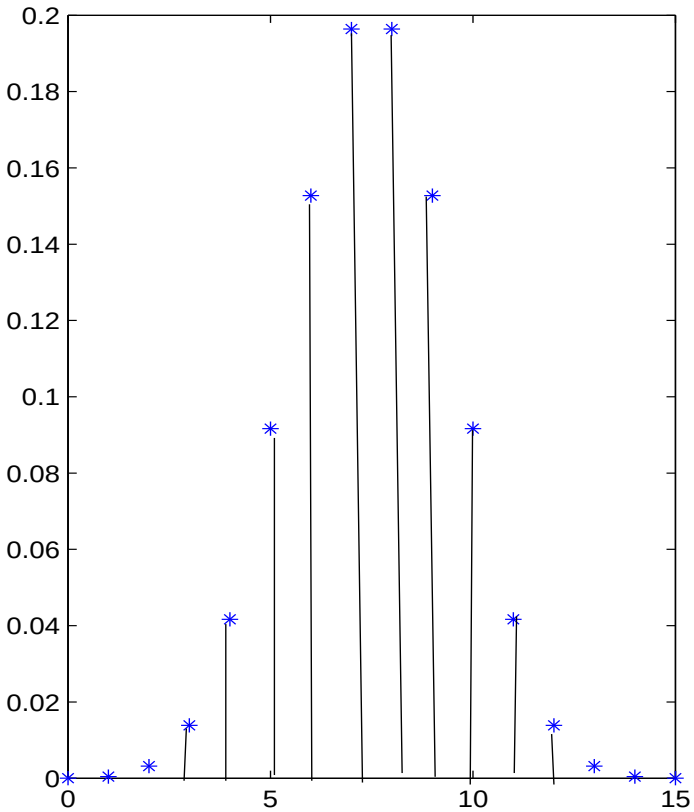
$$P(X < k) \approx \Phi\left(\frac{k-\frac{1}{2}-np}{\sqrt{npq}}\right).$$

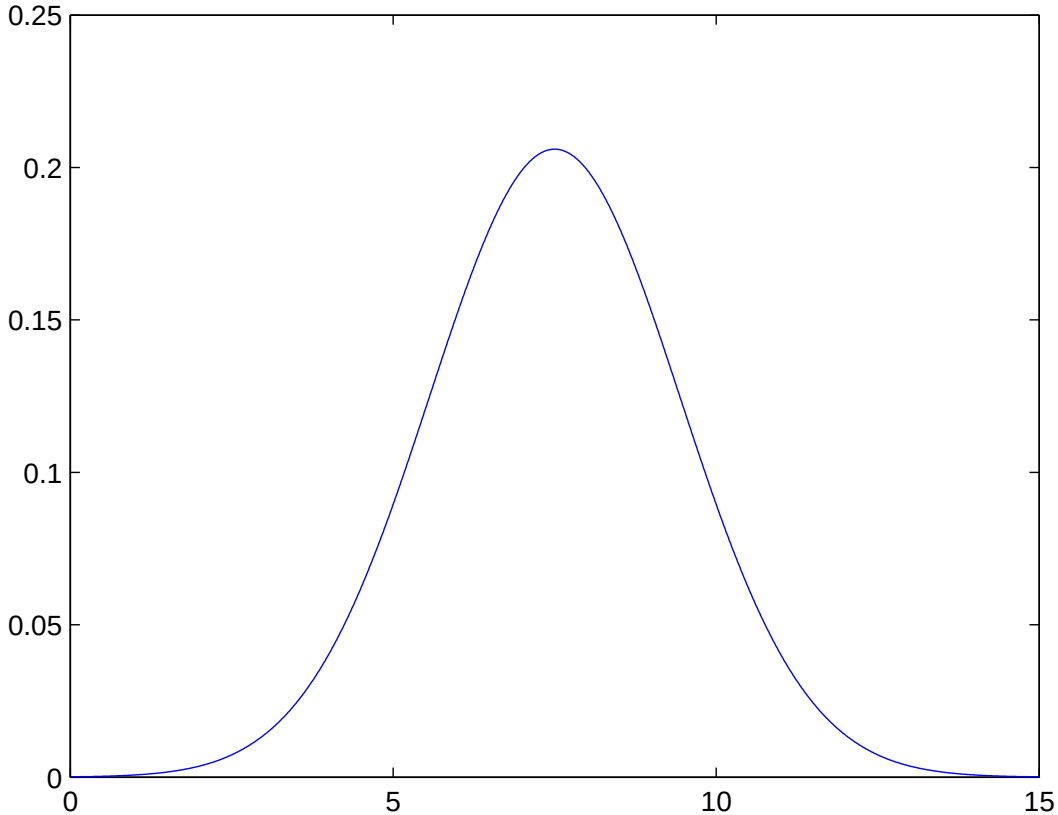
KTH Matematik

Binomial distribution: approximation

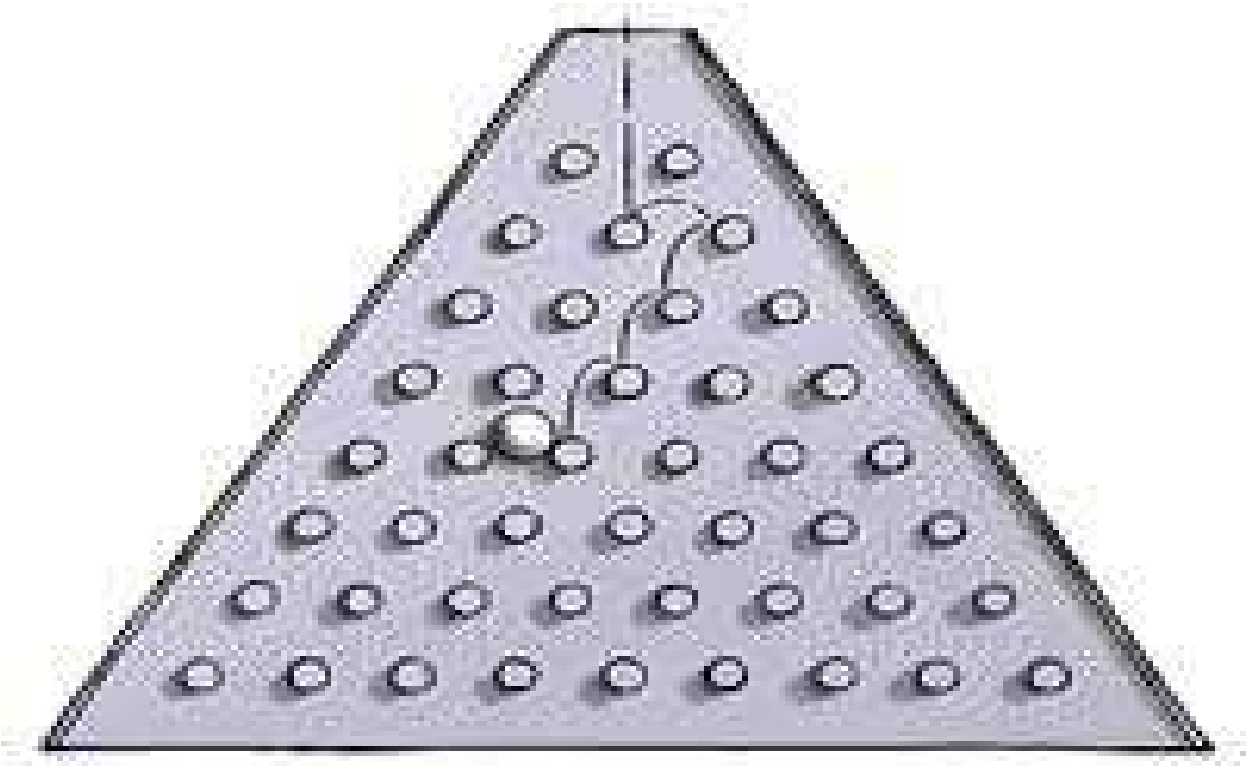
For $p = 0.5$ and $n = 15$ the Binomial distribution with mean $15 \cdot 0.5 = 7.5$, and variance $15 \cdot 0.5 \cdot 0.5 = 3.75$. The curve in the right hand figure $\mu = 7.5$ and $\sigma^2 = 3.75$.

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$





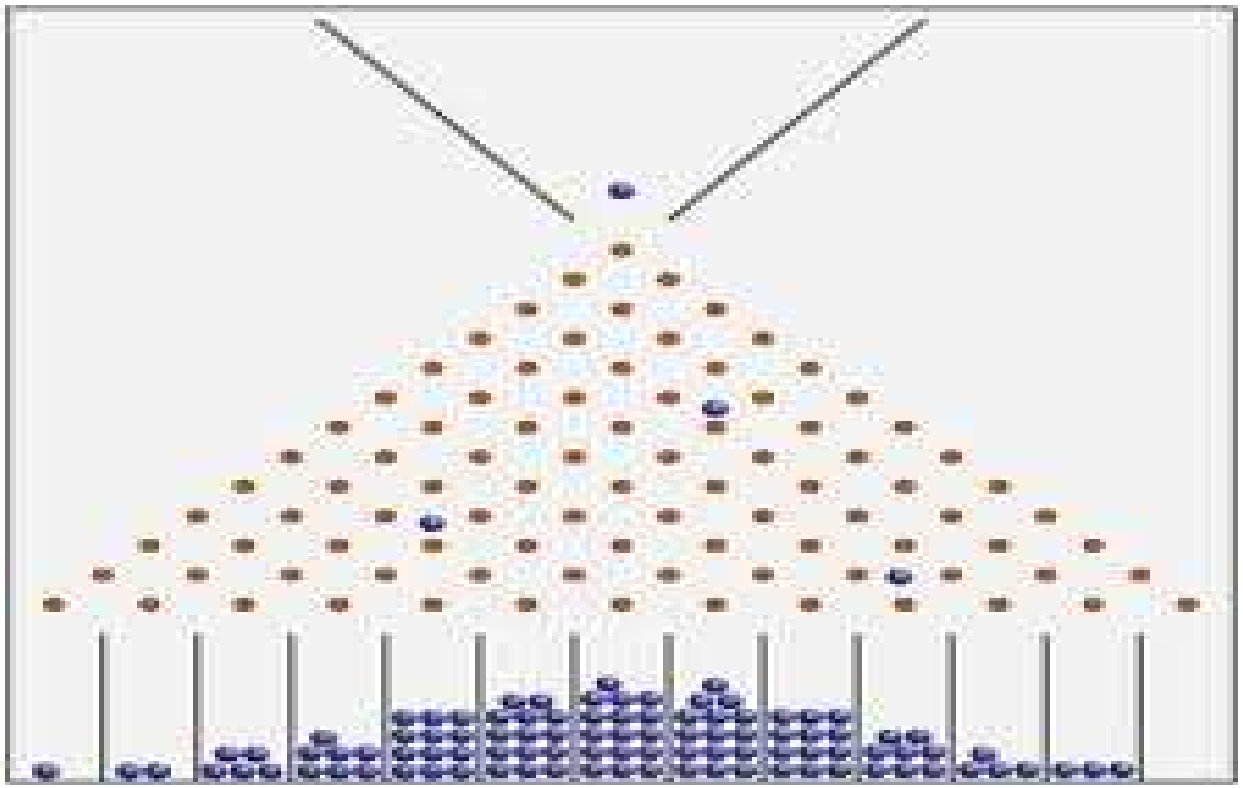
The Galton Board



KTH Matematik

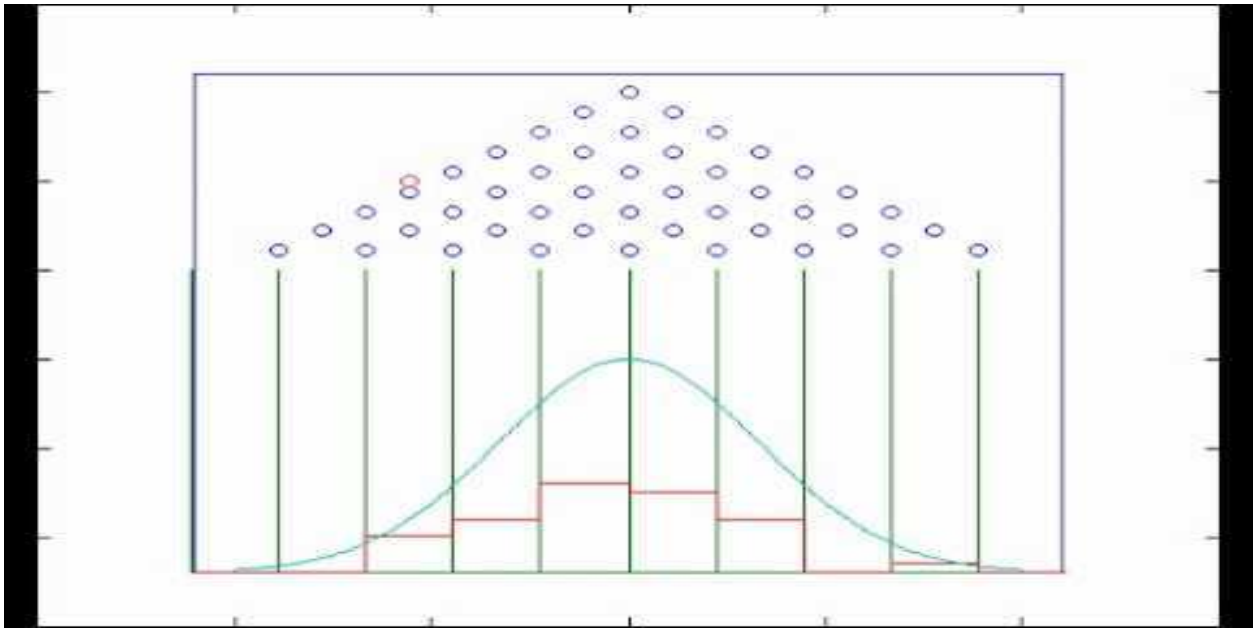
The Galton Board Experiment

The Galton board experiment consists of performing n trials with probability of success p. The trial outcome are represented graphically as a path in the Galton board: success corresponds to a bounce to the right and failure to a bounce to the left.



KTH Matematik

The Galton Board Experiment: Bell Shape

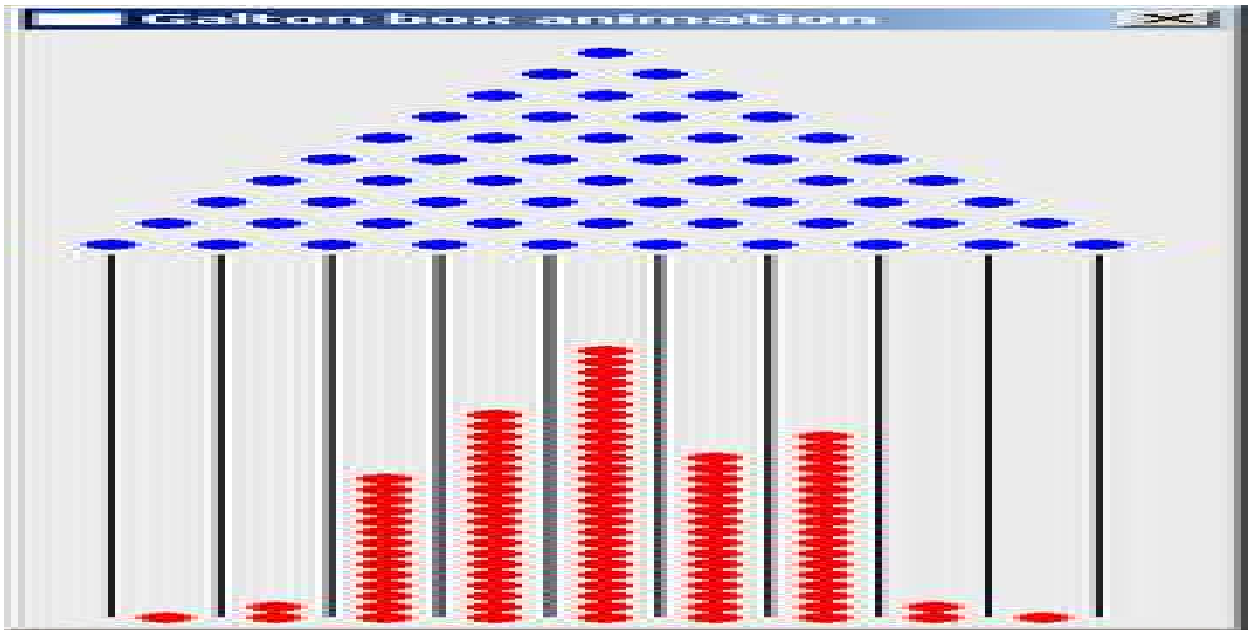


KTH Matematik

The Galton Board Experiment

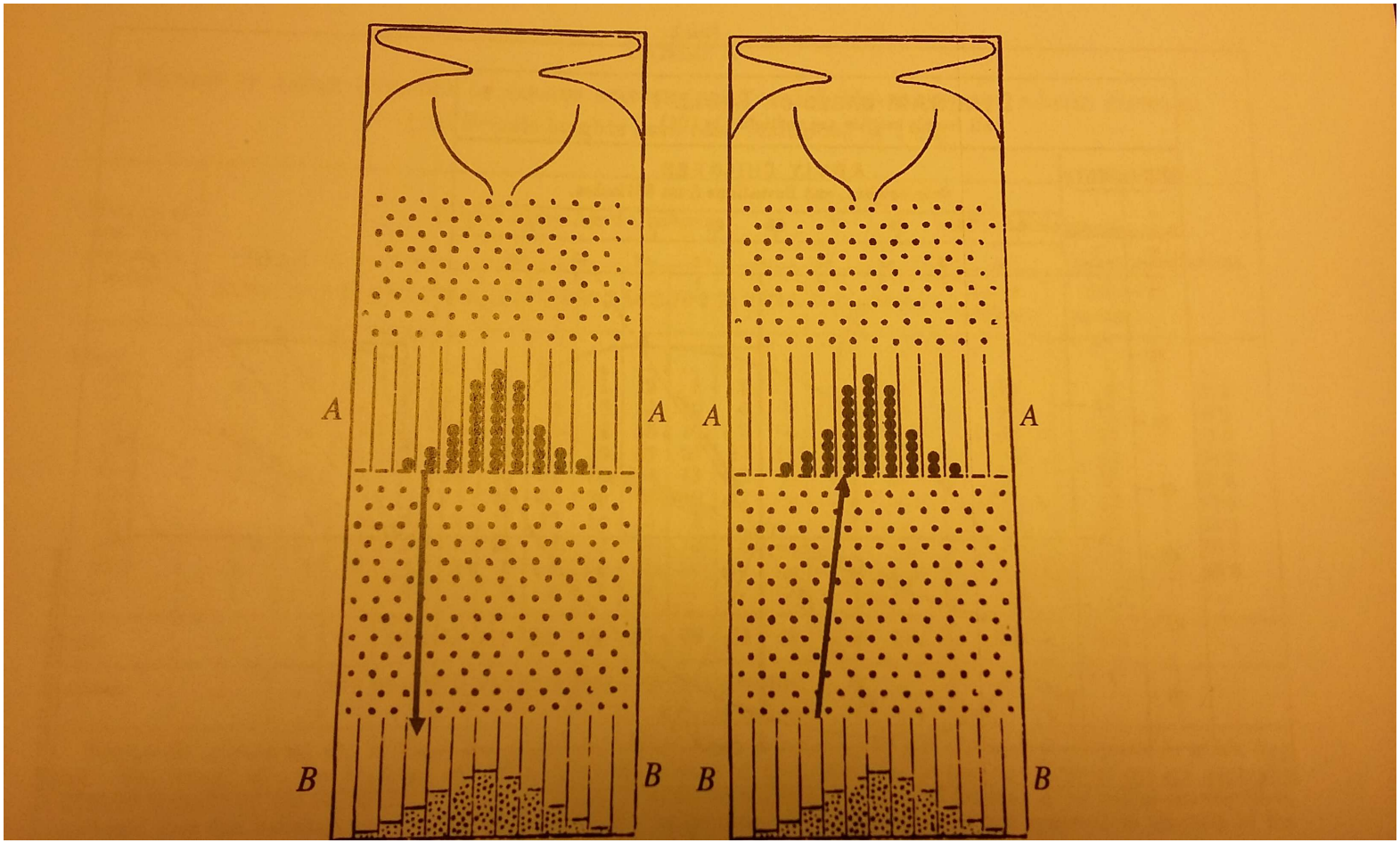
An applet for the Galton board experiment

`http://www.math.uah.edu/stat/applets/GaltonBoardExperiment.html`



KTH Matematik

Galtons Quincunx



CGS: en tillämpning på multiplikativa processer

Sats

Låt $X_1, X_2, \ldots, X_n$ vara oberoende och likafördelade s.v. och positiva dvs. för all i har vi $P(X_i > 0) = 1$. Om n är tillräckligt stort, så har

$$Y = X_1 \cdot X_2 \cdots X_n$$

approximativt en lognormalfördelning.

Bevis.

$$\ln Y = \ln X_1 + \ln X_2 + \ldots + \ln X_n$$

Således är $\ln Y$ en summa av oberoende likafördelade s.v. er (dvs. $\ln X_i$na) och därmed enligt CGS approximativt normalfördelad om n är tillräckligt stort, och därför är Y approximativt lognormalfördelad, om n är tillräckligt stort.

KTH Matematik

Extra: Lognormalfördelning

Antag att Z är en positiv stokastisk variabel, dvs. $P(Z > 0) = 1$. Vi sätter $Y = \ln Z$ och ANTAR att $Y \in N(\mu, \sigma)$. Vad är fördelningen för Z ? Tag $z > 0$.

$$F_Z(z) = P(Z \leq z) = P(\ln Z \leq \ln z)$$

$$= P(Y \leq \ln z) = P\left(\frac{Y - \mu}{\sigma} \leq \frac{\ln z - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{\ln z - \mu}{\sigma}\right)$$

Derivering ger

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \varphi\left(\frac{\ln z - \mu}{\sigma}\right) \cdot \frac{1}{\sigma z} =$$

$$= \frac{1}{\sigma z \sqrt{2\pi}} e^{-(\ln z - \mu)^2 / 2\sigma^2}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Nytt: Lognormalfördelning

Z är en positiv stokastisk variabel, d.v.s. $P(Z > 0) = 1$ och $Y = \ln Z \in N(\mu, \sigma) \Rightarrow$

$$f_Z(z) = \begin{cases} \frac{1}{\sigma z \sqrt{2\pi}} e^{-(\ln z - \mu)^2/2\sigma^2} & \text{för } z \geq 0, \\ 0 & \text{för } z < 0. \end{cases}$$

Vi säger att Z är lognormalfördelad, $Z \in LN(\mu, \sigma)$.

KTH Matematik

Lognormalfördelning

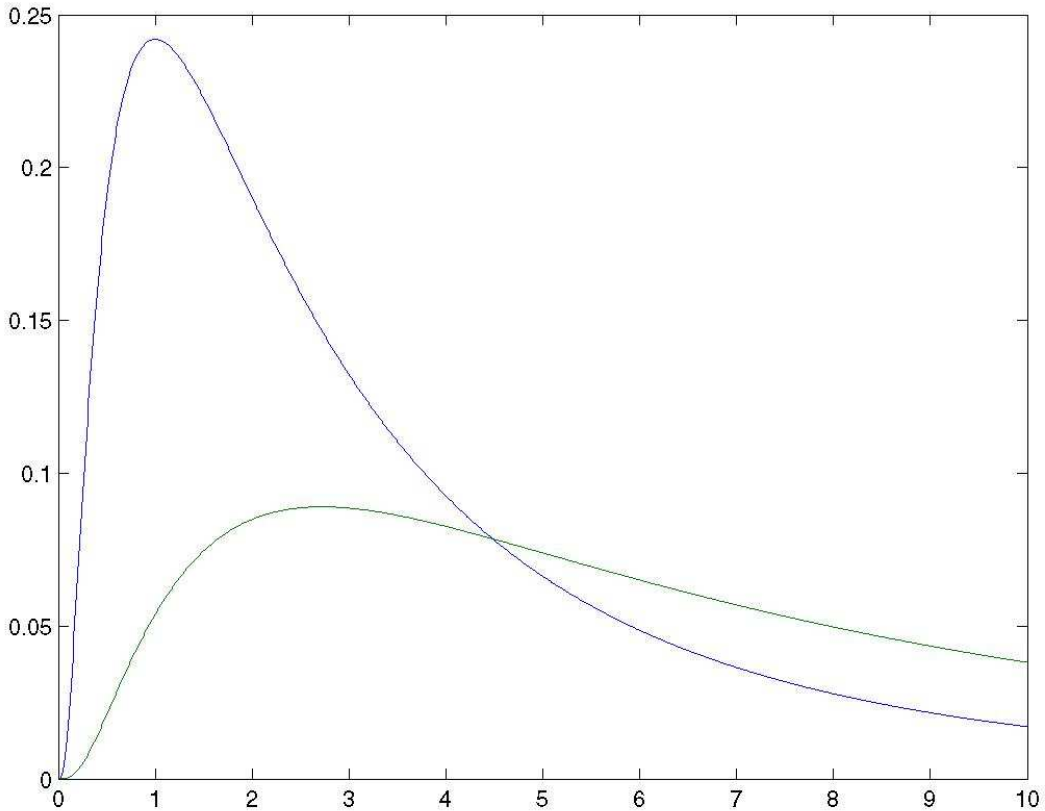
$Z \in LN(\mu, \sigma)$. Observera att μ och σ INTE är väntevärde och standardavvikelse för Z. Man kan i själva verket härleda att

$$E(Z) = e^{\mu+\sigma^2/2}, V(Z) = e^{2\mu+\sigma^2} \cdot \left(e^{\sigma^2} - 1\right)$$

KTH Matematik

Lognormalfördelning

$Z \in LN(1,1) \Rightarrow E(Z) = 4.48,\ V(Z) = 34.5$
$Z \in LN(2,1) \Rightarrow E(Z) = 12.18,\ V(Z) = 255.0$



Poissonfördelningen

Definition

En diskret s.v. X säges vara Poissonfördelad med parameter μ, $\mathrm{Po}(\mu)$*-fördelad, om*

$$p_X(k) = \frac{\mu^k}{k!} e^{-\mu}, \text{ för } k = 0, 1 \ldots .$$

Vi påminner om att om X är $\mathrm{Po}(\mu)$-fördelad, så gäller

$$E(X) = \mu \quad \text{och} \quad V(X) = \mu.$$

KTH Matematik

Summan av två oberoende diskreta: faltningsformeln

Ett hjälpresultat:
X och Y vara oberoende diskreta variabler. k är ett icke-negativt heltal,

$$P(X+Y=k) = P\left(\cup_{i=0}^{k}\{X=i \cap Y=k-i\}\right)$$

$$= \sum_{i=0}^{k} P\left(\{X=i \cap Y=k-i\}\right),$$

ty händelserna i unionen är diskjunkta. Oberoendet ger

$$= \sum_{i=0}^{k} P\left(\{X=i\}\right) P\left(\{Y=k-i\}\right),$$

och vi har den s.k. *faltningsformeln*

$$P(X+Y=k) = \sum_{i=0}^{k} p_X(i)p_Y(k-i) = \sum_{i=0}^{k} p_X(k-i)p_Y(i).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Summan av två oberoende Poissonfördelade

Poissonfördelningen är den viktigaste diskreta fördelningen, och har t.ex. följande trevliga egenskap.

Sats

Om X och Y vara oberoende $\mathrm{Po}(\mu_X)$- resp. $\mathrm{Po}(\mu_Y)$-fördelade s.v. Då gäller att $X+Y$ är $\mathrm{Po}(\mu_X+\mu_Y)$-fördelad.

Bevis. Hjälpresultatet (faltningsformeln) ovan ger

$$P(X+Y=k) = \sum_{i=0}^{k} P(X=i)P(Y=k-i) = \sum_{i=0}^{k} \frac{\mu_X^i}{i!} e^{-\mu_X} \frac{\mu_Y^{(k-i)}}{(k-i)!} e^{-\mu_Y}$$

$$= e^{-(\mu_X+\mu_Y)} \sum_{i=0}^{k} \frac{\mu_X^i \mu_Y^{(k-i)}}{i!(k-i)!}$$

$$= e^{-(\mu_X+\mu_Y)} \frac{(\mu_X+\mu_Y)^k}{k!} \underbrace{\sum_{i=0}^{k} \binom{k}{i} \left(\frac{\mu_X}{\mu_X+\mu_Y}\right)^i \left(\frac{\mu_Y}{\mu_X+\mu_Y}\right)^{(k-i)}}_{=1, \text{ enl. Binomialsatsen.}}$$

Approximation av $\mathrm{Po}(\mu)$

Om bägge villkoren $p \leq 0.1$ och $npq \geq 10$ är uppfyllda kan vi välja om vi vill Poissonapproximera eller normalapproximera. Detta är ingen motsägelse, som följande sats visar.

Sats

Om X är $\mathrm{Po}(\mu)$-fördelad med $\mu \geq 15$ så är X approximativt $N(\mu, \sqrt{\mu})$-fördelad.

Diskreta st. v:er: Poissonfördelningen

Ofta när det är rimligt att anta att en s.v. X är $\mathrm{Bin}(n,p)$-fördelad, så är det även rimligt att anta att p är liten och att n är stor. Låt oss anta att $p = \mu/n$, där n är "stor" men μ är "lagom". Då gäller

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\mu}{n}\right)^k \left(1-\frac{\mu}{n}\right)^{n-k}$$

$$= \frac{\mu^k}{k!} \underbrace{\left(1-\frac{\mu}{n}\right)^n}_{\approx e^{-\mu}} \underbrace{\frac{n(n-1)\dots(n-k+1)}{n^k}}_{\approx 1} \underbrace{\left(1-\frac{\mu}{n}\right)^{-k}}_{\approx 1} \approx \frac{\mu^k}{k!} e^{-\mu}.$$

KTH Matematik

Approximation med $\mathrm{Po}(np)$

Vi kan alltså införa Poissonfördelningen som en approximation av binomialfördelningen. Detta kan vi formalisera till följande sats.

Sats

Om X är $\mathrm{Bin}(n,p)$-fördelad med $p \leq 0.1$ så är X approximativt $\mathrm{Po}(np)$-fördelad.

I vår approximation antog vi även att n var stor. Detta är inte nödvändigt, men vårt enkla resonemang fungerar inte utan denna extra förutsättning. Man kan visa att om X är $\mathrm{Bin}(n,p)$ och Y är $\mathrm{Po}(np)$ så gäller att

$$|P(X=k) - P(Y=k)| \leq np^2.$$

KTH Matematik

Ett exempel : approximation med Po(np),

I en datalänk växlar de binära siffrorna eller *bitarna* (0 och 1) polaritet dvs. 0 övergår i 1 eller omvänt med sannolikheten 10^{-7}. En viss bit i en bitgrupp växlar polaritet oberoende av alla de andra bitarna. Protokollet i datalänken kontrollerar en bitgrupp som omfattar 8000000 (≈ 1 *Mbyte*) bitar.

Felkontrollen upptäcker om fem eller flera bitar fått sina polariteter omkastade i bitgruppen. Om fem eller flera polaritetsväxlingar har upptäckts, begärs av protokollet en upprepad överföring av denna bitgrupp.

Vi säger att *en överföring lyckas* om en upprepad överföring av bitgruppen *inte* kommer att begäras. De olika överföringarna av en och samma bitgrupp antas vara oberoende av varandra.

KTH Matematik

Approximation med Po(np), ett exempel: forts.

Betrakta den stokastiska variabeln

X = antalet överföringar av en bitgrupp om 8000000 bitar som behövs för att överföringen skall lyckas för första gången, den lyckade överföringen medräknad.

Sökt: väntevärdet för X.

KTH VETENSKAP OCH KONST
KTH Matematik

Exempel: forts.

Låt p_s beteckna sannolikheten för att en ny överföring av en bitgrupp om 8000000 bitar lyckas dvs. p_s = sannolikheten för högst fyra växlingar av polaritet.

Den stokastiska variabeln X har $ffg\,(p_s)$ - fördelningen eller "För-första-gången"-fördelningen p.g.a. att de olika överföringarna är oberoende och p.g.a. att den lyckade överföringen medräknas. Dvs.

$$P(X = k) = (1 - p_s)^{k-1} p_s, \quad k = 1, 2, 3, \ldots,$$

En enkel kalkyl ger väntevärdet $E\,(X) = \frac{1}{p_s}$ (se kursens formelsamling). Därmed behöver vi p_s.

KTH Matematik

Exempel: forts.

Låt Z vara antalet bitar som får sin polaritet omkastad, när en bitgrupp om 8000000 bitar överförs. Då fås

$$p_s = P\left(Z \leq 4\right).$$

P.g.a. att bitarna i en bitgrupp växlar polaritet oberoende av varandra är vi ledda till den statiska modellen $Z \in Bin(8000000, 10^{-7})$, dvs. att Z är binomialfördelad med parametrarna $n = 8000000$ och $p = 10^{-7}$.

KTH Matematik

Exempel: forts.

Vi beräknar p_s med en approximation av binomialfördelningen $Bin(8000000, 10^{-7})$ med Poissonfördelningen med parametern $n \cdot p = 0.8$. Denna approximation är rimlig, ty $p = 10^{-7} \leq 0.1$. Vi har alltså att Z är approximativt $\mathrm{Po}(0.8)$- fördelad. Tabellsamlingen ger med Poissonfördelningen $\mathrm{Po}(0.8)$ approximationen $p_s = P(Z \leq 4) \approx 0.99859$.

SVAR: $\underline{E(X) \approx \frac{1}{0.99859} \approx 1.0014.}$

KTH Matematik

Exempel: uppskattningen ovan

$X \in Bin(8000000, 10^{-7})$ och Y är $\mathrm{Po}(0.8)$ så gäller att

$$|P(X = k) - P(Y = k)| \leq np^2.$$

$$|P(X = k) - P(Y = k)| \leq 8000000 \cdot 10^{-14} = 8.0000e - 08$$

KTH Matematik

Exempel: check mot MATLAB Statistics Toolbox

>> help binocdf

BINOCDF Binomial cumulative distribution function.
Y=BINOCDF(X,N,P) returns the binomial cumulative distribution function with parameters N and P at the values in X.

The size of Y is the common size of the input arguments. A scalar input functions as a constant matrix of the same size as the other inputs.

The algorithm uses the cumulative sums of the binomial masses.

See also binofit, binoinv, binopdf, binornd, binostat, cdf.

>> binocdf$(4, 8000000, 10\widehat{\ }(-7))$

ans = 0.9986

KTH Matematik

Urnor igen

Vi påminner om urnmodellerna. Vi hade en urna med kulor av två slag: v vita och s svarta. Vi drog n kulor ur urnan slumpmässigt.
Sätt $A =$ "Man får k vita kulor i urvalet".
Dragning utan återläggning:

$$P(A) = \frac{\binom{v}{k}\binom{s}{n-k}}{\binom{v+s}{n}}.$$

Dragning med återläggning:

$$P(A) = \binom{n}{k}\left(\frac{v}{v+s}\right)^k\left(\frac{s}{v+s}\right)^{n-k}.$$

KTH Matematik

Hypergeometrisk fördelning

Antag att vi har N enheter, där proportionen p, dvs Np stycken, har egenskapen A. Drag ett slumpmässigt urval om n stycken enheter. Sätt $X =$ antalet enheter i urvalet med egenskapen A.
I termer av urnmodellen för dragning utan återläggning gäller $Np = v$ och $N(1-p) = s$ om $A =$ "vit kula". Således fås

$$p_X(k) = P(X = k) = \frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}},$$

för $0 \leq k \leq Np$ och $0 \leq n - k \leq N(1-p)$.
Man säger att X är $\text{Hyp}(N, n, p)$-fördelad.

Hypergeometrisk fördelning

X är Hyp(N, n, p)-fördelad, man kan visa att

$$E(X) = np \quad \text{och} \quad V(X) = \frac{N-n}{N-1}\, np(1-p).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Approximationer, Hyp(N, n, p)

Om n/N är någolunda liten, så verkar det troligt att det inte spelar så stor roll om vi drar med återläggning eller ej.
Vi har

$$\frac{\binom{Np}{k}\binom{N(1-p)}{n-k}}{\binom{N}{n}} = \frac{Np!}{k!(Np-k)!}\frac{N(1-p)!}{(n-k)![N(1-p)-(n-k)]!}\frac{n!(N-n)!}{N!}$$

$$= \frac{n!}{k!(n-k)!}\frac{Np!(N(1-p)!(N-n)!}{(Np-k)![N(1-p)-(n-k)]!N!}$$

$$\approx \frac{n!}{k!(n-k)!}\frac{(Np)^k(N(1-p))^{n-k}}{N^n} = \binom{n}{k}p^k q^{n-k}.$$

Approximationer, Hyp(N, n, p)

Sats

Om X är Hyp(N, n, p)*-fördelad med $n/N \leq 0.1$ så är X approximativt* Bin(n, p)*-fördelad.*

Sats

Om X är Bin(n, p)*-fördelad med $npq \geq 10$ så är X approximativt $N(np, \sqrt{npq})$-fördelad.*

Detta innebär att

$$\left.\begin{matrix} P(X \leq k) \\ P(X < k) \end{matrix}\right\} \approx \Phi\left(\frac{k - np}{\sqrt{npq}}\right).$$

KTH VETENSKAP OCH KONST
KTH Matematik

Approximation av Hyp(N, n, p)

Av detta följer att Hyp$(N, n, p) \approx N(np, \sqrt{npq})$ om $n/N \leq 0.1$ och $npq \geq 10$. Det räcker dock att kräva $\frac{N-n}{N-1}np(1-p) \geq 10$.

KTH VETENSKAP OCH KONST
KTH Matematik

Sammanfattning

$$\overbrace{\approx}^{npq\geq 10} N(np, \sqrt{npq})$$

$$\text{Hyp}(N, n, p) \overbrace{\approx}^{n/N\leq 0.1} \text{Bin}(n, p)$$

$$\overbrace{\approx}^{\frac{N-n}{N-1}np(1-p)\geq 10} N(np, \sqrt{npq})$$

$$\overbrace{\approx}^{p\leq 0.1} \text{Po}(\underbrace{np}_{=\mu}) \overbrace{\approx}^{\mu\geq 15} N(\mu, \sqrt{\mu})$$