Statistikk 1, 10.03.08
Nye begreper

Diskret sannsynlighetsfordelinger

- Binomisk fordeling
  - Uavhengige Bernoulliforsøk
  - Dummyvariable for suksess
  - Antallsparameter n og sannsynlighetsparameter p
  - Binomisk fordeling for antall suksesser
  - Tabell
  - Forventning og varians
- Hypergeometrisk fordeling
  - Tilfeldig utvalg uten tilbakelegging på n fra populason N hvorav M er ”suksesser”
  - Hypergeometriske sannsynligheter for antall ”suksesser” i utvalget
  - Forventing og varians
- Poissonfordeling
  - Mange Bernoulliforsøk med lav p
  - Poissonprosess, punktintensitet
  - Poissonsannsynligheter
  - Forventning og varians

Binomisk fordeling

•Bernoulliforsøk er uavhengige forsøk med to mulige utfall: suksess og ikke-suksess.
•Antall suksesser i *n* Bernoulliforsøk med suksess-sannsynlighet *p* er binomisk fordelt med antallsparameter *n* og sannsynlighetsparameter *p*.

$D_i =$ dummyvariabel for suksess i Bernoulliforsøk i, $P\left(D_i = 1\right) = p$, $P\left(D_i = 0\right) = 1 - p$.

$D_1, D_2, \cdots, D_n$ er stokastisk uavhengige og identisk fordelt.

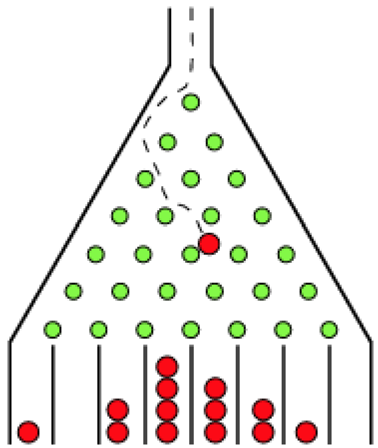
$X = D_1 + \cdots + D_n$ er antall suksesser i n Bernoulliforsøk med suksess-sannsynlighet p

$$P\left(X = x\right) = f(x; n, p) = \binom{n}{x} p^x \left(1 - p\right)^{n-x} \quad x = 0, 1, \cdots, n$$

$$E\left(D_i\right) = p \quad \Rightarrow \quad E\left(X\right) = E\left(\sum_{i=1}^{n} D_i\right) = \sum_{i=1}^{n} E\left(D_i\right) = np$$

$$\operatorname{var}\left(D_i\right) = p\left(1 - p\right), \quad \operatorname{cov}\left(D_i, D_j\right) = 0 \; i \neq j \quad \Rightarrow$$

$$\operatorname{var}\left(X\right) = \operatorname{var}\left(\sum_{i=1}^{n} D_i\right) = \sum_{i=1}^{n} \operatorname{var}\left(D_i\right) + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{cov}\left(D_i, D_j\right) = np\left(1 - p\right)$$



Galtonbrett (quincunx)

Tolvbarnsfamilier etter antall jenter (x) i Saksen i Tyskland 1889 (A. Geissler)

x	observert	forventet	avvik
0	7	2.3	4.7
1	45	26.1	18.9
2	181	132.8	48.2
3	478	410.0	68.0
4	829	854.2	-25.2
5	1112	1265.6	-153.6
6	1343	1367.3	-24.3
7	1033	1085.2	-52.2
8	670	628.1	41.9
9	286	258.5	27.5
10	104	71.8	32.2
11	24	12.1	11.9
12	3	0.9	2.1

Det er 7+45+…+3=6115 familier i materialet. De har til sammen 12x6115=73380 barn. Av disse er 0x7+1x45+…+12x3=35280 jenter. Estimert sannsynlighet for jentefødsel er dermed *p*=35280/73380=0.4808.

Under binomisk modell er sannsynligheten for at en tolvbarnsfamilie skal ha x jenter:

$$P\left(X = x\right) = f(x;12,p) = \binom{12}{x} p^x \left(1-p\right)^{12-x} \quad x = 0,1,\cdots,12.$$

Forventet antall tolvbarnsfamilier med x jenter er dermed $6115 \cdot f(x;12,p)$.

Disse forventete antallene er gitt i kolonne 3, mens de observerte antallene er i kolonne 2.

Avviket er differensen mellom kolonne 2 og 3.

Passer den binomiske modellen til Geisslers data? Hva karakteriserer avviket?

Kan kjønn være uavhengig fra fødesel til fødsel i Saksen, og ha samme sannsynlighet 0.48 for jente?

Hypergeometrisk fordeling

Alle de $\binom{N}{n}$ mulige utvalgene på n fra en populasjon på N hvorav M er "spesielle" er like sannsynlige

$D_i =$ dummyvariabel for "spesiell" i trekning i, $P\left(D_i = 1\right) = \frac{M}{N}$, $P\left(D_i = 0\right) = \frac{N-M}{N}$.

$D_1, D_2, \cdots, D_n$ er stokastisk avhengige variable, men de er identisk fordelt.

$X = D_1 + \cdots + D_n$ er antall "spesielle" i det tilfeldige utvalget på n

$$P\left(X = x\right) = h(x;n,p) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{n}{M}\binom{N-n}{M-x}}{\binom{N}{M}} \quad x = 0,1,\cdots,n.$$

$$E\left(D_i\right) = \frac{M}{N} = p \quad \Rightarrow \quad E\left(X\right) = E\left(\sum_{i=1}^{n} D_i\right) = \sum_{i=1}^{n} E\left(D_i\right) = n\frac{M}{N} = np$$

$$\operatorname{var}\left(D_i\right) = \frac{M}{N}\frac{N-M}{N} = p\left(1-p\right),$$

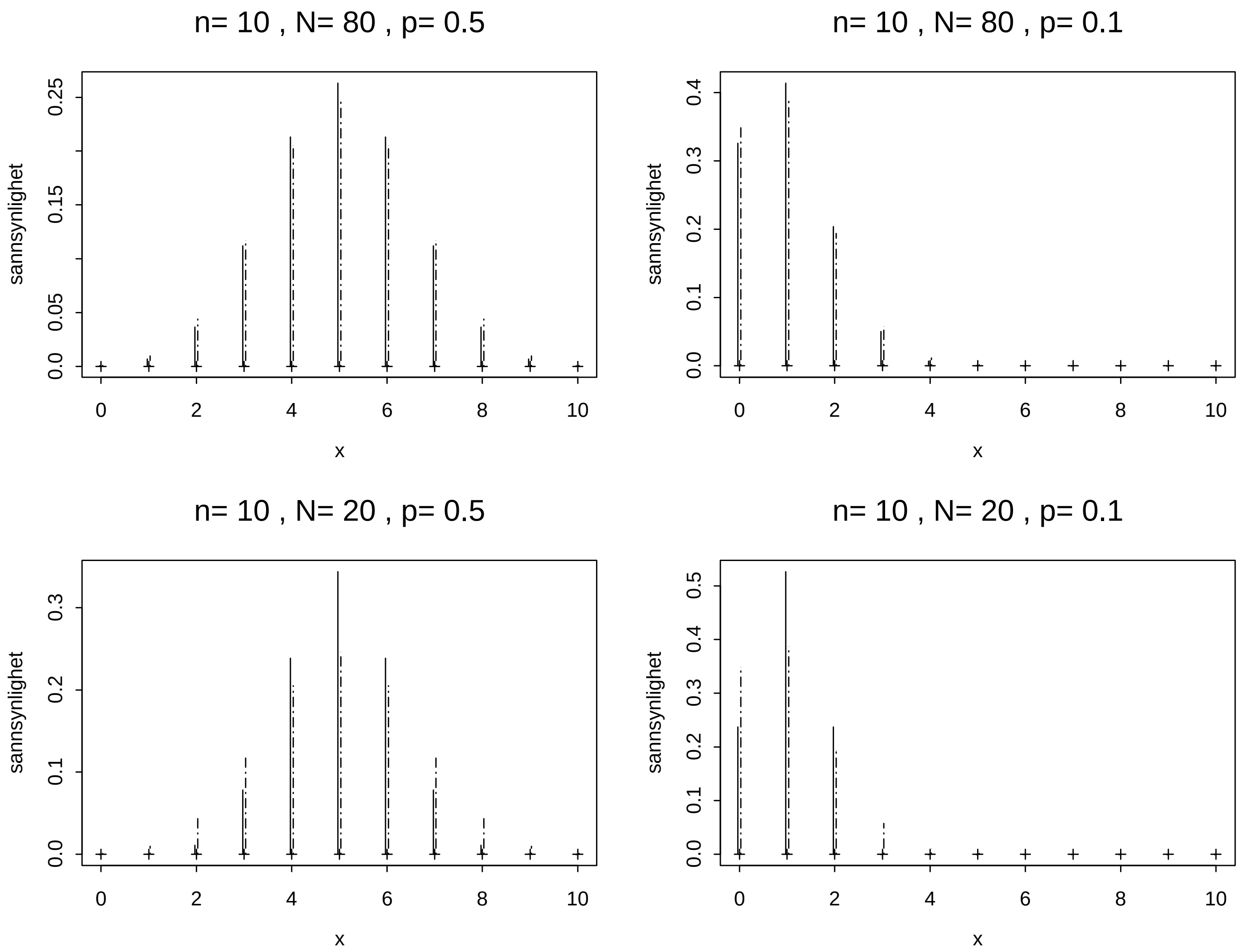
$$P(D_1 = 1, D_2 = 1) = P(D_i = 1, D_j = 1) = \frac{M\left(M-1\right)}{N\left(N-1\right)} = E\left(D_i D_j\right) \quad i \neq j$$

$$\Rightarrow \operatorname{cov}\left(D_i, D_j\right) = \frac{M\left(M-1\right)}{N\left(N-1\right)} - \left(\frac{M}{N}\right)^2 = -\frac{1}{N-1}\frac{M}{N}\frac{N-M}{N} = -\frac{1}{N-1}p\left(1-p\right),$$

Dermed

$$\operatorname{var}\left(X\right) = \operatorname{var}\left(\sum_{i=1}^{n} D_i\right) = \sum_{i=1}^{n}\operatorname{var}\left(D_i\right) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\operatorname{cov}\left(D_i, D_j\right) = np\left(1-p\right) - 2\frac{n\left(n-1\right)}{2}\frac{1}{N-1}p\left(1-p\right)$$

$$= np\left(1-p\right)\frac{N-n}{N-1}.$$



Hypergeometrisk fordeling (hele pinner) sammenliknet med binomisk fordeling (delte pinner).

Poissonfordeling

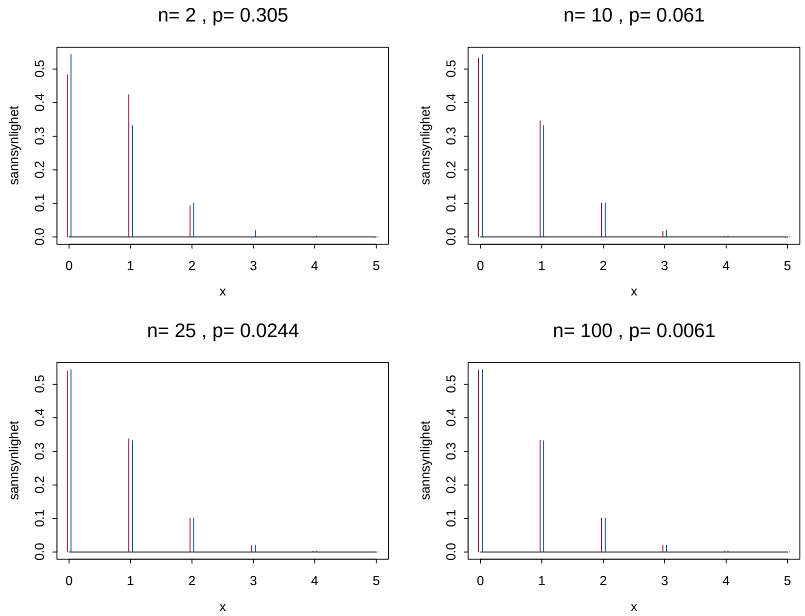
Når $np = \lambda$ mens $n \to \infty,\ p \to 0$ vil de binomiske sannsynlighetene

$$\binom{n}{x} p^x (1-p)^{n-x} = \frac{np(n-1)p\cdots(n-x+1)p}{x!}\left(1-\frac{\lambda}{n}\right)^n (1-p)^{-x} \to \frac{\lambda^x}{x!}e^{-\lambda} = f(x;\lambda) \quad x = 0,1,2,\cdots.$$

$$EX = np = \lambda,$$

$$\text{var}(X) = np(1-p) \to \lambda.$$

Pinnediagram for binomisk fordeling (røde pinner til venstre) og Poissonfordeling (blå pinner til høyre) med samme forventning np=lambda=0.61:



Botkiewicz (1898) telte opp årlig antall døde *x* av hestespark i hvert av 10 armekorps i den Prøysiske hær over en 20-årsperiode. Observert antall armekorpsår med *x* døde er gitt i kolonne 2. Det er i alt 122 slike dødsfall, og estimert forventet antall døde pr. armekorpsår er 122/200=0.61:

x	observert	forventet	avvik
0	109	108.7	0.3
1	65	66.3	-1.3
2	22	20.2	1.8
3	3	4.1	-1.1
4	1	0.6	0.4
5	0	0.1	-0.1