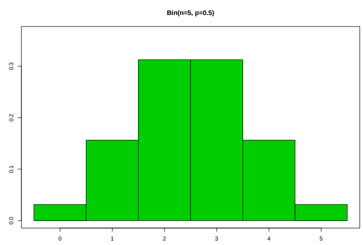
Kapittel 5: Diskrete sannsynlighetsfordelinger

TMA4240 Statistikk (F2 og E7)

5.1-5.4: Uniform, binomisk og hypergeometrisk fordeling:
onsdag 8.september 2004

5.5-5.6: Negativ binomisk og Poisson fordeling:
mandag 13.september 2004.



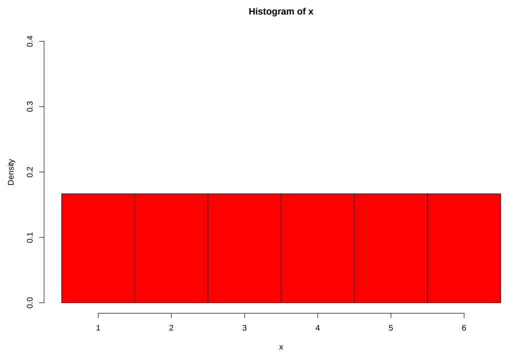
Arbeidshverdag for statistiker

- Samle inn data (subjektive eller objektive) under usikkerhet.
- Beskrive, forstå, trekke konklusjoner og gjøre beslutninger fra data.
- Trenger å bruke en SV med tilhørende fordeling som beskriver vårt fenomen.
- Hvilken fordeling?
  - se på prosessen som har “skapt dataene”.
  - se grafisk på data og studere fordelingens form.
- Derfor: kap. 5 og 6: beskrive viktige fordelinger for å
  - lære situasjoner er fordelingen passer
  - forstå hvordan $f(x)$ fremkommer
  - se hva $E(X)$ og $\text{Var}(X)$ er og forstå hvorfor
  - lære å regne ut $f(x)$, $F(x) = P(X \leq x)$, $P(X \geq x)$.
- Deretter: anslå parametere i fordelingene og trekke konklusjoner under usikkerhet (kap. 9-11).

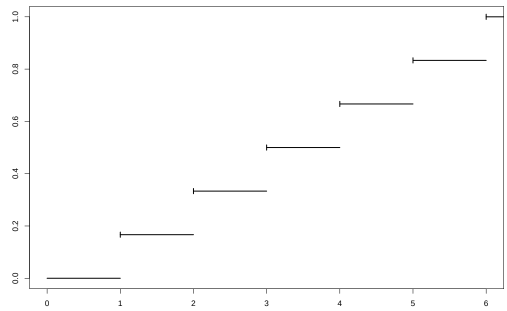
5.2 Diskret uniform fordeling

Diskret uniform fordeling: Hvis den stokastiske variabelen X antar verdiene $x_1, x_2, ..., x_k$ med lik sannsynlighet så er X diskret uniformt fordelt med fordeling

$$f(x;k) = \frac{1}{k}, \qquad x = x_1, x_2, ..., x_k$$



$f(x)$



$F(x)$

TEO 5.1: Forventning og varians i den diskrete uniforme fordelingen $f(x;k)$ er

$$\mu = E(X) = \frac{\sum_{i=1}^{k} x_i}{k} \quad \text{og} \quad \sigma^2 = \text{Var}(X) = \frac{\sum_{i=1}^{k}(x_i-\mu)^2}{k}$$

Midtveiseksamen

- Fra høsten 2004 vil det i TMA4240 bli innført tellende skriftlig midtveiseksamen.
- Denne vil bli gitt i form av en flervalgsoppgave (“multiple choice”) bestående av n spørsmål som alle har m svaralternativer. Studentene må velge et svaralternativ for hvert spørsmål (det er således ikke lov å svare “blankt” på et spørsmål).
- For å få karakter bedre enn F (36%) må minst 8 spørsmål være korrekt besvart.
- Ole lurer på om han skal la være å lese til midtveiseksamen og heller velge tilfeldige svaralternativer på alle spørsmålene (han vil da ikke engang lese oppgaveteksten før han svarer). Før han bestemmer seg, ber han en studiekamerat regne ut hvor stor sannsynlighet han da har for få bedre enn F.
- La X være antall korrekte svar Ole får på de n spørsmålene ved å velge tilfeldig.
- Hvilken fordeling har X?
- Dersom antall spørsmål er 20 ($n = 20$) og antall svaralternativer er 2 ($m = 2$), hva er sannsynligheten for at Ole får bedre enn F hvis han velger å svare tilfeldig på alle spørsmålene (dvs. $P(X \geq 8)$)? Finn også $P(X \geq 8)$ for $m = 4$ og $m = 5$.

Tabell over binomisk fordeling

n	x	p: .05	.10	.15	.20	.25	.30	.40	.50	.60	.70	.80	.90
20	0				.012								
	1				.069								
	2				.206								
	3				.411								
	4				.630								
	5				.804								
	6				.913								
	7	1.000	1.000	.994	.968	.898	.772	.416	.132	.021	.001	.000	.000
	8				.990								
	9				.997								
	10				.999								
	⋮				⋮								
	20				1.000								

$$f(x) = P(X = x) = b(x; n, p) \quad, n = 20, p = 0.20$$

$$P(X \leq 7) = 0.968$$

5.3 Binomisk fordeling

Bernoulli prosess: Et Bernoulli eksperiment (prosess) har følgende egenskaper:

1. Eksperimentet består av n gjentatte forsøk.
2. Hvert forsøk undersøker man om en hendelse A inntreffer (suksess) eller ikke (A'=failure).
3. Sannsynligheten for hendelsen A (suksess) kaller vi p, og denne er den samme fra forsøk til forsøk.
4. De n gjentatte forsøkene er uavhengige av hverandre.

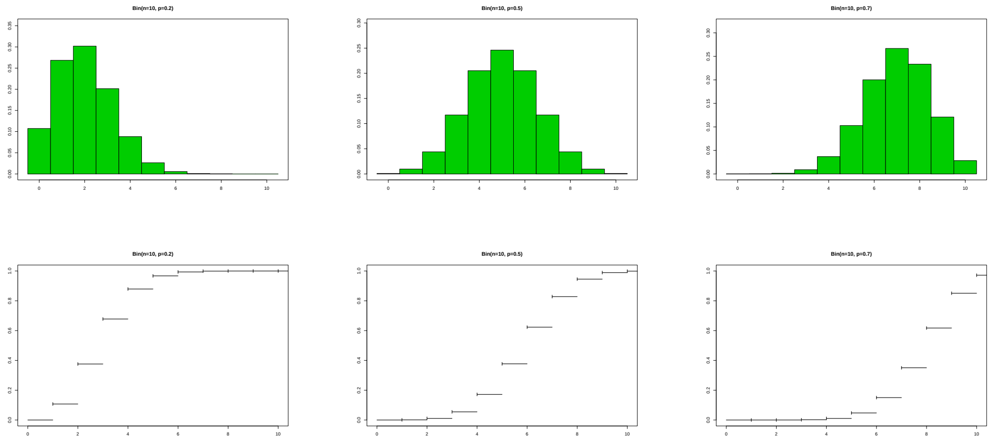
- Et Bernoulli eksperiment kan resultere i
  - hendelsen A (suksess) med sannsynlighet p og
  - komplementet av hendelsen A (A'=failure) med sannsynlighet $1 - p$.
- La den stokastiske variablen X være antall ganger hendelsen A (suksess) inntreffer på de n uavhengige forsøkene.
- Sannsynlighetsfordelingen til X kalles *binomisk fordeling* og er gitt ved

$$b(x; n, p) = \binom{n}{x} p^x (1-p)^{(n-x)}, \qquad x = 0, 1, ..., n$$

- Kumulativ fordeling: $F(x) = P(X \leq x)$ finnes ved tabelloppslag.

Binomisk fordeling (forts.)

$n = 10, p = 0.2$ $\quad$ $n = 10, p = 0.5$ $\quad$ $n = 10, p = 0.7$



TEO 5.2: Forventning og varians i binomisk fordeling $b(x; n, p)$ er

$$\mu = E(X) = np \quad \text{og} \quad \sigma^2 = \text{Var}(X) = np(1-p)$$

Urne med kuler [v1]

- Defi nisjon: $p = \dfrac{\text{antall røde kuler}}{\text{antall kuler}}$
- Prosedyre: Utfør n ganger
  - trekk en kule tilfeldig
  - registrer fargen
  - legg kula tilbake
- Da er antallet røde kuler binomisk fordelt.

Urne med kuler [v2]

- Definisjoner:
  - $p_1 = \frac{\text{antall hvite kuler}}{\text{antall kuler}}$
  - $p_2 = \frac{\text{antall sorte kuler}}{\text{antall kuler}}$
  - $p_3 = \frac{\text{antall blå kuler}}{\text{antall kuler}}$
  - $p_4 = \frac{\text{antall røde kuler}}{\text{antall kuler}}$
- Prosedyre: Utfør n ganger
  - trekk en kule tilfeldig
  - registrer fargen
  - legg kula tilbake
- Da er
  - antallet hvite kuler
  - antallet sorte kuler
  - antallet blå kuler
  - antallet røde kuler
- multinomisk fordelt.

Multinomisk fordeling

Multinomisk fordeling: Et forsøk kan resultere i

- k mulige utfall $A_1, A_2, \ldots, A_k$, med sannsynligheter
- $p_1, p_2, \ldots, p_k$.

La de stokastiske variablene $X_1, X_2, \ldots, X_k$ representere antall ganger utfallene $A_1, A_2, \ldots, A_k$ opptrer i n uavhengige forsøk. Sannsynlighetsfordelingen til $X_1, X_2, \ldots, X_k$ kalles *multinomisk fordeling* og er gitt ved

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \binom{n}{x_1, x_2, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

med $\sum_{i=1}^{k} x_i = n$ og $\sum_{i=1}^{k} p_i = 1$.

Urne med kuler [v3]

- Urne med N kuler, der n av de er røde.
- Trekk kuler uten å legge de tilbake, og la X være antall røde kuler som er trukket.
- Hva er fordelingen til X?
  - Bernoulli forsøk?
  - Gjentatte forsøk - JA
  - "Suksess eller fiasko" - Trekker rød eller ikke - JA
  - Sannsynlighet for å trekke rød konstant - NEI
    - første trekning - $P(R) = \frac{n}{N}$
    - andre trekning - $P(R) = \frac{n-1}{N-1}$ eller $P(R) = \frac{n}{N-1}$
- Gitt at det er 20 kuler ($N = 20$), der 7 av de er røde ($n = 7$). Hva er sannsynligheten for å trekke 4 røde på 10 forsøk?

5.4 Hypergeometrisk fordeling

Hypergeometrisk eksperiment: har følgende egenskaper:

1. Vi har en mengde av N enheter. Av de N enhetene så klassifiseres k som hendelsen A (suksess) og $N - k$ som komplementet av hendelsen A (A'=failure).
2. Et tilfeldig utvalg av størrelse n trekkes uten tilbakelegging fra de N enhetene.

Antallet ganger, X, som hendelsen A (suksess) inntreffer er da en *hypergeometrisk stokastisk variabel*.

Hypergeometrisk fordeling: En hypergeometrisk stokastisk variabel,

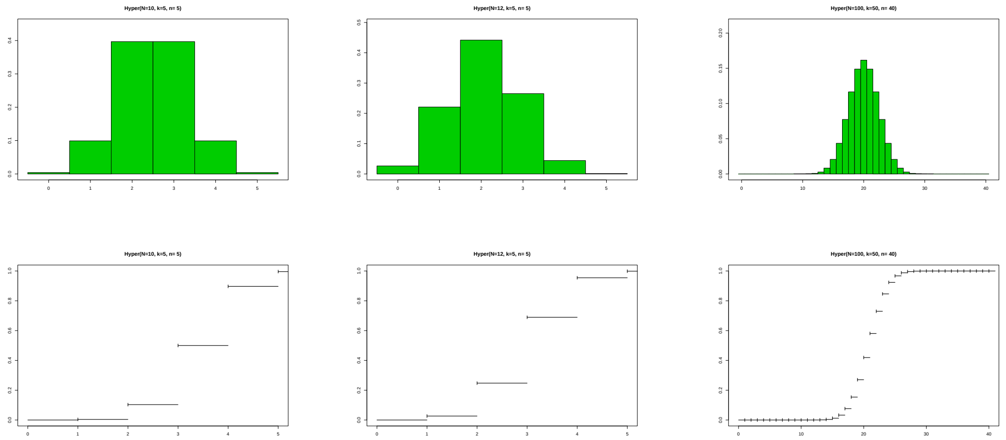
- X, angir antallet ganger hendelsen A (suksess) inntreffer i et hypergeometrisk eksperiment
- der n enheter trekkes fra N enheter,
- der k av de N enheter er klassifisert som hendelsen A (suksess) og
- $N - k$ som komplementet av hendelsen A (A'=failure).

Sannsynlighetsfordelingen til X kalles en *hypergeometrisk fordeling* og er gitt ved

$$h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad x = 0, 1, 2, ..., n$$

Hypergeometrisk fordeling (forts.)

$N = 10, k = 5, n = 5$ $\quad$ $N = 12, k = 5, n = 5$ $\quad$ $N = 100, k = 50, n = 40$



TEO 5.3: Forventning og varians i den hypergeometriske fordelingen $h(x; N, n, k)$ er

$$\mu = E(X) = \frac{nk}{N} \quad \text{og} \quad \sigma^2 = \text{Var}(X) = \frac{N-n}{N-1} \cdot n \cdot \frac{k}{N}(1 - \frac{k}{N})$$

Hypergeometisk og binomisk fordeling

- Hvis n er liten i forhold til N ($\frac{n}{N} \leq 0.05$), så vil sammensetningen av de N enhetene endres lite under trekningen.
- Dermed kan $\frac{k}{N}$ sees på som den binomiske sannsynligheten p.
- Dermed kan binomisk fordeling sees på som en "stor populasjon" versjon av hypergeometrisk fordeling.

Multivariabel hypergeometrisk fordeling

Multivariabelt hypergeometrisk eksperiment: har følgende egenskaper:

1. Et tilfeldig utvalg av størrelse n trekkes uten tilbakelegging fra N enheter.
2. Av de N enhetene så klassifi seres a_1 i cellen A_1, a_2 i cellen A_2,..., a_k i cellen A_k.
3. Av de n enhetene så klassifi seres x_1 i cellen A_1, x_2 i cellen A_2,..., c_k i cellen A_k.

Sannsynlighetsfordelingen til $X_1, X_2, ..., X_k$ kalles *multivariabel hypergeometrisk fordeling*

$$f(x_1, x_2, ..., x_k; a_1, a_2, ..., a_k, n) = \frac{\binom{a_1}{x_1}\binom{a_2}{x_2}\cdots\binom{a_n}{x_n}}{\binom{N}{n}}$$

med $\sum_{i=1}^{k} x_i = n$ og $\sum_{i=1}^{k} a_i = N$.

Kakelotteri

- 300 lodd fordelt på 3 farger (100 av hver)
- 9 vinnerlodd, 3 av hver farge (33,66,99)
- Vi kjøper 5 lodd.
- To strategier:
  - trekk 5 lodd blant de 300 loddene
  - trekk 5 lodd av samme farge
- Hvilken strategi gir størst sjanse for å vinne?