5. Diskrete sannsynlighetsfordelinger.

5.1. Definisjoner og begreper.

Vi har nå vært gjennom to hovedemner:

- Bearbeiding av *innsamlede* data, med grafiske framstillinger og beregning av viktige parametre.
- Sannsynlighetsregning.

Vi skal nå begynne å kombinere kunnskaper fra disse to områdene. Istedenfor å *samle inn* data og sette opp frekvenstabeller, skal vi bruke våre kunnskaper i sannsynlighetsregning til å sette opp *forventede* frekvenstabeller ut fra teoretiske betraktninger. Deretter skal vi beregne statistiske parametre basert på denne teoretiske frekvenstabellen.

Men først må vi definere et par begreper, som helt sikkert vil virke tåkete i starten. Du husker kanskje at et *stokastisk eksperiment* er en aktivitet der vi ikke kan forutsi utfallet (kaste terning, trekke kort fra en kortstokk, osv.). Da definerer vi:

En ***stokastisk variabel*** er en funksjon som tilordner ett bestemt tall til ethvert utfall av et stokastisk eksperiment.

Her er det på sin plass med et par merknader:

- Siden utfallet av et stokastisk eksperiment ikke kan forutsies, er det heller ikke mulig å forutsi verdien av en stokastisk variabel.
- Formuleringen "…en *variabel* er en *funksjon* som…" virker påfallende. Egentlig burde vi sagt *stokastisk funksjon* istedenfor *stokastisk variabel*. Men *variabel*-navnet er så innarbeidet at det ikke har noen hensikt å prøve å endre terminologien.
- Hvilket *tall* er det som tilordnes de enkelte utfallene? Det er faktisk opp til oss å tilordne fornuftige tall til hvert utfall.
- Vi bruker ofte X (stor bokstav) til å betegne en stokastisk variabel, mens x (liten bokstav) betegner *verdien* til variabelen. Skrivemåten " $X = x$ " betyr da at den stokastiske variabelen X har verdien x.

La oss se på et eksempel. Vårt statistiske eksperiment går ut på å kaste to pengestykker. Vi bestemmer oss for at en stokastisk variabel X skal være antall kron. Vi får da denne sammen-hengen mellom utfall og X, der K står for "kron" og M står for "mynt":

Utfall	M - M	M - K	K - M	K - K
Verdi for X	0	1	1	2

Siden X kun kan ha de heltallige verdiene 0, 1 og 2, sier vi at X er en *diskret stokastisk variabel*:

En stokastisk variabel som kun kan ha bestemte (vanligvis heltallige) verdier, sies å være en ***diskret stokastisk variabel***. Den kan imidlertid ha *uendelig mange* diskrete verdier.

Her kommer et viktig begrep til:

Sannsynlighetsfordelingen til en diskret stokastisk variabel *X* skrives $P(X = x)$ eller kortere $P(x)$, og angir sannsynligheten for at den stokastiske variabelen *X* skal ha verdien *x*.

I vårt kron-mynt-eksempel kan vi anta at hvert utfall er like sannsynlig, slik at hvert utfall har sannsynlighet 0.25. Sannsynlighetsfordelingen for *X* (antall kron) blir da:

Verdi av *X*	0	1	2
Sannsynlighet $P(X = x)$	0.25	$2 \cdot 0.25 = 0.50$	0.25

Du innser sikkert at vi må stille to krav til en sannsynlighetsfordeling $P(X = x)$:

$0 \le P(X = x) \le 1$ for alle verdier av *X* fordi en sannsynlighet må ligge i området $[0,1]$.
$\sum P(X = x) = 1$ når vi summerer over alle verdier av *X*, fordi samlet sannsynlighet for alle utfallene må være lik 1.

Dessuten har vi at:

Når $P(X = x)$ er en diskret sannsynlighetsfordeling, er

$$P(a \le X \le b) = \sum_{x=a}^{x=b} P(X = x).$$

Den *kumulative sannsynligheten* er sannsynligheten for at en stokastisk variabel *X* skal ha en verdi som er mindre enn eller lik en bestemt verdi *x*. Den er gitt ved:

$$P(X \le x) = \sum_{n=-\infty}^{x} P(X = n)$$

5.2. Forventningsverdi og varians.

Et lite innledende eksempel: Dersom du slår i hjel litt dødtid med å kaste to pengestykker noen hundre ganger, og notere antall ”Kron” i hvert forsøk, kan du beregne middelverdien av antall ”Kron” for forsøket ditt som

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

der x_i er antall ”Kron” i kast nr. *i* (d.v.s. at x_i er enten 0, 1 eller 2) og *n* er antall kast. Men hvis du er litt smart, vil du først telle opp antall ganger det blir 0 ”Kron” og kalle dette antallet for f_0. Antall ganger det blir 1 ”Kron” kalles f_1 og antall ganger det blir 2 ”Kron” kalles f_2. Da kan beregningen av $\overline{x}$ forenkles til

$$\overline{x} = \frac{1}{n}\left(f_0 \cdot 0 + f_1 \cdot 1 + f_2 \cdot 2 \right).$$

Men dette kan omformes til

$$\overline{x} = \frac{f_0}{n} \cdot 0 + \frac{f_1}{n} \cdot 1 + \frac{f_2}{n} \cdot 2 = P(0) \cdot 0 + P(1) \cdot 1 + P(2) \cdot 2$$

fordi $\frac{f_0}{n}$ er sannsynligheten for å få 0 ”Kron” og tilsvarende for de andre sannsynlighetene, alt basert på *eksperimentelle* data.

Nå har du neppe tid til å utføre dette eksperimentet i praksis. Du vil heller sette opp en *teoretisk modell* av forsøket, der du benytter at $P(0) = 0.25$, $P(1) = 0.50$ og $P(2) = 0.25$. Da kan du beregne middelverdi etter samme formel som ovenfor, og får

$$\overline{x} = 0.25 \cdot 0 + 0.50 \cdot 1 + 0.25 \cdot 2 = \underline{\underline{1.00}}.$$

Den middelverdien av *X* som du får på grunnlag av din *teoretiske* modell, kaller vi gjerne ***forventningsverdi*** (***Expected Value***) og bruker symbolet $\mathrm{E}(X)$ eller μ for å skille den fra middelverdi som er beregnet på grunnlag av *eksperimentelle* data. På samme måte beregner vi *variansen* av *X*, som ofte skrives $\mathrm{Var}(X)$ eller σ^2.

Vi kan samle betraktingene ovenfor slik:

La *X* være en stokastisk variabel med sannsynlighetsfordeling $P(X = x)$. Da er:

Forventningsverdi: $\mathrm{E}(X) = \mu = \sum x \cdot P(X = x)$

Varians: $\mathrm{Var}(X) = \sigma^2 = \sum (x-\mu)^2 \cdot P(X = x) = \sum \left(x^2 \cdot P(X = x) \right) - \mu^2$

I begge formlene summerer vi over alle de verdiene som *X* kan ha. Prøv selv om du får til omformingen av varians-formelen! Standardavviket (som gjerne skrives $\mathrm{Std}(X)$ eller σ) blir som vanlig kvadratrota av variansen.

Et lite eksempel vil kanskje klargjøre bruken av disse formlene:

Eksempel 5.1: Du ”fikser” en gammel terning slik at ”5” ser ut som ”3”, og ”6” ser ut som ”4”. La *X* være det tallet som terningen viser etter ”fiksingen”. Sett opp sannsynlighets-fordelingen til *X*, og beregn $\mathrm{E}(X)$, $\mathrm{Var}(X)$ og $\mathrm{Std}(X)$.

Løsning: Etter å ha ”fikset” terningen, får vi denne sannsynlighetsfordelingen for *X*:

X	1	2	3	4
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$

Forventningsverdien for X blir

$$\mu = \mathrm{E}(X) = 1\cdot\tfrac{1}{6} + 2\cdot\tfrac{1}{6} + 3\cdot\tfrac{2}{6} + 4\cdot\tfrac{2}{6} = \underline{\underline{\tfrac{17}{6}}}.$$

Variansen for X blir

$$\sigma^2 = \mathrm{Var}(X) = \left(1^2\cdot\tfrac{1}{6} + 2^2\cdot\tfrac{1}{6} + 3^2\cdot\tfrac{2}{6} + 4^2\cdot\tfrac{2}{6}\right) - \left(\tfrac{17}{6}\right)^2 = \underline{\underline{\tfrac{41}{36}}}$$

slik at standardavviket for X blir

$$\sigma = \mathrm{Std}(X) = \sqrt{\tfrac{41}{36}} \approx \underline{\underline{1.07}}.$$

Vi tar et eksempel til:

Eksempel 5.2: En gambler tilbyr deg dette spillet: du betaler en innsats til gambleren. Så slås to ("ærlige") terninger. Dersom begge terningene viser 6, får du 10 ganger innsatsen tilbake. Dersom begge terningene viser samme tall (unntatt 6), får du 4 ganger innsatsen tilbake. I alle andre tilfeller taper du innsatsen. Hvor mye kan du forvente å vinne eller tape på dette spillet?

Løsning: Anta at din innsats i et spill er a. La X være din gevinst. Da kan X ha disse verdiene:

$X = -a$ dersom terningene viser ulike tall (du taper innsatsen).

$X = 4a - a = 3a$ dersom terningene viser like tall unntatt to 6-ere.

$X = 10a - a = 9a$ dersom terningene viser to 6-ere.

I de to siste tilfellene må du huske at du allerede har betalt inn innsatsen. Den får du ikke tilbake.

Det er i alt $6^2 = 36$ utfall når vi kaster 2 terninger. Det er kun ett av disse utfallene som gir 2 seksere. Det er 5 utfall som gir like tall (unntatt seksere) på de to terningene. Da er det

$$36 - 1 - 5 = 30$$

utfall som fører til at du taper innsatsen. Vi kan da sette opp denne sannsynlighetsfordelingen for X:

X	$-a$	$3a$	$9a$
$P(X=x)$	$\frac{30}{36} = \frac{5}{6}$	$\frac{5}{36}$	$\frac{1}{36}$

Din forventede gevinst blir da

$$\mu = -a\cdot\tfrac{5}{6} + 3a\cdot\tfrac{5}{36} + 9a\cdot\tfrac{1}{36} = \underline{\underline{-\tfrac{1}{6}a}}.$$

Med andre ord: I det lange løp vil du tape en sjettedel av innsatsen (du hadde vel aldri ventet at gambleren ville la deg vinne i det lange løp?).

5.3. Binomial-fordelingen.

I de eksemplene vi har sett på hittil, har vi laget "skreddersydde" sannsynlighetsfordelinger tilpasset hver enkelt situasjon. Men noen situasjoner er såpass vanlige at det er lurt å lage ferdige sannsynlighetsfordelinger som vi kan hente fram når vi har bruk for dem. ***Binomial-fordelingen*** er en slik sannsynlighetsfordeling.

Vi starter med et innledende eksempel:

Eksempel 5.3: I en urne har du 5 kuler: 2 hvite og 3 svarte. Du trekker 3 kuler *med tilbakelegging*. La X være antall *hvite* kuler som trekkes. Finn $\mathrm{E}(X)$ og $\mathrm{Var}(X)$.

Løsning: Du er sikkert enig i at sannsynligheten for å trekke en hvit kule er $P(\text{Hvit}) = \frac{2}{5}$, og at denne sannsynligheten er konstant fordi vi har tilbakelegging. Sannsynligheten for å trekke en svart kule blir da $P(\text{Svart}) = 1 - \frac{2}{5} = \frac{3}{5}$. Når vi gjentar trekkingen 3 ganger med tilbakelegging, får vi at sannsynlighetsfordelingen for X (antall hvite kuler) blir:

X	0	1	2	3
$P(X = x)$	$\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5} = \frac{27}{125}$	$3\cdot\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5} = \frac{54}{125}$	$3\cdot\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{3}{5} = \frac{36}{125}$	$\frac{2}{5}\cdot\frac{2}{5}\cdot\frac{2}{5} = \frac{8}{125}$

Begrunnelser: Sannsynligheten for å trekke 0 hvite kuler (d.v.s. 3 svarte) må være $\frac{3}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}$.
Du kan trekke 1 hvit (og to svarte) kuler på 3 måter: Du kan trekke hvit kule etterfulgt av to svarte, eller svart – hvit – svart, eller svart – svart – hvit. Sannsynligheten blir

$$\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5} + \frac{3}{5}\cdot\frac{2}{5}\cdot\frac{3}{5} + \frac{3}{5}\cdot\frac{3}{5}\cdot\frac{2}{5} = 3\cdot\frac{2}{5}\cdot\frac{3}{5}\cdot\frac{3}{5}.$$

Tilsvarende resonnement gir de to siste sannsynlighetene i tabellen.
Kontroll: $\frac{27}{125} + \frac{54}{125} + \frac{36}{125} + \frac{8}{125} = 1$.

Da får vi at:

Forventningsverdi $\mu = \sum_{x=0}^{3} x \cdot P(X = x) = 0\cdot\frac{27}{125} + 1\cdot\frac{54}{125} + 2\cdot\frac{36}{125} + 3\cdot\frac{8}{125} = \frac{150}{125} = \underline{\frac{6}{5}}$.

Varians $\sigma^2 = \sum_{x=0}^{3} x^2 \cdot P(X = x) - \mu^2 = 0^2\cdot\frac{27}{125} + 1^2\cdot\frac{54}{125} + 2^2\cdot\frac{36}{125} + 3^2\cdot\frac{8}{125} - \left(\frac{6}{5}\right)^2 = \underline{\frac{18}{25}}$.

Dette er en helt vanlig situasjon. Dersom en andel p av en populasjon har en bestemt egenskap, og du tilfeldig trekker ut n objekter fra populasjonen med tilbakelegging slik at p holder seg konstant, kan du få at 0, 1, 2, …, n av de uttrukne objektene har egenskapen.

Vi oppsummerer og generaliserer situasjonen slik:

Vi får en ***binomisk sannsynlighetsfordeling*** dersom alle disse kravene er oppfylt:

1. Et eksperiment kan kun ha to mulige utfall.
2. Eksperimentet gjentas på samme måte n ganger.
3. Utfallet i ett eksperiment er uavhengig av tidligere utfall.
3. Sannsynligheten for å få ett bestemt utfall er konstant lik p.

Dersom betingelsene for en binomisk fordeling er oppfylt, kan vi uten videre sette opp:

For en ***binomisk sannsynlighetsfordeling*** gjelder:
Et eksperiment gjentas *n* ganger med konstant sannsynlighet *p* for ”gunstig” utfall. La *X* være antall ganger vi får ”gunstig” utfall. Da er sannsynlighetsfordelingen for *X* gitt ved

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

Begrunnelse for formelen: Sannsynligheten for å få *x* ”gunstige” utfall på rad, etterfulgt av $n-p$ ”ikke gunstige” utfall er $p^x (1-p)^{n-x}$. Men kombinasjonen av *x* ”gunstige” og $n-x$ ”ikke gunstige” utfall kan du få på $\binom{n}{x}$ ulike måter. Sannsynligheten for å få *x* ”gunstige” utfall (og $n-x$ ”ikke gunstige”) av *n* forsøk blir da $\binom{n}{x} p^x (1-p)^{n-x}$.

Vi skal uten bevis sette opp at:

Dersom *X* er binomisk fordelt med sannsynlighet *p* i hvert forsøk, og vi utfører *n* forsøk, er:

Forventningsverdi $\mathrm{E}(X) = \mu = n \cdot p$.

Varians $\mathrm{Var}(X) = \sigma^2 = n \cdot p \cdot (1-p)$.

Hvis vi bruker formlene for forventningsverdi og varians på dataene i eksempel 5.3, får vi at

$$\mu = 3 \cdot \tfrac{2}{5} = \underline{\tfrac{6}{5}}$$

og

$$\sigma^2 = 3 \cdot \tfrac{2}{5} \cdot \left(1 - \tfrac{2}{5}\right) = \underline{\tfrac{18}{25}}.$$

Vi har forutsatt at vi har trekking med tilbakelegging. Men dersom *n* er svært liten sammenliknet med størrelsen av hele populasjonen, kan vi trygt anta at *p* ikke påvirkes i særlig grad av at vi trekker ut noen få objekter fra populasjonen selv om vi ikke har tilbakelegging. Vi bruker derfor binomisk fordeling også i situasjoner der vi har en svært stor populasjon og trekker ut noen få objekter fra denne. Dersom vi trekker ut såpass mange objekter at det er fare for at *p* endres vesentlig, må vi bruke den *hypergeometriske* fordelingen som vi skal se på senere i dette notatet.

La oss ta et lite eksempel på denne bruken av binomisk fordeling:

Eksempel 5.4: La *populasjonen* være alle norske velgere. Anta at *i populasjonen* er 40 % mot norsk medlemskap i EU. Vi velger ut $n = 600$ tilfeldige velgere. La *X* være antall EU-motstandere vi trekker. Finn $\mathrm{E}(X)$, $\mathrm{Var}(X)$ og $\mathrm{Std}(X)$.

Løsning: Sannsynligheten for at en tilfeldig valgt velger skal være EU-motstander er $p = 0.4$. Siden et utvalg på $n = 600$ er mye mindre enn hele populasjonen, kan vi anta at *p* ikke endres selv om vi her i prinsippet har trekking uten tilbakelegging. Da er:

$$\mathrm{E}(X) = \mu = n \cdot p = 600 \cdot 0.4 = \underline{\underline{240}}.$$

$$\mathrm{Var}(X) = \sigma^2 = n \cdot p \cdot (1-p) = 600 \cdot 0.4 \cdot 0.6 = \underline{\underline{144}} \quad \Leftrightarrow \quad \mathrm{Std}(X) = \sigma = \sqrt{144} = \underline{\underline{12}}.$$

Dersom vi utfører en slik spørreundersøkelse i praksis, vil vi altså *i gjennomsnitt* trekke 240 EU-motstandere. Men antallet vil variere rundt denne verdien, og standardavviket for dette antallet er altså 12. Nå har vi tidligere sagt at det er ”relativt normalt” å få verdier innenfor $\mu \pm 2\sigma$. Mer presist kan vi si at i ca. 96 % av alle slike spørreundersøkelser vil vi få en verdi i dette intervallet. I vår spørreundersøkelse må vi derfor forvente ”normale” verdier i intervallet

$$240 \pm 2 \cdot 12 = [216,\ 264].$$

Regnes dette om til prosent, får vi

$$\left[\frac{216}{600} \cdot 100\ \%,\ \frac{264}{600} 100\ \%\right] = \underline{[36\%,\ 44\%]}.$$

Det vil si at selv om andelen EU-motstandere i *populasjonen* er konstant lik 40 %, kan vi ved å trekke ut et tilfeldig utvalg på 400 velgere forvente å få verdier mellom 36 % og 44 %. Vi skal i senere notat gå nærmere inn på slike analyser.

Vi tar et eksempel til, der vi ser på bruk av formelen for binomisk sannsynlighet.

Eksempel 5.5: Ved en eksamen er det $n = 10$ spørsmål. Til hvert spørsmål er det 4 alternative svar, og kun ett av dem er rett. Kandidaten krysser av for det svaret som han tror er rett. For å stå til eksamen, må han svare rett på minst 5 av spørsmålene. Vår kandidat er fullstendig blank. Han har ikke peiling på stoffet, og gjetter vilt på alle de 10 spørsmålene. Hva er sannsynligheten for at han skal stå?
Hint: Bruk funksjonen **`binom.fordeling`** i Excel.

Løsning: Med 4 svar-alternativer på hvert spørsmål, er sannsynligheten for å svare rett $p = \frac{1}{4}$ for hvert spørsmål. La *X* være antall spørsmål der kandidaten gjetter rett. Sannsynligheten for å stå blir da $P(X = 5) + P(X = 6) + \cdots + P(X = 10)$. Du innser sikkert at det er brysomt å beregne dette for hand. Vi bruker heller dataverktøy. Med Excel har vi funksjonen

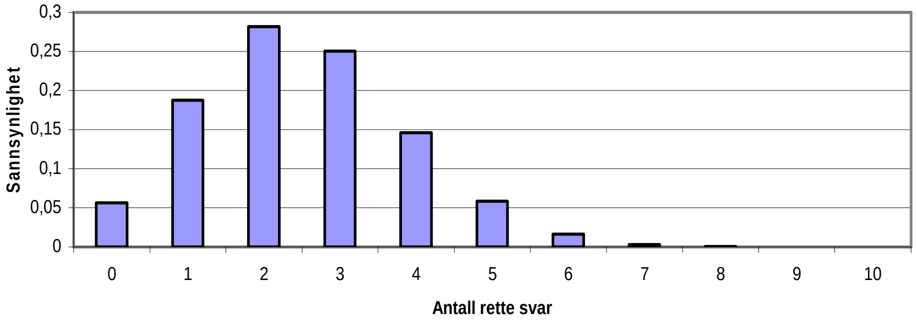
`=binom.fordeling(`*`x`*`;`*`n`*`;`*`p`*`;kumulativ)`

De tre første parametrene er greie. Den siste parameteren, **`kumulativ`**, gir vi verdien **`USANN`** dersom vi ønsker sannsynligheten $P(X = x)$ og verdien **`SANN`** dersom vi ønsker sannsynligheten $P(X \le x)$. I vårt tilfelle er det lurest å si at sannsynligheten for å stå til eksamen er $P(X \ge 5) = 1 - P(X \le 4)$. Vi taster da inn i Excel:

=1-BINOM.FORDELING(4;10;1/4;SANN)

og får svaret 0.078. Det er altså 7.8 % sannsynlighet for at vår totalt blanke kandidat skal stå til en slik eksamen.

Figuren nedenfor viser grafisk sannsynlighetene for å få 0, 1, …, 10 rette svar til eksamen med ren gjetting:



De fleste lærebøker i statistikk har tabeller med binomiske sannsynligheter. Disse tabellene er vanligvis bygd opp med ***kumulative*** sannsynligheter (vår lærebok er et unntak). Hvis man skal finne $P(X=x)$ med en slik tabell, beregner man $P(X \le x) - P(X \le (x-1))$.

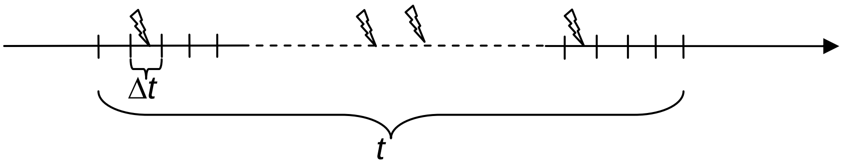
Eksempel: $P(X=5) = P(X \le 5) - P(X \le 4)$.

5.4. Poisson-fordelingen.

Dette er en annen standard-fordeling som ofte forekommer i praksis. Vi bruker denne fordelingen dersom vi har hendelser som fordeler seg *tilfeldig* over et bestemt intervall. Dette intervallet kan være tidsintervall, strekning, flate, eller volum. Mer presist skal vi kreve at sannsynligheten for at en hendelse skal inntreffe i et *lite* intervall Δt er:

- Proporsjonalt med størrelsen av intervallet.
- Uavhengig av om en tilsvarende hendelse har inntruffet i tidligere intervall.

Videre skal vi kreve at sannsynligheten for at to eller flere hendelser skal inntreffe i samme *lille* intervall er neglisjerbar.



Et større intervall *t* kan tenkes bygd opp av mange slike små del-intervall Δt. Dersom forekomsten av hendelser i disse små del-intervallene tilfredsstiller kravene ovenfor, så er antall hendelser *X* i intervallet *t* ***Poisson-fordelt***. Vi kan vise at:

Dersom en stokastisk variabel *X* er Poisson-fordelt, så er

$$P(X=x) = \frac{\mu^x e^{-\mu}}{x!}.$$

Forventningsverdi $\mathrm{E}(X) = \mu$.

Varians $\mathrm{Var}(X) = \sigma^2 = \mu$ slik at standardavviket er $\mathrm{Std}(X) = \sigma = \sqrt{\mu}$.

Du finner en utledning av disse formlene (og litt mer om Poisson-fordelingen) i et eget notat.

Jeg er klar over at denne innledningen kan virke litt ullen. La meg derfor prøve å klargjøre med et par eksempler:

Eksempel 5.6: Du sitter og ser en tippekamp på TV. Dersom det scores i en annen kamp, gis det et scoringsvarsel i form av et "pling". Vi antar at disse "pling"-ene er *tilfeldig* fordelt ut over alle de 90 minuttene kampen varer, og at det "pling"-es 30 ganger i løpet av kampen. Du må dessverre forlate TV-en i 5 minutter. Hva er sannsynlighetene for at du skal gå glipp av 0, 1, 2, 3, … "pling" i løpet av disse 5 minuttene?

Løsning: Det er rimelig å anta at antall "pling" i løpet av de 5 minuttene er Poisson-fordelt. Forventet antall "pling" pr minutt blir da $\frac{30}{90}=\frac{1}{3}$, slik at forventet antall "pling" i løpet av 5 minutter blir $\mu = 5\cdot\frac{1}{3}=\frac{5}{3}$. Da setter vi i gang med formelen ovenfor:

$$P(X=0)=\frac{\left(\frac{5}{3}\right)^0 e^{-\frac{5}{3}}}{0!}\approx \underline{0.189}.$$

$$P(X=1)=\frac{\left(\frac{5}{3}\right)^1 e^{-\frac{5}{3}}}{1!}\approx \underline{0.315}.$$

$$P(X=2)=\frac{\left(\frac{5}{3}\right)^2 e^{-\frac{5}{3}}}{2!}\approx \underline{0.262}.$$

$$P(X=3)=\frac{\left(\frac{5}{3}\right)^3 e^{-\frac{5}{3}}}{3!}\approx \underline{0.146}.$$

$$P(X=4)=\frac{\left(\frac{5}{3}\right)^4 e^{-\frac{5}{3}}}{4!}\approx \underline{0.061}.$$

osv… osv…

Et annet eksempel:

Eksempel 5.6: Vi teller opp 20 ugrasrøtter i en ellers velstelt gressplen på 25 m^2. Vi antar at disse røttene er tilfeldig fordelt utover plenen. La *X* være antall ugrasrøtter innenfor et tilfeldig avmerket areal på 1 m^2. Sett opp sannsynlighetsfordelingen for *X*.

Løsning: Det er rimelig å anta at antall ugrasrøtter er Poisson-fordelt med

$$\mu=\frac{20}{25}=0.8.$$

Vi får:

$$P(X=0)=\frac{0.8^0 e^{-0.8}}{0!}\approx \underline{0.449}.$$

$$P(X=1)=\frac{0.8^1 e^{-0.8}}{1!}\approx \underline{0.359}.$$

$$P(X=2)=\frac{0.8^2 e^{-0.8}}{2!}\approx \underline{0.144}.$$

$$P(X=3)=\frac{0.8^3 e^{-0.8}}{3!}\approx \underline{0.038}.$$

osv… osv…

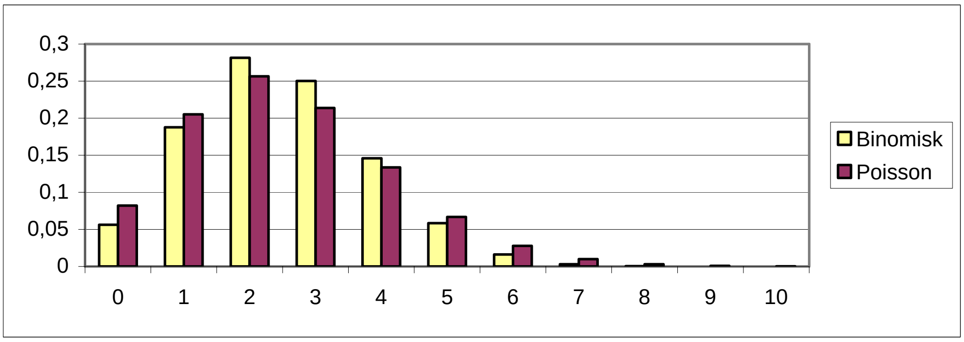
Vi kan også komme fram til Poisson-fordelingen ved å ta utgangspunkt i binomisk fordeling. La *n* være antall små-interavall, mens *p* er sannsynligheten for å få en hendelse i et slikt lite intervall. Så gjør vi små-intervallene stadig mindre, slik at *n* stadig øker samtidig som *p* avtar mens $\mu = n \cdot p$ holdes konstant. Da kan vi vise at når $n \to \infty$ samtidig som $p \to 0$ mens $\mu = n \cdot p$ er konstant, vil binomialfordelingen nærme seg Poisson-fordelingen.

Som en røff regel kan vi anta at Poisson-fordelingen er en god tilnærmelse til binomisk fordeling dersom $n \geq 100$ samtidig som $\mu = n \cdot p \leq 10$. Disse kravene kan variere noe fra lærebok til lærebok.

Som eksempel på bruk av Poisson-fordeling som tilnærming til binomisk fordeling, kan vi gå tilbake til eksempel 5.5 med studenten som stilte fullstendig blank til eksamen med flervalgs-spørsmål. I dette eksemplet er $n = 10$ mens $\mu = n \cdot p = 10 \cdot \frac{1}{4} = \frac{5}{2}$, slik at kravene for "god" tilnærming egentlig ikke er oppfylt. Ved hjelp av Excel setter vi opp grafen nedenfor, der søylene for Poisson-fordeling er framkommet fra funksjonen

=POISSON(*x*;5/2;USANN)

der *x* er antall rette svar.



Du ser at Poisson-fordelingen gir rimelig bra tilnærmelse til binomisk fordeling.

Poisson-fordelingen brukes til å studere en mengde situasjoner, for eksempel:

- Antall branner i en by i et bestemt tidsrom.
- Antall trafikkuhell på en bestemt veistrekning.
- Antall defekte enheter i en produksjonsserie.
- Antall bakterier i en ml vann.
- Antall parasitter på en oppdrettslaks.
- …

Ved hjelp av Poisson-fordelingen kan man bl.a. vurdere om en endring i antallet eller et avvik fra tillatte verdier ligger innenfor de grensene for statistiske fluktuasjoner som man normalt må forvente, eller om det er grunn til å tro at det virkelig foreligger en reell endring.

5.5. Andre diskrete sannsynlighetsfordelinger.

Vi har nå sett på to av de vanligste diskrete sannsynlighetsfordelingene. Men det fins en mengde andre slike fordelinger, tilpasset bestemte formål. Jeg skal kort omtale to slike.

5.5.1. Uniform fordeling.

Når du kaster en "ærlig" terning, er alle utfall like sannsynlig. Med 6 mulige utfall blir

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{for } x \in \{1,2,3,4,5,6\} \\ 0 & \text{ellers} \end{cases}.$$

En slik sannsynlighetsfordeling der sannsynligheten er lik for alle verdier innenfor et intervall og null ellers kalles ***uniform***.

For treningens skyld kan vi beregne forventningsverdi og standardavvik for "terning-fordelingen". Vi får:

$$\mathrm{E}(X) = \mu = \sum x \cdot P(X = x) = 1\cdot\tfrac{1}{6} + 2\cdot\tfrac{1}{6} + 3\cdot\tfrac{1}{6} + 4\cdot\tfrac{1}{6} + 5\cdot\tfrac{1}{6} + 6\cdot\tfrac{1}{6} = \tfrac{1}{6}\cdot 21 = \underline{\underline{\tfrac{7}{2}}}.$$

$$\mathrm{Var}(X) = \sigma^2 = \sum\left(x^2 \cdot P(X = x)\right) - \mu^2$$

$$= 1^2\cdot\tfrac{1}{6} + 2^2\cdot\tfrac{1}{6} + 3^2\cdot\tfrac{1}{6} + 4^2\cdot\tfrac{1}{6} + 5^2\cdot\tfrac{1}{6} + 6^2\cdot\tfrac{1}{6} - \left(\tfrac{7}{2}\right)^2 = \underline{\tfrac{35}{12}}$$

slik at standardavviket blir

$$\mathrm{Std}(X) = \sqrt{\tfrac{35}{12}} \approx \underline{\underline{1.7}}.$$

5.5.2. Hypergeometrisk fordeling.

Binomisk fordeling framkom ved trekking *med* tilbakelegging. Noen ganger har vi bruk for sannsynlighetsfordelingen for trekking *uten* tilbakelegging. Denne fordelingen kalles ***hypergeometrisk fordeling***, og er slik:

Vi starter med *N* objekter, der *r* av disse har en bestemt egenskap.
Vi trekker *n* objekter tilfeldig uten tilbakelegging. La *X* være antall av de uttrukne objektene som har egenskapen. Sannsynlighetsfordelingen til *X* er da

$$P(X = x) = \frac{\binom{r}{x}\cdot\binom{N-r}{n-x}}{\binom{N}{n}}.$$

Vi kan vise at:

$$\mathrm{E}(X) = \mu = \frac{r}{N}\cdot n.$$

$$\mathrm{Var}(X) = \sigma^2 = \frac{r}{N}\cdot\frac{N-r}{N}\cdot\frac{N-n}{N-1}\cdot n.$$

Det er ikke meningen at du skal gå rundt og huske disse formlene! Men du bør vite at de eksisterer.