Temaopgave i statistik for matematik B og A

Indhold

Temaopgave i statistik – matematik B og A

Denne temaopgave tager udgangspunkt i opgaver om terningekast og meningsmålinger med omdrejningspunkt i binomialfordelingen. Målet med opgaven er at føre dig gennem stoffet og at ruste dig til den mundtlige eksamen.

Temaopgaven består i at besvare de 10 opgaver der står i dette materiale.

Kapitlet 'Statistik' s. 177-204 i Matema10k for B-niveau danner grundlag for rapporten. Vi anbefaler at du anvender kapitlet ved at slå begreber m.m. op når du støder på noget ubekendt i rapporten.

Undervejs får du bl.a. brug for nogle Excel-ark som du finder færdiglavede i vores Fronterrum. For at kunne lave rapporten skal du derfor kunne arbejde med Excel på din computer.

God fornøjelse.

Introduktion: Hvad er en binomialfordeling?

Binomialfordelingen er i statistik en måde at beregne den teoretiske fordeling af antallet af udfald i et eksperiment der opfylder:

1. Der skal være hvad man betegner som et "basiseksperiment".
2. Basiseksperimentet skal kunne fortolkes som havende præcis to udfald (som man kan betegne som "succes" og "fiasko").
3. Man kender sandsynligheden for "succes" (og dermed også for "fiasko" fordi summen af de to sandsynligheder giver 1).
4. Man gentager basiseksperimentet et antal gange.

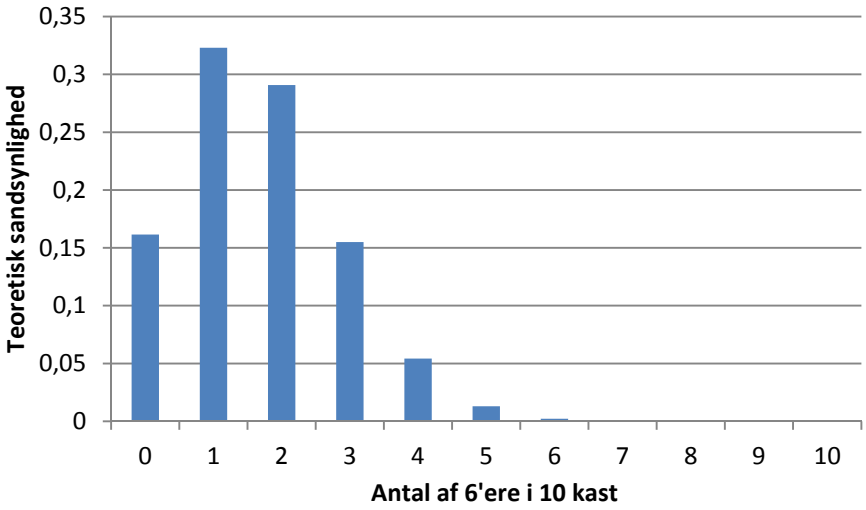
Et klassisk eksempel på en binomialfordeling er at beregne fordelingen af sandsynligheden for antal 6'ere hvis man kaster fx 10 gange med en terning. Vi ved at sandsynligheden for at få en 6'er er $\frac{1}{6}$ ved kast med en regulær terning. Vi kan kalde udfaldet "en 6'er" for succes. Dermed er det fiasko at få 1, 2, 3, 4 eller 5.

Man kan fx tegne et søjlediagram over hvor stor sandsynligheden er for at få 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 eller 10 seksere hvis man kaster 10 gange med en terning. Med søjlediagrammet anskueliggør man fordelingen af succeser (se nedenfor).

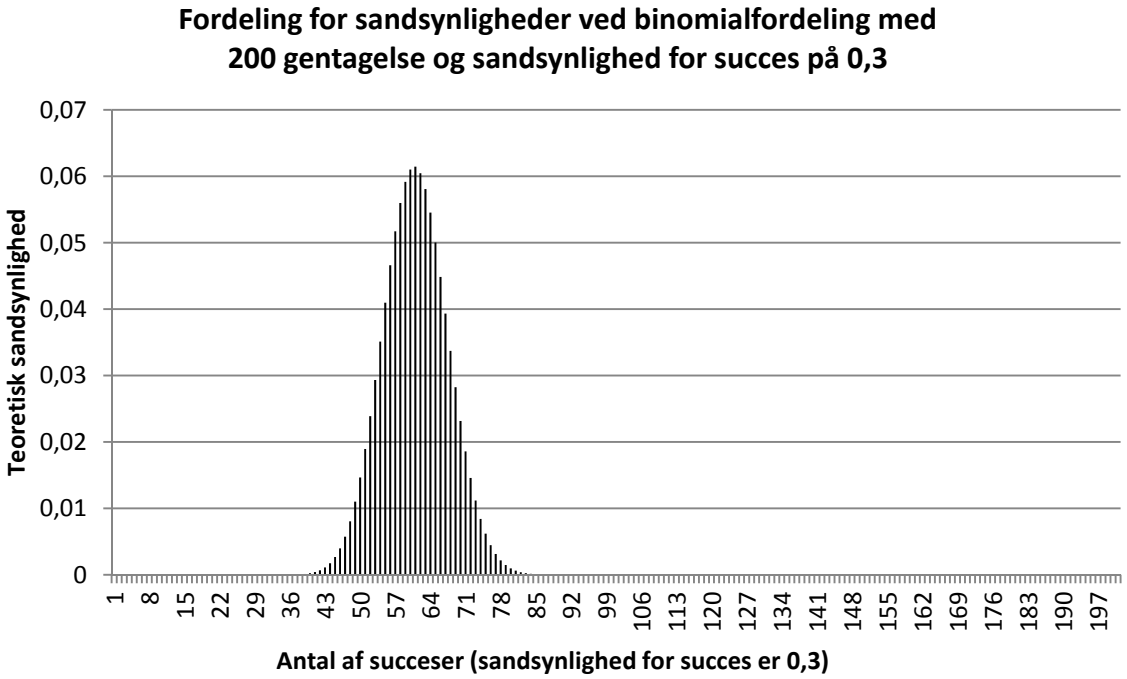
Fordi fordelingen kun gælder hele antal kast, kalder vi det en diskret fordeling. En kontinuert fordeling vil kræve at udfaldene udgøres af alle reelle tal.

Hvis man har mange gentagelser, betragter man hyppigt fordelinger som kontinuerte på trods af at de egentlig er diskrete. Ved at betragte dem som kontinuerte kan man opskrive en funktionsforskrift for fordelingsfunktionen, og det giver store fordele i bearbejdningen.

Sandsynlighederne for de forskellige antal 6'ere hvis man kaster 10 gange med en regulær terning, bliver:



Følgende figur viser en fordeling for en anden binomialfordeling:



Her er søjlerne placeret så tæt så det nærmest ser ud som om der er en kurve. Hvad der egentlig er en diskret fordeling, tager sig nærmest ud som en kontinuert fordeling.

Opgave 1. Kast med 12 terninger 20 gange – i praksis

Find 12 (rigtige) terninger, og kast dem på en gang. Tæl 1'ere og 2'ere. Hvis du kun har én terning, så kast den 12 gange efter hinanden.

a) Hvor mange 1'ere og 2'ere (tilsammen) vil du forvente, der er ud af de 12 terninger? Begrund svaret.
b) Vil du få det samme antal 1'ere og 2'ere (tilsammen) hver gang du kaster de 12 terninger?
c) Kast de 12 terninger 20 gange i alt, og tæl op hver gang hvor mange 1'ere og 2'ere du opnår.

d) Lav en tabel i et regneark (Excel) over hyppigheden for de enkelte mulige tilfælde.

e) Udarbejd dernæst et histogram som kopieres over i den egentlige rapport. Kommentér histogrammets udseende.
Er du førstegangsbruger af regneark, er der bagerst i rapportformuleringen vedlagt en tastevejledning – og i Fronter ligger flere vejledninger til Excel.

Sandsynligheder

Intuitivt kender vi sandsynligheder. Vi ved intuitivt at sandsynligheden for at få en 1'er ved at kaste en almindelig terning er $\frac{1}{6} = 16{,}7\%$. Tilsvarende ved vi at sandsynligheden for at trække et billedkort fra et almindeligt kortspil uden jokere er $\frac{12}{52} = \frac{3}{13} = 23{,}1\%$, mens sandsynligheden for at trække en spar er $\frac{13}{52} = \frac{1}{4} = 25\%$.

En sandsynlighed er en teoretisk størrelse. Sandsynligheden er et bud på hvad resultatet af en hændelse vil blive. Derfor skelner vi i materialet her mellem

- statistiske resultater af forsøg eller simuleringer
- teoretiske sandsynligheder.

Opgave 2. Kast med 12 terninger – teoretisk

I denne opgave skal du undersøge et kast med 12 terninger hvor man *spørger om hvor mange 1'ere og 2'ere der tilsammen* ***forventes - teoretisk set***.

a) Forsøget kan beskrives ved binomialfordelingen *b(12,1/3)*. Forklar hvorfor det er binomialfordelingen med sandsynligheden 1/3 der skal anvendes.

b) Åbn *"Excelark til illustration af binomialsandsynligheder"*, sæt parameterværdierne til *n=12* og *p=1/3*. Kopier det pindediagram som du får frem, ind i din besvarelse. Kommentér diagrammet.

c) Alle de forskellige sandsynligheder kan beregnes ved hjælp af binomialformlen $P(k) = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$. Brug formlen til at eftervise at *P(2) = 0,1272* og *P(3) = 0,212* (disse værdier fremkommer af regnearket). Beregn den teoretiske middelværdi og spredning.
(Bemærk at der er en trykfejl i binomialformlen i bogen s. 194)

d) Forklar hvad der skal forstås ved: *"forventes - teoretisk set"*?

Opgave 3. Kast med 12 terninger 20 gange - simulering

Hvis man skulle kaste med flere terninger eller lave mange flere gentagelser end blot 20 som i opgave 1, er det en fordel at have en tilfældighedsgenerator som kan generere tilfældige hele tal mellem 1 og 6. Dette gøres her ved hjælp af en simulering i Excel.

Åbn regnearket "20 kast" - det ligger i Fronter. Vælg den udgave der passer til din udgave af Excel. Hvis du åbner arket med Excel 2003, bliver du spurgt om en bestemt makro skal aktiveres, og du skal svare ja (der kan opstå problemer pga. sikkerhedsniveauet på din pc – spørg i så fald). Hvis du åbner med Excel 2007, skal du først svare "Ja" til at åbne filen. Derefter skal du klikke på "Indstillinger" over regnearket. Her skal du vælge "Aktiver indholdet".

Når du har åbnet Excelarket, skal du vælge "Gem som" og gemme arket på din egen pc (fx med navnet "Opgave 3").

a) Orienter dig i regnearkets konstruktion, og besvar følgende spørgsmål: Hvorfor er der 20 tal i søjle D? Hvad betyder tallet i celle D7? Hvad betyder tallene i søjle H? Hvordan er frekvenserne i søjle J beregnet?
b) Lad regnearket finde gennemsnittet af de 20 kast – se tastevejledningen bagerst. Noter resultatet.
c) Lav nu 10 simuleringer. En simulering køres ved at skrive et tilfældigt tal i cellen B5 og trykke ENTER. For hver simulering skal du notere gennemsnittet. Beregn i alle tilfælde hvor mange procent gennemsnittet afviger fra den sande (teoretiske) middelværdi $\mu = n \cdot p$. Her benyttes formlen $\frac{|\bar{x}-\mu|}{\mu} \cdot 100\%$ hvor $\bar{x}$ er det beregnede gennemsnit. Lav en tabel over de procentvise afvigelser.
d) Klip to (gerne vidt forskellige) pindediagrammer for simuleringer ind i besvarelsen, og kommentér forskellen mellem disse og det teoretiske pindediagram fra opgave 2.

Opgave 4. Kast med 12 terninger 1000 gange – simulering

Åbn regnearket "1000 kast".

a) Kør simuleringen 10 gange. For hver simulering skal du notere gennemsnittet.
b) Beregn i alle tilfælde hvor mange procent gennemsnittet afviger fra den sande middelværdi.
c) Kommentér disse værdier og sammenhold med resultatet fra opgave 3c.
d) Udvælg vilkårligt to af pindediagrammerne for simuleringerne ind i besvarelsen, og kommentér forskellen mellem disse og det teoretiske pindediagram. Kommentér endvidere forskellen mellem disse og simuleringerne med de 20 kast.

Opgave 5. Spørgeskema til 50 personer - simulering

I denne opgave (og i opgave 6) skal du undersøge et spørgeskema til 50 personer. Det kunne eksempelvis være at man spurgte 50 elever på et gymnasium om de kunne tænke sig at der blev spillet livemusik til den næste fest. Svarene kan være interessante i sig selv, men matematisk set er det mest interessante om de adspurgte i virkeligheden svarer i overensstemmelse med hele populationens mening. Populationen er samtlige elever på det pågældende gymnasium. Opgaven går derfor ud på at finde svar på spørgsmålet: Med hvilken sikkerhed er det forsvarligt kun at adspørge et lille udsnit af populationen?

Da spørgeskemaet kun rummer svarmulighederne ja eller nej, er dette problem matematisk set det samme problem som kast af de 12 terninger hvor vi talte 1'ere og 2'ere (jf. opgave 3 og 4). I denne opgave skal vi se på hvordan svarene kan variere fra én meningsmåling til en anden. Bemærk at vi i regnearket der omtales nedenfor, spørger 50 personer i alt 1000 gange.

Åbn regnearket "Spørgeskema til 50 personer", og orienter dig i regnearkets konstruktion.

a) Hvad forstås der ved tallene i celle D11, H11 og J11?
b) Tegn et pindediagram (se tastevejledningen bagerst i denne rapport) som viser hyppigheden af de forskellige antal ja-svar. Kør simuleringen to gange, og kopier de to pindediagrammer over i din besvarelse. Kommentér diagrammerne.

c) Beregn den matematiske middelværdi, og sammenlign gennemsnittene med middelværdien i de to simuleringer i b (samme fremgangsmåde som i opgave 4).
d) Gør rede for hvad skal der laves om i regnearket hvis du skal simulere en situation hvor 20 % af befolkningen svarer ja. Foretag ændringen, og kør simuleringen to gange. Kopier et pindediagram for hver simulering ind i din besvarelse, og kommentér kort resultaterne.
e) Hvordan vil det se ud hvis kun 4 % svarer ja? Her skal du igen ændre i regnearket, og du skal derefter køre simuleringen to gange. Kopier et pindediagram for hver simulering ind i din besvarelse, og kommentér kort resultaterne.

Vurdering af data

I opgave 3 og 4 har du forhåbentlig kunnet konkludere at man får en mere sikker vurdering af middelværdien (og dermed også af sandsynligheden for at slå en 1´er eller en 2´er) når man gennemfører forsøget 1000 gange end hvis man gennemfører det 20 gange. I det ene tilfælde har man kastet 12 000 terninger og i det andet $20 \cdot 12 = 240$ terninger. Jo flere kast, jo mere vil de tilfældige udsving udligne hinanden, og jo tættere vil man dermed komme på den sande middelværdi. I det følgende skal vi se på hvordan ***usikkerheden*** i bestemmelse af sandsynligheder ud fra stikprøver knytter sig til antal stikprøver og spredningen.

Boks 1: Beregning af spredning

Som du har set i ovenstående øvelser, får man ikke det samme antal 1'ere og 2'ere hver gang man kaster med tolv terninger. Vi skulle forvente fire 1'ere eller 2'ere.

For at angive hvor langt væk fra gennemsnittet vore resultater kan forventes at ligge, definerer man en størrelse der hedder *spredningen*. Denne betegnes med det græske bogstav σ (sigma) og beregnes ved

$$\sigma = \sqrt{n \cdot p \cdot (1-p)}$$

hvor *n* her er antallet af terninger og *p* er sandsynligheden (i decimaltal, eksempelvis 33 % = 0,33) for at få det ønskede resultat.

Denne formel gælder altid for en *binomialfordeling*.

Boks 2: 95%-konfidensinterval

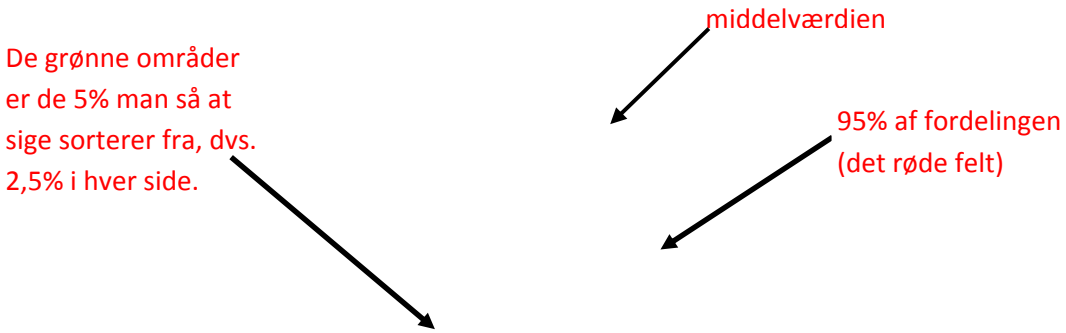
Lad os antage at vi udfører et *binomialforsøg* som et terningekast. Ud fra dette forsøg bestemmes et gennemsnit $\bar{x}$ og spredning σ. Hvis man har *tilstrækkeligt mange* observationer (*tilstrækkeligt mange* kast), gælder med god tilnærmelse at 95 % af observationerne vil ligge i intervallet

$$[\bar{x} - 2 \cdot \sigma \,;\, \bar{x} + 2 \cdot \sigma]$$

Med andre ord er det 95 % sikkert at den sande middelværdi ligger i $[\bar{x} - 2 \cdot \sigma \,;\, \bar{x} + 2 \cdot \sigma]$.

Dette interval kaldes 95%-*konfidensintervallet* for μ.

Man præciserer ikke hvad man mener med "tilstrækkeligt mange" – men det viser sig at være en rimelig tilnærmelse når $n > 30$, forudsat at p ikke er for tæt på 0 eller 1.



Boks 4: Usikkerheden på *p*

Usikkerheden på sandsynlighedsparameteren p er defineret som $\Delta p = \pm\frac{\sigma}{n}$.

Konfidensintervallet for sandsynlighedsparameteren p er direkte knyttet til konfidensintervallet for μ idet $\mu = n \cdot p$. Dermed kan p med 95 % sikkerhed siges at ligge i intervallet

$$[p_{min}, p_{max}] = [p - 2\Delta p, p + 2\Delta p] = \left[\frac{\bar{x} - 2\sigma}{n}, \frac{\bar{x} + 2\sigma}{n}\right]$$

Eksempel

Ved en gallupundersøgelse spørges 1500 mennesker om de ser alt kongestof på TV2. 300 rødmer dybt og indrømmer. Her har vi $n = 1500$, og det bedste bud på middelværdien er netop $\bar{x} = 300$. Det bedste bud på antal ja-sigere i hele befolkningen er derfor $p = \frac{300}{1500} = 0{,}2 = 20\%$.

For at finde usikkerheden på p beregner vi først spredningen: $\sigma = \sqrt{1500 \cdot 0{,}2 \cdot 0{,}8} \cong 15{,}5$. Usikkerheden er da

$$\Delta p = \pm\frac{\sigma}{n} = \pm\frac{15{,}5}{1500} = \pm 0{,}0103.$$

95%-sikkerhedsintervallet er $[0,20 - 2 \cdot 0,0103; 0,20 + 2 \cdot 0,0103] \cong [0{,}18; 0{,}22]$.

Vi kan med 95% sikkerhed konkludere at mellem ca. 18% og 22% af befolkningen ser alt kongestof på Tv 2.

Meningsmålinger hvor stikprøven er repræsentativ for populationen, er et *binomialforsøg* ligesom terningekastene. Antallet af adspurgte svarer til antallet af terninger ved kast. Ja-procenten svarer til sandsynligheden for at få det ønskede.

Boks 5: Gallupundersøgelser

Dette er netop det en Gallupundersøgelse af befolkningens holdning til de politiske partier (og mange andre forhold) går ud på. Gallup udspørger en repræsentativ gruppe af befolkningen, udregner gennemsnitssvaret og beregner intervallet $[\bar{x} - 2 \cdot \sigma\ ;\ \bar{x} + 2 \cdot \sigma]$ for svarene. Man kan nu forvente at hvis hele befolkningen blev spurgt om det samme, så ville man få et svar som ligger inden for det beregnede interval med 95% sikkerhed.

Opgave 6. Usikkerheden på en undersøgelse med 50 adspurgte

Vi ser nu på situationen hvor 50 elever på et gymnasium bliver adspurgt om de kunne tænke sig livemusik til næste fest (se opgave 5).

a) Forklar med ord hvad skal der gælde om udvælgelsen af de 50 elever?

Vi antager nu at de 50 udspurgte er repræsentative for hele gymnasiet med hensyn til spørgsmålet. Antag endvidere at 10 elever svarer ja til spørgsmålet.

b) Da eleverne netop er repræsentative, hvor mange procent af skolens elever ville da svare ja?

Spørgsmålet er nu hvor nøjagtigt vores bud er på at hele skolen vil svare dette. Antag derfor at ja-procenten er 20%.

c) Beregn usikkerheden på antal ja-sigere, og beregn konfidensintervallet. Undersøg om 20 positive svar ligger i 95%-konfidensintervallet.

Opgave 7. Usikkerheden og antallet af adspurgte

I forlængelse af debatten om unges sundhed udvælges 12 gymnasier på Sjælland hvor man spørger eleverne om de er tilfredse med kantinens mad set ud fra et sundhedsmæssigt synspunkt. Vi antager at der på alle 12 gymnasier tilsammen går 10.000 elever. Fra 2 af de 12 gymnasier udvælges 1000 studerende hvor 250 i alt svarer positivt til spørgsmålet.

a) Forklar med ord hvad der skal gælde om udvælgelsen af de 1000 elever.
b) Beregn usikkerheden på ja-procenten idet vi antager at de udvalgte er repræsentative.
c) Nu spørges 2000 elever og ja-procenten er stadigvæk den samme. Beregn usikkerheden på ja-procenten.
d) Samme spørgsmål som i c) med henholdsvis 4000 og 8000 elever.
e) Er der et system i hvordan usikkerheden formindskes? Prøv at beskrive det så præcist muligt med ord.

Opgave 8. Usikkerheden og antallet af adspurgte – teoretisk

a) Brug formlen for spredningen og formlen for usikkerheden, og vis at usikkerheden kan beskrives ved $\Delta p = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}}$.
b) Hvis man ønsker at halvere usikkerheden, hvor mange flere skal man spørge?
c) Passer dette med de opnåede resultater i opgave 7?

Opgave 9. Binomialfordelingens konfidensinterval – teoretisk

En binomialfordeling *b(n,p)* vil for store værdier af *n* ligne en normalfordeling meget. I det foregående har vi anvendt at 95%-konfidensintervallet for middelværdien μ er givet ved $[\bar{x} - 2 \cdot \sigma \,;\, \bar{x} + 2 \cdot \sigma]$. Et relevant spørgsmål at stille sig selv er hvorfor konfidensintervallet egentlig ser sådan ud? Hvorfor er konfidensintervallet for eksempel ikke defineret direkte ud fra standardafvigelsen som $[\bar{x} - \sigma;\ \bar{x} + \sigma]$?

Fortsættes næste side

a) Vis ved beregning at frekvensfunktionen er symmetrisk om 2. aksen.
Vink: en funktion f er lige (dvs. symmetrisk omkring 2. aksen) hvis $f(x) = f(-x)$.

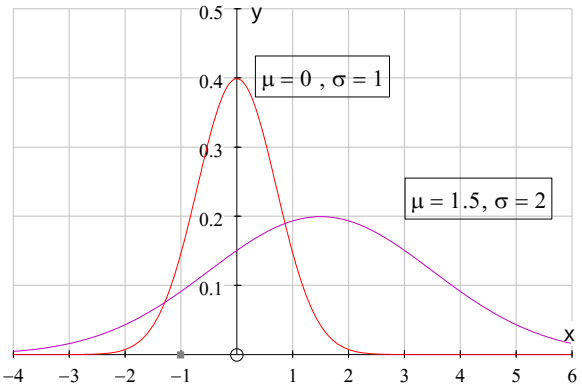
b) Eftervis $\int_{-\infty}^{\infty} f_N(x)dx = 1$ på din grafregner, og fortolk betydningen heraf. Forklar hvordan du bestemmer integralet på din grafregner.

c) Kommentér ud fra velvalgte pindediagrammer fra de tidligere besvarede opgaver hvorfor denne (i form) ligner frekvensfunktionen for normalfordelingen.

d) Bestem for *N(0,1)*
- $P(\mu - \sigma \leq x \leq \mu + \sigma)$
- $P(\mu - 2\sigma \leq x \leq \mu + 2\sigma)$
- $P(\mu - 3\sigma \leq x \leq \mu + 3\sigma)$

e) Redegør dernæst for hvorfor 95%-konfidensintervallet for binomialfordelingen (for store *n*) ser ud som $[\bar{x} - 2 \cdot \sigma\ ;\ \bar{x} + 2 \cdot \sigma]$.

Da standardnormalfordelingen er defineret symmetrisk om 2. aksen, kan det synes mystisk at binomialfordelingen kan tilnærmes normalfordelingen. Men normalfordelingen kan modificeres ved at vælge spredningen og middelværdien. Når middelværdien er μ og spredningen er σ, er frekvensfunktionen for normalfordelingen givet ved



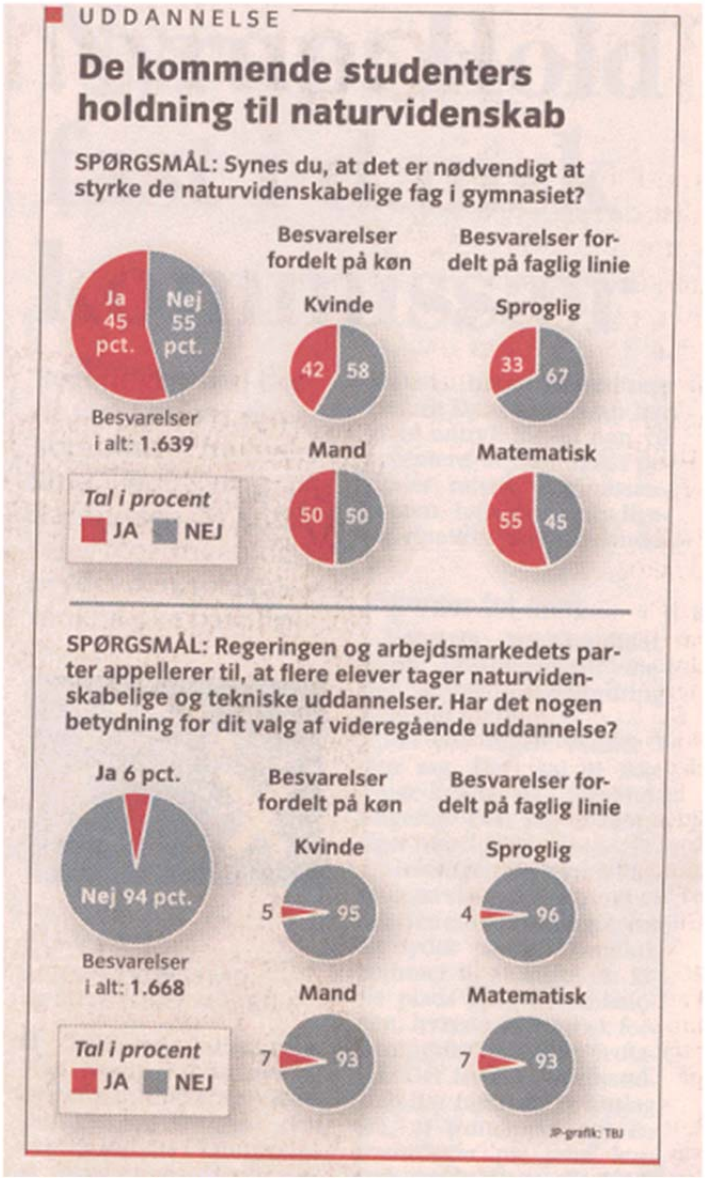
$$f_N(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

På figuren ovenfor ses at frekvensfunktionen har flyttet sig en smule ved den lille justering på spredningen og middelværdien.

Opgave 10. Skal naturvidenskab styrkes?

Materialet til opgaven er udklippene på næste side fra Jyllands-Posten 27.5.2004.

a) Find fejlen i artiklen om undersøgelsen (nærlæs lagkagediagrammerne).
b) Kan man forvente at det er en repræsentativ undersøgelse?
c) Hvor mange elever har svaret ja/nej til spørgsmålet om hvorvidt det er nødvendigt at styrke de naturvidenskabelige fag i gymnasiet?
d) Udregn usikkerheden og konfidensintervallet for både 'ja' og 'nej'.
e) Redegør for at kvinder udgør 62,5 % og mænd udgør 37,5 % af de adspurgte.
Vink: Lad x repræsentere brøkdelen af kvinder og y repræsentere brøkdelen af mænd. På baggrund af informationerne (overvej selv) kan man opstille to ligninger $0{,}55 = 0{,}58 \cdot x + 0{,}50 \cdot y$ og $0{,}45 = 0{,}42 \cdot x + 0{,}50 \cdot y$ (løs de to ligninger).
f) Hvor mange procent udgør hhv. sproglige og matematiske studerende?
g) Hvad kan man konkludere på baggrund af undersøgelsen?


UDDANNELSE
De kommende studenters holdning til naturvidenskab
SPØRGSMÅL: Synes du, at det er nødvendigt at styrke de naturvidenskabelige fag i gymnasiet?
Besvarelser fordelt på køn
Besvarelser fordelt på faglig linie
Ja 45 pct.
Nej 55 pct.
Besvarelser i alt: 1.639
Kvinde
42
58
Sproglig
33
67
Mand
50
50
Matematisk
55
45
Tal i procent
JA
NEJ
SPØRGSMÅL: Regeringen og arbejdsmarkedets parter appellerer til, at flere elever tager naturvidenskabelige og tekniske uddannelser. Har det nogen betydning for dit valg af videregående uddannelse?
Ja 6 pct.
Nej 94 pct.
Besvarelser i alt: 1.668
Besvarelser fordelt på køn
Besvarelser fordelt på faglig linie
Kvinde
5
95
Sproglig
4
96
Mand
7
93
Matematisk
7
93
Tal i procent
JA
NEJ
JP-grafik: TBJ


STUDENT 2004
Morgenavisen Jyllands-Posten fokuserer i en række artikler på de kommende studenter.
Avisen har kontaktet samtlige gymnasier i Vejle og Fyns amter. Undersøgelsen omfatter alene elever fra det almene gymnasium.
I Vejle Amt har 6 ud af 8 gymnasier deltaget, og i Fyns Amt har 10 ud af 11 gymnasier deltaget.
Deltagelse:
Antal 3.g elever på de 16 gymnasier, der har deltaget i undersøgelsen: 2.210
Antal 3.g elever der har besvaret spørgeskemaet: 1.677
Svarprocent: 76 pct.
Tidligere artikler i serien er blevet bragt: 6/5, 13/5, 24/5, 25/5, 26/5
Læs mere på www.jp.dk
Artikler og spørgeskemaundersøgelsen er udarbejdet af: SANNE GRAM, MIA FRANCIS NIELSEN og LISBETH BJERRE

Tastevejledning til Excel

Gennemsnit

Skal man beregne gennemsnit af alle tal i en kolonne, fx fra E2 til E20, markeres en tom celle, fx nedenunder, og der skrives "=middel(e2:e20)"

Kopiering af formel

Marker cellen hvor din formel står. Sæt markøren i nederste højre hjørne på cellen og markøren bliver til et plus-tegn. Klik med musens venstre knap og hold knappen nede mens du trækker musen ned til alle de celler der skal indeholde formlen og slip så museknappen. Nu beregnes formlen i alle cellerne.

Diagram – histogram

Når du skal indsætte et histogram (stolpediagram) i et regneark skal du have to de kolonner med x'erne og y'erne stående i regnearket. I nedenstående eksempel står tallene i kolonne G3 til 15 og H3 til 15.

Placer markøren i en celle til venstre for dine måledata og tast følgende

- vælg **Indsæt**
- vælg **Diagram**
- Vælg diagramtype **søjle** og den øverste til venstre som undertype
- Vælg **Næste**
- Som dataområde markeres H3 til H15
- Øverst i diagramvinduet vælges fanen **Serie**
- I feltet Navn skrives **Hyppighed**
- Som Kategoriakseetiketter markeres G3 til G15
- Vælg **Næste**
- Som diagramtitel skrives "**Spørgeskema til 50 personer**"
- Vælg **Udfør**