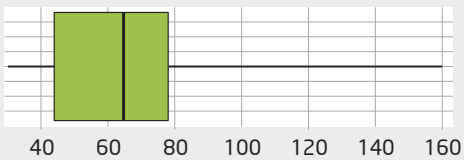
eksemplet med budsjettstørrelse for tippeligaklubbene. Minimumsverdien er her 30, og maksimumsverdien er 159. Vi har dermed følgende fem tall: 30, 44, 64.5, 78, 159. Denne versjonen av boksplott er vist på figur ~~6.9~~ 5.9. Den midterste boksen går fra Q_1 til Q_3 og viser spennet for den midterste halvdelen av verdiene. Så går det to streker ut fra boksen, en ned fra Q_1 til minimumsverdien og en opp fra Q_3 til maksimumsverdien.

Figur 5.9 Vanlig boksplott for tippeligabudsjett



Utbryter: En observasjon som skiller seg ut fra flertallet. På engelsk: «outlier».

Den andre typen boksplott tar også med spesielle observasjoner som kalles *utbrytere*. En utbryter er en «vill» observasjon, en verdi som skiller seg ut ved å ligge langt over eller under storparten av verdiene. Slike verdier kan skyldes eksepsjonelle individer, målefeil eller feil inntasting av data. I noen tilfeller velger vi å fjerne utbryterobservasjonene, mens vi beholder dem i andre tilfeller. Uansett er det nyttig å skanne dataene for slike ekstreme observasjoner. En vanlig måte å definere utbrytere på er å si at det er verdier som ligger lavere enn en gitt avstand ned fra Q_1, eller høyere enn den samme avstanden opp fra Q_3. Hvor langt under nedre kvartil eller over øvre kvartil må observasjonen være for å kalles en utbryter? Jo, en vanlig regel er å bruke 1.5 ganger bredden på boksen i boksplottet. Denne bredden er $Q_3 - Q_1$ og kalles *interkvartilbredden*, forkortet *IK*. En observasjon som ligger høyere enn $Q_3 + 1.5 \cdot IK$, er med andre ord en utbryter. Og tilsvarende er en observasjon som ligger lavere enn $Q_1 - 1.5 \cdot IK$, en utbryter.

Interkvartilbredden: Avstanden mellom øvre og nedre kvartil: $IK = Q_3 - Q_1$

EKSEMPEL 5.2

For budsjettene i tippeligaen fant vi at nedre kvartil var $Q_1 = 44$ millioner kroner, og øvre kvartil var $Q_3 = 78$ millioner kroner. Interkvartilbredden er da $IK = 78 - 44 = 34$ millioner kroner. Vi regner ut 1.5 ganger IK som $1.5 \cdot 34 = 51$. For tippeligaen var $Q_3 = 78$ millioner kr, så utbrytergrensen er $78 + 51 = 129$ millioner kr i øvre ende og $44 - 51 = -7$ millioner kr i nedre ende. Det er altså ingen utbrytere for de lave budsjettverdiene. Men for høye budsjettverdier ser vi at det er en klubb (Rosenborg, RBK) som har et budsjett på 159 millioner kr. Rosenborg representerer altså en utbryter i forhold til budsjettall.

$$P(\text{vinne ved å beholde}) = P(\text{bil bak dør 1} \mid \text{åpner dør 3})$$

$$= \frac{P(\text{bil bak dør 1} \cap \text{åpner dør 3})}{P(\text{åpner dør 3})}$$

$$= \frac{1/6}{1/3+1/6} = \frac{1/6}{1/2} = \frac{2}{6} = \frac{1}{3}$$

7.6

Sannsynligheten for $P(\text{åpner dør 3})$ finner vi ved å lese av siste kolonne på figur ~~7.4~~. Det er to muligheter for at dette inntreffer. Enten er bilen bak dør 2 (med sannsynlighet $\frac{1}{3}$), og programlederen må derfor åpne dør 3 (med sannsynlighet 1). Dette har derfor sannsynligheten $\frac{1}{3} \cdot 1 = \frac{1}{3}$. Den andre muligheten er at bilen er bak dør 1 (med sannsynlighet $\frac{1}{3}$), og programlederen åpner dør 3 (med sannsynlighet $\frac{1}{2}$ fordi han velger tilfeldig mellom dørene 2 og 3). Dette har derfor sannsynligheten $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$. De to disjunkte hendelsene har total sannsynlighet lik $1/3 + 1/6 = 2/6 + 1/6 = 3/6 = 1/2$. Av figur ~~7.4~~ ser vi videre at sannsynligheten er $\frac{1}{6}$ for hendelsen «bil bak dør 1» ∩ «åpner dør 3» = «bilen er bak dør 1, og programlederen åpner dør 3».

7.6

Vi analyserer nå dørbyttet. Vi leser av figur ~~7.4~~ at sannsynligheten er $\frac{1}{3}$ for hendelsen «bilen er bak dør 2, og programlederen åpner dør 3», og får

7.6

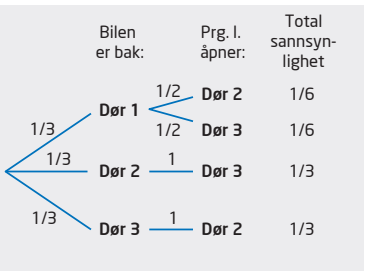
$$P(\text{vinne ved å bytte}) = P(\text{bil bak dør 2} \mid \text{åpner dør 3})$$

$$= \frac{P(\text{bil bak dør 2} \cap \text{åpner dør 3})}{P(\text{åpner dør 3})}$$

$$= \frac{1/3}{1/3+1/6} = \frac{1/3}{1/2} = \frac{2}{3}$$

Som nevnt er framgangsmåten helt tilsvarende hvis spilleren starter med å velge dør 2 eller 3. Argumentet viser at vi øker sjansen for å vinne dersom vi bytter dør.

Figur 7.6 Utfall og sannsynligheter hvis spilleren velger dør 1 fra start



Dette resultatet er intuitivt, siden p er andelen av forsøkene som ender med suksess i det lange løp, og n er antall forsøk vi utfører, forventer vi å få $n \cdot p$ suksesser en binomisk forsøksrekke.

I en binomisk fordeling forventer vi $n \cdot p$ suksesser.

Variansen til en binomisk fordelt tilfeldig variabel X finner vi ved å bruke variansregelen på side 193 og addere alle variansene $p \cdot (1 - p)$, siden de binomiske forsøkene er uavhengige. Dette gir oss

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) = n \cdot p(1 - p)$$

Forventning og varians for binomisk fordeling

Hvis X er en binomisk fordelt variabel hvor $0 \leq p \leq 1$ er sannsynligheten for suksess i hvert av n forsøk, så er

$$\text{E}(X) = n \cdot p \quad \text{og} \quad \text{Var}(X) = n \cdot p \cdot (1 - p)$$

EKSEMPEL 10.8

For X binomisk fordelt med $n = 25$ og $p = 0.2$ så er:

$$\text{E}(X) = 25 \cdot 0.2 = 5 \quad \text{og} \quad \text{Var}(X) = 25 \cdot 0.2 \cdot (1 - 0.2) = 4$$

10.3 Hypergeometrisk fordeling

10.3

Den neste typen diskrete tilfeldige variabler vi skal studere, kan illustreres med krukka i figur 10.5. Vi trekker et visst antall baller fra krukka, *uten tilbakelegging*, og *teller opp* hvor mange av disse kulene som er røde. Den tilfeldige variabelen er altså

$X =$ «antall røde baller når vi trekker n baller uten tilbakelegging»

Sannsynlighetsfordelingen til X kan utledes ved å bruke gunstige delt på mulige. Hvert utvalg er jo like sannsynlig, så det er lett å telle opp antall mulige utvalg totalt. Men hvordan teller vi antall gunstige utvalg? Hvor mange utvalg inneholder det ønskede antallet røde baller? Jo, vi teller først antall utvalg med ønsket antall røde baller, og multipliserer så dette med antall utvalg med de resterende grønne ballene av det totale antallet grønne baller. Det er lettere å forklare dette med et eksempel:

Se side 107 for gunstige delt på mulige.

Forventning og varians for hypergeometrisk fordeling

Hvis X er en hypergeometrisk fordelt variabel, beskrevet av N, n, K og k som over, da er:

$$\mathrm{E}(X) = n \cdot \frac{K}{N} \quad \text{og} \quad \mathrm{Var}(X) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

EKSEMPEL 10.13

For X hypergeometrisk fordelt med $N = 52$, $n = 5$ og $K = 13$, så er:

$$\mathrm{E}(X) = 5 \cdot \frac{13}{52} = 1.25 \quad \text{og} \quad \mathrm{Var}(X) = 5 \cdot \frac{13}{52} \cdot \frac{52-13}{52} \cdot \frac{52-5}{52-1} \approx 0.86$$

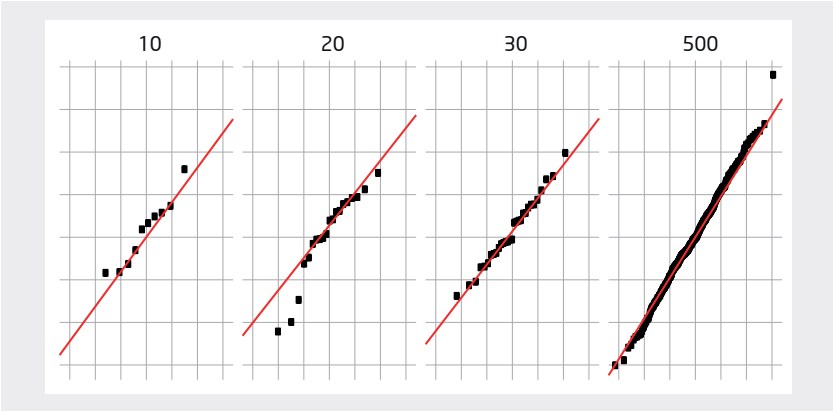
10.3.1 En anvendelse med hypergeometrisk fordeling (*)

Vi har hittil illustrert hypergeometrisk fordeling med enkle eksempler. Nå tar vi et mere praktisk eksempel.

Vi skal estimere hvor stor populasjon av abbor vi har i et tjern. Dette er et viktig, siden en lav bestand gjør at abboren er truet. Dersom det er høy sannsynlighet for at bestanden er mindre enn 100 individer, blir det fiskeforbud i en toårsperiode slik at bestanden kan bygge seg opp igjen. For å undersøke abborbestanden fanger viltforvalteren totalt 30 abbor i tjernet, som merkes og slippes ut igjen. En stund senere fanger viltforvalteren 30 abbor på nytt. Av de tretti hun fanget i andre runde, har syv merke fra første runde.

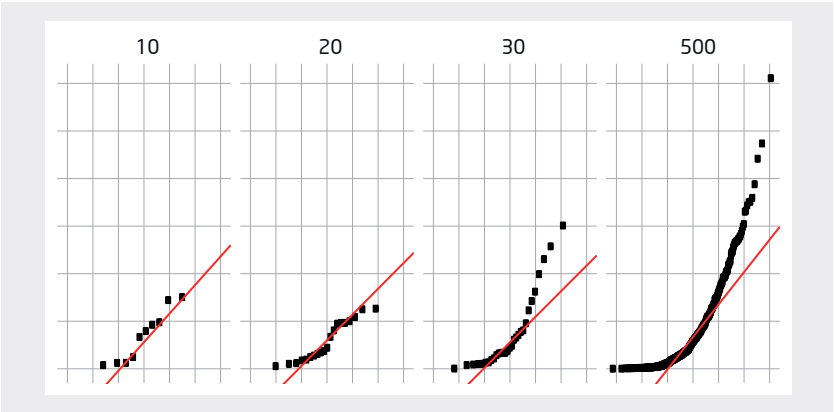
Vi kan nå bruke hypergeometrisk fordeling til å anslå den totale bestanden i tjernet. Den totale bestanden, som er ukjent, er tallet N i den hypergeometriske fordeling. Vi ønsker altså å anslå størrelsen til N. Vi kjenner antall gunstige, $K = 30$ (det er fisken vi har merket), $n = 30$ (antall fisk fanget i andre runde) og $k = 7$, som er antall fisk merket av de totalt $n = 30$ som ble fanget i andre runde. Vi regner ut formelen for hypergeometrisk fordeling for ulike verdier av N (med $K = 30$, $n = 30$ og $k = 7$). Anslaget vårt på den totale bestandsstørrelsen er den verdien N som gir høyest sannsynlighet for å gjenfange syv av tretti fisk. Vi bruker datamaskin til å regne ut formelen for mange verdier av N. Resultatet er vist på figur ~~10.6~~. Det er $N = 128$ som gir høyest sannsynlighet for å gjenfange 7 av 30, og dette er vårt beste anslag for N basert på informasjonen vi har tilgjengelig. Ifølge dette estimatet trenger vi derfor ikke innføre fiskeforbud.

10.4



Figur 11.11 QQ-plott for normalfordelte observasjoner. De fire diagrammene viser tilfeldige utvalg med størrelsene n = 10, 20, 30 og 500.

Vi kan sammenlikne de normalfordelte observasjonene fra figur 11.11 med QQ-plott der utvalgene er trukket fra en fordeling som *ikke* er normalfordelt. Se figur 11.12.



Figur 11.12 QQ-plott for observasjoner som ikke er normalfordelte. De fire panelene viser tilfeldige utvalg med størrelsene n = 10, 20, 30 og 500.

I de to største utvalgene ser vi klart at QQ-plottet ikke er lineært. For en utvalgsstørrelse på $n = 500$ er det hevet over enhver tvil at dataene ikke kan være nær normalfordelte. Men for små utvalg, $n = 10$ og $n = 20$, er det ikke store forskjellen mellom figurene ~~3 og 4~~. Dette innebærer at for svært små utvalg, la oss si med $n < 30$, er det vanskelig å avgjøre normalfordeling ved hjelp av QQ-plott.