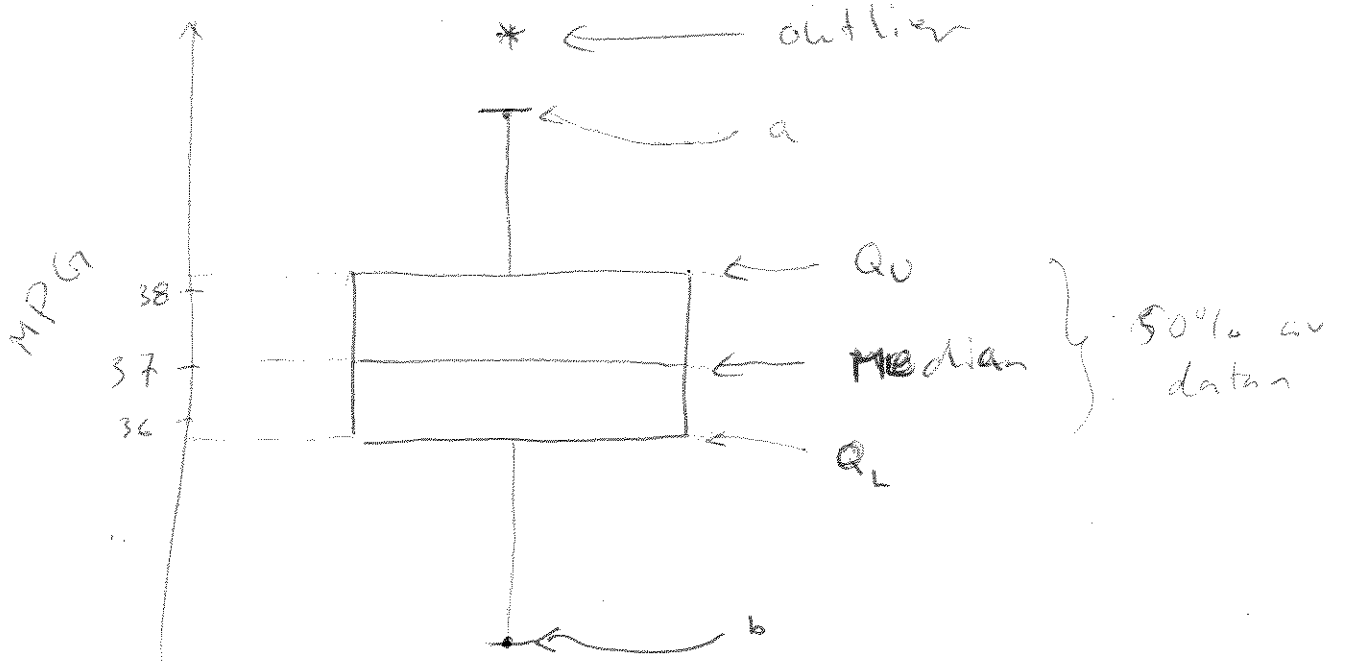
Boxplots (lådadiagram) (kap 2.8)

Boxplots är ett väldigt bra sätt att illustrera kontinuerlig data. En boxplot kan se ut på följande sätt (Figur 31)



Punkten a är den datapunkt (observationen) med högst värde som är mindre än $Q_U + 1.5 \times IQR = Q_U + 1.5(Q_U - Q_L)$

Punkten b är den datapunkt med lägst värde som är större än $Q_L - 1.5 \cdot IQR = Q_L - 1.5(Q_U - Q_L)$

PLOTTAR!

Outliers är (data)punkter som ligger väldigt långt ifrån resten av datasetet. I en boxplot är outliers "utanför" a och b och betecknas med en stjärna eller liknande.

Skevhet

I bland kan det vara viktigt att känna igen ett dataset som inte är symmetriskt. Dataset som inte är symmetriska kallas för skeva. I ett skevt dataset är de vanligare med extrema observationer åt ena hållet än åt det andra hållet.

PLOT!

Ett tecken på att ett dataset är skevt är att medianen och medelvärdet ligger en bit ifrån varandra. PLOT!

En boxplot av ett skevt dataset kännetecknas av att den ena delen av lådan är större än den andra och att motsvarande "morrhår" är större än det andra.

PLOT!

Exempel på typiska skeva dataset är löner i viss, befolkningsmängd.

Punktskattningar (kap 6)

En population har en parameter vars värde är okänt. Vi vill gissa värdet utan att undersöka hela populationen. Hur ska vi göra gissningarna och vad har gissningarna (skumpmässiga!) för egenskaper?

Ex 1: Marknadsundersökning. Coca cola skall lansera den nya colan CCX som utmanare till Pepsi max. Hur stor del av populationen föredrar CCX före Pepsi max?

(population = alla i Sverige/USA/världen

parameter: p = proportionen som föredrar CCX)

Ex 2: Kvalitetskontroll. Pigens frallor har normalfördelad vikt med okända parametrar μ och σ^2. Vad är värdet på dessa?

Hur ska vi gissa parametrarna p, μ och σ^2?

1) Vi samlar in data (stickprov)

2) Data sammanfattas med en skattning

Fördelningen för skattningen är avgörande för hur bra vår gissning blir. Det är skattningen som är vårran gissning!

Ex 1 (forts): Coca Cola frågar 1000 personer. Låt

$$X_k = \begin{cases} 1 & \text{om person } k \text{ föredrar CCX} \\ 0 & \text{—} \quad \text{—} \quad \text{—} \quad \text{Pepsi max} \end{cases}$$

Vår data är $x_1, x_2, \ldots, x_{1000}$. För att gissa värdet på p så tar vi $\bar{x} = \frac{x_1 + \ldots + x_{1000}}{1000}$

dvs andelen som föredrog CCX i undersökningen.

Är $p = \bar{x}$? Nej!!! Är p nära $\bar{x}$? Förhoppningsvis!

Ex 2 (forts): Pigen väger 500 av sina fiskar. Vikterna ~~blir~~ blir $x_1, x_2, \ldots, x_{500}$ (data).

$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{500}}{500}$ är skattaren för μ

$s^2 = \frac{1}{499} \sum_{k=1}^{500} (x_k - \bar{x})^2$ är skattaren för σ^2

OBS! Det finns ett väldigt viktigt före/efter perspektiv.

1) När modellen sätts upp (före mätning) är observationerna slumpvariabler (X_k)

2) Efter att vi har observerat vårt stickprov (efter mätning) så är observationerna numeriska värden (x_k)

Ex 1 (forts.): Innan mätningarna är gjorda så har vi $\bar{X} = \frac{X_1 + X_2 + \dots + X_{1000}}{1000}$ som är en slumpvariabel. Efter mätningarna har vi $\bar{x} = \frac{x_1 + x_2 + \dots + x_{1000}}{1000}$ som är ett numeriskt värde.

Skattningarna $\bar{x}, \bar{x}, s^2$ i exemplen ovan kallas för punktskattningar eftersom vi får ett värde. Senare kommer vi prata om intervallskattningar.

I "före" perspektivet så har vi slumpvariabler $\bar{X}$ och S^2, dessa kallas för skattare medan de numeriska värdena $\bar{x}$ och s^2 i "efter" perspektivet kallas för skattningar.

Väntevärdesriktighet

En skattare kallas för väntevärdesriktig om väntevärdet av skattaren är lika med värdet på parametern.

Ex 1 (forts): $E(\bar{X}) = E\left(\frac{X_1 + X_2 + \dots + X_{1000}}{1000}\right) =$

$= \frac{E(X_1) + E(X_2) + \dots + E(X_{1000})}{1000} = \frac{p + p + p + \dots + p}{1000} =$

$= \frac{1000p}{1000} = p$

Ex 2 (forts). Även här har vi

$$E(\bar{X}) = E\left(\frac{X_1 + \dots + X_{500}}{500}\right) = \frac{E(X_1) + \dots + E(X_{500})}{500} =$$

$$= \frac{\mu + \mu + \dots + \mu}{500} = \frac{500\mu}{500} = \mu$$

Vi har också

$$E(S^2) = E\left(\frac{1}{499}\sum_{k=1}^{500}(X_k - \bar{X})^2\right) = \dots = \sigma^2$$

Lång räkning

Låt oss fokusera på vår skattare för väntevärdet, dvs $\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

1) Som ovan gäller att $E(\bar{X}) = \mu$

2) Vi har även att $\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\sigma^2}{n}$

Med notationen $\sigma_{\bar{X}}^2 = \mathrm{Var}(\bar{X})$ har vi att $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$

Observera att om $\sigma_{\bar{X}}^2$ är "stor" så har vi stor spridning. Därmed är sannolikheten stor att vår skattning $\bar{X}$ missar målet rejält, dvs det är stor sannolikhet att $\bar{X}$ faktiskt inte är nära μ.

Observera att om n är stort, dvs om vårt stickprov är stort, så är $\frac{\sigma^2}{n} = \sigma_{\bar{X}}^2$ litet så $\bar{X}$ är nära μ med större sannolikhet!

PLOT!