26th March '2024

Continuing with Sec 6.4 : Straight Line Model.

Recall that, we assume $Y_i \sim N(\alpha + \beta x_i, \delta^2)$ independent, $i = 1, 2, \cdots, n$

Y_i: response; x_i: predictor.

Error terms: $\varepsilon_i = y_i - E(Y_i) = y_i - \alpha - \beta x_i$, $\varepsilon_i \overset{iid}{\sim} N(0, \delta^2)$.

Assuming δ^2 is unknown, we found:

MLE's

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$,

$\bar{x}$ → sample mean of x's

$\bar{y}$ → Sample mean of y's

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = \frac{1}{n}\sum_i \hat{\varepsilon}_i^{\,2}$$

$\hat{\varepsilon}_i$ is the i^{th} residual.

$$\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$$

LSE's

$\hat{\alpha}$ = same as MLE

$\hat{\beta}$ = same as MLE

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^{\,2}$$

$n-2$ → Sample Size − # parameters estimated.

Let's Look at pg 33 of ch. 6 Lecture Notes (R-output of Ex: 6.4.1)

Returning to Example 6.4.1, the fitted model from R is given below.

$\widehat{E(y)} = \hat{\alpha} + \hat{\beta} X$

```
> Sal.lm<-lm(SalMonth~WTNumN, data=salarynz)
> summary(Sal.lm)

Call:
lm(formula = SalMonth ~ WTNumN, data = salarynz)

Residuals:
    Min      1Q  Median      3Q     Max
-2960.8  -522.6  -157.4   406.4  4136.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2887.40      56.26   51.33   <2e-16 ***
WTNumN        234.99      21.06   11.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 874.7 on 1149 degrees of freedom
Multiple R-squared:  0.09779,Adjusted R-squared:  0.097
F-statistic: 124.5 on 1 and 1149 DF,  p-value: < 2.2e-16
```

Function call:

5 number summary of residuals

t obs

p-value

α, $\hat{\alpha}$

β, $\hat{\beta}$

Ho: $\alpha = 0$ given β in the model

Ho: $\beta = 0$ given α in the model

1, 2, 3

Std error is a part of CI calculation.

Figure 6.7: R Output: Linear regression for salary data

- The estimated relationship between monthly salary and work term number is:

$$\text{Salary} = 2887.40 + 234.99 \times \text{Work Term number}.$$

- The estimate of σ is $s =$"Residual standard error" $= 874.7$ on 1149 degrees of freedom.
- We estimate that monthly salary increases by \$234.99 for each additional work term.
- The intercept estimate is the estimated monthly salary for zero work terms, but this is not meaningful here. Instead, we could quote the estimated monthly salary for work term 1, \$2887.40 + \$234.99.

$H_0: \alpha = 0$ given β in Model.

$$t\text{-value} = t_{obs} = \frac{\hat{\alpha} - 0}{\sqrt{Var(\hat{\alpha})}} = \frac{\hat{\alpha}}{\text{Std. Error}} = 51.33 = \frac{2887.4}{56.26}$$

t-value under Ho follows a $t_{(n-2)}$

(1) Residual Standard Error = Sample Standard deviation = S

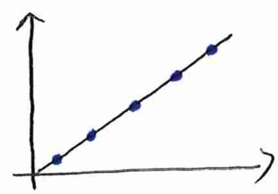
1149 Degrees of freedom = Denominator in S^2 formula $= n-2$.

(2) Multiple R-Squared : Coefficient of Determination, R^2

From output, $R^2 = 0.09779$

↳ 9.779 or 10% of variation in salary (y) is explained by the linear model $(\hat{\alpha} + \hat{\beta} x_i)$

$R^2 \in [0, 1]$



$R^2 = 1$, or 100%.

R^2 is naturally going to increase with # of predictors included in model.

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

$x_1, x_2, \ldots, x_k$ are different predictors.

Adjusted R-Squared : 0.097

↳ includes a penalty term for # of predictors included in model.

↳ It increases when the new term improves the model more than would be expected by chance.

(3) TBD..... we'll come back to this for ANOVA discussions

What is the estimated relationship between Salary and Work term #? Estimated Salary = \$ 2887.4 + \$ 234.99 Work Term #

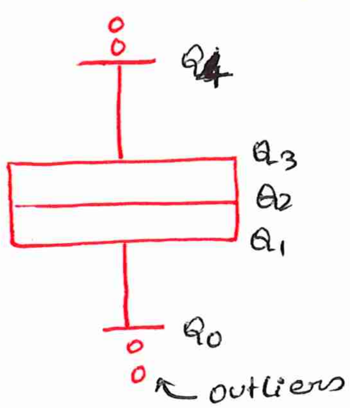
$$E(\hat{y}) = \hat{\alpha} + \hat{\beta} x_i$$

↔ **Interpretation** :→

- For each additional work term, monthly Co-op salary is expected to increase by \$ 234.99.
- The estimated monthly salary for one work term is \$ 2887.40 + \$ 234.99 (1) = \$ 3122.39

(the intercept is not a meaningful / appropriate quantity in this context. Zero work terms means nothing)

Revisiting Boxplots.



$Q_3 - Q_1$ = Interquartile range (IQR)

Question :

If, $Q_4 = Q_3 + 1.5(IQR)$, and $Q_0 = Q_1 - 1.5(IQR)$, why aren't the whiskers always symmetric?

Answer : Q_0 and Q_4 are "notched" at the most extreme data point which extends to no more than 1.5 (IQR) away from Q_1 and Q_3 (the box), respectively.

For Example ↪ $Q_1 = -1$, $Q_2 = 0$, $Q_3 = 1$, then $IQR = Q_3 - Q_1 = 2$

⟹ Q_4 = longest data point within $[Q_3, Q_3 + 1.5\,IQR]$
$= [1, 4]$

Q_1 = Smallest data point within $[Q_1 - 1.5\,IQR, Q_1]$
$= [-4, -1]$

☆ Exercise : Make Boxplots with the following Data :

```
x1 <- c(-4, -1, 0, 1, 4)
x2 <- c(-2, -1, 0, 1, 4)
x3 <- c(-2, -1, 0, 1, 5)
x4 <- c(-6, -1, 0, 1, 5)
```

See how the plots change !!!

```
x5 <- rexp(100, rate = 0.5)
```

Skewed data, see how the plot changes.

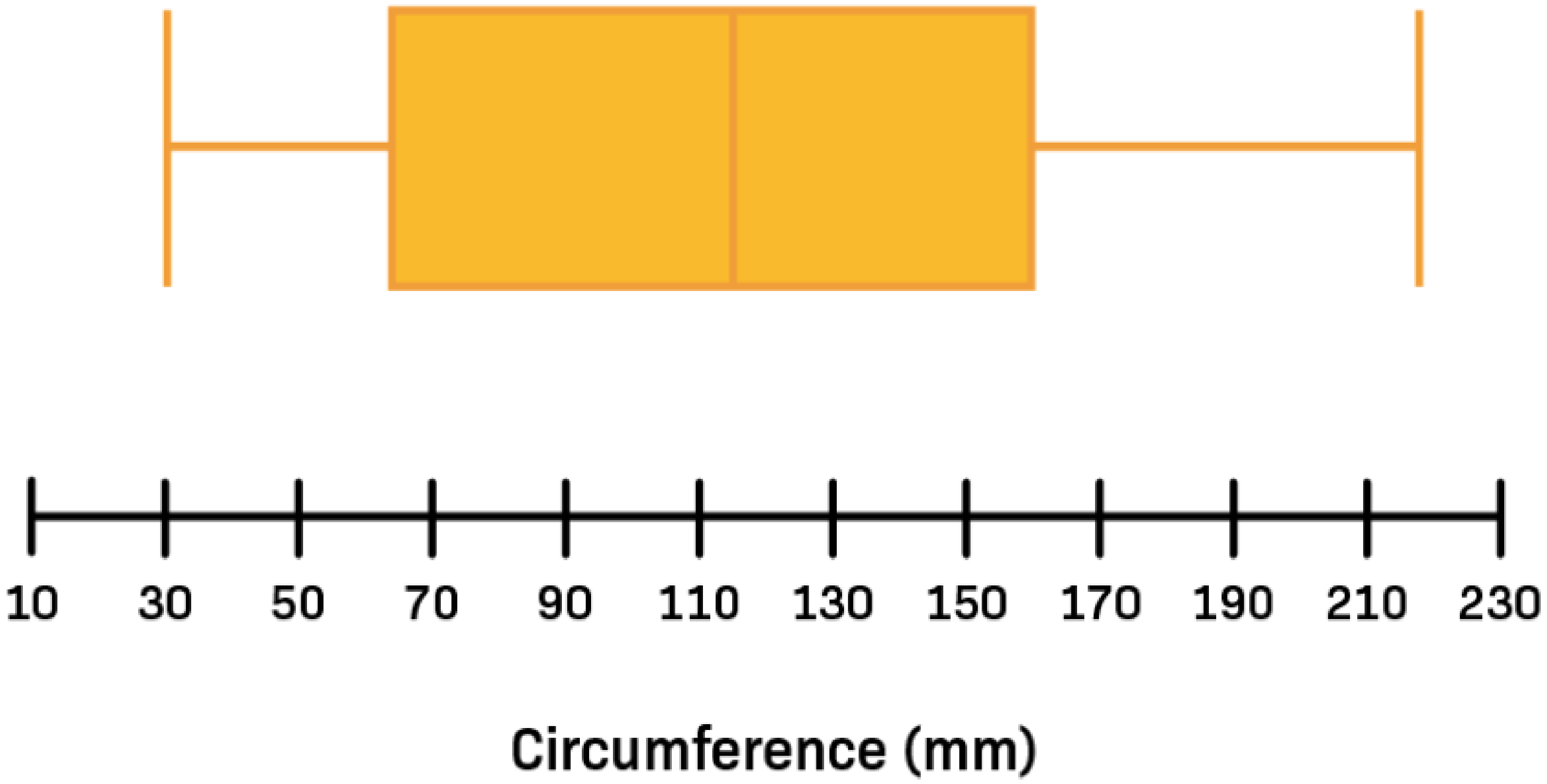
Revisiting Boxplots

STAT 261

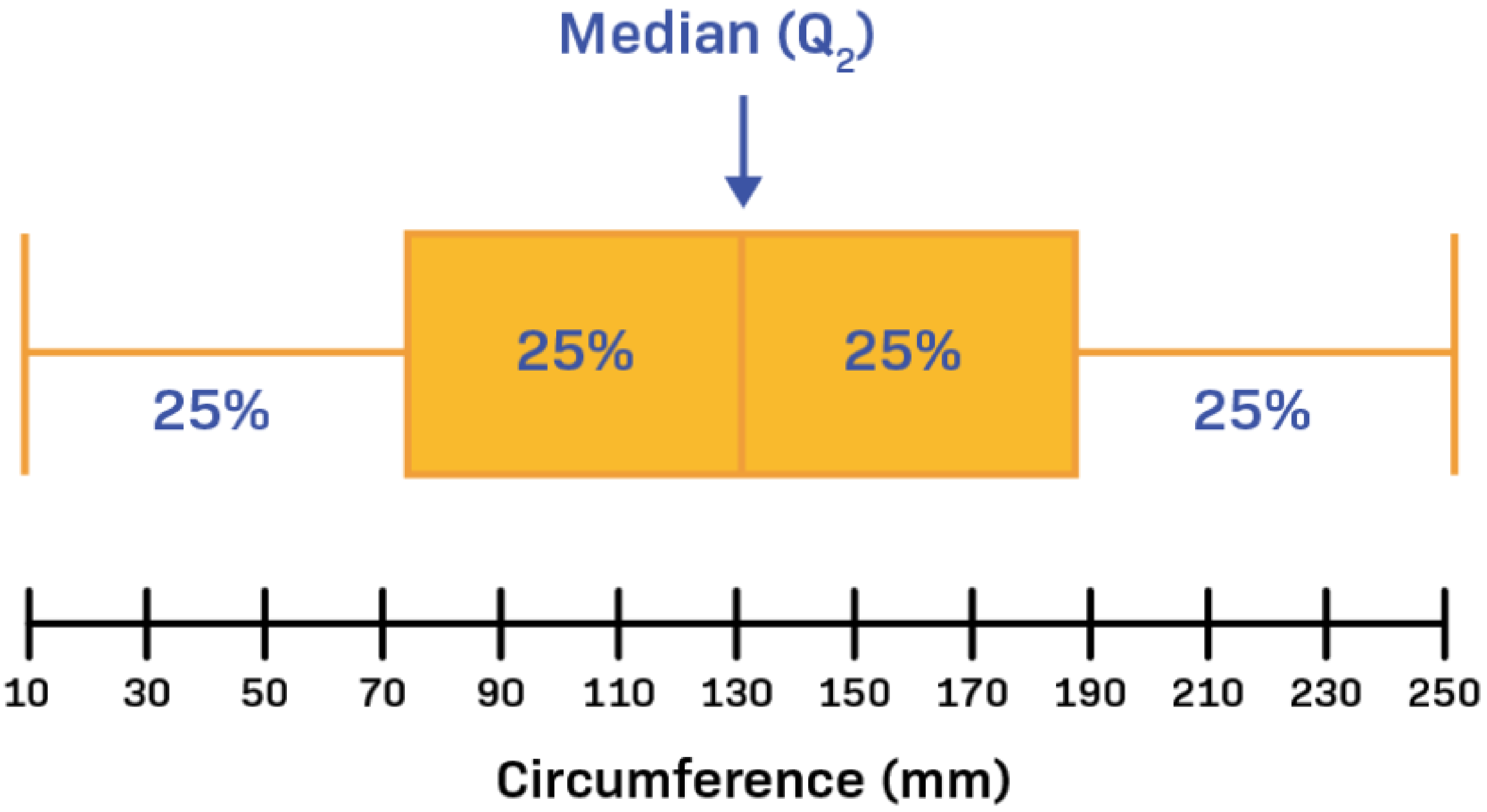
Let’s have the 5-point summary of a tree circumference data set

- Minimum: 30 mm
- First Quartile: 65.5 mm
- Median: 115 mm
- Third Quartile: 161.5 mm
- Maximum: 214 mm

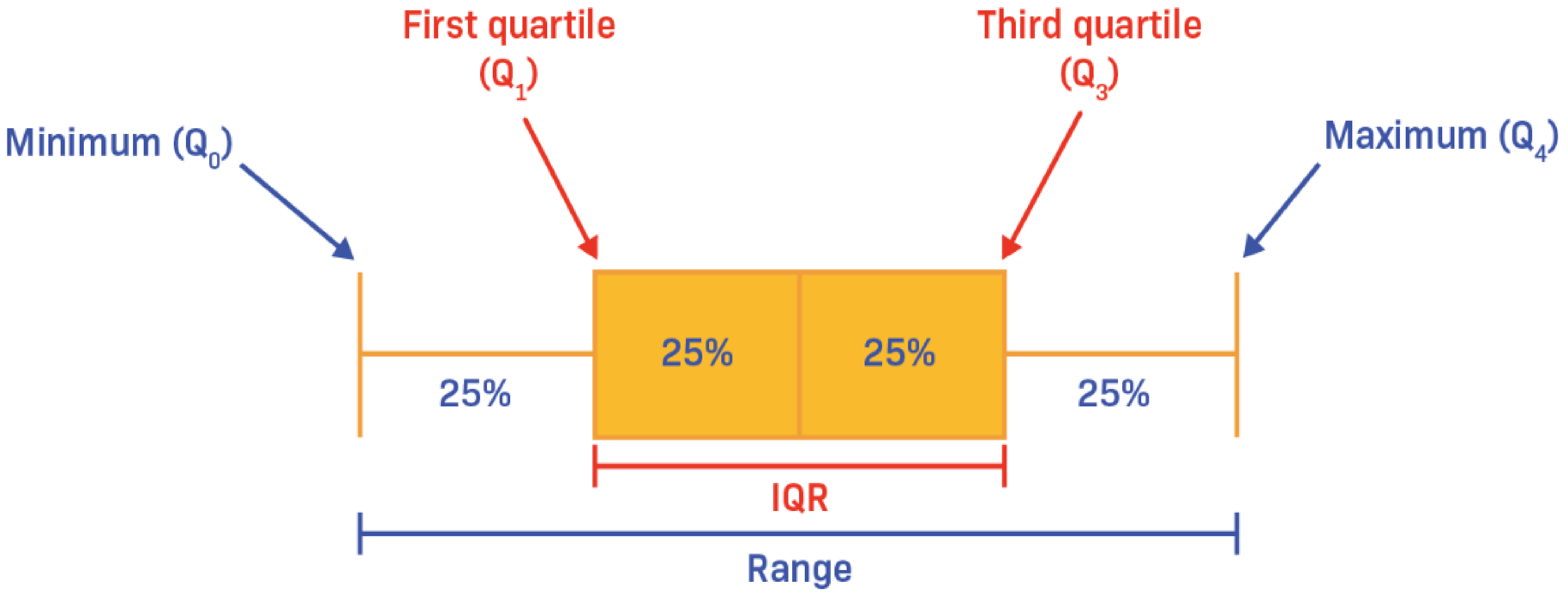
Boxplot of the tree circumference data set



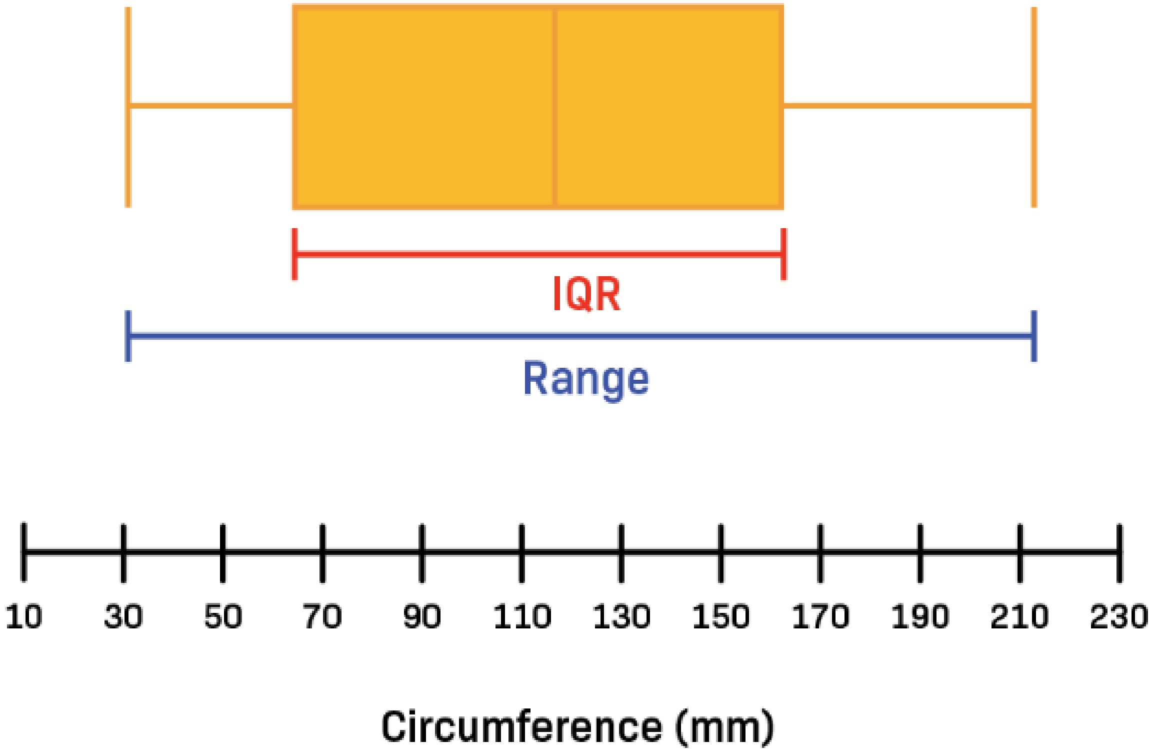
Boxplots use median to describe the center of a data set



Boxplots use IQR and range to describe the spread of a data set



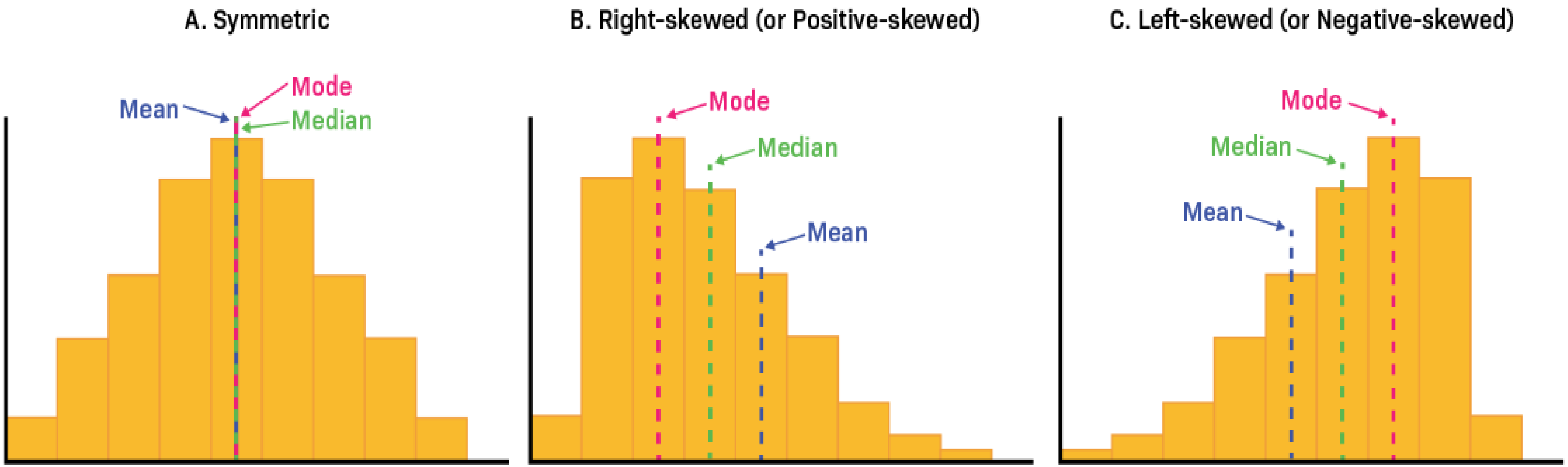
Boxplot to measure the spread using the data set



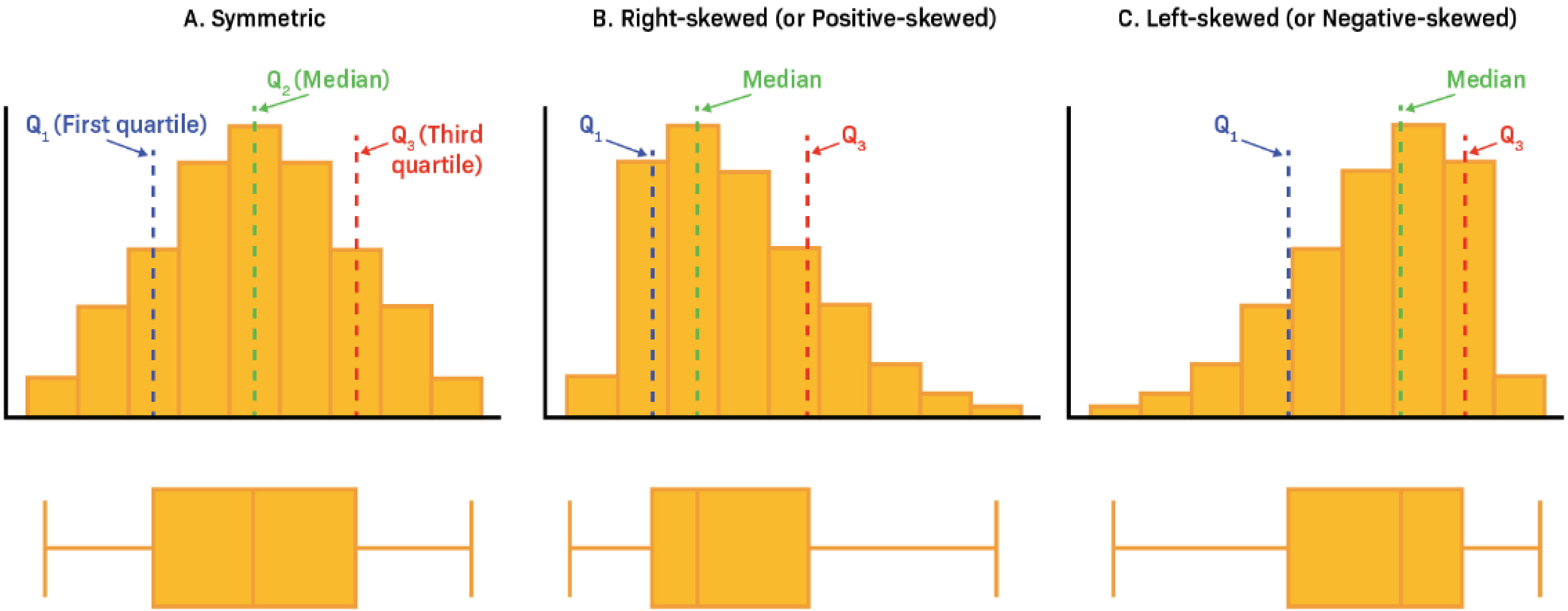
IQR = 161.5 mm - 65.5 mm = 96 mm

Range = 214 mm - 30 mm = 184 mm

Histograms of a symmetric, right-skewed, and left-skewed data set



Histograms and Boxplots of symmetric, right-skewed, and left-skewed data sets



Boxplots do not show modes