Characteristic shapes of a distribution

Symmetric:

see next page

Right-skewed:

Left-skewed:

Alternative measure of spread [to σ^2]:

Interquartile range (IQR):

Population: $IQR = Q_3 - Q_1$ = range of the middle 50% of the dist.

Sample: $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$ = " " " " of the data.

Properties:

- $\hat{Q}_3 - \hat{Q}_1$ is an estimate of $Q_3 - Q_1$
- A consistent est. • has bias but exactly how much is bias – difficult
- SE is hard to compute.

$\Big\}$ Later: using "bootstrap"

Rule of thumb for "outlier" detection: An observation may be considered an "oulier" if it falls outside the interval from $\hat{Q}_1 - 1.5 * \widehat{IQR}$ to $\hat{Q}_3 + 1.5 * \widehat{IQR}$.

Ex: (CPU data): Estimated (or sample) IQR=? Could the observation 139 be an outlier?

using R

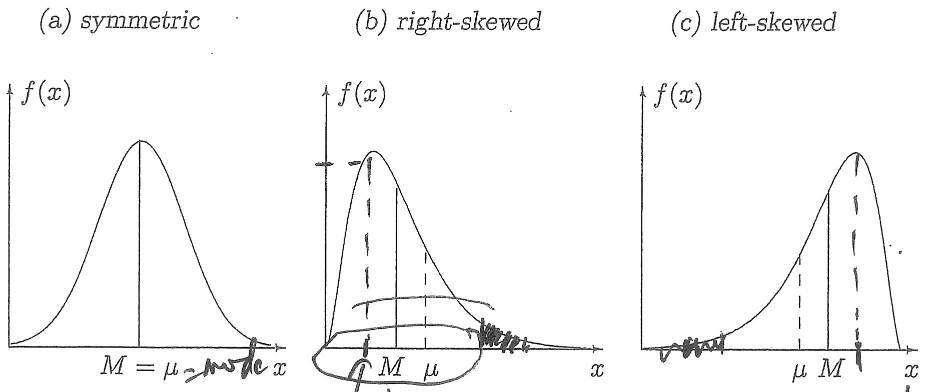


FIGURE 8.2: *A mean μ and a median M for distributions of different shapes.*

DEFINITION 8.6

> **Median** means a "central" value.
>
> **Sample median** $\hat{M}$ is a number that is exceeded by at most a half of observations and is preceded by at most a half of observations.
>
> **Population median** M is a number that is exceeded with probability no greater than 0.5 and is preceded with probability no greater than 0.5. That is, M is such that
>
> $$\begin{cases} P\{X > M\} & \leq & 0.5 \\ P\{X < M\} & \leq & 0.5 \end{cases}$$

Understanding the shape of a distribution

Comparing the mean μ and the median M, one can tell whether the distribution of X is right-skewed, left-skewed, or symmetric (Figure 8.2):

$$\begin{array}{lcl} \text{Symmetric distribution} & \Rightarrow & M = \mu \\ \text{Right-skewed distribution} & \Rightarrow & M < \mu \\ \text{Left-skewed distribution} & \Rightarrow & M > \mu \end{array}$$

Computation of a population median

For continuous distributions, computing a population median reduces to solving one equation:

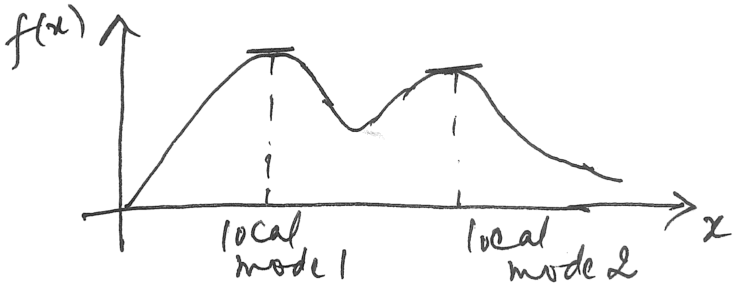
$$\begin{cases} P\{X > M\} & = & 1 - F(M) & \leq & 0.5 \\ P\{X < M\} & = & F(M) & \leq & 0.5 \end{cases} \Rightarrow F(M) = 0.5.$$

Example 8.8 (Uniform, Figure 8.3a). Uniform(a, b) distribution has a cdf

$$F(x) = \frac{x-a}{b-a} \quad \text{for } a < x < b.$$

Mode = most frequent value in the data / distribution

Bimodal or multi-modal:



Which measure of center to use — mean or median?:

mean: if the dist. is more or less symmetric

median: dist. is clearly skewed (or use mean after applying a symmetrizing transformation, e.g., log, sqrt or cube roots)

Which measure of spread to use — SD or IQR?:

- with mean

with median

Graphical Statistics

"Plot the data before you do anything with it."

Boxplot: Displays the 5-number summary of the data, i.e., $(\min, \hat{Q}_1, \hat{Q}_2, \hat{Q}_3, \max)$. It shows

- the data distribution (e.g., symmetric, right-skewed or left-skewed)
- outliers

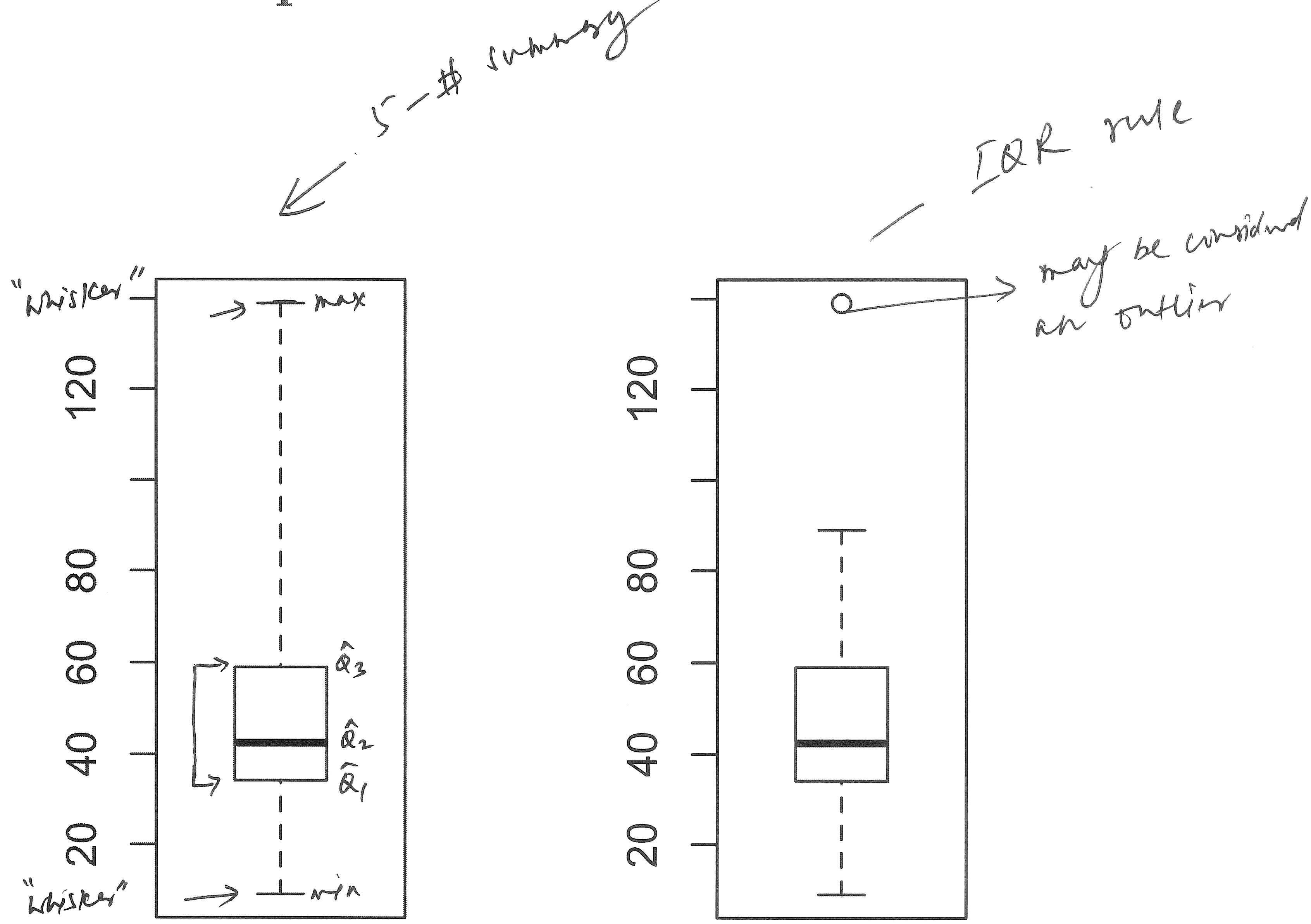
Alternative form: The bottom whisker extends from $\hat{Q}_1$ to $\max\{\min, \boxed{\hat{Q}_1 - 1.5 \times I\hat{Q}R}\}$ and the top whisker extends from $\hat{Q}_3$ to $\min\{\max, \hat{Q}_3 + 1.5 \times I\hat{Q}R\}$

Side-by-side boxplots: Draw side-by-side boxplots on the same scale to compare distributions of more than one data set — see Figure 8.10 in the textbook.

Ex: CPU data

```
?boxplot # see help
par(mfrow=c(1,2)) # 2 plots in 1 row
# plot of 5-number summary
boxplot(cpu, range=0)
# uses 1.5 (IQR) rule (also default), i.e.,
# same as boxplot(cpu)
boxplot(cpu, range=1.5)
par(mfrow=c(1,1)) # back to the default, 1 plot per row
```

Boxplots for CPU data



Histogram

Show the data distribution and suggests possible outliers. Its shape is similar to the population pdf/pmf, especially if the sample size is large.

Frequency histogram: Consists of bars, one over each bin, whose heights represent the *number* of observations in the bins.

Relative frequency histogram: Consists of bars, one over each bin, whose heights represent the *proportion* of observations in the bins.

How to construct a histogram?

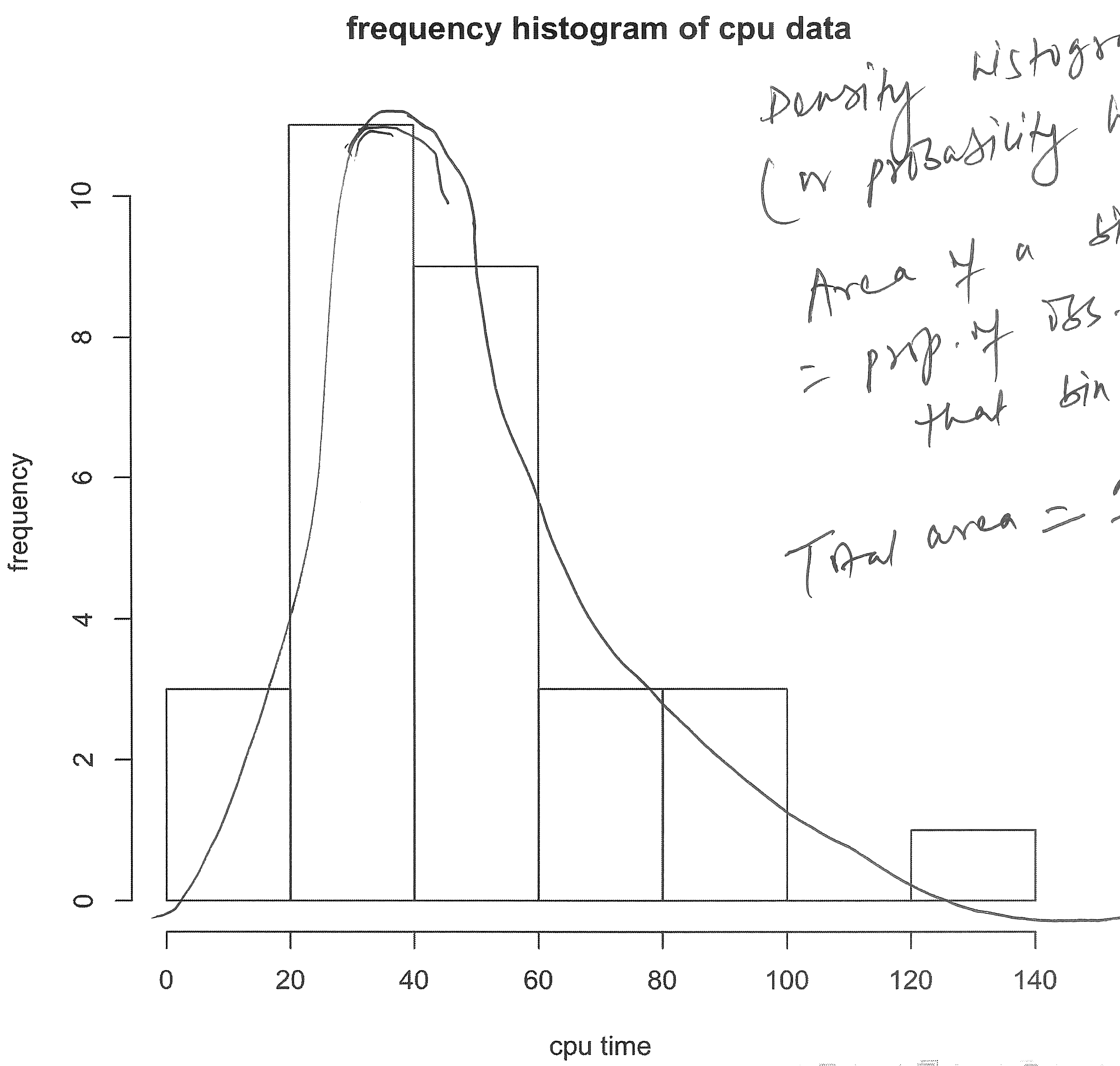
- effect of number of bins (too many or too few)
- bins of unequal sizes

```
# frequency histogram by default
hist(cpu, xlab="cpu time", ylab="frequency",
main="histogram of cpu data")
```

cpu

Find a way to make a rel. freq. hist. in R.

```
# probability (density) histogram
hist(cpu, freq=FALSE, xlab="cpu time",
ylab="density", main="histogram of cpu data")
```



Density histogram (or probability histogram):

Area of a bin = prop. of obs. of in that bin.

Total area = 1.