Chapter 5: Understanding and comparing distributions

This chapter is a continuation of previous chapter with a concentration on the comparison of distributions.

The *boxplot* is a particularly useful visual tool for comparison; it is graphical display of a 5-number summary with one modification: outliers are identified. Boxplots are useful for visually displaying the center, spread, range, and any outliers of the distribution. They are also useful for *comparing* several distributions simultaneously. The components are shown:

- The central *box* shows $Q_1$, the median, and $Q_3$.
- The *whiskers* extend to the most extreme values that are within the fences.
- The *lower fence* is $Q_1 - 1.5 \times \text{IQR}$ and the *upper fence* is $Q_3 + 1.5 \times \text{IQR}$.
- The fences are not shown; instead *whiskers* are drawn at largest datum less than the upper fence and the smallest datum greater than the lower fence.
- Any points outside the fences are *outliers* and are plotted individually.

**Example** Was Tyrannosaurus Rex cold-blooded?

The question may be answered by examining the isotopic concentrations of oxygen in bone phosphate of fossilized skeletons. Isotopic absorption depends upon body temperature, implying that there will be variation in the mean isotopic concentration among bones from different locations if the animal is cold-blooded. On the other hand, there are minor temperature variations in warm-blooded animals. Substantial variation in the mean isotopic concentration among different bones constitutes evidence supporting the position that T. Rex was cold-blooded.

The data shown in the figure above and right are isotopic

![Diagram of boxplot with data points](image-url)
concentration from different bones from a single T. Rex fossil.\textsuperscript{1} Between 3 and 6 observations were obtained from each bone. The boxplots indicate that there are substantial differences among bones with respect to the mean isotopic concentrations of oxygen. However, the data are few in number (per bone), and so we need quantitative measure of the strength of evidence supporting the position that there are differences among bones in isotope concentration.\textsuperscript{2}

\textit{Outlier Detection using the IQR}: A common rule for detecting outlying values is called the 1.5 IQR rule: values at least \(1.5 \times \text{IQR}\) greater than \(Q_3\) or at least \(1.5 \times \text{IQR}\) less than \(Q_1\) are outliers. Hence, outliers lie outside the interval:

\[ [Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}] \]

Histograms and boxplots yield similar, but not identical impressions of the same data. Below are histograms and boxplots representing the same data.

\begin{itemize}
  \item Skewed right
  \item Unimodal, symmetric
  \item Skewed left
  \item Biomodal
\end{itemize}

\textsuperscript{1}See Ramsey, F.L., Schafer, D.W. \textit{The Statistical Sleuth}, 2nd Ed., p. 146.
\textsuperscript{2}Discussed in STAT 452.
Remark: The boxplots show that $Q_3$ is further from $M$ than $Q_1$ if the distribution is skewed right. $Q_3$ is closer to $M$ than $Q_1$ if the distribution is skewed left. $Q_1$ and $Q_3$ are equidistant to $M$ if the distribution is symmetric.

Example: Female mice ($n = 349$) were randomly assigned to six treatment groups to investigate whether restricting dietary intake increases life expectancy. Diet treatments were:

1. NP: mice ate unlimited amount of nonpurified, standard diet
2. N/N85: mice fed normally before and after weaning. After weaning, ration was controlled at 85 kcal/wk.
3. N/R50: normal diet before weaning and reduced calorie diet (50 kcal/wk) after weaning.
4. R/R50: reduced calorie diet of 50 kcal/wk both before and after weaning.
5. lopro: low protein diet before weaning, restricted diet (50 kcal/wk) after weaning and dietary protein content decreased with advancing age.
6. N/R40: normal diet before weaning and reduced diet (40 Kcal/wk) after weaning.

Which group had the greatest spread? Greatest median lifetime? Smallest median? Which greatest number of outliers? Least?

Time Plots: Time series plots are useful graphical descriptions for quantitative variables collected over time. Below and to the left is a set of time plots showing the annual number of TB cases per 100,000 persons from 1982 to 2005.

In three countries, there is a slow but consistent decline in the number of cases. In Armenia, there has been an increase since the fall of the Soviet Union (1991). The trend

---

4lopro - using the IQR as the measure of spread.
5N/R40
6NP
7left
8N/N85 and lopro, respectively.
in Angola is very unusual; the data may be incorrect because of problems in monitoring and reporting TB cases during the civil war (1995 to 1997). HIV/AIDS is also potentially responsible for the increase.

The plot to the right (and above) shows three countries from southern Africa. The data are plotted differently: individual points are shown and a smooth or smoother is graphed. The smooth shows the general trend of the data more clearly than the individual points. It is apparent that the TB rate has increased dramatically, most likely attributable to the development of HIV/AIDS epidemics.

**Transformations (re-expressions) of data:** Sometimes, when the distribution of a quantitative variable is skewed, the data values can be transformed to a different scale to make the distribution more symmetric. Symmetrically distributed data are, loosely speaking, a prerequisite for accurate hypothesis tests and confidence intervals.

**Example** Ozone pollution is believed to be a source of mortality, increasing the yearly risk of death from respiratory diseases by 40% to 50% in heavily polluted cities like Los Angeles and Riverside and by about 25% throughout the rest of the country.\(^9\) Consequently, forecasting daily ozone levels in large cities is an important use of statistics. The data shown in the histogram are \( n = 366 \) measurements on daily maximum one-hour-average ozone reading from 1976 in Los Angeles.

The panels show the original data and after transformation to a log scale. If the original observations are denoted by \( y_1, \ldots, y_{366} \) then the log-transformed data are \( \log(y_1), \ldots, \log(y_{366}) \). The log-transformed data are more symmetric in distribution. There are two logarithm transformations. The common logarithm uses 10 as the base whereas the natural logarithm uses Euler’s number \( e = 2.718282\ldots \) as the base.

The common logarithm of a positive number \( y \) is the exponent \( x \) to which 10 must be raised to obtain \( y \). Thus, \( y = 10^x \). Some examples are

\[
\begin{align*}
\log_{10}(10) &= 1 \text{ since } 10^1 = 10 \\
\log_{10}(100) &= 2 \text{ since } 10^2 = 100.
\end{align*}
\]

Viewed another way, the log base 10 function extracts the exponent on the 10. Similarly,

\[
\log_{10}(1000) = 3
\]

Another example is \( \log_{10}(500) = 2.699 \) since \( 10^{2.699} = 500 \).

**Remarks:**

1. The effect of transforming to the logarithmic scale is to shrink the right tail of a distribution towards the median.

2. Log functions cannot be applied to negative numbers, for instance, \( \log_{10}(-2) \) is undefined. If a log transformation is to be applied to a data set with negative numbers, an expedient solution is to add a constant \( c \) to all values (this is called *shifting* the data. The constant \( c \) is chosen so that all values are positive after shifting.
3. Logarithmic transformations preserve the original order of the observations. That is, if $y_i < y_j$, then $\log_{10}(y_i) < \log_{10}(y_j)$, and if $\log_{10}(y_i) < \log_{10}(y_j)$ then $y_i < y_j$. Order preserving transformations are nearly always better than a transformation that is not order preserving. The transformation $y \rightarrow y^2$ is not order preserving since $-2 < -1$ but $(-2)^2 > (-1)^2$. 