Describing data from its shape

Dr. Abbas Adigun(PhD)

Biostatistician

May 19, 2017

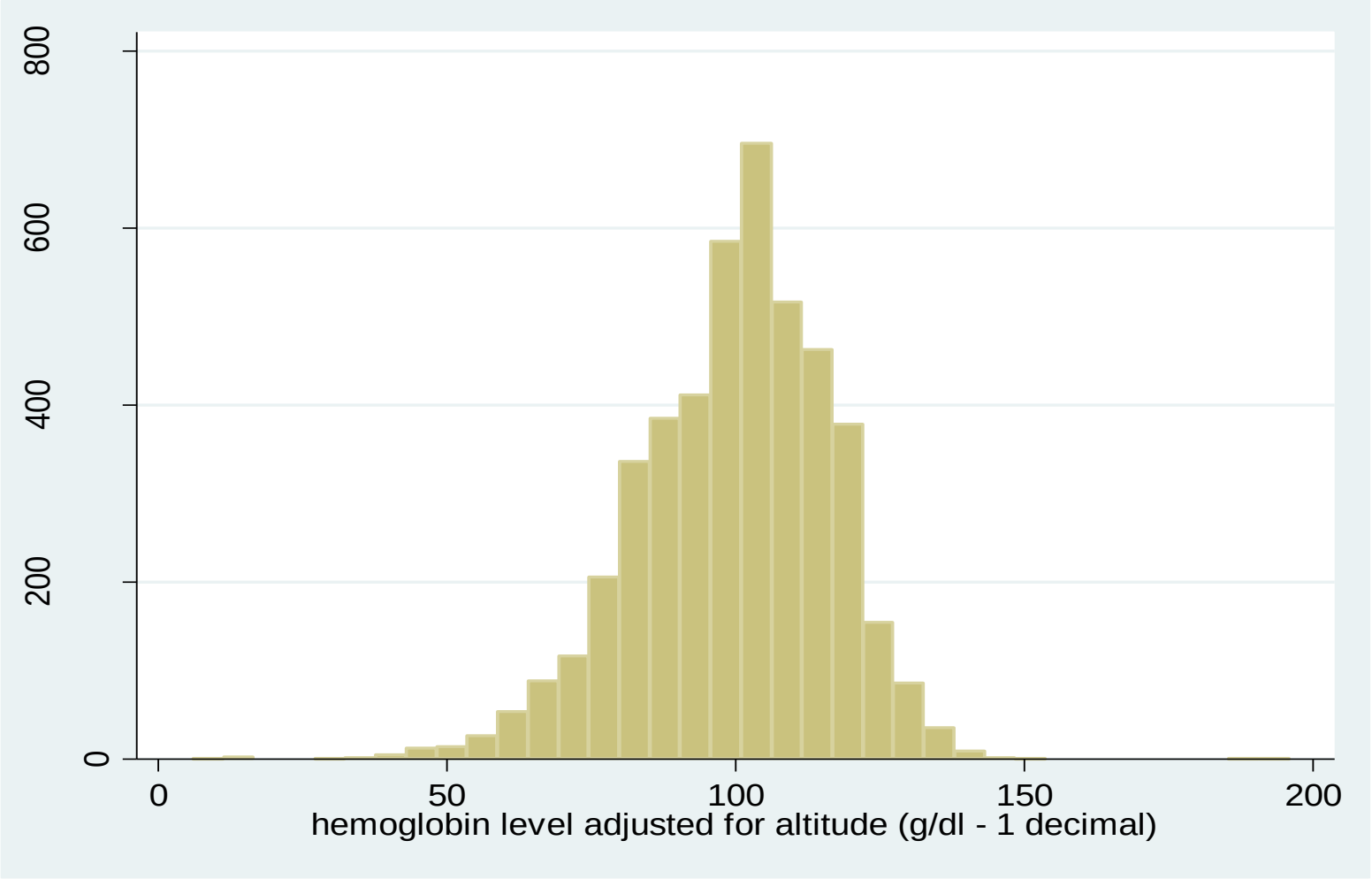
Describing data from the shape

Although, the choice of the most appropriate procedures for summarizing and analyzing data will depend on the type of variable involved. However, the way the data are distributed – the *shape of the distribution*, can also be influential.

Symmetric or Normal distribution

Do most of the values clump together around *one* particular value, with progressively fewer values both below and above this value? This is a *symmetric* or *mound-shaped* distribution.

Example



Stata commands

```
use "C:\Users\ADIGUM\Desktop\abuja\hemoglo.dta", clear
hist hc56,freq
```

Positive skew

Are most of the values concentrated towards the bottom of the range, with progressively fewer values towards the top of the range? This is a *right or positively skewed* distribution.

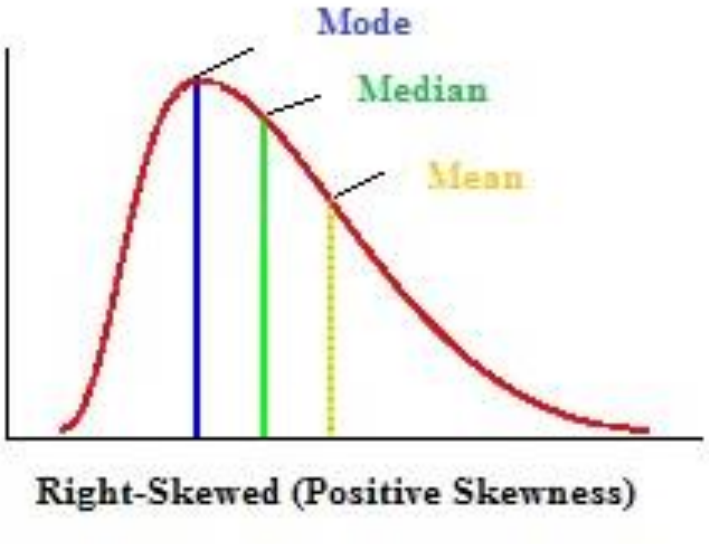
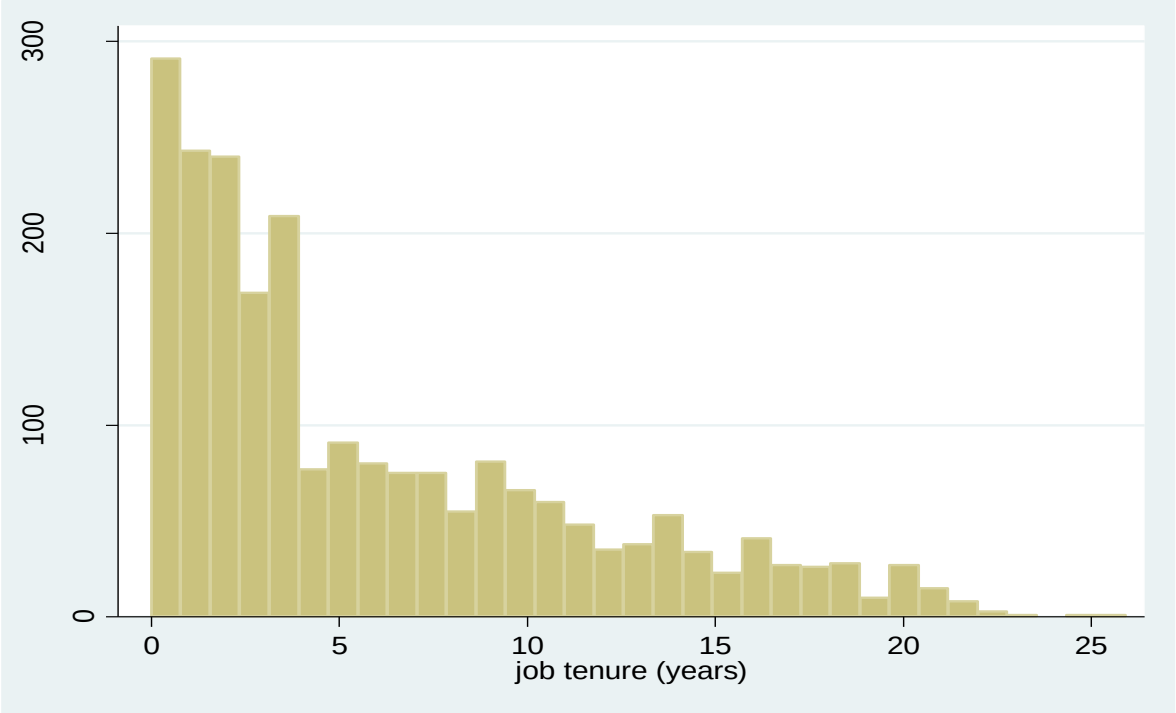


Figure 4.2 shows the job tenure in a particular FMC, Nigeria.



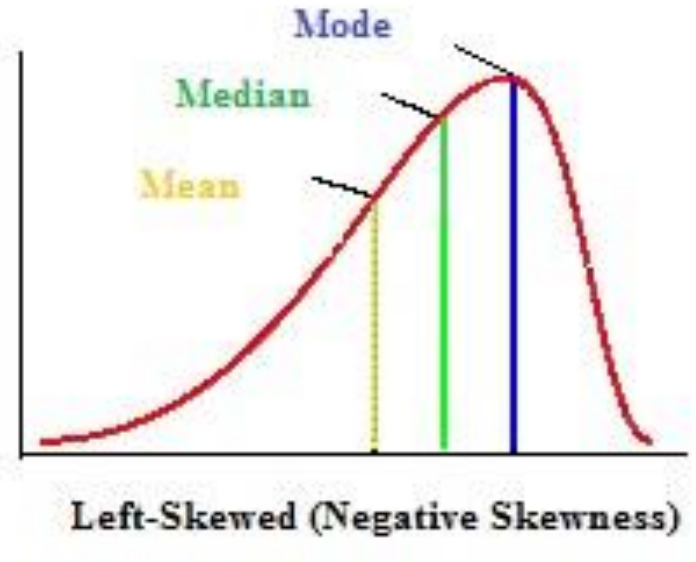
Stata commands

use "C:\Users\ADIGUM\Desktop\abuja\shape.dta", clear

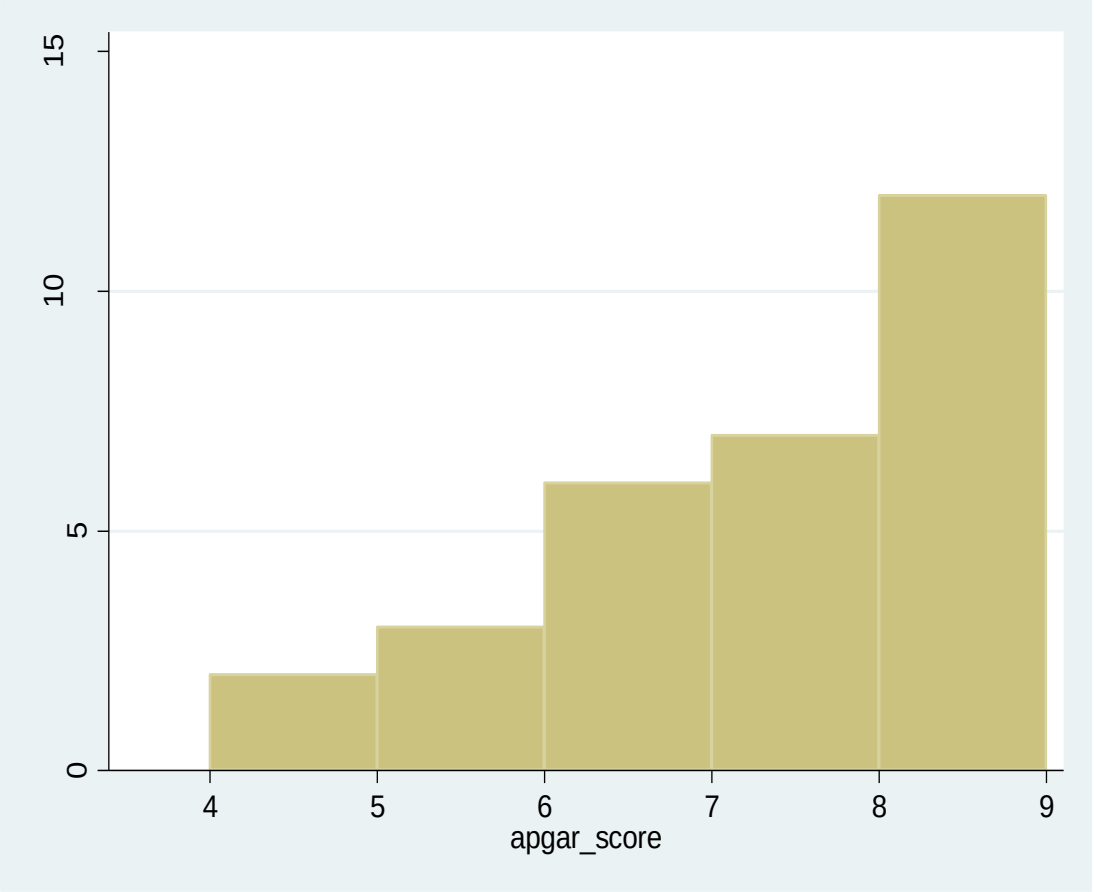
hist tenure, freq

Negative skewed

For left or negatively skewed distribution most values are found towards the top of the range, with progressively fewer values towards the bottom of the range.



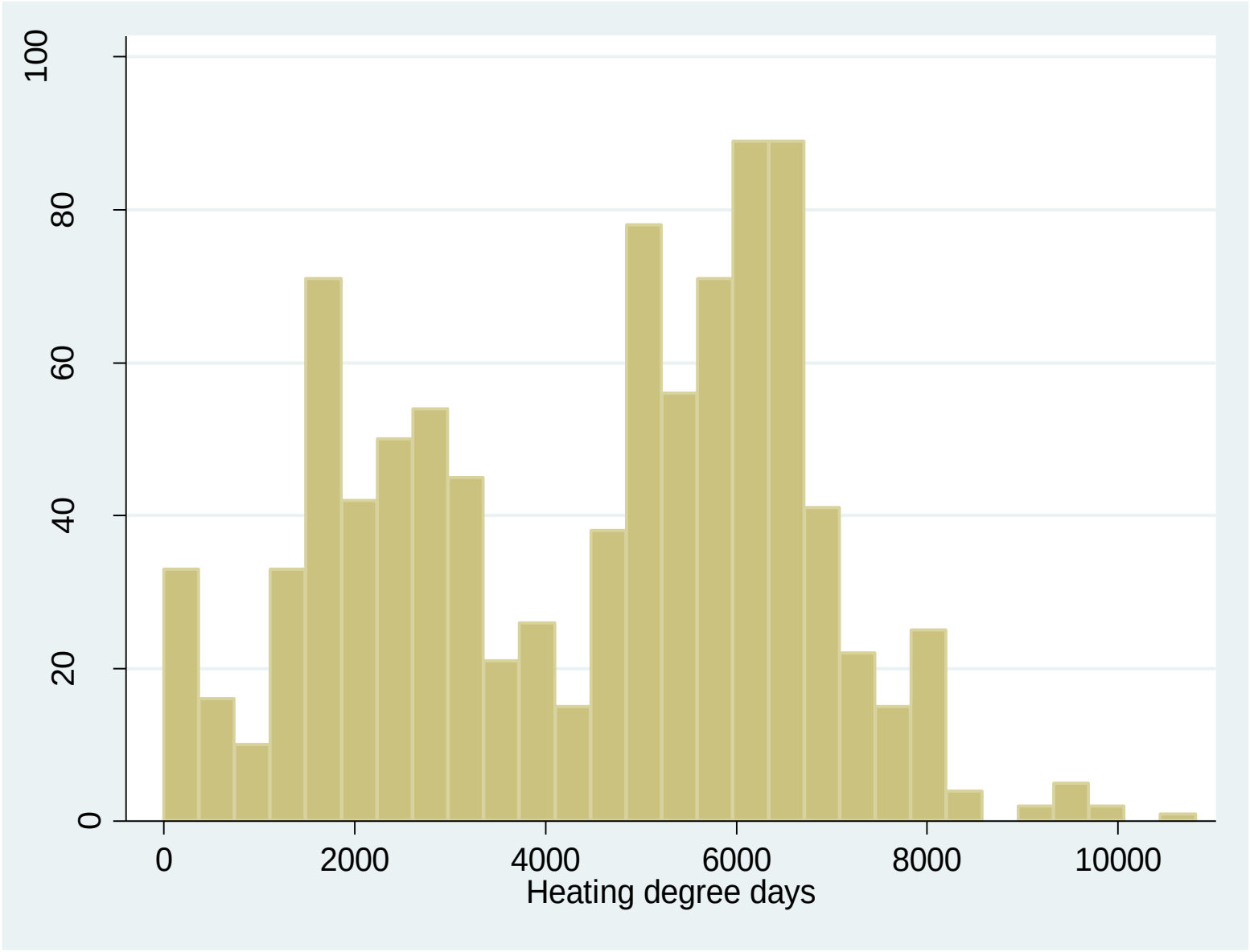
Example: Apgar score from the infant data



Bimodal distribution

A bimodal distribution is one with two distinct humps. These are less common than the shapes described above, and are sometimes the result of two separate distributions, which have not been disentangled.

Figure 2.3 shows a hypothetical bimodal distribution of heating degree days


100
80
60
40
20
0
0
2000
4000
6000
8000
10000
Heating degree days

Data transfomations

In data analysis **transformation** is the replacement of a variable by a function of that variable: for example, replacing a variable x by the square root of x or the logarithm of x.

In a stronger sense, a transformation is a replacement that changes the shape of a distribution or relationship.

Common transformation

The most useful transformations in data analysis are the reciprocal, logarithm, cube root, square root, and square.

Reciprocal transformation

The **reciprocal**, x to 1/x, with its sibling the **negative reciprocal**, x to -1/x, is a very strong transformation with a drastic effect on distribution shape.

It can not be applied to zero values.

The reciprocal of a ratio may often be interpreted as easily as the ratio itself:

Example

population density (people per unit area) becomes area per person;

persons per doctor becomes doctors per person

The **logarithm**, x to log base 10 of x, or x to log base e of x (ln x), or x to log base 2 of x, is a strong transformation with a major effect on distribution shape.

It is commonly used for reducing right skewness and is often appropriate for measured variables.

It can not be applied to zero or negative values

Cube root transformation

The **cube root**, x to x^(1/3). This is a fairly strong transformation with a substantial effect on distribution shape:

It is also used for reducing right skewness, and has the advantage that it can be applied to zero and negative values.

Square root

The **square root**, x to x^(1/2) = sqrt(x), is a transformation with a moderate effect on distribution shape:

It is also used for reducing right skewness, and also has the advantage that it can be applied to zero values.

It is commonly applied to counted data, especially if the values are mostly rather small.

Square

The **square**, x to x^2, has a moderate effect on distribution shape and it could be used to reduce left skewness.

CORBIS/Brian Leng (05065)