

Shifting and rescaling data distributions

It is useful to consider the effect of systematic alterations of all the values in a data set.

The simplest such systematic effect is a **shift** by a fixed constant. Suppose a certain data set is given, and a second data set is obtained from the first by adding the same number c (positive or negative) to each value. Then

- any measure of center (median or mean) of the new data set is *shifted* by the same constant value c ;
- any measure of spread (*IQR* or standard deviation) is *unchanged*; and
- any measure of relative standing (percentile value or *z*-score) is *unchanged*

Another common alteration is a **rescaling** of the data. Suppose a certain data set is given, and a second data set is obtained from the first by rescaling each value to a different unit of measure (every one of the original values x is replaced with a scaled value kx , k being the *scale factor*). Then

- any measure of center (median or mean) of the new data set is *rescaled* by the same scale factor k ;
- any measure of spread (*IQR* or standard deviation) is *rescaled* by the same scale factor k ; and
- any measure of relative standing (percentile value or *z*-score) is *unchanged* by the rescaling

The Normal Model

- The most commonly occurring distributions in practice are symmetric and “bell-shaped”. Mathematicians have devised a theoretical model for such distributions, the **normal model**. It faithfully describes many real data sets and is the basis for most statistical inference techniques.
- The **normal curve**, a curve meant to describe the contour of a symmetric and bell-shaped histogram, is characterized by the location of its mean, labeled μ (the Greek letter ‘m’), and the size of its standard deviation, labeled σ (the Greek letter ‘s’). These two numbers, which determine all the information about the distribution, are called **parameters** of the model.
- Note that we generally use Roman characters (\bar{x}, s) to represent statistics, which are computed from the actual data measurements, while we use Greek letters (μ, σ) to represent parameters, which are theoretical assumptions about what happens in general.

- The mean μ of the distribution lies on the scale axis at the position of the central peak of the curve. The points on either side of the mean at which the curve *changes concavity* are located exactly one standard deviation σ away from the mean; that is, they are located on the axis at the values $\mu - \sigma$ and $\mu + \sigma$.
- The normal distribution with mean μ and standard deviation σ is denoted $N(\mu, \sigma)$.

The 68-95-99.7 Rule

For the normal model, the following approximations are useful:

- about 68% of the data will lie within one standard deviation of the mean (between $\mu - \sigma$ and $\mu + \sigma$);
- about 95% of the data will lie within two standard deviations of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$);
- nearly all (about 99.7%) of the data will lie within three standard deviation of the mean (between $\mu - 3\sigma$ and $\mu + 3\sigma$).

Working with the Normal Model

- In situations where the normal model is being applied to a given situation, sketch a graph of the model and identify the appropriate scale by marking on the horizontal axis the seven values

$$\mu - 3\sigma, \mu - 2\sigma, \mu - \sigma, \mu, \mu + \sigma, \mu + 2\sigma, \mu + 3\sigma$$

- More specific percentages associated with the normal model $N(\mu, \sigma)$ can be found with your calculator: the percentage of the data lying between two particular values a and b ($a \leq y \leq b$) is computed as

DISTR normalcdf (a , b , μ , σ)

(If no upper bound b is given, it is understood that $b = \infty$ – use 1E99 for ∞ ; if no lower bound is given, it is understood that $a = -\infty$ – use -1E99 for $-\infty$.)

- The standard normal distribution $N(0, 1)$ has mean 0 and standard deviation 1. If a normal model $N(\mu, \sigma)$ applies to a data set, then the corresponding standardized values z will follow the standard normal distribution $N(0, 1)$. Percentages associated with the standard normal model $N(0, 1)$ can also be found with your calculator by omitting entry of the values of μ and σ : the percentage of the data lying between two particular values a and b ($a \leq y \leq b$) of a standard normal model is computed as

DISTR normalcdf(a , b)

- Technology can also be used to work with the inverse problem: to determine the *critical z-score*, labeled z^* , that lies above p % of all the data in the standard normal model, compute

DISTR invNorm(p)

More generally, to determine the *critical value*, labeled x^* , that lies above p % of all the data in the normal model $N(\mu, \sigma)$, compute

DISTR invNorm(p , μ , σ)

- **The Nearly Normal Condition**

Only if a data set has unimodal and symmetric shape may be appropriately apply the normal model to its study; check this by consulting a histogram or boxplot to verify the relevant features of the distribution.