Chapter 2 Probabilities and Comparing Data

2.1 Comparing Data with Z-Scores

2.1.1 Introduction to z-scores

In Ghana, the mean height of young adult women is normally distributed with mean 159.0 cm and standard deviation 4.9 cm. (Monden & Smits, 2009) Serwa, a female BYU-Idaho student from Ghana, is 169.0 cm tall. Her height is $169.0 - 159.0 = 10$ cm greater than the mean. When compared to the standard deviation, she is about two standard deviations ($\approx 2 \times 4.9$ cm) taller than the mean.

The heights of men are also normally distributed. The mean height of young adult men in Brazil is 173.0 cm ("Oramento," 2010), and the standard deviation for the population is 6.3 cm. (Castilho & Lahr, 2001) A Brazilian BYU-Idaho student, Gustavo, is 182.5 cm tall. Compared to other Brazilians, he is taller than the mean height of Brazilian men.

When we examined the heights of Serwa and Gustavo, we compared their height to the standard deviation. If we look carefully at the steps we did, we subtracted each individual's height from the mean height for people of the same gender and nationality.

2.1.2 Computing z-scores

This shows how much taller or shorter the person is than the mean height. In order to compare the height difference to the standard deviation, we divide the difference by the standard deviation. This gives the number of standard deviations the individual is above or below the mean.

For example, Serwa's height is 169.0 cm. If we subtract the mean and divide by the standard deviation, we get

$$z = \frac{169.0 - 159.0}{4.9} = 2.041$$

We call this number a z-score. The z-score for a data value tells how many standard deviations away from the mean the observation lies. If the z-score is positive, then the observed value lies above the mean. A negative z-score implies that the value was below the mean.

We compute the z-score for Gustavo's height similarly, and obtain

$$z = \frac{182.5 - 173.0}{6.3} = 1.508$$

Gustavo's z-score is 1.508. As noted above, this is about one-and-a-half standard deviations above the mean. In general, if an observation x is taken from a random process with mean μ and standard deviation σ, then the z-score is

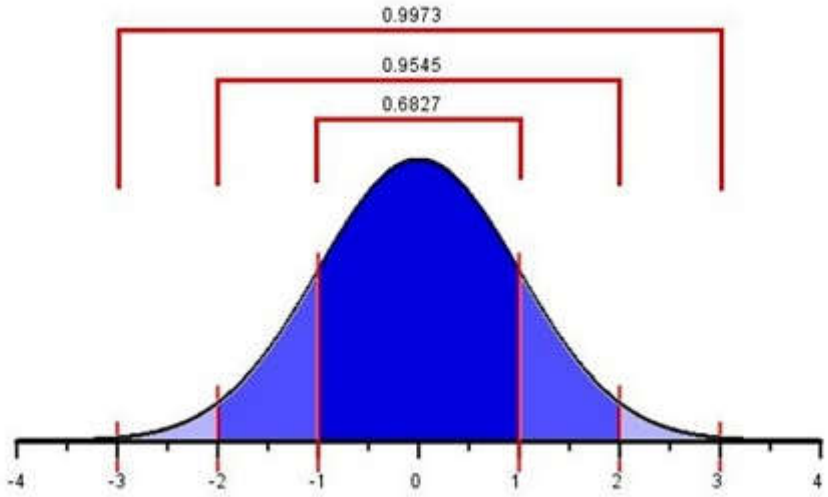
$$z = \frac{x - \mu}{\sigma}$$

The z-score can be computed for data from any distribution, but it is most commonly applied to normally distributed data.

2.1.2.1 68-95-99.7% Rule for Bell-shaped Distributions

Heights of women (or men) in a particular population follow a normal distribution. Most people's heights are close to the mean. A few are very tall or very short. We would like to make a more precise statement than this.

For any bell-shaped distribution, - 68% of the data will lie within 1 standard deviation of the mean, - 95% of the data will lie within 2 standard deviations of the mean, and - 99.7% of the data will lie within 3 standard deviations of the mean.



68-95-99.7% Rule

This is called the 68-95-99.7% Rule for Bell-shaped Distributions. Some statistics books refer to this as the Empirical Rule.

Approximately 68% of the observations from a bell-shaped distribution will be between the values of $\mu - \sigma$ and $\mu + \sigma$. Consider the heights of young adult women in Ghana. We expect that about 68% of Ghanian women have a height between the values of

$$\mu - \sigma = 159.0 - 4.9 = 154.1 \text{ cm}$$

and

$$\mu + \sigma = 159.0 + 4.9 = 163.9 \text{ cm}.$$

So, if a female is chosen at random from all the young adult women in Ghana, about 68% of those chosen will have a height between 154.1 and 163.9 cm. Similarly, 95% of the women's heights will be between the values of

$$\mu - 2\sigma = 159.0 - 2(4.9) = 149.2 \text{ cm}$$

and

$$\mu + 2\sigma = 159.0 + 2(4.9) = 168.8 \text{ cm}.$$

Finally, 99.7% of the women's heights will be between

$$\mu - 3\sigma = 159.0 - 3(4.9) = 144.3 \text{ cm}$$

and

$$\mu + 3\sigma = 159.0 + 3(4.9) = 173.7 \text{ cm}.$$

2.1.3 Unusual Events

If a z-score is extreme (either a large positive number or a large negative number), then that suggests that that observed value is very far from the mean. The 68-95-99.7% rule states that 95% of the observed data values will be within two standard deviations of the mean. This means that 5% of the observations will be more than 2 standard deviations away from the mean (either to the left or to the right).

We define an **unusual observation** to be something that happens less than 5% of the time. For normally distributed data, we determine if an observation is unusual based on its z-score. We call an observation unusual if $z < -2$ or if $z > 2$. In other words, we will call an event unusual if the

absolute value of its z-score is greater than 2.

2.2 Probability

2.2.1 Probability Notation

You may already have a good understanding of the basics of probability. It is worth noting that there is a special notation used to denote probabilities. The probability that an event, x, will occur is written $P(x)$. As an example, the probability that you will roll a 6 on a die can be written as

P (Roll a 6 on a die)= $\frac{1}{6}$

2.2.2 Rules of Probability

Probabilities follow patterns, called **probability distributions,** or distributions, for short. There are three rules that a probability distribution must follow.

The three rules of probability are:

- **Rule 1**: The probability of an event X is a number between 0 and 1.

$$0 \leq P(X) \leq 1$$

- **Rule 2**: If you list all the outcomes of an experiment (such as rolling a die) the probability that one of these outcomes will occur is 1. In other words, the sum of the probabilities of all the possible outcomes of any experiment is 1.

$$\sum P(X) = 1$$

- **Rule 3**: (Complement Rule) The probability that an event X will not occur is 1 minus the probability that it will occur.

$$P(\text{not } X) = 1 - P(X)$$

You may have noticed that the Complement Rule is just a combination of the first two rules.