2.4 The Normal Distributions

Idea:
Instead of describing data by a histogram, we like to describe the overall pattern by a smooth curve, the **density curve** $f(x)$. A density curve is a curve that

- is always on or above the horizontal axis, i.e., $f(x) \geq 0$
- has an area of exactly 1.0 underneath the curve and above the horizontal axis, i.e., $\int_{-\infty}^{\infty} f(x)dx = 1.0$.

Examples:
These are density curves:

These are no density curves:

Mean and Median of a Density Curve:
The **median** of a density curve is the point that divides the area under the curve in half.

The **mean** of a density curve is the balance point at which the curve would balance if it were made of solid material.

Note:

- Since a density curve relates to an entire (idealized) population, we use the symbols μ and σ to denote the mean and standard deviation of a density curve.
- For a symmetric density curve, the mean and the median are identical.

Normal Distributions:
The **Normal (Probability) Distribution** is a bell-shaped curve that occurs very frequently in real life. Examples we have seen so far that could be considered as bell-shaped are ...

The Empirical Rule (or 68–95–99.7 Rule):
For a normal distribution with mean μ and standard deviation σ it holds that

- 68% of the observations fall into the interval $\mu - \sigma$ to $\mu + \sigma$
- 95% of the observations fall into the interval $\mu - 2\sigma$ to $\mu + 2\sigma$
- 99.7% of the observations fall into the interval $\mu - 3\sigma$ to $\mu + 3\sigma$

Note:
The same percentages (and the same rule) hold if we have a reasonable large number of observations that orginate from a normal distribution. In the notation above, we have to replace μ by $\overline{x}$ and σ by s, i.e.,

- 68% of the observations fall into the interval $\overline{x} - s$ to $\overline{x} + s$
- 95% of the observations fall into the interval $\overline{x} - 2s$ to $\overline{x} + 2s$
- 99.7% of the observations fall into the interval $\overline{x} - 3s$ to $\overline{x} + 3s$

Example: "The Cost of Victories"

Recall that $\overline{x} = 18.83, \quad s = 4.43$.
Thus,

$$18.83 \pm 4.43 = [\qquad\quad , \qquad\quad] = I_1$$

$$18.83 \pm 2 \cdot 4.43 = [\qquad\quad , \qquad\quad] = I_2$$

$$18.83 \pm 3 \cdot 4.43 = [\qquad\quad , \qquad\quad] = I_3$$

When assuming that the data originates from a bell–shaped distribution, we can assume that the 68–95–99.7 rule applies:

- 68% of $30 \approx 20.4$ observations are expected in interval I_1
- 95% of $30 \approx 28.5$ observations are expected in interval I_2
- 99.7% of $30 \approx 29.91$ observations are expected in interval I_3

Now look at the data and see if it there are really as many observations in each of the 3 intervals as predicted by the 68–95–99.7 rule. We have

- .... observations that are ≥ 15 and ≤ 23, i.e., fall in interval I_1
- .... observations that are ≥ 10 and ≤ 27, i.e., fall in interval I_2
- .... observations that are ≥ 6 and ≤ 32, i.e., fall in interval I_3

Conclusion:
The empirical rule holds very well for this data set. Our first visual impression that this data set is bell–shaped has been confirmed by this rule.

Note:
Even small departures from expected number of observations in an interval and true number of observations in an interval does not necessarily violate the rule (i.e., expecting 20.4 observations in an interval but having only 18 might still be OK — 20.4 and 10 certainly would not be OK).

The Standard Normal Distribution:

The **standard Normal distribution** is a Normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$, in symbols $N(0, 1)$.

If a variable X has any Normal distribution with mean μ and standard deviation $\sigma > 0$, in symbols

$$X \sim N(\mu, \sigma^2),$$

then the **standardized variable**

$$Z = \frac{X - \mu}{\sigma}$$

has the standard Normal distribution, in symbols

$$Z \sim N(0, 1).$$

Note:

- When we write $X \sim N(\mu, \sigma^2)$, we think of a Normal distribution with mean μ and standard deviation $\sigma > 0$. So, in the symbolic notation, the variance σ^2 is indicated and not the standard deviation σ. This is the common notation among almost all books — unfortunately, our textbook is one of the few exceptions. In Moore/McCabe (page 73), the notation $X \sim N(\mu, \sigma)$ is used to denote a Normal distribution with mean μ and standard deviation $\sigma > 0$. In this course, we will use the common notation (and **not** the Moore/McCabe notation), but I will usually state explicitly what the variance (or standard deviation) is.
- We can **standardize** each value x_i of a variable X through the transformation

  $$z_i = \frac{x_i - \overline{x}}{s}$$

  and call the value z_i the **z–score** of the values x_i
- If the original (X) data comes from a bell–shaped distribution, then the following version of the 68–95–99.7 rule holds for the z–scores:
  - 68% of the observations have z–scores between -1 and 1
  - 95% of the observations have z–scores between -2 and 2
  - 99.7% of the observations have z–scores between -3 and 3
- If we calculate the z–score z_i for all x_i $(i = 1, \ldots, n)$ and look at the mean and variance of our transformed data set z, we always get $\overline{z} = 0$ and $s_z = 1$, i.e., z has mean 0 and a standard deviation of 1.

Normal Distribution Calculations:

The **Standard Normal Table** (Table A) is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve to the left of z:

There is an interactive "table" on the Web at

`http://www.ruf.rice.edu/~lane/hyperstat/z_table.html`

When we are interested in the area left of a value z, we write

$$P(Z < z) = ?? = \Phi(z)$$

which reads "What is the probability (or proportion) that a standard Normal variable Z is less than a value z". In this setting, we implicitely assume that $Z \sim N(0, 1)$. To determine the ?? in the equation above, we make use of Table A (or the Web page) which displays values $\Phi(z)$ for many different z's.

Examples:

(i) $P(Z < 1.0) = ...$

$P(Z < 2.85) = ...$

$P(Z < -2.32) = ...$

(ii) $P(0.5 < Z < 1.0) = ...$

(iii) $P(Z > -0.3) = ...$

(iv) $P(Z = x) = ...$ for any possible value x,
e.g.,

(v) The Normal distribution is symmetric, i.e., often only one side of the table is printed. We can reproduce the other side (if missing) using the formula

$$\Phi(-z) = 1 - \Phi(z),$$

e.g., $\Phi(-1.55) = ...$

$1 - \Phi(1.55) = ...$

(vi) $P(\mid Z \mid > 1.32) = ...$

(vii) Find a number # such that $P(Z < \#) = 0.25$:

(viii) Find a number # such that $P(\mid Z \mid > \#) = 0.7$:

(ix) $X \sim N(1, 2^2)$:

$P(X < 3) = ...$

(x) $X \sim N(180, 30^2)$:

$P(180 < X < 200) = ...$

(xi) $X \sim N(180, 30^2)$:

Find the number # such that $P(X < \#) = 0.30$: