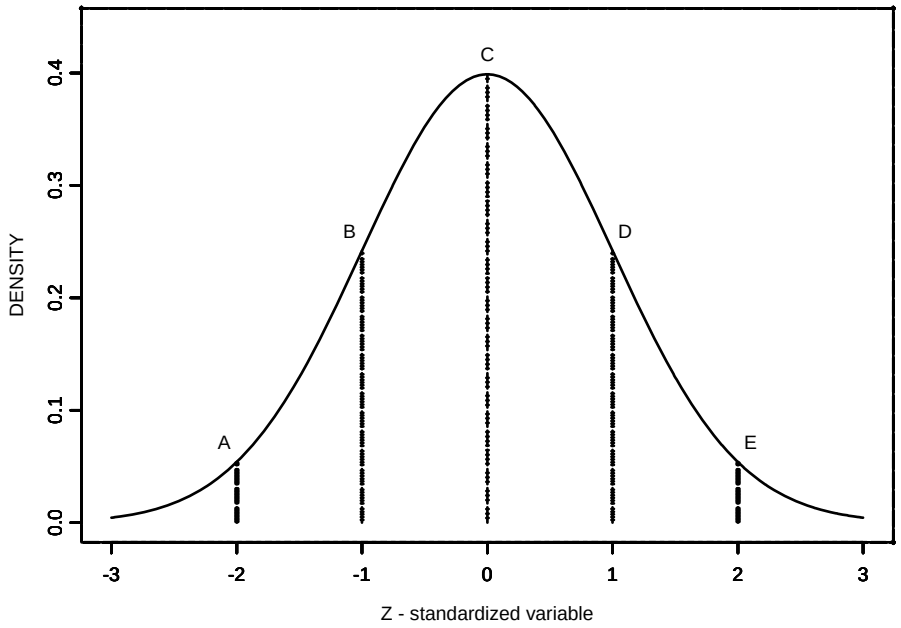
Chapter 4 Class Notes – The Normal Distribution

When the continuous random variable Y has the Normal distribution with mean μ and SD σ, we write $Y \sim N(\mu, \sigma)$, and its *density function* for $-\infty < y < \infty$ is given by

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

In this expression, $Z = \frac{Y-\mu}{\sigma}$ is the Standardization Formula



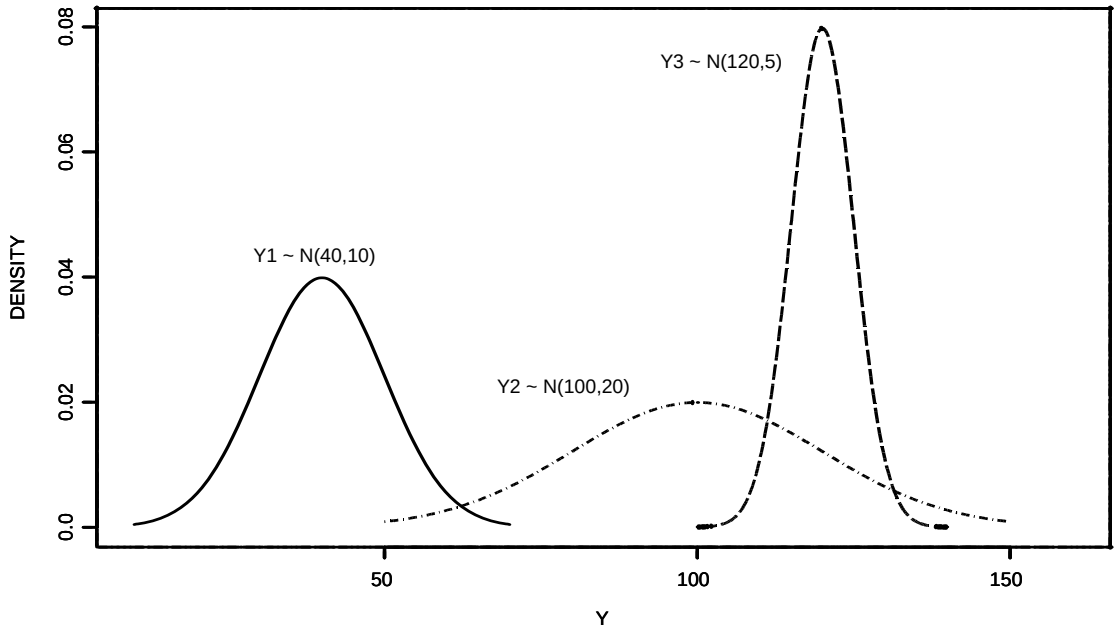
In the above graph, note that the point Z = -3 on the horizontal axis corresponds to Y = μ - 3σ, … the point Z = 3 corresponds to Y = μ + 3σ. Thus, there is a one-to-one relationship between the Z and Y horizontal axes, and we will find it advantageous to switch back and forth between these. The above curve is symmetric, and the area under it between points B and D is about 68%, between

the points A and E is about 95%, and under the graphed curve (i.e., between -3 and 3) is 99.7%.

The text (pp. 121-123) states that histograms for certain specific variables resemble Normal curves:

- **serum cholesterol for children (μ = 162, σ = 28 mg/dLi)**
- **egg-shell thickness (μ = 0.38, σ = 0.03 mm)**
- **inter-spike times in housefly nerve cells (μ = 15.6, σ = 0.4 ms)**
- **measurement error in an electronic instrument for counting white blood cells (μ = 7000, σ = 100 cells/mm^3)**

Beyond this, we'll find the Normal curve (and associated probabilities) of paramount importance for other reasons (Ch.5).



In the above graph (from p.124), note the roles of μ (center) and σ (spread) – in each case the area under the curve is 1.0 = 100%.

To find *Areas under the Normal curve* (i.e., *probabilities*), the first step is to draw the Normal curve, the second step is to standardize using the above Standardization Formula, and the third step is to use Table 3 on pp.616-17. We'll reverse this process when we seek percentiles and p-values. Be sure to also carefully work through the examples on pp. 127-132.

We'll consider here a new example: serum cholesterol for 17 year olds (Y in mg/dLi) and suppose that Y ~ N(176, 30):

a. Randomly select one such 17-year old, what is the probability that her serum cholesterol is less than 206 mg/dLi?

Draw the normal curve centered at μ = 176, and locate y = 206 on the horizontal axis. The associated standardized value is $z = \frac{206-176}{30} = 1.00$ Then, from Table 3, the area is 0.8413.
The probability is 84.13%.

b. Randomly select one such 17-year old, what is the probability that his serum cholesterol exceeds 221 mg/dLi?

Draw the normal curve centered at μ = 176, and locate y = 221 on the horizontal axis. The associated standardized value is $z = \frac{221-176}{30} = 1.50$. Then, from Table 3, the area is 1 - 0.9332 = 0.0668. The probability is 6.68%.

c. Randomly select one such 17-year old, what is the probability that her serum cholesterol is between 150 and 170 mg/dLi?

Draw the normal curve centered at μ = 176, and locate y_1 = 150 and y_2 = 170 on the horizontal axis. The associated standardized values

are $z_1 = \frac{150-176}{30} = -0.8667 \approx -0.87$ and $z_2 = \frac{170-176}{30} = -0.20$. Then, from Table 3, the area is 0.4207 - 0.1922 = 0.2285. The probability is 22.85%.

d. Find the 90th percentile of the serum cholesterol distribution.

Draw the normal curve centered at $\mu = 176$, and approximately locate the 90th percentile on the horizontal axis. Locate 0.9000 (approximately) in the *center* of Table 3, and note $z \approx 1.28$. From the Standardization Formula, $y = \mu + z\sigma = 176 + 1.28 \times 30 = 214.4$. Thus, the 90th percentile serum cholesterol value is 214.4 mg/dLi. Be sure to plot (at least in one's mind) the answer on the x-axis to make sure it makes sense.

e. Find the symmetric interval containing 90% of the serum cholesterol values.

Draw the normal curve centered at $\mu = 176$, and approximately locate the center 90% interval on the horizontal axis. Locate 0.9500 and 0.0500 (approximately) in the center of Table 3, and note $z_1 \approx -1.645$ and $z_2 \approx 1.645$. Again, from the Standardization Formula, $y_1 = 176 - 1.645 \times 30 = 126.65$ and $y_2 = 176 + 1.645 \times 30 = 225.35$. Thus, the symmetric interval containing 90% of the serum cholesterol values is from 126.65 mg/dLi to 225.35 mg/dLi. Again, plot the points on the x-axis to make sure the answer makes sense.

For a given sample, Normality can be assessed by:

- **checking the histogram to see if it resembles a Normal curve**
- **calculate $\bar{y}$ and s, the three intervals $\bar{y}$ ± s, $\bar{y}$ ± 2s, $\bar{y}$ ± 3s, count how many of the sample values are in each of these intervals and compare these actual percentages with the theoretical values from the empirical rule: 68.26%, 95.44% and 99.74%**
- **have the computer calculate the Normal Probability Plot (NPP) and check the NPP plot for linearity. Be sure to read pp.134-138 regarding NPPs.**

For a given sample, the NPP is obtained by plotting actual values versus 'normal scores' – expected quantile values which assume the (best fitting) Normal distribution – and so a straight line in this plot implies that the data resemble a Normal distribution. Note the relationship between the skews of the original histograms and the NPPs in Figs. 4.4.5-4.4.7 on pp.137-8. Repeated values produce the NPP in Fig. 4.4.8. Transformations (such as log, square root, reciprocal, etc.) can make the resulting histogram and NPP look more Normal; see Figures 4.4.9 and 4.4.10 (for log transform).

Students are asked to work the homework problems for Chapter 4, and check the answer below for the GPA exercise.

GPA = Y ~ N(μ = 2.50, σ = 0.50)

(a) Y = 2.10 gives Z = -0.80, and Y = 2.90 gives Z = 0.80, so the probability is 0.7881 – 0.2119 = 0.5762 or 57.62%.

(b) Y = 1.80 gives Z = -1.40, so the probability is 0.0808 (8.08%)

(c) Y = 3.40 yields Z = 1.80, so the probability is 1 – 0.9641 = 0.0359 or 3.59%.

(d) The 90th percentile means z* = 1.28, so y* = 2.50 + 1.28*0.50 = 3.14. The GPA of 3.14 is the 90th percentile.

(e) 2% in the RH tail means the 98th percentile → z* = 2.05 → y* = 2.50 + 2.05*0.50 = 3.525. The GPA of 3.525 has only 2% of the students above it.