

**STAT 325 – Handout 18**  
**Normal Distribution (6.2)**

- A random variable  $X$  is said to have a **normal** (or Gaussian) distribution with parameters

$\mu$  and  $\sigma$  if the pdf of  $X$  is:  $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$  for  $-\infty < x < \infty$ .

- Notation:  $X \sim N(\mu, \sigma)$
- $E(X) = \mu$ ;  $\text{Var}(X) = \sigma^2$ ;  $\text{SD}(X) = \sigma$

**Example 18-1: Exam scores**

Suppose that Professors V, W, X, and Y have very different distributions of scores on their final exams, but all follow a normal distribution:

$$V \sim N(70, 10) \qquad W \sim N(70, 5) \qquad X \sim N(80, 5) \qquad Y \sim N(80, 10).$$

a) Use Minitab to sketch this pdf for these normal distributions, and reproduce the sketches (roughly) below:

b) Describe the common shape of these pdf's.

c) Describe the impact of each parameter value ( $\mu$  and  $\sigma$ ) on the pdf's.

- The pdf of a normal distribution is a bell-shaped curve with peak at  $x = \mu$ , inflection points at  $x = \mu \pm \sigma$ .

d) For each of the four distributions above, shade the region corresponding to the probability that a randomly selected student scores 90 or higher. Which professor has the highest probability of such a score? Which has the lowest probability? Also make a guess for the values of these probabilities.

e) In principle how would you calculate these probabilities?

- The normal pdf has no closed-form expression for its integral.
  - So there is no closed-form expression for the normal cdf
  - Three options for probability calculations
    - Numerical integration
    - Normal probability table
    - Software
- Minitab calculations: Graph> Probability Distribution Plot> Normal
- R: cdf `pnorm( $\mu$ ,  $\sigma$ )`, inverse cdf `qnorm( $\mu$ ,  $\sigma$ )`, simulation `rnorm( $\mu$ ,  $\sigma$ )`

f) Use Minitab to calculate the four probabilities in d).

g) Now calculate the probability that a randomly selected student scores 70 or lower for each professor, *without* using Minitab. [*Hint*: You already know enough to calculate all four probabilities.]

h) Use Minitab to determine how high your score would have to be in order to be in the top 10%, for each professor.

- A normal distribution with mean 0 and standard deviation 1 is called a **standard normal** distribution.
  - Denoted by  $Z$
  - More notation: pdf  $\phi(z)$ , cdf  $\Phi(z)$
  - Normal probability table gives  $\Phi(z)$  for positive values of  $z$
- Key result: If  $X \sim N(\mu, \sigma)$ , then  $Z = (X - \mu) / \sigma \sim N(0,1)$ 
  - Process called **standardization**, or calculating a **z-score**
  - $z$ -score reports how many SDs above/below mean the value of interest is

i) Suppose that Ben scores 90 on Professor W's final exam and Sofia scores 90 on Professor Y's final exam. Who did better relative to his classmates? Explain, based on  $z$ -score calculations.

j) Suppose that Peter scores 65 on Professor V's final exam and Kelly scores 65 on Professor Y's final exam. Who did worse relative to his classmates? Explain, based on  $z$ -score calculations, and indicate how these  $z$ -scores differ from those in i).

**Example 18-2: Birthweights**

Birthweights of newborn babies in the U.S. can be modeled by a normal distribution with mean 3300 grams (which is about 7.3 pounds) and SD 570 grams (which is about 1.3 pounds).

a) Sketch the pdf for this normal distribution.

b) Shade the region corresponding to the probability that a randomly selected newborn weighs less than 10 pounds, which is 4536 grams. Based on your shaded region, make a guess for this probability.

c) How many SDs above the mean is a newborn who weighs 4536 grams?

d) Use the normal probability table to find  $\Phi(z)$  for the  $z$ -score that you found in c). This is the probability of what?

e) According to the normal model, what proportion of newborns in the U.S. weigh more than 10 pounds? Explain how you can calculate this from the normal probability table.

f) Babies weighing less than 2500 grams (which is about 5.5 pounds) are officially considered to be of low birth weight. How many SDs below the mean is this?

g) Express the probability that a randomly selected newborn is officially of low birth weight in terms of both functions  $\phi(z)$  and  $\Phi(z)$ .

h) Use the normal probability table to determine the probability that a randomly selected newborn is officially of low birth weight. [*Hint*: You will have to deal with the fact that the table only reports probabilities for positive values of  $z$ .]

i) Use Minitab to verify these normal probability calculations.

j) Determine how small a newborn's birthweight must be in order to be among the lightest 1% of all newborns. Also report the  $z$ -score for this weight. Do this both with Minitab and with the normal probability table.

k) Use Minitab to determine the probability that a newborn weighs within 1 SD of the mean. Then report for 2 and 3 SDs.

- These normal probability calculations form the theoretical basis for the empirical rule, which states that with data that roughly follow a bell-shaped curve:
  - $\approx 68\%$  of the data fall within 1 SD of the mean
  - $\approx 95\%$  of the data fall within 2 SDs of the mean
  - $\approx 99.7\%$  (i.e., almost all) of the data fall within 3 SDs of the mean
- This also forms the theoretical basis for the use of 1.96 in the confidence interval expression that we learned earlier.

**Example 18-3: Watching paint dry**

Suppose that the drying time for a certain type of paint under specified test conditions is known to be normally distributed with mean 75 minutes and standard deviation 4 minutes. Suppose that chemists have devised a new additive that is hoped will reduce the mean drying time (without changing the standard deviation). Suppose that a test is conducted to measure the drying time for a test specimen, and suppose that company executives decide that they will be convinced that the additive is effective only if the drying time on this specimen is less than 70 minutes.

a) If the additive actually has no effect at all on the drying time, what is the probability that the company executives will mistakenly conclude that it is effective? Include a shaded sketch with your calculation.

Now suppose that the additive really is effective and that it reduces the mean drying time to 65 minutes, without changing the standard deviation of 4 minutes.

b) Produce a sketch of the two normal curves on the same scale.

c) What is the probability that this test will fail to convince the executives that the additive is effective, even though it actually is?

d) If you want alter the cut-off value from 70 in order to reduce the error probability in a) to .025, what cut-off value should you choose?

e) Using this new cut-off value from d), what is the probability that that the test will fail to convince the executives that the additive is effective, even though it actually is?

f) How does the probability in e) compare to that in c)? Explain why this makes sense.

- With many statistical decision problems, there are two types of errors that can be made.
  - Reducing one error probability often involves a trade-off of increasing the other error probability.

Recall:

- If  $X \sim N(\mu, \sigma)$ , then  $Z = (X - \mu) / \sigma \sim N(0,1)$
- A more general result: If  $X \sim N(\mu, \sigma)$  and  $Y = aX + b$ , then  $Y \sim N(\text{_____}, \text{_____})$

a) Fill in the first blank by determining  $E(Y)$  using rules of expected values.

b) Fill in the second blank by determining  $SD(Y)$  using rules of variances.

An even more general result applies to linear combinations of random variables:

- If  $X_i \sim N(\mu_i, \sigma_i)$  for  $i = 1, \dots, k$ , and the  $X_i$ 's are independent, then  $W = \sum_{i=1}^k a_i X_i + b \sim N(\text{_____}, \text{_____})$

c) Fill in the first blank by determining  $E(W)$  using rules of expected values.

d) Fill in the second blank by determining  $SD(W)$  using rules of variances.

e) Which of the above is/are true even if the  $X_i$ 's are not independent?

**Example 18-4: SAT scores**

Suppose that SAT scores of applicants to a particular university can be modeled as a normal random variable with mean 600 and SD 60 on the Math portion and with mean 550 and SD 70 on the Verbal portion.

a) Determine the probability that the combined score exceeds 1200. Also state any assumptions that you must make in order to perform this calculation.

b) Determine the probability that the Math score exceeds the Verbal score. Also state any assumptions that you must make in order to perform this calculation.

c) Do you think the assumption you had to make is reasonable in this context? Explain.

**Example 18-5: Random Rendezvous**

Suppose that you and a friend agree to meet for lunch at a certain restaurant, but both of your arrival times are random variables, independent of each other. Each of you agrees to wait exactly 15 minutes before giving up and leaving if the other has not yet shown up. Suppose that each of your arrival times, in minutes after noon, follows a normal distribution with mean 30 and SD 10.

a) Make a guess for the probability that the two of you will actually meet.

b) Approximate this probability with an R simulation. Start with only 10,000 repetitions, so you can examine a scatterplot of the pair of arrival times. Then proceed to 1,000,000 repetitions for a more accurate approximation. Include a margin-of-error with your approximation.

c) Determine the probability that the two of you actually meet. Use good notation, and explain every step of your calculation. [*Hint*: Express the event of interest in terms of a particular random variable.]