Normal Distributions

Section 15.4

Dr. John Ehrke
Department of Mathematics

Fall 2012

Measures of Relative Standing

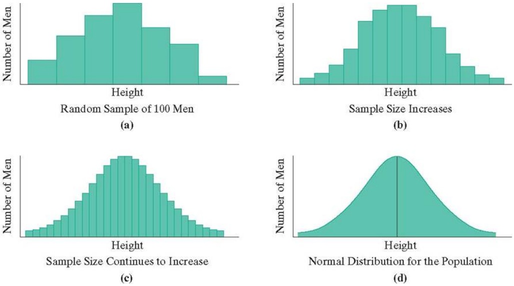
Sometimes you need to know the position of one data point relative to all other data points within a given population. For example, if you took a standardized test in high school, like the SAT or the ACT, you, but especially the college you applied to for entrance, will want to know how your score of 1500 on the SAT compares to someone else's score of 1450.

- Is 1500 really that much higher relatively speaking?
- What if the school wanted to compare your score versus another student's ACT score? They are different exams, so is there a way to compare them?
- If the school required you to score in the top 30% of all those who took the exam, does your score make the cut?

We will answer these and other questions like them in this section.

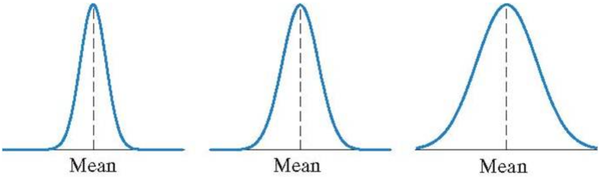
Probability Distribution

Suppose you have a set of measurements on a continuous random variable, and you create a relative frequency histogram to describe the data set. For a small number of measurements, only a few classes is needed, but as more and more measurements are collected, the need for more classes (and therefore shorter class widths) becomes necessary. As the number of measurements becomes very large, and the class widths become very narrow, the relative frequency histogram appears more and more like a smooth curve. This smooth curve describes the *probability distribution of the continuous random variable*.



Shape of a Normal Distribution

The shape of the normal distribution depends on the mean and standard deviation. These three graphs have the same mean, but different standard deviations. As the standard deviation increases, the distribution becomes more spread out.



Sample Z-Score

The mean and standard deviation of a sample can be used to calculate a *z-score*, which measures the relative standing of a measurement in a data set.

Definition

The **sample z-score** is a measure of relative standing defined by

$$\text{z-score} = \frac{x - \bar{x}}{s_x}.$$

If the population mean, μ, and the population standard deviation σ are known, we use these in place of $\bar{x}$ and s_x, respectively.

What does a z-score tell you?

College-Bound Student Data for SAT 2012

Data in this report are for high school graduates in the year 2012. Information is summarized for seniors who took the SAT at any time during their high school years through June 2012. If a student took the test more than once, the most recent score is used. The overall mean scores for the major areas are listed below.

Table 1: Overall Mean Scores

								Writing Subscores			
SAT	Test-Takers	Critical Reading		Mathematics		Writing *		Multiple Choice		Essay	
	Number	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total	1,664,479	496	114	514	117	488	114	48.9	11.5	7.2	1.6

Example

Draw distributions for each of the three main areas: Critical Reading, Mathematics, and Writing and determine the z-score associated with a reading score of 720, math score of 600, and writing score of 400.

College-Bound Student Data for SAT 2012

Polling Question #19

Suppose you are given a set of data that is normally distributed with mean, $\bar{x} = 18$ and standard deviation, $s_x = 5$. Which of the following is true?

(a) The z-score corresponding to the data item $x = 7$ is 2.2.

(b) The data item $x = 25$ has a smaller z-score than $x = 7$.

(c) The data item $x = 25$ is 1.4 standard deviations above the mean.

(d) The data item $x = 7$ is 2.2 standard deviations above the mean.

Polling Question #20

A student scores 60 on a vocabulary test and 80 on a grammar test. The data items for both tests are normally distributed. The vocabulary test has a mean of 50 and a standard deviation of 5. The grammar test has a mean of 72 and standard deviation of 6. Which of the following best describes the student's relative performance on the exams?

(a) The student's relative performance was better on the vocabulary test.

(b) The student's relative performance was better on the grammar test.

(c) The student did equally well on both exams.

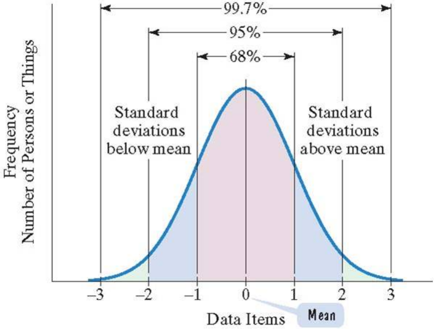
(d) The student's relative performance cannot be determined.

Empirical Rule

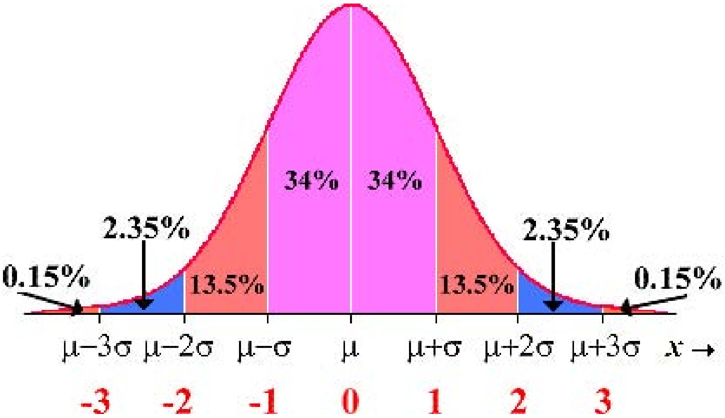
Theorem (Empirical Rule)

Given a distribution of measurements that is approximately mound-shaped:

- *The interval* $(\mu \pm \sigma)$ *contains approximately 68% of the data.*
- *The interval* $(\mu \pm 2\sigma)$ *contains approximately 95% of the data.*
- *The interval* $(\mu \pm 3\sigma)$ *contains approximately 99.7% of the data.*



Modified Empirical Rule



Applying the Empirical Rule

Example

Load the "`bodytemp.txt`" file on your calculator. In this example, we will use the empirical rule to determine if the population is normally distributed.

1. Calculate the sample mean, $\bar{x}$ and the sample standard deviation, s_x.
2. Use the Empirical rule to describe the distribution of data points between $\bar{x} \pm s_x$, $\bar{x} \pm 2s_x$, and $\bar{x} \pm 3s_x$.
3. How does this estimation compare to the actual data?
4. Based on your observation, do you think the data is normally distributed?

There are a wide variety of "normality tests" used by statisticians to answer the same kind of question we are attempting to answer in this question. While such tests are beyond the level of this course all require a large sample size, i.e. $n > 20$ and more realistically $n > 50$ to be of any great value.

Polling Question #21

The scores on a test are normally distributed with a mean of 70 and standard deviation of 7. What percentage of students scored between a 63 and 84?

(a) 13.5%

(b) 84%

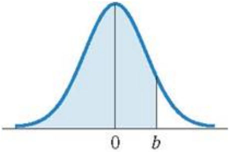
(c) 47.5%

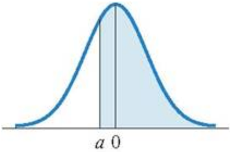
(d) 81.5%

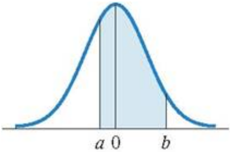
The Normal Cumulative Distribution Function (CDF)

There are three basic types of questions we are interested in concerning areas underneath the probability density function.

- $P(x \leq b)$
- $P(x \geq a)$
- $P(a \leq x \leq b)$







The Basic Idea

Example

Consider a normally distributed set of data with mean, $\bar{x} = 10$ and standard deviation, $s_x = 2$.

(a) Find the probability that x lies between 11 and 13.6.

(b) Find the probability that x is greater than 12.2.

(c) Find the probability that x is less than 8.7.

Normal CDF Question

Example

Studies show that the gasoline usage of compact cars sold in the United States is normally distributed with mean 25.5 miles per gallon (mpg) and a standard deviation of 4.5 mpg.

(a) What percentage of compacts get 20 mpg or more?

(b) Which is more likely? Finding a compact car that gets more than 50 mpg, or finding a compact car that gets less than 10 mpg?

(c) What percentage of compacts get between 16.5 and 34.5 mpg?

Calculating Percentiles

Suppose we changed the previous question up slightly and rather than ask what percentage of compact cars fell in certain intervals we asked you to find the miles per gallon a compact car must have to be considered in the top 5% among all other compacts produced during that year. In other words, find x_0 such that $P(x \leq x_0) = 0.95$.

This is the *95th percentile* of the distribution, and can be found as follows:

Solution: We begin by trying to calculate the z-score for such a point.

$$z_0 = \frac{x_0 - 25.5}{4.5}$$

Since the value of z_0 corresponds to x_0, it must also have area .95 to its left. In the z-score table, we find that the z-score which corresponds to this point is $z_0 = 1.64$. Thus we have

$$z_0 = \frac{x_0 - 25.5}{4.5} = 1.64.$$

Solving for x_0 we obtain $x_0 = 32.88$. So a car must get at least approximately 33 mpg to be in the top 5% among all compact cars.

Inverse Normal Command

(a) What z-score is the 80th percentile?

(b) What z-score is the 30th percentile?

(c) Between what two z-scores is the middle 40% of the data?

(d) If a distribution has a mean of 40 and a standard deviation of 4, what is the 75th percentile?

(e) If you know that the mean salary for your profession is \$53,000 with a standard deviation of \$2500, to what percentile does your salary \$57,000 correspond?

```
1  // To find the score corresponding to a certain percentile, you can
   use the InverseNormal(percentile, mean, standard deviation)
   command.
2  InverseNormal(0.8,0,1)
   0.84162                                                        D
3  //This calculates the z-score corresponding to the 80th percentile. (
   z= 0.84)
```

Grading on the Curve

Example

Sometimes grading on a curve isn't always a good thing. Suppose an instructor grades on a curve by assuming the test scores are normally distributed. If the average grade is 70 and the standard deviation is 8, answer the questions below if the instructor wishes to assign grades as follows: 10% A's, 20% B's, 40% C's, 20% D's, and 10% F's.

(a) If you made a 75 on the exam, and the exam was graded on the curve described above, what grade would you receive?

(b) What is the lowest grade you could make and still receive an A on the exam?

(c) What is the cutoff for passing with a C?

(d) To be considered in the top quarter of the class, what grade would you have to make?

Grading on the Curve

Polling Question #22

An honors program requires a student score in the top 2.5% on a particular exam to be considered for entry into the program. From past experience, out of all those who have taken the exam, the average test score is 100 and the standard deviation is 5. What is the minimum score a student must make to be considered by the honors program? Answers are rounded to the nearest whole number.

(a) 110

(b) 108

(c) 90

(d) 100

Margin of Error for the Sample Mean

The margin of error formula to calculate the margin of error of a sample mean, provided that we have a sample from a population that is normally distributed and know the population (or sample) standard deviation is given by

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and the associated confidence interval is

$$CI = \bar{x} \pm z_{\alpha/2} \cdot E$$

$\alpha =$ level of confidence, usually $\alpha = 0.05$ represents 1 - 0.05 = .95 = 95% confidence level
$z_{\alpha/2} =$ this is the point on the standard normal for which $\alpha/2$ area lies above this point
$\sigma =$ population standard deviation
$n =$ sample size

Margin of Error and Confidence Intervals

Example

Use the margin of error formula above to calculate the margin of error in the following presidential poll where 700 likely voters (LV) were sampled:

Obama	Romney
48 %	52%