Working with z-Scores 2:
Surveys & the Central Limit Theorem

By now, you should be well familiar with the normal distribution and the normal curve. You should also be familiar with the z-score table and how it works. If you aren't comfortable working with these things, review those subjects before starting this worksheet. This worksheet will help you to take what you've learned about z-scores and apply it to surveys.

In statistical studies, there's always a chance that when we pick a single item at random from a population that it won't be close to the average for the population. If we're trying to measure the height of the average American male, randomly picking Tom Cruise or Shaquille O'Neal isn't going to give us a true indication of what that value is.

We minimize the risk of picking a statistical weirdo by picking multiple members of the population to study. A batch of members of the sample space, studied together, is called a **survey**.

The mean of a sample should be the same as the mean of the individual members of the population, or close to it, but the standard deviation will be smaller. After all, even if my survey of American men gives me Tom Cruise *and* Shaquille O'Neal, at least they'll help to average each other out.

For a population that has a mean of μ and a standard deviation of σ for some statistical variable x, we can take a sample of size n and take the mean of the measurements in the sample to get $\bar{x}$ (which we read "x-bar"). We can define a second statistical variable concerning all possible samples of that size from that population, and look at all those possibilities as a new sample space. This new variable has a distribution called a **sampling distribution**. When n is large enough (we'll define "large enough" in a moment), the sampling distribution is approximately normal, with a mean equal to μ, the mean of the population, and a standard deviation of $\frac{\sigma}{\sqrt{n}}$, derived from the standard deviation of the population and the sample size. This result is called the **Central Limit Theorem**.

The good news for statisticians is that the Central Limit Theorem is still true even if the distribution of x, the variable for the individuals, isn't distributed normally! Taking a sample evens out the distribution and makes it more normal the bigger n gets.

This brings us back to "large enough" and what it means. The minimum usable value for n will vary depending on what the distribution of the individuals is. If that's normal, then any n will do. If the distribution is skewed or bimodal, or strange in some other way, n must be much higher. An n higher than 30 is safe, according to your textbook.

If the survey is large enough to give a normal sampling distribution of $\bar{x}$, we can use the z-table to find out things about the mean of the population from the mean of the survey.

The formula for a z-score for x, if the distribution of X is normal, is:

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

This means that a z-score for a survey mean, $\bar{x}$, would be:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Example 1: For a study of bone brittleness, the ages of people at the onset of osteoporosis following a normal distribution with a mean age of 71 and a standard deviation of 2.8 years. What is the probability of:

a) selecting one person who had the onset of osteoporosis at age 68 or less?

b) having a sample of 5 people who had an average age of onset of osteoporosis at age 68 or less?

c) having a sample of 50 people who had an average age of onset of osteoporosis at age 68 or less?

Solution: a) First, note that if the question hadn't specified that if the age of onset of osteoporosis wasn't distributed normally, we couldn't even answer this question! Since the ages are normal, X ~ N(71, 2.8), so we use the z-score:

$$z = \frac{x - \mu}{\sigma} = \frac{68 - 71}{2.8} = -1.07\ldots$$

We look this up in the z-score table and get the answer 0.1423. Not likely, but it will happen about 1 time in 7.

b) Now we have a sample. Since x is normal, $\bar{x}$ is also, even at this low sample size. The mean for the sample is still 71 and the standard deviation is now $\frac{2.8}{\sqrt{5}} = 1.252\ldots$ The standard deviation has gone down. The z-score is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{68 - 71}{2.8/\sqrt{5}} = -2.40\ldots$$

This time we get a probability of 0.0082. This means that the probability of getting a sample mean of 68 years or less will occur less than 1% of the time.

c) The sample has gotten larger again, and our new standard deviation is $\frac{2.8}{\sqrt{50}} = 0.396\ldots$ The z-score is:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{68 - 71}{2.8/\sqrt{50}} = -7.58\ldots$$

The probability of getting a sample mean that's off the real mean for the population by 3 years is now so remote it isn't even on the table. That's how powerful even a small sample can be.

EXERCISES

A. In a certain school district, the distribution of the heights of eighth-graders who play the tuba is approximately normal, with a mean of 146.1 cm and a standard deviation of 8.4 cm. Determine the probability that

1) a randomly chosen tuba player from the eighth grade is shorter than 150 cm?

2) two random eighth-grade tuba players have an average height less than 150 cm?

3) ten random eighth-grade tuba players have an average height less than 148 cm?

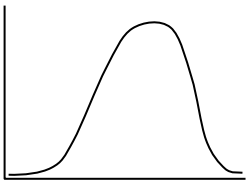
4) 30 random eighth-grade tuba players have an average height less than 148 cm?

B. In a study for an acne medication, dermatologists count the number of blemishes on patients' faces.

1) Could we calculate the probability that a patient has more than 10 blemishes using the z-score table? Why or why not?

2) The study only has 28 test subjects. What would have to be true about the distribution of the individuals' numbers of blemishes to justify using the Central Limit Theorem with a number this low?

C. A statistical variable, X, has a mean, μ, of 121 and a standard deviation, σ, of 7.31. A curve showing the distribution of this variable is on the right. What is the probability that:



1) a randomly chosen population member has a value for x that is 123 or more?

2) when we randomly pick 25 members of the population and average their x-values, we get an $\bar{x}$ of 123 or more?

3) when we randomly pick 75 members of the population and average their x-values, we get an $\bar{x}$ of 123 or more?

D. In the example from the main section of the worksheet, we saw that a survey size of 5 was not large enough to guarantee that our survey's mean was not 3 years less than the sample mean, and 50 was.

1) If we want to guarantee that the survey's mean being more than 3 years below the true population mean is statistically impossible, i.e. the z-score is −3.50 or less, what is the smallest value for n we can choose?

2) If we want a guarantee of not more than 1 year below the true population mean, what is the smallest value for n we can choose?

SOLUTIONS

A. (1) 67.72% (2) 74.54% (3) 76.42% (4) 89.25%

B. (1) No, because the z-table can only be used with the normal distribution. Since counting something makes the variable discrete, it can't be normally distributed.

(2) The distribution of the data would have to be close to normal: nearly symmetrical, unimodal, and shaped like a bell curve.

C. (1) We cannot know, since the population distribution is not normal. Since it's close to normal, we can still do (2) and (3). (2) 8.53% (3) 0.89%

D. (1) $n \geq 11$ (2) $n \geq 96$