

Fall 2015

Mathematics in Contemporary Society Chapter 8

Patrick J. Wallach

Queensborough Community College

[How does access to this work benefit you? Let us know!](#)

Follow this and additional works at: http://academicworks.cuny.edu/qb_oers

 Part of the [Mathematics Commons](#)

Recommended Citation

Wallach, Patrick J., "Mathematics in Contemporary Society Chapter 8" (2015). *CUNY Academic Works*.
http://academicworks.cuny.edu/qb_oers/9

This Textbook is brought to you for free and open access by the Queensborough Community College at CUNY Academic Works. It has been accepted for inclusion in Open Educational Resources by an authorized administrator of CUNY Academic Works. For more information, please contact AcademicWorks@cuny.edu.

Chapter 8

More Normal Distribution Problems

We'll look at another problem similar to what we did in Week 7's notes, with some new questions. Here is the z-score percentile chart from last week:

| z-score | Percentile | z-score | Percentile | z-score | Percentile |
|---------|------------|---------|------------|---------|------------|
| -4.0 | 0.003% | -1.1 | 13.57% | 1.4 | 91.92% |
| -3.5 | 0.02% | -1.0 | 15.87% | 1.5 | 93.32% |
| -3.4 | 0.03% | -0.9 | 18.41% | 1.6 | 94.52% |
| -3.3 | 0.05% | -0.8 | 21.19% | 1.7 | 95.54% |
| -3.2 | 0.07% | -0.7 | 24.20% | 1.8 | 96.41% |
| -3.1 | 0.10% | -0.6 | 27.43% | 1.9 | 97.13% |
| -3.0 | 0.13% | -0.5 | 30.85% | 2.0 | 97.72% |
| -2.9 | 0.19% | -0.4 | 34.46% | 2.1 | 98.21% |
| -2.8 | 0.26% | -0.3 | 38.21% | 2.2 | 98.61% |
| -2.7 | 0.35% | -0.2 | 42.07% | 2.3 | 98.93% |
| -2.6 | 0.47% | -0.1 | 46.02% | 2.4 | 99.18% |
| -2.5 | 0.62% | 0.0 | 50.00% | 2.5 | 99.38% |
| -2.4 | 0.82% | 0.1 | 53.98% | 2.6 | 99.53% |
| -2.3 | 1.07% | 0.2 | 57.93% | 2.7 | 99.65% |
| -2.2 | 1.39% | 0.3 | 61.79% | 2.8 | 99.74% |
| -2.1 | 1.79% | 0.4 | 65.54% | 2.9 | 99.81% |
| -2.0 | 2.28% | 0.5 | 69.15% | 3.0 | 99.87% |
| -1.9 | 2.87% | 0.6 | 72.57% | 3.1 | 99.90% |
| -1.8 | 3.59% | 0.7 | 75.80% | 3.2 | 99.93% |
| -1.7 | 4.46% | 0.8 | 78.81% | 3.3 | 99.95% |
| -1.6 | 5.48% | 0.9 | 81.59% | 3.4 | 99.97% |
| -1.5 | 6.68% | 1.0 | 84.13% | 3.5 | 99.98% |
| -1.4 | 8.08% | 1.1 | 86.43% | 4.0 | 99.997% |
| -1.3 | 9.68% | 1.2 | 88.49% | | |
| -1.2 | 11.51% | 1.3 | 90.32% | | |

Example 1

Suppose the North Atlantic Sea Bass has a length that is normally distributed with a mean of 29.7" and a standard deviation of 5.7". Answer the following questions:

- 1) Find the percentile of a catch of 35".
 - a) First we calculate the z-score using the formula:

$$z = \frac{\text{data value} - \text{mean}}{\text{standard deviation}}$$

$$= \frac{(35 - 29.7)}{5.7}$$

$$= \frac{5.3}{5.7}$$

$$= 0.929824561$$

$$= 0.9 \text{ (rounded to nearest hundredth on the chart)}$$

b) We can then look up $z = 0.9$ as a percentile. (See the table.) We find the corresponding percentile is **81.59%**. This means that **81.59%** of the North Atlantic Sea Bass are less than 35" in length.

2) What percentage of bass has a length below 20"?

a) First we calculate the z-score using the formula:

$$z = \frac{\text{(data value - mean)}}{\text{standard deviation}}$$

$$= \frac{(20 - 29.7)}{5.7}$$

$$= \frac{-9.7}{5.7}$$

$$= -1.701754386$$

$$= -1.7 \text{ (rounded to nearest tenth)}$$

b) We can then look up $z = -1.7$ as a percentile. The corresponding percentile is **4.46%**. This means that **4.46%** of the bass are less than 20" in length.

3) What percentage of bass has a length above 40"?

a) First we calculate the z-score using the formula:

$$z = \frac{(40 - 29.7)}{5.7}$$

$$= \frac{10.3}{5.7}$$

$$= 1.8 \text{ (rounded to nearest tenth)}$$

b) We can then look up $z = 1.8$ as a percentile. We see 96.41%. This means that 96.41% of the lengths are below 40. But we want the percentage above. If 96.41% of the values are below 40, then **100.00% - 96.41% = 3.59%** of the bass are over 40" in length.

4) What percentage of bass has a length above 20”?

a) If you know the z-score from previous example. Use it:

We saw before that $z = -1.7$ when the length is 20”

b) We know that when $z = -1.7$, the percentile is 4.46%. If 4.46% of the values are below 20”, then $100.00\% - 4.46\% = 95.54\%$ of the bass are over 20” in length.

5) What percentage of percentage of bass has a length between 25” and 37”?

a) We have two values, so we need to calculate two z-scores:

$$\text{For data value}=25 \quad z = \frac{(25 - 29.7)}{5.7}$$

$$= \frac{-4.7}{5.7}$$

$$= -0.824561404$$

$$= -0.8 \text{ (rounded to nearest hundredth on the chart)}$$

$$\text{For data value}=37 \quad z = \frac{(37 - 29.7)}{6.8}$$

$$= \frac{7.3}{6.8}$$

$$= 1.3 \text{ (rounded to nearest tenth)}$$

b) We can then look up both percentiles. For $z = -0.8$, 21.19% of the scores are below 25; for $z = 1.3$, 90.32% of the scores are below 37. If 90.32% of the scores are below 37, and 21.19% of that represents scores below 25, then we can take the difference between the two. $90.32\% - 21.19\% = 69.13\%$ of the bass have lengths between 25” and 37”.

6) We can use what we know about the normal distribution and the table to examine a claim made about a member of the population.

Suppose Pete the Fisherman claims that he once caught a bass that was 50” in length. What do you think of his claim?

a) We can calculate the z-score of 50”:

$$z = \frac{(50 - 29.7)}{5.7}$$

$$= \frac{20.3}{5.7}$$

$$= 3.561403509$$

$$= 3.5 \text{ (rounded to nearest value on the chart)}$$

- b) We can look up the percentile for $z = 3.5$, which is 99.98%. Are we willing to believe that Pete caught a fish at the 99.98th percentile? I don't know. Is it an impossible feat? No. Is it very unlikely? Yes. We have to make a judgment here about the reliability of Pete's claim. (We may also wonder if the population mean and standard deviation are correct. But that's another story for now.)

- 7) What bass length corresponds to the 95th percentile?

This is obviously a different sort of problem. We're given the percentile so we need to work in reverse.

- a) First we should find the z-score that has a percentile of 95%. We see that $z = 1.6$ has a percentile of 94.52% and $z = 1.7$ has a percentile of 95.54%. We shall use $z = 1.7$ then. (*We tend to round upward to get at least the percentile desired, particularly when the percentile is right in the middle.*)

- b) We plug the information that we know into the z-score formula and solve for the missing data value:

$$z = \frac{(\text{data value} - \text{mean})}{\text{standard deviation}}$$

$$1.7 = \frac{(\text{data value} - 29.7)}{5.7}$$

$$9.7 = \text{data value} - 29.7 \text{ (multiply both sides by 5.7)}$$

$$\mathbf{39.4 = \text{data value}}$$
 (add 29.7 to both sides)

So 39.4" is the length that corresponds to the 95th percentile. 95% of the bass have lengths below 39.4"

Note: This shortcut formula works even better:

$$\text{data value} = (z)(\text{standard deviation}) + \text{mean}$$

(Using the values from the example:)

$$\text{length} = (1.7)(5.7) + 29.7$$

$$\text{length} = 9.7 + 29.7$$

$$\text{length} = \mathbf{39.4}$$

- 8) What bass length corresponds to the 99th percentile?

This will be much easier the second time around.

a) First we should find the z-score that has a percentile of 99%. We see that $z = 2.3$ has a percentile of 98.93% and $z = 2.4$ has a percentile of 99.18%. We shall use $z = 2.3$ then. (*In this case 2.3 is much closer to 99%. But using 2.4 would be OK as well.*)

b) We plug the information that we know into the shortcut formula and obtain the missing data value:

$$\text{data value} = (z)(\text{standard deviation}) + \text{mean}$$

$$\text{length} = (2.3)(5.7) + 29.7$$

$$\text{length} = 13.1 + 29.7$$

$$\text{length} = \mathbf{42.8}$$

So 42.8" is the length that corresponds to the 99th percentile. 99% of the bass have lengths below 42.8"

If $z=2.4$ was used:

$$\text{data value} = (z)(\text{standard deviation}) + \text{mean}$$

$$\text{length} = (2.4)(5.7) + 29.7$$

$$\text{length} = 13.7 + 29.7$$

$$\text{length} = \mathbf{43.4}$$

So 43.4" is the length that corresponds to the 99th percentile. 99% of the bass have lengths below 43.4" We see that using $z=2.3$ or $z=2.4$ leads to slight differences.

- 9) We often are interested in the 90th, 95th and 99th percentile. But any other percentile may come up in a problem.

Suppose the Department of Fishing decides the size restriction for bass should be set at the 80th percentile (Anything caught that is shorter should be thrown back). What would the restriction be?

a) First we should find the z-score that has a percentile of 80%. We see that $z = 0.8$ has a percentile of 78.81 and $z = 0.9$ has a percentile of 81.59. We shall use $z = 0.9$ (We decide to round up because the value is in the middle.)

b) We plug the information that we know into the shortcut formula and obtain the missing data value:

$$\text{length} = (0.9)(5.7) + 29.7$$

$$\text{length} = 5.1 + 29.7$$

$$\text{length} = \mathbf{34.8}$$

Anything bass under 34.8" should be thrown back.

Question 1: A normal distribution of MA-321 test scores has a mean of 82.7 and a standard deviation of 5.3. Answer the following:

- a) What percentage of the test scores are below 78?
- b) What percentage of the test scores are above 90?
- c) What percentage of students has test scores between 80 and 95?
- d) What test score is at the 95th percentile?
- e) What test score is at the 99th percentile?

You should be able to handle any normal distribution problem now.

Design your own normal distribution!

Make up your own problem!!! Fill in the blanks (make up values that seem appropriate) and answer the questions:

The GPA of a college student is a normal distribution with a mean of ____ and a standard deviation of _____.

- 1) Use the 68-95-99.7 Rule to find the values in the normal distribution that are within:
 - a) One standard deviation (68% of data)
 - b) Two standard deviations (95% of data)
 - c) Three standard deviations (99.7% of data)

(For 2) –4), fill in your own values and solve.)

- 2) Your GPA is _____. What is your percentile rank?

- 3) Find the percentage of students with a GPA over _____.

- 4) Find the percentage of students with a GPA between _____ and _____.

(Use what we learned about solving for data values to answer this.)

- 5) 95% of the GPAs are below what value?

- 6) 99% of the GPAs are below what value?

Lab Assignment #7 – Z-scores & Percentiles
Due _____

Use the data set of the 50 values that you have gathered for your project.
You can place these values in cells A2 to A51. In cell A1, put **Values**.

- 1) Calculate the:
 - a) mean
 - b) standard deviation(These should be the same results from Lab #6.) Place these results in columns B and C, or in column A below your data set.

- 2) We can use the function **STANDARDIZE**(value, mean, standard deviation) to convert any value to a z-score when the mean and standard deviation are supplied. Here's how we do it:
 - a) Go to the top of an empty column, let's say cell D1. Type the word **z-scores**. Go to D2.
 - b) Click on f_x
 - c) Choose the **Statistical** category and the function name **STANDARDIZE**. You will be asked for three values (X, Mean and Standard_dev).
 - d) For X, click on the location of your first value (A2).
 - e) For Mean, click on the location of the mean (or type in the value itself).
 - f) For Standard_dev, click on the location of the standard deviation (or type in the value itself).
 - g) The result that appears is the z-score.
 - h) Copy this result for the other 49 values to get 50 z-scores total.

***Note:** This will work just fine if you just type in the values for the mean and standard deviation in e) & f). Otherwise the formula location of the mean and standard deviation shift and will not work properly. The other trick is to use \$ in the*

mean and standard_dev field—if C1 is the location of the mean, type \$C\$1, if C2 is the location of the standard deviation, type \$C\$2. When you copy the formula, the other 49 calculations will have the correct mean and standard deviation. I will demonstrate this in the lab.

- 3) We can use the function **NORMDIST**(value, mean, standard deviation,TRUE) to convert any value to a percentile (as a decimal).
 - a) Go to the top of an empty column, let's say cell E1. Type the word **percentile**. Go to E2.
 - b) Click on f_x
 - c) Choose the **Statistical** category and the function name **NORM.DIST**. You will be asked for four values (X, Mean, Standard_dev and Cumulative).
 - d) Repeat the procedure from 2) to obtain the values for X, Mean, and Standard_dev.
 - e) For Cumulative, type the word **TRUE**
 - f) The result that appears is the percentile.
 - g) Click on % to convert the value to a percentage.
 - h) Copy this result for the other 49 values to get 50 percentiles total. (Consult the note from earlier.)

- 4) The function **NORM.INV**(percentile as percentage, mean, standard deviation) will provide the value corresponding to a specific percentile.
 - a) Go to the top of an empty column, let's say cell F1. Type the words **90th percentile, 95th percentile and 99th percentile** in cells F1 to F3. Go to G1.
 - b) Click on f_x
 - c) Choose the **Statistical** category and the function name **NORM.INV**. You will be asked for three values (Probability, Mean and Standard_dev).
 - d) For probability, type in the probability desired as a percentage. Type 90%

e) Repeat the procedure from 2) to obtain the values for Mean, and Standard_dev.

f) The value that appears is the value that corresponds to the 90th percentile. Any values from your data set that are close to this value should have a percentile near 90%.

g) Copy this result in cells G2 and G3 to obtain the value of the 95th and 99th percentile (use 95% and 99%).

Your results will look something like this (with 50 values, not 10).

| <u>Values</u> | | | <u>z-</u> <u>scores</u> | <u>percentile</u> | 90th percentile | 92.76186 |
|---------------|-------------|----------|----------------------------|-------------------|----------------------------|----------|
| 15 | Mean | 51.09091 | -1.11 | 13.4% | 95th percentile | 104.575 |
| 24 | Stan | | | | 99th percentile | 126.7344 |
| 24 | Dev | 32.51601 | -0.83 | 20.2% | | |
| 25 | | | -0.80 | 21.1% | | |
| 45 | | | -0.19 | 42.6% | | |
| 86 | | | 1.07 | 85.8% | | |
| 62 | | | 0.34 | 63.1% | | |
| 43 | | | -0.25 | 40.2% | | |
| 115 | | | 1.97 | 97.5% | | |
| 79 | | | 0.86 | 80.5% | | |
| 56 | | | 0.15 | 56.0% | | |
| 12 | | | -1.20 | 11.5% | | |

You will use the results from Lab#6 and Lab #7 for your project report. (Details will be given out on in class.)

Answer the following questions:

- Does your sample have high or low variation? Why? What is suggested by the five-number summary, range (from lab #6) and z-scores?
- Identify any outliers in your data set. (Outliers are values in your data set with particularly high/low z-scores (less than -2 or more than 2) or percentiles (less than 3% or above 97%).)

The following will conclude our work with statistics. This material is more difficult to absorb. This will be used for extra credit on the exam.

Statistical Inference

We have explored many ways to measure the values of a data set. There are measures of average, the mean, median and mode, which summarize data set in terms of one central value. Measures of variation, the range, five-number summary and standard deviation, supply an overall understanding of whether data sets values are close together or far apart. The shape of a distribution, whether it's normal, uniform, left-skewed, right-skewed or bimodal give a visual sense of how the data values are arranged. Percentiles and z-scores provide an understanding of how individual data values relate to the normal distribution of the population.

At the center of statistical research is the use of sample results to infer conclusions about the population. This process is referred to as **statistical inference**.

Your statistical project is all about statistical inference. By gathering a sample of values from a population, calculating measures of average and variation, creating graphical displays, and calculating z-scores and percentiles you are obtaining an understanding of the larger population itself.

If you believe your sample is a fair representation of the population, sample statistics should be close to population parameters. If your sample mean is 92.3, you should expect the population to be close to 92.3. If your sample shows high variation, you can expect the same from the population. If your sample seems to be a left-skewed distribution, you may believe the population is also left-skewed.

You may have found that your data set contained outliers, which are extremely different from the other values in your data set. How significant are these outliers when they occur? What is the idea of "significance" anyway?

What's Significant?

One of the struggles that a statistical researcher must undergo is whether or not a particular observation is meaningful. A result is said to be **statistically significant** if it is unlikely to have occurred under normal, ordinary circumstances. Statistically significant events usually necessitate an explanation.

Example 1

If you flip a coin 4 times and end up with 4 heads and 0 tails, you would not consider such an event to be statistically significant, although a result of 2 heads and 2 tails seems the most probable.

However, if you flipped the same coin 100 times and obtained a result of 90 heads and 10 tails, you would consider such an event to be statistically significant. You would probably conclude that the coin was unfair and improperly favored heads over tails.

But is it impossible to flip a coin 100 times and obtain 90 heads? No, it's not impossible. But it is considered so improbable that we would have serious doubts about its occurrence.

Question 2: Suppose you roll a six-sided die 100 times. Give an example of an event that is not statistically significant and another example of an event that would be statistically significant.

Example 2

A baseball team rallies from 0-3 deficit in a championship series and wins the series 4-3. Is this statistically significant? Probably not. It requires the team to win four games in a row, which is a difficult but hardly impossible task. What makes it seem more interesting to fans is the timing of the event (during the playoffs). It is less interesting during the regular course of the season.

Winning streaks are not unusual in sports, so when a team or player wins several games in a row, it is not necessarily an outstanding achievement. But as the wins pile up: 10, 20, 30, etc., the event reaches statistical significance.

Question 3: The New England Patriots once won 21 games in a row. Is this statistically significant? Why or why not?

Example 3

We saw many presidential polls in 2012, and the numbers seemed to keep changing. Different polls had shown Barack Obama to have 51%, 53%, 56%, 50% and 49% of the vote at different times. We saw Mitt Romney generally with smaller figures, but not always. When does a change in these percentages become statistically significant?

For one thing, these polls generally have some **margin of error**. If Barack Obama is said to have 51% of the vote with a margin of error of 4%, the actual percentage is 47% to 55%. If a later poll shows that he has 50% of the vote with a margin of error of 2%, the actual percentage is 48% to 52%. Since the intervals overlap, there may be no real change at all. A statistically significant result should have no such overlap. If a still later poll shows Barack Obama to have 55% of the vote with a margin of error of 1%, then we could say there has been a statistically significant change between the polls.

Question 3: Consider 2004 presidential polls at <http://www.cnn.com/ELECTION/2004/special/polls/index.html>

How did all of the polling over the months compare to the actual result? Did their seem to be statistically significant changes in opinion over the time displayed?

Example 4

Suppose we conduct an experiment with one of those Smart Pills we've heard about in the past.

Suppose 25 people in a treatment group are given a Smart Pill before an exam, and another 25 people in a control group are given a placebo. They are all then subjected to a test of reasoning skills.

If 68% of the people in the treatment group pass the test and only 64% of the people in the control group pass the test, is this statistically significant?

A difference of only 4% is probably not significant enough to make us believe the Smart Pill has any true effect. However, if we were performed a similar experiment with 2,500 people in the treatment group and 2,500 people in the control group, we would probably see a 4% difference as being statistically significant. How do we know exactly?

Measuring Statistical Significance

Statisticians have to try to draw the line between results that are statistically significant and results that are not.

If a result or event is at the **0.05 significance level** it means that the event has a probability of occurring of only 5% (1 out of 20 times) or less.

For example, the probability of flipping a coin five times and getting heads five times in a row is only $(\frac{1}{2})^5 = 0.03125$. If a researcher flips a coin five times and this happens, this outcome is statistically significant at the 0.05 level. Such a result would lead the research to question the validity of the coin. (Keep in mind that this result is not impossible, just unlikely.)

If a result is at the **0.01 significance level** it means that the event has a probability of occurring of only 1% (1 out of 100 times) or less.

For example, the probability of picking four hearts in a row from four different decks of cards is only $(\frac{13}{52})^4 = 0.0039$. If a researcher picks four cards from four separate decks and this happens, this is statistically significant at the 0.01 level. Such a result would lead the research to question the validity of the decks of cards.

Question 4: You roll a six-sided die a certain number of times. When does repeatedly getting the number 6 become statistically significant at the 0.05 level? At the 0.01 level? (Hint: The probability of rolling a 6 is $\frac{1}{6}$).

Margin of Error and Confidence Intervals

Suppose you are told by the school that 40% of the QCC student population plans to vote in the November election, with a margin of error of 4%. We know from earlier examples that this really means that the actual population percentage is between $40\% - 4\% = 36\%$ and $40\% + 4\% = 44\%$. But how were these numbers obtained?

We suppose that a number of students were polled and 40% of those student indicated they would vote in the November election. But it may have been 4 students out of a sample of 10; or 40 students out of a sample of 100; or 400 students out of a sample of 1000. Which was it?

The larger the sample, the more confident we can be about the results. I would be more satisfied with the results of a sample of 200 students than only twenty students. This is where the margin of error comes in.

The margin of error measures the accuracy of a particular sample size, at certain level of significance.

If you take a sample of size n from a large population and measure some characteristic (such as who will vote next week or how many people own a car) as a percentage, then the **margin of error (E) at 95% confidence** is:

$$E \approx \frac{1}{\sqrt{n}}$$

The **95% confidence interval** is then:

Sample Percentage – E < Actual Population Percentage < Sample Percentage + E

Example 5

If you take a sample of 100 QCC students and find that 35% of them take the bus to school, the margin of error (E) at 95% confidence is:

$$E \approx \frac{1}{\sqrt{100}} = \frac{1}{10} = 0.1 = 10\%$$

This means you can be 95% certain that the actual percentage is between $35\% - 10\% = 25\%$ and $35\% + 10\% = 45\%$. We call 35% to 45% a **95% confidence interval**, which conveys a 95% degree of certainty that the actual population percentage is between the values given.

If you want a smaller margin of error, you use a smaller sample.

Example 5

If you take a sample of 900 QCC students and find that 20% of them plan to transfer to York College, the margin of error (E) at 95% confidence is:

$$E \approx \frac{1}{\sqrt{900}} = \frac{1}{30} = 0.033 = 3.3\%$$

This means you can be 95% certain that the actual percentage is between $20\% - 3.3\% = 16.7\%$ and $20\% + 3.3\% = 23.3\%$. We call 16.7% to 23.3% a 95% confidence interval, which conveys a 95% degree of certainty that the actual population percentage is between the values given.

Question 5: Suppose a survey of 1000 high school students reveals that 80% of them plan to attend a college when they graduate. Calculate the margin of error, and construct a 95% confidence interval.

Why This Works

A mathematical rule called **The Central Limit Theorem** states that in terms of population parameters that are percentages, sample percentage results are themselves normally distributed around the population parameter percentage.

As such, there is a mathematical expectation that any sample result will be relatively close to the population percentage. How close? Depending on the size of the sample, we can apply the 68-95-99.7 Rule and say that 95% of the time a sample result is two standard deviations away from the population percentage. Two standard deviations from the population percentage turn out mathematically to be:

$$\frac{1}{\sqrt{n}}$$

which is the margin of error.

You may encounter the Central Limit Theorem again. It is used to justify the statistical process of using sample results to estimate population parameters, which is central to all statistics.

Hypothesis Testing

We just saw how a sample percentage can be used to give a confidence interval for the actual population percentage by taking the margin of error percentage (E):

$$E \approx \frac{1}{\sqrt{n}}$$

(where n is the sample size) and determining a 95% confidence interval:

$$\text{Sample Percentage} - E < \text{Actual Population Percentage} < \text{Sample Percentage} + E$$

We can then be 95% certain the actual population percentage is between those two percentages.

But what if we already know the population percentage parameter, or think that we know it? What is the use of sampling then?

If we believe we know the claimed population percentage, we can test the validity of it using hypothesis testing.

Hypothesis Testing

We use the statistical process of **hypothesis testing** to determine the validity of a claim about a population parameter.

Example 1

Suppose Medico Pharmaceuticals claims that its drug Growefest will cause hair growth in 60% of bald men. How should we test the validity of this claim?

The idea, it would seem, would be to conduct your own test with your own sample and see what percentage of bald men experienced hair growth. That result would then be compared to the 60% claim. But we want to put it in the language of hypothesis testing.

The Process

First we construct a **null hypothesis**, which is what is being proposed as being true.

The null hypothesis here is that Growefest causes hair growth in 60% of all bald men.

From the null hypothesis, we construct an **alternative hypothesis**, which is what we're willing to accept if we decide to reject the null hypothesis.

The alternative hypothesis here is that Growefest causes hair growth in less than 60% of all bald men. Notice that we don't care to consider if the percentage is actually more than 60% (that's up to Medico to look at—they claimed 60% and it should be at least that). We focus only on what we care about as an alternative.

After the null hypothesis and alternative hypothesis have been created, we move on to the rest of the hypothesis test.

We decide to test Growefest on a sample of 500 men. We remember that for a sample of size 500, the margin of error is:

$$E \approx \frac{1}{\sqrt{500}} = \frac{1}{22.36} = 0.045 = 4.5\%$$

This means that we are 95% confident the population value is within 4.5% of the sample result:

$$\text{Sample Percentage} - E < \text{Actual Population Percentage} < \text{Sample Percentage} + E$$

This becomes:

$$\text{Sample Percentage} - 4.5\% < \text{Actual Population Percentage} < \text{Sample Percentage} + 4.5\%$$

Following the same idea, we should be 95% confident the sample result should be within $E=4.5\%$ of the population percentage:

$$\text{Population Percentage} - E < \text{Sample Percentage} < \text{Population Percentage} + E$$

or:

$$\text{Population Percentage} - 4.5\% < \text{Sample Percentage} < \text{Population Percentage} + 4.5\%$$

The result of these calculations (for our hypothesis test) is that we want to know if our sample percentage is below $60\% - 4.5\% = 55.5\%$ or not.

If it is below 55.5%, we “**reject the null hypothesis**” that 60% of all bald men will grow hair using Growefest and accept the alternative hypothesis that the percentage is some amount less (You’re not saying you know how much less.) than 60%

If the result is 55.5% or more, we “**fail to reject the null hypothesis**” that 60% of all bald men will grow hair using Growefest. Notice the use of language here. By failing to reject the null hypothesis, you’re not saying that you actually believe it. You’re saying you weren’t able to disprove it.

Therefore, if your test of 500 men using Growefest reveals that only 263 grew hair, the percentage is only $263/500 = 52.6\%$, which is less than 55.5%. This would cause us to reject the null hypothesis and accept the alternative hypothesis.

However, if your test of 500 men using Growefest reveals that 289 grew hair, the percentage is $289/500 = 57.8\%$, which is more than 55.5%. This would cause us to fail to reject the null hypothesis. We have not obtained enough evidence to reject the claim of Medico Pharmaceuticals.

We should note here that making use of the 95% confidence interval means that our hypothesis test is being conducted at the 0.05 level of significance (because $100\% - 95\% = 5\% = 0.05$). To conduct the test at a different level of significance, we need to have more knowledge about calculating different margins of error. This is beyond the scope of this course, but it is something you may see in the future.

Here is an overall summary of hypothesis testing:

- 1) Construct a null hypothesis, which is what is being proposed as being true. The null hypothesis is what the researcher is trying to disprove. Usually, the null hypothesis is that a population parameter is equal to some value.
 - 2) From the null hypothesis, construct an alternative hypothesis, which is what we're willing to accept if we decide to reject the null hypothesis. The alternative hypothesis is what the researcher believes to be the truth. The alternative hypothesis is often that the population parameter is more or less than the value given (sometimes it's that the population parameter is simply not equal to the value given).
 - 3) A sample is gathered to test the validity of the null hypothesis.
 - 4) Using what is known about the margin of error of the given sample and the corresponding confidence interval, we compare our sample results to the confidence interval.
 - 5) a) If our sample results are outside of the confidence interval and in favor of the alternative hypothesis, we reject the null hypothesis and accept the alternative hypothesis.
- c) If our sample results are within the confidence interval or not in favor of the alternative hypothesis, we "fail to reject" the null hypothesis. Further study may be warranted.

Question 6: Construct your own hypothesis test regarding a percentage claim about CUNY students. Include the null hypothesis, alternative hypothesis, sample results, margin of error, 95% confidence interval, and final conclusion, using the format given above. (Use your imagination.)

Hypothesis testing can also be performed on claims of population values, as opposed to percentages. An example is given below, with the more vigorous mathematics excluded.

Example 2

Suppose the vice president of the college claims that the average IQ of CUNY students is 103.8. Thinking that CUNY students are even better, we want to test the validity of her claim.

First we construct the null hypothesis, which would be that the average IQ of CUNY students is 103.8.

The alternative hypothesis here is that the average IQ of CUNY students is actually greater than 103.8.

After the null hypothesis and alternative hypothesis have been created, we move on to the rest of the hypothesis test.

We would then gather a sample of CUNY students of some size and calculate the mean IQ and standard deviation. The values of sample size, sample mean and sample standard deviation are all used to decide whether we should reject the null hypothesis and accept the alternative hypothesis, or fail to reject the null hypothesis.

Obviously, if the null hypothesis is true, the sample mean should be close to the claimed population mean (103.8). And the larger the sample is, the closer we should expect the sample mean to the claimed population mean.

If the sample mean higher than 103.8 and not “close enough”, we reject the null hypothesis and accept the null hypothesis. If it is close enough to 103.8, or lower than 103.8, we fail to reject the null hypothesis.

(The mathematics of hypothesis testing of values is somewhat more complicated than it is for percentages. I'm willing to share it with anyone who is truly interested. It is something you will see more of if you go further with statistics.)

Next week, you will be writing about your data set and the work you did on Lab #6 and Lab #7. Details to follow....

This concludes our work with statistics. The next topic will be problem solving, starting next week.