AP Statistics Summer Assignment

For students who have *completed* Algebra 2 and are **entering** AP Statistics

This review packet is designed to help you prepare for AP Statistics. Begin working on the packet two weeks before the school year starts. It consists of four parts, with part 1 requiring internet access. Dedicate approximately one hour per day to complete the packet. You can write directly on this sheet or attach additional pages. Feel free to use your graphing calculator for the assignment.

You should expect to submit this completed packet on the first day of school and plan to be assessed on these skills at the beginning of the school year. An answer key is provided; however, ensure your work is shown to receive full credit. If you would like additional resources to support your practice, we recommend Khan Academy as a great first step.

Part 1: Reading and Writing.
Read the two articles at the end of this packet ("Research Basics: Interpreting Change" and "Overstating Aspirin's Role in Breast Cancer Prevention") from the Washington Post and then answer the following questions in complete sentences.

1. What was the story that the newspapers wrote after the research was published by the Journal of the American Medical Association?

2. What other information needed to be added to the story so that people could make decisions for themselves about the use of aspirin to prevent breast cancer?

3. How was the data collected to perform this study?

4. What type of study was performed?

5. Can this type of study be used to prove the aspirin prevents breast cancer?

6. What type of study must be done in order to 'prove' something?

7. What is the difference between 'cause' and 'association'?

8. You may have heard the statement "you can prove anything with statistics". Using what you have learned reading this article, explain what you think is meant by this statement.

Go on the internet to www.gapminder.org, select "Gapminder World" panel, and the scatterplot should load. You are looking at worldwide data of Life Expectancy vs. Per Capita Income. Point your cursor at the x-axis or y-axis labels to get more information about these variables. Every colored circle on the graph represents a country. Point the cursor at various circles and the name of the country will appear. The size of each circle is proportion to that country's population—look in the lower right corner to see each country's population as you point the cursor at it. If you would like, slide the year indicator back to the first year that data was recorded (1950 for this combination of variables), and then click on "Play" to watch the change in the scatterplot, year by year, from that year to the present. Even more fun is to select one or more countries (this causes all the other countries to dim into the background), and watch the track made by the selected countries over time.

9. What is the relationship between Per Capita Income and Life Expectancy in the world?

10. Which countries are the farthest from the pattern shown by the rest of the world?

11. Which country has the highest life expectancy now? _____ Which has the highest per capita income? _____
    Which has the lowest income? _____ The lowest life expectancy? _____

12. Which group of countries (by color) has gained most since 1950 relative to the rest of the world, in both income and life expectancy?

13. Watch the "track" of Rwanda from 1950 – 2010. What events in Rwanda might explain the unusual changes that happened?
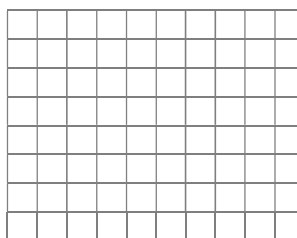
Part 2: Algebra Review
14. Find the slope and y-intercept of the line. Sketch the line.

   a. $y = \frac{2}{3}(2x - 4)$
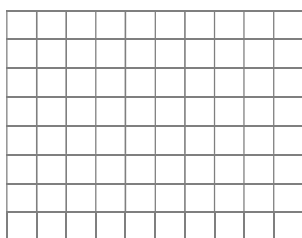
   b. $3x + 2y = 14$

15. Write the equation of the line containing the points $(8, -5)$ and $(3, 4)$

16. Given $f(x) = 4x - 2$ and $g(x) = 3x^2$, find the following:

   a. $f(-3) =$

   b. $g(-3) =$

   c. $f\big(g(-1)\big) =$

   d. $g\left(f\left(\frac{3}{2}\right)\right) =$

17. Solve each inequality for x.

   a. $0.05 < 2.5\sqrt{\frac{0.6}{x}}$

   b. $0.01 \geq 1.78\sqrt{\frac{.5(.5)}{x}}$

18. Expand each expression.

   a. $\log\left(\frac{2x}{3}\right)^2$

   b. $\ln\left(\frac{x\sqrt[3]{y}}{z^2}\right)$

19. Solve each equation for x.

   a. $\log_3 81 = x$

   b. $\log_3 x = 5$

   c. $\log_x 256 = 8$

   d. $\log_5(x - 4) = 0$

Part 3: Basic Statistics Review

There are two categories of numbers that are used to describe a set of data: measures of center and measures of spread.

Measures of CENTER:
- The **mean** is the average number. It is the sum of all the data values divided by the number (n) of values.
  Ex. {4, 36, 10, 22, 9}
  $$Mean = \bar{x} = \frac{\sum x_i}{n} = \frac{4 + 36 + 10 + 22 + 9}{5} = 16.2$$

- The **median** is the value that separates the bottom 50% of data from the top 50% of data. It is the middle element of an ordered set of data that is odd in number. It is the average of the two middle elements of an ordered set of data that is even in number.
  Ex. {4, 36,10,22, 9} = {4, 9, **10**, 22, 36}   median = 10

  Ex. {4, 36, 10, 22, 9, 43} = {4, 9, **10, 22**, 36, 43}   median $= \frac{10+22}{2} = 16$

- The **mode** is the value that occurs most often in a set of data. If the data occurs with the same frequency, then there is no mode. If two (or more) values occur the most then they are both the mode.

Measures of SPREAD:
- The **range** is a measure of the spread of the entire data. It is calculated by subtracting the minimum value from the maximum value.
  Ex. {4, 36, 10, 22, 9} = {4, 9, 10, 22, 36} range $= 36 - 4 = 32$

- The **interquartile range** (IQR) is a measure of the spread of the middle 50% of the data. It is calculated by subtracting the **25th percentile** (Q1) from the **75th percentile** (Q3). Q1 is the median of the lower half of the data. It separates the bottom 25% of values from the top 75% of values. Q3 is the median of the upper half of the data. It separates the top 75% of values from the bottom 25% of values. In neither of these cases is the median considered in the top half or the bottom half of the data.

  Ex. {28, 1, 9, 4, 36, 18, 10, 22, 9} = {1, 4, 9, 9, 10, 18, 22, 28, 36}
      median = 10, $Q1 = \frac{4+9}{2} = 6.5$, $Q3 = \frac{22+36}{2} = 29$ $IQR = 29 - 6.5 = \mathbf{22.5}$

  Ex. {4, 36, 10, 22, 9, 43} = {4, 9, 10, 22, 36, 43}
      median = 16, $Q1 = 9$, $Q3 = 36$ IQR $= 36 - 9 = \mathbf{27}$

Qualitative data can be stored on the TI-83/84 in lists. Use STAT 1: Edit to go to the list editor and enter the data into columns (Figure 1.2). Alternatively, from the home screen braces can be used to define a data set, which then can be stored in one of the list named L_1 through L_6 (Figure 1.1).
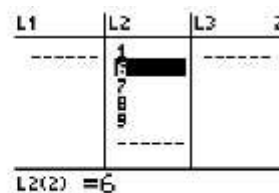

Figure 1.1

Figure 1.2

In either case, new lists can be created from existing lists, such as $L_2 + 5$ (Figure 1.3). Make sure when you enter the new list that you are on the L_3 icon and not within the list of numbers.
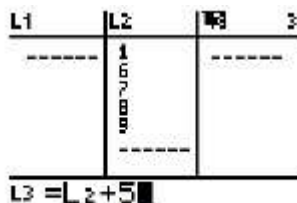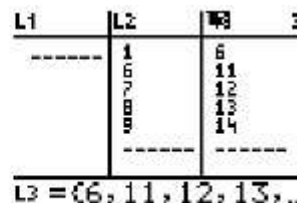

Figure 1.3

Figure 1.4

To see statistical results including the quartiles and standard deviation, use STAT CALC 1: 1-Var Stats (Figure 3.1), followed by the list name (Figure 3.2).

```
EDIT CALC TESTS        1-Var Stats L1█      1-Var Stats
██1-Var Stats                               x̄=16.2
2:2-Var Stats                               Σx=81
3:Med-Med                                   Σx²=1977
4:LinReg(ax+b)                              Sx=12.8918579
5:QuadReg                                   σx=11.53082824
6:CubicReg                                  ↓n=5
7↓QuartReg                                  █
```

Figure 3.1   Figure 3.2   Figure 3.3

If you push the down arrow key, you can see the rest of the statistics (Figure 3.4).

```
1-Var Stats
↑n=5
  minX=4
  Q1=6.5
  Med=10
  Q3=29
  maxX=36
█
```

Figure 3.4

20. Given the following data on parents ages when theirs sons were born, answer the following questions

    Dad: 41 27 23 31 30 33 26 32 43 25 34 27 25 34 27 26 28 32 32 35 27 33 34 34 34 35
    Mom: 39 26 23 30 28 33 23 32 38 23 35 24 24 33 24 23 24 32 23 30 24 29 34 35 26 31

    Enter these as two lists in your calculator and use the 1-Var Stat option to calculate the following:

    a. Find the mean and median for the Dad data: Mean ($\bar{x}$)_____ Median_____ Which is larger? _____

    b. Find the mean and median for the Mom data: Mean ($\bar{x}$)_____ Median_____ Which is larger? _____

    c. Now compare the two means you calculated. Which is larger? _____

        Is this what you expected? _____. Explain why or why not. _____

        _____

    d. Calculate the standard deviations (S_x) for both sets of data: Dad_____ Mom_____

        Why might these values be different? Explain. _____

        _____

    e. Find Q1, Q3, IQR, and range for the Dad data. Q1_____ Q3_____ IQR_____ Range _____

    f. Find Q1, Q3, IQR, and range for the Mom data. Q1_____ Q3_____ IQR_____ Range _____

21. A company has two machines that fills cans of soft drinks. The desired number of ounces per can is 12. Samples from each machine show the following number of ounces per can:
        Machine A: 11.1, 12.0, 11.4, 12.1, 11.7, 11.5, 12.2, 11.4, 11.3, 11.9
        Machine B: 10.9, 12.4, 12.7, 11.8, 12.3, 11.9, 12.0, 12.5, 12.7, 11.6

   Find the mean and standard deviation for both machines.
    $\bar{x}_A$ =_____ S_A =_____ $\bar{x}_B$ =_____ S_B =_____

22. Using your answer to #21, explain which machine is "better" at filling soft drink cans.

    _____
    _____
    _____

23. The following data is the number of jump shots made in 25 attempts from various distances.

Distance from basket (ft)	25	15	10	20	15	5	10	5	25
Jump shots made	6	10	15	7	9	18	12	20	5

    a. Make a scatterplot of the distance from basket vs. jump shots made data. Be sure to label the axes with numbers and words.
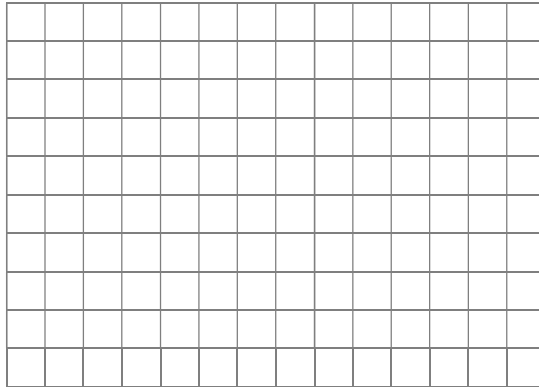
    b. Does your graph show a positive correlation, negative correlation, or no correlation?

    c. According to your model, estimate the distance away from the basket that will result in zero jump shots being made.

24. The scores on an Algebra 1 final were 90, 86, 78, 82, 96, 100, and 45.

    a. Create a boxplot (box-and-whisker plot) for this data. Be sure to include all necessary labels.

    b. Create a stemplot (stem-and-leaf plot) for this data. Be sure to include a key.

25. Twelve students were surveyed concerning their summer plans. Four students will be taking a summer class. Three students will be working. Three students will be taking a family vacation. Two students will be watching TV all summer.

    a. Create a bar graph for this data. Be sure to include all necessary labels.

    b. Create a pie chart for this data. Be sure to include all necessary labels.

Part 4: Probability Review

26. You have an equally likely change of rolling any value on each of two dice. Find the probability of the given event.
    a. rolling a sum of either 7 or 9

    b. rolling a sum greater than 5

    c. rolling a 6 on exactly one die

    d. rolling doubles

27. There are three language classes that the 200 Upper School students at The Bear Creek School may take: Spanish, French, and Latin. 121 students take Spanish, 40 students take French, and 28 students take Latin. 3 students take both Latin and Spanish, 4 students take both French and Latin, 2 students take both Spanish and French, and 1 student takes all three classes. Draw a Venn diagram to represent this information.

    a. What is the probability that a student selected at random will take Spanish only?

    b. What is the probability that a student selected at random takes none of these languages?

    c. What is the probability that a student selected at random takes Latin but not French?

28. Thirty-five students in an AP Statistics class took a test: 9 received A's, 18 received B's, and 8 received C's. Find the probability of each event.
    a. If a student from the class is chosen at random, what is the probability that the student did not receive a C?

    b. If the teacher randomly chooses 3 test papers, what is the probability that the teacher chose tests with grades A, B, and C in that order?

Selected Answers

9. Logarithmic; as income increases, life expectancy increases

14a. $m = \frac{4}{3}; b = -\frac{8}{3}$

14b. $m = -\frac{3}{2}; b = 7$

15. $y = -\frac{9}{5}x + \frac{47}{5}$

16a. -14 b. 27   c. 10   d. 48

17a. $x < 1500$ b. $x \geq 7921$

18a. $2\log 2 + 2\log x - 2\log 3$

18b. $\ln x + \frac{1}{3}\ln y - 2\ln z$

19a. $x = 4$

19b. $x = 243$

19c. $x = 2$

19d. $x = 5$

20a. mean = 31.08, median = 52, median is larger

20b. mean = 2 8.69, median = 28.5, mean is larger

20c. Dad

20d. Dad: 4.87, Mom: 5.12; Answers will vary.

Sample answer: The moms' ages are more spread out from the mean than the dads' ages.

20e. Q1 = 27, Q3 = 34, IQR = 7, Range = 20

20f. Q1 = 24, Q3 = 33, IQR = 9, Range = 16

21. $\bar{x}_A = 11.66, s_A = 0.375; \bar{x}_B = 12.08, s_B = 0.561$

22. Answers will vary. Sample answer: Machine B is better at filling soft drink cans. Although the standard deviation is higher, meaning the data varies more from the mean than machine A's data, the mean is closer to 12 ounces than the mean of machine A. Machine A's mean is almost an entire standard deviation from 12 ounces, making it more a more biased statistic.
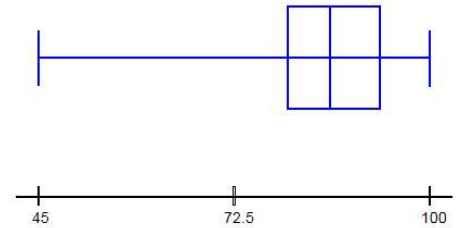
23a.
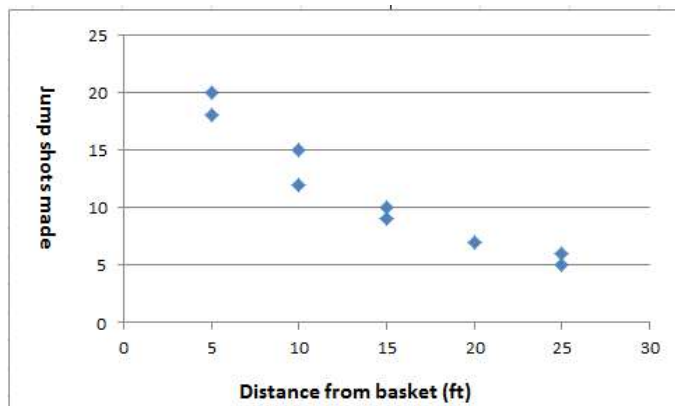
23b. negative correlation   c. 32 ft

24a.

Minimum: 45
Lower quartile: 80
Median: 86
Upper quartile: 93
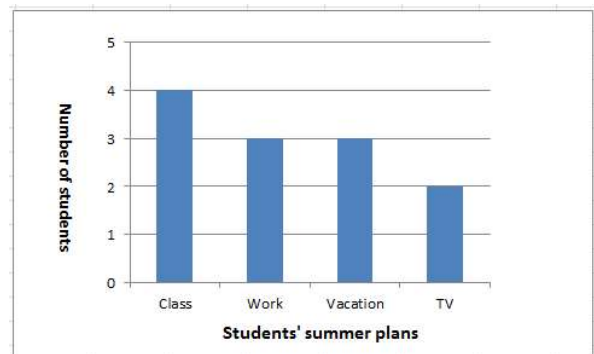Maximum: 100

24b.

Scores on Algebra 1 Final

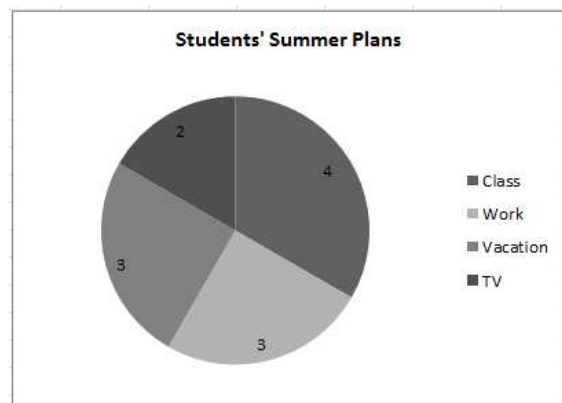|	
4	5
5	
6	
7	8
8	2 6
9	0 6
10	0

Key: 1 | 2 = 12

25a.

25b.

26a. 5/18   b. 13/18   c. 5/18   d. 1/6

27a. 0.585   b. 0.095   c. 0.12

28a. 27/35   b. 0.033 or 216/6545

The Washington Post

Research Basics: Interpreting Change

Tuesday, May 10, 2005

How Big Is the Difference?

Many medical studies end up concluding that two groups have different health outcomes -- death rates, heart attack rates, cholesterol levels and so forth. This difference is typically expressed as a *relative change* , as in the statement: "The treatment group had 50 percent fewer cases of eye cancer than the control group." The problem with this comparison is that it provides no information about how common eye cancer is in either group.

Thinking about relative changes in risk is like deciding when to use a coupon at a store. Imagine you have a coupon that says "50 percent off any one purchase." You go to the store to buy a pack of gum for 50 cents and a large Thanksgiving turkey for $35. Will you use the coupon for the gum or the turkey? Most people would use it for the turkey.

Why? Because paring half the price off $35 reaps a bigger savings --$17.50 --than cutting half off 50 cents -- or $0.25.

The analogy in health is that "50 percent fewer cases" is a very different number when applied to eye cancer -- a rare problem accounting for about 2,000 new cases in the U.S. each year -- than when applied to heart attacks -- a common problem accounting for about 800,000 new cases annually.

To really understand how big a difference is, you need to find out the *starting* and *ending points* -- sometimes called " *absolute risks* ." In the coupon example, the start and end points are the regular and the sales price. In a study about medical treatment, the start and end points are the chances of something happening in the untreated and treated groups.

Presenting the starting and ending point requires a few more words than presenting relative changes. For example, "In a year, two of 100,000 untreated people developed eye cancer; in contrast, one of 100,000 treated people developed eye cancer." For the price of a few more words you gain perspective: The chance of developing eye cancer is small.

Cause or Association?

Many important insights into human health come from *observational studies* -- studies in which the researcher simply records what happens to people in different situations, without intervening. Such studies first linked cigarette smoking to lung cancer and high cholesterol to heart disease. But not all observed associations represent cause and effect. And problems can occur when this key point is overlooked.

An example may help make the distinction clear. A man thought his rooster made the sun rise. Why? Because each morning when he woke up while it was still dark, he would hear his rooster crow as the sun rose. He confused association with causation until the day his rooster died, when the sun rose without any help.

A more serious example involves the long-held belief that most women should take estrogen after menopause. That idea, only recently discredited, also came from observational studies. The observation -- shown in more than 40 studies involving hundreds of thousands women -- was that women who took estrogen supplements also had less heart disease. But it turned out that estrogen was not the reason why this was the case. Instead, women taking estrogen tended to be healthier and wealthier. Their health and wealth -- not their estrogen supplements -- were responsible for the lower risk of heart disease.

The only way to reliably distinguish a cause from an association is to conduct a true experiment -- a *randomized trial* . In this type of study, patients are assigned randomly --that is, by chance--to receive a therapy or not receive it. This study design is the best way to construct two groups that are similar in every way except one -- whether they get the therapy being studied. That means any differences observed afterward must be caused by the therapy. In the case of estrogen and heart disease, such a study showed that the long-held beliefs were wrong.

Unfortunately, it is not always possible to do a randomized trial. For example, it is extremely unlikely that we could get people to agree to be randomly assigned to either eating only fast food or only organic food every day for a year (and that they would actually adhere to the diet if they did agree to be randomized). In such cases, scientists have to rely on observational studies. But when new tests or treatments are proposed, randomized trials ought to be conducted prior to their widespread use. Doctors prescribed estrogen to millions of women for many years until the randomized trial showed that intuition and dozens of observational studies were wrong.

-- Lisa M. Schwartz, Steven Woloshin and H. Gilbert Welch

The Washington Post

A May 10 Health section story about a study exploring aspirin use and breast cancer prevention incorrectly labeled hormone receptor positive cancers the most dangerous kind. That description applies to hormone receptor negative breast cancers.

Overstating Aspirin's Role in Breast Cancer Prevention
How Medical Research Was Misinterpreted to Suggest Scientists Know More Than They Do

By Lisa M. Schwartz, Steven Woloshin and H. Gilbert Welch
Special to The Washington Post
Tuesday, May 10, 2005

Medical research often becomes news. But sometimes the news is made to appear more definitive and dramatic than the research warrants. This series dissects health news to highlight some common study interpretation problems we see as physician researchers and show how the research community, medical journals and the media can do better.

Preventing breast cancer is arguably one of the most important priorities for women's health. So when the Journal of the American Medical Association published research a year ago suggesting that aspirin might lower breast cancer risk, it was understandably big news. The story received extensive coverage in top U.S. newspapers, including The Washington Post, the Wall Street Journal, the New York Times and USA Today, and the major television networks. The headlines were compelling: "Aspirin May Avert Breast Cancer" (The Post), "Aspirin Is Seen as Preventing Breast Tumors" (the Times).

In each story, the media highlighted the change in risk associated with aspirin -- noting prominently something to the effect that aspirin users had a "20 percent lower risk" compared with nonusers. The implied message in many of the stories was that women should consider taking aspirin to avoid breast cancer.

But the media message probably misled readers about both the size and certainty of the benefit of aspirin in preventing breast cancer. That's because the reporting left key questions unanswered:

· Just how big is the potential benefit of aspirin?

· Is it big enough to outweigh the known harms?

· Does aspirin really prevent breast cancer, or is there some other difference between women who take aspirin regularly and those who don't that could account for the difference in cancer rates?

This article offers a look at how the message got distorted, what the findings really signify--and some broader lessons about interpreting medical research.

How Big a Benefit?

Just how big is the potential benefit of aspirin?

The 20 percent reduction in risk certainly sounds impressive. But to really understand what this statistic means, you need to ask, "20 percent lower than what?" In other words, you need to know the chance of breast cancer for people who do not use aspirin. Unfortunately, this information did not appear in any of the media reports. While it might be tempting to fault journalists for sloppy, incomplete reporting, it is hard to blame them when the information was missing from the journal article itself.

In the study, Columbia University researchers asked approximately 3,000 women with and without breast cancer about their use of aspirin in the past. The typical woman in this study was between the ages of 55 and 64. According to the National Cancer Institute, about 20 out of 1,000 women in this age group will develop breast cancer in the next five years. Therefore, the "20 percent lower

chance" would translate into a change in risk from 20 per 1,000 women to 16 per 1,000 -- or four fewer breast cancers per 1,000 women over five years.

For people who prefer to look at percentages, this translates as meaning that 2 percent develop breast cancer without aspirin, while 1.6 percent develop it with aspirin, for an absolute risk reduction of 0.4 percent over five years.

Another way to present these results would be to say that a woman's chance of being free from breast cancer over the next five years was 98.4 percent if she used aspirin and 98 percent if she did not. Seeing the actual risks leaves a very different impression than a statement like "aspirin lowers breast cancer risk by 20 percent." (See "Research Basics: How Big Is the Difference?")

Against What Size Harms?

Is the potential benefit of aspirin big enough to outweigh its known harms?

Unfortunately, aspirin, like most drugs, can have side effects. These, according to the U.S. Preventive Services Task Force, include a small risk of serious (and possibly fatal) bleeding in the stomach or intestine, or strokes from bleeding in the brain -- harms briefly noted but not quantified in the original study or in most media reports. To decide whether aspirin is worth taking, women need to know how the potential size of aspirin's benefit in reducing breast cancer compares with the drug's potential harms.

Sound medical practice dictates doing the same kind of calculation -- of potential benefits against potential harms -- anytime you consider taking a drug.

We provide the relevant information in the "Aspirin Study Facts," below. The first column shows the health outcome being considered (e.g., getting breast cancer, having a major bleeding event). The second column shows the chance of the outcome over five years for women *not* taking aspirin. The third column shows the corresponding chance for women taking aspirin. And the fourth column shows the difference -- the possible effect of aspirin.

As the table shows, the size of the known risk for stomach bleeding to a woman taking aspirin daily nearly matches the size of the still-hypothetical benefit in terms of breast cancer protection. That kind of comparison might lead some women to conclude that the tradeoff doesn't warrant the risk.

While it may take you some time to become familiar with this table, we think this sort of presentation would be helpful in many situations; for example, whenever people are deciding about taking a new medication or undergoing elective surgery.

Is It Really Aspirin?

Does aspirin really prevent breast cancer, or is there some other difference between women in the study that could account for the difference in cancer rates?

Can we be sure that aspirin was responsible for the "20 percent fewer" breast cancers that the Columbia researchers found among aspirin users compared with nonusers?

To understand why not, it is necessary to know some of the details about how the study was conducted.

The researchers collected information from all of the women in New York's Nassau and Suffolk counties on Long Island, who were diagnosed with breast cancer in 1996 and 1997. For comparison, they matched these women with others who did not have breast cancer, but who were about the same age and from the same counties. The researchers asked all the women about their use of aspirin.

They found that aspirin use was more common among the women without breast cancer. While the researchers were careful to report that the use of aspirin was "associated" with reduced risk of breast cancer, the media used stronger language, suggesting aspirin played a role in preventing breast tumors.

Unfortunately, this kind of study -- an observational study -- cannot prove that it was the aspirin that lowered breast cancer risk. Strictly speaking, the researchers demonstrated only that there is an association between aspirin and breast cancer.

Consider how an association between aspirin and breast cancer could exist even if aspirin has no effect on breast cancer.

It could be that women who use aspirin regularly are already at a lower risk of breast cancer. Imagine, for example, there was a gene that protected against breast cancer but also made people more susceptible to pain. Women who carried this gene would be more apt to use aspirin for pain relief. The lower breast cancer risk in aspirin users might simply reflect the fact that they had this gene. In other words, aspirin might have nothing to do with the findings. To really know if aspirin lowers breast cancer risk would require a different kind of study -- a randomized trial. (See "Research Basics: Cause or Association?")

Nonetheless, observational studies are important (and often crucial) in building the case for doing a randomized trial. In this instance, the researchers had a theory for how aspirin might prevent breast cancers. They predicted that it would only be true for certain kinds of cancers (so-called hormone receptor positive cancers, the most dangerous kind, which account for about 60 percent of all breast cancers). And that is just what they observed: The association between aspirin and breast cancer was not seen in hormone receptor negative cancers. That the researchers' prediction was correct supports (but does not prove) the idea that aspirin reduces risk. The next logical step would be a randomized trial.

The difference between "cause" and "association" may seem subtle, but it is actually profound. Even so, people -- like the headline writers in this case -- often go beyond the evidence at hand and assume that an association is causal. Readers should know that many associations do not reflect cause and effect.

The Bottom Line

In a large observational study, researchers found slightly fewer breast cancers among women who took aspirin regularly compared with women who did not. Because aspirin's benefit in reducing breast cancer (assuming it can be proven) was small, it may not outweigh the drug's known harms. While it is possible that aspirin itself reduces the risk of breast cancer, we cannot be sure from this study. It would take a randomized trial to be certain. Fortunately, one has just been completed by researchers at Harvard Medical School, and the results are expected in the very near future. Until then, it is too soon to recommend taking aspirin to prevent breast cancer. ·

Lisa Schwartz, Steven Woloshin and Gilbert Welch are physician researchers in the VA Outcomes Group in White River Junction, Vt., and faculty members at the Dartmouth Medical School. They conduct regular seminars on how to interpret medical studies. (See http://www.vaoutcomes.org.) The views expressed do not necessarily represent the views of the Department of Veterans Affairs or the United States Government.