In Steven Zaillian's film *A Civil Action*, the attorney for the companies accused of dumping toxic chemicals tells the following to one of the town residents:

> *If I took a hundred pennies and threw them up in the air, about half of them would land heads, and the other half tails, right? Now if I looked around closely, I'd probably find some heads grouped together in a cluster. What does that mean? Does that mean anything?*

In this assignment, we will explore these questions.

Let p_1 represent the underlying rate of birth defects during the time the contaminated wells were being used, and let p_2 represent the underlying rate of birth defects during the time the contaminated wells were not being used. We want to know whether these underlying (but unknown) rates p_1 and p_2 are equal.

(1) State appropriate null and alternative hypotheses. Since we believe *a priori* that water pollution should increase (rather than decrease) the rate of birth defects, use a one-sided alternative hypothesis.

The true values of p_1 and p_2 are unknowable. Instead, we observe a sample of births (from the population of all births that "could have occurred" in Woburn during these time periods), and we observe the proportion of birth defects when the wells were or were not being used. Let $\hat{p}_1$ represent the observed proportion of birth defects when the contaminated wells were being used, and let $\hat{p}_2$ represent the observed proportion of birth defects when the contaminated wells were not being used.

According to one source, there were 16 birth defects out of 414 births when the contaminated wells were being used, and 3 birth defects out of 228 births when the contaminated wells were not being used.

(2) Calculate the values of $\hat{p}_1$ and $\hat{p}_2$ from these data. Also calculate $(\hat{p}_1 - \hat{p}_2)$, the difference between the two proportions.

(3) If the null hypothesis were true, what would be the sampling distribution of the quantity $(\hat{p}_1 - \hat{p}_2)$? Give the name of the sampling distribution, its mean, and its SD. Show your work.

(4) If the null hypothesis were true, the quantity $(\hat{p}_1 - \hat{p}_2)$ would lie within what range 95% of the time? Show your work.

(5) Calculate the *P*-value, the probability of observing a value of $(\hat{p}_1 - \hat{p}_2)$ as large as (or larger than) the value you observed, if the null hypothesis were true. (Since we used a one-sided alternative hypothesis, be sure to calculate a one-sided *P*-value.) Show your work.

(6) Based on your *P*-value, do you reject or not reject the null hypothesis? Does it appear that the rate of birth defects differed in times when the contaminated wells were or were not used?

(more on the back)

Download the following article, available in the Stat 1 Classes folder:

> Lagakos, Wessen, and Zelen (1986): An Analysis of Contaminated Well Water and Health Effects in Woburn, Massachusetts. *Journal of the American Statistical Association*, volume 81 issue 395, pp. 583–596.

Read sections 1, 2, 5, and 6, and skim sections 3 and 4. The *Journal of the American Statistical Association* is one of the most prestigious American journals for statistics research. This article is written for an audience of professional statisticians, so some parts may be a tough read, but most of this article is fairly readable.

(7) In two sentences or fewer, summarize the major findings of this article.

(8) What additional explanatory variable do the authors attempt to control for that is not accounted for in your analysis in **(1)–(6)** above?

(9) Describe briefly how the authors tried to ensure that the data gathered in their survey was accurate. (Section 5.2)

(10) Explain whether this is an observational study or a controlled experiment. What do the authors say about potential lurking variables ("confounding factors")? (Section 6)

(11) The results of this article were known by the families' attorney (the character played by John Travolta) and his partners fairly early on in the case, but they play only a limited role in the trial. Why is the type of statistical evidence in this article insufficient to serve as "proof" in a legal proceeding? How does the nature of statistical evidence differ from the type of evidence used by an attorney in a trial?