# Examining Potential Predictors for Completion of the Gardasil Vaccine Sequence Based on Data Gathered at Clinics of Johns Hopkins Medical Institutions

Christopher E. Barat
Courtney Wright
Stevenson University

Betty Chou
Johns Hopkins Medical Institutions

*Journal of Statistics Education* Volume 19, Number 1 (2011),
www.amstat.org/publications/jse/v19n1/barat.pdf

---

**Key Words**: Categorical data; Chi-square test of homogeneity; Confounding variables; Inference for proportions; Logistic regression; Odds ratio.

## Abstract

This paper presents categorical data that were gathered at two urban clinics and two suburban clinics of Johns Hopkins in an effort to identify characteristics of young female patients who successfully complete the three-injection sequence of the Gardasil quadrivalent human papillomavirus vaccine (HPV4). Available categorical correlates included patient age group, patient race, clinic location type, type of insurance provider, and clinic practice type. The data may be used to illustrate graphical display techniques and inference procedures for categorical data, as well as illuminate ways in which relationships between categorical variables may be hidden or behave differently than expected in the presence of confounding variables. The methods used to gather the data may also serve to illustrate the limitations of drawing conclusions from observational studies.

## 1. Introduction and Background

Most students in introductory statistics classes are introduced at some point to the notion of analyzing relationships between categorical variables by means of two-way tables and bar graphs for marginal and conditional distributions. The data set presented in this paper provides an excellent source of "raw material" for introductory lessons and assignments in categorical data analysis. The data set includes observations of multiple categorical (predictor) variables – all of which are intended to predict membership in the categories of another categorical (response) variable – and therefore can be analyzed in many different ways, including:

- analyzing the relationship between any one predictor and the response variable;
- analyzing the relationship between a set of predictors and the response variable (including the ways in which one predictor in the set may confound the relationship between another predictor in the set and the response variable);
- identifying the "best" possible predictors or sets of predictors for predicting membership in the categories of the response variable.

The methods of analysis that may be used with these data range from the simple (relative frequency distributions, tables, and bar graphs) to the relatively complex (logistic regression).

The data were gathered by researchers at Johns Hopkins Medical Institutions (JHMI) as part of an attempt to characterize young female patients who successfully complete the anti-HPV Gardasil vaccination sequence. HPV (human papillomavirus) is a virus that has been linked to the development of cervical cancer, other anogenital cancers, and genital warts. It is regarded as a sufficient cause of cervical cancer, rather than a necessary cause, due to other contributing factors; however, 100% of patients with cervical cancer have been found to have been infected with HPV (Markowitz, Dunne, Saraiya, Lawson, Chesson, and Unger, 2007). Gardasil, developed by Merck Laboratories, was licensed for use by the U.S. Food and Drug Administration in 2006. The FDA recommended Gardasil for use by women aged 9-26. Gardasil works by creating immunities to certain strains of HPV. The "typical" Gardasil "regimen" consists of a sequence of three 0.5-ml shots, the second being given at Month 2 (two months after the first) and the third being given at Month 6. The vaccine label, however, recommends completion of the three-shot sequence within 12 months (as reported in Chao, Velicer, Slezak, and Jacobsen 2009).

Dunne and Markowitz (2006, p. 628) argue that the vaccine "may have a significant impact on the prevalence and incidence of HPV infection, genital warts, cervical cancer precursor lesions, and HPV-associated cancers," citing the results of numerous studies conducted in the years immediately prior to the FDA's licensing of Gardasil. Another randomized, placebo-controlled, double-blind experiment conducted by Garland, Hernandez-Avila, Wheeler, Perez, Harper, Leodolter, Tang, Ferris, Steben, Bryan, Taddeo, Railkar, Esser, Sings, Nelson, Boslego, Sattler, Barr, and Koutsky (2007) and involving women aged 16-24 found that Gardasil significantly reduced the cases of anogenital disease associated with HPV. However, despite Gardasil's apparent benefits, the adoption of the vaccine has been a slow process. A 2007 study conducted by the Center for Disease Control found that only 25.1% of women aged 13-17 had received one or more of the Gardasil doses (DHHS/CDC 2008). Moreover, even if a woman begins the three-shot sequence, she may not complete it. Follow-through appears to be particularly lacking in poor, minority-heavy communities and among those lacking some form of health insurance. However, an early study found that those patients seen by pediatricians were more likely to follow through, even in urban areas (Boughton, 2008).

In promoting the use of Gardasil, it is important to identify subpopulations that may be at risk for failure to complete the three-shot sequence. A large-scale prospective study conducted with female members of the managed care group Kaiser Permanente in Southern California (Chao et al. 2009; henceforth, this study will simply be referred to as "the Kaiser Permanente study") examined the correlation between completion rate for the Gardasil vaccine and such potential categorical predictors as patient demographics (including race), socioeconomic status, primary care physician characteristics, historical health service utilization, other health-related conditions, and patient age group (9-17 and 18-26). The study concluded that race and socioeconomic status were most strongly affiliated with successful completion within the 12-month period recommended on the vaccine label.

JHMI operates four clinics in the Baltimore area at which the Gardasil vaccine is available: two in suburban locations (White Marsh and Odenton) and two in Baltimore City (Johns Hopkins Hospital Medical Center and Bayview Medical Center). Our research project represented a retrospective attempt to identify "good" categorical predictors for regimen completion based on electronic and written records from pediatric, family practice, and OB-GYN clinics at these four locations. While the set of potential predictors was much smaller, several of the predictors addressed issues that were not focuses of the Kaiser Permanente study. Patients at JHMI clinics arrive with various forms of insurance, including government-sponsored medical assistance (hitherto referred to simply as "medical assistance"). We wanted to know whether those receiving medical assistance were more likely to fail to complete the sequence than those who had some form of insurance. Also, metro Baltimore differs from Southern California, the site of the Kaiser Permanente study, in that there is a well-defined, independently-governed urban "core" (Baltimore City). We therefore wished to determine whether patients attending suburban clinics were more likely to complete the regimen than those attending clinics in the urban "core." Accordingly, we identified race, type of insurance used, type of practice attended, and clinic location as the four categorical predictors that we wanted to use to predict membership in the "completion" category of the response variable "Did you complete the three-shot sequence within 12 months, or not?".

The balance of this paper is structured as follows. Following a general description of how the "raw" JHMI data were converted into usable form for our study (Section 2) and a list of the preliminary hypotheses that served as a starting point for our investigations (Section 3), we proceed in Section 4 to enumerate the specific pedagogical uses to which our data may be put. The topics that the data are used to illuminate are ordered roughly according to the sequence in which an introductory statistics course might cover these topics. Thus, Sections 4.1 and 4.2 concern data collection and exploratory data analysis, while Sections 4.3 and 4.4 display how the data may be used to illustrate certain inference procedures. In Section 4.5, we discuss in some detail several different ways in which interactions between predictors may cause an observed relationship between a predictor and the response variable (completion rate) to change. Sections 4.6 and 4.7 briefly take up the more advanced topics of odds ratios and logistic regression, which are most commonly covered in a second class on statistical methods but may also be touched upon in sufficiently advanced introductory classes. Section 4 concludes with a list of suggested assignments that extend the explorations covered in the paper and are suitable for use by students who have been exposed to exploratory and inferential techniques.

## 2. The Data

Our data consist of measurements taken from 1413 cases, young female patients aged 11-26 years, who:

- made their first "Gardasil visit" to a JHMI clinic between 2006 (the year of introduction of the Gardasil vaccine) and 2008;
- had 12 months (as noted above, the recommended time period for completion according to the vaccine label) to complete the regimen;
- provided a complete set of the following information: age in years, total number of shots completed (1, 2, or 3), specific clinic attended (Odenton, White Marsh, Outpatient Center, or Bayview), race (white, black, Hispanic, "unknown", or "other"), type of insurance used (private-payer, hospital-based, military, or medical assistance), type of practice (pediatric, family practice, OB-GYN). (Note: Since fewer than 10 individuals in the original data set had no insurance of any sort and self-paid, these cases were removed from the final version of the data set. Further descriptions of the insurance categories may be found in the Appendix.)

We converted the race, insurance, and practice type data into categorical form using numerical labels and also created the following new categorical variables:

- Completion indicator (Did the patient complete the three-shot regimen within the label-recommended period of 12 months?).
- Age group. (Note: We used the age groups 11-17 and 18-26 to match the age categories in the Kaiser Permanente study as closely as possible. However, the instructor may choose to group differently using the available raw age data.)
- Medical assistance indicator (Did the patient receive medical assistance or not?).
- Clinic location type (suburban vs. urban).

During the process of screening and cleaning the data, we removed any cases with incomplete or obviously incorrect information (e.g., mistakenly recording a patient's date of birth as the date of completion of the first shot) and checked to make sure that there was no duplication of patients at different clinics (e.g., a patient getting the first shot at one clinic and subsequent shots at other clinics). As stated previously, the few self-paying cases were also removed.

Additional details on data coding are provided in the Appendix.

Files associated with this article may be downloaded from the JSE server by clicking on the links below.

| Name of File | Description |
|---|---|
| gardasil.txt | Documentation file describing data set |
| gardasil.dat | Data file – tab delimited |
| gardasil.xls | Excel data file |

## 3. Our Preliminary Hypotheses

We began our analysis with the following preliminary hypotheses, based on previous studies and our own subjective assessment of the situation:

- The completion rate for patients at suburban clinics should be higher than the completion rate for patients at urban clinics. (Patients at suburban clinics are more likely to have insurance to pay for their shots and are less likely to be minorities, who tend to have lower completion rates.)
- Patients with some type of insurance should have a higher completion rate than patients receiving medical assistance. (Insurers are more likely to cover most or all of the cost of the shots.)
- Patients attending pediatric clinics should have a higher completion rate than patients attending OB-GYN clinics, with patients attending family practice clinics having a completion rate somewhere in the middle. (Pediatric practitioners are more used to giving shots and therefore are more likely to remind patients to complete their regimen.)

> ***Helpful Hint 1:*** *We suggest that the instructor discuss the reasonableness of all of these hypotheses with students and ask them what they think might be other appropriate hypotheses involving these predictors. (For example, which subgroups of insurance type might be more likely to complete the regimen?)*

Forming a hypothesis based on age group is a bit more difficult. Based on results seen in the Kaiser Permanente study, where 47% of patients aged 18-26 completed the regimen compared with 41% of patients aged 9-17, it would be reasonable to argue that older patients are more likely to complete than younger patients.

> ***Helpful Hint 2:*** *We would encourage the instructor to regard the age-group hypothesis issue as an "open question" and encourage students to construct their own hypothesis based on exploratory data analysis of both the age-group data and the age data itself. See Section 4.2.*

## 4. Pedagogical Uses of the Data

### 4.1. Issues Related to Data Collection

An immediate issue regarding data collection is that it is very likely that the data are subject to numerous confounding factors, due to the fact that this was a retrospective study. For example, we had no way of knowing whether a patient with an incomplete shot regimen might have completed the regimen at a clinic not affiliated with JHMI. Likewise, a patient living in Baltimore City may have "crossed over" to visit a suburban clinic, or a patient living in the suburbs may have visited an urban clinic. Lacking information on patient addresses for reasons of confidentiality, we had no way of identifying such "crossovers."

> ***Helpful Hint 3:*** *The instructor should spend some time discussing the effects of the aforementioned confounding factors on the resulted obtained. For example, what might be the characteristics of a patient who comes from the city to visit a suburban clinic? Might this person be more affluent than normal (since she can afford to travel the extra distance to a suburban clinic location) and thus more likely to receive the support required to complete the regimen?*

> ***Helpful Hint 4:*** *This is also a good opportunity for the instructor to discuss with students how this observational study might be redesigned to reduce the effects of confounding variables. For instance, in the prospective Kaiser Permanente study, the cohort of subjects was restricted to members of the Kaiser Permanente health system, which the authors argued were "broadly representative of the diverse racial/ethnic and socioeconomic backgrounds of the source population in southern California" (Chao et al 2009, p. 865). The Kaiser Permanente study also made extensive use of the company's large information database to get detailed information on subjects' medical histories. Students should be encouraged to consider how they might redesign the Johns Hopkins study to take extra information about the subjects into account, perhaps using the description of the Kaiser Permanente study as a starting point for the discussion.*

## 4.2 Exploratory Data Analysis

Frequency distributions, bar graphs, and pie charts may be easily constructed for each individual predictor variable. The pie charts shown below (Figures 1 and 2) afford a good opportunity for the instructor to point out that "completion of the regimen" and "completion of all three shots" are not identical in this situation. The 537 individuals who completed all three shots included 68 patients who did not complete the regimen within 12 months. These 68 patients were **not** counted as successes when we constructed the completion indicator predictor.

***Figure 1.*** *Pie chart of Number of Shots Completed*

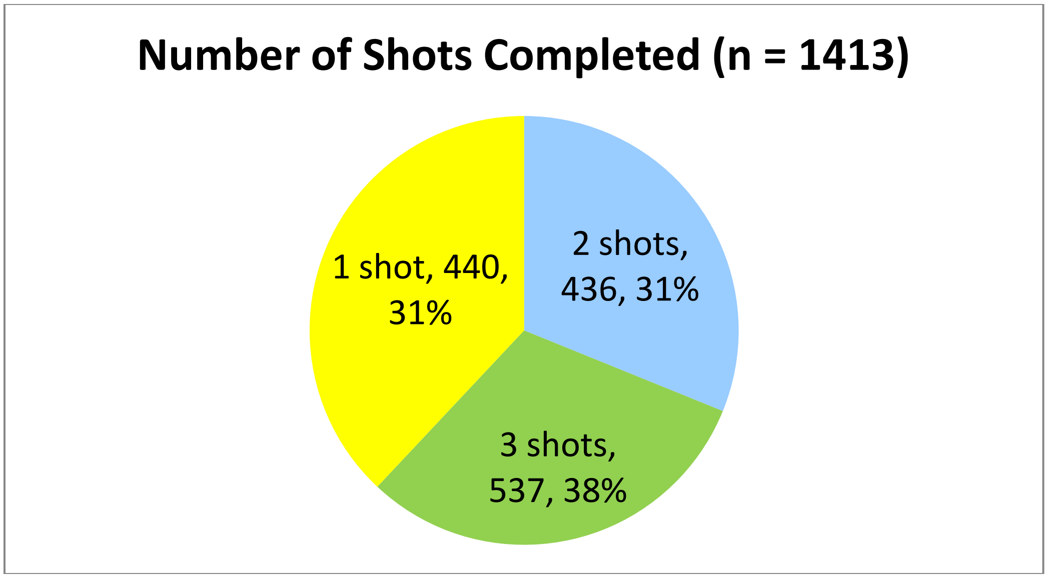
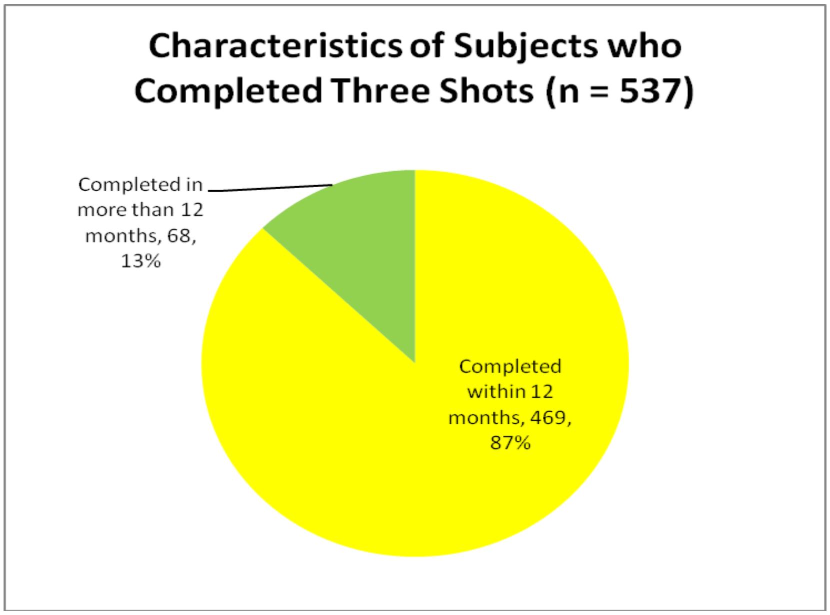


***Figure 2.*** *Pie chart of Characteristics of Subjects Who Completed Three Shots*



A complete demographic breakdown and a summary of completion rates for each category of each predictor may be found in Table 1.

***Table 1.*** *Demographic Breakdown and Summary of Completion Rates*

| **Category** | **Count** | **Overall percentage (*n* = 1413)** | **Count completed within 12 months (*n* = 469)** | **Completion Percentage (within 12 months)** |
|---|---|---|---|---|
| Age 11-17 | 701 | 49.6% | 247 | 35.2% |
| Age 18-26 | 712 | 50.4% | 222 | 31.2% |
| Medical assistance | 275 | 19.5% | 55 | 20% |
| No medical assistance | 1138 | 80.5% | 414 | 36.4% |
| Private payer | 723 | 51.2% | 253 | 35% |
| Hospital based | 84 | 5.9% | 39 | 46.4% |
| Military | 331 | 23.4% | 122 | 36.9% |
| Suburban | 963 | 68.2% | 355 | 36.9% |
| Odenton | 798 | 56.5% | 275 | 34.5% |
| White Marsh | 165 | 11.7% | 80 | 48.5% |
| Urban | 450 | 31.8% | 114 | 25.3% |
| Johns Hopkins | 89 | 6.3% | 19 | 21.3% |
| Bayview | 361 | 25.5% | 95 | 26.3% |
| Pediatrics | 515 | 36.4% | 162 | 31.5% |
| Family practice | 365 | 25.8% | 106 | 29.0% |
| OB-GYN | 533 | 37.7% | 201 | 37.7% |
| White | 732 | 51.8% | 280 | 38.3% |
| Black | 443 | 31.4% | 105 | 23.7% |
| Hispanic | 52 | 3.7% | 17 | 32.7% |
| Other | 61 | 4.3% | 17 | 27.9% |
| Unknown | 125 | 8.8% | 50 | 40% |
| **Total** | **1413** | | **469** | |

Contingency tables and segmented bar graphs may be used to analyze the relationship between any two predictors. Figure 3 displays the relationship between practice type and age group as given in Table 2. The graph and table clearly suggest that (1) completion rates tend to be higher for patients aged 11-17 than for patients aged 18-26, regardless of the type of practice the patient visits; (2) for each age group, completion rates are highest for patients visiting OB-GYN clinics and lowest for patients visiting pediatric clinics. Note that this last observation is not in accord with our preliminary hypothesis on practice type completion rates as stated in Section 3: the relationship between OB-GYN rate and pediatric rate is actually the opposite of what we had expected. Also note that, due to the widely varying sample sizes in the cells (combinations of practice type and age group), it is not clear from graphical analysis alone that the differences we see are statistically significant. See Sections 4.4 and 4.5 for a further discussion of these results.

***Figure 3.*** *Conditional Distribution of Completion Rate Given Each Combination of Age Group and Practice Type*

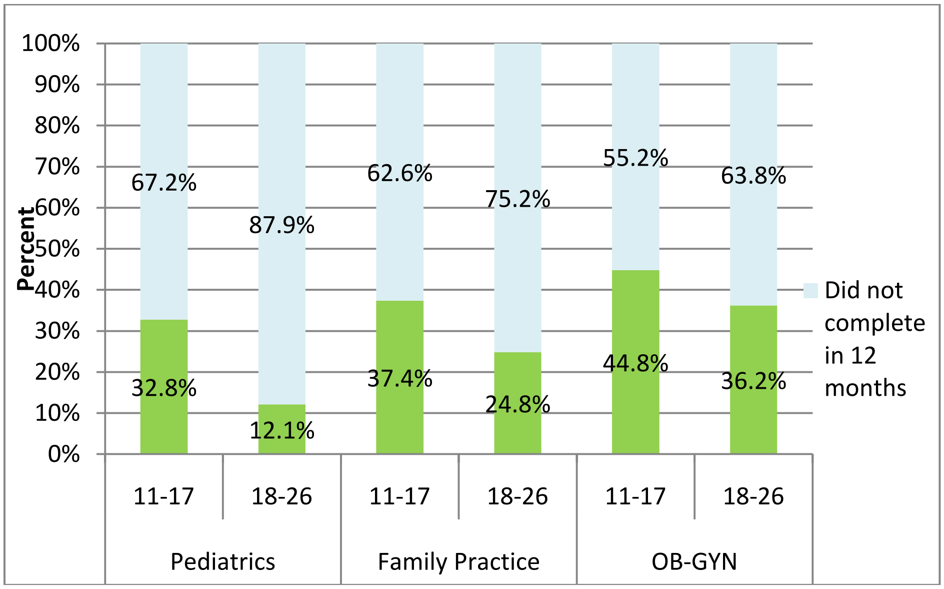


***Table 2.*** *Contingency Table of Conditional Distribution of Figure 3*

| Practice Type | Age Group | Completed | Did not complete | Total |
|---|---|---|---|---|
| Pediatrics | 11-17 | 158 | 324 | 482 |
|  | 18-26 | 4 | 29 | 33 |
| Family Practice | 11-17 | 46 | 77 | 123 |
|  | 18-26 | 60 | 182 | 242 |
| OB-GYN | 11-17 | 43 | 53 | 96 |
|  | 18-26 | 158 | 279 | 437 |
| Total |  | 469 | 944 | 1413 |

## 4.3. Inference for Two Proportions

The predictors of location type, medical assistance indicator, and age group each have two categories. In order to compare completion rates for all categories of these predictors, a two-sample test of the equality of population proportions may therefore be used. For example, suppose that we wish to determine whether the completion rate at suburban clinics is (as we originally hypothesized) greater than the completion rate at urban clinics. Letting $p_1$ = the proportion of those who complete at suburban clinics and $p_2$ = the percentage of those who complete at urban clinics, we therefore test

$H_0$: $p_1 = p_2$
$H_a$: $p_1 > p_2$

Using StatCrunch v5.0 software yields $z = 4.29$, $P < 0.0001$. Our preliminary hypothesis is clearly supported by the available evidence. Likewise, those patients without medical assistance (i.e., those who have some form of medical insurance) have a significantly higher completion rate ($z = 5.18$, $P < 0.0001$) than those who have such assistance. However, the two age groups do **not** have significantly different completion rates ($z = 1.62$, $P = 0.1055$ for the two-sided test). This seems to contradict what we saw in Figure 3, where completion rates for patients aged 11-17 were consistently higher for each practice type. (Note: In this case, and in all other subsequent discussions of statistical significance, "significance" is taken to mean "significance at the 5% level.")

> ***Helpful Hint 5:*** *At this point, the instructor should lead the students to consider the possibility that a confounding variable is affecting the true relationship between age group and completion rate. See Section 4.5 below for a more detailed investigation of this issue.*

> ***Potential Pitfall 1:*** *Note that here we are doing multiple two-sample tests on the same set of data. In other situations in which multiple pairwise comparisons can be done, such as in a one-factor analysis of variance, the overall Type I error rate can be controlled by reducing the desired level of significance $\alpha$ used in each individual comparison. For example, the Bonferroni method in ANOVA changes pairwise $\alpha$ to $\alpha/k$, where $k$ = the number of possible comparisons. Students should be made aware that a similar adjustment may need to be done here.*

When analyzing the relationship between completion rates for two different predictors (cf. Table 2 above), two-sample $z$ tests for proportions may be used to compare completion rates for any two categories of one predictor, given that they are in a particular category of the other predictor. (Note that these rates would be considered conditional probabilities and the set of all such rates would form a conditional distribution.) Based on the data in Table 2, for example, the completion rate for 11-17 year olds is significantly higher in pediatric clinics ($z = 2.47$, $P = 0.007$) and family practice clinics ($z = 2.51$, $P = 0.006$), but not significantly higher in OB-GYN clinics ($z = 1.58$, $P = 0.057$).

> ***Potential Pitfall 2:*** *In certain cases, such as the first (pediatrics) test discussed in the previous paragraph, violation of the assumption of a sufficiently large observed number of successes and failures may be an issue. In our analysis, we used the assumption that each sample contain at least five successes and at least five failures (cf. Baldi and Moore 2009, p. 510). Depending on the text the instructor is using, this assumption may be different. The instructor should make students aware of this issue and the consequences of its violation (i.e., the potential loss of the assumption that the difference in sample proportions of successes is approximately normally distributed) and also how violations might be addressed. Making a Wilson "plus-four" adjustment for a two-proportion z test (i.e., adding one success and one failure to each set of observations) is one way of addressing the issue of violation of the observed success and failure count criterion. (For the pediatrics test above, the Wilson adjustment yields $z = 2.28$, $P = 0.011$.) Some texts do not address the Wilson method, but it has become an*

*increasingly popular way of handling all cases of inference for proportions. The instructor should consider introducing and discussing the idea here.*

### 4.4. Chi-Square Test of Homogeneity

Completion rates for predictors with more than two categories – type of insurance, practice type, and race – may be analyzed using a chi-square test of homogeneity, with the predictor variable categories providing one dimension of the table (row or column) and "completion"/ "non-completion" counts providing the other. In the specific case of practice type, letting $p_1$, $p_2$, and $p_3$ denote completion rates for pediatric, family practice, and OB-GYN clinics, respectively, our hypotheses for a test of homogeneity would be

$H_0$: $p_1 = p_2 = p_3$
$H_a$: at least two of $p_1$, $p_2$, $p_3$ are different

Table 3 displays observed and expected counts for the test (expected counts are in parentheses). The corresponding chi-square test yields $\chi^2 = 8.444$ and P = 0.015, indicating that completion rate is different for at least two of the three practice types.

***Table 3.*** *Observed and Expected Counts for Practice Type Chi-Square Test*

| | **Completed** | **Did not complete** |
|---|---|---|
| **Pediatrics** | 162 (170.9) | 353 (344.1) |
| **Family practice** | 106 (121.2) | 259 (243.8) |
| **OB-GYN** | 201 (176.9) | 332 (356.1) |

The test of homogeneity may also be used to compare conditional completion rates for the aforementioned predictors, given that they are in a particular category of some other predictor. Table 4 displays observed and expected cell counts for a test of homogeneity of practice type completion rates among individuals receiving medical assistance. The test reveals that there is no significant difference ($\chi^2 = 0.255$, P = 0.881) in completion rates among the different practice types for patients receiving assistance. By contrast, it can be shown that there **is** a significant difference ($\chi^2 = 10.6$, P = 0.005) in completion rates among practice types for patients not receiving assistance.

***Table 4.*** *Observed and Expected Counts for Practice Type Chi-Square Test for Individuals Receiving Medical Assistance*

| | **Completed** | **Did not complete** |
|---|---|---|
| **Pediatrics** | 40 (40.8) | 164 (163.2) |
| **Family practice** | 2 (2.4) | 10 (9.6) |
| **OB-GYN** | 13 (11.8) | 46 (47.2) |

A commonly used method of checking the conditions under which this test may be safely used (e.g., Baldi and Moore 2009, p. 572) is that all cell counts be at least 1 and that no more than 20% of all expected cell counts should be less than 5. In the case of Table 4, the conditions are satisfied: only 16% (1/6) of the expected counts are less than 5.

> ***Potential Pitfall 3:*** *The instructor should make sure that students understand the consequences of violating the aforementioned conditions (e.g. losing the key assumption that the test statistic has a chi-square distribution), as well as ways of remedying the situation (e.g., combining rows or columns of the table to increase one or more expected cell counts and adjusting the number of degrees of the chi-square statistic as needed).*

> ***Potential Pitfall 4:*** *Students often have trouble distinguishing between a chi-square test of homogeneity, such as the one used here, and a chi-square test of independence, since both involve analysis of a two-way table. The reasoning should be emphasized here. We started the study with independent samples of patients in each of the "practice type" categories, with the size of each sample being predetermined due to the retrospective nature of the study, and compared the percentages of individuals in each sample who completed the regimen. This is different from a case in which we select a single sample from a single population and classify it according to the categories of two categorical variables, which would necessitate a test of independence. Since the tests' mechanics are the same, the instructor's emphasis here should be on how the test hypotheses should be written and how the results should be interpreted.*

## 4.5. The Effects of Confounding Variables

It is well known that observed associations between variables can sometimes be misleading when a confounding variable is present. Simpson's paradox (cf. Baldi and Moore 2009, p. 138) is one famous example of this phenomenon, in which associations between two variables can reverse direction when data are aggregated across the categories of a third variable. While our data do not illustrate Simpson's paradox, they do provide illustrations that such aggregations may also cause associations to vanish, or to otherwise behave in a manner different than what one would expect. Below, we show how the data illustrate several different ways in which the relationship between a predictor and the completion rate can be affected by the presence or absence of a confounding predictor.

**a. "Masking."** We saw earlier (cf. in the paragraph prior to Table 2 and Figure 3, and also in Section 4.3) that patients aged 11-17 consistently have higher completion rates for all three different practice types, yet the two age groups do not have significantly different completion rates when practice types are combined or pooled together. In this case, aggregation across practice type is causing an existing association between age group and completion rate to be hidden or "masked."

b. "**Cancellation.**" Aggregation over location type, like aggregation over practice type, causes the relationship between age group and completion rate to disappear. However, the specific effect on the relationship is different. Figure 4 and Table 5 show that the relationship between age group and completion rate changes direction when separate location types are considered. In suburban clinics, where the completion rate is 36.9%, the completion rate for patients aged 11-17 is significantly **higher** than that of patients aged 18-26 ($z$ = 4.42, P < 0.0001). However, in urban clinics, where the completion rate is 25.3%, the completion rate for patients aged 11-17 is significantly **lower** than that of patients aged 18-26 ($z$ = -2.28, P = 0.011). In other words, aggregation over location type is causing the two opposing relationships to cancel one another out. The explanation for this phenomenon may be found in Tables 1 and 5. 42% of all patients aged 11-17 go to urban clinics, where the completion rate is much lower, compared with only 22% of patients aged 18-26. Since Figure 4 shows that the completion rate for patients aged 18-26 is not greatly affected by the location of the clinic where the treatment is taking place (31% suburban vs. 31.8% urban), but the completion rate for patients aged 11-17 clearly is (44.9% suburban vs. 22% urban), the relationships between age group and completion rate cancel upon aggregation over location type. Note that this analysis indicates that younger patients in urban clinics may be at particular risk for non-completion.

***Figure 4.*** *Conditional Distribution of Completion Rate Given Each Combination of Age Group and Location Type*

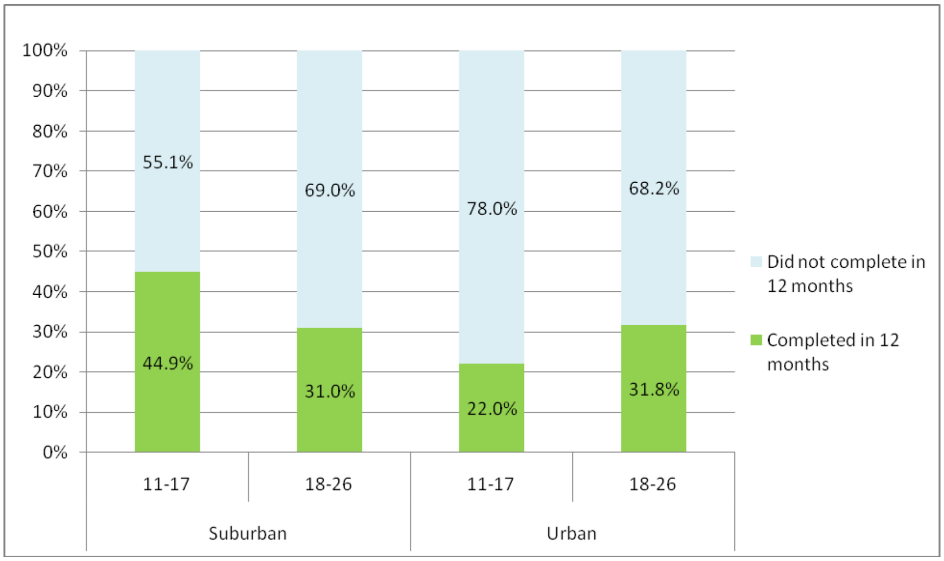


***Table 5.*** *Contingency Table of Conditional Distribution of Figure 4*

| Location Type | Age Group | Completed | Did not complete | Total |
|---|---|---|---|---|
| Suburban | 11-17 | 182 | 223 | 405 |
| | 18-26 | 173 | 385 | 558 |
| Urban | 11-17 | 65 | 231 | 296 |
| | 18-26 | 49 | 105 | 154 |
| Total | | 469 | 944 | 1413 |

**c. Unexpected Relationships.** A confounding predictor may also sometimes help to explain why an expected relationship between another predictor and the completion rate does not behave as we would anticipate. We saw in the discussion of the relationship between age group and practice type in Section 4.2 that the pediatric completion rate, contrary to our preliminary hypothesis in Section 3, was lower than the OB-GYN completion rate. Including the third variable of location type as part of the analysis of the relationship between practice type and completion rate (Figure 5 and Table 6) reveals that, while the pediatrics completion rate was significantly **higher** than the OB-GYN completion rate ($z$ = 2.26, P = 0.012) in suburban clinics, it was significantly **lower** ($z$ = -3.84, P < 0.0001) in urban clinics. However, 72% of all OB-GYN patients attended suburban clinics, while only 42% of pediatric patients did so (Table 6). While the OB-GYN completion rate was not greatly affected by the location of the clinic (38.2% suburban vs. 36.4% urban), the extremely low completion rate for urban pediatric patients, combined with the fact that over half of all pediatric patients attended urban clinics, caused the overall pediatric completion rate to be "pulled" significantly below the overall OB-GYN completion rate.

***Figure 5.*** *Conditional Distribution of Completion Rate Given Each Combination of Practice Type and Location Type*

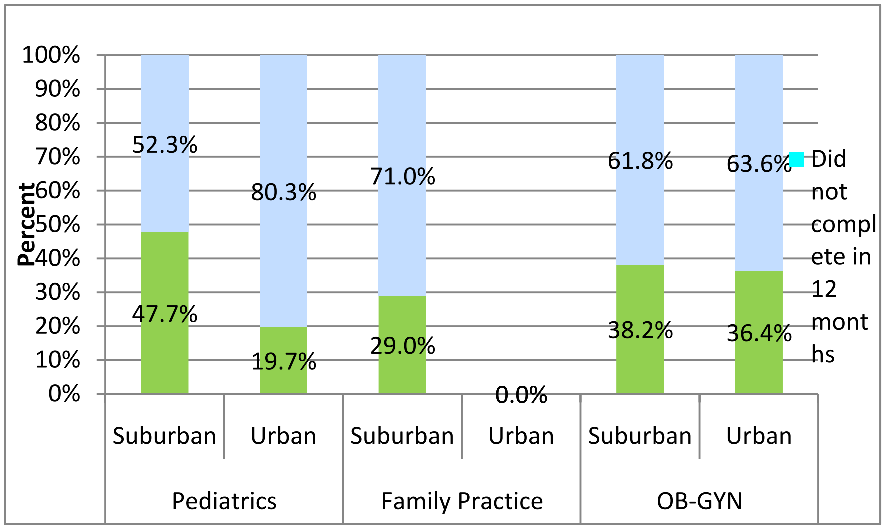


***Table 6.*** *Contingency Table of Conditional Distribution of Figure 5*

| | | Pediatrics | Family practice | OB-GYN | Total |
|---|---|---|---|---|---|
| Suburban | Completed | 103 | 106 | 146 | 355 |
| | Did not complete | 113 | 259 | 236 | 608 |
| Urban | Completed | 59 | | 55 | 114 |
| | Did not complete | 240 | | 96 | 336 |
| Total | | 515 | 365 | 533 | 1413 |

Taken together with the data in Table 1, the results in this section suggest that, while patients attending urban clinics are at greater risk for non-completion in general (Table 1; 25.3% completed), younger urban patients appear to be at particularly high risk (Figure 4 and Table 5; 22.0% completed), especially those going to pediatric clinics (Figure 5 and Table 6; 19.7% completed).

> ***Helpful Hint 7:*** *The results described above, taken together with similar results that can be obtained for other predictors (cf. the suggested assignments in Section 4.8), naturally lend themselves to a classroom discussion of an appropriate "action plan" for targeting groups that are at risk for non-completion. How might the statistical analysis of these data be incorporated into such efforts? (For example, in brochures or posters made available by health officials?)*

## 4.6 Odds Ratios

These retrospective data easily lend themselves to the calculation of odds ratios, either by hand or by using calculator or computer software, as well as to the interpretation of results in terms of odds ratios. Suppose that we wish to compare completion rates in suburban and urban areas using odds ratios. 355 of 963 suburban patients completed the regimen, compared to 114 of 450 urban patients. The odds ratio for completion, suburban vs. urban, is therefore (355)(336) ÷ (608)(114), or approximately 1.72. A corresponding 95% confidence interval for the OR is [1.341, 2.209]. Since the confidence interval contains only values greater than 1, we can extend the interpretation of the confidence interval to say that we are 95% confident that the odds in favor of completion are higher for suburban patients than for urban patients.

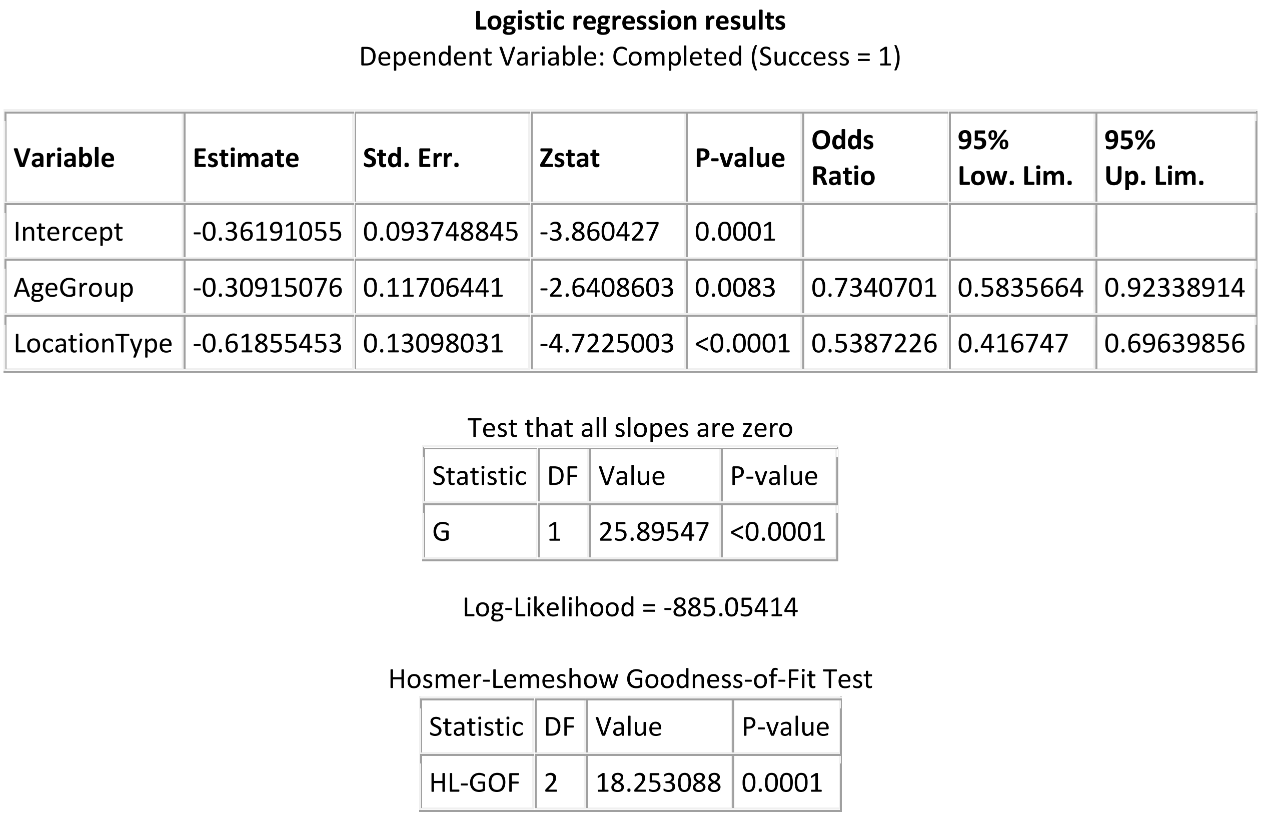
## 4.7 Logistic Regression

The dichotomous nature of the response variable in this situation – a patient either completes the regimen within 12 months (success) or does not (failure) – gives the instructor the opportunity to introduce the concept of logistic regression using one, some, or all of the available predictors. The depth of coverage will depend upon the course being taught. In an elementary course, the instructor may wish to use a simple software package, such as StatCrunch, to display and discuss logistic regression output and how it differs from the output for simple linear regression. A more in-depth course, such as a course in biostatistics, could address issues of model fit and the accuracy of predictions using the logistic model.

As an illustration of how logistic regression results may be presented, consider the data in Table 5, which displayed the relationship between location type and age group. Table 7 displays StatCrunch v5.0 output for a logistic regression model in which age group and location type are used as the predictors and "completion (yes/no)" as the response. The output suggests that the model as a whole is useful for prediction of completion (cf. the result of the test "Test that all slopes are zero") and that each individual predictor (main effect) is also useful for prediction (cf.

the P-values in the table below "Dependent Variable"). The odds ratio results indicate that patients aged 11-17 (AgeGroup = 0) and suburban patients (LocationType = 0) have higher completion rates than patients aged 18-26 (AgeGroup = 1) and urban patients (LocationType = 1). However, the goodness-of-fit test result provided by StatCrunch v5.0, the Hosmer-Lemeshow test, suggests that the model's actual predictive ability is rather poor. The model in which age group and insurance type ("medical assistance (yes/no)") are used as predictors turns out to have better predictive ability (Hosmer-Lemeshow P = 0.906), as does the three-predictor model in which age group, insurance type ("medical assistance (yes/no)"), and practice type are the predictors (Hosmer-Lemeshow P = 0.929).

***Table 7.*** *Logistic Regression Output with Age Group and Location Type as Predictors*



**Logistic regression results**

Dependent Variable: Completed (Success = 1)

| **Variable** | **Estimate** | **Std. Err.** | **Zstat** | **P-value** | **Odds Ratio** | **95% Low. Lim.** | **95% Up. Lim.** |
|---|---|---|---|---|---|---|---|
| Intercept | -0.36191055 | 0.093748845 | -3.860427 | 0.0001 | | | |
| AgeGroup | -0.30915076 | 0.11706441 | -2.6408603 | 0.0083 | 0.7340701 | 0.5835664 | 0.92338914 |
| LocationType | -0.61855453 | 0.13098031 | -4.7225003 | <0.0001 | 0.5387226 | 0.416747 | 0.69639856 |

Test that all slopes are zero

| Statistic | DF | Value | P-value |
|---|---|---|---|
| G | 1 | 25.89547 | <0.0001 |

Log-Likelihood = -885.05414

Hosmer-Lemeshow Goodness-of-Fit Test

| Statistic | DF | Value | P-value |
|---|---|---|---|
| HL-GOF | 2 | 18.253088 | 0.0001 |

## 4.8 Suggested Assignments

We have touched upon only a few of the possible specific ways in which these data may be analyzed using the methods described in Section 4. Here are several other suggestions that may make for good student assignments or projects.

- ***Are race and socioeconomic status, as was suggested in the Kaiser Permanente study, truly useful predictors for successful completion of the Gardasil regimen?*** Table 1 suggests that suburban completion rate is higher than urban completion rate; the student then may be asked to show that this relationship holds regardless of the patient's age, race, or clinic type. Similarly, appropriate analyses can be used to indicate that any patient receiving medical assistance should be considered at greater risk for non-completion, regardless of the level of any other predictor. The same holds true for any black or Hispanic patient. All of these results support the notion that race and socioeconomic status are useful predictors for completion.

- ***Students should be encouraged to investigate completion rates for additional combinations of predictors using appropriate graphical and inference methods and to identify situations in which there appears to be an interaction between predictor categories.*** Here are some possible examples. (In all cases, "significance" means at the 5% level.)
    - Completion rates for patients aged 11-17 are significantly higher than those for patients aged 18-26 in pediatric clinics and family practice clinics, but not significantly higher in OB-GYN clinics.
    - Among patients aged 11-17, completion rates for the three practice types are not significantly different. Among patients aged 18-26, however, such rates are significantly different.
    - Among those receiving medical assistance, there is significant difference in completion rate for those going to suburban and urban clinics. Among those not receiving medical assistance, however, the completion rate at suburban clinics was significantly higher than the completion rate at urban clinics.
    - Among white patients, suburban and urban completion rates are not significantly different. However, among blacks and Hispanics, the suburban completion rate is significantly higher than the urban completion rate.
    - Among patients going to pediatric and OB-GYN clinics, the completion rate for whites is significantly higher than it is for blacks and Hispanics, and the completion rate for those not receiving medical assistance is significantly higher than for those receiving medical assistance. For patients attending family practice clinics, however, these pairs of completion rates are not significantly different.

## 5. Conclusion

We believe that these data furnish an interesting and highly relevant way of exploring categorical data analysis. Female students will particularly appreciate how the knowledge gleaned from the analysis may affect decisions related to their own health. At the same time, the fact that the data were not generated as the result of a planned experiment or prospective study allows the instructor to append a discussion as to how the data collection process might be improved. Students can discuss and design a "better" data-gathering method, possibly using clinics in their own neighborhood in place of the JHMI clinics.

## APPENDIX - Data Coding

See the documentation file www.amstat.org/publications/jse/v19n1/gardasil.txt for a more complete description.

- **Age:** Patient's age in years
- **AgeGroup:** Patient's age group (0 = 11-17, 1 = 18-26)
- **Race:** Patient's race (0 = white, 1 = black, 2 = Hispanic, 3 = other/unknown)
- **Shots:** Number of shots completed
- **Completed:** Did the patient complete the three-shot sequence within a 12-month period? (0 = no, 1 = yes)
- **InsuranceType:** Type of insurance (0 = medical assistance, 1 = private payer [Blue Cross Blue Shield, Aetna, Cigna, United, Commercial, CareFirst], 2 = hospital based [EHF], 3 = military [USFHP, Tricare, MA])
- **MedAssist:** Medical assistant indicator variable (0 = patient does not have medical assistance, 1 = patient has medical assistance)
- **Location:** Clinic that patient attended (1 = Odenton, 2 = White Marsh, 3 = Johns Hopkins Outpatient Center, 4 = Bayview)
- **LocationType:** Location type indicator variable (0 = suburban, 1 = urban)
- **PracticeType:** Type of practice patient visited (0 = pediatric, 1 = family practice, 2 = OB-GYN)

## Acknowledgement

This paper derives from an undergraduate research project conducted by Courtney Wright at Stevenson University during the Spring Term of 2010. Dr. Barat mentored the project, while Dr. Chou provided the JHMI data and the necessary background information and references on the nature of HPV and the Gardasil vaccine. Our data analysis was the basis for a Contributed Paper, *Establishing Best Predictors for Completion of the HPV4 Vaccine Sequence by Female Patients at Clinics of Johns Hopkins Medical Institutions*, to the Session on Health Policy Statistics at the August 2010 Joint Statistical Meetings in Vancouver, BC. Ms. Wright also presented a poster summarizing the results of the analysis at the Stevenson University School of the Sciences' Spring Poster Session in May 2010.

The authors thank Dexter Whittinghill, Editor of the Data Sets and Stories section of the Journal of Statistics Education, and the other individuals who reviewed the paper for JSE and provided numerous helpful suggestions and corrections. We also wish to acknowledge the support and encouragement of the late Dr. Susan Slattery, former Chair of the Department of Mathematics at Stevenson University, who was killed in an accident as the first draft of this paper was being completed.

## References

Baldi, B. and Moore, D.S. (2009). The Practice of Statistics in the Life Sciences (1st ed.). New York: W.H. Freeman. 513-514.

Boughton, B. (2008). "Survey hints at uneven adoption of cancer vaccine." Nature Medicine 14, 5.

Chao, C., Velicer, C., Slezak, J.M., and Jacobsen, S.J. (2009). "Correlates for Completion of 3-Dose Regimen of HPV Vaccine in Female Members of a Managed Care Organization." Mayo Clinical Proceedings 84, 864-870.

Department of Health and Human Services, Centers for Disease Control and Prevention (2008). "Vaccination Coverage Among Adolescents Aged 13-17 Years - United States, 2007." Morbidity and Mortality Weekly Report 57, 1100-1103.

Dunne, E.F., and Markowitz, L.E. (2006). "Genital Human Papillomavirus Infection." Clinical Infectious Diseases 43, 624-629.

Garland, S.M., Hernandez-Avila, M., Wheeler, C.M., Perez, G., Harper, D.M., Leodolter, S., Tang, G.W.K., Ferris, D.G., Steben, M., Bryan, J., Taddeo, F.J., Railkar, R., Esser, M.T., Sings, H.L., Nelson, M., Boslego, J., Sattler, C., Barr, E., and Koutsky, L.A. (2007). "Quadrivalent Vaccine against Human Papillomavirus to Prevent Anogenital Diseases." New England Journal of Medicine 356, 1915-1927.

Markowitz, L.E., Dunne, E.F., Saraiya, M., Lawson, H.W., Chesson, H., and Unger, E.R. (2007). "Quadrivalent Human Papillomavirus Vaccine: Recommendations of the Advisory Committee on Immunization Practices." Morbidity and Mortality Weekly Report, 56, 1-23.

---

Christopher E. Barat, PhD.
Associate Professor of Mathematics
Stevenson University
1525 Greenspring Valley Road
Stevenson, MD 21153
mailto:cbarat@stevenson.edu
(443) 334-2329

Courtney Wright
c/o Department of Mathematics
Stevenson University
1525 Greenspring Valley Rd.
Stevenson, MD 21153

Betty Chou, M.D.
Assistant Professor of Obstetrics and Gynecology

Johns Hopkins University Bayview Medical Center
4940 Eastern Avenue
Baltimore, MD 21224

Volume 19 (2011) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Guidelines for Readers/Data Users | Home Page | Contact JSE | ASA Publications