CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: https://www.tandfonline.com/loi/ucha20

Taking a Chance in the Classroom: Speed Dating: Exploring Initial Romantic Attraction

Kari Lock Morgan, Mine Çetinkaya-Rundel & Dalene Stangl

To cite this article: Kari Lock Morgan, Mine Çetinkaya-Rundel & Dalene Stangl (2014) Taking a Chance in the Classroom: Speed Dating: Exploring Initial Romantic Attraction, CHANCE, 27:2, 44-49, DOI: 10.1080/09332480.2014.914752

To link to this article: https://doi.org/10.1080/09332480.2014.914752

Kari Lock Morgan, Mine Çetinkaya-Rundel, and Dalene Stangl
Column Editors

Speed Dating: Exploring Initial Romantic Attraction

What predicts initial romantic attraction in the opposite sex? Asking questions like this that are intrinsically interesting to students helps them see that data analysis is not just about finishing an assignment but that it's a powerful tool for daily life and a means of coming to understand the world in which we live.

The Data

The data come from speed dating experiments conducted at Columbia University between 2002–2004. Participants were students from Columbia's graduate and professional schools: 276 males and 276 females. Each person met with 10–20 people of the opposite sex for four minutes (all questions asked pertain only to heterosexual dating). The order of pairings and session assignments was random. The data include responses to an initial questionnaire, ratings of partners after each four-minute speed date, and responses to follow-up surveys the day after and 3–4 weeks later.

We present a simplified version of the data, as used in *Statistics: Unlocking the Power of Data* by Lock, Lock, Lock Morgan, Lock, and Lock. This data set is organized such that each case (row) is a date and includes both male and female responses to basic questions. To avoid dependencies in the data (an issue to bring up with your students), only first pairings are included, yielding 276 observations. Each variable listed in Table 1 appears twice, once with the variable name followed by an M, indicating male answers, and once followed by an F, indicating female answers. For example, AgeM is the age of the male, and AttractiveF is the female's rating of the male's attractiveness. This data set can be found at *http://chance.amstat.org/2014/04/supplemental-vol-27-no-2*.

Dating is a topic of inherent interest to students, and this data set is rich in variables both categorical and quantitative. In this article, we present ways in which these data can be used in introductory, second-level (regression), and more advanced undergraduate and graduate courses.

Introductory Statistics

This data set has been used by the authors at various points throughout the semester in an introductory statistics course: week one when students are seeing data for the first time, early in the course when they are learning descriptive statistics, midstream as a review/synthesis of confidence intervals and hypothesis testing, and late in the course when covering multiple regression.

When presented during the first week of class, this data set helps students gain comfort and familiarity with observations and variables,

as well as with thinking about relevant questions upon which the data can shed light. Students are asked to identify basic properties of the data set such as observations (each observation is a date, not a person), the types of variables, and which are categorical and quantitative. Students are also asked to come up with questions that these data can help answer, and they return to these initial questions later in the course. This exercise provides a powerful reminder to students of how much they have learned and how far they have come since the beginning of the course.

The speed dating data set is particularly useful for teaching descriptive statistics or reviewing basic methods of inference because it has both quantitative and categorical variables from which interesting univariate and bivariate questions arise. A few are listed below, but we encourage you to make up your own, or even better, have your students come up with their own.

- What proportion of speed daters want to see their partner again?
- What proportion of speed dates result in a "match"?
- Are males or females more selective?
- Are males or females more optimistic about their partner's opinions of them?
- Is a match more likely to result between people of the same race?
- Is there a correlation between intelligence and attractiveness ratings?
- Do females prefer males older than themselves?
- Are people more interested if they think the interest is mutual? (Are people more likely to want to see their partner again if they think their partner is more likely to want to see them?)

Table 1—Speed Dating Variables

Decision	1 = Yes (want to see the date again), 0 = No (do not want to see date again)
Like	Overall, how much do you like this person? (1 = don't like at all, 10 = like a lot)
PartnerYes	How probable do you think it is that this person will say 'yes' for you? (1=not probable, 10=extremely probable)
Age	Age
Race	Race (Caucasian, Asian, Black, Latino, or Other)
Attractive	Rate attractiveness of partner on a scale of 1–10 (1 = awful, 10=great)
Sincere	Rate sincerity of partner on a scale of 1–10 (1 = awful, 10=great)
Intelligent	Rate intelligence of partner on a scale of 1–10 (1 = awful, 10=great)
Fun	Rate how fun partner is on a scale of 1–10 (1 = awful, 10=great)
Ambitious	Rate ambition of partner on a scale of 1–10 (1 = awful, 10=great)
Shared Interests	Rate the extent to which you share interests/hobbies with partner on a scale of 1–10 (1 = awful, 10=great)

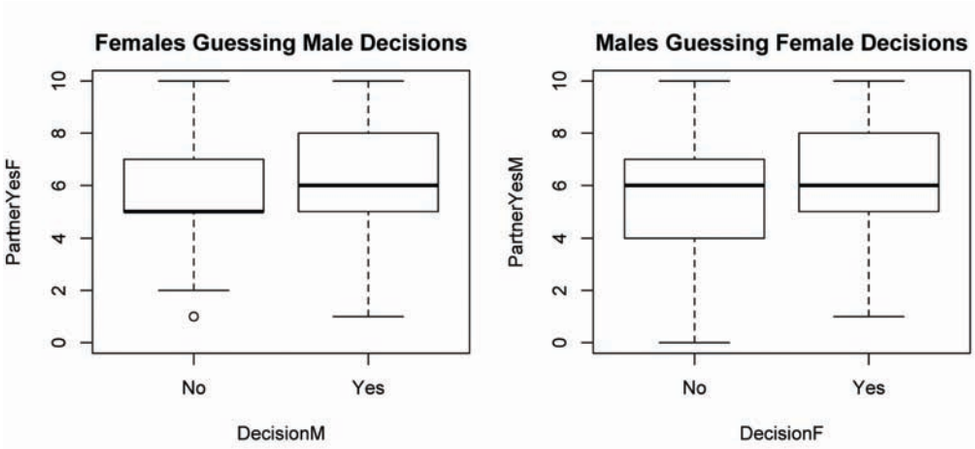

Figure 1. Comparing guesses at partner's chance of saying yes by actual partner decision

- Is how much date partners like each other reciprocal?
- Do different races tend to be more selective, or tend to have different average overall liking scores?
- Which of the six attribute ratings are associated with whether the person wants to see their partner again? Does this differ by gender?
- Are older (or younger) people more selective?
- Can people predict whether their partner is interested in them?

As an example, we explore the last question: Can people predict whether their partner is interested in them? The variables PartnerYesM and PartnerYesF measure how probable a person thinks it is that their partner will say "yes" to wanting to see them again, and the variables DecisionM and DecisionF record whether each person actually wants to see their partner again. The relationships between these variables are shown in Figure 1.

Note that for PartnerYes, although students were asked to rate how probable they think it is that their partner will say "yes" to them on the scale of (1=not probable, 10=extremely probable), one

Table 2—Sample Mean PartnerYes Score for Each Group, and *P*-Value Testing for a Difference Within Each Gender

	Partner did not want second date	Partner did want second date	One-sided *p*-value
Average male guess at how probable (1-10) female was to want second date	5.4	6.1	0.005
Average female guess at how probable (1-10) male was to want second date	5.4	6.2	0.002

MALES

Variable	Coefficient	SE	95% CI	*p*-value
(Intercept)	-3.43	1.05	(-5.57, -1.41)	0.001
AttractiveM	1.02	0.16	(0.73, 1.34)	6×10^{-11}
SincereM	-0.16	0.14	(-0.44, 1.34)	0.23
IntelligentM	-0.27	0.17	(-0.62, 0.07)	0.12
FunM	-0.02	0.13	(-0.28, 0.23)	0.85
AmbitiousM	-0.09	0.13	(-0.35, 0.16)	0.47
SharedInterestsM	0.17	0.09	(-0.01, 0.36)	0.07

FEMALES

Variable	Coefficient	SE	95% CI	*p*-value
(Intercept)	-5.18	1.12	(-7.50, -3.10)	4×10^{-6}
AttractiveF	0.33	0.10	(0.13, 0.54)	0.002
SincereF	-0.02	0.12	(-0.26, 0.22)	0.89
IntelligentF	0.07	0.16	(-0.24, 0.39)	0.67
FunF	0.25	0.11	(0.03, 0.47)	0.03
AmbitiousF	-0.11	0.11	(-0.34, 0.11)	0.32
SharedInterestsF	0.30	0.09	(0.14, 0.48)	0.0005

Figure 2. Output of logistic regression models with decision as a response

male thought his female partner was so unlikely to say yes that he answered 0, or of course it could be a typo in an overly confident response of 10.

The sample means for each group are displayed in Table 2, as well as a one-sided *p*-value testing for a difference in means for each gender.

For both genders, there is a relationship between whether a person wants a second date and how probable he/she believes it is that his/her partner will want a second date—students *can* predict whether their partner is interested in them, in the sense that guesses at a partner's chance of saying yes are higher, on average, for partners who actually say yes. Also, this relationship and the mean PartnerYes scores in each group are surprisingly consistent for males and females.

There are many other interesting questions that can be asked just from this simple subset of the data, and many more if you extend to the full data set described below. If doing inference, some questions require confidence intervals and some require hypothesis tests. Because race/ethnicity has multiple categories, questions involving this variable also bring in chi-square tests and ANOVA. This makes the data set excellent for review and synthesis at the end of the course.

Regression

This data set also works well for introducing multiple regression and logistic regression. The issue of multicollinearity is clear. Various regression models are appropriate, but a natural choice is to regress either Decision (for logistic regression) or Like (for multiple regression) on the six attributes (Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests). This uses a person's ratings of their partner on these attributes to predict either how much they like their partner overall, or

whether they decide they want to see their partner again. Running these models for males and for females (or with interaction variables) allows for interesting comparisons. The output of logistic regression models with Decision as a response, for males and females, is shown in Figure 2.

Not surprisingly, attractiveness is very significant in both cases, especially for men. Attractiveness is clearly the easiest quality to judge after just a four-minute speed date, and pointing out how this might change in the real dating world where interactions last longer than four minutes is a good way to start a discussion on generalizability of findings from an experimental setting to the real world. For men, based solely on the output from this regression, it appears to be basically all about physical attractiveness. For females, how much they share interests with males, and how fun the males are (or are perceived to be) are also significant. However, in bivariate logistic regressions for men, FunM, AmbitousM, and SharedInterestM are also significant, and for the women, all except AmbitiousF are significant.

These variables are all highly correlated, providing a great opportunity to discuss multicollinearity. Figure 3 is a scatterplot matrix for males (the corresponding plot for females is similar), basically showing that, if someone likes someone else, they tend to rate them more highly regarding all of the attributes (which is interesting in its own right).

In particular, we can focus on one of the negative coefficients, such as Ambitious for females. We pose the following questions to students for discussion: "Does this mean that in this sample, females tend to prefer males who they perceive as less ambitious?" By running a regression on Ambitious alone (shown in Figure 4), we see this not to be the case. The coefficient for Ambitious is only negative when other variables

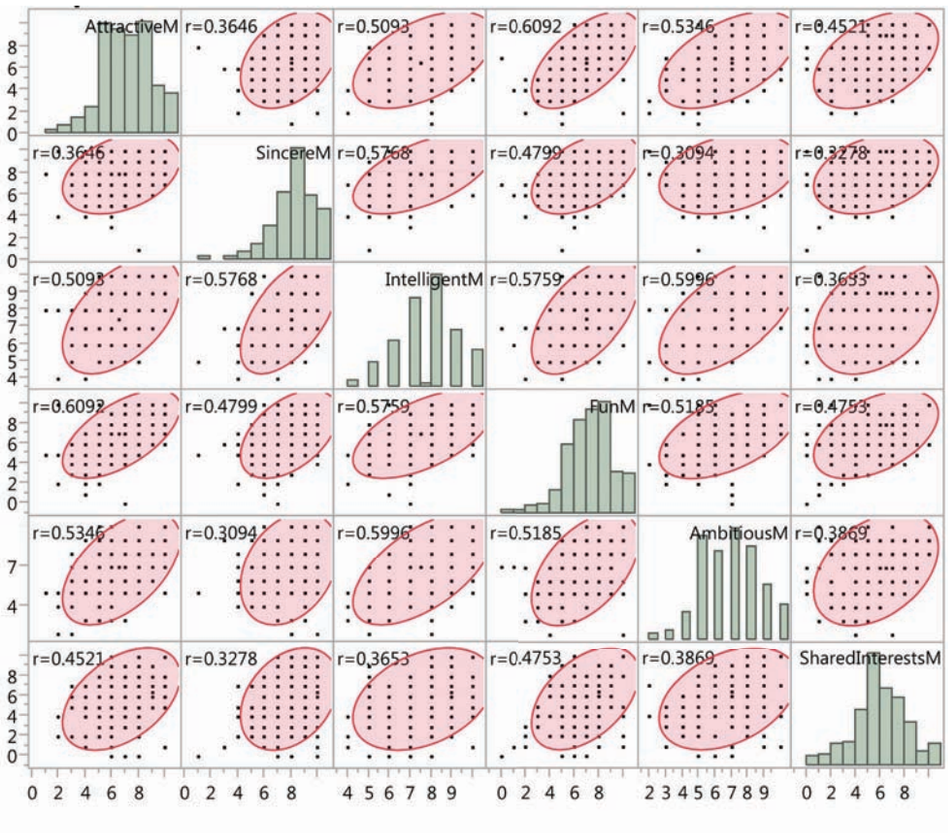

Figure 3. Scatterplot matrix of explanatory variables

Figure 4. Logistic regression of Decision on Ambitious for females

correlated with Ambitious are also included in the model.

This example illustrates that individual coefficients can be misleading when variables are highly correlated, and that coefficients of variables depend on the other variables included in the model.

More Advanced Statistics Courses

In more advanced statistics courses, a complex data structure may be an advantage, as it calls for more sophisticated modeling techniques. Here we describe a more complete version of the data (the original version of the data set), as used in *Data Analysis Using Regression and Multilevel/Hierarchical Models* by Andrew Gelman and Jennifer Hill. The data and codebook are available on Andrew Gelman's website at *www.stat.columbia.edu/~gelman/arm/examples/speed.dating*. Here, each row represents a date from the perspective of one partner, thus each date appears twice.

This version includes data on all dates, not just first dates, as well as additional variables, yielding a data set with 8,379 rows and 169 variables.

On an initial questionnaire, participants filled out basic information such as age, race/ethnicity, field of study, intended career, where they are from, their undergraduate institution, how frequently they date and go out, and their interest in a variety of activities. They also answered questions pertaining to the speed dating experiment, such as their goal and expectations for the event. The data set also includes variables corresponding to the pair, such as whether the participants were the same race, the correlation between their interest ratings, etc. Other variables include information pertaining to the wave of the experiment and the date (such as position and order).

After each speed date, participants filled out a scorecard rating their partner, with questions such as the following:

- Would you like to see him or her again (yes/no)
- Rating of attributes: Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests (1–10)
- Overall, how much do you like this person? (1–10)
- How probable do you think it is that this person will say 'yes' for you? (1–10)
- Have you met this person before? (yes/no)

If two people both responded they wanted to see each other again, the pair was designated as a "match," and they exchanged email addresses. Follow-up surveys gathered data on whether they had contacted or been contacted by any of their matches and whether they had been on any dates resulting from the speed dating.

At multiple times (before speed dating, halfway through the session, the day after, and 3–4 weeks later), participants were asked to consider the six attributes (Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests) and rate the importance of each attribute on a scale of 1–10, or else distribute 100 points among the attributes according to importance (the instructions differed slightly with different waves of the experiment). Each time, they rated these attributes from each of the following perspectives:

- We want to know what you look for in the opposite sex.
- What do you think MOST of your fellow men/women look for in the opposite sex?
- What do you think the opposite sex looks for in a date?
- How do you think you measure up?
- And finally, how do you think others perceive you?
- (follow-up only) What best reflects the actual importance of attributes in your decisions?

This data set can be used to practice data wrangling and munging. The provided data set is organized such that only the responses from one partner appear in each row; however, answering questions involving responses from both partners (e.g., Do males prefer females younger than them?) requires finding the partner information according to the partner ID number. Completing this task requires some coding ability on the part of the students and hence provides a great exercise in data manipulation for more advanced students.

Another way in which this more complex data set differs from the simpler version is the obvious violation of the independence condition. Each date corresponds to two cases, and each person corresponds to as many cases as he/she had dates (or twice that, if you count them as the partner as well). While this makes analysis of this form of the data set beyond the scope of introductory courses, the dependencies in the data are exactly what make these data statistically interesting for more advanced students.

One way to deal with the dependencies is multilevel modeling. One level is the person (including all variables constant for a person, such as age and race), and another level is each of that person's dates (including ratings of each partner). For example, if using regression, each person differs slightly in how they rate people, suggesting allowing intercepts and/or slopes in the regression to vary by person. We also could take this a step further and allow intercepts and/or slopes to vary as well based on the person being rated (e.g., some people are more attractive than others).

Alternatively, the dependencies could be handled by viewing this data set as a network. These data represent a bipartite graph (actually a collection of several complete bipartite graphs), and thus can be analyzed from this framework. In this setting, the different types of variables fall naturally into place; there are variables pertaining to each node (person) and each edge (date). These different types of variables, as well as the dependencies, can be explicitly brought into the analysis via network modeling.

Using the Original Paper

While educational to let students generate their own research questions, it also can be good to provide students with the original paper and ask them to verify some of the findings. This paper in particular has some interesting results. Below are direct quotes from the paper your students can explore:

- "Women put greater weight on the intelligence and the race of partner, while men respond more to physical attractiveness."
- "Women put more emphasis on the partner's race."
- "On average, men do not value women's intelligence or ambition when it exceeds their own; moreover, a man is less likely to select a woman whom he perceives to be more ambitious than he is."
- "We find that women exhibit a preference for men who grew up in affluent neighborhoods."
- "We find that male selectivity is invariant to the number of potential partners, while female selectivity is strongly increasing in it. Surprisingly, female subjects are no more selective than males in small groups; rather, it is the female elasticity of the number of acceptances (i.e., the number of males a female subject wishes to meet again) with respect to group size that is lower than the male elasticity."

Showing the students published literature and asking them to verify (or contradict) the results can be empowering for them and demonstrate the usefulness of what they are learning.

Conclusion

College students tend to be inherently interested in exploring what makes a person desirable and other considerations surrounding dating. Illustrating statistical techniques and concepts with speed dating data takes advantage of this intrinsic interest. This data set is sufficiently complex to be useful for asking a variety of interesting questions and illustrating a variety of statistical techniques, yet it can be simplified to work in the first week of an introductory statistics course.

Further Reading

Fisman, R., S. Iyengar, E. Kamenica, and I. Simonson. 2006. Gender differences in mate selection: Evidence from a speed dating experiment. *Quarterly Journal of Economics* 121(2): 673–697.

Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Lock, Lock, Lock Morgan, Lock, and Lock. 2013. *Statistics: Unlocking the power of data*. Hoboken, NJ: John Wiley & Sons.

About the Authors

Mine Çetinkaya-Rundel is an assistant professor of the practice at Duke University. Her research interests include statistics pedagogy, spatial statistics, small-area estimation, and survey and public health data. She is a co-author of *OpenIntro Statistics* and a contributing member of the OpenIntro project, whose mission is to make educational products that are open-licensed, transparent, and help lower barriers to education.

Kari Lock Morgan earned her PhD in statistics from Harvard University and is an assistant professor of the practice in the department of statistical science at Duke University. Her primary interests are causal inference and statistics education. She is coauthor of the book *Statistics: Unlocking the Power of Data*.

Dalene Stangl is professor of the practice of statistical science and public policy and associate chair of the department of statistical science at Duke University in North Carolina. She has served in editorial positions for the *Journal of the American Statistical Association*, *The American Statistician*, and *Bayesian Analysis* and has co-edited two books with Donald Berry, *Bayesian Biostatistics* and *Meta-Analysis in Medicine and Health Policy*. Her primary interest is promoting Bayesian ideas in the reform of statistics education and statistical practice.