Comparing Income across Gender among New Coders

Erik Rauer & Audrey Le Meur

5/14/2021

Abstract

In 2016, Free Code Camp conducted a survey of its new users, collecting information such as gender, last year's income, expected future income, events attended, and bootcamp participation (Larson and Yitbarek 2017). Using hypothesis testing in R, we compared last year's earnings and expected earnings across gender groups. Given that the average woman made more than the average man, we then investigated the relationship between women's income and their participation in women-centered events and bootcamps. Our analysis showed that women who participated in women-focused events and bootcamps earned more income than women who did not. Future research could evaluate the existence of a causal relationship between participation in these events and increased earnings.

Introduction

The gap between men and women's earning is a well documented phenomenon. The average woman makes 82% of the salary of the average man (Labor Statistics 2019). A similar trend occurs among computer programmers, where women earn on average 90.7% the income of men (Labor Statistics 2019). Given these statistics, we hypothesize that the new female users in the Free Code Camp survey will have both a lower income and expected earnings than those identifying as male.

Method

Data & Cleaning

The set of data we used for our analysis came from a survey conducted in 2016 by Quincy Larson, the creator of freecodecamp.org, and Saron Yitbarek, the creator of codenewbie.org, two online websites created to teach beginners how to code (Larson 2019). The survey's goal was to get data on a wide range of demographic and socioeconomic questions from new coders. With this data, Larson and Yitbarek hoped to better understand the end goals of those beginning to learn how to code, in the hopes that they can better meet their users' needs and to help "understand... the global movement toward coding" (Larson 2019).

Larson and Yitbarek designed the survey to get as many responses as possible while still asking a lot of questions. To do so, they took a variety of precautions. For one, they made the survey completely anonymous. This ensures that those responding would answer more accurately, since they would not have to worry about answers being tracked back to them. Another precaution taken was to make all questions optional, allowing users to skip any they did not want to answer. While this means that some questions might have less answers, it raises the overall response rate by preventing users from being stuck on one question they don't want to answer and not submitting any answers because of it. In addition to these precautions, the survey was kept as short as possible, preventing users from getting bored and not finishing. To assist in this attempt, Larson and Yitbarek would only display certain questions, if a previous one was answered in a certain way. For example, if a user indicated that they had attended a coding boot camp, they might then be asked which one they attended. Users who responded that they had not been to any boot camps would not see this second

question. Finally, Larson and Yitbarek had professional data scientists take a look at the survey and critique the questions, offering advice on how to better word them to minimize bias.

All these precautions proved successful, resulting in the survey receiving more than 15,000 responses. However, all this data was not perfect and had to be cleaned up quite a bit. To start, the survey data was split into two parts. The first consisted of questions related to the past experience, current employment information, and future goals of the participants. The second part was made of questions regarding the demographic and socioeconomic status of the participants. The data from these two parts had to be combined into one large dataset, associating each user's responses to the first part with those of the second. In addition to this combination, various other steps were done to clean the data. First, obvious outliers were removed. Since this survey was open to the general public, there existed a variety of responses that were clearly not honest, such as a user who reported an income of $20,000,000. Several questions allowed users to respond with ranges of values, such as "200-210". These ranges were replaced with their average in order to make analysis of the data simpler. Additionally, some questions, such as "How long have you been coding for?", permitted responses in terms of months or years. Any response that was measured in years was converted to months, allowing the category to simply be in numbers without messing with the scales of the responses. Finally, text answers were normalized to make similar answers the same. For example, the responses "Back-end Web Developer" and "back end web developer", while being the same answer, might be counted differently in analysis of the data. To prevent this from occurring, they were normalized into one answer, for example both responses might have been changed to be "Back-End Web Developer" instead.

Investigation

We decided to take a closer look at the result of two specific questions from the new coder survey. The first is the question "What's your gender?" which had five possible responses: female, male, agender, trans, and genderqueer. We will compare how these five different groups responded to the following two questions: "About how much money did you make last year (in US dollars)?" and "About how much money do you expect to earn per year at your first developer job (in US dollars)?". We will generally refer to the responses to these two questions as the person's income and their expected earnings, respectively. In doing these comparisons we hope to see whether the general trends regarding gender and income are reflected in new coders.

After analyzing the results of these comparisons, we took a closer look at the results of users identifying as female. Specifically, we examine responses to the questions "Which types of in-person coding events have you attended?" and "Which full-time coding bootcamp have you attended?". From the various possible answers to these two questions we identified 8 boot camps and events focused on helping women learn to code. We then split all female-identifying responders into two categories: those who have attended a women focused coding event and those who have not attended one. From there, we compared both the current income and the expected earnings of the respondents to see whether attending a women focused coding event makes a significant difference in these values. Finally, we compare the income and expected earnings of women who attended a women focused coding event with the income and expected earnings of men who attended a coding event, as well as the income and expected earnings of women who did not attend a coding event with men who did not either. In doing these comparisons, we hope to determine how helpful these women focused events are.

Data Analysis

Comparing last year's earnings across gender

Examining Figure 1 does not immediately reveal any major differences among the reported incomes of the various genders. Agender-identifying responders appear to mostly have reported a fairly low income, most making below $40,000 with one outlier at the $100,000 mark. The distributions for those identifying as male

and female are relatively similar, with almost identical ranges and medians. Performing an ANOVA test gives a p-value of 0.00988, indicating that there might exist some significant difference in the average income among gender groups. However, running Tukey's Honest Significance Test gives the results in Table 1. As can be seen, none of the p-values are below our alpha value of 0.05, suggesting that there is no significant difference between the means of the various distributions. These results indicate that our hypothesis was wrong, the women who participated in this survey did not, on average, have a significantly lower income than the men. In fact, there was no significant difference in the average income of any genders reported in the survey.

Figure 1: Last Year's Income by Gender

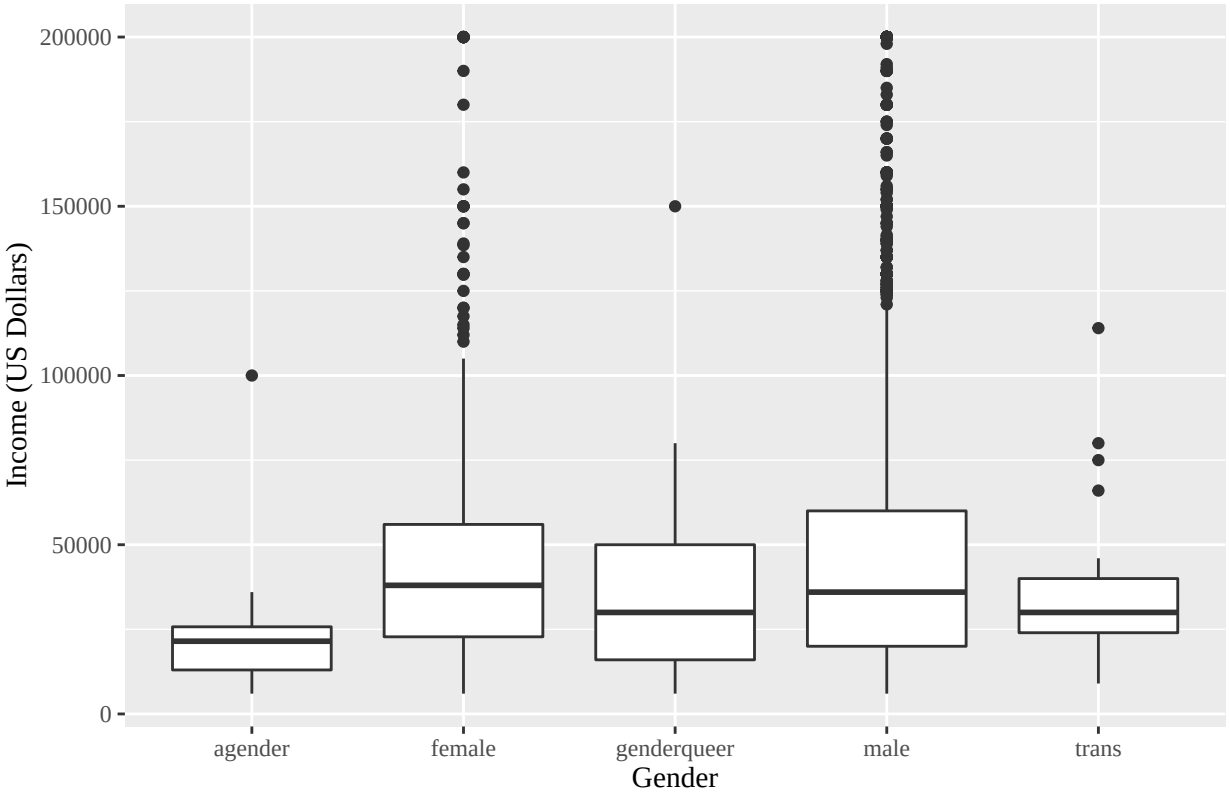


Table 1: Tukey Results for Previous Year's Income by Gender

Comparison	Estimate	Lower Bound	Upper Bound	P-Value
genderqueer-agender	11270.556	-19170.1401	41711.252	0.8508049
trans-agender	12678.571	-20794.4355	46151.578	0.8398987
female-agender	17947.025	-8107.7542	44001.805	0.3285178
male-agender	20485.001	-5474.2716	46444.274	0.1979513
trans-genderqueer	1408.015	-25097.5979	27913.629	0.9999010
female-genderqueer	6676.469	-9477.6906	22830.629	0.7919334
male-genderqueer	9214.445	-6785.2169	25214.107	0.5159178
female-trans	5268.454	-16056.7213	26593.629	0.9620218
male-trans	7806.430	-13401.9508	29014.810	0.8535115
male-female	2537.976	-326.4174	5402.369	0.1106883

Comparing expected earnings across gender

Similar to the distributions of income, examining Figure 2 does not immediately suggest major differences in the expected earnings of the various genders. While they appeared to have a relatively low income, the same cannot be said of the expected earnings for users who identify as agender, whose distribution is fairly similar to those of the other genders. The distributions of those identifying as male and female are not as similar as they were for income. While the median of men appears to be below that of women, there seems to be more men who expect to earn a six figure income than women who expect to do so. Performing an ANOVA test results in a p-value of 0.0000135, which indicates that there may be a difference between average expected income across gender groups. Running Tukey's Honest Significance Test presents the results displayed in Table 2. Of the various distributions, there is only a significant difference between the means of the users identifying as male and those identifying as female. In fact, the women who took this survey had a significantly higher expected earning than the men did, which is the complete opposite of what we had hypothesized.

Figure 2: Expected Earnings by Gender

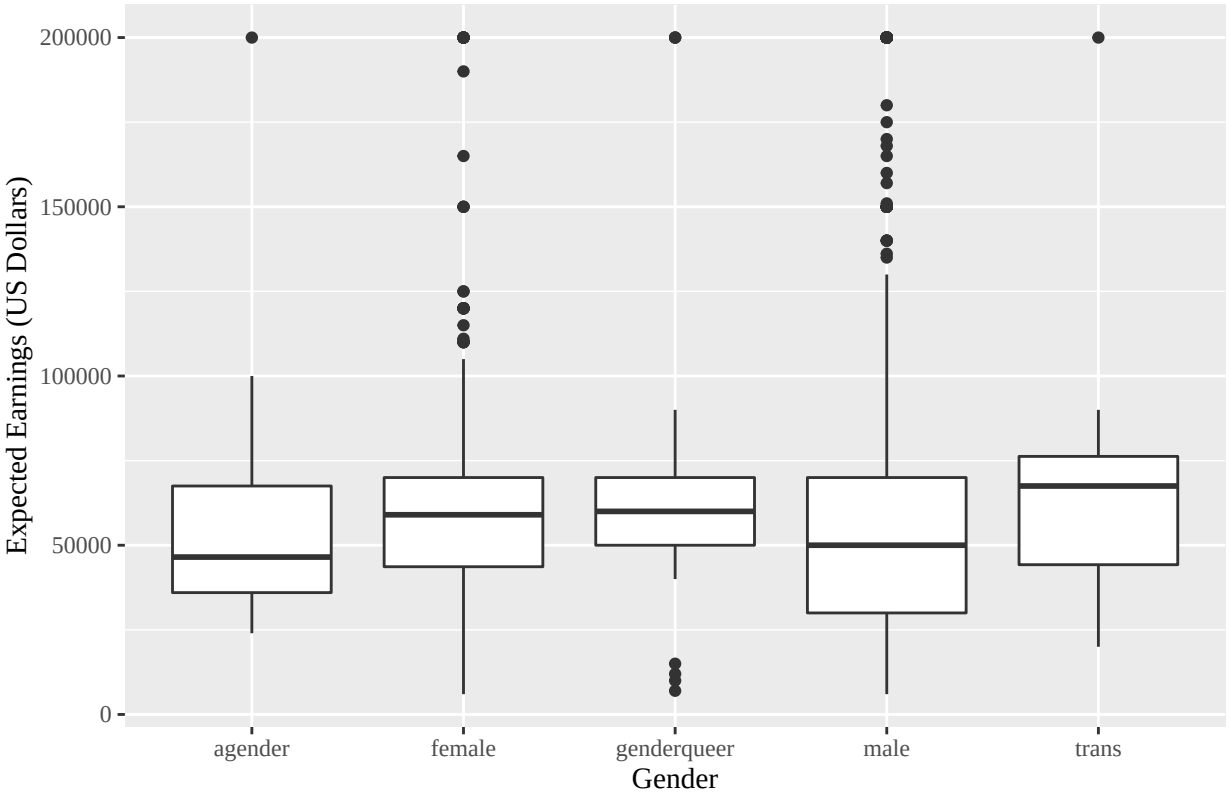


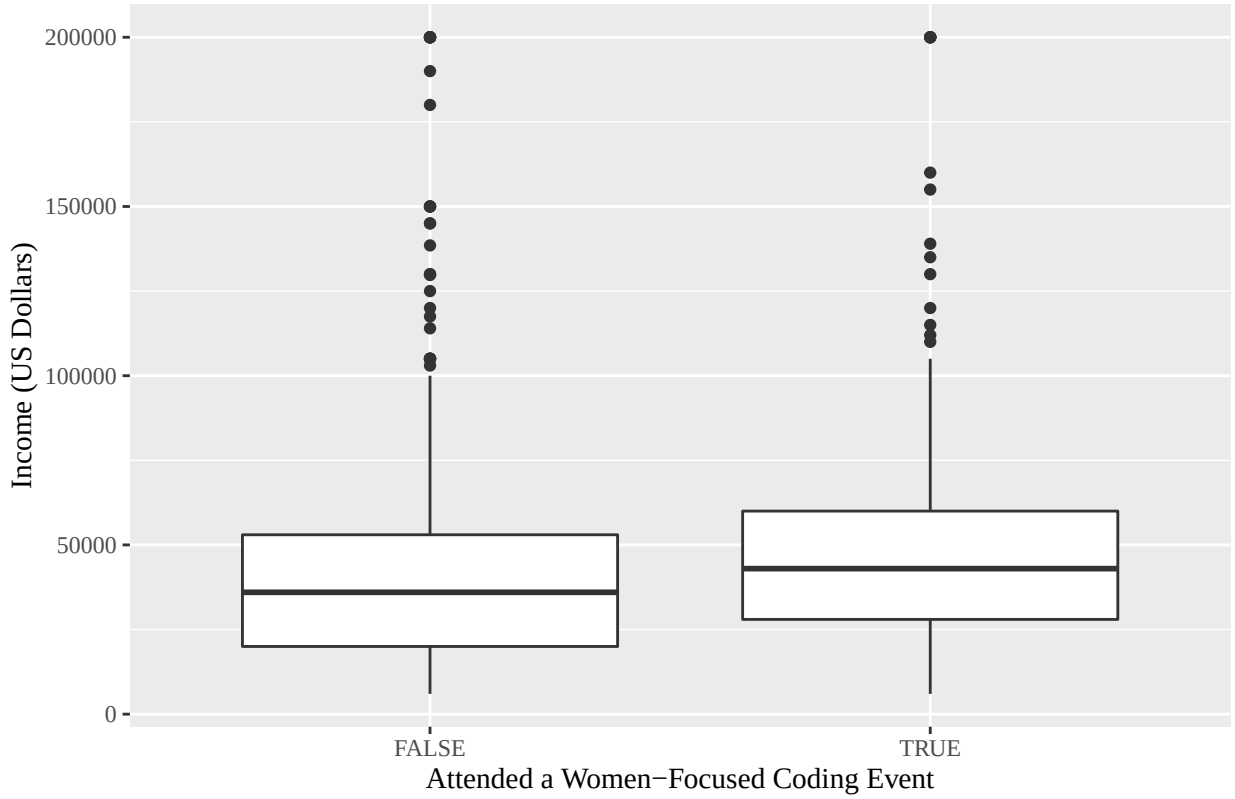
Table 2: Tukey Results for Expected Earnings by Gender

Comparison	Estimate	Lower Bound	Upper Bound	P-Value
female-male	3999.8767	1438.9711	6560.782	0.0002000
agender-male	5607.0316	-13388.0023	24602.066	0.9290263
genderqueer-male	14350.3266	-635.1497	29335.803	0.0681207
trans-male	14616.6844	-5525.5804	34758.949	0.2758067
agender-female	1607.1549	-17475.2770	20689.587	0.9993832
genderqueer-female	10350.4500	-4745.6553	25446.555	0.3332412
trans-female	10616.8077	-9607.8981	30841.514	0.6066513
genderqueer-agender	8743.2950	-15384.3959	32870.986	0.8604822

Comparison	Estimate	Lower Bound	Upper Bound	P-Value
trans-agender	9009.6528	-18618.0833	36637.389	0.9007574
trans-genderqueer	266.3578	-24774.5086	25307.224	0.9999998

Comparing income and expected earnings on whether or not they attended a female coding event

Figure 3: Previous Year's Income by Whether a Women's Coding Event was Attende

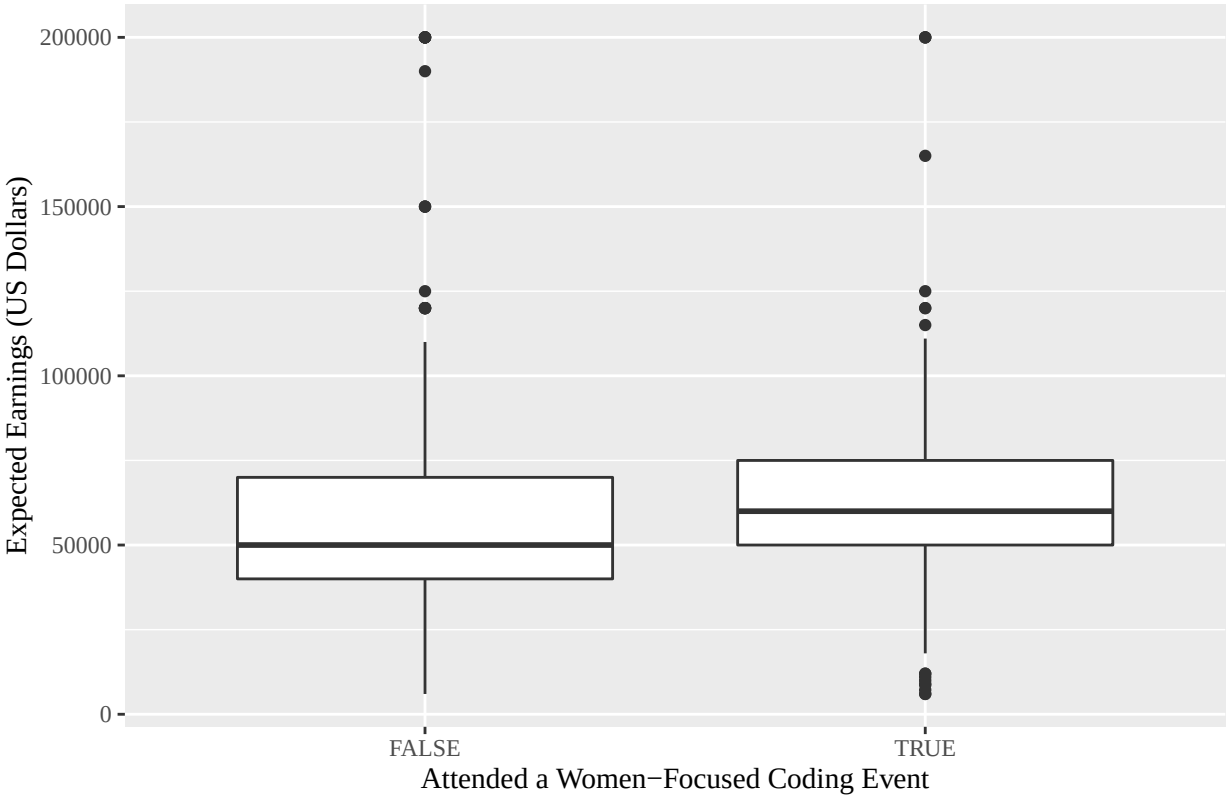


Comparing the distribution of the income of women that attended a women-focused coding event with the distribution of the income of women who did not attend such an event shows some differences between the two. Primarily of note is that the first, second, and third quartiles of the distribution of women who attended an event appear to be roughly $10,000 higher than those who didn't. Performing a student's t-test confirms what these observations seem to show: the average women who attended a female focused event had an income of $8,000 more than those who didn't, which is a significantly higher amount (p-values of 5.482x10-6). This suggests two possibilities: either women who have a higher income tend to be the ones who go to these women focused coding events, or alternatively, these coding events do help increase the income of their attendees, likely by increasing the attendee's skillset and thus their quality in the eyes of employers.

Similar to the distributions of previous year's income, Figure 4 shows a difference in the distribution of expected earnings of women who attended women focused coding events and women who did not. In fact, just like the distributions of income, the first, second and third quartiles of the distribution of expected earnings of women who attended a coding event appear to be $10,000 higher than their counterparts in the distribution of women who did not attend an event. Once again, a student's t-test indicates that there is a significant difference in the average of these two distributions resulting in a p-value of 2.399x10-11. On average, women who attended a women focused coding event expected to earn an income $10,000 more than

those who didn't. Once again these results suggest two possibilities. Either the women who attend these coding events have a higher confidence in their abilities and the worth of those abilities, or these events lead to their attendees becoming more self-assured in their coding skills, thus increasing their self-estimation of their worth.

Figure 4: Expected Earnings by Whether a Female Coding Event was Attended



Conclusions

Possible Future Work

References

Labor Statistics, US Bureau of. 2019. *Highlights of Women's Earnings in 2019.*

Larson, Quincy. 2019. "How We Crafted a Survey for Thousands of People Who Are Learning to Code." *freeCodeCamp.org.* Free Code Camp. https://www.freecodecamp.org/news/we-just-launched-the-biggest-ever-survey-of-people-learning-to-code-cac81dadf1ea/#.8g9ts8gm5.

Larson, Quincy, and Saron Yitbarek. 2017. "FreeCodeCamp/2016-New-Coder-Survey." *GitHub.* Free Code Camp. https://github.com/freeCodeCamp/2016-new-coder-survey/tree/c7eaf6b3da8e874e94d71960b6812a3e0ced0704.