MSAN 628 - Homework 1

Andre Guimaraes Duarte

March 24, 2017

Reading Questions

1. The biggest issue facing the field of "data science", in my opinion, is the fact that it is currently very ill-defined. Depending on whom you ask, their definition of the term and/or field may vary drastically. Some may think of "data science" as a glorification of "statistics", while other may define it as working with specific tools and programs, and others yet may imagine it as a way to get insights from any and all types of information available. This misunderstanding leads to the data scientist role to be vastly different across companies and even within companies.

   I believe that this issue will slowly fix itself, as more data scientists emerge in the market and themselves define what they do. I think different branches within data science will appear, which will have a clearer indication of the specificities of each role within the field of data science.

2. In my opinion, the biggest difference between traditional statistics and data science is that data science is broader in terms of tools. For example, we can think of machine learning, or NLP as "things" in data science, but they are not part of statistics per se. Statistics continue to be a major component of data science, and it is helped by the fact that we often deal with "big data" (which makes statistical tests more powerful for example). Technology plays a role in this difference mainly due to the fact that we are not obtaining "big data" thanks to tools that did not exist even 10-20 years ago.

3. DJ Patil and Jeff Hammerbacher (then at LinkedIn and Facebook, respectively) coined the term "data scientist" in 2008. This is considered the date when the "data scientist" role appeared officially. However, William Cleveland wrote a position paper about data science called "Data Science: An action plan to expand the field of statistics" in 2001, so this term ("data science") existed at least as early as then.

4. In my opinion, data science in academia is still a very nascent thing. Only now are we starting to see professors and researchers calling themselves "data scientists", while this role has existed in the industry (officially and unofficially) for many years now. I think data science in academia is more focused on theoretical analyses and finding novel tools and algorithms. In the industry, there is a lot of effort towards this same goal of advancing the field. However, data scientists in the industry also need to be data engineers and understand how to gather, clean, store the data, find the most efficient solutions, then analyze the data in many ways to finally get results and drive business decisions.

5. I think Mike Driscoll's answer from 2010 is very good at explaining what data science is, as the "civil engineering of data". Data scientists know the tools and materials to use, and must have an incredible knowledge of the theoretical background for all of them. A data scientist is someone who can not only get, store, and analyze data, but also understand why they are doing those things and demystify the sometimes cryptic results.

Quantitative Questions

1. (a) The event where both E and G, but not F , occur can be written $E \cap G \cap F^c$.

   (b) The event where at least two of the events occur can be written $(E \cap F) \cup (E \cap G) \cup (F \cap G)$.

   (c) The event where at most one of the events occurs can be written $(E \cap F^c \cap G^c) \cup (E^c \cap F \cap G^c) \cup (E^c \cap F^c \cap G) \cup (E^c \cap F^c \cap G^c)$.

2. $$\begin{aligned}(E\cup F)\cap(E^c\cup F)\cap(E\cup F^c) &= ((E\cup F)\cap E^c)\cup((E\cup F)\cap F)\cap(E\cup F^c)\\ &= (E\cap E^c)\cup(F\cap E^c)\cup F\cap(E\cup F^c)\\ &= (F\cap E)\cup(F\cap F^c)\\ &= E\cap F\end{aligned}$$

3. E and $\emptyset$ are disjoint ($P(E\cap\emptyset)=0$). Therefore, we have $P(E)=P(E\cup\emptyset)=P(E)+P(\emptyset)$. Therefore, we get $P(\emptyset)=0$.

   E and E^c are disjoint ($P(E\cap E^c)=0$). Also, we have $S=E\cup E^c$. Therefore, we get $P(S)=P(E\cup E^c)=P(E)+P(E^c)=1$ using the first and second axioms. Rearranging, we get $P(E^c)=1-P(E)$.

4. (a) We have $P(S)=1$, $P(\emptyset)=0$, $P(\{H\})=p\in(0,1)$, and $P(\{T\})=1-p\in(0,1)$. So axiom 1 is verified.

   We have $P(S)=1$. So axiom 2 is verified.

   We know that the events $\{H\}$ and $\{T\}$ are disjoint. In addition, $\{H\}\cup\{T\}=S$. We have $P(\{H\}\cup\{T\})=P(S)=1=p+1-p=P(\{H\})+P(\{T\})$. Also, since $P(\emptyset)=0$ and $\emptyset$ and all other events are disjoint, axiom 3 is verified.

   (b) If $P(\{T\})=1-\frac{p}{2}$, axioms 2 and 3 are not verified because $P(S)=P(\{H\}\cup\{T\})=P(\{H\})+P(\{T\})=p+1-\frac{p}{2}\geq 1$.

5. (a) The probability that she passes all three exams is $P(F)\cap P(S|F)\cap P(T|F\cap S)=0.9\times0.8\times0.7=0.504$.

   (b) The conditional probability she failed the 2nd exam, if we know that she does not pass all three exams is $P(S^c|F^c\cup S^c\cup T^c)=\frac{1\times P(S^c)}{P(F^c\cup S^c\cup T^c)}=\frac{0.9\times0.2}{1\times0.1+0.9\times0.2+0.9*0.8\times0.3}\approx 0.363$.

6. (a) There is only one possibility out of 36 that the sum is 6 and the first roll is 4, so $P(E_1\cap F)=\frac{1}{36}$. In addition, $P(F)=\frac{1}{6}$ and $P(E_1)=\frac{5}{36}$, so $P(E_1)\times P(F)\neq P(E_1\cap F)$. In other words, E_1 and F are not independent.

   (b) There is only one possibility out of 36 that the sum is 7 and the first roll is 4, so $P(E_2\cap F)=\frac{1}{36}$. In addition, $P(F)=\frac{1}{6}$ and $P(E_2)=\frac{6}{36}=\frac{1}{6}$, so $P(E_2)\times P(F)=P(E_2\cap F)$. In other words, E_2 and F are independent.

7. We have $P(\textit{cancer})=0.008$, $P(\textit{positive}|\textit{cancer})=0.9$, $P(\textit{positive}|\textit{not cancer})=0.07$. So $P(\textit{not cancer})=1-P(\textit{cancer})=0.992$. In addition, we have $P(\textit{positive})=P(\textit{positive}|\textit{cancer})\times P(\textit{cancer})+P(\textit{positive}|\textit{notcancer})\times P(\textit{not cancer})$. Therefore, we get $P(\textit{cancer}|\textit{positive})=\frac{P(\textit{positive}|\textit{cancer})\times P(\textit{cancer})}{P(\textit{positive})}=\frac{0.9\times0.008}{0.9\times0.008+0.07\times0.992}\approx 0.093$.

Computational Questions

1. (a) We use a bootstrap method to generate a distribution of the probability of winning a Monty Hall type problem by using the strategy of switching doors after the first one is opened. To calculate the probability of winning, I run the Monty Hall simulation 1,000 times (so the probability of winning is roughly equivalent to the number of wins over the total simulations, which in this case is 1,000). To generate the distribution, I run this process 1,000 times (so I get 1,000 probabilities). The resulting histogram is shown in Figure 1. We can see that our best guess for our probability of winning is around 66.6% ($\frac{2}{3}$).

   (b) The probability of choosing the correct door at the beginning of the problem is $\frac{1}{3}$, since we choose at random and there are 3 doors. Let's say we choose Door A. Then, the probability that the prize is behind Doors B or C is $P(B\cup C)=\frac{2}{3}$. Once one Door is opened, (which will be one of B and C, let's say C), the probability that the prize is behind Door A continues to be $P(A)=\frac{1}{3}$. Since events B and C are disjoint (the prize is either behind Door B or Door C), we have $P(B\cup C)=P(B)+P(C)=\frac{2}{3}$. But we know that the prize is not behind Door C! So $P(C)=0$. Therefore, we now have $P(B)=\frac{2}{3}$.

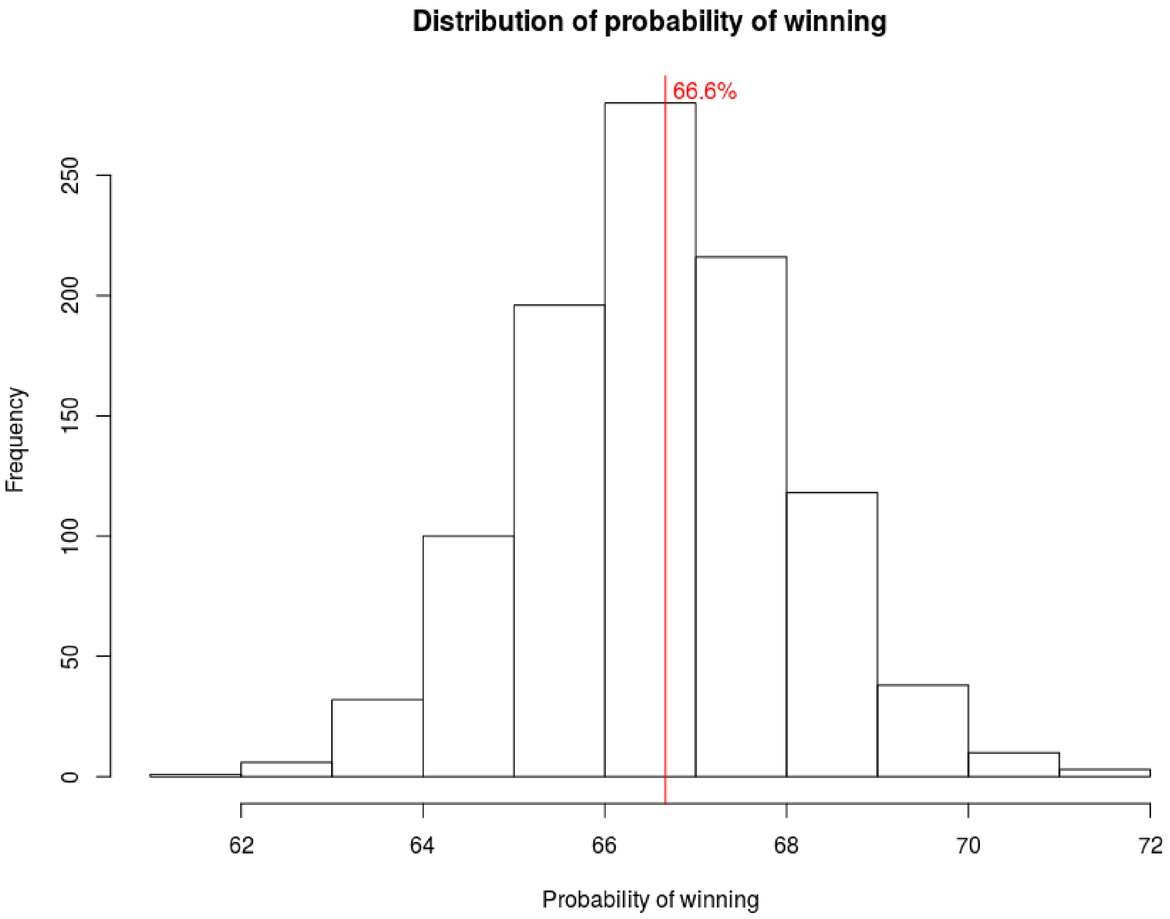

Figure 1: Monty Hall probability of winning by switching doors using bootstrap.

2. (a) We use a bootstrap method to generate a distribution of the probability of at least two people sharing a birthday. To calculate this probability, I run the simulation 1,000 times (so the probability of at least two people sharing a birthday is roughly equivalent to the number of times there is a duplicate date in the sample divided by the number of total simulations, which is 1,000). To generate the distribution, I run this process 1,000 times (so I get 1,000 probabilities). The resulting histogram is shown in Figure 2 for n=30 students. We can see that our best guess for at least two people sharing the same birthday is around 70.5%.

   (b) Generalizing this simulation for any class size, and plotting all the histograms for $n = 1 \dots 30$, we see that the minimum class size required for there to be a probability of 50% or higher of having at least two students with the same birthday is n=23. The corresponding histogram is shown in Figure 3.

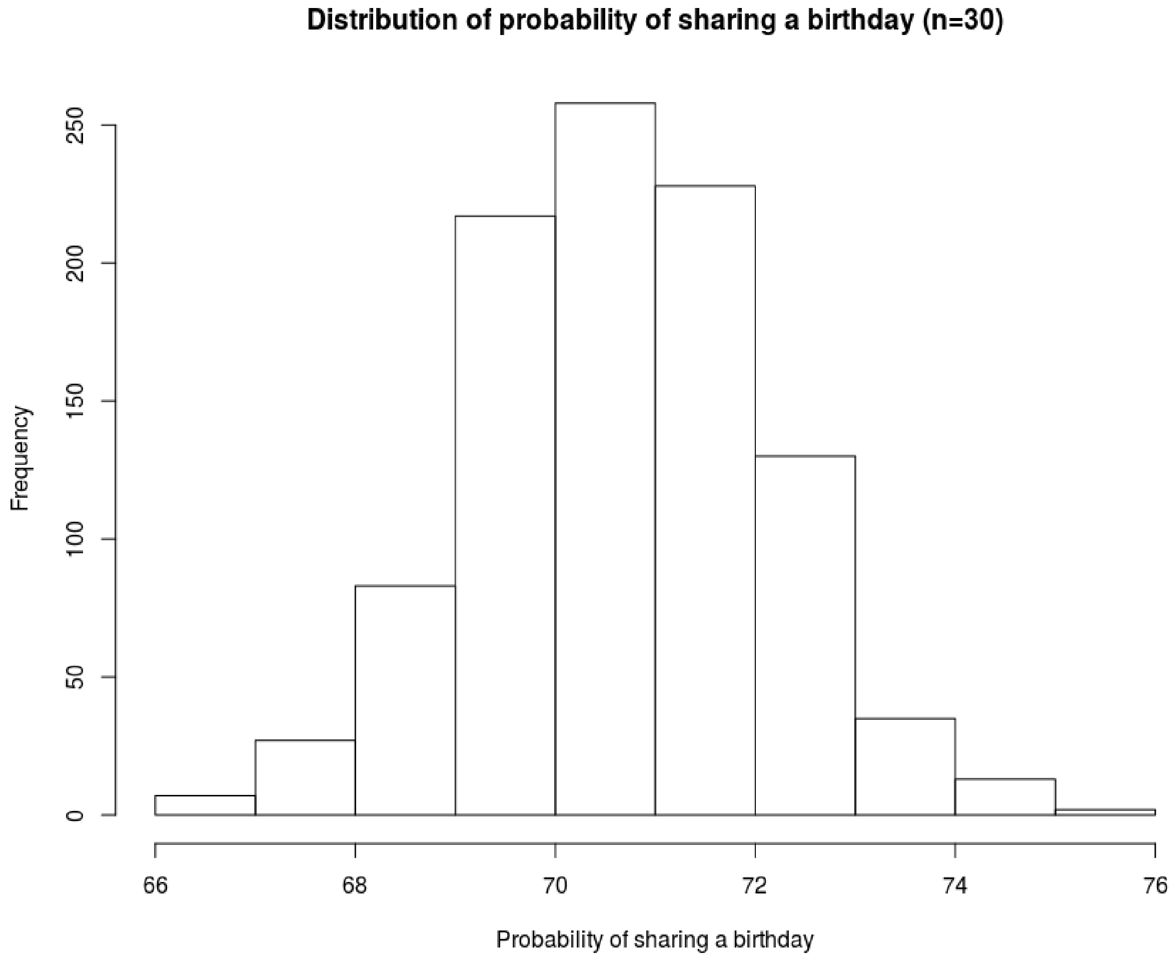

Figure 2: Probability of at least two students (n=30) sharing a birthday using bootstrap.

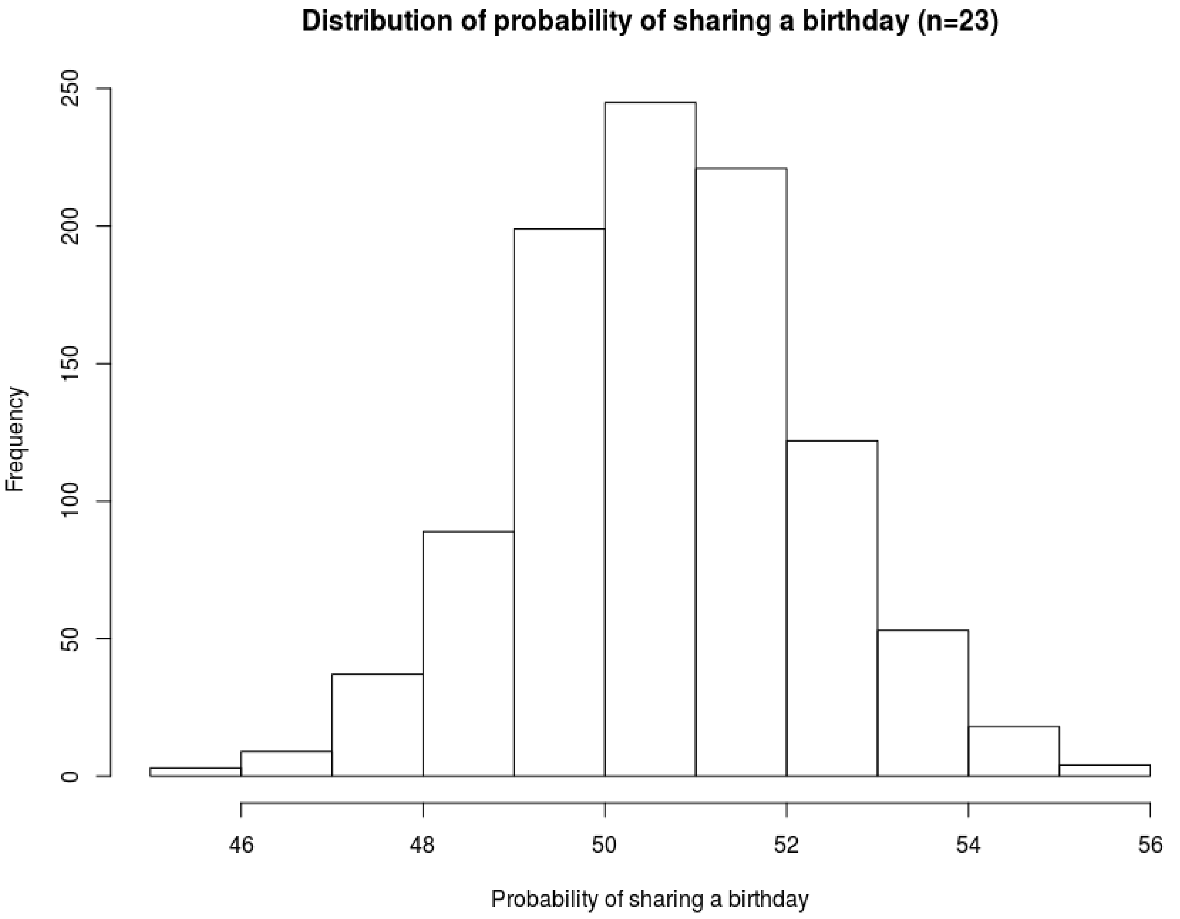

Figure 3: Probability of at least two students (n=23) sharing a birthday using bootstrap.