Yury Park
CS7633 – Human-Robot Interaction

Written Case Study #1

I. Introduction

This case study is centered around an experiment involving a group of older adults with hypertension and the regularity with which they may be induced to take their blood pressure. Given the two main research questions that the insurance company wants answered, the corresponding hypotheses to be tested are as follows:

1. The people using the mobile robot take their blood pressure more regularly than those using a stationary reminder alarm.
2. Adding social capabilities to the mobile robot increases the regularity with which people take their blood pressure.

There is also arguably a third hypothesis which, although not directly addressed in the two given research questions, is nonetheless implicit, namely:

3. The people using the socially capable robot (regardless of mobility) take their blood pressure more regularly than those using the non-socially capable counterpart.

The null hypothesis, in turn, would be as follows:

1. The people using the mobile robot take their blood pressure with the same regularity as those using a stationary reminder alarm.
2. Adding social capabilities to the mobile robot has no impact on the regularity with which people take their blood pressure.
3. The people using the socially capable robot take their blood pressure with the same regularity as those using the non-socially capable counterpart.

Regarding the experiment, we will first interpret its results assuming that it had been perfectly designed and executed, before going on to point out any potential issues with the experiment and how they may impact the interpretation of the results.

II. Analysis (Assuming a Perfect Experiment)

According to the gathered data and the ANOVA summary, we can observe that the p-values for both the first and second independent variables (labeled A and B, where A = mobile / not mobile, and B = social / not social) are well below the common threshold value of 0.05; in a vacuum, this would indicate that the null hypotheses #1 and #3 as noted above should be rejected. The results also show that there likely exists an interaction between A and B, as the p-value of 0.002295 for A x B is also well below the threshold of 0.05. The existence of an interaction in this case means that the effect of variable A on the regularity with which people take blood pressure is not the same across both levels of variable B (or vice versa). In other words, the data suggests that either or both of the following conclusions may be true: 1) adding or not adding social capabilities a robot makes a difference in terms of the impact that the robot's mobility (or lack thereof) has on people's regularity of taking blood pressure, and/or 2) adding or not adding mobility to a robot makes a difference in terms of the impact that the robot's social capability (or lack thereof) has on people's regularity of taking blood pressure. The ANOVA table does not, however, tell us which of these two conclusions is the correct one; what it tells us is that a statistically significant interaction exists between A and B.

Next, observing the data pertaining to the dependent variable (i.e. the total number of times in a given week that a person took his/her blood pressure, with a range of 0 thru 21), the tables indicate that using a socially capable, mobile robot fared best, with a mean of 15.6. Coming in at a close second was a non-social mobile robot with a mean of 15. Both of the non-mobile robots performed worse, with the non-social robot managing a mean of 11.3 and the social robot, somewhat surprisingly, coming in at dead last with a mean of 3.9.

From this information, we may be tempted to conclude that adding mobility always has a positive impact on people's willingness to take blood pressure measurements regularly, while social capability may have a beneficial or detrimental impact depending, respectively, on whether it is implemented on a mobile or non-mobile robot. However, it would at best be premature to jump to such conclusions. First, as mentioned previously, we do not yet know the precise nature of the interaction between variables A and B; we only know that an interaction exists, and thus we cannot even reject the null hypothesis #2 based on this finding alone (for example, it may turn out that mobility is the one affecting the social capability's impact, and not the other way around). Second, even if we could reject one or more null hypotheses right away, doing so does not automatically equate to proving the alternate hypotheses (for example, rejecting null hypothesis #1 only implies that people using the mobile robot do *not* take their blood pressure with the same regularity as those using a stationary reminder alarm; it does not prove that the former group takes the blood pressure *more* regularly), and a post-hoc analysis would be useful in this regard. Third, although the ANOVA table indicates that the interaction as well as the main effects (A and B individually) are statistically significant, it may in fact be the case that the interaction is significant and dominates the main effects, or the interaction itself is dominated by the main effects.[1] That is, the fact that significance was found across the board in this case makes it difficult to interpret the main effects in isolation, and we would be well-advised to re-examine the main effects (e.g. through a post-hoc analysis) to see if they are worthy of interpretation in their own right or whether the interaction alone is the overriding factor.

III. Critique of the Experiment Design and Execution

The analysis thus far has been based on the assumption that the experiment is perfect in every way. In fact, there are a number of potentially serious issues with the experiment design and execution that may have compromised the integrity of the data and violated certain key assumptions that a valid 2-way ANOVA is expected to satisfy.

One of the most striking oversights involving the experiment is that the same group of 10 adults is used for all of the four different experiment types over the 4-week period. Not only does this imply a complete lack of a control group, but it also violates a key assumption of a valid 2-way ANOVA: independence of observations, i.e. that there be no relationship between the observations in each group or between the groups. Because the current experiment fails this assumption, the results are

[1] In his book, Design and Analysis: A Researcher's Handbook, 4th ed. (2004), Keppel states that, in interpreting a two-way design, there are four basic outcomes: 1) nothing is significant; 2) the interaction is not significant, but at least one main effect is significant; 3) the interaction is significant, but is dominated by the main effects; or 4) the interaction is significant and dominates the main effects.

likely invalid, and a different statistical test (such as a repeated measures design) may need to be used instead of the 2-way ANOVA.

The procedure via which the 10 adults were chosen for the experiment is also problematic. The research questions to which the health insurance company seeks answers refer to older people with hypertension in general; and yet, the interns began the recruiting process by specifically selecting a pool of 100 adults all of whom have previously participated in a robotics study. This violates another assumption of the 2-way ANOVA: that the groups be sampled randomly. More concretely, this may skew the quality of the results, for instance by dampening the role of the novelty effect especially for the first week (this effect may be mitigated by a training or familiarization period, but this study does not address whether such a period was provided to the subjects). As HRI is a relatively new and emerging field and most people are not accustomed to the presence of robots in their everyday lives, the impact of having taken part in a previous robotics study is probably non-negligible and may be a hidden independent variable that needs to be accounted for. Furthermore, gender (also not accounted for in the experiment) may be hidden independent variable as well.

Another byproduct of using the same group of 10 people throughout the experiment that makes this experiment likely inappropriate for a 2-way ANOVA is that it is particularly difficult to interpret the results meaningfully by isolating the effects of the various robot interactions from the effects of maturation (e.g. people may develop a growing familiarity or bond over time with one type of robot, and when the robot is switched with another type, may react to the change behaviorally to a greater (or lesser) extent than they would have had they begun the experiment with that latter robot) and of other possible social interventions over time (e.g. friends or family members preferring one robot or another, or making comments that may introduce bias into the subjects).

This brings us to another issue with the experiment: the particular fixed order in which the different types of robots were introduced to the people each week. What if, for example, the robotic head mounted on a non-mobile device were introduced in the first week instead of the fourth? Is the reason that performance (in terms of the regularity of blood pressure measurement) took a nosedive in the fourth week due solely to the social / mobile capabilities of the robot in a vacuum, or can it be attributed to the fact that people were introduced to this robot after having spent considerable time with the other varieties? One may theorize that, after having formed a bond with the social mobile robot during the second week (which received an "upgrade" of talking / smiling capabilities after the first week), and after having internalized the non-mobile, beeping device as just an inanimate object during the third week, some of the subjects may have been put off by seeing the smiling robot's head suddenly decapitated and placed on that same non-mobile body in the final week. There may be other plausible theories and explanations; the point is, the experiment as currently conducted is utterly powerless to refute or confirm these alternate theories. After all, there isn't even a post-experiment survey to gather the subjective experiences of these people, let alone sets of studies implementing the different permutations of the order by which the robots are introduced.

Finally, the experiment does not provide a tangible measurement of the volumes at which the robots spoke or beeped, only noting that the robots in the 3rd and 4th weeks respectively beeped and spoke "loudly", and making no such mention regarding robots in the first two weeks. As we do not know whether some of the participants may be hard of hearing, the precise volume of the robot's sounds from week to week may be relevant to the interpretation of the data. In the same vein, the experiment does not provide any information regarding the distance between a robot and each participant at the moment the robot spoke or beeped, or the placement of walls or other obstacles between them, or what the participant was doing at the time (e.g. napping, watching TV, etc.), all of which may impact the participants' responsiveness to the robot.

The manifold issues with the experiment as outlined in this section, along with the points raised in section II, mean that there is little, if any, in the way of conclusive findings that can be drawn from the data.

IV. Further Discussion and Future Social Capabilities

The insurance company states it is interested in learning more about the various social capabilities that may be integrated into the robot to improve results. First and foremost, however, it is important to consider whether even adding further social capabilities would be a wise decision from a cost-benefit perspective. The goal or intention of the robot in the present case is simple: to induce people to take their blood pressure measurements on time. In a sense, the robot is little more than a glorified alarm clock that is further constrained by its subject matter (it is not even meant to serve as a general reminder device, e.g. for appointment reminders). Would such a narrowly tailored robotic device even be attractive to prospective consumers, whether they are individuals or health clinics? The company would need to consider whether it may be a better idea to generalize or expand the robot's set of intended functions before implementing more advanced social capabilities.

That said, one suggestion for improving the current mobile robot's social capacity and function would be to endow it with the ability to grasp and hand over the blood pressure device to the human along with its spoken reminder – or perhaps simpler yet, equip the robot with its own blood pressure monitor, such that all the person would have to do once the robot approaches is to place his/her arm into a slot mounted on the robot's body (at an intuitive and convenient spot), allowing it to take the measurement automatically and display and/or speak out the results. One may conjecture that a major reason people "forget" to regularly take their blood pressure may not merely be the lack of a reminder, but rather the inconvenience of having to do it themselves, esp. when they're in the middle of doing something else, or perhaps just plain laziness. (This conjecture could be tested in another experiment, one that is properly designed of course.)

This improvement could then be further amplified by implementing certain head and/or eye movements to gaze toward the direction of the blood pressure monitor at the appropriate time, to simulate shared attention or turn-taking.[2] It is already well-established that humans tend to treat computers as they treat other people,[3] and implementing gaze in such a manner may help further elicit engagement, trust, and compliance from the human.

We will end by mentioning a number of other improvements meriting consideration and additional research. First, it has been demonstrated that people tend to view robots that display a range of

---

[2] In "Meet Me Where I'm Gazing: How Shared Attention Gaze Affects Human-Robot Handover Timing" (2014), Moon et al. found that a robot's gaze at the right place and time can help improve the efficiency and/or subjective experience of human-robot handovers.

[3] See, e.g. Fong, et al. "A Survey of Socially Interactive Robots" (2003); A. Paiva, "Affective Interactions: Towards a New Generation of Computer Interfaces" (2000)

emotions (in terms of facial expressions, tone of voice, and/or body language) as social creatures with feelings, desires and intents, and respond to them emotionally.[4] Hence, outfitting the robot in our case with certain basic emotional outlets (such as happy facial expression, gesture and/or sound when the person takes blood pressure with regularity vs. sad expression, gesture and/or sound otherwise) may provide the resulting data an extra boost. Second, certain learning algorithms may be implemented for the robot to better facilitate interaction[5] – by, for instance, learning how many times to ask the person to take the measurements for maximum persuasive effect before it becomes counterproductive, or learning a number of different ways and combinations to repeat essentially the same message. Third, sensors and algorithms that aid in situational awareness, such as face / body detection and tracking (to help figure out whether the person is napping or otherwise indisposed, whether the person is paying attention to the robot, whether the person is actually a visiting guest / a stranger, and so on), may help in preventing the robot from committing social faux pas or other undesirable behaviors.[6] Fourth, adding the option for the robot to put in a reminder remotely via text, email or the like may aid in situations in which the person is at work or is otherwise absent from home.

---

[4] Breazeal. "Emotion and Sociable Humanoid Robots" (2002)

[5] Learning has been found to be often important for the facilitation of communication, interaction, and sharing of knowledge. See, e.g. Klingspor, Kaiser. "Human-Robot Communication and Machine Learning" (1997)

[6] See Bohus, Horvitz. "Dialog in the Open World: Platform and Applications" (2009), which lists situational awareness as a core competency of conducting interaction in the open world.