AI v.s. Expert Advice:

A Comparative Study on Confidence Levels in Crowd Estimation Size

Annie Chen

2023-12-15

Load the packages we want.

```
### Load all packages needed
library(dplyr)
library(knitr)
library(ggplot2)
```

1. Overview

This experiment aims to understand how advice from an AI tool affects people's confidence of their answer when estimating the crowd size.

In the experiment, participants are shown a picture of a dense crowd in a stadium, and asked to estimate the crowd size. Then participants are randomly assigned to one of the three groups:

- **Control**: Shown the picture again, asked if they want to change their guess, then asked confidence in their guess
- **Treatment 1**: Shown ChatGPT's advice on how to estimate crowd size, asked if they want to change their guess, next asked confidence in their guess. Then asked how similar was their strategy to ChatGPT's
- **Treatment 2**: Shown an expert's advice on how to estimate crowd size, asked if they want to change their guess, next asked confidence in their guess. Then asked how similar was their strategy to the experts'

Finally all participants are asked how familiar they are with ChatGPT.

The first step is to load the data:

```
chatgpt_data = read.csv("/Users/anniechen/Desktop/chat_gpt_experiment_data.csv", header = TRUE)
```

Looks like it loaded!

2. Calculate summary statistics

```
summary = summary(chatgpt_data)

kable(summary, digits = 2)
```

id	duration	guess1	condition	guess2	similar	confidence	use_gpt
Min. : 1.0	Min. : 35.0	Min. : 6.0	Min. :0.0000	Min. : 1.5	Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.: 476.5	1st Qu.: 144.5	1st Qu.: 811.8	1st Qu.:0.0000	1st Qu.: 900.0	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:1.000
Median : 952.0	Median : 196.0	Median : 1500.0	Median :1.0000	Median : 1595.0	Median :3.000	Median :2.000	Median :2.000
Mean : 952.0	Mean : 236.4	Mean : 7976.7	Mean :0.9942	Mean : 5252.9	Mean :2.671	Mean :2.453	Mean :2.418
3rd Qu.:1427.5	3rd Qu.: 277.0	3rd Qu.: 3000.0	3rd Qu.:2.0000	3rd Qu.: 3000.0	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:3.000
Max. :1903.0	Max. :1523.0	Max. :2545000.0	Max. :2.0000	Max. :900000.0	Max. :4.000	Max. :5.000	Max. :5.000
NA	NA	NA's :1	NA	NA's :407	NA's :642	NA	NA

There are a total of 1903 participants in this experiment. The treatment takes value 0 for control, 1 for Treatment 1 (ChatGPT), 2 for Treatment 2 (Expert). The two guesses of the crowd size, which are around 1500-1600 people, are quite similar and even close to the actual number in this picture, which is 1567 people. Moreover, an interesting finding is that people are not very confidence in their estimations, with an average of 2.453 in their confidence level.

3. Check for Balance

```
balance_table = chatgpt_data %>% group_by(condition) %>%      summarise(mean_guess1 = mean(guess1),
          mean_use = mean(use_gpt),
          mean_duration = mean(duration),
          sd_guess1 = sd(guess1),
          sd_use = sd(use_gpt),
          sd_duration = sd(duration),
          N = n(),
          prop = n()/1903
          )

kable(balance_table, digits = 2)
```

condition	mean_guess1	mean_use	mean_duration	sd_guess1	sd_use	sd_duration	N	prop
0	7307.40	2.34	206.69	52455.88	1.16	144.54	641	0.34
1	NA	2.48	241.91	NA	1.27	151.45	632	0.33
2	8984.12	2.43	261.00	107792.22	1.21	173.74	630	0.33

The data look pretty balanced across treatment arms. However, there is a missing value in treatment 1 's guess1 column.

```
chatgpt_data <- chatgpt_data %>% filter(!is.na(guess1))
```

```
  balance_table = chatgpt_data %>% group_by(condition) %>%      summarise(mean_guess1 = mean(guess1),
          mean_use = mean(use_gpt),
          mean_duration = mean(duration),
          sd_guess1 = sd(guess1),
          sd_use = sd(use_gpt),
          sd_duration = sd(duration),
          N = n(),
          prop = n()/1902
          )

kable(balance_table, digits = 2)
```

condition	mean_guess1	mean_use	mean_duration	sd_guess1	sd_use	sd_duration	N	prop
0	7307.40	2.34	206.69	52455.88	1.16	144.54	641	0.34
1	7650.66	2.48	242.05	51467.01	1.27	151.53	631	0.33
2	8984.12	2.43	261.00	107792.22	1.21	173.74	630	0.33

Now the data looks balanced regarding the pre-treatment variables. One thing we should be careful is that the standard deviations for the initial guess are quite large, which means that there are significant variability in the guesses across participants. We should consider this when process the data.

4. Hypothesis

I have two hypothesis:

- 1. AI tools might affect people's confidence in their answers when estimating the crowd size.
- 2. People who use ChatGPT more frequently (in treatment group 1) tend to believe in ChatGPT's method more; however, people who are more familiar with ChatGPT (in control group and treatment group 2) may be less confident in their answers because they tend to reply more on ChatGPT's support.

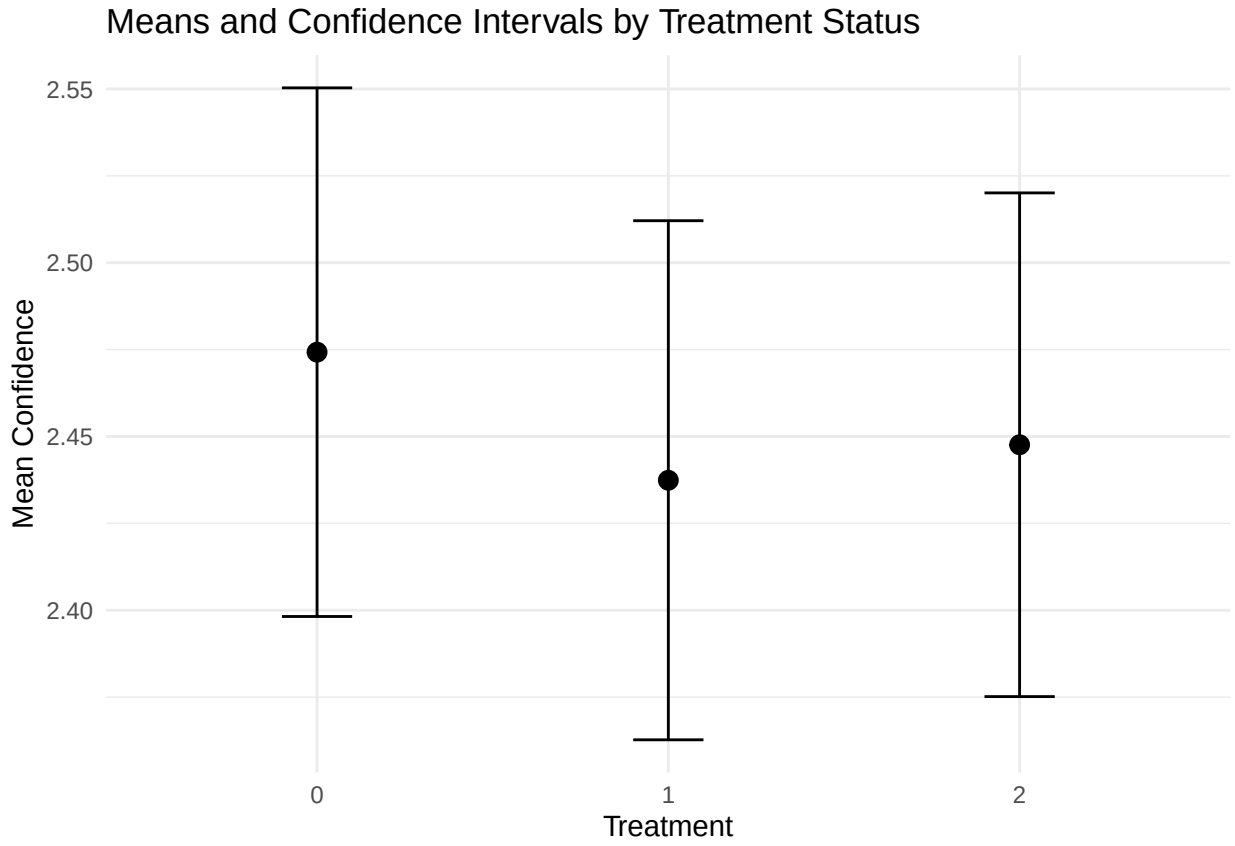
5. Plot Outcome

Plot the means and confidence intervals of the main outcome "confidence".

```
# Generating summary statistics by treatment status
summary_data <- chatgpt_data %>%
  group_by(condition) %>%
  summarise(
    mean_confidence = mean(confidence),
    sd_confidence = sd(confidence),
    n = n(),
    se_confidence = sd_confidence / sqrt(n)
  )

# Calculating 95% confidence intervals
summary_data <- summary_data %>%
  mutate(
    lower_ci = mean_confidence - 1.96 * se_confidence,  # Using Z-score for 95% CI
```

```
    upper_ci = mean_confidence + 1.96 * se_confidence    # Using Z-score for 95% CI
  )
```

```
# Plotting means and confidence intervals
ggplot(summary_data, aes(x = factor(condition), y = mean_confidence)) +
  geom_errorbar(aes(ymin = lower_ci, ymax = upper_ci), width = 0.2) +
  geom_point(size = 3) +
  labs(title = "Means and Confidence Intervals by Treatment Status",
       x = "Treatment", y = "Mean Confidence") +
  theme_minimal()
```



Based on the above chart, we can conclude that different types of advice (either ChatGPT or the expert) did not severely change people's confidence levels on their answers. Surprisingly, I though that expert's or ChatGPT's advice would at least one increase people's confidence, but it did not.

6. Calculate ATEs and CIs

The formula for the 95% Confidence Interval is: $\overline{x} \pm 1.96 s_x / \sqrt{N}$.

Moreover, the formula for standard error of ATE is: $\sqrt{\frac{sd_t^2}{N_t} + \frac{sd_c^2}{N_c}}$

```
arm_info = chatgpt_data %>% group_by(condition) %>% summarise(
            mean_c = mean(confidence),
            sd_c = sd(confidence),
```

```
            N = n(),
            lb_c = mean_c - 1.96 * sd_c / sqrt(N),
            ub_c = mean_c + 1.96 * sd_c / sqrt(N)
            )
meanCI = arm_info %>% select(condition, mean_c, lb_c, ub_c)

ate_info = arm_info %>% mutate(
    ate = mean_c - mean_c[1],
    se_ate = sqrt( sd_c^2 / N + sd_c[1]^2 / N[1]),
    lb = ate - 1.96 * se_ate,
    ub = ate + 1.96 * se_ate
    )
  ate_info = ate_info %>% filter(condition != 0) %>%
    select(condition, ate, lb, ub)

  kable(ate_info, digits = 2)
```

condition	ate	lb	ub
1	-0.04	-0.14	0.07
2	-0.03	-0.13	0.08

I calculated the Average Treatment Effect (ATE) for the two treatments on the confidence levels comparing to the control group and got a surprised result!

- Both ChatGPT and expert group have very small and negative ATE, suggesting that people slightly have less confidence compared to the control group.
- However, the 95% confidence intervals for both group include 0, suggestion that the effects are not statistically significant. Therefore, we cannot confidently state it.

In conclusion, it is a precise zero, which rules out a practically significant effect size.

7. Subgroup Analysis

I want to test whether my second hypothesis - "People who use ChatGPT more frequently (in treatment group 1) tend to believe in ChatGPT's method more; however, people who are more familiar with ChatGPT (in control group and treatment group 2) may be less confident in their answers because they tend to reply more on ChatGPT's support." is right or wrong. I conduct a subgroup analysis based on people's familiarity (column "use_gpt") to separate participants into subgroups.

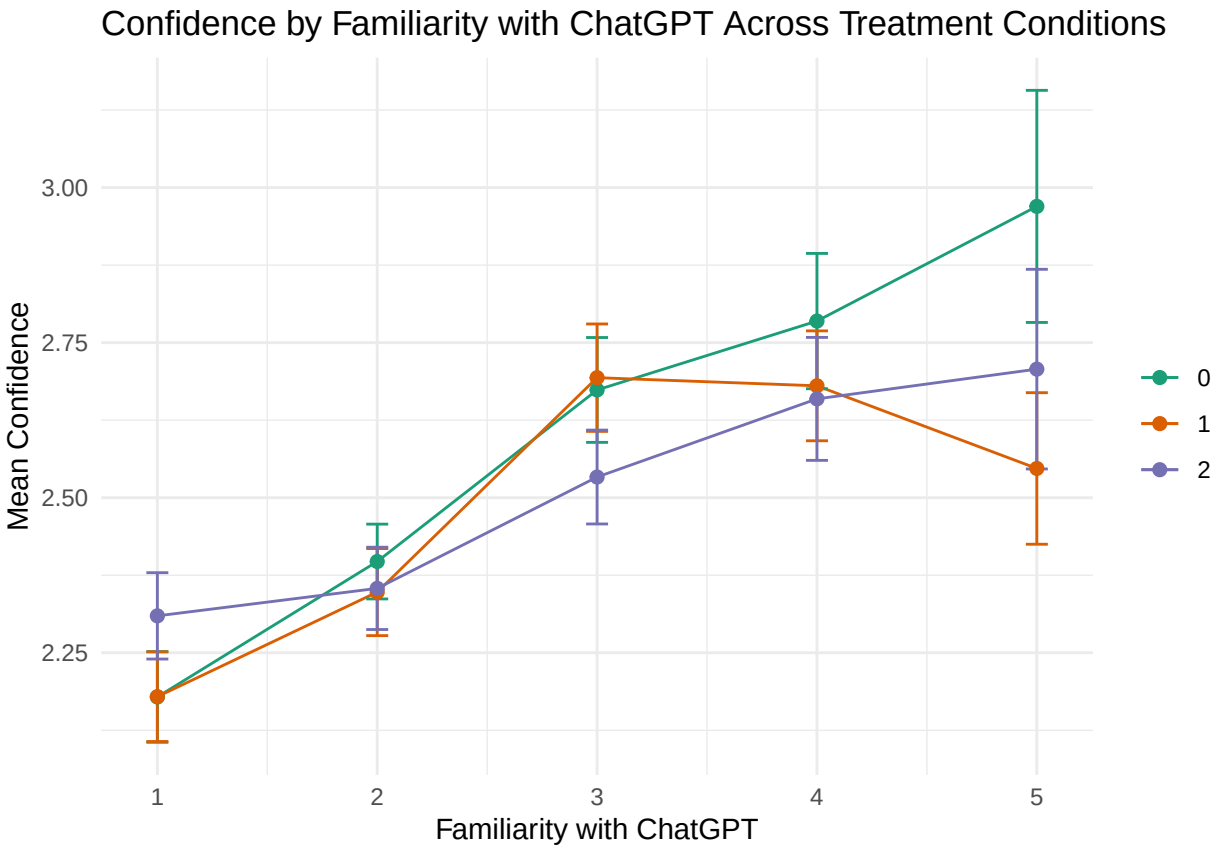
```
# make treatment as a factor variable
chatgpt_data = chatgpt_data %>% mutate(condition = as.factor(condition))

# group by familiarity of ChatGPT, treatment group, and summarizing statistics
sub_table <- chatgpt_data %>%
  group_by(use_gpt, condition) %>%
  summarise(
    mean_c = mean(confidence),
    sd_c = sd(confidence),
    N = n(),
  ) %>%
  arrange(use_gpt)
```

```
## `summarise()` has grouped output by 'use_gpt'. You can override using the
## `.groups` argument.
```

```
kable(sub_table, digits = 2)
```

use_gpt	condition	mean_c	sd_c	N
1	0	2.18	0.98	179
1	1	2.18	0.95	173
1	2	2.31	0.90	168
2	0	2.40	0.87	209
2	1	2.35	0.95	184
2	2	2.35	0.93	195
3	0	2.67	1.00	141
3	1	2.69	0.96	124
3	2	2.53	0.88	135
4	0	2.78	0.97	79
4	1	2.68	0.87	97
4	2	2.66	0.95	91
5	0	2.97	1.07	33
5	1	2.55	0.89	53
5	2	2.71	1.03	41

```
ggplot(sub_table, aes(x = use_gpt, y = mean_c, group = condition, color = condition)) +
  geom_line() +
  geom_errorbar(aes(ymin = mean_c - sd_c / sqrt(N), ymax = mean_c + sd_c / sqrt(N)), width = 0.1) +
  geom_point(size = 2) +
  labs(title = "Confidence by Familiarity with ChatGPT Across Treatment Conditions",
       x = "Familiarity with ChatGPT", y = "Mean Confidence") +
  theme_minimal() +
  scale_color_brewer(type = "qual", palette = "Dark2") +
  theme(legend.title = element_blank())
```



It's also a surprising result!!

My initial goal was to see if people who had seen ChatGPT's advice would show more confident when their familiarity with ChatGPT increased.

However, for treatment 1 (ChatGPT) - despite expecting this group to have higher confidence with increased familiarity, the graph shows a downward trend. I think this unexpected result might mean that those those more familiar with ChatGPT might be more critical to the advice or have higher standards, so they have decrease in confidence.

Contrast to my hypothesis, there was a general upward trend in confidence with greater familiarity in treatment 2 and control group. I think this suggests that maybe there's a general increase in confidence related to the use of AI tools, which is not necessarily linked to the exposure of the advice.

8. Challenges to Solve

- **Non-compliance**: We don't know how to make sure that people in treatment group actually take up ChatGPT's and expert's advice. We cannot know did they read or even adopt the methods or not.
  - I'm thinking of calculating intent-to-treat (ITT) and treatment-on-the-treated (ToT) in this experiment, but lack of data stating about people who really read or consider the advice (only text describing their strategy).
- **Sample Attrition**: There are 407 missing values in guess 2 (92 from control, 165 from treatment 1, and 150 from treatment 2). I think this is a problem of people did not change their guess even after seeing ChatGPT or an expert's advice. Although it's a non-differential attrition happened both in treatment and control group, it's still a sign that we might lose some ability to track outcomes for some participants.

9. Conclusion

In conclusion, ATE shows a trend that AI tools may influence people's confidence in their answers when estimating the crowd size, but the confidence interval shows us a precise zero so this observed effect is not statistically significant. Therefore, we cannot conclusively state that AI tools decrease confidence based on this data.

From subgroup analysis - In treatment group 1, those who are more familiar with ChatGPT did not believe in ChatGPT's method more because they might have higher standard on ChatGPT's answer; however, in control group and treatment group 2, those who are more familiar with ChatGPT have more confident in their answers. Maybe being familiar with AI tools seems to make people be more confident in their decisions, whether they received specific advice from it or not.

Although there are some indications from our findings, we still need further research to fully understand these insights.

10. ChatGPT

For the subgroup analysis - I made the sub_table but still want to make visualization of it, so I ask ChatGPT for help to how to a plot by adjusting the color and grid lines.