

Beyond the Bot: Customer Happiness with Human vs AI Customer Service

Introduction

The project aims to compare customer satisfaction and response accuracy between chatbot and human support in customer service settings. The study employs a mixed-design approach, incorporating both between-subjects and within-subjects factors.

Study Design

1. **Analysis:**

- Statistical tests: Paired sample t-tests, two-way repeated measures ANOVA, Pearson correlation

2. **Data Collection:**

- Quantitative: Satisfaction scores, usage frequency, accuracy ratings
- Qualitative: Open-ended questions for deeper insights

3. **Design:**

- Participants interact with both support types sequentially
- Order of interaction is randomized to assess its impact
- The independent variables (IVs) include the type of support (chatbot or human), the order in which participants encounter them, and the frequency of use of customer support.
- Dependent variables (DVs) include customer satisfaction scores and the rating of chatbot response accuracy.

4. **Participants:**

- Target: 40+ individuals with experience using customer services (e.g., Amazon)

Data Collection

- **Screening Question :** Frequency of use to eliminate users who have used it chatbot never or have had minimal experience(one or twice).
- **Data Integration :** Merged datasets from different collection sources (e.g., online platforms, email surveys) into a unified format, ensuring consistency in variable names and scales.
- **Quality Assurance :**
 - Verified the accuracy of data.
 - Ensured consistency across all datasets in terms of scoring scales and response formats.
 - Checked for statistical assumptions relevant to the planned analyses (t-tests, ANOVA) such as normality and homoscedasticity.

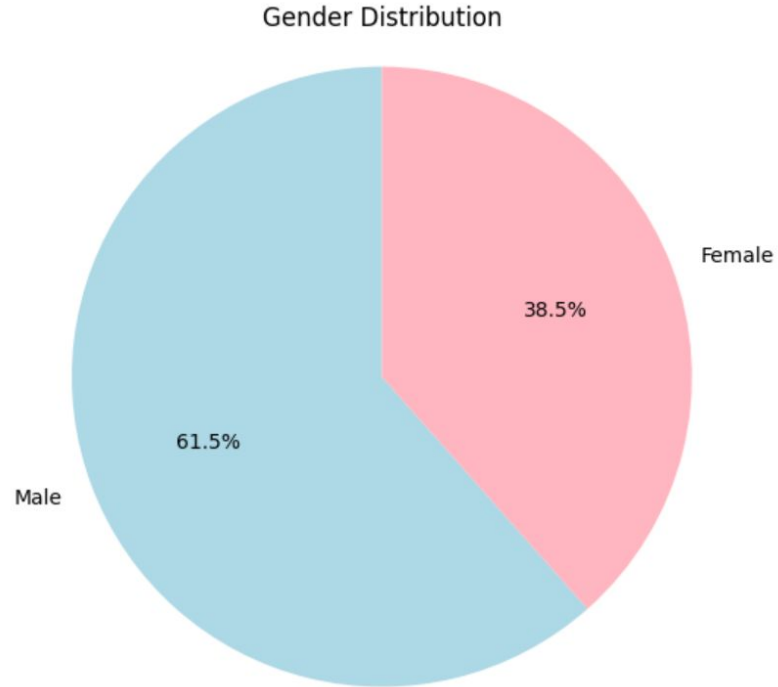
Operational definition of variables

- **Type of support (IV):** Whether the customer received support from a chatbot or a human representative.
- **Order of encountering support (IV):** The sequence in which customers interacted with the chatbot or human during the study.
- **Customer Satisfaction Scores (DV):** Quantitative scores given by customers on a scale of 1-5 based on their satisfaction with support received.
- **Rating of chatbot response accuracy(DV):** Scores given by customers on a scale of 1-5 based on how accurate they perceive the chatbot response.

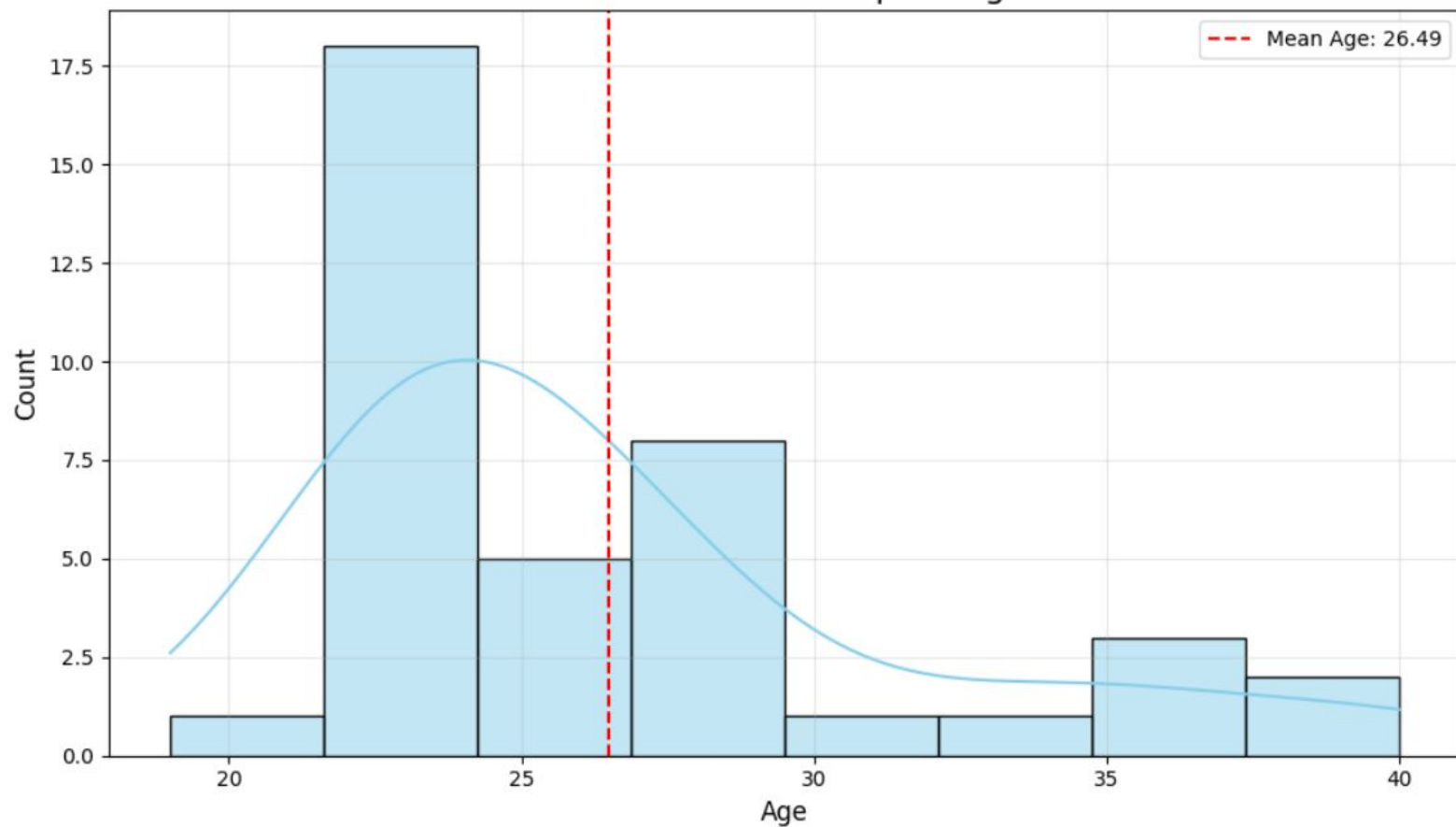
Data Cleaning

- **Converted categorical variables** like type of support (chatbot, human) and order of encountering support into a numerical format for analysis.
- **Checked for missing values** in satisfaction scores, accuracy ratings, and frequency of use data. Used mean or median imputation for continuous variables (satisfaction and accuracy scores) and mode imputation for order of support as it is a categorical variable.
- Filled empty responses for interaction with chatbot as No. This is done after the screening questions. A no here means low interaction and/or referring to interacting with automated caller on customer service before reaching the correct department for support.
- **Identified outliers** in satisfaction and accuracy scores.
- **Standardized** satisfaction and accuracy scores to ensure uniformity across different scales.

Exploratory Data Analysis



Distribution of Participant Ages



Mean Age: 26.49
Median Age: 25.00
Age Range: 19 to 40
Standard Deviation: 5.10

Analysis



```
# 1. Repeated Measures ANOVA
```

```
df_long = pd.melt(df[['Satisfaction_Chatbot', 'Satisfaction_Human']],  
                  var_name='Support_Type', value_name='Satisfaction')  
df_long['Subject'] = np.repeat(range(len(df)), 2)
```


```
aov_rm = AnovaRM(df_long, 'Satisfaction', 'Subject', within=['Support_Type']).fit()  
print("Repeated Measures ANOVA Results:")  
print(aov_rm.summary())
```

```
# 2. Paired Sample t-test
```

```
t_stat, p_value = stats.ttest_rel(df['Satisfaction_Chatbot'], df['Satisfaction_Human'])  
print("\nPaired Sample t-test Results:")  
print(f"t-statistic: {t_stat}, p-value: {p_value}")
```

```
# 3. Correlation Tests
```

```
pearson_r, pearson_p = stats.pearsonr(df['Chatbot_Accuracy'], df['Satisfaction_Chatbot'])  
print("\nPearson Correlation (Chatbot Accuracy vs Satisfaction):")  
print(f"r: {pearson_r}, p-value: {pearson_p}")
```



```
# Spearman correlation between frequency of use and satisfaction
spearman_rho, spearman_p = stats.spearmanr(df['Frequency_of_Use'], df['Satisfaction_Overall'])
print("\nSpearman Correlation (Frequency of Use vs Overall Satisfaction):")
print(f"rho: {spearman_rho}, p-value: {spearman_p}")

# 4. Order Effect Analysis
t_stat_order, p_value_order = stats.ttest_ind(
    df[df['Order'] == 0]['Satisfaction_Overall'],
    df[df['Order'] == 1]['Satisfaction_Overall']
)
print("\nIndependent t-test for Order Effect:")
print(f"t-statistic: {t_stat_order}, p-value: {p_value_order}")

# 5. Descriptive Statistics
print("\nDescriptive Statistics:")
print(df[['Satisfaction_Chatbot', 'Satisfaction_Human']].describe())
```

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
TypeofSupport	Sphericity Assumed	38.771	1	38.771	72.761	<.001	.663
	Greenhouse-Geisser	38.771	1.000	38.771	72.761	<.001	.663
	Huynh-Feldt	38.771	1.000	38.771	72.761	<.001	.663
	Lower-bound	38.771	1.000	38.771	72.761	<.001	.663
TypeofSupport * Whichtypeofsupportdidyou encounterfirstinyourmostre centcustomerse	Sphericity Assumed	.002	1	.002	.004	.950	.000
	Greenhouse-Geisser	.002	1.000	.002	.004	.950	.000
	Huynh-Feldt	.002	1.000	.002	.004	.950	.000
	Lower-bound	.002	1.000	.002	.004	.950	.000
Error(TypeofSupport)	Sphericity Assumed	19.716	37	.533			
	Greenhouse-Geisser	19.716	37.000	.533			
	Huynh-Feldt	19.716	37.000	.533			
	Lower-bound	19.716	37.000	.533			

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	1114.693	1	1114.693	1540.757	<.001	.977
Which type of support did you encounter first in your most recent customer service	.026	1	.026	.037	.849	.001
Error	26.768	37	.723			

Estimated Marginal Means

Type of Support

Measure: MEASURE_1

Type of Support	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.487	.090	4.304	4.670
2	3.076	.155	2.762	3.391

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	How satisfied were you with the chatbot support ?	3.08	39	.957	.153
	How satisfied were you with the human support ?	4.49	39	.556	.089

Paired Samples Correlations

		N	Correlation	Significance	
				One-Sided p	Two-Sided p
Pair 1	How satisfied were you with the chatbot support ? & How satisfied were you with the human support ?	39	.175	.143	.286

Paired Samples Test

		Paired Differences							Significance	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	One-Sided p	Two-Sided p
					Lower	Upper				
Pair 1	How satisfied were you with the chatbot support ? - How satisfied were you with the human support ?	-1.410	1.019	.163	-1.740	-1.080	-8.645	38	<.001	<.001

Paired Samples Effect Sizes

			Standardizer ^a	Point Estimate	95% Confidence Interval	
					Lower	Upper
Pair 1	How satisfied were you with the chatbot support ? - How satisfied were you with the human support ?	Cohen's d	1.019	-1.384	-1.821	-.938
		Hedges' correction	1.039	-1.357	-1.785	-.920

a. The denominator used in estimating the effect sizes.

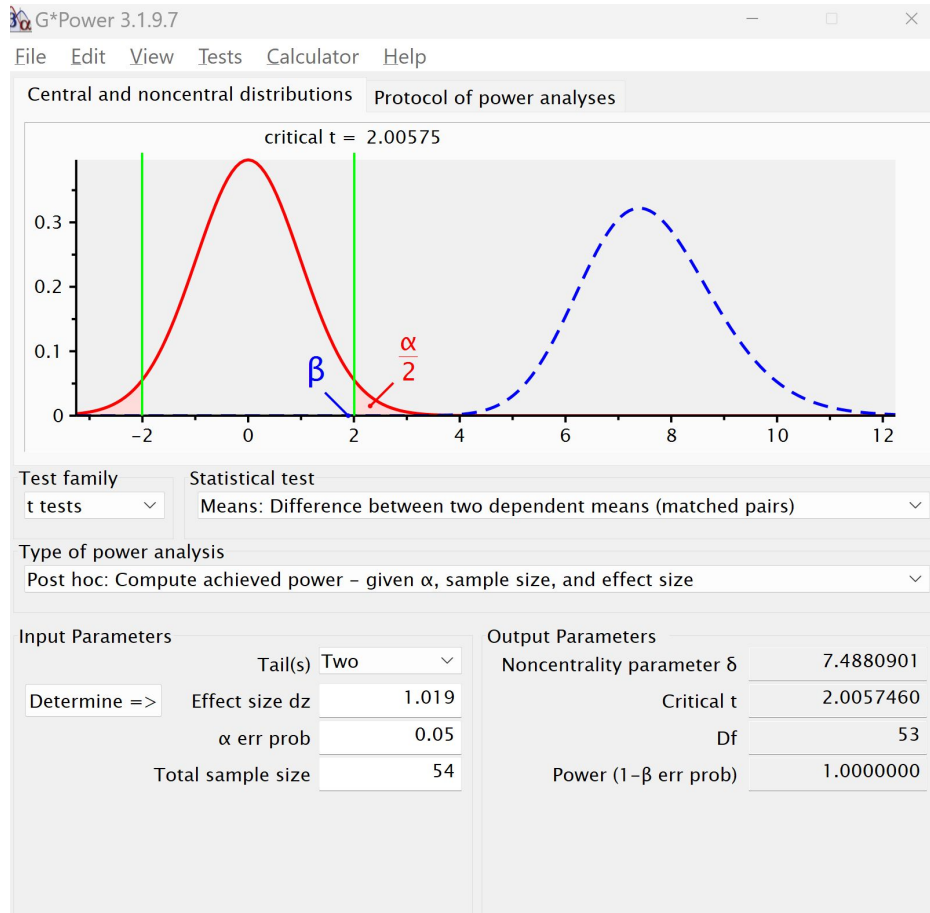
Cohen's d uses the sample standard deviation of the mean difference.

Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

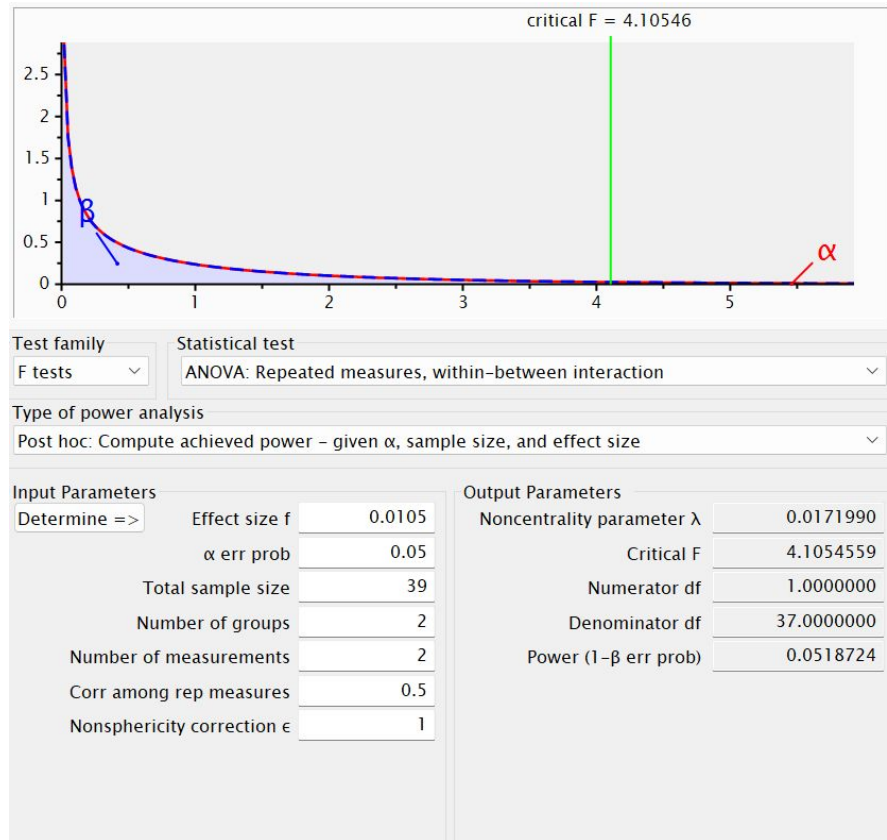
Power Analysis

Using G-Power

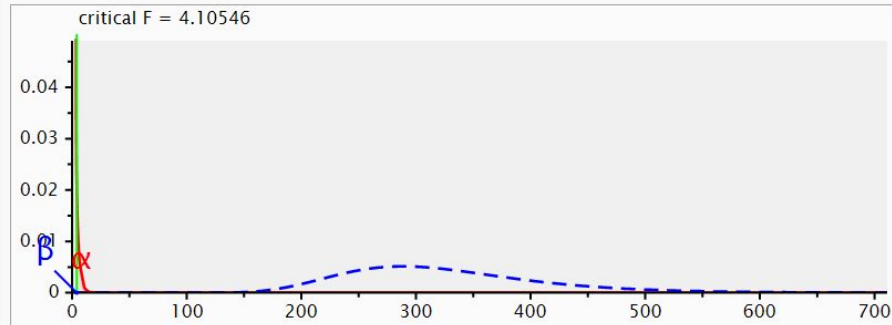
T-test



ANOVA



Central and noncentral distributions Protocol of power analyses



Test family

F tests

Statistical test

ANOVA: Repeated measures, within factors

Type of power analysis

Post hoc: Compute achieved power - given α , sample size, and effect size

Input Parameters

Determine =>

Effect size f 1.4026

α err prob 0.05

Total sample size 39

Number of groups 2

Number of measurements 2

Corr among rep measures 0.5

Nonsphericity correction ϵ 1

Output Parameters

Noncentrality parameter λ 306.8967

Critical F 4.105459

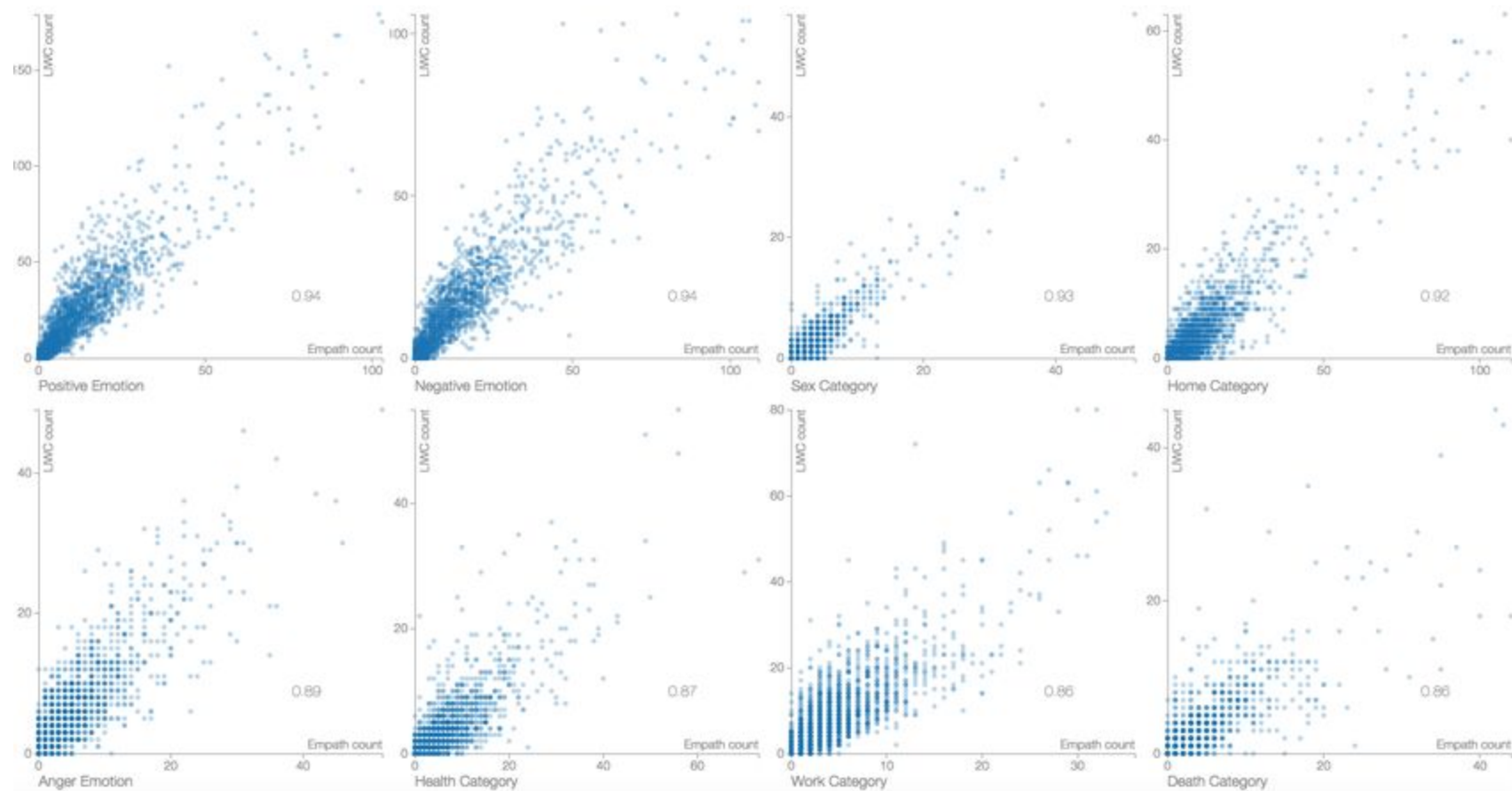
Numerator df 1.0000000

Denominator df 37.0000000

Power ($1 - \beta$ err prob) 1.0000000

LIWC vs Empath

- **EMPATH:**
 - Covers a broader set of categories than LIWC
 - Lacks some of LIWC's condensed summary variables
 - Offers hundreds of predefined lenses for text analysis
 - Is available for free use
- Despite their differences, studies have shown that EMPATH and LIWC results are **highly correlated** with an average **Pearson correlation of 0.90**
- For specific categories like positive emotion, correlations can be as high as 0.944



EMPATH ON CHATBOT SUPPORT

Liking

- The high score for "**time**" suggests that **users** particularly **appreciate** the **quick response time** of **chatbots**.
- "**Technology**" reflects an appreciation for the technological aspect of chatbot support.
- "**Speed**" further reinforces the appreciation for fast responses.
- The presence of "**business**" **might indicate** that users see chatbots as an **efficient business solution**.

Improvements

- "**Understand**" being the top category strongly suggests that users want chatbots to better comprehend their queries.
- "**Communicate**" indicates a **desire for improved communication** abilities from chatbots.
- "**Cognitive_processes**" and "**intelligence**" suggest users want chatbots to be smarter and more capable of complex reasoning.
- "**Improve**" directly reflects users' desire for enhancements in chatbot functionality.

EMPATH ON CUSTOMER SUPPORT

Liking

- **"Understand"** being the **top category** indicates that users **highly value the human ability to comprehend** their issues.
- **"Solve"** suggests that human support is **effective in resolving problems**.
- **"Positive_emotion"** and **"emotional"** indicate that **users appreciate the empathy and emotional connection** with human support.
- **"Help"** reinforces the perception that human support is **beneficial in addressing user needs**.

Improvements

- **"Time"** and **"respond"** being the top categories suggest that users **want faster response times** from human support.
- **"Timely"** further reinforces the **desire for quicker service**.
- **"Positive_emotion"** indicates that users generally have a positive attitude towards human support but see room for improvement.
- The presence of **"irritation"** **might suggest some frustration** with current human support experiences, possibly related to response times.

Result

Higher Satisfaction with Human Support: There is a clear preference for human customer support over chatbots. This is evident from the significant difference in satisfaction scores favoring human support ($M=4.50$ for humans vs. $M=3.08$ for chatbots).

It highlights the importance of human elements, such as empathy and personalization, in customer service as highlighted by the text analysis using EMPATH, which might not be adequately replicated by chatbots.

Order of Interaction Not Influential: The order in which customers interact with the support types (human or chatbot first) does not significantly affect their satisfaction. This suggests that customer satisfaction is more influenced by the support type than by any sequence effect.

Negative Correlation of Chatbot Accuracy with Satisfaction: Interestingly, better chatbot response accuracy predicts lower customer satisfaction. It could indicate that factors other than accuracy, such as the conversational style or the perceived empathy of responses, are crucial in customer satisfaction. Additionally only accurate response might not meet other customer needs or expectations.

Frequency of Use and Satisfaction: There is a negligible correlation between how often customers use the support service and their satisfaction levels, indicating that the frequency of interaction does not majorly impact satisfaction.

The negligible impact of the frequency of use on satisfaction may reflect that customers' satisfaction is more a function of the quality of each interaction rather than cumulative experiences.

Limitations and Critical Evaluation

- **Sample Representativeness:** The sample, drawn from users of services like Amazon, may not represent all demographics of customer service users. This could limit the generalizability of the findings.
- **Type II Error in Some Tests:** While power analysis suggests a low chance of Type II error in most tests, the high chance (94.8%) in the interaction effect between support type and order is a significant limitation. It might have been the reason of failing to detect an existing effect due to the study design or sample size.
- **Unexpected Findings in Chatbot Accuracy:** The negative correlation between chatbot accuracy and satisfaction is counterintuitive and warrants further investigation. It could be due to unmeasured variables like customer expectations or the nature of queries handled by chatbots.
- **High power (1.0) in detecting differences between human and chatbot support** and in the paired samples t-test suggests that our study was well-equipped to detect true effects. However, the **lower power** in the **correlation test** (0.757) implies a moderate risk of missing a true effect.

Recommendations for Future Research

- **Expand Sample Diversity:** Include a more diverse participant pool to enhance the generalizability of the findings.
- **Investigate the Paradox in Chatbot Accuracy:** Explore qualitative factors like user expectations or specific aspects of chatbot interactions that might explain the paradoxical relationship between chatbot accuracy and satisfaction.
- **Longitudinal Studies:** Conduct longitudinal studies to understand how repeated interactions with chatbots and human support over time affect customer satisfaction.
- **Examine Underlying Factors :** Delve deeper into why human support is preferred, focusing on qualitative aspects like empathy, personalization, and emotional intelligence in customer service interactions.

References

- https://www.researchgate.net/figure/Empath-categories-strongly-agreed-with-LIWC-at-an-average-Pearson-correlation-of-090_fig3_301872654
- <https://www.ijcai.org/proceedings/2017/677>
- <https://github.com/Ejhfast/empath-client>

Appendix

Repeated Measures ANOVA

Repeated measure ANOVA with the type of support (Human or Chatbot) being the within subject factor and the order of support as the between subject factor. We analyzed the main effects of type of support and encounter order as well as the interaction effect.

In within subjects comparison, customers report significantly greater satisfaction with human customer support representatives ($M = 4.487, SE = 0.089$) than chatbot based systems ($M = 3.077, SE = 0.144$; $F(1,37) = 72.761, p < 0.001$). There is no significant effect or interaction of the order with which participants reported their satisfaction score ($F_s < 0.037, p_s < 0.85$).

Paired samples t-test

Customers report a significantly higher satisfaction with human customer support representatives ($M = 4.50, SD = 0.56$) than chatbot based system ($M = 3.08, SD = 0.96$); $t(38) = -8.65, p < 0.001$.

Pearson Correlation for Chatbot Response Accuracy and Satisfaction

After aggregating the independent variables into chatbot response accuracy, this overall chatbot response accuracy was strongly and significantly negatively correlated with customer satisfaction such that lower customer satisfaction was predicted by better chatbot response accuracy ($r(N=39) = -0.411, p = 0.009$).

Pearson Correlation for Frequency of use and Customer Satisfaction

After aggregating the independent variables into frequency of use, this overall frequency of use was weakly and negatively correlated with customer satisfaction such that little or no change in customer satisfaction was predicted by frequency of use ($r(N=39) = -0.033, p = 0.841$).

Repeated Measures ANOVA

Fail to reject the null hypothesis that there is no significant effect of order in which participants interact with human and chatbot customer service representatives. The chance of type II error is zero.

We fail to reject the null hypothesis that there is no interaction between the type of support being offered and the order. The chance of type II error is 94.8%.

We reject the null hypothesis that there is no significant difference between the type of support being offered. Type I error is <0.001 (p-value associated with the test statistic (F) at $\alpha = 0.05$).

Paired samples t-test

The results show that customers prefer human customer representatives compared to chatbot support. We reject the null hypothesis as $p < 0.001$. The chance of making Type I error is <0.001 (p-value associated with the test statistic (t) at $\alpha = 0.05$).

Pearson Correlation for Chatbot Response Accuracy and Satisfaction

The results show that customers are not satisfied with chatbot response accuracy. We reject the null hypothesis as $p = 0.009$ which is less than 0.05. The chance of making Type I error is 9 in 1000 (p-value associated with the test statistic at $\alpha = 0.05$).

Pearson Correlation for Frequency of use and Customer Satisfaction

The results show that there is little or no relation between frequency of use and customer satisfaction. We fail to reject the null hypothesis as $p = 0.841$ which is greater than 0.05. Chance of Type II error is 94.5%.