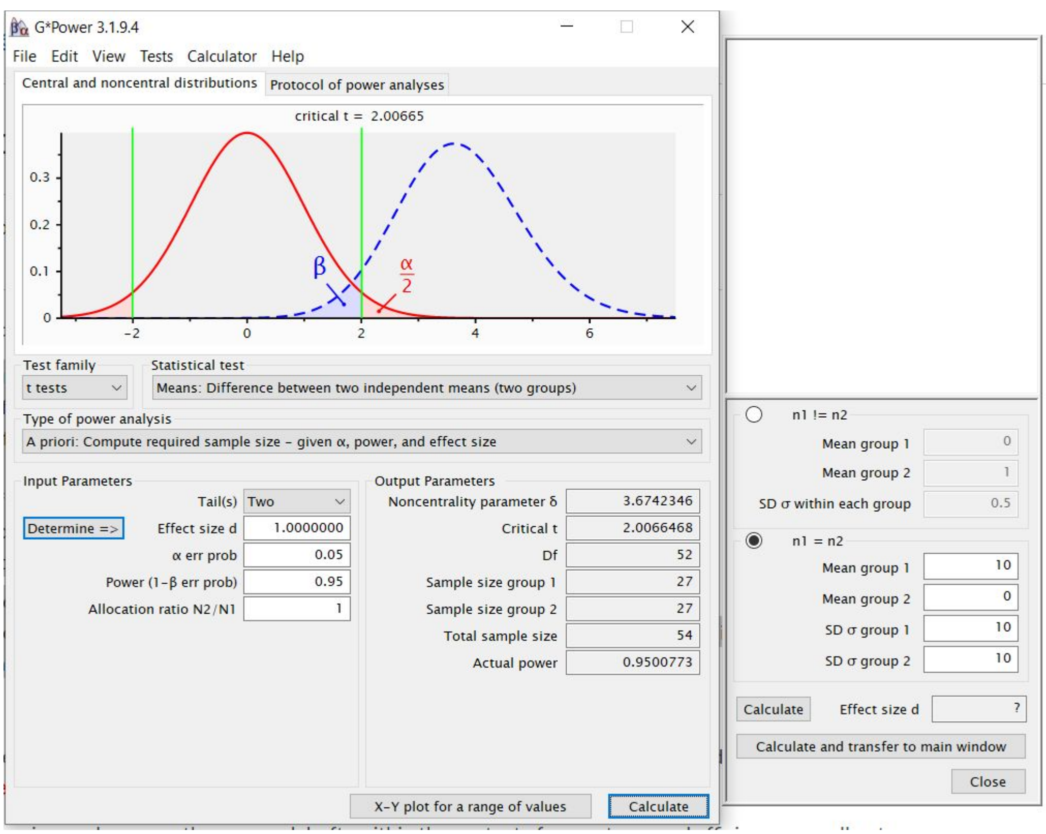
Sex Differences in Identifying A.I. Generated Music

By Alec Moldovan

Complete Approved Project Proposal

- **the research question (including a brief description of why this question is of interest);**
    - Will a male or female be any better at distinguishing whether a musical piece is generated from Artificial Intelligence or a Human?
        - We want to see if either males or females may have some innate ability to identify AI vs. human creations.
- **identification of variable(s), factor(s), population and groups to be studies;**
    - Variables: Gender
    - Factor: Audio Clip ('I am AI')
- **a formulation of the null and alternative hypothesis, along with with a choice of type I error rate;**
    - Males and Females cannot distinguish between AI and human-generated music.
    - Type I error rate = 0.05
- **a table (or plot) of the power of the test for selected sample sizes (N = 10, 20, 30, 50);**



- **a description of the statistical tests planned;**
    - Dr Bellofiore foreshadowed we will be using a chi-squared test.

- **a brief discussion of the methods implemented in this study to reduce/avoid bias.**
  - The survey will be given to similarly aged, educated participants
  - We will block (experimental unit) is based on age and gender.
- **a textbox including the following items:**
  - **one (or more, if needed) survey question aiming at collecting the data you need;**
    - Is this piece AI-generated? (Note: This audio clip is 1 of 2 of human- and AI-generated music)
  - **one (or more, if needed) factor or grouping variable pertaining to demographic or academic information (e.g. age, sex, academic level, GPA)**
    - Age, Sex

Design of experiment

A Turing test is a test of a computer's ability to mimic intelligent behaviour. In order to pass, the machine must perform in a way that makes it indistinguishable from a human. Turing tests pair a computer, a human, and a human evaluator. The computer passes the test if the elevator cannot reliably tell the difference between the performance of the computer from the human.

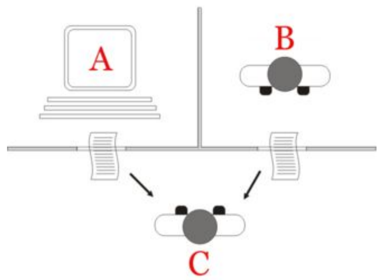


Fig. 1 Turing test between a computer (A), a human conversationalist (B), and a human evaluator (C).

Turing tests can be used for any type of task. One interesting advancement in A.I. is the ability of computer programs to piece together original musical compositions. Using machine learning, an algorithm can compose its own rhythm, melody, and harmony. Advanced A.I.'s can even put together musical scores that sound like something a human composer would create. This case is the inspiration behind this project.

Regardless of the task, a Turing test is only as useful as the evaluator is competent. In his original description of the test Alan Turing never specified the skill or knowledge level required to be an interrogator in his description of the test, however, the success of these types of test is determined by the skill, or naïveté of the evaluator.

The question this project address is, do women perform better as the human evaluator for Turing tests of musical composition than men? There is a "female advantage at recognizing familiar melodies" which may help women to distinguish between human and computer pattern making in music [1].

The projects null hypothesis is as follows: Male and Female ability to distinguish between AI and human-generated music is equal with an α of 0.05. To examine this, a survey was given to participants to act as the evaluator in a Turing test comparing two musical computations, composed by an A.I. and human respectively. The Independent variables considered were the two sexes (Male and Female). A sample of 54 students of SJSU BME majors was used as participants to represent to similarly aged and educated participants.

Design of Experiment Matrix

Test			
Sex Differences in Identifying A.I. Generated Music			
Gender	#Pass	#Fail	Sample Size
Female			28
Male			26

Statistical analysis plan

Logistic Regression in R

Logit Function & Inverse Logit Function

The logit function or also known as the log-odds function computes the logarithm of the odds $\frac{p}{1-p}$ where p is equal to probability, which also equals $\beta_0 + \beta_1 x_1$ where β_0 *and* β_1 equal to the intercept and slope of the logit graph and x_1 is equal to the binary response variable [2].

Below is the equation:

$$logit(p) \ = \ log(\frac{p}{1-p}) \ = \beta_0 + \beta_1 x_1 \ = \ -0.31015 \ + \ 0.02247x_1$$

Eq. 1 Logit function relation to general linear models. The project's actual coefficients are used here.

So if x_1 equals 0, if it is not a female, then the log(odds of passing) are -0.31015, which means, in this case, it is basically 50% chance since the logit(1) equals 0.5 probability. This should foreshadow that gender is not a good predictor for the response variable (Pass for successfully detecting correctly the AI-generated song vs. Fail for failing to detect the AI-generated song).

Here is an example of how Logistic Regression is part of Generalized Linear Models

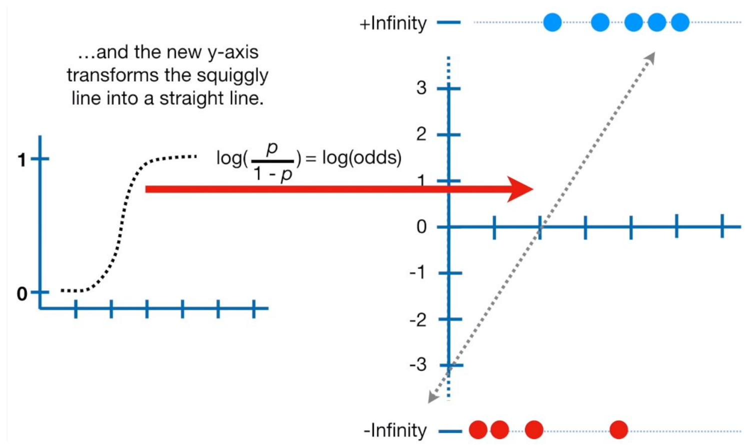


Fig 1. Graphical representation of how a logistic regression is graphed using the logit function, probabilities, and odds [4].

Chi-Squared (Significance Testing)

Our null hypothesis is that Males and Females cannot distinguish between AI and human-generated music. We chose a Type I error rate = 0.05 with a sample of 54 participants.

The expected values, assuming there is no relationship between being female and being better at detecting AI-generated music, is calculated as such:

1. Calculate the probability of guessing the correct

$$Degrees\ of\ Freedom\ =\ (no.\ of\ rows\ -1)\times(no.\ of\ columns-1)$$

Eq.2 The Degrees of Freedom is calculated by this formula.

The test statistic should be designed to describe, with a single number, how much the "observed" frequencies differ from the "expected" frequencies (i.e, the frequencies we would expect if the null hypothesis is true). The study employed the Yate's Correction for Continuity because chi-square tests are biased upward when used on 2X2 contingency tables [5].

$$X^2 = \sum \frac{(|observed\ value - expected\ value| - 0.5)^2}{expected\ value}$$

Eq. 3 Yates' correction for continuity version of Pearson's chi-squared statistics.

The chi-squared critical value is calculated by calculating the chi-statistic when alpha = 0.05 (aka p-value = 0.05) as shown below:

	P				
DF	0.995	0.975	0.20	0.10	0.05
1	0.0000393	0.000982	1.642	2.706	3.841

Fig.

To test if our hypothesis is statistically significant, we need to compare the Chi-Statistic value to the Chi-critical value.

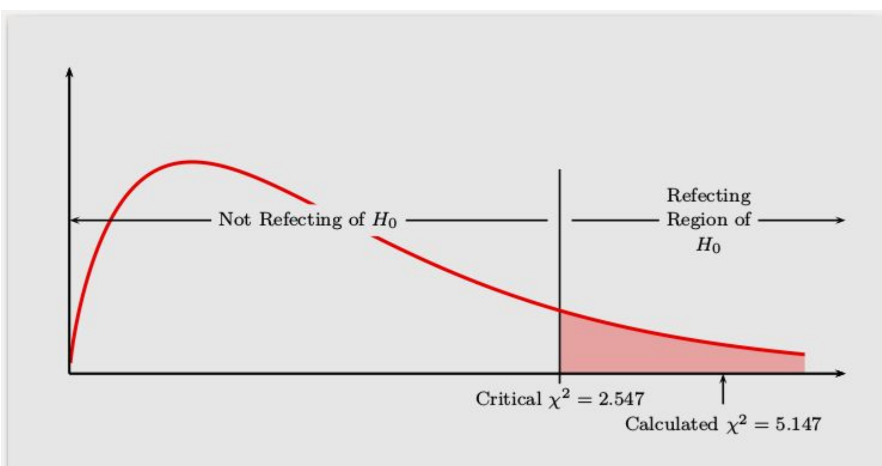


Fig. A graphical example of how the Chi-Square distribution is associated with hypothesis testing [3].

If X^2 -statistic value is > X^2 -critical value	If X^2 -statistic value is < X^2 -critical value
Reject H_o ; Accept H_a	Cannot reject H_o ; Reject H_a

Table 1. Hypothesis Testing.

Results

Coefficients	Estimate	Standard Error	Z value	Pr(>|z|)
Intercept	-0.31015	0.39696	-0.781	0.435
femmale1	0.02247	0.55083	0.041	0.937

Table 1. The Logistic Regression Model Fit Coefficients and model fit outputs from R (programming language).

‘Actual’ Values	Pass	Fail	Total
Female	12	16	28
Male	11	15	26
Total	23	31	54

Table 2. The actual values.

‘Expected’ Values	Pass	Fail	Total
Female	11.926*	16.074*	28
Male	11.074*	14.926*	26
Total	23	31	54

Table 3. The expected values.

	Value	Associated Formula
Degrees of Freedom	1	Equation 2

X^2 - continuity corrected statistic	7.609e-31	Equation 3
X^2 - critical value	3.841	

Table 4. Chi-Square Test Results

Conclusions and recommendations

Since the chi-squared statistic of 7.609e-31 is less than the critical chi-square value at 3.841 we cannot reject the null hypothesis that Males and Females cannot distinguish between AI and human-generated music. Based on our sample size of 54 there is no association between gender and ability to detect AI produced music. Increasing the N value would amplify any effect with the current hypothesis. Another alternative is reevaluating the null hypothesis and choosing a different independent variable that could offer results that have an effect that is statistically significant with the current N of 54. Based on these results, when determining qualifications for a Turing test based on composed musical scores there are no benefits to choose human evaluators based on their sex.

Discussion of limitations

The study design is a source of bias for two reasons: (1) The songs selected may have been too simple (i.e. How complex a song is played maybe a better indicator for interrogators). (2) Confederate Effect is when a human to be misidentified by the interrogator as a machine due to preconceived beliefs how a machine should perform in contrast to a human from the interrogator [6].

References

1. Miles, S. A., Miranda, R. A., & Ullman, M. T. (2016). Sex Differences in Music: A Female Advantage at Recognizing Familiar Melodies. *Frontiers in psychology*, *7*, 278. doi:10.3389/fpsyg.2016.00278
2. The Pennsylvania State University. (2018). 6.2 - Binary Logistic Regression with a Single Categorical Predictor. Retrieved from https://newonlinecourses.science.psu.edu/stat504/node/150/
3. Mehmet OZC. (2016, January 6). *Plotting chi-square distribution*. Retrieved from https://tex.stackexchange.com/questions/409124/plotting-chi-square-distribution?rq=1
4. StatQuest with Josh Starmer. (2018, June 4). *YouTube* [Video file]. Retrieved from https://www.youtube.com/watch?v=vN5cNN2-HWE
5. Stephanie. (2017, October 12). Yates Correction: What is it used for in Statistics? Retrieved from https://www.statisticshowto.datasciencecentral.com/what-is-the-yates-correction/

6. Wikipedia. (2008, October 21). Confederate effect. Retrieved from https://en.wikipedia.org/wiki/Confederate_effect

Appendix
Raw Data

Gender	Response	Result
Male	B	0
Male	B	0
Female	B	0
Male	B	0
Male	B	0
Female	B	0
Female	B	0
Female	B	0
Male	A	1
Female	A	1
Female	A	1
Male	B	0
Female	A	1
Female	B	0
Female	B	0
Male	A	1
Female	A	1
Male	A	1
Male	B	0
Female	B	0
Male	B	0
Female	B	0
Male	A	1
Female	A	1

Male	A	1
Male	B	0
Female	B	0
Female	B	0
Female	B	0
Male	A	1
Male	B	0
Male	B	0
Male	A	1
Male	B	0
Male	A	1
Female	A	1
Male	B	0
Male	B	0
Female	B	0
Female	B	0
Female	A	1
Male	A	1
Female	B	0
Male	A	1
Female	B	0
Female	B	0
Female	A	1
Female	A	1
Female	A	1
Female	A	1
Male	B	0
Female	A	1
Male	A	1
Male	B	0

Python and R Coding
Python

Logistic Regression.py

```python
import pandas as pd
import numpy as np
from sklearn import preprocessing
import statsmodels.api as sm
import matplotlib.pyplot as plt
plt.rc("font", size=14)
from sklearn.linear_model import LogisticRegression
import seaborn as sns
sns.set()
from scipy.stats import chi2_contingency
from scipy.stats import chi2

fig, axs = plt.subplots(3)
fig.suptitle('Results', fontsize=14, fontweight='bold')
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)

desired_width = 320
pd.set_option('display.width', desired_width)

pd.set_option('display.max_columns', 10)
# ______________________________________________________________________________________________________
# Pulling Data

data = pd.read_csv('Turing-ish Test_December 6, 2019_12.14.csv')
# ______________________________________________________________________________________________________
# Data Wrangling

# Columns of interest selected.
data = data[['Q2', 'Q1']]

data.drop([0, 1], inplace=True)
data.reset_index(inplace=True)
####################
data = data[['Q2', 'Q1']]
data.columns = ['Gender', 'Pass/Fail']
####################
# Convert answers to 0 or 1
data.replace(to_replace='B', value=0, inplace=True) # All who chose B is converted to a numerical data type of 0
data.replace(to_replace='A', value=1, inplace=True) # All who chose A is converted to a numerical data type of 1
####################
# Is a Female? True = 1 False = 0
data.replace(to_replace='Male', value=0, inplace=True) # If not a female then by boolean logic it is False = 0
data.replace(to_replace='Female', value=1, inplace=True) # If a female then by boolean logic it is True = 1
# ______________________________________________________________________________________________________
# ______________________________________________________________________________________________________

"""Chi-Squared"""
# For testing if there is any statistical significance.
alpha = 0.05
matrix_data = [[12,16],
               [11,15]]
"""
Observed Data
_________________________________________
          |   Pass   |   Fail   |   Total
Female    |    12    |    16    |    28
Male      |    11    |    15    |    26
Total     |    23    |    31    |    54
________________________________________
Note: A 2X2 matrix, therefore we need to do Yate's Continuity Correction
"""
# We want to get the expected values that ASSUME there is no relationship between the being FEMALE and PASSING.
pass_odds_ratio = 23/54  # This is the probability of passing
fail_odds_ratio = 31/54  # This is the probability of failing

tot_fem_pop = 28
tot_male_pop = 26

# Getting expected values, if we assume there is no relationship being female and passing the test.
expected_pass_fem = tot_fem_pop * pass_odds_ratio
print(f'The number of females passing: {expected_pass_fem}')
print(f'The number of females not passing: {28-expected_pass_fem} ')

expected_pass_male = tot_male_pop *pass_odds_ratio
print(f'The number of males passing: {expected_pass_male}')
print(f'The number of males not passing: {26-expected_pass_male} ')
```

```


"""
Expected Data
______________________________________________
            |    Pass     |    Fail     |   Total
Female      |   11.926*   |   16.074*   |    28
Male        |   11.074*   |   14.926*   |    26
Total       |     23      |     31      |    54
_____________________________________________
Note: A 2X2 matrix, therefore we need to do Yate's Continuity Correction
"""
print(f'\nThis is the observed data below:\n\n{matrix_data}\n')
stat, p, dof, expected = chi2_contingency(matrix_data, correction=True)
print(f'Using scipy chi2 function below is the expected table:\n\n{expected}\n')
"""
Chi-Squared Statistic
The test statistic should be designed to describe, with a single
number, how much the "observed" frequencies differ from the
"expect" frequencies (i.e, the frequencies we would expect if the null hypothesis is true)

χ2 = Σ((O - E)^2)/E

If Statistic >= Critical Value: significant result, reject null hypothesis (H0), dependent.
If Statistic < Critical Value: not significant result, fail to reject null hypothesis (H0), independent.

degrees of freedom: (rows - 1) * (cols - 1)
"""
rows = 2
columns = 2
DOF = (rows - 1) * (columns - 1)
print(f'The Degrees of Freedom for the chi-squared distribution is: {DOF}\n')

# χ2 = Σ((O - E)^2)/E
print(f'The chi-statistic is: {stat}')  # Chi-statistic
prob = 1 - alpha
critical = chi2.ppf(prob, DOF)
print(f'The critical value is: {critical}\n')

if abs(stat) >= critical:
    print('Dependent (reject H0)')
else:
    print('Independent (fail to reject H0)\nThere is no relationship of being female and being able to detect AI '
          'better than males.')


plt.show()
#
"""
Calculating expected values if there is no relationship of being female and being able to detect AI better.

The number of females passing: 11.925925925925926
The number of females not passing: 16.074074074074076
The number of males passing: 11.074074074074074
The number of males not passing: 14.925925925925926

################################################################

This is the observed data below:

Observed Data

[[12, 16], [11, 15]]
______________________________________________
            |  Pass  |  Fail  |   Total
Female      |   12   |   16   |    28
Male        |   11   |   15   |    26
Total       |   23   |   31   |    54
_____________________________________________

################################################################

Using scipy chi2 function below is the expected table:

Expected Data

[[11.92592593 16.07407407]
 [11.07407407 14.92592593]]
______________________________________________
            |    Pass     |    Fail     |   Total
Female      |   11.926*   |   16.074*   |    28
```

```
Male       |    11.074*    |    14.926*    |    26
Total      |      23       |      31       |    54
________________________________________

The Degrees of Freedom for the chi-squared distribution is: 1

The chi-statistic is: 0.05503367600141795
The critical value is: 3.841458820694124

Independent (fail to reject H0)
There is no relationship of being female and being able to detect AI better than males.
"""
```

R

main.R

```r
# Title     : TODO
# Objective : TODO
# Created by: alec_
# Created on: 12/6/2019
femmale <- as.factor(c(1,0))

response<-cbind(Pass=c(12,11), Fail=c(16,15))
response

turing.logistic<-glm(response~femmale, family=binomial(link=logit))

# Ouput
turing.logistic
summary(turing.logistic)
anova(turing.logistic)



plot(response~femmale)


```