wordle_analysis

2025-08-14

Preamble

This is a template for extracting and graphing important statistics and doing comparison/multiple test.

Claude

```
claude_control <- fromJSON("predictions/claude_control.json")
claude_control_guesses = claude_control$guess_results$num_guesses
print(mean(claude_control_guesses))
```

```
## [1] 5.32
```

```
print(var(claude_control_guesses))
```

```
## [1] 1.56
```

```
claude_experiment <- fromJSON("predictions/claude_hypothesis.json")
claude_experiment_guesses = claude_experiment$guess_results$num_guesses
print(mean(claude_experiment_guesses))
```

```
## [1] 6.68
```

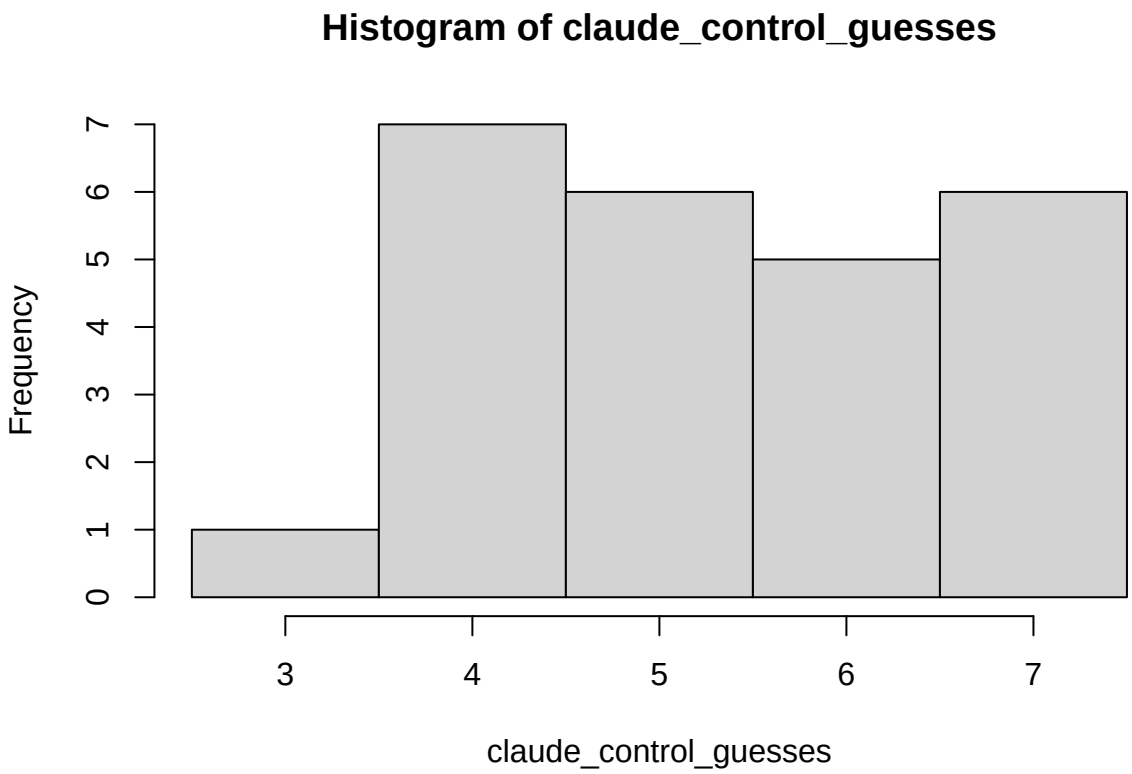
```
print(var(claude_experiment_guesses))
```

```
## [1] 0.9766667
```

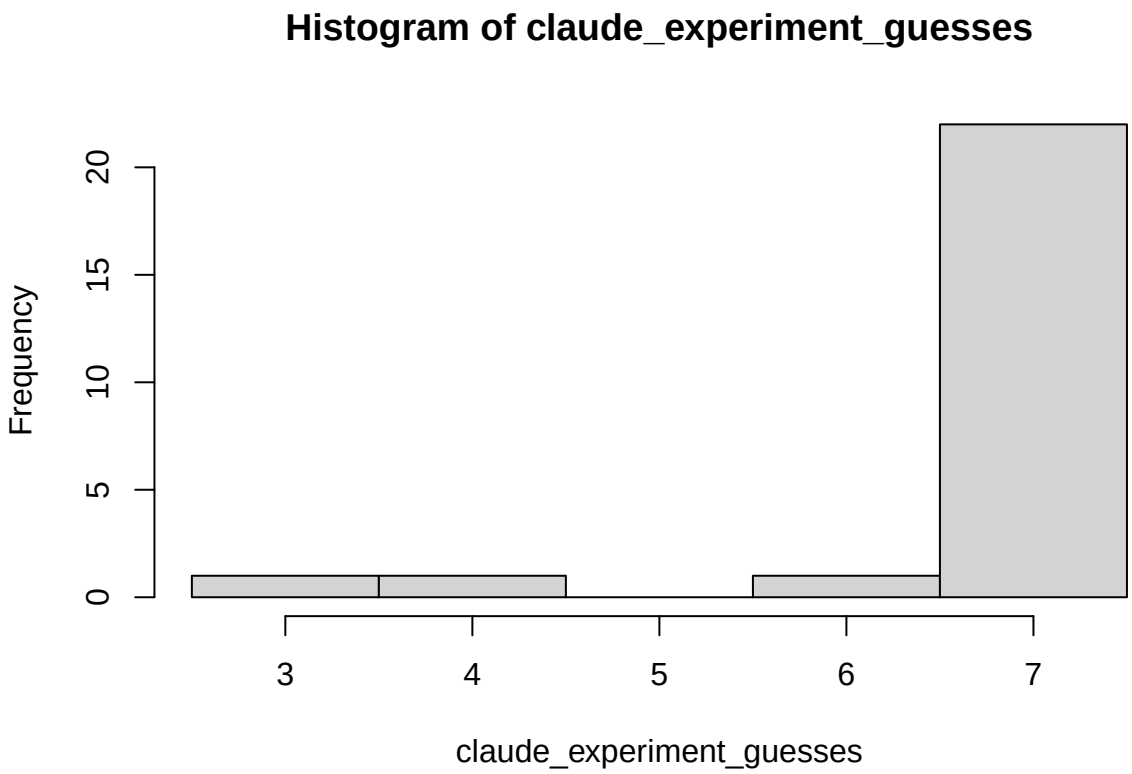
H0: experiment set is no higher than control set HA: experiment set is higher

Catch: Cannot assume normal distribution of data, see graph for reason

```
hist(claude_control_guesses, breaks = 2.5:7.5)
```




```
hist(claude_experiment_guesses, breaks = 2.5:7.5)
```



Hypothesis Testing:

```
wilcox.test(claude_experiment_guesses, claude_control_guesses, alternative = "greater", exact = TRUE)
```

```
## Warning in wilcox.test.default(claude_experiment_guesses,
## claude_control_guesses, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  claude_experiment_guesses and claude_control_guesses
## W = 505.5, p-value = 1.798e-05
## alternative hypothesis: true location shift is greater than 0
```

```
# Still need to do Permutation Test here.
```

GPT

```
gpt_control <- fromJSON("predictions/gpt_control.json")
gpt_control_guesses = gpt_control$guess_results$num_guesses
print(mean(gpt_control_guesses))
```

```
## [1] 5.04
```

```
print(var(gpt_control_guesses))
```

```
## [1] 3.706667
```

```r
gpt_experiment <- fromJSON("predictions/gpt_hypothesis.json")
gpt_experiment_guesses = gpt_experiment$guess_results$num_guesses
print(mean(gpt_experiment_guesses))
```

```
## [1] 6.48
```

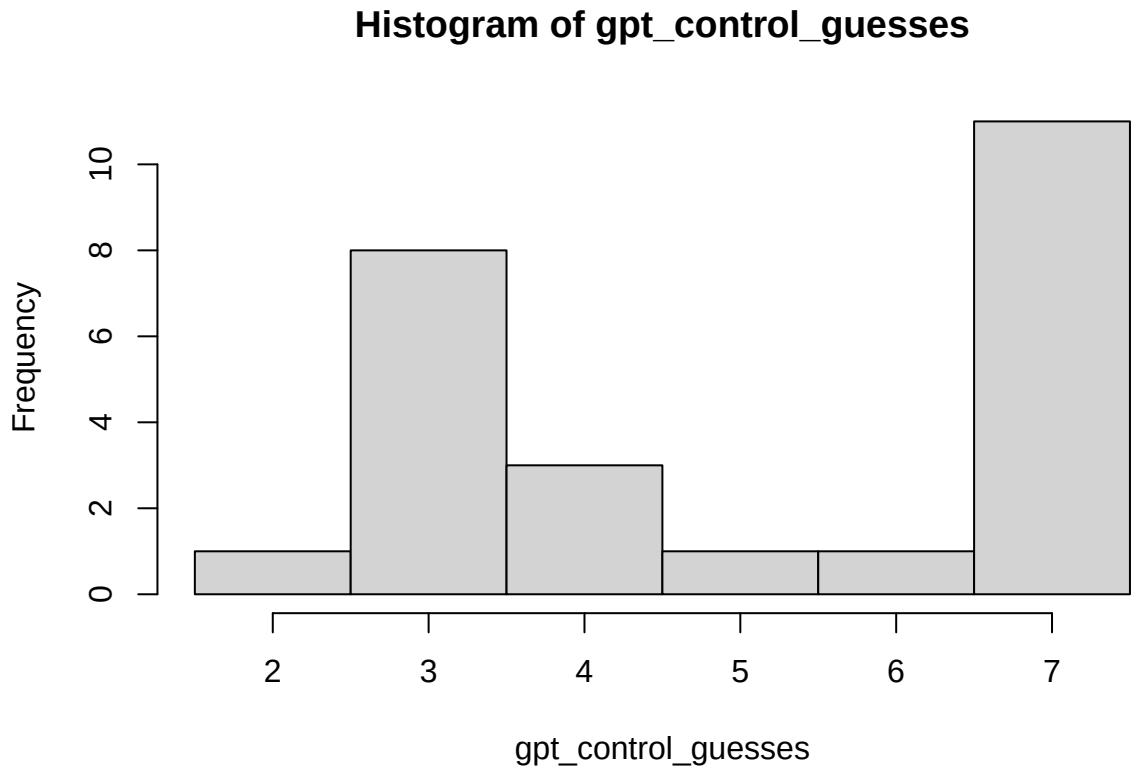
```r
print(var(gpt_experiment_guesses))
```

```
## [1] 0.9266667
```

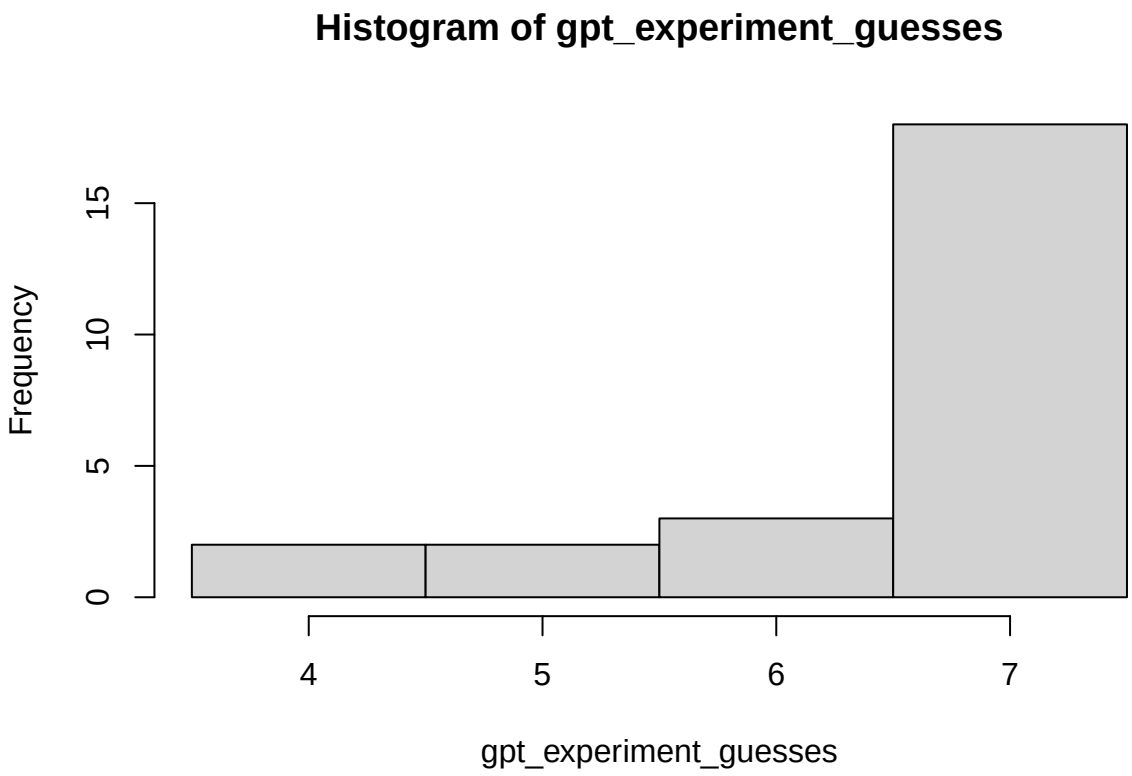
H0: experiment set is no higher than control set HA: experiment set is higher

Catch: Cannot assume normal distribution of data, see graph for reason

```r
hist(gpt_control_guesses, breaks = 1.5:7.5)
```




```r
hist(gpt_experiment_guesses, breaks = 3.5:7.5)
```



Hypothesis Testing:

```
wilcox.test(gpt_experiment_guesses, gpt_control_guesses, alternative = "greater", exact = TRUE)
```

```
## Warning in wilcox.test.default(gpt_experiment_guesses, gpt_control_guesses, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  gpt_experiment_guesses and gpt_control_guesses
## W = 437.5, p-value = 0.003448
## alternative hypothesis: true location shift is greater than 0
```

```
# Still need to do Permutation Test here.
```

Mistral

```
mistral_control <- fromJSON("predictions/mistral_control.json")
mistral_control_guesses = mistral_control$guess_results$num_guesses
print(mean(mistral_control_guesses))
```

```
## [1] 5.96
```

```
print(var(mistral_control_guesses))
```

```
## [1] 2.123333
```

```
mistral_experiment <- fromJSON("predictions/mistral_hypothesis.json")
mistral_experiment_guesses = mistral_experiment$guess_results$num_guesses
print(mean(mistral_experiment_guesses))
```

```
## [1] 6.8
```

```
print(var(mistral_experiment_guesses))
```

```
## [1] 0.5
```

H0: experiment set is no higher than control set HA: experiment set is higher

Catch: Cannot assume normal distribution of data, see graph for reason

```
hist(mistral_control_guesses, breaks = 2.5:7.5)
```


```
hist(mistral_experiment_guesses, breaks = 3.5:7.5)
```

Hypothesis Testing:

```
wilcox.test(mistral_experiment_guesses, mistral_control_guesses, alternative = "greater", exact = TRUE)
```

```
## Warning in wilcox.test.default(mistral_experiment_guesses,
## mistral_control_guesses, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  mistral_experiment_guesses and mistral_control_guesses
## W = 412.5, p-value = 0.004917
## alternative hypothesis: true location shift is greater than 0
```

```
# Still need to do Permutation Test here.
```

Discuss

Why do we use not use a t-test? Because data is not normal. Does it look like normal to you?

And what tests do we use? Wilcoxon and Permutation. Explained here for why it is appropriate.