Lab 1

Stat 5000Lab #1
Fall 2024 Due Tue Sept 3rd Name: Sam Olson

Directions: Complete the exercises below. When you are finished, turn in any required files online in Canvas, then check-in with the Lab TA for dismissal.

Assignment

Q: 1. Calculate the sample mean score for each treatment group. What is the difference in the two sample means?

A:

Continuous Variable Tabulations				
Variable	trt	NumObs	Mean	Standard Deviation
y	extrinsic	23	15.7391	5.2526
y	intrinsic	24	19.8875	4.4418

Figure 1: Sample Means by Intrinsic/Extrinsic

Sample mean for extrinsic: ≈ 15.7391

Sample mean for intrinsic: ≈ 19.8875

Difference in sample means: Intrinsic Mean - Extrinsic Mean $= 19.888 - 15.739 = 4.1484$

Q: 2. Use SAS to create a comparative box-plot for the sample mean score for each treatment group. Describe what you see.

A:

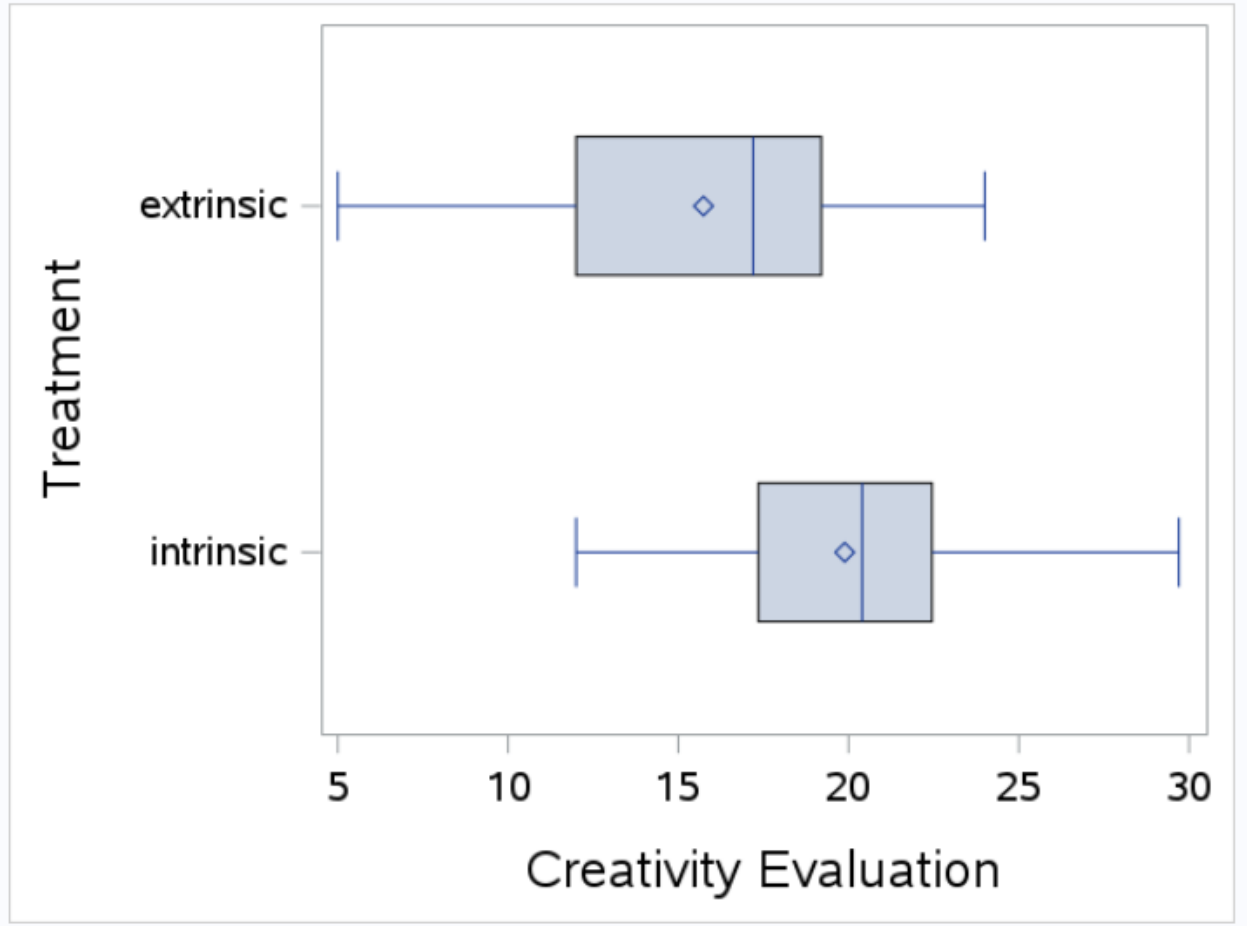

Figure 2: Boxplot of Samples by Treatment Type

Broadly speaking, we see that the scores of the intrinsic group tend to be larger than the scores of the extrinsic group. The first quartile, mean, and third quartile are all comparatively higher values than their corresponding statistic in the extrinsic group. We see from the above box and whisker plots that the IQR of the intrinsic group is roughly equal to or greater than the upper quartile of the extrinsic group (the values of the box from the mean to Q3 of extrinsic are values that are contained in Q1 to the mean of the intrinsic group).

Q: 3. What are the null and alternative hypotheses for the randomization test necessary to explore the research question?

A:

The null hypothesis is that there is no difference between the average instrinic score and the average extrinsic score, with the alternate hypothesis being that the difference between the average scores in each group is not zero. We denote this as $H_0 : \mu_{intrinsic} = \mu_{extrinsic}$, with $H_A : \mu_{intrinsic} \neq \mu_{extrinsic}$, where $\mu_{intrinsic}$ is the average score for study participants receiving the intrinsic treatment and similarly $\mu_{extrinsic}$ is the average score for study participants receiving the extrinsic treatment.

Q: 4. Conduct a randomization test for these data in SAS (be sure to keep the random seed set at 500 so everyone gets the same answer) and study the reference distribution for the difference in the sample means for the 10,000 random assignments of treatments to subjects. Describe the shape, center and variability of this distribution.

A:

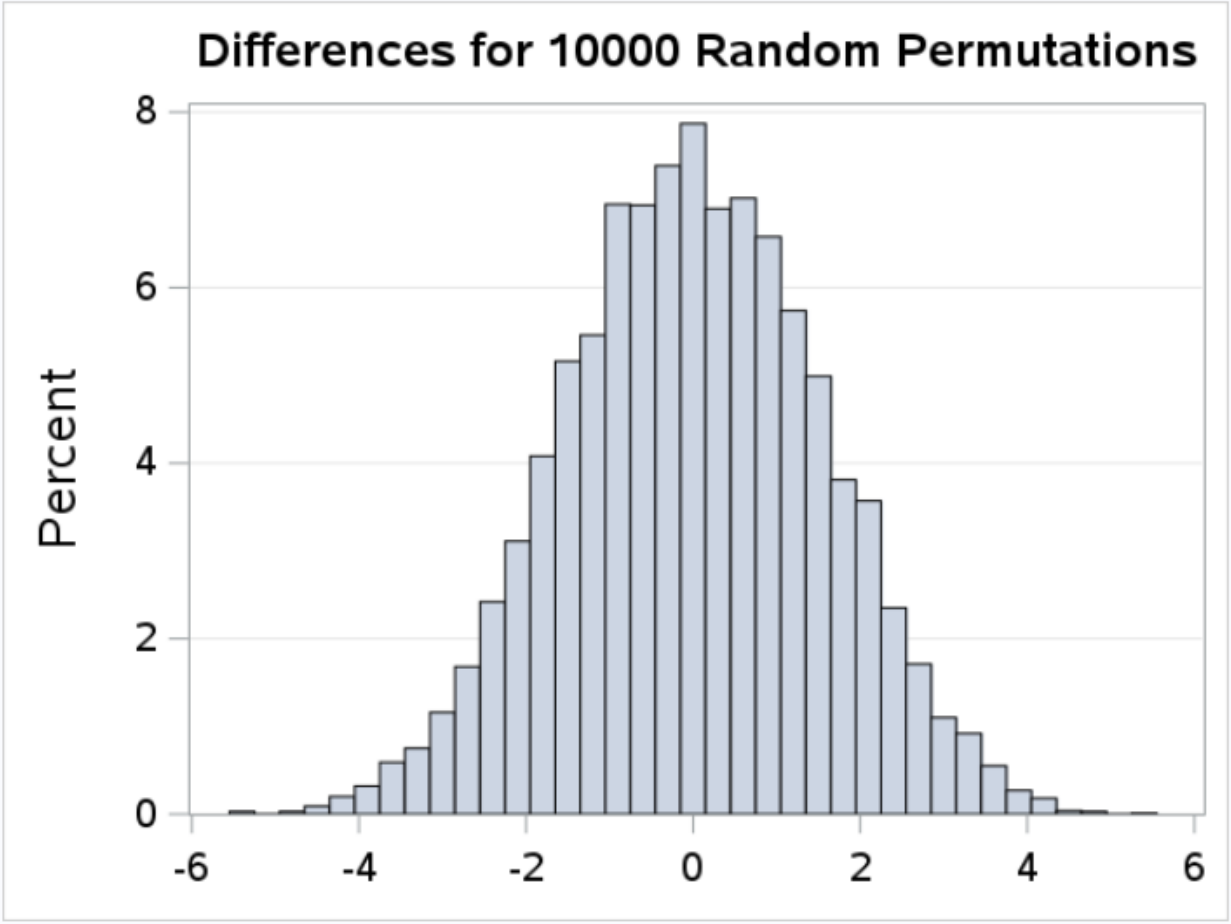

Figure 3: Permutation Test Results

The shape of the above distribution is approximately normal centered at 0. There appears to be near zero in both skewness and kurtosis as well, and symmetric variance about the center (roughly equal number of values greater than 0 as there are values less than 0); the distribution also resembles the standard normal distribution. However, as no statistics were calculated on the distribution of differences, I don't feel comfortable describing the "mean" or "variance" of the above distribution.

Q: 5. Locate the observed difference in the sample means from part (a) on the reference distribution. Given the observed difference in the sample means from part (a), what is the p-value for this randomization test?

A:

There were 45 results as extreme as the observed value in the study (from the total 10,000 studies simulated during the permutation test). This gives us the following (one-sided) p-value calculation:

$$\text{p-value} = \frac{\text{simulated results as Extreme as observed}}{total simulated observations} = \frac{45}{10{,}000} = 0.0045$$

Differences as Extreme as the Observed Difference (4.1484)

Obs	_sample_	mean1	mean2	diff
1	6077	15.1917	20.6391	-5.44746
2	5267	15.2583	20.5696	-5.31123
3	4889	15.2708	20.5565	-5.28569
4	3929	15.5167	20.3000	-4.78333
5	1302	15.5333	20.2826	-4.74928
6	4148	15.5333	20.2826	-4.74928
7	6742	15.5833	20.2304	-4.64710
8	835	15.6000	20.2130	-4.61304
9	5905	15.6333	20.1783	-4.54493
10	3965	15.6458	20.1652	-4.51938
11	3317	15.6667	20.1435	-4.47681
12	2063	15.6708	20.1391	-4.46830
13	8936	15.7000	20.1087	-4.40870
14	837	15.7083	20.1000	-4.39167
15	3161	15.7125	20.0957	-4.38315
16	5083	15.7333	20.0739	-4.34058
17	3907	15.7375	20.0696	-4.33207
18	2284	15.7417	20.0652	-4.32355
19	9212	15.7458	20.0609	-4.31504
20	4942	15.7542	20.0522	-4.29801
21	9391	15.7708	20.0348	-4.26395
22	5572	15.7958	20.0087	-4.21286
23	3486	15.8000	20.0043	-4.20435
24	5706	15.8000	20.0043	-4.20435
25	4879	15.8000	20.0043	-4.20435
26	6382	15.8000	20.0043	-4.20435
27	5219	15.8083	19.9957	-4.18732
28	9529	15.8125	19.9913	-4.17880
29	9851	15.8250	19.9783	-4.15326
30	2956	19.9000	15.7261	4.17391
31	5699	19.9167	15.7087	4.20797
32	5732	19.9250	15.7000	4.22500
33	5178	19.9417	15.6826	4.25906
34	7482	19.9667	15.6565	4.31014
35	5554	19.9667	15.6565	4.31014
36	942	19.9833	15.6391	4.34420
37	8026	19.9833	15.6391	4.34420
38	2095	20.0583	15.5609	4.49746
39	3327	20.0833	15.5348	4.54855
40	330	20.1000	15.5174	4.58261
41	1453	20.1042	15.5130	4.59112
42	4095	20.1625	15.4522	4.71033
43	8275	20.1625	15.4522	4.71033
44	6831	20.2375	15.3739	4.86359
45	254	20.4875	15.1130	5.37446

Figure 4: Table of Results as Extreme as Observed

6. Interpret the results of the test in the context of the research question.

A:

Having a p-value of 0.0045 means that we calculate the probability that randomization alone leads to a test statistic as extreme as or more extreme than the one observed in the study (of the mean score of intrinsic group being 4.1484 higher than the mean score of the extrinsic group) is 0.0045, or slightly less than half of one percent. Thus, we saw evidence that there is a treatment effect between the intrinsic and extrinsic treatment, which in turn lends support to the belief that receiving the intrinsic treatment during the study contributed to an average increase in the score (of the haiku). (**I am not saying there is a 0.0045 probability of the null hypothesis being correct.**)

Q: 7. What aspects of the data collection in this experiment would need special attention by the researcher?

A:

Primarily I would want the researcher to be certain of the exchangeability of the data, as it would be vital to the appropriateness of the randomization test (permutation) done above. If in some way the independence of observations was not assured then the researcher may be led to false conclusions (certainly related to and impacted by the additional notes given below!)

Some additional asides Of particular note: Researchers should be especially weary of generalizing and applying the results of this study to a broader population of interest. I also have concerns (and project those concerns to the researcher) with regards to the selection of participants for the study. Specifically: Though participants are all "creative writers", there are many different ways in which one can be a "creative writer", and we would want to ensure our sample is representative of all possible (or typical) creative writers as well as having that representation within each treatment group.

I am also concerned about how the scores were generated: Though the scores are being generated from non-participants in the study (*"12 poets then scored these poems on a 40-point scale based on the creativity shown in the writing"*, meaning the score of a participant is the average score given by the poets), the score is given as a composite of multiple scores. This leads inevitably to some fuzziness.

There are two points I want to illustrate with this, one being the meaning of a score and the second being the potential variability of scores for an individual participant. First, what's the difference between a 14 and a 15? Is that difference the same as the difference between a 16 and a 17? Poets grade on a 40-point scale based on the creativity of a haiku, but how consistent is that scoring, and how vague are the underlying scoring criteria? Secondly, since we are using the mean value of the scores of an individual, these average scores can be dramatically affected by outliers. By comparison the median score would not be as affected by outliers, i.e. if there was one particular poet who tended to score participants especially low or high, then the overall distribution would be biased (not using the technical definition) by that one judge. This means if a participant was scored $\{2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 38\}$, they would have mean score 5 and median score 2. As a researcher I would want to at least be aware of this possibility and think through if the mean score is the proper metric of the data to collect.

Total: 20 points **# correct: %:**