1 Comparing two populations

1.1 (10 points)

The boxplots are shown in Figure 1, and the histograms in Figure 2. The responses do indeed look different. Specifically, the treated group seems to have a higher mean and smaller variance.

![Boxplots for Problem 1.1](image-url)

Figure 1: Boxplots for Problem 1.1
Figure 2: Histograms for Problem 1.1
1.2 (20 points)

(a) We find $\bar{Y}_2 - \bar{Y}_1 = 10.0$. Note that it is meaningless to report more digits of accuracy since the original data only contained 2 digits of accuracy.

(b) We use the following R code to sample from the distribution of $\bar{Y}_2 - \bar{Y}_1$ under the null hypothesis:

```
# First create a big matrix, each row with a copy of the data
bigm = matrix(data = Response, nrow=10000, ncol=44, byrow=T)
# Use the sample command to shuffle each of the rows
bigm = apply(bigm, 1, sample)
# Form a vector equal to the mean of the first 21 columns
# minus the mean of the latter 23 columns
md = apply(bigm[1:21,], 2, mean) - apply(bigm[22:44,],2,mean)
```

The resulting histogram is shown in Figure 3.

(c) We are interested in testing the null hypothesis $H_0: \mu_2 = \mu_1$ with the alternative $H_1: \mu_2 > \mu_1$. So we use a one-sided test on the test statistic $\bar{Y}_2 - \bar{Y}_1$. If we chose the alternative hypothesis to be $\mu_2 \neq \mu_1$, then it would a two-sided test.

After computing the approximate distribution above, we can use the following R command to estimate the probability that the $\bar{Y}_2 - \bar{Y}_1$ is greater than or equal to the observed value under the null hypothesis. (Here omd is the observed mean difference from part (a).)

```
length(md[md >= omd])/length(md)
```

We get an approximate p-value of 1.4%. This is significant evidence to reject the null hypothesis in favor of the alternative. This implies that the special activities help.

1.3 (10 points)

First we assume that all the scores are independent, which is reasonable since presumably each student took the test individually. Also we assume the scores in each population are identically distributed, which is not only reasonable but necessary since there is no other data to distinguish them by. Lastly, the strongest assumption we need for the t-test is that we would like to treat the scores as being normally distributed.

We can actually make an appeal to the Central Limit Theorem to explain why test scores often are normally distributed. Suppose a test has many problems, each equal the same number of points and each problem tests a different fact or skill. If $X_{ij}$ is a random variable equal to the number of points that
Figure 3: For Problem 1.2.b. Histogram of $\bar{Y}_2 - \bar{Y}_1$ under $H_0$. 

$Y_2bar - Y_1bar$, Frequency
$-15 -10 -5 0 5 10 15$

0 500 1000 1500
student $i$ scores on problem $j$ of a test, and $Y_i$ is that student’s score, then of course $Y_i = \sum_j X_{ij}$. So as long as the $X_{ij}$ are not too strongly correlated, then $Y_i$ will tend to be normally distributed. (There are much weaker forms of the CLT that don’t require independence.) The key is that the scores are sums of many different relatively independent variables.

Even more crudely, we can simply look at the histograms and see that they at least look roughly normal, or at least unimodal.

Since there is no reason to believe that the variances of the two populations are equal, we use Welch’s test. This is the test that R computes by default for a two-sample t test. But it’s not difficult to do the same calculation by hand. We compute the test statistic:

$$t = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

For our data, we have $N_1 = 21$, $N_2 = 23$, $s_1^2 = 121$, and $s_2^2 = 294$. So we compute $t = 2.3$. If $Y_2$ and $Y_1$ are normal, under the null hypothesis this test statistic can be approximated as having a t-distribution with $\nu$ degrees of freedom where $\nu$ is given by the formula:

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1(N_1-1)} + \frac{s_2^4}{N_2(N_2-1)}}$$

For this data, we compute $\nu \approx 38$. Again, we are interested in a one-sided test, so we compute $p = P(t_{38} \geq 2.3) = 1.3\%$. Based on this, once more we reject $H_0$ in favor of $H_1$.

1.4 (10 points)

Both tests lead to the same conclusion: there is clear evidence that the activities have a beneficial effect on the students’ test scores. The fact that they both give the same result does not strengthen the conclusion too much since both tests use the same data, just slightly different assumptions (the permutation test makes no assumption of normality).

2 Random Numbers

2.1 (10 points)

The histograms are shown in Figure 4. They look reasonably uniform, given that we only have 1000 points for a histogram with 10 bins. For this much random data, it is to be expected that the histogram doesn’t perfectly match the distribution function.

To be a bit more precise, if the data were truly uniform [0, 1] distributed, the height of a bar of our histogram would be Bernoulli ($n = 1000, p = 1/10$)
Figure 4: Histograms of randu columns
distributed. Such a random variable has a mean of \( np = 100 \) and a standard deviation of \( \sqrt{np(1-p)} \approx 9.5 \). While the bar heights of the histogram are slightly correlated, this calculation suggests it’s reasonable to expect a deviation of 9.5% of the mean height. By contrast, if we had a million points and plotted a histogram with 10 bins, we’d expect a deviation of .3% of the mean height; in that case the histogram should be much more flat.

2.2 (20 points)

If Figure 5, we plot the histograms against the beta distribution for particular values of \( a \) and \( b \). For columns 1,2,3 we choose respectively \((a, b) = (0.94, 1.05), (1.12, 1.11), \) and \((0.97, 1.00)\).

![Histograms with fitted beta distributions. Parameters chosen by method of moments.](image)

While it was enough for this assignment to just eyeball the fit, there are some standard methods which are systematic. Since there were a lot of questions about this, we’ll go over two such methods.
2.2.1 Method of Moments

One strategy is to make use of the relationship between the parameters of our family of random variables and their first few moments. In our case, if $X$ is $B(a, b)$ distributed, we have:

$$EX = \frac{a}{a + b}$$

$$\text{Var } X = \frac{ab}{(a + b)^2(a + b + 1)}$$

Note that the above system of equations can be inverted; we can solve for $a$ and $b$ in terms of $EX$ and $\text{Var } X$. Suppose we do that and then in the resulting formulas replace the mean and variance with the sample mean and sample variance respectively. Then we get the following estimates for $a$ and $b$:

$$\hat{a} = \bar{x} \left( \frac{(1 - \bar{x})}{s^2} - 1 \right)$$

$$\hat{b} = (1 - \bar{x}) \left( \frac{(1 - \bar{x})}{s^2} - 1 \right)$$

2.2.2 Maximum Likelihood Estimate

Another standard estimation procedure is to calculate the values of $a$ and $b$ which maximize the density function at the observed values. Since the data points are presumed independent, this means maximizing the following function:

$$F(a, b) = \prod_{i=1}^{n} f(x_i | a, b)$$

Here $f$ is the density of the beta distribution as given in the problem statement. There is no closed form expression for the maximizing values of $a$ and $b$, so this needs to be computed numerically. Fortunately, the function $F$ does have a nice property that makes it conducive to such numerical optimization (the log of it is concave).

2.3 (10 points)

The scatterplots are shown in Figure 6. The points look uniformly distributed throughout each square area, which means they are pairwise independent and uncorrelated.

2.4 (10 points)

The plot is shown in Figure 7. Let’s define $w = 9x - 6y + z$. First, while it is clear from the plot $w$ and $x$ are correlated, that’s not very interesting since it’s to be expected by the definition of $w$. What is interesting is that $w$ has a discrete distribution. Specifically, it always equals an integer. But if $x, y, z$ were
Figure 6: Pairwise scatterplots
truly uniformly distributed independent random variables, then $w$ would have a continuous distribution. So we conclude that $x, y, z$ are in fact not independent. It’s not quite accurate to say they’re correlated, since correlation only involves linear relations between variables, and relation we observe is nonlinear.

Figure 7: Plot of $x$ vs. $w = 9x - 6y + z$

Aside: why is it true that $w$ should have a continuous distribution? Remember that if $a$ and $b$ are independent random variables the density of $a + b$ is given by the convolution of the density for $a$ with the density for $b$. So the exact formula for what the density of $w$ should be is given by the convolution of the densities for $9x$ and $-6y$ and $z$. Furthermore, the convolution of continuous densities is continuous.

There are a couple of lessons to be learned here. First, looking at the pairwise scatterplots of data is not enough to determine whether any patterns exist in higher dimensions. In this case, we can visualize what going on here by thinking about the points $(x, y, z)$ in a cube. All the points lie on a set of equidistant parallel planes that are normal to the vector $(9, -6, 1)$. (See http://en.wikipedia.org/wiki/Image:Randu.png for a picture of this.) Secondly, one needs to be mindful of how one generates and uses pseudorandom numbers in computations. The dataset comes from an infamous algorithm for generating pseudorandom numbers called ‘randu’. If we used these numbers for a Monte Carlo method, we might be led to some blatantly wrong conclusions, like that $P(.1 \leq w \leq .9) = 0$. 