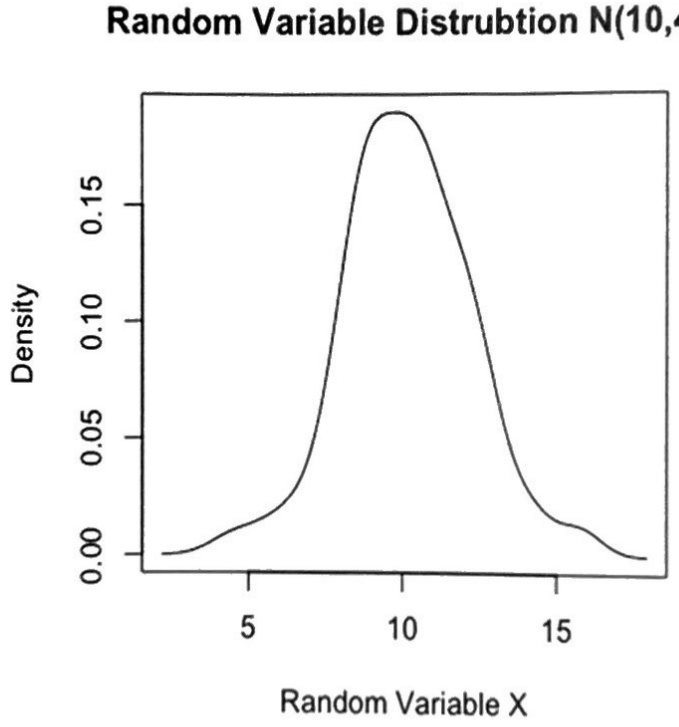
David Levy
11/15/17

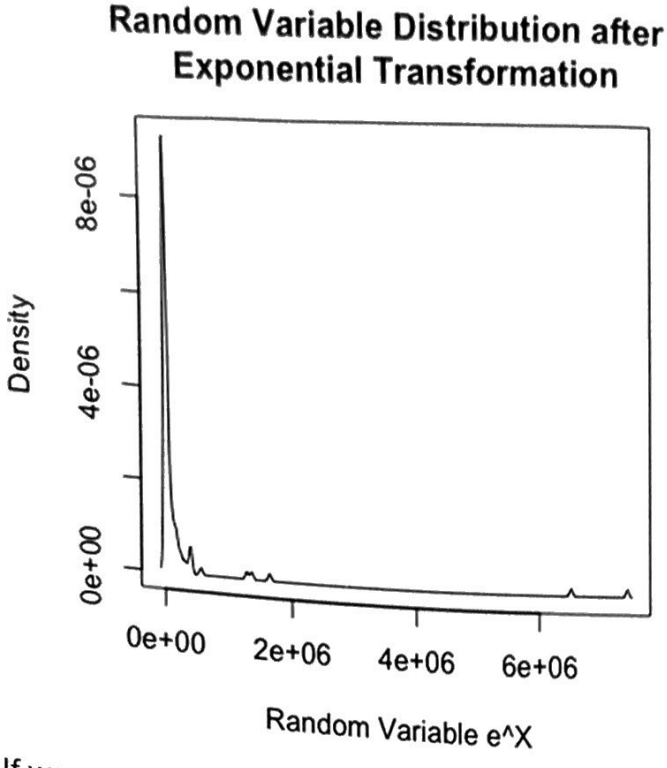
Biostatistics 200A
Problem Set 4

1.

a.

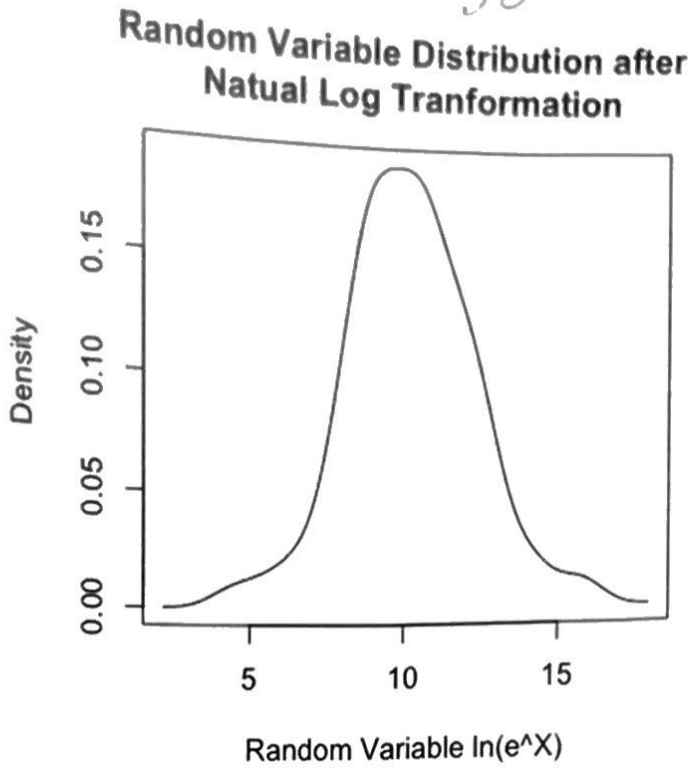
i. 100 random variables with mean 10 and $\sigma = 2$ were generated using R, and then transformed exponentially. The distributions of the untransformed random variable and the transformed random variable are shown in the graphs below. Clearly, the original data look normal, but the exponentially transformed data do not. The transformed distribution is positively skewed.





If we were given data that resembled the second plot shown above, especially since we know how this data was obtained, it would be wise to apply the natural log function to the data, and then display the resulting distribution to determine whether or not it is in fact lognormal. As demonstrated by the graph shown below, when the data is transformed back to its original state by applying a natural logarithm function, the original normally distributed random variable is restored.

45/50



ii. The R output for the calculations requested are shown below.

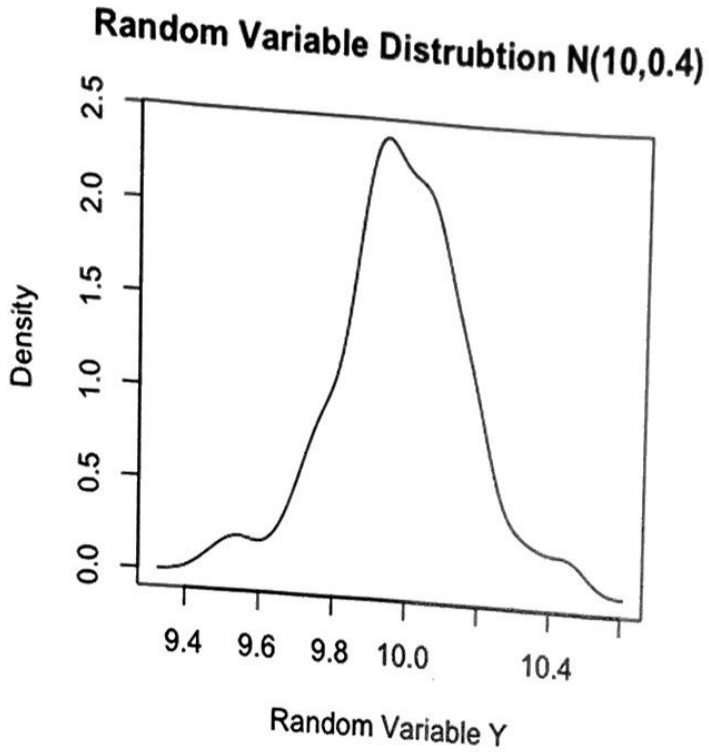
```
> samp_mean_x
[1] 10.18272
> samp_mean_etrans_x
[1] 248194.6
>
> samp_var_x
[1] 4.224834
> samp_var_etrans_x
[1] 1.002408e+12
>
> se_x
[1] 0.205544
> se_etrans_x
[1] 100120.3
```

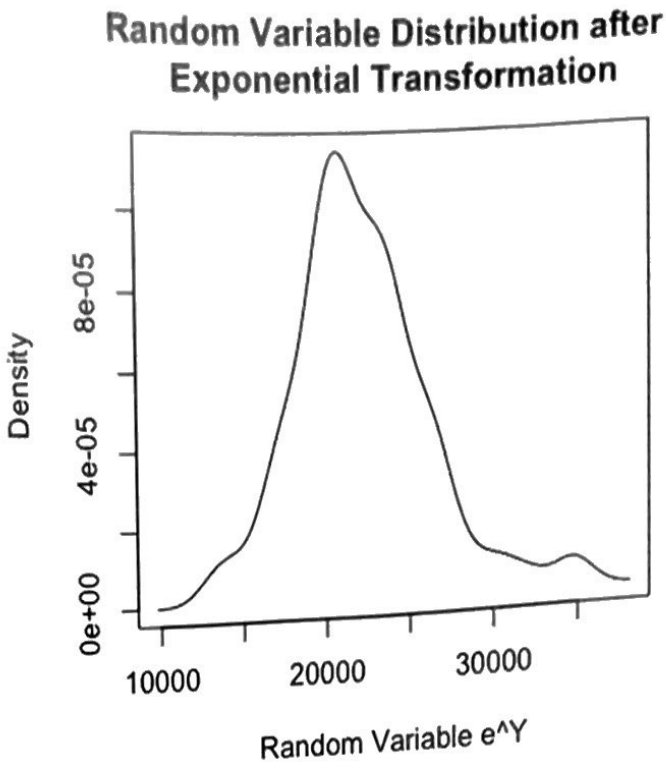
It is very easy to see that the sample variances are widely different. This is caused by the random variables lying on the far right of the original distribution. When we apply an exponential function, those values have a much greater effect on the mean of the transformed data than they did on the mean of the untransformed data. Thus, the

variance of the transformed data is huge, while the variance of the untransformed data is approximately 4, as we would expect given the population parameter we specified to generate our random variables.

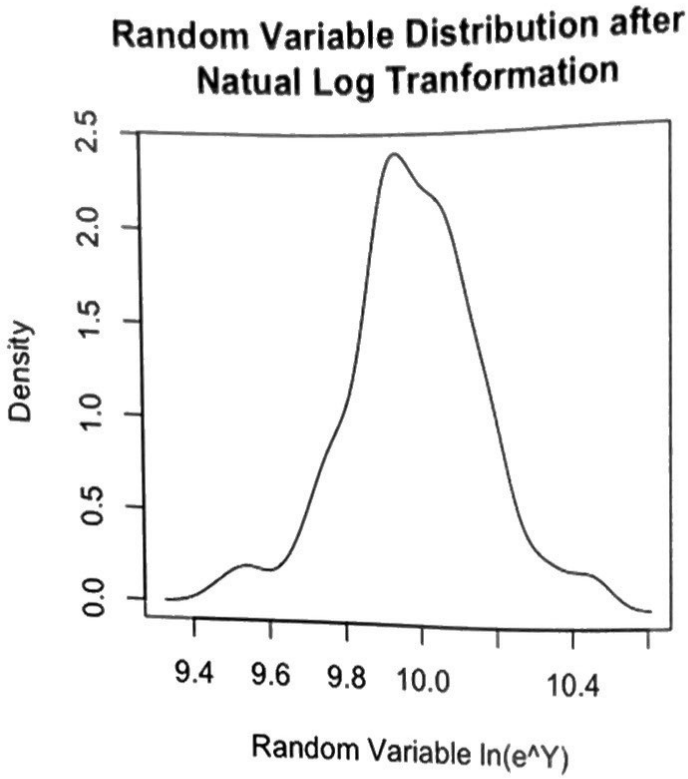
b.

i. 100 random variables with mean 10 and $\sigma = 0.2$ were generated using R, and then transformed exponentially. The distributions of the untransformed random variable and the transformed random variable are shown in the graphs below. Unlike our results in part (a), the original data *and* the transformed data look normal, though they clearly have different means and variances.

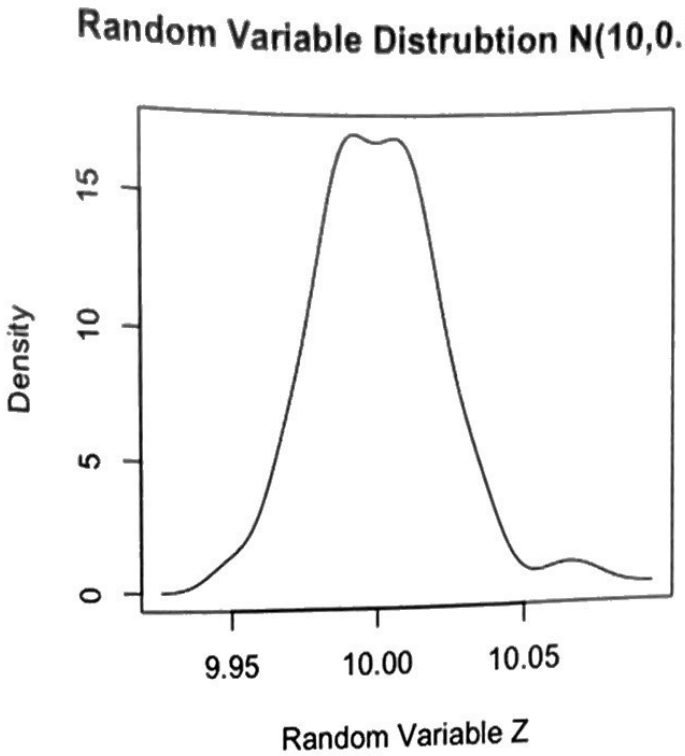


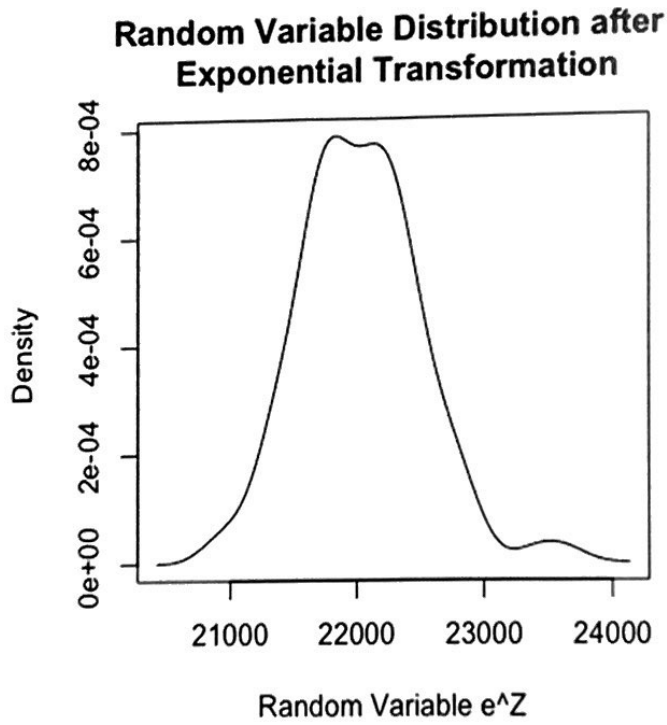


If we were given data that resembled the second plot shown above, it would not be immediately necessary to transform the data, as it already appears normal. However, the graph shown below demonstrates that when the data is transformed back to its original state by applying a natural logarithm function, the original normally distributed random variable is still restored.

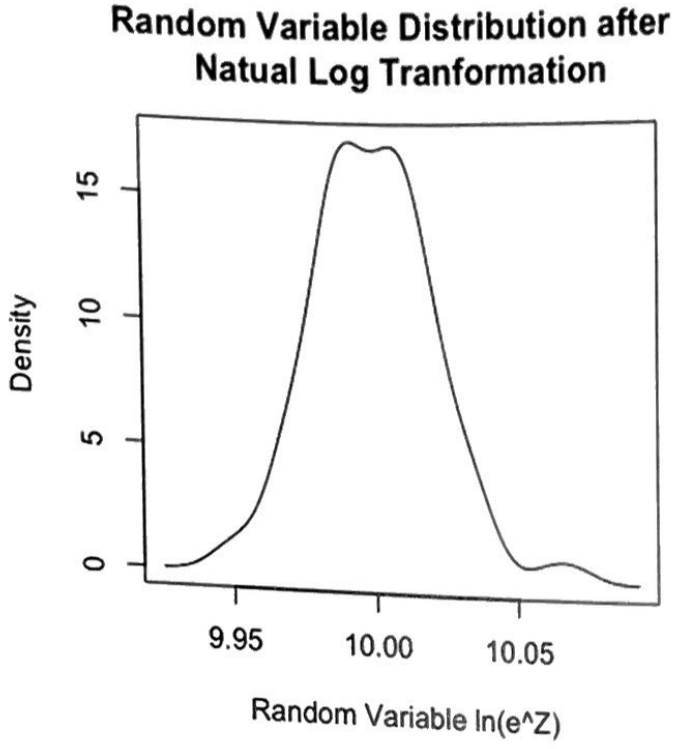


Next, 100 random variables with mean 10 and $\sigma = 0.02$ were generated using R, and then transformed exponentially. The distributions of the untransformed random variable and the transformed random variable are shown in the graphs below. Unlike our results in part (a), and similar to the results shown for the previous case, the original data *and* the transformed data look normal, though they clearly have different means and variances.


Random Variable Distrubtion N(10,0.04)
Density
0
5
10
15
9.95
10.00
10.05
Random Variable Z


Random Variable Distribution after
Exponential Transformation
Density
0e+00
2e-04
4e-04
6e-04
8e-04
21000
22000
23000
24000
Random Variable e^Z

If we were given data that resembled the second plot shown above, it would not be immediately necessary to transform the data, as it already appears normal. However, the graph shown below demonstrates that when the data is transformed back to its original state by applying a natural logarithm function, the original normally distributed random variable is still restored.



ii. The R output for the calculations requested are shown below. Note that as above, Y represents the random variable for the case where $\sigma = 0.2$ and Z represents the random variable for the case where $\sigma = 0.02$.

```
> samp_mean_y
[1] 9.986961
> samp_mean_z
[1] 9.999911
> samp_mean_etrans_y
[1] 22093.63
> samp_mean_etrans_z
[1] 22029.55
>
> samp_var_y
[1] 0.03254347
> samp_var_z
[1] 0.0004607994
> samp_var_etrans_y
[1] 16166737
> samp_var_etrans_z
[1] 225382.3
> se_y
[1] 0.01803981
> se_z
[1] 0.002146624
> se_etrans_y
[1] 59745.86
> se_etrans_z
[1] 47.47445
```

Clearly, the sample variances are different for both Y and Z. In both instances, the sample variance of the untransformed data is very close to the parameter we used to generate our vector of random variables. The variance of the transformed data is huge in both instances. There is, however, a general decreasing trend in the variances of the transformed data as we decrease the population variance of the untransformed random variable.

2.

a. Testing the equivalence of variances requires an F-test. For an F-test, we present the hypotheses: H_0: ${\sigma_1}^2 = {\sigma_2}^2$; H_1: ${\sigma_1}^2 \neq {\sigma_2}^2$. The F-statistic is calculated as follows: $F = \frac{{s_1}^2}{{s_2}^2}$. The R output for this calculation is shown below:

```
> F_stat
[1] 1.404664
```

In order to obtain a p-value for our F-test, we should observe that the mean of the F-distribution is 1, so that our p-value should be twice the probability of obtaining an F-statistic as large or larger than the one displayed above. The output for the p-value we calculated in R is displayed below:

```
> p_val_F
[1] 0.3518248
```

Alternatively, we could observe that for a 95% confidence-level, we would need an F-statistics greater than or equal to the one displayed below:

```
> comp_F
[1] 1.826764
```

In either case, we can tell that we are unable to reject the null hypothesis, and we can conclude that the assumption of equal variance is reasonable at a 95% confidence-level.

b. In order to test whether the mean scores are the same, we will elect to perform a two-sided, two-sample t-test. The conclusion of our F-test above supports the legitimacy of our assumption of equal variance. We do, however, have to assume independence of the samples and normality of the underlying probability distributions as well. For the purposes of completing the task at hand, we will accept those assumptions without further investigation. Our hypotheses are: $H_0: \bar{x}_1 = \bar{x}_2$; $H_1: \bar{x}_1 \neq \bar{x}_2$. We proceed by calculating a t-statistic, $t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, where $s = \sqrt{\frac{{s_1}^2(n_1-1) + {s_2}^2(n_2-1)}{n_1+n_2-2}}$ represents the pooled estimate of variance for two independent samples. The p-value is determined by finding the probability of obtaining a t-statistic as extreme as or more extreme than the one we calculated, and multiplying it by two. The R output for our p-value calculation is shown below:

```
> #b
>   diff_means <- RA_mean - OA_mean
> pooled_var <- ((((RA_n-1)*RA_sd^2) + ((OA_n-1)*OA_sd^2))/deg_frdm)
> t_stat <- diff_means/(sqrt(pooled_var*((1/RA_n)+(1/OA_n))))
> p_val_t <- 2*pt(t_stat, deg_frdm, lower.tail = FALSE)
> p_val_t
[1] 0.2268912
```

The p-value we calculated is not reasonable for rejecting the null hypothesis at the 95% confidence level, so we fail to reject the null hypothesis that the means of the two groups are equal.

c. In order to determine the number of subjects needed to detect a difference of 1 unit at 90% power with equal sample sizes and a two-sided test with significance level $\alpha = 0.05$, we use the following formula: $n = \frac{(\sigma_1{}^2+\sigma_2{}^2)(Z_{1-\frac{\alpha}{2}}+Z_{1-\beta})^2}{\Delta^2}$, where n represents the sample size per group and Δ represents the desired difference to detect. All other valuables are self-explanatory. The R output for our per-group sample size calculation is shown below:

```
> #c
>   desired_diff <- 1
> beta <- 0.1
> alpha <- 0.05
> n <- ((RA_sd^2+OA_sd^2)*(qnorm(1-alpha/2)+qnorm(1-beta))^2)/(desired_diff^2)
> n
[1] 184.1951
```

Thus, we round up to the nearest whole number and recommend a sample size of 185 subjects per group.

3.

a. We observe that two-sided power can be written as the probability of rejecting the null hypothesis given that the alternative is true. We should also observe that the test being performed in this instance would be a t-test for independent samples with unequal variances. And since both variances are known, we would compute a z-statistic. So $\text{Power} = 1 - \beta = P\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_A{}^2}{fN}+\frac{\sigma_B{}^2}{(1-f)N}}} > 1.96 \middle| \mu_A - \mu_B = \Delta\right) + P\left(\frac{\bar{X}+\bar{Y}}{\sqrt{\frac{\sigma_A{}^2}{fN}+\frac{\sigma_B{}^2}{(1-f)N}}} < -1.96 \middle| \mu_A - \mu_B = \Delta\right)$. Clearly, in order to maximize power, we need the $\sqrt{\frac{\sigma_A{}^2}{fN}+\frac{\sigma_B{}^2}{(1-f)N}}$ term to be as small as possible. We achieve this by minimizing $\frac{\sigma_A{}^2}{fN}+\frac{\sigma_B{}^2}{(1-f)N}$. If we take this to be a function of f, we can use calculus methods to determine a minimum. More specifically, we can state that a function for the slope of the tangent line (the derivative) is

$\frac{{\sigma_B}^2}{f^2N} - \frac{{\sigma_A}^2}{(1-f)^2N}$. We find a minimum when $\frac{{\sigma_B}^2}{f^2N} = \frac{{\sigma_A}^2}{(1-f)^2N}$. Algebra yields $f = \frac{\sigma_A}{\sigma_A+\sigma_B}$.

b. Using equal sample sizes for each group can be a good choice when the variances of independent samples are equal (or close to equal). The maximization of power in this instance is demonstrated above. When variances are significantly different, however, it is clear that power suffers, and that we would do better to assign unequal numbers of subjects to each group in order to maximize power, using the fraction calculated above. Equal sample size also carries with it the benefits of reducing the complexity of the study design, and allows for the possibility of certain designs that require pairing of comparable individuals, which benefit from the fact that if groups are assigned randomly, confounders will likely play equal roles in each group. All in all, equal sample sizes per group can help to maximize power, efficiency and study protocols, but may also have negative effects on the quality of the study if certain conditions are not met, and if selection bias is not carefully avoided.

4.

In the S.U.D. group, the median duration of labor was 10 minutes. In the control group, the median duration of labor was 20 minutes. In this situation, medians are preferred because the data is imprecise and clearly has outliers. Outliers will have a much stronger effect on means than medians, making median the preferred measure of central tendency in this case. While we perform a Wilcoxon rank-sum test, it is important to note that our hypotheses are: H_0: $f_{SUD} = f_{Control}$; H_1: $f_{SUD} \neq f_{Control}$.

Rank	Data Value	Rank	Data Value
17	60	12	20
14	25	10.5	15
2	6	3	7
4	8	18	75
1	<5	20	120*
7	10	7	10
14	25	19	100
10.5	15	5	9
7	10	14	25
9	13	16	30

The rank-sum of the S.U.D. and control groups are $W_{SUD} = 76.5$ and $W_{Control} = 133.5$, which correctly sum to $\frac{20 \cdot 21}{2} = 210$. We know that $E(W_{SUD}|H_0) = \frac{m(N+1)}{2} = \frac{9(21)}{2} = 94.5$ and $E(W_{Control}|H_0) = \frac{n(N+1)}{2} = \frac{11(21)}{2} = 115.5$. Without a p-value, we can intuitively state that this data likely represents a significant deviation from what we would expect to see under the null hypothesis. Finding a p-value would be difficult without using the normal approximation, which requires that both groups' sample sizes be greater than 10. Our data do not meet that requirement. R confirms our intuition:

```
        Wilcoxon signed rank test with continuity correction

data:  comb
V = 210, p-value = 9.449e-05
alternative hypothesis: true location is not equal to 0

Warning message:
In wilcox.test.default(comb, conf.level = 0.95) :
  cannot compute exact p-value with ties
```

-2

p-value should be around .18

5.

a. When the null hypothesis that the distributions of X and Y are identical is true, we know that $E(W_X|H_0) = \frac{m(N+1)}{2} = \frac{4(7)}{2} = 14$ and $E(W_Y|H_0) = \frac{n(N+1)}{2} = \frac{2(7)}{2} = 7$. These expected values give us the distribution of the Wilcoxon rank-sum test statistics when the null hypothesis is true.

-3

b. According to Table H in the class notes, we see that if we use the entry where $n = 4$, $m = 2$, and $T_x = 10$, we obtain a p-value of 0.133 from the table. Thus, our two-sided p-value is $p = 0.266$. We would conclude that there is no significant difference between the two groups at a 95% confidence-level. In other words, we are unable, given this data, to reject the null hypothesis that the distributions for X and Y are identical.

✓

Problem is asking for distribution of W_x i.e.

$P(W_x = 3) = ?$ $P(W_x = 4) = ?$ ...