1. (a) From the histogram, the two samples roughly follow normal distribution. We would like to test if the variances are same using

\[
\text{var.test}(x,y, \text{alternative="two.sided"})
\]

However, the given code is

\[
\text{var.test}(x,y, \text{alternative="less"})
\]

The two tests have the same F statistic with same F distribution under null, so the one-sided test has all the information needed for the two-sided test. You either compare \( F = 0.0494 \) to rejecting region

\[
c(qf(0.025,99,99), qf(0.975,99,99))
\]

Or directly use the connection between p-values

\[
p_{2\text{-sided}} = 2 \times \min(p_{1\text{-sided}}, 1 - p_{1\text{-sided}}) < 4.4e - 16
\]

Conclusion should be that they don’t have the same variance.

(b) The two sample t-test statistic with unpooled variances

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{100} + \frac{s_y^2}{100}}} \\
\sim t_{df}
\]

The degree of freedom is computed by Welch’s approximation.

(c) \[
\bar{X} - \bar{Y} = 1.449709 - (-49.42237) = 50.872079 \\
s_x^2 = 8.804076^2 = 77.5117542
\]

From part a), we know

\[
F = \frac{s_x^2}{s_y^2} = 0.04938247
\]

Then,

\[
s_y^2 = s_x^2/0.04938247 = 1569.6208435
\]

The test statistic is

\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{100} + \frac{s_y^2}{100}}} = 3.0885236
\]

The degree of freedom is \( 108.7539428 \). The test is one-sided, so the rejecting region is \( \geq 1.6589856 \), the 95% quantile of \( t_{108.7539428} \). As the conclusion, \( H_0 \) is rejected. \( \mu_X \) is significantly bigger than \( \mu_Y \).

(d) \[
CI = \bar{X} - \bar{Y} \pm t_{108.7539428,0.01}\sqrt{\frac{s_x^2}{100} + \frac{s_y^2}{100}} \\
= [11.9812057, 89.7629523]
\]
To give a rigorous asymptotic proof is beyond the context of this class. The Central Limit Theorem idea is the normal approximation for \( \bar{X} \) and \( \bar{Y} \) with large sample sizes.

\[
\bar{X} \approx N(\mu_X, \sigma_X^2/m), \quad \bar{Y} \approx N(\mu_Y, \sigma_Y^2/n)
\]

Since \( \bar{X} \) and \( \bar{Y} \) assumed independent, the difference of the two independent Normal r.v.s above is

\[
\bar{X} - \bar{Y} \approx N(\mu_X - \mu_Y, \sigma_X^2/m + \sigma_Y^2/n)
\]

Properly standardize it

\[
\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/m + \sigma_Y^2/n}} \approx N(0, 1)
\]

Under \( H_0 \), \( \mu_X - \mu_Y = 0 \). And we can use consistent estimators \( s_X^2 \) and \( s_Y^2 \) to replace unknown \( \sigma_X^2 \) and \( \sigma_Y^2 \). The resulting statistic is exactly

\[
T = \frac{\bar{X} - \bar{Y} - 1}{s_p \sqrt{1/25 + 1/25}} \sim t_{48}
\]

2. This is a Pearson’s \( \chi^2 \) independence test for contingency table. Construct the table as following
The row marginal probabilities are (0.3, 0.5, 0.2). The column marginal probabilities are (0.4, 0.4, NA).
The expected counts are
Oops, the minimum expected count drop beyond 5. As a math student, you should refuse to
give a conclusion before more data are provided.

3. (a) For this part, suppose \( m = n = 25 \). From the story, the two groups are independent.
The hypothesis testing problem is

\[
H_0 : \mathbb{E}X - \mathbb{E}Y \leq 1 \text{ versus } H_1 : \mathbb{E}X - \mathbb{E}Y > 1
\]

In the pearson’s test for goodness-of-fit (by default, null is Normal distribution), both p-values are more than 0.05, so we may conclude that both samples follow roughly Normal distribution. In the F-test, we can calculate the 2-sided p-value 0.7752 > 0.05 following
the argument in problem 1. The variances can be assumed to be the same.
Therefore, it is a independent two sample for Normal mean with equal variance. The
test statistic is

\[
T = \frac{\bar{X} - \bar{Y} - 1}{s_p \sqrt{1/25 + 1/25}} \sim t_{48}
\]

\[
s_p = \sqrt{(3.2^2 \times 24 + 2.8^2 \times 24)/48} = 3.0066593
\]

The test statistic is 1.2294153. The 95% quantile of \( t_{48} \) is 1.6772242. \( T < t_{48,.05} \), so we
fail to reject \( H_0 \).

(b) \[ 1-\text{pt}(t3, 48) \]

The p-value is 0.1124543.
2. (c) Missing
(d) Under null, $T$ follow $t_{48}$. It is bell-shaped and centered at 0. Therefore, the right one best corresponds to the distribution under null.

4. (a) The two histograms have similar shape under similar scale, so they should have similar variances.
(b) The two samples roughly follow the Normal distribution. They have similar variances. But we still cannot decide which test is proper for the hypothesis testing problem. The missing element here is if the two samples are paired or independent.
Suppose they are independent samples. Then the $t$ statistic with pooled variance should be used
\[ T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{1/m + 1/n}} \sim t_{m+n-2} \]
(c) $\bar{X} - \bar{Y} < 0$, so the observed statistic $T_{obs} < 0$. The test rejects for big $T$, so the p-value for the is the right-tail probability $1 - F(T_{obs})$, $F(\cdot)$ the c.d.f. for $t_{m+n-2}$. It should be greater than 0.5, as $F(0) = 0.5$.
Also, from the two histograms, the difference of mean is very big in reference to standard deviation. $T_{obs}$ must be far in the negative direction. The p-value should be expect to be close to 1.
(d) By definition, the p-value is
\[ 1 - pt(T_{obs}, m+n-2) \]
Or, you may use std Normal distribution instead of $t_{m+n-2}$ because $t$-distribution converges to std Normal when sample size getting big.
(e) Same as problem 3 d)

5. (a) According to the story, the same 50 bags are measured twice. If the labels for each bag can be traced, this should be a paired-test. Otherwise, we can assume independence between samples. (Under random reshuffle, you may show the correlation of $X_i$ and $Y_i$ are less than $1/50$)
From the histogram, the samples are roughly Normal. So we may assume $X$, measure before shipment, and $Y$ measure after shipment, are from Normal with mean $\mu_X$ and $\mu_Y$. By default, the null should be the status quo $H_0 : \mu_X = \mu_Y$. Since we are interested in the loss of concentration, the alternative should be $H_1 : \mu_X > \mu_Y$.
(b) The output is the F-test for
\[ H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 < \sigma_Y^2 \]
Well this is not what we care about. Following the same argument in problem 1 a), we can learn about the data from this output and do the test:
\[ H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2 \]
The two tests have the same F statistic with same F distribution under null, so the one-sided test has all the information needed for the two-sided test.
You either compare \( F = 1.4507 \) to rejecting region
\[
c(qf(0.025,49,49), qf(0.975,49,49))
\]
Or directly use the connection between p-values
\[
p_{2\text{-sided}} = 2 \times \min(p_{1\text{-sided}}, 1 - p_{1\text{-sided}}) = 0.1964
\]
Conclusion should be that they have the same variance.

(c) The output is the T-test with unpooled variances for
\[
H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X < \mu_Y
\]
Well this is not what we care about. we can learn about the data from this output and do the test with pooled variance:
\[
H_0 : \mu_X = \mu_Y \text{ versus } H_1 : \mu_X > \mu_Y
\]
If you insist doing the paired test, there is no information about correlation, so you may refuse to work any further.
Otherwise, from the previous part, we learn
\[
s_X^2 = 42.6801151
\]
and
\[
F = \frac{s_X^2}{s_Y^2} = 1.450689
\]
so we can calculate
\[
s_Y^2 = 29.4205823
\]
The pooled variance is
\[
s_p^2 = \frac{49s_X^2 + 48s_Y^2}{98} = 36.0503487
\]
From the t-test output, we learn
\[
T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_X^2/50 + s_Y^2/50}} = 8.3788
\]
So
\[
\bar{X} - \bar{Y} = T\sqrt{s_X^2/50 + s_Y^2/50} = 10.0615886
\]
The t statistic with pooled variance is
\[
T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{1/50 + 1/50}} = 8.3788 \sim t_{98}
\]
The p-value is \( 1.9595436 \times 10^{-13} < .05. \)
(d) The \( H_0 \) is rejected. There is a loss of concentration.

4
6. You may either use a Pearson’s test of independence for contingency table, or a two sample Binomial test.

The contingency table is

The row marginal probabilities are (0.6448598, 0.3551402, NA). The column marginal probabilities are (0.970405, 0.970405, NA).

The expected counts are

The minimum expected count is above 5. The Pearson’s $\chi^2 = 2.4978728$ out of 1 degree of freedom. The 95% quantile for $\chi^2_1$ is 3.8414588. We fail to reject $H_0$. The birth defects are independent of the contaminated well.

Or, we can do a two sample binomial test. $X = 16 \sim Binom(414, p_1), Y = 3, \sim Binom(228, p_2)$.

Want to test $H_0 : p_1 = p_2$ versus $H_1 : p_1 \neq p_2$

\[
\hat{p}_1 = 16/414 = 0.039, \quad \hat{p}_2 = 3/228 = 0.013, \quad \hat{p} = (3 + 16)/(414 + 228) = 0.03
\]

The test statistic is

\[
Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/414 + 1/228)}} = 1.8237895 \approx N(0, 1)
\]

At level $\alpha = 0.05$, $Z_{\alpha/2} = 1.96$. $|Z| < Z_{\alpha/2}$, so we fail to reject $H_0$. The birth defects are independent of the contaminated well.