Problem 1. (10 pts.) (a) Let $x =$ number of heads
Model: $x \sim \text{binomial}(12, \theta)$. 
Null distribution $\text{binomial}(12, 0.5)$. 
Data: 3 heads in 12 tosses. 
Since $H_A$ is one-sided the rejection region is one-sided. Since $H_A$ says that $\theta$ is small it predicts a small number of heads in 12 tosses. That is, we reject $H_0$ on a small number of heads. 
So, rejection region = left tail of null distribution. 
\[ c_{0.95} = \text{qbinom}(0.05, 12, 0.5) - 1 = 2 \]
Rejection region is $0 \leq x \leq 2$. 
\[ p = \text{pbinom}(3, 12, 0.5) = 0.072998 \]

![Binomial(12,15) null distribution and rejection region $x \leq 2$.](image)

Ani concludes that there is not enough evidence to reject the null hypothesis at the significance level 0.05.

(b) Let $n =$ number of tosses that were tails before the third that is heads 
Probability model: Choose two tosses in the first $n+2$ for heads; the $n + 3$rd toss must be heads: 
\[ p(n) = \binom{n+2}{2}(1-\theta)^n \theta^3. \]
This is called the negative binomial distribution with parameters 3 and $\theta$. Data: 9 tails to get 3 heads. 
Since $H_A$ is one-sided the rejection region is one-sided. Since $H_A$ is that $\theta$ is small it predicts a large number of tails before 3 heads. So we reject on a large number of tails. 
Note: \text{pbinom}, \text{qbinom}, \text{etc.} are for the negative binomial distribution in R. 
Rejection region = right tail of null distribution. We compute $c_{0.05} = \text{qnbinom}(0.95, 3, 0.5) + 1 = 9$. So the rejection region is $n \geq 9$. Our data is $x = 9$, being careful
not to make an off-by-one error we get

\[ p = 1 - \text{pnbinom}(8, 3, 0.5) = 0.032715 \]

Negative binomial null distribution and rejection region

Efrat rejects the null hypothesis in favor of \( H_A \) at significance level 0.05.

(c) No. Computing a \( p \)-value requires that the experiment be fully specified ahead of time so that the definition of ‘data at least as extreme’ is clear.

Having said that, it’s a shame to waste good data. You can still analyze the data for suggestive results. As long as you report everything honestly people can reach their own conclusions.

(d) Prior: \( \text{beta}(n, m) \) has pdf \( c \theta^{n-1}(1 - \theta)^{m-1} \)

Likelihood experiment 1: \( \left( \frac{12}{3} \right) \theta^3 (1 - \theta)^9 \)

Likelihood experiment 2: \( \left( \frac{11}{2} \right) \theta^3 (1 - \theta)^9 \)

Since the likelihoods are the same up to a constant factor the posterior has the same form

\[ c \theta^{n+3-1}(1 - \theta)^{m+9-1} \]

which is the pdf of a \( \text{beta}(n + 3, m + 9) \) distribution.

The two posteriors are identical. In the Bayesian framework the same data produces the same posterior.

(e) The main point is that in the frequentist framework the decision to reject or accept \( H_0 \) depends on the exact experimental design because it uses the probabilities of unseen data as well as those of the actually observed data.

Problem 2. (10 pts.) We use the \( \chi^2 \) test for \( \sigma^2 = 1 \).

Call the data \( x \).

\( \sigma^2 = 1 \)
\[ s^2 = \text{sample variance} = \text{var}(\text{data}) = 2.34 \]

\[ n = 12 = \text{number of data points} \]

\[ \chi^2\text{-statistic: } X^2 = \frac{(n-1)s^2}{\sigma^2} = 25.74 \]

The one-sided \( p \)-value is
\[ p = 1 - \text{pchisq}(X^2, n-1) = 0.0071. \]

Since \( p < 0.05 \) we reject the null hypothesis \( H_0 \) that \( \sigma^2 = 1 \) in favor of the alternative that \( \sigma^2 > 1 \).

**Problem 3.** (10 pts.) We do a \( \chi^2 \) test of goodness of fit comparing the observed counts with the counts expected from Benford’s distribution.

You can use either test statistic
\[ G = 2 \sum O_i \ln \left( \frac{O_i}{E_i} \right) \]
or
\[ X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]
where \( O_i \) are the observed counts and \( E_i \) are the expected counts from Benford’s distribution. The total count = 100.

<table>
<thead>
<tr>
<th>First digit ( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td>23</td>
<td>16</td>
<td>3</td>
<td>12</td>
<td>13</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>expected</td>
<td>30.1</td>
<td>17.6</td>
<td>12.5</td>
<td>9.7</td>
<td>7.9</td>
<td>6.7</td>
<td>5.8</td>
<td>5.1</td>
<td>4.6</td>
</tr>
<tr>
<td>( X^2 ) components</td>
<td>1.67</td>
<td>0.15</td>
<td>7.22</td>
<td>0.55</td>
<td>3.29</td>
<td>0.07</td>
<td>3.04</td>
<td>2.98</td>
<td>2.51</td>
</tr>
</tbody>
</table>

The \( \chi^2 \)-statistics are \( G = 22.72 \) and \( X^2 = 21.49 \).

There are 9 cells that must sum to 100 so the degrees of freedom = 8.

The \( p \)-value using \( G \) is
\[ p = P(G \text{ test stat } > G \mid H_0) = 1 - \text{pchisq}(G, 8) = 0.0037 \]

The \( p \)-value using \( X^2 \)
\[ p = P(X^2 \text{ test stat } > X^2 \mid H_0) = 1 - \text{pchisq}(X2, 8) = 0.0060 \]

Since (for both \( G \) and \( X^2 \)) \( p > \alpha \) we don’t reject \( H_0 \). There is not enough evidence to indict the 18.05 staff for embezzlement!

**Problem 4.** (10 pts.)

(a) We let \( x \) = the first set of 20 numbers and \( y \) the second. R makes it almost too easy. We give the command \( \text{var.test}(x, y) \). R then prints out

\[
\begin{align*}
F &= 0.9703, \text{ num df } = 19, \text{ denom df } = 19, \text{ p-value } = 0.9484 \\
\text{alternative hypothesis: true ratio of variances is not equal to 1}
\end{align*}
\]

95 percent confidence interval:
0.3840737 2.4515249
sample estimates:
ratio of variances
0.9703434

The $p$-value is 0.9484 with $F$-statistic 0.9703.

(b) We found the formula for the $F$ statistic for this test at

\[ s_x^2 = \text{var}(x) = 1.1302 \]
\[ s_y^2 = \text{var}(y) = 1.1647 \]

Our $F$-statistic is

\[ fstat = \frac{s_x^2}{s_y^2} = 0.9703 \]

The degrees are freedom are both 19. Since the $F$-statistic is less than 1, the $p$-value is

\[ p = 2*\text{pf}(fstat, 19, 19)) = 0.9484 \]

which matches our result in part (a).

Problem 5. (10 pts.)(a) Let’s specify the assumptions and hypotheses for this test.
We have 4 groups of data: Clean, 5-day, 10-day, full
Assumptions: Each group of data is drawn from a normal distribution with the same variance $\sigma^2$; all data is drawn independently.
$H_0$: the means of all the normal distributions are equal.
$H_A$: not all the means are equal.

The test compares the between group variance with the within group variance. Under the null hypothesis both are estimates of $\sigma^2$, so their ratio should be about 1. We’ll reject $H_0$ if this ratio is far from 1.

We used R to do the computation. Here’s the code.

```r
mns = c(1.32, 1.26, 1.53, 1.39)
v = c(0.56, 0.80, 0.93, 0.82)
m = 351 # number of samples per group
n = length(mns) # number of groups
msb = m*var(mns) # between group variance
msw = mean(v) # within group variance
fstat = msb/msw
df1 = n-1;
df2 = n*(m-1)
p = 1 - pf(fstat, df1,df2)
```
print(fstat)
print(p)

This produced an $F$-statistic of 6.09 and $p = 0.00041$. Since the $p$-value is much smaller than 0.05 we reject $H_0$.

(b) We compare 10-day beards with each of the others. In each case we have: $H_0$: the means are the same $H_A$: the 10-day mean is greater than the other mean.

Note carefully that this is a one-sided test while the $F$-test in part (b) is a two-sided test.

From the class 19 reading we have the $t$-statistic for two samples. Since both samples have the same size $m = 351$ the formula looks a little simpler.

$$t = \frac{\bar{x} - \bar{y}}{s_{x-y}},$$

where the pooled sample variance is

$$s_P^2 = \frac{s_x^2 + s_y^2}{m}.$$ 

Note: the test assumes equal variances which we should verify in each case. This raises the issue of multiple tests from the same data, but it is legitimate to do this as exploratory analysis which merely suggests directions for further study.

The following table shows the one-sided, 2-sample $t$-test comparing the mean of the 10-day growth against the other three states.

<table>
<thead>
<tr>
<th></th>
<th>$t$-stat</th>
<th>one-sided $p$-value</th>
<th>$F$-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>3.22</td>
<td>0.00066</td>
<td>10.39</td>
</tr>
<tr>
<td>5-day</td>
<td>3.85</td>
<td>0.00007</td>
<td>14.79</td>
</tr>
<tr>
<td>full</td>
<td>1.98</td>
<td>0.0239</td>
<td>3.93</td>
</tr>
</tbody>
</table>

We also give the $F$-statistic for the two samples. You can check that the $F$-statistic for two-samples is just the square of the $t$-statistic.

The significance level of this experiment is greater than 0.01. That is, since we ran 3 tests each with probability 0.01 of a type I error the total probability of type I error is greater than 0.01—it will be close to 0.03.

We reiterate that with multiple testing the true significance level of the test is larger than the significance level for each individual test.