Stat 310 Homework 8 Key

Chapter 8, problems 47, 48, 50. Chapter 9, problems 1, 2, 5, 6, 9 12, 13. Due 11/23/99.

8.47 For two factors — starchy or sugary and green base leaf or white base leaf — the following counts for progeny were observed (Fisher 1958):

Type	Count
Starchy green	1997
Starchy white	906
Sugary green	904
Sugary white	32

According to genetic theory, the cell probabilities are $.25(2+\theta)$, $.25(1-\theta)$, $.25(1-\theta)$, and $.25\theta$, where $\theta(0 < \theta < 1)$ is a parameter related to the linkage of the factors.

a) Find the mle of θ and its asymptotic variance. We'll deal with this problem using generic n_1, n_2, n_3, n_4 for the cell counts, plugging in the observed values at the end.

$$\begin{aligned}
L(\theta) &= [.25(2+\theta)]^{n_1}\,[.25(1-\theta)]^{n_2}\,[.25(1-\theta)]^{n_3}\,[.25\theta]^{n_4} \\
l(\theta) &= -(n_1+n_2+n_3+n_4)\log(4) + n_1\log(2+\theta) + (n_2+n_3)\log(1-\theta) + n_4\log(\theta) \\
l'(\theta) &= \frac{n_1}{2+\theta} - \frac{n_2+n_3}{1-\theta} + \frac{n_4}{\theta} \\
&= \frac{n_1\theta(\theta-1) + (n_2+n_3)\theta(2+\theta) + n_4(\theta-1)(2+\theta)}{\theta(\theta-1)(2+\theta)} \\
&= \frac{\theta^2(n_1+n_2+n_3+n_4) + \theta(-n_1+2n_2+2n_3+n_4) - 2n_4}{\theta(\theta-1)(2+\theta)}
\end{aligned}$$

To find the mle, we simply solve the quadratic in the numerator, keeping only the positive root (we know $\theta > 0$). There must be just one positive root as the quadratic coefficient is positive and the constant coefficient is negative, so $-4ac$ is positive. Plugging in the numbers gives

$$3839\theta^2 + 1655\theta - 64$$

for which the positive root is $\hat{\theta} = 0.0357$. For the asymptotic variance, we need the second derivative of the log-likelihood function. For the asymptotic variance, we should use minus the inverse of the curvature of the log-likelihood function evaluated at the mle. A common variant of this procedure uses the expected curvature of the log-likelihood function in place of the observed curvature. We'll look at both, beginning with the expectation.

$$\begin{aligned}
l''(\theta) &= -\left[\frac{n_1}{(2+\theta)^2} + \frac{n_2+n_3}{(1-\theta)^2} + \frac{n_4}{\theta^2}\right] \\
E(l''(\theta)) &= -\frac{n}{4}\left[\frac{2+\theta}{(2+\theta)^2} + \frac{(1-\theta)+(1-\theta)}{(1-\theta)^2} + \frac{\theta}{\theta^2}\right] \\
E(l''(\theta)) &= -\frac{n}{4}\left[\frac{1}{(2+\theta)} + \frac{2}{(1-\theta)} + \frac{1}{\theta}\right] \\
&= -\frac{n}{4}\left[\frac{\theta(1-\theta) + 2\theta(2+\theta) + (2+\theta)(1-\theta)}{\theta(1-\theta)(2+\theta)}\right]
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{n}{4}\left[\frac{2+4\theta}{\theta(1-\theta)(2+\theta)}\right] \\
-1/E(l''(\theta)) &= \frac{2\theta(1-\theta)(2+\theta)}{n(1+2\theta)}.
\end{aligned}
$$

When this is evaluated at the mle, we get

$$V(\hat{\theta}) \approx \frac{2(0.0357)(0.9643)(2.0357)}{3839(1.0714)} = .000034076$$

leading to an estimated standard error, $s_{\hat{\theta}}$, of 0.005838. If we use the observed curvature instead, we get

$$
\begin{aligned}
l'(\theta) &= \frac{\theta^2(n_1+n_2+n_3+n_4)+\theta(-n_1+2n_2+2n_3+n_4)-2n_4}{\theta(\theta-1)(2+\theta)} \\
l''(\theta) &= \frac{2\theta(n_1+n_2+n_3+n_4)+(-n_1+2n_2+2n_3+n_4)}{\theta(\theta-1)(2+\theta)} - \frac{\text{denom}' * \text{numer}}{\text{denom}^2}, \\
l''(\hat{\theta}) &= \frac{2\hat{\theta}(n_1+n_2+n_3+n_4)+(-n_1+2n_2+2n_3+n_4)}{\hat{\theta}(\hat{\theta}-1)(2+\hat{\theta})} + 0 \\
V(\hat{\theta}) &= -\frac{\theta(\theta-1)(2+\theta)}{2\hat{\theta}(n_1+n_2+n_3+n_4)+(-n_1+2n_2+2n_3+n_4)} \\
&= -\frac{(0.0357)(-0.9643)(2.0357)}{2(0.0357)(3839)+1655} = .000036328 \\
s_{\hat{\theta}} &= 0.006027.
\end{aligned}
$$

The two answers are slightly different. If you have the option, it's better to use the answer based on the observed curvature, but we'll accept both answers. The expected variance is often easier to derive.

b) Form an approximate 95% confidence interval for θ based on part (a). To do this, we use the asymptotic normality of the distribution of the mle about the true value of the parameter, so that 95% confidence intervals will be of the form

$$\hat{\theta} \pm 1.96 * s_{\hat{\theta}}.$$

Using the answers from a), the intervals (based on expected and observed curvature, respectively) are

$$0.0357 \pm 1.96 * 0.005838 = 0.0357 \pm 0.0114 = (0.0243, 0.0471)$$

and

$$0.0357 \pm 1.96 * 0.006027 = 0.0357 \pm 0.0118 = (0.0239, 0.0475).$$

c) Use the bootstrap to find the approximate standard deviation of the maximum likelihood estimate and compare to the result of part (a). Ok, the bootstrap gives us the option of approximating the standard errors of estimates through simulation as opposed to algebra. We do this by taking many samples from the best fitted model, "refitting" the parameters for each new sample, and computing the standard deviation of the resulting estimates. We did this in matlab, using the code given below. This code is not optimized for running performance, but hopefully is somewat optimized for conceptual clarity.

```
B = 10000; % This is the number of bootstrap "resamples".

thetavec = zeros(B,1); % storage for the results of our simulations

thetaML = 0.0357; % our initial ML estimate

% set up some constants used for generating new samples

probvec = [2 + thetaML, 1 - thetaML, 1 - thetaML, thetaML]/4;
cumprobvec = [2 + thetaML, 3, 4 - thetaML, 4]/4;
n = 3839;

% now for the big loop; we generate B samples of size n,
% and compute the mle of theta based on the new sample.
% This requires solving the quadratic equation that we
% found for the numerator of the derivative of the log-likelihood
% function in part a).

for(i = 1:B)
x = rand(n,1); % generate random values

n1 = sum(x < cumprobvec(1));
n2 = sum(x < cumprobvec(2)) - n1;
n3 = sum(x < cumprobvec(3)) - n1 - n2;
n4 = n - n1 - n2 - n3;

a = n;
b = -n1 + 2*(n2 + n3) + n4;
c = -2*n4;

thetavec(i) = (-b + sqrt(b*b - 4*a*c))/(2*a);
end

varest = var(thetavec);
```

Upon running this program, we get an estimated variance for $\hat{\theta}$ of 0.000034184 with a corresponding estimated standard error of 0.005847. Your answers, of course, will differ somewhat from this due to random sampling, but will not differ by much. This is pretty close to what we found above, which might be expected since we're working with a fairly large set of data (several thousand observations). Note that the "large set of data" refers to the size of n, not B. For estimating standard errors, using a few hundred boostrap resamples is generally sufficient. We used quite a few more because we were looking ahead to part (d).

d) Use the bootstrap to find an approximate 95% confidence interval for θ and compare to part (b). There are a few ways to attack this problem. The first is to simply use the standard error for $\hat{\theta}$ and construct the confidence interval as

$$\hat{\theta} \pm 1.96 * s_{\hat{\theta}} = 0.0357 \pm 0.115 = (0.0242, 0.0472)$$

which is virtually identical to the result found in (b). This procedure works by assuming that the estimate of interest is approximately normally distributed about the true parameter value. The bootstrap does allow for another way of computing confidence intervals that works even if there is some asymmetry. We note that we are using $\hat{\theta}$ to estimate θ. Similarly, we can think of using our resample estimate, $\hat{\theta}^*$, as an estimate of $\hat{theta}$. We can put these two together to get a new guess for θ using $\hat{\theta}$ and $\hat{\theta}^*$ as follows:

$$\begin{aligned}
\theta &= \hat{\theta} + bias, \\
\hat{\theta} &= \hat{\theta}^* + bias^*, \\
bias^* &= \hat{\theta} - \hat{\theta}^*, \\
\theta &\approx \hat{\theta} + bias^* \\
&= 2\hat{\theta} - \hat{\theta}^*.
\end{aligned}$$

Now, in the course of doing the bootstrap for part (b), we computed B values for $\hat{\theta}^*$, which we can think of as allowing us to make B guesses as to the value of θ. We can then form a confidence interval for θ by sorting these guesses and choosing an interval that contains the central 95% of the guesses. In this case, since $B = 10000$, we could do this by choosing the 250th and 9750th sorted guesses as bounds.

```
stheta = sort(2*thetaML - thetavec);
bootconfint = [stheta(250), stheta(9750)];
```

In this instance, the boostrap confidence interval is

$$(0.0240, 0.0470)$$

which is quite close to what we found above. If we look at a histogram of the bootstrap resample values,

```
hist(thetavec,20) % histogram with 20 bins
print -deps boothist
```

we can see that the histogram looks bell-shaped with a mean near $\hat{\theta}$, so we would expect the normal approximation to be ok. Again, your exact numerical values will differ somehwat from those found above, but the qualitative conclusions (that the intervals don't differ that much) should be the same.

8.48 Referring to problem 47, consider two other estimates of θ. (1) The expected number of counts in the first cell is $n(2+\theta)/4$; if this expected number is equated to the count X_1, the following estimate is obtained:

$$\tilde{\theta}_1 = \frac{4X_1}{n} - 2.$$

(2) The same procedure done for the last cell yields

$$\tilde{\theta}_2 = \frac{4X_4}{n}.$$

Compute these estimates. Using the fact that X_1 and X_4 are binomial random variables, show that these estimates are unbiased, and obtain expressions for their variances. Evaluate the estimated standard errors and compare them to the estimated standard error of the mle.

First, a qualitative observation. While these estimators may be unbiased, they waste information. The first divides the observations into "cell 1" and "everything else" and ignores the fact that we have some knowledge as to subdivisions of the data within "everything else" that might be pertinent to estimating θ. The second pulls a similar trick. Hence, as they waste information, these estimates are inferior to the mle. Now, to the computation. Showing that these estimates are unbiased is fairly easy:

$$\begin{aligned} E(\tilde{\theta}_1) &= \frac{4E(X_1)}{n} - 2 = 2 + \theta - 2 = \theta, \\ E(\tilde{\theta}_2) &= \frac{4E(X_4)}{n} = \theta. \end{aligned}$$

As to the variances, the variance of a binomial random variable is $np(1-p)$. Carrying this through yields

$$\begin{aligned} V(\tilde{\theta}_1) &= V\left(\frac{4X_1}{n} - 2\right) \\ &= \frac{16}{n^2}V(X_1) = \frac{16}{n^2}n\frac{2+\theta}{4}\frac{2-\theta}{4} = \frac{4-\theta^2}{n}. \\ V(\tilde{\theta}_2) &= V\left(\frac{4X_4}{n}\right) \\ &= \frac{16}{n^2}V(X_4) = \frac{16}{n^2}n\frac{\theta}{4}\frac{4-\theta}{4} = \frac{\theta(4-\theta)}{n}. \end{aligned}$$

Plugging in numbers gives

$$\begin{aligned} \tilde{\theta}_1 &= 4\frac{n_1}{n} - 2 = 4(1997/3839) - 2 = 0.0808 \\ V(\tilde{\theta}_1) &= \frac{4 - 0.0808^2}{3839} = 0.00104 \\ s_{\tilde{\theta}_1} &= 0.0323 \\ \tilde{\theta}_2 &= 4\frac{n_4}{n} = 4\frac{32}{3839} = 0.0333 \\ V(\tilde{\theta}_2) &= \frac{0.0333(4 - 0.0333)}{3839} == 0.000034408 \\ s_{\tilde{\theta}_2} &= 0.0059. \end{aligned}$$

The standard error of the first alternative is quite a bit larger than that of the mle, but the standard error of the second is about the same.

Side note: Why does the latter estimate look plausible? We know that the mle is the one to use, but this one looks just as good. This is due to the fact that in this instance θ is fairly small. Looking back at the expected second derivative of the log-likelihood function,

$$E(l''(\theta)) = -\frac{n}{4}\left[\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta}\right],$$

we can see that the last term in the brackets, which corresponds to the last cell, contributes the largest fraction of the "information" in the estimate when θ is small. Thus, focusing our attention solely upon this cell loses less information than focusing on the first, second, or third cells. When θ is large, however, the situation changes dramatically. For θ near 1, the counts in cell 4 will supply much less information than is contained in the mle; the counts in cells 2 and 3 will be the most informative ones.

8.50 If gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2, respectively. Plato et al. (1964) published the following data on haptoglobin type in a sample of 190 people:

Haptoglobin Type		
Hp1-1	Hp1-2	Hp2-2
10	68	112

a) Find the mle of θ.

$$\begin{aligned}
L(\theta) &= (1-\theta)^{2n_1}(2\theta(1-\theta))^{n_2}\theta^{2n_3} \\
l(\theta) &= 2n_1\log(1-\theta) + n_2\log(2) + n_2\log(\theta) + n_2\log(1-\theta) + 2n_3\log(\theta) \\
l'(\theta) &= -\frac{2n_1}{1-\theta} + \frac{n_2}{\theta} - \frac{n_2}{1-\theta} + \frac{2n_3}{\theta} \\
&= \frac{2n_3+n_2}{\theta} - \frac{2n_1+n_2}{1-\theta}.
\end{aligned}$$

Setting this to zero, we get

$$\begin{aligned}
(2n_3+n_2)(1-\hat{\theta}) &= (2n_1+n_2)\hat{\theta} \\
2n_3+n_2 &= 2(n_1+n_2+n_3)\hat{\theta} \\
\hat{\theta} &= \frac{2n_3+n_2}{2n} = \frac{2(112)+68}{2(10+68+112)} = \frac{292}{380} = 0.7684.
\end{aligned}$$

b) Find the asymptotic variance of the mle.

$$\begin{aligned}
l''(\theta) &= -\left[\frac{2n_3+n_2}{\theta^2} + \frac{2n_1+n_2}{(1-\theta)^2}\right] \\
&= -\frac{(2n_3+n_2)(1-\theta)^2 + (2n_1+n_2)\theta^2}{\theta^2(1-\theta)^2} \\
&= -\frac{(2n_3+n_2)(1-2\theta) + 2n\theta^2}{\theta^2(1-\theta)^2} \\
l''(\hat{\theta}) &= -\frac{2n\hat{\theta}(1-2\hat{\theta}) + 2n\hat{\theta}}{\hat{\theta}^2(1-\hat{\theta})^2} \\
&= -\frac{2n}{\hat{\theta}(1-\hat{\theta})} \\
V(\hat{\theta}) &= \frac{\hat{\theta}(1-\hat{\theta})}{2n} = \frac{(0.7684)(0.2316)}{380} = 0.0004683 \\
s_{\hat{\theta}} &= 0.0216
\end{aligned}$$

As a side note, in this case the expected value of $l''(\theta)$ is the same as the value of $l''(\hat{\theta})$, so the observed and expected approaches to the variance of the mle yield the same results.

c) Find an approximate 99% confidence interval for θ. Using the asymptotic normality of the mle, an approximate 99% c.i. is

$$\hat{\theta} \pm (2.58) * s_{\hat{\theta}} = 0.7684 \pm 0.0557 = (0.7127, 0.8241).$$

d) Use the bootstrap to find the approximate standard deviation of the maximum likelihood estimate and compare your answer to that found in part (b). The matlab code for this simulation is very similar to that used in problem 47; the distinction is mostly in the computation of the estimates. The code is given below:

```
B = 10000; % This is the number of bootstrap "resamples".
thetavec = zeros(B,1); % storage for the results of our simulations
thetaML = 0.7684; % our initial ML estimate

% set up some constants used for generating new samples

probvec = [(1-thetaML)^2, 2*thetaML*(1-thetaML), thetaML^2];
cumprobvec = cumsum(probvec);
n = 190;

for(i = 1:B)
x = rand(n,1); % generate random values
n1 = sum(x < cumprobvec(1));
n2 = sum(x < cumprobvec(2)) - n1;
n3 = n - n1 - n2;
thetavec(i) = (2*n3 + n2)/(2*n);
end
varest = var(thetavec);
```

A histogram of thetavec shows that the estimates are roughly normally distributed about the mle. The standard error that I found here was $s_{\hat{\theta}} = 0.0218$; essentially the same as the answer found in (b).

e) Use the bootstrap to find an approximate 99% confidence interval and compare the result to the answer to part (c). Using the same approach as in problem 47, we compute $2\hat{\theta} - \hat{\theta}^*$ for all of the bootstrap resamples, sort the resulting values, and pick off appropriate quantiles. In this case, for a 99% confidence interval using $B = 10000$ we would use the 50th and 9950th sorted values. Doing this, I got

$$(0.7157, 0.8263),$$

which is very close to the answer found in (c).

9.1 A coin is thrown independently 10 times to test the hypothesis that the probability of heads is $\frac{1}{2}$ versus the alternative that the probability is not $\frac{1}{2}$. The test rejects if either 0 heads or 10 heads are observed.

a) What is the significance level of the test? The significance level of a test, α, is equal to the probability of a type I error, that we reject H_0 when H_0 is in fact true. In the context of this problem, it is the probability that we declare the coin unfair when the coin is actually fair.

$$\begin{aligned}\alpha &= P(\text{reject } H_0 | H_0 \text{ true}) \\ &= P(X = 0|H_0) + P(X = 10|H_0).\end{aligned}$$

Under H_0, the distribution of X is binomial$(10, 0.5)$, so

$$\begin{aligned}\alpha &= \binom{10}{0}(0.5)^0(1-0.5)^{10} + \binom{10}{10}(0.5)^{10}(1-0.5)^0 \\ &= \frac{1}{1024} + \frac{1}{1024} = \frac{1}{512} = 0.0020.\end{aligned}$$

b) If in fact the probability of heads is 0.1, what is the power of the test? The power of a test, $1 - \beta$, is the chance that we reject H_0 when the alternative hypothesis H_a holds. In the context of this problem, it is the chance that we correctly detect that the coin is unfair. Under $H_a : p = 0.1$, X has a binomial$(10, .1)$ distribution.

$$\begin{aligned}1-\beta &= P(\text{reject } H_0 | H_a \text{ true}) \\ &= P(X = 0|H_a) + P(X = 10|H_a) \\ &= \binom{10}{0}(0.1)^0(1-0.1)^{10} + \binom{10}{10}(0.1)^{10}(1-0.1)^0 = 0.3487.\end{aligned}$$

The value of α is so low that even under this extreme assumption for p we don't have very good power.

9.2 Which of the following hypotheses are simple, and which are composite?

a) X follows a uniform distribution on $[0, 1]$. A hypothesis is simple if, under that hypothesis, the distribution of all random variables involved is completely specified; all of the parameters are known. In this case, given that X is uniform, we need to know the range of values it can take on; as this range is given the distribution is completely specified. Ergo, this is a simple hypothesis.

b) A die is unbiased. Assuming that we are working with a standard 6-sided die, the assertion that the die is unbiased tells us the exact probabilities of all possible outcomes, so the distribution is completely specified. This is a simple hypothesis.

c) X follows a normal distribution with mean 0 and variance $\sigma^2 > 10$. In this case, the distribution is not completely specified: both $X \sim N(0, 11)$ and $X \sim N(0, 12)$ are allowed. This is a composite hypothesis.

d) X follows a normal distribution with mean $\mu = 0$. Once more, with feeling! As with part c), the distribution is not completely specified (any value of $\sigma^2 > 0$ is allowed) so this is a composite hypothesis.

9.5 True or false:

a) The significance level of a statistical test is equal to the probability that the null hypothesis is true.

False. The significance level of a statistical test is the probability that we reject the null hypothesis when the null hypothesis is in fact true. Our action, whether we accept or reject, is based on the sample and is thus a random variable. The truth or falsity of a given hypothesis does not change from sample to sample, so the "probability that the null hypothesis is true" has no meaning.

b) If the significance level of a test is decreased, the power would be expected to increase.

False. The significance level of a test, α, and the probability of a type II error, β, are typically balanced against one another: for a given set of observations (n fixed) decreasing one means increasing the other. Hence, if we decrease α, β should increase, and the power, $1 - \beta$, should decrease.

c) If a test is rejected at the significance level α, the probability that the null hypothesis is true equals α.

False. If a test is rejected at the significance level α, then either 1) the null hypothesis is true but an event of probability less than α has occurred, or 2) the null hypothesis is false. Again, it does not make sense to speak of the probability that the null hypothesis is true.

d) The probability that the null hypothesis is falsely rejected is equal to the power of the test.

False. This is basic terminology; α is the probability that the null hypothesis is falsely rejected. The power of a test, $1 - \beta$, is the probability that the null hypothesis is correctly rejected in favor of the true alternative.

e) A type I error occurs when the test statistic falls in the rejection region of the test.

False. More precisely, not always true. The statement above is correct if the null hypothesis holds; if the alternative hypothesis holds and we reject the null then no error has occurred.

f) A type II error is more serious than a type I error.

False. The relative import of errors of both types is something that varies from problem to problem and must be assessed by the practitioner. In shipping parts to major consumers, for example, one might test for flaws in the lot of equipment being shipped, with the null hypothesis that there are none or very few flaws. A type II error here, believing that there are few flaws and hence shipping a flawed product that will later have to be recalled, is more expensive than a type I error.

g) The power of a test is determined by the null distribution of the test statistic.

False. The power of a test is a function of the alternative hypothesis and the rejection region, and does not involve the null distribution directly.

h) The likelihood ratio is a random variable.

True. The likelihood ratio is the relative likelihood of the observed sample under the null hypothesis to the likelihood of the observed sample under the alternative hypothesis. As it is a function of the sample, its value changes from sample to sample and it is thus a random variable.

9.6 Following the reasoning of Example A of Section 9.3, develop a likelihood ratio test of $H_0 : p = 0.6$ versus $H_A : p = 0.7$ based on observing a binomial random variable with 10 trials.

To develop the likelihood ratio test, we first compute the likelihood of the various values that the random variable can take on under both the null and alternative hypotheses. In this case, under the null X is binomial$(10, .6)$, and under the alternative X is binomial$(10, .7)$. Thus,

X	0	1	2	3	4	5	6	7	8	9	10
$p_{X\|H_0}$	0.0001	0.0016	0.0106	0.0425	0.1115	0.2007	0.2508	0.2150	0.1209	0.0403	0.0060
$p_{X\|H_A}$	0.0000	0.0001	0.0014	0.0090	0.0368	0.1029	0.2001	0.2668	0.2335	0.1211	0.0282
$\frac{p_{X\|H_0}}{p_{X\|H_A}}$	17.76	11.42	7.34	4.72	3.03	1.95	1.25	0.806	0.518	0.333	0.214

Under a likelihood ratio test, we would reject H_0 in favor of H_A when the likelihood ratio is too small. In the context of this problem, that corresponds to rejecting the null when our test statistic is very large, as large values of X ($X \geq 7$) are more likely under H_A than under H_0. Conversely, small values of X are more likely under H_0 than under H_A. Let us say that we choose to reject H_0 if X is 9 or 10. Then the significance level of this test is $0.0403 + 0.0060 = 0.0463$, and the power is $0.1211 + 0.0282 = 0.1493$. Where the boundary of the rejection region is set is something that must be chosen by the experimenter, but the shape of the rejection region should always correspond to rejecting H_0 for values of X above a threshold.

9.9 Let $X_1, \ldots, X_{25}$ be a sample from a normal distribution having a variance of 100. Find the rejection region for a test at level $\alpha = 0.10$ of $H_0 : \mu = 0$ versus $H_A : \mu = 1.5$. What is the power of the test? Repeat for $\alpha = 0.01$.

We begin by computing the likelihood ratio:

$$\begin{aligned}\frac{L_0(\mu)}{L_A(\mu)} &= \frac{\prod_{i=1}^{n}(2\pi * 100)^{-1/2}\exp(-\frac{1}{2*100}(X_i - 0)^2)}{\prod_{i=1}^{n}(2\pi * 100)^{-1/2}\exp(-\frac{1}{2*100}(X_i - 1.5)^2)} \\ &= \exp\left(-\frac{1}{200}\left[\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n}(X_i - 1.5)^2\right]\right) \\ &= \exp\left(-\frac{1}{200}\left[-\sum_{i=1}^{n}(-3X_i + 2.25)\right]\right)\end{aligned}$$

This ratio decreases as the sum of the X_i's (and hence $\bar{X}$) increases, so we should choose the rejection region so that we reject H_0 for all values of the mean above a set threshold. Now we need to determine that threshold. Given that we want $\alpha = 0.10$,

$$\begin{aligned}P(\bar{X} > k|H_0) &= 0.10 \\ P\left(\frac{\bar{X} - 0}{\sqrt{100/25}} > \frac{k}{2}\middle| H_0\right) &= 0.10 \\ k &= 2 * \Phi^{-1}(1 - 0.1) = 2 * z_{0.9} = 2 * 1.2816 = 2.56.\end{aligned}$$

Now, the power of this test versus the specified alternative is

$$\begin{aligned}\text{Power} &= P(\bar{X} > 2.56|H_A)\\ &= P\left(\frac{\bar{X}-1.5}{2} > \frac{2.56-1.5}{2}\middle| H_A\right)\\ &= P(Z > 0.53) = 1 - P(Z < 0.53) = 1 - 0.7019 = 0.2981.\end{aligned}$$

If we now decide that we want to use $\alpha = 0.01$,

$$\begin{aligned}P(\bar{X} > k|H_0) &= 0.01\\ P\left(\frac{\bar{X}-0}{\sqrt{100/25}} > \frac{k}{2}\middle| H_0\right) &= 0.01\\ k &= 2 * \Phi^{-1}(1-0.01) = 2 * z_{0.99} = 2 * 2.3263 = 4.65.\end{aligned}$$

Now, the power of this test versus the specified alternative is

$$\begin{aligned}\text{Power} &= P(\bar{X} > 4.65|H_A)\\ &= P\left(\frac{\bar{X}-1.5}{2} > \frac{4.65-1.5}{2}\middle| H_A\right)\\ &= P(Z > 1.58) = 1 - P(Z < 1.58) = 1 - 0.9429 = 0.0571.\end{aligned}$$

For a starting variance this large, the alternative is too close to the null to be discerned easily.

9.12 Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with the density function $f(x|\theta) = \theta \exp[-\theta x]$. Derive a likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_A : \theta \neq \theta_0$, and show that the rejection region is of the form $\{\bar{X} \exp[-\theta_0 \bar{X}] \leq c\}$.

To determine the form of the rejection region, we construct a likelihood ratio, where we use the value of θ specified by the null hypothesis in the numerator and the maximum likelihood value of θ in the denominator. To do this, we need the mle of θ:

$$\begin{aligned}L(\theta) &= \prod_{i=1}^{n} \theta \exp[-\theta X_i]\\ l(\theta) &= n\log(\theta) - \theta \sum_{i=1}^{n} X_i\\ l'(\theta) &= \frac{n}{\theta} - \sum_{i=1}^{n} X_i\\ \hat{\theta} &= 1/\bar{X}.\end{aligned}$$

The likelihood ratio is given by

$$\begin{aligned}\frac{L_0(\theta)}{L(\hat{\theta})} &= \frac{\theta_0^n \exp[-\theta_0 \sum_{i=1}^n X_i]}{\hat{\theta}^n \exp\left[-\hat{\theta}\sum_{i=1}^n X_i\right]}\\ &= \left(\frac{\theta_0}{\hat{\theta}}\right)^n \exp\left[-(\theta_0 - \hat{\theta})n\bar{X}\right]\\ &= \left(\theta_0 \bar{X} \exp[-\theta_0 \bar{X} + 1]\right)^n\end{aligned}$$

As we choose the rejection region to be all sets of observations such that the likelihood ratio is less than a specified level $c_1 > 0$ (this by definition of a likelihood ratio test) this corresponds to

$$\begin{aligned}\left(\theta_0 e^{-1}\bar{X}\exp\left[-\theta_0\bar{X}\right]\right)^n &\leq c_1\\ \theta_0 e^{-1}\bar{X}\exp\left[-\theta_0\bar{X}\right] &\leq c_1^{1/n}\\ \bar{X}\exp\left[-\theta_0\bar{X}\right] &\leq e\theta_0^{-1}c_1^{1/n}.\end{aligned}$$

At this point, noting that θ_0 is a constant and redefining the quantity on the right hand side of the inequality as "c", we see that the rejection region of the likelihood ratio test corresponds to

$$\bar{X}\exp\left[-\theta_0\bar{X}\right] \leq c.$$

9.13 Suppose, to be specific, that in problem 12, $\theta_0 = 1$, $n = 10$, and that $\alpha = 0.05$. In order to use the test, we must find the appropriate value of c.

a) Show that the rejection region is of the form $\{\bar{X} \leq x_0\} \cup \{\bar{X} \geq x_1\}$, where x_0 and x_1 are determined by c.

As we saw in problem 12, the rejection region is of the form

$$\bar{X}\exp\left[-\theta_0\bar{X}\right] \leq c.$$

The random variable $\bar{X}$ can take on values in the range $(0, \infty)$, and the likelihood ratio function is continuous in this range. If we differentiate the likelihood ratio function with respect to $\bar{X}$, getting

$$(1 - \theta_0\bar{X})\exp\left[-\theta_0\bar{X}\right],$$

we note that the likelihood ratio function has just a single critical point on the interior of $(0, \infty)$, at $\bar{X} = 1/\theta_0$, and that it is monotonically increasing on $(0, \theta_0^{-1})$ and monotonically decreasing on (θ_0^{-1}, ∞). Thus, for cutoff values of c less than that associated with the maximum, the rejection region must be of the form

$$\{\bar{X} \leq x_0\} \cup \{\bar{X} \geq x_1\}$$

where x_0 and x_1 are determined by c.

b) Explain why c should be chosen so that $P(\bar{X}\exp(-\bar{X}) \leq c) = 0.05$ when $\theta_0 = 1$.

The rejection region is of the form

$$\bar{X}\exp\left[-\theta_0\bar{X}\right] \leq c,$$

which reduces to

$$\bar{X}\exp\left[-\bar{X}\right] \leq c$$

when $\theta_0 = 1$. As this is the value of θ given by the null hypothesis,

$$P(\bar{X}\exp(-\bar{X}) \leq c) = P(\text{reject } H_0 | H_0 \text{ true})$$

or the significance level of the test. As we have specified that $\alpha = 0.05$, we must choose c so that

$$P(\bar{X}\exp(-\bar{X}) \leq c) = 0.05.$$

c) Explain why $\sum_{i=1}^{10} X_i$ and hence $\bar{X}$ follow gamma distributions when $\theta_0 = 1$. How could this knowledge be used to choose c?

Under H_0, $X_i \sim$ Exponential(1). An exponential random variable is also a Gamma($\alpha =$ 1,λ) random variable by definition, as the exponential is a special case of the Gamma, so $X_i \sim$ Gamma(1,1). If we add independent Gamma random variables with the same value of λ together, the sum is also a Gamma random variable with $\alpha_{sum} = \sum \alpha_i$. We proved this in class using moment generating functions. Thus, $\sum_{i=1}^{n} X_i$ has a Gamma$(n, 1)$ distribution. When we multiply a Gamma random variable by a constant, we showed in a previous homework (focusing on change of variables) that only the λ parameter is changed; multiplication by k results in a new Gamma random variable with the same α and $\lambda_{new} = \lambda/k$. Thus, when we take the mean here (multiplying the sum by n^{-1}) we see that the sample mean $\bar{X}$ has a Gamma(n, n) distribution.

In terms of using this knowledge to choose c, knowing the distribution of $\bar{X}$ should allow us to calculate the exact value of c such that 95% coverage is attained. In most cases, we don't know the exact distribution. To do this, we could proceed in matlab as follows. Let's say that we're guessing that about 2.5% will be excluded on either side (eventually, due to the asymptotic normality of the mle about the true value, the intervals will be symmetric). Matlab will compute the q quantiles of a Gamma(α, λ) distribution, invoked as $gaminv(q, \alpha, 1/\lambda)$. The last argument is inverted because matlab uses a different parameterization of the gamma (and exponential) distributions than your text.

```
gaminv(0.025,10,0.1) % find the 2.5 percentile
> 0.4795
0.4795 * exp(-0.4795)
> 0.2969
gaminv(0.975,10,0.1) % find the 97.5 percentile
> 1.7085
1.7085 * exp(-1.7085)
> 0.3095 % the two LR values don't agree; modify!
delta = 0.005;
[gaminv(0.025+delta,10,.1),gaminv(0.975+delta,10,.1)]
> 0.4949 1.7510
ans .* exp(-ans)
> 0.3017 0.3040
delta = 0.006;
[gaminv(0.025+delta,10,.1),gaminv(0.975+delta,10,.1)]
> 0.4977 1.7606
ans .* exp(-ans)
> 0.3026 0.3027
delta = 0.0061;
[gaminv(0.025+delta,10,.1),gaminv(0.975+delta,10,.1)]
> 0.4980 1.7616
ans .* exp(-ans)
```

```
> 0.3026 0.3026
```

This is a trial and error approach; if we were going to do this repeatedly we could probably come up with a better way of coding it using something akin to Newton's method. Here, we see that the value of c is 0.3026, and that the confidence interval that we get is slightly asymmetric - we will reject the null for the 3.11 percent smallest values and the 1.89 percent largest values.

d) Suppose that you hadn't thought of the preceding fact. Explain how you could determine a good approximation to c by generating random numbers on a computer (simulation).

This is actually easier and of more general applicability. As we know the distribution of the observations under the null hypothesis (exponential(1)), we generate multiple samples of size $n = 10$ from this distribution and compute $\bar{x}\exp(-\bar{x})$ for each of the samples. Say we generate 10000 samples and record the results. If we sort the results, and set the 500th value as our cutoff, then using this cutoff we would have rejected 5% of our observations.

```
B = 10000;
n = 10;
resultvec = zeros(B,1);
xbarvec = zeros(B,1);
lambdastart = 1;

x = exprnd(lambdastart,B,n);
xbarvec = mean(x')';
resultvec = xbarvec.*exp(-xbarvec);
sres = sort(resultvec);
sres(500)
> 0.3017
```

Close, but not exact. If we do this a few times, the results are 0.3015, 0.3037, 0.3036, 0.3029, 0.3021, etc. This will get us into the correct ballpark!