prepared by Professor Jenny Baglivo

3 MATH4427 Notebook 3

This notebook is concerned with the second core topic of mathematical statistics, namely, hypothesis testing theory. The notes include material from Chapter 9 of the Rice textbook.

3.1 Hypotheses: Definitions and Examples

1. *Hypothesis:* An *hypothesis* is an assertion about the distribution of a random variable or a random k-tuple.

2. *Simple vs Compound Hypotheses:* There are two types of hypotheses:
   - A *simple hypothesis* specifies the distribution completely. For example,

     H: X is a Bernoulli random variable with parameter $p = 0.45$.
     H: (X_1, X_2, X_3) has a multinomial distribution with $n = 10$ and $(p_1, p_2, p_3) = (0.3, 0.5, 0.2)$.

   - A *compound hypothesis* does not specify the distribution completely. For example,

     H: X is a Bernoulli random variable with parameter $p \geq 0.45$.
     H: (X_1, X_2, X_3) has a multinomial distribution with $n = 10$.

3. *Neyman-Pearson Framework:* In the *Neyman-Pearson framework* of hypothesis testing, there are two competing assertions:

   (a) The null hypothesis, H_O, and
   (b) The alternative hypothesis, H_A.

   The null hypothesis is accepted as true unless sufficient evidence is provided to the contrary. If sufficient evidence is provided to the contrary, then the null hypothesis is rejected in favor of the alternative hypothesis.

Note: If there is insufficient evidence to reject a null hypothesis, researchers often say that they have "failed to reject the null hypothesis" rather than they "accept the null hypothesis".

Example: Efficacy of a proposed new treatment. Suppose that the standard treatment for a given medical condition is effective in 45% of patients. A new treatment promises to be effective in more than 45% of patients.

In testing the efficacy of the proposed new treatment, the hypotheses could be set up as follows:

H_O: The new treatment is no more effective than the standard treatment.

H_A: The new treatment is more effective than the standard treatment.

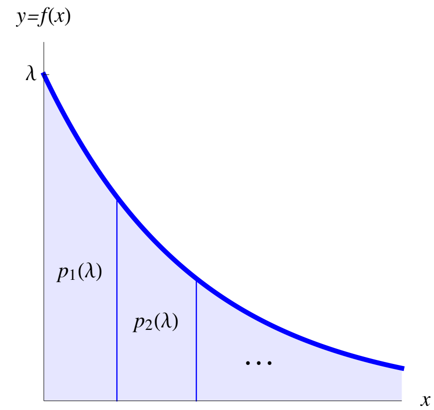
If p is the proportion of patients for whom the new proposed treatment would be effective, then the hypotheses could be set up as follows:

$$H_O : p = 0.45 \quad \text{versus} \quad H_A : p > 0.45.$$

Example: Determining goodness-of-fit to an exponential distribution. In testing whether or not an exponential model is a reasonable model for sample data, the hypotheses would be set up as follows:

H_O: The distribution of X is exponential.

H_A: The distribution of X is not exponential.



If Pearson's test is used to determine goodness-of-fit, then the range of X would be divided into k subintervals, for some k. Let λ be the parameter of the exponential distribution, and

$$p_i = p_i(\lambda)$$

be the probability that an observation falls in the i^{th} subinterval, for $i = 1, 2, \ldots, k$.

The list of probabilities is a parameter for a multinomial experiment with k outcomes. The hypotheses would then be set up as follows:

$$H_O : (p_1, p_2, \ldots, p_k) = (p_1(\lambda), p_2(\lambda), \ldots, p_k(\lambda)) \quad \text{versus}$$

$$H_A : (p_1, p_2, \ldots, p_k) \neq (p_1(\lambda), p_2(\lambda), \ldots, p_k(\lambda)).$$

3.2 Hypothesis Tests: Definitions and Examples

1. *Test:* A *test* is a decision rule allowing the user to choose between competing assertions.

2. *Components of a Test:* Let X_1, X_2, ..., X_n be a random sample from the X distribution. To set up a test:

   (a) A *test statistic*, $T = T(X_1, \ldots, X_n)$, is chosen and

   (b) The range of T is subdivided into the *rejection region* (RR) and the complementary *acceptance region* (AR).

   (c) If the observed value of T is in the acceptance region, then the null hypothesis is accepted. Otherwise, the null hypothesis is rejected in favor of the alternative.

   The test statistic, and the acceptance and rejection regions, are chosen so that the probability that T is in the rejection region is small (that is, near 0) when the null hypothesis is true. Hopefully, the probability that T is in the rejection region is large (that is, near 1) when the alternative hypothesis is true, but this is *not* guaranteed.

3. *Upper/Lower Tail Tests:* In an *upper tail test*, the null hypothesis is rejected when the test statistic is in the upper tail of distributions satisfying the null hypothesis.

   A test which rejects the null hypothesis when the test statistic is in the lower tail of distributions satisfying the null hypothesis is called a *lower tail test*.

4. *Two Tailed Tests:* In a *two tailed test*, the null hypothesis is rejected if the test statistic is either in the upper tail or the lower tail of distributions satisfying the null hypothesis.

5. *One/Two -Sided Tests:* Upper and lower tail tests are also called *one sided tests*. Two tail tests are also called *two sided tests*.

Example: Upper tail test of a Bernoulli proportion. Let Y be the sample sum of a random sample of size 25 from a Bernoulli distribution with success probability p.

Consider the following decision rule for a test of the null hypothesis that $p = 0.45$ versus the alternative hypothesis that $p > 0.45$:

$$\text{Reject } p = 0.45 \text{ in favor of } p > 0.45 \text{ when } Y \geq 16.$$

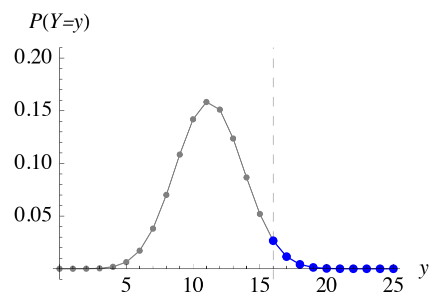
The test statistic Y is a binomial random variable with $n = 25$. The acceptance and rejection regions are

$$AR = \{0, 1, 2, \ldots, 15\} \quad \text{and} \quad RR = \{16, 17, \ldots, 25\}.$$

1. *Null Hypothesis:* If the null hypothesis is true, then the probability that the test statistic is in the rejection region is

$$P_{0.45}(Y \in RR) \approx 0.044,$$

as illustrated on the right.



2. *Alternative Hypothesis:* If the alternative hypothesis is true, then the probability depends on the correct value of p.

For example,

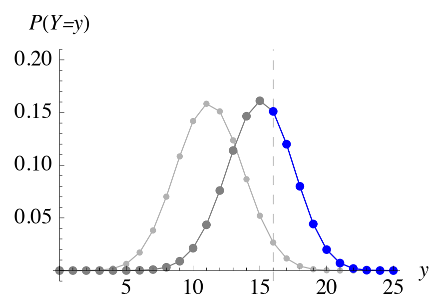
- If $p = 0.60$, then

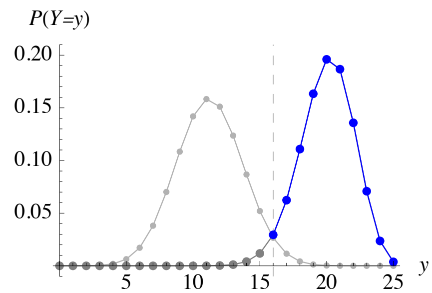
$$P_{0.60}(Y \in RR) \approx 0.425;$$

- If $p = 0.80$, then

$$P_{0.80}(Y \in RR) \approx 0.983.$$

Each case is illustrated on the right.





Example: Two tail test of a Normal mean. Let $\overline{X}$ be the sample mean of a random sample of size 16 from a normal distribution with mean μ and standard deviation 10.

Consider the following decision rule for a test of the null hypothesis that $\mu = 85$ versus the alternative hypothesis that $\mu \neq 85$:[1]

$$\text{Reject } \mu = 85 \text{ in favor of } \mu \neq 85 \text{ when } |\overline{X} - 85| \geq 4.7.$$

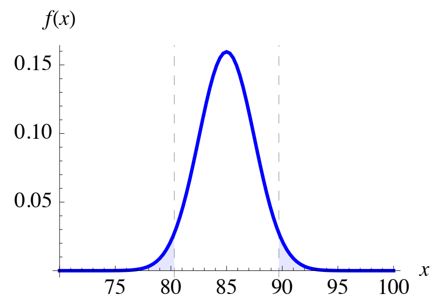
The test statistic $\overline{X}$ has a normal distribution with standard deviation $\frac{10}{\sqrt{16}} = 2.5$. The acceptance and rejection regions are:

$$AR = (80.3, 89.7) \quad \text{and} \quad RR = (-\infty, 80.3] \cup [89.7, \infty).$$

1. *Null Hypothesis:* If the null hypothesis is true, then the probability that the test statistic is in the rejection region is

   $$P_{85}(\overline{X} \in RR) \approx 0.06,$$

   as illustrated on the right.



2. *Alternative Hypothesis:* If the alternative hypothesis is true, then the probability depends on the correct value of μ.

   For example,

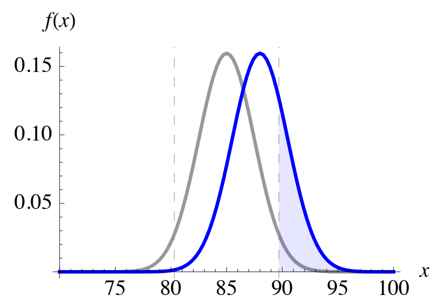
   - If $\mu = 88$, then

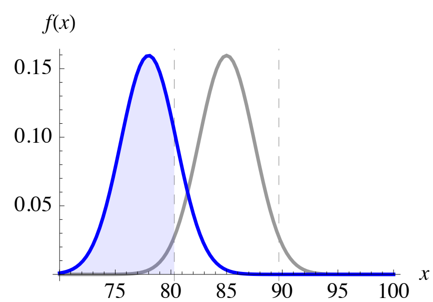
     $$P_{88}(\overline{X} \in RR) \approx 0.249;$$

   - If $\mu = 78$, then

     $$P_{78}(\overline{X} \in RR) \approx 0.821.$$

   Each case is illustrated on the right.





[1] A possible scenario that would make this set up interesting is the following: The mean score for certain subject matter in a certain population of students is 85. Those who participate in a review class have the potential of changing their score, although it is assumed that only the mean can change. X represents the score earned by those who participate in the review class.

Equivalent tests. Consider two tests, each based on a random sample of size n:

1. A test based on statistic T with rejection region RR_T, and
2. A test based on statistic W with rejection region RR_W.

The tests are said to be *equivalent* if

$$T \in RR_T \quad \Longleftrightarrow \quad W \in RR_W.$$

That is, given the same information, either both tests accept the null hypothesis or both reject the null hypothesis. Equivalent tests have the same properties.

The two examples above can be used to illustrate this concept:

1. _*Bernoulli proportion example, continued:*_ Let $\widehat{p}$ be the sample proportion,

   $$\widehat{p} = \frac{Y}{25}, \quad \text{where } Y \text{ is the sample sum.}$$

   Then the test with decision rule

   "Reject $p = 0.45$ in favor of $p > 0.45$ when $\widehat{p} \geq 0.64$"

   is equivalent to the test with decision rule

   "Reject $p = 0.45$ in favor of $p > 0.45$ when $Y \geq 16$"

   since $\frac{Y}{25} \geq 0.64$ if and only if $Y \geq 16$.

2. _*Normal mean example, continued:*_ Let Z be the standardized mean under the null hypothesis that $\mu = 85$,

   $$Z = \frac{\overline{X} - 85}{10/\sqrt{16}}, \quad \text{where } \overline{X} \text{ is the sample mean.}$$

   Then the test with decision rule

   "Reject $\mu = 85$ in favor of $\mu \neq 85$ when $|Z| \geq 1.88$"

   is equivalent to the test with decision rule

   "Reject $\mu = 85$ in favor of $\mu \neq 85$ when $\overline{X} \leq 80.3$ or $\overline{X} \geq 89.7$"

   since $\left|\frac{\overline{X}-85}{10/\sqrt{16}}\right| \geq 1.88$ if and only if $\overline{X} \leq 80.3$ or $\overline{X} \geq 89.7$.

3.3 Measuring the Performance of Hypothesis Tests

3.3.1 Parameter Space, Null Space

Null and alternative hypotheses are often written as assertions about parameters. In these cases, we use the following notation:

$$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o$$

where Ω is the set of all parameters, and $\omega_o \subset \Omega$ is the subset under the null hypothesis.

For example,

1. If X is a Bernoulli random variable with success probability p, and we are interested in testing the null hypothesis $p \leq 0.30$ versus the alternative hypothesis $p > 0.30$, then

2. If X is an exponential random variable with parameter λ, and we are interested in testing the null hypothesis $\lambda = 5$ versus the alternative hypothesis $\lambda < 5$, then

3. If X is a normal random variable with parameters μ and σ^2, with both unknown, and we are interested in testing the null hypothesis $\mu = 0$ versus the alternative $\mu \neq 0$, then

3.3.2 Errors of Type I and II

When carrying out a test, two types of errors can occur:

1. An *error of type I* occurs when a true null hypothesis is rejected.
2. An *error of type II* occurs when a false null hypothesis is accepted.

The Greek letter α is often used to denote the probability of committing an error of type I, and β is used to denote the probability of committing an error of type II.

For example,

1. Consider again the upper tail test of the Bernoulli proportion from page 5.

   The probability of committing a type I error is (about) ____________________.

   The probability of committing a type II error
   when the true proportion is 0.60 is (about) ____________________.

2. Consider again the two tail test of the normal mean from page 6.

   The probability of committing a type I error is (about) ____________________.

   The probability of committing a type II error
   when the true mean is 78 is (about) ____________________.

> *Note:* Since the null hypothesis was a simple hypothesis in both examples, there was no need to specify a parameter value when finding the type I error.
>
> Since the alternative hypothesis was a compound hypothesis in both examples, I needed to know the specific true alternative model in order to find the type II error.

3.3.3 Size, Significance Level, and Power

Consider the test with decision rule:

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } T \in RR$$

where $T = T(X_1, X_2, \ldots, X_n)$ is a test statistic based on a random sample of size n.

1. *Size/Significance Level:* The *size* or *significance level* of the test is defined as follows:

$$\alpha = \sup_{\theta \in \omega_o} P_\theta(T \in RR).$$

   (The size or significance level is the maximum type I error, or the least upper bound of type I errors, if a maximum does not exist).

   Note that

   (a) A test with size α is called a "$100\alpha\%$ test."

   (b) If the significance level is α, and the observed data lead to rejecting the null hypothesis, then the result is said to be *statistically significant* at level α.

2. *Power:* The *power* of the test at the parameter θ is $P_\theta(T \in RR)$.

   Note that

   (a) If $\theta \in \omega_o$, then the power at θ is the same as the type I error.

   (b) If $\theta \in \Omega \setminus \omega_o$, then the power corresponds to the test's ability to correctly reject the null hypothesis in favor of the alternative hypothesis.

   (c) Power is a function of the unknown parameter.

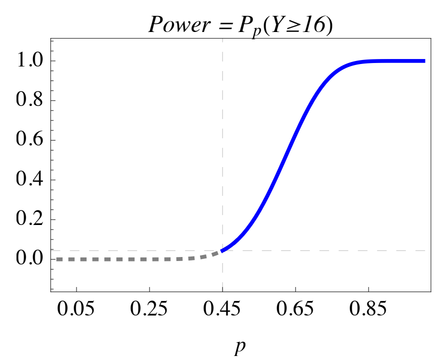
For example,

1. Power as a function of p for the test of

$$p = 0.45 \quad vs \quad p > 0.45$$

   introduced on page 5 is shown on the right.

   Power increases as p increases.

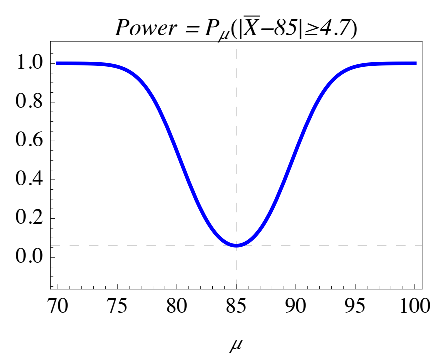


2. Power as a function of μ for the test of

$$\mu = 85 \quad vs \quad \mu \neq 85$$

   introduced on page 6 is shown on the right.

   Power increases as you move away from 85 in either direction.



Exercise. Let Y be the sample sum of a random sample of size 12 from a Bernoulli distribution with success probability p. Consider tests of the null hypothesis $p = 0.45$ versus the alternative hypothesis $p > 0.45$ using Y as test statistic.

(a) Find the decision rule for a test whose size is as close to 5% as possible, and report the exact size of the test. [2]

(b) Find the power of the test from part (a) when $p = 0.65$.

(c) Find the type II error of the test from part (a) when $p = 0.85$.

[2] You may find the following table of probabilities when $p = 0.45$ useful:

y	0	1	2	3	4	5	6
$P_{.45}(Y = y)$	0.0008	0.0075	0.0339	0.0923	0.1700	0.2225	0.2124

y	7	8	9	10	11	12	
$P_{.45}(Y = y)$	0.1489	0.0762	0.0277	0.0068	0.0010	0.0001	

Exercise. Let Y be the sample sum of a random sample of size 6 from a Poisson distribution with parameter λ. Consider tests of the null hypothesis $\lambda = 2$ versus the alternative hypothesis $\lambda < 2$ using Y as test statistic.

(*a*) Find the decision rule for a test whose size is as close to 10% as possible, and report the exact size of the test.[3]

(*b*) Find the power of the test from part (*a*) when $\lambda = 1.2$.

(*c*) Find the type II error of the test from part (*a*) when $\lambda = 1.8$.

[3] You may find the following partial table of probabilities when $\lambda = 2$ useful:

y	0	1	2	3	4	5	6	7
$P_2(Y = y)$	0.0000	0.0001	0.0004	0.0018	0.0053	0.0127	0.0255	0.0437

y	8	9	10	11	12	13	14	15
$P_2(Y = y)$	0.0655	0.0874	0.1048	0.1144	0.1144	0.1056	0.0905	0.0724

y	16	17	18	19	20	21	22	23
$P_2(Y = y)$	0.0543	0.0383	0.0255	0.0161	0.0097	0.0055	0.0030	0.0016

3.3.4 Observed Significance Level, P Value

The *observed significance level*, or p *value*, is the minimum significance level for which the null hypothesis would be rejected for the given data.

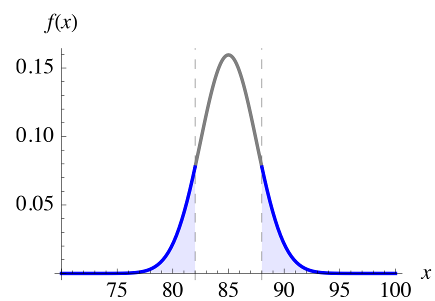
For example,

1. Continuing with the test introduced on page 11, suppose that 8 successes are observed. Then, the p value is

2. Continuing with the test introduced on page 12, suppose that 4 events are observed. Then, the p value is

Exercise. Let $\overline{X}$ be the sample mean of a random sample of size 16 from a normal distribution with mean μ and standard deviation 10, and consider testing $\mu = 85$ versus $\mu \neq 85$ using $\overline{X}$ as test statistic.

Find the p value if a sample mean of 82 is observed.



3.3.5 Comparing Tests: Power for Fixed Size

Consider two $100\alpha\%$ tests of $\theta \in \omega_o$ versus $\theta \in \Omega \setminus \omega_o$:

1. The test with decision rule

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } T \in RR_T,$$

   where $T = T(X_1, X_2, \ldots, X_n)$ is a test statistic based on a random sample of size n.

2. The test with decision rule

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } W \in RR_W,$$

   where $W = W(X_1, X_2, \ldots, X_n)$ is a test statistic based on a random sample of size n.

Then the test based on T is *uniformly more powerful* than the test based on W iff

$$P_\theta(T \in RR_T) \geq P_\theta(W \in RR_W) \text{ for all } \theta \in \Omega \setminus \omega_o,$$

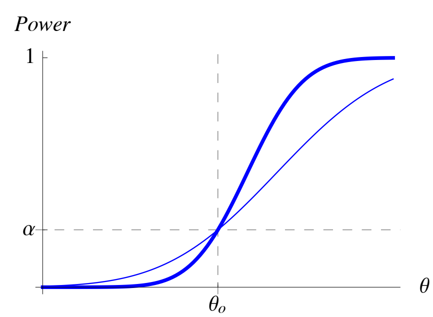
with strict inequality (that is, with "$>$" replacing "$\geq$") for at least one θ.

To illustrate the concept, let θ be a parameter with values anywhere on the real line.

1. *Upper Tail Test:* The curves to the right are typical power curves for an upper tail test of

$$\theta \leq \theta_o \text{ versus } \theta > \theta_o,$$

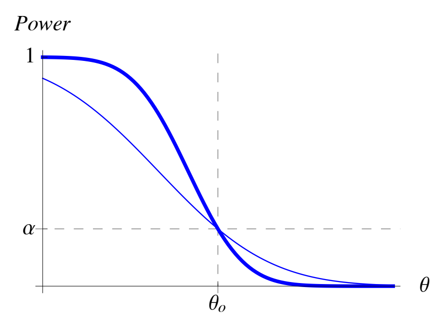
   where the null hypothesis is rejected when the test statistic is in the upper tail of its sampling distribution. Each test has size α.



2. *Lower Tail Test:* The curves to the right are typical power curves for a lower tail test of

$$\theta \geq \theta_o \text{ versus } \theta < \theta_o.$$

   where the null hypothesis is rejected when the test statistic is in the lower tail of its sampling distribution. Each test has size α.



In each case, the test with the thicker power curve is the one that is uniformly more powerful.

A natural question to ask is whether a uniformly most powerful test exists. Formally,

> ***UMPT:*** Let $T = T(X_1, X_2, \ldots, X_n)$ be the test statistic for a size α test of
>
> $$\theta \in \omega_o \text{ versus } \theta \in \Omega - \omega_o.$$
>
> Then the test based on T is said to be a *uniformly most powerful* test of size α if it is uniformly more powerful than any competitor test of the same size.

We will consider the question of finding a UMPT (if one exists) in Section 3.7 (page 24).

3.4 Hypothesis Tests for Normal Distributions

3.4.1 Tests for the Mean Parameter

Size α Tests of $H_O : \mu = \mu_o$

	σ *Known*	σ *Estimated*
Test Statistic	$Z = \dfrac{\overline{X} - \mu_o}{\sqrt{\sigma^2/n}}$	$T = \dfrac{\overline{X} - \mu_o}{\sqrt{S^2/n}}$
RR for $H_A : \mu < \mu_o$	$Z \leq -z(\alpha)$	$T \leq -t_{n-1}(\alpha)$
RR for $H_A : \mu > \mu_o$	$Z \geq z(\alpha)$	$T \geq t_{n-1}(\alpha)$
RR for $H_A : \mu \neq \mu_o$	$\lvert Z\rvert \geq z(\alpha/2)$	$\lvert T\rvert \geq t_{n-1}(\alpha/2)$

Notes:

1. *Upper Tail Notation:* It is common practice to use upper tail notation for the quantiles used at cutoffs in hypothesis tests. Thus, we use "$x(p)$" to denote the quantile "x_{1-p}."
2. *Left Column:* When σ is known, the exactly standardized statistic Z has a standard normal distribution.
3. *Right Column:* When σ is estimated, the approximately standardized statistic T has a Student t distribution with $n - 1$ df.

Exercise. In recent years, the yield (in bushels per acre) of corn in the United States has been approximately normally distributed with a mean of 120 bushels per acre and with a standard deviation of 10 bushels per acre.

A survey of 40 farmers this year gives a sample mean yield of 123.8 bushels per acre. We want to know whether this evidence is sufficient to say that the national mean has changed.

Assume this information summarizes the values of a random sample from a normal distribution with mean μ and standard deviation 10. Test $\mu = 120$ versus $\mu \neq 120$ at the 5% significance level. Report the conclusion and the observed significance level (p value).

Exercise (Hand et al, 1994, page 229). The table below gives information on a *before-and-after* experiment of a standard treatment for anorexia.

Before	After	After–Before	Before	After	After–Before	Before	After	After–Before
70.5	81.8	11.3	72.3	88.2	15.9	74.0	86.3	12.3
75.1	86.7	11.6	77.3	77.3	0.0	77.5	81.2	3.7
77.6	77.4	−0.2	78.1	76.1	−2.0	78.1	81.4	3.3
78.4	84.6	6.2	79.6	81.4	1.8	79.7	73.0	−6.7
80.6	73.5	−7.1	80.7	80.2	−0.5	81.3	89.6	8.3
84.1	79.5	−4.6	84.4	84.7	0.3	85.2	84.2	−1.0
85.5	88.3	2.8	86.0	75.4	−10.6	87.3	75.1	−12.2
88.3	78.1	−10.2	88.7	79.5	−9.2	89.0	78.8	−10.2
89.4	80.1	−9.3	91.8	86.4	−5.4			

After−Before summaries: $n = 26$, $\overline{x} = -0.45$, $s = 7.9887$.

Twenty-six young women suffering from anorexia were enrolled in the study. The table gives their weights (in pounds) before treatment began and at the end of the fixed treatment period. We analyze the differences data (after − before).

Assume the differences data are the values of a random sample from a normal distribution with mean μ and standard deviation σ. Use these data to test the null hypothesis $\mu = 0$ (the treatment has not affected mean weight) versus the two-sided alternative $\mu \neq 0$ (the treatment has changed the mean, for better or for worse) at the 5% significance level. Report your conclusion, and comment on your analysis.

3.4.2 Tests for the Variance Parameter

<u>*Size α Tests of $H_O : \sigma^2 = \sigma_o^2$*</u>

	μ Known	*μ Estimated*
Test Statistic	$V = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \mu)^2$	$V = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)S^2}{\sigma_o^2}$
RR for $H_A : \sigma^2 < \sigma_o^2$	$V \leq \chi^2_n(1-\alpha)$	$V \leq \chi^2_{n-1}(1-\alpha)$
RR for $H_A : \sigma^2 > \sigma_o^2$	$V \geq \chi^2_n(\alpha)$	$V \geq \chi^2_{n-1}(\alpha)$
RR for $H_A : \sigma^2 \neq \sigma_o^2$	$V \leq \chi^2_n(1-\alpha/2)$ or $V \geq \chi^2_n(\alpha/2)$	$V \leq \chi^2_{n-1}(1-\alpha/2)$ or $V \geq \chi^2_{n-1}(\alpha/2)$

<u>*Notes:*</u>

1. *Upper Tail Notation:* It is common practice to use upper tail notation for the quantiles used at cutoffs in hypothesis tests. Thus, we use "$x(p)$" to denote the quantile "x_{1-p}."

2. *Left Column:* When μ is known, and $\sigma = \sigma_o$, then

$$V = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma_o}\right)^2 = \sum_{i=1}^{n} Z_i^2$$

   has a chi-square distribution with n degrees of freedom.

3. *Right Column:* When μ is estimated, and $\sigma = \sigma_o$, then

$$V = \frac{1}{\sigma_o^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)S^2}{\sigma_o^2}$$

   has a chi-square distribution with $n - 1$ degrees of freedom, by the sampling distribution theorem we learned in the first notebook.

4. *Standard Deviation Tests:* A test of the null hypothesis $\sigma^2 = \sigma_o^2$ for normal distributions is equivalent to a test of $\sigma = \sigma_o$.

Exercise. In recent years, the yield (in bushels per acre) of corn in the United States has been approximately normally distributed with a mean of 120 bushels per acre and with a standard deviation of 10 bushels per acre.

In a recent survey of 16 farmers, the following yields were reported:

151.401	132.122	104.008	121.201	102.828	109.120	112.403	142.736
124.752	121.322	120.030	105.243	105.567	128.974	138.625	115.289

Sample summaries: $n = 16$, $\overline{x} = 120.976$, $s = 14.7486$.

Assume these data are the values of a random sample from a normal distribution.

(a) Conduct a test of the null hypothesis that $\mu = 120$. Use a two-sided alternative and the 5% significance level. Clearly state your conclusion.

(b) Conduct a test of the null hypothesis that $\sigma = 10$. Use a two-sided alternative and the 5% significance level. Clearly state your conclusion.

(c) Comment on the results of parts (a) and (b).

3.5 Hypothesis Tests for Bernoulli Distributions

Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum of a random sample of size n from a Bernoulli distribution with success probability p.

Approximate Size α Tests of $H_O : p = p_o$

	Small Sample Test	*Large Sample Test*
Test Statistic	$Y = \sum_{i=1}^{n} X_i$	$Z = \frac{Y - np_o}{\sqrt{np_o(1-p_o)}}$
RR for $H_A : p < p_o$	$Y \le c_\alpha$	$Z \le -z(\alpha)$
RR for $H_A : p > p_o$	$Y \ge c_{1-\alpha}$	$Z \ge z(\alpha)$
RR for $H_A : p \ne p_o$	$Y \le c_{\alpha/2}$ or $Y \ge c_{1-\alpha/2}$	$\lvert Z \rvert \ge z(\alpha/2)$

Notes:

1. *Small Sample Test:* The cutoff c_q is chosen so that $P_{p_o}(Y \le c_q) \approx q$ based on the exact binomial distribution of Y when $p = p_o$. See page 11 for an example.

2. *Large Sample Test:* If $p = p_o$, then Z has an approximate standard normal distribution. The quantiles of the standard normal distribution are used as cutoffs for the tests.

   Rule of thumb: The approximation is good when $E_{p_o}(Y) \ge 10$ and $E_{p_o}(n - Y) \ge 10$.

3. *Equivalent Large Sample Test:* The large sample Z statistic can be written equivalently using the sample proportion $\widehat{p} = Y/n$ as follows:

$$Z = \frac{Y - np_o}{\sqrt{np_o(1-p_o)}} = \frac{\widehat{p} - p_o}{\sqrt{p_o(1-p_o)/n}}.$$

4. *Survey Sampling:* The binomial approximation to the hypergeometric distribution can be used to study $p = M/N$, the proportion of individuals in a subpopulation of interest when the population size is N, and the number of individuals in the subpopulation of interest is M. The approximation is good when n is large and

$$n < 0.01N \quad \text{and} \quad 0.05 < \frac{M}{N} < 0.95.$$

Exercise. Suppose that the standard treatment for a given medical condition is effective in 45% of patients. A new treatment promises to be effective in more than 45% of patients.

In order to determine if the proportion of patients for which the new treatment would be effective (p) is just 0.45 or if it is greater, a group of researchers decide to administer the treatment to 250 patients with the condition and to use the 1% significance level.

Assume that the treatment was effective in 51.2% (128/250) of patients in the study.

(a) Use the information above to test $p = 0.45$ versus $p > 0.45$ at the 1% significance level. Clearly state the conclusion, and report the observed significance level (p value).

(b) What assumptions have you made in your analysis in part (a)?

3.6 Hypothesis Tests for Poisson Distributions

Let $Y = \sum_{i=1}^{n} X_i$ be the sample sum of a random sample of size n from a Poisson distribution with parameter λ.

Approximate Size α Tests of $H_O : \lambda = \lambda_o$

	Small Sample Test	*Large Sample Test*
Test Statistic	$Y = \sum_{i=1}^{n} X_i$	$Z = \frac{Y - n\lambda_o}{\sqrt{n\lambda_o}}$
RR for $H_A : \lambda < \lambda_o$	$Y \leq c_\alpha$	$Z \leq -z(\alpha)$
RR for $H_A : \lambda > \lambda_o$	$Y \geq c_{1-\alpha}$	$Z \geq z(\alpha)$
RR for $H_A : \lambda \neq \lambda_o$	$Y \leq c_{\alpha/2}$ or $Y \geq c_{1-\alpha/2}$	$\lvert Z\rvert \geq z(\alpha/2)$

Notes:

1. *Small Sample Test:* The cutoff c_q is chosen so that $P_{\lambda_o}(Y \leq c_q) \approx q$ based on the exact Poisson distribution of Y when $\lambda = \lambda_o$. See page 12 for an example.

2. *Large Sample Test:* If $\lambda = \lambda_o$, then Z has an approximate standard normal distribution. The quantiles of the standard normal distribution are used as cutoffs for the tests.

   Rule of thumb: The approximation is good when $E_{\lambda_o}(Y) > 100$.

3. *Equivalent Large Sample Test:* The large sample Z statistic can be written equivalently using the sample mean rate $\widehat{\lambda} = Y/n$ as follows:

$$Z = \frac{Y - n\lambda_o}{\sqrt{n\lambda_o}} = \frac{\widehat{\lambda} - \lambda_o}{\sqrt{\lambda_o/n}}.$$

Exercise. Phone calls to a medical hotline during the night shift (midnight to 8AM) in a large metropolitan area follow a Poisson process with an average rate of 2 calls per hour.

A medical administrator in a nearby community is interested in determining if the rate for her community (λ) is the same or if it is different. She decides to use information gathered over 10 night shifts (80 hours) and the 5% significance level.

Assume that a total of 134 calls were recorded during the 10-night period.

(a) Use the information above to test $\lambda = 2$ versus $\lambda \neq 2$ at the 5% significance level. Clearly state the conclusion, and report the observed significance level (p value).

(b) What assumptions have you made in your analysis in part (a)?

3.7 Likelihood Ratio Tests

Likelihood ratio tests were introduced by J. Neyman and E. Pearson in the 1930's.

In many practical situations, likelihood ratio tests are uniformly most powerful. In situations where no uniformly most powerful test (UMPT) exists, likelihood ratio tests are popular choices because they have good statistical properties.

3.7.1 Likelihood Ratio Test and Neyman-Pearson Lemma For Simple Hypotheses

Let X_1, X_2, ..., X_n be a random sample from a distribution with parameter θ, and let $Lik(\theta)$ be the likelihood function based on this sample. Consider testing the null hypothesis $\theta = \theta_o$ versus the alternative hypothesis $\theta = \theta_1$, where θ_o and θ_1 are constants.

1. The *likelihood ratio statistic*, Λ, is the ratio of the likelihood functions:

$$\Lambda = \frac{Lik(\theta_o)}{Lik(\theta_1)}.$$

2. A *likelihood ratio test* based on Λ is a test whose decision rule has the following form

$$\text{Reject } \theta = \theta_o \text{ in favor of } \theta = \theta_1 \text{ when } \Lambda \leq c,$$

   for some cutoff, c.

   Note that

   (a) If $\theta = \theta_o$, then the numerator of Λ is likely to be larger than the denominator.
   (b) If $\theta = \theta_1$, then the numerator of Λ is likely to be smaller than the denominator.

The following theorem states that the likelihood ratio test is a uniformly most powerful test for a simple null hypothesis versus a simple alternative hypothesis.

Neyman-Pearson Lemma. Given the situation above, if the critical value c is chosen so that $P_{\theta_o}(\Lambda \leq c) = \alpha$, then the test with decision rule

$$\text{Reject } \theta = \theta_o \text{ in favor of } \theta = \theta_1 \text{ when } \Lambda \leq c$$

is a uniformly most powerful test of size α.

Note that the Neyman-Pearson Lemma is often applied as a first step in attempts to find a uniformly most powerful test (a UMPT) when one or both hypotheses are not simple.

Further, a likelihood ratio test is not implemented as shown above. Instead, an equivalent test (with a simpler statistic and rejection region) is used.

It is instructive to give an outline of a key idea Neyman and Pearson used to prove the theorem above. I will consider continuous random variables here; sums replace integrals when considering discrete random variables.

Let the statistic T be a competitor for the likelihood ratio statistic Λ. The following table gives information about the acceptance and rejection regions of each test:

	RR_T	AR_T
RR_Λ $(\Lambda \leq c)$: $f_o(\boldsymbol{x}) \leq c f_1(\boldsymbol{x})$	E_1	E_2
AR_Λ $(\Lambda > c)$: $f_o(\boldsymbol{x}) > c f_1(\boldsymbol{x})$	E_3	E_4

The notation $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ refers to a data list. E_1 corresponds to data lists leading to rejection under both tests, E_2 corresponds to data lists leading to rejection when using Λ but acceptance when using T, and so forth. Now,

1. We know that $P_o(\Lambda \in RR_\Lambda) = P_o(T \in RR_T) = \alpha$.

   Since RR_Λ corresponds to $E_1 \cup E_2$ and RR_T corresponds to $E_1 \cup E_3$, we can write:

   $$\int_{E_1 \cup E_2} f_o(\boldsymbol{x})\ d\boldsymbol{x} = \int_{E_1 \cup E_3} f_o(\boldsymbol{x})\ d\boldsymbol{x} \quad \Rightarrow \quad \int_{E_2} f_o(\boldsymbol{x})\ d\boldsymbol{x} = \int_{E_3} f_o(\boldsymbol{x})\ d\boldsymbol{x}$$

2. We would like to show that $P_1(\Lambda \in RR_\Lambda) \geq P_1(T \in RR_T)$.

   Since RR_Λ corresponds to $E_1 \cup E_2$ and RR_T corresponds to $E_1 \cup E_3$, this is equivalent to showing that

   $$\int_{E_2} f_1(\boldsymbol{x})\ d\boldsymbol{x} \geq \int_{E_3} f_1(\boldsymbol{x})\ d\boldsymbol{x}.$$

   (a) Since $f_o(\boldsymbol{x}) \leq c f_1(\boldsymbol{x})$ on E_2:

   $$\int_{E_2} f_1(\boldsymbol{x})\ d\boldsymbol{x} \geq \frac{1}{c} \int_{E_2} f_o(\boldsymbol{x})\ d\boldsymbol{x}.$$

   (b) Since $f_o(\boldsymbol{x}) > c f_1(\boldsymbol{x})$ on E_3:

   $$\int_{E_3} f_1(\boldsymbol{x})\ d\boldsymbol{x} \leq \frac{1}{c} \int_{E_3} f_o(\boldsymbol{x})\ d\boldsymbol{x}.$$

   (c) Since the integrals on the right are equal by assumption, we have

   $$\int_{E_2} f_1(\boldsymbol{x})\ d\boldsymbol{x} \geq \frac{1}{c} \int_{E_2} f_o(\boldsymbol{x})\ d\boldsymbol{x} = \frac{1}{c} \int_{E_3} f_o(\boldsymbol{x})\ d\boldsymbol{x} \geq \int_{E_3} f_1(\boldsymbol{x})\ d\boldsymbol{x}.$$

   Thus, the power of the test based on the likelihood ratio statistic Λ is at least as large as the power of any competitor T whose size is equal to the size of Λ.

Example: Upper tail test for a Bernoulli proportion. Let Y be the sample sum of a random sample of size n from a Bernoulli distribution, and consider testing

$$p = p_o \text{ versus } p = p_1, \text{ where } p_1 > p_o.$$

The computations below demonstrate that the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p = p_1 \text{ when } \Lambda \le c$$

is equivalent to the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p = p_1 \text{ when } Y \ge k$$

where c and k satisfy $P_{p_o}(\Lambda \le c) = P_{p_o}(Y \ge k) = \alpha$, for some α.

1. Since $Lik(p) = p^Y(1-p)^{n-Y}$, the likelihood ratio statistic is

$$\Lambda = \frac{Lik(p_o)}{Lik(p_1)} = \frac{(p_o)^Y(1-p_o)^{n-Y}}{(p_1)^Y(1-p_1)^{n-Y}} = \left(\frac{p_o}{p_1}\right)^Y \left(\frac{1-p_o}{1-p_1}\right)^{n-Y}.$$

2. The following inequalities are equivalent:

$$\begin{aligned}
\Lambda \le c &\iff \log(\Lambda) \le \log(c) \\
&\iff Y\log\left(\frac{p_o}{p_1}\right) + (n-Y)\log\left(\frac{1-p_o}{1-p_1}\right) \le \log(c) \\
&\iff Y\Big(\log\left(\frac{p_o}{p_1}\right) - \log\left(\frac{1-p_o}{1-p_1}\right)\Big) \le \log(c) - n\log\left(\frac{1-p_o}{1-p_1}\right) \\
&\iff Y\log\left(\frac{p_o(1-p_1)}{p_1(1-p_0)}\right) \le \log(c) - n\log\left(\frac{1-p_o}{1-p_1}\right) \\
&\iff Y \ge k \qquad \text{where } k = \Big(\log(c) - n\log\left(\frac{1-p_o}{1-p_1}\right)\Big)\Big/\log\left(\frac{p_o(1-p_1)}{p_1(1-p_0)}\right)
\end{aligned}$$

The inequality switches since $p_1 > p_o$ implies that the ratio $\left(\frac{p_o(1-p_1)}{p_1(1-p_0)}\right)$ is less than 1, and its logarithm is a negative number.

Thus, by the Neyman-Pearson lemma, the test based on Y is uniformly most powerful.

Further, since the form of the test based on Y depends only on the fact that $p_1 > p_o$, we can now conclude that the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p > p_o \text{ when } Y \ge k$$

is a uniformly most powerful size α test of the null hypothesis $p = p_o$ versus the one-sided alternative $p > p_o$, where k is chosen so that $P_{p_o}(Y \ge k) = \alpha$.

Exercise: Lower tail test for a Bernoulli proportion. Let Y be the sample sum of a random sample of size n from a Bernoulli distribution. Use the Neyman-Pearson Lemma (and the example above) to demonstrate that the test with decision rule

$$\text{Reject } p = p_o \text{ in favor of } p < p_o \text{ when } Y \leq k$$

is a uniformly most powerful size α test of the null hypothesis $p = p_o$ versus the one-sided alternative $p < p_o$, where k is chosen so that $P_{p_o}(Y \leq k) = \alpha$.

Solution:

(1) As a first step, we fix a value of $p_1 < p_o$ and consider testing $p = p_o$ versus $p = p_1$ using a likelihood ratio test. Now (please complete),

(2) Since p_1 was arbitrarily chosen in step (1), (please complete)

Question: Do you think a UMPT exists for $p = p_o$ versus $p \neq p_o$? Why?

Exercise: Upper tail test for an exponential mean. Let Y be the sample sum of a random sample of size n from an exponential distribution with parameter $\frac{1}{\theta}$ and PDF

$$f(x) \;=\; \frac{1}{\theta}\, e^{-x/\theta} \quad \text{when } x > 0, \text{ and 0 otherwise.}$$

Use the Neyman-Pearson lemma to demonstrate that the test with decision rule

$$\text{Reject } \theta = \theta_o \text{ in favor of } \theta > \theta_o \text{ when } Y \geq k$$

is a uniformly most powerful size α test of the null hypothesis $\theta = \theta_o$ versus the one-sided alternative $\theta > \theta_o$, where k is chosen so that $P_{\theta_o}(Y \geq k) = \alpha$.

Solution:

(1) The likelihood function is as follows (please complete):

(2) Fix a value of $\theta_1 > \theta_o$ and consider testing $\theta = \theta_o$ versus $\theta = \theta_1$ using a likelihood ratio test. Now (please complete),

(3) Since λ_1 was arbitrarily chosen in step (2), (please complete)

3.7.2 Generalized Likelihood Ratio Tests

The methods in this section generalize the approach above to tests with compound hypotheses. Generalized likelihood ratio tests are not guaranteed to be uniformly most powerful. In fact, in many situations (e.g. two tailed tests) uniformly most powerful tests do not exist.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ, and let $Lik(\theta)$ be the likelihood function based on this sample. Consider testing

$$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o$$

where Ω is the set of all parameters, and $\omega_o \subset \Omega$ is the subset under the null hypothesis.

1. The *(generalized) likelihood ratio statistic*, Λ, is the ratio of the maximum value of the likelihood function for models satisfying the null hypothesis to the maximum value of the likelihood function for <u>*all*</u> models under consideration:

   $$\Lambda = \frac{\max_{\theta \in \omega_o} Lik(\theta)}{\max_{\theta \in \Omega} Lik(\theta)}$$

   Note that

   (a) $\Lambda \leq 1$ since the maximum over a subset of Ω must be less than or equal to the maximum over the full parameter space.

   (b) The denominator is maximized at the maximum likelihood (ML) estimator, $\widehat{\theta}$.

2. A *(generalized) likelihood ratio test* based on this statistic is a test whose decision rule has the following form

   $$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } \Lambda \leq c.$$

   Note that

   (a) If $\theta \in \omega_o$, then the numerator and denominator of Λ are likely to be approximately equal, and the value of Λ is likely to be close to 1.

   (b) If $\theta \in \Omega \setminus \omega_o$, then the numerator of Λ is likely to be much smaller than the denominator, and the value of Λ is likely to be close to 0.

Note that, in practice, the term "likelihood ratio test" refers to the generalized test given above and not to the form given in the Neyman-Pearson lemma.

Further, as in the simple case given in the Neyman-Pearson lemma, the (generalized) likelihood ratio test is not implemented as shown above. Instead, an equivalent test (with a simpler statistic and rejection region) is used.

Example: Two tailed test of a Normal mean, when σ is estimated. Let $\overline{X}$ be the sample mean and S^2 be the sample variance of a random sample of size n from a normal distribution with both parameters unknown. Consider testing

$$\mu = \mu_o \text{ versus } \mu \neq \mu_o$$

using a generalized likelihood ratio test.

In this case, $\theta = (\mu, \sigma^2)$, the

- Full parameter space is $\Omega = \{(\mu, \sigma^2) | \mu \in \mathbf{R}, \sigma^2 > 0\}$,
- Null parameter space is $\omega_o = \{(\mu_o, \sigma^2) | \sigma^2 > 0\}$,
- Likelihood function is

$$Lik(\theta) = Lik(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu)^2/(2\sigma^2)} = \sigma^{-n}\,(2\pi)^{-n/2}\, e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2},$$

- Generalized likelihood ratio statistic is

$$\Lambda = \frac{\max_{\theta\in\omega_o} Lik(\theta)}{\max_{\theta\in\Omega} Lik(\theta)} = \frac{Lik\left(\mu_o, \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_o)^2\right)}{Lik\left(\overline{X}, \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)}.$$

Further, since

$$\Lambda \leq c \iff \cdots \iff \frac{(\overline{X}-\mu_o)^2}{S^2/n} \geq k_* \iff \left|\frac{\overline{X}-\mu_o}{\sqrt{S^2/n}}\right| \geq k,$$

for appropriately chosen c and k, the test of the null hypothesis $\mu = \mu_o$ versus the two-sided alternative hypothesis $\mu \neq \mu_o$ we used earlier is a generalized likelihood ratio test.

Footnotes:

1. This test is not uniformly most powerful:

   (a) If $\mu_1 > \mu_o$, then the appropriate upper tail test would be more powerful than the two tailed test;

   (b) if $\mu_1 < \mu_o$, then the appropriate lower tail test would be more powerful than the two tailed test.

2. Similar methods could be used to show that other two tailed tests mentioned earlier are generalized likelihood ratio tests (or approximate generalized likelihood ratio tests). None of these would be uniformly most powerful.

3.7.3 Large Sample Theory: Wilks' Theorem

In many situations, the exact distribution of Λ (or any equivalent form) is not known.

In the 1930's, Wilks proved a useful large sample approximation to the distribution of $-2\log(\Lambda)$ that can be used to construct approximate likelihood ratio tests.

> ***Wilks' Theorem.*** Let Λ be the (generalized) likelihood ratio statistic for a test of the following null versus alternative hypotheses
>
> $$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o$$
>
> based on a random sample of size n from the X distribution. Under smoothness conditions on the X distribution, and when n is large, the test statistic
>
> $-2\log(\Lambda)$ has an approximate chi-square distribution with $r - r_o$ *df*,
>
> where
>
> 1. r is the number of free parameters in Ω,
> 2. r_o is the number of free parameters in ω_o, and
> 3. log() is the natural logarithm function.

Note that the smoothness conditions mentioned in Wilks' theorem are the same as the ones mentioned for the Cramer-Rao lower bound and Fisher's theorem.

3.7.4 Approximate Likelihood Ratio Tests; Special Cases

Under the conditions of Wilks' Theorem, an approximate $100\alpha\%$ test of

$$H_O\colon\ \theta \in \omega_o \qquad \text{versus} \qquad H_A\colon\ \theta \in \Omega \setminus \omega_o$$

has the following decision rule:

$$\text{Reject } \theta \in \omega_o \text{ in favor of } \theta \in \Omega \setminus \omega_o \text{ when } -2\log(\Lambda) \geq \chi^2_{r-r_o}(\alpha)$$

where

(i) $\chi^2_{r-r_o}(\alpha)$ is the $100(1-\alpha)\%$ point on the chi-square distribution with $r - r_o$ *df*,

(ii) r is the number of free parameters in Ω, and

(iii) r_o is the number of free parameters in ω_o.

Example: Comparing k Bernoulli proportions. Consider testing the null hypothesis that k Bernoulli parameters are equal,

$$H_O : p_1 = p_2 = \cdots = p_k,$$

versus the alternative hypothesis that not all parameters are equal, using independent random samples of size n_i from Bernoulli distributions with success probabilities p_i, for $i = 1, 2, \ldots, k$.

In this case, $\theta = (p_1, p_2, \ldots, p_k)$, the

- Full parameter space is $\Omega = \{(p_1, p_2, \ldots, p_k) : 0 < p_1, p_2, \ldots, p_k < 1\}$,
- Null parameter space is $\omega_o = \{(p, p, \ldots, p) : 0 < p < 1\}$,
- There are k free parameters in Ω and 1 free parameter in ω_o, and
- The statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k} \left[2Y_i \log\left(\frac{Y_i}{n_i\widehat{p}}\right) + 2(n_i - Y_i)\log\left(\frac{n_i - Y_i}{n_i(1-\widehat{p})}\right) \right],$$

  where Y_i is the sample sum of the i^{th} sample, for $i = 1, 2, \ldots, k$, $\widehat{p}$ is the estimate of the common parameter under the null hypothesis:

$$\widehat{p} = \frac{Y_1 + Y_2 + \cdots + Y_k}{n_1 + n_2 + \cdots + n_k}$$

  and log() is the natural logarithm function.

If each n_i is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k-1)$ df.

Note that, by independence, the likelihood function can be written as the product of likelihood functions for each sample,

$$Lik(\theta) = Lik(p_1, p_2, \ldots, p_k) = \prod_{i=1}^{k} Lik_i(p_i),$$

and the ratio of likelihoods can also be written as a product. Now, since the logarithm of a product equals the sum of the logarithms, the $-2\log(\Lambda)$ statistic is the sum of k components.

Question: What is the form of the likelihood $Lik_i(p_i)$ for each i?

Example (Source: Hand et al, 1994, page 237). As part of a study on depression in adolescents (ages 12 through 18), researchers collected information on 465 individuals who were seriously emotionally disturbed (SED) or learning disabled (LD). The following table summarizes one aspect of the study:

Group	*LD/SED*	*Gender*	*#Depressed* (y_i)	*#Individuals* (n_i)	*Proportion Depressed* ($\widehat{p}_i = y_i/n_i$)
1	LD	Male	41	219	0.1872
2	LD	Female	26	102	0.2549
3	SED	Male	13	95	0.1368
4	SED	Female	17	49	0.3469
			$y = 97$	$n = 465$	$\widehat{p} = y/n = 0.2086$

Let p_i be the probability that an adolescent in the i^{th} group is severely depressed, and assume that the information in the table summarizes independent random samples from Bernoulli distributions.

We are interested in conducting a test of $p_1 = p_2 = p_3 = p_4$ versus the alternative hypothesis that not all proportions are equal using the generalized likelihood ratio test and the 5% significance level. The rejection region for this test is

$$-2\log(\Lambda) \geq \chi^2_3(0.05) = 7.81.$$

The components of the test statistic, and its observed value are as follows:

	LD/Male	*LD/Female*	*SED/Male*	*SED/Female*	
Component of $-2\log(\Lambda)$*:*	0.623	1.260	3.273	5.000	10.156

Thus, (please state conclusions and comments)

Example: Comparing k Poisson means. Consider testing the null hypothesis that k Poisson means are equal,

$$H_O : \lambda_1 = \lambda_2 = \cdots = \lambda_k,$$

versus the alternative hypothesis that not all means are equal, using independent random samples of size n_i from Poisson distributions with means λ_i, for $i = 1, 2, \ldots, k$.

In this case, $\theta = (\lambda_1, \lambda_2, \ldots, \lambda_k)$,

$$\Omega = \{(\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_1, \lambda_2, \ldots, \lambda_k > 0\} \text{ and } \omega_o = \{(\lambda, \lambda, \ldots, \lambda) : \lambda > 0\}.$$

There are k free parameters in Ω and 1 free parameter in ω_o.

The statistic $-2\log(\Lambda)$ simplifies to

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2Y_i \log\left(\frac{Y_i}{n_i\widehat{\lambda}}\right),$$

where Y_i is the sample sum of the i^{th} sample, for $i = 1, 2, \ldots, k$, $\widehat{\lambda}$ is the estimate of the common parameter under the null hypothesis:

$$\widehat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_k}{n_1 + n_2 + \cdots + n_k}$$

and log() is the natural logarithm function. If each mean ($E(Y_i) = n_i\lambda$) is large, then $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k - 1)$ df.

Note that, by independence, the likelihood function can be written as the product of likelihood functions for each sample,

$$Lik(\theta) = Lik(\lambda_1, \lambda_2, \ldots, \lambda_k) = \prod_{i=1}^{k} Lik_i(\lambda_i),$$

and the ratio of likelihoods can also be written as a product. Now, since the logarithm of a product equals the sum of the logarithms, the $-2\log(\Lambda)$ statistic is the sum of k components.

Question: What is the form of the likelihood $Lik_i(\lambda_i)$ for each i?

Example (Source: Waller et al, in Lange et al, Wiley, 1994, pages 3-24). As part of a study of incidence of childhood lukemia in upstate New York, data were collected on the number of children contracting the disease in the five-year period from 1978 to 1982.

The following table summarizes the results using six geographic regions of equal total population size (175,000 people per area):

	Region 1	*Region 2*	*Region 3*	*Region 4*	*Region 5*	*Region 6*	
#Cases:	89	86	97	96	120	102	590

The geographic regions run from west to east.

Let λ_i be the average number of childhood leukemia cases for a five-year period in i^{th} geographic region, and assume that the information in the table gives independent observations from Poisson distributions with parameters λ_i (thus, using $n_i = 1$ in each case).

We are interested in conducting a test of the null hypothesis $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 = \lambda_6$ versus the alternative hypothesis that not all means are equal using the generalized likelihood ratio test and the 5% significance level. The rejection region for this test is

$$-2\log(\Lambda) \geq \chi_5^2(0.05) = 11.07.$$

The estimated common rate is $\widehat{\lambda} =$ __________________________.

The components of the test statistic are given in the following table:

	Region 1	*Region 2*	*Region 3*	*Region 4*	*Region 5*	*Region 6*
Component of $-2\log(\Lambda)$						

The sum of the components is __________________________.

Thus, (please state conclusions and comments)

Example: Multinomial goodness-of-fit. Consider a multinomial experiment with k outcomes and probabilities p_i, for $i = 1, 2, \ldots, k$, and consider testing

$$H_O : (p_1, p_2, \ldots, p_k) \in \omega_o \quad \text{versus} \quad H_A : (p_1, p_2, \ldots, p_k) \in \Omega \setminus \omega_o$$

using the results of n independent trials of the multinomial experiment, where

$$\Omega = \left\{ (p_1, p_2, \ldots, p_k) : 0 < p_1, p_2, \ldots, p_k < 1, \textstyle\sum_{i=1}^{k} p_i = 1 \right\},$$

and ω_o depends on the particular family of models of interest.

The parameter is $\theta = (p_1, p_2, \ldots, p_k)$, and there are $k - 1$ free parameters in Ω. There are two cases to consider:

1. *Known Model:* If the null hypothesis is $H_O : (p_1, p_2, \ldots, p_k) = (p_{1_o}, p_{2_o}, \ldots, p_{k_o})$, where the p_{i_o}'s are known, then ω_o contains a single k-tuple and has 0 free parameters.

   The statistic $-2\log(\Lambda)$ simplifies to

   $$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{np_{i_o}}\right),$$

   where log() is the natural logarithm function, and X_i is the number of occurrences of the i^{th} outcome, for $i = 1, 2, \ldots, k$. If n is large, then

   $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k - 1)$ df.

2. *Estimated Model:* If e parameters need to be estimated, then ω_o has e free parameters.

   The statistic $-2\log(\Lambda)$ simplifies to

   $$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{n\widehat{p_{i_o}}}\right),$$

   where log() is the natural logarithm function, X_i is the number of occurrences of the i^{th} outcome, and $\widehat{p_{i_o}}$ is the estimated value of p_{i_o}, for $i = 1, 2, \ldots, k$. If n is large, then

   $-2\log(\Lambda)$ has an approximate chi-square distribution with $(k - 1 - e)$ df.

Example (Source: Plackett, Macmillan, 1974, page 134). In a study of spatial dispersion of houses in a Japanese village, the number of homes in each of 1200 squares of side 100 meters was recorded. There were a total of 911 homes.

The following table summarizes the results:

	0 Homes	*1 Home*	*2 Homes*	*3+ Homes*	
#Squares:	584	398	168	50	1200

We are interested in conducting a test of the null hypothesis that the proportions follow a grouped Poisson distribution,

$$H_O : (p_1, p_2, p_3, p_4) = (P_\lambda(X = 0), P_\lambda(X = 1), P_\lambda(X = 2), P_\lambda(X \geq 3))$$

(where X is a Poisson random variable), versus the alternative hypothesis that the proportions do not follow such a model using the generalized likelihood ratio test and the 5% significance level. Since ω_o has 1 free parameter, the rejection region for this test is

$$-2\log(\Lambda) \geq \chi_2^2(0.05) = 5.99.$$

Using $\widehat{\lambda} = \frac{911}{1200} = 0.7592$, the estimated probabilities ($\widehat{p_i}$) under the null hypothesis are

	0 Homes	*1 Home*	*2 Homes*	*3+ Homes*
Estimated Probability				

Further, the components of the likelihood ratio statistic are as follows:

	0 Homes	*1 Home*	*2 Homes*	*3+ Homes*
Component of $-2\log(\Lambda)$				

The sum of the components is __________________________.

Thus, (please state conclusions and comments)

3.7.5 Pearson's Goodness-of-Fit Test Revisited

Consider the problem of testing goodness-of-fit to a multinomial model with probabilities $(p_1, p_2, \ldots, p_k)$. Pearson's statistic for goodness-of-fit,

$$\boldsymbol{X^2} = \sum_{i=1}^{k} \frac{(X_i - n\widehat{p}_i)^2}{n\widehat{p}_i},$$

is closely related to the likelihood ratio statistic discussed in the last section,

$$-2\log(\Lambda) = \sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{n\widehat{p}_i}\right).$$

Specifically, $\boldsymbol{X^2}$ is a second order Taylor approximation to $-2\log(\Lambda)$.

> *Note:* Several important results follow from knowing that $\boldsymbol{X^2}$ is a second order Taylor approximation to $-2\log(\Lambda)$: (1) we know that their values are close when n is large, (2) we know that both have approximate chi-square distributions with $(k-1-e)$ *df*, where e is the number of estimated parameters, and (3) we know that the test based on $\boldsymbol{X^2}$ is an approximate generalized likelihood ratio test.

Spatial dispersion example, continued: The observed value of the likelihood ratio statistic for the spatial dispersion example on the previous page is ______________, and the observed value of Pearson's statistic is 3.01. The values of the two statistics are close in this case.

Exercise. To demonstrate that $\boldsymbol{X^2}$ is a second order Taylor approximation to $-2\log(\Lambda)$ when the multinomial model has known parameters,

1. Let a be a positive constant, and let $f(x) = 2x\ln\left(\frac{x}{a}\right)$.

   Find simplified expressions for $f'(x)$ and $f''(x)$.

2. Find a simplified expression for the 2^{nd} order Taylor polynomial of $f(x)$ centered at a.

3. Use the results of the first two steps to demonstrate that $-2\log(\Lambda) \approx \boldsymbol{X^2}$, that is, use the results of the first two steps to demonstrate that

$$\sum_{i=1}^{k} 2X_i \log\left(\frac{X_i}{np_i}\right) \approx \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}.$$