1. The factorization theorem is the powerful tool that we use to identify sufficient statistics. For each of the following cases, find the sufficient statistic. In some cases, no simplification works, and you’ll have to say “the whole sample is needed for the sufficient statistic.”

(a) \( X_1, X_2, \ldots, X_n \) is a sample from the exponential distribution with mean \( \lambda \). The probability density is
\[
\frac{1}{\lambda} e^{-\frac{x}{\lambda}} 1(x > 0).
\]

(b) \( X_1, X_2, \ldots, X_n \) is a sample from the Cauchy distribution with median \( \omega \). The probability density is
\[
\frac{1}{\pi} \frac{1}{1 + (x - \omega)^2}.
\]

(c) \( X_1, X_2, \ldots, X_n \) is a sample from the logistic distribution with median \( \mu \). The cumulative distribution function is
\[
F(x) = \frac{1}{1 + e^{-(x-\mu)}}
\]
and the density is
\[
f(x) = \frac{e^{-(x-\mu)}}{(1 + e^{-(x-\mu)})^2}. \quad \text{(You do not need } F \text{ to do this problem.)}
\]

(d) \( X_1, X_2, \ldots, X_n \) is a sample from the lognormal distribution with parameters \( \mu \) and \( \sigma \). (Do not think of \( \mu \) and \( \sigma \) as the mean and standard deviation.) The density is
\[
f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}.
\]

(e) The values \( c_1, c_2, \ldots, c_n \) are \( n \) known positive numbers. The random variables \( X_1, X_2, \ldots, X_n \) are independent, and the distribution of \( X_i \) is Poisson with mean \( c_i \theta \).

**SOLUTION:**

(a) The likelihood is
\[
\prod_{i=1}^{n} \left\{ e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \right\} = \left( e^{-\lambda} \right)^n \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.
\]
The only part of the likelihood in which the \( x_i \)'s and \( \lambda \) cannot be separated is the third factor. There appears \( \sum x_i \), so this must be the sufficient statistic. In random variable form, this is \( \sum X_i \). It would also be reasonable to use \( \bar{x} \) as the sufficient statistic.

(b) The likelihood is
\[
\frac{1}{\pi^n} \prod_{i=1}^{n} \left( 1 + (x_i - \omega)^2 \right).
\]
There is no way to separate the \( x_i \)'s from \( \omega \), so the whole sample is needed for the sufficient statistic.
(c) The likelihood is \( \prod_{i=1}^{n} \frac{e^{-(x_i - \mu)}}{(1 + e^{-(x_i - \mu)})^2} \). The numerator will factor neatly, but the denominator hopelessly entangles the \( x_i \)'s with \( \mu \). The whole sample is needed for the sufficient statistic.

(d) The likelihood is \( \prod_{i=1}^{n} \frac{1}{\sigma x_i \sqrt{2\pi}} e^{-(\frac{(\log_x - \mu)^2}{2\sigma^2})} = \left( \frac{1}{\sigma^n (2\pi)^n/2} \right) \left( \prod_{i=1}^{n} \frac{1}{x_i} \right) e^{-(\frac{1}{2\sigma^2}) \sum_{i=1}^{n} (\log_x - \mu)^2} \). The first factor has parameters but no \( x_i \)'s, and the second factor has \( x_i \)'s but no parameters. The sufficient statistic must therefore come from the exponent in the third factor. Write now

\[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log_x - \mu)^2 = -\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^{n} (\log x_i)^2 - 2 \mu \sum_{i=1}^{n} \log x_i + n \mu^2 \right\} \]

This reveals that the sufficient statistic is \( \left( \sum_{i=1}^{n} (\log x_i)^2 , \sum_{i=1}^{n} \log x_i \right) \).

The recommended way to deal with lognormal data is by taking logarithms. Let \( Y_i = \log X_i \). Then \( Y_1, Y_2, \ldots, Y_n \) will be a normal sample, for which the sufficient statistic is \( \left( \sum_{i=1}^{n} Y_i^2 , \sum_{i=1}^{n} Y_i \right) \).

(e) The likelihood is \( \prod_{i=1}^{n} \frac{e^{-c_i \frac{\theta (\sqrt{c_i} \theta)^{x_i}}{x_i!}}}{\prod_{i=1}^{n} x_i!} = \left( e^{-\theta \sum_{i=1}^{n} c_i} \right) \left( \prod_{i=1}^{n} c_i^{x_i} \right) \theta^{\sum_{i=1}^{n} x_i} \). The only part of the likelihood in which the \( x_i \)'s and \( \theta \) cannot be separated is the third factor. There appears \( \sum_{i=1}^{n} x_i \), so this must be the sufficient statistic. In random variable form, this is \( \sum_{i=1}^{n} X_i \). It would also be reasonable to use \( \bar{X} \) as the sufficient statistic.
2. Suppose that \( x_1, x_2, \ldots, x_n \) are known numbers. The random variables \( Y_1, Y_2, \ldots, Y_n \) are independent, and the distribution of \( Y_i \) is normal \( N(0, \frac{\sigma^2}{x_i^2}) \). The standard deviation of \( Y_i \) is \( \frac{\sigma}{x_i} \). The only unknown parameter is \( \sigma \).

(a) Construct the likelihood for \( Y_1, Y_2, \ldots, Y_n \).
(b) Identify the sufficient statistic. If the sufficient statistic is complicated, indicate why this is so.
(c) Find the maximum likelihood estimate for \( \sigma \).
(d) Show that \( E \hat{\sigma}_{ML}^2 = \sigma^2 \). That is, the maximum likelihood estimate of \( \sigma^2 \) is unbiased.

**SOLUTION:**

For part (a), just write

\[
L = \prod_{i=1}^{n} \frac{x_i}{\sigma \sqrt{2\pi}} e^{- \frac{x_i^2}{2\sigma^2 x_i^2}} = \frac{1}{\sigma^n (2\pi)^{n/2}} \left( \prod_{i=1}^{n} x_i \right) e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 / x_i^2}
\]

(b) The likelihood involves only one expression in the \( y_i \)'s, and that must be the sufficient statistic. It is \( \sum_{i=1}^{n} x_i^2 / y_i^2 \). In random variable form, this would be \( \sum_{i=1}^{n} x_i^2 Y_i^2 \).

(c) Just solve \( \frac{d}{d\sigma} \log L \) \( \text{let} = 0 \). This is

\[
-\frac{n}{\sigma} + \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 y_i^2 \text{let} = 0
\]

This reduces to \( n = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 y_i^2 \), and thus \( \hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 y_i^2 \). In random variable form, this is \( \hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 Y_i^2 \). The maximum likelihood estimate of the standard deviation is of course the square root of this.

(d) Since \( E Y_i^2 = \text{Var} Y_i = \frac{\sigma^2}{x_i^2} \), it follows quickly that \( E \hat{\sigma}_{ML}^2 = \sigma^2 \).
3. The only data you’ve got is \( X \), and it follows a binomial distribution with \( n = 6 \) and with unknown \( p \). You are going to test \( H_0: p = 0.40 \) versus \( H_1: p \neq 0.40 \). You have decided to use rejection set \( R = \{ x = 0 \} \cup \{ x = 6 \} \).

(a) Find the probability of type I error for this test.
(b) Find the probability of type II error if really \( p = 0.90 \).
(c) Find the probability of type II error if really \( p = 1.00 \).
(d) Find the probability of type II error if really \( p = 0.50 \).

SOLUTION:

(a) This is \( P[ X = 0 \mid p = 0.40 ] + P[ X = 6 \mid p = 0.40 ] \) = \( 0.60^6 + 0.40^6 \) = \( 0.046656 + 0.004096 \) = 0.050752. This is very close to 5%.

(b) This is \( 1 - \{ P[ X = 0 \mid p = 0.90 ] + P[ X = 6 \mid p = 0.90 ] \} \) = \( 1 - \{ 0.106 + 0.906 \} \) = \( 1 - \{ 0.000001 + 0.531441 \} \) = \( 1 - 0.531442 \) = 0.468558 \( \approx \) 47%.

(c) The type II error probability is zero. If \( p = 1.00 \), you are certain to see \( X = 7 \), and you cannot make an error.

(d) This is \( 1 - \{ P[ X = 0 \mid p = 0.50 ] + P[ X = 6 \mid p = 0.50 ] \} \) = \( 1 - \{ 0.50^6 + 0.50^6 \} \) = \( 1 - 2 \times 0.015625 \) = \( 1 - 0.03125 \) = 0.96875 \( \approx \) 97%.

You are likely to make this error!

4. Greene County has undertaken a survey related to recycling of beverage bottles. A random sample was taken of \( n \) households, asking whether they recycle beverage bottles. The survey modeled \( X \), the number responding “yes,” as binomial \( (n, p) \), meaning

\[
P[ X = x ] = \binom{n}{x} p^x (1 - p)^{n-x}.
\]

Each household that responded “yes,” was further asked for the number of bottles recycled. It was assumed that the number of bottles follows a Poisson distribution with mean \( \lambda \). Let \( T \) be the total number of bottles resulting from this sampling process. Find the mean and variance of \( T \).

SOLUTION: The conditional law of \( T \), given \( X = x \), is Poisson with parameter \( x \lambda \). It’s just the sum of \( x \) independent Poisson random variables. Thus

\[
E( T \mid X = x ) = x \lambda \quad \text{Var}( T \mid X = x ) = x \lambda
\]

In random variable form, these are

\[
E( T \mid X ) = X \lambda \quad \text{Var}( T \mid X ) = X \lambda.
\]
Then \( E(T) = E\{E(T|X)\} = E(X\lambda) = np\lambda. \)

Also \( \text{Var}(T) = \text{Var}\{E(T|X)\} + E\{\text{Var}(T|X)\} \)

\[
= \text{Var}\{X\lambda\} + E\{X\lambda\} \\
= \lambda^2 \text{Var}(X) + \lambda E X = \lambda^2 np(1-p) + \lambda np \\
= \lambda np(\lambda(1-p) + 1)
\]

5. \( X_1, X_2, \ldots, X_n \) is a sample from the probability law which is uniform on the interval \([-\theta, 5\theta]\). The density is \( f(x|\theta) = \frac{1}{6\theta} 1(-\theta \leq x \leq \theta) \).

(a) Find the method of moments estimate for \( \theta \).
(b) Find the maximum likelihood estimate for \( \theta \).

SOLUTION: For part (a), just notice that for any individual \( X_i \), \( E(X_i) = 2\theta \) as \( 2\theta \) is midway between \(-\theta\) and \( 5\theta \). Replace \( E(X_i) \) by \( \bar{X} \), leading to \( \bar{X} = 2\hat{\theta}_{MM} \). This leads to \( \hat{\theta}_{MM} = \frac{\bar{X}}{2} \).

For (b), write the likelihood for the sample as

\[
L = \prod_{i=1}^{n} \frac{1}{6\theta} 1(-\theta \leq x_i \leq 5\theta) = \frac{1}{(6\theta)^n} 1(-\theta \leq x_{\min}) 1(x_{\max} \leq 5\theta) \\
= \frac{1}{(6\theta)^n} 1(-\theta \leq x_{\min}) 1\left(\frac{x_{\max}}{5} \leq \theta\right) = \frac{1}{(6\theta)^n} 1(-x_{\min} \leq \theta) 1\left(\frac{x_{\max}}{5} \leq \theta\right) \\
= \frac{1}{(6\theta)^n} 1\left(\max\left\{-x_{\min}, \frac{x_{\max}}{5}\right\} \leq \theta\right)
\]

In maximizing \( L \), you want the estimate for \( \theta \) to be as small as possible. This results in \( \hat{\theta}_{ML} = \max\left\{-x_{\min}, \frac{x_{\max}}{5}\right\} \). In random variable form, this is \( \hat{\theta}_{ML} = \max\left\{-X_{\min}, \frac{X_{\max}}{5}\right\} \).