18.443 Exam 1 Spring 2015
Statistics for Applications
3/5/2015

1. **Log Normal Distribution:** A random variable X follows a $Lognormal(\theta, \sigma^2)$ distribution if $Y = \ln(X)$ follows a $Normal(\theta, \sigma^2)$ distribution.

   For the normal random variable $Y = \ln(X)$

   - The probability density function of Y is
   $$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}}, \quad -\infty < y < \infty.$$
   - The moment-generating function of Y is
   $$M_Y(t) = E[e^{tY} \mid \theta, \sigma^2] = e^{t\theta + \frac{1}{2}\sigma^2 t^2}$$

   (a). Compute the first two moments of a random variable $X \sim Lognormal(\theta, \sigma^2)$.

   $\mu_1 = E[X \mid \theta, \sigma^2]$ and $\mu_2 = E[X^2 \mid \theta]$

   Hint: Note that $X = e^Y$ and $X^2 = e^{2Y}$ where $Y \sim N(\theta, \sigma^2)$ and use the moment-generating function of Y.

   (b). Suppose that $X_1, \ldots, X_n$ is an i.i.d. sample from the $Lognormal(\theta, \sigma^2)$ distribution of size n. Find the method of moments estimates of θ and σ^2.

   Hint: evaluate μ_2/μ_1^2 and find a method-of-moments estimate for σ^2 first.

   (c). For the log-normal random variable $X = e^Y$, where

   $Y \sim Normal(\theta, \sigma^2)$,

   prove that the probability density of X is
   $$f(x \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} (\frac{1}{x}) e^{-\frac{1}{2}\frac{(\ln(x)-\theta)^2}{\sigma^2}}, \quad 0 < x < \infty.$$

   (d). Suppose that $X_1, \ldots, X_n$ is an i.i.d. sample from the $Lognormal(\theta, \sigma^2)$ distribution of size n. Find the mle for θ assuming that σ^2 is known to equal σ_0^2.

   (e). Find the asymptotic variance of the mle for θ in (d).

   Solution:

(a).

$$\mu_1 = E[X] = E[e^Y] = M_Y(1) = e^{\theta+\sigma^2/2}$$
$$\mu_2 = E[X^2] = E[e^{2Y}] = M_Y(2) = e^{2\theta+2\sigma^2}$$

(b). First, note that:

$$\mu_2/(\mu_1^2) = e^{\sigma^2}$$

It follows that a method-of-moments estimate for σ^2 is

$$\hat{\sigma}^2 = \ln(\hat{\mu}_2/\hat{\mu}_1^2)$$

where

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i$$
$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

Substituting $\hat{\sigma}^2$ for σ^2 in the formula for μ_1 we get

$$\begin{aligned} \hat{\mu}_1 &= e^{\theta+\hat{\sigma}^2/2} \\ \Longrightarrow \hat{\theta} &= \ln(\hat{\mu}_1) - \hat{\sigma}^2/2 \end{aligned}$$

(c). Consider the transformation

$$X = e^Y.$$

which has the inverse: $y = \ln(x)$ and $dy/dx = 1/x$.

It follows that

$$f_X(x) = f_Y(\ln(x))|dy/dx| = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}e^{-\frac{1}{2\sigma^2}(\ln(x)-\theta)^2}$$

(d). The log of the density function for single realizations x is

$$\ln[f(x \mid \theta, \sigma_0^2] = -\frac{1}{2}\ln(2\pi\sigma_0^2) - \ln(x) - \frac{1}{2}\frac{(\ln(x)-\theta)^2}{\sigma_0^2}$$

For a sample $x_1, \ldots, x_n$, the likelihood function is

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^n \ln[f(x_i \mid \theta, \sigma_0^2] \\ &= -\frac{1}{2\sigma_0^2}\sum_{i=1}^n (\ln(x_i)-\theta)^2 + (\text{terms not depending on } \theta) \end{aligned}$$

$\ell(\theta)$ is minimized by $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n \ln(x_i)$ – the mle from the sample of $Y_i = \ln(X_i)$ values.

(e). The asymptotic variance satisfies

$$E[-\frac{d^2\ell(\theta)}{d\theta^2}] \approx 1/Var(\hat{\theta})$$

Since $\frac{d^2\ell(\theta)}{d\theta^2} = \frac{n}{\sigma_0^2}$ is constant

$$Var(\hat{\theta}) \approx \sigma_0^2/n$$

This asymptotic variance is in fact the actual variance of $\hat{\theta}$.

2. The Pareto distribution is used in economics to model values exceeding a threshhold (e.g., liability losses greater than \$100 million for a consumer products company). For a fixed, known threshhold value of $x_0 > 0$, the density function is

$$f(x \mid x_0, \theta) = \theta x_0^{\theta} x^{-\theta-1}, \quad x \geq x_0, \text{ and } \theta > 1.$$

Note that the cumulative distribution function of X is

$$P(X \leq x) = F_X(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}.$$

(a). Find the method-of-moments estimate of θ.

(b). Find the mle of θ.

(c). Find the asymptotic variance of the mle.

(d). What is the large-sample asymptotic distribution of the mle?

Solution:

(a) Compute the first moment of a Pareto random variable X :

$$\begin{aligned}
\mu_1 &= \int_{x_0}^{\infty} x f(x \mid x_0, \theta) dx \\
&= \int_{x_0}^{\infty} x \times \theta x_0^{\theta} x^{-\theta-1} dx \\
&= \theta x_0^{\theta} \int_{x_0}^{\infty} x^{-\theta} dx \\
&= \theta x_0^{\theta} (\tfrac{1}{\theta-1}) x_0^{-(\theta-1)} \\
&= x_0 \left(\frac{\theta}{\theta-1}\right)
\end{aligned}$$

Solving $\mu_1 = \hat{\mu}_1 = \overline{x}$ for θ gives:

$$\hat{\theta} = \left(\tfrac{\overline{x}}{\overline{x}-x_0}\right)$$

(b). For a single observation $X = x$, we can write

$$\begin{aligned}
\log[f(x \mid \theta)] &= \ln(\theta) + \theta \ln(x_0) - (\theta-1)\ln(x) \\
\tfrac{\partial \log[f(x|\theta)}{\partial \theta}] &= \tfrac{1}{\theta} + \ln(x_0) - \ln(x) \\
\tfrac{\partial^2 \log[f(x|\theta)}{\partial \theta^2}] &= -\tfrac{1}{\theta^2}
\end{aligned}$$

The mle for θ solves

$$\begin{aligned}
0 = \tfrac{\partial \ell(\theta)}{\partial \theta} &= \tfrac{\partial}{\partial \theta}(\textstyle\sum_{i=1}^{n} \ln[f(x_i \mid \theta)]) \\
&= \textstyle\sum_1^n [\tfrac{1}{\theta} + \ln(x_0) - \ln(x_i)] \\
&= \tfrac{n}{\theta} + n\ln(x_0) - \textstyle\sum_1^n \ln(x_i) \\
\Longrightarrow \hat{\theta} &= \tfrac{n}{\sum_1^n \ln(x_i) - n\ln(x_0)} = [\tfrac{1}{n}\textstyle\sum_1^n \ln(x_i/x_0)]^{-1}
\end{aligned}$$

(c). The asymptotic variance of $\hat{\theta}$ is

$$Var(\hat{\theta}) \approx \tfrac{1}{nI(\theta)} = \tfrac{\theta^2}{n}$$

Because $I(\theta) = E[-\frac{\partial^2 \ln[f(x|\theta)]}{\partial\theta^2}] = \frac{1}{\theta^2}$

(d) The asymptotic distribution of $\hat{\theta}$ is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{D}} N(0, \frac{1}{I(\theta)}) = N(0, \theta^2)$$

or

$$\hat{\theta} \xrightarrow{\mathcal{D}} N(\theta, \frac{\theta^2}{n})$$

3. Distributions derived from Normal random variables. Consider two independent random samples from two normal distributions:

- $X_1, \ldots, X_n$ are n i.i.d. $Normal(\mu_1, \sigma_1^2)$ random variables.
- $Y_1, \ldots, Y_m$ are m i.i.d. $Normal(\mu_2, \sigma_2^2)$ random variables.

(a). If $\mu_1 = \mu_2 = 0$, find two statistics

$$T_1(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$
$$T_2(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

each of which is a t random variable and which are statistically independent. Explain in detail why your answers have a t distribution and why they are independent.

(b). If $\sigma_1^2 = \sigma_2^2 > 0$, define a statistic

$$T_3(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

which has an F distribution.

An F distribution is determined by the numerator and denominator degrees of freedom. State the degrees of freedom for your statistic T_3.

(c). For your answer in (b), define the statistic

$$T_4(X_1, \ldots, X_n, Y_1, \ldots, Y_m) = \frac{1}{T_3(X_1, \ldots, X_n, Y_1, \ldots, Y_m)}$$

What is the distribution of T_4 under the conditions of (b)?

(d). Suppose that $\sigma_1^2 = \sigma_2^2$. If $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, and $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2$, are the sample variances of the two samples, show how to use the F distribution to find

$$P({S_X}^2/S_Y^2 > c).$$

(e). Repeat question (d) if it is known that $\sigma_1^2 = 2\sigma_2^2$.

Solution:

(a). Consider

$$T_1 = \frac{\sqrt{n}\overline{X}}{\sqrt{S_X^2}}$$
$$T_2 = \frac{\sqrt{m}\overline{Y}}{\sqrt{S_Y^2}}$$

where

$$\begin{array}{rclcl}
\overline{X} & = & \frac{1}{n}\sum_1^n X_i & \sim & N(\mu_1, \sigma_1^2/n) \\
S_X^2 & = & \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2 & \sim & (\frac{\sigma_1^2}{n-1}) \times \chi^2_{n-1} \\
\overline{Y} & = & \frac{1}{m}\sum_1^m Y_i & \sim & N(\mu_2, \sigma_2^2/n) \\
S_X^2 & = & \frac{1}{m-1}\sum_1^m (Y_i - \overline{Y})^2 & \sim & (\frac{\sigma_2^2}{m-1}) \times \chi^2_{m-1}
\end{array}$$

We know from theory that $\overline{X}$ and S_X^2 are independent, and $\overline{Y}$ and S_Y^2 are independent, and all 4 are mutually independent because they depend on independent samples.

For $\mu_1 = 0$, we can write

$$T_1 = \frac{\sqrt{n}\overline{X}/\sigma_1}{\sqrt{S_X^2/\sigma_1^2}} \sim t_{n-1}$$

a t distribution with $(m-1)$ degrees of freedom, because the numerator is $N(0,1)$ random variable independent of the denominator which is $\sqrt{\chi^2_{m-1}/(m-1)}$.

And for $\mu_2 = 0$, we can write

$$T_2 = \frac{\sqrt{m}\overline{Y}/\sigma_2}{\sqrt{S_Y^2/\sigma_2^2}} \sim t_{m-1}$$

a t distribution with $(n-1)$ degrees of freedom, because the numerator is $N(0,1)$ random variable independent of the denominator which is $\sqrt{\chi^2_{n-1}/(n-1)}$.

(b). For $\sigma_1^2 = \sigma_2^2$ consider the statistic:

$$\begin{array}{rcl}
T_3 & = & \frac{S_X^2}{S_Y^2} \\
& = & \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2}
\end{array}$$

The numerator is a χ^2_{n-1} random variable divided by its degrees of freedom $(n-1)$ and the denominator is an independent χ^2_{m-1} random variable divided by its degrees of freedom $(m-1)$. By definition the distribution of such a ratio is an F distribution with $(n-1)$ and $(m-1)$ degrees of freedom in the numerator/denominator.

(c). The inverse of an F random variable is also an F random variable – the degrees of freedom for numerator and denominator reverse.

(d). In general we know:

$$\begin{array}{rcl}
\frac{(n-1)S_X^2}{\sigma_1^2} & \sim & \chi^2_{n-1} \\
\frac{(m-1)S_Y^2}{\sigma_2^2} & \sim & \chi^2_{m-1}
\end{array}$$

which are independent.

So, we can develop the expression:

$$\begin{aligned}
P(\tfrac{S_X^2}{S_Y^2} > c) &= P(\frac{(n-1)S_X^2/\sigma_1^2}{(m-1)S_Y^2/\sigma_2^2} > \frac{(n-1)/\sigma_1^2}{(m-1)\sigma_2^2} \times c) \\
&= P(F_{(n-1),(m-1)} > \tfrac{(n-1)}{(m-1)} \times (\tfrac{\sigma_2^2}{\sigma_1^2}) \times c)
\end{aligned}$$

The answer is the upper-tail probability of an F distribution with $(n-1), (m-1)$ degrees of freedom, equal to the probability of exceeding $(\frac{(n-1)}{(m-1)} \times (\frac{\sigma_2^2}{\sigma_1^2}) \times c)$

For (d), use $\frac{\sigma_2^2}{\sigma_1^2} = 1$ and for (e) use $\frac{\sigma_2^2}{\sigma_1^2} = 1/2$

4. **Hardy-Weinberg (Multinomial) Model of Gene Frequencies**

For a certain population, gene frequencies are in equilibrium: the genotypes AA, Aa, and aa occur with probabilities $(1-\theta)^2$, $2\theta(1-\theta)$, and θ^2. A random sample of 50 people from the population yielded the following data:

Genotype Type		
AA	Aa	aa
35	10	5

The table counts can be modeled as the multinomial distribution:

$$(X_1, X_2, X_3) \sim Multinomial(n = 50, p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2).$$

(a). Find the mle of θ

(b). Find the asymptotic variance of the mle.

(c). What is the large sample asymptotic distribution of the mle?

(d). Find an approximate 90% confidence interval for θ. To construct the interval you may use the follow table of cumulative probabilities for a standard normal $N(0,1)$ random variable Z

$P(Z < z)$	z
0.99	2.326
0.975	1.960
0.950	1.645
0.90	1.182

(e). Using the mle $\hat{\theta}$ in (a), 1000 samples from the

$$Multinomial(n = 50, p = ((1-\hat{\theta})^2, 2\hat{\theta}(1-\hat{\theta}), \hat{\theta}^2))$$

distribution were randomly generated, and mle estimates were computed for each sample: $\hat{\theta}_j^*, j = 1, \ldots, 1000$.

For the true parameter θ_0, the sampling distribution of $\Delta = \hat{\theta} - \theta_0$ is approximated by that of $\tilde{\Delta} = \hat{\theta}^* - \hat{\theta}$. The 50-th largest value of $\tilde{\Delta}$ was $+0.065$ and the 50-th smallest value was -0.067.

Use this information and the estimate in (a) to construct a (parametric) bootstrap confidence interval for the true θ_0. What is the confidence level of the interval? (If you do not have an answer to part (a), assume the mle $\hat{\theta} = 0.25$).

Solution:

(a). Find the mle of θ

- $(X_1, X_2, X_3) \sim Multinomial(n, p = ((1-\theta)^2, 2\theta(1-\theta), \theta^2))$

- Log Likelihood for θ

$$\begin{aligned}
\ell(\theta) &= log(f(x_1, x_2, x_3 \mid p_1(\theta), p_2(\theta), p_3(\theta))) \\
&= log(\tfrac{n!}{x_1!x_2!x_3!} p_1(\theta)^{x_1} p_2(\theta)^{x_2} p_3(\theta)^{x_3}) \\
&= x_1 log((1-\theta)^2) + x_2 log(2\theta(1-\theta)) \\
&\qquad + x_3 log(\theta^2) + (\text{non-}\theta\ terms) \\
&= (2x_1 + x_2) log(1-\theta) + (2x_3 + x_2) log(\theta) + (\text{non-}\theta\ terms)
\end{aligned}$$

- First Differential of log likelihood:

$$\ell'(\theta) = -\frac{(2x_1 + x_2)}{1-\theta} + \frac{(2x_3 + x_2)}{\theta}$$

$$\Longrightarrow \quad \hat{\theta} = \frac{2x_3 + x_2}{2x_1 + 2x_2 + 2x_3} = \frac{2x_3 + x_2}{2n} = \frac{2(5) + 10}{2(50)} = 0.2$$

(b). Find the asymptotic variance of the mle.

- $Var(\hat{\theta}) \longrightarrow \dfrac{1}{E[-\ell''(\theta)]}$
- Second Differential of log likelihood:

$$\begin{aligned}
\ell''(\theta) &= \frac{d}{d\theta}[-\frac{(2x_1 + x_2)}{1-\theta} + \frac{(2x_3 + x_2)}{\theta}] \\
&= -\frac{(2x_1 + x_2)}{(1-\theta)^2} - \frac{(2x_3 + x_2)}{\theta^2}
\end{aligned}$$

- Each of the X_i are $Binomial(n, p_i(\theta))$ so

$$\begin{aligned}
E[X_1] &= np_1(\theta) = n(1-\theta)^2 \\
E[X_2] &= np_2(\theta) = n2\theta(1-\theta) \\
E[X_3] &= np_3(\theta) = n\theta^2
\end{aligned}$$

- $E[-\ell''(\theta)] = \dfrac{2n}{\theta(1-\theta)}$
- $\hat{\sigma}^2_{\hat{\theta}} = \dfrac{\hat{\theta}(1-\hat{\theta})}{2n} = \dfrac{0.8(1-0.8)}{2 \times 50} = 0.16/100 = (.4/10)^2 = (.04)^2$

(c) The asymptotic distribution of $\hat{\theta}$ is $N(\theta, \frac{\theta(1-\theta)}{2n})$

(d) An approximate 90% confidence interval for θ is given by

$$\{\theta : \hat{\theta} - z(\alpha/2)\sqrt{Var(\hat{\theta})} < \theta < \hat{\theta} + z(\alpha/2)\sqrt{Var(\hat{\theta})}\}$$

where $\alpha = 1 - 0.90$ and $z(.05) = 1.645$, and $\sqrt{Var(\hat{\theta})} \approx (.04)$.

So the approximate 90% confidence interval is:

$$\{\theta : 0.20 - .06580 < \theta < 0.20 + .06580\}$$

(e). For the bootstrap distribution of the errors $\Delta = \hat{\theta} - \theta_0$, (where θ_0 is the true value), the approximate 5% and 95% quantiles are

$\underline{\delta} = -0.067$ and $\overline{\delta} = 0.065$.

The approximate 90% confidence interval is

$$\begin{aligned} & \{\theta : \hat{\theta} - \overline{\delta} < \theta < \hat{\theta} - \underline{\delta}\} \\ = \; & [0.2 - 0.065, 0.2 + 0.067] \end{aligned}$$