

LECTURE 4

Two especially useful statistical tools

The estimation of statistical model parameters and their evaluation are major components of survival analysis (next topics). The following outlines two technical issues central to statistical estimation in general, maximum likelihood estimation and the variances of statistical functions. These somewhat theoretical topics are not critical to understanding the application of survival analysis methods but provide insight into the origins of the model parameter estimates and their variances.

Maximum likelihood estimation

Maximum likelihood techniques are used in the vast majority of statistical analyses to estimate the parameters of models representing the relationships within sampled data. The complexity of this technique lies in the technical details and not the underlying principle. Maximum likelihood estimation is conceptually simple. A small example introduces the fundamental considerations at the heart of maximum likelihood estimation. Suppose a thumb tack tossed in the air has an unknown probability of landing with the point up (denoted p). Furthermore, three tacks are tossed and one lands point up and the

other two land point down. The probability that this event occurs is $3p(1 - p)^2$. When two values are proposed as an estimate of p , it is not hard to decide which is the most likely to have produced the observed result (one up and two down) and, therefore, is the better estimate of the unknown parameter p . For example, the likelihood that one up-tack occurs out of three tossed when $p = 0.2$ (0.384) is four times greater than the probability of the same event when $p = 0.8$ (0.096). The "maximum-likelihood" question becomes: Which of these two postulated probabilities is the best estimate of the unknown underlying value p ? Although no unequivocal answer exists, the answer that the probability $p = 0.2$ is more sensible. The observed data "best" support this answer when the choice is between 0.2 and 0.8.

Maximum likelihood estimation is an extension of this logic. The data are considered as fixed and all possible parameters are considered (not just two outcomes). The parameter that makes the observed data the most likely (maximizes the likelihood of its occurrence) is chosen as the "best" estimate. It is the value most consistent with the observed data. For the three tack example, this value is 0.333. No other choice of p makes the data (one up and two down) more likely. For all other possible values of the parameter p , the probability $3p(1 - p)^2$ is less than $3(0.333)(0.667)^2 = 0.444$. For the tack data (one-up and two-down), the maximum likelihood estimate $\hat{p} = 0.333$ is not the correct value but simply the best available conjecture in light of the observed data.

A slightly more extensive example continues to indicate the logic of the estimation process. Say $n = 50$ tacks are tossed in the air and $x = 15$ land point up. That is, the data are

up, up, down, up, down, down, ... , down and up.

The probability that this event occurred is

$$L = p \times p \times (1 - p) \times p \times (1 - p) \times (1 - p) \times \dots \times (1 - p) \times p$$

or more succinctly

$$L = p^{15} (1 - p)^{35}$$

and is called the *likelihood function*. As with the first example, the value of the parameter p is unknown. The question becomes: Out of all possible values for p , which value makes the observed result ($x = 15$ up-tacks) most likely to have occurred? The answer is found by calculating the likelihood function for all possible values of p . Because sums are easier to describe conceptually and deal with mathematically, instead of the likelihood L (a product), the logarithm of L (a sum) is used [denoted $\log(L)$]. For the tack example, the likelihood value L is the product $L = p^{15} (1 - p)^{35}$ and the log-likelihood value is the sum $\log(L) = 15 \log(p) + 35 \log(1 - p)$. Any value that maximizes the log-likelihood function also maximizes the likelihood function. For the thumb tack data, 12 selected values of p produce the log-likelihood values in Table 1.0.

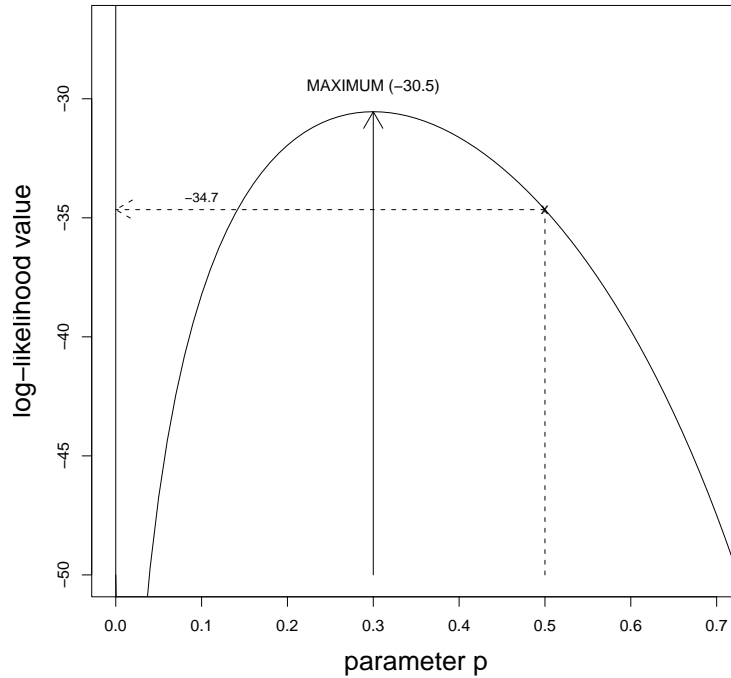
Table 1.0 Values of the parameter p and the corresponding log-likelihood summary values for $x = 15$ and $n = 50$

p	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
$\log(L)$	-46.7	-38.2	-34.1	-32.0	-30.9	-30.5	-30.8	-31.6	-32.9	-34.7	-36.9	-39.7

In fact, the possible values of p range continuously from 0 to 1. Figure 1.0 displays $\log(L)$ for the relevant range of p (0 to 0.7) showing increasing log-likelihood values

until $p = 0.3$ and then decreasing log-likelihood values after $p = 0.3$.

Figure 1.0 The log-likelihood function for the thumb tack data ($n = 50$ and $x = 15$)



The value 0.3 is the value of p that maximizes the log-likelihood summary $\log(L)$ and, therefore, maximizes the likelihood function L . It is denoted \hat{p} and is called *the maximum likelihood estimate of the parameter p*. No other value is more consistent with the data. The occurrence of 15 up-tacks (35 down-tacks) is most likely when p is 0.30 making $\hat{p} = 0.3$ the maximum likelihood estimate. That is, the log-likelihood value $\log(L_{\hat{p}=0.3}) = -30.5$ is greater than $\log(L)$ for all other values of p .

A natural and commonly used estimate of the probability an up-tack is the proportion of up-tacks among the total number tossed or $\hat{p} = 15/50 = 0.3$. An amazing property of maximum likelihood estimation is that it frequently provides a rigorous justification

for "everyday" estimates. For example, mean values, proportions and rates are frequently maximum likelihood estimates.

A maximum likelihood estimate is typically derived with a calculus argument. The thumb tack example continues to illustrate. The maximum of a single valued function is that point where the derivative of the function is zero. In symbols, the maximum of a function represented by $f(x)$ occurs at the value of x that is the solution to the equation

$\frac{d}{dx} f(x) = 0$. For example, when x tacks land up out of n tosses, then

$$\frac{d}{dp} f(p) = \frac{d}{dp} \log(L) = \frac{d}{dp} \left[x \log(p) + (n - x) \log(1 - p) \right] = 0.$$

Thus,

$$\frac{x}{\hat{p}} - \frac{n - x}{1 - \hat{p}} = 0 \quad \text{yields the solution} \quad \hat{p} = \frac{x}{n}$$

where $f(p) = \log(L) = x \log(p) + (n - x) \log(1 - p)$ is the log-likelihood function for all possible parameter values p ($0 \leq p \leq 1$). Again, the estimated value \hat{p} maximizes the likelihood function and is also the natural estimate of the parameter p (proportion of up-tacks).

In addition, the variance of a maximum likelihood estimate can be estimated from the log-likelihood function. For the example, the variance of the distribution of \hat{p} is estimated by $\hat{p}(1 - \hat{p})/n$. In general, maximum likelihood estimates are found with a computer program so the details of the estimation process as well as the derivation of the

variance expression are rarely issues when analyzing data and are left to more theoretical presentations.

When more than one estimated parameter is involved, the notation and computation become more elaborate but the maximum likelihood principle remains the same. Regardless of the complexity of the likelihood function, the chosen estimates are the values that are most likely to have produced the sampled data. Suppose that l parameters are to be estimated, then the maximum likelihood estimates are the l parameters that make the likelihood function or the log-likelihood function as large as possible. In symbols, the l parameters represented by $\theta_1, \theta_2, \theta_3, \dots, \theta_l$ have maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots, \hat{\theta}_l$ when the likelihood value L evaluated at these values is larger than the likelihood values calculated from all other possible values of the parameters $\theta_1, \theta_2, \theta_3, \dots, \theta_l$ or $L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_l) > L(\theta_1, \theta_2, \dots, \theta_l)$. The computational process of finding these specific estimates and their variances is an extremely tedious for more than two or three parameters. A computer algorithm is almost always used.

For most statistical techniques, the parameter values are thought of as fixed and the data are subject to sampling variation. Maximum likelihood estimation reverses the situation. The data are considered fixed and the parameters are varied to determine the value or values that maximize the likelihood function.

Four key properties of maximum likelihood estimates are:

1. Maximum likelihood estimates based on large numbers of observations have approximately normal distributions. Often, as few as 20 or 30 observations are sufficient to

produce estimates with approximate normal distributions. Therefore, the evaluation of maximum likelihood estimates in terms of confidence intervals and statistical tests follow typical statistical patterns. For example, if $\hat{\theta}$ represents a maximum likelihood estimate, then an approximate 95% confidence interval is $\hat{\theta} \pm 1.960 \sqrt{\text{variance}(\hat{\theta})}$. In addition, the test statistic

$$X^2 = z^2 = \frac{(\hat{\theta} - \theta_0)^2}{\text{variance}(\hat{\theta})}$$

has an approximate chi-square distribution with one degree of freedom when θ_0 is the "true" underlying parameter. The word "true" in this statistical context means that the difference between the estimate $\hat{\theta}$ and the parameters θ_0 is due entirely to random variation. The maximum likelihood estimated variance, represented by $\text{variance}(\hat{\theta})$, serves as an estimate the variance of the approximate normal distribution that describes the variation of the estimated value, $\hat{\theta}$. This chi-square assessment of a maximum likelihood estimated parameter is frequently called *Wald's test*.

2. A maximum likelihood estimate is optimal in the sense that it usually has a smaller variance than competing estimates. When the sample size is large, the maximum likelihood estimate is generally the most precise estimate available. Thus, for a wide variety of analyses, the estimates that most efficiently utilize the sampled data are found by maximizing the likelihood function.

3. As noted, the estimated variance of the approximate normal distribution of a

maximum likelihood estimate is part of the estimation process. The computer algorithm that produces the estimate, produces an estimate of its variance.

4. A function of a maximum likelihood estimate is itself a maximum likelihood estimate and has properties 1, 2 and 3. For example, when the estimate $\hat{\theta}$ is a maximum likelihood estimate, then $e^{\hat{\theta}}$ or $\sqrt{\hat{\theta}}$ or $n\hat{\theta}$ or $1/\hat{\theta}$ are also maximum likelihood estimates. These estimates also have approximate normal distributions and minimum variances for large sample sizes. For example, because $\hat{p} = 0.3$ is the maximum likelihood estimate of the probability that a tack lands up, then $n\hat{p} = 100(0.3) = 30$ is the maximum likelihood estimate of the number of up-tacks that would occur among 100 tosses. Furthermore, the probability $\hat{q} = 1 - \hat{p} = 1 - 0.3 = 0.7$ making \hat{q} the maximum likelihood estimate of the probability of a tack landing point down.

Likelihood statistics

Producing an optimal estimate from sampled data is only one of the valuable properties of a likelihood function. The likelihood function or the logarithm of the likelihood function (not just the maximum) reflects the probability that the collected data occurred for a specific set of parameters. For example, if $p = 0.2$, then the likelihood of one-up and two-down tacks is $L = 3p(1 - p)^2 = 3(0.2)(0.8)^2 = 0.384$.

Another example, from Table 1.0, shows the value $\log(L)$ is -34.7 if p was 0.5 . This value is clearly not the maximum but, nevertheless, reflects the probability that fifteen up-tacks occurs when p is 0.5 . The maximum value of the log-likelihood occurs at $\hat{p} = 0.3$ where is $\log(L) = -30.5$. These two log-likelihood values (-34.7 and -30.5) differ for one of two distinct reasons. Because the estimate \hat{p} is subject to random variation, the two log-likelihood values possibly differ simply by chance. Alternatively, the value $p = 0.5$ may not be the underlying parameter, causing the two log-likelihood to systematically differ.

To choose between these two alternatives (random versus systematic?) based on the observed difference between two log-likelihood values generated from two statistical models, a theorem from theoretical statistics is used. It is a statistical fact that the difference between two log-likelihood values multiplied by -2 has an approximate chi-square distribution when three conditions hold. The first condition is that the two models generating the log-likelihood values are calculated from exactly the same data. Second, the compared models must be nested. Nested means that one model is a special case of the

other (examples follow). Third, the two log-likelihood values differ by chance alone. When the first two conditions apply, a chi-square test statistic (called the *likelihood ratio test statistic*) produces an assessment of the third condition in terms of a probability. The question becomes: Is the observed difference between log-likelihood values calculated from two nested models likely random? To help answer this question, the two log-likelihood values and a chi-square distribution produce a significance probability (*p*-value).

The thumb tack tossing example conforms to all three requirements if the actual underlying probability that a tack lands up is $p = 0.5$ (null hypothesis). Then, the likelihood ratio test statistic (using the log-likelihood values from Table 1.0 and Figure 1.0)

$$X^2 = -2[\log(L_{p=0.5}) - \log(L_{p=0.3})] = -2[(-34.657) - (-30.543)] = 8.228$$

is a single observation from a chi-square distribution with one degree of freedom. The degrees of freedom are the difference between the number of parameters estimated to calculate each likelihood value. For $\log(L_{p=0.3})$, one estimate is made ($\hat{p} = 0.3$) and for $\log(L_{p=0.5})$, no estimate is made ($p = 0.5$ was selected) yielding one degree of freedom. The probability of a more extreme difference between log-likelihood values arising by chance alone is then $p\text{-value} = P(X^2 \geq 8.228 | p = 0.5) = 0.004$ from a chi-square distribution with one degree of freedom. The conjecture that $p = 0.5$ is the underlying parameter of the distribution that produced the thumb tack data is not plausible, indicating that the actual value of p is not likely 0.5 and more likely closer to 0.3.

Calculating the differences between two log-likelihood values is a fundamental statistical tool and applies to comparing a large variety of models with any number of

parameters. In general, for a model based on k variables and l parameters, the log-likelihood value is represented by

$$\log(L_1) = \log - \text{likelihood} = \log[L(x_1, x_2, x_3, \dots, x_k | \theta_1, \theta_2, \theta_3, \dots, \theta_l)].$$

A second log-likelihood value based on a nested model created by removing m parameters (set equal to zero) is represented by

$$\log(L_0) = \log - \text{likelihood} = \log[L(x_1, x_2, x_3, \dots, x_k | \theta_1 = 0, \dots, \theta_m = 0, \theta_{m+1}, \dots, \theta_l)]$$

or

$$\log(L_0) = \log - \text{likelihood} = \log[L(x_1, x_2, x_3, \dots, x_k | \theta_{m+1}, \dots, \theta_l)].$$

As long as these two log-likelihood values are calculated from the same data, compare nested models and differ only because of random variation, the likelihood ratio statistic $X^2 = -2[\log(L_0) - \log(L_1)]$ has a chi-square distribution with m degrees of freedom. The degrees of freedom are m because m parameters are deleted from the more complex model (most parameters) to form a simpler and, as required, a nested model.

The comparison of log-likelihood values reflects the relative "goodness-of-fit" between two sets of conditions (two nested models). The observed difference indicates the effectiveness of the simpler model (based on fewer parameters) relative to the more complex model. When a parameter value or a set of parameter values is eliminated from a model and the log-likelihood value remains essentially unaffected (only a slight increase), the inference is made that the values eliminated are unimportant and likely have only random influences. The difference is said to be consistent with random

variation. Conversely, when a parameter value or a set of parameter values is eliminated from a model and the log-likelihood value strikingly increases, the inference is made that the values eliminated are important and likely have systematic influences. The comparison of log-likelihood values, therefore, produces a chi-square test statistic X^2 and a significance probability that allows a statistical evaluation of the difference induced between two log-likelihood values created by eliminating model parameters (random or systematic?). The strategy of comparing log-likelihood values is fundamental to evaluating the influence of model parameters associate with a specific variable or a specific set of variables in a multivariable analysis.

A log-likelihood value by itself is not a useful assessment of goodness-of-fit (a comparison between model and data). The magnitude of a single log-likelihood value is primarily determined by the sample size; the larger the sample size, the larger the log-likelihood statistic. Because a difference between two likelihood values from nested models is not influenced by the sample size (same data for both calculations), it exclusively reflects differences between compared models.

A sometimes handy "rule of thumb" states: when a chi-square test statistic X^2 that results from comparing two log-likelihood values is less than m (the number of parameters eliminated), no evidence exists that these parameters play a systematic role in the model. The rule is simply an application of the fact that the mean value of a chi-square distribution is its degrees of freedom. The likelihood ratio test statistic has a chi-square distribution with m degrees of freedom when the parameters eliminated have only random influences. Therefore, for an observed chi-square statistic less than its mean value m

($X^2 < m$), the p -value will always be greater than 0.3 and usually in the neighborhood of 0.4 and never greater than 0.5. Of course, exact probabilities exist in tables and are part of computer analysis programs.

The statistical properties of the function $f(X)$

An observed variable denoted x frequently has known or postulated properties but questions arise concerning a function of x , denoted $f(x)$. Two important questions are: What is the mean and what is the variance of the distribution of $f(x)$? Or, in symbols,

$$\text{mean of the distribution of } f(x) = ? \quad \text{and} \quad \text{variance}[f(x)] = ?$$

Two rules allow the estimation of the mean and the variance of the distribution of the variable $f(x)$ derived from the mean and variance of the distribution of the variable x .

Rule 1: the mean of the distribution of the variable $f(x)$ (denoted μ_f) is approximately the value of the function evaluated at the mean of the distribution of x (denoted μ). In symbols,

$$\mu_f = \text{mean of the distribution of } f(x) \approx f(\mu)$$

where μ_f represents the mean of the distribution of the $f(x)$ -values and μ represents the mean of the distribution of the x -values.

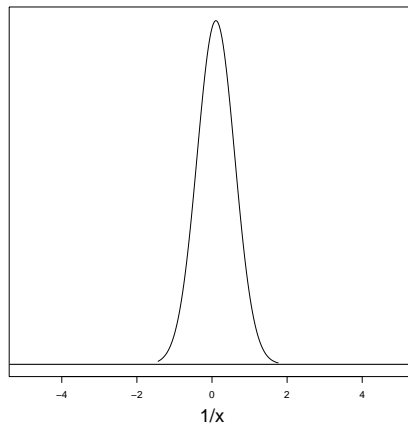
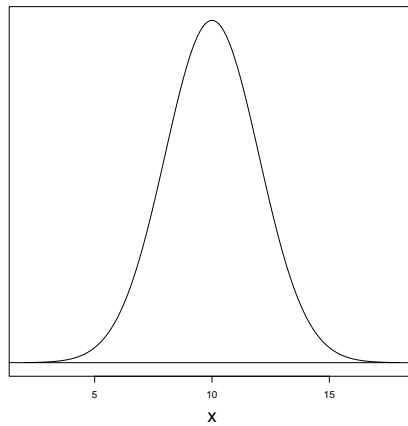
Rule 2: the variance of the distribution of the variable $f(x)$ (denoted $\text{variance}[f(x)]$) is approximately

$$\text{variance of } f(x) = \text{variance}[f(x)] \approx \left[\frac{d}{dx} f(\mu) \right]^2 \text{variance}(x).$$

The symbol $\frac{d}{dx} f(\mu)$ represents the derivative of $f(x)$ with respect to x evaluated at the

mean value μ . Both rules 1 and 2 are an application of a basic mathematical tool called a *Taylor series expansion* and their application to statistics is sometimes referred to as "the delta method."

Figure 2.0 The distributions of x and $f(x) = \frac{1}{x}$



For example, suppose the variable x has a symmetric distribution (Figure 2.0 -- top) with a mean value $= \mu = 10$ and a $variance(x) = 4$. The distribution of $f(x) = 1/x$ (Figure 2.0 -- bottom) then has an approximate mean value $\mu_f \approx 1/\mu = 1/10 = 0.1$ (rule 1)

and an approximate $\text{variance}[f(x)] \approx \text{variance}(x)/\mu^4 = 4/10^4 = 0.0.0004$ because

$$\frac{d}{dx} f(x) = \frac{d}{dx} \left[\frac{1}{x} \right] = -\frac{1}{x^2} \quad \text{and} \quad \left[\frac{d}{dx} f(\mu) \right]^2 = \frac{1}{\mu^4} = \frac{1}{10^4} \quad (\text{rule 2}).$$

Application 1.

A Poisson distributed variable is sometimes transformed by taking the square root to produce a more normal-like distribution. For a Poisson distribution, the mean value (represented by $\mu = \lambda$) and the variance (represented by $\text{variance}(x) = \lambda$) are equal. The function $f(x) = \sqrt{x}$ produces a more symmetric and approximate normal distribution (if λ is not too small) with mean value $= \sqrt{\lambda}$ and variance $= 1/4$. Specifically, applying rules 1 and 2, the approximate mean of the normal-like distribution is

$$\mu_f = \text{mean of the distribution of } \sqrt{x} = \mu_{\sqrt{x}} \approx \sqrt{\mu} = \sqrt{\lambda}.$$

with approximate variance

$$\text{variance of } f(x) = \text{variance}(\sqrt{x}) = \frac{1}{4\lambda} \text{variance}(x) = \frac{1}{4\lambda} \lambda = \frac{1}{4}$$

because again the derivative of \sqrt{x} is

$$\frac{d}{dx} f(x) = \frac{d}{dx} \sqrt{x} = -\frac{1}{2\sqrt{x}} \quad \text{and} \quad \left[\frac{d}{dx} f(\lambda) \right]^2 \approx \frac{1}{4\lambda} \quad (\text{rule 2}).$$

Application 2.

Applying rules 1 and 2 to the logarithm of a variable x again yields an expression for the approximate mean value and variance of the distribution of the transformed

variable $\log(x)$. Thus, when $f(x) = \log(x)$, then

$$\mu_f = \text{mean of the logarithm of } f(x) = \mu_{\log(x)} \approx \log(\mu)$$

where again μ represents the mean of the distribution of the variable x . The derivative of $\log(x)$ is

$$\frac{d}{dx} f(x) = \frac{d}{dx} \log(x) = \frac{1}{x}$$

making

$$\text{variance of } f(x) = \text{variance}[f(x)] = \text{variance}[\log(x)] \approx \frac{1}{\mu^2} \text{variance}(x).$$

Corollary:

$$\text{variance}(x) \approx \mu^2 \text{variance}[\log(x)].$$

Quantities are frequently transformed to have more symmetric (normal-like) distributions by using logarithms; thus, creating more accurate test statistics and confidence intervals. The mean and variance of the distribution of the logarithm of a variable then become necessary parts of the statistical evaluation. Note: all logarithms used are natural logarithms (base $e = 2.718281828 \dots$), sometimes called Napier logarithms in honor of John Napier (*b.* 1550) who pioneered the use of logarithms.

Application 3. A somewhat complicated application -- Greenwood's variance

A not exactly straight-forward application of the expression for the variance of the logarithm of a variable produces an estimate for the variance of an estimated survival

probability (Greenwood's variance formula).

Step 1. When \hat{p}_i represents an estimated probability from a binomial distribution, then

$$\text{variance}[\log(\hat{p}_i)] \approx \left[\frac{1}{\hat{p}_i} \right]^2 \text{variance}(\hat{p}_i) = \left[\frac{1}{\hat{p}_i} \right]^2 \frac{\hat{p}_i(1 - \hat{p}_i)}{n_i} = \frac{\hat{q}_i}{n_i \hat{p}_i} \quad (\text{rule 2})$$

where $\hat{q}_i = 1 - \hat{p}_i$ and the variance of the distribution of \hat{p}_i is estimated by $\hat{p}_i(1 - \hat{p}_i)/n_i$.

Step 2. For the estimated variance of a product-limit estimated survival probability (\hat{P}_k),

$$\text{variance}(\hat{P}_k) \approx \hat{P}_k^2 \text{variance}[\log(\hat{P}_k)]. \quad (\text{rule 2, corollary})$$

Step 3. The logarithm of the product-limit estimate of \hat{P}_k is a sum of the logarithms of k specific conditional survival probabilities $\hat{p}_1, \hat{p}_2, \hat{p}_3, \dots, \hat{p}_k$ and is

$$\log(\hat{P}_k) = \log(\prod \hat{p}_i) = \sum \log(\hat{p}_i) \quad i = 1, 2, \dots, k.$$

Step 4. Putting steps 1, 2 and 3 together gives

$$\begin{aligned} \text{variance}(\hat{P}_k) &\approx \hat{P}_k^2 \text{variance}[\log(\hat{P}_k)] = \hat{P}_k^2 \text{variance}[\sum \log(\hat{p}_i)] \\ &= \hat{P}_k^2 \sum \text{variance}[\log(\hat{p}_i)] \approx \hat{P}_k^2 \sum \frac{\hat{q}_i}{n_i \hat{p}_i} \quad i = 1, 2, \dots, k. \end{aligned}$$