

Advanced Theory

Survival Analysis 2005

Problems for February 8, 2004

2) Look at page 6 of the Kaplan Meier notes.

Prove:

Complete data (n=d)

$$\hat{\mu} = \frac{1}{n} \sum_i^n t_i = \sum_i^n P_{i-1}(t_i - t_{i-1})$$

Solution:

$$\begin{aligned} & \sum_i^n P_{i-1}(t_i - t_{i-1}) \\ &= \sum_i^n \frac{n-i+1}{n}(t_i - t_{i-1}) \\ &= \frac{n}{n}(t_1) + \frac{n-1}{n}(t_2 - t_1) + \frac{n-2}{n}(t_3 - t_2) \dots \\ &= \frac{1}{n}t_1 + \frac{1}{n}t_2 + \frac{1}{n}t_3 \dots \\ &= \bar{t} \end{aligned}$$

3) Look at page 15 of the Kaplan Meier notes.

Prove:

Complete data (n=d)

$$\text{variance}(\widehat{P}_k) = \widehat{P}_k^2 \sum_i^k \frac{q_i}{n_i p_i} = \widehat{P}_k(1 - \widehat{P}_k)/n$$

Assume that $p_1, p_2 \dots p_k$ are defined over fixed quantile ranges that do not change as n increases. This is a not quite the assumption that is made, but assuming it

as an approximation makes the proof a bit easier. By the δ method and central limit theorem,

$$\begin{aligned} \sqrt{n}(\widehat{P}_k - P_k) &\xrightarrow{d} N(0, \Sigma_k) \\ \Sigma_k &= \begin{bmatrix} \frac{\partial}{\partial p_1} P_k & \dots & \frac{\partial}{\partial p_k} P_k \end{bmatrix} \begin{bmatrix} \frac{p_1 q_1}{\theta_1} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{p_k q_k}{\theta_k} \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial p_1} P_k \\ \vdots \\ \frac{\partial}{\partial p_k} P_k \end{bmatrix} \\ &= \begin{bmatrix} \frac{P_k}{p_1} & \dots & \frac{P_k}{p_k} \end{bmatrix} \begin{bmatrix} \frac{p_1 q_1}{\theta_1} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{p_k q_k}{\theta_k} \end{bmatrix} \begin{bmatrix} \frac{P_k}{p_1} \\ \vdots \\ \frac{P_k}{p_k} \end{bmatrix} \\ &= \widehat{P}_k^2 \sum_i^k \frac{q_i}{\theta_i p_i} \end{aligned}$$

where $\theta_i = \lim_{n \rightarrow \infty} \frac{n_i}{n}$. We can hence use the asymptotic variance to approximate the finite sample variance. The above proof is a bit lacking because in reality the quantile range over which p_i is defined changes as n increases under the complete data assumption. Hence, for a more complete proof under the complete data assumption, we would need to appeal to a more general central limit theorem, such as the central limit theorem for triangular arrays, to more rigorously derive the asymptotic distribution.

Also note that in the complete data case we have $n_i = n - i + 1$ and $p_i = \frac{n_i - 1}{n_i}$, hence,

$$\begin{aligned} &\widehat{P}_k^2 \sum_i^k \frac{\widehat{q}_i}{n_i \widehat{p}_i} \\ &= \widehat{P}_k^2 \sum_i^k \frac{1}{n - i} - \frac{1}{n - i + 1} \end{aligned}$$

which, by subtraction of like terms,

$$\begin{aligned} &= \left(\frac{\widehat{P}_k^2}{n - k} - \frac{\widehat{P}_k^2}{n} \right) \\ &= \left(\frac{\widehat{P}_k^2 \frac{n-1}{n} \frac{n-2}{n-1} \dots \frac{n-k}{n-k+1}}{n - k} - \frac{\widehat{P}_k^2}{n} \right) \end{aligned}$$

$$\begin{aligned}
&= \left(\frac{\widehat{P}_k}{n} - \frac{\widehat{P}_k^2}{n} \right) \\
&= \frac{\widehat{P}_k(1 - \widehat{P}_k)}{n}
\end{aligned}$$

4) Look at page 19 of the Kaplan Meier notes.

Prove:

Complete data (n=d)

$$\text{variance}(\widehat{\mu}) = \frac{1}{n} \sum_i^N (t_i - \bar{t})^2$$

Notice that with the complete data, $\widehat{P}_i = \frac{n-i}{n}$. Hence, we can cancel terms to see that,

$$\begin{aligned}
\widehat{\mu} &= \sum_i^n \frac{n-i}{n} (t_i - t_{i-1}) \\
&= \sum_i^n \frac{1}{n} t_i = \widehat{t}
\end{aligned}$$

The MLE estimator of the variance of the sample mean is

$$= \sum_i^n \frac{1}{n} (t_i - \widehat{t})^2$$

which is a biased estimator.

5) Look at page 20 of the Kaplan Meier notes.

Prove:

Complete data (n=d)

$$\text{variance}(\widehat{t}_m) \approx \left[\frac{t_l - t_u}{\widehat{P}_l - \widehat{P}_u} \right]^2 \widehat{V}_m$$

$$V_m = \widehat{P}_m \sum \frac{q_i}{n_i p_i} \quad (\text{Greenwood's Formula})$$

The proof of this is quite complicated, and covered in both the Probability Theory classes and the Advanced Statistical Theory class. Basically, we know from the proofs in those classes that,

$$\sqrt{n}(\widehat{Q} - Q) \xrightarrow{d} N(0, P(1 - P)f(Q)^{-2})$$

Where \widehat{Q} is the estimator of the population quantile, $Q = S^{-1}(P)$. Notice that \widehat{V}_m is simply $P_m(1 - P_m)/n$ and $\left[\frac{t_l - t_u}{\widehat{P}_l - \widehat{P}_u}\right]^2$ is an empirical estimate of $f(S^{-1}(P_m))^{-2}$.

Explore the accuracy (computationally) when

$$P(t) = \exp(-\lambda t)$$

Below is a little program that compares the estimator of the variance with the variance of the estimates. You can adjust λ , the sample size, and number of iterations in the program to see how various assumptions change the results. Running the program a few times reveals that the estimates for various variances seem to be a bit too large. Of course, I could have made a programming mistake.

```
lambda=.2
iterations=1000
#Make sure the sample size is odd (for program purposes only).
samplesize=31

#True median
qexp(.5, rate=lambda)

variances=medians=rep(NA,iterations)

for(i in 1:iterations){
randexp < - sort(rexp(samplesize,rate=lambda))
```

```

variances[i] < - .5**2/samplesize*((randexp[(samplesize-1)/2+2]-
  randexp[(samplesize-1)/2])/
  ((1-((samplesize-1)/2+2)/samplesize)-(1-((samplesize-1)/2)/samplesize))**2
medians[i] < - randexp[(samplesize-1)/2+1] }

```

#Comparison of mean of estimated variances with variance of estimates.

```
mean(variances)
```

```
var(medians)
```