C. Estimating location measures:

KM: 4.5a

1. Define $p$ quantile $\nu$ to satisfy
   
   a. $F(\nu) \geq p$, $F(\nu-) \leq p$

   b. $S(\nu) \leq 1 - p$, $S(\nu-) \geq 1 - p$

   c. $\nu = S^{-1}(1 - p)$ as long as distribution is continuous; Assume this.

2. Estimating quantiles
   
   a. Behaves better than mean because it does not depend on extremes: robust

   b. Estimate as $U$ such that $\hat{S}(U) \leq 1 - p$, $\hat{S}(U-) \geq 1 - p$

      i. Uniquely defined unless curve has a flat spot with value $p$

      ii. Value is place where you jump through $p$

      iii. Otherwise can be any value on that flat spot.

      iv. We’ll estimate as midpoint

   c. Quantile Confidence Interval is inversion of Survival Interval

      i. Draw (point-wise) CI for $S$

         • May be on log or arcsine scale

      ii. Draw horizontal line at $p$
Lecture 3

iii. CI is parts of horizontal line inside confidence band

d. May approximate interval using normal distribution

i. For true quantile \( \nu \),

\[
P \left[ \frac{\hat{S}(\nu) - S(\nu)}{\sqrt{\text{Var}[\hat{S}(\nu)]}} \geq 1.96 \right] \approx 0.05
\]

- We say this statistic is *pivotal*.

e. Intervals for \( S(t) \) give vertical lines

f. Intervals for \( \nu \) give horizontal lines

g. Shouldn’t we be doing this with simultaneous intervals?

i. No, because we only use half an interval at two places

h. No guarantee that interval is connected

i. Since upper confidence bound might go back up

i. Alternative approach via standard error

i. \( \hat{F}(\hat{\nu}) - \hat{F}(\nu) \approx f(\nu)(\hat{\nu} - \nu) \)

ii. \( p - \hat{F}(\nu) \approx f(\nu)(\hat{\nu} - \nu) \)

iii. \( \text{Var} \left[ \hat{F}(\nu) \right] \approx f(\nu)^2 \text{Var} [\hat{\nu}] \)

iv. \( \text{Var} [\hat{\nu}] \approx \hat{f}(\hat{\nu})^{-2} \text{Var} \left[ \hat{F}(\hat{\nu}) \right] \)

- Need estimator of density.

KM: 4.5b

R Code

SAS Code

3. Mean
Lecture 3

a. Defined as \( E[X] = \int_0^\infty t f(t) \, dt \)

b. Integration by parts
   i. \( \int_a^b u \, dv = uv \bigg|_a^b - \int_a^b v \, du \)
   ii. In this case, let \( v = -S(t), u = t \).

iii. gives
    \[
    E[X] = (-S(t))t \bigg|_0^\infty - \int_0^\infty (-S(t)) \, dt = \int_0^\infty S(t) \, dt
    \]

iv. Argument can be extended to discrete distributions

c. Estimate as \( \int_0^\infty \hat{S}(t) \, dt \)

d. \( \infty \) if \( \hat{S} > 0 \) at last event
   i. Finite if \( \hat{S} \) hit zero at last event
      • Could fix using parametric estimate of rest of curve
   ii. Can estimate standard error
      • Estimators typically wrong if last event not a events.

e. Common fix: Estimate restricted mean life \( \int_0^K S(t) \, dt \) for some \( K \) by \( \int_0^K \hat{S}(t) \, dt \)

KM: 5.4

D. Other Sampling Schemes

1. For data grouped in intervals:
   a. Summarize data by number \( m_j \) censored and \( D_j \) with event in
Lecture 3

intervals \((a_{j-1}, a_j]\)

b. Number at risk should be measured somewhere in \((a_{j-1}, a_j]\)

c. Raise \(Y_j\) to \(Y'_j = Y_j + m_j/2 + D_j\)
   i. Approx. equivalent to estimating \(h(t)\) on as
      \[D_j/(Y_j + (m_j + D_j)/2)\].
   ii. \(Y_j\) still number at risk at time \(a_j\)

d. Called the actuarial estimate or life table estimate.

KM: 4.6

2. With truncation:
   a. Estimates are conditional on inclusion
   b. If truncation \(\perp\) event time, hazard rate still the same
   c. Hence can still get \(h\) from slope of Nelson–Aalen estimator \(\hat{H}\).

KM: 7.1

V. Testing with survival curves

A. Introduction to survival curve hypothesis testing

1. Notation
   a. Suppose that \(K\) populations have survival functions \(S_k\),
Suppose we have samples of size $n_k$ respectively.

Generally assume that they are independent.

2. Two questions
   a. Do (population) survival curves have some relation?
      i. If $K = 1$, does it take some simple form like exponential?
      ii. If $K > 1$, are they all the same?
      iii. Generally $K = 2$ simpler than $K > 2$
   b. Do (population) survival curves have relation with some attribute?
      i. Attributes might be
         • Mean
         • Median
         • Value at some time
      ii. Relation might be
         • Equals a null value if $K = 1$
         • Equal to each other if $K > 1$

3. One-sample Hypothesis Test
a. Comparing observed number of events to expected number.

b. Expected value for $o_i = D_i/Y_i$ is $e_i = 1 - S(t_{i+1})/S(t_i) = 1 - \exp(-\int_{t_i}^{t_{i+1}} h(s) \, ds) \approx \int_{t_i}^{t_{i+1}} h(s) \, ds$.

c. Test statistic is then $T = \sum_{i=1}^{D} W(t_i) (o_i - e_i) = O - E$ for some weight function $W$ depending on time.

d. $O = \sum_{i=1}^{D} W(t_i) o_i$, $E = \sum_{i=1}^{D} W(t_i) e_i$.

e. Often use $W(t_i) = Y_i$ to give test statistic $T = (\sum_{i=1}^{D} D_i - Y_i e_i)$

f. Estimate of variance like chi-square example:

   i. Variance is approximately what gets subtracted off (call it $E$) from first part (call it $O$)

   ii. $(0 - E)^2/E$ approximately $\chi_1^2$

   iii. Variance for $K > 1$ also $\approx E$ but we don’t need this

h. Often $h$ is determined empirically from a very large sample and is effectively non-random

L: 8.2

4. The $K > 1$ testing task

   a. Pick one group labeled $k$ from $\{1, \ldots, K\}$.

   b. Ask whether $S_k$ is different from the rest.
Lecture 3

i. Maybe $K = 2$ and $k = 1$ or $2$

ii. Intuition doesn’t require this

iii. More generality at this stage will be useful.

c. under the Proportional Hazards Alternative

5. Null and alternative hypotheses for $K$ samples, $K > 1$. 

a. Postulate a common $S$

b. Postulate how distns will differ under $H_A$

i. Shift for standard $t$ tests, etc.

ii. Assume $-\frac{d}{dt} \log(S_k(t)) = \alpha_k \times (-\frac{d}{dt} \log(S(t)))$:

iii. Called proportional hazards in survival analysis lingo, and Lehmann alternative in nonparametrics literature

- Implies $\log(S_k(t)) = \alpha_k \log(S(t))$

iv. Implies $S_k(t) = S(t)^{\alpha_k}$

6. Useful Consequences of the Lehmann Alternative

a. If null and alternative distributions for $T$ are related by the Lehmann relation, then so are those for $U = g(T)$ any increasing invertible transformation of $T$

i. Because $P_A[U > u] = P_A[T > g^{-1}(u)] = P_0[T > g^{-1}(u)]^{\alpha_k} = P_0[U > u]^{\alpha_k}$
b. Lehmann alternative for exponential is also a shift alternative on the log scale:
   i. Suppose $U = \log(T)$,
   ii. Alternative distribution of $U$ has survival curve
       \[ P_A[U \geq u] = P[T \geq \exp(u)] = S_k(\exp(u)) = \exp(-\lambda \exp(u))^{\alpha_k} = \exp(-\lambda \exp(u + \log(\alpha_k))). \]

7. No-censoring general rank test
   a. Suppose that $T_j$ are times associated with groups $g_j$,
      \[ g_j \in \{1, \ldots, K\}, \ j \in \{1, \ldots, K\}. \]
   b. Make $n$ scores $a_j$
      i. Nondecreasing in $j$
      ii. Sum to zero.
   c. Let $R_j$ be the rank of $T_j$ from the entire sample.
   d. General rank statistic for testing group $k$ different from rest is
      \[ W_k = \sum_{j: g(j) = k} a_j R_j. \]

8. Optimal Test for Equality of Distributions
   a. Best means giving best power for very large $n$ and $\alpha$ near 1
      i. Under a shift alternative, then best scores are expected values of order statistics evaluated at $g'(\cdot)/g(\cdot)$ for $g$ the density.
ii. This works out to expected order statistic of exponential if null distribution is that of $\log(T)$ for $T$ exponential

b. Best scores are $a_j = -1 + \log(1 - j/(n + 1))$

c. $a_j \approx -1 + \sum_{i=n+1}^{n} (1/i)$

d. Approximately Log rank statistic $\sum_{i=1}^{n} [D_{ki} - (D_i/Y_i)Y_{ki}]$

ii. Interpretation: Expected # of events in stratum $l$ if they were distributed $\propto$ # at risk $-$ # who actually had event.

ii. $D_{ki} | D_i, Y_{ki}, Y_i$ is hypergeometric:

<table>
<thead>
<tr>
<th>$D_{ki}$</th>
<th>$D_i - D_{ki}$</th>
<th>$D_i$</th>
<th>$Y_{ki}$</th>
<th>$Y_i - Y_{ki}$</th>
<th>$Y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{ki}$</td>
<td>$Y_{ki} - D_{ki}$</td>
<td>$Y_i - D_i$</td>
<td>$(Y_{ki} - D_{ki})$</td>
<td>$Y_i - Y_{ki}$</td>
<td>$Y_i$</td>
</tr>
</tbody>
</table>

9. Phrase in terms of survival quantities.

a. Let $Y_j$ be # at risk when item $j$ of joint sample has event.

b. $Y_{kj}$ be # at risk from group $l$

c. $D_{ki} = \#$ from strata $l$ dying at time $i$

d. Then $R_j = n + 1 - Y_j$

e. Use as test statistics $W_k = \sum_{j:g_j=k} \alpha R_j$

f. Big when the observations for sample $l$ tend to be bigger than those in other groups.

g. In multi-group case, you can calculate statistic for either group.
Lecture 3

1. \[ \sum_k W_k = \sum_k \sum_i (D_{ik} - D_i Y_{ik}/Y_i) = \sum_i \sum_{k=1}^2 (D_{ik} - D_i Y_{ik}/Y_i) = \sum_i (D_i - D_i Y_i/Y_i) \equiv 0, \]

ii. so when \( K = 2 \) it doesn’t matter which group you use

iii. Constraints also holds for larger \( K \).

10. Null Distribution:

a. Mean zero

b. Shape is approximately normal (Háyek 1960).

\[ KM: 7.3b \]

c. At each time, under \( H_0 \), contribution to test \( D_{ki} \sim \]

Hypergeometric.

i. Marginals are \( D_i, Y_i - D_i \) and \( Y_{ki}, Y_i \).

ii. Table:

| \( D_{ki} \) | \( D_i - D_{ki} \) | \( D_i \) |
| \( Y_{ki} - D_{ki} \) | \( Y_i - D_i \) | \( Y_i - D_i \) |
| \( Y_{ki} \) | \( Y_i - Y_{ki} \) | \( Y_i \) |

iii. Variance contribution is Hypergeometric variance

\[ \frac{Y_{ki}(Y_i - Y_{ki})Y_i - D_i}{Y_i^2 Y_i - 1} D_i. \]

d. \( \text{Var}_{H_0} [W_k] \approx \sum_{i=1}^D \frac{Y_{ki}(Y_i - Y_{ki})Y_i - D_i}{Y_i^2 Y_i - 1} D_i \)

i. Test statistic is equivalent to Mantel–Haenzel test.

ii. Variances add
Lecture 4

- as with Greenwood’s formula

iii. Variance formula is exact in no–censoring case, but depends on censoring mechanism with censoring

[R Code] [SAS Code]