14] Let $w_{ij} = K(\underline{x}_i, \underline{x}_j) = w_{ji}$, $i \neq j$.

Let $H_n = \sum\sum_{i<j} w_{ij}$. Then

$U_n = \frac{1}{\binom{n}{2}} H_n$ is a U-statistic.

$\frac{n(n-1)}{2}$ sample mean of the w_{ij}

Note that $T_n = \frac{2 H_n}{n(n-1)} = U_n$ with $w_{ij} = \underline{x}_i^T \underline{x}_j$.

Aug 2023!

15] Theorem: Let $\underline{x}_1, \ldots, \underline{x}_n$ be iid and $w_{ij} = \underline{x}_i^T \underline{x}_j$ for $i \neq j$. Let $E(\underline{x}_i) = \underline{\mu}$ and $V(w_{ij}) = \sigma_w^2$. Let $\theta = \text{cov}(w_{ij}, w_{id})$ where $j \neq j$, $i < j$ and $i < d$. Then

usually this term dominates

a) $V(T_n) = \frac{2\sigma_w^2}{n(n-1)} + \frac{4(n-2)}{n(n-1)}\theta$

with $\theta \geq 0$.

b) If Ho $\underline{\mu} = \underline{0}$ is true, then

$V(T_n) = \frac{2\sigma_w^2}{n(n-1)}$.

since $\theta = 0$

16] $S_m^2 = \frac{1}{m-1} \sum_{i=1}^{m} (W_i - \bar{W})^2 \xrightarrow{P} \sigma_W^2$

where the W_i are as in 13].

$$\hat{\sigma}_W^2 = \frac{1}{n(n-1)} \sum\sum_{i \neq j} (W_{ij} - T_n)^2$$

is nearly the sample variance of the W_{ij}, which are not ind.

from U statistics theory

17] Partial proof of 15].

$V(H_n) = \text{cov}(H_n, H_n) =$

$$\text{cov}\left(\sum\sum_{i<j} W_{ij}, \sum\sum_{k<d} W_{kd} \right) =$$

$$\sum\sum_{i<j} \sum\sum_{k<d} \text{cov}(W_{ij}, W_{kd}). \qquad (*)$$

The covariances are of 3 types:

I) $(ij) = (kd)$ with $i<j$:

$\text{cov}(W_{ij}, W_{kd}) = V(W_{ij}) = \sigma_W^2$

II) i, j, k, d are distinct with $i<j$, $k<d$.

Then $W_{ij} \perp\!\!\!\perp W_{kd}$ and $\text{cov}(W_{ij}, W_{kd}) = 0$.

III) Exactly three of the four subscripts are distinct which have $cov(w_{ij}, w_{id}) = \theta$ for $j \neq d$, $i < j$, and $i < d$ or $i \neq k$, $i < j$, $k < d$ which have $cov(w_{ij}, w_{id}) = \theta$ since $w_{ij} = w_{ji}$.

The number of ways to get 3 distinct subscripts is $a - b - c$

done in U statistics theory

choose ik choose kd

$$= \binom{n}{2}^2 - \binom{n}{2}\binom{n-2}{2} - \binom{n}{2} = n(n-1)(n-2)$$

a = # terms in $(*)$

b = # terms where i, j, k, d are distinct with $i < j$ and $k < d$

c = # terms with $(ij) = (kd)$ and $i < j$.

Thus $V(H_n) = 0.5\, n(n-1)\sigma_w^2 + n(n-1)(n-2)\theta$

and $V(T_n) = \frac{4}{[n(n-1)]^2} V(H_n) = \frac{2\sigma_w^2}{n(n-1)} + \frac{4(n-2)\theta}{n(n-1)}$;

$\theta \geq 0$ also comes from U-statistics theory.

18] Let $s_0 = \sqrt{\frac{2\hat{\sigma}_w^2}{n(n-1)}}$ or $\sqrt{\frac{2 s_m^2}{n(n-1)}}$ from 16].

These formulas are simpler than the ones in the published literature.

When Ho is not true, suppose $s_n^2 \xrightarrow{P} V(T_n)$.

$\frac{T_n}{s_0} \approx \frac{T_n}{s_n}$ under Ho,

but $\frac{|T_n|}{s_0} > \frac{|T_n|}{s_n}$ since $\theta > 0$ when Ho is false. Thus $\frac{|T_n|}{s_0}$ has more power when Ho is false, where power = P(reject Ho). (want power near level (eg. .05) when Ho is true, power near 1 if Ho is false)

19] HD tests for Ho $\underline{\mu}_1 = \underline{\mu}_2$ vs H, $\underline{\mu}_1 \neq \underline{\mu}_2$.

want to estimate $\|\underline{\mu}_1 - \underline{\mu}_2\|^2 =$

$\cdot (\underline{\mu}_1 - \underline{\mu}_2)^T (\underline{\mu}_1 - \underline{\mu}_2) = \underline{\mu}_1^T \underline{\mu}_1 + \underline{\mu}_2^T \underline{\mu}_2 - 2\underline{\mu}_1^T \underline{\mu}_2$.

Suppose we have 2 independent (iid) random samples

$\underline{X}_{1,1}, \ldots, \underline{X}_{1,n_1}$

$\underline{X}_{2,1}, \ldots, \underline{X}_{2,n_2}$.

Let $\underline{a} = \sum_{i=1}^{n_1} \underline{X}_{1,i}$, $\mathbb{X}_1 = (\underline{X}_{1ij}) = \begin{pmatrix} \underline{X}_{1,1}^T \\ \vdots \\ \underline{X}_{1,n_1}^T \end{pmatrix}$

$\underline{c} = (\underline{X}_{1,1}^T, \ldots, \underline{X}_{1,n_1}^T)^T$,

$\underline{b} = \sum_{i=1}^{n_2} \underline{X}_{2,i}$, $\mathbb{X}_2 = (\underline{X}_{2ij}) = \begin{pmatrix} \underline{X}_{2,1}^T \\ \vdots \\ \underline{X}_{2,n_2}^T \end{pmatrix}$

$\underline{d} = (\underline{X}_{2,1}^T, \ldots, \underline{X}_{2,n_2}^T)^T$.

Let $T_n = \frac{1}{n_1(n_1-1)} [\underline{a}^T\underline{a} - \underline{c}^T\underline{c}] + \frac{1}{n_2(n_2-1)} [\underline{b}^T\underline{b} - \underline{d}^T\underline{d}]$

$- \frac{2\,\underline{a}^T\underline{b}}{n_1 n_2}$. Under $H_0: \underline{\mu}_1 = \underline{\mu}_2$

and $s_n^2 \xrightarrow{P} V(T_n)$, $\frac{T_n}{s_n} \xrightarrow{D} N(0,1)$.

PhD topic:

{13-19} try to get better S_n for 19}.

20] "m out of n" bootstrap with $m_i = \left[\frac{2n_i}{3}\right]$ simulated well.

21] For the 1-sample problem,

$$T_m^* = \frac{1}{m(m-1)} \sum_{k \neq d}^{m} \sum^{m} \underline{x}_{i_k}^T \underline{x}_{i_d} \quad \text{where}$$

$\underline{x}_{i_1}, \ldots, \underline{x}_{i_m}$ are the cases selected without replacement.

Hence $T_m^* = \frac{1}{m(m-1)} \sum_{i \neq j}^{n} \sum^{n} w_i w_j \underbrace{\underline{x}_i^T \underline{x}_j}_{\text{constant wrt bootstrap dist.}}$ ($n(n-1)$ terms in the sum)

where the indicator $w_k = \begin{cases} 1 & \underline{x}_k \text{ in the sample} \\ 0 & \text{else.} \end{cases}$

The w_k are identically distributed but dependent since $\sum_{k=1}^{n} w_k = m$.

Since the W_k are binary $\in \{0,1\}$,

$W_k^a = W_k$ for $a > 0$.

A product of indicators is an indicator $= 1$ iff all of the indicators $= 1$.

Hence $W_{i_1}^{a_1} \cdots W_{i_K}^{a_K} = W_{i_1} \cdots W_{i_K}$

$= W_{i_1, i_2 \cdots i_K}$ is the indicator for whether $x_{i_1}, \ldots, x_{i_K}$ are all in the sample of size m. Here $i_1, \ldots, i_K$ are K distinct indices.

Thus $E\left[W_{i_1}^{a_1} \cdots W_{i_K}^{a_K}\right] = E\left(W_{i_1} \cdots W_{i_K}\right)$

$$= \underset{1}{\frac{m}{n}} \quad \underset{2}{\frac{m-1}{n-1}} \quad \cdots \quad \underset{K}{\frac{m-K+1}{n-K+1}}$$

} slots

permutations

Note that $m(m-1)\cdots(m-K+1)$ ordered K tuples were selected from $n(n-1)\cdots(n-K+1)$ possible ordered K tuples (sequences of length K).

2] Thus $E[T_m^*] = \frac{1}{m(m-1)} \sum\sum_{i \neq j} E(w_i w_j)\, \underset{\sim}{x}_i^T \underset{\sim}{x}_j$

$= \frac{1}{m(m-1)} \cdot \frac{m(m-1)}{n(n-1)} \sum_{i \neq j} \underset{\sim}{x}_i^T \underset{\sim}{x}_j = T_n.$

multiple testing

1] Suppose there is a single gene g and there are n measurements of "expression of gene g"

$x_1^A, \ldots, x_n^A$ A = normal cells (control)

$x_1^B, \ldots, x_n^B$ B = cancer cells (patient)

H_0 $\mu_{Ag} = \mu_{Bg}$ H_1 $\mu_{Ag} \neq \mu_{Bg}$

2] Suppose there is a test statistic for gene g that produces (an estimated) p-value $\hat{p}_g$ (eg reject H_0 at level α if $\hat{p}_g < \alpha$).

3] with DNA microarrays and next generation sequencing techniques,

Can measure the expression levels of m genes simultaneously, where m is in the thousands:

$g \in \{1, \ldots, m\}$.

H_{0g} the mean expression levels of gene g in conditions A and B are the same

H_{1g} " " different.

4) If $m > 20000$, $\alpha = .05$, and H_{0g} is true for 20000 genes g, then on average, there are $20000(.05) = 1000$ false positives (discoveries).

When H_{0g} is rejected further experiments need to be done in order to validate or not the "discovery" (that gene g is involved in the response to change of "environment" between A and B). These experiments cost a lot of time and money. Want few false discoveries without losing too much in power.

5] Given m tests and m pvalues $\hat{p}_1, \ldots, \hat{p}_m$, let $\hat{R} = \hat{R}(\hat{p}_1, \ldots, \hat{p}_m)$ give a set of indices i, for hypotheses H_{0i}, that are rejected. $\hat{R} = \phi$ is possible.

6] Let $I_0 = \{i \in \{1, \ldots, m\} : H_{0i} \text{ is true}\}$.

we call <u>false positive</u> the indices $i \in \hat{R} \cap I_0$ and <u>true positive</u> the indices $i \in \hat{R} \cap I_0^c = \hat{R} \setminus I_0$.

Let $FP = \text{card}(\hat{R} \cap I_0) = \#$ of false positives

and $TP = \text{card}(\hat{R} \cap I_0^c) = \#$ of true positives.

Here $I_0^c = \{i \in I : i \notin I_0\}$.

7] The Bonferonni correction uses $\hat{R}_B = \{i : \hat{p}_i \leq \frac{\alpha}{m}\}$.

8] Some assumptions on the p-value $\hat{P}_i$ are needed, such as $\sup_{\theta \in \Theta_{0i}} P_\theta(\hat{P}_i \leq v) \leq v \quad \forall v \in [0,1]$,

9] If 8] holds, then for $\hat{R}_B$ ←7]

$$P(FP_i > 0) \leq \alpha.$$

Thus the existence of false positives $\leq \alpha$, but TP is often small.

10] The false discovery proportion

$$FDP = \frac{FP}{FP+TP} \quad \text{with } \frac{0}{0} = 0.$$

The false discovery rate

$$FDR = E\left[\frac{FP}{FP+TP} I(FP+TP \geq 1)\right]$$

and is widely used in biostatistics.

11] To make FP as small as possible and TP as large as possible, the $i \in \hat{R}$ correspond to the smallest pvalues. Need to determine how many of the smallest

$\hat{P}_i$ should be in $\hat{R}$.

want card($\hat{R}$) $= \hat{K}$ "large" with FDR $\leq \alpha$,

so $\hat{P}_{(1)}, \ldots, \hat{P}_{(\hat{K})} \in \hat{R}$ for some $\hat{K}$,

where $\hat{R} = \phi$ is possible ($\hat{K} = 0$).

[2] Consider the line

$$\hat{P} = Y = 0 + \frac{\alpha}{m} k \qquad \left(\text{where } 0 < \alpha < 1 \text{ is fixed, eg } \alpha = .05\right).$$

Let $\hat{R}_{BH} = \phi$ if $\left\{ k \in I : \hat{P}_{(k)} \leq \frac{\alpha k}{m} \right\} = \phi$.

otherwise, let $\hat{K} = \max\left\{ k \in I : \hat{P}_{(k)} \leq \frac{\alpha k}{m} \right\}$

and $\hat{R}_{BH} = \left\{ i \in I : \hat{P}_i \leq \alpha \frac{\hat{K}}{m} \right\}$

$= \left\{ i_{(1)}, \ldots, i_{\hat{K}} \text{ corresponding to } \hat{P}_{(1)}, \ldots, \hat{P}_{(\hat{K})} \right\}$.

under fairly strong regularity conditions,

FDR $\leq \alpha$ for $\hat{R}_{BH}$.