14.3 Tukey's T-procedure

John Wilder Tukey (1915 - 2000)
method initially published in 1949

14.3.1 Tukey's pairwise comparisons theorem

Lemma 14.1 Studentized range

Let $T_1, \dots, T_m$ be a random sample from $N(\mu, \sigma^2)$, $\sigma^2 > 0$. Let $R = \max_{j=1,\dots,m} T_j - \min_{j=1,\dots,m} T_j$ be the range of the sample. Let S^2 be the estimator of σ^2 such that S^2 and $\underline{T} = (T_1, \dots, T_m)^T$ are independent and $\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu$ for some $\nu > 0$.
Let $Q = \frac{R}{S}$. The distribution of Q then depends on neither μ, nor σ.

Proof: $$\frac{R}{S} = \frac{\frac{1}{\sigma}\max(T_j - \mu) - \frac{1}{\sigma}\min(T_j - \mu)}{\frac{S}{\sigma}} = \frac{\max\left(\frac{T_j - \mu}{\sigma}\right) - \min\left(\frac{T_j - \mu}{\sigma}\right)}{\frac{S}{\sigma}}$$

- $\forall j = 1, \dots, m \quad \frac{T_j - \mu}{\sigma} \sim N(0,1)$: distribution of numerator of Q depends neither on μ, nor on σ
- $\frac{S}{\sigma}$ = transformation of random variable having χ^2_ν → distribution depends on neither μ, nor on σ
- numerator & denominator of Q are independent
  → joint distribution of numer. & denom. of Q depends on neither μ, nor on σ

$\Rightarrow$ distribution of $\frac{\text{numer.}}{\text{denom.}}$ depends on neither μ, nor on σ

11

Distribution of $\frac{R}{S}$ still depends on m and ν.

m: sample size of T; ν: degr. of freedom of χ^2

Def 14.5 *Studentized range*

The random variable $Q = \frac{R}{S}$ from Lemma 14.1 will be called *studentized range* of size m with ν degrees of freedom and its distribution will be denoted as $q_{m,\nu}$.

14

NOTATION:

- $0 < p < 1$: $q_{m,\nu}(p) = p \cdot 100\%$ quantile of distribution $q_{m,\nu}$
- $CDF_{q,m,\nu}(\circ)$ = cdf of a random variable $Q \sim q_{m,\nu}$

Illustrations $q_{m,\nu}$

15-16

Theorem 14.2 Tukey's pairwise comparisons theorem, balanced version

Let $T_1, \dots, T_m$ be independent random variables and let $T_j \sim N(\mu_j, \nu^2\sigma^2)$, $j = 1, \dots, m$, where $\nu^2 > 0$ is a known constant. Let S^2 be the estimator of σ^2 such that S^2 and $T = (T_1, \dots, T_m)^T$ are independent and

$$\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu \quad \text{for some } \nu > 0.$$

Then

$$P\left(\forall j \neq l : |T_j - T_l - (\mu_j - \mu_l)| < q_{m,\nu}(1-\alpha)\cdot\sqrt{\nu^2 S^2}\right) = 1 - \alpha.$$

Proof:

- $\frac{T_j - \mu_j}{\nu} \overset{iid}{\sim} N(0, \sigma^2)$
- $R := \max_j \frac{T_j - \mu_j}{\nu} - \min_j \frac{T_j - \mu_j}{\nu}$

$$\Rightarrow \frac{R}{S} \sim q_{m,\nu} \quad \text{(studentized range)}$$

• For any $0<\alpha<1$

$$1-\alpha = P\left(\frac{\max_j \left(\frac{T_j-\mu_j}{\nu}\right) - \min_j \left(\frac{T_j-\mu_j}{\nu}\right)}{S} < q_{m,\nu}^{(1-\alpha)} \right)$$

$$= P\left(\frac{\max_j (T_j-\mu_j) - \min_j (T_j-\mu_j)}{\nu \cdot S} < q_{m,\nu(1-\alpha)} \right)$$

$$= P\left(\max_j (T_j-\mu_j) - \min_j (T_j-\mu_j) < \nu \cdot S\, q_{m,\nu(1-\alpha)} \right)$$

$$= P\left(\forall_{j\neq l} \; |(T_j-\mu_j)-(T_l-\mu_l)| < \nu \cdot S\, q_{m,\nu(1-\alpha)} \right)$$

$$= P\left(\forall_{j\neq l} \; |T_j - T_l - (\mu_j-\mu_l)| < \sqrt{\nu^2 S^2} \cdot q_{m,\nu(1-\alpha)} \right)$$

LM

Theorem 14.3 Tukey's ..., general version

18

$T_j \sim N(\mu_j, \nu^2\sigma^2) \rightarrow T_j \sim N(\mu_j, \nu_j^2 \cdot \sigma^2)$

$\sqrt{\nu^2 S^2} \rightarrow \sqrt{\frac{\nu_j^2+\nu_l^2}{2} \cdot S^2}$

$1-\alpha = P(\ldots) \rightarrow 1-\alpha \leqq P(\ldots)$

• Tukey (1953), Kramer (1956) formulated it.

• Hayter (1984) proved.

Proof is not an easy adaptation of the balanced version proof.

14.3.2 Tukey's honest significance differences (HSD)

→ application of Tukey's pairwise comparisons theorem to practical problems under the following assumptions:

ASSUMPTIONS

* $T = (T_1, \dots, T_m)^\top \sim N_m(\mu, \sigma^2 V)$
  - $\mu = (\mu_1, \dots, \mu_m)^\top \in \mathbb{R}^m$, $\sigma^2 > 0$ · unknown param.
  - $V = \operatorname{diag}(v_1^2, \dots, v_m^2)$: known diag. matrix

* S^2: estimate of σ^2
  - S^2 and T independent
  - $\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu$ for some $\nu > 0$

Example: G-sample problem (homoscedastic)

$Y_{ij} \overset{iid}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \dots, G$, $j = 1, \dots, n_i$ $\quad (m = G)$ (samples independ.)

* $T_{i\cdot} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \bar{Y}_{i\cdot}$

$$(T_1, \dots, T_G)^\top = (\bar{Y}_{1\cdot}, \dots, \bar{Y}_{G\cdot})^\top \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_G \end{pmatrix}, \sigma^2 V\right)$$

$$V = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_G} \end{pmatrix}$$

* $S^2 := \frac{1}{n-G} \sum_{i=1}^{G} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$

$$\star\ \frac{(n-G)\cdot S^2}{\sigma^2} \sim \chi^2_{n-G}\ ,\quad S^2 \perp\!\!\!\perp (T_1, \dots, T_G)^T$$
$$(\bar{Y}_{1\cdot}, \dots, \bar{Y}_{G\cdot})^T$$

End example

Interest in solving the <u>multiple comparison problem</u>:

$$H_{j,l}: \underbrace{\mu_j - \mu_l}_{\theta_{j,l}} = \theta^0_{j,l} \qquad j = 1, \dots, m-1,\ l = j+1, \dots, m$$

$\rightarrow$ $m^* = \binom{m}{2}$ elementary hypotheses

$$\theta^0 = (\theta^0_{1,2}, \theta^0_{1,3}, \dots, \theta^0_{m-1,m})^T \in \mathbb{R}^{m^*}$$

given

<u>most common</u> $\theta^0 = 0_{m^*}$

i.e. (global) hypothesis H_0 is

$$H_0: \mu_1 = \dots = \mu_m$$

$$H_{j,l}: \mu_j = \mu_l \qquad j = 1, \dots, m-1,\ l = j+1, \dots, m$$

Let us use Tukey's theorem:

$$1-\alpha \leqq P\left(\forall j \neq l \;\; |T_j - T_l - (\mu_j - \mu_l)| < q_{m,\nu}(1-\alpha) \cdot \sqrt{\frac{v_j^2 + v_l^2}{2} S^2}\right)$$

balanced case

$$1-\alpha \leqq P\left(\forall j \neq l \;\; \left| \frac{T_j - T_l - (\mu_j - \mu_l)}{\sqrt{\frac{v_j^2 + v_l^2}{2} S^2}} \right| < q_{m,\nu}(1-\alpha)\right)$$

Let for given $\theta_{j,l}^0$ $T_{j,l}(\theta_{j,l}^0) := \dfrac{T_j - T_l - \theta_{j,l}^0}{\sqrt{\frac{v_j^2 + v_l^2}{2} S^2}}$ $(=: T_{j,l}^0)$

Then $\forall \theta^0 = (\theta_{1,2}^0, \ldots, \theta_{m-1,m}^0)^T$, $\forall 0 < \alpha < 1$

$$1-\alpha \leqq P(\forall j \neq l \;\; |T_{j,l}^0| < q_{m,\nu}(1-\alpha), \; \theta = \theta^0) =$$

$$= P\left(\forall j \neq l \;\; \left| \frac{T_j - T_l - \theta_{j,l}^0}{\sqrt{\;}} \right| < q_{m,\nu}(1-\alpha), \; \theta = \theta^0\right) =$$

$$= P(\forall j \neq l \;\; (\theta_{j,l}^{TL}(\alpha), \theta_{j,l}^{TU}(\alpha)) \ni \theta_{j,l}^0, \; \theta = \theta^0),$$

where $\theta_{j,l}^{TL}(\alpha) = T_j - T_l - q_{m,\nu}(1-\alpha)\sqrt{\frac{v_j^2 + v_l^2}{2} S^2}$

$$\theta_{j,l}^{TU}(\alpha) = T_j - T_l + q_{m,\nu}(1-\alpha)\sqrt{\frac{v_j^2 + v_l^2}{2} S^2}$$

It seems that we have just derived a set of simultaneous confidence intervals for $\theta_{j,l} = \mu_j - \mu_l$, $j = 1, \ldots, m-1$, $l = j+1, \ldots, m$

Theorem 14.4 Tukey's honest significance differences

Short:
- simultaneous confidence intervals for $\vartheta_{j,\ell} = \mu_j - \mu_\ell$, $j = 1, \dots, m-1$, $\ell = j+1, \dots, m$

- balanced case

$$1-\alpha = P\left(\vartheta_{j,\ell} \notin \left(\vartheta_{j,\ell}^{TL}(\alpha), \vartheta_{j,\ell}^{TU}(\alpha)\right) \forall \vartheta_{j,\ell}^{0} : \vartheta = \vartheta^0\right)$$

- unbalanced case

$$1-\alpha \leq P\left(\vartheta_{j,\ell} \notin \left(\vartheta_{j,\ell}^{TL}(\alpha), \vartheta_{j,\ell}^{TU}(\alpha)\right) \forall \vartheta_{j,\ell}^{0} : \vartheta = \vartheta^0\right)$$

- P-values adjusted for multiple comparison:

$$p_{j,\ell}^{T} = 1 - CDF_{q,m,\nu}\left(|t_{j,\ell}^{0}|\right), \quad j < \ell$$

$t_{j,\ell}^{0}$ = value of $T_{j,\ell}(\vartheta_{j,\ell}^{0}) = \dfrac{T_j - T_\ell - \vartheta_{j,\ell}^{0}}{\sqrt{\frac{v_j^2 + v_\ell^2}{2} s^2}}$ attained with given data

$H_{j,\ell}$: $\vartheta_{j,\ell} = \vartheta_{j,\ell}^{0}$, $\vartheta_{j,\ell} = \mu_j - \mu_\ell$

Proof: The only issue to clarify are the p-values adjusted for multiple comparism.

We have

$$\forall j,i \quad (\theta_{j,i}^{TL}(\alpha), \theta_{j,i}^{TL}(\alpha)) \not\ni \theta_{j,i}^{0}$$

$\Leftrightarrow$ $H_{j,i}$ rejected by MCP

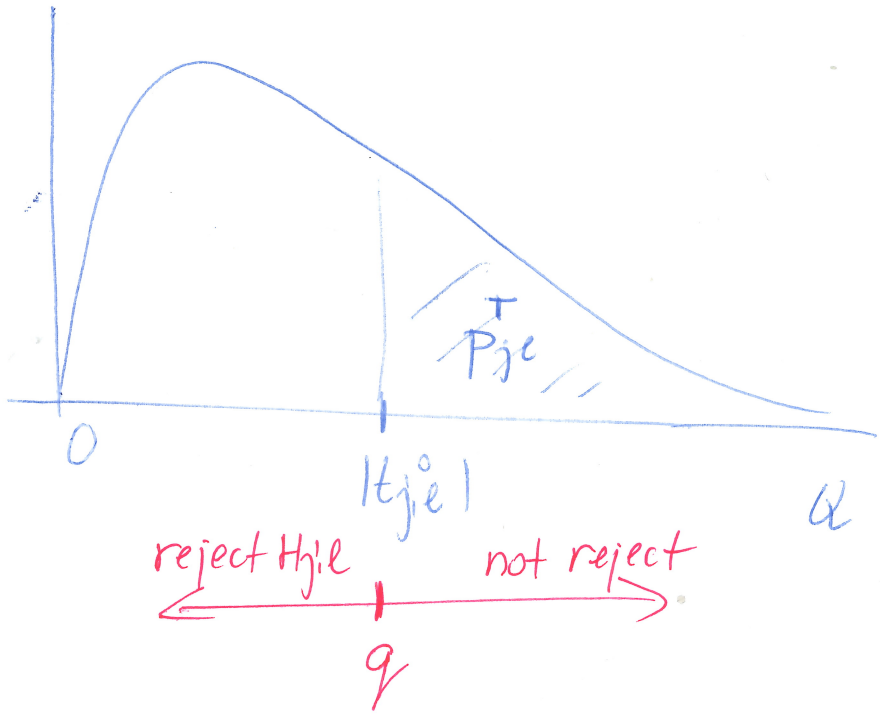
$\Leftrightarrow$ $|T_{j,i}(\theta_{j,i}^{0})| \geq q_{m,\nu}(1-\alpha)$

Hence $p_{j,i}^{T} = \inf\{\alpha : (\theta_{j,i}^{TL}(\alpha), \theta_{j,i}^{TU}(\alpha)) \not\ni \theta_{j,i}^{0}\}$

$= \inf\{\alpha : |T_{j,i}(\theta_{j,i}^{0})| \geq q_{m,\nu}(1-\alpha)\}$.

Let $t_{j,i}^{0}$ = value of $T_{j,i}(\theta_{j,i}^{0})$ attained with given data

$p_{j,i}^{T} = \inf\{\alpha : |t_{j,i}^{0}| \geq q_{m,\nu}(1-\alpha)\}$



$p_{j,i}^{T} = 1 - CDF_{q,m,\nu}(|t_{j,i}^{0}|)$.

14.3.3 Tukey's HSD in a linear model

Consider a linear model $Y|X \sim N_n(X\beta, \sigma^2 I_n)$

$\text{rank}(X_{n\times k}) = k < n$

- $L = \begin{pmatrix} \ell_1^T \\ \vdots \\ \ell_m^T \end{pmatrix}$ given matrix

- $\eta := L\beta = (\ell_1^T\beta, \dots, \ell_m^T\beta)^T$ = parameters of interest

$\eta_1, \dots, \eta_m$

- L must be such that

$(v_{jl}) = V := L(X^TX)^{-1}L^T$ is diagonal

→ properties of LSE under normality

- $T = \hat{\eta} := L\hat{\beta} = (\ell_1^T\hat{\beta}, \dots, \ell_m^T\hat{\beta})^T \sim N_m(\eta, \sigma^2 V)$ (given X)

- $\frac{(n-k)\,MSe}{\sigma^2} \sim \chi^2_{n-k}$ (both given X as well as unconditionally)

- T and MSe independent (given X)

→ all subsequent inference will be conditional given X

Tukey's pairwise comparisons theorems allow for inference on

$$\theta_{j,l} := \eta_j - \eta_l \;,\; j<l$$
$$= (l_j - l_l)^T \beta$$

- <u>simultaneous CI's</u> for $\theta_{j,l}$ are

$$\theta_{j,l}^{TL}(\alpha) = \hat{\eta}_j - \hat{\eta}_l - q_{m,n-k}(1-\alpha)\sqrt{\frac{v_j^2+v_l^2}{2} MSe}$$

$$\theta_{j,l}^{TU}(\alpha) = \hat{\eta}_j - \hat{\eta}_l + q_{m,n-k}(1-\alpha)\sqrt{\frac{v_j^2+v_l^2}{2} MSe}$$

- p-values adjusted for multiple comparison to test (elementary) hypotheses $H_{j,l}: \theta_{j,l} = \theta_{j,l}^0$, $j<l$ based on statistics

$$T_{j,l}^0 = T_{j,l}^0(\theta_{j,l}^0) = \frac{\hat{\eta}_j - \hat{\eta}_l - \theta_{j,l}^0}{\sqrt{\frac{v_j^2+v_l^2}{2} MSe}}$$

$$p_{j,l}^T = 1 - CDF_{q,m,n-k}(|t_{j,l}^0|)$$

One-way classification

(Homoscedastic) G-sample problem

$Y_{gj} \overset{iid}{\sim} N(m_g, \sigma^2)$, $g = 1, \dots, G$, $j = 1, \dots, n_g$

Y_{gj} all independent

Lemma 13.1 (no previous knowledge)

$$\overline{Y} = \begin{pmatrix} \overline{Y}_{1\cdot} \\ \vdots \\ \overline{Y}_{G\cdot} \end{pmatrix} \sim N_G\left(\begin{pmatrix} m_1 \\ \vdots \\ m_G \end{pmatrix}, \sigma^2 \underbrace{\begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_G} \end{pmatrix}}_{V} \right)$$

$$\frac{(n-G)\, MSe}{\sigma^2} \sim \chi^2_{n-G}$$

MSe and $\overline{Y}$ independent (given covariates, given groups allocation)

Tukey's simultaneous conf. intervals for $m_g - m_h$, $g < h$:

$$\overline{Y}_{g\cdot} - \overline{Y}_{h\cdot} \pm q_{G, n-G}(1-\alpha) \sqrt{\frac{1}{2}\left(\frac{1}{n_g} + \frac{1}{n_h}\right) MSe}$$

(lead coverage of $1-\alpha$ in a balanced case)

→ R function TukeyHSD

Two-way classification, BALANCED data

$Y_{g,h,j} \overset{iid}{\sim} N(m_{g,h}, \sigma^2)$, $g = 1, \dots, G$

$Y_{g,h,j}$ all independent $h = 1, \dots, H$, $j = 1, \dots, n_{gh} = J$ (balanced)

Perhaps <u>additivity</u> can be assumed for effect of the grouping factors.

<u>Lemma 13.2</u>

- under both <u>additivity</u> and interaction model (= no structure)

$$\overline{Y} := \begin{pmatrix} \overline{Y}_{1\cdot} \\ \vdots \\ \overline{Y}_{G\cdot} \end{pmatrix} \sim N_G\left(\begin{pmatrix} \overline{m}_{1\cdot} \\ \vdots \\ \overline{m}_{G\cdot} \end{pmatrix}, \sigma^2 \begin{pmatrix} \frac{1}{J \cdot H} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{J \cdot H} \end{pmatrix} \right)$$

$$\overline{Y}_{g\cdot} = \frac{1}{J \cdot H} \sum_{h=1}^{H} \sum_{j=1}^{J} Y_{g,h,j} \qquad \overline{m}_{g\cdot} = \frac{1}{H} \sum_{h=1}^{H} m_{g,h}$$

$$\frac{\nu_e^* MS_e^*}{\sigma^2} \sim \chi^2_{\nu_e^*} \quad , \quad MS_e^* \text{ and } \overline{Y} \text{ independent (given the grouping factors)}$$

depending on which structure is assumed for the group means:

$\nu_e^* = n - (G + H - 1)$: additive model

$\nu_e^* = n - G \cdot H$: interaction model

Tukey's simultaneous confidence intervals

for $\overline{m}_{g_1\cdot} - \overline{m}_{g_2\cdot}$, $g_1 < g_2$

$$\overline{y}_{g_1\cdot} - \overline{y}_{g_2\cdot} \pm q_{G,\nu_e^*}(1-\alpha)\sqrt{\frac{1}{J\cdot H}\cdot MS_e^*}$$

Remember that in the additive model

$$\overline{m}_{g_1\cdot} - \overline{m}_{g_2\cdot} = m_{g_1,h} - m_{g_2,h}$$

for any $h = 1, \ldots, H$

$\Rightarrow$ R function TukeyHSD

UNBALANCED data

- TukeyHSD is still producing something
- What? $\rightarrow$ see notes