Southern Illinois University
CARBONDALE

BOOTSTRAPPING ANALOGS OF THE ONE WAY MANOVA TEST

Hasthika S. Rupasinghe
Department of Mathematics

Southern Illinois University
CARBONDALE

Contents

Two Sample Hotelling's T^2 Test

Suppose there are two independent random samples $x_{1,1}, \ldots, x_{n_1,1}$ and $x_{1,2}, \ldots, x_{n_2,2}$ from two populations or groups, and that it is desired to test

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 \neq \mu_2$$

where the μ_i are $p \times 1$ vectors.

Assume that T_i satisfies a central limit type theorem

$$\sqrt{n}(T_i - \mu_i) \xrightarrow{D} N_p(0, \Sigma_i) \text{ for } i = 1, 2$$

where the Σ_i are positive definite.

Two Sample Hotelling's T^2 Test...

To simplify large sample theory, assume $n_1 = kn_2$ for some positive real number k. Let $\hat{\Sigma}_i$ be a consistent nonsingular estimator of Σ_i. Then

$$\begin{pmatrix} \sqrt{n_1}\,(T_1 - \mu_1) \\ \sqrt{n_2}\,(T_2 - \mu_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right],$$

or

$$\begin{pmatrix} \sqrt{n_2}\,(T_1 - \mu_1) \\ \sqrt{n_2}\,(T_2 - \mu_2) \end{pmatrix} \xrightarrow{D} N_{2p}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\Sigma_1}{k} & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right].$$

Hence

$$\sqrt{n_2}\,[(T_1 - T_2) - (\mu_1 - \mu_2)] \xrightarrow{D} N_p(0, \frac{\Sigma_1}{k} + \Sigma_2).$$

Two Sample Hotelling's T^2 Test...

Using $nB^{-1} = \left(\frac{B}{n}\right)^{-1}$ and $n_2 k = n_1$, if $\mu_1 = \mu_2$, then

$$n_2(T_1 - T_2)^T \left(\frac{\Sigma_1}{k} + \Sigma_2\right)^{-1} (T_1 - T_2) =$$

$$(T_1 - T_2)^T \left(\frac{\Sigma_1}{n_1} + \frac{\Sigma_2}{n_2}\right)^{-1} (T_1 - T_2) \xrightarrow{D} \chi_p^2.$$

Hence

$$T_0^2 = (T_1 - T_2)^T \left(\frac{\hat{\Sigma}_1}{n_1} + \frac{\hat{\Sigma}_2}{n_2}\right)^{-1} (T_1 - T_2) \xrightarrow{D} \chi_p^2.$$

Note that k drops out of the above result.

Two Sample Hotelling's T^2 Test...

If the sequence of positive integers $d_n \to \infty$ and $Y_n \sim F_{p,d_n}$, then $Y_n \xrightarrow{D} \chi^2_p/p$.
Instead of rejecting H_0 when $T_0^2 > \chi^2_{p,1-\alpha}$, reject H_0 when

$$T_0^2 > pF_{p,d_n,1-\alpha} = \frac{pF_{p,d_n,1-\alpha}}{\chi^2_{p,1-\alpha}}\chi^2_{p,1-\alpha}.$$

The term $\dfrac{pF_{p,d_n,1-\alpha}}{\chi^2_{p,1-\alpha}}$ can be regarded as a small sample correction factor that improves the test's performance for small samples.
For example, use $d_n = (n_1 - p, n_2 - p)$. Here $P(Y_n \leq \chi^2_{p,\alpha}) = \alpha$ if Y_n has a χ^2_p distribution, and $P(Y_n \leq F_{p,d_n,\alpha}) = \alpha$ if Y_n has an F_{p,d_n} distribution.

One Way MANOVA test

The one way MANOVA is used to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p \text{ Vs. } H_a : not H_0.$$

Note that if $m = 1$ the one way MANOVA model becomes the one way ANOVA model. One might think that performing m ANOVA tests is sufficient to test the above hypotheses. But the separate ANOVA tests would not take the correlation between the m variables into account. On the other hand the MANOVA test will take the correlation into account.

One Way MANOVA test...

The one way MANOVA model is $y_{ij} = \mu_i + \epsilon_{ij}$ where the ϵ_{ij} are iid with $E(\epsilon_{ij}) = 0$ and $Cov(\epsilon_{ij}) = \Sigma_\epsilon$. The summary One Way MANOVA table is shown bellow.

Source	matrix	df
Treatment or Between	B_T	$p-1$
Residual or Error or Within	W	$n-p$
Total (Corrected)	T	$n-1$

One Way MANOVA test...

There are three commonly used test statistics to test above hypotheses. Namely,

1. Hotelling Lawley trace statistic: $U = tr(B_T W^{-1}) = tr(W^{-1} B_T)$
2. Wilks' lambda: $\Lambda = \frac{|W|}{|B_T + W|}$
3. Pillai's trace statistic: $\boldsymbol{V} = tr(B_T \boldsymbol{T}^{-1}) = tr(\boldsymbol{T}^{-1} B_T)$

One Way MANOVA test...

If the $y_{ij} - \mu_j$ are iid with common covariance matrix Σ_ϵ, and if H_0 is true, then under regularity conditions Fujikoshi (2002) showed

1. $(n - m - p - 1)U \xrightarrow{D} \chi^2_{m(p-1)}$,
2. $-[n - 0.5(m + p - 2)]log(\Lambda) \xrightarrow{D} \chi^2_{m(p-1)}$, and
3. $(n - 1)V \xrightarrow{D} \chi^2_{m(p-1)}$.

Contents

Brief introduction to bootstrapping.

Use the information of a number of resamples from the sample to estimate the population distribution.

Given a sample of size n:

1. Treat the sample as population.
2. Draw B samples of size n with replacement from your sample - the bootstrap samples.
3. Compute for each bootstrap sample the statistic of interest - for examples: the mean, the median.
4. Estimate the sample distribution of the statistic by the bootstrap sample distribution.

Bootstrap R Example

Suppose the data is 1, 2, 3, 4, 5, 6, 7, 8, 9 ,10. Then $n = 10$ and the sample median $T_n = 5.5$. R was used to draw $B = 2$ bootstrap samples (Samples of size n drawn with replacement from the original data) and computed the sample median $T^*_{1,n} = 6.5$, $T^*_{2,n} = 5.5$.

```
b1 <- sample(1:10, replace = T)
b1
[1] 3 8 7 4 1 7 6 10 9 3
median(b1)
[1] 6.5


b2 <- sample(1:10, replace = T)
b2
[1] 6 3 9 5 8 8 4 2 1 9
median(b2)
[1] 5.5
```

Contents

Some Notation

Suppose $w_1, ..., w_n$ are iid $r \times 1$ random vectors with mean μ and nonsingular covariance matrix Σ_w. Let a future test observation w_f be independent of the w_i but from the same distribution. Let $(\overline{w}, S)$ be the sample mean and sample covariance matrix where

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \text{ and } \mathrm{S} = \mathrm{S_W} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{w_i} - \overline{\mathrm{w}})(\mathrm{w_i} - \overline{\mathrm{w}})^{\mathsf{T}}. \quad (1)$$

Then the ith *squared sample Mahalanobis distance* is the scalar

$$D^2_{w_i} = D^2_{w_i}(\overline{w}, S) = (w_i - \overline{w})^T S^{-1}(w_i - \overline{w}). \quad (2)$$

Bootstrapping a hypothesis test

Let $D_i^2 = D^2_{w_i}$ for each observation w_i. Let $D_{(c)}$ be the cth order statistic of $D_1, ..., D_n$. Consider the hyperellipsoid

$$\mathcal{A}_n = \{w : D^2_w(\overline{w}, S) \leq D^2_{(c)}\}. \tag{3}$$

Bootstrapping a hypothesis test...

Olive (2013) showed that;

$$\{w : D^2_w(\overline{w}, S) \le D^2_{(U_n)}\} \tag{4}$$

is a large sample $100(1-\delta)\%$ **non parametric prediction region for a future value** w_f given iid data $w_1, ..., , w_n$. Let $D_{(U_n)}$ be the $100q_n$th percentile of the D_i;
where $q_n = (1-\delta+0.05, 1-\delta+p/n)$ for $\delta > 0.1$ and

$$q_n = (1-\delta/2, 1-\delta+10\delta p/n), \text{ otherwise.} \tag{5}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$.

while the classical large sample $100(1-\delta)\%$ prediction region is

$$\{w : D^2_w(\overline{w}, S) \le \chi^2_{p,1-\delta}\}. \tag{6}$$

Bootstrapping a hypothesis test...

Following Olive (2015);

Theorem 1. Let the $(1-\delta)$th percentile $D^2_{1-\delta}$ be a continuity point of the distribution of D^2. Assume that $D^2_{\mu}(T_n, \Sigma_T) \xrightarrow{D} D^2$, $D^2_{\mu}(T_n, \hat{\Sigma}_T) \xrightarrow{D} D^2$, and $\hat{D}^2_{1-\delta} \xrightarrow{P} D^2_{1-\delta}$ where $P(D^2 \le D^2_{1-\delta}) = 1-\delta$.

1. Then $R_c = \{w : D^2_w(T_n, \hat{\Sigma}_T) \le \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ confidence region for μ.
2. If μ is known, $R_p = \{w : D^2_w(\mu, \hat{\Sigma}_T) \le \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic $T_{f,n}$.
3. Region R_c contains μ iff region R_p contains T_n.

Bootstrapping a Hypothesis Test...

Hence if there was an iid sample $T_{1,n}, ..., T_{B,n}$ of the statistic, the prediction region (4) for $T_{f,n}$ contains $E(T_n) = \mu$ with asymptotic coverage $\geq 1 - \delta$.

The Olive (2015) prediction region method bootstraps this procedure by using a bootstrap sample of the statistic $T^*_{1,n}, ..., T^*_{B,n}$.

The prediction region method for testing $H_0 : \mu = c$ versus $H_1 : \mu \neq c$ is simple. Let $\hat{\mu}$ be a consistent estimator of μ and make a bootstrap sample $w_i = \hat{\mu}^*_i - c$ for $i = 1, ..., B$. Make the nonparametric prediction region (4) for the w_i and fail to reject H_0 if 0 is in the prediction region, reject H_0 otherwise.

Contents

Why new Tests?

1. The two sample Hotelling's T^2 test is the classical method. If it is not assumed that the population covariance matrices are equal, then this test uses the sample mean and sample covariance matrix $T_i = \overline{x}_i$ and $\hat{\Sigma}_i = S_i$ applied to each sample. This test is robust to assumptions;
   - Both populations are multivariate normally distributed.
   - The populations have a common population covariance matrix.

   But the test can be very poor if outliers are present.

2. The classical one way MANOVA model assumes that the covariance matrix of each group is the same. This test is also not an outlier resistant test.

Bootstrapping Analogs of the Hotelling's T^2 Test

Recall:

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_1 : \mu_1 \neq \mu_2$$

For illustrative purposes, the simulation study will take T_i to be the coordinatewise median, sample mean, 25% trimmed mean or the Olive and Hawkins (2010) RMVN estimator T_{RMVN}.

Let the TEST 1, 2, 3 and 4 use,

1. T_1 - Coordinatewise median applied to the i^{th} sample.
2. T_2 - Sample mean applied to the i^{th} sample.
3. T_3 - 25% trimmed mean applied to the i^{th} sample.
4. T_4 - RMVN location estimator applied to the i^{th} sample.

respectively.

Bootstrapping Analogs of the Hotelling's T^2 Test...

TEST 1:

Definition 1: *Coordinatewise Median* is defined as $MED(X) = (Med(X_1), ..., Med(X_p))'$

Steps to perform Test 1:

1. Make a bootstrap sample of $w_j = MED(X_j) - MED(Y_j)$ for $j = 1, ..., B$.
2. Makes the prediction region described above for the w_j.
3. Determine whether 0 is in the prediction region.
4. Make the decision,
   - If 0 is in the prediction region, the test 1 fails to reject H_0.
   - Reject H_0 if 0 is not in the prediction region.

To get the steps for the test 2 and test 3, simply replace w_j above by difference of $\bar{X}s'$ or 25% trimmed $\bar{X}s'$

Bootstrapping Analogs of the Hotelling's T^2 Test...

TEST 4:

Steps to perform Test 4:

1. Make a bootstrap sample w_j : the difference of the RMVN location estimators of the two groups for $j = 1, ..., B$.
2. Makes the prediction region described above for the w_j.
3. Determine whether 0 is in the prediction region.
4. Make the decision,
   - If 0 is in the prediction region, the test 1 fails to reject H_0.
   - Reject H_0 if 0 is not in the prediction region.

Simulation: Distributions

Simulation uses 5000 runs with 1000 bootstrap samples. Four types of data distribution have considered.

Distributions

Use AX where $A = diag(1, \sqrt{2}, ..., \sqrt{p})$ and X from;

- Multivariate normal: $N_p(\mu, I)$.
- Multivariate t_4.
- Mixture distribution: $(0.6)N_p(0, I) + 0.4N_p(0, 25I)$.
- Multivariate lognormal dist. shifted to have nonzero mean $\mu = 0.6491$, but a population coordinatewise median of 0.

Note that Cov(x_2) = σ^2 Cov(x_1), and for the first three distributions, $E(x_i) = E(w_i) = 0$ if $\delta = 0$.

Simulation: Outliers

Five outlier types have considered. $100\gamma\%$ of the data;

Outliers

1. A tight cluster at major axis: $(0, ..., 0, pm)^{'}$
2. A tight cluster at minor axis: $(pm, 0, ..., 0)^{'}$
3. Point mass: $N((pm, ..., pm)^{'}, diag(1, ..., p))$
4. x_{1p} replaced by pm
5. x_{11} replaced by pm

Let the *coverage* be the proportion of times that H_0 is rejected. We want the *coverage* near 5% when H_0 is true and the coverage close to 100% for good power when H_0 is false. A *coverage* outside of (4%, 6%) suggests that the true *coverage* is not 5%.

Simulation: Outputs

Type I error rates for clean multivariate normal data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0470	0.0554	0.0568	0.0402	0.0560
				1000	0.0440	0.0606	0.0540	0.0414	
			2	250	0.0472	0.0550	0.0574	0.0422	0.0498
				1000	0.0420	0.0568	0.0538	0.0392	
5	250	500	1	250	0.0490	0.0524	0.0496	0.0394	0.0552
				1000	0.0462	0.0588	0.0584	0.0448	
			2	250	0.0460	0.0540	0.0524	0.0436	0.0070
				1000	0.0470	0.0500	0.0534	0.0386	
15	750	750	1	750	0.0462	0.0626	0.0622	0.0466	0.0450
				1000	0.0390	0.0514	0.0470	0.0378	
			2	750	0.0492	0.0598	0.0608	0.0464	0.0516
				1000	0.0474	0.0556	0.0568	0.0446	
15	750	1500	1	750	0.0466	0.0538	0.0550	0.0466	0.0480
				1000	0.0492	0.0556	0.0548	0.0444	
			2	750	0.0424	0.0538	0.0520	0.0454	0.0014
				1000	0.0514	0.0532	0.0542	0.0426	

Simulation: Outputs

Type I error rates for clean $0.6N_p(0,I) + 0.4N_p(0,25I)$

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0420	0.0560	0.0480	0.0394	0.0462
				1000	0.0386	0.0532	0.0464	0.0336	
			2	250	0.0454	0.0550	0.0476	0.0416	0.0476
				1000	0.0370	0.0484	0.0400	0.0368	
5	250	500	1	250	0.0460	0.0542	0.0538	0.0416	0.0470
				1000	0.0368	0.0502	0.0416	0.0404	
			2	250	0.0480	0.0600	0.0474	0.0390	0.0060
				1000	0.0416	0.0598	0.0498	0.0416	
15	750	750	1	750	0.0434	0.0536	0.0540	0.0448	0.0496
				1000	0.0406	0.0598	0.0474	0.0396	
			2	750	0.0468	0.0626	0.0518	0.0456	0.0464
				1000	0.0456	0.0566	0.0490	0.0454	
15	750	1500	1	750	0.0456	0.0584	0.0568	0.0488	0.0502
				1000	0.0426	0.0550	0.0478	0.0438	
			2	750	0.0456	0.0576	0.0508	0.0442	0.0004
				1000	0.0416	0.0572	0.0488	0.0510	

Simulation: Outputs

Type I error rates for clean multivariate t_4 data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0442	0.0574	0.0570	0.0266	0.0456
				1000	0.0426	0.0570	0.0530	0.0282	
			2	250	0.0496	0.0618	0.0614	0.0328	0.0542
				1000	0.0480	0.0558	0.0578	0.0292	
5	250	500	1	250	0.0484	0.0512	0.0540	0.0346	0.0504
				1000	0.0420	0.0488	0.0494	0.0310	
			2	250	0.0408	0.0580	0.0526	0.0348	0.0058
				1000	0.0410	0.0492	0.0510	0.0348	
15	750	750	1	750	0.0470	0.0550	0.0562	0.0232	0.0414
				1000	0.0382	0.0526	0.0476	0.0228	
			2	750	0.0472	0.0572	0.0542	0.0248	0.0442
				1000	0.0502	0.0496	0.0556	0.0258	
15	750	1500	1	750	0.0482	0.0556	0.0528	0.0224	0.0446
				1000	0.0464	0.0496	0.0528	0.0254	
			2	750	0.0442	0.0534	0.0502	0.0314	0.0016
				1000	0.0452	0.0508	0.0554	0.0262	

Simulation: Outputs

Type I error rates for clean lognormal data

p	n_1	n_2	σ	B	Median	Mean	Tr.Me	RMVN	Class
5	250	250	1	250	0.0408	0.0460	0.0514	0.0274	0.0470
				1000	0.0388	0.0494	0.0474	0.0254	
			2	250	0.0436	0.9816	0.0858	0.1108	0.9968
				1000	0.0398	0.9846	0.0788	0.1168	
5	250	500	1	250	0.0398	0.0540	0.0496	0.0316	0.0472
				1000	0.0368	0.0588	0.0446	0.0292	
			2	250	0.0418	0.9998	0.1192	0.2492	0.9964
				1000	0.0424	0.9994	0.1158	0.2520	
15	750	750	1	750	0.0402	0.0506	0.0480	0.0216	0.0502
				1000	0.0410	0.0444	0.0490	0.0238	
			2	750	0.0506	1.0000	0.3670	1.0000	1.0000
				1000	0.0510	1.0000	0.3748	1.0000	
15	750	1500	1	750	0.0420	0.0580	0.0514	0.0258	0.0514
				1000	0.0478	0.0558	0.0608	0.0284	
			2	750	0.0446	1.0000	0.6110	1.0000	1.0000
				1000	0.0464	1.0000	0.6256	1.0000	

Simulation: Outputs

Type I error rates and cutoffs with outliers for $p = 4$

Dist.	Otype	γ	pm		Med	Mean	Tr.Me	RMVN	Class
MVN	1	0.4	10	Cov	0.6946	1.0000	1.0000	0.0330	1.0000
				cut	10.158	9.769	9.798	10.701	
	2	0.4	20	Cov	0.5232	1.0000	1.0000	0.0382	1.0000
				cut	9.836	9.776	9.809	9.268	
	3	0.4	20	Cov	0.8578	1.0000	1.0000	0.0402	1.0000
				cut	10.214	9.761	9.760	9.288	
	4	0.1	10	Cov	0.0980	0.8654	0.1450	0.0382	0.8684
				cut	9.898	9.771	9.777	9.851	
Mix	2	0.4	20	Cov	0.0828	1.0000	1.0000	0.0144	1.0000
				cut	10.542	9.788	9.878	11.300	
	5	0.1	10	Cov	0.0820	0.5306	0.1228	0.0184	0.5276
				cut	9.933	9.779	9.881	11.056	
MVT	1	0.4	10	Cov	0.0854	0.6700	0.1548	0.0204	1.0000
				cut	10.232	9.799	9.787	10.200	
	5	0.1	20	Cov	0.0864	1.0000	0.1418	0.0304	1.0000
				cut	9.924	9.795	9.795	9.830	
Log	3	0.4	20	Cov	0.0778	1.0000	1.0000	0.0162	1.0000
				cut	13.689	9.822	9.827	12.607	
	4	0.1	10	Cov	0.0842	0.3158	0.1482	0.0234	0.3044
				cut	10.013	9.875	9.872	10.416	

Simulation: Power

- In the power simulation, $\delta > 0$ was used.
- Hence for the first three distributions $\mu_2 = 0$ and $\mu_1 = \delta(1, ..., 1)^T$.
- Then the Euclidean distance between the two means was $\sqrt{p}$, where p is the number of parameters.
- Therefore the distance increases as p increase.
- The value of δ had to be fairly small so that the simulated power was not always 1.

Power Simulation Results

$Coverages$ **when H_0 is false for MVN data.**

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.35	0.9598	0.9990	0.9928	0.9942	0.9988
			1000	0.35	0.9684	0.9994	0.9970	0.9978	
		2	250	0.35	0.5958	0.8442	0.7672	0.7604	0.8402
			1000	0.35	0.5832	0.8346	0.7438	0.7470	
15	750	1	750	0.15	0.7394	0.9552	0.9012	0.9268	0.9556
			1000	0.15	0.7474	0.9522	0.8984	0.9178	
		2	750	0.15	0.3078	0.5318	0.4550	0.4468	0.5156
			1000	0.15	0.3118	0.5218	0.4430	0.4464	

Power Simulation Results

$Coverages$ **when H_0 is false for mixture data.**

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.45	0.8826	0.4062	0.9304	0.9938	0.4032
			1000	0.45	0.8858	0.4058	0.9338	0.9948	
		2	250	0.45	0.4458	0.1910	0.5222	0.7454	0.1642
			1000	0.45	0.4656	0.1890	0.5386	0.7626	
15	750	1	750	0.20	0.6204	0.2274	0.7148	0.9492	0.2114
			1000	0.20	0.6316	0.2228	0.7190	0.9494	
		2	750	0.20	0.2318	0.1154	0.2894	0.5034	0.1042
			1000	0.20	0.2438	0.1092	0.2916	0.4980	

Power Simulation Results

$Coverages$ when H_0 is false for multivariate t_4 data.

p	$n_1 = n_2$	σ	B	δ	Med	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.38	0.9642	0.9562	0.9916	0.9878	0.9548
			1000	0.38	0.9728	0.9572	0.9944	0.9880	
		2	250	0.38	0.5958	0.5960	0.7198	0.6488	0.6074
			1000	0.38	0.6188	0.6152	0.7490	0.6636	
15	750	1	750	0.20	0.9418	0.9270	0.9868	0.9714	0.9232
			1000	0.20	0.9422	0.9304	0.9860	0.9724	
		2	750	0.20	0.4934	0.4932	0.6422	0.5384	0.4754
			1000	0.20	0.4842	0.4916	0.6362	0.5252	

Power Simulation Results

$Coverages$ when H_0 is false for lognormal data.

p	$n_1 = n_2$	σ	B	δ	Median	Mean	Tr.Me	RMVN	Class
5	250	1	250	0.45	0.9982	0.8256	0.9994	0.8790	0.8208
			1000	0.45	0.9980	0.8324	0.9996	0.8830	
		2	250	0.45	0.8210	0.4704	0.6488	0.0914	0.4630
			1000	0.45	0.8378	0.4646	0.6624	0.1038	
15	750	1	750	0.30	1.0000	0.9186	1.0000	0.8514	0.9120
			1000	0.30	1.0000	0.9178	1.0000	0.8544	
		2	750	0.30	0.9436	1.0000	0.5042	0.9438	1.0000
			1000	0.30	0.9484	1.0000	0.5022	0.9424	

Real data example

Data description:

- The Johnson (1996) STATLIB bodyfat data consists of 252 observations on 15 variables including the density determined from underwater weighing and the percent body fat measurement.
- Consider these two variables with two age groups: age ≤ 50 and age > 50.

Real data example

Data description:

- The Johnson (1996) STATLIB bodyfat data consists of 252 observations on 15 variables including the density determined from underwater weighing and the percent body fat measurement.
- Consider these two variables with two age groups: age ≤ 50 and age > 50.

Classical test results vs. new test(s) results:

- The test with the RMVN estimator had $D_0 = 1.78$ while the test with the coordinatewise median had $D_0 = 1.35$.
- Both tests had cutoffs near 2.37 and fail to reject H_0.
- The classical two sample Hotelling's T^2 test rejects H_0 with a test statistic of 4.74 and a p-value of 0.001.

Real data example

The DD plots, shown in Figures 1 and 2, reveal five outliers.

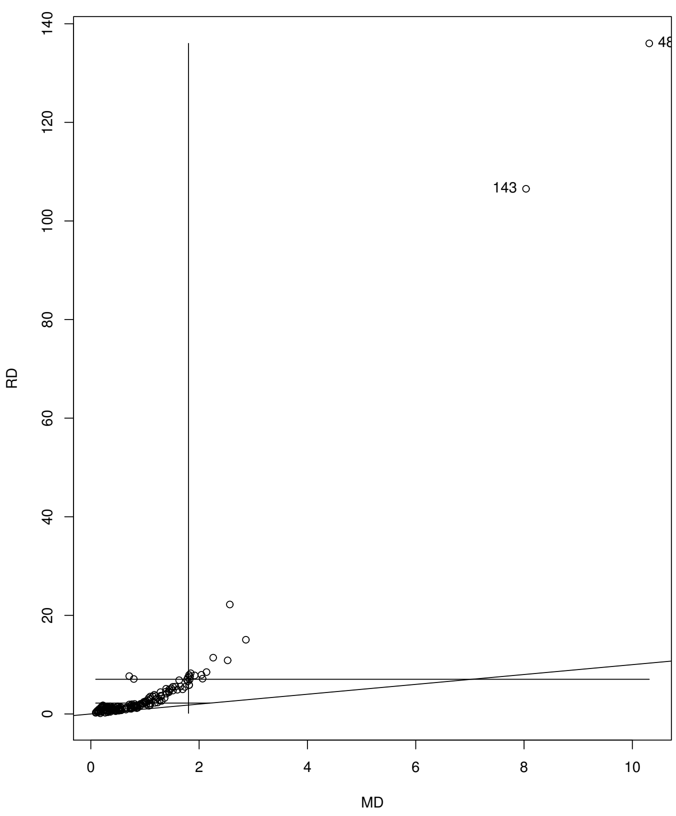

Figure:

DD plot for the age ≤ 50 group.

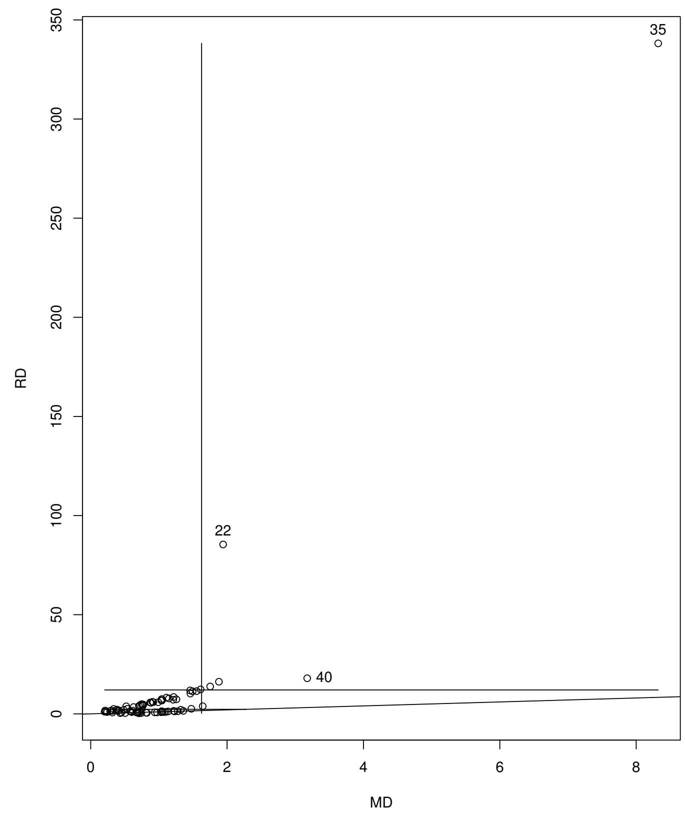

Figure:

DD plot for the age > 50 group.

Real data example

- After deleting the outliers, all three tests fail to reject H_0.
- The RMVN test had $D_0 = 1.63$ with cutoff 2.25, the coordinatewise median test had $D_0 = 1.22$ with cutoff 2.38.
- Classical test had test statistic 2.39 with a p-value of 0.09.

Alternative tests for the classical MANOVA

Theorem

If

$$\begin{pmatrix} \sqrt{n_1}\,(T_1-\mu_1) \\ \vdots \\ \sqrt{n_p}\,(T_p-\mu_p) \end{pmatrix} \xrightarrow{D} N_{mp}\left[\begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_p \end{pmatrix}\right],$$

then under $H_0 : \mu_1 = \cdots = \mu_p$

$$\sqrt{n}\begin{pmatrix} T_1-T_p \\ \vdots \\ T_{p-1}-T_p \end{pmatrix} \xrightarrow{D} N_{(m-1)p}\left[\begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\Sigma_1}{k_1}+\frac{\Sigma_p}{k_p} & \frac{\Sigma_p}{k_p} & & \frac{\Sigma_p}{k_p}\ldots & \frac{\Sigma_p}{k_p} \\ \frac{\Sigma_p}{k_p} & \frac{\Sigma_2}{k_2}+\frac{\Sigma_p}{k_p} & \frac{\Sigma_p}{k_p} & \ldots & \frac{\Sigma_p}{k_p} \\ \vdots & & \ddots & & \vdots \\ \frac{\Sigma_p}{k_p} & \frac{\Sigma_p}{k_p} & \frac{\Sigma_m}{k_p} & \ldots & \frac{\Sigma_{p-1}}{k_{p-1}}+\frac{\Sigma_p}{k_p} \end{pmatrix}\right].$$

Proof.

See Rupasinghe Arachchige Don (2017) for the proof.

□

Deriving a better test.

Large sample theory can be used to derive a better test that does not need the equal population covariance matrix assumption $\Sigma_i \equiv \Sigma_\epsilon$.
To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^p \pi_i = 1$.
Assume H_0 is true, and let $\mu_i = \mu$ for $i = 1, ..., p$. Suppose the $\mu_i = \mu$ and $\sqrt{n_i}(T_i - \mu) \xrightarrow{D} N_m(0, \Sigma_i)$, and $\sqrt{n}(T_i - \mu) \xrightarrow{D} N_m\left(0, \frac{\Sigma_i}{\pi_i}\right)$. Let

$$w = \begin{bmatrix} T_1 - T_p \\ T_2 - T_p \\ \vdots \\ T_{p-2} - T_p \\ T_{p-1} - T_p \end{bmatrix}. \tag{7}$$

Deriving a better test...

Then $\sqrt{n}w \xrightarrow{D} N_{m(p-1)}(0, \Sigma_w)$ with $\Sigma_w = (\Sigma_{ij})$ where $\Sigma_{ij} = \dfrac{\Sigma_p}{\pi_p}$ for $i \neq j$, and $\Sigma_{ii} = \dfrac{\Sigma_i}{\pi_i} + \dfrac{\Sigma_p}{\pi_p}$ for $i = j$.

Hence

$$t_0 = nw^T \hat{\Sigma}_w^{-1} w = w^T \left(\frac{\hat{\Sigma}_w}{n}\right)^{-1} w \xrightarrow{D} \chi^2_{m(p-1)}$$

as the $n_i \to \infty$ if H_0 is true.

Deriving a better test...

Here

$$\frac{\hat{\Sigma}_w}{n} = \begin{bmatrix} \frac{\hat{\Sigma}_1}{n_1} + \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_p}{n_p} & \dots & \frac{\hat{\Sigma}_p}{n_p} \\ \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_2}{n_2} + \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_p}{n_p} & \dots & \frac{\hat{\Sigma}_p}{n_p} \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_p}{n_p} & \frac{\hat{\Sigma}_p}{n_p} & \dots & \frac{\hat{\Sigma}_{p-1}}{n_{p-1}} + \frac{\hat{\Sigma}_p}{n_p} \end{bmatrix}$$

is a block matrix where the off diagonal block entries equal $\hat{\Sigma}_p/n_p$ and the ith diagonal block entry is $\frac{\hat{\Sigma}_i}{n_i} + \frac{\hat{\Sigma}_p}{n_p}$ for $i = 1, ..., (p-1)$.

Deriving a better test...

Reject H_0 if $t_0 > m(p-1)F_{m(p-1),d_n}(1-\alpha)$ where $d_n = min(n_1, ..., n_p)$.
This test may start to outperform the one way MANOVA test if $n \geq (m+p)^2$ and $n_i \geq 20m$ for $i = 1, ..., p$.

A useful one way MANOVA model

- A useful one way MANOVA model is $Z = XB + E$.
- where X is the full rank matrix where the first column of X is $v_1 = 1$
- ith column v_i of X is an indicator for group $i - 1$ for $i = 2, ..., p$.
- For example, $v_3 = (0^T, 1^T, 0^T, ..., 0^T)^T$
- where the p vectors in v_3 have lengths $n_1, n_2, ..., n_p$, respectively.

A useful one way MANOVA model

Let,

$$Y_{ij} = \begin{pmatrix} Y_{ij1} \\ \vdots \\ Y_{ijm} \end{pmatrix} = \mu_i + e_{ij}, \qquad EY_{ij} = \mu_i = \begin{pmatrix} \mu_{ij1} \\ \vdots \\ \mu_{ijm} \end{pmatrix}$$

for $i = 1, \dots, p$ and $j = 1, \dots, n_i$

$$Z = \begin{pmatrix} Y_{11}^T \\ \vdots \\ Y_{1n_1}^T \\ Y_{21}^T \\ \vdots \\ Y_{2n_2}^T \\ \vdots \\ Y_{p-1,1}^T \\ \vdots \\ Y_{p-1,n_{p-1}}^T \\ \vdots \\ Y_{p,1}^T \\ \vdots \\ Y_{p,n_p}^T \end{pmatrix} \quad and \quad X = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

A useful one way MANOVA model

$$B = \begin{bmatrix} \mu_p^T \\ (\mu_1 - \mu_p)^T \\ (\mu_2 - \mu_p)^T \\ \vdots \\ (\mu_{p-2} - \mu_p)^T \\ (\mu_{p-1} - \mu_p)^T \end{bmatrix}.$$

Thus testing $H_0 : \mu_1 = \cdots = \mu_p$ is equivalent to testing $H_0 : LB = 0$ where $L = [0 \; I_{p-1}]$

Test that is even better?

Test H_0 when $\hat{\Sigma}_w$ is unknown or difficult to estimate.

Since the common covariance matrix assumption in classical MANOVA test Cov$(\epsilon_k) = \Sigma_\epsilon$ for $k = 1, ..., n$ is extremely strong, using the prediction region method to test $H_0 : LB = 0$ may be a useful alternative.

Test that is even better?...

Steps

- Take a sample of size n_i with replacement from the n_i cases for each group for $i = 1, 2, ..., p$.
- Let $\hat{B}_i^*$ be the ith bootstrap estimator of B for $i = 1, ..., B$.
- Let the $(p-1)m \times 1$ vector $w_i = vec(L\hat{B}_i^*) = ((\hat{\mu}_1^* - \hat{\mu}_p^*)^T, ..., (\hat{\mu}_{p-1}^* - \hat{\mu}_p^*)^T)^T$ for $i = 1, ..., B$, where $vec(A)$ stacks columns of a matrix into a vector.
- For a robust test use $w_i = ((T_1^* - T_p^*)^T, ..., (T_{p-1}^* - T_p^*)^T)^T$ where T_i is a robust location estimator, such as;
  - The coordinatewise median;
  - Trimmed mean; applied to the cases in the ith treatment group.
- The prediction region method fails to reject H_0 if 0 is in the resulting confidence region. We likely need $n \geq 20mp$, $n \geq (m+p)^2$, and $n_i \geq 20m$.

Test that is even better?...

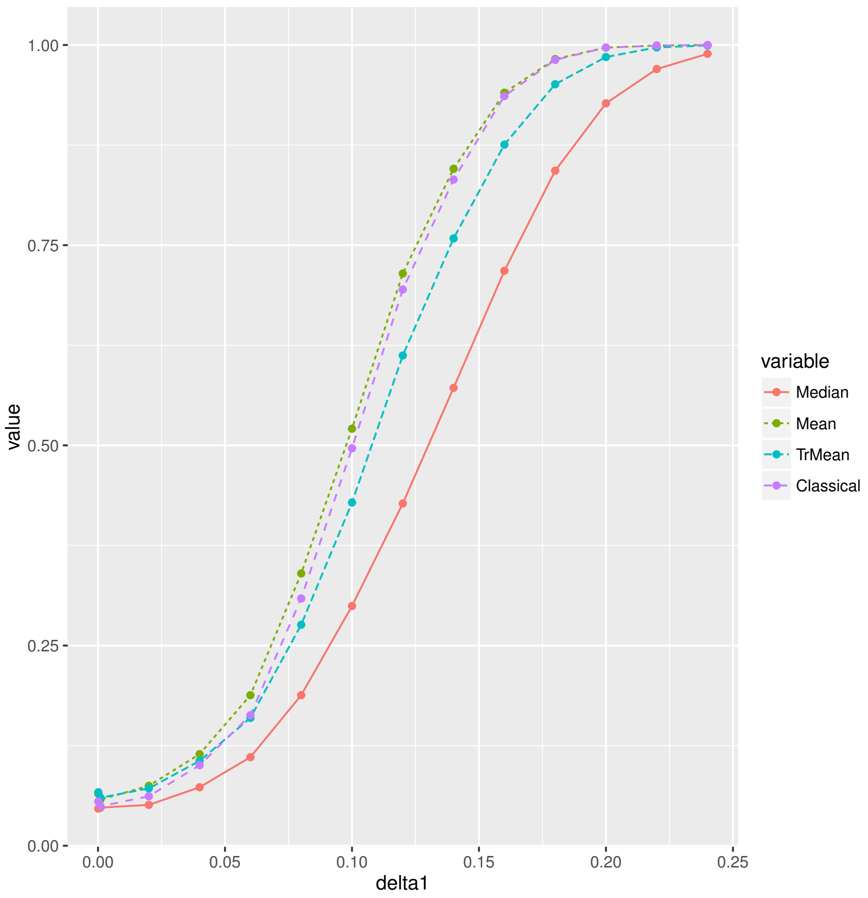

Figure: Power Curve for clean MVN Data, $m = 5$ with a balanced design

Test that is even better?...

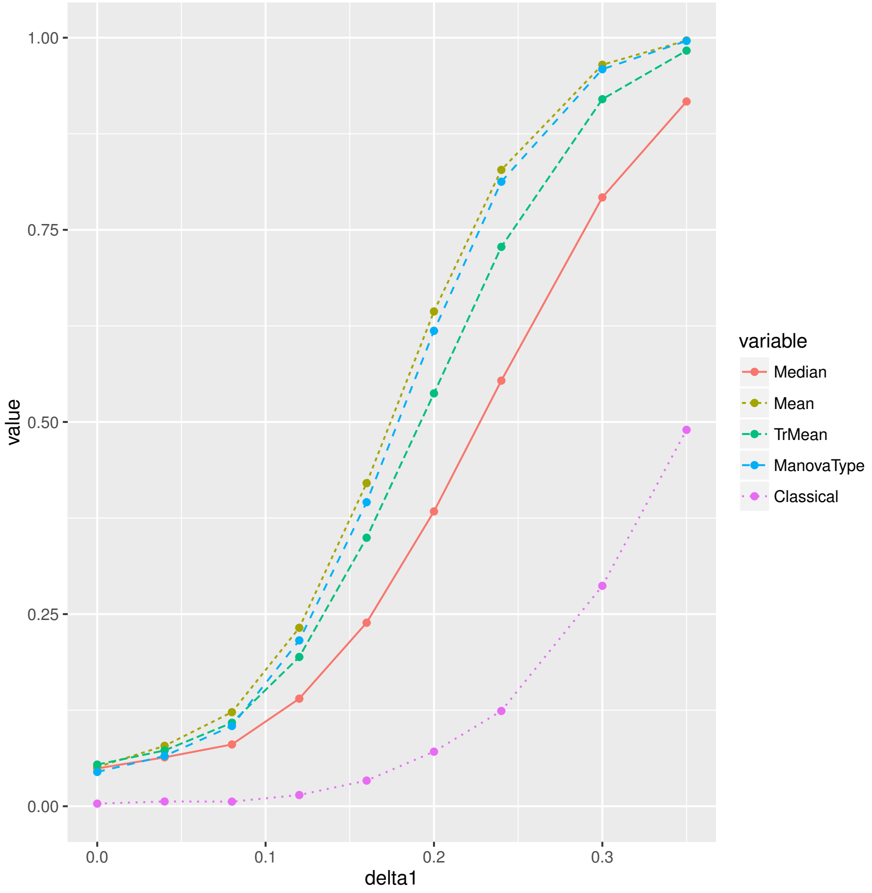

Figure: Power Curve for clean MVN Data, $m = 5$ with an unbalanced design

Test that is even better?...

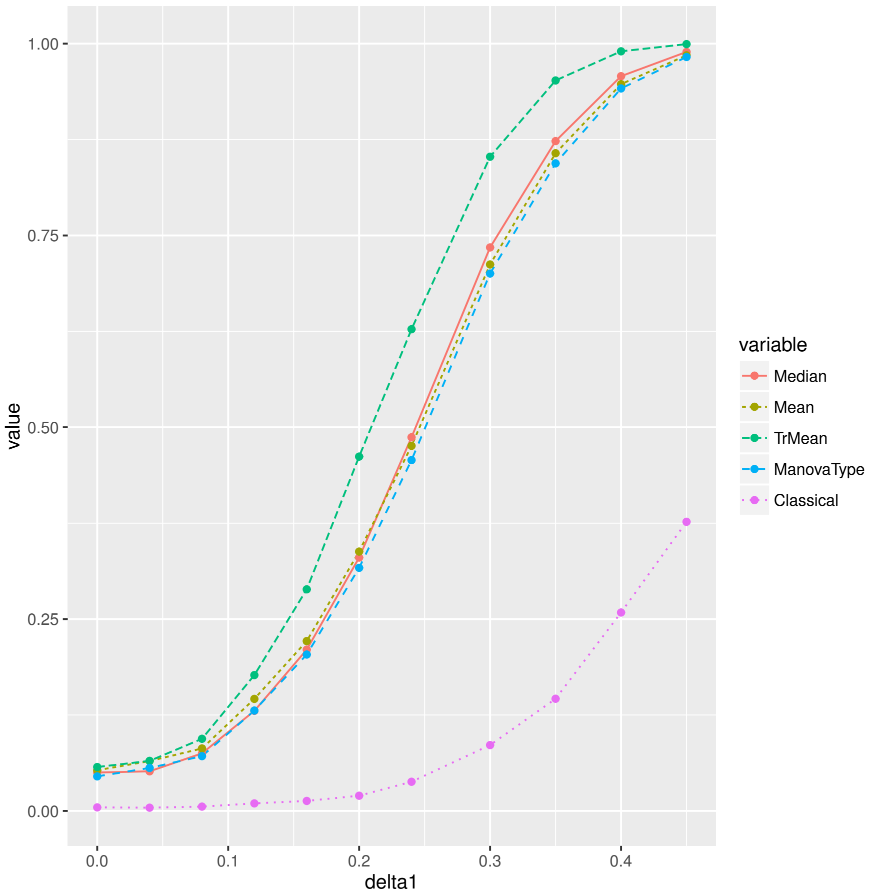

Figure: Power Curve for clean Multivariate t_5 Data, $m = 5$ with an unbalanced design

Real data example

- North Carolina Crime data consists of $n = 630$ observations on 24 variables.
- Region is a categorical variable with three categories viz.
  1. Central
  2. West
  3. Other
- Number of observations $n_1 = 232$, $n_2 = 146$ and $n_3 = 245$ respectively and has considered as the three groups.
- This example uses;
  1. "wsta" - weekly wage of state employees,
  2. "avgsen" - average sentence days,
  3. "prbarr" - 'probability' of arrest,
  4. "prbconv" - 'probability' of conviction,
  5. "taxpc" - tax revenue per capita

  as variables.

Real data example...

The DDplots reveals a few outliers.

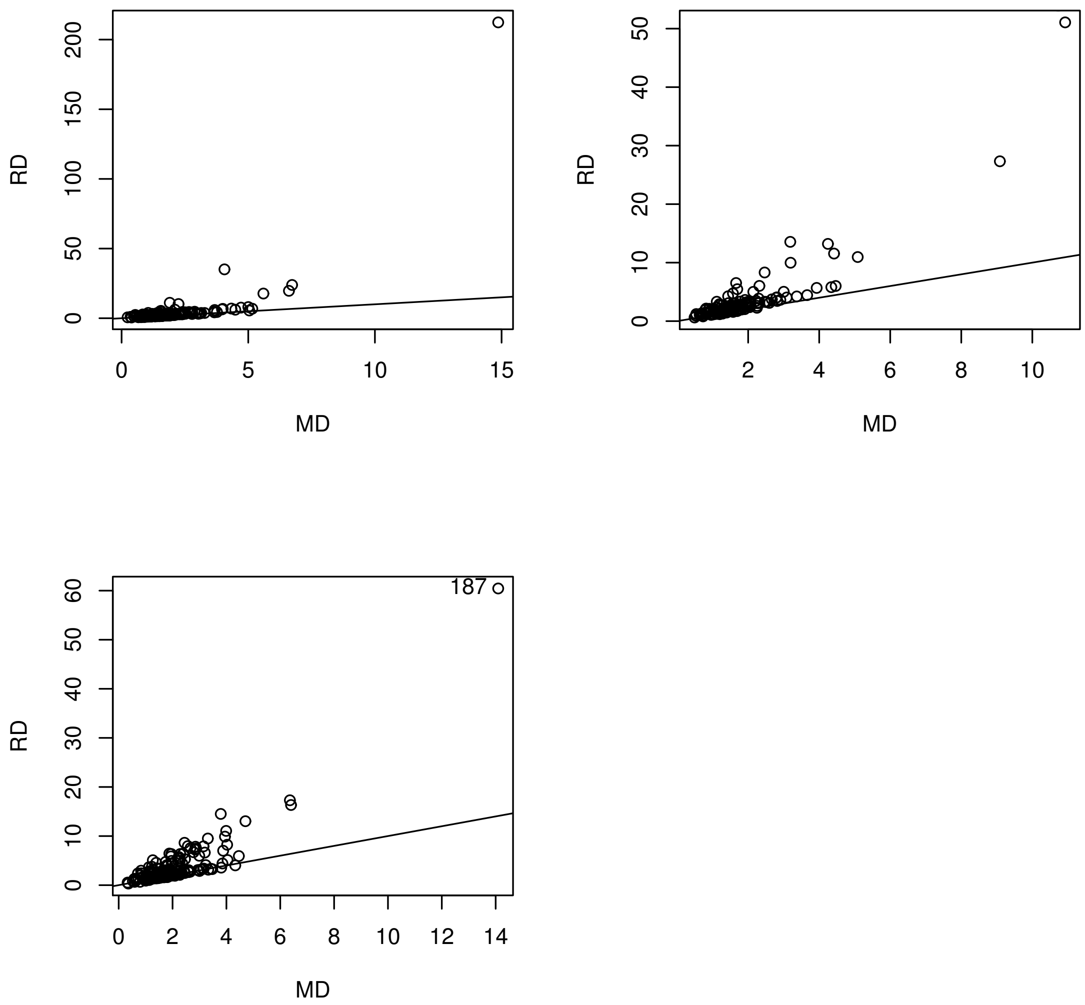

Figure: DDplots for Crime data

Real data example...

Furthermore the boxplots shows that the data are highly skewed.

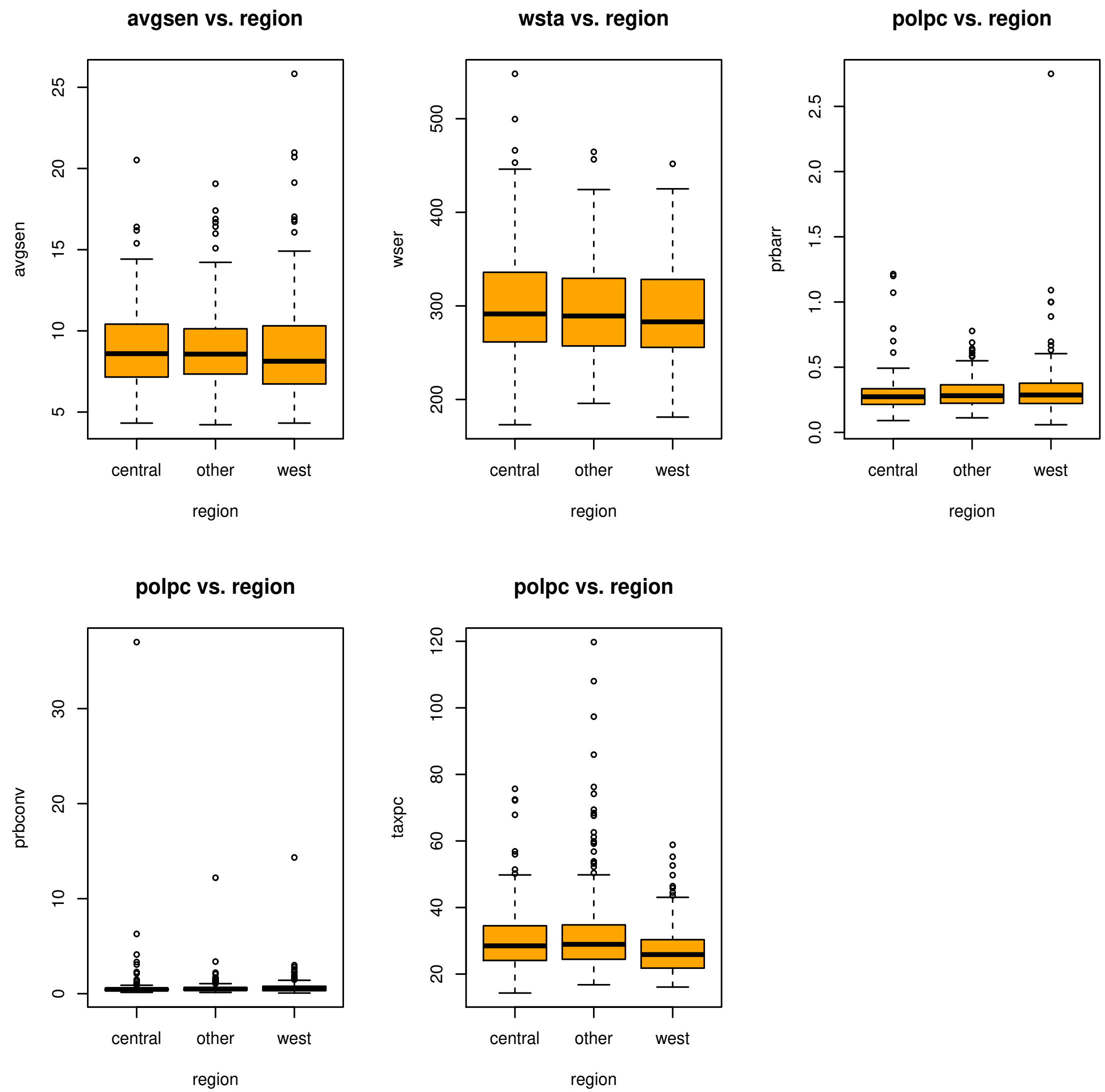

Figure: Side by side boxplots for Crime data

Real data example...

Classical vs. The New Test...

- **New test:** The test with the median had $D_0 = 4.086$ with the cutoff of 4.32 and failed to reject H_0.
- **Classical:** The classical one-way MANOVA test had a p-value of 0.001 and rejected the null hypothesis.

References

Bickel, P. J., Ren, J. –J. (2001). *The bootstrap in hypothesis testing.* In: de Gunst, M., Klaassen, C., van der Vaart, A. Eds. State of the Art in Probability and Statistics: Festschrift for William R. van Zwet. The Institute of Mathematical Statistics, CA: Hayward, 91-112.

Clarke, B. R. (1986). *Nonsmooth analysis and Fréchet differentiability of M functionals.* Probability Theory and Related Fields 73:137-209.

Clarke, B. R. (2000). *A review of differentiability in relation to robustness with an application to seismic data analysis.* Proceedings of the Indian National Science Academy, A 66:467-482.

Curran, J. M. (2013). *Hotelling: Hotelling's T-squared test and variants. R* Package version 1.0-2, (https://cran.r-project.org/package=Hotelling).

Fernholtz, L. T. (1983). *von Mises Calculus for Statistical Functionals*. New York: Springer.

Finch, H., French, B. (2013). *A "Monte Carlo" comparison of robust MANOVA test statistics.* Journal of Modern Applied Statistical Methods 12:35-81.

Fujikoshi, Y. (2002), *Asymptotic Expansions for the Distributions of Multivariate Basic Statistics and One-Way MANOVA Tests Under Nonnormality,* Journal of Statistical Planning and Inference, 108, 263-282.

Gill, R. D. (1989). *Non- and semi-parametric maximum likelihood estimators and the von Mises method, part 1.* Scandinavian Journal of Statistics 16:97-128.

Johnson, R.A., and Wichern, D.W. (1988), *Applied Multivariate Statistical Analysis,* 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

Johnson, R. W. (1996). *Fitting percentage of body fat to simple body measurements.* Journal of Statistics Education *4*, online at (www.amstat.org/publications/jse/).

Kakizawa, Y. (2009), *Third-Order Power Comparisons for a Class of Tests for Multivariate Linear Hypothesis Under General Distributions,* Journal of Multivariate Analysis, 100, 473-496.

Konietschke, F., Bathke, A. C., Harrar, S. W., Pauly, M. (2015). *Parametric and nonparametric bootstrap methods for general MANOVA.* Journal of Multivariate Analysis 140:291-301.

Machado, J.A.F., and Parente, P. (2005), *Bootstrap Estimation of Covariance Matrices Via the Percentile Method,* Econometrics Journal, 8, 70–78.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis,* Academic Press, London, UK.

References...

Olive, D. J. (2004). *A resistant estimator of multivariate location and dispersion.* Computational Statistics $\&$ Data Analysis 46:99-102.
Olive, D. J. (2013). *Asymptotically optimal regression prediction intervals and prediction regions for multivariate data.* International Journal of Statistics and Probability 2:90-100.
Olive, D. J. (2016). *Bootstrapping hypothesis tests and confidence regions.* Unpublished Manuscript with the bootstrap material from Olive (2017b) at (http://lagrange.math.siu.edu/Olive/ppvselboot.pdf).
Olive, D. J. (2017a). *Applications of hyperellipsoidal prediction regions.* Statistical Papers. To appear.
Olive, D. J. (2017b). *Robust Multivariate Analysis.* New York: Springer. To appear.
Olive, D. J., Hawkins, D. M. (2010). *Robust multivariate location and dispersion.* Unpublished Manuscript available from (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).
Olive, D.J., Pelawa Watagoda, L.C.R., and Rupasinghe Arachchige Don, H.S. (2015), *Visualizing and Testing the Multivariate Linear Regression Model,* International Journal of Statistics and Probability, 4, 126-137.
Rupasinghe Arachchige Don, H.S. and Pelawa Watagoda, L.C.R., (2017), *Bootstrapping Analogs of the Two Sample Hotelling's T^2 Test,* Preprint at (http://lagrange.math.siu.edu/Olive/stwosample.pdf).
Press, S.J. (2005), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Dover, Mineola, NY.
R Core Team (2016). R: *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (www.R-project.org).
Ren, J.-J. (1991), *On Hadamard Differentiability of Extended Statistical Functional,* Journal of Multivariate Analysis, 39, 30-43.
Ren, J.-J., and Sen, P.K. (1995), *Hadamard Differentiability on D[0,1$]^p$,* Journal of Multivariate Analysis, 55, 14-28.
Todorov, V., Filzmoser, P. (2010). *Robust statistics for the one-way MAVOVA.* Computational Statistics $\&$ Data Analysis 54:37-48.
Van Aelst, S., Willems, G. (2011). *Robust and efficient one-way MANOVA* tests. Journal of the American Statistical Association 106:706-718.
Wilcox, R. R. (1995). *Simulation results on solutions to the multivariate "Behrens-Fisher" problem via trimmed means.* The Statistician 44:213-225.
Willems, G., Pison, G., Rousseeuw, P. J., Van Aelst, S. (2002). *A robust "Hotelling" test.* Metrika 55:125-138.
Zhang, J., Olive, D. J., Ye, P. (2012). *Robust covariance matrix estimation with canonical correlation analysis.* International Journal of Statistics and Probability 1:119-136.

Southern Illinois University
CARBONDALE

Thank you!

Hasthika S. Rupasinghe
hasthika@siu.edu