APPENDIX A

Methods for determining exact confidence intervals for δ, δ_D and ψ.

1. Two independent groups:

If x_1 and x_2 are measurements from normally distributed populations, then

$T_v(ncp) = \dfrac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{1/n_1 + 1/n_2}}$ has a non-central t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom and non-centrality parameter $ncp = \delta\sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$, with $\delta = (\mu_1 - \mu_2)/\sigma$.

The lower and upper bounds ncp_L and ncp_U are now chosen that

$$P\left(T_v(ncp_L) \geq \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}\right) = \alpha/2$$

and

$$P\left(T_v(ncp_U) \leq \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{1/n_1 + 1/n_2}}\right) = \alpha/2,$$

which means that

$P(ncp_L \leq ncp \leq ncp_U) = 1\text{-}\alpha$,

so that (ncp_L, ncp_U) is the exact $100(1\text{-}\alpha)\%$ *CI* for ncp.

The TNONCT(x, v, P) function of the SAS system (SAS Institute Inc., 2002-2003) calculates the non-centrality parameter value such that $P(T_v(ncp) \leq x) = P$ and can be used to determine ncp_L and ncp_U. *CI* boundaries for δ are then

$$\delta_L = \sqrt{\frac{n_1+n_2}{n_1 n_2}} ncp_L \quad \text{and} \quad \delta_U = \sqrt{\frac{n_1+n_2}{n_1 n_2}} ncp_U \,. \tag{A1}$$

The SAS-program **VI_delta** (available on the web page of this manual) makes use of this method.

2. Two dependent groups:

To determine an exact *CI* for δ_D, let

$$T_\nu\left(ncp\right) = \frac{\overline{x}_D - \mu_D}{s_D / \sqrt{n}} \quad , \quad \nu = n\text{-}1 \quad \text{and} \quad ncp = \delta_D \sqrt{n} \quad ,$$

so that

$$\delta_{DO} = ncp_L / \sqrt{n} \quad \text{and} \quad \delta_{DO} = ncp_U / \sqrt{n} \quad . \tag{A2}$$

The SAS-program **VI_delta_D** (available on the web page of this manual) makes use of this method.

3. Contrasts

If a contrast of means $\psi = \sum_{i=1}^{k} c_i \mu_i$ is estimated by $\hat{\psi} = \sum_{i=1}^{k} c_i \overline{x}_i$, where $\overline{x}_i$ is the mean of a random sample drawn from a normal population with mean μ_i and SD σ, then $t_{\hat{\psi}} = \hat{\psi} / s_{\hat{\psi}}$ (where $s_{\hat{\psi}}$ is the standard error of $\hat{\psi}$ as defined in (6.30)) follows a non-central t-distribution with n_m-m degrees of freedom and non-centrality parameter ncp_ψ (as defined in (6.41)). The quantity n_m is the total number of observations of the m samples involved with $\hat{\psi}$. The exact $\left(1-\alpha\right)100\%$ *CI* for ncp_ψ is obtained in exactly the same way as for ncp in Appendix A.2 above, but now $T_\nu\left(ncp\right)$ is replaced with $t_{\hat{\psi}}$, $\nu = n_m$-m, and $ncp = ncp_\psi$.

The SAS programs *VI_delta_Kontras* and *VI_delta_Kontras_D* (available on the web page of this manual) make use of this method.

APPENDIX B
Estimation and confidence intervals of the Mahalanobis D (see Steyn & Ellis, 2009)

For two populations with sizes N_A, N_B and $N = N_A + N_B$, then

$$D^2 = \frac{N^2}{N_A N_B}\frac{1-\Lambda}{\Lambda} \qquad \text{(B1)}$$

Further, according to Johnson & Kotz (1985: 177), we find that if the two populations follow an *m*-variable normal distribution, then

$\frac{1-\hat{\Lambda}}{\hat{\Lambda}} = T^2/(n-2) = \hat{U}^{(1)}$ is approximately $c(\gamma)F$ distributed, where

F is a central $F_{a(\gamma),b(\gamma)}$ -distribution. Taking $\gamma = \frac{n_A n_B}{n} D^2$ and $n = n_A + n_B$,

$$\mathrm{g} = (1+2\gamma/\mathrm{m})/(1+\gamma/\mathrm{m}),$$

$$\mathrm{h} = (1+\gamma/\mathrm{m})^2/(1+2\gamma/\mathrm{m}),$$

$$\ell = \mathrm{n}-\mathrm{p}-3,$$

and then define

$$\mathrm{a}(\gamma) = \mathrm{mh},\ \text{and}$$

$$\mathrm{b}(\gamma) = 4+(\mathrm{mh}+2)/(\mathrm{B}-1),\ \text{where}$$

$$\mathrm{B} = \frac{(\ell+\mathrm{h})(\ell+\mathrm{m})}{(\ell-2)(\ell+1)}$$

and $\mathrm{c}(\gamma) = \mathrm{mgh}\left(\mathrm{b}-2\right)/\left(\mathrm{b}\ell\right)$.

Further, $E\left(\left(1-\hat{\Lambda}\right)/\hat{\Lambda}\right) \doteq \frac{m+\gamma}{n-m-3}$, so that $\hat{\gamma} \doteq \left(n-m-3\right)\frac{1-\hat{\Lambda}}{\hat{\Lambda}} - m$, (B2)

and equation (7.7) Chapter 7 follows.

The values of γ_{L1} and γ_{U1} can be chosen in such a way that

$$P\left(F_{a\left(\gamma_{L1}\right),b\left(\gamma_{L1}\right)} \geq \hat{U}^{(1)} / c\left(\gamma_{L1}\right) \right) = \alpha/2$$

and

$$P\left(F_{a\left(\gamma_{U1}\right),b\left(\gamma_{U1}\right)} \leq \hat{U}^{(1)} / c\left(\gamma_{U1}\right) \right) = \alpha/2, \quad \text{so that}$$

$P\left(\gamma_L < \gamma < \gamma_U\right) = 1-\alpha$, which is the $\left(1-\alpha\right)100\%$ *CI* for γ .

For D is the approximate *CI* is thus:

$$\left(\sqrt{\frac{n}{n_A n_B}\gamma_L} \; ; \; \sqrt{\frac{n}{n_A n_B}\gamma_U} \right) \qquad \text{(B3)}$$

The same SAS-program as in Appendix C.1, namely *VI_zeta_kwadr1* (available on the web page of the manual) can be used.
Also see Zou (2007), where a slightly different approach is adopted to obtain the non-central *F*.

It can happen that no appropriate values for γ_L can be found because $\left(1-\hat{\Lambda}\right)/c\hat{\Lambda}$ is too small. In these cases the lower bound in (B3) is set to zero.

APPENDIX C

Estimation and confidence intervals for the ζ^2 index based on the Hotelling-Lawley-Statistic for the m variable MANOVA (see Steyn & Ellis, 2009)

For k populations, each containing m variables, the index ζ^2 is defined as

$$\zeta^2 = \frac{U^{(s)}}{s + U^{(s)}} \quad ,$$

with $U^{(s)} = \sum_{i=1}^{s} \lambda_i = trace\left(\sum_{\mu} \sum^{-1} \right)$, where $\sum_{\mu}$ is the multivariate matrix analogoue of σ_{μ}^2, (the between-population variance) and $\sum$ is the multivariate matrix analogue of σ^2 (within-population variance). The λ-values are the characteristic roots of the matrix $\sum_{\mu} \sum^{-1}$ while $s = min(k-1, m)$.

The Hotelling-Lawley-statistic $\hat{U}^{(s)}$ is used in MANOVA to test the equality of the mean vectors, $\mu_1, ..., \mu_k$, of k populations and is defined as

$$\hat{U}^{(s)} = \sum_{i=1}^{s} \hat{\lambda}_i = trace\left(\boldsymbol{H}\boldsymbol{E}^{-1} \right) \quad ,$$

where ***H*** and ***E*** are the 'between' and 'within' sample sum of squares som matrices, which are the analogues of the between and within sample sum of squares in the univariate case.

C.1 Approximate estimator and confidence interval:

According to Betz (1987: 3172), if the k populations are m-variable multivariate normally distributed, then $\hat{U}^{(s)}/c(\gamma)$ is approximately $F_{a(\gamma),b(\gamma)}$-distributed, where γ is defined as:

$$\gamma = nU^{(s)} \quad \text{and}$$

$$a(\gamma) = mh \quad \text{with} \quad h = \frac{(k-1+\gamma/m)^2}{k\text{-}1+2\gamma/m} \quad ,$$

and

$$b(\gamma) = 4+(mh+2)/(B-1) \quad \text{with} \quad B = \frac{(\ell+h)(\ell+m)}{(\ell\text{-}2)(\ell+1)} \quad ,$$

where

$$\ell = n-k-m-1 \ .$$

Further,

$$c(\gamma) = mgh(b-2)/(b\ell) \quad \text{with} \quad g = \frac{k-1+2\gamma/m}{k-1+\gamma/m} \ .$$

This approximation with a F-distribution is, according to Betz, even good for small samples.

The following approximate result is

$$E\left(\hat{U}^{(s)}\right) \doteq \frac{c(\gamma)b(\gamma)}{b-2} = \frac{mgh}{\ell} = \frac{m(k-1)+\gamma}{n\text{-}k\text{-}m\text{-}1} \quad ,$$

so that a approximate-unbiased estimator for γ is

$$\hat{\gamma}_1 = (n-k-m-1)\hat{U}^{(s)} - m(k-1) \ . \qquad \text{(C1)}$$

Because $\zeta^2 = \frac{\gamma}{ns+\gamma}$, an approximately unbiased estimator for it is:

$$\hat{\zeta}_1^2 = \frac{(n\text{-}k\text{-}m\text{-}1)\hat{U}^{(s)} - m(k-1)}{ns+(n\text{-}k\text{-}m\text{-}1)\hat{U}^{(s)} - m(k-1)} \ . \qquad \text{(C2)}$$

As in Appendix B γ_{L1} and γ_{U1} can be chosen in such a way that

$$P\left(F_{a(\gamma_{L1}),b(\gamma_{L1})} \geq \hat{U}^{(s)} / c(\gamma_{L1}) \right) = \alpha/2$$

and

$$P\left(F_{a(\gamma_{U1}),b(\gamma_{U1})} \leq \hat{U}^{(s)} / c(\gamma_{U1}) \right) = \alpha/2, \quad \text{so that}$$

$P(\gamma_{L1} < \gamma < \gamma_{U1}) = 1-\alpha$, which is an approximate $(1-\alpha)100\%$ *CI* for $\gamma = nU^{(s)}$.

The approximate *CI* for ζ^2 is then:

$$\left(\frac{\gamma_{L1}}{ns+\gamma_{L1}}; \frac{\gamma_{U1}}{ns+\gamma_{U1}} \right) \quad . \tag{C3}$$

The SAS-program *VI_zeta_kwadr1* (available on the web page of the manual) makes use of this method.

If $\hat{U}^{(s)} / c$ is too small, $\hat{\zeta}_1^2$ can be negative and the lower bound γ_O can not be established. In this situation we set $\hat{\zeta}_1^2$ = 0 and the lower bound for (C3) is then 0.

C.2 Asymptotic estimator and confidence interval:

According to Seber (1984: 39) then, asymptotically, i.e., as $n \to \infty$, the quantity $(n-k)\hat{U}^{(s)}$ follows a non-central Chi-squared distribution $X_\nu^{2\prime}$ with non-centrality parameter $\gamma = nU^{(s)}$ and $\nu = (k-1)m$ degrees of freedom.

Here the asymptotically unbiased estimator for ζ^2 is:

$$E\left((n-k)\hat{U}^{(s)}\right) \doteq m(k-1)+\gamma \quad , \text{ so that}$$

$$\hat{\gamma}_2 = (n-k)\hat{U}^{(s)} - m(k-1)$$

an unbiased estimator for γ and for ζ^2 is:

$$\hat{\zeta}_2^2 = \frac{(n\text{-}k)\hat{U}^{(s)} - m(k-1)}{ns + (n\text{-}k)\hat{U}^{(s)} - m(k-1)} \qquad \text{(C4)}$$

As in C.1 above, γ_{02} and γ_{B2} can be chosen such that

$$P\left(X_v^{2\prime}\left(\gamma_{L2}\right) \geq (n-k)\hat{U}^{(s)}\right) = \alpha/2$$

and

$$P\left(X_v^{2\prime}\left(\gamma_{U2}\right) \leq (n-k)\hat{U}^{(s)}\right) = \alpha/2 \quad , \text{ so that}$$

$(\gamma_{L2}\ , \gamma_{U2})$ is an asymptotic $(1-\alpha)100\%$ *CI* for γ.

The CNONCT (x,v,P) function in SAS system (SAS Institute Inc., 2003) provides the value of the non-centrality parameter such that $P\left(X_v^{2\prime}(ncp) \leq x\right) = P$ and can be used to determine γ_{L2} and γ_{U2}.

If n is large then we have the following approximate interval

$$\left(\frac{\gamma_{L2}}{ns+\gamma_{L2}}; \frac{\gamma_{U2}}{ns+\gamma_{U2}}\right)$$

which is the *CI* for ζ^2.

The SAS-program *VI_zeta_kwadr2* (available on the web page of this manual) makes use of this method.

As with previous intervals, if $\hat{U}^{(s)}$ is too small then $\hat{\zeta}_2^2$ can be negative and the lower bound γ_{O2} can not be determined. In these cases we set γ_{02} = 0 and $\hat{\zeta}_2^2$ = 0.