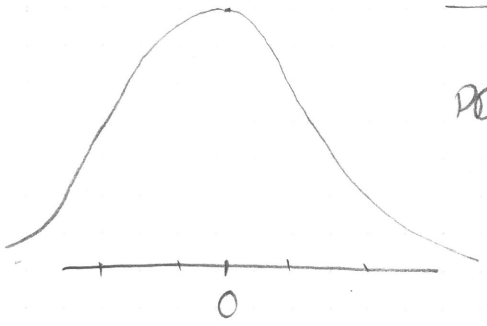
T^2 test Statistic

A classic problem is to test that a population mean is equal to a given value μ.

In the univariate case, the t-statistic $t = \frac{\sqrt{n}(\bar{x} - \mu)}{s}$ is used where $\bar{x}$ is the sample mean, s is the sample standard deviation.

When $x_1, x_2, \ldots, x_n$ drawn from $N(0, \sigma^2)$, t has a student-t distribution with $N = n-1$ degrees of freedom. n = sample size.



PDF $f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}(1 + x^2/\nu)^{-\frac{\nu+1}{2}}$

Mean 0 Median 0 Variance: $\begin{cases} \frac{\nu}{\nu-2} & \nu > 2 \\ \infty & 1 < \nu \leq 2 \end{cases}$

ν: degrees of freedom.

See Wikipedia

In the multidimensional case, Hotelling proposed his "T^2 statistic"

$$T^2 = n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu)$$

when $x_1, \ldots, x_n \sim N(\mu_0, \Sigma)$ the distribution is known exactly.

Proposition: Let $x_1, x_2, \ldots x_n \sim N_p(\mu_0, \Sigma)$ then the distribution of

$$\frac{(n-p)T^2}{pN}$$

is non-central F distributed with p and $n-p$ degrees of freedom

and non-centrality parameter $\Delta = n(\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)$

If $\mu = \mu_0$ then the F-distribution is central $(\Delta = 0)$.

The more interesting problem is: We have two sets of observations
$x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_m$

Do these sets have different (population) means?

This is called a <u>two-sample</u> problem.

Suppose $x_1, \dots, x_{n_1} \sim N_p(\mu_1, \Sigma)$ and $y_1, \dots, y_{n_2} \sim N_p(\mu_2, \Sigma)$.
The null hypothesis is $\mu_1 = \mu_2$

The sample means $\bar{x}$ and $\bar{y}$ are distributed like $N_p(\mu_i, n_i^{-1}\Sigma)$
for $i = 1, 2$.

Hence $\bar{x} - \bar{y} \sim N_p(0, \tau\Sigma)$ $\quad \tau = \frac{1}{n_1} + \frac{1}{n_2}$

$$S := \frac{1}{n_1 + n_2 - 2} \left[\sum_{k=1}^{n_1} (x_k - \bar{x})(x_k - \bar{x})' + \sum_{k=1}^{n_2} (y_k - \bar{y})(y_k - \bar{y})' \right]$$

Then $(n_1 + n_2 - 2)S$ is distributed like $\sum_{k=1}^{n_1+n_2-2} Z_k Z_k'$

$$Z_k \sim N_p(0, \Sigma)$$

So $T^2 = \frac{n_1 n_2}{n_1+n_2} (\bar{X} - \bar{Y})' S^{-1} (\bar{X} - \bar{Y})$

is distributed like the assumptions of the Proposition

$\Rightarrow T^2$ has n_1+n_2-2 degrees of freedom

and $\frac{(n_1+n_2-p-1) T^2}{p(n_1+n_2-2)} \sim F$ with degrees of freedom p and n_1+n_2-p-1.

This T^2 statistic is optimal when populations are Normal and $p \ll n$.

Bai - Sarandasa Test [1996] See Section 4.

Consider a two sample test

$H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$.

Assumptions:

1) $X_{ij} = \Gamma Z_{ij} + \mu_j \quad i = 1, \dots, n_j \quad j = 1, 2$.

s.t. Γ is $p \times m$ matrix $(m \geq p)$ $\Gamma\Gamma' = \Sigma$

Z_{ij} m-dimensional iid random vectors such that $Z_{ij} = (Z_{ijk})$

$\mathbb{E}[Z_{ij}] = 0$ $\mathrm{Var}[Z_{ij}] = I_m$ $\mathbb{E}[Z_{ijk}^4] = 3 + \Delta < \infty$

In addition,

$$\mathbb{E}\left[\prod_{k=1}^{m} z_{ijk}^{\nu_k}\right] = 0 \quad \text{(and 1)} \quad \text{when there is}$$

at least one $\nu_k = 1$ (two ν_k's equal to 2, correspondingly) whenever $\nu_1 + \nu_2 + \cdots + \nu_m = 4$.

2) $p/n \to y > 0$ and $n_1/n \to \kappa \in (0,1)$ where $n = n_1 + n_2$.

3) $\mu'\Sigma\mu = o(\tau\, tr\Sigma^2)$ and $\lambda_{max} = o\left(\sqrt{tr\Sigma^2}\right)$

where $\tau = \frac{1}{n_1} + \frac{1}{n_2}$

Consider the statistic.

$$M_n = \|\bar{x}_1 - \bar{x}_2\| - \tau\, tr S_n.$$

where

$$S_n = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \qquad \bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

Under the null hypothesis, it can be shown that

$$\mathbb{E}[M_n] = 0.$$

Under assumptions (1)-(3), it can also be shown that when $n \to \infty$ and H_0,

$$Z_n = \frac{M_n}{\sqrt{Var(M_n)}} \xrightarrow{\mathcal{D}} N(0,1) \qquad (*)$$

If we assume that the populations are Normal, then under H_0, $\text{var}\, M_n = \sigma_M^2$ where

$$\sigma_M^2 := 2\tau^2\left(1+\frac{1}{n}\right)\text{tr}\,\Sigma^2$$

Also, by (1)-(3), it can be shown that

$$\text{Var}\, M_n = \sigma_M^2\,(1+o(1))$$

Therefore, we substitute σ_M^2 for $\text{Var}(M_n)$ in (*) and the statistic Z_n remains asymptotically Normal.

The term σ_M^2 requires us to calculate $\text{tr}\,\Sigma^2$. We need to estimate this from the data. A natural choice might be to replace this term by $\text{tr}\,S_n^2$ but this is neither unbiased nor consistent.

If $nS_n \sim W(n,\Sigma)$ ie. Wishart. Then

$$B_n^2 = \frac{n^2}{(n+2)(n-1)}\left(\text{tr}\,S_n^2 - \frac{1}{n}(\text{tr}\,S_n)^2\right)$$

is an unbiased and consistent estimator of $\text{tr}\,\Sigma^2$. [Proof in paper]

Plugging this all into Z_n gives the Bai-Saranadasa test statistic

$$Z = \frac{\|\bar{X}_1 - \bar{X}_2\| + \tau \operatorname{tr} S_n}{\sqrt{\operatorname{Var}(M_n)}}$$

$$\sim \frac{\|\bar{X}_1 - \bar{X}_2\| + \tau \operatorname{tr} S_n}{\sqrt{2\tau^2\left(1+\frac{1}{n}\right)\operatorname{tr}\Sigma^2}}$$

$$\sim \frac{n_1 n_2}{n_1+n_2} \frac{\|\bar{X}_1 - \bar{X}_2\| - \tau \operatorname{tr} S_n}{\sqrt{\frac{2(n+1)}{n} B_n}} \xrightarrow{\mathcal{D}} N(0,1)$$

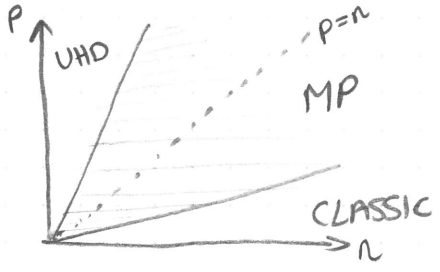
Therefore the test rejects H_0 if $Z > z_\alpha$

Chen - Qin Test [2010]

Bai - Saranadasa requires finite fourth moment and $p/n \to y < \infty$

Chen & Qin propose a different test statistic that also works in the "ultra high dimensional" case. when

UHD : $n \to \infty$, $p/n \to \infty$.



The UHD regime is common in genetic studies such as testing for gene expression.

Their paper starts with an interesting analysis of Hotelling and Bai-Saranadasa, that we repeat here.
[BS]

1) [BS] proposed replacing the term $(\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2)$ in Hotelling's test by the term $\|\bar{X}_1 - \bar{X}_2\| = (\bar{X}_1 - \bar{X}_2)'(\bar{X}_1 - \bar{X}_2)$ where S was the pooled covariance matrix by assuming $\Sigma_1 = \Sigma_2 = \Sigma$.

2) Hotelling's T^2 test works well in the case where the data dimension p is small and fixed. We need $p \ll n := n_1 + n_2 - 2$.

3) Hotelling's test has decreasing power (ie. $P(\text{reject } H_0 \mid H_1 \text{ true}) \searrow 0$) as $p/n = y \in [0,1)$ gets larger.

A reason for this effect is that the statistic requires the inverse of the sample covariance matrix. Standardising by the covariance has benefits for fixed p but becomes a liability when p becomes large.

$$S \not\to \Sigma \qquad p, n \to \infty \qquad p/n \to y \in (0,1)$$

$$\text{As} \quad \lambda_1^S \not\to \lambda_1^\Sigma \quad \& \quad \lambda_{max}^S \not\to \lambda_{max}^\Sigma$$

4) The key idea in [BS] is to remove S^{-1} as it is no longer beneficial when $p/n \to y > 0$. To compensate they subtracted $tr(S_n)$ so that $\mathbb{E}[M_n] = \|\mu_1 - \mu_2\|^2$

5) Carefully looking at the M_n term in [BS], we see that it requires terms of the form

$$\sum_{j=1}^{n_i} X_{ij}' X_{ij} \qquad i=1,2.$$

but these terms are not needed and we can consider

$$T_n = \frac{\sum_{i\neq j}^{n_1} X_{1j}' X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}' X_{2i}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}' X_{2j}}{n_1 n_2}$$

instead of M_n. And $\mathbb{E}[T_n] = \|\mu_1 - \mu_2\|^2$

Further $\operatorname{tr}(S_n)$ is no longer needed which itself imposes demands on the dimensionality.

The setup is similar to [BS]:

$$X_{ij} = \Gamma_i Z_{ij} + \mu_i \qquad j=1,\cdots,n_i \qquad i=1,2.$$

Γ_i is $p\times m$ matrix $(m\geq p)$ s.t. $\Gamma_i \Gamma_i' = \Sigma_i$

$Z_{ij} = (Z_{ijk})$ m-dimensional iid random vectors.

$$\mathbb{E}Z_{ij} = 0 \quad \operatorname{Var}[Z_{ij}] = I_m \quad \mathbb{E}[Z_{ijk}^4] = 3+\Delta < \infty$$

s.t. $\mathbb{E}[Z_{ij\ell_1}^{\alpha_1} Z_{ij\ell_2}^{\alpha_2} \cdots Z_{ij\ell_q}^{\alpha_q}] = \mathbb{E}[Z_{ij\ell_1}^{\alpha_1}]\mathbb{E}[Z_{ij\ell_2}^{\alpha_2}]\cdots\mathbb{E}[Z_{ij\ell_q}^{\alpha_q}]$

for integers $\alpha_1,\cdots,\alpha_q$ satisfying $\sum_{\ell=1}^q \alpha_\ell \leq 8$ and $\ell_1 \neq \ell_2 \neq \cdots \neq \ell_q$.

(This holds if coordinates are independent).

Notice that a requirement that $\Sigma_1 = \Sigma_2 = \Sigma$ is not needed. This is useful as checking this in high dimensions is not easy.

The precise assumptions are:

(A) $n \to \infty$, $n_1/(n_1+n_2) \to k \in (0,1)$

(B) $\mu_i' \Sigma_i \mu_i = o\left(n^{-1} \operatorname{tr}\left[(\Sigma_1+\Sigma_2)^2\right]\right)$ $i=1,2$.

(B) is true under $H_0: \mu_1 = \mu_2$.

Theorem (Chen, Qin 2010)

$$\frac{T_n - \|\mu_1 - \mu_2\|^2}{\sqrt{\operatorname{Var}(T_n)}} \xrightarrow{\mathcal{D}} N(0,1)$$

∎

See workshop on how these tests are applied.