Tests Concerning two Univariate Normal populations

Suppose, we have two populations characterised by $N(\mu_1, \sigma_1^2)$ & $N(\mu_2, \sigma_2^2)$ distributions respectively.

Let us draw two random samples of sizes n_1 & n_2 independently from the ~~first~~ population. Let, $x_{11}, x_{12}, \ldots, x_{1n_1}$ be the observations of the first sample &

$x_{21}, x_{22}, \ldots, x_{2n_2}$ be the observations from the second sample.

Let us consider,

$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ [mean of the i^{th} sample], $i=1,2$.

$s_{i0}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)^2$ [i^{th} sample variance when, μ_i is known], $i=1,2$.

$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ [i^{th} sample variance when, 'μ_i' is unknown], $i=1,2$.

One may, then, consider the following tests:

(1) To test $H_0 : \mu_1 - \mu_2 = \xi_0$

against $H_{11} : \mu_1 - \mu_2 > \xi_0$

$H_{12} : \mu_1 - \mu_2 < \xi_0$

$H_{13} : \mu_1 - \mu_2 \neq \xi_0$,

where, 'ξ_0' is the ~~specie~~ specified value of $(\mu_1 - \mu_2)$.

Since the hypothesis concerns the population means, common sense suggests that the test should be based on the sample means.

Case - I: σ_1 & σ_2 are known

We first consider the difference $(\bar{x}_1 - \bar{x}_2)$. Observe that

$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$; $V(\bar{x}_1 - \bar{x}_2) = V(\bar{x}_1) + V(\bar{x}_2)$ [$\because$ the two samples are drawn independently]

$= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$ will follow a standard normal distribution.

[$(\bar{x}_1 - \bar{x}_2)$, being a linear function of normal variates, is itself normally distributed].

As such, we may take as our test statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}, \quad \overset{H_0}{\sim} N(0,1).$$

The test procedure is given as follows:

a) We reject H_0 against H_{11} iff $t > \tau_\alpha$, where τ_α is the upper α-point of a standard normal distribution.

b) We reject H_0 against H_{12} iff $t < -\tau_\alpha$, $-\tau_\alpha$ being the lower α-point of a standard normal distribution.

c) We reject H_0 against H_{13} iff $t > \tau_{\alpha/2}$ or $t < -\tau_{\alpha/2}$ ie iff $|t| > \tau_{\alpha/2}$, $\tau_{\alpha/2}$ being the upper $\frac{\alpha}{2}$-point of a standard normal distribution.

Here, α is the desired level of significance.

Here, ... level ... significance.

Case - II: σ_1 & σ_2 are unknown [Fisher's 't'- dist^n]

Since, σ_1^2 & σ_2^2 are unknown, we replace them by their unbiased estimators s_1^2 & s_2^2 respectively & consider

$$\frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

as our test statistic. However, this does not have a a simple distribution. In order to avoid complications, we shall assume that σ_1^2 & σ_2^2, although unknown individually, are known to be equal.

We, thus, make the so called homo-scedasticity assumption: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say.

We now have, $V(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$,

Since, the two samples are drawn independently.

Further, we replace σ^2 by the pooled estimator

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}{n_1+n_2-2}.$$

We, thus, arrive at our test statistic $t = \frac{(\bar{x}_1 - \bar{x}_2) - \xi_0}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$.

Observe that 't' may be written as

$$\frac{\bar{x}_1 - \bar{x}_2 - \xi_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \Bigg/ \sqrt{\frac{(n_1+n_2-2)s^2}{\sigma^2} \Big/ n_1+n_2-2}$$

As such, under H_0, 't' is of the form $\frac{\tau}{\sqrt{\chi^2_{n_1+n_2-2}/(n_1+n_2-2)}}$,

where, τ & $\chi^2_{n_1+n_2-2}$ are independently distributed.

Thus, under H_0 't' follows a 't' distribution with (n_1+n_2-2) d.f.

The test procedure is then given as below:

a) We reject H_0 against H_{11} iff $t > t_{\alpha; n_1+n_2-2}$; $t_{\alpha; n_1+n_2-2}$ being the upper α-point of a 't' distribution with (n_1+n_2-2) d.f.

b) We reject H_0 against H_{12} iff $t < -t_{\alpha, n_1+n_2-2}$, $-t_{\alpha, n_1+n_2-2}$ being the lower α-point of a 't' distribution with (n_1+n_2-2) d.f.

c) We reject H_0 against H_{13} iff $t > t_{\frac{\alpha}{2}; n_1+n_2-2}$ & $t < -t_{\frac{\alpha}{2}; n_1+n_2-2}$ ie iff $|t| > t_{\alpha/2; n_1+n_2-2}$ where, $t_{\alpha/2, n_1+n_2-2}$ is the upper $\frac{\alpha}{2}$-point of a t-distribution with (n_1+n_2-2) d.f.

Here, ~~in each case~~, α is the desired ~~let~~ level of significance.

Here, in each case, the t-test performed is called a Fisher's 't'-test.

(*)

Remark: σ_1^2 & σ_2^2 are supposed to be unequal.

Fisher-Behran Problem:

We shall consider the situation of homo-scedasticity is not valid. Here the problem of hypothesis testing or of interval estimation become some what difficult to handle. There are a no. of procedures which can be suggested but none of them can be said to be completely satisfactory.

Procedure ①: [Based on period t-statistic].

Neyman's Approach: Suppose, w.L.G, $n_1 \leq n_2$.
Neyman's approach is that we get n_1 pairs of observations by pairing each obs. of the 1st sample with some obs. of the 2nd, where, in considering any later obs. of the 1st sample, we reject the obs. of the 2nd sample that have already been paired. In this way, of course $(n_2 - n_1)$ obs. of the 2nd sample will be left out. Then, the n_1 pairs of obs, thus formed, are looked upon as a sample of size n_1 from a bivariate normal popl^n with mean μ_1 & μ_2; variances σ_1^2 & σ_2^2 & Correlation Coefficient $\rho \neq 0$. Suppose, x_{1j} is paired with $x_{2j'}$ where $x_{2j'}$ is one of the values of $x_{21}, x_{22}, \ldots, x_{2n_2}$.

x_{1j}	x_{2j}
x_{11}	x_{21}
x_{12}	x_{22}
$\vdots$	$\vdots$
x_{1n_1}	$\vdots$
	x_{2n_2}

If we put $z_j = x_{1j} - x_{2j'}$, $(j = 1, 2, \ldots, n_1,\ j' = 1, 2, \ldots, n_2)$, then, a test or a confidence interval to $\mu_1 - \mu_2$ may then be obtained by using the statistic

$\dfrac{\sqrt{n_1}(\bar{z}-\mu_z)}{s_z}$, where $\bar{z} = \sum_{j=1}^{n_1} z_j / n_1$,

$s_z^2 = \frac{1}{n-1} \sum_{j=1}^{n_1} (z_j - \bar{z})^2 \overset{H_0}{\sim} t_{n_1-1}$. $H_0: \mu_1-\mu_2 = \mu_z$.

This procedure is the simplest. However it has two drawbacks

i) The test is not based on all the observations since we exclude some obs. of the 2nd sample. So if we include some other set of n_2 obs., the decision regarding acceptance may be different.

ii) The result of the test of the C.I. ultimately obtained will not depend ~~a~~ solely on the nature of the data but will depend also on the way the obs. of the 1st sample is paired with those of the 2nd.

• Procedure → 2: Cohran & Cox's Approximate test:

Cohran & Cox recomanded the use of the statistic $t' = \dfrac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\sigma_2^2/n_2}}$ which under H_0 will be $\dfrac{(\bar{x}_1-\bar{x}_2)-\xi_0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$

However since the exact distn. of the statistic t' is complicated & will exactly depend on the ratio of the two popln var's, i.e. σ_1^2 & σ_2^2, they suggest that we should take as the upper α-point of t' under H_0, say t'_α, the value $\dfrac{w_1 t_{\alpha,n_1-1} + w_2 t_{\alpha,n_2-1}}{w_1+w_2}$, $w_i = \dfrac{\sigma_i^2}{n_i}$, ~~where~~

~~$\dfrac{w_1 t_{\alpha,n_1-1} + w_2 t_{\alpha,n_2-1}}{w_1+w_2}$~~ ~~$w_i =$~~ Test procedure

① Reject $H_0: \mu_1-\mu_2=\xi_0$ ag. $H_1: \mu_1-\mu_2 > \xi_0$ if $t' > \dfrac{w_1 t_{\alpha,n_1-1} + w_2 t_{\alpha,n_2-1}}{w_1+w_2}$

(2) To test $H_0 : \frac{\sigma_1}{\sigma_2} = \xi_0$ [or, $H_0' : \frac{\sigma_1^2}{\sigma_2^2} = \xi_0^2$].

Against $H_{11} : \frac{\sigma_1}{\sigma_2} > \xi_0$ [or, $H_{11}' : \frac{\sigma_1^2}{\sigma_2^2} > \xi_0^2$].

$H_{12} : \frac{\sigma_1}{\sigma_2} < \xi_0$ [or, $H_{12}' : \frac{\sigma_1^2}{\sigma_2^2} < \xi_0^2$].

$H_{13} : \frac{\sigma_1}{\sigma_2} \neq \xi_0$ [or, $H_{13}' : \frac{\sigma_1^2}{\sigma_2^2} \neq \xi_0^2$].

Since, the hypothesis concerns ~~the~~ the population variances, common sense suggests that the test should be based on the sample variances.

Case-I: μ_1 & μ_2 are known

Here, the sample variances are appropriately taken as s_{10}^2 & s_{20}^2 respectively.
observe that,

$$\frac{s_{10}^2/\sigma_1^2}{s_{20}^2/\sigma_2^2} = \frac{s_{10}^2}{s_{20}^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = \frac{\sum_{j=1}^{n_1}\left(\frac{x_{1j}-\mu_1}{\sigma_1}\right)^2/n_1}{\sum_{j=1}^{n_2}\left(\frac{x_{2j}-\mu_2}{\sigma_2}\right)^2/n_2}$$

which is of the form $\frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$.

Also the two χ^2's are independently distributed. ~~Hence, under H_0,~~
Hence, we may take as our test statistic $F = \frac{s_{10}^2}{s_{20}^2} \cdot \frac{1}{\xi_0^2}$, which, under H_0 follows an F-distribution with d.f. (n_1, n_2).
The test procedure is given as follows:

a) ~~Reje~~ We reject H_0 against H_{11} iff $F > F_{\alpha; n_1, n_2}$, where, $F_{\alpha; n_1, n_2}$ is the upper α-point of an F distribution with (n_1, n_2) d.f.

b) We reject H_0 against H_{12} iff $F < F_{1-\alpha; n_1, n_2}$, where, $F_{1-\alpha; n_1, n_2}$ is the lower α-point of an F distribution with (n_1, n_2) d.f.

c) We reject H_0 against H_{13} iff $F > F_{\alpha/2; n_1, n_2}$ & $F < F_{(1-\alpha)/2; n_1, n_2}$,
where, $F_{\alpha/2; n_1, n_2}$ & $F_{\frac{1-\alpha}{2}; n_1, n_2}$ ~~is the~~ are, respectively, the upper & lower α-point of the F distribution with (n_1, n_2) d.f.
Here, α is the desired level of significance.

152 Case → II : μ_1 & μ_2 are unknown

Here, the sample variances are approximately taken as s_1^2 & s_2^2 respectively. Observe that

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} = \frac{\sum_{j=1}^{n_1}\left(\frac{x_{1j}-\bar{x}_1}{\sigma_1}\right)^2/(n_1-1)}{\sum_{j=1}^{n_2}\left(\frac{x_{2j}-\bar{x}_2}{\sigma_2}\right)^2/(n_2-1)},$$

which is of the form $\frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)}$.

Also, the two χ^2's are independently distributed. Hence, we may take as our test statistic, $F = \frac{s_1^2}{s_2^2} \cdot \frac{1}{\xi_0^2}$, which under H_0 follows F distribution with (n_1-1, n_2-1) d.f.

The test procedure is given as below:

a) We reject H_0 against H_{11} iff $F > F_{\alpha; n_1-1, n_2-1}$, $F_{\alpha; n_1-1, n_2-1}$ being the upper α-point of a F distribution with (n_1-1, n_2-1) d.f.

b) We reject H_0 against H_{12} iff $F < F_{1-\alpha; n_1-1, n_2-1}$ i.e. $F < \frac{1}{F_{\alpha; n_1-1, n_2-1}}$, $\frac{1}{F_{\alpha; n_1-1, n_2-1}}$ being the lower α-point of a 'F' distribution with (n_1-1, n_2-1) d.f.

c) We reject H_0 against H_{13} iff $F > F_{\frac{\alpha}{2}; n_1-1, n_2-1}$ & $F < F_{1-\frac{\alpha}{2}; n_1-1, n_2-1}$ i.e. $F < \frac{1}{F_{\frac{\alpha}{2}; n_1-1, n_2-1}}$, where, $F_{\alpha/2; n_1-1, n_2-1}$ & $\frac{1}{F_{\alpha/2; n_1-1, n_2-1}}$ are, respectively, the upper & lower $\frac{\alpha}{2}$-point of a F dist^n with (n_1-1, n_2-1) d.f.

Here, α is the desired level of significance.