Introduction to Statistical Method

Comparing Two Means and Two Variances

Comparing Two Means - A Point Estimator

We have two populations with different means μ_1 and μ_2, the goal is to estimate the difference $\mu_1 - \mu_2$ by taking a sample from each population in independent way.

Natural point estimator: $\mu_1 \hat{-} \mu_2 := \hat{\mu_1} - \hat{\mu_2} = \overline{X}_1 - \overline{X}_2$

To determine confidence intervals and to test hypothesis we need to know the distribution $\overline{X}_1 - \overline{X}_2$

Theorem

The $\overline{X}_1$ and $\overline{X}_2$ be the sample means based on independent random samples of size n_1 and n_2 drawn from normal distributions with mean μ_1 and μ_2 and variance σ_1^2 and σ_2^2.

The $\overline{X}_1 - \overline{X}_2$ is normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$

$\dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ is standard normal random variable.

(Central Limit Theorem allows us to apply this result even to non-normal populations if we have really large sample sizes)

OC Curve Application

$d = \dfrac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$, if $n = n_1 = n_2$, unchanged, else $n = \dfrac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

> If the variances are unknown, we need some more sophisticated methods.
>
> The unknown variances are equal, situation is much easier.

Comparing Two Variances

Consider test types of this:

- $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2$ (right-tailed test)
- $H_0 : \sigma_1^2 = \sigma_2^2$, $H_1 : \sigma_1^2 \neq \sigma_2^2$ (two-tailed test)

we move the σ to one side to consider only about the quotient.

> $(n-1)S^2/\sigma^2$ follows a chi-squared distribution with $n-1$ degree of freedom.
>
> If the variance is put into quotient, it is easier to handle

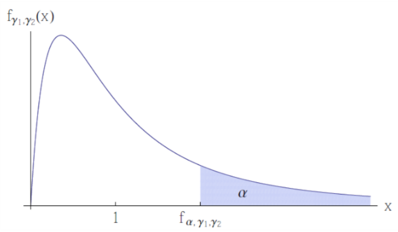
F-Distribution

$X_{\gamma_1}^2$ and $X_{\gamma_2}^2$ are independent chi-squared random variables with γ_1 and γ_2 degrees of freedom.

Random variable $F_{\gamma_1,\gamma_2} = \dfrac{X_{\gamma_1}^2/\gamma_1}{X_{\gamma_2}^2/\gamma_2}$ is said to follow a F-distribution with γ_1 and γ_2 degree of freedom.

$P[F_{\gamma_1,\gamma_2} < x] = P[1/F_{\gamma_1,\gamma_2} > 1/x] = 1 - P[F_{\gamma_2,\gamma_1} < 1/x]$

Then $f_{\gamma_1,\gamma_2}(x) = \gamma_1^{\gamma_1/2}\gamma_2^{\gamma_2/2} \dfrac{\Gamma(\frac{\gamma_1+\gamma_2}{2})}{\Gamma(\frac{\gamma_1}{2})\Gamma(\frac{\gamma_2}{2})} \dfrac{x^{\gamma_1/2-1}}{(\gamma_1 x + \gamma_2)^{(\gamma_1+\gamma_2)/2}}$ for $x \geq 0$



Define $f_{\alpha,\gamma_1,\gamma_2}$ by $P[F_{\gamma_1,\gamma_2} > f_{\alpha,\gamma_1,\gamma_2}] = \alpha$

Then $\begin{aligned} 1-\alpha &= P[F_{\gamma_1,\gamma_2} \geq f_{1-\alpha,\gamma_1,\gamma_2}] \\ &= 1 - P[F_{\gamma_1,\gamma_2} < f_{1-\alpha,\gamma_1,\gamma_2}] \\ &= P[F_{\gamma_2,\gamma_1} < 1/f_{1-\alpha,\gamma_1,\gamma_2}] \\ &= 1 - P[F_{\gamma_2,\gamma_1} \geq 1/f_{1-\alpha,\gamma_1,\gamma_2}] \end{aligned}$ also we can see $\alpha = P[F_{\gamma_2,\gamma_1} \geq f_{\alpha,\gamma_2,\gamma_1}]$.

So $f_{1-\alpha,\gamma_1,\gamma_2} \cdot f_{\alpha,\gamma_2,\gamma_1} = 1$.

Remark

Let S_1^2 and S_2^2 be sample variance based on independent random samples of size n_1 and n_2 from normal populations with means μ_1 and μ_2 and variance σ_1^2 and σ_2^2.

If $\sigma_1^2 = \sigma_2^2$ then the statistic S_1^2/S_2^2 follows F-distribution with $n_1 - 1$ and $n_2 - 1$ distribution.

Since $F_{n_1-1,n_2-1} = \dfrac{[(n_1-1)S_1^2/\sigma_1^2]/(n_1-1)}{[(n_2-1)S_2^2/\sigma_2^2]/(n_2-1)} = \dfrac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$, so it is trivial to require $\sigma_1^2 = \sigma_2^2$

F-Test

We can derive F-Test from F-distribution that:

$H_0 : \sigma_1 = \sigma_2$ based on $F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2}$ is a F-Test

We reject H_0 at significance level α

- in favor of $H_1 : \sigma_1 > \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha,n_1-1,n_2-1}$
- in favor of $H_1 : \sigma_1 < \sigma_2$ if $\frac{S_2^2}{S_1^2} > f_{\alpha,n_2-1,n_1-1}$
- in favor of $H_1 : \sigma_1 \neq \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha/2,n_1-1,n_2-1}$ or $\frac{S_2^2}{S_1^2} > f_{\alpha/2,n_2-1,n_1-1}$

When testing to see whether two population variances are equal for the purpose of comparing their means, one hopes to not reject H_0.

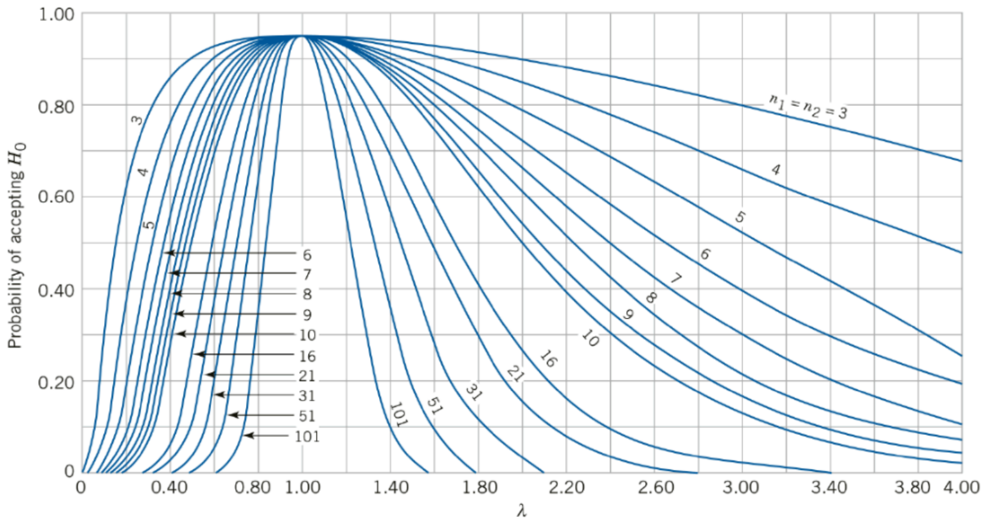
If H_0 is not rejected, one can assume that the variances are in fact equal and continue with the test for equality for means.

In this case, a small Type II error β is more important than α small.

OC Curves for F-Test

For case $n = n_1 = n_2$, the OC curves plotting β against the parameter $\lambda = \frac{\sigma_1}{\sigma_2}$.

The curves are for both one- two- sided alternatives.



Comparing Two Means - Equal Variances

$\dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ follows standard normal distribution.

We now want to estimate σ^2.

The pooled estimator is $S_p^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

$$X^2_{n_1+n_2-2} = \frac{(n + 1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

Furthermore, $T_{n_1+n_2-2} = \dfrac{Z}{\sqrt{X^2_{n_1+n_2-2}/(n_1 + n_2 - 2)}} = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$ follows T-distribution with $n_1 + n_2 - 2$ degree of freedom.

So the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is $(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2, n_1+n_2-2}\sqrt{S_p^2(1/n_1 + 1/n_2)}$

Pooled T-Test - Variance Equal

Let $X_1^{(i)} .. X_{n_i}^{(i)}, i = 1, 2$ be random samples of size n_i from two normal distributions with means μ_i and identical σ^2.

S_p^2 be the pooled sample variance and $(\mu_1 - \mu_2)_0$ a null value for difference of means.

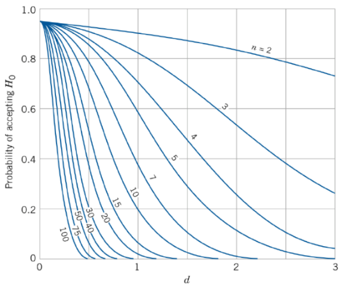
Then Test $H_0 : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ based on $T_{n_1+n_2-2} = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$ is a pooled test for equality of means.

We reject H_0 at significance level α

- in favor of $H_1 : \mu_1 - \mu_2 \neq (\mu_1 - \mu_2)_0$ if $|T_{n_1+n_2-2}| > t_{\alpha/2, n_1+n_2-2}$
- in favor of $H_1 : \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} > t_{\alpha, n_1+n_2-2}$
- in favor of $H_1 : \mu_1 - \mu_2 < (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} < -t_{\alpha, n_1+n_2-2}$

OC Curves T Test - Variance Equal

Equal variance σ^2 and equal sample size $n_1 = n_2 = n$, $d = \frac{|\mu_1 - \mu_2|}{2\sigma}$, we must use the modified sample size $n^* = 2n - 1$. The σ can be substitute with an estimated one or express the deviation in terms of σ.



Unequal Variances

$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$ for the unequal variance σ_1 and σ_2, we can estimate the variance to get the statistic $T_\gamma = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$ where the γ for the degree of freedom is $\gamma = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$

Pooled T-Test - Variances Unequal

We reject H_0 at significance level α

- in favor of $H_1 : \mu_1 - \mu_2 \neq (\mu_1 - \mu_2)_0$ if $|T_\gamma| > t_{\alpha/2,\gamma}$
- in favor of $H_1 : \mu_1 - \mu_2 > (\mu_1 - \mu_2)_0$ if $T_\gamma > t_{\alpha,\gamma}$
- in favor of $H_1 : \mu_1 - \mu_2 < (\mu_1 - \mu_2)_0$ if $T_\gamma < -t_{\alpha,\gamma}$

Paired T-Test

> In some situations, we do not take independent samples from two different populations, but rather the samples are naturally related to each other.

X and Y are two random variables, we define $D = X - Y$, then
$\mu_D = E[D] = E[X - Y] = E[X] - E[Y] = \mu_X - \mu_Y$

Then we can analyze D using method for the mean of a single random variable.

$H_0 : \mu_D = 0$ and $H_1 : \mu_D < 0$, with n for size, $\overline{D}$ and s_D^2, we derive test statistic $T = \frac{\overline{D}}{s_D/\sqrt{n}}$

Wilcoxon Rank - Sum Test

> If the sample sizes are small, the variances unequal, or the populations are not normally distributed, the T-tests may not yield good results.

The Wilcoxon Rank-sum test ranks observed measurements by size and the sum of the ranks is used to decide whether to reject the null hypothesis.

It assumes that two random variables X and Y the null hypothesis is that they follow same distribution:

$H_0 : P[X > Y] = 1/2$.

Then it is assumed that they follow continuous distributions that differs only on location.

$f_X(x) = f_Y(x + \delta)$ for all $x \in \mathbb{R}$

Let X and Y be two random samples following the same continuous distribution but with possibly shifted location. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, $m \leq n$ be random sample from X and Y and associated with rank R_i, $1 \leq i \leq m + n$ from smallest to largest.

So hypothesis changed to $H_0 : M_X = M_Y$ and $H_1 : M_X \neq M_Y$, the test is based on the statistic
$W_m := \text{ sum of the ranks of } X_1, \ldots X_m$

$E[W_m] = \frac{1}{2}m(m + n + 1)$ and $VarW_m = \frac{1}{12}mn(m + n + 1)$

The paired test, the Wilcoxon signed rank test can be applied to $D = X - Y$.