Sejam X, Y duas variáveis de interesse representando duas sub-populações. Estamos interessados em verificar se a média populacional de X é menor, maior ou igual à de Y. Sendo assim, precisamos considerar os casos em que X é independente de Y e o caso em que não são independentes (pareados).

Dados independentes e dependentes

Um pesquisador propôs um novo método de investimento para aumentar o rendimento mensal. Selecionou 20 investidores aleatoriamente de um universo de investidores cadastrados.
Em um primeiro momento, o pesquisador deixou os investidores investirem do jeito que sabem e ao final verificou a renda obtida.
X é o rendimento dos investidores sem ter o conhecimento do método.

Então, o pesquisador ensinou seu método aos investidores, onde Y passou a ser o rendimento dos investidores após a aplicação do método ensinado.

Claramente, X, Y são dependentes.

O mesmo pesquisador testará o mesmo método de forma diferente. Para testar o seu método, o pesquisador selecionou 20 indivíduos com características similares do universo de investidores, dos quais

1. 10 foram designados aleatoriamente a não receber o método X: rendimento de um indivíduo que não recebeu o método
2. 10 foram designados aleatoriamente a receberem o método Y: rendimento de um indivíduo que recebeu o método
   Nesse caso, as variáveis X, Y são independentes.

Em ambas abordagens, temos as mesmas hipóteses de interesse:

$$1. \begin{cases} H_0 : \mu_X \geq \mu_y \\ H_1 : \mu_X < \mu_Y \end{cases} \quad 2. \begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases} \quad 3. \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Podemos definir $\mu_D = \mu_X - \mu_Y$ e reescrever as hipóteses

1. $\begin{cases} H_0 : \mu_D \geq 0 \\ H_1 : \mu_D < 0 \end{cases}$ 2. $\begin{cases} H_0 : \mu_D \leq 0 \\ H_1 : \mu_D > 0 \end{cases}$ 3. $\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$

Para os próximos exemplos, assumiremos normalidade para X, Y

Caso pareado (variáveis dependentes)

No caso em que X, Y são dependentes, as amostras são (X_n) a.a de X e (Y_n) a.a de Y tais que X_i, Y_i são dependentes.

Para este caso, fazemos $D_i = X_i - Y_i, i = 1, 2, \dots, n$. Temos que (D_n) é uma amostra aleatória de
$D = X - Y \sim N(\mu_D, \sigma_D^2) = N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y)$

Observe ainda que

$$\bar{D}_{\text{Par}} = \sum_{i=1}^{n} \frac{D_i}{n} \sim N\left(\mu_D, \frac{\sigma_D^2}{n}\right)$$

Note ainda que as variância e covariância de X, Y estão embutidas em σ_D^2.
Podemos usar

$$s_D^2(\underset{\sim}{D}) = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D}_{\text{Par}})^2$$

Para estimar σ_D^2
Podemos construir as decisões como já vimos anteriormente em testes sob normalidade com variância desconhecida.

Exemplo

Foram coletados os rendimentos (em mil reais) antes e após a aplicação o método para 12 investidores. Queremos verificar se o método aumentou o rendimento médio. Chamaremos de X o rendimento anterior ao treinamento e Y o rendimento após. Isto é, queremos verificar se $\mu_D = \mu_X - \mu_Y \leq 0$. Portanto, nossa hipótese nula é de que o treinamento não tem efeito positivo no rendimento:

$$\begin{cases} H_0 : \mu_D \geq 0 \\ H_1 : \mu_D < 0 \end{cases}$$

Indíce	Antes(X)	Depois(Y)	Dif(D)	Dif$^2(D^2)$
1	2.4	4.3	-1.9	3.61
2	2.8	3.4	-0.6	0.36
3	4.6	3.2	1.4	1.96
4	3.1	3.3	-0.2	0.04
5	3.1	3.3	-0.2	$0,04$
6	4.7	5.8	-1.1	1.21
7	3.5	3.8	-0.3	0.09
8	1.7	3.5	-1.8	3.24
9	2.3	3.2	-0.9	0.81
10	2.6	3.9	-1.3	1.69
11	4.2	3.6	0.6	0.36
12	3.4	4.3	-0.9	0.81
Média	3.2	3.8	-0.6	
s^2	0.87	0.55	0.9	

Temos que $s^2(\underset{\sim}{D}) = 0.9$. (Podemos calcular diretamente ou usando a coluna D^2 e substituindo no somatório $s_D^2 = \left(\frac{\sum_{i=1}^n d_i^2}{n} - \bar{d}^2\right)\frac{n}{n-1}$)
Rejeitamos H_0 se $\frac{\bar{d}_{\text{Par}}}{\sqrt{\frac{s_D^2}{n}}} < -t_{\alpha,n-1}$ onde $t_{\alpha,n-1}$ é tal que

$P(t_{n-1} < -t_{\alpha,n-1}) = \alpha$, onde t_{n-1} é a distribuição T de Student com $n-1$ graus de liberdade.

Temos que $\frac{\bar{d}_{\text{Par}}}{\sqrt{\frac{s_D^2}{n}}} = -2.19$ e, a $\alpha = 0.05$, $-t_{0.05,11} = -1.796$.
Como $-2.19 < -1.796$, podemos dizer que há evidências de que o método aumenta o rendimento médio dos investidores a 5% de significância. Por outro lado, com $\alpha = 0.01, -t_{0.01,11} = -2.718$ e, por $-2.19 \geq -2.718$, dizemos que não há evidências para rejeitar a hipótese de que o método não aumenta o rendimento (rejeitar a hipótese nula) a 1% de significância estatística.

Caso de independência

Sejam X, Y variáveis aleatórias *independentes* tais que

$$\begin{cases} X \sim N(\mu_X, \sigma_X^2) \\ Y \sim (\mu_Y, \sigma_Y^2) \end{cases}$$

Considere (X_n) amostra aleatória de X e (Y_m). Estamos interessados em testar as hipóteses:

$$1. \begin{cases} H_0 : \mu_X \geq \mu_Y \\ H_1 : \mu_X < \mu_Y \end{cases} \quad 2. \begin{cases} H_0 : \mu_X \leq \mu_Y \\ H_1 : \mu_X > \mu_Y \end{cases} \quad 3. \begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Podemos definir $\mu_D = \mu_X - \mu_Y$ e obter a equivalência dessas hipóteses

$$1. \begin{cases} H_0 : \mu_D \geq 0 \\ H_1 : \mu_D < 0 \end{cases} 2. \begin{cases} H_0 : \mu_D \leq 0 \\ H_1 : \mu_D > 0 \end{cases} 3. \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

Considere

$$\bar{D}_{\text{NPar}} = \bar{X} - \bar{Y}$$

Como $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right), \bar{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$, temos que $\bar{D}_{\text{NPar}} \sim N\left(\mu_D, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$

Ambas variâncias conhecidas

Podemos substituir o valor numérico das variâncias na distribuição de $\bar{D}_{\text{NPar}}$, obtendo uma distribuição normal com pontos de corte para as decisões:

$$1. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -z_\alpha \underbrace{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}_{\text{Var}(\bar{D}_{\text{NPar}})}, \sigma_2 = \sigma_X^2 = \sigma_Y^2$$

$$2. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} > z_\alpha \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}$$

$$3. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}} \text{ ou } \bar{d}_{NPar} > z_\alpha \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}$$

Variâncias desconhecidas e iguais

Temos que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ é conhecido.

Estimando via t-Student

Através da distribuição t-Student com $n+m-2$ graus de liberdade, podemos estimar os pontos de corte. Temos nossas decisões:

$$1. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -t_{n+m-2,\alpha} \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$$

$$2. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} > t_{n+m-2,\alpha} \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$$

$$3. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -t_{n+m-2,\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}} \text{ ou } \bar{d}_{NPar} > t_{n+m-2,\frac{\alpha}{2}} \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$$

Onde $s_p^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$, com s_X^2, s_Y^2 sendo os estimadores não enviesados para as variâncias de X e Y, respectivamente. (Ponderamos os estimadores com base no tamanho de sua amostra, assim favorecendo os estimadores mais precisos)

Variâncias desconhecidas e diferentes (caso geral)

Temos que σ_X^2, σ_Y^2 são desconhecidos e diferentes.

Estimando via t-Student

usaremos a distribuição t-Student com n' graus de liberdade. Temos nossas decisões:

$$1. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -t_{n',\alpha}\sqrt{\frac{s_p^2}{n}+\frac{s_p^2}{m}}$$
$$2. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} > t_{n',\alpha}\sqrt{\frac{s_p^2}{n}+\frac{s_p^2}{m}}$$
$$3. \text{ Rejeita } H_0 \text{ se } \bar{d}_{\text{NPar}} < -t_{n',\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n}+\frac{s_p^2}{m}} \text{ ou } \bar{d}_{NPar} > t_{n',\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n}+\frac{s_p^2}{m}}$$

Onde $s_p^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}$.

Encontrando n'

Na fórmula acima, temos os graus de liberdade da t-Student dado por

$$n' \approx \frac{\left(\frac{s_X^2}{n}+\frac{s_Y^2}{m}\right)^2}{\frac{\left(\frac{s_X^2}{n}\right)^2}{n-1}+\frac{\left(\frac{s_Y^2}{m}\right)^2}{m-1}}$$

Esse valor, caso não inteiro, deverá ser arredondado.

Exemplo (Importante)

Queremos testar a resistência de dois tipos de viga de aço, A e B. Tomando-se $n = 15$ vigas do tipo A e $m = 20$ vigas do tipo B. de um teste f, conseguimos com 10% de significância que as variâncias não são iguais. Obtemos os valores da tabela:

Tipo	Média	Variância(s^2)
A	70.5	81.6
B	84.3	210.8
$\bar{d}_{\text{NPar}}$	-13.8	$--$

Teste a hipótese

$$\begin{cases} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

Com significância $\alpha = 0.05$ para os casos

Caso 1. Variâncias conhecidas

Temos do produtor que $\sigma_X^2 = 81, \sigma_Y^2 = 209$
$\bar{d}_{\text{NPar}} = 70.5 - 84.3 = -13.8$. Logo,
$z_{\frac{\alpha}{2}} \sqrt{\frac{81}{15} + \frac{209}{20}} = 7.8$. Como $-13.8 < -7.8$, concluímos que há evidências para rejeitarmos a hipótese nula de que as resistências médias das vigas A, B são iguais a $\alpha = 5\%$ de significância estatística

Caso 2. Variâncias desconhecidas e iguais.

Para $\alpha = 0.05$, $t_{33,0.025} = 2.03$. Encontrando $s_P^2 = 155.988$. Finalmente,
$t_{33,0.025} \sqrt{\frac{s_P^2}{n} + \frac{s_P^2}{m}} = 8.65$.
Como $-13.8 < -8.65$, concluímos que há evidências para rejeitarmos a hipótese nula a $\alpha = 5\%$ de significância estatística.

Caso 3. Variâncias desconhecidas e diferentes

Primeiro calculamos $n' = 32.08 \overset{\text{Arrendonda}}{=} 32$. Assim, $t_{32,0.025} = 2.037$.

Portanto, $t_{32,0.025} \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} = 8.14$. Como $-13.8 < -8.15$, concluímos que há evidências para rejeitarmos a hipótese nula a $\alpha = 5\%$ de significância estatística.