STA 114: Statistics

Notes 15. Comparing two groups by their means

A large number of statistical applications boil down to comparing two populations through their means. For example, suppose you have to decide which of the two sites, site A and site B, is to be excavated in a copper mine. Your decision is to be based on copper specimens $X_1, \cdots, X_n$ from site A and $Y_1, \cdots, Y_m$ from site B. A reasonable data model is given by $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma_1^2)$ and $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2)$ with X_i's and Y_j's independent of each other. Your decision on which site to excavate should depend on your assessment of the quantity $\eta = \mu_1 - \mu_2$.

Similar tasks arise in clinical trials when comparing efficacy of a treatment against control, in comparing income or achievement between two groups (split by gender or race or training received, etc.), and so on. Note that, what we are interested in here is the difference between the group specific expected values (means) of the outcome variable. Another interesting variable to look at would be $D = Y_{m+1} - X_{n+1}$, the difference in the outcome value between future (hypothetical) samples drawn from each group. However, we won't address this today.

The two means problem with equal variance

In some applications it is reasonable to assume that the two groups have identical variability around their respective means, i.e., the model simplifies to $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma^2)$, $Y_1, \cdots, Y_m \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma^2)$, X_i's and Y_j's are independent, with model parameters $\mu_1 \in (-\infty, \infty)$, $\mu_2 \in (-\infty, \infty)$, $\sigma^2 \in (0, \infty)$.

We shall denote $X = (X_1, \cdots, X_n)$, $Y = (Y_1, \cdots, Y_n)$, so, our data is (X, Y) and an observation on this data is denoted (x, y), with $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_m)$. Because of the assumed independent between the X_i's and Y_j's, the log-likelihood function is given by

$$\begin{aligned}
\ell_{x,y}(\mu_1, \mu_2, \sigma^2) &= \log f(x, y \mid \mu_1, \mu_2, \sigma^2) \\
&= \log \left[\frac{e^{-\frac{1}{2}(x_1-\mu_1)^2}}{\sqrt{2\pi\sigma^2}} \times \cdots \times \frac{e^{-\frac{1}{2}(x_n-\mu_1)^2}}{\sqrt{2\pi\sigma^2}} \times \frac{e^{-\frac{1}{2}(y_1-\mu_2)^2}}{\sqrt{2\pi\sigma^2}} \times \cdots \times \frac{e^{-\frac{1}{2}(y_m-\mu_2)^2}}{\sqrt{2\pi\sigma^2}} \right] \\
&= \text{const} - \frac{n+m}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (x_i - \mu_1)^2}{2\sigma^2} - \frac{\sum_{j=1}^m (y_j - \mu_2)^2}{2\sigma^2} \\
&= \text{const} - \frac{n+m}{2} \log \sigma^2 - \frac{(n-1)s_x^2 + (m-1)s_y^2 + n(\bar{x} - \mu_1)^2 + m(\bar{y} - \mu_2)^2}{2\sigma^2}
\end{aligned}$$

To perform ML inference on $\eta = \mu_1 - \mu_2$, we first need to derive the profile likelihood function of this quantity. Recall that this function is defined to be:

$$\ell^*_{x,y}(\eta) = \max_{(\mu_1, \mu_2, \sigma^2) \text{ such that } \mu_1 - \mu_2 = \eta} \ell_{x,y}(\mu_1, \mu_2, \sigma^2), \quad -\infty < \eta < \infty.$$

That is, the profile likelihood for η equals $\ell_{x,y}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)$ for the $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)$ that enjoys the best support from observed data (x, y) among all (μ_1, μ_2, σ^2) satisfying $\mu_1 - \mu_2 = \eta$. To find this point $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)$ we must use a Lagrange multiplier approach.

Define

$$\tilde{\ell}(\mu_1, \mu_2, \sigma^2, \lambda) = \ell_{x,y}(\mu_1, \mu_2, \sigma^2) + \lambda(\mu_1 - \mu_2 - \eta)$$

over $-\infty < \mu_1, \mu_2 < \infty$, $\sigma^2 > 0$ and $-\infty < \lambda < \infty$. Then$\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2$, for some $\hat{\lambda}$ must be the solutions of

$$\begin{aligned}
0 &= \frac{\partial}{\partial \mu_1}\tilde{\ell}(\mu_1, \mu_2, \sigma^2, \lambda) = \frac{n(\bar{x} - \mu_1)}{\sigma^2} + \lambda \\
0 &= \frac{\partial}{\partial \mu_2}\tilde{\ell}(\mu_1, \mu_2, \sigma^2, \lambda) = \frac{m(\bar{y} - \mu_2)}{\sigma^2} - \lambda \\
0 &= \frac{\partial}{\partial \sigma^2}\tilde{\ell}(\mu_1, \mu_2, \sigma^2, \lambda) = -\frac{n+m}{2\sigma^2} + \frac{(n-1)s_x^2 + (m-1)s_y^2 + n(\bar{x} - \mu_1)^2 + m(\bar{y} - \mu_2)^2}{2\sigma^4} \\
0 &= \frac{\partial}{\partial \lambda}\tilde{\ell}(\mu_1, \mu_2, \sigma^2, \lambda) = \mu_1 - \mu_2 - \eta.
\end{aligned}$$

Some algebra shows that the solutions must equal

$$\begin{aligned}
\hat{\mu}_1 &= \frac{n\bar{x} + m\bar{y}}{n+m} + \frac{m}{n+m}\eta \\
\hat{\mu}_2 &= \frac{n\bar{x} + m\bar{y}}{n+m} - \frac{n}{n+m}\eta \\
\hat{\sigma}^2 &= \frac{(n-1)s_x^2 + (m-1)s_y^2 + n(\bar{x} - \hat{\mu}_1)^2 + m(\bar{y} - \hat{\mu}_2)^2}{n+m} \\
&= \frac{(n-1)s_x^2 + (m-1)s_y^2 + \frac{nm^2}{(n+m)^2}(\bar{x} - \bar{y} - \eta)^2 + \frac{mn^2}{(n+m)^2}(\bar{x} - \bar{y} - \eta)^2}{n+m} \\
&= \frac{(n-1)s_x^2 + (m-1)s_y^2 + \frac{nm}{n+m}(\bar{x} - \bar{y} - \eta)^2}{n+m}
\end{aligned}$$

and consequently,

$$\begin{aligned}
\ell^*_{x,y}(\eta) &= \ell_{x,y}(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) \\
&= \text{const} - \frac{n+m}{2}\log\hat{\sigma}^2 - \frac{n+m}{2} \\
&= \text{const} - \frac{n+m}{2}\log\hat{\sigma}^2 \\
&= \text{const} - \frac{n+m}{2}\log\left(1 + \frac{\frac{nm}{n+m}(\bar{x} - \bar{y} - \eta)^2}{(n-1)s_x^2 + (m-1)s_y^2}\right).
\end{aligned}$$

MLE and ML intervals of η

The MLE $\hat{\eta}_{\text{MLE}}(x, y)$ of η is found by maximizing the profile likelihood function in η, which is same as minimizing the log term on the above right. Because log is a monotone increasing function, $\hat{\eta}_{\text{MLE}}(x)$ then must minimize $(\bar{x} - \bar{y} - \eta)^2$ in η. This happens at

$$\hat{\eta}_{\text{MLE}}(x) = \bar{x} - \bar{y}.$$

For any positive constant c, the ML interval $B_c(x,y) = \{\eta : \ell^*_{x,y}(\eta) \geq \ell^*_{x,y}(\hat{\eta}_{\text{MLE}}(x)) - c^2/2\}$ must equal

$$B_c(x,y) = \left\{ \eta : \frac{n+m}{2} \log\left(1 + \frac{\frac{nm}{n+m}(\bar{x}-\bar{y}-\eta)^2}{(n-1)s_x^2 + (m-1)s_y^2}\right) \leq c^2 \right\}$$
$$= (\bar{x}-\bar{y}) \mp c' \sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}$$

for some $c' > 0$ that depends on c and n, m [more precisely, $c' = \sqrt{(n+m-2)\{\exp(\frac{c^2}{n+m}) - 1\}} \approx c$ for large n, m.].

ML confidence intervals for η

Let's calculate the coverage probability of $B(x,y) = (\bar{x}-\bar{y}) \mp c\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}$ at a given point (μ_1, μ_2, σ^2) in the parameter space. This equals

$$\begin{aligned}
\gamma(\mu_1,\mu_2,\sigma^2; B) &= P_{[X,Y|\mu_1,\mu_2,\sigma^2]}\left(\mu_1 - \mu_2 \in B(X,Y)\right) \\
&= P_{[X,Y|\mu_1,\mu_2,\sigma^2]}\left(-c \leq \frac{(\bar{X}-\mu_1)-(\bar{Y}-\mu_2)}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}}} \leq c\right) \\
&= 2\Phi_{n+m-2}(c) - 1
\end{aligned}$$

because of the following result.

RESULT 1. Suppose $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma^2)$, $Y_1, \cdots, Y_m \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma^2)$, X_i's and Y_j's are independent. Then

$$T = \frac{(\bar{X}-\mu_1)-(\bar{Y}-\mu_2)}{\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2}}}$$

has a $t(n+m-2)$ distribution.

Proof. By our old results on normal data, $\bar{X} \sim \mathsf{Normal}(\mu_1, \sigma^2/n)$, $\frac{(n-1)s_X^2}{\sim}\chi^2(n-1)$, $\bar{Y} \sim \mathsf{Normal}(\mu_2, \sigma^2/m)$, $\frac{(m-1)s_Y^2}{\sigma^2} \sim \chi^2(m-1)$ and these four random variables are independent of each other. Hence $U = \frac{(\bar{X}-\mu_1)-(\bar{Y}-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim \mathsf{Normal}(0,1)$ and $V = \frac{(n-1)s_X^2+(m-1)s_Y^2}{\sigma^2} \sim \chi^2(n+m-2)$ with U and V independent [two independent chi-square variables add to form another chi-square variable, with the parameters added]. Therefore, $T = U/\sqrt{V/(n+m-2)} \sim t(n+m-2)$. □

From the above coverage calculation, it's clear that the confidence coefficient of $(\bar{x}-\bar{y}) \mp c\sqrt{\left(\frac{1}{n}+\frac{1}{m}\right)\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}$ equals $2\Phi_{n+m-2}(c) - 1$. Hence we can form a $100(1-\alpha)\%$ ML

confidence interval for η by taking $c = z_{n+m-2}(\alpha)$. So an ML $100(1-\alpha)\%$-CI for η equals

$$(\bar{x} - \bar{y}) \mp z_{n+m-2}(\alpha)\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

Unequal variances

In the more general setting, we should allow the two groups to have different variabilities around their respective means, i.e., we cannot assume $\sigma_1^2 = \sigma_2^2$. So now our model is $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma_1^2)$, $Y_1, \cdots, Y_m \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2)$, X_i's and Y_j's are independent. The model parameters are $-\infty < \mu_1, \mu_2 < \infty$, $\sigma_1^2, \sigma_2^2 > 0$.

Rather surprisingly exact $100(1-\alpha)\%$ confidence intervals for $\eta = \mu_1 - \mu_2$ are not known for this problem. Instead, the following approximately $100(1-\alpha)\%$ confidence interval (known as Welch's method) is widely popular:

$$(\bar{x} - \bar{y}) \mp z_{r(x,y)}(\alpha)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$$

where the degrees of freedom $r(x, y)$ depends on data as

$$r(x,y) = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}}.$$

Example (Soporific drug)**.** In a sleep study, 10 patients (group 1) received a soporific drug while 10 other patients (group 2) received a placebo. For every patient, their increase in nightly sleep hours was recorded. Let X_i denote the measurements from group 1 and Y_j's those from group 2. Model $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma_1^2)$, $Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2)$, X_i's and Y_j's are independent. We are interested in confidence intervals for $\eta = \mu_1 - \mu_2$ based on observations (1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4) from group 1 and (0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0.0, 2.0) from group 2. For these observations $n = m = 10$, $\bar{x} = 2.33$, $s_x = 2$, $\bar{y} = 0.75$ and $s_y = 1.79$.

If we assume $\sigma_1^2 = \sigma_2^2$, then a 95%-confidence interval for η is

$$1.58 \mp z_{18}(.05) \times 0.85 = 1.58 \mp 2.1 \times 0.85 = [-0.205, 3.365].$$

On the other hand, if we didn't assume equality and the variance, then we first calculate $r(x, y) = 17.78$ (fairly close to $n + m - 2 = 18$). Therefore a 95% (approximate) confidence interval is

$$1.58 \mp z_{17.78}(0.05) \times 0.85 = [-0.205, 3.365].$$

Confidence coefficient of Welch's interval

Let $W_\alpha(x, y)$ denote Welch's $100(1-\alpha)\%$ approximate confidence interval for $\eta = \mu_1 - \mu_2$. Can we calculate its exact confidence coefficient? For any $-\infty < \mu_1, \mu_2 < \infty$, $\sigma_1^2, \sigma_2^2 > 0$,

it's coverage is (by definition):

$$\gamma(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2; W_\alpha) = P_{[X,Y|\mu_1,\mu_2,\sigma_1^2,\sigma_2^2]}\left(\mu_1 - \mu_2 \in (\bar{X} - \bar{Y}) \mp z_{r(X,Y)}(\alpha)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}\right)$$

To get handle on this quantity, we first notice that

$$\gamma(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2; W) = \gamma(0, 0, 1, \sigma_2^2/\sigma_1^2; W).$$

To see this, define $X' = (X_1', \cdots, X_n')$ and $Y' = (Y_1', \cdots, Y_m')$ where $X_i' = (X_i - \mu_1)/\sigma_1$ and $Y_j' = (Y_j - \mu_2)/\sigma_1$. Then $X_i' \overset{\text{IID}}{\sim} \mathsf{Normal}(0, 1)$, $Y_j' \overset{\text{IID}}{\sim} \mathsf{Normal}(0, \sigma_2^2/\sigma_1^2)$ and X_i', Y_i' are independent. Also,

$$\mu_1 - \mu_2 \in (\bar{X} - \bar{Y}) \mp z_{r(X,Y)}(\alpha)\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}} \iff 0 \in (\bar{X}' - \bar{Y}') \mp z_{r(X',Y')}\sqrt{\frac{s_{X'}^2}{n} + \frac{s_{Y'}^2}{m}}$$

because $\bar{X}' = (\bar{X} - \mu_1)/\sigma_1$, $\bar{Y}' = (\bar{Y} - \mu_2)/\sigma_1$, $s_{X'}^2 = s_X^2/\sigma_1^2$, $s_{Y'}^2 = s_Y^2/\sigma_1^2$ and $r(X', Y') = r(X, Y)$.

Therefore the confidence coefficient of W_α can be found as

$$\gamma(W_\alpha) = \min_{\tau^2 \in (0,\infty)} \gamma(0, 0, 1, \tau^2; W_\alpha).$$

Unfortunately, it is not possible to get $\gamma(0, 0, 1, \tau^2; W_\alpha)$ in closed form for a given τ^2 (and α). But we can simulate! Below are some results of a massive simulation in R where we report $\tilde{\gamma}(\tau^2) = \gamma(0, 0, 1, \tau^2; W_{.05})$.

n	m	$\tilde{\gamma}(1)$	$\tilde{\gamma}(5)$	$\tilde{\gamma}(10)$	$\tilde{\gamma}(100)$	$\tilde{\gamma}(10^3)$	$\tilde{\gamma}(10^4)$	$\tilde{\gamma}(10^5)$	$\tilde{\gamma}(10^6)$
2	2	0.98	0.97	0.96	0.94	0.94	0.95	0.95	0.95
2	4	0.95	0.96	0.96	0.95	0.95	0.95	0.95	0.95
4	4	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95
10	10	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
10	20	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
20	20	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95