# STA 250: Statistics

# Notes 15. Comparing two normal populations

# Book chapters: 9.6

## 1 Background

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.

## 2 Classical analysis of 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 $(X, Y)$.

A typical (classical) inference task for this model is to test

$$H_0 : \mu_1 = \mu_2 \text{ against } H_1 : \mu_1 \neq \mu_2.$$

Because $\bar{X}$ is expected to be close $\mu_1$ and $\bar{Y}$ is expected to be close to $\mu_2$, the statistic $|\bar{X} - \bar{Y}|$ gives a reasonable measure of evidence against $H_0$, i.e., we will be inclined to think $H_0$ is false if $|\bar{X} - \bar{Y}|$ was observed to be large. Note that under the model,

$$\bar{X} - \bar{Y} \sim \mathsf{Normal}\left(\mu_1 - \mu_2, \sigma^2\left\{\frac{1}{n} + \frac{1}{m}\right\}\right). \tag{1}$$

So if we knew the value of $\sigma$, we could use the test statistic:

$$T(X,Y) = \frac{|\bar{X} - \bar{Y}|}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

and reject $H_0$ for large values of $T(X,Y)$. But $\sigma$ too is unknown and we need to use an estimate $\hat{\sigma}(X,Y)$ of $\sigma$ to construct our test statistic:

$$T(X,Y) = \frac{|\bar{X} - \bar{Y}|}{\hat{\sigma}(X,Y)\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

and reject $H_0$ for large values of $T(X,Y)$. A typical estimate we use is:

$$\hat{\sigma}^2(x,y) = \frac{n+m}{n+m-2}\hat{\sigma}^2_{\text{MLE}} = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}.$$

## 3 ML Theory

It turns out that any ML interval for $\eta = \mu_1 - \mu_2$ equals

$$\bar{X} - \bar{Y} \mp c \cdot \hat{\sigma}(X,Y)\sqrt{\frac{1}{n} + \frac{1}{m}} \tag{2}$$

with $\hat{\sigma}$ as above, for some $c > 0$. To calculate coverage of this interval, we can state the following analog of (1)

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\hat{\sigma}(X,Y)\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2) \tag{3}$$

and therefore choosing $c = z_{n+m-2}(\alpha)$ guarantees a coverage of $1-\alpha$ at any parameter value. Two consequences of this result are:

Result 1. A $100(1-\alpha)\%$ ML confidence interval for $\eta = \mu_1 - \mu_2$ is

$$\bar{x} - \bar{y} \mp z_{n+m-2}(\alpha) \cdot \hat{\sigma}(x,y)\sqrt{\frac{1}{n} + \frac{1}{m}}$$

with

$$\hat{\sigma}(x,y) = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}}$$

and a size-$\alpha$ ML test for $H_0 : \mu_1 = \mu_2$ rejects $H_0$ if $0 \notin \bar{x} - \bar{y} \mp z_{n+m-2}(\alpha) \cdot \sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$.

Recall that we love ML tests and confidence intervals because they are optimal a large class of reasonable testing and interval rules.

## 4 Default Bayesian Analysis

For a Bayesian analysis of the model and inference on $\eta = \mu_1 - \mu_2$, a common default prior choice on $(\mu_1, \mu_2, \sigma^2)$ is the reference prior:

$$\xi(\mu_1, \mu_2, \sigma^2) = \frac{1}{\sigma^2}, \quad -\infty < \mu_1, \mu_2 < \infty, \sigma > 0$$

which leads to an extended normal-inverse-chi-square posterior pdf on $(\mu_1, \mu_2, \sigma^2)$ that we can describe as:

1. $\sigma^2 | (X = x, Y = y)$ has an inverse-$\chi^2$ distribution: $\frac{(n+m-2)\hat{\sigma}^2(x,y)}{\sigma^2} \sim \chi^2(n + m - 2)$
2. $\mu_1$ and $\mu_2$ are conditionally independent given $\sigma^2$ and $(X = x, Y = y)$ with: $\mu_1 | \sigma^2 \sim \mathsf{Normal}(\bar{x}, \sigma^2/n)$ and $\mu_2 | \sigma^2 \sim \mathsf{Normal}(\bar{y}, \sigma^2/m)$.

A big consequence is that the posterior pdf of $\eta = \mu_2 - \mu_2$ can be described as:

$$\frac{\eta - (\bar{x} - \bar{y})}{\hat{\sigma}(x, y)\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n + m - 2)$$

and consequently, a 95% posterior range for $\eta = \mu_1 - \mu_2$ given data is

$$\bar{x} - \bar{y} \mp z_{n+m-2}(\alpha) \cdot \hat{\sigma}(x, y)\sqrt{\frac{1}{n} + \frac{1}{m}}$$

which is numerically the same as the 95% ML confidence interval for $\eta$.

## 5 Unequal variances: Classical

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

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

The associated Welch's t-test for $H_0 : \mu_1 = \mu_2$ rejects $H_0$ if $0 \notin (\bar{x} - \bar{y}) \mp z_{r(x,y)}(\alpha)\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}$.

## 6 Unequal variance: Default Bayes

There is no standard default prior for the full model with $\sigma_1 \neq \sigma_2$. It is possible to do default Bayes analysis of the two samples separately:

$$X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_1, \sigma_1^2), \xi_1(\mu_1, \sigma_1^2) = \frac{1}{\sigma_1^2} \implies \xi_1(\mu_1, \sigma_1^2|x) = \mathsf{N}\chi^{-2}(\bar{x}, n, n-1, s_x^2)$$

$$Y_j \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu_2, \sigma_2^2), \xi_2(\mu_2, \sigma_2^2) = \frac{1}{\sigma_2^2} \implies \xi_2(\mu_2, \sigma_2^2|y) = \mathsf{N}\chi^{-2}(\bar{y}, m, m-1, s_y^2)$$

and combine them together to draw inference on $\eta = \mu_1 - \mu_2$. In particular if took the "product prior" $\xi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \xi_1(\mu_1, \sigma_1^2) \times \xi_2(\mu_2, \sigma_2^2)$ then the posterior pdf of $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ also retains the product structure:

$$\xi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2|x, y) = \xi_1(\mu_1, \sigma_1^2|x)\xi_2(\mu_2, \sigma_2^2|y).$$

Unfortunately, it is still not possible to write the posterior pdf of $\eta = \mu_1 - \mu_2$ in a recognizable form. But we can use Monte Carlo! Here is how we could construct a 95% posterior range for $\eta$ with a large Monte Carlo sample size $M$ (say $M = 10,000$):

1. Sample $M$ values of $\mu_1$ from its marginal posterior given $X = x$. From our single normal Bayesian theory, this posterior pdf is given by: $\sqrt{n}(\mu_1 - \bar{x})/s_x \sim t(n-1)$. That is samples $\mu_1^{(1)}, \cdots, \mu_1^{(M)}$ of $\mu_1$ can be drawn by first drawing $w^{(1)}, \cdots, w^{(M)} \overset{\text{IID}}{\sim} t(n-1)$ and setting $\mu_1^{(k)} = \bar{x} + s_x w^{(k)}/\sqrt{n}$, $k = 1, \cdots, M$.

2. Sample $M$ values of $\mu_2$ from its marginal posterior give $Y = y$. Now we draw $v^{(k)} \overset{\text{IID}}{\sim} t(m-1)$ and set $\mu_2^{(k)} = \bar{y} + s_y v^{(k)}/\sqrt{m}$, $k = 1, \cdots, M$.

3. Combine the two samples to get $M$ samples of $\eta$: $\eta^{(k)} = \mu_1^{(k)} - \mu_2^{(k)}$, $k = 1, \cdots, M$ and report its 95% range by finding their 2.5-th and 97.5-th sample percentiles.

**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 $\eta$ is

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

which is also the 95% posterior range under the default Bayes analysis.

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

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

Under the product prior default Bayes analysis a 95% posterior range for $\eta$ equals: $[-0.33, 3.48]$.