Experimental Setup

Here we are going through three main proofs regarding ANOVA we introduced during the lecture. We provide some details that we did not cover. Most of the stuff here is basically some algebraic manipulations of expressions, which is not that important and too lengthy to show during the lecture.

First, let's clarify our data. Generally, we have drawn different samples from different populations. Let's say we have k different populations. If $k = 2$, we are essentially dealing with the case of $\boldsymbol{t}$-tests. In ANOVA, $k \geqslant 3$.

Now we let μ_i and σ_i^2 be the mean and the variance, respectively, of the population i, where $i = 1, 2, 3, \cdots, k$. We assume all those populations follow normal distributions:

$$\text{Population } i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

Then we draw samples from each populations. We obtain sample i of size n_i from population i, and the sample mean and the sample variance are $\bar{x}_i$ and s_i^2, respectively. In summary, the data is like this:

Sample	1	2	3	$\cdots$	k
Data	x_{11} x_{12} x_{13} x_{14} $\vdots$	x_{21} x_{22} x_{23} x_{24} $\vdots$	x_{31} x_{32} x_{33} x_{34} $\vdots$	$\cdots$ $\cdots$ $\cdots$ $\cdots$ $\vdots$	x_{k1} x_{k2} x_{k3} x_{k4} $\vdots$
Sample mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\cdots$	$\bar{x}_k$
Sample variance	s_1^2	s_2^2	s_3^2	$\cdots$	s_k^2
Sample size	n_1	n_2	n_3	$\cdots$	n_k

We also denote the total number of data points as n, that is, $n = \sum_{i=1}^{k} n_i$.

1 SST = SSB + SSW

Proof. We start with the definition of SST:

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{\bar{x}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij}-\bar{x}_i)+(\bar{x}_i-\bar{\bar{x}})]^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(x_{ij}-\bar{x}_i)^2+(\bar{x}_i-\bar{\bar{x}})^2+2(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}})] \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{\bar{x}})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}})
\end{aligned}$$

Note that the red term is just **SSW**. We also notice that the inner sum is with respect to j, so any terms regarding i can be treated as a constant term and taken to the font of the inner sum. Therefore, the blue term becomes:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i-\bar{\bar{x}})^2=\sum_{i=1}^{k}n_i(\bar{x}_i-\bar{\bar{x}})^2$$

which is just **SSB**. The last term becomes:

$$\begin{aligned}
2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(\bar{x}_i-\bar{\bar{x}}) &= 2\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i) \\
&= 2\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})\left[\sum_{j=1}^{n_i}x_{ij}-\sum_{j=1}^{n_i}\bar{x}_i\right] \\
&= 2\sum_{i=1}^{k}(\bar{x}_i-\bar{\bar{x}})\left[n_i\cdot\bar{x}_i-n_i\cdot\bar{x}_i\right] \\
&= \mathbf{0}
\end{aligned}$$

Now, we see that SST = SSB + SSW. □

2 The Distribution Related To SSW

Let's write SSW in an estimator format. Recall that another way of writing SSW is:

$$\text{SSW} = \sum_{i=1}^{k} df_i S_i^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2$$

Remember from **Lecture 16**, we derived that:

$$\frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \boldsymbol{\chi^2}(n_i - 1)$$

If we sum them up, we have:

$$\sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{\sigma_i^2} \sim \boldsymbol{\chi^2}(n - k)$$

Again, let's consider the simpler case where all populations have an equal variance $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. The above formula becomes:

$$\sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{\sigma_i^2} = \sum_{i=1}^{k} \frac{(n_i - 1)S_i^2}{\sigma^2} = \frac{\sum_{i=1}^{k} (n_i - 1)S_i^2}{\sigma^2} \sim \boldsymbol{\chi^2}(n - k)$$

Note the numerator is **SSW**, so:

$$\frac{\text{SSW}}{\sigma^2} \sim \boldsymbol{\chi^2}(n - k) \tag{1}$$

One thing we want to emphasise is that the above formula (1) is always true regardless of whether the null hypothesis is true or not.

3 The Distribution Related To SSB

Similarly, let's first write SSB in an estimator format:

$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{\boldsymbol{X}}_i - \bar{\bar{\boldsymbol{X}}})^2$$

From the central limit theorem, we have:

$$\bar{\boldsymbol{X}}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma_i^2}{n_i}\right)$$

Again, we do the same thing as SSW, that is, assuming all population variances are equal: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. Therefore,

$$\bar{\boldsymbol{X}}_i \sim \mathcal{N}\left(\mu_i, \frac{\sigma^2}{n_i}\right) \tag{2}$$

Now let's have a look at $\bar{\bar{\boldsymbol{X}}}$, which is the grand mean calculated by all data points:

$$\bar{\bar{\boldsymbol{X}}} = \frac{\sum_{i=1}^{k} n_i \cdot \bar{\boldsymbol{X}}_i}{n} = \sum_{i=1}^{k} \frac{n_i}{n} \bar{\boldsymbol{X}}_i$$

Apparently, $\bar{\bar{\boldsymbol{X}}}$ is a weighted average of each of the sample mean $\bar{\boldsymbol{X}}_i$, and it also follows a normal distribution. Now, let's figure out its mean and variance.

$$\begin{aligned}
\mathbb{E}\left[\bar{\bar{\boldsymbol{X}}}\right] &= \mathbb{E}\left[\sum_{i=1}^{k} \frac{n_i}{n} \bar{\boldsymbol{X}}_i\right] = \sum_{i=1}^{k} \mathbb{E}\left[\frac{n_i}{n} \bar{\boldsymbol{X}}_i\right] = \sum_{i=1}^{k} \frac{n_i}{n} \mathbb{E}\left[\bar{\boldsymbol{X}}_i\right] \\
&= \sum_{i=1}^{k} \frac{n_i}{n} \mu_i = \bar{\mu}
\end{aligned}$$

where $\bar{\mu}$ is just a weighted average of each population mean μ_i . In terms of its variance:

$$\begin{aligned}
\mathbb{V}\text{ar}\left(\bar{\bar{\boldsymbol{X}}}\right) &= \mathbb{V}\text{ar}\left(\sum_{i=1}^{k} \frac{n_i}{n} \bar{\boldsymbol{X}}_i\right) = \sum_{i=1}^{k} \mathbb{V}\text{ar}\left(\frac{n_i}{n} \bar{\boldsymbol{X}}_i\right) \qquad \text{due to i.i.d. of } \boldsymbol{X}_i \\
&= \sum_{i=1}^{k} \frac{n_i^2}{n^2} \mathbb{V}\text{ar}\left(\bar{\boldsymbol{X}}_i\right) = \sum_{i=1}^{k} \frac{n_i^2}{n^2} \cdot \frac{\sigma^2}{n_i} = \sum_{i=1}^{k} \frac{n_i \sigma^2}{n^2} \\
&= \frac{\sigma^2}{n^2} \sum_{i=1}^{k} n_i = \frac{\sigma^2}{n^2} \cdot n = \frac{\sigma^2}{n}
\end{aligned}$$

Therefore, we see that:

$$\bar{\bar{\boldsymbol{X}}} \sim \mathcal{N}\left(\bar{\mu}, \frac{\sigma^2}{n}\right) \tag{3}$$

If you think about it, it actually makes sense due to the central limit theorem.

Now let's get back to see SSB:

$$\begin{aligned}
\text{SSB} &= \sum_{i=1}^{k} n_i(\bar{\boldsymbol{X}}_i - \bar{\bar{\boldsymbol{X}}})^2 = \sum_{i=1}^{k} n_i \left[(\bar{\boldsymbol{X}}_i - \mu_i) + \left(\mu_i - \bar{\bar{\boldsymbol{X}}}\right)\right]^2 \\
&= \sum_{i=1}^{k} \left[n_i \left(\bar{\boldsymbol{X}}_i - \mu_i\right)^2 + n_i \left(\bar{\bar{\boldsymbol{X}}} - \mu_i\right)^2 + 2n_i \left(\bar{\boldsymbol{X}}_i - \mu_i\right) \left(\mu_i - \bar{\bar{\boldsymbol{X}}}\right)\right] \\
&= \sum_{i=1}^{k} n_i \left(\bar{\boldsymbol{X}}_i - \mu_i\right)^2 + \sum_{i=1}^{k} n_i \left(\bar{\bar{\boldsymbol{X}}} - \mu_i\right)^2 + 2\sum_{i=1}^{k} n_i \left(\bar{\boldsymbol{X}}_i - \mu_i\right) \left(\mu_i - \bar{\bar{\boldsymbol{X}}}\right)
\end{aligned} \tag{4}$$

Now let's take a closer look at the blue term:

$$\begin{aligned}
&\sum_{i=1}^{k} n_i \left(\bar{\bar{\boldsymbol{X}}} - \mu_i\right)^2 + 2\sum_{i=1}^{k} n_i \left(\bar{\boldsymbol{X}}_i - \mu_i\right) \left(\mu_i - \bar{\bar{\boldsymbol{X}}}\right) \\
&= \sum_{i=1}^{k} \left(n_i \bar{\bar{\boldsymbol{X}}}^2 - 2n_i\mu_i \bar{\bar{\boldsymbol{X}}} + n_i\mu_i^2\right) \\
&\quad + 2\sum_{i=1}^{k} \left(n_i \bar{\boldsymbol{X}}_i \mu_i - n_i \bar{\boldsymbol{X}}_i \bar{\bar{\boldsymbol{X}}} - n_i\mu_i^2 + n_i\mu_i \bar{\bar{\boldsymbol{X}}}\right) \\
&= \bar{\bar{\boldsymbol{X}}}^2 \sum_{i=1}^{k} n_i - 2\bar{\bar{\boldsymbol{X}}} \sum_{i=1}^{k} n_i\mu_i + \sum_{i=1}^{k} n_i\mu_i^2 \\
&\quad + 2\sum_{i=1}^{k} n_i \bar{\boldsymbol{X}}_i \mu_i - 2\bar{\bar{\boldsymbol{X}}} \sum_{i=1}^{k} n_i \bar{\boldsymbol{X}}_i - 2\sum_{i=1}^{k} n_i\mu_i^2 + 2\bar{\bar{\boldsymbol{X}}} \sum_{i=1}^{k} n_i\mu_i
\end{aligned}$$

Note that

$$\sum_{i=1}^{k} n_i = n \text{ and } \sum_{i=1}^{k} n_i \bar{\boldsymbol{X}}_i = n\bar{\bar{\boldsymbol{X}}}$$

so we can further simplify the expression as:

$$\begin{aligned} & n\bar{\bar{\boldsymbol{X}}}^2 - 2\bar{\bar{\boldsymbol{X}}}\sum_{i=1}^{k} n_i\mu_i + \sum_{i=1}^{k} n_i\mu_i^2 \\ & + 2\sum_{i=1}^{k} n_i\bar{\boldsymbol{X}}_i\mu_i - 2\bar{\bar{\boldsymbol{X}}} \cdot n\bar{\bar{\boldsymbol{X}}} - 2\sum_{i=1}^{k} n_i\mu_i^2 + 2\bar{\bar{\boldsymbol{X}}}\sum_{i=1}^{k} n_i\mu_i \end{aligned}$$

Merging the terms in the same colour, we can further simplify the expression as:

$$-n\bar{\bar{\boldsymbol{X}}}^2 + 2\sum_{i=1}^{k} n_i\bar{\boldsymbol{X}}_i\mu_i - \sum_{i=1}^{k} n_i\mu_i^2$$

If the **null hypothesis $\boldsymbol{H_0}$** is true, which means $\mu_1 = \mu_2 = \cdots = \mu_k$, then we have $\mu_i = \bar{\mu}$, where $i = 1, 2, 3, \cdots, k$. Threfore, the above expression can simplified again as:

$$\begin{aligned} -n\bar{\bar{\boldsymbol{X}}}^2 + 2\sum_{i=1}^{k} n_i\bar{\boldsymbol{X}}_i\bar{\mu} - \sum_{i=1}^{k} n_i\bar{\mu}^2 &= -n\bar{\bar{\boldsymbol{X}}}^2 + 2\bar{\mu}\sum_{i=1}^{k} n_i\bar{\boldsymbol{X}}_i - \bar{\mu}^2\sum_{i=1}^{k} n_i \\ &= -n\bar{\bar{\boldsymbol{X}}}^2 + 2\bar{\mu} \cdot n\bar{\bar{\boldsymbol{X}}} - n\bar{\mu}^2 \\ &= -n\left(\bar{\bar{\boldsymbol{X}}} - \bar{\mu}\right)^2 \end{aligned} \tag{5}$$

which is the final form of the blue term when H_0 is true. Putting the formula (5) back to formula (4), we have:

$$\text{SSB} = \sum_{i=1}^{k} n_i\left(\bar{\boldsymbol{X}}_i - \mu_i\right)^2 - n\left(\bar{\bar{\boldsymbol{X}}} - \bar{\mu}\right)^2 \tag{6}$$

which is only true **if the null hypothesis $\boldsymbol{H_0}$** is true.

Now things become clearer. From formula (2), we can get:

$$\begin{aligned} & \frac{\sqrt{n_i}\left(\bar{\boldsymbol{X}}_i - \mu_i\right)}{\sigma} \sim \mathcal{N}(0,1) \Rightarrow \frac{n_i\left(\bar{\boldsymbol{X}}_i - \mu_i\right)^2}{\sigma^2} \sim \boldsymbol{\chi^2}(1) \\ & \Rightarrow \sum_{i=1}^{k} \frac{n_i\left(\bar{\boldsymbol{X}}_i - \mu_i\right)^2}{\sigma^2} \sim \boldsymbol{\chi^2}(k) \end{aligned}$$

Similarly, from formula (3), we can get:

$$\frac{\sqrt{n}\left(\bar{\bar{\boldsymbol{X}}}-\bar{\mu}\right)}{\sigma} \sim \mathcal{N}(0,1) \Rightarrow \frac{n\left(\bar{\bar{\boldsymbol{X}}}-\bar{\mu}\right)^2}{\sigma^2} \sim \boldsymbol{\chi}^{\mathbf{2}}(1)$$

Now we divide by σ^2 at both sides of equation (6), we get:

$$\frac{\text{SSB}}{\sigma^2} = \underbrace{\sum_{i=1}^{k} \frac{n_i\left(\bar{\boldsymbol{X}}_i-\mu_i\right)^2}{\sigma^2}}_{\sim\boldsymbol{\chi}^{\mathbf{2}}(\boldsymbol{k})} - \underbrace{\frac{n\left(\bar{\bar{\boldsymbol{X}}}-\bar{\mu}\right)^2}{\sigma^2}}_{\sim\boldsymbol{\chi}^{\mathbf{2}}(\mathbf{1})}$$

$$\Rightarrow \frac{\text{SSB}}{\sigma^2} \sim \boldsymbol{\chi}^{\mathbf{2}}(k-1) \tag{7}$$

Here we want to emphasise that equation (7) is valid **only when** the null hypothesis H_0 is true.

4 The $\boldsymbol{F}$-test

By definition, we have:

$$\text{MSB} = \frac{\text{SSB}}{k-1} \text{ and } \text{MSW} = \frac{\text{SSW}}{n-k}$$

Therefore, under the null hypothesis H_0, we can see that:

$$\frac{\text{MSB}}{\text{MSW}} = \frac{\frac{\text{SSB}}{k-1}}{\frac{\text{SSW}}{n-k}} = \frac{\frac{\text{SSB}}{\sigma^2}\cdot\frac{1}{k-1}}{\frac{\text{SSW}}{\sigma^2}\cdot\frac{1}{n-k}} = \frac{\frac{\boldsymbol{\chi}^{\mathbf{2}}(k-1)}{k-1}}{\frac{\boldsymbol{\chi}^{\mathbf{2}}(n-k)}{n-k}} \sim \boldsymbol{\mathcal{F}}(k-1, n-k)$$

That's why we use the $\boldsymbol{F}$-tests for ANOVA.