1 Analysis of Variance (ANOVA)

1.1 Hypothesis Test

Suppose that we have k sources of random variables. Each source i provides us with n_i i.i.d observations. We have a total of $N = \sum_{i=1}^{k} n_i$ observations:

$$X_{ij} = \mu_i + \epsilon_{ij}$$

where $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, and we assume that $\forall(i,j), \epsilon_{ij} \sim N(0, \sigma^2)$. Thus

$$\forall(i,j) \left(X_{ij} \sim N(\mu_i, \sigma^2) \right)$$

We wish to test whether our sources are the same, e.g., whether their means are equal. We are assuming that their standard deviations, σ, are equal (but we don't know them). Let $\theta = (\{\mu_i\}, \sigma^2)$ be our unknown (latent) variables. If we knew the prior distribution of θ, P_θ, which takes nonzero values on $\mathbb{R}^k \times \mathbb{R}^+$, we could form the following hypothesis:

$$\begin{aligned} \mathcal{H}_0 &: \mu_1 = \ldots = \mu_k \\ \mathcal{H}_1 &: \text{not } \mathcal{H}_0 \end{aligned}$$

and achieve optimality (minimum probability of error) via the given likelihood ratio test:

$$L(\mathbf{X}) = \frac{P_1(\mathbf{X})}{P_0(\mathbf{X})} \gtrless \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}$$

where

$$\begin{aligned} P(\mathcal{H}_0) &= \int_{\{\sigma^2 \geq 0, 1^T\mu = 0\}} \int_{\sigma \geq 0} P_\theta(\mu, \sigma) d\sigma d\mu \\ P(\mathcal{H}_1) &= 1 - P(\mathcal{H}_0) \\ P_0(\mathbf{X}) &= \int_{\{\sigma^2 \geq 0, 1^T\mu = 0\}} P(\mathbf{X}|\theta') P_\theta(\theta') d\theta' \\ P_1(\mathbf{X}) &= \int_{\{\sigma^2 \geq 0, 1^T\mu \neq 0\}} P(\mathbf{X}|\theta') P_\theta(\theta') d\theta' \\ P(\mathbf{X}|\theta') &= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{ij} (X_{ij} - \mu_i)^2 \right) \end{aligned}$$

however, since the prior distribution on θ is unavailable, we run into two problems: we can't form a bayesian test, and we can't calculate the conditional probabilities of the observations. Instead, we form a Neyman Pearson test and a Generalized MAP likelihood ratio:

$$L(\mathbf{X}) = \frac{\max_{\theta': \sigma^2 \geq 0, 1^T\mu \neq 0} P(\mathbf{X}|\theta') P_\theta(\theta')}{\max_{\theta': \sigma^2 \geq 0, 1^T\mu = 0} P(\mathbf{X}|\theta') P_\theta(\theta')} \gtrless \tau$$

where τ will be chosen such that $P_F = P_0(\text{decide } \mathcal{H}_1) < \alpha$. We simplify again by assuming that $P_\theta(\theta')$ is "uniform" on its (albeit infinite) support, e.g., we force $P_\theta(\theta') = c$.

We arrive at the following ML estimates for the means and variances under the two hypotheses:

$$\log L(\mathbf{X}) = -\frac{1}{2\hat{\sigma}^2}\left(\sum_{ij}(X_{ij}-\hat{\mu}_i)^2 - \sum_{ij}(X_{ij}-\hat{\mu})^2\right) \gtrless \tau'$$

where

$$\begin{aligned}
\hat{\mu} &= \arg\max_{\{\mu_i\}|\mathcal{H}_0} P(\mathbf{X}|\theta) = \frac{1}{N}\sum_{ij} X_{ij} \\
\hat{\mu}_i &= \arg\max_{\{\mu_i\}|\mathcal{H}_1} P(\mathbf{X}|\theta) = \frac{1}{n_i}\sum_{ij} X_{ij} \\
\hat{\sigma}^2 &= \arg\max_{\sigma|\mathcal{H}_1\vee\mathcal{H}_0} P(\mathbf{X}|\theta) = \frac{1}{N}\sum_{ij}(X_{ij}-\hat{\mu})^2
\end{aligned}$$

and note that $\hat{\mu} = \frac{1}{N}\sum_i n_i\hat{\mu}_i$. We can reduce the ratio testing problem again; we divide through by $\sum_{ij}(X_{ij}-\hat{\mu}_i)^2$ to get

$$\frac{s_0}{s_1} = \frac{\sum_{ij}(X_{ij}-\hat{\mu})^2}{\sum_{ij}(X_{ij}-\hat{\mu}_i)^2} \gtrless t$$

Now, we can reduce this to a comparison between the means under the two hypotheses:

$$\begin{aligned}
s_0 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\hat{\mu}_i+\hat{\mu}_i-\hat{\mu})^2 \\
&= \sum_{ij}(X_{ij}-\hat{\mu}_i)^2 + 2(X_{ij}-\hat{\mu}_i)(\hat{\mu}_i-\hat{\mu}) + (\hat{\mu}_i-\hat{\mu})^2 \\
&= \sum_{ij}(X_{ij}-\hat{\mu}_i)^2 + \sum_i n_i(\hat{\mu}_i-\hat{\mu})^2 \\
&= s_1 + s_2
\end{aligned}$$

Thus $\frac{s_0}{s_1} = 1 + \frac{s_2}{s_1}$; and it suffices for us to accept $\mathcal{H}_1$ when

$$P_0\left(\frac{\sum_i n_i(\hat{\mu}_i-\hat{\mu})^2}{\sum_{ij}(X_{ij}-\hat{\mu}_i)^2} \geq t\right) \leq \alpha$$

where α is chosen to be a small probability of false alarm, e.g., .05.

1.2 Useful Distributions

1.2.1 χ^2 distribution