Suppose a data analyst for John Toffee's coffee company wants to assess the mean weekly earnings at two different locations in the city. Location 1 is located on the university campus, while Location 2 is located in the Northwestern Hospital cafeteria.

The two objectives of this analysis are

- to test whether the average weekly earnings for Locations 1 and 2 are the same
- If there is evidence of a difference, obtain a 98% confidence interval for this difference.

Motivating Example: Assumptions

- Let $Y_{11}, Y_{12}, ..., Y_{1n_1}$ be a random sample of earnings from Location 1, where $Y_{1i} \sim G(\mu_1, \sigma_1)$, $i = 1, ..., n_1$
- Let $Y_{21}, Y_{22}, ..., Y_{2n_2}$ be a random sample of earnings from Location 2, where $Y_{2i} \sim G(\mu_2, \sigma_2)$, $i = 1, ..., n_2$
- We assume **both populations** have the **same population variance** σ^2, i.e. $\sigma_1^2 = \sigma_2^2$

Stack two sets of observations in a vector of length n = n1 + n2

$$[Y_{11} \cdots Y_{1n_1}, Y_{21} \cdots Y_{2n_2}]$$

Then Since we assume independence, we can then construct the likelihood function for μ_1, μ_2, and σ^2 as follows:

$$\underbrace{\Pi_{j=1}^{2}}_{\text{For both Locations}}\Pi_{i=1}^{n_j}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2\sigma^2}(y_{ji}-\mu_j)^2\right]$$

- From the likelihood function on the previous slide, we can obtain the following estimates for μ_1, μ_2, and σ^2:

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} = \bar{y}_1$$
$$\hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{2i} = \bar{y}_2$$
$$\hat{\sigma}^2 = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1}[(y_{1i}-\bar{y}_1)^2 + (y_{2i}-\bar{y}_2)^2]$$

Last line: MLE of variance

We can instead use a *pooled estimate* of variance

$$s_p^2 = \frac{1}{n_1+n_2-2}\sum_{i=1}^{n_1}[(y_{1i}-\bar{y}_1)^2 + (y_{2i}-\bar{y}_2)^2]$$

t_{-2}

can also be written as

$$s_p^2 = \frac{w_1 s_1^2 + w_2 s_2^2}{w_1 + w_2}$$

: w_1, w_2 are weights
⇒ weighted average

We note that the maximum likelihood *estimator* of $\mu_1 - \mu_2$ is $\bar{Y}_1 - \bar{Y}_2$

- $E(\bar{Y}_1 - \bar{Y}_2) = \mu_1 - \mu_2$
- $Var(\bar{Y}_1 - \bar{Y}_2) = \sigma^2(\frac{1}{n_1} + \frac{1}{n_2})$: assuming populations have same variance

If we know $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but the actual value of σ^2 is unknown, we may consider an *estimator* of $Var(\bar{Y}_1 - \bar{Y}_2)$ from the pooled data:

$$S_p^2(\frac{1}{n_1} + \frac{1}{n_2})$$

Sampling Distribution for $\bar{Y}_1 - \bar{Y}_2$ when $\sigma_1 = \sigma_2$

- **Theorem**: If $Y_{11}, ..., Y_{1n_1}$ is a random sample from a $G(\mu_1, \sigma)$ distribution and independently $Y_{21}, ..., Y_{2n_2}$ is a random sample from a $G(\mu_2, \sigma)$ distribution then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

and

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n_j}(Y_{ji} - \bar{Y}_j)^2 \sim \chi^2(n_1+n_2-2)$$

Then, to test $H_0: \mu_1 - \mu_2 = \mu_0$:

$$D = \frac{|(\bar{Y}_1 - \bar{Y}_2) - \mu_0|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Ex.

Hypothesis Testing for $\mu_1 - \mu_2$ when $\sigma_1 = \sigma_2$: Coffee Shop Example

- From the motivating example, suppose we observe a sample of $n_1 = n_2 = 12$, with $\bar{y}_1 = 1250$, $\bar{y}_2 = 1244$, $\sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 = 30.5$, and $\sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2 = 32.7$. Test $H_0: \mu_1 - \mu_2 = 0$ and state your conclusion in the context of the problem.

: similar sample variance
⇒ we can assume $\sigma_1^2 = \sigma_2^2$

$$s_p^2 = \frac{1}{12+12-2}(30.5 + 32.7) = 3.16 \quad ; \quad s_p = 1.78$$

$$\Rightarrow P(D \geq d) = P\left(D \geq \frac{|1250 - 1244|}{1.78\sqrt{1/12 + 1/12}}\right) \quad : d \sim t(n_1 + n_2 - 2)$$

$$= 2 - 2P(T \le 11.3)$$

Populations with unequal variances

If the population variances are known, our estimator is

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim G(0, 1)$$

If not, there is no exact pivotal quantity we can use. However, for large values of n1 and n2, we can use the approximation

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim G(0, 1)$$

Ex. Going back to the coffee shop example, suppose we instead observe a sample of $n_1 = n_2 = 80$, with $\bar{y}_1 = 1250$, $\bar{y}_2 = 1244$, $\sum_{i=1}^{12}(y_{1i} - \bar{y}_1)^2 = 30.5$, and $\sum_{i=1}^{12}(y_{2i} - \bar{y}_2)^2 = 44.7$. Test $H_0 : \mu_1 - \mu_2 = 0$ and state your conclusion in the context of the problem.

$$P(D \ge d; H_0) = P\left[D \ge \frac{|(1250 - 1244) - 0|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right]$$

$$s_1^2 = \frac{30.5}{11} \quad ; \quad s_2^2 = \frac{44.7}{11}$$

$$\Rightarrow P(|D| \ge 4.368) \quad : \quad D \sim G(0,1)$$