18 Inference about Differences in Means

18.B Appendix: The Distribution of the Pooled Sample Variance

In this appendix, we prove that the t-statistic for the difference in means $\mu_X - \mu_Y$ in the two-sample model with normal trials and equal variances (Section 18.1.3) has a t distribution.

To proceed with our small-sample techniques based on the t distribution, we need to be able to find an estimator of the variance of $\bar{X}_n - \bar{Y}_m$ that is described by a χ^2 distribution and is independent of $\bar{X}_n - \bar{Y}_m$. This is where the equal variance assumption comes into play. If σ_X^2 and σ_Y^2 differ, there is no estimator with the properties we seek.

Distribution of the pooled sample variance and the *t*-statistic.

In the normal two-sample model with equal variances,

(i) $\bar{X}_n - \bar{Y}_m \sim N(\mu_X - \mu_Y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$,
(ii) $\frac{n+m-2}{\sigma^2} S_{\text{pool}}^2 \sim \chi^2(n+m-2)$, *and*
(iii) $\bar{X}_n - \bar{Y}_m$ *and* S_{pool}^2 *are independent.*

$$\textit{Thus}\quad \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{S_{\text{pool}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

Part (i) of the statement was noted in the text. For part (ii) we need the following fact about χ^2 distributions:

Sums of independent χ^2 random variables.

If $C \sim \chi^2(c)$ and $D \sim \chi^2(d)$ are independent, then $C + D \sim \chi^2(c+d)$.

This fact is easy to explain. Remember that a $\chi^2(c)$ random variable can be represented as the sum of the squares of c independent $N(0,1)$ random variables. If we represent both C and D in this way, making sure that all of the $N(0,1)$ random variables used are independent, then $C + D$ will be the sum of the squares of $c + d$ independent $N(0,1)$ variables, and so will have a $\chi^2(c+d)$ distribution.

We first derive the distribution of $\mathcal{S}^2_{\text{pool}}$. Since $\sigma^2_X = \sigma^2_Y = \sigma^2$, we have

$$\frac{n-1}{\sigma^2}\mathcal{S}^2_X \sim \chi^2(n-1) \quad \text{and} \quad \frac{m-1}{\sigma^2}\mathcal{S}^2_Y \sim \chi^2(m-1)$$

(see Section 17.A.2). It will be useful to express these facts in the form

$$\mathcal{S}^2_X \sim \sigma^2\frac{\chi^2(n-1)}{n-1} \quad \text{and} \quad \mathcal{S}^2_Y \sim \sigma^2\frac{\chi^2(m-1)}{m-1}. \tag{18.B.1}$$

Here, the notation $\sigma^2\frac{\chi^2(n-1)}{n-1}$ represents the distribution of a $\chi^2(n-1)$ random variable that has been multiplied by $\frac{\sigma^2}{n-1}$. Since $\mathcal{S}^2_X = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ and $\mathcal{S}^2_Y = \frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y}_m)^2$, multiplying the relations in (18.B.1) by $n-1$ and by $m-1$ yields

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2\chi^2(n-1) \quad \text{and} \quad \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \sim \sigma^2\chi^2(m-1). \tag{18.B.2}$$

Because the x and y samples are independent of each other, the two sums in (18.B.2) are independent of one another, so the fact yields

$$\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^m (Y_i - \bar{Y}_m)^2 \sim \sigma^2\chi^2(n+m-2).$$

Dividing both sides by $n+m-2$ gives us

$$\frac{1}{n+m-2}\left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{j=1}^m (Y_j - \bar{Y}_m)^2\right) \sim \sigma^2\frac{\chi^2(n+m-2)}{n+m-2}.$$

The left-hand side of this expression is the definition of $\mathcal{S}^2_{\text{pool}}$, so we are done.

Next, we establish part (iii), that the difference $\bar{X}_n - \bar{Y}_m$ and the pooled sample variance $\mathcal{S}^2_{\text{pool}}$ are independent random variables. We know from Section 17.A.4 that $\bar{X}_n$ and $\mathcal{S}^2_X$ are independent random variables, as are $\bar{Y}_n$ and $\mathcal{S}^2_Y$. Moreover, since the x and y samples are independent, we also know that $\bar{X}_n$ is independent of $\mathcal{S}^2_Y$, and $\bar{Y}_m$ of $\mathcal{S}^2_X$. Finally, we saw in the text that we can write $\mathcal{S}^2_{\text{pool}}$ as $\frac{n-1}{n+m-2}\mathcal{S}^2_X + \frac{m-1}{n+m-2}\mathcal{S}^2_Y$. Since each of the two terms in this sum is independent of each of the two terms in $\bar{X}_n - \bar{Y}_m$, it follows (see "New independent random variables from old" in Section 4.A.1) that $\bar{X}_n - \bar{Y}_m$ and $\mathcal{S}^2_{\text{pool}}$ are independent.