8.4: Estimation in the Two-Sample Normal Model

As we have noted before, the normal distribution is perhaps the most important distribution in the study of mathematical statistics, in part because of the central limit theorem. As a consequence of this theorem, measured quantities that are subject to numerous small, random errors will have, at least approximately, normal distributions. Such variables are ubiquitous in statistical experiments, in subjects varying from the physical and biological sciences to the social sciences.

In this section, we will study estimation problems in the two-sample normal model and in the bivariate normal model. This section parallels the section on Tests in the Two-Sample Normal Model in the Chapter on Hypothesis Testing.

The Two-Sample Normal Model

Preliminaries

Suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$ is a random sample of size m from the normal distribution with mean μ and standard deviation σ, and that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ is a random sample of size n from the normal distribution with mean ν and standard deviation τ. Moreover, suppose that the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Usually, the parameters are unknown, so the parameter space for our vector of parameters (μ, ν, σ, τ) is $\mathbb{R}^2 \times (0, \infty)^2$.

This type of situation arises frequently when the random variables represent a measurement of interest for the objects of the population, and the samples correspond to two different treatments. For example, we might be interested in the blood pressure of a certain population of patients. The $\boldsymbol{X}$ vector records the blood pressures of a control sample, while the $\boldsymbol{Y}$ vector records the blood pressures of the sample receiving a new drug. Similarly, we might be interested in the yield of an acre of corn. The $\boldsymbol{X}$ vector records the yields of a sample receiving one type of fertilizer, while the $\boldsymbol{Y}$ vector records the yields of a sample receiving a different type of fertilizer.

Usually our interest is in a comparison of the parameters (either the means or standard deviations) for the two sampling distributions. In this section we will construct confidence intervals for the difference of the distribution means $\nu - \mu$ and for the ratio of the distribution variances τ^2 / σ^2. As with previous estimation problems, the construction depends on finding appropriate pivot variables.

For a generic sample $\boldsymbol{U} = (U_1, U_2, \ldots, U_k)$ from a distribution with mean a, we will use our standard notation for the sample mean and for the sample variance.

$$M(\boldsymbol{U}) = \frac{1}{k} \sum_{i=1}^{k} U_i \tag{8.4.1}$$

$$S^2(\boldsymbol{U}) = \frac{1}{k-1} \sum_{i=1}^{k} [U_i - M(\boldsymbol{U})]^2 \tag{8.4.2}$$

We will need to also recall the special properties of these statistics when the sampling distribution is normal. The special pivot distributions that will play a fundamental role in this section are the standard normal, the student t, and the Fisher F distributions. To construct our interval estimates we will need the quantiles of these distributions. The quantiles can be computed using the special distribution calculator or from most mathematical and statistical software packages. Here is the notation we will use:

Let $p \in (0, 1)$ and let $j,\ k \in \mathbb{N}_+$.

1. $z(p)$ denotes the quantile of order p for the standard normal distribution.
2. $t_k(p)$ denotes the quantile of order p for the student t distribution with k degrees of freedom.
3. $f_{j,k}(p)$ denotes the quantile of order p for the student f distribution with j degrees of freedom in the numerator and k degrees of freedom in the denominator.

Recall that by symmetry, $z(p) = -z(1-p)$ and $t_k(p) = -t_k(1-p)$ for $p \in (0, 1)$ and $k \in \mathbb{N}_+$. On the other hand, there is no simple relationship between the left and right tail probabilities of the F distribution.

Confidence Intervals for the Difference of the Means with Known Variances

First we will construct confidence intervals for $\nu - \mu$ under the assumption that the distribution variances σ^2 and τ^2 are known. This is not always an artificial assumption. As in the one sample normal model, the variances are sometime stable, and hence are at least approximately known, while the means change under different treatments. First recall the following basic facts:

The difference of the sample means $M(\boldsymbol{Y}) - M(\boldsymbol{X})$ has the normal distribution with mean $\nu - \mu$ and variance $\sigma^2/m + \tau^2/n$. Hence the standard score of the difference of the sample means

$$Z = \frac{[M(\boldsymbol{Y}) - M(\boldsymbol{X})] - (\nu - \mu)}{\sqrt{\sigma^2/m + \tau^2/n}} \tag{8.4.3}$$

has the standard normal distribution. Thus, this variable is a pivotal variable for $\nu - \mu$ when σ, τ are known.

The basic confidence interval and upper and lower bound are now easy to construct.

For $\alpha \in (0, 1)$,

1. $\left[M(\boldsymbol{Y}) - M(\boldsymbol{X}) - z\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}, M(\boldsymbol{Y}) - M(\boldsymbol{X}) + z\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}\right]$ is a $1 - \alpha$ confidence interval for $\nu - \mu$.
2. $M(\boldsymbol{Y}) - M(\boldsymbol{X}) - z(1 - \alpha)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}$ is a $1 - \alpha$ confidence lower bound for $\nu - \mu$.
3. $M(\boldsymbol{Y}) - M(\boldsymbol{X}) + z(1 - \alpha)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}$ is a $1 - \alpha$ confidence upper bound for $\nu - \mu$.

Proof

The variable T given above has the standard normal distribution. Hence each of the following events has probability $1 - \alpha$ by definition of the quantiles:

1. $\left\{-z\left(1 - \frac{\alpha}{2}\right) \le Z \le z\left(1 - \frac{\alpha}{2}\right)\right\}$
2. $\{Z \ge z(1 - \alpha)\}$
3. $\{Z \le -z(1 - \alpha)\}$

In each case, solving the inequality for $\nu - \mu$ gives the result.

The two-sided interval in part (a) is the *symmetric interval* corresponding to $\alpha/2$ in both tails of the standard normal distribution. As usual, we can construct more general two-sided intervals by partitioning α between the left and right tails in anyway that we please.

For every α, $p \in (0, 1)$, a $1 - \alpha$ confidence interval for $\nu - \mu$ is

$$\left[M(\boldsymbol{Y}) - M(\boldsymbol{X}) - z(1 - \alpha p)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}, M(\boldsymbol{Y}) - M(\boldsymbol{X}) - z(\alpha - p\alpha)\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}}\right] \tag{8.4.4}$$

1. $p = \frac{1}{2}$ gives the symmetric two-sided interval.
2. $p \to 1$ gives the interval with the confidence lower bound.
3. $p \to 0$ gives the interval with confidence upper bound.

Proof

From the distribution of the pivot variable and the definition of the quantile function,

$$\mathbb{P}\left[z(\alpha - p\alpha) < \frac{[M(\boldsymbol{Y}) - M(\boldsymbol{X})] - (\nu - \mu)}{\sqrt{\sigma^2/m + \tau^2/n}} < z(1 - p\alpha)\right] = 1 - \alpha \tag{8.4.5}$$

Solving for $\nu - \mu$ in the inequality gives the confidence interval.

The following theorem gives some basic properties of the length of this interval.

The (deterministic) length of the general two-sided confidence interval is

$$L = [z(1 - \alpha p) - z(\alpha - \alpha p)]\sqrt{\frac{\sigma^2}{m} + \frac{\tau^2}{n}} \tag{8.4.6}$$

1. L is a decreasing function of m and a decreasing function of n.
2. L is an increasing function of σ and an increasing function of τ
3. L is an decreasing function of α and hence an increasing function of the confidence level.
4. As a function of p, L decreases and then increases, with minimum value at $p = \frac{1}{2}$.

Part (a) means that we can make the estimate more precise by increasing either or both sample sizes. Part (b) means that the estimate becomes less precise as the variance in either distribution increases. Part (c) we have seen before. All other things being equal, we can increase the confidence level only at the expense of making the estimate less precise. Part (d) means that the symmetric, equal-tail confidence interval is the best of the two-sided intervals.

Confidence Intervals for the Difference of the Means with Unknown Variances

Our next method is a construction of confidence intervals for the difference of the means $\nu - \mu$ without needing to know the standard deviations σ and τ. However, there is a cost; we will assume that the standard deviations are the same, $\sigma = \tau$, but the common value is unknown. This assumption is reasonable if there is an inherent variability in the measurement variables that does not change even when different treatments are applied to the objects in the population. We need to recall some basic facts from our study of special properties of normal samples.

The *pooled estimate* of the common variance $\sigma^2 = \tau^2$ is

$$S^2(\boldsymbol{X}, \boldsymbol{Y}) = \frac{(m-1)S^2(\boldsymbol{X}) + (n-1)S^2(\boldsymbol{Y})}{m+n-2} \tag{8.4.7}$$

The random variable

$$T = \frac{[M(\boldsymbol{Y}) - M(\boldsymbol{X})] - (\nu - \mu)}{S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{1/m + 1/n}} \tag{8.4.8}$$

has the student t distribution with $m + n - 2$ degrees of freedom

Note that $S^2(\boldsymbol{X}, \boldsymbol{Y})$ is a weighted average of the sample variances, with the degrees of freedom as the weight factors. Note also that T is a pivot variable for $\nu - \mu$ and so we can construct confidence intervals for $\nu - \mu$ in the usual way.

For $\alpha \in (0, 1)$,

1. $\left[M(\boldsymbol{Y}) - M(\boldsymbol{X}) - t_{m+n-2}\left(1 - \frac{\alpha}{2}\right) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}, M(\boldsymbol{Y}) - M(\boldsymbol{X}) + t_{m+n-2}\left(1 - \frac{\alpha}{2}\right) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}\right]$ is a $1 - \alpha$ confidence interval for $\nu - \mu$.
2. $M(\boldsymbol{Y}) - M(\boldsymbol{X}) - t_{m+n-2}(1 - \alpha) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}$ is a $1 - \alpha$ confidence lower bound for $\nu - \mu$.
3. $M(\boldsymbol{Y}) - M(\boldsymbol{X}) + t_{m+n-2}(1 - \alpha) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}$ is a $1 - \alpha$ confidence upper bound for $\nu - \mu$.

Proof

The variable T given above has the standard normal distribution. Hence each of the following events has probability $1 - \alpha$ by definition of the quantiles:

1. $\left\{-t_{m+n-2}\left(1 - \frac{\alpha}{2}\right) \le T \le t_{m+n-2}\left(1 - \frac{\alpha}{2}\right)\right\}$
2. $\{T \ge t_{m+n-2}(1 - \alpha)\}$
3. $\{T \le -t_{m+n-2}(1 - \alpha)\}$

In each case, solving the inequality for $\nu - \mu$ gives the result.

The two-sided interval in part (a) is the *symmetric interval* corresponding to $\alpha/2$ in both tails of the student t distribution. As usual, we can construct more general two-sided intervals by partitioning α between the left and right tails in anyway that we please.

For every $\alpha,\ p \in (0, 1)$, a $1 - \alpha$ confidence interval for $\nu - \mu$ is

$$\left[M(\boldsymbol{Y}) - M(\boldsymbol{X}) - t_{m+n-2}(1 - \alpha p) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}, M(\boldsymbol{Y}) - M(\boldsymbol{X}) - t_{m+n-2}(\alpha - p\alpha) S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}}\right] \tag{8.4.9}$$

1. $p = \frac{1}{2}$ gives the symmetric two-sided interval.
2. $p \to 1$ gives the interval with the confidence lower bound.
3. $p \to 0$ gives the inteval with confidence upper bound.

Proof

From the distribution of the pivot variable and the definition of the quantile function,

$$\mathbb{P}\left[t_{m+n-2}(\alpha - p\alpha) < \frac{[M(\boldsymbol{Y}) - M(\boldsymbol{X})] - (\nu - \mu)}{S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{1/m + 1/n}} < t_{m+n-2}(1 - p\alpha)\right] = 1 - \alpha \tag{8.4.10}$$

Solving for $\nu - \mu$ in the inequality gives the confidence interval.

The next result considers the length of the general two-sided interval.

The (random) length of the two-sided interval above is

$$L = [t_{m+n-2}(1-p\alpha) - t_{m+n-2}(\alpha - p\alpha)]S(\boldsymbol{X}, \boldsymbol{Y})\sqrt{\frac{1}{m} + \frac{1}{n}} \tag{8.4.11}$$

1. L is an decreasing function of α and hence an increasing function of the confidence level.
2. As a function of p, L decreases and then increases, with minimum value at $p = \frac{1}{2}$.

As in the case of known variances, part (c) means that all other things being equal, we can increase the confidence level only at the expense of making the estimate less precise. Part (b) means that the symmetric, equal-tail confidence interval is the best of the two-sided intervals.

Confidence Intervals for the Ratio of the Variances

Our next construction will produce interval estimates for the ratio of the variances τ^2/σ^2 (or by taking square roots, for the ratio of the standard deviations τ/σ). Once again, we need to recall some basic facts from our study of special properties of random samples from the normal distribution.

The ratio

$$U = \frac{S^2(\boldsymbol{X})\tau^2}{S^2(\boldsymbol{Y})\sigma^2} \tag{8.4.12}$$

has the F distribution with $m-1$ degrees of freedom in the numerator and $n-1$ degrees of freedom in the denominator, and hence this variable is a pivot variable for τ^2/σ^2.

The pivot variable U can be used to construct confidence intervals for τ^2/σ^2 in the usual way.

For $\alpha \in (0,1)$,

1. $\left[f_{m-1,n-1}\left(\frac{\alpha}{2}\right)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}, f_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}\right]$ is a $1-\alpha$ confidence interval for τ^2/σ^2.
2. $f_{m-1,n-1}(1-\alpha)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}$ is a $1-\alpha$ confidence lower bound for τ^2/σ^2.
3. $f_{m-1,n-1}(\alpha)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}$ is a $1-\alpha$ confidence upper bound for $\nu - \mu$.

Proof

The variable U given above has the F distribution with $m-1$ degrees of freedom in the numerator and $n-1$ degrees of freedom in the denominator. Hence each of the following events has probability $1-\alpha$ by definition of the quantiles:

1. $\left\{f_{m-1,n-1}\left(\frac{\alpha}{2}\right) \le U \le f_{m-1,n-1}\left(1-\frac{\alpha}{2}\right)\right\}$
2. $\{U \ge f_{m-1,n-1}(1-\alpha)\}$
3. $\{U \le fm-1,n-1(\alpha)\}$

In each case, solving the inequality for τ^2/σ^2 gives the result.

The two-sided confidence interval in part (a) is the *equal-tail* confidence interval, and is the one commonly used. But as usual, we can partition α between the left and right tails of the distribution of the pivot variable in any way that we please.

For every α, $p \in (0,1)$, a $1-\alpha$ confidence set for τ^2/σ^2 is

$$\left[f_{m-1,n-1}(\alpha - p\alpha)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}, f_{m-1,n-1}(1-p\alpha)\frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})}\right] \tag{8.4.13}$$

1. $p = \frac{1}{2}$ gives the equal-tail, two-sided interval.
2. $p \to 1$ gives the interval with the confidence lower bound.
3. $p \to 0$ gives the inteval with confidence upper bound.

Proof

From the F pivot variable and the definition of the quantile function,

$$\mathbb{P}\left[f_{m-1,n-1}(\alpha - p\,\alpha) < \frac{S^2(\boldsymbol{X},\mu)\tau^2}{S^2(\boldsymbol{Y},\nu)\sigma^2} < f_{m-1,n-1}(1-p\,\alpha)\right] = 1-\alpha \tag{8.4.14}$$

Solving for τ^2/σ^2 in the inequality.

The length of the general confidence interval is considered next.

The (random) length of the general two-sided confidence interval above is

$$L = [f_{m-1,n-1}(1-p\alpha) - f_{m-1,n-1}(\alpha - p\alpha)] \frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})} \tag{8.4.15}$$

Assuming that $m > 5$ and $n > 1$,

1. L is an decreasing function of α and hence an increasing function of the confidence level.
2. $\mathbb{E}(L) = \frac{\tau^2}{\sigma^2} \frac{m-1}{m-3}$
3. $\operatorname{var}(L) = 2 \frac{\tau^4}{\sigma^4} \left(\frac{m-1}{m-3} \right)^2 \frac{m+n-4}{(n-1)(m-5)}$

Proof

Parts (b) and (c) follow since $\frac{\sigma^2}{\tau^2} \frac{S^2(\boldsymbol{Y})}{S^2(\boldsymbol{X})^2}$ as the F distribution with $n-1$ degrees of freedom in the numerator and $m-1$ degrees of freedom in the denominator.

Optimally, we might want to choose p so that $\mathbb{E}(L)$ is minimized. However, this is difficult computationally, and fortunately the equal-tail interval with $p = \frac{1}{2}$ is not too far from optimal when the sample sizes m and n are large.

Estimation in the Bivariate Normal Model

In this subsection, we consider a model that is superficially similar to the two-sample normal model, but is actually much simpler. Suppose that

$$((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \tag{8.4.16}$$

is a random sample of size n from the bivariate normal distribution of a random vector (X, Y), with $\mathbb{E}(X) = \mu$, $\mathbb{E}(Y) = \nu$, $\operatorname{var}(X) = \sigma^2$, $\operatorname{var}(Y) = \tau^2$, and $\operatorname{cov}(X, Y) = \delta$.

Thus, instead of a *pair of samples*, we have a *sample of pairs*. This type of model frequently arises in *before and after experiments*, in which a measurement of interest is recorded for a sample of n objects from the population, both before and after a treatment. For example, we could record the blood pressure of a sample of n patients, before and after the administration of a certain drug. The critical point is that in this model, X_i and Y_i are measurements made on the same underlying object in the sample. As with the two-sample normal model, the interest is usually in estimating the difference of the means.

We will use our usual notation for the sample means and variances of $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$. Recall also that the sample covariance of $(\boldsymbol{X}, \boldsymbol{Y})$, is

$$S(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n-1} \sum_{i=1}^{n} [X_i - M(\boldsymbol{X})][Y_i - M(\boldsymbol{Y})] \tag{8.4.17}$$

(not to be confused with the pooled estimate of the standard deviation in the two sample model).

The vector of differences $\boldsymbol{Y} - \boldsymbol{X} = (Y_1 - X_1, Y_2 - X_2, \ldots, Y_n - X_n)$ is a random sample of size n from the distribution of $Y - X$, which is normal with

1. $\mathbb{E}(Y - X) = \nu - \mu$
2. $\operatorname{var}(Y - X) = \sigma^2 + \tau^2 - 2\,\delta$

The sample mean and variance of the sample of differences are given by

1. $M(\boldsymbol{Y} - \boldsymbol{X}) = M(\boldsymbol{Y}) - M(\boldsymbol{X})$
2. $S^2(\boldsymbol{Y} - \boldsymbol{X}) = S^2(\boldsymbol{X}) + S^2(\boldsymbol{Y}) - 2\,S(\boldsymbol{X}, \boldsymbol{Y})$

Thus, the sample of differences $\boldsymbol{Y} - \boldsymbol{X}$ fits the normal model for a single variable. The section on Estimation in the Normal Model could be used to obtain confidence sets and intervals for the parameters $(\nu - \mu, \sigma^2 + \tau^2 - 2\,\delta)$.

In the setting of this subsection, suppose that $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ are independent. Mathematically this fits both models—the two-sample normal model and the bivariate normal model. Which procedure would work better for estimating the difference of means $\nu - \mu$?

1. If the standard deviations σ and τ are known.
2. If the standard deviations σ and τ are unknown.

Answer

1. The two methods are equivalent.
2. The bivariate normal model works better.

Although the setting in the last problem fits both models *mathematically*, only one model would make sense in a real problem. Again, the critical point is whether (X_i, Y_i) makes sense as a pair of random variables (measurements) corresponding to a given object in the sample.

Computational Exercises

A new drug is being developed to reduce a certain blood chemical. A sample of 36 patients are given a placebo while a sample of 49 patients are given the drug. Let X denote the measurement for a patient given the placebo and Y the measurement for a patient given the drug (in mg). The statistics are $m(\boldsymbol{x}) = 87$, $s(\boldsymbol{x}) = 4$, $m(\boldsymbol{y}) = 63$, $s(\boldsymbol{y}) = 6$.

1. Compute the 90% confidence interval for τ / σ.
2. Assuming that $\sigma = \tau$, compute the 90% confidence interval for $\nu - \mu$.
3. Based on (a), is the assumption that $\sigma = \tau$ reasonable?
4. Based on (b), is the drug effective?

Answer

1. $(1.149, 1.936)$
2. $(-24.834, -23.166)$
3. Perhaps not.
4. Yes

A company claims that an herbal supplement improves intelligence. A sample of 25 persons are given a standard IQ test before and after taking the supplement. Let X denote the IQ of a subject before taking the supplement and Y the IQ of the subject after the supplement. The before and after statistics are $m(\boldsymbol{x}) = 105$, $s(\boldsymbol{x}) = 13$, $m(\boldsymbol{y}) = 110$, $s(\boldsymbol{y}) = 17$, $s(\boldsymbol{x}, \boldsymbol{y}) = 190$. Do you believe the company's claim?

Answer

A 90% confidence lower bound for the difference in IQ is 2.675. There may be a vary small increase.

In Fisher's iris data, let X denote consider the petal length of a Versicolor iris and Y the petal length of a Virginica iris.

1. Compute the 90% confidence interval for τ / σ.
2. Assuming that $\sigma = \tau$, compute the 90% confidence interval for $\nu - \mu$.
3. Based on (a), is the assumption that $\sigma = \tau$ reasonable?

Answer

1. $(0.8, 1.3)$
2. $(10.5, 14.1)$
3. Yes

A plant has two machines that produce a circular rod whose diameter (in cm) is critical. Let X denote the diameter of a rod from the first machine and Y the diameter of a rod from the second machine. A sample of 100 rods from the first machine as mean 10.3 and standard deviation 1.2. A sample of 100 rods from the second machine has mean 9.8 and standard deviation 1.6.

1. Compute the 90% confidence interval for τ / σ.
2. Assuming that $\sigma = \tau$, compute the 90% confidence interval for $\nu - \mu$.
3. Based on (a), is the assumption that $\sigma = \tau$ reasonable?

Answer

1. $(1.127, 1.578)$
2. $(0.832, 0.168)$
3. Perhaps not.