Homework 8

Aleck Zhao

November 29, 2016

4. When measuring variability, we often think of it in relative terms. Suppose we define the *relative variability* of a random variable to be the ratio of its standard deviation to its mean:

$$r = \frac{\sqrt{\text{Var}(X)}}{E[X]}$$

Suppose we observe $X_1, \cdots, X_n$ that are iid $N(\mu, \sigma^2)$ where μ and σ re both unknown constants. If we denote the *population* relative variability $r = \sigma/\mu$, we want to test the hypothesis

$$H_0 : r = r_0 \quad \text{v.s.} \quad H_a : r \neq r_0$$

where r_0 is a specified constant. Under the null hypothesis the parameter μ is only constrained to be positive. Let θ denote the pair (μ, σ).

(a) Write down the parameter space Θ for θ under $H_0 \cup H_a$ and Θ_0 for θ under H_0.

Solution. Under $H_0 \cup H_a$, the value r can take on any real value, since for normal random variables,

$$r = \frac{\sqrt{\text{Var}(X)}}{E[X]} = \frac{\sigma}{\mu}$$

which can be anything. Thus, under $H_0 \cup H_a$, we have $\theta \in \mathbb{R}^2$.
Under H_0, μ is positive, and

$$r = \frac{\sigma}{\mu} = r_0$$

so the parameter space is restricted to the line in $\mathbb{R}^2$ corresponding to the line $\sigma = r_0\mu$. □

(b) What are the dimensions of the parameter spaces in a) and what is the difference in the two dimensions?

Answer. The dimension of Θ is 2, and the dimension of Θ_0 is 1, and their difference is 1.

(c) Write down the likelihood function under the null hypothesis H_0 as a function of the unknown parameter μ and to keep it in a simple form, express it in terms of the quantities

$$S_1 = \sum_{i=1}^{n} X_i, \qquad S_2 = \sum_{i=1}^{n} X_i^2$$

Solution. Under H_0, we have

$$\sigma = r_0\mu \implies \sigma^2 = r_0^2\mu^2$$

so the distribution of X_i is $N\left(\mu, r_0^2\mu^2\right)$. Thus, the likelihood function is

$$\begin{aligned}\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}(r_0\mu)} \exp\left(-\frac{(X_i-\mu)^2}{2r_0^2\mu^2}\right) &= \frac{1}{(2\pi)^{n/2} r_0^n \mu^n} \exp\left[-\frac{1}{2r_0^2\mu^2}\left(\sum_{i=1}^{n}(X_i-\mu)^2\right)\right] \\ &= \frac{1}{(2\pi)^{n/2} r_0^n \mu^n} \exp\left(\frac{-S_2 + 2\mu S_1 - n\mu^2}{2r_0^2\mu^2}\right)\end{aligned}$$

□

(d) Differentiate the log-likelihood in c) and show that when the derivative is equated to zero, we need to take μ to be the solution to a certain quadratic equation.

Proof. The log-likelihood is given by

$$\begin{aligned}\ell(\mu) &= \log\left[\frac{1}{(2\pi)^{n/2} r_0^n \mu^n} \exp\left(\frac{-S_2 + 2\mu S_1 - n\mu^2}{2r_0^2\mu^2}\right)\right] \\ &= -\frac{n}{2}\log(2\pi) - n\log r_0 - n\log\mu + \frac{-S_2 + 2\mu S_1 - n\mu^2}{2r_0^2\mu^2}\end{aligned}$$

and its derivative with respect to μ is given by

$$\frac{\partial}{\partial\mu}\ell(\mu) = -\frac{n}{\mu} + \frac{S_2 - \mu S_1}{r_0^2\mu^3} = \frac{-nr_0^2\mu^2 - \mu S_1 + S_2}{r_0^2\mu^3}$$

If we equate this with zero, then the numerator must be 0, which is a quadratic in μ, and μ is the solution to this quadratic, as desired. □

(e) Which root in the previous part must be the desired MLE for μ under H_0 and why? Show that the second derivative of the log-likelihood in c) is negative when evaluated at the MLE.

Proof. The solutions to the quadratic are given by

$$\mu = \frac{-S_1 \pm \sqrt{S_1^2 + 4nr_0^2 S_2}}{2nr_0^2}$$

and under the null hypothesis, we restrict $\mu > 0$, so the desired MLE is the positive root, or

$$\hat{\mu} = \frac{-S_1 + \sqrt{S_1^2 + 4nr_0^2 S_2}}{2nr_0^2}$$

The second derivative of the log-likelihood evaluated at $\hat{\mu}$ is given by

$$\begin{aligned}\frac{\partial}{\partial\mu}\left[\frac{-nr_0^2\mu^2 - \mu S_1 + S_2}{r_0^2\mu^3}\right]\Bigg|_{\hat{\mu}} &= \frac{r_0^2\mu^3(-2nr_0^2 - S_1) - 3r_0^2\mu^2(-nr_0^2\mu^2 - \mu S_1 + S_2)}{(r_0^2\mu^3)^2}\Bigg|_{\hat{\mu}} \\ &= \frac{-2nr_0^2 - S_2}{r_0^2\hat{\mu}^3} < 0\end{aligned}$$

since the right part of the numerator evaluates to 0. This is less than 0 because S_2 is a sum of squares, and nr_0^2 is also positive. Thus, the second derivative of the log-likelihood is negative when evaluated at the MLE, as desired. □

(f) Explain why the MLE for μ under H_0 must be the root found in e)

Answer. Since the second derivative of the log-likelihood is negative and the first derivative is 0 when evaluated at $\hat{\mu}$, this is the value that maximizes the log-likelihood. Since the logarithm is an increasing function, this is equivalent to maximizing the likelihood function, which is the definition of an MLE.

(g) Write down an expression for the MLE for μ under H_0 in terms of the first two sample moments $\hat{\mu}_1$ and $\hat{\mu}_2$.

Answer. We have

$$S_1 = \sum_{i=1}^n X_i = n\hat{\mu}_1, \qquad S_2 = \sum_{i=1}^n X_i^2 = n\hat{\mu}_2$$

so substituting these into our expression for $\hat{\mu}$, we have

$$\hat{\mu} = \frac{-n\hat{\mu}_1 + \sqrt{(n\hat{\mu}_1)^2 + 4nr_0^2(n\hat{\mu}_2)}}{2nr_0^2} = \frac{-\hat{\mu}_1 + \sqrt{\hat{\mu}_1^2 + 4r_0^2\hat{\mu}_2}}{2r_0^2}$$

(h) Describe how you could, in principle, compute a delta method approximation to the mean and variance of the MLE for μ under H_0 using the expression you get in h). Express the required variances and covariances in terms of the quantities μ and r_0.

Solution. We would calculate

$$E[\hat{\mu}] = E\left[\frac{-\hat{\mu}_1 + \sqrt{\hat{\mu}_1^2 + 4r_0^2\hat{\mu}_2}}{2r_0^2}\right] = -\frac{\mu}{2r_0^2} + \frac{1}{2r_0^2}E\left[\sqrt{\hat{\mu}_1^2 + 4r_0^2\hat{\mu}_2}\right]$$

For the expectation, we can use a Taylor expansion of $\sqrt{\hat{\mu}_1^2 + 4r_0^2\hat{\mu}_2}$ about the point (μ, σ^2). Let $f(\hat{\mu}_1, \hat{\mu}_2) = \sqrt{\hat{\mu}_1^2 + 4r_0^2\hat{\mu}_2}$. The Taylor expansion about $(\mu, r_0^2\mu^2)$ is approximately

$$\begin{aligned}
&f(\mu,\sigma^2) + (\hat{\mu}_1 - \mu)\frac{\partial}{\partial\hat{\mu}_1}f(\mu,\sigma^2) + (\hat{\mu}_2 - \sigma^2)\frac{\partial}{\partial\hat{\mu}_2}f(\mu,\sigma^2) \\
&+ (\hat{\mu}_1 - \mu)^2\frac{\partial^2}{\partial\hat{\mu}_1^2}f(\mu,\sigma^2) + (\hat{\mu}_2 - \sigma^2)^2\frac{\partial^2}{\partial\hat{\mu}_2^2}f(\mu,\sigma^2) + 2(\hat{\mu}_1 - \mu)(\hat{\mu}_2 - \sigma^2)\frac{\partial^2}{\partial\hat{\mu}_1\hat{\mu}_2}f(\mu,\sigma^2)
\end{aligned}$$

If we pull the expectation through this, the first degree terms go to zero. Then we have

$$\begin{aligned}
E\left[(\hat{\mu}_1 - \mu)^2\right] &= \mathrm{Var}(\hat{\mu}_1) = \frac{\sigma^2}{n} = \frac{r_0^2\mu^2}{n} \\
E\left[(\hat{\mu}_2 - \sigma^2)^2\right] &= \mathrm{Var}(\hat{\mu}_2) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i^2\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i^2\right) \qquad (1) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i^2) = \frac{1}{n^2}\sum_{i=1}^n \left(E[X_i^4] - (E[X_i^2])^2\right)
\end{aligned}$$

For the expectations, we use the fact that $E[X^k] = E[(\mu + \sigma Z)^k]$ and the various moments of the standard normal variable Z :

$$\begin{aligned}
E[X_i^2] &= E[(\mu + \sigma Z)^2] = E\left[\mu^2 + 2\mu\sigma Z + \sigma^2 Z^2\right] \\
&= \mu^2 + 2\mu\sigma E[Z] + \sigma^2 E[Z^2] \\
&= \mu^2 + \sigma^2 = \mu^2(1 + r_0^2) \\
E[X_i^4] &= E[(\mu + \sigma Z)^4] = E\left[\mu^4 + 4\mu^3\sigma Z + 6\mu^2\sigma^2 Z^2 + 4\mu\sigma^3 Z^3 + \sigma^4 Z^4\right] \\
&= \mu^4 + 4\mu^3\sigma E[Z] + 6\mu^2\sigma^2 E[Z^2] + 4\mu\sigma^3 E[Z^3] + \sigma^4 E[Z^4] \\
&= \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 = \mu^4(1 + 6r_0^2 + 3r_0^4)
\end{aligned}$$

where we replaced σ^2 by $r_0^2\mu^2$ since we are under H_0. Thus, the expectation in (1) is given by

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^n \left[\mu^4(1 + 6r_0^2 + 3r_0^4) + \mu^4(1 + r_0^2)^2\right] &= \frac{\mu^4\left[(1 + 6r_0^2 + 3r_0^4) + (1 + 2r_0^2 + r_0^4)\right]}{n} \\
&= \frac{\mu^4(2 + 8r_0^2 + 4r_0^4)}{n}
\end{aligned}$$

Next, we have

$$\begin{aligned}
E[(\hat{\mu}_1 - \mu)(\hat{\mu}_2 - \sigma^2)] = \mathrm{Cov}(\hat{\mu}_1, \hat{\mu}_2) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{j=1}^n X_j^2\right) = \frac{1}{n^2}\mathrm{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathrm{Cov}(X_i, X_j^2) = \frac{1}{n^2}\sum_{i=j}\mathrm{Cov}(X_i, X_j^2) = \frac{1}{n^2}\sum_{k=1}^n \mathrm{Cov}(X_k, X_k^2) \\
&= \frac{1}{n^2}\sum_{k=1}^n (E[X_k^3] - E[X_k]E[X_k^2])
\end{aligned}$$

In a similar method as above, we have

$$\begin{aligned} E[X_k^3] = E[(\mu+\sigma Z)^3] &= E\left[\mu^3 + 3\mu^2\sigma Z + 3\mu\sigma^2 Z^2 + \sigma^3 Z^3)\right] \\ &= \mu^3 + 3\mu^2\sigma E[Z] + 3\mu\sigma^2 E[Z^2] + \sigma^3 E[Z^3] \\ &= \mu^3 + 3\mu\sigma^2 = \mu^3(1+3r_0^2) \end{aligned}$$

so the covariance is given by

$$\frac{1}{n^2}\sum_{k=1}^{n}\left[\mu^3(1+3r_0^2) - \mu^3(1+r_0^2)\right] = \frac{1}{n^2}\sum_{k=1}^{n} 2r_0^2\mu^3 = \frac{2r_0^2\mu^3}{n}$$

Thus, the approximate mean of the MLE is given by

$$E[\hat{\mu}] \approx f(\mu,\sigma^2) + \frac{r_0^2\mu^2}{n}\frac{\partial}{\partial\hat{\mu}_1}f(\mu,\sigma^2) + \frac{\mu^4(2+8r_0^2+4r_0^4)}{n}\frac{\partial}{\partial\hat{\mu}_2}f(\mu,\sigma^2) + \frac{4r_0^2\mu^3}{n}\frac{\partial}{\partial\hat{\mu}_1\hat{\mu}_2}f(\mu,\sigma^2)$$

where we may substitute $r_0^2\mu^2$ for σ^2 since we are under H_0 when evaluating f and its derivatives. To calculate the variance of the MLE, use just the first order expansion:

$$f(\mu,\sigma^2) + (\hat{\mu}_1-\mu)\frac{\partial}{\partial\hat{\mu}_1}f(\mu,\sigma^2) + (\hat{\mu}_2-\sigma^2)\frac{\partial}{\partial\hat{\mu}_2}f(\mu,\sigma^2)$$

If we pull the variance through this, we get

$$\begin{aligned} &\mathrm{Var}\left(f(\mu,\sigma^2) + (\hat{\mu}_1-\mu)\frac{\partial}{\partial\hat{\mu}_1}f(\mu,\sigma^2) + (\hat{\mu}_2-\sigma^2)\frac{\partial}{\partial\hat{\mu}_2}f(\mu,\sigma^2)\right) \\ &= \frac{\partial}{\partial\hat{\mu}_1}f(\mu,\sigma^2)\mathrm{Var}(\hat{\mu}_1) + \frac{\partial}{\partial\mu_2}f(\mu,\sigma^2)\mathrm{Var}(\hat{\mu}_2) + 2\frac{\partial}{\partial\mu_1}f(\mu,\sigma^2)\frac{\partial}{\partial\mu_2}f(\mu,\sigma^2)\mathrm{Cov}(\hat{\mu}_1,\hat{\mu}_2) \end{aligned}$$

since constant terms vanish within a variance. Using our results from above, we conclude that

$$\mathrm{Var}(\hat{\mu}) \approx \frac{r_0^2\mu^2}{n}\frac{\partial}{\partial\mu_1}f(\mu,\sigma^2) + \frac{\mu^4(2+8r_0^2+4r_0^4)}{n}\frac{\partial}{\partial\mu_2}f(\mu,\sigma^2) + \frac{4r_0^2\mu^3}{n}\frac{\partial}{\partial\mu_1}f(\mu,\sigma^2)\frac{\partial}{\partial\mu_2}f(\mu,\sigma^2)$$

□

(i) What is the MLE for $\theta = (\mu,\theta)$ under $H_0 \cup H_a$?

Solution. The likelihood function under this parameter space is given by

$$f(X_1,\cdots,X_n \mid \mu,\sigma) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)$$

and the log-likelihood is given by

$$\begin{aligned} \ell(\mu,\sigma) &= \log\left[\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right] \\ &= \sum_{i=1}^{n}\log\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\right] \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 \end{aligned}$$

Taking the partial derivatives and setting equal to 0, we have

$$\frac{\partial}{\partial\sigma}\ell(\mu,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$
$$\frac{\partial}{\partial\mu}\ell(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

The second equation gives us

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

and substituting this into the first equation, we get

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Thus, the MLE for θ is given by

$$\hat{\theta} = \left(\bar{X}, \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)$$

□

(j) Write an R function that takes as input a data vector X and hypothesized relative variability r_0 and outputs the MLE for $\theta = (\mu, \sigma)$ under H_0.

(k) Do the following 1000 times. Sample $X_1, \cdots, X_{25}$ from the $N(10, 2.5^2)$ distribution and compute the MLE for μ under $H_0 : r = 0.25$. Estimate the bias and variance of the MLE under the assumption that H_0 is true and compare with the bias and variance of the usual MLE for μ. Is it the case that knowing the relative variability leads to a better estimator of μ than would otherwise be available?

Answer. Under H_0, the bias was -0.0081, whereas the bias with the usual estimator (the sample mean) was 0.0189. The variance of the MLE under H_0 was 0.2072, whereas the bias with the usual estimator (the sample variance) was 0.2389. In both cases, the MLE under H_0 is better because it has less variance and less bias than the usual estimator.

(l) Assume that the data in `data.hw8.csv` are a sample of size 25 from a $N(\mu, \sigma^2)$ distribution. Compute the sample mean, the same standard deviation and estimate the relative mean using the ratio of the two. Does it seem plausible that σ/μ is not 0.25?

Answer. The sample mean is 10.29126. The sample SD is 3.873. The relative mean is

$$3.873/10.29126 = 0.376 \neq 0.25.$$

(m) Write R functions to compute the MLE for (μ, σ) under $H_0 \cup H_a$, and a function to compute the log-likelihood as a function of the data vector and parameter vector θ. Use this to compute the GLRT Λ and $-2\log\Lambda$ and use this to find an approximate p-value for the test of the hypothesis for the relative variability

$$H_0 : r = 0.25 \quad \text{v.s.} \quad H_a : r \neq 0.25$$

Answer. From the data, I computed $-2\log\Lambda = 8.159$. This is approximately a chi-square distribution with 1 degree of freedom, which has p-value 0.004.

Chapter 11: Comparing Two Samples

1. A computer was used to generated four random numbers from a normal distribution with a set mean and variance: 1.1650, 0.6268, 0.0751, 0.3516. Five more random normal numbers with the same variance but perhaps a different mean were then generated (the mean may or may not actually be different): 0.3035, 2.6961, 1.0591, 2.7971, 1.2641.

   a. What do you think the means of the random normal number generators were? What do you think the difference of the means was?

   Solution. Using the sample mean, the mean of the first RNG would be 0.5546. The mean of the second RNG would be 1.6240. The difference in means would be 1.0694. □

   b. What do you think the variance of the random number generator was?

   Solution. The variance of the first batch of numbers was 0.4651, and the sample variance of the second batch was 1.086. The pooled sample variance is given by

   $$\frac{3(0.4651) + 4(1.086)}{7} = 0.8199$$

   which would be the variance of the RNG. □

   c. What is the estimated standard error of your estimate of the difference of the means?

   Solution. The estimated standard error of the difference of means is given by

   $$s_p\sqrt{\frac{1}{n} + \frac{1}{m}} = \sqrt{0}.8199 \cdot \sqrt{\frac{1}{4} + \frac{1}{5}} = 0.6074$$

   □

   d. Form a 90% confidence interval for the difference of the means of the random number generators.

   Solution. The difference of means from part a was 1.0694, and we have $t_7(10/2) = 1.895$, so the 90% confidence interval is given by

   $$1.0694 \pm 1.895(0.6074) = (-0.0816, 2.2204)$$

   □

   e. In this situation, is it more appropriate to use a one-sided test or a two-sided test of the equality of the means?

   Answer. It is more appropriate to use a two-sided test because we don't know anything about the means, and we already have a confidence interval.

   f. What is the p-value of a two-sided test of the null hypothesis of equal means?

   Solution. Under the null hypothesis, the means are equal, so the test statistic is

   $$t = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{1.0694}{0.6074} = 1.7606$$

   Then since we have a t-distribution, the p-value is

   $$P(t < -1.7606) + P(t > 1.7606) = 2(0.0608) = 0.1216$$

   □

g. Would the hypothesis that the means were the same versus a two-sided alternative be rejected at the significance level $\alpha = 0.1$?

Answer. No, since the p-value is not less than α, we do not reject the null hypothesis.

h. Suppose you know that the variance of the normal distribution was $\sigma^2 = 1$. How would your answers to the preceding questions change?

Answer. Instead of using the pooled sample variance, we could directly use the standard deviation. Then the test statistic would be a standard normal variable instead of a t.

2. The difference of the means of two normal distributions with equal variance is to be stimated by sampling an equal number of observations from each distribution. If it were possible, would it be better to halve the standard deviation of the populations or double the sample sizes?

Answer. Let X_i have distribution $N(\mu_X, \sigma_X^2)$ and Y_i have distribution $N(\mu_Y, \sigma_Y^2)$. If we draw n from each sample, then $\bar{X} - \bar{Y}$ has distribution $N\left(\mu_X - \mu_Y, \frac{\sigma_X^2 + \sigma_Y^2}{n}\right)$. If we halve the SD of each sample, then this has the effect of reducing the variance of $\bar{X} - \bar{Y}$ by a factor of 4, while if we double the sample sizes, this only reduces the variance by a factor of 2. Thus, it would be better to halve the SD of the populations.

3. In Section 11.2.1, we considered two methods of estimating $\text{Var}(\bar{X} - \bar{Y})$. Under the assumption that the two population variances were equal, we estimated this quantity by

$$s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right)$$

and without this assumption by

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

Show that these two estimates are identical if $m = n$.

Proof. If $n = m$, then we have

$$\begin{aligned} s_p^2 \left(\frac{1}{n} + \frac{1}{m}\right) &= \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2} \left(\frac{1}{n} + \frac{1}{m}\right) \\ &= \frac{(n-1)s_X^2 + (n-1)s_Y^2}{2(n-1)} \cdot \frac{2}{n} \\ &= \frac{s_X^2 + s_Y^2}{n} = \frac{s_X^2}{n} + \frac{s_Y^2}{n} \\ &= \frac{s_X^2}{n} + \frac{s_Y^2}{m} \end{aligned}$$

as desired. □

10. Verify that the two-sample t test at level α of $H_0 : \mu_X = \mu_Y$ versus $H_A : \mu_X \neq \mu_Y$ rejects if and only if the confidence interval for $\mu_X - \mu_Y$ does not contain zero.

Proof. Suppose we are at the significance level α with $n + m - 2$ degrees of freedom. The confidence interval for $\mu_X - \mu_Y$ is given by

$$(\bar{X} - \bar{Y}) \pm t_{n+m-2}(\alpha/2) s_{\bar{X}-\bar{Y}}$$

If the confidence interval does not contain 0, WLOG $\bar{X} > \bar{Y}$, then we have

$$\frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} > t_{n+m-2}(\alpha/2)$$

However, this is exactly the condition where we reject the null hypothesis. Since each step here was invertible, the reverse direction holds as well. □

11. Explain how to modify the t test of Section 11.2.1 to test $H_0 : \mu_X = \mu_Y + \Delta$ versus $H_A : \mu_X \neq \mu_Y + \Delta$ where Δ is specified.

Answer. In this case, we have

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_{\bar{X}-\bar{Y}}} = \frac{\bar{X} - \bar{Y} - \Delta}{s_{\bar{X}-\bar{Y}}}$$

under the null hypothesis. Everything else is the same.

Chapter 12: The Analysis of Variance

2. Verify that if $I = 2$, the estimate s_p^2 of Theorem A of Section 11.2.1 is the s_p^2 given in Section 12.2.1.

Proof. The estimate s_p^2 of Theorem A Section 11.2.1 is given by

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

and in section 12.2.1 it is

$$s_p^2 = \frac{SS_W}{I(J-1)} = \frac{1}{2(J-1)} \sum_{i=1}^{2} (J-1)s_i^2 = \frac{s_1^2 + s_2^2}{2}$$

if $I = 2$. Then since $n = m$ in this assumption, the first estimate becomes

$$s_p^2 = \frac{s_X^2 + s_Y^2}{2}$$

and these two expressions are identical, as desired. □

3. For a one-way analysis of variance with $I = 2$ treatment groups, show that the F statistic is t^2, where t is the usual t statistic for a two-sample case.

Proof. We have

$$F = \frac{SS_B/(I-1)}{SS_W/(I(J-1)}$$

and if $I = 2$, then the numerator is a chi-square variable with 1 degree of freedom, which is exactly Z^2. A t statistic is a ratio between a Z and the square-root of a normalized chi-square variable, so squaring a t yields this F statistic, as desired. □

4. Prove the analogues of Theorems A and B in Section 12.2.1 for the case of unequal numbers of observations in the cells of a one-way layout.

Theorem A. Suppose the i-th treatment has $f(i)$ observations. Then we have

$$\begin{aligned} E[SS_W] &= \sum_{i=1}^{I} \sum_{j=1}^{f(i)} E[(Y_{ij} - \bar{Y}_{i\cdot})^2] \\ &= \sum_{i=1}^{I} \sum_{j=1}^{f(i)} \frac{f(i)-1}{f(i)} \sigma^2 = \sum_{i=1}^{I} (f(i)-1)\sigma^2 \end{aligned}$$

Similarly, we have

$$\begin{aligned} E[SS_B] &= \sum_{i=1}^{I}\sum_{j=1}^{f(i)} E[(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2] \\ &= \sum_{i=1}^{I}\sum_{j=1}^{f(i)} \left[\alpha_i^2 + \frac{I-1}{If(i)}\sigma^2\right] = \sum_{i=1}^{I}\left(f(i)\alpha_i^2 + \frac{I-1}{I}\sigma^2\right) \\ &= (I-1)\sigma^2 + \sum_{i=1}^{I} f(i)\alpha_i^2 \end{aligned}$$

□

Theorem B. We have

$$\frac{1}{\sigma^2}\sum_{j=1}^{f(i)}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

follows a chi-square distribution with $f(i)-1$ degrees of freedom. We sum from $i=1$ to $i=I$ and these comprise SS_W, and each is independent. The total number of degrees of freedom of these independent chi-square variables is

$$\sum_{i=1}^{I}(f(i)-1) = -I + \sum_{i=1}^{I} f(i)$$

Then the proof of the second part is identical to the proof in the book. □