MS Module 9: Two-Sample *t* Test and Confidence Interval – practice problems

(The attached PDF file has better formatting.)

Exercise 1.2: Difference of means for small samples

Samples from two groups have the following samples sizes, means, and standard deviations:

Group	*Sample Size*	*Sample Mean*	*Sample SD*
Group 1	9	10	1.2
Group 2	8	15	2.4

μ_1 = the mean of Group #1; μ_2 = the mean of Group #2.
The null hypothesis is H_0: $\mu_1 = \mu_2$; the alternative hypothesis is H_a: $\mu_1 \neq \mu_2$.

A. What is the variance of the estimated mean of each group?
B. What is the variance of the estimated difference of the group means?
C. What is the standard deviation of the estimated difference of the group means?
D. What are the degrees of freedom for a *t* test of each group's mean?
E. What are the degrees of freedom for a *t* test of the difference of the group means?
F. What is the *t* value for a 95% two-sided confidence interval of the difference of the group means?
G. What is the 95% two-sided confidence interval of the difference of the group means ($\mu_1 - \mu_2$)?

Part A: The variance of the estimate of the mean is the variance of the group sample / the sample size:

! Group 1: 1.2^2 / 9 = 0.16
! Group 2: 2.4^2 / 8 = 0.72

Question: The mean is a single value; how does it have a variance?

Answer: The mean of the population is a single value, and the mean of any sample is a single value. The mean of a sample is an estimate of the mean of the population. This estimate is a random variable with a variance.

Part B: The variance of the sum of independent random variables is the sum of the variances. The variance of *k* × a random variable (where *k* is a scalar) is k^2 × the variance of the random variable. The variance of the difference of two random variables = 1^2 × the variance of the first random variable + $(-1)^2$ × the variance of the second random variable, which is the sum of the variances.

For the practice problem, the variance of the difference in the means is 0.16 + 0.72 = 0.88.

Part C: The standard deviation is the square root of the variance: $0.88^{0.5}$ = 0.938083.

Part D: The degrees of freedom for a *t* test of each group's mean is 9 – 1 = 8 for Group 1 and 8 – 1 = 7 for Group 2.

Part E: The textbook has a formula for the approximate degrees of freedom for a *t* test of the difference of the group means. We show the computation and then discuss the rationale for the formula.

The degrees of freedom is approximated by a ratio:

The numerator = $(s^{21}/m + s^{22}/n)^2$ = (variance of group 1 mean + variance of group 1 mean)^2 =

$$(0.16 + 0.72)^2 = 0.7744$$

The denominator = $(s^{21}/m)^2/(m-1) + (s^{22}/n)^2/(n-1)$ =

(square of variance of group 1 mean / (group 1 observations – 1)
\+ square of variance of group 2 mean / (group 2 observations – 1)) =

$$0.16^2 / (9 – 1) + 0.72^2 / (8 – 1) = 0.077257.$$

The approximate degrees of freedom = 0.7744 / 0.077257 = 10.02369.

We truncate the degrees of freedom to 10.

Question: The table of *t* values has degrees of freedom that are integers. Why does the textbook truncate? Why not interpolate?

Answer: This formula is a rough approximation. The textbook uses the next lowest integer to be conservative.

If a group has a normal distribution whose variance is not known but is estimated from the sample data, the estimate of the group mean has a *t* distribution. We use *t* values to test hypotheses.

If a group has a normal distribution with a known variance, the estimate of the group mean has a normal distribution. If the sample size is large and the central limit theorem applies, the estimate of the group mean has close to a normal distribution. We use *z* values to test hypotheses. The *t* value for a large sample is not materially different from the *z* value.

The difference of two normal distributions is also a normal distribution. We use *z* values for testing hypotheses about the difference of the group means.

The difference of two *t* distributions, especially if they have different degrees of freedom, is not a *t* distribution,. The *t* distribution used in the textbook estimates *p* values, but the estimates are not precise.

The *t* distribution is a family of distributions: for lower degrees of freedom, the distribution is more heavy-tailed. We want to choose the *t* distribution that is best for estimating *p* values when testing the difference of means. Several procedures have been suggested for choosing the degrees of freedom.

The textbook uses one such procedure. It gives justification for the procedure; it does not prove it.

It forms a ratio derived from the variances of the estimated group means, or s^{21}/m and s^{22}/n, where s^{21} and s^{22} are the variances of the observations in the groups.

! The numerator of the ratio is $(s^{21}/m + s^{22}/n)^2$
! The denominator of the ratio is $(s^{21}/m)^2/(m-1) + (s^{22}/n)^2/(n-1)$

This ratio is not a whole number. The *t* distribution uses whole numbers for the degrees of freedom. In theory, one might form a smooth distribution of *t* distributions for non-integer degrees of freedom. In practice, one might interpolate between two *t* distributions. But the procedure for differences of means is a rough estimate, and the true distribution of the difference of the group means is not exactly a *t* distribution.

The textbook uses a conservative approach, choosing the integer part of the computed degrees of freedom. The theoretical *p* value is equal to or less than the *p* value computed with the textbook's approach.

Part F: For a two-sided 95% confidence interval, we want the value at which a *t* distribution with 10 degrees of freedom gives a 97.5% cumulative distribution, which is 2.228139 (table lookup or spreadsheet function).

Part G: The two-sided 95% confidence interval is

! lower bound: –5 – 0.938083 × 2.228139 = -7.090179
! upper bound: –5 + 0.938083 × 2.228139 = -2.909821

Question: For final exam problems about differences in means, do we use $\mu_1 - \mu_2$ or $\mu_2 - \mu_1$?

Answer: Statistical inferences about the difference in means do not depend which group mean is μ_1 vs μ_2. The confidence interval is centered on the difference in the group means. The proper choice is usually clear from the context: only one choice gives the confidence interval in the exam question. If in doubt, use $\mu_1 - \mu_2$, or the mean group #1 – the mean of group #2.