Week 8 Problem 2

Luis Park

March 25, 2025

Question 2

Suppose $Y_{11}, ..., Y_{1n_1} \sim^{i.i.d} N(\mu_1, \sigma_1^2)$ are independent of $Y_{21}, ..., Y_{2n_1} \sim^{i.i.d} N(\mu_2, \sigma_2^2)$. Let $\overline{Y_1}$ and $\overline{Y_2}$ be the two averages and s_1^2 and s_2^2 the two sample variances. Consider a test of $H_o : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$ and assume $\sigma_1 = \sigma_2 = \sigma$.

Question 2 Part A

Problem:

The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n_1+n_2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} + \overline{Y_2})^2)$. Find the bias of this estimate as a function of n_1 and n_2. The restricted maximum likelihood estimate is calculated by first integrating the joint likelihood function with respect to the two mean parameters μ_1 and μ_2, and then maximizing the resulting function over σ^2. Show that this leads to the unbiased pooled sample variance $s_p^2 = \hat{\sigma}^2 = \frac{1}{n_1+n_2-2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} + \overline{Y_2})^2) = \frac{(n-1)s_1^2+(n_2+1)s_2^2}{n_1+n_2-2}$

Answer:

Bias of $\hat{\sigma^2}$

We are given that the maximum likelihood estimator for σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n_1+n_2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} + \overline{Y_2})^2)$$

To find the bias of this estimator, we must find what $E[\hat{\sigma^2}] - \sigma^2$ is equal to.

So our first step is to find $E[\hat{\sigma^2}]$

$$E[\hat{\sigma^2}] = E[\frac{1}{n_1+n_2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} + \overline{Y_2})^2)] = \frac{1}{n_1+n_2}(E[\sum(Y_{1i} - \overline{Y}_1)^2] + E[\sum(Y_{2i} + \overline{Y_2})^2])$$

We also know that $s_1^2 = \frac{1}{n_1-1}\sum(Y_{1i} - \overline{Y}_1)$ and $s_2^2 = \frac{1}{n_2-1}\sum(Y_{2i} - \overline{Y}_2)$

Also, we know that

$$E[s_1^2] = \sigma^2 \text{ and } E[s_2^2] = \sigma^2, \text{ so this means}$$
$$\frac{1}{n_1-1}E[\sum(Y_{1i} - \overline{Y}_1)^2] = \sigma^2 \rightarrow E[\sum(Y_{1i} - \overline{Y}_1)^2] = \sigma^2(n_1 - 1), \text{ similarly,}$$
$$\frac{1}{n_2-1}E[\sum(Y_{2i} - \overline{Y}_2)^2] = \sigma^2 \rightarrow E[\sum(Y_{2i} - \overline{Y}_2)^2] = \sigma^2(n_2 - 1)$$
$$\text{So } \frac{1}{n_1+n_2}(E[\sum(Y_{1i} - \overline{Y}_1)^2] + E[\sum(Y_{2i} + \overline{Y_2})^2]) = \frac{1}{n_1+n_2}((n_1 - 1)\sigma^2 + (n_2 - 1)\sigma^2) = \sigma^2\frac{n_1+n_2-2}{n_1+n_2}$$

Thus the bias of $\hat{\sigma}^2$ is:

$$bias(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2 = \sigma^2(\frac{n_1+n_2-2}{n_1+n_2} - 1) = -\sigma^2\frac{2}{n_1+n_2}$$

REML Estimate

First, we need the joint likelihood of both groups

$$L(\mu_1, \mu_2, \sigma^2) = \prod \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{(Y_{1i}-\mu_1)^2}{2\sigma}) + \prod \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{(Y_{2i}-\mu_2)^2}{2\sigma})$$

Then we need to integrate out μ_1, μ_2

$$\int\int L(\mu_1, \mu_2, \sigma^2)d\mu_1 d\mu_2 =$$
$$L_{REM}(\mu_1, \mu_2, \sigma^2) =$$
$$(\sigma^2)^{-(n_1+n_2-2)/2}exp(-\frac{1}{2\sigma^2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} - \overline{Y}_2)^2))$$

And once we maximize L_{REM} with respect to σ^2, we get

$$\hat{\sigma}^2 = \frac{1}{n_1+n_2-2}(\sum(Y_{1i} - \overline{Y}_1)^2 + \sum(Y_{2i} + \overline{Y_2})^2)$$

The REML procedure adjusts for the degrees of freedom lost in estimating the means, leading to an improved estimate of σ^2.

Question 2 Part B

Problem:

Show s_p^2 is a Gamma random variable

Show, using results we have already proved, that s_p^2 is a Gamma random variable, independant of $\overline{Y}_1$ and $\overline{Y}_2$. Show that the pivot $W = \frac{(n_1+n_2-2)s_p^2}{\sigma^2} \sim \chi^2_{(n_1+n_2-2)}$, when conditioning on μ_1, μ_2, and σ^2

Answer:

The pooled variance is defined as

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

Recall that $s_1^2 = \frac{1}{n_1-1}\sum(Y_{1i} - \overline{Y_1})^2$ and $s_2^2 = \frac{1}{n_2-1}\sum(Y_{2i} - \overline{Y_2})^2$

Since the sum of $(Y_{1i} - \mu_1)^2$ divided by σ^2 follows a χ^2_n, we can find what distribution $\sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2}$ follows:

We know $\sum \frac{(Y_{1i}-\mu_1)^2}{\sigma^2} = \sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2} + n\frac{(\overline{Y_1}-\mu_1)^2}{\sigma^2} = \sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2} + (\frac{\overline{Y_1}-\mu_1}{\sigma/\sqrt{n}})^2$

We know $(\frac{\overline{Y_1}-\mu_1}{\sigma/\sqrt{n_1}})^2$ follows a χ_1^2 because it is simply squaring one standard normal

And we also know $(\frac{\overline{Y_1}-\mu_1}{\sigma/\sqrt{n_1}})^2$ is independent of $\sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2}$ because $Y_{1i} - \overline{Y_1}$ is independent of $\overline{Y_1}$ (proved this in the presentation of last week)

Because of this independence, the chi-squared have an additive property. Thus, we have

$\sum \frac{(Y_{1i}-\mu_1)^2}{\sigma^2} = \sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2} + (\frac{\overline{Y_1}-\mu_1}{\sigma/\sqrt{n_1}})^2 \longrightarrow \chi^2_{n_1} = \sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2} + \chi^2_1$, which means $\sum \frac{(Y_{1i}-\overline{Y_1})^2}{\sigma^2} \sim \chi^2_{n_1-1}$

With the same methodology, we can show that $\sum \frac{(Y_{2i}-\overline{Y_1})^2}{\sigma^2} \sim \chi^2_{n_2-1}$

Now to summarize clearly, we have here that

$$\sum (Y_{1i} - \overline{Y_1})^2 \sim \sigma^2 \chi^2_{(n_1-1)}, \sum (Y_{2i} - \overline{Y_2})^2 \sim \sigma^2 \chi^2_{(n_2-1)} \Longrightarrow$$
$$(n_1 - 1)s_1^2 \sim \sigma^2 \chi^2_{(n_1-1)},\ (n_2 - 1)s_2^2 \sim \sigma^2 \chi^2_{(n_2-1)}$$

And by the additive property of the chi-square distribution, their sum is (because independent):

$$(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \sim \sigma^2 \chi^2_{(n_1+n_2-2)}$$

So this means:

$$s_p^2 = \frac{\sigma^2 \chi^2_{(n_1+n_2-2)}}{n_1+n_2-2}$$

Since a chi-square random variable with k degrees of freedom follows a gamma distribution with shape k/2 and scale 2, we conclude that :

$$s_p^2 \sim \frac{\sigma^2}{n_1+n_2-2} Gamma(\frac{n_1+n_2-2}{2}, 1/2) = Gamma(\frac{n_1+n_2-2}{2}, \frac{n_1+n_2-2}{2\sigma^2})$$

This proves that s_p^2 follows a gamma distribution

Pivot of W

Prove that the pivot $W = \frac{(n_1+n_2-2)s_p^2}{\sigma^2}$

We know that $s_p^2 = \frac{\sigma^2 \chi^2_{(n_1+n_2-2)}}{n_1+n_2-2}$ so this means

$$W = \frac{(n_1+n_2-2)\sigma^2 \chi^2_{n_1+n_2-2}}{(n_1+n_2-2)(\sigma^2)} = \chi^2_{n_1+n_2-2}$$

Question 2 Part C

Problem:

Show that the pooled two sample t statistic T satisfies the definition of a $t_{n_1+n_2-2}$ random variable for any hypothesized value of $\mu_1 - \mu_2 = 0$. What values of T would lead you to reject H_o at level α? What values of T^2 would lead you to reject? What is the null sampling distribution of T^2 ? What is a CI for $\mu_1 - \mu_2$? Make the distinction between the pooled standard deviation estimate (root mean square error) and the standard error.

Answer:

Pooled Two-Sample T Statistic

We are given the test statistic as

$$T = \frac{\overline{Y}_1 - \overline{Y_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Recall that

$T = \frac{Z}{\sqrt{W/(n_1+n_2-2)}} \sim t_{n_1+n_2-2}$ where Z is independent of W because $\overline{Y_i}$ is independent of s_i^2

Furthermore W is a χ^2 with df of $n_1 + n_2 - 2$

We defined previously that $W = \frac{(n_1+n_2-2)s_p^2}{\sigma^2}$ and we also know $Z = \frac{\overline{Y_1}-\overline{Y_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$

So this means $T = \frac{\overline{Y_1}-\overline{Y_2}}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\sqrt{\frac{\sigma^2(n_1+n_2-2)}{(n_1+n_2-2)s_p^2}} = \frac{\overline{Y_1}-\overline{Y_2}}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

And as you can see, this precisely is the T statistic we observed earlier. So this means the two pooled two sample t statistic T satisfies the definition of $t_{n_1+n_2-2}$ for any hypothesized value of $\mu_1 - \mu_2 = 0$

Rejecting H_o at Level α

We would reject for values of T where

$|T| > t_{\alpha/2,n_1+n_2-2}$ where t is the critical value

Values of T^2 Leading to Rejection

$T^2 \sim F(1, n_1 + n_2 - 2)$ under H_o

For F distribution, we would reject T^2 if $T^2 > F_{\alpha,1,n_1+n_2-2}$. Note that the null sampling distribution of T^2 would be the F distribution stated above.

Confidence Interval of $\mu_1 - \mu_2$

$CI = (\overline{Y_1} - \overline{Y_2}) \pm t_{1-\alpha/2,n_1+n_2-2} * s_p\sqrt{1/n_1 + 1/n_2}$

Pooled Standard vs. Standard Error

Pooled Standard is combining the variances of both samples into a single estimate of σ

Standard error measures the variability of the sampling distribution of the difference in sample means

Question 2 Part D

Problem:

As an example, imagine dividing N = 200 subjects into two equal-sized treatment groups and administering a treatment to one group and a placebo to the other. What values of T would lead you to reject the null hypothesis? What values would lead you to conclude there is a positive difference in means? Explain why it is justifiable to claim to have shown a positive difference when the alternative hypothesis does not specify a direction. How would the test change if you assumed the target null mean ($\mu_o = 0$) and variances $\sigma_1 = \sigma_2 = 1$) were correct?

Answer:

Rejecting Null Hypothesis:

We would reject the null hypothesis if

$$|\frac{\overline{Y_1}-\overline{Y_2}}{s_p\sqrt{\frac{1}{100}}}| = |\frac{\overline{Y_1}-\overline{Y_2}}{\sqrt{\frac{(99)s_1^2+(99)s_2^2}{198}}\sqrt{\frac{1}{100}}}| > t_{\alpha/2,198}$$

Rejecting Null Hypothesis with Positive Difference:

$$\frac{\overline{Y_1}-\overline{Y_2}}{s_p\sqrt{\frac{1}{100}}} = \frac{\overline{Y_1}-\overline{Y_2}}{\sqrt{\frac{(99)s_1^2+(99)s_2^2}{198}}\sqrt{\frac{1}{100}}} > t_{\alpha,198}$$

Why Can We Claim a Positive Difference in a Two-Tailed Test?

We can still claim a positive difference in a two-tailed test because a two-sample t-test is more selective in rejecting the null hypothesis. Unlike a one-tailed test, which rejects the null when the test statistic falls in the top $1-\alpha$ quantile, a two-tailed test rejects only if the statistic falls in the top or bottom $1-\alpha/2$ quantiles. This makes rejection more difficult in either direction. This in result allows us to claim a positive or negative difference in the two tailed test

Question 2 Part E

Problem:

Find an expression for the power of the test if $\mu_1 - \mu_2 = c\sigma$. For example, suppose $c = 0.25$ would be on the lower boundary of being an important (practically significant) difference in means. Find the smallest n to have power at least 0.99 of detecting such a difference at $\alpha = 0.1$. Explain what having such a high power allows you to say if you fail to reject H_o

Answer:

Expression of Power

Reminder that power is the probability of rejecting H_o when H_1 is true

If we assume H_o is true when H_1 is actually true, then our new test statistic becomes

$$T = \frac{\overline{X_1}-\overline{X_2}-(\mu_1-\mu_2)}{SE} = \frac{\overline{X_1}-\overline{X_2}-0}{SE}$$

Note that $\mu_1 - \mu_2 = 0$ only because we **assume** H_o is true

Now we need solve for SE

$$SE = \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sigma\sqrt{\frac{2}{n}} \text{ if we assume } n_1 = n_2 = n$$

So this gives us a T statistic of

$$T = \frac{\overline{X_1}-\overline{X_2}-0}{SE} \sim N(\frac{c}{\sqrt{2/n}}, 1)$$

What does this test statistic tell us? It tells us how many standard deviations away the true difference is from 0. Now, to calculate the power, we should use a standard normal. This make the calculation easier and it well approximates the power as n increases.

Under the alternative, T is approximately a $N(\frac{c}{\sqrt{2/n}}, 1)$. Let us say $Z' \sim N(\frac{c}{\sqrt{2/n}}, 1)$

$$\text{So } power \approx P(Z' > z_{1-\alpha/2}|H_a) + P(Z' < -z_{1-\alpha/2}|H_a) = P(Z' - \frac{c}{\sqrt{2/n}} > z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}}) + P(Z' - \frac{c}{\sqrt{2/n}} < -z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}}) = P(Z > z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}}) + P(Z < -z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}})$$

$$power \approx P(Z > z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}}) + P(Z < -z_{1-\alpha/2} - \frac{c}{\sqrt{2/n}})$$

Solving for sample size N

We want to find the smallest n such that the power is greater than or equal to 0.99. We are given that $c = 0.25$ and $\alpha = 0.1$

We can solve for this using the power equation from above.

We know that for an $\alpha = 0.1$, our critical is $z_{1-0.1/2} = z_{0.95} = 1.645$

So our power is

$$power = P(Z > 1.645 - \frac{0.25}{\sqrt{2/n}}) + P(Z < -1.645 - \frac{0.25}{\sqrt{2/n}})$$

For a large power, $P(Z < -1.645 - \frac{0.25}{\sqrt{2/n}})$ is negligible

So our power approximation now becomes:

$$power \approx P(Z > 1.645 - \frac{0.25}{\sqrt{2/n}}) = 0.99$$

This implies

$$1.645 - \frac{0.25}{\sqrt{2/n}} = -2.326$$

Now solve for n

$$n \approx 504$$

Implication of High Power

A high power of 99 percent tells us 99 percent chance of detecting a true effect (difference in means) if it exists. If we fail to reject with this high of a power, it strongly suggests that there is no practically significant difference

For smaller powers like 50 percent, the chances of detecting a difference is only 50 percent, so rejecting the null hypothesis could be due to the lower sensitivity.