2. **The pooled 2-Sample t Test** Suppose $Y_{11}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ are independent of $Y_{21}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$. Let $\bar{Y}_1$ and $\bar{Y}_2$ 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 $\sigma_1 = \sigma_2 = \sigma$.

a) The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n_1+n_2}(\sum(Y_{1i} - \bar{Y}_1)^2 + \sum(Y_{2i} - \bar{Y}_2)^2)$. Find the bias of this estimate as a function of n_1 and n_2. The *restricted* maximum likelihood (REML) 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 $s_p^2 = \frac{\sum(Y_{1i}-\bar{Y}_1)^2+\sum(Y_{2i}-\bar{Y}_2)^2}{n_1+n_2-2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$. Note that the only justification for this procedure from a frequentist perspective is that it leads to an improved estimate of σ that reflects the degrees of freedom lost to estimating the two mean parameters (it makes much more sense as an objective Bayes procedure).

Pooled 2 - sample t-test

Begin by looking at the bias of the MLE's.

Recall: $\text{Bias}(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \sigma^2$

0 = unbiased, biased otherwise

To make this easier to think about, we are going to start with 1 sample.

$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

MLE : $\hat{\sigma}^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{n}$

$\text{Bias}(\hat{\sigma}^2_{MLE}) = E\left[\frac{\Sigma(Y_i-\bar{Y})^2}{n}\right] - \sigma^2 = \frac{n-1}{n}\sigma - \sigma^2 = \frac{-\sigma^2}{n}$ → bias

Two sample:

$Y_{1i}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ and $Y_{2i}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$

note: $E[\Sigma(Y_i - \bar{Y})^2] = (n-1)\sigma^2$ * previously shown

MLE : $\frac{1}{n_1+n_2}\left[\Sigma(Y_{1i} - \bar{Y}_1)^2 + \Sigma(Y_{2i} - \bar{Y}_2)^2\right] = \frac{(n+1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2}$

Bias:

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

$= \frac{n_1+n_2-2}{n_1+n_2}\sigma^2$

$\text{Bias}(\hat{\sigma}^2) = \frac{n_1+n_2-2}{n_1+n_2}\sigma^2 - \sigma^2 = \left(\frac{n_1+n_2-2}{n_1+n_2} - 1\right)\sigma^2 = \left(\frac{-2}{n_1+n_2}\right)\sigma^2$ * biased

2. **The pooled 2-Sample t Test** Suppose $Y_{11}, \ldots, Y_{1n_1} \overset{\text{iid}}{\sim} N(\mu_1, \sigma_1^2)$ are independent of $Y_{21}, \ldots, Y_{2n_2} \overset{\text{iid}}{\sim} N(\mu_2, \sigma_2^2)$. Let $\bar{Y}_1$ and $\bar{Y}_2$ 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 $\sigma_1 = \sigma_2 = \sigma$.

a) The MLE for σ^2 is $\hat{\sigma}^2 = \frac{1}{n_1+n_2}(\sum(Y_{1i} - \bar{Y}_1)^2 + \sum(Y_{2i} - \bar{Y}_2)^2)$. Find the bias of this estimate as a function of n_1 and n_2. The *restricted* maximum likelihood (REML) 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 $s_p^2 = \frac{\sum(Y_{1i}-\bar{Y}_1)^2+\sum(Y_{2i}-\bar{Y}_2)^2}{n_1+n_2-2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$. Note that the only justification for this procedure from a frequentist perspective is that it leads to an improved estimate of σ that reflects the degrees of freedom lost to estimating the two mean parameters (it makes much more sense as an objective Bayes procedure).

MLE : biased low as doesn't correct for estimating averages

Restricted maximum likelihood estimate (REML): frequentist procedure to integrate the means out of the function and then maximize in terms of σ

→ will get an unbiased estimate that reflects the loss of degrees of freedom

Generally, REML: $L(\sigma^2) = \int_{-\infty}^{\infty} L(\mu, \sigma^2)\,d\mu$ wrt σ^2

1 sample: $\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{\sum(y_i-\mu)^2}{2\sigma^2}}$

$= (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum(y_i-\bar{y})^2}{2\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}}\,d\mu$

* Recall: $\sum(y_i-\mu)^2 = \sum(y_i-\bar{y})^2 + n(\bar{y}-\mu)^2$

$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}} \sqrt{2\pi\frac{\sigma^2}{n}}$

$\propto (\sigma^2)^{-\frac{n-1}{2}} e^{-\frac{(n-1)s^2}{2\sigma^2}}$

maximize: Integrated log likelihood

like the previous ways we have maximized, we are going to take the log of the likelihood, differentiate, and then set to ∅.

$\ell(\sigma^2) = -\frac{n-1}{2}\log(\sigma^2) - \frac{(n-1)s^2}{2\sigma^2}$

$\ell'(\sigma^2) = -\frac{n-1}{2\sigma^2} + \frac{(n-1)^2 s^2}{2(\sigma^2)^2} = 0$

$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum(y_i-\bar{y})^2 \rightarrow E\left[\frac{1}{n-1}\sum(y_i-\bar{y})^2\right] = \sigma^2$ → so no bias anymore

change n to $n-1$ fixes the issue with the bias

REML Objective Bayes Perspective

$L(\mu_1, \mu_2, \sigma^2) = f(y_1, \ldots, y_n \mid \mu, \sigma^2)$

$p(\mu_1, \mu_2) \propto c$

$\rightarrow L(\mu_1, \mu_2, \sigma^2)\,p(\mu_1, \mu_2)$

$= f(y_1, \ldots, y_n, \mu_1, \mu_2 \mid \sigma^2)$

$\iint d\mu_1\,d\mu_2 = f(y_1, \ldots, y_1 \mid \sigma^2)$

maximizes over σ^2

* Marginalized likelihood function for σ^2

2 sample: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L(\mu_1, \mu_2, \sigma^2)\,d\mu_1\,d\mu_2$

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

compared to $\hat{\sigma}^2_{MLE}$, we subtract 2 in the denominator which now makes this unbiased as $E[s_p^2] = \sigma^2$

The (-2) accounts for the 2 degrees of freedom we are accounting for with the 2 means, one from each sample

*more samples would increase the number of integrals but also the degrees of freedom accounted for in the denominator

b) Show, using results we have already proved, that s_p^2 is a Gamma random variable, independent of $\bar{Y}_1$ and $\bar{Y}_2$. 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.

Using the sample variance (s_p^2) we just found for the 2-sample test, we are going to show this is a Gamma variable independent of $\bar{Y}_1$ and $\bar{Y}_2$

parameters that the distribution is reliant on

s_p^2 is the random variable

Recall: $$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-n}$$

$\frac{E(\Sigma y_i - \bar{y})}{\sigma^2}$ → mean / sample variance

In each sample:

$$\frac{(n_1-1)s_1^2}{\sigma^2} = \frac{\Sigma(y_{iy}-\bar{Y}_i)}{\sigma^2} \sim X^2_{n-1}$$ *Previously shown as chi-square distribution

split into the two samples we have

$$\frac{(n_1-1)s_1^2}{\sigma^2} \sim X^2_{n_1-1}, \quad \text{[independent] and} \quad \frac{(n_2-1)s_2^2}{\sigma^2} \sim X^2_{n_2-1}$$

$$\frac{(n_1-1)s_1^2}{\sigma^2} + \frac{(n_2-1)s_2^2}{\sigma^2} \sim X^2_{n_1+n_2-1}$$ → sum of two individual chi-squares is chi-square with the sum of the degrees of freedom

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{\sigma^2} \sim X^2_{n_1+n_2-2}$$ → only works because of the same denominator (we can add the seperate distributions)

Recall: all chi-squares are gamma

if: $X \sim X^2_{(m)} \longrightarrow X \sim \text{Gamma}\left(\frac{m}{2}, \frac{1}{2}\right)$ *all chi-squares are gamma

RV, df

* also works in reverse: $Y \sim \text{Gamma}(\alpha, \lambda) \rightarrow 2\lambda Y \sim X^2(2\alpha)$

look at the individual samples

$(n_1-1)s_1^2 \sim \text{Gamma}\left(\frac{n_1-1}{2}, \frac{1}{2\sigma^2}\right)$

and

$(n_2-1)s_2^2 \sim \text{Gamma}\left(\frac{n_2-1}{2}, \frac{1}{2\sigma^2}\right)$

* sum of two gamma's is gamma only if λ's are the same

λ for a chi-square distribution = ½

so,

$(n_1-1)s_1^2 + (n_2-1)s_2^2 \sim \text{Gamma}\left(\frac{n_1+n_2-2}{2}, \frac{1}{2\sigma^2}\right)$ *unscaled

multiply by denominator (degrees of freedom) to adjust variance $\frac{\alpha}{\lambda}$

$$\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = S_p^2 \sim \text{Gamma}\left(\frac{n_1+n_2-2}{2}, \frac{n_1+n_2-2}{2\sigma^2}\right)$$

→ divided by adjusted observations (df) to be unbiased (corrected for means)

Through this, we have shown the pivot W is chi-square

$$W = \frac{(n_1+n_2-2)s_2^2}{\sigma^2} \sim X^2_{(n_1+n_2-2)}$$

c) 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 $\mu_1 - \mu_2$ (e.g. $\mu_1 - \mu_2 = 0$).

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

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.

Part c asks us to show that the pooled two sample t statistic T satisfies the definition of a t-distribution with (n_1+n_2-2) df R.V. for any hypothesized value of $\mu_1-\mu_2$.

Let $Z \sim N(0,1)$ independent of $W \sim \chi^2_{(m)}$

Recall: $T = \frac{Z}{\sqrt{W/m}} \sim t_{(m)}$ $\quad m = n_1 + n_2 - 2$

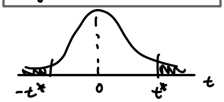
and $Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N(0,1)$ * independent of W (because $\bar{Y}_i$ independent of S_i^2)

$$T = \frac{Z}{\sqrt{W/m}} = \frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \cdot \sqrt{\frac{(n_1+n_2-1)\,\sigma^2}{(n_1+n_2-1)S_p^2}} = \frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{(n_1+n_2-2)}$$

Often 0 in null, works to center the statistic

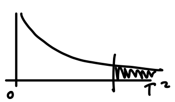
What values of T would lead you to reject H_0 at level α?

$\boxed{\text{Reject for } |T| > t^*}$ $\to 1-\frac{\alpha}{2}$ quantile of $t_{(n_1+n_2-2)}$



What values of T^2 would lead you to reject?

$\boxed{T^2 > (t^*)^2}$



(takes both $\pm t^*$)

these are values close to the null

F distribution is the ratio of two gammas
Thus, two different degrees of freedom.

T^2 follows an F-distribution with 1 numerator degree of freedom and n_1+n_2-2 denominator degrees of freedom

simple t-distribution

null sampling distribution: $T^2 \sim F(1, n_1+n_2-2)$ under H_0

reject based on F value (from table or R) * reject for large F

Confidence interval for $(\mu_1-\mu_2)$:

$0.95 = P(-t^* < T < t^*)$

$= P(\bar{Y}_1-\bar{Y}_2 - t^* SE(\bar{Y}_1-\bar{Y}_2) < \mu_1-\mu_2 < \bar{Y}_1-\bar{Y}_2 + t^* \cdot SE(\bar{Y}_1-\bar{Y}_2))$

$= \bar{Y}_1-\bar{Y}_2 \pm t^* \cdot \underbrace{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}_{SE(\bar{Y}_1-\bar{Y}_2)}$

distinction difference between pooled standard deviation estimate (RMSE) and standard error:

$MSE = \overset{\text{sample SD}}{S_p} \to$ estimates $SD(Y_i)$ single y

SE estimates $SD(\bar{Y}_1-\bar{Y}_2)$ $\to$ relies of MSE

$= S_p \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $\to$ accounts for differences between means of the groups

d) 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?

Part D has us look at an example of a pooled 2-sample t-test.

$N = 200$. with $n_1 = 100$ = treatment and $n_2 = 100$ = placebo

Hypothesis testing in this case has to do with whether or not there is a significant effect from the treatment

$H_0: \mu_1 - \mu_2 = 0$ (no difference in means between treatment and placebo)

$H_a: \mu_1 - \mu_2 \neq 0$ (there is a difference in means between treatment and placebo)

Given what we have just discussed about values that would lead us to reject, we can say that the values of T that would lead us to reject the null hypothesis would be when $\bar{Y}_1 - \bar{Y}_2$ is considerably greater than 0 (or the null) or considerable less than 0 (the null)

→ based on whatever cutoff we have set from α level.

→ provides evidence that the two means are not the same.

Phil, and we as a class, claim it is justifiable to claim to have shown a positive difference when the alternative does not show a difference.

We can say this because we know that a one-sided test is less sensitive than a two sided test (only 1.64 SD's compared to 1.96 SD's)

So, if the results are significant in the 2-sided test, they could be further from 0/null compared to significant differences in a one-sided - st.

So if we have data that is extreme enough in one direction that it would be too weird to be zero, then it would be even more weird to see data on the other side of zero (even further from the obtained results).

→ rejecting counter to data beyond zero

optional last question:

How would the test change if you assumed the target null mean ($\mu_0 = 0$) and variances ($\sigma_1 = \sigma_2 = 1$) were correct?

* ends up a z-test
* $\bar{Y}_2$ only R.V. left → testing if it has a mean of zero.

e) 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 high power allows you to say if you fail to reject H_o (as compared to having, say, 50% or 10% power).

Part e asks us to find an expression for the power of the test if $\mu_1 - \mu_2 = c\sigma$, where c is a constant = 0.25

find the smallest n to have a power = 0.99

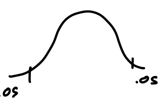
2 sample t-test $\quad n_1 = n_2 = 100 \quad$ assume $\sigma_1 = \sigma_2 = 1$

df = 100 + 100 − 2 = 198

Question tells us to reject at $\alpha = .1$ for $|T| > t^* = 1.6526$

0.95 quartile

$t^* = 1.6526$



power: $P(|T| > 1.6526 \mid \mu_1 - \mu_2 = c\sigma)$

Plug in T Statistic
$$= P\left(\frac{\bar{Y}_1 - \bar{Y}_2 - \overbrace{0}^{\text{null}}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > 1.6526 \;\Big|\; \mu_1 - \mu_2 = c\sigma\right)$$

reject region
$$= P\left(\bar{Y}_1 - \bar{Y}_2 > 1.6526 \cdot \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \;\Big|\; \mu_1 - \mu_2 = c\sigma\right) \quad * S_p \approx \sigma$$

Subtract alternative mean difference = $c\sigma$ and divide by standard deviation

standardize based on alternative hypothesis
$$\approx P\left(z > \frac{t^*\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} - c\sigma}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right)$$

$$= P\left(z > t^* - c\sqrt{\frac{n}{2}}\right) > .99$$

assume the n's are equal
So $n_1 + n_2 = n$

need $\quad 1.6526 - 0.25\sqrt{\frac{n}{2}} < -2.326$

$$\sqrt{n} > (1.6526 + 2.326)\frac{\sqrt{2}}{.25} = 22.5$$

$$\boxed{n \geq 507}$$

Having such a high power allows you to say you fail to reject H_0 because it means the test has a 99% chance of detecting if the difference is significant.

→ makes the non-rejections meaningful → looked hard and didn't find it