homework04

zhewei xie

2024-08-08

Problem 1: NBC pilot survey

Part A

Question

Is there evidence that one show consistently makes people happier among viewers?

Approach

I use both de Moivre's equation as well as the "Pythagorean theorem" for the difference in sample means between "Living with Ed" and "My Name is Earl." To apply the correct formulas, we need the sample mean, sample standard deviation, sample size in each group, and the difference between the two means. After applying the formula below, we could easily find out the 95% confidence interval for the difference in mean viewer response to the Q1_Happy question for these two shows.

$se(\overline{x}_1 - \overline{x}_2) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$

Results

Table 1: The confidence intervals with a 95% confidence level for the difference in mean viewer response to the Q1_Happy question for the shows "Living with Ed" and "My Name is Earl," based on the de Moivre's equation and the Pythagorean theorem.

Different Mean	Standard Error	Lower	Upper
-0.1490515	0.12766	-0.3992651	0.1011621

Conclusion

This 95% confidence interval contains zero, so we do not have a statistically significant difference in mean viewer response to the Q1_Happy question for these two shows. Although the average rating for "Living with Ed" is slightly higher, the difference is not significant enough to rule out the possibility of chance.

Part B

Question

Is there evidence that one show consistently makes people feel more annoyed among viewers?

Approach

First, I stored the differences in the means of 10,000 bootstrapped samples from "The Biggest Loser" and "The Apprentice: Los Angeles." Then, I ask for a confidence interval based on this bootstrapped sampling distribution.

Results

Table 2: The confidence intervals with a 95% confidence level for the difference in mean viewer response to the Q1_Annoyed question for the shows "The Biggest Loser" and "The Apprentice: Los Angeles," based on 10,000 bootstrap sampling.

lower	upper	level	method	estimate
-0.5204678	-0.0215524	0.95	percentile	-0.3293413

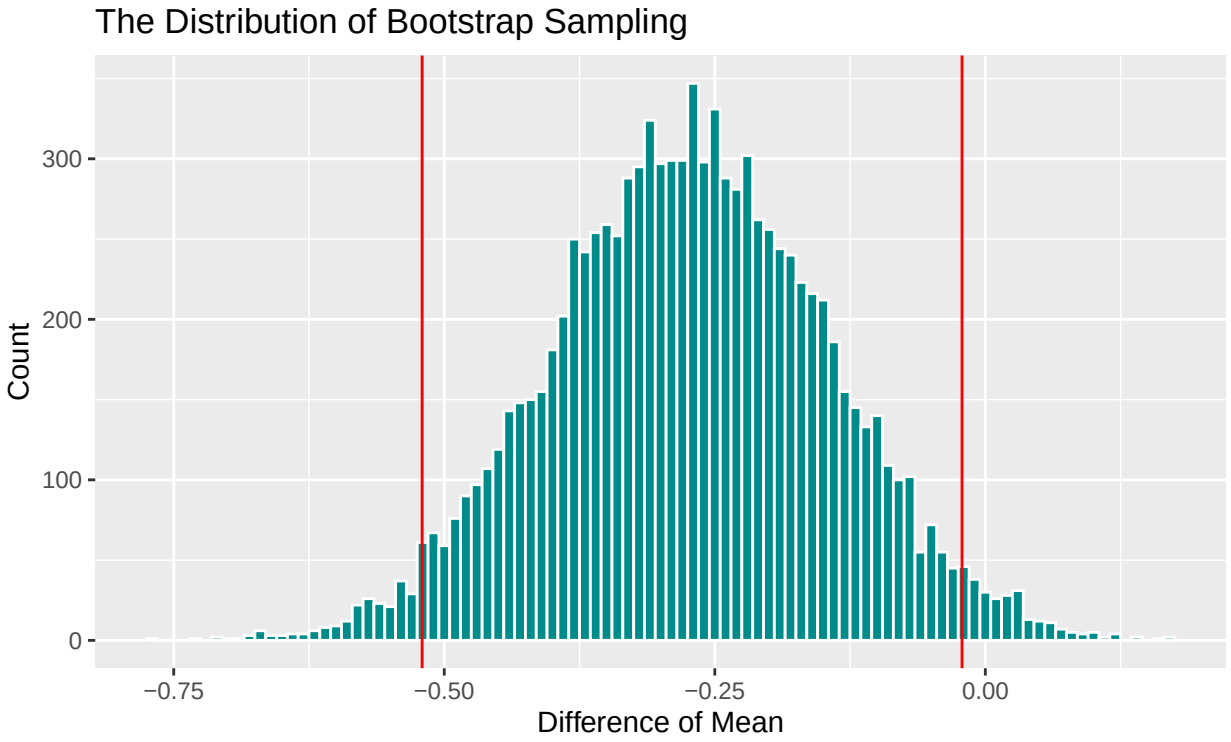

Figure 1. The distribution of 10,000 bootstrap sampling for the difference in mean viewer response to the "Q1_Annoyed" question for the shows "The Biggest Loser" and "The Apprentice: Los Angeles," with a 95% confidence level verticle line. The 2.5% confidence level is approximately −0.520, and the 97.5% confidence level is approximately −0.022.

Conclusion

This 95% confidence interval doesn't contain zero, so we have a statistically significant difference in mean viewer response to the Q1_Annoyed question for these two shows. We have a 95% confidence to conclude that "The Apprentice: Los Angeles" made people feel more annoyed.

Part C

Question

What proportion of American TV watchers would we expect to give a response of 4 or greater to the "Q2_Confusing" question for the show "Dancing with the Stars."

Approach

First, I stored the proportions of American TV watchers who gives as response of 4 or greater to the "Q2_Confusing" question of 10,000 bootstrapped samples for the show "Dancing with the Stars." Then, I ask for a confidence interval based on this bootstrapped sampling distribution.

Results

Table 3: The confidence intervals with a 95% confidence level for the proportion of American TV watchers expected to give a response of 4 or greater to the "Q2_Confusing" question for the show "Dancing with the Stars," based on 10,000 bootstrap sampling.

lower	upper	level	method	estimate
0.038674	0.1160221	0.95	percentile	0.0773481

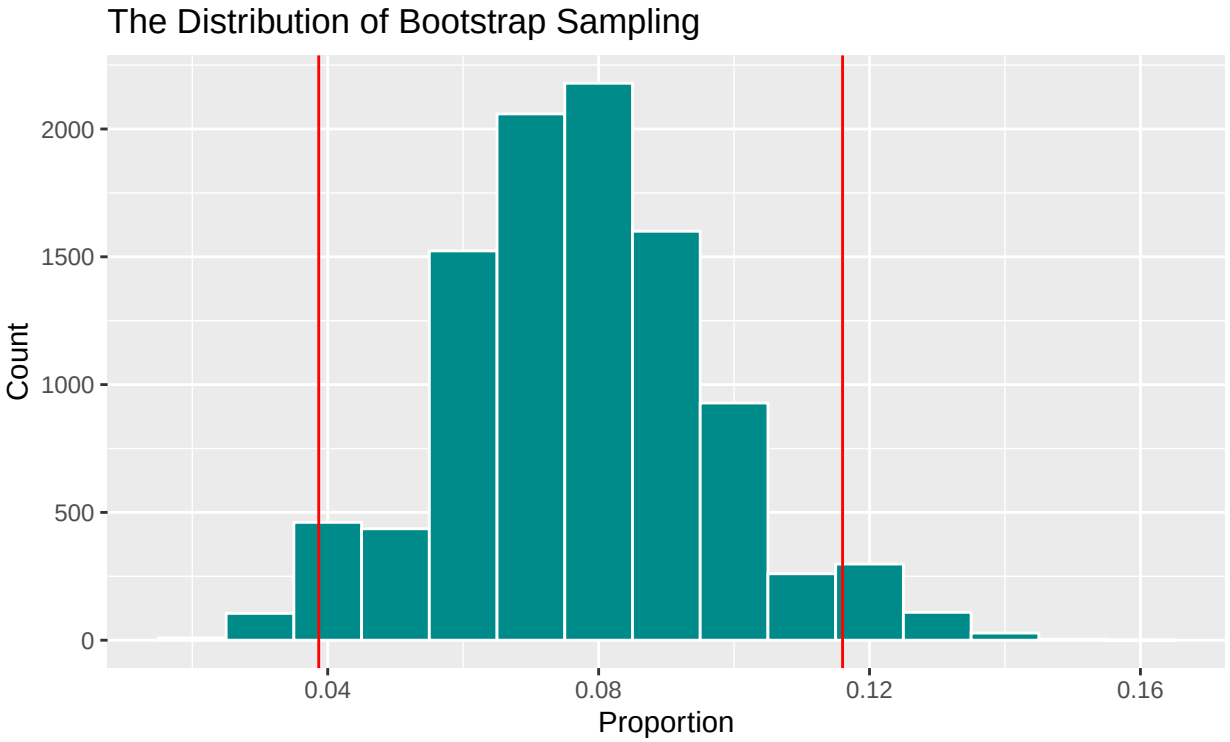

Figure 2. The distribution of 10,000 bootstrap sampling for the proportion of American TV watchers expected to give a response of 4 or greater to the "Q2_Confusing" question for the show "Dancing with the Stars," with a 95% confidence level verticle line. The 2.5% confidence level is approximately 0.039, and the 97.5% confidence level is approximately 0.116.

Conclusion

I have a 95% confidence level to report that there is approximately 3.9% to 11.6% American TV watchers expected to give a response of 4 or greater to the "Q2_Confusing" question for the show "Dancing with the Stars."

Problem 2: EBay

Question

My task is to calculate the difference in revenue ratios between the treatment and control DMAs and provide a 95% confidence interval for this difference. These results will be used to evaluate whether the evidence supports the hypothesis that the revenue ratio is the same in the treatment and control groups, or if the data instead suggests that paid search advertising on Google generates additional revenue for eBay.

Approach

First, I stored the differences in the means of 10,000 bootstrapped samples from the results of the experiment ran by EBay in May of 2013. Then, I ask for a confidence interval based on this bootstrapped sampling distribution.

Results

Table 4: The confidence intervals with a 95% confidence level for the difference in mean revenue ratio at the DMA level for eBay, i.e. the ratio of revenue after to revenue before for each DMA, based on 10,000 bootstrap samling.

lower	upper	level	method	estimate
-0.0909335	-0.0129423	0.95	percentile	-0.0550873

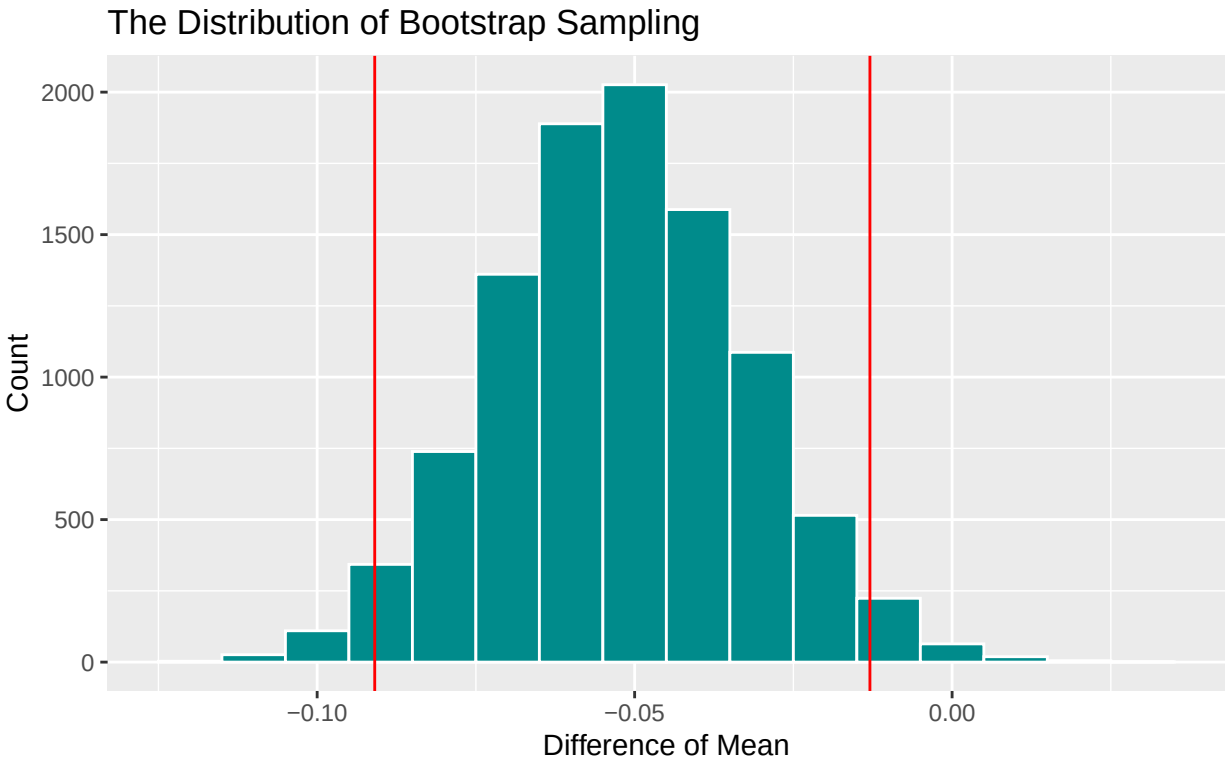

Figure 3. The distribution of 10,000 bootstrap sampling for the difference in mean revenue ratio at the DMA level for eBay. The 2.5% confidence level is approximately −0.091, and the 97.5% confidence level is approximately −0.013.

Conclusion

This 95% confidence interval doesn't contain zero, so we have a statistically significant difference in mean revenue ratio at the DMA level for eBay. We have a 95% confidence to conclude that paid search advertising on Google generates additional revenue for eBay.

Problem 3 - Iron Bank

Step 1: **Our null hypothesis is that the cluster of trades by the Iron Bank is as rare as the baseline probability of any legal trade being flagged by the SEC's detection algorithm, which is 2.4%.**

Step 2: **Our test statistic is the number of trades cases which are being flagged. Higher numbers of cases imply stronger evidence against the null hypothesis. In our data, 70 were flagged by the SEC's detection algorithm of the last 2021 trades by Iron Bank employees.**

Step 3: we must calculate the probability distribution of the test statistic, assuming that the null hypothesis is true. This distribution provides context for the observed data. It allows us to check if the observed data looks plausible under this distribution, or instead whether the null hypothesis looks too implausible to be believed. To do this, I repeated Monte Carlo simulation 100,000 times and store the result.

Let's now visualize the distribution.

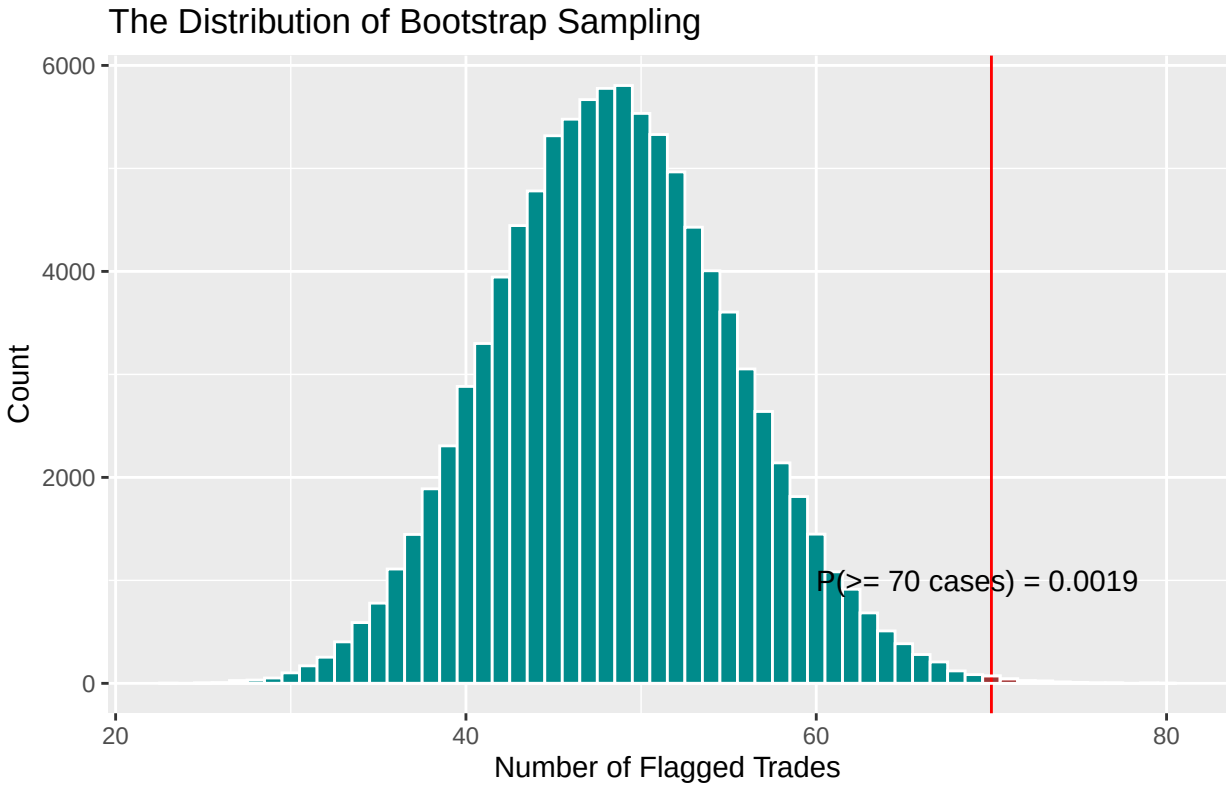

Figure 4. The distribution of 100,000 bootstrap sampling of the number of trades of the Iron Bank flagged by the SEC´s detection algorithm.

Step 4: calculate a p-value. The number of simulations yielded 70 cases of trades or more is 192 out of 100,000, or about 0.192%. **So our p-value is $p = 0.00192$.**

Step 5: **Based on this p-value, we reached a conclusion that there is a 0.2% chance that the null hypothesis is right, which implies that the clustering of flagged trades at Iron Bank is unlikely to be due to chance, warranting further investigation by the SEC.**

Problem 4: milk demand, revisited

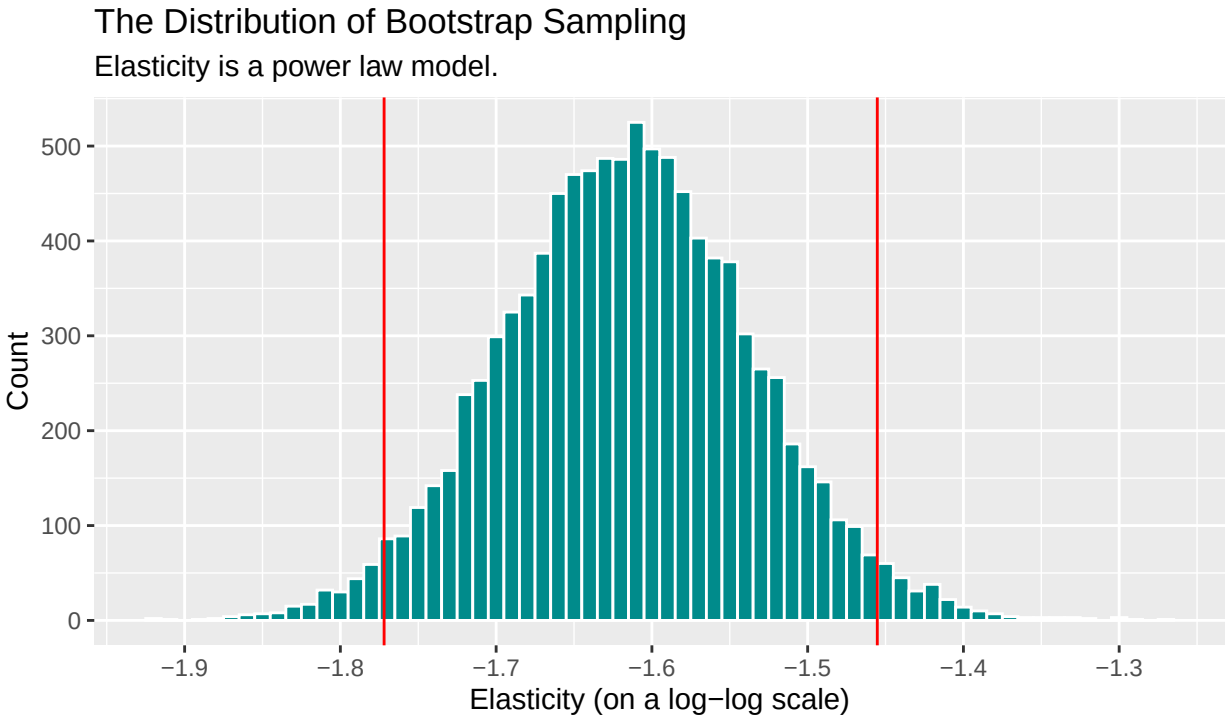

Figure 5. The distribution of 10,000 boostrap sampling of the price elasticity of demand for milk. The 95% bootstrap confidence interval for the elasticity is showed in the figure, with the 2.5% confidence level approximately at −1.77 and the 97.5% confidence level approximately at −1.46, both rounded to two decimal places. Therefore, for a confidence interval of 95%, the price elasticity of demand for milk (calculated as log(sales)~log(price)) is approximately from −1.77 to −1.46.

Problem 5: standard-error calculations

Part A

First of all, there are some assumptions:

$\because X_1, ..., X_N \sim \text{Bernoulli}(p)$ and $Y_1, ..., Y_M \sim \text{Bernoulli}(q)$ (all independent)

$\because \hat{p} = \overline{X}_N$ and $\hat{q} = \overline{Y}_M$

Question i

$\because E(\hat{p}) = E(\frac{\sum_{i=1}^{N} X_i}{N})$ and $E(\hat{q}) = E(\frac{\sum_{j=1}^{M} Y_j}{M})$

$\therefore E(\hat{p} - \hat{q}) = E(\hat{p}) - E(\hat{q}) = \frac{\sum_{i=1}^{N} E(X_i)}{N} - \frac{\sum_{j=1}^{M} E(Y_j)}{M} = p - q$

Question ii

$\because var(\hat{p}) = var(\frac{\sum_{i=1}^{N} X_i}{N}) = \frac{1}{N^2} var(\sum_{i=1}^{N} X_i) = \frac{1}{N^2} \sum_{i=1}^{N} var(X_i)$

$\because var(X_i) = p(1-p)$

$\therefore var(\hat{p}) = \frac{1}{N}p(1-p)$

$\therefore \text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{N}}$

Question iii

$\because var(\hat{\Delta}) = var(\hat{p} - \hat{q}) = var(\frac{\sum_{i=1}^{N} X_i}{N} + (-1) \times \frac{\sum_{j=1}^{M} Y_j}{M}) = \frac{1}{N^2}\sum_{i=1}^{N} var(X_i) + \frac{1}{M^2}\sum_{j=1}^{M} var(Y_j)$

$\because var(X_i) = p(1-p)$ and $var(Y_j) = q(1-q)$

$\therefore var(\hat{\Delta}) = \frac{1}{N}p(1-p) + \frac{1}{M}q(1-q)$

$\therefore \text{SE}(\hat{\Delta}) = \sqrt{\frac{p(1-p)}{N} + \frac{q(1-q)}{M}}$

Part B

First of all,

$\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$

$\overline{Y}_M = \frac{1}{M}\sum_{j=1}^{M} Y_j$

$\Delta = \mu_X - \mu_Y$

$\hat{\Delta} = \overline{X}_N - \overline{Y}_M$

Assume $X \perp Y$

Expected value

$\because E(\hat{\Delta}) = E(\overline{X}_N - \overline{Y}_M) = E(\frac{1}{N}\sum_{i=1}^{N} X_i - \frac{1}{M}\sum_{j=1}^{M} Y_j)$

$\therefore E(\hat{\Delta}) = \frac{1}{N}\sum_{i=1}^{N} E(X_i) - \frac{1}{M}\sum_{j=1}^{M} E(Y_j) = E(X_i) - E(Y_j) = \mu_X - \mu_Y = \Delta$

Standard error

$\because var(\hat{\Delta}) = var(\overline{X}_N - \overline{Y}_M) = var(\frac{\sum_{i=1}^{N} X_i}{N} + (-1) \times \frac{\sum_{j=1}^{M} Y_j}{M}) = \frac{1}{N^2}\sum_{i=1}^{N} var(X_i) + \frac{1}{M^2}\sum_{j=1}^{M} var(Y_j)$

$\therefore \text{SE}(\hat{\Delta}) = \sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}$