Stat 23400 Lecture 15: Analysis of Two Sample Data (Unpaired), Comparing Two Proportions

Outline for Today

- Topics:
  - Analysis of Two Sample (Unpaired) Data (MMSA 10.1, 10.2 and OIS 7.3)
  - Comparing Two Proportions (MMSA 10.4 and OIS 6.2)
- Hw8 will be posted on Canvas.

Analysis of Two Sample (Unpaired) Data

Two Sample Problems: Motivation

- Is the air more polluted in Chicago than in LA?
- Are smokers suffering less from depression than non-smokers?
- Are the response in the treatment group different from that in the control group?

Two Sample Problems: Motivation

- The goal of two-sample inference is to compare the responses in two groups.
- Each group is considered to be an $i.i.d.$ sample from a distinct population.
- The responses in each group are independent of those in the other group (in addition to being independent of each other).

For example, suppose we have a SRS of size n_1 drawn from a $N(\mu_1, \sigma_1^2)$ population and an independent SRS of size n_2 drawn from a $N(\mu_2, \sigma_2^2)$ population.

Example: The first sample might be heights of male students and the second heights of female students.

We want to know make inference on the difference of the population mean $\mu_1 - \mu_2$.

Two Sample Problems: Formulation

- To compare μ_1 and μ_2, i.i.d. samples from each of the two populations are taken.

  i.i.d. sample of size n_1 from population 1 : $X_{1,1}, X_{1,2}, \ldots X_{1,n_1}$

  i.i.d. sample of size n_2 from population 2 : $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$

- **The responses in each group need to be independent of those in the other group**
- Unlike the matched pairs design, there is **no matching** of the observations in the two samples and the two samples may be of different sizes

Two Sample Problems: Statistic

- How do we estimate $\mu_1 - \mu_2$?

- How close is your estimate to $\mu_1 - \mu_2$? How to get CI and do test?

Case 1: σ_1 and σ_2 Are Known

Assuming both populations are normal, we can derive the distribution of the difference of two means

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Therefore the two sample z-statistic is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Case 1: Comparing Two Means when σ's are known

- Find a $(1-\alpha)$ CI for $\mu_1 - \mu_2$.

- How to test the hypothesis $H_0 : \mu_1 = \mu_2$?

Case 2: Comparing Two Means with σ_1 and σ_2 are unknown and different

- Recall the two sample z-statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \text{ under } H_0$$

- Of course, σ_1^2 and σ_2^2 are often unknown. Thus we substitute them by the sample variances S_1^2 and S_2^2, where

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_1)^2}{n_1 - 1} \quad \text{and} \quad S_2^2 = \frac{\sum_{i=1}^{n_2}(X_{2,i} - \overline{X}_2)^2}{n_2 - 1}.$$

- The **two-sample t-statistic** is

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Approximate Distribution of the Two-Sample t-Statistic

- Unfortunately, the two-sample t-statistic does NOT have a t-distribution (because the difference of two T-RVs is NOT a T-RV!)
- The two-sample t-statistic has an **approximate T_k distribution**. For the degrees of freedom k we have two formulas:
  1. Satterthwaite-Welch formula:

$$k = \frac{(w_1 + w_2)^2}{w_1^2/(n_1 - 1) + w_2^2/(n_2 - 1)}, \quad \text{where} \quad \begin{array}{l} w_1 = s_1^2/n_1, \\ w_2 = s_2^2/n_2. \end{array}$$

  2. simple formula: $\boxed{k = \min(n_1 - 1, n_2 - 1)}$

- Comparison of the two formulas:
  - The Satterthwaite-Welch formula is more accurate. It gives larger d.f. and yields shorter CIs and smaller p-value
  - The simple formula is conservative. It yields wider CIs and larger p-values than the actual p-value
  - It is fine to **just use the simple formula**.

Case 2: CI and Hypothesis Testing

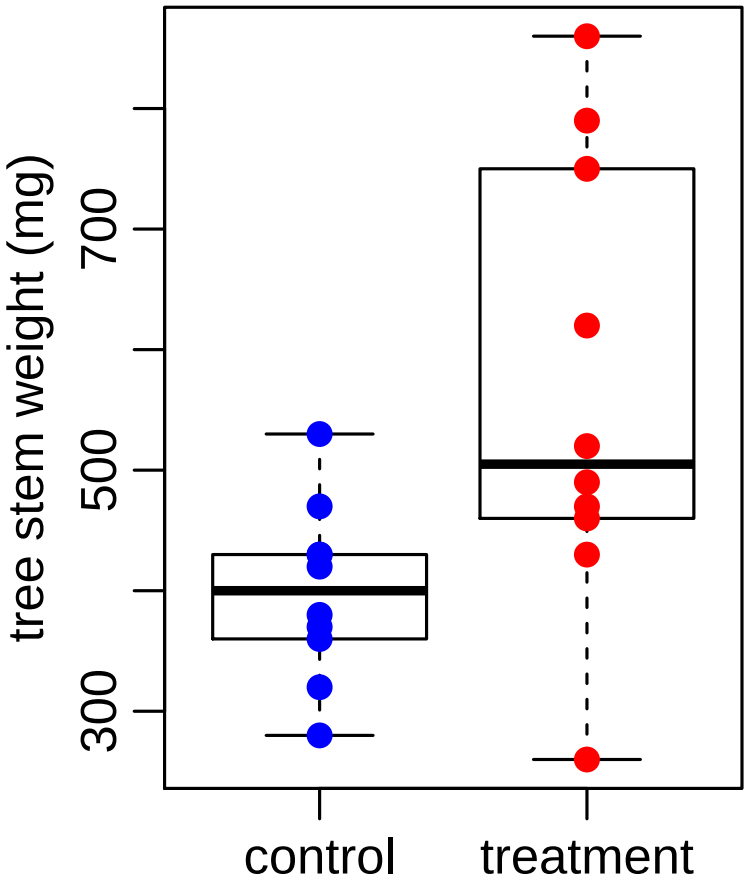
- Find a $(1-\alpha)$ CI for $\mu_1 - \mu_2$.

- How to test the hypothesis $H_0 : \mu_1 = \mu_2$?

Case 2 Example: Nitrogen Effect on Tree Growth

There are 20 northern red oak seedlings, all grown in the same type of soil in the same greenhouse.
Half received nitrogen, and half did not.
After 140 days, stem weights (in milligrams) were:

Control no nitrogen		Treatment nitrogen	
320	430	260	750
530	360	430	790
280	420	470	860
370	380	490	620
470	430	520	460
$\bar{x}_C = 399$		$\bar{x}_T = 565$	
$s_C = 72.79$		$s_T = 186.74$	
$n_C = 10$		$n_T = 10$	



Case 2 Example: CI for the Nitrogen Effect on Tree Growth

Find a 95% CI for the nitrogen effect on stem weight.

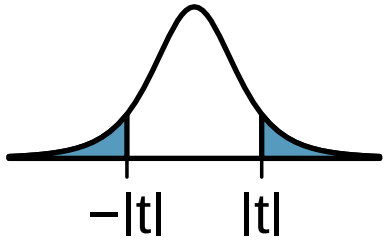
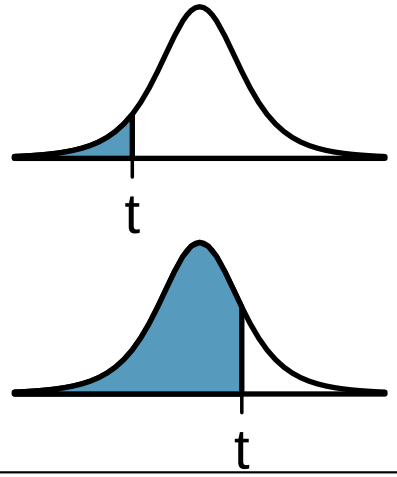
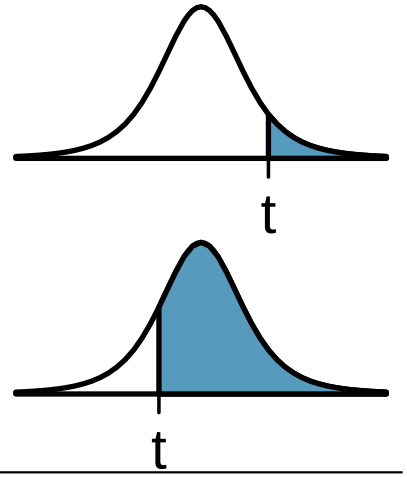
one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	9	1.38	1.83	2.26	2.82	3.25

Case 2: Hypothesis Tests for $\mu_1 - \mu_2$

To test H_0: $\mu_1 - \mu_2 = 0$, the two-sample t-statistic is

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}, \quad \text{Under } H_0\,,\ T \overset{approx}{\sim} t_k$$

where the df is $k = \min(n_1 - 1, n_2 - 1)$, or the one given by the software formula, and the p-value is computed as follows depending on H_A.

H_a	$\mu_1 - \mu_2 \neq 0$	$\mu_1 - \mu_2 < 0$	$\mu_1 - \mu_2 > 0$
p-value			

The bell curve above is the t-curve with k degrees of freedom.

Case 2 Example: Test for the Nitrogen Effect on Tree Growth

How to testing $H_0 : \mu_T - \mu_C = 0$ v.s. $H_A : \mu_T - \mu_C \neq 0$?

one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 9	1.38	1.83	2.26	2.82	3.25

Two-Sample Tests/CIs in R

```
> ctrl = c(320,430,530,360,280,420,370,380,470,430)
> trt = c(260,750,430,790,470,860,490,620,520,460)
```

By default, the R command `t.test` does NOT assume $\sigma_1 = \sigma_2$.

```
> t.test(ctrl, trt)

Welch Two Sample t-test

data:  ctrl and trt
t = -2.6191, df = 11.673, p-value = 0.02286
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-304.52438  -27.47562
sample estimates:
mean of x mean of y
399       565
```

Note the df $= 11.673$ given above is based on the software formula, which is more accurate than the simple formula.

Robustness of Two-Sample t-Procedures

As long as the sample sizes are not too small, the two-sample t-statistic

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

can still be well-approximated by t-distribution even when the populations are not normal. This is the so-called **robustness** of the two-sample t-procedures.

- It is generally good if $n_1 + n_2$ is not too small (**both** $n_1, n_2 \geq 15$), the data are **not strongly skewed, and there are no outliers**. (Check histograms or side-by-side boxplots of the data.)
- With $n_1 + n_2$ sufficiently large (**both** $n_1, n_2 \geq 30$), the approximation is good **even when the data are clearly skewed**.
- Given a fixed sum of the sample sizes $n = n_1 + n_2$ the t-approximation works **the best when the sample sizes are equal** $n_1 = n_2$ (choose equal sample sizes if you can).

Case 3: What if $\sigma_1 = \sigma_2$?

- **Inference:** the difference of the population mean $\mu_1 - \mu_2$.
- **Data:**
  i.i.d. sample of size n_1 from population 1 : $X_{1,1}, X_{1,2}, \ldots X_{1,n_1}$
  i.i.d. sample of size n_2 from population 2 : $X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$

- When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we can combine s_1^2 and s_2^2 to get a better **estimator for** σ^2, which is called **pooled sample variance**.
- **Pooled sample variance** S_p^2

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$
$$= \frac{\sum_i (X_{1,i} - \overline{X}_1)^2 + \sum_i (X_{2,i} - \overline{X}_2)^2}{n_1 + n_2 - 2}.$$

Case 3: The Pooled Two-Sample t-Statistic (When $\sigma_1 = \sigma_2$)

When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the two-sample t-statistic then becomes

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which is called **the pooled two-sample t-statistic**.

- $T \sim T_{n_1+n_2-2}$ exactly when the two populations are normal.
- $T \sim T_{n_1+n_2-2}$ approximately as long as the sample size n_1, n_2 are not too small.

Remark: since we need to estimate only ONE variance parameter σ^2, the degrees of freedom, $n_1 + n_2 - 2$ are greater than the degrees of freedom that we have when we need to estimate TWO variances σ_1^2, σ_2^2.

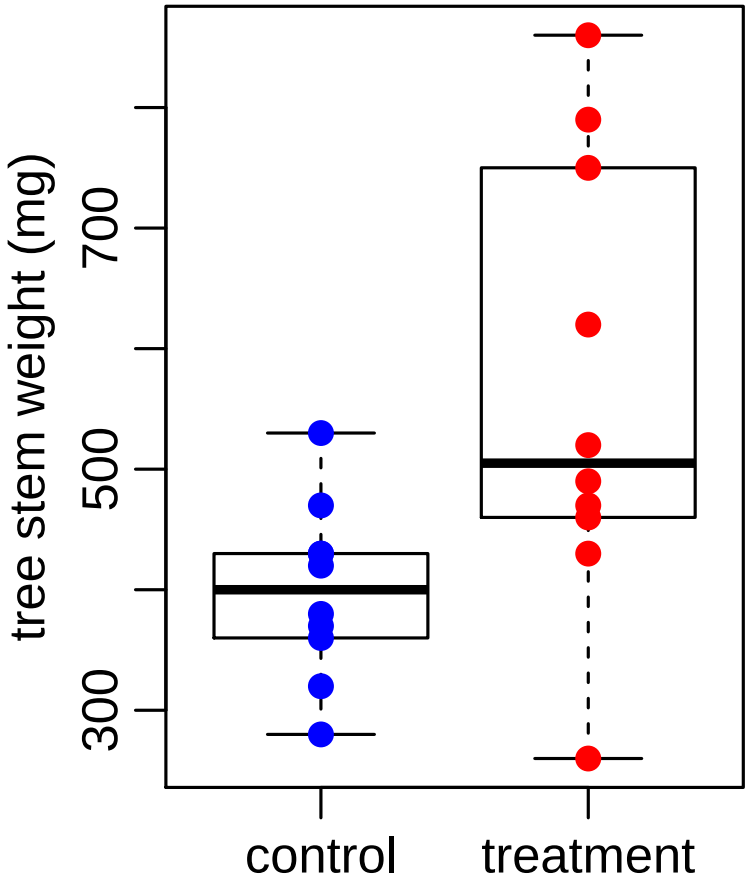
Case 3: Inference for Two Sample Problems w/ Equal but Unknown σs

- Find a $(1-\alpha)$ CI for $\mu_1 - \mu_2$.

- How to test the hypothesis $H_0 : \mu_1 = \mu_2$?

Example - Tree Growth Revisit: Assuming $\sigma_1 = \sigma_2$

Control no nitrogen		Treatment nitrogen	
320	430	260	750
530	360	430	790
280	420	470	860
370	380	490	620
470	430	520	460
mean $= 399$		mean $= 565$	
SD $= 72.79$		SD $= 186.74$	
$n_C = 10$		$n_T = 10$	



Assuming $\sigma_1 = \sigma_2$, find the pooled SD.

Example - Tree Growth Revisit: CI and Hypothesis Test

- Find a 95% CI for $\mu_T - \mu_C$.

- How to test $H_0 : \mu_T - \mu_C = 0$ against $H_A : \mu_T - \mu_C \neq 0$, assuming $\sigma_1 = \sigma_2$?

Case 3 Two-Sample Tests/CIs in R

One can force σ_1, σ_2 to be equal by the argument `var.equal = T`.

```
> t.test(ctrl, trt, var.equal = T)

Two Sample t-test

data:  ctrl and trt
t = -2.6191, df = 18, p-value = 0.01739
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-299.15788  -32.84212
sample estimates:
mean of x mean of y
399       565
```

Which Two-Sample Tests/CIs to Use?

We have introduced two different two-sample tests/CIs:

- the one assuming $\sigma_1 = \sigma_2$ used the **pooled SD**.
- the one assuming $\sigma_1 \neq \sigma_2$ is called **Welch's method**.

They usually provide different answers when:

- the sample SDs are very different, and
- the sizes of the groups are also very different

So which method should I use?

- When σ_1 and σ_2 are indeed equal, the method based on pooled SD is more powerful.
- However, it is usually hard to check whether $\sigma_1 = \sigma_2$. So it's safer to use Welch's method.
- The pooled SD method is only appropriate when background research indicates the population SDs are nearly equal.

Comparing Two Proportions

Comparing Two Proportions

- Sometimes, we need to compare two proportions. For example, we may be interested in knowing how effective the new antiviral drug, Paxlovid, at preventing severe COVID-19 symptoms.
- Suppose we have two populations A and B with unknown proportions p_1 and p_2 respectively.
- We may choose an SRS of size n_1 from a large population having proportion p_1 of successes, and an independent SRS of size n_2 from another population having proportion p_2 of successes.

Population	Population Proportion	Sample Size	Count of Successes	Sample Proportion
A	p_1	n_1	X_1	$\widehat{p}_1 = X_1/n_1$
B	p_2	n_2	X_2	$\widehat{p}_2 = X_2/n_2$

Large Sample Confidence Intervals for $p_1 - p_2$

- Derive the large sample $(1 - \alpha)$ CI.
- How large the sample sized need to be?

Example: Aspirin and Heart Attacks

The Physicians' Health Study was a 5-year randomized study published testing whether regular intake of aspirin reduces mortality from cardiovascular disease[1].

- Participants were male physicians 40-84 years old in 1982 with no prior history of heart attack, stroke, and cancer, no current liver or renal disease, no contraindication of aspirin, no current use of aspirin
- Every other day, one group of physicians took one aspirin tablet, the other group took a placebo.
- Response: whether the participant had a heart attack (including fatal or non-fatal) during the 5 year period.

[1] *Source*: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262-64,1988.

Example: Aspirin and Heart Attacks

Result:

	Heart Attack?		
Group	Yes	No	Sample Size
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow \hat{p}_1 = \frac{189}{11034} \approx 0.0171$

$\Rightarrow \hat{p}_2 = \frac{104}{11037} \approx 0.0094$

Find a 99% CI for $p_1 - p_2$, where p_1 and p_2 are the proportions of getting a heart attack in the placebo group, and that in the treatment group, respectively.

Example: Aspirin and Heart Attacks

Conclusion:

- We are 99% confidence that the incidence rate of heart attack in the aspirin group is between 0.0037 and 0.0117 lower than in the placebo group.
- As the 99% CI does not contain 0, the incidence rate of heart attack was significantly lower in aspirin group than in the placebo group
- Can we claim that taking aspirin every other day is effective in reducing the chance of heart attack?

  Yes, because it was a randomized, double-blind, placebo-controlled experiment.

Testing the Equality of Two Proportions

How to test $H_0 : p_1 = p_2$?

- Pooled estimator of p:
- SE under H_0:
- z-statistic under H_0:
- Check conditions:

Example: Aspirin and Heart Attacks

Group	Sample Size	Heart Attack
Placebo	11034	189
Aspirin	11037	104

Perform the hypothesis testing $H_0 : p_1 = p_2$ v.s. $H_A : p_1 > p_2$.

Example: Partisanship 2015

A Gallop poll in 2015 based on a random sample of 12137 adults in U.S. (aged ≥ 18), found that 29% self-identified as Democrats, 26% as Republicans, and 45% as independent or other.
True or False and explain: a 95% confidence interval for the difference of proportions of American adults self-identified as Democrats and Republicans $p_D - p_R$ is

$$0.29-0.26 \pm 1.96\sqrt{\frac{0.29(1-0.29)}{12137}+\frac{0.26(1-0.26)}{12137}} = (0.019, 0.041)$$

- How many samples are there? One or two?
- The two sample percentages, 29% and 26%, are calculated based on the same sample. They were not independent, but negatively correlated. The more people identified as Democrats, the fewer identified as Republicans. One cannot use a two-sample CI here.

---

http://http://news.gallup.com/poll/188096/democratic-republican-identification-near-historical-lows.aspx

Example: Partisanship 2015 v.s. 2011

Continue the previous example.
Another survey of 15,000 American adults in 2011 found that 35.3% identified as Democrats, 34.0% as Republicans, and 30.7% as independent or other. Assume both surveys in 2011 and 2015 were both based on simple random samples. Can we test whether there were more American adults self-identified as independent or other in 2015 than in 2011 using a two-sample z-test for proportions?

Example: Partisanship 2015 v.s. 2011

Let p_{2011} and p_{2015} be the percentages identified as independent or others in 2011 and in 2015, respectively.

Test $\mathrm{H}_0 : p_{2011} = p_{2015}$ vs $\mathrm{H}_A : p_{2011} < p_{2015}$.

Statistical Inference for Proportions in R

Example: Aspirin and Heart Attacks Revisited

Data:

	Heart Attack?		
Group	Yes	No	Sample Size
Placebo	189	10845	11034
Aspirin	104	10933	11037

$\Rightarrow \hat{p}_1 = \frac{189}{11034} \approx 0.0171$

$\Rightarrow \hat{p}_2 = \frac{104}{11037} \approx 0.0094$

Next, we will discuss how to use the function prop.test in R to do inference for a single proportion or two proportions.

```
# set up the data in R
> xc -> 189
> xt -> 104
> nc -> 11034
> nt -> 11037
```

Using "prop.test" for a single proportion

Let p_t be the proportion of treatment group developing heart attack. Consider the following hypothesis test:

$$H_0 : p_t = 0.01 \quad \text{v.s.} H_A : p_t < 0.01 \,.$$

```
> prop.test(xt, nt, p = 0.01, alternative = "less")

1-sample proportions test with continuity correction

data:  xt out of nt, null probability 0.01
X-squared = 0.31535, df = 1, p-value = 0.2872
alternative hypothesis: true p is less than 0.01
95 percent confidence interval:
 0.00000000 0.01110918
sample estimates:
         p
0.00942285
```

Note that "prop.test" is using a Chi-squared test. The C.I. is calculated based on Wilson Score Method.

Testing the Equality of Two Proportions

Consider the following hypothesis testing problem,

$$H_0 : p_t = p_c \quad v.s. H_A : p_t < p_c \, .$$

```
> prop.test(c(xt, xc), c(nt, nc), alternative = "less")

2-sample test for equality of proportions with continuity
correction

data:  c(xt, xc) out of c(nt, nc)
X-squared = 24.429, df = 1, p-value = 3.855e-07
alternative hypothesis: less
95 percent confidence interval:
 -1.000000000 -0.005082393
sample estimates:
    prop 1     prop 2
0.00942285 0.01712887
```