Stat 411/511

The Randomization Test

Oct 16 2015

Charlotte Wickham

stat511.cwick.co.nz

Today

Review randomization model

Conduct randomization test

What about CIs?

Using a t-distribution as an approximation to the randomization distribution.

Creativity scores in two motivation groups, and their summary statistics

Does intrinsic motivation improve creativity?

	Motivation Group			
	Assigned randomly by researcher			
	Intrinsic		***Extrinsic***	
	12.0	20.5	5.0	17.4
	12.0	20.6	5.4	17.5
	12.9	21.3	6.1	18.5
	13.6	21.6	10.9	18.7
	16.6	22.1	11.8	18.7
	17.2	22.2	12.0	19.2

The intrinsic group has an average creativity score 4.1 points higher than the extrinsic group

Sample Size:	24		23	
Average:	19.88	–	15.74	= 4.1
Sample Standard Deviation:	4.44		5.25	

Read: 1.1.1 in Sleuth

Questionnaires given creative writers, to rank intrinsic and extrinsic reasons for writing

INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

__ You get a lot of pleasure out of reading something good that you have written.
__ You enjoy the opportunity for self-expression.
__ You achieve new insights through your writing.
__ You derive satisfaction from expressing yourself clearly and eloquently.
__ You feel relaxed when writing.
__ You like to play with words.
__ You enjoy becoming involved with ideas, characters, events, and images in your writing.

*List of **extrinsic** reasons for writing*

*List of **intrinsic** reasons for writing*

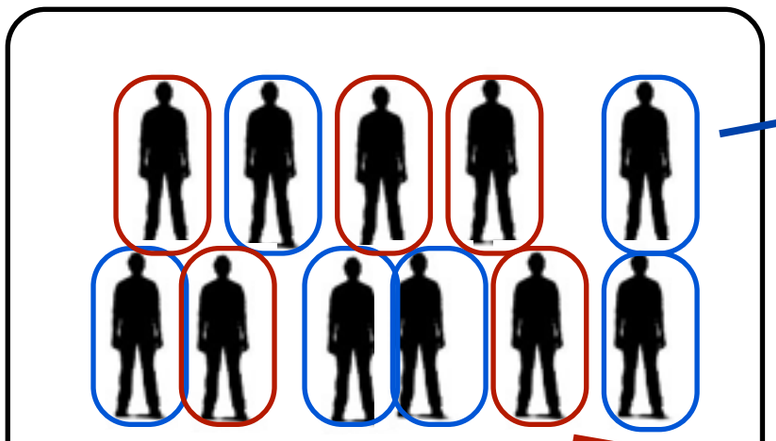
INSTRUCTIONS: Please rank the following list of reasons for writing, in order of personal importance to you (1 = highest, 7 = lowest).

__ You realize that, with the introduction of dozens of magazines every year, the market for free-lance writing is constantly expanding.
__ You want your writing teachers to be favorably impressed with your writing talent.
__ You have heard of cases where one bestselling novel or collection of poems has made the author financially secure.
__ You enjoy public recognition of your work.
__ You know that many of the best jobs available require good writing skills.
__ You know that writing ability is one of the major criteria for acceptance into graduate school.
__ Your teachers and parents have encouraged you to go into writing.

The randomized experiment model

Key idea: there is no population, and no sampling!

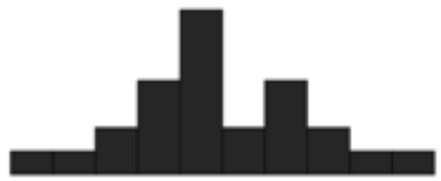
Some experimental units



Doesn't matter where they came from.

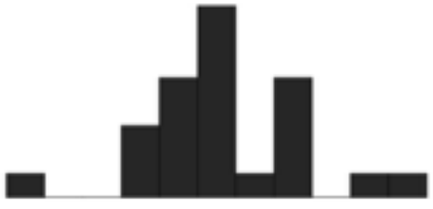
assigned at random

get treatment 1



observed responses for units assigned to treatment 1

get treatment 2



observed responses for units assigned to treatment 2

Chance only enters through the random assignment of units to treatments

Remember: Statistical testing

1. Set up the null hypothesis
(and alternative hypothesis)

2. Calculate the **test statistic**

3 Evaluate the evidence against the null hypothesis by comparing the test statistic to test statistics expected under the null hypothesis, the **null distribution**.

To do a test all we really need to know is the null distribution.
I.e. the randomization distribution if the null was true.

The evidence is summarized by a **p-value**, the probability we would see such an extreme test-statistic if the null hypothesis is true.

4. **If the p is low, the null must go!**

Reject or fail to reject the null hypothesis

Randomization Distribution

The randomization distribution is the histogram of all values for the statistic from all possible ways the experimental units could have been randomly assigned to groups.

In the sampling model, the reason there is variability in a sample statistic is because we induced variability by taking a random sample. We describe the variability using the sampling distribution of the statistic.

In the randomized experiment model, the only reason we see variability in group statistics is because we induced variability by randomly assigning people to groups. We describe the variability using the randomization distribution of the statistic.

In randomized experiments it's the relationship between the randomization distribution and the effect of the treatment that allow us to make inferences.

DISPLAY 1.7 A different group assignment for the creativity study, and a different result

	Creativity score	Actual grouping	Another grouping	Creativity score	Actual grouping	Another grouping
subject 1	12.0	Intrinsic(2)	1	5.0	Extrinsic(1)	2
subject 2	12.0	Intrinsic	2	5.4	Extrinsic	2
...	12.9	Intrinsic	1	6.1	Extrinsic	1
	13.6	Intrinsic	2	10.9	Extrinsic	2
	16.6	Intrinsic	2	11.8	Extrinsic	1
	17.2	Intrinsic	1	12.0	Extrinsic	1
	17.5	Intrinsic	2	12.3	Extrinsic	1
	18.2	Intrinsic	2	14.8	Extrinsic	2
	19.1	Intrinsic	1	15.0	Extrinsic	2
	19.3	Intrinsic	2	16.8	Extrinsic	2
	19.8	Intrinsic	2	17.2	Extrinsic	2
	20.3	Intrinsic	2	17.2	Extrinsic	1
	20.5	Intrinsic	1	17.4	Extrinsic	2
	20.6	Intrinsic	2	17.5	Extrinsic	2
	21.3	Intrinsic	1	18.5	Extrinsic	2
	21.6	Intrinsic	2	18.7	Extrinsic	1
	22.1	Intrinsic	1	18.7	Extrinsic	1
	22.2	Intrinsic	2	19.2	Extrinsic	1
	22.6	Intrinsic	1	19.5	Extrinsic	1
	23.1	Intrinsic	1	20.7	Extrinsic	1
	24.0	Intrinsic	1	21.2	Extrinsic	1
	24.3	Intrinsic	1	22.1	Extrinsic	2
	26.7	Intrinsic	1	24.0	Extrinsic	2
	29.7	Intrinsic	1			

Averages from actual grouping

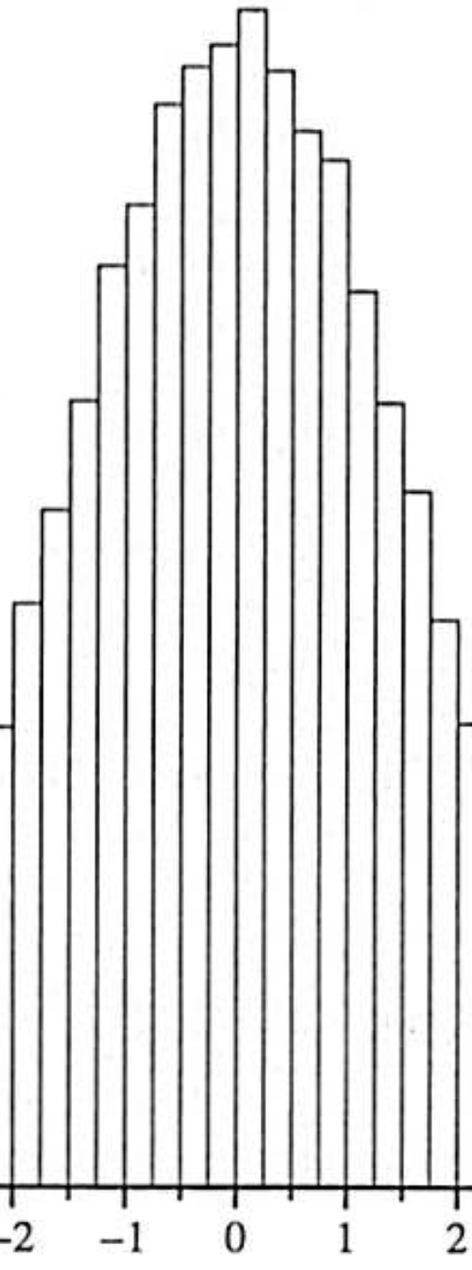
Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

Averages from another grouping

Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

500,000 test-statistics, from 500,000 random regroupings

DISPLAY 1.8 Histogram of differences between group averages, from 500,000 regroupings of the creativity study data



An approximation to the distribution we would expect from chance alone

null distribution

1302/500000 are as small or smaller than -4.14

1335/500000 are as large or larger than 4.14

The value from the data

two-sided p-value = (1302 + 1335)/500000 = 0.005274

A statistical summary

There is strong evidence that the effect of the intrinsic questionnaire is not the same as the extrinsic questionnaire in **this set of subjects** (randomization test, p-value = 0.005).

no population inference

Randomization test

Procedure

We pick a test-statistic and calculate the observed value.

To get a p-value we compare our observed test-statistic to the randomization distribution of test-statistics obtained by assuming the null is true.

The p-value will be the proportion of test-statistics in the randomization distribution that are as or more extreme than the observed test-statistic.

Explain the steps in a randomization test for testing for a treatment effect in a controlled experiment.

The Randomization test

No sampling from a population, so no assumptions on a population.

Assumed random allocations to groups.

We used the difference in sample averages as our test statistic, but we could have used something else.

Null hypothesis: there is no difference between treatments (for any subject)

What's the alternative?

Alternative hypothesis: there is some difference between treatments for at least one subject.

Some ways the alternative could be true:

one treatment induces a fixed additive change in response, δ, for all subjects (a.k.a the additive treatment model)

one treatment induces a larger mean response across subjects

one treatment induces a larger variance in response across subjects

one treatment induces more skewness in response across subjects

We might tailor our test statistic to the type of deviation from the null we expect to see, but different test statistics don't change the alternative hypothesis

Confidence Intervals

There are no parameters of interest so, there are no confidence intervals of interest.

We could assume a particular type of alternative **that is** parameterized. Then we could make confidence intervals on that parameter.

(e.g. additive treatment model).

this is what the Sleuth does
Section 2.4.1

The additive treatment model

The **additive treatment model**, says:

A subject's response on treatment 2 is their response on treatment 1 plus some fixed number, δ, that is the same for everyone.

unknown parameter

In math, consider subject *i*

Y_{i1} = Observed value of subject i under treatment 1

Y_{i2} = Observed value of subject i under treatment 2

$Y_{i2} = Y_{i1} + \delta$ for all i

If we have random allocation to groups **and** we are willing to assume the additive treatment model, then our hypotheses in the randomization test become:

Null hypothesis: the treatment effect is zero, $\delta = 0$

Alternative hypothesis: the treatment effect is not zero, $\delta \neq 0$

Creativity case study

Let’s assume the additive treatment model.

Creativity score given Intrinsic Questionnaire =

Creativity score given Extrinsic Questionnaire + δ

Let’s also use the t-statistic, instead of the difference in sample averages, as our test statistic.

$$\frac{\overline{Y}_1 - \overline{Y}_2}{\mathrm{SE}_{\overline{Y}_1 - \overline{Y}_2}}$$

DISPLAY 1.7 A different group assignment for the creativity study, and a different result

	Creativity score	Actual grouping	Another grouping	Creativity score	Actual grouping	Another grouping
subject 1	12.0	Intrinsic(2)	1	5.0	Extrinsic(1)	2
subject 2	12.0	Intrinsic	2	5.4	Extrinsic	2
	12.9 ...	Intrinsic	1	6.1	Extrinsic	1
	13.6	Intrinsic	2	10.9	Extrinsic	2
	16.6	Intrinsic	2	11.8	Extrinsic	1
	17.2	Intrinsic	1	12.0	Extrinsic	1
	17.5	Intrinsic	2	12.3	Extrinsic	1
	18.2	Intrinsic	2	14.8	Extrinsic	2
	19.1	Intrinsic	1	15.0	Extrinsic	2
	19.3	Intrinsic	2	16.8	Extrinsic	2
	19.8	Intrinsic	2	17.2	Extrinsic	2
	20.3	Intrinsic	2	17.2	Extrinsic	1
	20.5	Intrinsic	1	17.4	Extrinsic	2
	20.6	Intrinsic	2	17.5	Extrinsic	2
	21.3	Intrinsic	1	18.5	Extrinsic	2
	21.6	Intrinsic	2	18.7	Extrinsic	1
	22.1	Intrinsic	1	18.7	Extrinsic	1
	22.2	Intrinsic	2	19.2	Extrinsic	1
	22.6	Intrinsic	1	19.5	Extrinsic	1
	23.1	Intrinsic	1	20.7	Extrinsic	1
	24.0	Intrinsic	1	21.2	Extrinsic	1
	24.3	Intrinsic	1	22.1	Extrinsic	2
	26.7	Intrinsic	1	24.0	Extrinsic	2
	29.7	Intrinsic	1			

Averages from actual grouping

Group	Average	Difference
Intrinsic (2)	19.88	4.14
Extrinsic (1)	15.74	

Averages from another grouping

Group	Average	Difference
Group 1	18.87	2.07
Group 2	16.80	

Actual grouping

	Extrinsic	Intrinsic
sample avg	15.74	19.88
sample sd	5.25	4.44
sample n	23	24

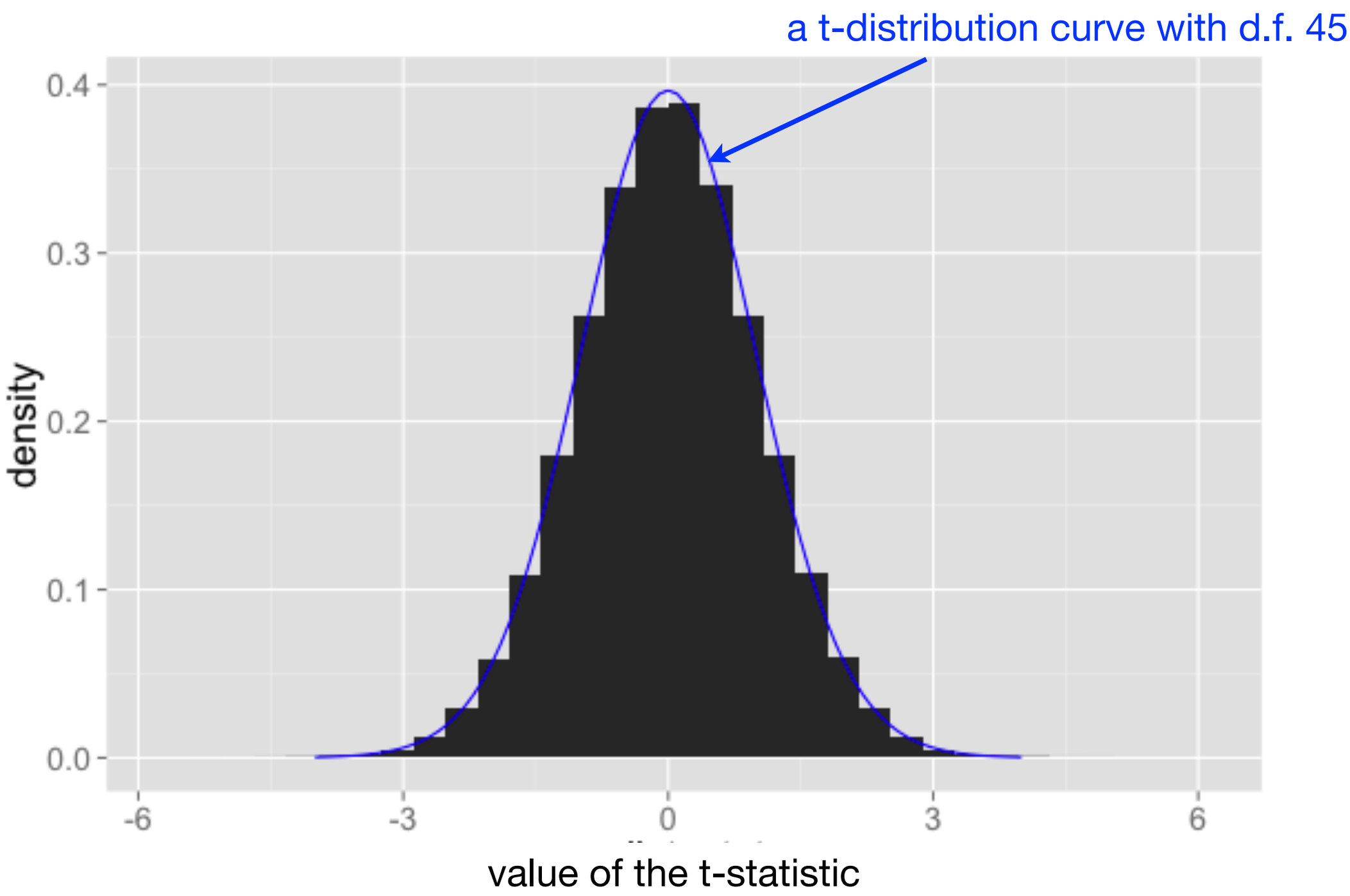
two sample t-stat = 2.92

Another grouping

	1	2
sample avg	18.87	16.80
sample sd	5.46	4.88
sample n	24	23

two sample t-stat = 1.37

A histogram of 500,000 t-statistics, from random regroupings of the creativity study.



The t-distribution is a very good approximation to the randomization distribution of the t-statistic

In a randomized experiment,

The result from a two sample t-test **is approximately the same as** a randomization test, when:

- you assume the additive treatment model
- the observed responses aren't too non-Normal

This is pretty amazing! The two sample t-test arose from a completely different model,
random sampling from populations.

This means we have increased the number of situations we can do a t-test.

We can do a two sample t-test when we have samples from Normal populations,

and when we have a randomized experiment of two treatments, with data that isn't too non-Normal. more on too non-Normal later...

The scope of inference (population or causal) is still completely restricted by the study design.

```
> t.test(Score ~ Treatment, data = case0101,
var.equal = TRUE)

   Two Sample t-test

data:  Score by Treatment
t = -2.9259, df = 45, p-value = 0.005366
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -6.996973 -1.291432
sample estimates:
mean in group Extrinsic mean in group Intrinsic
                15.73913                 19.88333
```

A statistical summary

based on the t-test

There is strong evidence that the effect of the intrinsic questionnaire is not the same as the extrinsic questionnaire in this set of subjects (two sample t-test, p-value = 0.005).

We estimate **the effect of the intrinsic questionnaire** is **to add 4.14 points** to the creativity score **compared to the extrinsic questionnaire**.

With 95% confidence, the effect of the intrinsic questionnaire is to add between 1.29 and 7.00 points to the creativity score compared to the extrinsic questionnaire.

note the language of an **additive treatment model**

"the effect … is to add …"