Tea: A High-level Language and Runtime System for Automating Statistical Analysis

Eunice Jun

with Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke

To appear at ACM UIST in October 2019.
Hi all,

I'm having a bit of trouble working out answers to percentage questions about normal distributions.

Question [Q] Is classification in ML the opposite of ANOVA in classical stats?

So ANOVA and Mixed Models tell you whether a certain factor had a significant effect on the response and whether levels of a factor had a significantly different effect on the response.

From what I understand, things like Logistic Regression, discriminant analysis, kNN, SVM etc seem to use the response to try to predict the classes the data points belong to.

So are these approaches basically opposites of each other?

If the ANOVA contrasts is significant, would one of the classification approaches also be expected to perform well?

And if a classification approach has high accuracy, sensitivity, specificity then can you conclude that Mixed Models or ANOVA can tell different things about the response?
**Question**: What percentage of scores fall below one standard deviation above the mean?

Hi all,

I'm having a bit of trouble working out answers to percentage questions about normal distributions.

**Question**: Is classification in ML the opposite of ANOVA in classical stats?

So ANOVA and Mixed Models tell you whether a certain factor had a significant effect on the response and whether levels of a factor had a significantly different effect on the response.

From what I understand, things like Logistic Regression, discriminant analysis, k-etc seem to use the response to try to predict the classes the data points belong to.

**Question**: What statistical test should I use?

Hi all, I'm looking for guidance from someone who knows what I don't! I have a semester of stats 101 under my belt (that I college) and that's more or less the extent of my knowledge.

Project background: my workplace has a moth problem. We have a moth building and I check them once a month. Each trap location is I also have a combined data set with monthly moth catch of all I've got 1.5 years' worth of data, but since I check monthly, it's relative. Also, since I'm recording moth catch, the data is relatively skewed if I'm wrong) because many traps have caught 0 moths dur

**Question**: Can I do anything with this data?

Hello everyone! I've been reviewing some data for parents and children who received therapy. The way their mental health is measured is with 2 tests, so both parents and children should complete both of these tests before and after the treatment.

However... Even though 20 children received therapy, there are few cases where there is both pre and post treatment data (between 5-8 for both tests for both parents and children). I had many ideas for how I could analyse this data before, but now I'm not sure I can do anything with this aside from a few graphs (which I've already done).

**Question**: What is the best way to analyze my dataset?

Hi there,

Statistics is not my strongest point, so I was wondering if some of you could help me out a little bit.

So far I have almost 200 respondents who filled in my survey. They were given 6 sets, each having 3 statements, prior to being asked where they would buy a certain product (e.g. offline or online), which my moderator being the price of the product (high priced vs. low priced). Each set measures certain characteristics (e.g. price-conscious, time-conscious etc.). Now I want to test my hypotheses that price-conscious consumers buy high priced products rather offline than online.

What would be the best way to do this in SPSS?
Statistics is hard

Hi all, I'm looking for guidance. I have a seminar at college and that's more than sure don't! I have a seminar at college and that's more than sure don't! I have a seminar at college and that's more than sure don't! I have a seminar at college and that's more than sure don't! I have a seminar at college and that's more than sure don't!

Project background: my building and I check the rice. I also have a combined ~1.5 years' worth of data. Also, since I'm recording me if I'm wrong) because I'm recording what I don't know.

What percentage of scores fall below one standard deviation above the mean?

Hi, I'm having a bit of trouble with distributions.

Question: [Q] Is class stats?

So ANOVA and Mixed models response and the response. From what I understand, etc. seem to use the response.

Posted by u/IzzyBeel 14 hot

Question: [Q] What statistics can I do anything with this data?

Also, since I'm recording, I've been keeping track of sets. Of you could help me out.

I had given 6 sets, I'd like to buy a certain product. The product (high priced rice-conscious, time-sensitive consumers buy...
Which statistical test should I use?

Does an optimization make my program run faster?

$H_1$: Optimized code runs faster

$H_0$: Difference between run times due to chance
Which statistical test should I use?

Does an optimization make my program run faster?

<table>
<thead>
<tr>
<th>Pearson’s r</th>
<th>Welch’s</th>
<th>Fisher’s Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointbiserial</td>
<td>F-test</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Kendall’s T</td>
<td>Repeated measures</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Spearman’s p</td>
<td>one-way ANOVA</td>
<td>MANOVA</td>
</tr>
<tr>
<td>Student’s t-test</td>
<td>Factorial ANOVA</td>
<td>ANCOVA</td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Two-way ANOVA</td>
<td>MANCOVA</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>Kruskal Wallis</td>
<td>McNemar</td>
</tr>
<tr>
<td>Wilcoxon signed rank</td>
<td>Friedman</td>
<td>Chi Square</td>
</tr>
</tbody>
</table>

$H_1$: Optimized code runs faster

$H_0$: Difference between run times due to chance
Which statistical test should I use?

Does an optimization make my program run faster?

It depends!

- Pearson’s r
- Welch’s
- Fisher’s Exact
- Wilcoxon signed rank
- Friedman
- Chi Square

$H_1$: Optimized code runs faster

$H_0$: Difference between run times due to chance
Which statistical test should I use?

How do financial incentives affect users’ performance?

$H_1$: Higher financial incentives, better user performance
$H_0$: Difference in performance due to chance
<table>
<thead>
<tr>
<th>Statistical Test</th>
<th>Pearson’s r</th>
<th>Spearman’s p</th>
<th>Student’s t-test</th>
<th>Paired t-test</th>
<th>Mann-Whitney U</th>
<th>Wilcoxon signed rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointbiserial</td>
<td>F-test</td>
<td>one-way ANOVA</td>
<td>Factorial ANOVA</td>
<td>Two-way ANOVA</td>
<td>Kruskal Wallis</td>
<td>Friedman</td>
</tr>
<tr>
<td>Kendall’s T</td>
<td>Repeated measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fisher’s Exact</td>
<td>Linear regression</td>
<td></td>
<td></td>
<td></td>
<td>McNemar</td>
<td>Chi Square</td>
</tr>
</tbody>
</table>

**H₁:** Higher financial incentives, better user performance  
**H₀:** Difference in performance due to chance
Which statistical test should I use?

How do financial incentives affect users’ performance?

Pearson’s r  Welch’s  Fisher’s Exact

It depends!

Wilcoxon signed rank  Friedman  Chi Square

$H_1$: Higher financial incentives, better user performance

$H_0$: Difference in performance due to chance
Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

$H_1$: Tea first tastes better

$H_0$: Difference in taste due to chance
Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

<table>
<thead>
<tr>
<th>Test</th>
<th>Test</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson’s r</td>
<td>Welch’s</td>
<td>Fisher’s Exact</td>
</tr>
<tr>
<td>Pointbiserial</td>
<td>F-test</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Kendall’s T</td>
<td>Repeated measures</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Spearman’s p</td>
<td>one-way ANOVA</td>
<td>MANOVA</td>
</tr>
<tr>
<td>Student’s t-test</td>
<td>Factorial ANOVA</td>
<td>ANCOVA</td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Two-way ANOVA</td>
<td>MANCOVA</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>Kruskal Wallis</td>
<td>McNemar</td>
</tr>
<tr>
<td>Wilcoxon signed rank</td>
<td>Friedman</td>
<td>Chi Square</td>
</tr>
</tbody>
</table>

$H_1$: Tea first tastes better

$H_0$: Difference in taste due to chance
Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

| Pearson’s $r$ | Welch’s $F$-test | **Fisher’s Exact**
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pointbiserial</td>
<td>Linear regression</td>
<td></td>
</tr>
<tr>
<td>Kendall’s $T$</td>
<td>Repeated measures</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Spearman’s $p$</td>
<td>One-way ANOVA</td>
<td>MANOVA</td>
</tr>
<tr>
<td>Student’s $t$-test</td>
<td>Factorial ANOVA</td>
<td>ANCOVA</td>
</tr>
<tr>
<td>Paired $t$-test</td>
<td>Two-way ANOVA</td>
<td>MANCOVA</td>
</tr>
<tr>
<td>Mann-Whitney $U$</td>
<td>Kruskal Wallis</td>
<td>McNemar</td>
</tr>
<tr>
<td>Wilcoxon signed rank</td>
<td>Friedman</td>
<td>Chi Square</td>
</tr>
</tbody>
</table>

$H_1$: Tea first tastes better
$H_0$: Difference in taste due to chance
Which statistical test should I use?

Does tea taste better with milk-then-tea or tea-then-milk?

- Fisher’s Exact
- Linear regression
- Logistic regression
- MANOVA
- ANCOVA
- MANCOVA
- McNemar
- Chi Square

$H_1$: Tea first tastes better

$H_0$: Difference in taste due to chance
EASY

Does this optimization make my program execute faster?

How do financial incentives affect users’ performance on a task?

Does tea taste better with milk poured first then tea or tea first then milk?
Does this optimization make my program execute faster?
How do financial incentives affect users’ performance on a task?
Does tea taste better with milk poured first then tea or tea first then milk?

<table>
<thead>
<tr>
<th>EASY</th>
<th>HARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson’s r</td>
<td>Welch’s F-test</td>
</tr>
<tr>
<td>Pointbiserial</td>
<td>Repeated measures</td>
</tr>
<tr>
<td>Kendall’s T</td>
<td>one-way ANOVA</td>
</tr>
<tr>
<td>Spearman’s p</td>
<td>Factorial ANOVA</td>
</tr>
<tr>
<td>Student’s t-test</td>
<td>Two-way ANOVA</td>
</tr>
<tr>
<td>Paired t-test</td>
<td>Kruskal Wallis</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>Friedman</td>
</tr>
<tr>
<td>Wilcoxon signed rank</td>
<td>McNemar</td>
</tr>
<tr>
<td></td>
<td>Fishers’ Exact</td>
</tr>
<tr>
<td></td>
<td>Linear regression</td>
</tr>
<tr>
<td></td>
<td>Logistic regression</td>
</tr>
<tr>
<td></td>
<td>MANOVA</td>
</tr>
<tr>
<td></td>
<td>ANCOVA</td>
</tr>
<tr>
<td></td>
<td>MANCOVA</td>
</tr>
<tr>
<td></td>
<td>Chi Square</td>
</tr>
</tbody>
</table>
Does this optimization make my program execute faster?
How do financial incentives affect users’ performance on a task?
Does tea taste better with milk poured first then tea or tea first then milk?

**EASY**
- Pearson’s r
- Pointbiserial
- Kendall’s T

**HARD**
- Spearman’s p
- Student’s t-test
- Paired t-test
- Mann-Whitney U
- Wilcoxon signed rank

- Welch’s F-test
- Repeated measures
- one-way ANOVA
- Two-way ANOVA
- Kruskal Wallis
- Friedman

- Fisher’s Exact
- Linear regression
- Logistic regression
- MANOVA
- ANCOVA
- MANCOVA
- McNemar
- Chi Square
<table>
<thead>
<tr>
<th>EASY</th>
<th>MEDIUM</th>
<th>HARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson's r</td>
<td>Wilcoxon</td>
<td>Fisher's Exact</td>
</tr>
<tr>
<td>Partial correlation</td>
<td>Friedman</td>
<td>Linear regression</td>
</tr>
<tr>
<td>Kendall's T</td>
<td>Repeated measures</td>
<td>Logistic regression</td>
</tr>
<tr>
<td>Spearman's p</td>
<td>One-way ANOVA</td>
<td>MANOVA</td>
</tr>
<tr>
<td>Student's t</td>
<td>Factorial ANOVA</td>
<td>ANCOVA</td>
</tr>
<tr>
<td>Non-parametric</td>
<td>Two-way ANOVA</td>
<td>MANCOVA</td>
</tr>
<tr>
<td>Mann-Whitney U</td>
<td>Kruskal-Wallis</td>
<td>McNemar</td>
</tr>
<tr>
<td>Wilcoxon signed rank</td>
<td>Friedman</td>
<td>Chi Square</td>
</tr>
</tbody>
</table>

t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
t.test(x = NULL, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

Difference between
Student's t-test and
Paired t-test

Tea eliminates this problem entirely.
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

Tea eliminates this problem entirely.
```
t.test(x = y = NULL
alternative = c("two.sided", "less", "greater")
mu = 0 paired = FALSE var.equal = FALSE
conf.level = 0.95 ...)
```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), 
mu = 0, paired = FALSE, var.equal = FALSE, 
conf.level = 0.95, ...)

Difference between Student’s t-test and Paired t-test

Each participant contributes exactly one data point

Very bad to mix up

Tea eliminates this problem entirely.
t.test(x = y, NULL, alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)

Difference between Student's t-test and Paired t-test

Each participant contributes exactly one data point

Each participant contributes exactly two data points

VERY BAD to mix up

Violate study design

Tea eliminates this problem entirely.
\texttt{t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"))}
\texttt{mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95 ...}

**Difference between Student’s t-test and Paired t-test**

- Each participant contributes **exactly one** data point
- Each participant contributes **exactly two** data points

**VERY BAD to mix up**

Violate study design

Returns **wrong statistical results**!

Tea eliminates this problem entirely.
Stats is better with Tea

Tea is correct by construction.

Tea is high-level.

Tea infers tests.

Tea improves upon expert choices, prevents common mistakes.
In [ ]: import tea

In [ ]: # Load data
tea.data("./datasets/UScrime.csv")

In [ ]: # Declare and annotate the variables of interest
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0,1]
    }
]

tea.define_variables(variables)

In [ ]: assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05,
}
import tea
tea.data('UScrime.csv')

variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0, 1]
    }
]
tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}
tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)

hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)

** ** NO STATISTICAL TEST ** **
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0, 1]
    }
]

tea.define_variables(variables)
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0,1]
    }
]

tea.define_variables(variables)
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0,1]
    }
]

tea.define_variables(variables)
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0, 1]
    }
]

tea.define_variables(variables)
variables = [
    {
        'Nominal':
            'name': 'So',
            'data type': 'nominal',
            'categories': [0, 1]
    },
    {
        'Ordinal':
            'name': 'Prob',
            'data type': 'ratio',
            'range': [0, 1]
    }
]

tea.define_variables(variables)
variables = [
    {
        'name': 'So',
        'data type': 'nominal',
        'categories': ['0', '1']
    },
    {
        'name': 'Prob',
        'data type': 'ratio',
        'range': [0,1]
    }
]

tea.define_variables(variables)
study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}

tea.define_study_design(study_design)
study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}

tea.define_study_design(study_design)
study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}

tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}

tea.assume(assumptions)
hypothesis = 'So: 1 > 0'

tea.hypothesize([['So', 'Prob'], hypothesis])
hypothesis = 'So:1 > 0'
tea.hypothesize(['So', 'Prob'], hypothesis)

Nominal, Ordinal:
Chocolate > Mint
Grade 1 < Grade 2

Ordinal, Ratio, Interval:
Grade ~ Temperature
Time of day ~ Temperature
```python
import tea
tea.data('USCrime.csv')

variables = [
    
    {'name': 'So',
     'data type': 'nominal',
     'categories': ['0', '1']
    },
    
    {'name': 'Prob',
     'data type': 'ratio',
     'range': [0,1]
    }
]

tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}

tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}

tea.assume(assumptions)

hypothesis = 'So:1 > 0'

tea.hypothesize([['So', 'Prob'], hypothesis])
```
import tea
tea.data('03crime.csv')

variables = [
    {'name': 'So',
     'data type': 'nominal',
     'categories': ['0', '1']},
    {'name': 'Prob',
     'data type': 'ratio',
     'range': [0,1]}
]
tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}
tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': [['So', 'Prob'],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)

hypothesis = 'So:1 > 0'
tea.hypothesis(['So', 'Prob', hypothesis])

✓ completeness
✓ syntax
✓ well-formed hypotheses
Statistical test selection as constraint satisfaction

```python
import tea
tea.data('03crime.csv')
variables = {
    'name': 'So',
    'data type': 'nominal',
    'categories': ['0', '1']
}
variables = {
    'name': 'Prob',
    'data type': 'ratio',
    'range': [0,1]
}
tea.define_variables(variables)
study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}
tea.define_study_design(study_design)
assumptions = {
    'groups normally distributed': [(['So', 'Prob']),
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
hypothesis = 'So: 1 > 0'
tea.hypothesis([['So', 'Prob'], hypothesis])
```

- completeness
- syntax
- well-formed hypotheses
Statistical test selection as constraint satisfaction

```python
import tea
tea.data('05crime.csv')
variables = [
    {'name': 'So',
     'data type': 'nominal',
     'categories': ['0', '1']},
    {'name': 'Prob',
     'data type': 'ratio',
     'range': [0, 1]}
]
tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}
tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': [['So', 'Prob']],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)

hypothesis = 'So:1 > 0'
tea.hypothesis(['So', 'Prob', hypothesis])
```

**logical constraints**

\[
\text{continuous}(x) \land \neg \text{categorical}(x)
\]

\[
\text{normal}(x) \rightarrow \neg \text{categorical}(x)
\]

\[
\downarrow
\]

**MaxSat**

**Z3**

- completeness
- syntax
- well-formed hypotheses
Statistical test selection as constraint satisfaction

import tea
tea.data('00crime.csv')
variables = [
    {'name': 'So',
     'data type': 'nominal',
     'categories': ['0', '1']},
    {'name': 'Prob',
     'data type': 'ratio',
     'range': [0, 1]}]
tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob'}
tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed':[['So', 'Prob']],
    'Type I (false positive) error rate': 0.05}
tea.assume(assumptions)

hypothesis = 'So:1 > 0'
tea.hypothesis(['So', 'Prob'], hypothesis)

✓ completeness
✓ syntax
✓ well-formed hypotheses

logical constraints

$$\text{continuous}(x) \land \neg \text{categorical}(x)$$

$$\text{normal}(x) \rightarrow \neg \text{categorical}(x)$$

MaxSat

$$\mathbf{Z3}$$

$$\{\text{valid statistical tests}\}$$
Statistical test selection as constraint satisfaction

import tea
tea.data('05crime.csv')
variables = [
    {'name': 'So',
     'data type': 'nominal',
     'categories': ['0', '1']},
    {'name': 'Prob',
     'data type': 'ratio',
     'range': [0,1]},
]
tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob',
}
tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': ['So', 'Prob']
}
tea.assume(assumptions)

hypothesis = 'So:1 > 0'
tea.hypothesis(['So', 'Prob'], hypothesis)

\[\text{completeness} \quad \checkmark\]
\[\text{syntax} \quad \checkmark\]
\[\text{well-formed hypotheses} \quad \checkmark\]
Statistical test selection as constraint satisfaction

```
import tea

variables = {
    'name': 'So',
    'data type': 'nominal',
    'categorical': ['0', '1']
}

variables = {
    'name': 'Prob',
    'data type': 'ratio',
    'range': [0, 1]
}

tea.define_variables(variables)

study_design = {
    'study type': 'observational study',
    'contributor variables': 'So',
    'outcome variables': 'Prob'
}

tea.define_study_design(study_design)

assumptions = {
    'groups normally distributed': [['So', 'Prob']]
}

healthy = tea.assume(assumptions)

hypothesis = So1 > 0

tea.hypothesise(healthy, hypothesis)
```

- completeness
- syntax
- well-formed hypotheses

logical constraints

\[ \text{continuous}(x) \land \neg \text{categorical}(x) \]

\[ \text{normal}(x) \rightarrow \neg \text{categorical}(x) \]

\[ \ldots \]

\[ \text{MaxSat} \]

\[ \text{Z3} \]

\{valid\ statistical\ tests\}

How do we logically represent statistical knowledge?
Statistical test applies iff all preconditions apply

Student’s t-test ↔

bivariate
one_x_variable
one_y_variable
independent_obs
categorical
two_categories
continuous
equal_variance
groups_normal
Statistical test applies iff all preconditions apply

Student’s t-test

\[\leftrightarrow\]

test properties
bivariate
one_x_variable
one_y_variable
independent_obs

variable properties
categorical
two_categories
continuous
equal_variance
groups_normal
Statistical test applies iff all preconditions apply

Student’s t-test

\[ \text{bivariate}(xy) \]
\[ \text{one}_x\_variable(xy) \]
\[ \text{one}_y\_variable(xy) \]
\[ \text{independent}_obs(xy) \]

\[ \text{categorical}(x) \]
\[ \text{two}_\text{categories}(x) \]
\[ \text{continuous}(y) \]
\[ \text{equal}_\text{variance}(xy) \]
\[ \text{groups}_\text{normal}(xy) \]
Statistical test applies iff all preconditions apply

test properties
\[ \text{bivariate}(xy) \land \text{one}_x\text{variable}(xy) \land \text{one}_y\text{variable}(xy) \land \text{independent}_\text{obs}(xy) \land \]

variable properties
\[ \text{categorical}(x) \land \text{two}_\text{categories}(x) \land \text{continuous}(y) \land \text{equal}_\text{variance}(xy) \land \text{groups}_\text{normal}(xy) \land \]

Student's t-test

- Pearson's r
- Pointbiserial
- Kendall's T
- Spearman's p
- Student's t-test
- Paired t-test
- Mann-Whitney U
- Wilcoxon signed rank
- Welch's t-test
- Repeated measures
  - one-way ANOVA
  - Factorial ANOVA
  - Two-way ANOVA
  - Kruskal Wallis
  - Friedman
  - Chi Square
  - Fisher's Exact
  - Bootstrapping
Statistical test selection as constraint satisfaction

\[
\text{logical constraints} \\
\text{continuous}(x) \land \neg \text{categorical}(x) \\
\text{normal}(x) \rightarrow \neg \text{categorical}(x)
\]

\[\cdots\]

MaxSat

\[
\text{Z3}
\]

\{\text{valid statistical tests}\}

✓ completeness
✓ syntax
✓ well-formed hypotheses

How do we formulate a MaxSat problem?
Satisfiability of logical formulas
Z3

Satisfiability of logical formulas

boolean, real number, integer, uninterpreted functions
Test clauses

Students_t_test \land bivariate(xy) \land one_x_variable(xy) \land one_y_variable(xy) \land independent_obs(xy) \land categorical(x) \land two_categories(x) \land continuous(y) \land equal_variance(xy) \land groups_normal(xy)
Other data/statistical analysis constraints

\[(\text{continuous}(x) \lor \text{categorical}(x)) \land \neg(\text{continuous}(x) \land \text{categorical}(x))\]
\[\text{normal}(x) \rightarrow \neg\text{categorical}(x)\]
\[\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x)\]

---

Test clauses

\[
\text{Students}_t_{\text{test}} \land \\
\text{bivariate}(xy) \land \\
\text{one}_x_{\text{variable}}(xy) \land \\
\text{one}_y_{\text{variable}}(xy) \land \\
\text{independent}_o_{\text{bs}}(xy) \land \\
\text{categorical}(x) \land \\
\text{two}_c_{\text{ategories}}(x) \land \\
\text{continuous}(y) \land \\
\text{equal}_v_{\text{ariance}}(xy) \land \\
\text{groups}_n_{\text{ormal}}(xy)
\]
Other data/statistical analysis constraints

\[(\text{continuous}(x) \lor \text{categorical}(x)) \land \neg(\text{continuous}(x) \land \text{categorical}(x))\]
\[\text{normal}(x) \rightarrow \neg\text{categorical}(x)\]
\[\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x)\]

Test clauses

\[\text{Students\_t\_test} \land \text{bivariate}(xy) \land \text{one\_x\_variable}(xy) \land \text{one\_y\_variable}(xy) \land \text{independent\_obs}(xy) \land \text{categorical}(x) \land \text{two\_categories}(x) \land \text{continuous}(y) \land \text{equal\_variance}(xy) \land \text{groups\_normal}(xy)\]

User assumptions

```python
assumptions = {
    'groups normally distributed': ['eq', 'Prob'],
    'Type I (False Positive) Error Rate': 0.05
}
tea.assume(assumptions)
```
Other data/statistical analysis constraints

\[(\text{continuous}(x) \lor \text{categorical}(x)) \land \lnot (\text{continuous}(x) \land \text{categorical}(x))\]

\[\text{normal}(x) \rightarrow \lnot \text{categorical}(x)\]

\[\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x)\]

---

Test clauses

\[\text{Students_t_test} \land\]
\[\text{bivariate}(xy) \land\]
\[\text{one_x_variable}(xy) \land\]
\[\text{one_y_variable}(xy) \land\]
\[\text{independent_obs}(xy) \land\]
\[\text{categorical}(x) \land\]
\[\text{two_categories}(x) \land\]
\[\text{continuous}(y) \land\]
\[\text{equal_variance}(xy) \land\]
\[\text{groups_normal}(xy)\]

---

User assumptions

```python
assumptions = {
    'groups normally distributed': ['So', 'P<0.05'],
    'Type I (false positive) Error Rate': 0.05
}

# assume(assumptions).
```

---

Z3

UNSAT

Test is invalid.
Remove test.
**Other data/statistical analysis constraints**

\[
\text{continuous}(x) \lor \text{categorical}(x) \land \neg (\text{continuous}(x) \land \text{categorical}(x)) \\
\text{normal}(x) \rightarrow \neg \text{categorical}(x) \\
\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x) \\
\ldots
\]

**User assumptions**

\[
\text{assumptions} = \{
\begin{array}{l}
\text{groups normally distributed': ["80, 0.05"],} \\
\text{Type I (False Positive) Error Rate': 0.05}
\end{array}
\}
\]

**Check test assumptions hold**

For each property:
- If property holds:
  - Add clause (property == TRUE)
- Else:
  - Add clause (property == FALSE)
  - Remove last test added

**Test clauses**

- Students_t_test ∧
- bivariate(xy) ∧
- one_x_variable(xy) ∧
- one_y_variable(xy) ∧
- independent_obs(xy) ∧
- categorical(x) ∧
- two_categories(x) ∧
- continuous(y) ∧
- equal_variance(xy) ∧
- groups_normal(xy)

**Z3**

**UNSAT**

**Test is invalid. Remove test.**
Other data/statistical analysis constraints

\[(\text{continuous}(x) \lor \text{categorical}(x)) \land \neg(\text{continuous}(x) \land \text{categorical}(x))\]

\[\text{normal}(x) \rightarrow \neg \text{categorical}(x)\]

\[\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x)\]

User assumptions

- 'groups normally distributed': [['Sex', 'Prob'], ['Type I (False Positive) Error Rate', 0.05]]
- tea, assume(assumptions)

Test clauses

- \(\text{Students}_t\_\text{test} \land \text{bivariate}(xy) \land \text{one}_x\_\text{variable}(xy) \land \text{one}_y\_\text{variable}(xy) \land \text{independent}_\text{obs}(xy) \land \text{categorical}(x) \land \text{two}\_\text{categories}(x) \land \text{continuous}(y) \land \text{equal}\_\text{variance}(xy) \land \text{groups}\_\text{normal}(xy)\)

Check test assumptions hold

For each property:
- If property holds:
  - Add clause (property == TRUE)
- Else:
  - Add clause (property == FALSE)
  - Remove last test added

All test assumptions are True

Add test to {valid tests}
Other data/statistical analysis constraints

\((\text{continuous}(x) \lor \text{categorical}(x)) \land \neg(\text{continuous}(x) \land \text{categorical}(x))\)
\n\(\text{normal}(x) \rightarrow \neg\text{categorical}(x)\)
\n\(\text{continuous}(x) \lor \text{ordinal}(x) \rightarrow \text{continuous}(x)\)

...

User assumptions

assumptions = {
  "groups normally distributed": ["Se", "Pco2"],
  "Type I (False Positive) Error Rate": 0.05
}

t3a.assume(assumptions)

Test clauses

\(\text{Students_t_test} \land \)
\(\text{bivariate}(xy) \land \)
\(\text{one_x_variable}(xy) \land \)
\(\text{one_y_variable}(xy) \land \)
\(\text{independent_obs}(xy) \land \)
\(\text{categorical}(x) \land \)
\(\text{two_categories}(x) \land \)
\(\text{continuous}(y) \land \)
\(\text{equal_variance}(xy) \land \)
\(\text{groups_normal}(xy)\)

...

Check test assumptions hold

For each property:
  If property holds:
    Add clause (property == TRUE)
  Else:
    Add clause (property == FALSE)

Remove last test added

All test assumptions are True

Add test to \{valid tests\}

If {} bootstrap!

Test is invalid.
Remove test.
Tea Output

Test: students_t

***Test assumptions:
Exactly two variables involved in analysis: So Prob
Exactly one explanatory variable: So
Exactly one explained variable: Prob
Independent (not paired) observations: So
Variable is categorical: So
Variable has two categories: So
Continuous (not categorical) data: Prob
Equal variance: So Prob
Groups are normally distributed: So Prob: NormalTest(W=0.8997463583946228
p_value=0.07962072640657425)

***Test results:
name = Student’s T Test
test_statistic = 4.20213
p_value = 0.00012
adjusted_p_value = 0.00006
alpha = 0.05
dof = 45

Effect size:
Cohen’s d = 1.24262
A12 = 0.83669

Null hypothesis = There is no difference in means between So = 0 and So = 1 on Prob.
Interpretation = t(45) = 4.20213 p = 0.00006. Reject the null hypothesis at alpha = 0.05. The mean of Prob for So = 1 (M=0.06371 SD=0.02251) is significantly greater than the mean for So = 0 (M=0.03851 SD=0.01778). The effect size is Cohen’s d = 1.24262 A12 = 0.83669. The effect size is the magnitude of the difference which gives a holistic view of the results [1].
Evaluation

12 tutorials
code snippets + text
Evaluation

12 tutorials
code snippets + text

I. How does Tea compare to experts?
I. How does Tea compare to experts?

Replicate, even improve upon expert choices
Evaluation

12 tutorials
code snippets + text

I. How does Tea compare to experts?
   Replicate, even improve upon expert choices

II. How does Tea compare to novices?
Evaluation

12 tutorials
code snippets + text

I. How does Tea compare to experts?
   Replicate, even improve upon expert choices

II. How does Tea compare to novices?
    Avoid common mistakes and false conclusions
Tea automates statistical test selection and execution.

Tea can aid with experimental design.

Tea programs can act as a format for pre-registration.
Tea automates statistical test selection and execution.
Tea can aid with experimental design.
Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.
Tea automates statistical test selection and execution.
Tea can aid with experimental design.
Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.

Internal validity!
Tea automates statistical test selection and execution. Tea can aid with experimental design. Tea programs can act as a format for pre-registration.

Tea promotes validity and reproducibility in statistical analysis.

Internal validity!

pip install tealang
tea-lang.org
Ongoing work

Field deployment, user testing

Future development
  - linear modeling
scone
(behind the tea cup)
Scone: Smart Sampling for Smarter Statistics

with Laurel Orr, Emery Berger, and Ben Zorn

ongoing summer project
+ Speed
+ Familiarity
- Not representative
generalize???

+ Speed
+ Familiarity
- Not representative
scone

automatically generate and manage representative samples
**tea**

Automated statistical analyses

Internal validity

pip install tealang
tea-lang.org

**scone**

Automated sampling

External validity

Stay tuned!
tea

Automated statistical analyses

Internal validity

pip install tealang
tea-lang.org

scone

Automated sampling

External validity

Stay tuned!

COLLABORATION, USERS, FEEDBACK