Assessing evidence: hypothesis testing

Data Science Team

Stanford University, Department of Statistics
Evaluating observations with a reference distribution

<table>
<thead>
<tr>
<th>Div/Sep</th>
<th>Married</th>
<th>Never Married</th>
<th>Widowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>94</td>
<td>157</td>
<td>111</td>
</tr>
<tr>
<td>M</td>
<td>61</td>
<td>359</td>
<td>122</td>
</tr>
</tbody>
</table>
We are going to step back and see how this exercise fits into a more general picture.

**Scientific method**

- Formulate an hypothesis
- Test it on the basis of data

*What does it mean to test one hypothesis?*

*Can the data prove it true?*
Hume and the problem of induction

- The problem of induction: no matter how many instances of white swans we might have observed, this does not justify the conclusion that all swans are white.
- Natural instinct, rather than reason, explains the human practice of making inductive inferences.
- How can we learn from data then?

David Hume (1711 – 1776)
Deductive method of testing

An hypothesis can only be empirically tested, and only after it has been advanced;

Predictions are deduced from the theory… and compared with the results of practical applications and experiments… they can be falsified or corroborated

The Logic of Scientific Discovery, 1934
What happens when the answer is not black or white?

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

The Design of Experiments, 1935
The scientific method in practice

- Commonly, the hypothesis being tested is referred to as a “null hypothesis,” to connote that it is the “default,” status quo understanding of how things work.

- The complement of the null – i.e. the state of nature that would prevail if the null hypothesis were false – is usually referred to as the alternative (though in some cases we specify a more restrictive set of alternatives)

- Concluding that our test statistic is very unlikely to be observed under the null is often called rejecting the null

- When we reject the null we make a discovery: we find out that the current understanding of the world is not enough
Evidence against the null

Even if the null is true (marital status and sex are independent), it is entirely possible for us to observe $d(O, E) = 70.04$.

*It is open to the experimenter to be more or less exacting in respect to the smallness of the probability he would require before he would be willing to admit that his observations have disproved the null.*

Fisher, “The design of experiments, 1935”

- The evidence against independence can be summarized by seeing how many of our reference tables were as extreme or more than our observed one.
**The p-value**

P-value is the **probability** of observing a statistic as extreme or more extreme as the one we actually observed **given that the null was true**

\[ \text{mean(\( \text{how\_far\_sample} > \text{how\_far}(\text{cross, expected\_counts}) \))} \]

[1] 0

- Since we took \( M = 2000 \) samples, we estimate that the p-value is less than 0.0005 = 1/2000 (ignoring sampling variability – will discuss this in a moment)
- It seems that our observed table \text{cross} is quite extreme.
Monte Carlo approximation

- Our *p-value* is based only on 2000 tables we generated (and depends on these particular 2000).

- It is an approximation of the true proportion of tables (with those marginal frequencies) that are more extreme than our observed table using the scores `how_far`.

- Such an approximation is called Monte Carlo approximation.
Evidence against independence

- Statisticians have developed many measures like $d(O, E)$.
- Here is one that only looks at the difference in one of the cells:

$$d(O, E) = O_{11} - E_{11}$$
Evidence against independence

In R, one of these common measures can be easily computed

The following function generates many tables *almost* the same way we did.

```r
chisq.test(cross, simulate.p.value=TRUE)
```

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  cross
X-squared = 111.33, df = NA, p-value = 0.0004998

It uses Pearson’s $X^2$ as a measure of discrepancy

$$\chi^2(O, E) = \sum_i \sum_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}}$$
Evidence against independence

This function does not simulate at all, but computes a similar p-value (lots of the work of early statisticians has been in figuring out what the p-values would be without having the power of computing.)

\[
\text{chisq.test}(\text{cross}, \text{simulate.p.value}=\text{FALSE})
\]

Pearson's Chi-squared test

data: cross
X-squared = 111.33, df = 3, p-value < 2.2e-16
What is a *p*-value?

- Suppose marital_status and sex really were independent and our marginal frequencies were good estimates.
- How would our p-values behave?

```r
pvalue_sample = numeric(10000)
for (i in 1:10000) {
    pvalue_sample[i] = mean(how_far_sample > how_far(sample_table, expected_counts))
}
```
What is a $p$-value?

We call this density *uniform* on the interval $[0,1]$.
What is a *p*-value?

- A random variable has a **uniform distribution** on the unit interval if

\[ P[X \in [a, b]] = |b - a| \]

for any \(0 \leq a \leq b \leq 1\).

- We say that *p*-values have a **uniform distribution under the null**

- This is an example of the general result referred to as the probability integral transform (you don’t need to know this, just putting it here in case of interest)
What is a \textit{p-value}?

- When marital\_status and sex really are independent, then the sampling distribution of our p-value is known.
- Here is a visualization of its cumulative distribution function (or \textbf{CDF}), the function given by $F(x) = P[X \leq x]$

\begin{verbatim}
ECDF = ecdf(pvalue_sample)
grid = seq(0, 1, length=101)
qplot(grid, ECDF(grid))
\end{verbatim}
What can we do with a *p*-value?

- Suppose we decide that a *p*-value less than 0.05 is strong evidence against independence.
- What do we know?
- Well, 5% of the time, even when marital_status and sex are independent, we will say that there is strong evidence they are not independent.
- This type of error is called a **Type I error**.
- In words: a **Type I error** is rejecting the null when the null is true.
- Thresholding the *p*-value at 5% controls the **Type I error rate** at 5%.
- In other words, suppose we test *many* hypotheses by thresholding the *p*-value at 5%, and in **every case**, the null was actually true.
- Then in about 5 percent of those cases, we will **incorrectly reject the null**.
Lady Tasting Tea

- This is a famous example of R. A. Fisher illustrating the idea of a **hypothesis test**.
- A lady declares that by tasting a cup of tea made with milk she can determine whether or not the tea was added before the milk.

**Experiment**

- She is asked to taste eight cups, four of each type.
- Cups are presented in a random order.
- She correctly identifies the type of 6 out of 8 cups.
- Do you think she can really tell the difference?
A representation of the experiment

- Let her choices either T or M. She must make 8 choices, 4 of which will be T and 4 of which will be M.
- The true types of the tea are also T or M.
- Here is an example of how we might describe the experiment

```r
lady = c('M', 'T', 'M', 'T', 'T', 'T', 'M', 'M')
truth = c('M', 'M', 'M', 'T', 'T', 'T', 'T', 'M')
N_correct = sum(lady == truth)
N_correct
```

[1] 6
A mental model

- The variables lady and truth are just two outcomes for the experiment in which the lady correctly identifies exactly 6 of the cups of tea.
- Is this unusual? How many other outcomes for her choices are there?
- How many other outcomes for her choices are there where she correctly identifies 6 or more?
A permutation of a list is a reordering of the list.

The set of all possible outcomes for lady is the set of all reorderings of the list lady.

```r
library(iterpc)
original_list = c(1:3)
original_list

[1] 1 2 3

I = iterpc(3, labels=original_list, ordered=TRUE)
getall(I)

[,1] [,2] [,3]
[1,] 1 2 3
[2,] 1 3 2
[3,] 2 1 3
[4,] 2 3 1
```
Let us count the ways...

```
N_outcomes = 0
exactly_six = 0
six_or_more = 0

library(foreach)
I = iterpc(8, labels=lady, ordered=TRUE)
permutations_lady = iter_wrapper(I)
results = (foreach(lady_permuted=permutations_lady, .combine=rbind)
    %do% { c(1, sum(lady_permuted == truth) == 6, sum(lady_permuted == truth) >= 6)})

totals = apply(results, 2, sum)

N_outcomes = totals[1]
extactly_six = totals[2]
six_or_more = totals[3]
data.frame(N_outcomes, exactly_six, six_or_more)
```
What if there was no difference between the cups of tea?

- If the lady could really tell no difference between the two, then any one of 40320 possible outcomes for her choices should reasonably be considered equally likely.
- There were 9792 such outcomes in which she would have correctly identified 6 or more.
- As the choices are equally likely, the chances that she would correctly identify 6 or more are

\[
\frac{9792}{40320} \approx 24\%.
\]

- Not that rare an occurrence. The chances she would correctly identify exactly 6 is about 23%!

\[
c(six_{or\_more} / N_{outcomes}, exactly\_six / N_{outcomes})
\]

[1] 0.2428571  0.2285714
More data?

- The percentage of 75% seemed pretty impressive until we computed the chances we would see such an impressive rate.
- What if she had correctly identified 60 out of 80 cups of tea?
- In this case, the number of reorderings is huge, about $10^{118}$!
- Our mental model is still valid: if the lady actually has no ability to tell the difference between the two types of teas, then any of these orderings is equally likely, whatever truth is.
- We can get a sense of how impressive this is by choosing several reorderings at random and computing the number of matches.
More data?

\begin{verbatim}
lady = c(rep('T', 40), rep('M', 40))
truth = c(rep('T', 40), rep('M', 40))
lady[1:10]
\end{verbatim}

[1] "T" "T" "T" "T" "T" "T" "T" "T" "T"
A Helper Function...

```r
more_than_60 = function(N_permutations) {

  exactly_sixty = 0; sixty_or_more = 0
  matches = numeric(N_permutations)

  for (i in 1:N_permutations) {
    lady_reordering = sample(lady, length(lady), replace=FALSE)
    N_match = sum(lady_reordering == truth)
    exactly_sixty = exactly_sixty + (N_match == 60)
    sixty_or_more = sixty_or_more + (N_match >= 60)
    matches[i] = N_match
  }

  return(list(matches=matches,
               exactly_sixty=exactly_sixty,
               sixty_or_more=sixty_or_more))
}
```
Results for 50,000 Permutations

```
results = more_than_60(50000)
c(results$exactly_sixty, results$sixty_or_more)

[1] 0 0
```
Maybe she does know what she’s doing... 

- We would really have been impressed by a 75% rate if she had tasted 80 cups of tea!
- We sometimes (though not always) saw 1 reorderings out of 50000 with a success rate of 75%.
- The probability 1/50000 might not be a great estimate of how likely she would be to achieve a success rate of 75% or higher if she really could not distinguish between the cups of tea.
- BUT, it certainly gives strong evidence that our mental model may be wrong...
Sampling distribution

Instead of just computing the chances above, we could record the number of matches for each reordering of lady and produce a histogram.

```r
(ggplot(data.frame(results), aes(x=matches)) + geom_histogram(aes(y=..density..), bins=20))
```
The null hypothesis

- Our mental model above represents how we might model the experiment under the assumption that the lady really cannot distinguish between different the different types.
- This is the null hypothesis.
- The permutations above represented different outcomes for our experiment.
- Under our null hypothesis, each of these outcomes was equally likely. This allowed us to compute the chances that the lady would have a success rate of 75% or higher if she really could not tell the difference (i.e. assuming the null hypothesis was correct.)
- For 6/8, the chances were about 25%, we were not very impressed. If she had achieved 60/80, we really would have been impressed.
- Observing something that is rare under the null hypothesis is evidence against the null hypothesis.
- How rare? This is measured by the p-value.