

# Topic8: Hypothesis Testing



# Outline

Topic8: Hypothesis Testing

Example: Famous Court Cases

The Western legal system

Framework for Hypothesis Testing


Defining Terms

## Example: Famous Court Cases

Do you remember this court case (Sept 2014, Dec 2015, July 2016)? [▶ Pistorius](#)



## Example: Famous Court Cases

Or this one? (1995, 1997, 2007, 2008) 



# The Western legal system

In a legal court case, the defendant is either innocent (not guilty) or guilty. But unless the defendant pleads guilty, we never know what the truth is. In fact, even if the defendant pleads guilty, there may be more going on!

Our modern Western legal system is based on the principle of being 'innocent until proven guilty' or 'proof beyond a reasonable doubt'. Hence we assume  $H_0$ : defendant is innocent. Unless there is strong evidence for  $H_1$ : defendant is guilty.

**What the pros and cons of this type of legal system?**

In any legal case, there are 4 possible outcomes:

	Truth: $H_0$ is true	Truth: $H_0$ is false
Decision: Retain $H_0$	Acquit innocent person	Acquit guilty person. (Type II error)
Decision: Reject $H_0$	Convict innocent person (Type I error)	Convict guilty person.

We can generalise this terminology, where  $\alpha$  and  $\beta$  are probabilities:

	Truth: $H_0$ is true	Truth: $H_0$ is false
Decision: Retain $H_0$	Specificity = $1 - \alpha$	False negative = $\beta$ (Type II error)
Decision: Reject $H_0$	False positive = $\alpha$ (Type I error)	Power or sensitivity = $1 - \beta$

We set the Type I error to be small, typically  $\alpha = 0.05$  (which is called the 'significance level'). Ideally we want the Power to be large. [▶ Article](#)

## Other contexts:

Context	Type I error	Type II error
$H_0$ : Patient is healthy $H_1$ : Patient has Diabetes	Wrong diagnosis ▶ Breast Cancer	Undiagnosed condition
$H_0$ : iPhone works $H_1$ : iPhone is faulty	Wastage for Apple	Ruins Apple reputation

# Framework for Hypothesis Testing

For each Hypothesis Test, we use the following framework:

**H** Set up the two hypotheses:  $H_0$  and  $H_1$ .

**A** State the assumption(s) of the test, and justify whether they are valid from the sample.

**T**

- ▶ State the Test Statistic, and it's distribution assuming  $H_0$  is true.
- ▶ State what values argue against  $H_0$ .
- ▶ Find the observed value of the Test Statistic.

**P** Calculate the  $P$ -value, which represents the probability of observing this sample (or more extreme) assuming  $H_0$  is true.

**C** Weigh up the conclusion, based on the size of the  $P$ -value.



# Defining Terms

## H

- ▶ The Null hypothesis  $H_0$  is the default hypothesis: what we currently believe to be true.
- ▶ The Alternate hypothesis  $H_1$  is a new claim about the population.
- ▶ The hypotheses are commonly articulated in terms of the unknown population parameter. Eg  $H_0 : \mu = 5$ .
- ▶ If so, then the alternate hypothesis can take 2 forms:  
1 sided ( $H_1 : \mu > 5$  or  $H_1 : \mu < 5$ ) or 2-sided ( $H_1 : \mu \neq 5$ ).
- ▶ How to decide between a 1 or 2 sided test? The decision must not be influenced by the data ('data snooping') – we must specify the hypotheses before we do the actual test. Hence, we always use a 2 sided test, unless we have prior evidence (eg a previous report) which suggests a 1 sided test.

## A

The assumptions are necessary for the test to be valid. We check whether they appear valid from the sample.

## T

- ▶ The Test Statistic  $\tau$  is a random variable, with a distribution which depends on the unknown parameter.
- ▶ The observed value of the Test Statistic  $\tau_{obs}$  (or  $\tau_0$ ) is calculated from the sample.
- ▶ Look at the distribution of  $\tau$  to determine what values will argue against  $H_0$  for  $H_1$ .
- ▶ Hypothesis testing involves some theory about the random variable  $\tau$  where every possible value  $\{\tau_0\}$  counts as some evidence about  $H_0$ . The Hypothesis Test weighs up the evidence against  $H_0$  based on the observed value.

## P

- ▶ The  $P$ -value is the probability of observing  $\tau_0$  or something more extreme (or unusual) under  $H_0$ .
- ▶ A small  $P$ -value either means that  $H_0$  is true but the sample is highly rare, or that  $H_0$  is false.
- ▶ The smaller the  $P$ -value, the stronger the evidence against  $H_0$  for  $H_1$ .
- ▶ A large  $P$ -value means that the sample is consistent with  $H_0$ .
- ▶ The critical region is the set of  $\tau$  such that  $H_0$  would be rejected.

$P$ -value	Correct language	Unhelpful Language
Small	Evidence against $H_0$ Reject $H_0$ for $H_1$	$H_0$ is false or $H_1$ is true.
Large	Data are consistent with $H_0$ Retain $H_0$	$H_0$ is true or $H_1$ is false.

## C

'The (null hypothesis) is ... never proved or established, but is possibly disproved, in the context of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.' (Ronald Fisher, Design of Experiments, 1935, p19).

- ▶ There is no final proof that  $H_0$  is true or false.
- ▶ The conclusion is not 'Accept'  $H_0$  or  $H_1$ , as we have assumed  $H_0$  to be true. That is, we have not proved  $H_0$  true, rather we look for evidence about whether it is false.
- ▶ If the  $P$ -value is small, it suggests there is evidence against  $H_0$ . If the  $P$ -value is not small, then it suggests the data are consistent with  $H_0$
- ▶ By 'small', a common convention is  $\alpha = 0.05$ . That is, for  $P$ -values under 0.05, we suggest there is evidence against  $H_0$ .

more certain of success, Luo says. The spacecraft could launch in 15–20 years, he adds, around the same time as the Taiji group says that it could launch. Luo thinks that a simpler project is more realistic now, but says that TianQin could lay the groundwork for a Taiji-like project in the future.

Wu Ji, director-general of the Chinese Academy of Sciences' National Space Science Center, says that the TianQin and Taiji teams should merge. "If China decides to have a space gravitational mission, there should be an integrated one, with a new name probably. There is no way to support two missions at the same time."

Both Wu Yue-Liang and Luo are confident that their proposals will move forward to the concrete design phase in the next five years. Taiji currently receives money from the Chinese Academy of Sciences and TianQin from the city of Zuhai — but both need much more cash. The LIGO discovery could increase their chances of success. "The government will know more the importance of fundamental research" in gravitational waves, says Wu Ji. "China should catch up in this area," he adds.

On 5 March, the Chinese central government released a draft list of 100 strategic projects that will be emphasized in the country's next five-year plan, which includes "a new generation of heavy launch vehicles, satellites, space platforms and new payload" and a "deep-space station".

Chinese researchers could also end up collaborating with Europe. As well as its main project, the Taiji group has outlined the possibility of a direct collaboration with eLISA: it would either contribute 1.5 billion yuan directly or develop its own scaled-down, 8-billion-yuan version of eLISA that would coordinate closely with the European effort, sharing data. Heinzl recommends that a united Chinese group work on one of these less ambitious options.

The direct contribution from China in particular could be a boon for eLISA. Originally, NASA collaborated with ESA on a planned space-based gravitational-wave observatory, named LISA. But the United States pulled out of LISA five years ago and ESA had to pare down the mission, resulting in the eLISA proposal. China's entry into the project could fill that hole, says Rainer Weiss,

#### REPRODUCIBILITY

## Statisticians issue warning on *P* values

*Statement aims to halt missteps in the quest for certainty.*

BY MONYA BAKER

**M**isuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions purely on the basis of *P* values (R. L. Wasserstein and N. A. Lazar *Am. Stat.* <http://doi.org/bc4d>; 2016). Researchers should describe not only the data analyses that produce statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

Véronique Kiermer, executive editor of the Public Library of Science journals, says that the ASA's statement lends weight and visibility to longstanding concerns over undue reliance on the *P* value. "It is also very important in that it shows statisticians, as a profession, engaging with the problems in the literature outside of their field," she adds.

*P* values are commonly used to test (and dismiss) a 'null hypothesis', which generally states that there is no difference between two groups, or that there is no correlation between a pair of characteristics. The smaller the *P* value, the less

cannot indicate the importance of a finding: for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

#### FRUSTRATION ABOUND

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gained an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). More controversially, in 2015, a set of documentary filmmakers published conclusions from a purposely shoddy clinical trial — supported by a robust *P* value — to show that eating chocolate helps people to lose weight. (The article has since been retracted.)

But Simine Vazire, a psychologist at the University of California, Davis, and editor of the journal *Social Psychological and Personality Science*, thinks that the ASA statement could help to convince authors to disclose all of the statistical analyses that they run. "To the extent that people might be sceptical, it helps to have statisticians saying, 'No, you can't interpret *P* values without this information,'" she says.

More drastic steps, such as a ban on publishing *P* values in articles instituted by at least one journal, could be counter-productive, says Andrew Vickers, a biostatistician at Memorial Sloan Kettering Cancer Center in New York City. He compares attempts to bar the use of *P* values to addressing the risk of automobile