Week 12 Lecture: Hypothesis Testing

Yilin

DigiPen

HYPOTHESIS TESTING

Example 1: Prevalence of an allele in diseased people

Suppose it is well-known that 17% of the population as a particular allele **A**. One hundred individuals from the population, all having a certain disease, *D*, are randomly tested, and our test shows that 25 of them have allele **A**.

Question

Because of this difference, is there a connection between the disease *D* and the allele **A**?

Definition: Null Hypothesis

Definition

Suppose we run an experiment on a sample of the population to determine if two properties, X and Y, are related. The **null hypothesis**, H_0, is the hypothesis that there is no statistical relation.

In the above example, the null hypothesis is that the presence of allele A is **not** related to the presence of disease D.

- We found that 25 of the 100 people with disease D had allele **A**.
- We know that 17% of the total population has allele **A**.
- The **null hypothesis** is that finding 25 people out of 100 people with allele **A** is "close enough" to 17 as to indicate **no significant** relation between D and **A**.

What is the null hypothesis?

We can discuss this in more detail. Let q be the **true probability** of an individual with disease D having allele **A**, so $q = P(\mathbf{A}|D)$. We are given that $P(A) = 0.17$. The **null hypothesis** is then

$$H_0 : \ q = 0.17.$$

We need to run what are called **statistical tests** to investing how "likely" a result of 25 would be assuming $q = 0.17$.

Definition: Alternative Hypothesis

Definition

Suppose we run an experiment on a sample of the population to determine if two properties, X and Y, are related. The **alternative hypothesis**, H_a, is a hypothesis that makes the **null hypothesis** false, i.e. that there is **some** statistical relation between X and Y.

If the **null hypothesis** H_0 is $q = 0.17$, then one good **alternative hypothesis** is $q \neq 0.17$. Another, more precise, **alternative** hypothesis H_a is that $q > 0.17$, as the data suggests it is not less than 0.17.

Note

All we need from an alternative hypothesis is that it is **mutually exclusive** from the null hypothesis.

Method

We will use **statistical tests** on data to either **accept** or **reject** the null hypothesis.

- **Accepting** the null hypothesis just means we cannot “rule it out”, this is not an assertion of whether it is true or false, it is more of a way to say “our data supports this”.
- **Rejecting** the null hypothesis means that we can rule it out “ with certainty”.
- Rejecting H_0, the claim that X and Y are unrelated is equivalent to supporting the assertation that X and Y are related, i.e. a good alternative hypothesis H_a.

Type I and Type II Errors

There are four possible outcomes from such a statistical test:

	H_0 accepted	H_0 rejected
H_0 true	Test is correct	Type I Error
H_0 false	Type II error	Test is correct

Definition

We define the **significance level**, α, and the **power**, β, of a statistical test.

- A **Type I error** rejects a true null hypothesis. We define the **significance level**, α, to be the probability of a Type I error
- A **Type II error** accepts a false null hypothesis. We define the **power**, β, to be the probability that that the statistical test **rejects** a false null hypothesis, i.e. $1 - \beta$ is the probabilty of a Type II error.

p-VALUES

Definition: p-value

Definition

The **p-value** is the probability of observing a result at least as extreme as the measured result if the null hypothesis is true.

Remember,"at least as extreme", roughly means "greater than or equal to" by some metric.

- We know 17% of the whole population has allele **A**
- From a sample of 100 people with disease D, 25 people have allele **A**.
- H_0 (the null hypothesis) is that the true probability of an individual with disease D having allele **A** is $q = 0.17$.

One-tailed test

The **one-tailed test** for the p-value is as follows:

- Set H_a, the **alternative hypothesis**, to be that $q > 0.17$,
- Compute the p-value, the probability of a result **at least as extreme** than our data, if we assume H_0 is true.
- The p-value is the probability that out of 100 people with disease D, 25 **or more** have allele **A**.
- If N is the number of the people from our sample with allele **A**, then

$$\begin{aligned} p\text{-value} = P(N \geq 25) &= \sum_{k=25}^{100} \binom{100}{k} (0.17)^k (0.83)^{100-k} \\ &\approx 0.027 = 2.7\%. \end{aligned}$$

This is found using an online calculator.

Summary

- We showed that if H_0 is true, that is $q = 0.17$, then the probability of 25 or more out of 100 with disease D having allele **A** is approximately 2.7%.
- In other words, if H_0 is true then we would see our result or worse 2.7% of the time.
- If we reject H_0 (belive H_0 is false) in favor of H_a, then there is a 2.7% chance that we are wrong, since our result or worse can happen 2.7% of the time if H_0 is true.
- In other words, we will make a Type I error with probability 0.027.
- Conclusion: The p-value is the probability of **rejecting** a **true** null hypothesis given that we observe data that is at or more extreme than the current data, i.e. a Type I error.

General rule of thumb for p-values

p-value	significance of data
$p > 0.1$	not significant
$0.1 > p > 0.05$	trends towards significant
$0.05 > p > 0.01$	significant
$0.01 > p > 0.001$	highly significant
$0.001 > p$	extremely significant

- In our example, the p-value was 0.027, so we would say the data that 25 people with disease D have allele **A**, is "significant".
- We would likely reject the null hypothesis because the probability of making a Type I error is 2.7%.
- Note that this is not at all fool-proof.

Example 2: 27 Instead of 25

Suppose we had different data, that 27 people instead of 25 people out of 100 tested had allele **A**? Our one-tailed test yields

$$\begin{aligned} p = P(\text{Type I error}) &= P(N \geq 27 \text{ if null hypothesis true}) \\ &= \sum_{k=27}^{100} \binom{100}{k} (0.17)^k (0.83)^{100-k} \approx 0.008. \end{aligned}$$

So a result of 27 with allele **A** would be **highly significant** since $0.001 < p < 0.01$.

Example 3: 22 Instead of 25

Suppose instead, that 2 people people out of 100 tested had allele **A**? Then

$$p = P(\text{Type I error}) = P(N \geq 22 \text{ if null hypothesis true})$$
$$= \sum_{k=22}^{100} \binom{100}{k} (0.17)^k (0.83)^{100-k} \approx 0.117$$

So a result of 22 with allele **A** would be **not significant** since $p > 0.1$.

Two Tails

Note that the one-tailed test, computing the probability of a result being "at least as extreme" as the data, 25, makes sense since our alternative hypothesis was $q > 0.17$.

Going back to the original experiment, take the null hypothesis to be $q = 0.17$, but suppose the alternative hypothesis, H_a, is $q \neq 0.17$. Now the term "at least as extreme as 17" has a different meaning. Now we want to find the probability that a result is at least *as far* from 17 as 25 is, i.e. 8 away from 17.

- One tail: $N \geq 25$
- Two tails: $N \geq 25$ or $N \leq 9$.

Two-tailed test

The p-value using a two-tailed test is then:

$$
\begin{aligned}
p &= P(\text{Result at least as extreme as data}) \\
&= P(N \leq 9) + P(N \geq 25) \\
&= \sum_{k=0}^{9} \binom{100}{k} (0.17)^k (0.83)^{100-k} + \sum_{k=25}^{100} \binom{100}{k} (0.17)^k (0.83)^{100-k} \\
&= 1 - \sum_{k=10}^{24} \binom{100}{k} (0.17)^k (0.83)^{100-k} \\
&\approx 0.0445.
\end{aligned}
$$

Since $0.01 < 0.0445 < 0.05$, the result 25 is **significant** as before.

Two-tailed test on 27 instead of 25

If the data was 27 instead of 25, then

$$\begin{aligned} p &= P(\text{Result at least as extreme as data}) \\ &= P(N \leq 7) + P(N \geq 27) \\ &= \sum_{k=0}^{7} \binom{100}{k} (0.17)^k (0.83)^{100-k} + \sum_{k=27}^{100} \binom{100}{k} (0.17)^k (0.83)^{100-k} \\ &\approx 0.0437. \end{aligned}$$

Since $0.01 < 0.0437 < 0.05$, the result 27 is **significant**, which is different from before, since the one-tailed test told us 27 was **highly significant**.

When to use the one-tailed test?

Question

We see that the one-tailed and two-tailed tests have the potential to yield different p-values. So which should we choose?

We use the one-tailed test provided

- There is evidence that the results will fall on one side of the null hypothesis, e.g. $25 > 17$ so we would not consider looking at the probability a result is any number less than 17.
- We establish (publicly) that we are using a one-tailed est **before** recording the data.

Both tests on a potentially biased coin

Suppose we a coin 10 times and get 8 heads, 2 tails. We suspect the coin is **not fair**. To test for this, let the null hypothesis be that the coin is fair, that the probability of heads, $q = 0.5$.
Since we suspect the coin is biased, take the alternative hypothesis to be $q > 0.5$. The one-tailed test yields

$$p = P(\text{heads } \geq 8) = \sum_{k=8}^{10} \binom{10}{k} (0.5)^k (0.5)^{10-k} = \frac{56}{1024} \approx 0.055.$$

Since $0.05 < 0.055 < 0.1$, the result of 8 heads "trends toward" but does not meet the threshold for significance, since the probability of not making a type I error is less thna 95%.
The two-tailed test would yield

$$p = P(\text{heads } \geq 8) + P(\text{heads } \leq 2) \approx 0.109.$$

So by the two-tailed test, the result of 8 is not significant.