Queen Mary
University of London

MTH4106 Introduction to Statistics

Notes 6 Spring 2012

Testing Hypotheses about a Proportion

Example Pete's Pizza Palace offers a choice of three toppings. Pete has noticed that rather few customers ask for anchovy topping. He thinks that if fewer than $1/5$ want anchovy topping then he should scrap it and replace it by another topping. How should he proceed?

Let p be the proportion of the pizza-eating population who like anchovy topping. We do not know what p is. Pete does not need to know the exact value of p. He needs to know whether $p \geqslant 1/5$ or not.

So there are two hypotheses:

$$\begin{aligned} p &\geqslant 0.2 \\ p &< 0.2. \end{aligned}$$

Mathematically and logically, the roles of these two hypotheses are symmetric: each is the negation of the other. In practical terms, they have different roles:

- if $p \geqslant 0.2$ then Pete doesn't change anything;
- if $p < 0.2$ then he has to go to the trouble and expense of replacing anchovy topping by a new flavour, with the risk that the new one is even less popular.

If we could definitely say which hypothesis is true, then this practical difference would not matter too much. But unless Pete asks *all* of his customers for their preference, all that he can do is to gather *evidence* in favour of one hypothesis or the other; he can never be quite sure. Because of the expense and risk involved in changing the flavour, Pete will want fairly strong evidence that $p < 0.2$ before he decides to change.

Here

"$p \geqslant 0.2$"	is called the *null hypothesis*	(written H_0)
"$p < 0.2$"	is called the *alternative hypothesis*	(written H_1).

type of investigation	null hypothesis	alternative hypothesis
to see whether a procedure should be changed	the hypothesis that supports the status quo	the hypothesis that supports a change
to find out if an outrageous claim is true	claim is false	claim is true
to explain phenomena	simpler explanation	more complicated explanation

Example To investigate whether girls are as good at mathematics as boys are.

- 200 years ago, this would have been considered an absurd suggestion, so the null hypothesis would have been "boys are better";
- nowadays, we don't think that it is absurd, and it is *simpler* to assume that gender has no effect on mathematical ability, so the null hypothesis would be "gender makes no difference".

Example In a criminal court in England and Wales, the accused must be assumed innocent unless there is overwhelming evidence to the contrary, so the null hypothesis is "the accused is innocent".

Subject	What can we do?	Possible conclusions
Mathematics	Can prove (for example, prove a theorem) and disprove (for example, give a counterexample).	"hypothesis is true" or "hypothesis is false"
Physics	Can disprove (the facts are not consistent with the theory) but not prove.	"hypothesis is false" or "hypothesis is consistent with the known facts"
Statistics	Can neither prove nor disprove.	for example, "the hypothesis is most unlikely to be true, but we cannot be certain that it is false"

As we have seen, there are many types of statistical investigation. Each one produces some data. From the data, we calculate one or more statistics. We are going to see how to use these statistics to test hypotheses.

Example Pete takes a random sample of 20 of his customers and records which topping they ask for. Let

$$X = \text{the number who ask for anchovy topping.}$$

Then $X \sim \text{Bin}(20, p)$.

Whatever the value of p (apart from $p = 0$ and $p = 1$), X *may* take any integer value between 0 and 20, but if p is higher then X is less likely to be small. If we obtain a small value of X, this suggests that p is small.

Pete will use X to test his hypotheses (remember H_0: $p \geqslant 0.2$).

X is called the *test statistic*.

If $p \geqslant 0.2$ then $\mathbb{E}(X) \geqslant 4$, so Pete might decide to reject H_0 if $X \leqslant 3$.

Definition A *test* consists of a statistic X and *rejection region* R. We declare in advance that

- if $X \in R$ then we will reject the null hypothesis in favour of the alternative hypothesis;
- if $X \notin R$ then we will say that there is insufficient evidence to reject the null hypothesis.

Example For the pizza example, here are some possibilities.

rejection region	$\{0\}$	$\{0,1\}$	$\{0,1,2\}$	$\{0,1,2,3\}$
reject H_0 if:	$X = 0$	$X \leqslant 1$	$X \leqslant 2$	$X \leqslant 3$

Unknown real state of affairs	Possible conclusions	
	Do not reject H_0	Reject H_0
H_0 is true (e.g. $p \geqslant 0.2$)	Correct conclusion	Type I error
H_0 is false (e.g. $p < 0.2$)	Type II error	Correct conclusion

Because of the difference in roles between H_0 and H_1, a Type I error is usually considered more serious than a Type II error.

Example	Type I error	Type II error
Pete's pizza	Unnecessarily change to a new topping	Keep an unpopular topping
Court case	Imprison an innocent person	Fail to imprison a guilty person

Define $\gamma(p)$ to be $\mathbb{P}(X \in$ rejection region$)$ if p is the true value.
For p satisfying H_0, put

$$\alpha(p) = \gamma(p) = \mathbb{P}(\text{Type I error}) \quad \text{if } p \text{ is the true value.}$$

For p satisfying H_1, put

$$\beta(p) = 1 - \gamma(p) = \mathbb{P}(\text{Type II error}) \quad \text{if } p \text{ is the true value.}$$

For p satisfying H_1, the quantity $\gamma(p)$ is called the *power* of the test at alternative p.

Example In the pizza example, we obtain the following values of $\gamma(p)$ from the table for the binomial distribution with $n = 20$: see Table 1 of *New Cambridge Statistical Tables* [1].

test	true value of p						
	0.05	0.10	0.15	0.20	0.25	0.30	0.35
$X = 0$	0.3585	0.1216	0.0388	0.0115	0.0032	00008	0.0002
$X \leqslant 1$	0.7358	0.3917	0.1756	0.0692	0.0243	0.0076	0.0021
$X \leqslant 2$	0.9245	0.6769	0.4049	0.2061	0.0913	0.0355	0.0121
$X \leqslant 3$	0.9841	0.8670	0.6477	0.4114	0.2252	0.1071	0.0444

The so-called *power curves* plotting $\gamma(p)$ against p are shown in Figure 1.
From the above figures, we calculate the probabilities of error as follows.

test	true value of p						
	0.05	0.10	0.15	0.20	0.25	0.30	0.35
	$\beta(p)$ Type II			$\alpha(p)$ Type I			
$X = 0$	0.6415	0.8784	0.9612	0.0115	0.0032	00008	0.0002
$X \leqslant 1$	0.2641	0.6083	0.8244	0.0692	0.0243	0.0076	0.0021
$X \leqslant 2$	0.0755	0.3231	0.5951	0.2061	0.0913	0.0355	0.0121
$X \leqslant 3$	0.0159	0.1330	0.3523	0.4114	0.2252	0.1071	0.0444

Suppose that the test $X \leqslant 3$ is used:

- if H_0 is true, the probability of a Type I error may be as high as 41%;
- if H_1 is true, the probability of a Type II error is relatively small (compared to the other tests).

On the other hand, suppose that the test $X = 0$ is used:

- if H_0 is true, the probability of a Type I error is almost negligible;

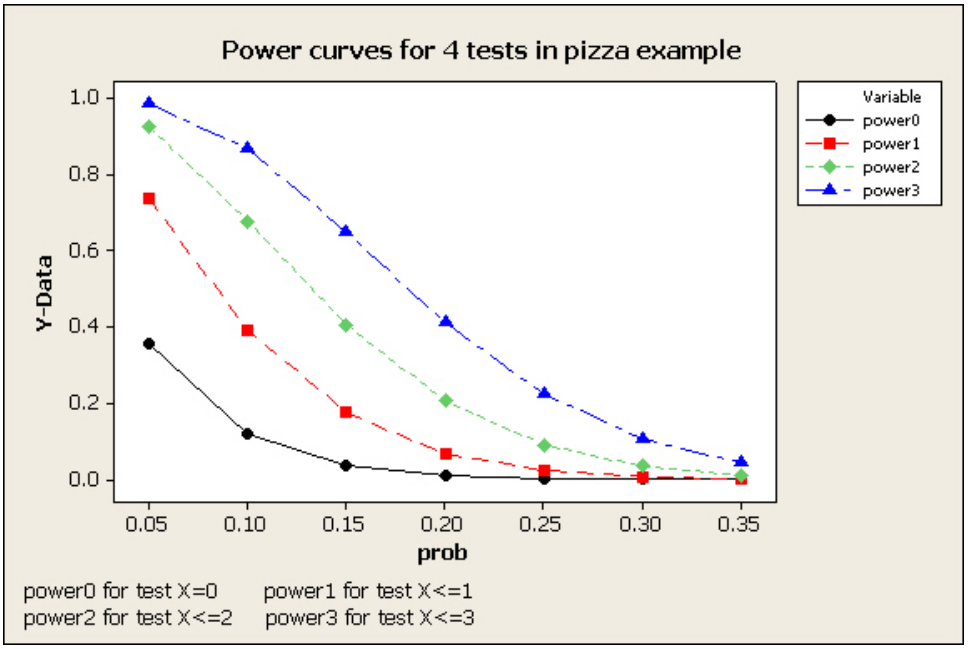

Figure 1: Power curves in pizza example

- if H_1 is true, the probability of a Type II error is large (compared to all the other tests), being more than 96% when $p = 0.15$.

Which test should we use? Ideally, we want α and β to be as small as possible, but one gets larger as the other gets smaller. Because a Type I error is usually more serious, we put a tolerance level first on α.

Suppose that Pete wants $\mathbb{P}(\text{Type I error}) \leqslant 0.1$, that is, $\alpha(p) \leqslant 0.1$ for all $p \geqslant 0.2$. From the foregoing table, the only tests he can use are $X = 0$ or $X = 1$.

Now we can consider Type II errors. Comparing these two tests, we see that the test $X \leqslant 1$ gives consistently lower values of $\beta(p)$ for $p < 0.2$. So we use the test $X \leqslant 1$.

For this test,

$$\max\{\alpha(p) : p \text{ satisfies } H_0\} = 0.0692,$$

which is actually rather less than 0.1. We say that 0.0692 is the *significance level* of the test. This means that if many people make similar trials with $n = 20$ and the test $X \leqslant 1$, for these hypotheses, a maximum of about 69 per 1000 will wrongly reject H_0.

Similarly, the test $X = 0$ has significance level 0.0115.

Suppose that we decide to use the test $X \leqslant 1$. When we find the value x from the sample (or trial), how should we report it?

(a) Suppose that $x = 0$. The value 0 is in the rejection region, so we report that "H_0 is rejected at 0.0692 significance level". Note that we do not say "H_0 is false".

This value would also have passed the stricter test $X = 0$, whose significance level is 0.0115. We do *not* change the significance level *after* finding the value of the statistic, but we do report the extra information.

The *P-value* of the observed value is defined to be the significance level of the strictest test that this value would have passed. In this case, we say "The P-value of this is 0.0115". This means that, if H_0 is true, then a value as extreme as, or more extreme than, the one we have observed would occur by chance in less than 1.2% of trials.

(b) Suppose that $x = 4$. This value is not in the rejection region, so we report that "H_0 is not rejected at 0.0692 significance level". This means that there is not enough evidence to reject H_0. It does not mean that H_0 is true.

The extreme value of p satisfying H_0 is $p = 0.2$. The test statistic is at least as extreme as our observed value 4 if $X \leqslant 4$. Table 1 of *New Cambridge Statistical Tables* shows that if $p = 0.2$ then $\mathbb{P}(X \leqslant 4) = 0.6296$. So the P-value is 62.96%. This means that if H_0 is true with $p = 0.2$ then we expect to find $x \leqslant 4$ in 63 cases out of 100, so it is not surprising that we do not reject H_0.

The smaller the P-value, the greater the evidence in favour of H_1.

In the pizza example, H_1 is called a *one-sided alternative hypothesis*, and the test is called a *one-sided test*.

Example I toss a coin 20 times to see if it is fair. Let $p = \mathbb{P}(\text{heads})$ and $X =$ number of heads out of 20. Now the two hypotheses are as follows:

$$\begin{aligned} H_0 &: \quad p = 0.5 \quad (\text{the null hypothesis}) \\ H_1 &: \quad p \neq 0.5 \quad (\text{two-sided alternative hypothesis}). \end{aligned}$$

Now rejection regions will have the from $\{0, 1, \ldots, c_1\} \cup \{c_2, \ldots, 19, 20\}$, so we will reject H_0 if $X \leqslant c_1$ or $X \geqslant c_2$. If the null hypothesis is true then the distribution of X is $\text{Bin}(20, 0.5)$, which is symmetric about 10, so it is sensible to take $10 - c_1 = c_2 - 10$; that is, $c_2 = 20 - c_1$.

Summary

Steps	Pizza example	Coin example
Unknown parameter	p	p
Two hypotheses	$p \geqslant 0.2$ vs. $p < 0.2$	$p = 0.5$ vs. $p \neq 0.5$
Decide which is the null hypothesis H_0	$p \geqslant 0.2$	$p = 0.5$
Decide on the type of trial or sample	Ask 20 customers which topping they prefer	Toss coin 20 times
Choose a suitable test statistic X	$X =$ number who like anchovy	$X =$ number of heads
Decide on the type of rejection region $(*)$, that is, a range of values of X which are unlikely if H_0 is true	$X \leqslant$ some number c	$X \leqslant k$ or $X \geqslant 20-k$
Find a reasonable probability model for the distribution of X	$X \sim \mathrm{Bin}(20, p)$	$X \sim \mathrm{Bin}(20, p)$
Set a tolerance limit $(\dagger)$ on the probability of Type I errors	$\alpha \leqslant 0.1$	$\alpha \leqslant 0.05$
Among tests whose rejection region has the form $(*)$ which satisfy $(\dagger)$, choose the one with the smallest probability of Type II errors (usually, this is a borderline case)	$X \leqslant 1$	$X \leqslant 5$ or $X \geqslant 15$
Find the significance level $(**)$ of the chosen test	$\alpha = 0.0692$	$\alpha = 0.0414$
Observe the value x of X	$x = 0$	$x = 14$
Report that H_0 is, or is not, rejected at significance level $(**)$	H_0 is rejected at the 6.92% level of significance	There is not enough evidence to reject H_0 at the 4.14% level of significance
Report the P-value, which is the maximum, over all parameter values which satisfy H_0, of the probability $\mathbb{P}(X$ is at least as extreme as $x)$ if that is the true value of the parameter	P-value $= 0.0115$	P-value $= 0.1154$

[1] D. V. Lindley and W. F. Scott, *New Cambridge Statistical Tables*, Cambridge University Press.