9 Hypothesis Testing

The two most common statistical applications are:

- **Estimation**: use data to find likely values of population mean, variance, proportion, etc.

- **Hypothesis Tests**: use data to say if a hypothesis about population mean, variance, proportion is true or false

- Hypotheses are usually about mean μ;

  "Mean height is 160cm"

  "Mean strength is ≥ 100,000 N"

  "Mean lifetime is ≤ 130 days"

  etc.

Example: (from Chatfield, p. 134)

- Company makes wire.

- It is known that wire strength has mean μ_0 = 1250 and s.d. σ = 150.

- Company has new production process.

- Measures strength of 25 new wires and finds $\bar{x}$ = 1312. (assume s.d. of new wire is also 150)

Question: does new process significantly increase strength of wires?

So, we want to test to see if $\mu = \mu_0 = 1250$

> The **Null Hypothesis** is the hypothesis that we want to test. It is denoted H_0

In example:

$$\mathbf{H_0: \mu = 1250}$$

New possibility is that mean strength is increased. So opposing theory to H_0 is that $\mu > 1250$

> The **Alternative Hypothesis** is the hypothesis that opposes H_0. We accept it only if we reject H_0. It is denoted H_1.

In example:

$$\mathbf{H_1: \mu > 1250}$$

Which Hypothesis is Which?

- With 2 hypotheses, how do we know which is H_0 and H_1?

H_0 is the hypothesis that

1) is assumed true now, and/or

2) must be disproved before we consider the other, and/or

3) we are most interested in departures from

Currently, $\mu = 1250$. We must disprove this before accepting that strength has increased.

We are more interested in seeing if μ has changed from 1250

So, H_0: $\mu = 1250$ ($= \mu_0$)
H_1: $\mu > 1250$

Another Example

A new drug is being tested.

We must be sure it works before production

Two hypotheses: drug works
drug does not work

H_0?

H_1?

<u>The **Null** hypothesis must specify the model fully.</u>

i.e. must have H_0: $\mu = \mu_0$

cannot have H_0: $\mu > \mu_0$ or H_0: $\mu < \mu_0$

If we want to test:

H_0: $\mu \leq \mu_0$
H_1: $\mu > \mu_0$

then in practice, we say

H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

We structure the procedure so that if we reject H_0, we would also reject $H_0 : \mu = \mu_1$ for any $\mu_1 < \mu_0$.

Test Statistics

- We have formed H_0 and H_1
- Now use data to decide if we can reject H_0
- To do this, we form a **test statistic** from data
- This shows us if the data is departing from H_0

Go back to wire strength example

1) Assume H_0 is actually true

2) If H_0 true, then $\bar{x}$ has mean 1250 and standard deviation $\sigma(\bar{x}) = \frac{150}{\sqrt{25}} = 30$.

3) If H_0 was true, would we expect to observe $\bar{x}$=1312 ?

   1312 is to the far right of the distribution (2.06 standard deviations) away from the mean (1250) – if H_0 true, it is an "extreme" value (less than 2% chance of observing a value as big as this, assuming $\bar{x}$ Normal).

   1312 is unlikely to be observed

So we might think that **H_0 mean=1250 is not true.**

This is pretty standard scientific method:

Construct a hypothesis (model): H_0
Compute the consequences: $\bar{x}$ ~N(1250,30^2).
If observation does not agree with the consequences reject the model (hypothesis)

Formal approach

This was developed in the 30s by Neyman and Pearson.

	The Universe	
Decision	H_0 true	H_0 False
Reject H_0	Error Type I	Correct
Do not reject H_0	Correct	Error Type II

Fix the P(Error Type I) at some small value called the *significance level* denoted by α.

Construct a test statistic T. Define a *critical region* *C* so that P(T belongs to *C* | H_0 true)=α.

If data returns a value of T that is in *C* reject H_0 .

There are a couple of things that should be noted:

We never demonstrate that H_0 is true – we either demonstrate it as false (with a chance of error = α) or we say we cannot reject the possibility it is true.

To use this approach we only need to calculate the distribution of T when H_0 is true. We don't need to know this when H_0 is false, this makes the approach usable in practice:

We can compute distribution of $\bar{x}$ if $\mu = 1250$, but this could not be done for the case $\mu > 1250$.

The choice of α is arbitrary – how sure you want to be that you do not reject H_0 when it is false. Typical values (thanks to available tables) are

5% - statistically significant (according to US Supreme Court),

1% - highly significant.

Why not 0! . Because it is a fact of life that then H_0 will never be rejected no matter how false it is. (Except for cases in which there is no randomness).

Note that the choice of T is not specified by the approach. This is very useful in complicated situations where working out the distribution of a "best T" might be impossible.

N-P theory goes on to tell us how to get the "best T" and the best critical region for most situations, but this is beyond this course.

The best T is defined as the one that minimizes: P(Type II Error) ie the probability of not rejecting H_0 when it is false.

$\beta = 1 - P(\text{Type II error})$ is called the *power* of the test. Power can only be computed if we can specify the distribution of T when H_0 is false.

<u>Applying the formal approach.</u>

The wire problem.

The company wants to be fairly sure that the wire is not of the same strength so we chose $\alpha = 0.05$.

The best T for testing this turns out to be based on the sample mean $\bar{x}$.

If H_0 is true and we assume a Normal distribution for $\bar{x}$. Then:

$$T = \frac{\bar{x} - 1250}{30} \sim N(0,1)$$

And the best C is T>c. Intuitively it is obvious that large values of $\bar{x}$ and hence T would lead us to reject H_0 in favor of H_1 $\mu > 1250$. N-P theory just confirms this.

All that is left is to compute c such that:

$P(T>c|H_0)=0.05$

From Normal tables we observe c=1.645 does the business.

The observed T=2.06 so we reject H_0 at 5%.

We note that:

$$\{T > c\} \Leftrightarrow \left\{\frac{\bar{x} - 1250}{30} > c\right\} \Leftrightarrow \{\bar{x} > 1250 + 30c\}$$

So in the end we reject H_0 if $\bar{x}$ is big enough.

In general, for testing whether the mean is larger than the value μ we have the decision rule:

Reject H_0 if:

$$T > c_\alpha \text{ where } T = \frac{\bar{x} - \mu}{sd(\bar{x})} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

where

c_α is such that $P(N(0,1) > c_\alpha) = \alpha$,

μ is the hypothesized mean,

σ^2 is the variance of the data and

n is the sample size.

<u>The alternative hypothesis H_1</u>

In the above we did not use H_1 explicitly, but we did use it to determine the form of the critical region: {T>c} i.e. $\bar{x}$ big enough, we would not reject H_0 if $\bar{x}$ was abnormally small.

If we had

H_1 $\mu < 1250$

The critical region would be {T<c}.

We need to think about the case:

H_1: $\mu \neq 1250$ the mean is different from 1250.

We want to reject H_0 if the data indicate the actual mean is bigger or smaller than 1250.

The critical region is of the form: $\{T < c_1\} \cup \{T > c_2\}$ to satisfy N-P theory we have:

$P(\{T < c_1\} \cap \{T > c_2\} | H_0) = \alpha$

There isn't a unique way of choosing c_1 and c_2 , conventionally (and sensibly) we take:

$P(T < c_1) = \frac{\alpha}{2}$ and $P(T > c_2) = \frac{\alpha}{2}$

If T is N(0,1) then $c_1 = - c_2$.

For $\alpha = 5\%$, c_1 =-1.96 and c_2 =1.96 (or effectively –2 and +2).

The first two tests are called *one tailed tests* and the last is a *two tailed test.*

We reject H_0 if the statistic is in the tails of the distribution.

Example of a two tailed test

Example: chemist wants to confirm that proportion of iron in a chemical is 0.12

Variance in prop. of iron known to be 0.01

Takes 10 samples: sample mean is 0.131

Question: is proportion of iron 0.12?

H_0: $\mu = 0.12$

H_1: $\mu \neq 0.12$

Take $\alpha = 0.05$.

$$T = \frac{\bar{x} - 0.12}{\sqrt{\frac{0.01}{10}}} = \frac{0.131 - 0.12}{0.031623} = \frac{0.011}{0.031623} = 0.348$$

As T is not less than –2 nor greater than +2 we cannot reject H_0 . The observations are consistent with the theory that $\mu = 0.12$.

Note however that this does not mean that $\mu = 0.12$. Just this data does not contain sufficient evidence to reject H_0 – with more data we might be able to do so. Suppose $\mu = 0.1200001$ i.e. H_0 is false but we would need millions of observations to have a hope of rejecting it.

A more pragmatic approach

The business of fixing a significance level for the test a priori is a philosophical requirement rather than a practical one. In practice we often quote the *observed significance level* . The α at which we would have just rejected H_0. This is usually called the *p-value*.

For the wire strength test we observed T=2.06

P(N(0,1)>2.06) = 0.019699 =p

For the iron content:

P(-0.348>N(0.1) or N(0,1) >0.348) = 0.72748 (verify as exx).

Discussion

The N-P approach is completely general – calculate the distribution of T when H_0 is true and reject H_0 is the observed T is unlikely.
In particular T does not have to be Normally distributed although it often is.

Choice of the significance level is somewhat arbitrary.

5% has been accepted as a criterion for scientific hypotheses. It was chosen when only a few values of the Normal distribution were accurately calculated but it has stuck.

1% is used for more critical situations such as medical or legal situations.

Quoting a p-value gets round this by ducking the issue.

In the context of clinical trials the US supreme court has recently stated that 2 sd (5%)(from mean) is importantly different and 3sd (~1%) is extremely different.

As we shall shortly see choosing α too small means that we rarely reject H_0 when it is false (i.e. frequently commit type II error) .

N-P theory does go on to tell us how to choose "best" T and critical regions – quite general procedures for situations when the exact distributions are known.