MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Hypothesis Tests

8.1 Introduction

A hypothesis H is a statement about a parameter – for example, that $\theta = 0$, $\mu = 4.2$ or $2 < \sigma < 5$. A test of a hypothesis is a procedure for deciding whether a *pre-conceived* hypothesis H is consistent with the data $x_1, x_2, \ldots, x_n$. This is not the same as deciding whether H is true or not. Data will *always* be consistent with two or more hypotheses that contradict each other!

Establishing the consistency of a hypothesis with the data is posed as a competition between this hypothesis, say H_0 and another one, H_1, although the two are not treated symmetrically. H_1 is present simply, or at least mainly, to define the direction of departures from H_0 that are regarded as interesting. For example, if testing whether grocery packages contain at least a specified amount μ_0, we would test $H_0 : \mu = \mu_0$ against the alternative $H_1 : \mu < \mu_0$, since the consumer does not care if s/he gets too much!

Thus a test of H_0 needs to provide an answer to the question: is the hypothesis H_0 consistent with the data $x_1, x_2, \ldots, x_n$ (or would a value of θ allowed by H_1 be preferable), or more precisely "is there significant evidence against H_0 in these data?". We call H_0 the null hypothesis and H_1 the alternative hypothesis, terminology that reinforces the asymmetry of the situation.

At its simplest, a hypothesis-testing procedure requires the following steps:

1. Statement of any model assumptions
2. Statement of the null hypothesis and the alternative hypothesis of interest
3. Calculation of the value of an appropriate test statistic

4a. Computation of the resulting p-value, or....

4b. Computation of the critical region for a specified significance level

5. Report on any conclusions.

Model Assumptions
As with any statistical procedure, we start with a probability model for the data. We will assume that the data is a simple random sample from the values of a particular population variable, whose population distribution is a member of a known parametric family.

We will first focus on the case when the parameter of interest is the population mean μ.

Null Hypothesis
Often the null hypothesis is that of *no difference* or *no effect* – i.e. there is no difference between the parameter value for this population and the parameter value for some previous reference population, reflecting the fact that the distribution of the variable in the current population is no different from that in the previous population, or that what differences there are have had no effect on this parameter value. That is why we call it the *null hypothesis*.

If we denote the known mean for the previous population by μ_0 and denote the unknown mean for the current population by μ, then the null hypothesis takes the form $H_0 : \mu = \mu_0$.

Alternative Hypothesis
We will usually have in mind some specific alternative hypothesis of interest, which we think might reasonably be true, and which we might accept if we reject H_0. We will denote the alternative hypothesis by H_1, and restrict attention to three standard cases:

(a) the current mean is greater than its previous value, i.e. $\mu > \mu_0$
(b) the current mean is less than its previous value, i.e. $\mu < \mu_0$
(c) the current mean differs from its previous value, i.e. $\mu \neq \mu_0$.

A shorthand way of writing the null and alternative hypotheses for case (a) is:

- $H_0 : \mu = \mu_0 \quad$ versus $\quad H_1 : \mu > \mu_0$,

with corresponding shorthands for the cases $H_1 : \mu < \mu_0$ and $H_1 : \mu \neq \mu_0$.

Test Statistic
To summarise the evidence provided by the data for or against H_0, we use the value of a suitable *test statistic* $T(X_1, \ldots, X_n)$, i.e. a function of the data with the following properties:
(a) 'extreme' values of the test statistic would be highly unlikely if H_0 were true and indicate evidence that H_0 is in fact false,
(b) when $\mu = \mu_0$ (i.e. when H_0 is true) the distribution of T is known and its distribution function is tabulated or can be easily calculated.

We have seen that the sample mean $\bar{X}$ is a natural estimator for an unknown population mean μ, so it is often sensible to base our test statistic on the function $\bar{X} - \mu_0$.

8.2 *p*-value approach: Consistency with H_0

If the observed value of our test statistic is relatively consistent with H_0 then it provides little or no evidence that H_0 is untrue. Thus, for a given value t, it is of interest to identify the set of values of the test statistic T which would be less consistent with H_0 and more consistent with H_1 than t. In later courses you will see that these are precisely the set of values whose relative likelihood of occurring under H_0 rather than H_1 is less than that for t.

Obviously, the set of values depends that are less consistent with H_0 and more consistent with H_1 depends on the particular alternative of interest H_1. In the three most common cases below, we can identify it by considering how the values we would expect to see for T would differ if H_1 rather than H_0 were true.

(a) $H_1 : \mu > \mu_0$ i.e. the alternative is that the current mean is greater than the reference mean. Here 'less consistent with H_0' corresponds to values such that $T(X_1, \ldots, X_n) > t$.

(b) $H_1 : \mu < \mu_0$ i.e. the alternative is that the current mean is less than the reference mean. Here 'less consistent with H_0' corresponds to values such that $T(X_1, \ldots, X_n) < t$.

(c) $H_1 : \mu \neq \mu_0$ i.e. the alternative is that the current mean differs from the reference mean. Here 'less consistent with H_0' corresponds to values such that $|T(X_1, \ldots, X_n)| > |t|$.

***p*-value**
Say our sample data $x_1, \ldots, x_n$ has resulted in an observed value $t_{obs} = T(x_1, \ldots, x_n)$ for the test statistic T. We measure the weight of evidence this provides by computing the probability, under the assumption that H_0 is true, of getting a value of the test statistic less consistent with H_0 (and more consistent with H_1) than the one actually observed. We call this probability the *p-value* corresponding to the observed value t_{obs}.

Thus, for each alternative, we calculate the p-value as follows:

(a) $H_1 : \mu > \mu_0 \Rightarrow p\text{-value} = P(T > t_{obs} \mid H_0 \text{ true})$

(b) $H_1 : \mu < \mu_0 \Rightarrow p\text{-value} = P(T < t_{obs} \mid H_0 \text{ true})$

(c) $H_1 : \mu \neq \mu_0 \Rightarrow p\text{-value} = P(|T| > |t_{obs}| \mid H_0 \text{ true})$.

Interpretation of the p-value
If the p-value is very small – i.e. the level of consistency with H_0 is very small – then we take that as strong evidence that either the null hypothesis $H_0 : \mu = \mu_0$ is false or that something very unlikely has happened. Thus, small p-values may well lead us to reject H_0 in favour of H_1.

Conversely, if the p-value is relatively large, then the this particular set of observations is relatively likely to occur when H_0 is true, and we conclude that there is no evidence to lead us to reject H_0.

8.3 Critical region approach: Type I and Type II Error

One way of evaluating the performance of a test procedure is to focus attention on some particular alternative value (say $\mu = \mu_1 > 0$) and ask how likely the procedure would be to detect that μ was not equal to μ_0 when in fact $\mu = \mu_1$. Denote this simple fixed alternative hypothesis by $H_1 : \mu = \mu_1$, and assume for the moment that μ can only take one of the two values $\mu = \mu_0$ or $\mu = \mu_1$. In this simplified context, there are only two possible errors, called type I error and type II error, where:

- Type I error is the error of deciding the null hypothesis H_0 is false when in fact H_0 is actually true,
- Type II error is the error of deciding the null hypothesis H_0 is true (and the alternative hypothesis H_1 is false) when in fact H_1 is actually true.

Significance level
There is a trade-off between type I and type II error. A change to the test procedure that reduces the type I error will usually increase the type II error, and vice-versa. Control of the type I error is often thought of as being more important, since H_0 represents in some sense the status-quo. Thus, a common way of applying a test procedure is to fix in advance some small acceptable threshold level α for the type I error. We call this level the *significance level* of the test, and speak of an α-level test. Typical values taken for α are 0.05 or 0.01.

Thus, for a test procedure with fixed type I error:

- Significance level α = P(Type I error) = P(Reject $H_0 | H_0$ true).

Note that, by definition, an α-level test procedure will reject H_0 if and only if the calculated p-value is less than or equal to α.

Note also that for large sample sizes, a small difference between the current parameter and the reference parameter may be statistically significant but not of any real practical importance.

Critical region
Fixing the significance level α in turn fixes the *critical region C* – the set of observations or values of the test statistic that would lead us to reject H_0 – and the *critical value* c^* – the value of the test statistic that is on the borderline between accepting and rejecting H_0. So

- Critical region C = set of values of the test statistic that would lead us to reject H_0.

When the test procedure has fixed type I error and the alternative of interest is $H_1 : \mu > \mu_0$, then this defines the critical value c^* as the value satisfying the condition $P(T > c^* | H_0 \text{ true}) = \alpha$. Corresponding conditions hold for the other two cases.

P(Type II error) and Power

Consider again the case of a simple fixed alternative and look first at $H_1 : \mu = \mu_1 > \mu_0$. A type II error occurs when we accept the null hypothesis H_0 as true when in fact H_1 is actually true. Under our test procedure for this alternative, we accept H_0 as true if and only if the value of our test statistic is less than or equal to c^*. Thus, for a test statistic T, critical value c^* and alternative hypothesis value μ_1, the probability of committing a type II error is just the probability that the value of our test statistic will be less than or equal to c^* when in fact H_1 is actually true, i.e.

- P(Type II error) = P(Accept $H_0 | H_1$ true) = $P(T \leq c^* | \mu = \mu_1)$.

We define the *power* of the test to be $1 - P(T \leq c^* | \mu = \mu_1)$, i.e. $1-$ P(Type II error). It gives a measure of how powerful the procedure would be in detecting that the alternative $\mu = \mu_1$ is true.

When the alternative of interest is $\mu < \mu_0$, so we are testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1 < 0$, the argument proceeds exactly as above, *except that* the orientation of the null and alternative values of μ has been reversed. Thus fixing the significance level at a given value α now fixes a *critical value* c^* such that we accept H_0 if the value of our test statistic is *greater* than or equal to c^* and we reject H_0 only if the value of our test statistic is *less* than c^*. It is still true that P(Type I error) = Significance level = α, but, for a test statistic T, critical value c^* and alternative hypothesis value μ_1, we now have

- P(Type II error) = P(Accept $H_0 | H_1$ true) = $P(T \geq c^* | \mu = \mu_1)$.

8.4 Confidence Intervals and Hypothesis Tests

Hypothesis tests are closely related to confidence intervals. In particular, the α-level test of $H_0 : \mu = \mu_0$ versus the two-sided alternative $H_1 : \mu \neq \mu_0$ will reject H_0 if and only if the corresponding two-sided $100(1-\alpha)\%$ confidence interval for μ does not contain μ_0.

Similar results connect one-sided tests and one-sided confidence intervals of the form $(-\infty, c_U)$ or (c_L, ∞). The α-level test of $H_0 : \mu = \mu_0$ versus the one-sided alternative $H_1 : \mu > \mu_0$ will reject H_0 if and only if μ_0 is not contained in the corresponding one-sided $100(1-\alpha)\%$ confidence interval (c_L, ∞), and the α-level test of $H_0 : \mu = \mu_0$ versus the one-sided alternative $H_1 : \mu < \mu_0$ will reject H_0 if and only if μ_0 is not contained in the corresponding one-sided $100(1-\alpha)\%$ confidence interval $(-\infty, c_U)$.

t-tests in R

We have already met the `t.test()` command in **R** in the context of calculating confidence intervals. The **R** command:

```
> t.test(data, mu = 0, alternative="greater", conf.level=0.9)
```

will compute a one-sample t-test on observations in an array `data`, with null hypothesis $H_0 : \mu = 0$, with alternative hypothesis $H_1 : \mu > 0$, and at significance level $\alpha = 0.1$. The numerical mean value 0 can be replaced by the value appropriate for your data, the alternative hypothesis "greater" can be replaced by the alternatives "less" or "two.sided" as desired, and the significance

level can be changed by setting `conf.level` to $1-\alpha$ (the default value is $\alpha = 0.05$).

8.5 Example - Normal distribution with known variance

As in §7.1, we start by consider the (unrealistically) simple case where we have a random sample of size n from a Normal $N(\mu, \sigma^2)$ distribution, where the population mean μ is an unknown parameter which we wish to test but the population variance σ^2 is known (say $\sigma^2 = \sigma_0^2$).

The following gives a typical example: When patients with a certain type of chronic illness are treated with the current standard medication, the mean time to recurrence of the illness is 53.3 days, with a standard deviation of $\sigma_0 = 26.4$ days. A new type of medication, that is thought to increase the time until recurrence, was tried by a randomly chosen sample of 16 patients. For this sample, the mean time to recurrence was $\bar{x}$ = 65.8 days.

Assuming the variance of the recovery time is the same for the new medication as for the current medication, we might want to test whether the new medication has increased the mean time to recovery, using a test with significance level, say $\alpha = 0.05$.

Model assumptions:
(a) $x_1, \ldots, x_n$ are the observed values of a random sample $X_1, \ldots, X_n, \quad \ldots$
(b) ...from a population with the Normal $N(\mu, \sigma^2)$ distribution, where μ is unknown but the value of σ^2 is known – say $\sigma^2 = \sigma_0^2$.

Thus in our medical example above we might assume:
(a) The recurrence times for the $n = 16$ patients are a random sample from the population of recurrence times for all patients that will use this new medication ...
(b) ...with distribution $N(\mu, \sigma^2)$, where μ is unknown but the value $\sigma^2 = \sigma_0^2 = (26.4)^2$ is known.

Hypotheses:
Say the past or 'status quo' value of the mean is some pre-assigned or known value μ_0 and we are interested in whether there is sufficient evidence to conclude the mean of the population from which the sample is taken has mean $\mu > \mu_0$. Then we take:

- Null hypothesis to be $H_0 : \mu = \mu_0$ (corresponding to no difference between the means)
- Alternative hypothesis $H_1 : \mu > \mu_0$ (corresponding to the new mean being greater)

Thus, in our medical example we would take: $H_0 : \mu = \mu_0 = 53.3$ versus $H_1 : \mu > 53.3$
The null hypothesis H_0 corresponds to *no difference* between the mean recurrence time μ for the new medication and the mean recurrence time $\mu_0 = 53.3$ for the standard medication. The alternative hypothesis H_1 corresponds to the mean recurrence time for the new medication being longer than the mean recurrence time for the standard medication.

Test Statistic:
Since $\bar{X}$ is the natural estimator of μ, we base our test statistic on $\bar{X} - \mu_0$. Since the population standard deviation σ_0 is assumed known we can take as our test statistic

$$T(X_1, \ldots, X_n) = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0$$

Then from §6, when H_0 is true (i.e. when $\mu = \mu_0$) we have $X \sim N(\mu_0, \sigma_0^2)$ and $T \sim N(0, 1)$.

In our medical example, this means we base our test statistic on $\bar{X} - 53.3$. Since the population standard deviation $\sigma_0 = 26.4$ is assumed known and $n = 16$, we can take as our test statis-

tic $T(X_1, \ldots, X_n) = \sqrt{n}(\bar{X} - \mu_0)/\sigma_0 = \sqrt{16}(\bar{X} - 53.3)/26.4$, where $\bar{X} \sim N(\mu, \sigma_0^2/n) = N(\mu, (26.4)^2/16)$. Thus, when H_0 is true (i.e. when $\mu = \mu_0 = 53.3$) we have $T = \sqrt{16}(\bar{X} - 53.3)/26.4 \sim N(0, 1)$.

The data gives $\bar{x} = 65.8$ so the observed test statistic is $t_{obs} = \sqrt{16}(65.8 - 53.3)/26.4 = 1.893$.

***p*-value**:
Since the alternative of interest is H_1: $\mu > \mu_0$, the values of T which are less consistent with H_0 than t_{obs} are the set of values $\{T > t_{obs}\}$. Also, when H_0 is true, $T \sim N(0, 1)$. Thus

$$p\text{-value} = P(T > t_{obs} | H_0 \text{ true}) = P(Z > t_{obs}) \text{ (where} Z \sim N(0,1)) = 1 - \Phi(t_{obs})$$

In the medical example, the values of T which are less consistent with H_0 than t_{obs} are the set of values $\{T > t_{obs} = 1.893\}$ so

$$p\text{-value} = P(T > t_{obs} | H_0 \text{ true}) = P(Z > 1.893) = 1 - \Phi(1.893) = 1 - 0.9708 = 0.0292.$$

Critical region:
Since the alternative of interest is H_1: $\mu > \mu_0$, the values of T which are less consistent with H_0 than a given value t are the set of values $\{T > t\}$ and the critical region of values for which the test would reject H_0 is of the form $C = \{T > c^*\}$.

To find c^* for a given significance level α, we recall that a test has significance level α if P(Reject $H_0 | H_0$ true) $= \alpha$. Thus, for a 0.05-level test, c^* is defined by the condition

$$0.05 = \alpha = P(\text{Reject } H_0 | H_0 \text{ true}) = P(T > c^* \mid H_0 \text{ true}) = P(Z > c^*) = 1 - \Phi(c^*),$$
$$\text{so } c^* = \Phi^{-1}(1 - \alpha) \quad \text{and for } \alpha = 0.05 \text{ this gives } c^* = \Phi^{-1}(0.95) = 1.645.$$

Thus in the medical example the critical region of values C has the form $C = \{T > c^*\}$, i.e. $C = \{T > 1.645\}$. Since $t_{obs} = 1.893$ is in C, the 0.05-level test would lead us to reject H_0.

NOTE: the form of the set of values of T which are less consistent with H_0 than a given value t depends crucially on the choice of the alternative hypothesis H_1. Here, for $H_1 : \mu < \mu_0$ it would have form $\{T < t\}$, and for $H_1 : \mu \neq \mu_0$ it would have form $\{T > |t|\}$.

Conclusions:
In giving conclusions we should (a) report the p-value and/or whether t_{obs} is in the critical region; and (b) interpret that to make practical conclusions about μ in the context of the example.

In the medical example, the p-value of 0.03 is quite small – if the mean for the new medication was really 53.3 we would only only observe data for which the consistency with H_0 was this small about 3 percent of the time. Thus there is reasonably strong evidence that H_0 is not true.

Similarly, the observed test statistic value $t_{obs} = 1.893$ falls well within the critical region of the 0.05-level test, so at this level we would reject H_0 in favour of H_1, and conclude that the new medication has increased the mean time to recovery.

As the required significance level decreases the borderline level of consistency also decreases. A level that was borderline for the 0.05-level test would be (well) above the borderline for, say, a 0.025-level test, and here $t_{obs} = 1.893$ would not be in the critical region of a 0.025-level test. Thus, if we only classified an observation as inconsistent with H_0 if it was outside this lower threshold, we would be more cautious and report that there is insufficient evidence to conclude that the new medication has increased the mean time to recovery.