2.4.3 General Procedure for Hypothesis Testing

We can bring together the various steps in taken in the examples above to give a generic guide for carrying out hypotheses tests in simple situations considered in this course.

1) Set up the **null and alternative hypotheses**, denoted by H_0 and H_A, respectively:–

i) The ***null hypothesis*** is usually the hypothesis that results in a more complete description of the probability model for the population. Typically it will specify the values of parameters in a statistical model and so allow numerical calculation of probabilities *under the null hypothesis* (i.e. if the null hypothesis is true).

Examples:–

a) The statistical model might be that our observations arise from a Normal distribution $N(\mu, \sigma^2)$ with σ a known value and H_0 might be that our hypothesis is that $\mu = 0$ (written for convenience as H_0: $\mu = 0$).

b) We have two groups and our model is that the observations of the two groups, X_i and Y_i say, come from Normal distributions with identical known variances $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ respectively. Our null hypothesis is that the two groups have identical means written as H_0: $\mu_X = \mu_Y$. Note that this does not specify the value of μ_X or μ_Y, only that they have the same (unknown) value. This will still allow us to make detailed numerical probability calculations since this hypothesis is equivalent to H_0: $\mu_X - \mu_Y = 0$, and we can thus say that the difference in sample means, *under the null hypothesis*, is an observation from $N(0, \{\sigma^2/n_X + \sigma^2/n_Y\}^{1/2})$ where n_X and n_Y are the

number of observations in the two groups and remember that σ^2 is assumed to be known numerically.

ii) The ***alternative hypothesis*** generally describes a departure of interest from the null hypothesis. Typically it is rather vaguer and does not specify values of parameters completely although in certain simple situations it might. It will generally not be possible to calculate probabilities numerically *under the alternative hypothesis*, instead it will indicate whether under the alternative hypothesis the test statistic is anticipated to be larger or smaller or merely different (i.e. either larger or smaller with equal evidence). In particular, it is important at this stage to decide whether the test is one or two-sided, that is, whether interest lies in departures from one or both sides of the relationship proposed in the null hypothesis.

Examples:–

a) H_0: $\mu = 0$ *vs* H_A: $\mu \neq 0$ — evidence of μ being either larger or smaller than 0 are of equal interest and a two-sided test is required.

b) H_0: $\mu = 0$ *vs* H_A: $\mu > 0$ — only evidence of μ being larger than 0 is of interest, any evidence of μ being smaller is not of interest. A one-sided test is required.

c) H_0: $\mu_X = \mu_Y$ *vs* H_A: $\mu_X \neq \mu_Y$ — evidence of μ_X being either larger or smaller than μ_Y are of equal interest and a two-sided test is required.

d) H_0: $\mu_X = \mu_Y$ *vs* H_A: $\mu_X < \mu_Y$ — only evidence of μ_X being smaller than μ_Y is of interest, any evidence of μ_X being larger is not of interest. A one-sided test is required.

2) Decide upon a **test statistic** and find its ***sampling distribution*** under the null hypothesis (i.e. assuming that the null hypothesis is true). In future modules you will meet methods for finding an appropriate test statistic for a given test. In this module we only make intuitive arguments for choices of test statistic based primarily upon the formulation of the null hypothesis; later modules will indicate how to use the alternative hypothesis in the construction of good test statistics. If the null hypothesis specifies the value of a parameter then an obvious *test statistic* is some estimator of it or else something closely related to it, e.g. **estimator/s.e.(estimator)** if the s.e.(estimator) is known numerically or else **estimator/e.s.e.(estimator)** if it isn't. The test statistic has to be a quantity whose sampling distribution can be determined under the null hypothesis.

Examples:–

a) Statistical model $X \sim N(\mu, 1)$, null hypothesis H_0: $\mu = \mu_0$ where μ_0 is some *specified value*, e.g. 0 or 39.3,…. . A suitable test statistic is the sample mean, $\bar{X}$, which has sampling distribution $N(\mu_0, n^{-1})$ where n is the sample size. If the observed value of the test statistic, $\bar{x}$, is very different from μ_0 *in the direction suggested by the alternative hypothesis* H_A, then it provides evidence against the null hypothesis.

An equally good test static would be $\bar{X} - \mu_0$ which has sampling distribution $N(0, n^{-1})$.

b) Statistical model $X \sim N(\mu, \sigma^2)$ where σ^2 is not known, null hypothesis H_0: $\mu = \mu_0$. An estimator for μ is $\bar{X}$ with standard error $\sigma/\sqrt{n}$ and estimated standard error $s/\sqrt{n}$ so a test statistic might be $\bar{X}/(s/\sqrt{n})$ or else $(\bar{X} - \mu_0)/(s/\sqrt{n})$ which have *approximate* sampling distributions $N(\mu_0, 1)$ and $N(0, 1)$ respectively.

3) Model assumptions. In finding the null distribution of the test statistic one always needs to make certain assumptions about the population and sample. In the case of the z-test above we must assume that we have a random sample, that the variance of the population is known and that either (i) the population is well modelled by a normal distribution, or (ii) the sample size n is large enough so that $\bar{X}$ is effectively normally distributed. At some point we must ask, are these assumptions likely to be met for this population? The step of checking model assumptions may itself involve carrying out one or more hypothesis tests. We will discuss this later in the module.

4) Observed value of test statistic. This usually involves a simple calculation.

5) Find the **p-value** of the investigation. Remember, it is the probability, under H_0, of a result as extreme as or more extreme than the sample result *in the direction of the alternative hypothesis*.

i) If the alternative hypothesis is two-sided (e.g. H_A: $\mu \neq \mu_0$) then we calculate the probability of obtaining a result equally far above (or further than) or equally far below (or further than) the hypothesised value.

ii) If the alternative is one-sided (e.g. H_A: $\mu > 0$) then we calculate the probability of obtaining a result equally far above or further above the hypothesised value.

6) Interpret and draw conclusions on model. The p-value of an investigation gives a precise numerical summary of the level of evidence against the null hypothesis provided by the data. The decision of whether to **'reject'** the null hypothesis should be based on the p-value and other criteria specific to the context of the problem.

A convention that has grown up through the use of statistical tables leads to a categorisation of the strength of evidence against H_0 based on a comparison of the p-value with commonly tabulated probabilities. If the p-value of an investigation is as small as or smaller than some value α (usually α = 0.10, 0.05, 0.01) we say that the data are "statistically significant at level α" or "statistically significant at the 100(1 – α)% level". The strength of evidence against H_0 is judged on this scale and interpreted as an adjective [preceding the word 'evidence'] describing the strength of the evidence as follows:–

p-value	**adjective**
p > 0.10	no
0.10 > p > 0.05	some
0.05 > p > 0.01	good
0.01 > p	strong

Some gradation of this is possible and is largely a matter of individual taste and habit, e.g.

p-value	adjective
0.15> p > 0.10	little
0.10 > p > 0.075	some slight
0.075 > p > 0.05	some suggestion of
0.05 > p > 0.025	fairly good
0.025 > p > 0.01	very good
0.01 > p > 0.001	strong
0.001> p >0.0001	very strong
0.0001 > p	overwhelming

(Remember all these adjectives precede the word *evidence* and it is usual to give the p-value as well.)

Examples:–

a) "The data are significant at the 5% level" ≡ "The data provide good evidence ($p < 0.05$) against the null hypothesis that ……"

[**NB:** it is presumed that if a level of significance is declared then this is the smallest of the conventional boundary values that can be quoted, i.e. if it is said to be significant at the 5% level then it is presumed that $0.05 > p > 0.01$]

[**NB again:** if the exact p-value is available then this would be quoted: "The data provide good evidence ($p = 0.0362$) against the null hypothesis that ……"]

b) "The data are significant at the 10% level" ≡ "The data provide only some evidence ($0.1 > p > 0.05$) against the null……"

c) "The data are significant at the 5% level" ≡ "The data provide very good evidence ($p = 0.0176$) against the null hypothesis that ……"

[**NB:** Common sense is required, especially in borderline cases, i.e. when the p-value is close to one of the conventional (but arbitrary nevertheless) boundary values of 0.1, 0.05, 0.01 etc. Common sense dictates that the conclusions should be the same whether the p-value is 0.051 or 0.049.]

7) The I.S.E.E. (If Significant Estimate Effect) principle. If the data do provide evidence against H_0 (conventionally the p-value is less than 0.1) then, in the case of hypothesis tests on parameters, you should include an estimate (with a confidence interval) of the effect to describe the size of the departure from H_0. If you do not then all you are providing is a negative statement that there is some/good/strong evidence that the null hypothesis is not true.

8) Interpret the conclusion within the context of the practical application at hand. Typically this means going back from the statistical model to the practical problem (see full example below).

2.4.3.1 Example:– (from §2.4.2.1)

Practical problem:– Sample of 25 mummy pots.

Hypothesis:– These pots come from Saqqara where the mean base circumference is 220mm

Statistical model:– measurements are Normally distributed with mean μ and variance 25^2, i.e. $X \sim N(\mu, 625)$

Null hypothesis:– H_0: $\mu = 220$mm

Alternative hypothesis:– H_A: $\mu \neq 220$mm

Test statistic & sampling distribution:– $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ where here $\mu_0 = 220$, $\sigma = 25$, $n = 25$. The sampling distribution of this is N(0, 1).

Model assumptions:– The 25 pots are a random sample (i.e. drawn independently with equal probability of selection from their source) and the measurements of base circumference are well-modelled by a Normal distribution with standard deviation 25.

Observed value of test statistic:– We calculate $z_{obs} = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ where we have $\bar{x} = 232.4$, $\mu_0 = 220$, $\sigma = 25$ and $n = 25$ so $z_{obs} = (232.4 - 220)/(25/\sqrt{25}) = 12.4/5 =$ **2.48**.

p–value:– If $Z \sim N(0, 1)$ then $P[Z > 2.48] = 0.00657$. Since the alternative hypothesis is two-sided (i.e. of form $\mu \neq 220$) we need to allow for the possibility of being equally or more extreme below the hypothesised value of 220 and so calculate the p-value as $2 \times 0.00657 = 0.0131$, i.e. about 1.3%.

Interpret and draw conclusions on model:– "The data are significant at the 5% level", or (better) "The data provide very good evidence (p=0.0131) that $\mu \neq 220$".

ISEE:– An estimate of μ is 232.4 with 95% confidence interval $232.4 \pm 1.96 \times 5 = (222.6, 242.2)$

Real-world interpretation and conclusion:– Measurements of the base circumferences of the sample of 25 mummy pots at the British Museum provide very good evidence (p=0.0131) that these pots do not originate in Saqqara because their base circumferences are approximately 12mm ($\pm$ 10mm) larger than is typical of Saqqara pots.

NB the $\pm$10mm statement is conventionally taken to indicate something equivalent to a 95% confidence interval (rather than say a 66% or 99.9% one) but this interpretation is very informal.

2.4.4 General Comments

1) It is important to get H_0 and H_A the right way around. Typically, the ***alternative hypothesis*** H_A is the **research hypothesis**,

e.g. the mean age of this population is different from that of another population;

e.g. the proportion in the population owning a car has increased from a known level;

whilst the null hypothesis H_0 is that there is no difference/no effect/no change statement.

2) Tests are usually two-sided unless there are very good prior reasons, not observation or data based, for making the test one-sided. If in doubt, then use a two-sided test.

i) Situations where a one-sided test is definitely called for are uncommon but one example is in a case of say two drugs A (the current standard and very expensive) and B (a new generic drug which is much cheaper). Then there might be a proposal that the new cheaper drug should be introduced ***unless there is evidence that it is very much worse than the standard***. In this case the model might have the mean response to the two drugs as $\mu_A = \mu_B$ and if low values are 'bad', high values 'good' then one might test H_0: $\mu_A = \mu_B$ against the one-sided alternative H_A: $\mu_A > \mu_B$. However, this example does raise further issues which will not be followed further here.

3) If testing hypotheses about a parameter, or function of parameters, then the test statistic is usually an *unbiased* estimator of the parameter or function.

4) If the sampling distribution of the test statistic is normal (or approx normal) then the test statistic may be given in its standardised form under H_0. For example a z-test for μ often has the test statistic given as $(\bar{X} - \mu_0)/\text{s.e.}(\bar{X})$ or $(\bar{X} - \mu_0)/\text{e.s.e.}(\bar{X})$ where μ_0 is the value of μ under H_0.

5) Logic of conclusions:–

i) There are always two explanations of the observed test statistic if the p-value is low:

a) H_0 is true but we happen to have observed a very unusual data set

b) H_0 is false and we have observed an 'ordinary' data set under the situation which is really true i.e. H_A.

We can never be certain which is the true case. A low p-value, i.e. a significant result, is not necessarily an important result. It just provides good **evidence** that an effect is present, but the effect may be of no practical importance. For example, a difference in population mean lifetime of one day may be important when comparing insects but not humans. However, hypothesis tests based on sample data may reveal 'significant' differences (that is, a non-zero difference) in two population mean lifetimes in both cases, 'significant' in the sense of 'statistically significant' or 'providing evidence' rather than important in the real world.

ii) **A large p-value does not guarantee that H_0 is true!** It just indicates that the data did not provide evidence against H_0.

Never talk about 'accepting' H_0, only of '**not rejecting H_0**'. For example, suppose when testing H_0: $\mu = 220$ *vs* H_A: $\mu \neq 220$ an observed sample mean yields a p-value of 0.43. This does not prove that $\mu = 220$, merely that the data are not inconsistent with sampling from such a population. There are infinitely many other values of μ which are consistent with the data. This emphasizes the importance of providing confidence intervals generally, not just in the case where the null hypothesis is rejected. In such a situation as this typically the confidence interval will be very wide, centred on the observed mean and easily including the hypothesised value 220.

6) Summary:– Hypothesis tests are a way of ***evaluating the level of the evidence*** provided by the data ***against*** the null hypothesis ***in favour*** of the alternative hypothesis (here 'in favour of' means 'in the direction of' or 'towards'). A highly significant result ($p < 0.05$) means that there is **good evidence** against the null hypothesis. Whether the actual difference from the null hypothesis is of any real-world importance is not a statistical issue — it is up to the scientist putting forward the problem who must judge the real-world importance:–

strength of evidence is not evidence of strength

Equally a statistically non-significant result ($p > 0.10$) does not mean the null hypothesis is true, only that the available data provide little or no evidence against it.

absence of evidence is not evidence of absence