UNIT 7: HYPOTHESIS TESTS AND P-VALUES

AIMS

To convey the ideas underlying the testing of hypotheses and how these ideas relate to confidence interval estimation; provide some understanding of which hypotheses test is appropriate; and to show how results of test are interpreted.

OBJECTIVES

At the end of Unit 7 you should be able to:

- Explain the role of hypothesis testing in the context of statistical inference.
- Explain what is meant by the null and alternate hypotheses.
- Explain what a level of significance is.
- Explain with a numeric example what a p-value is and how it is used in hypothesis testing.
- Distinguish between Type I and Type II errors.
- Demonstrate an understanding of the differences and similarities between hypothesis testing and confidence interval estimation.
- Describe some of the more common hypothesis tests and the conditions and assumptions which govern their use.

Reading: Bland: pp.133-9.
or Bowers-2: pp.66-76.

You don't need to follow the details of the examples, just try to understand the main line of reasoning.

Introduction

At the beginning of Unit 6 we discussed briefly the difference between estimation and hypothesis testing as two alternative approaches to statistical inference. For reasons which we will consider later in this unit, estimation, rather than hypothesis testing, is the preferred approach. However, there are three principal reasons why we need to spend some time discussing hypothesis tests. First, for some applications hypothesis testing offers a simpler, more convenient, approach. Second, published journal papers often include the results of hypothesis tests, so some familiarity with the most common tests is needed. Finally, some computer programs include hypothesis tests but not the confidence interval equivalent.

The null hypothesis

Researchers generally start with a *research question*. For example, "Is the bone mineral density of depressed and non-depressed women different?"; "Does pre-operative anaesthesia reduce post amputation pain?", and so on. These questions are often framed in the form of a *hypothesis* (which might not be stated explicitly) which the researchers are curious to test and thus answer their research question. In both of the above examples the authors studies used the confidence interval approach to resolving their research question.

However, the equivalent hypothesis test approach would have been to express the research question in what is called its **null** form (which usually the researchers wish to "disprove"). For example:

null hypothesis: there is *no* difference in bone density between depressed and non-depressed women

or,

null hypothesis: pre-operative anaesthesia has *no* effect on post-amputation pain levels

The null hypothesis is always framed in this negative way - no change, no effect, no difference, etc. The **alternate hypothesis** is invariably that there *is* a difference/effect/change.

With this approach we need some way of deciding for or against the null hypothesis. All we have available to help us decide one way or the other is the sample data. So the question becomes, "Does the evidence from the sample data suggest that we should *reject* the null hypothesis (because it appears *not* to be

true) or *not* reject it (because it appears to be *true*)?" We decide between these alternatives by using what is called the *p-value*. We'll see how in a moment.

Q. 7.1 A researcher believes that ginkgo biloba might be effective in improving cognitive functioning in elderly subjects. Express this belief in the form of: (a) a research question; (b) null and alternate hypotheses.

The p-value

Hypothesis tests and confidence intervals start with exactly the same information - data obtained from a sample. However, whereas the estimation approach uses the sample data to construct a confidence interval, in hypotheses testing the data is used to calculate what is known as a **p-value**. It is the value of this statistic which we use to decide whether to reject the null hypothesis or not.

> **The p-value is a *probability* - the probability of getting the sample outcome actually observed (or one more extreme) if the null hypothesis is true.**

To illustrate the idea, suppose you are a doctor in a sexual health clinic. The question is raised as to whether or not the same proportion of males and females use the clinic. Your impression from working there for the last year is that the proportions *are* the same. You decide to investigate more thoroughly. The research question is, "Do equal proportions of males and females use the clinic?" The null hypothesis becomes, "The proportions of males and females are the same."

You take as a sample the records of the last 1000 patients. Now if the null hypothesis is true you expect to find *about* 500 males (and about 500 females) in the sample. We have to say "about" because we know that we are extremely unlikely to get exactly 500 males and 500 females, even if the true proportions *are* the same - as we have already seen, a sample is never an *exact* replica of its population.

Let's look at a few possible outcomes. If you got 490 males and 510 females you would be inclined *not* to reject the null hypothesis of equal proportions. Why? Because the probability of such an outcome, or one more extreme* - i.e. the p-

---

* By more extreme we mean outcomes even further away than 490 from the null hypothesis value of 500 males, i.e. not only 490 but also 489, or 488, or 486, ... or 3, or 2, or 1, or 0 males.

value - is quite high if the null hypothesis is true. In other words, this is not a particularly unusual result. Suppose you got 10 males and 990 females. Clearly with this outcome we would happily reject the null hypothesis, Why? Because the probability of this outcome (the p-value) is very small *if* the null hypothesis is true.

Suppose you got 450 women and 550 men - about 50 fewer females than you might have expected (and 50 more males). The decision as to whether to reject or not reject the null hypothesis is now more difficult. Its quite possible, even with *equal* proportions in the population, to get outcome proportions such as these (bearing in mind the vagaries of sampling). How do we decide in cases like this? In the first case above, an outcome of 490 males has a high probability, i.e. a high p-value. In the second case, an outcome of 10 males has a very low probability, low p-value. Clearly there is a critical "line in the sand" - a p-value beyond which not to reject the null hypothesis would stretch credulity beyond what is reasonable. This critical value for the p-value has been determined by convention to be 0.05 (0.01 is also sometimes used) and is called the **significance level** of the hypothesis test (denoted α).

In other words, if the p-value (the probability of any particular outcome) is less than 0.05 we will reject the null hypothesis, because such an outcome is so unlikely that the null hypothesis is almost certain not to be true.

Another way of thinking of the p-value is as a **measure of the strength of the evidence against the null hypothesis**. If the evidence against is strong enough, the null hypothesis can be rejected in favour of the **alternate hypothesis** (that depression *does* reduce bone mineral density, or that pre-operative anaesthesia *does* reduce stump pain). A legal metaphor might be helpful here. When a person appears in court accused of a crime, there is a presumption of innocence (the null hypothesis is that the accused is innocent). The prosecution presents evidence (the sample data) to the jury. The jury assesses the evidence (calculates a p-value) and decides whether it is strong enough to reject the assumption of innocence (a low p-value - reject), or not (a high p-value – do not reject)).

In other words, the decision as to whether to reject or not reject the null hypothesis amounts to a comparison between the p-value associated with a particular outcome and the significance level, usually 0.05 (sometimes 0.01). The decision rule is:

If the p-value is less than 0.05, then the evidence is sufficiently strong against the null hypothesis for it to be rejected. If the p-value is not less than 0.05, the null hypothesis cannot be rejected.

In practice of course we do not have to calculate p-values by hand; most computer statistics programs will provide a p-value with every test they are asked to perform.

The hypothesis test procedure can be summed up as follows (where α is the level of significance - usually 0.05, sometimes 0.01):

- **Decide on the research question.**
- **Transform the research question into an appropriate null hypothesis.**
- **Decide on the α to be used in the test.**
- **Calculate a p-value for whatever outcome obtained.**
- **Compare this p-value with α.**
- **If the p-value is less than α - reject the null hypothesis; otherwise do not reject it.**

Q. 7.2 Suppose the researcher with the ginkgo biloba question conducts a randomised control trial and finds after three months that in the treatment group (those receiving ginkgo biloba) the median improvement in a cognitive function score (with a range of 0- 50) is 5 points compared to a zero increase in the placebo group. If significance levels are, (a) 0.05 and (b) 0.01, what decisions will be made about the null hypothesis if the p-value for this sort of improvement is: (i) 0.03; (ii) 0.50; or (iii) 0.009?

Types of error

Its important to note that when we make the decision to reject or not reject the null hypothesis on the basis of a p-value and a significance level of 0.05, there is a probability of 0.95 (i.e. 1 - 0.05) that our decision is correct, but we can never be absolutely *certain* of this because of the vagaries of samples. Two obvious possibilities for errors exist:

- Either we reject the null hypothesis when it is correct, i.e. it should *not* have been rejected - known to statisticians as a **Type I** error, and to clinicians as a **false positive**. We conclude there is an effect but there isn't.

- Or we do not reject the null hypothesis when it is incorrect, i.e. we *should* have rejected it-known to statisticians as a **Type II** error, and to clinicians as a **false negative**. We conclude there isn't an effect but there is.

Q. 7.3 Suppose in Q. 7.2(a)(i) that the researcher misreads the computer print-out p-value as 0.30 instead of 0.03, and accordingly does *not* reject the null hypothesis of no effect. What sort of error is committed?

Hypothesis tests and confidence intervals compared

Confidence intervals and hypothesis tests perform a similar function. For example, in the bone mineral density study (bmd) referred to above and considered in UNIT 6, the question the authors wished to answer was, "Is the bmd the same in depressed and non-depressed women?" This question can be answered in two ways, either by calculating a confidence interval for the difference in mean bmds and seeing if it contains 0 (if so, the means are the same; if not, the means are different).

Or we could test the null hypothesis of no difference in mean bmds, using the p-value as described above, i.e. seeing if its less than 0.05 (if not, there's no difference in means; if it is, there is a difference). The results from these two alternative approaches will be the same:

> **When testing the difference between two population measures (i.e. two means, medians, etc) if the confidence interval for the difference includes 0, then the p-value will equal 0.05 or more; if the confidence interval doesn't include 0, then the p-value will be less than 0.05.**

Confidence intervals are preferred to p-values for two reasons primarily. First, because the former provide a *range of plausible values* for the difference in two population parameters, whereas the hypothesis test only indicates whether the two parameters are equal or not, without any information on the magnitude of any difference. Second, the confidence interval is in clinically meaningful units.

In the study on bone mineral density (bmd), which we first encountered in Unit 6 (see Figure 6.3), the authors used both approaches. Figure 7.1 shows 95% confidence intervals for differences in mean bone mineral density (bmd) at six sites, along with the corresponding p-values.

The authors' unstated null hypothesis is that there is no difference in true (population) mean bmd between the two groups of women. From the table we can see that out of the six sites it is only at the radius, where the confidence interval (-0.01 to 0.04) includes 0, that the true difference in mean bmd is not statistically significant.

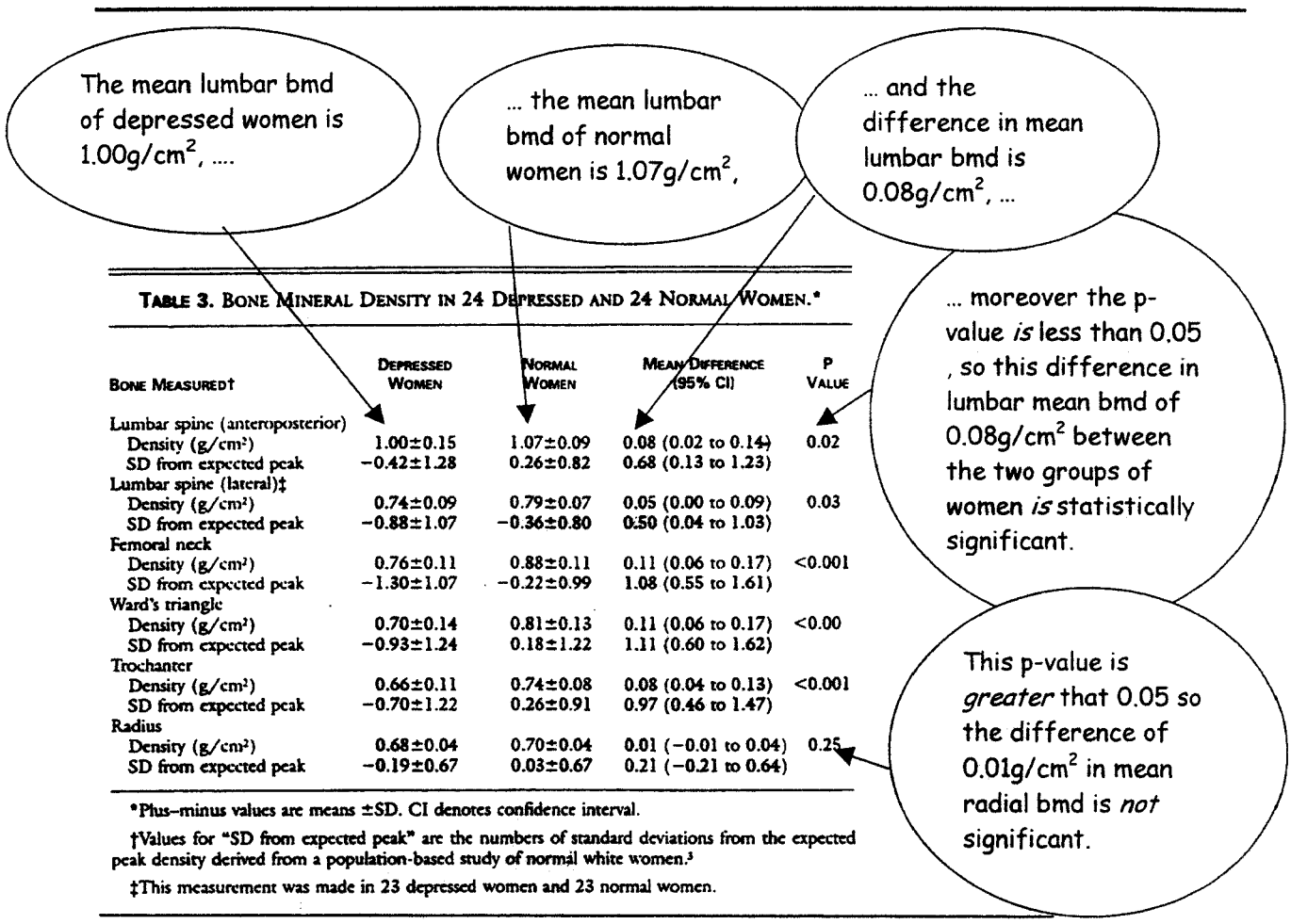

TABLE 3. BONE MINERAL DENSITY IN 24 DEPRESSED AND 24 NORMAL WOMEN.*

BONE MEASURED†	DEPRESSED WOMEN	NORMAL WOMEN	MEAN DIFFERENCE (95% CI)	P VALUE
Lumbar spine (anteroposterior)				
Density (g/cm²)	1.00±0.15	1.07±0.09	0.08 (0.02 to 0.14)	0.02
SD from expected peak	−0.42±1.28	0.26±0.82	0.68 (0.13 to 1.23)	
Lumbar spine (lateral)‡				
Density (g/cm²)	0.74±0.09	0.79±0.07	0.05 (0.00 to 0.09)	0.03
SD from expected peak	−0.88±1.07	−0.36±0.80	0.50 (0.04 to 1.03)	
Femoral neck				
Density (g/cm²)	0.76±0.11	0.88±0.11	0.11 (0.06 to 0.17)	<0.001
SD from expected peak	−1.30±1.07	−0.22±0.99	1.08 (0.55 to 1.61)	
Ward's triangle				
Density (g/cm²)	0.70±0.14	0.81±0.13	0.11 (0.06 to 0.17)	<0.00
SD from expected peak	−0.93±1.24	0.18±1.22	1.11 (0.60 to 1.62)	
Trochanter				
Density (g/cm²)	0.66±0.11	0.74±0.08	0.08 (0.04 to 0.13)	<0.001
SD from expected peak	−0.70±1.22	0.26±0.91	0.97 (0.46 to 1.47)	
Radius				
Density (g/cm²)	0.68±0.04	0.70±0.04	0.01 (−0.01 to 0.04)	0.25
SD from expected peak	−0.19±0.67	0.03±0.67	0.21 (−0.21 to 0.64)	

*Plus–minus values are means ±SD. CI denotes confidence interval.

†Values for "SD from expected peak" are the numbers of standard deviations from the expected peak density derived from a population-based study of normal white women.[3]

‡This measurement was made in 23 depressed women and 23 normal women.

Figure 7.1 Comparison of bone mineral density (bmd) in depressed and normal women

Notice the p-value given for the radius is 0.25, which is greater than the significance level of 0.05, and therefore indicates no strong evidence against the null hypothesis of zero difference. So the confidence interval and the

hypothesis test support the same conclusion of no statistically significant difference. For the other five sites the confidence intervals don't contain zero and the p-values are all less than 0.05, both results indicating significant differences in mean bmd.

We can see why if possible it is preferable to use a confidence interval, if we look at the lumbar spine result in Figure 7.1. The p-value is 0.02 so we know that the is a statistically significant difference in bmd between the two groups and we can reject the null hypothesis of no difference. But that's all it tells us. However, the confidence interval of (0.02 to 0.14) g/cm^2, tells us not only that the difference is significant (because the confidence interval does not include 0), but *in addition* it provides us with the range of values (in clinically meaningful units - g/cm^2 in this example) within which we can reasonably assume the true difference will lie.

ie 2 gps independent

Figure 7.2 is from randomised controlled trial comparing ambulatory versus conventional blood pressure measurement in the management of hypertensive patients. The authors used two tests to compare the characteristics of the two groups, the **t test** and the **chi-squared (χ^2) test**. These two tests are perhaps the most widely used in the literature. We will discuss their appropriate use at the end of this unit.

Table 1.—Baseline Characteristics of Patients Randomized to Antihypertensive Drug Treatment Based on Conventional Blood Pressure (CBP) or Ambulatory Blood Pressure (ABP) Measurements

Characteristics	CBP Group (n=206)	ABP Group (n=213)	P
Age, mean (SD), y	51.3 (11.9)	53.8 (10.8)	.03
Body mass index, mean (SD), kg/m²	28.5 (4.8)	28.2 (4.4)	.39
Women, No. (%)	102 (49.5)	124 (58.2)	.07
Receiving oral contraceptives, No. (%)*	14 (13.7)	10 (8.1)	.17
Receiving hormonal substitution, No. (%)*	19 (18.6)	19 (15.3)	.51
Previous antihypertensive treatment, No. (%)†	134 (65.0)	139 (65.3)	.95
Diuretics, No. (%)*	47 (35.1)	59 (42.4)	.26
β-Blockers, No. (%)*	65 (48.5)	80 (57.6)	.17
Calcium channel blockers, No. (%)*	45 (33.6)	38 (27.3)	.32
Angiotensin-converting enzyme inhibitors, No. (%)*	50 (37.3)	48 (34.5)	.72
Multiple-drug treatment, No. (%)*	62 (46.3)	65 (46.8)	.97
Smokers, No. (%)	42 (20.5)	35 (16.4)	.29
Alcohol use, No. (%)	115 (55.8)	102 (47.9)	.10
Serum creatinine, mean (SD), μmol/L‡	85.75 (15.91)	86.4 (16.80)	.25
Serum total cholesterol, mean (SD), mmol/L‡	6.00 (1.03)	6.10 (1.19)	.32

*Percentages and values of *P* computed considering only women receiving antihypertensive drug treatment before their enrollment.
†Defined as antihypertensive drug treatment within 6 months before the screening visit.
‡Divide creatinine by 88.4 and cholesterol by 0.02586 to convert milligrams per deciliter.

Figure 7.2 Baseline characteristics of patients in blood pressure measurement study. JAMA, 1997, 278, 1065-71.

Q. 7.4 Is there a statistically significant difference in the two groups in, (a) mean age; (b) mean bmi; and (c) % of smokers?

The next example (Figure 7.3) is from a study of the merits of three different prescribing strategies for sore throat. Group 1 were given a prescription for antibiotics for 10 days. Group 2 were not given a prescription. Group 3 were given a prescription for antibiotics if symptoms were not starting to settle down *after* 3 days. Apart from the clinical aspects of the study, patients were also given a questionnaire and asked to score, "very", "moderately", "slightly", or "not at all", to a number of questions. The differences between the groups in the percentage scoring ("very" or "moderately") as compared to ("slightly" or "not at all") to the various questions, were tested using the chi-squared test (χ^2) and the results are shown in Figure 7.3.

96% in Group 1 answered "very" or "moderately" to consultation question,

... 90% in Group 2 answered "very" or "moderately" to consultation question, ...

... and 93% in Group 3 answered "very" or "moderately" to consultation question ....

Table 4 Satisfaction, belief, and intention of patients consulting doctor for sore throat. Values are numbers (percentages) scoring "very" or "moderate" on Likert scale

	Group 1 (antibiotics)	Group 2 (no antibiotics)	Group 3 (antibiotic offer delayed)	χ^2 test	P value
Satisfaction with consultation	202/211 (96)	166/184 (90)	165/177 (93)	4.7	0.09
Doctor dealt with worries	201/211 (95)	165/184 (90)	164/177 (93)	4.5	0.1
Likely to see doctor if sore throat recurs	148/187 (79)	87/162 (54)	92/162 (57)	27	0.001
Antibiotics are effective	181/207 (87)	95/173 (55)	99/165 (60)	55	0.001
Legitimation of illness					
Work or school*	128/209 (61)	117/184 (64)	96/177 (54)	3.56	0.17
Family or friends†	75/210 (36)	69/183 (38)	67/176 (38)	0.27	0.9

*Importance of seeing doctor to be able to take time off work or school.
†Importance of seeing doctor to be able to explain illness to family, friends, or acquaintances.

.... but these differences are *not* statistically significant, since the p-value for the χ^2 test is *not* less than 0.05. All three groups were equally satisfied.

Figure 7.3 Comparison of patient satisfaction, beliefs and intentions, in three different prescribing strategies for sore throat. BMJ, 1997, 314.

The figure shows, for example, that the sample percentages answering ("very" or "moderately") to the question, "Were you satisfied with your consultation?" is 96% in Group 1, 90% in Group 2, and 93% in Group 3. The null hypothesis is that

these %s are equal, and although the percentages do seem very similar, we would need the chi-squared test to confirm that the observed differences are statistically significant.

For the question, "Were you satisfied with your consultant", the p-value for the chi-squared test is given in the table as 0.09, which is not less than 0.05, so does not offer sufficiently strong evidence to enable us to reject the null hypothesis. There appears to be an equal degree of satisfaction with the consultation in all three groups.

Note that the p-value of 0.09 implies that we can be 91% certain (1.00 minus 0.09 = 0.91, or 91%) that there is a difference between the percentages in each group who are satisfied with their consultation. However, since we have set a benchmark p-value of 0.05 (which 0.09 exceeds) we cannot conclude that these differences are statistically significant. They may be due to chance alone.

Q. 7.5 What do the p-values in Figure 7.3 indicate about the differences in the percentages of patients across the three groups who: (a) believed that the doctor dealt with their worries; (b) would be likely to see the doctor again if sore throat re-occurs; (c) believed antibiotics to be effective; (d) felt that their illness had been legitimised either at work or school and/or among family and friends?

In the following study we see a different and unwelcome way of expressing p-values. Figure 7.4 is from a study comparing a community-based (in Sparkbrook), and a hospital-based (in Small Heath), service for patients suffering acute severe psychiatric illness in Birmingham. The two groups were independent. Included in the measurements was the level of distress experienced by the patients' relatives, as indicated by their score on the Social Behaviour Assessment Schedule (SBAS) which produces ordinal data.

The authors provide the results of measures on two components of the SBAS, distress due to the objective burden (caring for the patient), and that due to social performance (of the patient). They wished to compare the *median* SBAS scores for the relatives of patients in Sparkbrook with those in Small Heath. These two groups are independent, but since this data is ordinal the authors could not use the two-sample t test to compare means (which can only be used with metric data), but instead used the Mann-Whitney test (suited for comparing the medians of ordinal data from two independent groups). (See Table 7.1 below).

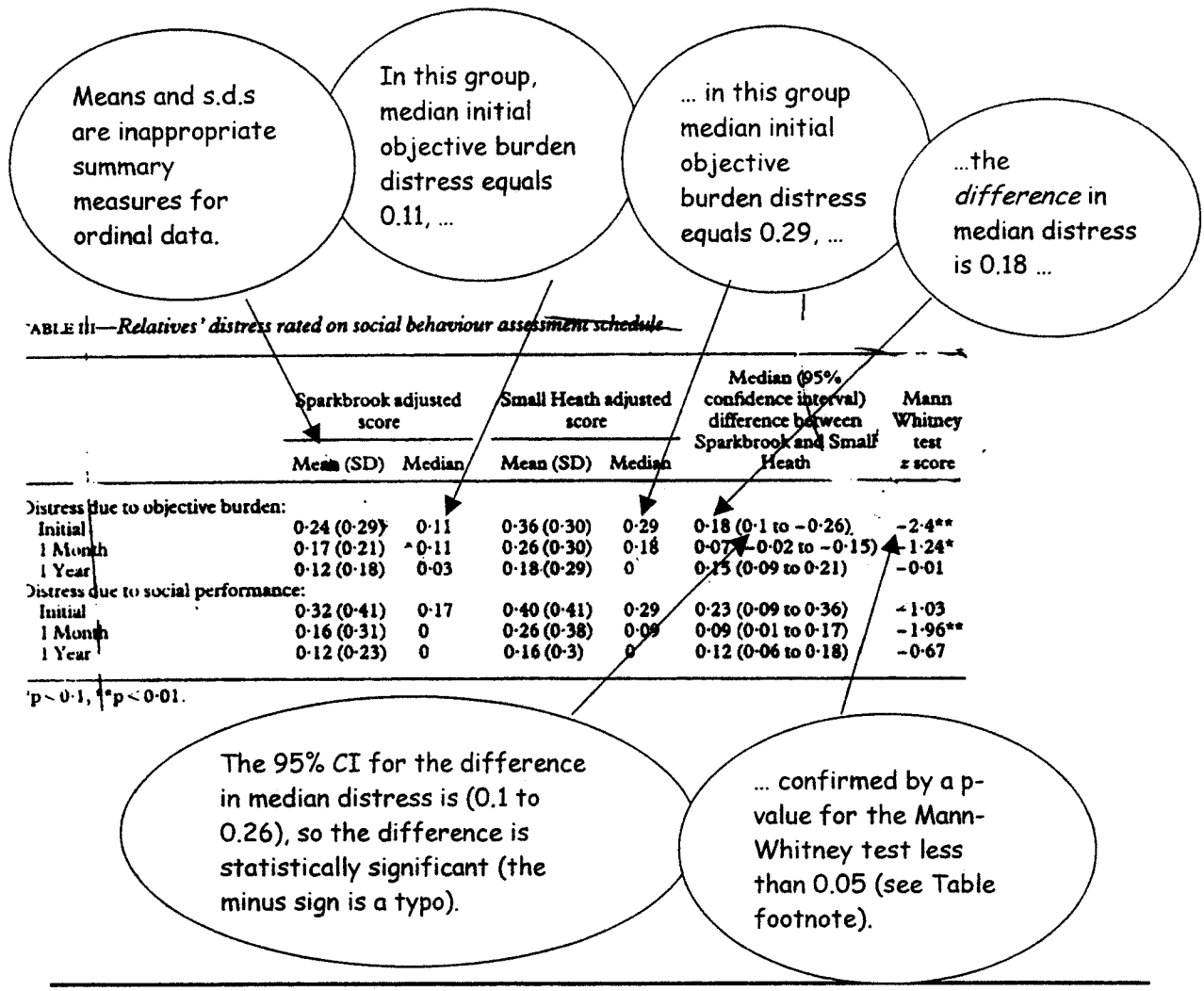

Figure 7.4 Distress scores in two groups of patients with acute mental illness. BMJ, 317, 1998.

Notice particularly that instead of giving the actual p-values in the table, the authors have chosen to provide the value of the z statistic* for each drug, along with a varying number of asterisks, depending on whether the corresponding p-value is less than 0.1 or 0.001. This is not good practice, the individual p-values are more informative and should be given.

Incidentally having provided confidence intervals there was no need to also give p-values. The authors have also calculated the means and s.d.s for this ordinal data which is inappropriate (see UNITS 1, 3 and 4 for a discussion on data types and appropriate summary measures). The only two significant difference in

* The z statistic derives from the Normal distribution and can be used with the Mann-Whitney test provided both groups are at least 10 in number. As a rule of thumb, z has to be larger than |2| to be significant.

distress are for initial objective burden, and social performance at 1 month, with both p-values less than 0.05 (in fact less than 0.01). The identification of objective burden at one month as significant (*) with a p-value less than 0.1 is not common practice.

There are dozens, perhaps hundreds, of statistical tests which at some time may be used by clinical researchers. Some of the most common are listed in Table 7.1. As you will see the choice of an appropriate test depends not only on data type and distributional shape but also on whether the groups are **independent** or **matched** (also known as paired).

Independent versus matched groups

Suppose, as part of an occupational health study, you wanted to compare the systolic blood pressures of male and female police officers employed in a particular police force. You could take a random sample from a list of female officers and a random sample from a list of male officers and measure the systolic blood pressure of each individual officer in each group. These two sample groups would be **independent** since the selection of female officers is *not* dependent on or influenced by the selection of male officers (and vice versa). The two selection processes are separate exercises, and could in fact be done by two different statisticians with no connection or reference to each other. The two groups don't even have to be the same size, although of course we would want them both to be large enough to ensure a reasonable representation of their respective populations.

On the other hand, suppose we suspect that the sample of female officers is going to be on the small side (and possibly not therefore completely representative). We might then want to make sure that the two groups of officers are similar in some critical ways, i.e. they have the same body mass index, the same age, the same years on the force, the same rank, etc. To achieve this we could, for *each* individual female officer selected, find a male officer of the same age, same body mass index, same length of service, rank, etc., and then measure their systolic blood pressures. These two groups are then said to be **individually matched** or **paired**. They necessarily have to be of the same size.

One common manifestation of matched groups is in *before and after studies*, where some characteristic of an individual is measure before and after some intervention. We thus have two groups of measurements, before and after.

These are of course necessarily matched because the *same* individual is measured twice.

Q. 7.6 With the help of Table 7.1 decide which is the most appropriate hypothesis test to determine: (a) in Figure 7.2, whether the %s receiving each of the five types of previous anti-hypertensive treatment is the same in both groups. (b) In Figure 7.2, whether the average level of serum creatinine is the same in both groups. (c) In Figure 7.2, whether the % of smokers is the same in both groups. (d) In Figure 7.1, whether the mean bone mineral density is the same in both groups at each site (the authors describe the groups as "individually matched"). State any assumptions needed in each case.

---

the two-sample t test: most often used to test for difference in means of two *independent* groups - *both sets of data must be metric and Normally distributed.*

the matched-pairs t test: most often used to test for difference in means of two *matched* groups - *data must be metric and differences between group scores Normally distributed*

the Mann-Whitney test: most often used to test difference in medians of two *independent* groups - *data can be either ordinal or metric with any shape distribution*

the Wilcoxon test: most often used to test difference in medians of two *matched* groups - *data can be either ordinal or metric with any shape distribution*

the chi-squared test: most often used to measure differences in proportions (or %s) for *independent* groups across categories - *data can be either nominal, ordinal, discreet metric (with only a few possible values), or grouped metric continuous (with only a few groups)*[*]

the McNemar test: most often used to measure the difference in proportions (or %s) of two *matched* groups across two categories - *data can be nominal or ordinal*

Table 7.1 Some of the more common hypothesis test and their uses

---

[*] When sample sizes are small, the chi-squared test may not be entirely appropriate. In a two-group two-category case Fisher's test may be used instead. The same data conditions apply.

Unit 7 Hypothesis testing

Solutions to questions

Q. 7.1 (a) Does ginkgo biloba improve cognitive functioning in elderly subjects?

(b) H_0: ginkgo biloba does not improve cognitive functioning in the elderly
H_1: ginkgo biloba does improve cognitive functioning in the elderly

Note: H_0 and H_1 conventionally denote the null and alternate hypotheses respectively.

Q. 7.2 (a) significance level of 0.05: (i) reject - since $0.03 < 0.05$; (ii) do not reject - since $0.50 \geq 0.05$; (iii) reject - since $0.009 < 0.05$. (b) significance level of 0.01: (i) do not reject - since $0.03 \geq 0.01$; (ii) do not reject - since $0.50 \geq 0.01$; (iii) reject - since $0.009 < 0.01$.

Bear in mind that an improvement of 5 points in a 50 point scale, although statistically significant, may not be *clinically* large enough to justify clinical intervention. Just because something is statistically significant does not mean it is also clinically significant.

Q. 7.3 In Q.7.2(a)(i) significance level $\alpha = 0.05$, so if the p-value is read as 0.30 the researcher does not reject the null hypothesis when he should have. This is a Type II error.

Q. 7.4 (a) Since $0.03 < 0.05$, the difference in the true mean ages of the two groups *is* statistically significant. (b) Since $0.39 \geq 0.05$, the difference in mean bmi is *not* statistically significant. (c) Since $0.07 \geq 0.05$, the difference in the proportions of women in the two groups is *not* statistically significant. (d) Since $0.29 \geq 0.05$, the difference in the % of smokers is *not* statistically significant.

Q. 7.5 (a) Since $0.10 \geq 0.05$ there is *no* statistically significant difference across the three groups (given antibiotics, not given antibiotics, and delayed antibiotics) in the % of patients who believed that the doctor dealt with their worries. (b) The groups did differ however in the true percentages who were likely to see their doctor if their sore throat re-occurs, p-value of $0.001 < 0.05$. In other words, we can be 99.9% certain (1.000 minus 0.001 = 0.999, or 99.9%) that the differences in the true percentage of patients in each of the three groups are real. (c) There was also a statistically significant difference across the groups in the %s who believed that antibiotics are effective ($0.001 < 0.05$).

(d) For both of the outcomes - legitimisation of illness at work or school and with family or friends – there is no statistically significant difference in the %s across the three treatment groups (0.17 and 0.90 both ≥ 0.05).

Note that although the chi-squared test may indicate that the three %s are not all equal, it does not by itself identify which are different. Inspection of the contingency table will usually suggest a candidate.

Q. 7.6 (a) The chi-squared test. Type of treatment is nominal (- order is arbitrary). The two groups are independent (they were randomised). We want to compare %s in each of the five categories between two or more groups.

(b) The two-sample t test. Serum creatinine is metric. The two groups are independent. We don't know whether both variables are Normally distributed but both sample sizes are quite large, and the t test is quite resilient against departures from Normality, particularly with large samples. Finally we want to compare means. Note that if we were unconvinced by the Normality argument, we might prefer the Mann-Whitney test - but this only compares medians.

(c) The chi-square test. Same arguments as in (a).

(d) The matched-pairs t test. Bone mineral density is metric. The groups are matched. We don't know whether the differences between the group measurements is Normal but total sample size is quite large, and the t test is quite resilient against departures from Normality, particularly with large samples. Finally we want to compare means. Note that if we were unconvinced by the Normality argument, we might prefer the Wilcoxon test - but this only compares medians.