9

Statistical Significance

Objectives Covered

24. Interpret statements of statistical significance with regard to comparisons of means and frequencies, explain what is meant by a statement such as $P < 0.05$ and distinguish between the statistical significance of a result and its importance in clinical application.
25. Explain the following regarding statistical tests of significance: the power of a test, the relationship between significance tests and confidence intervals, one versus two-tailed tests, and comparison-wise versus study-wise significance levels.

Study Notes

Interpretation of Comparison Results

The term *statistically significant* is often encountered in scientific literature, and yet its meaning is still widely misunderstood. The determination of statistical significance is made by the application of a procedure called a statistical test. Such procedures are useful for interpreting comparison results. For example, suppose that a clinician finds that in a small series of patients the mean response to treatment is greater for drug A than for drug B. Obviously the clinician would like to know if the observed difference in this small series of patients will hold up for a population of such patients. In other words he wants to know whether the observed difference is more than merely "sampling error." This assessment can be made with a statistical test.

To understand better what is meant by statistical significance, let us consider the three possible reasons for the observed drug A versus drug B difference:

1. Drug A actually could be superior to drug B.
2. Some confounding factor that has not been controlled in any way, for example, age of the patients, may account for the difference. (In this case we would have a biased comparison.)
3. Random variation in response may account for the difference.

Only after reasons 2 and 3 have been ruled out as possibilities can we conclude that drug A is superior to drug B. To rule out reason 2, we need a study design that does not permit any extraneous factors to bias the comparison, or else we must deal with the bias statistically, as for example by age-adjustment of rates. To rule out reason 3, we test for statistical significance. If the test shows that the observed difference is too large to be explained by random variation (chance) alone, we state that the difference is statistically significant and thus conclude that drug A is superior to drug B.

Significance Tests

Underlying all statistical tests is a *null hypothesis.* For tests involving the comparison of two or more groups, the null hypothesis states that there is no difference in population parameters among the groups being compared. In other words, the null hypothesis is consistent with the notion that the observed difference is simply the result of random variation in the data. To decide whether the null hypothesis is to be accepted or rejected, a test statistic is computed and compared with a *critical value* obtained from a set of statistical tables. When the test statistic exceeds the critical value, the null hypothesis is rejected, and the difference is declared statistically significant.

Any decision to reject the null hypothesis carries with it a certain risk of being wrong. This risk is called the significance level of the test. If we test at the 5% significance level, we are taking a 5% chance of rejecting the null hypothesis when it is true. Naturally we want the significance level of the test to be small. The 5% significance level is very often used for statistical tests. A statement such as "The difference is statistically significant at the 5% level" means that the null hypothesis was rejected at the 5% significance level.

The *P* Value

Many times the investigator will report the lowest significance level at which the null hypothesis could be rejected. This level is called the *P* value. The *P* value therefore expresses the probability that a difference as large as that observed would occur by chance alone. If we see the statement $P < 0.01$, this means the probability that random variation alone accounts for the difference is very small, and we are willing to say the result is statistically significant. On the other hand, the statement $P > 0.10$ implies that chance alone is a viable explanation for the observed

difference, and therefore the difference would be referred to as not statistically significant. Although arbitrary, the *P* value of 0.05 is almost universally regarded as the cutoff level for statistical significance. It should be taken only as a guideline, however, because, with regard to statistical significance, a result with a *P* value of 0.051 is almost the same as one with a *P* value of 0.049.

Commonly Used Tests

The type of data involved determines the specific procedure used to perform the significance test. When the individual observations are categorical (e.g., improved/not improved, smoked/did not smoke) and are summarized in frequency tables, the chi-square test is used. The chi-square test statistic (see the first exercise in this chapter) indicates how well the observed frequencies match those that are expected when the null hypothesis is true. When the observed frequencies are identical to the expected frequencies, the chi-square statistic has a value of 0, and the corresponding *P* value is 1. The more the observed frequencies differ from the expected, the larger the value of the chi-square statistic and the smaller the *P* value (hence, the more we doubt the null hypothesis). A chi-square table can be used to determine the *P* value from the chi-square statistic. Every chi-square statistic has an associated parameter, called the degrees of freedom, that is needed to find the *P* value from the table. Although the expected frequencies are derived from the null hypothesis, their total must equal the total observed frequency. This constrains the number of expected frequencies that can be ascertained independently, a number called the degrees of freedom. For example, suppose that a certain rare birth disorder has been reported so far in six cases, all male infants. The null hypothesis to be tested is that there is no association between the disorder and the sex of the infant. The null hypothesis thus predicts expected frequencies of three males and three females. Note that only one of the expected frequencies can be determined independently because their total must be six. Thus, the chi-square statistic for this test has one degree of freedom. Where 2×2 or larger dimensioned frequency tables are involved, the product

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

gives the degrees of freedom.

When the individual observations are measurements, such as weight or blood pressure, the primary focus for a two-group comparison is usually on the difference in means. Here, the *t* statistic is used to test the null hypothesis of no difference. The *t* statistic is determined as the difference in the means for the two groups divided by the standard error of this difference. Again, the farther the *t* statistic departs from 0, the smaller the *P* value becomes. A *t* table can be used to establish *P* from the value of *t* and its degrees of freedom. The degrees of freedom for the *t* statistic is given by the sum of the group sample sizes minus 2.

Sample Size and the Interpretation of Nonsignificance

A statistically significant difference is one that cannot be accounted for by chance alone. The converse is not true; that is, a nonsignificant difference is not necessarily attributable to chance alone. In the case of a nonsignificant difference, the sample size is very important. This is because, with a small sample, the sampling error is likely to be large, and this often leads to a nonsignificant test even when the observed difference is caused by a real effect. In any given instance, however, there is no way to determine whether a nonsignificant difference derives from the small sample size or because the null hypothesis is correct. It is for this reason that a result that is not statistically significant should almost always be regarded as inconclusive rather than an indication of no effect.

Sample size is an important aspect of study design. The investigator should consider how large the sample must be so that a real effect of important magnitude will not be missed because of sampling error. (Sample size determination for two-group comparisons is discussed by Bland [see Recommended Readings].)

Clinical Significance vs. Statistical Significance

It is important to remember that a label of statistical significance does not necessarily mean that the difference is significant from the clinician's point of view. With large samples, very small differences that have little or no clinical importance may turn out to be statistically significant. The practical implications of any finding must be judged on other than statistical grounds.

Power

Rejecting the null hypothesis when it is true is referred to as a type I error. Conversely, accepting the null hypothesis when it is false is a type II error. The type I error could be equated to a false positive in the context of diagnostic testing, and the type II error to a false negative. The significance level of a statistical test is the probability of making a type I error. If we think of statistical testing as analogous to screening for a particular disease with the null hypothesis that the disease is absent, then 1 minus the significance level corresponds to the specificity of the screening test. One minus the probability of a type II error is analogous to the sensitivity of the screening test. For statistical testing, this probability is called the power of the test.

Just as the sensitivity of a screening test indicates the likelihood of detecting a disease when it is present, the power of a statistical test indicates the likelihood of detecting a departure from the null hypothesis when such a departure exists. Once the significance level is set (usually at 5%), then the risk of making a type I error is established at that one specific value. However, with conventional statistical

tests, the risk of making a type II error, and thus the power, has an infinite number of possible values. This is because, in theory, there is a continuous range of possible departures from the null hypothesis. Suppose, for example, two antihypertensive drugs were being compared on their ability to reduce blood pressure, the null hypothesis would be that there was no difference in mean reductions while the departures from the null hypothesis would include innumerable possibilities (e.g., 1 mm Hg, 2, 5, 10, etc.). A specific departure is called the effect size, and the value of the power for a test increases with increases in the effect size.

The determination of power is beyond the scope of this text and, in truth, seldom is that calculation done after a study is completed. Most often the power calculation is done when a study is being designed. More specifically, the power is used to establish the adequacy of the sample size being considered for the study. Returning to the example in the previous paragraph, we shall consider it important to know if one drug provides a reduction of 10 mm Hg more than the other, meaning, the effect size we wish to detect is 10. The study is proposed to have 20 subjects on each drug. For a statistical test with a 5% significance level, the power to detect this effect size would be 53%. (The calculation also assumed that the between-subject standard deviation for reductions was 15 mm Hg.) This indicates that the study will have only about an even chance of demonstrating statistical significance when the effect size we wish to detect actually exists. It is usually desirable to have a study with at least 80% power. Recall from the previous discussion of nonsignificant differences that they are often caused by insufficient sample sizes. The reason is that power is primarily determined by sample size once an effect size is specified. If in the proposed study, the sample size were increased to 50 per drug group, the power becomes 90%.

Confidence Intervals on Effect Sizes

Suppose the study described in the preceding paragraph for the comparison of two antihypertensive drugs was done with 50 subjects per group and the result was that one drug provided a mean reduction of 12 mm Hg versus 10 mm Hg for the other drug. The difference in means of 2 mm Hg is the sample estimate of the effect size for one drug relative to the other. Again assuming the between-subject standard deviation for individual reductions was 15 mm Hg in each group, the standard error of the estimated effect size can be determined to be 3.0 mm Hg (see the exercise on labile hypertension at the end of this chapter for an example of this calculation). The 95% confidence interval on the effect size is –4 to 8 mm Hg. Since 0 (the null hypothesis value of the effect size) is within the interval, the result is not significant at the 5% level. The confidence interval thus provides a method for doing a significance test at a particular level and, in addition, gives an indication of the limits that can be put on the effect size. For this example, it is reasonable to conclude that one drug does not have an efficacy advantage of more than 8 mm Hg reduction in blood pressure over the other drug.

One- vs. Two-Tailed Tests

It is almost always the case that departures from the null hypothesis, in either direction (e.g., drug A is more efficacious than drug B or vice-versa), are of interest to detect. To keep the false-positive rate at 5%, the significance level has two equal components of 2.5% each to account for random variations in both directions. Tests that divide the significance level in this way are referred to as two-tailed.

Suppose, however, that drug A were a less-costly formulation of the conventional drug, that being drug B. It is reasonable to believe that drug A would be favored for use over drug B unless drug B proved to be more efficacious. Here the main interest would focus on a departure from the null hypothesis in one direction only. Now the significance level need only have one component of 5% to account for random departures in drug B's favor (false positives implying B is more efficacious). Such a test is called a one-tailed test.

A one-tailed test is always more powerful than a two-tailed test at the same significance level. The *P* value for a one-tailed test is half that of a two-tailed test. It is therefore sometimes possible to claim "statistical significance" for a result with a one-tailed test whereas such a claim could not be made for the same result with a two-tailed test. This would occur for any two-tailed test that yields $0.05 < P \leq 0.10$. A one-tailed test should be justified with reasoning similar to that given in the previous paragraph prior to any examination of the data. Any report including one-tailed *P* values needs to clearly identify them and justify their use.

Multiplicity of Significance Testing

Returning again to the example of the comparison of two antihypertensive drugs, consider now the evaluation of two different outcomes, such as blood pressure reduction and incident side effects. Once the data collection is completed, two tests of statistical significance could be performed, one regarding the difference in blood pressure reduction means and the other regarding the difference in side effect incidence rates. If both tests are performed at the 5% significance level, the risk of a type I error is 5% for each comparison. In this case, these 5% risks would be referred to as the comparison-wise significance levels. Because of the two outcomes involved, there is also a study-wise significance level, that is, the risk of making a type I error in one or both tests.

The study-wise significance level will always be larger than the comparison-wise levels. This can be shown by use of the addition rule of probability. For the example of two independent outcomes, Pr (type I error for the first outcome or type I error for the second outcome) = $0.05 + 0.05 - 0.05 \times 0.05 = 0.0975$. It can be shown, that when there is some degree of dependence of one outcome on the other, the study-wise significance level will be between 0.05 and 0.0975.

The study-wise significance level increases as the number of comparisons increases. It is possible to control the study-wise risk of a type I error at a low level, such as 0.05, by reducing the comparison-wise significance levels. This comes at the cost of lower power for the comparison-wise tests. The increased study-wise risk of a type I error with multiple significance testing arises in several different ways, the most common being: (1) with multiple outcomes as described above, (2) with multiple comparisons on the same outcome (e.g., when there are more than two treatment groups in the same study), and (3) with repeated tests on accumulating data for a given outcome at various stages of the completion of a study.

A number of statistical methods now exist to control the study-wise type I error rate at a specified level, typically 5%. The simplest method is the application of Bonferoni's rule. This rule determines the significance level used for each of the comparison-wise tests by dividing the desired study-wise significance level by the number of comparisons. To maintain a 5% study-wise significance level for the example of two outcomes, this would require that $P < 0.025$ to claim statistical significance for either outcome-specific test.

The loss of power that results from any multiplicity correction, such as Bonferoni's rule, is a serious disadvantage associated with these methods. Furthermore, the role of the study-wise significance level in scientific inquiry has often been questioned and remains highly controversial. Even so, the reader should keep in mind the problem of multiplicity when considering a report containing a multitude of *P* values. Remember, that if $P < 0.05$ is used as the criterion for statistical significance, with 100 tests at least 5 are expected to show significance even if no true effects are involved whatsoever.

Exercises

Proportionate Mortality Among Polyvinyl Chloride Workers

In February 1974, four fatal cases of cancer of the liver among men who worked in a polyvinyl chloride polymerization plant were reported (Monson, Peters, & Johnson, 1974). A proportionate mortality analysis of all deaths from 1947 to 1974 among workers in that plant is shown in Table 9–1.

To determine whether the excess of cancer deaths could be attributed to chance alone, we do a chi-square (χ^2) test. First, Table 9–1 is reduced to the form shown in Table 9–2.

Then the chi-square statistic is computed as

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2 = \frac{(41 - 27.9)^2}{27.9} + \frac{(120 - 133.1)^2}{133.1} = 7.44$$

Table 9–1 Observed and Expected Deaths in Polyvinyl Chloride Workers

Cause of Death	*Observed*	*Expected*	*Obs./Exp.*
All	161	161.0	1.0
All cancer	41	27.9	1.5
Digestive	13	8.3	1.6
Liver and biliary tract	8	0.7	11.0
Lung	13	7.9	1.6
Brain	5	1.2	4.2
Lymphatic and hemopoietic	5	3.4	1.5
Other cancer	5	7.1	0.7
Central nervous system/vascular	8	9.5	0.8
Circulatory	66	68.6	1.0
External	22	24.3	0.9
Suicide	10	5.3	1.9
All other cases	24	30.5	0.8

Note: Expected numbers based on age/time/cause-specific proportionate mortality ratios for U.S. white males.

Table 9–2 Reduced Version of Table 9–1

Cause of Death	*Observed*	*Expected*
Cancer	41	27.9
All other	120	133.1

This chi-square value has one degree of freedom (*df*), since only one of the expected numbers can be determined independently of the total number of deaths. To get the *P* value we now refer to a chi-square table. Table 9–3 is an abbreviated version of such a table, and from it we see that our computed value of chi-square exceeds that for $P = 0.01$ (for 1 *df*). We can thus report $P < 0.01$. Note that if the

Table 9–3 Abbreviated Table of Chi-Square Corresponding to Selected Values of *P*

df	*0.50*	*0.10*	*0.05*	*0.02*	*0.01*
1	0.45	2.71	3.84	5.41	6.63
2	1.39	4.61	5.99	7.82	9.21
3	2.37	6.25	7.82	9.84	11.34
4	3.36	7.78	9.49	11.67	13.28

computed chi-square value had been 6.63, we could report P as precisely 0.01. Such instances are unlikely, however, and if a chi-square table is used to determine P, it is usually sufficient to describe a range for it. Modern computer software is capable of translating any chi-square value into a precise P value, and the use of computers for significance testing has lead to the practice of reporting P values precisely instead of in ranges.

1. **Is the excess of cancer deaths statistically significant? Why?**
2. **The chi-square value for central nervous system vascular diseases is 0.04 (1 *df*). Use Table 9–2 to report a *P* value. What does the chi-square value tell you about the discrepancy of deaths in this category from the expected number?**
3. **What are the major difficulties with proportionate mortality analysis as a means of revealing the carcinogenic potential of polyvinyl chloride?**

Oral Contraceptives and Birth Defects

Exposure to exogenous sex steroids during pregnancy was investigated for 108 mothers of children with congenital limb-reduction defects and 108 mothers of normal controls (Janerich, Piper, & Glebatis, 1974). Unintentional use of oral contraceptives early in pregnancy was the primary source of exposure. Fifteen of the mothers of the affected children were found to have been exposed, whereas only four of the controls were exposed.

4. **Show, in the form of a 2×2 table, the results of this study.**
5. **The chi-square value for the comparisons of rates of exposure among cases and controls was 5.77 (1 *df*). (See recommended readings for computation of χ^2 from a 2×2 table.) Use Table 9–3 to find the corresponding *P* value. What is your interpretation of the finding?**

Labile Hypertension

In Exercise 7, Chapter 8, labile hypertensives were compared with normotensives regarding mean weight, mean heart rate, and proportion with a family history of hypertension (see Table 8–2). While a comparison of the 95% confidence intervals for means and proportions provides a rough assessment of whether sampling error might account for the differences, a better approach is to do significance tests. For example, the comparison of mean weights calls for a t test. The choice of a specific procedure for doing the t test is somewhat complicated by the assumptions we are willing to make about the standard deviations in the groups being compared (see the recommended readings for details). Unless the sample sizes are small (say, < 30 in either group), however, the procedure described below provides a reasonably accurate determination of whether $P < 0.05$.

Using the comparison of mean weights in Table 8–2 as an example, the first step is to determine the standard error of the difference in means as

$$\text{SED} = \sqrt{(\text{SE1}^2 + \text{SE2}^2)},$$

where, SE1 and SE2 are the standard errors of the mean in the two groups. For the example,

$$\text{SED} = \sqrt{\left((1.5)^2 + (1.9)^2\right)} = 2.42$$

Next, calculate the *t* statistic as the difference in the means for the two groups divided by SED,

$$t = (70.6 - 81.3)/2.42 = -4.42$$

When the absolute value of *t* is greater than or equal to 2 (as it is for the example), $P < 0.05$. In this case we would conclude that random sampling error does not account for the difference in means.

Having established that the difference in mean weight is statistically significant, we might be further interested in assessing the magnitude of the difference. The mean weight for the hypertensive group is 10.7 kg more than that for normotensive group, but this result alone does not take random sampling error into account. A 95% confidence interval would be helpful in this regard. The 95% confidence interval on the difference is $10.7 \pm 2(2.42) = 5.9, 15.5$ kg. Thus we conclude that the difference is probably at least 5.9 kg and could be as much as 15.5 kg. It should also be noted that the 95% confidence interval by itself is sufficient to establish the statistical significance of the difference. To do this we have only to determine whether the null hypothesized difference (zero) is included in the interval. If the interval excludes zero (as it does for this example), then $P < 0.05$.

6. **Do a *t* test to establish whether the difference in means for heart rate in Table 8–2 is statistically significant.**
7. **The chi-square test is more appropriate than the *t* test for the comparison of two proportions. Why? For the comparison of family history of hypertension in Table 8–2, use the data given to determine the corresponding 2×2 frequency table. This table generates a chi-square value of 2.69 (1 *df*). Find the *P* value. What do you conclude from it?**

Low-Tar/Nicotine Cigarettes

Cigarette consumption was studied (Turner, Sillett, & Ball, 1974) in 10 volunteers smoking cigarettes of progressively lower tar/nicotine content during three

consecutive periods of 1 week each. The subjects recorded on diary cards the number of cigarettes smoked daily. Approximately 30 cigarette butts were collected from each subject during each period. The mean consumption and butt length findings are given in Table 9–4.

Table 9–4 Cigarette Consumption and Butt Length (Means ± 2 Standard Errors) According to Tar/Nicotine Content

	Tar/Nicotine Content		
	Medium	*Low*	*Very Low*
Mean number of cigarettes consumed daily	25.7 ± 6.50	30.9 ± 8.30	29.2 ± 6.20
Mean butt length (mm)	8.84 ± 2.96	7.20 ± 2.82	4.54 ± 2.22

The following remarks are given in the paper:

> When changing from medium to low brands, nine subjects increased their consumption and one reduced slightly, mean consumption rising from 25.7 to 30.9 ($P < 0.01$). There was no significant change in consumption from low to very low.
>
> During the medium period the mean butt lengths were 8.84 mm, in the low 7.20 mm, and in the very low 4.54 mm. The difference between the low and very-low brands was statistically significant ($P < 0.01$).

8. **Did the subjects alter their smoking habits when changing to lower tar/nicotine cigarettes? How? Is there evidence for the assertion that lower tar/nicotine cigarettes cause smokers to consume more tobacco?**
9. **In the study the volunteers were informed of the tar/nicotine content of the cigarettes used during each period. What problems does this introduce? How could the study be done to avoid such problems?**

Propranolol Treatment in Parkinson's Disease

Propranolol was compared with a placebo in 18 patients with Parkinson's disease who had been taking stable doses of levodopa for 3 months or more but who still had tremor (Marsden, Parker, & Rees, 1974). Each patient was given propranolol for a 4-week period and a placebo for a similar period but was not aware of the identity of this treatment plan. A physician, who was also unaware of the treatment plan, scored each patient for total disability, tremor, rigidity, akinesia, posture, handwriting, and circle drawing. Results are given in Table 9–5.

Table 9–5 Effects of Propranolol (120 mg Daily) versus Placebo in 18 Patients with Parkinson's Disease on Levodopa

	*Start Scores**	*Placebo*	*Propranolol*	*Significance†*
Total disability	27.80	25.70	27.60	N.S.
Tremor	2.67	2.86	2.19	N.S.
Rigidity	4.25	2.92	2.94	N.S.
Akinesia	6.58	6.06	6.75	N.S.
Posture	3.53	4.11	3.86	N.S.
Writing	1.58	1.56	1.28	$P < 0.02$
Circle drawing	1.81	1.94	1.36	$P < 0.02$

*A high score indicates severe disability.
†Propranolol compared with placebo. N.S., not significant.

10. What were the apparent benefits of the propranolol treatment?

11. The investigators concluded that the changes that were noted were "not of clinical value and none of these patients have been maintained on propranolol." In view of the fact that certain of the findings were statistically significant, what kinds of considerations would lead the investigators to this conclusion?

References

Janerich, D. T., Piper, J. M., & Glebatis, D. M. (1974). Oral contraceptives and congenital limb-reduction defects. *New England Journal of Medicine, 291,* 697.

Marsden, D. C., Parker, J. D., & Rees, J. E. (1974). Propranolol in Parkinson's disease. Letter to the editor. *Lancet, 2,* 410.

Monson, R. R., Peters, J. M., & Johnson, M. N. (1974). Proportional mortality among vinylchloride workers. *Lancet 2,* 397.

Turner, J. A. M., Sillett, R. W., & Ball, K. P. (1974). Some effects of changing to low-tar and low-nicotine cigarettes. *Lancet 2,* 737.

Recommended Readings

Bland, M. (2000). *An introduction to medical statistics* (3rd ed.). Oxford, UK: Oxford Medical Publications. Chapter 9 gives a good introduction to the principles of significance tests; Chapter 10 describes procedures for t tests; and Chapter 13 describes procedures for chi-square tests.

Ingelfinger, J. A., Mosteller, R., Thibodeau, L. A., & Ware, J. H. (1994). *Biostatistics in clinical medicine* (3rd ed.). New York: McGraw-Hill. Chapter 7 emphasizes the interpretation of *P* values with many clinical examples.