Review: Frequentist inferences for means and proportions

Here we review some of the basic material from EPIB-607 (or equivalent). In particular, we will cover frequentist hypothesis testing and interval estimation for means and proportions, while reminding ourselves of the exact meanings and interpretations of p-values and confidence intervals, both in theory and in real practice.

For single means and single proportions, we have the analogy below between means and proportions, and a summary table for all possible situations for means follows on the next page.

	MEANS	PROPORTIONS
DATA	$\{x_1, x_2, \ldots, x_n\}$	$\{x_1, x_2, \ldots, x_n\} = \{0, 1, \ldots, 1\}$
ESTIMATOR	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\#\text{ of 1's}}{n}$
SD	$sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$	$sd = \sqrt{\hat{p}(1-\hat{p})}$
CI	$\overline{x} \pm 1.96 \times \frac{sd}{\sqrt{n}}$	$\hat{p} \pm 1.96 \times \frac{sd}{\sqrt{n}}$
TEST	$Z = \frac{\overline{x} - \mu_0}{\frac{sd}{\sqrt{n}}}$	$Z = \frac{\hat{p} - p_0}{\frac{sd}{\sqrt{n}}}$

1 or 2 sample	$\sigma_1 = \sigma_2$?	σ's known?	σ estimate	test	CI
1	N/A	Yes	N/A	$z = \frac{\overline{x}-x_0}{\sigma/\sqrt{n}}$	$\overline{x} \pm z\sigma/\sqrt{n}$
1	N/A	No	$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$	$t = \frac{\overline{x}-x_0}{s\sqrt{n}}$	$\overline{x} \pm ts/\sqrt{n}$
2	Yes	Yes	N/A	$z = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x}-\overline{y} \pm z\sqrt{\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})}$
2	Yes	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$ $s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}}$	$\overline{x}-\overline{y} \pm t\sqrt{s^2(\frac{1}{n_1}+\frac{1}{n_2})}$
2	No	Yes	N/A	$\frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}}$	$\overline{x}-\overline{y} \pm z\sqrt{(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2})}$
2	No	No	$s_1 = \sqrt{\frac{\sum_{i=1}^{n_1}(x_i-\overline{x})^2}{n_1-1}}$, $s_2 = \sqrt{\frac{\sum_{i=1}^{n_2}(y_i-\overline{y})^2}{n_2-1}}$	$t = \frac{(\overline{x}-\overline{y})-(x_0-y_0)}{\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}}$	$\overline{x}-\overline{y} \pm t\sqrt{(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2})}$

Table 1: Tests and confidence intervals required for one and two sample problems for continuous variables. In all cases, the data are assumed to be normally distributed, or the sample size large enough for the central limit theorem to apply. The data are assumed to be represented by x_i, $i = 1, \ldots, n$ for a single sample, or x_i, $i = 1, \ldots, n_1$ and y_i, $i = 1, \ldots, n_2$ for a two sample problem. Sample sizes are n for a single sample problem, and n_1 and n_2 for the two sample problem. z indicates a normal table is used, t indicates a t table is required. When a t-table is required, the degrees of freedom are equal to $n-1$ for a single sample problem, while the degrees of freedom are n_1+n_2-2 for a two sample problem with equal variances, and $\min(n_1-1, n_2-1)$ for unequal variances (conservative value). x_0 and y_0 indicate null values under the null hypothesis (usually but not always equal to zero). For paired two sample problems, form the within individual differences, and use the formulae for the one sample case.

Summary of Testing Procedures

There are two philosophies, state error rates in advance, or calculate p-values based on the data observed.

Philosophy 1:

1. State H_0 and H_A. State the α error you think is appropriate for the problem.
2. Find the rejection region.
3. From the data, check whether the observed data fall into the rejection region or not.
4. If the data fall into the rejection region, conclusion is that there is enough evidence to reject the null hypothesis H_0 in favour of the alternative H_A. If the data do not fall into the rejection region, can only say that there is no evidence to reject the null hypothesis.

Philosophy 2:

Definition: The p-value is the probability of obtaining a result as or more extreme than that observed *assuming that the null hypothesis is in fact true.*

It is very important to note that the p-value is **not** the probability that the null hypothesis is correct after having seen the data, even though many clinicians often falsely interpret it this way. The p-value does not directly or indirectly provide this probability, and in fact can be orders of magnitude different from it. In other words, it is possible to have a p-value equal to 0.05, when the probability of the null hypothesis is 0.5, different from the p-value by a factor of 10. Therefore, p-values are the answer to a rather obscure question, which, at best, indirectly helps the researcher in answering their scientific question.

Example of Testing for two means

Here we will illustrate the formula for testing equality of two means, i.e,

$$z = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right|.$$

For example, suppose we wish to look at the difference in mean tumor diameter between two groups of patients with brain cancer in a clinical trial setting, with subjects randomized into accelerated and standard schedule groups. Suppose we observe a mean tumor diameter of $\bar{x}_1 = 3.0$ cm ($\sigma_1 = 1.5$ cm) in 200 subjects under the new schedule, and a mean tumor diameter of $\bar{x}_2 = 3.7$ cm ($\sigma_2 = 1.4$ cm) in 200 subjects under the standard schedule. Plugging into the above formula, we get:

$$z = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| = \left| \frac{3.0 - 3.7}{\sqrt{\frac{1.5^2}{200} + \frac{1.4^2}{200}}} \right| = 4.82$$

Looking up 4.82 on normal tables gives a p-value of $2 \times (0.0000007) = 0.0000014$. Since this indicates a very rare event under H_0, we can reject the null hypothesis that the two means are equal.

- How is this p-value interpreted?
- How useful is knowing this p-value by itself?

Proportions

The situation with proportions is simpler, since one does not need to worry about estimating variances (since the variance of a proportion is fixed once the proportion itself is fixed). We have already seen the case of a single proportion, and for two or more proportions, we have:

Testing (two proportions exactly):

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1}$$

where $\hat{p}$ is the overall estimate of p under $H_0 : \ p_1 = p_2 = p_0$.

Confidence Intervals:

$$\left(\hat{p}_1 - \hat{p}_2 - z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + z\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

In the above formula, $\hat{p}_1$ and $\hat{p}_2$ are the observed proportions in the two groups out of sample sizes n_1 and n_2, respectively, and z is the relevant percentile from normal tables, chosen according to the desired level of the confidence interval. For example, for a 95% confidence interval $z = 1.96$, for a 90% interval $z = 1.64$, and so on.

Testing (two or more proportions):

The generic setup is:

	Category 1	Category 2	...	Category c
Population 1	n_{11}	n_{12}	...	n_{1c}
Population 2	n_{21}	n_{22}	...	n_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Population r	n_{r1}	n_{r2}	...	n_{rc}

Examples:

1. Use of Stroke Unit versus Medical Unit for acute stroke in the elderly (Taken from Garraway et al, British Medical Journal, 1980).

	Patient Independent	Patient Dependent
Stroke Unit	67	34
Medical Unit	46	45

2. Quality of Sleep before elective operation

	Bad	Reasonably Good	Very Good
Triazolam	2	17	12
Placebo	8	15	8

Example 1 can be handled by the methods for two proportions based on the binomial distribution, which we have already seen. However, it is not possible to directly extend these methods to the case when there are three (or more) outcome categories and/or more than two populations. Furthermore, we have been using the Normal distribution approximation to the binomial, which we know is only valid for "large enough" sample sizes. What can we do if we have a table larger than 2×2 or if the sample size is "small"?

Methods to Compare Two or More Proportions

Suppose we wish to test the null hypothesis that $\pi_1 = \pi_2 = \ldots = \pi_N$, that is, we have measured the frequency of occurrence of a dichotomous outcome in N populations, and wish to check if the frequencies are all equal. There are several candidate tests:

Normal approximation (Z) Test: We have seen this test when $N = 2$. The test does not apply when $N > 2$. Alternative hypothesis can be one or two-sided. Requires large samples sizes to be accurate. "Large"

is often stated as a criterion like

$$\text{sample size } \times \min\{\pi, (1-\pi)\} \geq 5.$$

This is somewhat arbitrary, but works reasonably well as a rough guide.

Chi-square (χ^2) Test: The χ^2 test does apply when $N > 2$, but the alternative hypothesis is always two-sided. Requires large samples sizes to be accurate. "Large" is often operationalized as "the expected number of subjects in each cell in the $r \times c$ table must be at least 5". We will see soon how to calculate these expected cell sizes.

Fisher's Exact Test: Both the χ^2 and Z tests require "large" sample sizes to be accurate, but the Fisher's Exact is "exact" for any sample size. The Fisher's Exact Test also applies when $N > 2$, but unlike the χ^2 test, the alternative hypothesis can be one or two-sided.

While it is common practice to use a χ^2 test for large sample sizes and Fisher's Exact Test for smaller sample sizes, a natural question is "Why not just use Fisher's Exact Test all the time, since it is always applicable?" There are two possible answers. The first is that, as we will see, it is computational "expensive" to use Fisher's Exact Test, compared to a χ^2 test. Second, there are different assumptions behind each. As will become clear from the examples on the next few pages, in the Fisher's Exact Test, all "margins" are held fixed ("conditioned upon"), while this is not the case for the Z and χ^2 tests. Thus there is a slightly different inferential philosophy behind each.

One sample χ^2 Test

Suppose we observe the following table of data:

	Success	Failure
Population	x	$n - x$

We would like to test the hypothesis $H_0 : \ \pi = \pi_0$. For example, we might observe patient survival rates one month following a particular sugery, and would lke to test if the survival rate is 80%. We observe the following data:

	Success	Failure
Population	60	40

We would like to test the hypothesis $H_0 : \pi = 0.80$, where π represents the true one month survival rate.

Procedure: Since we hypothesize $\pi = 0.80$, and since we have 100 subjects, we *expect* 80 survivors and 20 deaths. Observed dicrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(60-80)^2}{80} + \frac{(40-20)^2}{20} \\ &= 400/80 + 400/20 = 25 \end{aligned}$$

Comparing the $X^2 = 25$ value on χ^2 tables with 1 degree of freedom (1 df), we find that $p < 0.0005$, so that we have evidence to reject the null hypothesis.

Two sample χ^2 Test

Suppose we observe the following table of data, introduced previously:

	Patient Independent	Patient Dependent	Total
Stroke Unit	67	34	101
Medical Unit	46	45	91
Total	113	79	192

We would like to test the hypothesis $H_0 : \pi_1 = \pi_2$; that is, the proportion of independent patients is the same on Medical or Stroke Units.

Procedure: Since we hypothesize $\pi_1 = \pi_2$, we *expect* to observe the following table of data, on average:

	Patient Independent	Patient Dependent	Total
Stroke Unit	59.44	41.56	101
Medical Unit	53.56	37.44	91
Total	113	79	192

Once again, observed discrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(67-59.44)^2}{59.44} + \frac{(34-41.56)^2}{41.56} + \frac{(46-53.56)^2}{53.56} + \frac{(45-37.44)^2}{37.44} \\ &= 4.9268 \end{aligned}$$

Comparing the $X^2 = 4.9268$ value on χ^2 tables with 1 df, we find that $0.025 < p < 0.05$ (by computer the exact value is 0.0264), so that we have evidence to reject the null hypothesis.

The χ^2 Test for 2×3 table

Suppose we observe the following table of data, introduced previously:

	Bad	Reasonably Good	Very Good	Total
Triazolam	2	17	12	31
Placebo	8	15	8	31
Total	10	32	20	62

We would like to test the hypothesis that the proportions of patients that experience bad, reasonably good and very good outcomes are the same whether they were given the drug or the placebo.

Procedure: Since we hypothesize equal proportions in each treatment group, we *expect* to observe the following table of data, on average:

	Bad	Reasonably Good	Very Good	Total
Triazolam	5	16	10	31
Placebo	5	16	10	31
Total	10	32	20	62

As before, observed dicrepancies from these expected values are evidence against the null hypothesis. We calculate:

$$\begin{aligned} X^2 &= \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(2-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(17-16)^2}{16} + \frac{(15-16)^2}{16} \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} \\ &= 4.525 \end{aligned}$$

Comparing the $X^2 = 4.525$ value on χ^2 tables with 2 df, we find that $0.10 < p < 0.15$ (by computer the exact value is 0.104), so that we do not have sufficient evidence to reject the null hypothesis at either the 0.05 or 0.10 levels. **Note that in general for an $r \times c$ table,** $\text{df} = (\text{r} - 1) \times (\text{c} - 1)$**.**

Question: We note that the proportion on triazolam increases from 20% to 53% to 60% across outcomes, so it may be a good idea to test for a trend (covered in 607?).

Fisher's Exact Test

Suppose we observe the following table of data:

	Success	Failure	Total
Group A	4	2	6
Group B	1	6	7
Total	5	8	13

As with the Z and χ^2 tests, we would like to test the null hypothesis $H_0 : \pi_1 = \pi_2$. However, since the sample size is so small, there is doubt about

the applicability of these tests to this data set. An "exact" test can be constructed via the following reasoning:

We have observed a total of 5 successes. If groups A and B receive equally effective treatments, then the five successes should be eqaully distributed between the two groups. If the sample sizes were equal, we would expect 2.5 successes in each group, but since the sizes are not equal, we expect the successes to be divided in a 6:7 ratio (almost but not quite half/half). As in the previous tests, discrepancies from this "fair split" indicate departures from the null hypothesis. We calculate:

$$\frac{6}{13} \times 5 = 2.31, \text{ and } \frac{7}{13} \times 5 = 2.69$$

Therefore, approximately 2:3 or 3:2 split is expected, and more extreme splits are evidence against the null hypothesis. How extreme is too extreme to be compatible with the null hypothesis? We will calculate the probability of each possible split:

A	5	1	4	2	3	3	2	4	1	5	0	6
B	0	7	1	6	2	5	3	4	4	3	5	2
	5	8	5	8	5	8	5	8	5	8	5	8
Prob	0.005		0.082		0.326		0.408		0.163		0.016	

The tables with probabilities of $0.005 + 0.082 + 0.016 = 0.103$ have values equal to or more extreme than those observed, so by the definition of the p-value, $p = 0.103$ by the Fisher's Exact Test.

Calculating Probabilities for the Fisher's Exact Test

The probabilities on the previous page were calculated using the **hypergeometric distribution**. In general, if we observe

A	a	b	$a+b$
B	c	d	$c+d$
	$a+c$	$b+d$	N

where $N = a + b + c + d$, then the probability of observing that table is:

$$\text{Prob} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$$

A less simplified equivalent formulae provides a clue as to how the probability is calculated:

$$\text{Prob} = \frac{\frac{(a+c)!}{a!c!} \times \frac{(b+d)!}{b!d!}}{\frac{N!}{(a+b)!(c+d)!}}$$

Consider the A and B labels as random labels. In how many ways can one choose that all 5 (or 4 or 3 or 2 or 1 or 0) of the A labels happen to end up as "successes"?

How useful are p-values for medical decision making?

While p-values are still often found in the literature, there are several major problems associated with their use:

1. P-values are often misinterpreted as the probability of the null hypothesis given the data, when in fact they are calculated assuming the null hypothesis to be true. In fact, p-values discuss probabilities of the form $P(\text{data}|H_0)$, not $P(H_0|\text{data})$. The latter is only available from a Bayesian viewpoint.

2. Clinicians often use p-values to "dichotomize" results into "important" or "unimportant" depending on whether $p < 0.05$ or $p > 0.05$, respectively. However, there is not much difference between p-values of 0.049 and 0.051, so that the cutoff of 0.05 is arbitrary.

3. *P*-values concentrate attention away from the magnitude of treatment differences. For example, one could have a p-value that is very small, but is associated with a clinically unimportant difference. This is especially prone to occur in cases where the sample size is large. Conversely, results of potentially great clinical interest are not necessarily ruled out if $p > 0.05$, especially in studies with small sample sizes. Therefore, one should not confuse statistical significance (i.e., $p < 0.05$) with practical or clinical importance.

4. The null hypothesis is almost *never* exactly true. Does one seriously ever believe that the null hypothesis $\mu = \mu_0$ is correct (rather than, say, $\mu = \mu_0 + 0.0000001$)? Since one knows the null hypothesis is almost surely false to begin with, it makes little sense to test it. Instead, one should concern oneself with the question of "by how much are the two treatments different", or "what is a point and interval estimate of μ".

There are so many problems associated with p-values that most statisticians now recommend against their use, in favor of confidence intervals or Bayesian methods. In fact, some prominent journals have virtually banished p-values from publication (e.g., *Epidemiology*, see quote below), others strongly discourage their use in favor of confidence intervals and/or have published articles and editorials encouraging the use of Bayesian methodology (e.g. *American Journal of Epidemiology*). In this course we will focus mainly on these more informative techniques for drawing statistical inferences.

One of the most prominent epidemiologists, Kenneth Rothman, former editor of the premiere journal, *Epidemiology*, wrote in that journal (Rothman, K. Writing for Epidemiology, 1998;9(3):333-337.

> When writing for Epidemiology, you can also enhance your prospects if you omit tests of statistical significance. Despite a widespread belief that many journals require significance tests for publication, the Uniform Requirements for Manuscripts Submitted to Biomedical Journals discourages them [see Uniform Requirements for Manuscripts Submitted to Biomedical Journals. N Engl J Med 1997;336:309-315], and every worthwhile journal will accept papers that omit them entirely. **In Epidemiology, we do not**

publish them at all. Not only do we eschew publishing claims of the presence or absence of statistical significance, we discourage the use of this type of thinking in the data analysis, such as in the use of stepwise regression. We also would like to see the interpretation of a study based not on statistical significance, or lack of it, for one or more study variables, but rather on careful quantitative consideration of the data in light of competing explanations for the findings. For example, we prefer a researcher to consider whether the magnitude of an estimated effect could be readily explained by uncontrolled confounding or selection biases, rather than simply to offer the uninspired interpretation that the estimated effect is significant, as if neither chance nor bias could then account for the findings.

Many data analysts appear to remain oblivious to the qualitative nature of significance testing. Although calculations based on mountains of valuable quantitative information may go into it, statistical significance is itself only a dichotomous indicator. As it has only two values, significant or not significant, it cannot convey much useful information. Even worse, those two values often signal just the wrong interpretation. These misleading signals occur when a trivial effect is found to be significant, as often happens in large studies, or when a strong relation is found nonsignificant, as often happens in small studies. P-values, being more quantitative, are preferable to statements about statistical significance tests, and we do publish P-values on occasion. We do not publish them as an inequality, such as $P < 0.05$, but as a number, such as $\mathrm{P} = 0.13$. By giving the actual value, one avoids the problem of dichotomizing the continuous P-value into a two-valued measure. Nevertheless, P-values still confound effect size with study size, the two components of estimation that we believe need to be reported separately. Therefore, we prefer that P-values be omitted altogether, provided that point and interval estimates, or some equivalent, are available.

One arena in which P-values are the usual analytic tool is in the assessment of trends, such as the trend in rate across dose categories. Even here, we believe that they should be avoided. Slope estimates are better, and smoothed trend evaluations, such

as kernel smoothing or spline regression, are better yet; these presentations should ideally include some assessment of statistical precision to accompany the trend estimate.

Frequentist confidence intervals

While the p-value provides some information concerning the rarity of events as or more extreme than that observed assuming the null hypothesis to be exactly true, it provides no information about what the true parameter values might be. In the above two mean example, we have observed a tumor diameter difference of 0.7 cm, which was shown to be "statistically significant", with a p-value of about 0.000001. Although we have observed a difference of 0.7 cm, we know that our data are from a random sample of patients to whom this procedure could be applied, so that the true mean difference could in fact be higher or lower than our observed difference. How likely is it that the true mean difference in tumor diameter is clinically important?

One way to answer this question is with a confidence interval. We can calculate a 95% confidence interval for the difference in means for the two groups using the formula (see previous chart)

$$\left(\bar{x}_1 - \bar{x}_2 - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Plugging in the values we obtained from our clinical trial example given above, we find a CI of (-0.46,-0.94). Thus, roughly speaking, it is likely that the true tumor diameter difference between our two schedules is somewhere between about half a cm less under the new schedule (-0.46 cm) and up to almost a 1 cm reduction (-0.94). Although our p-value for this same data set was very small, which enabled us to reject the null hypothesis, we can see that the confidence interval provides more clinically useful information about the magnitude of the difference. We can also see that, in contrast to what may be believed after seeing the p-value, we are still uncertain about the clinical utility of the new schedule, since values near the lower limit of the CI would not be very interesting clinically (it would represent less than a 30% change from the mean baseline tumor size), while differences near 1

cm may be clinically interesting. Therefore, our conclusions from the CI are more detailed than those from the p-value. This is true in general, as we will now discuss.

Interpreting confidence intervals

Confidence intervals are derived from procedures that are set up to "work" 95% of the time (if a 95% CI is used). The two CI equations above provide procedures that, when used repeatedly across different problems, will capture the true value of the mean (or difference in means) 95% of the time, and fail to capture the true value 5% of the time. In this sense, we have confidence that the procedure works well in the long run, although in any *single application*, of course, the interval either does or does not contain the true mean. Note that we are careful **not** to say that our confidence interval has a 95% probability of containing the true parameter value. For example, we did not say that the true difference in mean tumor diameter is in the interval (-0.49, -0.94) with 95% probability. This is because the confidence limits and the true mean tumor diameters are both fixed numbers, and it makes no more sense to say that the true mean is in this interval than it does to say that the number 2 is inside the interval (1, 6) with probability 95%. Of course, 2 is inside this interval, just like the number 8 is outside of the interval (1, 6). However, the procedure used to calculate confidence intervals provides random upper and lower limits which depend on the data collected, and in repeated uses of this formula across a range of problems, we expect the random limits to capture the true value 95% of the time, and exclude the true mean 5% of the time. Refer to Figure 1. If we look at the set of confidence intervals as a whole, we see that about 95% of them include the true parameter value. However, if we pick out a single trial, it either contains the true value (about 95% of the time) or excludes this value (about 5% of the time).

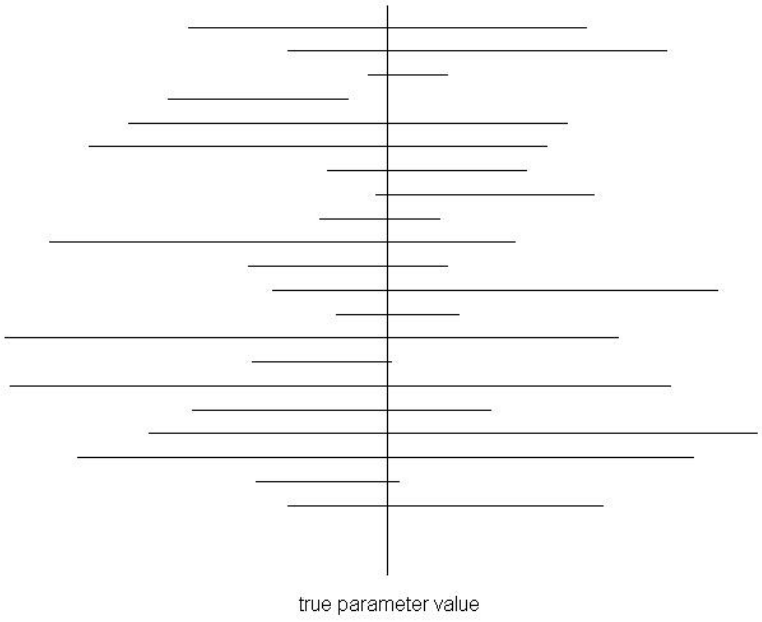

Figure 1: A series of 95% confidence intervals for an unknown parameter.

Despite their somewhat unnatural interpretation, confidence intervals are generally preferred to p-values. This is because they focus attention on the range of values compatible with the data, on a scale of direct clinical interest. Given a confidence interval, one can assess the clinical meaningfulness of the result, as can be seen in Figure 2.

Depending on where the upper and lower confidence interval limits fall in relation to the upper and lower limits of the region of clinical equivalence, different conclusions should be drawn. The region of clinical equivalence, sometimes called the region of clinical indifference, is the region inside of which two treatments, say, would be considered to be the same for all practical purposes. The point 0, indicating no difference in results between two treatments, is usually included in the region of clinical equivalence, but values above and below 0 are usually also included. How wide this region is depends on each individual clinical situation. For example, if one treatment schedule is much more expensive than another, one may want at least a 50% reduction in tumor diameter in order to consider it the preferred treatment. There are five different conclusions that can be made after a confidence interval has been calculated, as illustrated by the five hypothetical intervals displayed in Figure 2:

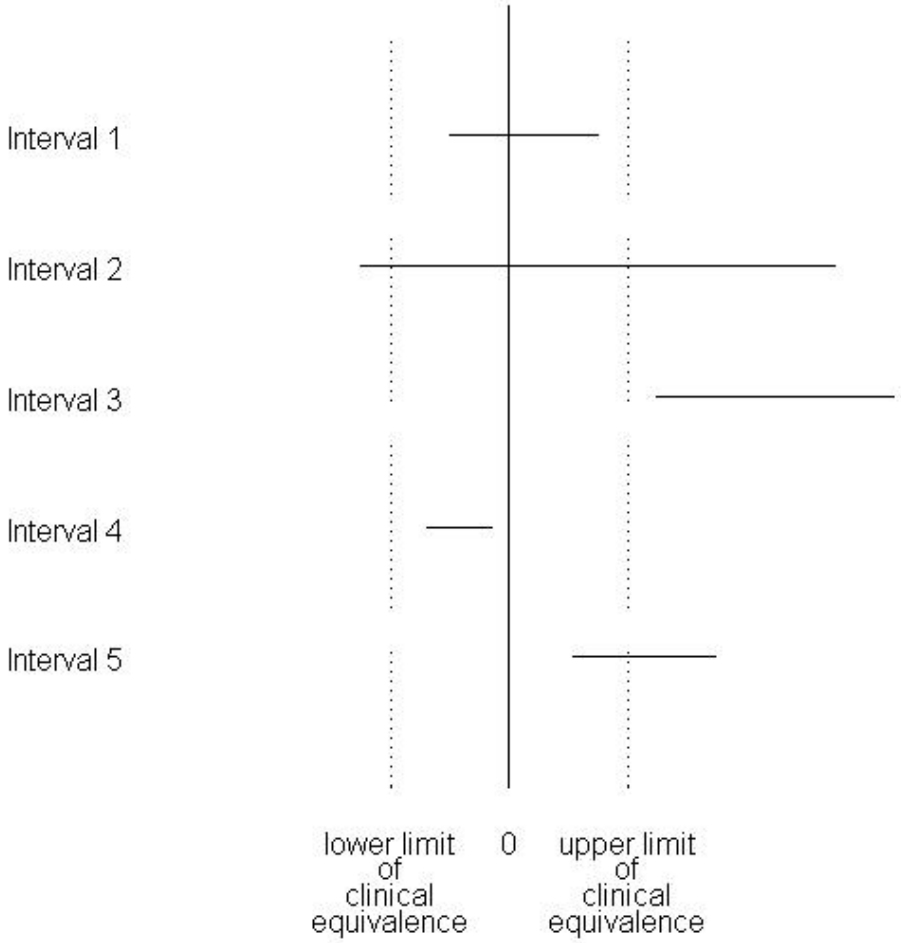

Figure 2: How to interpret confidence intervals. Depending on where the confidence interval lies in relation to a region of clinical equivalence, different conclusions can be drawn.

1. The CI includes zero, and both upper and lower CI limits, if they were the true values, would not be clinically interesting. Therefore, this variable has been shown to have no important effect.

2. The CI includes zero, but one or both of the upper or lower CI limits, if they were the true values, would be interesting clinically. Therefore, the results of this variable in this study is inconclusive, and further evidence needs to be collected.

3. The CI does not include zero, and all values inside the upper and lower CI limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable to be important.

4. The CI does not include zero, but all values inside the upper and lower CI limits, if they were the true values, would not be clinically interesting. Therefore, this study shows this variable, while having some small effect, is not clinically important.

5. The CI does not include zero, but only some of the values inside the upper and lower CI limits, if they were the true values, would be clinically interesting. Therefore, this study shows this variable has at least a small effect, and may be clinically important. Further study is required in order to better estimate the magnitude of this effect.

Revisiting the confidence interval discussed above in light of Figure 2, we see that the interval based on the two group clinical trial is of type 5. Once again, note that this confidence interval provides much more detailed conclusions compared to the information contained in a p-value. P-values group together intervals 1 and 2 as "nonsignificant" and intervals 3, 4, and 5 as "significant". This can lead to very misleading conclusions, from a clinical viewpoint. For example, quite similar clinical conclusions should be drawn from intervals 1 and 4, even though one is "significant", and the other is not. It should now be clear why many journals discourage reporting results in terms of p-values, and encourage confidence intervals.

Summary of frequentist statistical inference

The main tools for statistical inference from the frequentist point of view are p-values and confidence intervals. P-values have fallen out of favor among statisticians, and although they continue to appear in medical journal articles, their use is likely to greatly diminish in the coming years. Confidence intervals provide much more clinically useful information than p-values, so are to be preferred in practice. Confidence intervals still do not allow for the formal incorporation of pre-existing knowledge into any final conclusions. For example, in some cases there may be compelling medical reasons why a new technique may be better than a standard technique, so that faced with an inconclusive confidence interval, a researcher may still wish to switch to the new technique, at least until more data become available. On what basis could this decision be justified? We will return to this question later, where we look at Bayesian statistical inference.