Appendix - Statistical argument in detail

Model

The model considered is a 2-group clinical trial (treatment v control), with true (unknown) treatment effect denoted by μ. The outcome is continuous and normally distributed with variance assumed known and equal to 0.5 (this is for convenience, so that the variance of an estimate of treatment effect based on a pair of observations equals 1).

Note that the assumption of known variance does not weaken the argument: incorporating estimation of the variance into the model would have little effect in an experiment of reasonable size and in a small one would serve, if anything, to dilute any apparent *trends* by adding extra uncertainty into the model. Further, the assumption of any particular value for this variance serves only to set the scale of the observations.

The trial will have 100 observations in each group, considered as 100 pairs of differences (random variables X_i with realisations x_i : i=1…100). Each X_i is distributed $N(\mu,1)$, and their mean $\bar{X}$ - distributed $N(\mu,0.01)$ with realisation $\bar{x}$ - will be used to estimate the treatment effect. Note that the choice of 100 pairs is also merely a convenience (which serves to set the scale of the size of the experiment). When we consider later the addition of k% more data, this translates directly into k additional pairs of observations.

Analysis of the first 100 pairs of observations

We carry out the trial and perform a significance test of the usual null hypothesis (H_0: μ=0).

Suppose the outcome is that $\bar{x}$ is greater than 0 with 2-sided significance P (the 1-sided significance being P/2). P depends only on the ratio (which we denote by z_1) of $\bar{x}$ to its standard error, and the relationship between them is:

$z_1=\bar{x}/\sqrt{(1/100)}=10\bar{x}=\Phi^{-1}(1-P/2)$

(where Φ is the standard normal probability integral)

Confidence intervals for μ come directly from the likelihood function $L(\mu|\bar{x})$ which has the same mathematical form as a normal distribution centred round $\bar{x}$ with variance 0.01

Thus our current knowledge about μ is reasonably represented by this function, variously termed in related contexts the *fiducial distribution*[1] or, more recently, the *confidence distribution*.[2]

Whilst this distribution cannot, in general, be interpreted and manipulated like an ordinary probability distribution,[3] in a Bayesian approach when conjoined with a proper prior distribution, a valid posterior distribution results.

In the case considered here, the confidence distribution for μ [which is $N(\bar{x},0.01)$] is also the Bayesian posterior distribution of μ if an (improper) uniform prior over $(-\infty,\infty)$ is used. Therefore, to the extent that it makes sense - and is possible - to represent our prior knowledge about μ by a proper uniform prior that is (effectively) guaranteed to cover the support of the confidence distribution, it will be justifiable to treat the confidence distribution like an ordinary probability distribution also. Note that this condition is equivalent to saying that, prior to the experiment, we are adopting the (not unreasonable) position that, whilst we know that μ must lie within some (stated but wide) limits, we have no preference for any particular value or values within those limits. Such a prior is termed 'locally uniform'.[4] In contrast, it could certainly be argued that, under the scientific method, there should be an initial presumption in favour of the null hypothesis. This would be represented

here by a prior with a lump at $\mu=0$. In the interests of simplicity (it is hard to say how big such a lump should be) we have not followed this route. However, we do note that any modification in that direction would increase the chances that results become less significant with more data and so, (in terms of the analysis of this paper) our approach is conservative.

In the following, we do indeed manipulate the confidence distribution like an ordinary probability distribution, which is justified based on the above argument. Further, we describe the *probability* the P-value moves in a particular direction. Whilst this is convenient shorthand - and has the advantage of expressing the results clearly - a more correct term (given the use of the likelihood to express a distribution for μ) would probably be *confidence*.

The effect of additional data

Suppose we wish to estimate the effect that additional data is likely to have - specifically k% more data - which here means an extra k pairs of observations.

Denote the mean of the differences of these extra pairs by the random variable $\overline{Y}$ with realisation $\bar{y}$

Clearly, conditional on μ, $\overline{Y}$ is distributed $N(\mu,1/k)$

Adding in the distribution of μ (above), gives the (unconditional) Expectation of $\overline{Y}$ as equal to $\bar{x}$, with variance equal to $(0.01+1/k)$ - being the sum of the variances of $\overline{Y}|\mu$ and μ.

Or, in other words, before we collect these new data, $\overline{Y}$ is distributed $N(\bar{x}, 0.01+1/k)$

After we collect them, we will have an updated mean (our new best estimate of the treatment effect):

$(100\bar{x}+k\bar{y})/(100+k)$

and an updated standard error of the mean:

$1/\sqrt{[100+k]}$

The statistical significance will depend (only) on the ratio of these, which we denote by z_2:

$z_2=(100\bar{x}+k\bar{y})/\sqrt{(100+k)}$

z_2 itself can be regarded as the realisation of the random variable Z_2, where Z_2 is simply:

$Z_2=(100\bar{x}+k\bar{Y})/\sqrt{(100+k)}$

Interest now centres on whether z_2 exceeds a given target-value: z_t say.

If we want to calculate the probability that the updated results will be more (or less) statistically significant than they were, the target z_t will be the same as the z-value originally achieved (i.e. $z_t=z_1$).

Specifically, the probability($Z_2<z_t$) is the probability that the updated results will be *less* significant than they were.

Alternatively, if we are interested in the likelihood that the updated results will actually achieve significance at a given level (α say, 2-sided) then z_t is expressed in terms of α:

$z_t=\Phi^{-1}(1-\alpha/2)$

Here, the probability($Z_2<z_t$) is the probability that the updated results will *not* be statistically significant at the (2-sided) α level.

In either case, the event of interest is ($Z_2<z_t$) which can be written as:

$(100\bar{x}+k\bar{Y})/\sqrt{(100+k)}<z_t$

or equivalently:

$(\bar{Y}-\bar{x})/\sqrt{(1/k+1/100)}<[z_t\sqrt{(100+k)}-\bar{x}(100+k)]/[k\sqrt{(1/k+1/100)}]$

Now, from a previous result, $(\overline{Y}-\bar{x})$ is distributed N(0, 0.01+1/k), so the left hand side of the above is a standard normal variate, and the probability of the event is given by:

$$\Phi\{[z_t\sqrt{(100+k)}-\bar{x}(100+k)]/[k\sqrt{(1/k+1/100)}]\}$$

Writing (for convenience) $\bar{x}$ in terms of z_1 and putting K=k/100 (so K gives the amount of extra data envisaged as a fraction of that actually collected), the expression simplifies to:

$$\Phi\{[z_t-z_1\sqrt{(1+K)}]/\sqrt{K}\}$$

Then, inserting the respective values for z_t and expressing z values in terms of P and α as appropriate gives us:

(1). probability(results become *less* significant)

$$=\Phi\{[\sqrt{(1/K)}-\sqrt{(1/K+1)}]\Phi^{-1}(1-P/2)\}$$

(2). probability(results will not *be* significant at the (2-sided) α level)

$$=\Phi\{[\sqrt{(1/K)}]\Phi^{-1}(1-\alpha/2)-[\sqrt{(1/K+1)}]\Phi^{-1}(1-P/2)\}$$

Finally, we might wish to consider the probability that the additional observations, analysed as a new stand-alone experiment (i.e. not combined with the original data) achieve a significant result (again using a 2-sided test at the α level). Here, the event of interest can still be written as $(Z_2<z_t)$, where $z_t=\Phi^{-1}(1-\alpha/2)$, but with a new definition for Z_2 which now depends on $\overline{Y}$ and k only:

$$Z_2=\overline{Y}/\sqrt{(1/k)}$$

Following the algebra through in the same way as before leads quickly to the result:

(3). probability(results of extra data analysed alone will not *be* significant at the (2-sided) α level)

$$=\Phi\{\Phi^{-1}(1-\alpha/2)/\sqrt{(1+K)}-\Phi^{-1}(1-P/2)/\sqrt{(1+1/K)}\}$$

Here the most interesting case is where we envisage a new experiment the same size as the one carried out (so K=1). This gives

(4). probability(results of a repeat experiment will not be significant at the (2-sided) α level)

$=\Phi\{[\Phi^{-1}(1-\alpha/2)-\Phi^{-1}(1-P/2)]/\sqrt{2}\}$

Technical footnote, implications of the equations and limiting behaviour

As a technical point, the possibility that the results might become significant in the opposite direction (i.e. an apparent benefit of treatment over control turns into a benefit of control over treatment) on the addition of extra data has been counted as becoming *less* significant even if the P-value actually drops. This is because such an outcome could hardly be claimed as successful confirmation of *trends* apparent in the original analysis.

From (1), (2) and (3) above, it can be seen that it is only the ratio of sample sizes (K) that matters, not their individual values. (Of course, if the number of subjects is small, the precision of the variance estimate can suffer if based on too few degrees of freedom. However, that would only add to the overall uncertainty and, with reasonable sample sizes, has little effect). Further, in (1), it is clear that the expression in [] is always negative (as K>0), so the expression in {} is always negative also (as P<1), and hence the probability that the results become less significant is always less than 0.5. In contrast, from (4), it can be seen that the probability that the results of a repeat experiment (to the same design) are non-significant is less than 0.5 if and only if $P<\alpha$; the probability being equal to 0.5 if $P=\alpha$. In other words, if the first experiment just attained statistical significance at a given level, a direct attempt to replicate the results has (rationally) a 50-50 chance of producing results significant at the same level.

Looking now at limiting behaviour, as $K=>0$, $\{[1-\sqrt{(1+K)}]/\sqrt{K}\}=>-0.5\sqrt{K}=>0$

So, (from (1)), the probability that the results become less significant tends to 0.5 (from below) as K=>0.

In other words, although extra data are always more likely than not to increase the significance, as the relative amount of new data envisaged diminishes, the effect of sampling variation becomes increasingly important such that, in the limit, the P-value is equally likely to move in either direction.

Inspection of equation (2) shows that, as K=>0, the effect of the extra data on the significance test (at the α level) itself tends to zero. Specifically, if P<α the results stay (statistically) significant, whilst if P>α they remain non-significant - which of course is how it has to be.

Limiting behaviour as K=>∞ is easiest seen by considering the equation preceding (1) and (2):

$$\Phi\{[z_t - z_1\sqrt{(1+K)}]/\sqrt{K}\} = \Phi\{z_t/\sqrt{K} - z_1\sqrt{(1/K+1)}\}$$

Clearly, whatever the value of z_t, the above expression tends to $\Phi\{-z_1\}=P/2$ as K=>∞, and this can be seen to hold directly for (3) also. Or, in other words, as the amount of new data envisaged becomes arbitrarily large, the probability that the P-value becomes less significant, or indeed fails to become statistically significant (at any level at all) whether the new results are combined with the original or not, tends to P/2: the 1-sided significance from the original analysis. However, this convergence is very slow, and of theoretical rather than of practical interest.

References

1. Fisher RA. The design of experiments. 8th ed. Edinburgh : Oliver & Boyd, 1966 [1935].

2. Xie MG, Singh K. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. International Statistical Review 2013; 81: 3-39.

3. Cox D. Principles of Statistical Inference. Cambridge: Cambridge University Press, 2006, p66.

4. Lee PM. Bayesian Statistics: an introduction. New York: Oxford University Press, 1989, p47.

Table 1: Percent of times the P-value would be expected get less significant had extra data been collected, given the current P-value and amount of extra data

Amount of extra data as % of current	Current (2 tailed) P-value						
	0.001	0.01	0.05	0.06	0.08	0.10	0.15
1000%	0.8%	3.0%	7.6%	8.4%	10.0%	11.4%	14.6%
100%	8.6%	14.3%	20.8%	21.8%	23.4%	24.8%	27.5%
50%	14.8%	20.6%	26.7%	27.5%	28.9%	30.1%	32.4%
20%	24.1%	29.1%	33.8%	34.4%	35.4%	36.3%	37.9%
10%	30.6%	34.5%	38.1%	38.6%	39.3%	40.0%	41.2%
1%	43.5%	44.9%	46.1%	46.3%	46.5%	46.7%	47.1%
0.01%	49.3%	49.5%	49.6%	49.6%	49.7%	49.7%	49.7%

Table 2: Percent of times to expect a non-significant result (2-tailed test; α=0.05) on addition of extra data, given the current P-value and amount of extra data

Amount of extra data as % of current	Current (2 tailed) P-value						
	0.001	0.01	0.05	0.06	0.08	0.10	0.15
1000%	0.2%	1.9%	7.6%	8.8%	11.2%	13.5%	18.7%
100%	0.4%	4.6%	20.8%	24.2%	30.3%	35.7%	47.0%
50%	0.2%	4.6%	26.7%	31.4%	39.7%	46.9%	61.0%
20%	0.0%	2.7%	33.8%	41.1%	53.8%	63.8%	80.4%
10%	0.0%	1.0%	38.1%	48.4%	65.2%	77.1%	92.3%

Table 3: Percent of times to expect a non-significant result (2-tailed test; α=0.05) of a repeat experiment of the same size, analysed independently

Current (2 tailed) P-value						
0.001	0.01	0.05	0.06	0.08	0.10	0.15
17.3%	33.2%	50.0%	52.2%	55.9%	58.8%	64.4%