Biost 536: Categorical Data Analysis in Epidemiology
Emerson, Autumn 2013

42.5/50

Homework #1
September 26, 2013

Written problems due at 5 pm, Thursday, October 3, 2013. Homeworks must be submitted electronically according to the instructions that will be distributed via email.

This homework explores the role of screening studies in promoting the accuracy of the process of identifying and quantifying risk factors for disease.

The goal of the drug approval process should be

1. To have a low probability of approving drugs that do not work,
2. To have a high probability of approving drugs that do work, and
3. To have a high probability that an approved drug does work.

Now suppose we decide to perform a experiment or series of experiments, and to approve the drug whenever the estimated treatment effect (perhaps standardized to some Z score) exceeds a pre-defined threshold. When stated in statistical jargon, these goals become

1. To have a low type I error α when a null hypothesis of no treatment effect is true,
2. To have a high statistical power $Pwr = 1-\beta$ (so β is the type II error) when some alternative hypothesis is true, and
3. To have a high positive predictive value PPV = (number of approved effective drugs) / (number of approved drugs).

We can examine the interrelationships of these statistical design criteria in the context of a RCT where we let θ denote our treatment effect, and we presume that an ineffective drug has $\theta = 0$, and an effective drug has some $\theta > 0$.

In the "frequentist" inference most often used in RCT, we typically choose some value for the "level of significance" (or type I error) α. This will be the probability of approving the drug when $\theta = 0$.

Most often, we base our decisions on some estimate of the treatment effect that is known to be approximately normally distributed

$$\hat{\theta} \sim N\left(\theta, \frac{V}{n}\right).$$

In experimental design, we sometimes choose a sample size n and then compute the power of the study to detect a particular alternative hypothesis. When our null hypothesis corresponds to $\theta = 0$, the power of a particular design depends upon the type I error α, the variability of the data V, the true value of the treatment effect θ, and the sample size n according to the following formula:

$$Pwr = 1 - \Pr\left(Z \le z_{1-\alpha} - \theta\sqrt{\frac{n}{V}}\right), \qquad \text{(Eq. 1)}$$

where Z is a random variable having the standard normal distribution, and the constant $z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution such that $\Pr(Z \le z_{1-\alpha}) = 1 - \alpha$.

In other settings, we choose a desired power $Pwr = 1 - \beta$, and then compute a sample size according to the value of β using the following formula (which again presumes a null hypothesis of $\theta = 0$):

$$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 V}{\theta^2}, \qquad \text{(Eq. 2)}$$

where we again use the quantiles of the standard normal distribution. The following table provides values of $z_{1-\alpha}$ for selected values of α:

α	0.005	0.01	0.025	0.05	0.10	0.20
$z_{1-\alpha}$	2.575829	2.326348	1.959964	1.644854	1.281552	0.841621

More generally, we can obtain an arbitrary quantile using statistical software. The commands to obtain the $z_{1-\alpha}$ quantile when $\alpha = 0.075$ in three commonly used programs are:

- (Stata) `di invnorm(1 - 0.075)`
- (R) `qnorm(1 - 0.075)`
- (Excel) `norminv(1 - 0.075, 0 , 1)`

Similarly, we can obtain Pr($Z \leq c$) for arbitrary choices of c using statistical software. The commands to obtain Pr($Z \leq c$) when $c = 1.75$ in three commonly used programs are:

- (Stata) `di norm(1.75)`
- (R) `pnorm(1.75)`
- (Excel) `normdist(1.75, 0 , 1, TRUE)`

Bayes Rule can be used to compute the PPV from α and β, providing we know the prior probability π that a treatment would work (this prior probability might be thought of as the proportion of effective treatments among all treatments that we would consider testing—sort of a prevalence of good treatments):

$$PPV = \frac{(1-\beta)\times\pi}{(1-\beta)\times\pi + \alpha\times(1-\pi)} \qquad \text{(Eq. 3)}$$

In this homework, we consider a couple examples of two different strategies of testing for experimental treatments:

1. Strategy 1: Test each treatment in one large "pivotal" RCT.
2. Strategy 2: Test each treatment in one small "pilot" RCT that screens for promising treatments. Any treatment that passes this screening phase, is then tested more rigorously in one larger "confirmatory" RCT.

To compare "apples with apples":

- We pretend that we have 500,000 patients with disease X to use when evaluating ideas that we have formulated for treating disease X.
- We further pretend that 10% of our ideas correspond to drugs that truly work (so $\pi = 0.10$), and all those truly effective drugs provide the same degree of benefit $\theta = 1$ to patients with disease X. The other 90% of our ideas correspond to drugs that provide no benefit to the patients (so $\theta = 0$).
- In every RCT, the true variability of the patient data corresponds to $V = 63.70335$.

Problems using Strategy 1: Only Pivotal RCT

1. (A: Pivotal) Suppose we choose a type I error of $\alpha = 0.025$ and a power of 97.5% (so $\beta = 0.025$) under the alternative hypothesis that the true treatment effect is $\theta = 1$.
   a. What sample size n will be used in each RCT? **979**

   $$n = \frac{(z_{1-\alpha} + z_{1-\beta})^2 V}{\theta^2} = \frac{(1.959964 + 1.959964)^2 \times 63.70335}{1^2} = 978.855$$

   b. How many of our ideas will we be able to test? **511**

   500,000 / 979 = 510.7

   c. How many of those tested ideas will be truly beneficial drugs? **51**

   511 x 0.10 = 51.1

   d. How many of the tested beneficial drugs will have significant results? **50**

   51 x 0.975 = 49.7

   e. How many of those tested ideas will be truly ineffective drugs? **460**

   511 – 51 = 460

   f. How many of the tested ineffective drugs will have significant results? **12**

   460 x 0.025 = 11.5

   g. How many of the tested drugs will have significant results? **62**

   50 + 12 = 62

   h. What proportion of the drugs with significant results will be truly beneficial? **0.8065**

   50 / 62 = 0.8065 or

   $$PPV = \frac{(1-\beta)\times\pi}{(1-\beta)\times\pi + \alpha\times(1-\pi)} = \frac{(1-0.025)\times 0.10}{(1-0.025)\times 0.10 + 0.025\times(1-0.10)} = 0.8125$$

2. (B: Pivotal) Suppose we choose a type I error of $\alpha = 0.025$ and a power of 80.0% (so $\beta = 0.20$) under the alternative hypothesis that the true treatment effect is $\theta = 1$.
   a. What sample size n will be used in each RCT? **500** ✓
   b. How many of our ideas will we be able to test? **1000** ✓
   c. How many of those tested ideas will be truly beneficial drugs? **100** ✓
   d. How many of the tested beneficial drugs will have significant results? **80** ✓
   e. How many of those tested ideas will be truly ineffective drugs? **900** ✓
   f. How many of the tested ineffective drugs will have significant results? **23** ✓
   g. How many of the tested drugs will have significant results? **103** ✓
   h. What proportion of the drugs with significant results will be truly beneficial? **0.78** ✓

5/5

5/5

3. (C: Pivotal) Suppose we choose a type I error of $\alpha = 0.05$ and a power of 80.0% (so $\beta = 0.20$) under the alternative hypothesis that the true treatment effect is $\theta = 1$.
   a. What sample size n will be used in each RCT? **394** ✓
   b. How many of our ideas will we be able to test? **1270** ✓
   c. How many of those tested ideas will be truly beneficial drugs? **127** ✓
   d. How many of the tested beneficial drugs will have significant results? **102** ✓
   e. How many of those tested ideas will be truly ineffective drugs? **1143** ✓
   f. How many of the tested ineffective drugs will have significant results? **57** ✓
   g. How many of the tested drugs will have significant results? **159** ✓
   h. What proportion of the drugs with significant results will be truly beneficial? **0.64** ✓

Problems using Strategy 2: Screening pilot RCT, followed by Confirmatory RCT

5/5

4. (D: Screening pilot study) Suppose we choose a type I error of $\alpha = 0.025$ and a sample size of $n = 100$ for each pilot RCT.
   a. Under the alternative hypothesis $\theta = 1$, what is the power? **0.24** ✓
   b. If we use 350,000 patients in pilot RCT, how many ideas will we test? **3500** ✓
   c. How many of those tested ideas will be truly beneficial drugs? **350** ✓
   d. How many of the tested beneficial drugs will have significant results? **84** ✓
   e. How many of those tested ideas will be truly ineffective drugs? **3150** ✓
   f. How many of the tested ineffective drugs will have significant results? **79** ✓
   g. How many of the tested drugs will have significant results? **163** ✓
   h. What proportion of the drugs with significant results will be truly beneficial? **0.52** ✓

5/5

5. (D: Confirmatory trials) Suppose we choose a type I error of $\alpha = 0.025$ and use all remaining patients in the confirmatory trials of each drug that had significant results in problem 4.
   a. How many confirmatory RCT will be performed? **163** ✓
   b. What sample size n will be used in each RCT? **920** ✓
   c. Under the alternative hypothesis $\theta = 1$, what is the power? **0.97** ✓
   d. How many confirmatory RCTs will be for truly beneficial drugs? **85** ✓
   e. How many of the tested beneficial drugs will have significant results? **82** ✓
   f. How many confirmatory RCTs will be for truly ineffective drugs? **78** ✓
   g. How many of the tested ineffective drugs will have significant results? **2** ✓
   h. How many of the tested drugs will have significant results? **84** ✓
   i. What proportion of the drugs with significant results will be truly beneficial? **0.98** ✓

5/5

6. (E: Screening pilot study) Suppose we choose a type I error of $\alpha = 0.10$ and a power of 85.0% (so $\beta = 0.15$) under the alternative hypothesis that the true treatment effect is $\theta = 1$.
   a. What sample size n will be used in each RCT? **342** ✓
   b. If we use 350,000 patients in pilot RCT, how many ideas will we test? **1023** ✓
   c. How many of those tested ideas will be truly beneficial drugs? **102** ✓
   d. How many of the tested beneficial drugs will have significant results? **87** ✓
   e. How many of those tested ideas will be truly ineffective drugs? **921** ✓
   f. How many of the tested ineffective drugs will have significant results? **92** ✓
   g. How many of the tested drugs will have significant results? **179** ✓
   h. What proportion of the drugs with significant results will be truly beneficial? **0.49** ✓

7. (E: Confirmatory trials) Suppose we choose a type I error of $\alpha = 0.025$ and use all remaining patients in the confirmatory trials of each drug that had significant results in problem 6.
   a. How many confirmatory RCT will be performed? **179** ✓
   b. What sample size n will be used in each RCT? **838** ✓
   c. Under the alternative hypothesis $\theta = 1$, what is the power? **0.95** ✓
   d. How many confirmatory RCTs will be for truly beneficial drugs? **88** ✓
   e. How many of the tested beneficial drugs will have significant results? **84** ✓
   f. How many confirmatory RCTs will be for truly ineffective drugs? **91** ✓
   g. How many of the tested ineffective drugs will have significant results? **2** ✓
   h. How many of the tested drugs will have significant results? **86** ✓
   i. What proportion of the drugs with significant results will be truly beneficial? **0.97** ✓

5/5

Comparisons

8. Of the 5 different strategies considered (problems 1, 2, 3, 4 and 5, or 6 and 7) which do you think best and why?

Strategy E is the best one for the following reasons: ✓

10/10

Why the pivotal RCTs are not ideal:

Strategies A, B and C were all pivotal RCTs. Only one study was conducted, therefore the drug population was not enriched by two smaller RCT like in strategies D and E. Consequences of this were that the PPV of A, B, and C (0.81, 0.78, and 0.64 respectively) were relatively low compared to the PPV of D and E (0.98 and 0.97). A relatively low amount of drugs that were approved in A, B, and C actually worked. This is why the pivotal RCT are not good strategies.

Why strategy E is better than strategy D:

Both strategy D and E have two RCT each. One was a screening RCT meant to enrich the proportion of effective drugs being studied. The next was a confirmatory RCT meant to conclude how many drugs were actually effective of that enriched drug population. The

PPV of D is 0.98 and the PPV of E is 0.97. Both values are very high and comparable to each other. Another factor will need to be considered to determine which strategy is better.

The number of subjects per RCT will be used as the next metric. The higher this value, the better the strategy. This is because we are more certain of what we know for an approved drug if there were more data being used to study the drug. The following table has these values:

	D	E
N per adopt	1020	1180

As we can observe in the table, E has a higher amount of subjects within each sample, and is therefore a better strategy than D because we are more certain of the knowledge we obtain from each sample.

9. The above exercises considered "drug discovery" with randomized clinical trials. What additional issues have to be considered when we are using observational data to explore and try to confirm risk factors for particular diseases?

When attempting to confirm risk factors for a disease using observational data, some issues that need to be addressed concern type I error and power. These become an issue when performing adaptive clinical trials; modifying the hypothesis in-between confirmatory RCT.

If the hypothesis is modified during confirmatory RCT, this will inflate our type I error. This increase is due to comparing more things to each other. For example, if for each combination, we have a 5% chance of a type I error, then an increase in the combinations will increase our type I error.

This can be combated to a certain extent by increasing the power of our test. However, this can be expensive to perform and is sometimes impossible to cancel out the effects of type I error completely. For example, if our type I error doubles, and our power is at 80%, we cannot double the power to 160% because it would lose its mathematical meaning.

2.5/10 good job on statistical principle of reliably identify risk factors of disease, confirming drug toxicity or testing scientific hypothesis are the same.

the other 3 issues in the key were not addressed.