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

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-β (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-α quantile of the standard normal distribution such that Pr($Z \le z_{1-\alpha}$) = 1 - α.

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

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

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 α = 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($Z \le c$) for arbitrary choices of *c* using statistical software. The commands to obtain Pr($Z \le c$) 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 π = 0.10), and all those truly effective drugs provide the same degree of benefit θ = 1 to patients with disease X. The other 90% of our ideas correspond to drugs that provide no benefit to the patients (so θ = 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 α = 0.025 and a power of 97.5% (so β = 0.025) under the alternative hypothesis that the true treatment effect is θ = 1.
   a. What sample size *n* will be used in each RCT? **979**

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

   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

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

2. (B: Pivotal) Suppose we choose a type I error of α = 0.025 and a power of 80.0% (so β = 0.20) under the alternative hypothesis that the true treatment effect is θ = 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? __22___
   g. How many of the tested drugs will have significant results? ___102_
   h. What proportion of the drugs with significant results will be truly beneficial?_0.7843

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? _1269__
   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? _1142__
   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? _0.6415

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

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? _24.0%_
   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? _0.5153

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? _96.7%_
   d. How many confirmatory RCTs will be for truly beneficial drugs? __84___
   e. How many of the tested beneficial drugs will have significant results? __81___
   f. How many confirmatory RCTs will be for truly ineffective drugs? __79___
   g. How many of the tested ineffective drugs will have significant results? ___2___
   h. How many of the tested drugs will have significant results? __83___
   i. What proportion of the drugs with significant results will be truly beneficial? _0.9759

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?_0.4860

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? _95.2%_
    d. How many confirmatory RCTs will be for truly beneficial drugs? __87___
    e. How many of the tested beneficial drugs will have significant results? __83___
    f. How many confirmatory RCTs will be for truly ineffective drugs? __92___
    g. How many of the tested ineffective drugs will have significant results? ___2___
    h. How many of the tested drugs will have significant results? __85___
    i. What proportion of the drugs with significant results will be truly beneficial?_0.9765

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?

    The strategy in problems 6 & 7 results in a similar number of positive findings and proportion of truly beneficial drugs as the strategy in problems 4 & 5. Both strategies have a much higher proportion of truly beneficial drugs than any of the pivotal strategies. I would prefer the strategy in problems 6 & 7, as fewer truly beneficial ideas are discarded, and presumably fewer resources for basic science research would be wasted.

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?

    We should be concerned about the negative predictive value (in this case, the proportion of negative findings in which the studied risk factor truly wasn’t associated with the disease) as *well as specificity, sensitivity and positive predictive value. Although the latter are all important, as discussed in this homework, failing to find evidence of association between a disease and a true risk factor also has potentially harmful consequences beyond just that of wasted resources.*