Exercises 9 · Parametric inference

Due Monday, April 18, 2016

(1) The power function of a test

Consider the following (highly stylized) hypothesis testing probem. Suppose that we observe a binomial random variable $X \sim \text{Binomial}(n, w)$ for sample size n and success probability w. For example, X might be the number of heads we see in in n flips of a coin. The goal is to test whether the null hypothesis that $w = 0.5$ is plausible in light of the data.

(A) Suppose that $n = 100$. What should our rejection region R be if we want to conduct a Neyman–Pearson test at the $\alpha = 0.05$ level, using X as a test statistic? The R function `qbinom` will help here. Try reading up on the help file using `?qbinom`

(B) Suppose that the true success probability is actually $w = 0.55$, so that $X \sim \text{Binomial}(n = 100, w = 0.55)$. How likely is X to fall in the rejection region you calculated in Part A? This quantity $P(X \in R \mid w = 0.55)$ is called the *power* of the test at the alternatve hypothesis $w = 0.55$, because it is the chance we will reject the null hypothesis (which we want to do when $w = 0.55$, because the null hypothesis is false.) Note: you can calculate this using probability theory, but you might find it easier to approximate it using Monte Carlo simulation; R's `rbinom` function will allow you to simulation binomial random variables if you decide to take this route.

(C) Now consider range of alternative w values spanning the interval $(0, 1)$. For example, you can create a sequence $0, 0.01, 0.02, \ldots, 0.99, 1)$ using the R function `seq(0,1,by=0.01)`. For each w in this range, calculate the power $P(X \in R \mid w)$ and plot the power as a function of w. This is called a power curve or a power function.

(D) Use the technique you developed in Part C to determine how big (approximately) the sample size n must be in order for you to have an 80% chance of rejecting the null hypothesis of $w = 0.5$ if the truth is $w = 0.55$. What if the true w is actually 0.51, making it nearly indistinguishable from the null hypothesis?

(2) Cherry picking, a.k.a. multiple testing

(A) Consider the general problem of testing whether two population proportions are different on the basis of observed sample propor-

tions. That is: $x_1 \sim \text{Binomial}(n_1, w_1)$ and $x_2 \sim \text{Binomial}(n_2, w_2)$, and we want to know whether the difference $\Delta = w_1 - w_2 = 0$. Calculate (by hand) the mean and variance of the sampling distribution of the estimator $\hat{\Delta} = \hat{w}_1 - \hat{w}_2$, where $\hat{w}_i = x_i/n_i$ for $i = 1, 2$, assuming the null hypothesis that $\Delta = 0$. Show your work.

(B) Suppose we genotype $n_1 = 57$ people who possess some particular binary phenotypic trait (group 1). We also genotype $n_2 = 63$ people without the trait (group 2). The trait is complex, so many genes affect it. But we have a particular favorite gene that we believe may play a role in the trait. Thus for each sample, we calculate the number of people who are homozygous dominant (AA) at this locus in the genome. We find that there are $x_1 = 23$ people in group 1 who are AA, while there are $x_2 = 10$ people in group 2 who are AA. Choose an α level and use your result above to test whether we can reject the null hypothesis that the proportion of AA genotypes is the same in the two underlying subpopulations. You may use the large-sample theory here: i.e. assuming that the central limit theorem has kicked in, and that the ratio $z = \hat{\Delta}/\text{SE}(\hat{\Delta})$ has an asymptotic normal distribution under the null hypothesis.

(C) The above questions concern an idealized setting where we test a single pre-specified hypothesis about the difference between two proportions. What happens, however, when we use the same data both to generate *and* test a hypothesis?

Consider a modification to the above testing problem. Suppose that for each group, we compute the proportion of homozygous dominant alleles at 5 different loci (versus a single locus in the problem above). That is, of the n_1 people in group 1, x_{11} are homozygous dominant at locus 1, x_{12} are homozygous dominant at locus 2, and so forth. Similar, of the n_2 people in group 2, x_{21} are homozygous dominant at locus 1, x_{22} are homozygous dominant at locus 2, and so forth for all 5 loci.

For each allele, we compute the z-score corresponding to a test for a difference in proportions. Then we pick the single largest z score across the five loci and check whether it falls in our rejection region corresponding to $\alpha = 0.05$. If it does, we claim we've found a significant genomic predictor of the phenotype.[1]

(i.) Set up a Monte Carlo simulation (or just use probability theory directly) to assess the probability that this procedure generates a false positive under the "global" null hypothesis that

[1] You'll have to assume something here about the population-wide proportion of homozygous-dominant alleles at each locus (e.g. 25% regardless of whether one has the binary trait.) Just be clear about whatever assumptions you're making.

at each of the five alleles, the population proportion of homozygous dominant genotypes is no different in population 1 (has trait) than in population 2 (doesn't have trait). That is, the null hypothesis for each individual test of proportions is assumed true.

(ii.) Above you looked for associations of genotype with phenotype at $K = 5$ loci. Find (approximately) the smallest value of K for which the actual probability of a false positive exceeds 80% under the global null hypothesis. Comment on the following statement in light of your findings: using the same data set both to generate and to test hypotheses is dangerous.

(3) Uncertainty quantification using normality

Suppose, as Gauss did 200 years ago, that the residuals in a simple linear regression model follow a normal distribution with mean 0 and variance σ^2:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2)\,. \end{aligned}$$

This means we have specified a normal likelihood for the data, given the parameters. That is, for a data set of size n,

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \phi(y_i - \beta_0 - \beta_1 x_i \mid 0, \sigma^2)\,,$$

where $\phi(t \mid m, v)$ is the notation typically used to denote the probability density function of the normal distribution having mean m and variance v, evaluated at the point t.

Your task: use these assumptions to prove the explicit formulas given in the course packet for the mean and variance of the sampling distribution for the maximum likelihood estimators of β_0 and β_1. Recall that these are the same as the least-squares estimators:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\,. \end{aligned}$$

That is, there are four important numbers here: the mean and variance for the sampling distribution of two different quantities. You should give these expressions in terms of the observed data, the sample size, and the unknown parameters β_0, β_1, and σ^2. Remember that the design points x_i are given—that is, they are constant, not random. The only source of randomness is the residuals ϵ_i.

If you feel stuck, try assuming that the true $\beta_0 = 0$. This will simplify matters.