BST 676 – Spring 2011 – Dr. Charnigo

Unit IV: Techniques for Hypothesis Testing

a. Motivating Case Study #1: Testing whether a drug relieves pain in a very small sample

Suppose that, as part of a Phase II clinical trial, a pharmaceutical company administers a drug intended to relieve pain to 20 otherwise healthy subjects experiencing headaches. Suppose, moreover, that 8 of the subjects find the drug at least somewhat effective, while 12 of the subjects find the drug ineffective.

Now imagine that the pharmaceutical company representative has asked you whether the drug may be effective in 70% of headache sufferers or whether a 70% effectiveness figure is ruled out by the data. In other words, was the pharmaceutical company just a bit unlucky with these 20 subjects, or does the drug really not perform too well?

You can formalize the question by defining p as the proportion of headache sufferers for whom the drug is effective and proposing a hypothesis test of $H_0 : p = 0.70$ against $H_1 : p \neq 0.70$. Your initial instinct may be to calculate

$$z := \frac{8/20 - 0.70}{\sqrt{0.70(1 - 0.70)/20}} = -2.93$$

and, noting that this is less than -1.96, tell the representative that the drug does not perform too well. However, a moment's thought reveals a problem with this inference:

So now you need to find another way to test $H_0 : p = 0.70$ against $H_1 : p \neq 0.70$. To complicate matters further, the representative has just sent you a text message asking in what percentage of headache sufferers the drug may be effective, if a 70% effectiveness figure is ruled out by the data. You think about constructing a confidence interval, but you are uncomfortable reporting

$$8/20 \pm 1.96\sqrt{8/20(1 - 8/20)/20} = 0.40 \pm 0.21$$

for much the same reason that you were uncomfortable with the z statistic.

We will return to this motivating case study at the end of Unit IV.

b. Motivating Case Study #2: Why are there three tests for the global null in PROC LOGISTIC?

The data set in {NonSmokers.xls} is drawn from the National Center for Health Statistics Public-Use Perinatal Mortality Data Files and pertains to infants born to non-smoking mothers in the years 2001 and 2002.

I used PROC LOGISTIC in Version 9.1 of SAS to fit the model

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \alpha + \beta x,$$

where x represents the birthweight of an infant and $p(x)$ represents the risk of mortality for an infant of birthweight x.

The results are shown in {LogisticOutput.rtf}. Having used PROC LOGISTIC before, you are familiar with how the output is organized. In particular, you are not surprised to see that the section labeled "Testing Global Null Hypothesis" actually presents three results for the hypothesis test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$.

In this example, all three of the results are in qualitative agreement, yielding p-values less than 0.0001. This provides near definitive confirmation of an association between birthweight and mortality.

Yet, you may wonder: what is a likelihood ratio test? what is a score test? what is a Wald test?

We will return to this motivating case study at the end of Unit IV. However, let us now briefly pursue an aside of interest to data analysts.

You will notice that the odds ratio estimate is 0.998, which *seems* very close to the neutral value of 1. One may wonder, then, whether the association between birthweight and mortality is of any practical importance.

To get some idea, we can interpret the 0.998 as the estimated factor by which the odds of mortality are multiplied when birthweight increases by 1 gram. Yet, one can hardly argue that 2401 grams is importantly different from 2400 grams. One may then ask by what estimated factor the odds of mortality are multiplied when birthweight increases by, say, 100 grams. The answer is and illustrates the principle that odds ratio estimates for continuous explanatory variables in a logistic regression model should not necessarily be judged by their proximity to the neutral value of 1.

c. Exact tests

An exact test is one for which the significance level is exactly equal to that which is claimed. A few examples will clarify this point.

Example #1. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with both $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$ unknown. For fixed $\mu_0 \in (-\infty, \infty)$, consider testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$. We learn in an introductory methods course that we can define

$$T := \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

and reject H_0 at significance level α if $T > t_{n-1,1-\alpha}$, the $(1-\alpha)$ quantile of a T distribution on $(n-1)$ degrees of freedom.

To understand this testing procedure, recall that

$$U := \frac{\bar{X} - \mu}{S/\sqrt{n}} = T + \frac{\mu_0 - \mu}{S/\sqrt{n}}$$

follows the T distribution on $(n-1)$ degrees of freedom. If H_0 is true and $\mu = \mu_0$, then $U = T$ and we have

$$P_\mu(T > t_{n-1,1-\alpha}) = P_\mu(U > t_{n-1,1-\alpha}) = \alpha,$$

so that the probability of incorrectly rejecting H_0 is exactly α. Hence, this testing procedure is exact.

Some further remarks are warranted about this example.

Remark #1. If H_0 is true but $\mu < \mu_0$, then we have

$$P_\mu(T > t_{n-1,1-\alpha}) = P_\mu\left(U + \frac{\mu - \mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}\right) < P_\mu(U > t_{n-1,1-\alpha}) = \alpha,$$

so that the probability of incorrectly rejecting H_0 is less than α. This does not violate the principle of exactness because, when a null hypothesis contains an interval — such as $(-\infty, \mu_0]$ in this example — rather than a single point, the significance level is defined as the maximum probability of incorrectly rejecting the null hypothesis. Symbolically, we could write

$$\sup_{\mu \in (-\infty, \mu_0]} P_\mu(T > t_{n-1,1-\alpha}) = \alpha$$

for this example. The symbol "sup" stands for supremum, which equals the maximum of a set when a maximum exists and which equals the least upper bound of a set when a maximum does not exist.

Remark #2. If H_0 is false and $\mu > \mu_0$, then we have

$$P_\mu(T > t_{n-1,1-\alpha}) = P_\mu\left(U + \frac{\mu - \mu_0}{S/\sqrt{n}} > t_{n-1,1-\alpha}\right) > P_\mu(U > t_{n-1,1-\alpha}) = \alpha,$$

so that the probability of correctly rejecting H_0 — also called the power of the test — is greater than α. A testing procedure for which the probability of rejecting H_0 is greater for any point in the alternative hypothesis than for any point in the null hypothesis is called unbiased.

Example #2. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mu^{-1}\exp[-x\mu^{-1}]1_{\{x>0\}}$ with $\mu \in (0,\infty)$ unknown. For fixed $\mu_0 \in (0,\infty)$, consider testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.

Note that $\sum_{i=1}^n X_i$ has the gamma distribution with shape n and scale μ, so that $\sum_{i=1}^n X_i/\mu$ has the gamma distribution with shape n and scale 1 (verify this yourself at home, for practice). Let $g_{n,1-\alpha}$ denote the $(1-\alpha)$ quantile of the gamma distribution with shape n and scale 1. Put $T := \sum_{i=1}^n X_i/\mu_0$ and consider rejecting H_0 at level α if $T > g_{n,1-\alpha}$. If H_0 is true and $\mu = \mu_0$, then

$$P_\mu\left(\sum_{i=1}^n X_i/\mu > g_{n,1-\alpha}\right) = P_\mu(T > g_{n,1-\alpha}) = \alpha.$$

Hence, the testing procedure proposed in this paragraph is exact.

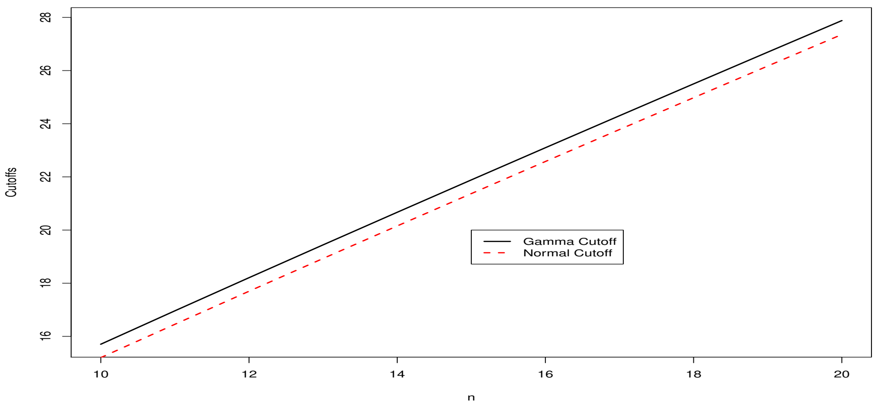
On the other hand, we can appeal to the Central Limit Theorem to conclude that $\sum_{i=1}^n X_i$ is approximately normally distributed with mean $n\mu$ and variance $n\mu^2$. Consider rejecting H_0 at level α if $T > n + \sqrt{n}z_{1-\alpha}$. If H_0 is true and $\mu = \mu_0$, then T is approximately normally distributed with mean n and variance n, so that

$$P_\mu(T > n + \sqrt{n}z_{1-\alpha}) = P_\mu\left(\frac{T-n}{\sqrt{n}} > z_{1-\alpha}\right) \approx \alpha.$$

The testing procedure proposed in this paragraph is not exact. Exactness would require that $n + \sqrt{n}z_{1-\alpha} = g_{n,1-\alpha}$, which is not in general true.

Below is a picture of $n + \sqrt{n} z_{0.95}$ and $g_{n,0.95}$ as a function of $n \in \{10, 11, \ldots, 20\}$. We see that the critical values from the approximate testing procedure are slightly too small, so that the probability of incorrectly rejecting the null hypothesis will be slightly

Figure 1:



Example #3. One of the most famous exact tests is Fisher's, which is used to analyze data from a 2×2 contingency table when one or more of the cells has a small count. Fisher's exact test is routinely employed to analyze adverse event data in Phase III clinical trials since some adverse events are quite rare. Importantly, using Fisher's exact test instead of the chi-square test is never wrong. The chi-square test is simply an approximate procedure that works well when none of the cells has a small count.

	Experimental Group	Control Group	Row Total
Adverse Event	a	b	M_1
No Adverse Event	c	d	M_2
Column Total	N_1	N_2	

The idea behind Fisher's exact test is as follows. If we consider all possible tables with the same row and column totals as the table actually observed, then we can assign a probability to each such table from the hypergeometric distribution. More specifically, let X be a discrete random variable with probablility mass function

$$f(x) := \frac{N_1!\ N_2!\ M_1!\ M_2!}{(N_1+N_2)!\ x!\ (N_1-x)!\ (M_1-x)!\ (M_2-N_1+x)!} \tag{1}$$

for any $x \in \{0, 1, \ldots, \min\{M_1, N_1\}\}$. Consider a table having first entry $x \in \{0, 1, \ldots, \min\{M_1, N_1\}\}$ but the same row and column totals as the table actually observed. Assign this table a probability of $f(x)$.

To test the null hypothesis that adverse events are no more likely under experimental conditions than under control conditions, calculate the probability associated with tables at least as extreme as the one actually observed:

$$f(a) + f(a+1) + \cdots + f(\min\{M_1, N_1\}). \tag{2}$$

We interpret (2) as a p-value, and so we reject the null hypothesis if (2) is less than a specified significance level. For a two-sided test, various definitions have been proposed for the p-value. One of them is

$$\sum_{x \in \{0,1,\ldots,\min\{M_1,N_1\}\}: f(x) \le f(a)} f(x). \tag{3}$$

In words, (3) is the probability associated with tables that are no more likely than the one actually observed.

d. Rank-based tests

Let n be a positive integer; let a and z be real-valued functions with domain $\{1, 2, \ldots, n\}$; and let $R_1, R_2, \ldots, R_n$ denote a random ordering of $\{1, 2, \ldots, n\}$. For example, if $n = 3$, then there are six equally likely possibilities:

$$R_1 = 1, R_2 = 2, R_3 = 3; \quad R_1 = 1, R_2 = 3, R_3 = 2;$$

$$R_1 = 2, R_2 = 1, R_3 = 3; \quad R_1 = 2, R_2 = 3, R_3 = 1;$$

$$R_1 = 3, R_2 = 1, R_3 = 2; \quad R_1 = 3, R_2 = 2, R_3 = 1.$$

Put

$$S := \sum_{j=1}^{n} a(j) z(R_j).$$

Using methods beyond the scope of BST 676, one can show that

$$T := \frac{S - n\bar{z}\bar{a}}{\sqrt{(n-1)^{-1}\sum_{j=1}^{n}[z(j)-\bar{z}]^2 \sum_{j=1}^{n}[a(j)-\bar{a}]^2}} \xrightarrow{L} N(0,1) \tag{4}$$

as $n \to \infty$, where $\bar{a} := n^{-1}\sum_{j=1}^{n} a(j)$ and $\bar{z} := n^{-1}\sum_{j=1}^{n} z(j)$, provided that

$$\frac{n \max_{j\in\{1,\ldots,n\}}[z(j)-\bar{z}]^2 \max_{j\in\{1,\ldots,n\}}[a(j)-\bar{a}]^2}{\sum_{j=1}^{n}[z(j)-\bar{z}]^2 \sum_{j=1}^{n}[a(j)-\bar{a}]^2} \to 0. \tag{5}$$

Result (4) is often used to justify rank-based tests, which are employed when we are reluctant to assume that data are normally distributed. For example, suppose that we observe continuous data $X_1, \ldots, X_{n_1}$ in an experimental group and continuous data $Y_1, \ldots, Y_{n_2}$ in a control group. If we suspect that the data are not normally distributed, then we may be reluctant to use a two-sample T test to compare groups. In this case, we may consider using the Wilcoxon rank sum test (sometimes called the Mann Whitney U test) to compare groups.

The Wilcoxon rank sum test entails ranking the observations in order from smallest to largest, calculating the sum of the ranks in the experimental group, and then deciding whether the sum of the ranks is significantly different from what would be expected if the two groups had a common probability density function.

To see how the Wilcoxon rank sum test emerges from result (4), put $n := n_1 + n_2$, $z(j) := j$, and $a(j) := 1_{\{j\in\{1,\ldots,n_1\}\}}$. If the two groups have a common probability density function, then the ranks $R_1, R_2, \ldots, R_n$ of the n subjects are a random ordering of $\{1, 2, \ldots, n\}$. We interpret $z(R_j)$ as the rank assigned to subject j and $a(j)$ as an indicator of whether subject j belongs to the experimental group. Thus, $S := \sum_{j=1}^{n} a(j)z(R_j)$ is the sum of the ranks in the experimental group.

We obtain a rank-based test with approximate significance level α by rejecting the null hypothesis of a common probability density function if $|T| > z_{1-\alpha/2}$, since T has an approximate standard normal distribution when the null hypothesis is true. Of course, if the null hypothesis is false, then the distribution of T may not be approximately standard normal. In fact, we count on the distribution of T not being approximately standard normal when the null hypothesis is false, because then we hope that $P(|T| > z_{1-\alpha/2})$ is much greater than α.

To relate the Wilcoxon rank sum test as described above to what one might see in an introductory methods textbook, we must evaluate some quantities in (4). Using the well-known results that $\sum_{j=1}^n j = n(n+1)/2$ and $\sum_{j=1}^n j^2 = n(n+1)(2n+1)/6$, we can show (do so at home, for practice) that

$$\bar{a} = n_1/n; \quad \bar{z} = n^{-1}\sum_{j=1}^n j = (n+1)/2;$$

$$\sum_{j=1}^n [a(j) - \bar{a}]^2 = n_1 - n_1^2/n;$$

$$\sum_{j=1}^n [z(j) - \bar{z}]^2 = n(n+1)(2n+1)/6 - n(n+1)^2/4 = n(n-1)(n+1)/12.$$

Hence, we have

$$\begin{aligned} T &= \frac{S - n\bar{z}\bar{a}}{\sqrt{(n-1)^{-1}\sum_{j=1}^n [z(j)-\bar{z}]^2 \sum_{j=1}^n [a(j)-\bar{a}]^2}} \\ &= \frac{S - n(n_1/n)(n+1)/2}{\sqrt{(n-1)^{-1} n_1(1-n_1/n)n(n-1)(n+1)/12}} \\ &= \frac{S - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2(n_1+n_2+1)/12}}. \end{aligned}$$

After all of this work, we still face one nagging question. Is (4) justified? We do not know yet, because so far we have not checked condition (5).

Noting that

$$\max_{j\in\{1,\ldots,n\}} [z(j) - \bar{z}]^2 \le \max_{j\in\{1,\ldots,n\}} z(j)^2 = z(n)^2 = n^2$$

and

$$\max_{j\in\{1,\ldots,n\}} [a(j) - \bar{a}]^2 \le \max_{j\in\{1,\ldots,n\}} a(j)^2 = a(1)^2 = 1,$$

we find that

$$\frac{n \max_{j\in\{1,\ldots,n\}}[z(j)-\bar{z}]^2 \max_{j\in\{1,\ldots,n\}}[a(j)-\bar{a}]^2}{\sum_{j=1}^n [z(j)-\bar{z}]^2 \sum_{j=1}^n [a(j)-\bar{a}]^2} \le \frac{n^3}{n(n-1)(n+1)n_1(1-n_1/n)/12}.$$

So, condition (5) holds if $n_1/n \to c \in (0,1)$ as $n \to \infty$. (What does that mean?)

e. Likelihood ratio tests (Cf. pp. 462-466 of Larsen and Marx)

When we use maximum likelihood for point estimation, there is a natural companion testing procedure called a likelihood ratio test. Suppose that $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ for some $\theta \in \Theta$. Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$, where $\Theta_0 \subset \Theta$. Let $L(\zeta; \mathbf{x})$ denote the likelihood as a function of $\zeta \in \Theta$. We define the likelihood ratio test statistic as

$$\lambda := \frac{\sup_{\zeta\in\Theta_0} L(\zeta;\mathbf{x})}{\sup_{\zeta\in\Theta} L(\zeta;\mathbf{x})} = \frac{L(\hat{\theta}_0;\mathbf{x})}{L(\hat{\theta};\mathbf{x})},$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ and $\hat{\theta}_0$ maximizes the likelihood over $\zeta \in \Theta_0$. We reject H_0 if $\lambda < c$, where $c \in (0, 1)$ is chosen so that the testing procedure has the desired significance level.

Example #4. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$ for some $\theta \in \Theta := \mathbb{R}$. Let θ_0 be an element of $\mathbb{R}$, and consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The likelihood function is

$$L(\zeta;\mathbf{x}) = (2\pi)^{-n/2} \exp\left[-\sum_{i=1}^{n}(x_i - \zeta)^2/2\right].$$

The maximum likelihood estimate $\hat{\theta}$ is readily seen to be $\bar{x}$, the sample mean (verify yourself at home, for practice). Since Θ_0 consists of the single point $\{\theta_0\}$, we trivially have $\hat{\theta}_0 = \theta_0$. Thus, the likelihood ratio test statistic is

$$\begin{aligned}
\lambda &= \frac{L(\theta_0;\mathbf{x})}{L(\bar{x};\mathbf{x})} \\
&= \frac{\exp[-\sum_{i=1}^{n}(x_i-\theta_0)^2/2]}{\exp[-\sum_{i=1}^{n}(x_i-\bar{x})^2/2]} \\
&= \frac{\exp[-\sum_{i=1}^{n}x_i^2/2 + \theta_0\sum_{i=1}^{n}x_i - n\theta_0^2/2]}{\exp[-\sum_{i=1}^{n}x_i^2/2 + \bar{x}\sum_{i=1}^{n}x_i - n\bar{x}^2/2]} \\
&= \exp[(\theta_0-\bar{x})n\bar{x} - n(\theta_0-\bar{x})(\theta_0+\bar{x})/2] \\
&= \exp[-n(\theta_0-\bar{x})^2/2].
\end{aligned}$$

Noting that $\lambda < c$ if and only if $n(\theta_0 - \bar{x})^2 > -2\log c$, we can readily choose c so that the testing procedure has the desired significance level. If the null hypothesis is true, then $n(\theta_0 - \bar{X})^2 \sim \chi_1^2$, so we should equate $-2\log c$ to $\chi^2_{1,1-\alpha}$. This yields $c =$

Example #5. Let $X_1, X_2, \ldots, X_n \stackrel{iid}{\sim} \theta^{-1}\exp[-x\theta^{-1}]1_{\{x>0\}}$ for some $\theta \in \Theta := (0,\infty)$. Let θ_0 be an element of $(0,\infty)$, and consider testing $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$. The likelihood function is

$$L(\zeta; \mathbf{x}) = \zeta^{-n} \exp\left[-\sum_{i=1}^{n} x_i \zeta^{-1}\right].$$

The maximum likelihood estimate $\hat{\theta}$ is readily seen to be $\bar{x}$, the sample mean (verify yourself at home, for practice). Now Θ_0 is the interval $(0, \theta_0]$, so we have to be a little more careful in identifying $\hat{\theta}_0$.

Consider two cases. If $\bar{x} \in \Theta_0$, then $\hat{\theta}_0 = \bar{x}$. (Why?) If $\bar{x} \notin \Theta_0$, then note that the log likelihood

$$-n \log \zeta - \sum_{i=1}^{n} x_i \zeta^{-1}$$

has derivative

$$-n/\zeta + \sum_{i=1}^{n} x_i/\zeta^2,$$

which is positive for all ζ less than its unique root $\bar{x}$. Since $\bar{x} > \theta_0$, the derivative must be positive for all ζ less than or equal to θ_0. This implies that the log likelihood is increasing on Θ_0, so that $\hat{\theta}_0 = \theta_0$.

Thus, the likelihood ratio test statistic equals 1 if $\bar{x} \in \Theta_0$ and otherwise equals

$$\frac{\theta_0^{-n}\exp[-\sum_{i=1}^{n} x_i \theta_0^{-1}]}{\bar{x}^{-n}\exp[-\sum_{i=1}^{n} x_i\ \bar{x}^{-1}]} = \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left[-\sum_{i=1}^{n} x_i \theta_0^{-1} + n\right].$$

We are not finished yet, though, because we have not described how to choose the critical value c. This example is more difficult than the preceding one because we will not be able to find a simple formula for c. However, some probabilistic reasoning and computing will aid us.

Put $T := \sum_{i=1}^{n} X_i/\theta_0$. If the null hypothesis is true and $\theta = \theta_0$, then T has the gamma distribution with shape n and scale 1. For any positive integer n and $c \in (0,1)$, we can estimate the significance level of the likelihood ratio test when c is used as the critical value,

$$P_\theta[\lambda < c] = P_\theta[(T/n)^n \exp[-T+n] < c\ \cap\ T > n], \tag{6}$$

by simulation. For instance, the R code below estimates (6) when $n = 20$ and $c = 0.10$.

```
T <- rgamma(10000,shape=20,scale=1)
sum( ((T/20)^20 * exp(-T+20) < 0.10)*(T > 20) ) / 10000
```

By trial and error, one can experiment with c until one finds the critical value at which the desired significance level is attained.

f. Wald tests

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with probability mass or density function $f(x;\theta)$ for an unknown parameter θ in a known parameter space $\Theta \subset \mathbb{R}$. Suppose, moreover, that $\hat{\theta}_n$ is a consistent sequence of estimators of θ and that there exists a sequence of positive numbers σ_n tending to 0 such that

$$\frac{\hat{\theta}_n - \theta}{\sigma_n} \xrightarrow{L} N(0,1).$$

Let θ_0 be a fixed point in the interior of Θ, and consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. If we put $Z := (\hat{\theta}_n - \theta_0)/\sigma_n$ and reject H_0 when $|Z| > z_{1-\alpha/2}$, then we have a test with approximate significance level α because

$$P_\theta(|Z| > z_{1-\alpha/2}) = P_\theta(Z > z_{1-\alpha/2}) + P_\theta(Z < -z_{1-\alpha/2}) \approx \alpha/2 + \alpha/2 = \alpha$$

under H_0, because Z has an approximate standard normal distribution under H_0.

What if H_0 is false? Then Z does not have an approximate standard normal distribution. Rather, $Z + (\theta_0 - \theta)/\sigma_n$ has an approximate standard normal distribution, and so

$$\begin{aligned}
& P_\theta(|Z| > z_{1-\alpha/2}) \\
= \; & P_\theta(Z > z_{1-\alpha/2}) + P_\theta(Z < -z_{1-\alpha/2}) \\
= \; & P_\theta\left(Z + \frac{\theta_0 - \theta}{\sigma_n} > z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right) + P_\theta\left(Z + \frac{\theta_0 - \theta}{\sigma_n} < -z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right) \\
\approx \; & 1 - \Phi\left[z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right] + \Phi\left[-z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right], \qquad (7)
\end{aligned}$$

where Φ denotes the standard normal cumulative distribution function.

Two observations can be made. First, if $\theta_0 - \theta < 0$, then $(\theta_0 - \theta)/\sigma_n \to -\infty$ so that expression (7) approaches 1. (How do we know this?) Likewise, if $\theta_0 - \theta > 0$, then $(\theta_0 - \theta)/\sigma_n \to +\infty$ so that expression (7) again approaches 1. Thus, if H_0 is false, the probability of its correct rejection — i.e., the power of the testing procedure — approaches 1 as $n \to \infty$. Such a testing procedure is called consistent.

Second, consider minimizing expression (7) with respect to θ. Differentiating in θ and setting the result to 0 yields

$$\phi\left[z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right] /\sigma_n - \phi\left[-z_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma_n}\right] /\sigma_n = 0, \tag{8}$$

where ϕ denotes the standard normal probability density function. By symmetry of the standard normal probability density function, the only way for condition (8) to hold is if

$$z_{1-\alpha/2} + (\theta_0 - \theta)/\sigma_n = -(-z_{1-\alpha/2} + (\theta_0 - \theta)/\sigma_n) = z_{1-\alpha/2} - (\theta_0 - \theta)/\sigma_n, \tag{9}$$

which implies that $\theta = \theta_0$. Thus, the approximate probability of rejecting H_0 is smaller when H_0 is true than when H_0 is false. Such a testing procedure is called asymptotically unbiased. The adjective asymptotically is necessary because the probabilistic statements in (7) are not exact.

Apart from its inexactness, the above testing procedure seems reasonable. However, a practical impediment is that σ_n may depend on the unknown parameter θ. For example, suppose that $f(x;\theta) := \theta^x(1-\theta)^{1-x}$ for $x \in \{0,1\}$ and $\theta \in \Theta := (0,1)$. Then $\hat{\theta}_n := n^{-1}\sum_{i=1}^n X_i$ and $\sigma_n := \sqrt{\theta(1-\theta)/n}$ meet all requirements stated above. Unfortunately, σ_n and hence Z cannot be calculated.

One solution to this dilemma is to find $\hat{\sigma}_n$ so that $\sigma_n/\hat{\sigma}_n$ converges in probability to 1 and then put $T := \sigma_n/\hat{\sigma}_n \times Z$. An approximate standard normal distribution for T under H_0 is justified by Slutsky's Theorem #3, and so we reject H_0 if $|T| > z_{1-\alpha/2}$. This is referred to as a Wald test. For the above example, we may take $\hat{\sigma}_n := \sqrt{\hat{\theta}_n(1-\hat{\theta}_n)/n}$.

g. Score tests

Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed with probability mass or density function $f(x;\theta)$ of the exponential family form, where the unknown parameter θ belongs to a known parameter space $\Theta \subset \mathbb{R}$.

The score statistic is defined as

$$S(\theta;\mathbf{X}) := \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i;\theta).$$

We have

$$E_\theta[S(\theta;\mathbf{X})] = \int_{\mathbb{R}} \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\{\log f(x;\theta)\}\, f(x;\theta)\, dx = \frac{d}{d\theta} \sum_{i=1}^{n} \int_{\mathbb{R}} f(x;\theta)\, dx =$$

if $X_1, X_2, \ldots, X_n$ are continuous, and a similar argument with summations replacing integrations works if they are discrete. Moreover,

$$Var_\theta[S(\theta;\mathbf{X})] = \sum_{i=1}^{n} Var_\theta \left[\frac{\partial}{\partial\theta} \log f(X_i;\theta) \right] =$$

Let θ_0 be a fixed point in the interior of Θ, and consider testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. If we put $Z := S(\theta_0;\mathbf{X})/\sqrt{J_n(\theta_0)}$ and reject H_0 when $|Z| > z_{1-\alpha/2}$, then we have a test with approximate significance level α. We refer to this as a score test.

To illustrate a score test, suppose that $f(x;\theta) := \theta^x(1-\theta)^{1-x}$ for $x \in \{0,1\}$ and $\theta \in \Theta := (0,1)$. We have

$$\begin{aligned}
S(\theta_0;\mathbf{X}) &= \sum_{i=1}^{n} \frac{\partial}{\partial\theta}[X_i \log\theta + (1-X_i)\log(1-\theta)]|_{\theta=\theta_0} \\
&= \frac{\sum_{i=1}^{n} X_i}{\theta_0} - \frac{n - \sum_{i=1}^{n} X_i}{1-\theta_0} \\
&= \frac{(1-\theta_0)\sum_{i=1}^{n} X_i - n\theta_0 + \sum_{i=1}^{n} X_i\theta_0}{\theta_0(1-\theta_0)} \\
&= \frac{n^{-1}\sum_{i=1}^{n} X_i - \theta_0}{n^{-1}\theta_0(1-\theta_0)}
\end{aligned}$$

and

$$J_n(\theta_0) = Var_{\theta=\theta_0}\left[\frac{\sum_{i=1}^{n} X_i}{\theta_0(1-\theta_0)} \right] = \frac{n}{\theta_0(1-\theta_0)}.$$

This yields

$$Z := \frac{n^{-1}\sum_{i=1}^n X_i - \theta_0}{n^{-1}\theta_0(1-\theta_0) \times \sqrt{n/[\theta_0(1-\theta_0)]}} = \frac{n^{-1}\sum_{i=1}^n X_i - \theta_0}{\sqrt{n^{-1}\theta_0(1-\theta_0)}},$$

which we recognize from our introductory methods course as the familiar one-sample test for a proportion.

h. Resolution of motivating case studies

Our first motivating case study asked, in the context of a Phase II clinical trial, how to test $H_0 : p = 0.70$ against $H_1 : p \neq 0.70$ using a sample of size 20. Having studied the concept of an exact test, we are now in a position to answer that question.

If the null hypothesis is true, then the probability of obtaining only 8 successes is $\binom{20}{8}(.70)^8(.30)^{12} = 0.0039$. The probability of obtaining any other number of successes between 5 and 20 is recorded in the table below; the probability of obtaining fewer than 5 successes is negligibly small.

Number	Probability	Number	Probability	Number	Probability	Number	Probability
5	0.0000	9	0.0120	13	0.1643	17	0.0716
6	0.0002	10	0.0308	14	0.1916	18	0.0278
7	0.0010	11	0.0654	15	0.1789	19	0.0068
8	0.0039	12	0.1144	16	0.1304	20	0.0008

Since a p-value is the probability of obtaining a result as least as extreme in its opposition to the null hypothesis as the one actually observed, if the null hypothesis were true, all we need to do is decide what constitutes a result at least as extreme as the one actually observed. This is easy for a one-sided alternative hypothesis. For a two-sided alternative hypothesis, one option is to include any result that is no more likely than the one actually observed.

In the present example, that would include the results of 0 through 8 and also the result of 20. If we add up the probabilities associated with these results, we obtain 0.0059. Again, the sum of these probabilities is interpreted as a p-value. So we reject $H_0 : p = 0.70$ in favor of $H_1 : p \neq 0.70$ at significance level 0.05.

We can proceed similarly to test $H_0 : p = p_0$ against $H_1 : p \neq p_0$ for any $p_0 \in (0, 1)$. A 95% confidence interval can then be defined as the set of all

$p_0 \in (0,1)$ such that $H_0 : p = p_0$ is not rejected in favor of $H_1 : p \neq p_0$.

I designed an Excel spreadsheet {ExactCI.xls} to illustrate the computations. Each entry in column B reports the probability of obtaining the corresponding number of successes in column A if $p = 0.21$. Each entry in column C indicates whether the corresponding entry in column B is less than or equal to 0.0282, the probability of obtaining 8 successes if $p = 0.21$. Each entry in column D is the product of the corresponding entries in columns B and C. Adding them up gives the p-value for testing $H_0 : p = 0.21$ against $H_1 : p \neq 0.21$.

Since the p-value of 0.0509 is greater than 0.05, we should include 0.21 in the 95% confidence interval for p. Likewise, we find that 0.20 should not be included, that 0.62 should be included, and that 0.63 should not be included.

Although there is some trial and error in finding the endpoints of the 95% confidence interval, a practical tactic is as follows. You can start with an initial guess of $\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ for the lower endpoint. If you obtain a p-value greater than 0.05, then your next guess should be smaller. (Why?) If you obtain a p-value less than 0.05, then your next guess should be larger. Similarly, you can start with an initial guess of $\hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ for the upper endpoint. If you obtain a p-value greater than 0.05, then your next guess should be larger. If you obtain a p-value less than 0.05, then your next guess should be smaller.

Our second motivating case study asked, in the context of logistic regression, what are a likelihood ratio test, a Wald test, and a score test?

Since presenting the second motivating case study, we have already given a general definition for a likelihood ratio test. To get some insight into what this looks like for logistic regression, let us suppose for simplicity that X is discrete with probability mass function $f_X(x)$. (This is not a terribly unrealistic supposition; naturally continuous variables such as birthweight are discretized, albeit very finely, by rounding.)

The probability that $Y = y$ and $X = x$, where $y \in \{0,1\}$, is

$$P(Y = y|X = x)P(X = x) = \left(\frac{\exp[\alpha + \beta x]}{1 + \exp[\alpha + \beta x]}\right)^y \left(\frac{1}{1 + \exp[\alpha + \beta x]}\right)^{1-y} f_X(x).$$

The corresponding likelihood function for observed data $(X_1, Y_1), \ldots, (X_n, Y_n)$

is

$$L(\zeta_1, \zeta_2; \mathbf{X}, \mathbf{Y}) := \prod_{i=1}^{n} \left(\frac{\exp[\zeta_1 + \zeta_2 X_i]}{1 + \exp[\zeta_1 + \zeta_2 X_i]} \right)^{Y_i} \left(\frac{1}{1 + \exp[\zeta_1 + \zeta_2 X_i]} \right)^{1-Y_i} f_X(X_i). \tag{10}$$

Let $\hat{\alpha}$ and $\hat{\beta}$ denote the values of ζ_1 and ζ_2 that maximize (10). (Practically speaking, these values must be approximated by an iterative algorithm; unlike in linear regression, there are no closed-form expressions for the maximum likelihood estimators in logistic regression.) Let $\hat{\alpha}_0$ denote the value of ζ_1 that maximizes (10) subject to the constraint that $\zeta_2 = 0$.

Then (negative twice) the (log) likelihood ratio test statistic for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ is

$$-2 \log L(\hat{\alpha}_0, 0; \mathbf{X}, \mathbf{Y}) + 2 \log L(\hat{\alpha}, \hat{\beta}; \mathbf{X}, \mathbf{Y}). \tag{11}$$

We will see in Unit V that there is a large-sample justification for taking critical values from the chi-square distribution on one degree of freedom. For now, there is one other point that should be noticed: the $f_X(X_i)$ terms play no role in the maximization of (10) and, in addition, cancel each other out in (11). This explains why we do not need to make any assumptions about the marginal distribution(s) of the predictor(s) in logistic regression.

The Wald test statistic for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ may be defined as either $\hat{\beta}$ divided by its standard error or the square of this quotient. In the former case, a standard normal distribution is used for critical values. In the latter case, a chi-square distribution on one degree of freedom is used. The standard error is derived from the Fisher information matrix, a generalization of the Fisher information from Unit III that applies when θ is a vector.

The score test statistic for $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ is defined as the vector-matrix product

$$\left[\frac{\partial}{\partial \zeta} \log L(\zeta_1, \zeta_2; \mathbf{X}, \mathbf{Y}) \right]^T \left[-E_{\zeta_1, \zeta_2} \left\{ \frac{\partial^2}{\partial \zeta^2} \log L(\zeta_1, \zeta_2; \mathbf{X}, \mathbf{Y}) \right\} \right]^{-1} \left[\frac{\partial}{\partial \zeta} \log L(\zeta_1, \zeta_2; \mathbf{X}, \mathbf{Y}) \right]$$

evaluated at $\zeta_1 = \hat{\alpha}_0$ and $\zeta_2 = 0$. Above, the notation $\frac{\partial}{\partial \zeta}$ represents a 2×1 column vector of first-order partial derivatives with respect to the components of ζ, while the notation $\frac{\partial^2}{\partial \zeta^2}$ represents a 2×2 matrix of second-order partial derivatives with respect to the components of ζ. A chi-square distribution on one degree of freedom is used for critical values.