BST 676 – Spring 2010 – Dr. Charnigo

Unit V: Evaluating Hypothesis Testing Procedures

a. Motivating Case Study #1: Why require such huge sample sizes in clinical trials?

A 1991 *New England Journal of Medicine* paper by Cohn and colleagues titled "A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure" reported that, among 403 male heart failure patients on digoxin and diuretic therapy who received enalapril and among 401 male heart failure patients on digoxin and diuretic therapy who received hydralazine with isosorbide dinitrate, there were 132 and 153 deaths during follow-up respectively.

The fractions $132/403 = 32.75\%$ and $153/401 = 38.15\%$ may be viewed as point estimates of population parameters p_1 and p_2, namely the short-term mortality risks among male heart failure patients on digoxin and diuretic therapy who receive enalparil and hydralazine with isosorbide dinitrate.

One may ask whether the sample sizes of 403 and 401 were adequate. From a methodological point of view, they seem large enough to defend a chi-square test of $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ since

$$403(132/403)(1 - 132/403) = 88.8 >> 10$$

and

$$401(153/401)(1 - 153/401) = 94.6 >> 10.$$

On the other hand, the chi-square test yields a p-value of 0.109, so that the null hypothesis cannot be rejected.

Yet, if the difference between 32.75% and 38.15% is "real", it is clinically important: For every 1000 patients treated using enalapril instead of hydralazine with isosorbide dintrate, we expect that 54 lives will be saved. The ratio $1000/54 = 18.5$ is often referred to as the "Number Needed to Treat" or "NNT".

Hence, the sample sizes of 403 and 401 were inadequate in that they were not large enough to flag a clinically important difference as statistically significant. We will return to this case study at the end of Unit V to address the question of what sample sizes would have been adequate. However, there are some interesting questions that we can answer now.

Discussion Questions.

1. From a methodological point of view, what might be more appropriate than a chi-square test for analyzing the mortality data from this clinical trial?

2. The authors could not reasonably claim that enalapril reduced the risk of mortality, because the observed difference between groups was not statistically significant. However, neither could the authors reasonably claim that enalapril did not reduce the risk of mortality, because the observed difference between groups was clinically important. So, their study appears to be inconclusive. Why, then, do you suppose that the authors went to press with their results?

b. Motivating Case Study #2: Can a clinical trial have a positive result but not be definitive?

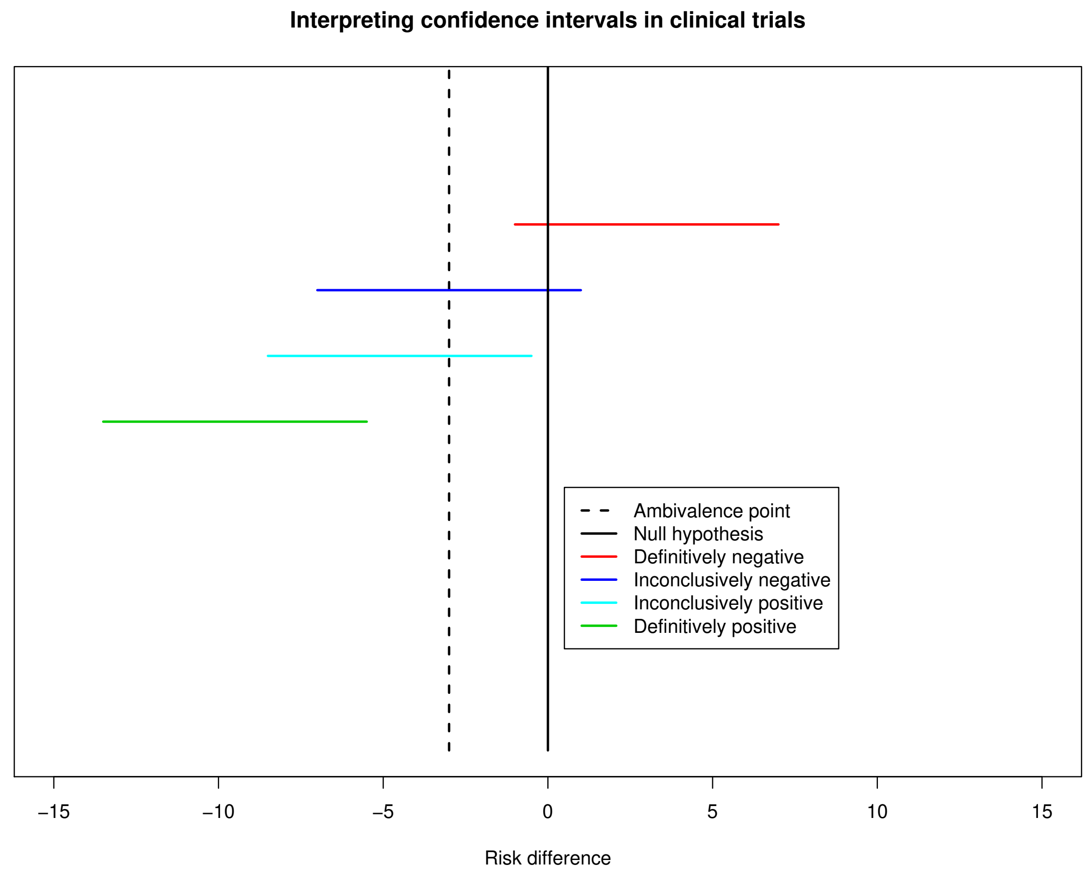
A 2010 *Journal of the American College of Cardiology* paper by Kaul and Diamond titled "Trial and error: how to avoid commonly encountered limitations of published clinical trials" cited the Treat Angina with aggrastat and determine Cost of Therapy with an Invasive or ConServative therapy — Thrombolysis In Myocardial Infarction 18 (TACTICS-TIMI 18) trial as an example of a trial in which a positive but not definitive result was reported.

This may run counter to your intuition. As the first motivating case study suggests, one can easily find situations in which a negative result is not definitive. In fact, I would venture to say that the majority of published negative results are not definitive. On the other hand, people usually regard a positive result as conclusive. Yet, as Kaul and Diamond among others have noted, a positive result can still be inconclusive.

Regarding TACTICS-TIMI 18, there were 177 events among 1114 patients assigned to early invasive management (15.9%) and 215 events among 1106 patients assigned to early conservative management (19.4%). The relative risk estimate of 0.82 was significantly different from unity ($p = 0.028$), which appears to warrant a recommendation of early invasive management over early conservative management.

However, the accompanying 95% confidence interval ranged from 0.68 to 0.98. While the 0.68 and even the 0.82 are convincing about the merits of early invasive management, the 0.98 is not. If the 0.98 were real, a possibility that should not be ruled out because of its presence in the 95% confidence interval, then the NNT would be 258. Most consumers of medical literature regard 50 as an upper bound for a clinically important NNT, since a treatment with a larger NNT may save a small number of lives but at a much higher cost to the rest of the patients — not just economic but also in terms of more non-fatal adverse events and other sources of discomfort.

In summary, the positive result from TACTICS-TIMI 18 is not definitive because it does not unambiguously warrant a recommendation of early invasive management over early conservative management. Some values in the 95% confidence interval for the relative risk clearly favor early invasive management, while others do not.



A visual schematic is provided in the figure above. The horizontal axis represents the risk difference as a percentage, $100(p_1 - p_2)\%$. The ambivalence point of -3% states that, in this example, we regard a result as clinically important if and only if the new treatment yields an absolute risk reduction of at least 3% compared to the old treatment (or placebo).

Four hypothetical confidence intervals, shown in different colors, illustrate results that are positive conclusive favoring the new treatment, positive inconclusive, negative inconclusive, and negative conclusive. Results that are positive conclusive favoring the old treatment are not shown. The figure makes clear that, whenever the ambivalence point entails superiority of the new treatment over the old treatment, a positive result need not be conclusive.

A reasonable question at this juncture is how to plan a study so that the result will be positive conclusive. We will address that question when we return to this case study at the end of Unit V. However, there are some interesting questions that we can answer now.

Discussion Questions.

1. Suppose the ambivalence point were placed at zero. How would our taxonomy of possible results change?

2. Suppose the ambivalence point were placed to the right of zero, so that it entailed superiority of the old treatment over the new treatment. Are there any circumstances under which this would make sense?

c. Power calculations

Let $X_1, X_2, \ldots, X_n$ be independently and identically distributed according to the probability mass or density function $f(x;\theta)$, where the unknown parameter θ belongs to the known parameter space Θ.

Consider testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$, where Θ_0 is a known subset of Θ. Suppose that we reject H_0 if and only if the data $\mathbf{X} := (X_1, X_2, \ldots, X_n)^T$ belong to a set R that we call the "rejection region".

<u>Example #1</u>. Put $f(x;\theta) := (2\pi)^{-1/2}\exp[-(x-\theta)^2/2]$, $\Theta := \mathbb{R}$, and $\Theta_0 := \{\theta_0\}$ for some fixed number $\theta_0 \in \mathbb{R}$. Then θ is the mean of the normal distribution generating the data and we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. To obtain a testing procedure with significance level 0.05, we may put $R := \{\mathbf{x} \in \mathbb{R}^n : |\bar{x} - \theta_0| > 1.96/\sqrt{n}\}$.

Define

$$G(\zeta) := P_{\theta=\zeta}(\mathbf{X} \in R) \quad \text{for } \zeta \in \Theta.$$

We refer to $G(\zeta)$ as the power function of the hypothesis testing procedure. If multiple procedures are being considered, then we may distinguish their power functions by subscripts, as in $G_1(\zeta)$, $G_2(\zeta)$, and so forth.

The interpretation of the power function is as follows. For $\zeta \in \Theta_0$, $G(\zeta)$ represents the probability of incorrectly rejecting the null hypothesis (committing a Type I error) if θ were equal to ζ. For $\zeta \notin \Theta_0$, $G(\zeta)$ represents the probability of correctly rejecting the null hypothesis (avoiding a Type II error) if θ were equal to ζ.

Example #1, continued. Let $\Phi(z) := \int_{-\infty}^{z} (2\pi)^{-1/2} \exp[-t^2/2]\, dt$ denote the standard normal cumulative distribution function. For $\zeta \in \mathbb{R}$, we have

$$\begin{aligned} G(\zeta) &= P_{\theta=\zeta}(|\bar{X} - \theta_0| > 1.96/\sqrt{n}) \\ &= P_{\theta=\zeta}(\bar{X} > \theta_0 + 1.96/\sqrt{n}) + P_{\theta=\zeta}(\bar{X} < \theta_0 - 1.96/\sqrt{n}) \\ &= P_{\theta=\zeta}\left(\sqrt{n}(\bar{X} - \zeta) > \sqrt{n}(\theta_0 - \zeta) + 1.96\right) \\ &\quad + P_{\theta=\zeta}\left(\sqrt{n}(\bar{X} - \zeta) < \sqrt{n}(\theta_0 - \zeta) - 1.96\right) \\ &= 1 - \Phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.96\right) + \Phi\left(\sqrt{n}(\theta_0 - \zeta) - 1.96\right). \end{aligned}$$

Suppose that we want 80% power to reject $H_0 : \theta = \theta_0$. We may then ask what is an appropriate sample size. However, we need one more piece of information, namely a presumed value for θ. For instance, if we presume that $\theta = \theta_0 + 1$, then the appropriate sample size is the smallest n for which $G(\theta_0 + 1) \geq 0.8$. Unfortunately, the equation

$$0.8 = 1 - \Phi(-\sqrt{n} + 1.96) + \Phi(-\sqrt{n} - 1.96)$$

cannot be solved analytically for n.

One option is to assume that $\Phi(-\sqrt{n} - 1.96)$ will be negligibly small. Then we can solve the equation

$$0.8 = 1 - \Phi(-\sqrt{n} + 1.96)$$

analytically for n. We obtain

$$\begin{aligned} & 0.2 = \Phi(-\sqrt{n}+1.96) \\ \iff \quad & -0.842 = \Phi^{-1}(0.2) = -\sqrt{n}+1.96 \\ \iff \quad & \sqrt{n} = 1.96+0.842 \\ \iff \quad & n = (1.96+0.842)^2 = 7.85. \end{aligned}$$

Since our sample size must be an integer, we round the 7.85 up to 8.

A second option is to perform the calculation numerically. Here is some R code that I used, followed by the R output.

```
n<-(1:10)
1-pnorm(-sqrt(n)+1.96)+pnorm(-sqrt(n)-1.96)
 [1] 0.1700658 0.2929765 0.4099541 0.5159909 0.6087657
 [6] 0.6877577 0.7535671 0.8074206 0.8508304 0.8853722
```

Again, we obtain an answer of 8.

Example #2. Put

$$f(x;\theta) = \theta_1^u(1-\theta_1)^{1-u}\theta_2^v(1-\theta_2)^{1-v}$$

for $x = (u,v)^T \in \{0,1\} \times \{0,1\}$ and $\theta = (\theta_1,\theta_2)^T \in \Theta := (0,1)\times(0,1)$, and

$$\Theta_0 := \{(\zeta_1,\zeta_2)^T \in (0,1)\times(0,1) : \zeta_1 = \zeta_2\}.$$

To obtain a testing procedure with approximate significance level 0.05, we may put

$$R := \{\mathbf{x} \in \mathbb{R}^{2n} : |\bar{u}-\bar{v}| > 1.96\sqrt{2\bar{x}(1-\bar{x})/n}\},$$

where $\bar{x} := (\bar{u}+\bar{v})/2$. The interpretation of this testing procedure is that we are asking whether two proportions, θ_1 and θ_2, are equal after collecting two independent samples of equal size.

We have

$$G(\zeta) = P_{\theta=\zeta}(|\bar{U}-\bar{V}| > 1.96\sqrt{2\bar{X}(1-\bar{X})/n}).$$

Put $\bar{\zeta} := (\zeta_1 + \zeta_2)/2$. For analytic tractability, we approximate $G(\zeta)$ by

$$
\begin{aligned}
& P_{\theta=\zeta}(|\bar{U} - \bar{V}| > 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n}) \\
= \;& P_{\theta=\zeta}(\bar{U} - \bar{V} > 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n}) \\
& + P_{\theta=\zeta}(\bar{U} - \bar{V} < -1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n}) \\
= \;& P_{\theta=\zeta}\left(\frac{\bar{U} - \bar{V} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}} > \frac{1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
& + P_{\theta=\zeta}\left(\frac{\bar{U} - \bar{V} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}} < \frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
\approx \;& 1 - \Phi\left(\frac{1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
& + \Phi\left(\frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
= \;& 1 - \Phi\left(\frac{1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} - \sqrt{n}(\zeta_1 - \zeta_2)}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}}\right) \qquad (1) \\
& + \Phi\left(\frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} - \sqrt{n}(\zeta_1 - \zeta_2)}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}}\right). \qquad (2)
\end{aligned}
$$

For a sample size calculation, we usually set line (1) to 0.80 if $\zeta_1 > \zeta_2$ and line (2) to 0.80 if $\zeta_1 < \zeta_2$. Either way, we obtain

$$
\begin{aligned}
& 0.8 = \Phi\left(\frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} + \sqrt{n}|\zeta_1 - \zeta_2|}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}}\right) \\
\iff \;& 0.842 = \Phi^{-1}(0.8) = \frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} + \sqrt{n}|\zeta_1 - \zeta_2|}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}} \\
\iff \;& 0.842\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)} + 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} = \sqrt{n}|\zeta_1 - \zeta_2| \\
\iff \;& n = \frac{\left(0.842\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)} + 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})}\right)^2}{|\zeta_1 - \zeta_2|^2}. \qquad (3)
\end{aligned}
$$

d. Unbiased, uniformly most powerful, and consistent tests

We say that a testing procedure with power function $G(\zeta)$ is unbiased if

$$\sup_{\zeta \in \Theta_0} G(\zeta) \leq \inf_{\zeta \notin \Theta_0} G(\zeta).$$

Again, "sup" is short for supremum. Recall that the supremum is the least upper bound of a set and generalizes the concept of a maximum. Likewise, "inf" is short for infimum. The infimum is the greatest lower bound of a set and generalizes the concept of a minimum. Thus, roughly speaking, an unbiased testing procedure is one for which the probability of rejecting the null hypothesis is higher when the null hypothesis is false than when it is true.

Example #1, continued. Instead of rejecting H_0 when $|\bar{X} - \theta_0| > 1.96/\sqrt{n}$, we can also reject H_0 when $\bar{X} - \theta_0 > 1.645/\sqrt{n}$. The power function of the former testing procedure was already seen to be

$$G_1(\zeta) := 1 - \Phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.96\right) + \Phi\left(\sqrt{n}(\theta_0 - \zeta) - 1.96\right),$$

while the power function of the latter testing procedure is readily seen to be

$$G_2(\zeta) := 1 - \Phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.645\right).$$

Differentiating $G_1(\zeta)$ in ζ yields

$$\sqrt{n}\left[\phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.96\right) - \phi\left(\sqrt{n}(\theta_0 - \zeta) - 1.96\right)\right], \tag{4}$$

where $\phi(z) := (2\pi)^{-1/2} \exp[-z^2/2]$ is the standard normal probability density function. Expression (4) can equal 0 only if

$$\sqrt{n}(\theta_0 - \zeta) + 1.96 = -[\sqrt{n}(\theta_0 - \zeta) - 1.96] = -\sqrt{n}(\theta_0 - \zeta) + 1.96,$$

which is true only if $\zeta = \theta_0$. Since $\lim_{\zeta \to \pm\infty} G_1(\zeta) = 1$ and $G_1(\theta_0) = 0.05 < 1$, we conclude that

$$G_1(\theta_0) \leq G_1(\zeta)$$

for all $\zeta \notin \Theta_0$. Hence,

$$\sup_{\zeta \in \Theta_0} G_1(\zeta) \leq \inf_{\zeta \notin \Theta_0} G_1(\zeta).$$

Thus, the former testing procedure is unbiased.

In contrast,

$$\lim_{\zeta \to -\infty} G_2(\zeta) = 0,$$

so that

$$0.05 = G_2(\theta_0) = \sup_{\zeta \in \Theta_0} G_2(\zeta) > \inf_{\zeta \notin \Theta_0} G_2(\zeta) = 0.$$

Thus, the latter testing procedure is not unbiased.

We say that a testing procedure with power function $G(\zeta)$ is uniformly most powerful if

$$G(\zeta) \geq \tilde{G}(\zeta) \quad \text{at all } \zeta \notin \Theta_0$$

compared to any $\tilde{G}(\zeta)$ representing the power function of another testing procedure with the same significance level.

In two-sided testing problems, there typically does not exist a uniformly most powerful test.

<u>Example #1, continued</u>. Suppose that $n = 1$. On the one hand, we have

$$G_1(\theta_0+1) = 1-\Phi(0.96)+\Phi(-2.96) = 0.170 < G_2(\theta_0+1) = 1-\Phi(0.645) = 0.259.$$

On the other hand, we have

$$G_1(\theta_0-1) = 1-\Phi(2.96)+\Phi(-0.96) = 0.170 > G_2(\theta_0-1) = 1-\Phi(2.645) = 0.004.$$

Thus, neither of the two testing procedures is uniformly most powerful.

Of course, the above computations do not prove nonexistence of a uniformly most powerful test. The standard technique to prove nonexistence in two-sided testing problems is to identify one testing procedure that has the greatest power at every $\zeta < \theta_0$, identify another testing procedure that has the greatest power at every $\zeta > \theta_0$, and show that the two testing procedures are different.

In one-sided testing problems, there often does exist a uniformly most powerful test.

Suppose that $\Theta \subset \mathbb{R}$ and

$$f(x;\theta) = a(x)b(\theta)\exp[c(x)d(\theta)].$$

Put $T := \sum_{i=1}^{n} c(X_i)$, and let $h(t;\theta)$ denote the probability mass or density function of T.

If $h(t;\theta_1)/h(t;\theta_2)$ is an increasing (resp., decreasing) function of t whenever $\theta_1 > \theta_2$, then a uniformly most powerful test of $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$ entails rejecting H_0 whenever T is greater than the $1-\alpha$ quantile (resp., is less than the α quantile) of its distribution under $\theta = \theta_0$.

If $h(t;\theta_1)/h(t;\theta_2)$ is an increasing (resp., decreasing) function of t whenever $\theta_1 > \theta_2$, then a uniformly most powerful test of $H_0 : \theta \ge \theta_0$ against $H_1 : \theta < \theta_0$ entails rejecting H_0 whenever T is less than the α quantile (resp., is greater than the $1-\alpha$ quantile) of its distribution under $\theta = \theta_0$.

<u>Example #3</u>. Put $f(x;\theta) := 1_{\{x\in(0,\infty)\}}\theta \exp[-x\theta]$, where $\theta \in \Theta := (0,\infty)$. Consider testing $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$, for some fixed number $\theta_0 \in \Theta$.

We will see presently that we obtain a uniformly most powerful test by rejecting H_0 when $T := \sum_{i=1}^n X_i$ is too extreme. The question is, does extreme mean too small or too large?

The probability density function of T is, for $t \in (0,\infty)$,

$$h(t;\theta) = \frac{\theta^n}{\Gamma[n]} t^{n-1} \exp[-\theta t].$$

Thus, for $\theta_1 > \theta_2$ we have

$$\frac{h(t;\theta_1)}{h(t;\theta_2)} = \frac{\theta_1^n}{\theta_2^n} \exp[(\theta_2 - \theta_1)t],$$

which is a decreasing function of t. Hence, a uniformly most powerful test is obtained by rejecting H_0 when T is too small. More specifically, we should reject H_0 when T is less than the α quantile of its distribution under $\theta = \theta_0$. This is equivalent to rejecting H_0 when $\theta_0 T$ is less than the α quantile of the gamma distribution with shape n and scale 1.

We say that a testing procedure with power function $G(\zeta)$ is consistent if

$$\lim_{n\to\infty} G(\zeta) = 1$$

at any fixed $\zeta \notin \Theta_0$.

<u>Example #1, continued</u>. If $\zeta < \theta_0$, then $\sqrt{n}(\theta_0 - \zeta) \to +\infty$ and

$$G_1(\zeta) = 1 - \Phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.96\right) + \Phi\left(\sqrt{n}(\theta_0 - \zeta) - 1.96\right) \to 1 - 1 + 1 = 1.$$

If $\zeta > \theta_0$, then $\sqrt{n}(\theta_0 - \zeta) \to -\infty$ and

$$G_1(\zeta) = 1 - \Phi\left(\sqrt{n}(\theta_0 - \zeta) + 1.96\right) + \Phi\left(\sqrt{n}(\theta_0 - \zeta) - 1.96\right) \to 1 - 0 + 0 = 1.$$

Thus, the testing procedure with power function $G_1(\zeta)$ is consistent.

e. Large sample behavior of likelihood ratio test statistics

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. Consider a likelihood ratio test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, where θ_0 is a fixed element of Θ. With the regularity conditions for maximum likelihood estimation in Unit III, we have

$$-2\log\lambda \xrightarrow{L} \chi_1^2 \tag{5}$$

under H_0, where

$$\lambda := \prod_{i=1}^{n} f(X_i;\theta_0) / \prod_{i=1}^{n} f(X_i;\hat\theta)$$

and $\hat\theta$ is the maximum likelihood estimator of θ.

While a detailed rigorous proof of this result is beyond the scope of BST 676, I can present a sketch of the proof that should render the result plausible.

Step #1. For $\zeta \in \Theta$, define the log likelihood

$$l(\zeta;\mathbf{X}) := \sum_{i=1}^{n} \log f(X_i;\zeta).$$

By Taylor expansion, we have

$$l(\theta_0;\mathbf{X}) = l(\hat\theta;\mathbf{X}) + (1/2)(\theta_0 - \hat\theta)^2 \frac{\partial^2}{\partial\zeta^2} l(\zeta;\mathbf{X})|_{\zeta=\hat\theta} + \text{Remainder}, \tag{6}$$

the Remainder presumably becoming negligible as $n \to \infty$.

Step #2. The Weak Law of Large Numbers yields

$$\begin{aligned} n^{-1}\frac{\partial^2}{\partial\zeta^2} l(\zeta;\mathbf{X}) &= n^{-1}\sum_{i=1}^{n} \frac{\partial^2}{\partial\zeta^2}\log f(X_i;\zeta) \\ \xrightarrow{P} \quad E_{\theta_0}\left[\frac{\partial^2}{\partial\zeta^2}\log f(X_1;\zeta)\right] &= \int_{\mathbb{R}} \frac{\partial^2}{\partial\zeta^2}\log f(x;\zeta)\ f(x;\theta_0)\ dx \end{aligned}$$

for any fixed $\zeta \in \Theta$. Recalling that $\hat\theta$ is consistent for θ_0, we may anticipate that

$$\begin{aligned} n^{-1}\frac{\partial^2}{\partial\zeta^2} l(\zeta;\mathbf{X})|_{\zeta=\hat\theta} &= n^{-1}\sum_{i=1}^{n} \frac{\partial^2}{\partial\zeta^2}\log f(X_i;\zeta)|_{\zeta=\hat\theta} \\ \xrightarrow{P} \quad E_{\theta_0}\left[\frac{\partial^2}{\partial\zeta^2}\log f(X_1;\zeta)|_{\zeta=\theta_0}\right] &= -J_1(\theta_0). \end{aligned} \tag{7}$$

In fact, statement (7) is true, although methods beyond the scope of BST 676 are required to prove it.

Step #3. We already know from Unit III that

$$n^{1/2}(\theta_0 - \hat{\theta}) \xrightarrow{L} N(0, J_1(\theta_0)^{-1}) = J_1(\theta_0)^{-1/2} N(0,1).$$

By the continuous mapping theorem, we have

$$n(\theta_0 - \hat{\theta})^2 \xrightarrow{L} J_1(\theta_0)^{-1}\chi_1^2. \tag{8}$$

Combining (7) and (8) yields

$$(1/2)(\theta_0 - \hat{\theta})^2 \frac{\partial^2}{\partial \zeta^2} l(\zeta; \mathbf{X})|_{\zeta=\hat{\theta}} \xrightarrow{L} -(1/2)J_1(\theta_0)^{-1}\chi_1^2 J_1(\theta_0) = -(1/2)\chi_1^2.$$

Assuming that the Remainder in (6) converges to zero in probability, which can be shown using methods beyond the scope of BST 676, we then have

$$\log \lambda = l(\theta_0; \mathbf{X}) - l(\hat{\theta}; \mathbf{X}) \xrightarrow{L} -(1/2)\chi_1^2$$

and hence

$$-2\log\lambda \xrightarrow{L} \chi_1^2.$$

In fact, result (5) generalizes to accommodate a k-dimensional vector parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_k)^T \in \Theta \subset \mathbb{R}^k$. Let Θ_0 be a subset of Θ in which $\theta_1 = \theta_{1,0}, \theta_2 = \theta_{2,0}, \ldots, \theta_r = \theta_{r,0}$ for some $r \le k$ and some fixed real numbers $\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{r,0}$. With appropriate k-dimensional analogues to the regularity conditions for maximum likelihood estimation in Unit III, we have

$$-2\log\lambda \xrightarrow{L} \chi_r^2$$

under H_0.

For example, in a logistic regression model with intercept α and partial slopes $\beta_1, \beta_2, \ldots, \beta_m$, a likelihood ratio test of the global null hypothesis $H_0 : \beta_1 = \beta_2 = \ldots = \beta_m = 0$ is linked to a chi-square distribution on m degrees of freedom. Here we have $k = m+1$, $r = m$, $\theta = (\beta_1, \ldots, \beta_m, \alpha)^T$, $\Theta = \mathbb{R}^{m+1}$, $\theta_{1,0} = \theta_{2,0} = \cdots = \theta_{r,0} = 0$, and $\Theta_0 = \{0\} \times \cdots \times \{0\} \times \mathbb{R}$. Note, in particular, that the degrees of freedom for the chi-square distribution match the dimension of the "full" parameter space Θ minus the dimension of the "reduced" parameter space Θ_0.

f. Comparing nominal and actual significance levels

For a hypothesis testing procedure that is not exact, another aspect relevant to its evaluation is how close the actual significance level is to the nominal significance level. In other words, if you *claim* to reject a true null hypothesis five percent of the time, do you *really* reject a true null hypothesis five percent of the time?

A testing procedure for which the nominal significance level exceeds the actual significance level is called conservative. Such a testing procedure is stilted toward acceptance of the null hypothesis.

A testing procedure for which the actual significance level exceeds the nominal significance level is called anticonservative. Such a testing procedure is stilted toward rejection of the null hypothesis.

What can you do if you are unwilling to tolerate a disparity between nominal and actual significance levels?

If the disparity occurs because of the discreteness of the test statistic, one option is to identify borderline values of the test statistic at which an auxiliary experiment is conducted to determine acceptance or rejection of the null hypothesis.

For example, consider testing $H_0 : p = 0.50$ against $H_1 : p \neq 0.50$ based on 100 Bernoulli trials. Rejecting H_0 when $|\hat{p} - 0.50|/\sqrt{0.50(1-0.50)/100} > 1.96$ is equivalent to rejecting H_0 when there are less than 41 or more than 59 successes in the 100 trials. So, the actual significance level is

$$1 - \sum_{k=41}^{59} \binom{100}{k} (0.50)^k (1-0.50)^{100-k} = 0.0569,$$

which differs from the nominal significance level of 0.05.

Rejecting H_0 when there are less than 40 or more than 60 successes does not fix the problem, since then the actual significance level is

$$1 - \sum_{k=40}^{60} \binom{100}{k} (0.50)^k (1-0.50)^{100-k} = 0.0352.$$

What we can do, however, is: (i) reject H_0 if there are less than 40 or more than 60 successes; (ii) accept H_0 if there are at least 41 but no more than

59 successes; and, (iii) reject H_0 with probability 0.682 — by, say, drawing a random number between 0 and 1 and rejecting H_0 if the random number is less than 0.682 — if there are exactly 40 or exactly 60 successes.

If the disparity occurs with a continuous test statistic, one option is to adjust the critical value for rejection of H_0 so that the actual significance level is 0.05. Sometimes the adjustment can be done analytically.

For instance, adjusting $n + 1.645\sqrt{n}$ to $g_{n,0.95}$ would fix the problem in Example #2 of Unit IV. (Recall that we were testing $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ based on $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mu^{-1} \exp[-x\mu^{-1}]1_{\{x>0\}}$ and rejecting H_0 if $\sum_{i=1}^{n} X_i/\mu_0$ were too large.)

Other times the adjustment must be informed by a simulation experiment. Here is R code that conducts such a simulation experiment for Example #2 of Unit IV with $\mu_0 = 5$ and $n = 20$.

```
Data <- matrix(rexp(100000*20,rate=1/5),nrow=100000,ncol=20)
TestStatistics <- apply(Data,1,sum)/5
quantile(TestStatistics,0.95)
```

g. Resolution of motivating case studies

Our first motivating case study asked what sample sizes were necessary for testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ under the supposition that H_1 held with $p_1 = 0.3275$ and $p_2 = 0.3815$. We were concerned not only with the sample sizes being large enough to defend a chi-square test methodologically but also with the sample sizes being large enough so that we would have a high probability of rejecting H_0 if our supposition were correct.

Assuming that equal numbers of patients would be randomized to enalapril and hydralazine-isosorbide dinitrate, we may apply formula (3) with $\zeta_1 = 0.3275$ and $\zeta_2 = 0.3815$. Noting that $\bar{\zeta} = 0.3545$ and that the answer must be rounded upward to the next integer, we obtain $n = 1232$ for each treatment group.

Discussion Questions.

1. What can you do if you anticipate 10% dropout in each treatment group?

2. We know from our introductory methods course that there is a generalization of formula (3) in which the planned sample sizes are not constrained to equality. While some dropout may be inevitable, leading to slight imbalances in the *actual* sample sizes available for analysis at the end of the study, can you think of any circumstances under which the *planned* sample sizes should not be constrained to equality?

Our second motivating case study asked for appropriate sample sizes in testing $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ if we wanted an 80% chance of not only rejecting H_0 but also ruling out that $p_1 - p_2 = t$, where $t < 0$ represents an ambivalence point.

The resolution of our second motivating case study will entail formulating an auxiliary hypothesis test with $H_0' : p_1 - p_2 = t$ and $H_1' : p_1 - p_2 \neq t$. More specifically, we will suppose that H_1' holds with $p_1 = \zeta_1$ and $p_2 = \zeta_2$ such that $\zeta_1 - \zeta_2 < t$ and then use an approximate power function for the auxiliary hypothesis test to determine appropriate sample sizes.

Consider the scenario from Example #2 of this Unit with Θ_0 redefined as

$$\Theta_0 := \{(\zeta_1, \zeta_2)^T \in (0,1) \times (0,1) : \zeta_1 - \zeta_2 = t\}$$

and R redefined as

$$R := \{\mathbf{x} \in \mathbb{R}^{2n} : |\bar{u} - \bar{v} - t| > 1.96\sqrt{2\bar{x}(1-\bar{x})/n}\}.$$

We have

$$G(\zeta) = P_{\theta=\zeta}(|\bar{U} - \bar{V} - t| > 1.96\sqrt{2\bar{X}(1-\bar{X})/n}).$$

Put $\bar{\zeta} := (\zeta_1 + \zeta_2)/2$. We approximate $G(\zeta)$ by

$$\begin{aligned}
& P_{\theta=\zeta}(\bar{U} - \bar{V} - t < -1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n}) \\
= \;& P_{\theta=\zeta}\left(\frac{\bar{U} - \bar{V} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}} < \frac{t - 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
\approx \;& \Phi\left(\frac{t - 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})/n} - \zeta_1 + \zeta_2}{\sqrt{\zeta_1(1-\zeta_1)/n + \zeta_2(1-\zeta_2)/n}}\right) \\
= \;& \Phi\left(\frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} - \sqrt{n}(\zeta_1 - \zeta_2 - t)}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}}\right). \qquad (9)
\end{aligned}$$

Setting line (9) to 0.80 yields

$$
\begin{aligned}
& 0.8 = \Phi\left(\frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} - \sqrt{n}(\zeta_1 - \zeta_2 - t)}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}}\right) \\
\iff \quad & 0.842 = \Phi^{-1}(0.8) = \frac{-1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} - \sqrt{n}(\zeta_1 - \zeta_2 - t)}{\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)}} \\
\iff \quad & 0.842\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)} + 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})} = \sqrt{n}|\zeta_1 - \zeta_2 - t| \\
\iff \quad & n = \frac{\left(0.842\sqrt{\zeta_1(1-\zeta_1) + \zeta_2(1-\zeta_2)} + 1.96\sqrt{2\bar{\zeta}(1-\bar{\zeta})}\right)^2}{|\zeta_1 - \zeta_2 - t|^2}. \qquad (10)
\end{aligned}
$$

As a numerical example, consider the supposition that $p_1 = 0.159$ and $p_2 = 0.194$. Appropriate sample sizes as a function of the ambivalence point are shown in Figure 1 below. Note two special cases. If the ambivalence point is 0%, then expression (10) reduces to expression (3). If the ambivalence point is -3.5% or less, then the difference between the supposed values of p_1 and p_2 is not clinically important and there is no sample size at which we have an 80% chance of reporting a clinically important difference.

Figure 1: