MthStat 465, Spring 2005, Lecture Number 19
A general procedure for choosing between two possible coins.

This is a repetition of Lecture 18, this time completely symbolically.

Suppose that $0 < p < q < 1$ are given, and that a coin is known to come up heads either with probablity p or with probability q. If the coin is tossed N times we are to come up with a decision rule as to which coin was tossed. N is to be determined.

The mathematical set-up is that we have a sample space S consisting of sequences of length N of the symbols H and T, representing Heads and Tails respectively in N consecutive tosses of the coin. We take the set of events, Σ, to be the set of all subsets of S. Our problem is to determine which of two possible probability measures, which we denote as $\Pr_p$and $\Pr_q$ we should have on the set of events. To define these probability measures consisely, for each $s \in S$ let $H(s) =$ the number of H's in S. Then

$$\begin{aligned}\Pr_p(\{s\}) &= p^{H(s)}(1-p)^{N-H(s)}\\ \Pr_q(\{s\}) &= q^{H(s)}(1-q)^{N-H(s)}\end{aligned}$$

Notice that everything in the model depends on N, but we are not indicating this explicitly. Also, H is a random variable, and for $x \in \{0, 1, \ldots, N\}$,

$$\begin{aligned}\Pr_p(\{s : H(s) = x\}) &= \binom{N}{x} p^x(1-p)^{N-x}\\ \Pr_q(\{s : H(s) = x\}) &= \binom{N}{x} q^x(1-q)^{N-x}\end{aligned}$$

Our decision rule is a random variable $D : S \to \{p, q\}$, where $D(s) = p$ means choose $\Pr_p$ for the model, and $D(s) = q$ means choose $\Pr_q$ for the model.

To evaluate the efficacy of our choice of D we choose two numbers e_p and $_q$, both between 0 and 1, and insist that

$$\begin{aligned}\Pr_p(\{s : D(s) = p\}) &\geq e_p\\ \Pr_q(\{s : D(s) = q\}) &\geq e_q\end{aligned}$$

The numbers e_p and e_q give us the probabilities of making the right decision, and we typically think of these numbers as begin near to 1.

A reasonable rule for D is to pick some number C and ask that $D(s) = p$ if $H(s) < C$ and $D(s) = q$ if $H(s) \geq C$. With any luck, C will not be an integer, so the equality case will not matter. Our problem is given e_p and e_q how to pick C and N so that

$$\begin{aligned}\Pr_p(\{s : H(s) < C\}) &\geq e_p\\ \Pr_q(\{s : H(s) \geq C\}) &\geq e_q\end{aligned}$$

The idea is to use the Central Limit Theorem/DeMoivre-Laplace Theorem to estimate C and N, and then check via the binomial distribution how well we have done, and revise our choices if need be. Recall the formulae for the mean and variance of a binomial random variable, Np and $Np(1-p)$.

$$\begin{aligned} e_p &\leq \Pr_p(\{s : H(s) < C\}) \\ (1) \qquad &= \Pr_p\left(\left\{s : \frac{H(s)-Np}{\sqrt{Np(1-p)}} < \frac{C-Np}{\sqrt{Np(1-p)}}\right\}\right) \\ (2) \qquad &\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(C-Np)/\sqrt{Np(1-p)}} \exp(-u^2/2)\,du. \end{aligned}$$

Let A_p satisfy

$$e_p = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{A_p} \exp(-u^2/2)\,du.$$

If we regard the approximation in (2) has an equality, then we get the inequality

$$(3) \qquad \frac{C-Np}{\sqrt{Np(1-p)}} \geq A_p$$

because increasing the upper limit of integration in (2) increases the integral.

$$\begin{aligned} e_q &\geq \Pr_q(\{s : H(s) \geq C\}) \\ (4) \qquad &= \Pr_q\left(\left\{s : \frac{H(s)-Nq}{\sqrt{Nq(1-q)}} \geq \frac{C-Nq}{\sqrt{Nq(1-q)}}\right\}\right) \\ (5) \qquad &\approx \frac{1}{\sqrt{2\pi}} \int_{(C-Nq)/\sqrt{Nq(1-q)}}^{\infty} \exp(-u^2/2)\,du. \end{aligned}$$

Let A_q satisfy

$$e_q = \frac{1}{\sqrt{2\pi}} \int_{A_q}^{\infty} \exp(-u^2/2)\,du.$$

If we regard the approximation in (5) has an equality, then we get the inequality

$$(6) \qquad \frac{C-Nq}{\sqrt{Nq(1-q)}} \leq A_q$$

because decreasing the lower limit of integration in (5) increases the integral.

To find provisionally acceptable values of C and N we then want to solve the simultaneous inequalities (3) and (6). We now make two additional assumptions: $e_p > 1/2$ and $e_q > 1/2$. These are reasonable as they mean that we want to be right at least half the time no matter whether $\Pr_p$ or $\Pr_q$ is the correct choice. The assumption $e_p < 1/2$ implies that $A_p > 0$, which in turn implies that $C - Np > 0$, and the assumption that $e_q > 1/2$ implies that $A_q < 0$, which in turn imples that $C - Nq < 0$. Solving the inequalities (3) and (6) together gives

$$(7) \qquad N \geq \left(\frac{A_p\sqrt{p(1-p)} - A_q\sqrt{q(1-q)}}{q-p}\right)^2$$

and

$$(8) \qquad Np + \sqrt{N}\sqrt{p(1-p)}A_p \leq C \leq Nq + \sqrt{N}\sqrt{q(1-q)}A_q.$$

Recall also that N must be an integer, and that all values of C with the same integer part will yield the same decision rule.

Here is a numerical example. Suppose $p = 0.4$ and $q = 0.65$, and that $e_p = 0.977$ and $e_q = 0.933$. This gives $A_p = 2$ and $A_q = -1.5$. This means $N \geq 46$, and

$$0.4N + 2\sqrt{0.24}\sqrt{N} \leq C \leq 0.65N - 1.5\sqrt{0.2275}\sqrt{N}.$$

For example, if $N = 50$, then

$$26.928 \leq C \leq 27.44$$

which leads to the same rule: Choose $\mathrm{Pr}_{0.4}$ if the number of heads in 50 tosses is less than 27, and choose $\mathrm{Pr}_{0.65}$ if the number of heads in 50 tosses is 27 or more.

If we choose $N = 100$ then

$$47.798 \leq C \leq 57.845$$

which leads to a variety of tests.

Now we have to ask ourselves if the approximation we made is justified. We see that for the values of N and p and q under consideration we should be on target, because $Np(1-p) > 10$ and $Nq(1-q) > 10$ for $N \geq 46$, $p = 0.4$ and $q = 0.65$.

We can also check this directly by looking at the probabilities exactly. For example, if we choose $\mathrm{Pr}_{0.4}$ if we get 26 or fewer heads in 50 tosses, and we choose $\mathrm{Pr}_{0.65}$ if we get 27 or more heads in 50 tosses, we look at

$$\begin{aligned}
\underset{0.4}{\mathrm{Pr}}\{H(s) \leq 26\} &= \sum_{x=0}^{26} \binom{50}{x} (.4)^x (0.6)^{50-x} \\
&\approx 0.9685944469 \\
\underset{0.65}{\mathrm{Pr}}\{H(s) \geq 27\} &= \sum_{x=27}^{50} \binom{50}{x} (.65)^x (0.35)^{50-x} \\
&\approx 0.9603588504
\end{aligned}$$

which is not quite the accuracy we wanted, as the first probability is a bit lower than the target of 0.977. If we take $N = 100$ and $C = 51$,

$$\begin{aligned}
\underset{0.4}{\mathrm{Pr}}\{H(s) \leq 50\} &= \sum_{x=0}^{50} \binom{100}{x} (.4)^x (0.6)^{100-x} \\
&\approx 0.9832383135 \\
\underset{0.65}{\mathrm{Pr}}\{H(s) \geq 51\} &= \sum_{x=51}^{100} \binom{100}{x} (.65)^x (0.35)^{100-x} \\
&\approx 0.9985494385
\end{aligned}$$

which are both quite a bit better than we wanted.