Part II
Decision theory

Chaos umpire sits,
And by decision more embroils the fray
By which he reigns: next him high arbiter
Chance governs all.
Milton, Paradise Lost, I. 907

The excitement that a gambler feels when making a bet is equal to the amount he might win times the probability of winning it.
Blaise Pascal

The only useful function of a statistician is to make predictions, and thus to provide a basis for action.
W.E.Deming

Enough probability; we have what we need now to start doing statistics.

"Decision theory" is about how to behave optimally under uncertainty, and as such provides a nice backbone for the various statistical procedures we'll be looking at.

Loss functions, expected loss, etc.

Figure 10: Wald.

The basic idea of decision theory is that we want to minimize our *expected loss*.

These things are always easier to think about with an example in mind. So, let's say I offer you a bet. I have two coins: one biased towards heads (70-30 heads) and one even more biased towards tails (80-20 tails) . I'll randomly choose one of the coins (not telling you which one), flip a few times, then let you make a bet on whether the coin was biased towards heads or tails. You can take heads or tails, but the odds are different:

	you bet heads	you bet tails
coin is heads	you win 1$	you lose 1$
coin is tails	you lose 10$	you win 10$

How do you decide whether to say heads (H) or tails (T), given the number of heads and tails that came up? I guess this depends on how lucky you feel...

Decision theory says, forget luck, look at the expected loss. First we need a "loss," or "cost function." This function quantifies how badly it hurts if we guess H when we should have guessed T, and vice versa. Let's take $C(truth, guess)$ to be the negative of the table above. (I.e., we feel pain exactly proportional to the amount of cash we lose, or feel happiness directly proportionally to how much cash we win. **Exercise 41**: Can you think of examples where this direct proportional rule might not be a good loss function?)

Now we want to compute the expected loss of whatever decision strategy we can think of, and then choose the strategy that minimizes the expected loss. Let's list our ingredients:

- A "decision strategy" is any possible function of the data

  $$guess(D) \rightarrow \mathcal{A} = \{H, T\}.$$

  (This is just a fancy way of saying, you see some data and make a choice, $guess(D)$, between the possible states of the coin, where the "action space" — the set of decisions, or actions, that make any sense — is $\mathcal{A} = \{H, T\}$.)

- The data D are the total number of heads and tails observed

- We need the likelihood of having observed D, given that the coin was actually H or T. This is given by our trusty binomial formula. Let's say I flipped N times, and the number of observed heads is n. Then

  $$p(n|\text{coin biased towards heads}) = \binom{N}{n}(0.7)^n(0.3)^{N-n},$$

  and

  $$p(n|\text{coin biased towards tails}) = \binom{N}{n}(0.2)^n(0.8)^{N-n}.$$

- We already defined our loss function $C(truth, guess)$.

- Finally, we have one last parameter that plays an important role: the probability that I chose the heads coin at the very beginning. Call this parameter θ.

So let's put everything together. We'll compute the expected loss as a function of θ, aka the "risk function"

$$R(\theta) \equiv E[C(truth, guess(D))|\theta].$$

Note that D is a r.v., therefore so is $guess(D)$, and it makes sense to compute this expectation.

We have

$$\begin{aligned}
R(\theta) &= \sum_{truth \in \mathcal{A}} p(truth|\theta) \sum_D p(D|truth) C(truth, guess(D)) \\
&= \theta \sum_D p(D|H) C(H, guess(D)) + (1-\theta) \sum_D p(D|T) C(T, guess(D)) \\
&= \theta \sum_{n=0}^{N} \binom{N}{n} (0.7)^n (0.3)^{N-n} C(H, guess(N,n)) \\
&\quad + (1-\theta) \sum_{n=0}^{N} \binom{N}{n} (0.2)^n (0.8)^{N-n} C(T, guess(N,n)) \\
&= \theta \sum_{n=0}^{N} \binom{N}{n} (0.7)^n (0.3)^{N-n} \left(\begin{cases} 1\$ & if\ guess(N,n) = H \\ -1\$ & if\ guess(N,n) = T \end{cases} \right) \\
&\quad + (1-\theta) \sum_{n=0}^{N} \binom{N}{n} (0.2)^n (0.8)^{N-n} \left(\begin{cases} -10\$ & if\ guess(N,n) = H \\ 10\$ & if\ guess(N,n) = T \end{cases} \right)
\end{aligned}$$

The point is, even though this looks a little ugly, we can in principle do these sums, and decide exactly which guess is optimal, given the data, as a function of θ. For example, if I tell you what my θ is — e.g., I promise to choose the heads coin with probability .95 — then the optimal guess will be heads, unless the data strongly argues against it (e.g., if N is large and n is small). And vice versa. We just have to turn the crank mechanically to come up with the optimal guess. **Exercise 42**: Compute the optimal decision rule for $N = 1$, as a function of θ, given this cost function. **Exercise 43**: Let's say $N = 100$ and $n = 10$. What is the optimal decision as a function of θ? **Exercise 44**: Let $N = 2$. Draw the risk function as a function of θ (feel free to use a calculator for help) for the commonsense decision rule: if $n/N > .5$, choose heads; otherwise choose tails. What can you say about the shape of $R(\theta)$ in general (i.e., for any N)?

Now, we just worked through a simple example, in which there were only two "states of nature." More generally, the states of nature could be continuous, instead. For example, imagine the situation in which I draw N i.i.d. Gaussian r.v.'s X_i from $\mathcal{N}(\theta, 1)$, and then offer to pay you proportionally to the negative of the square distance between your guess — call it $\hat{\theta}(\{X_1, X_2, ...X_N\})$ — and the true parameter θ. So in this case the cost function is

$$C(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2.$$

And all the theory goes through if we replace the binomial likelihood term above with the corresponding Gaussian likelihood, and the corresponding sums by integrals — we can still compute $R(\theta)$ for any choice of $\hat{\theta}(\{X_1, X_2, ...X_N\})$:

$$R(\theta) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} \mathcal{N}(\theta, 1)(X_i) \left(\theta - \hat{\theta}(\{X_1, X_2, ...X_N\})\right)^2 \prod_{i=1}^{N} dX_i.$$

Exercise 45: Compute the risk function of the decision rule

$$\hat{\theta}(\{X_1, X_2, ...X_N\}) \equiv \frac{1}{N} \sum_{i=1}^{N} X_i,$$

under this squared-error cost function. Is this decision rule optimal if we know $\theta = 2$? If not, what is the optimal rule? Does this make sense? Prove that the optimal rule is unique in this case.

To sum up, this minimum-expected-loss idea gives us a straightforward, principled way to behave optimally under uncertainty, if we have a cost function that corresponds well to reality and we know the true probability of all the events in sight.

But there's still one ingredient we haven't discussed adequately: doesn't it seem strange that our optimal strategy depends on the parameter θ — which in most real-world situations is unknown? How do we deal with this?

Domination, admissibility, Bayes and minimax

We saw in the last section that it's straightforward to calculate the optimal decision rule when you have the correct cost function, likelihoods, and parameter in hand. But we also noticed that the optimal decision rule when θ is known is kind of degenerate, and that the optimal rule for one parameter might be very different from what is optimal under another parameter.

What do we do when we don't know the true parameter θ? Life would be easy if there were a decision rule, $T(D)$, whose risk function $R_T(\theta)$ was smaller than that of any other risk function in a *uniform* sense:

$$R_T(\theta) \leq R_S(\theta) \quad \forall \theta \in \Theta,$$

for any possible other decision rule $S(D)$ (here Θ is the "parameter space," the set of all the parameters under consideration); then obviously we'd use T.

The problem is this almost never happens, except (as we saw above) in the trivial case that $\Theta = \{\theta\}$. So we need to back off a little bit and think about what's important.

This idea does lead to one useful strategy: get rid of all the really useless decision rules. That is, if U is a decision rule such that

$$R_S(\theta) \leq R_U(\theta) \quad \forall \theta \in \Theta,$$

with

$$R_S(\theta_1) < R_U(\theta_1)$$

for some $\theta_1 \in \Theta$ — that is, U is always at least as bad as S, and strictly worse for at least one parameter — then it seems reasonable to conclude that U is a pretty subpar strategy. In this case, we say:

- The decision rule S "dominates" U
- U is "inadmissible."

So we can restrict our attention to the set of *admissible* decision rules — that is, the set of rules which aren't dominated by another.

This seems like a good start, at least. But it turns out there are usually a whole bunch of admissible strategies. So we need some way to narrow things down further. The two most commonly accepted strategies for choosing admissible decision rules are as follows:

- The *minimax* strategy: choose a strategy T that minimizes the "worst-case" error:

  $$\max_\theta R(\theta)$$

  In other words, choose your decision rule T such that

  $$\max_\theta R_T(\theta) = \min_S \max_\theta R_S(\theta),$$

  where the minimization is taken over the class of all possible decision rules and the maximization gives you the worst-case error.

- The *Bayes* strategy: choose a strategy T that minimizes the *average* error according to your prior beliefs about θ:

  $$E_\theta R(\theta) = \int p(\theta) R(\theta) d\theta.$$

There has been a lot of discussion about the relative merits of these two approaches, and some interesting connections between the two have been established. But it still comes down to a matter of taste, really. If you have good prior information — that is, a good sense of which θ are more probable than others — then it is a good idea to encapsulate this information in a prior distribution $p(\theta)$ and make use of the corresponding Bayes decision rule. If, on the other hand, you have very limited *a priori* information about θ and you want to play it safe, then minimizing the worst-case risk seems like a good idea. (Of course, then the best case risk might not be very good. We'll think about this some more in the exercises.)

Exercise 46: Neither the Bayes nor the minimax strategies are unique in general. Can you think of any simple conditions on the cost function and parameter space that does make each strategy unique? (Hint: functions which are strictly convex and have convex domains have unique minima, and convexity is preserved by both addition and maximization; that is, the pointwise sum and maximum of any two convex functions is convex.)

Exercise 47: Is a Bayes strategy necessarily admissible? What about a minimax strategy? If not, think of a simple condition that ensures that the Bayes (or minimax) strategy will be admissible.

Exercise 48: Give a real-world example where the minimax strategy leads to bad results — i.e., playing it too safe (making sure the worst case is not too bad) leads to unhappiness even in the best case.

Bayes estimates under squared-loss: conditional expectation[15]

A concrete example might help to understand these somewhat abstract ideas. The following example will also take us directly into the next part of the course, on estimation theory.

Let's consider our Gaussian game again. Recall, I draw N i.i.d. Gaussian r.v.'s X_i from $\mathcal{N}(\theta, 1)$, and you want to guess θ given the data. The cost function is

$$C(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2.$$

The risk $R(\theta)$, for any choice of $\hat{\theta}(\{X_1, X_2, ...X_N\})$, is

$$R(\theta) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \prod_{i=1}^{N} \mathcal{N}(\theta, 1)(X_i) \Big(\theta - \hat{\theta}(\{X_1, X_2, ...X_N\})\Big)^2 \prod_{i=1}^{N} dX_i.$$

Now assume further that we have some good prior information about the true θ (for example, you might have played this game with me before, or know something about how I choose θ in the first place). Let's say your prior density on θ is Gaussian with mean zero and variance one: that is, you believe I'm drawing θ from $\mathcal{N}(0, 1)$. (If you don't like this "game" example, it should be easy enough to think of a problem in your own line of work where these or similar assumptions might be appropriate.) Then what is the optimal estimate given the data?

It turns out that in general, no matter what the prior $p(\theta)$ or likelihood of the data under θ, $p_\theta(D) = p(D|\theta)$, that under squared error loss the best thing to do is to simply look at the conditional mean. That is, under squared loss, the optimal Bayesian guess is simply

$$\hat{\theta}_{Bayes} = E(\theta|D).$$

This is incredibly useful: instead of solving the very abstract problem of minimizing some complicated expected risk to solve the Bayesian optimal decision problem, all we have to do is compute a conditional expectation. We won't prove this now (although it's not difficult to prove; try it yourself); we'll prove something very similar in a few lectures called the Rao-Blackwell theorem.

[15] HMC 11.2.2

Let's use this idea in this Gaussian example. Let's assume that our prior is itself Gaussian, with mean 0 and variance 1 for simplicity, i.e., assume that

$$p(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\theta^2/2}.$$

Now what is the conditional mean of θ given a single sample x? We have

$$E(\theta|x) = \int_{-\infty}^{\infty} \theta p(\theta|x) d\theta,$$

in general, and by Bayes' rule

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)},$$

with the normalization factor

$$p(x) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta)d\theta$$

and the "likelihood" term $p(x|\theta)$ given by a Gaussian function with (by assumption) unit variance,

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2}.$$

Each of these terms can be written out without too much trouble; the final result is that the optimal Bayes estimator is $x/2$.

Exercise 49: Prove this. (Hint: compute the posterior distribution, $p(\theta|x)$, and then just read off the mean.)

Exercise 50: Generalize the above result: what is the Bayes optimal estimator under squared loss given N i.i.d. samples, x_i, under the $\theta \sim \mathcal{N}(0, 1)$ prior? How does this relate to the sample mean estimator? Does this make intuitive sense? What if a Gaussian prior with different variance is used (i.e., $\theta \sim \mathcal{N}(0, \sigma^2)$?

Exercise 51: Compute the optimal Bayes estimator when the prior is exponential, $p(\theta) = \lambda e^{-\lambda\theta}, \theta > 0$. Interpret this result intuitively.