Pooling Intervals via Density Functions

Administrative Things

- Assignment 8 due Thursday.
- Reading from Axelrod this next week's classes.
- Handout reading for next class (Shed Load – Seen this?)

Today's Topics

- Biases in judgement that affect subjective intervals
- Pooling subjective confidence intervals
  - Contrast with normal-based intervals when disjoint
  - How to do it with disjoint/contradictory intervals?
- Density-based methods for pooling
  - A little bit of background on the simulation method for pooling
  - Replace simulation with some math
- Short-cut methods that work as well as more complex calculations

Review from Last Time

- Subjective intervals
  - Coverage tends to be about 50%, with length used to signal knowledge
  - Variation modeled as Cauchy (not normal)

    (But remember, not all 50% intervals indicate Cauchy variation)
- Simulation method for a subjective interval
  - Center sample at center of interval, with half-width of interval giving the scale of the Cauchy sample (i.e., formula is center + half-width*?cauchy).

- Check your results

  A big advantage of the simulation method is the ease of checking that you have represented the information source correctly.

  If, for example, you have a subjective interval from 100 to 200, then 50% of the simulated values should fall "into the box" (between the quartiles). Similarly, if you have a 95% normal interval, then 95% of the simulated values should fall into the range specified by the interval.

Improving Subjective Intervals

➢ Premise

- Define "better intervals" = better assessment of the uncertainty

- You will get better subjective intervals by avoiding some common human judgment pitfalls. Once you know what these pitfalls are, it is more likely that you can avoid them

- *Decision Traps* gives many more examples, references.

➢ Biases in judgment

- *Availability bias* occurs when we use the available information that is convenient to get your hands on, but not always the most informative.

  - Egocentric (Who does the cleaning?)

  - Bus arrivals (Which one comes first?)

  - Chance for deadly flood in California over next 20 years? (Context)

- *Accepting anecdotal evidence* rather than making a more careful study

  - Which kills more: pigs or sharks?

  - CD players in airplanes

  - Smoking while buckled up

- *Separating random from systematic* when we forget Bonferroni.

  - Streaks in sports

- Rewards versus punishment in manufacturing and education

- *Anchoring* our value in one problem to a number in an unrelated context.

  - What's your area code?

  - In what year was Helen of Troy born?

- *Functional bias* in our view and attitudes to a problem

  - Silo mentality

  - Rational? Source of rewards (like here at Wharton)

Pooling Subjective Information

➢ Bias in subjective intervals, background

- Group-think, conformity. Other forms of dependence.

  No central limit theorem when it comes to subjective intervals. Asking 1,000,000 people for the diameter of the moon will not converge to the right answer!

  Our ultimate procedure for pooling subjective intervals (later today) will reflect this recognition of systematic error in subjective intervals.

- Consultants as an independent, outside source of information

➢ Manipulating several *independent* subjective intervals

- For two, the simulation procedure works exactly as before.

  (1) Generate a large simulated sample for each interval. Remember that the subjective intervals have about 50% coverage (unless you have some reason to believe otherwise, such as from tracking the source).

  (2) Use the subset selection method to pool them, not regression. Since the data are not normal, we can no longer use the slick regression trick.

  JMP notes. You will need to re-scale the histogram to make the values near zero visible; also change the increments. I typically have used a range of –10 to +10 with an increment of 5. For the subset-selection method, identify those near zero in the histogram of the differences, then use the

Subset command from the *Tables* menu to build a new data set with just the selected rows. You might need to use the brush tool to find those close to zero if the plot is very sparse.

Examples of Pooling Subjective Information

➢ Two, *independent* sources in all examples

➢ Two similar sources (data in JMP data file available from web page)

- Estimates of project costs are [3,5] for source A and [4,8] for source B.
- As 50% Cauchy intervals, we obtain (using 10,000 simulated values) a pooled interval of [3.8 to 5.9], roughly the interval from the center of one to the center of the other.
- As 95% normal intervals, we obtain (again via simulation) the interval [3.3 to 5.3], shifted more in the direction of the narrow interval.

➢ Two contradictory, non-overlapping sources with A = [3,5] and C=[12,16]

- As Cauchy 50% intervals, I get an interval something like [4,12] with length 8. Since so few are close to zero in the simulation (even starting from 10,000), you need to be careful about sampling variation. Try several to see what happens.
- As normal, 95% intervals, you will not get any values near zero. They are simply so rare that you would need a much, much larger sample.

Another Look at the Simulation Pooling Method

➢ Questions to ask yourself

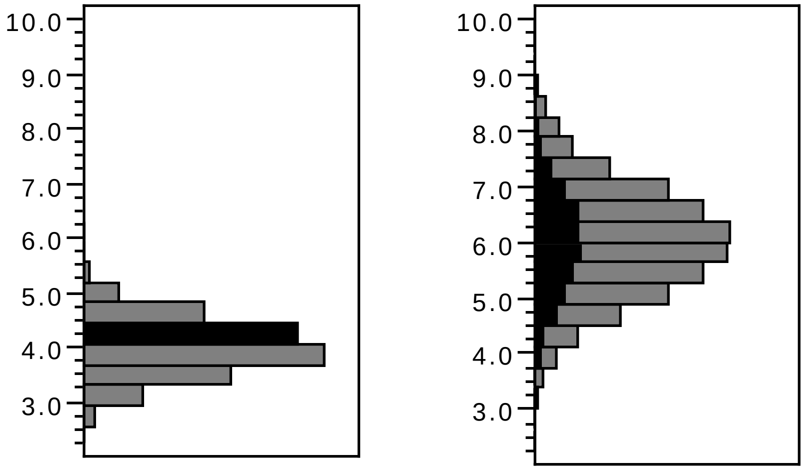
- What are we doing when we use the subset of observations that “match” in the two samples to define a pooled interval?
- We often only get to use a small subset of the simulated data. So, why do we need to restrict ourselves in this manner? What’s going on?

➢ Answer

- The matched observations reveal a “simultaneous” probability that combines the information in the two samples.

➢ Understanding this answer

- To keep the pictures nice looking, consider pooling the two sources above, represented as 95% normal intervals ([3,5] and [4,8]). (Cauchy samples have lots of outliers, and the histograms are not so nice to look at.)



- What are we learning from these histograms?

  The histograms indicate the probabilities that we are assigning to various ranges, and are best thought of in a relative sense. For example, on the left we believe that there is a much larger probability that the unknown value falls into the highlighted range, rather than from 7 to 8, say.

  Denote the probability of an observation in this subset by $P_1(4 \le S \le 4.5)$, where we let S denote the sought quantity.

- What do we learn by looking at both samples?

  The highlighted values on the right are the paired values from the second source. Since the two samples are *independent,* this subset on the right has the same shape as the overall source.

- What do we learn by selecting just those that approximately match?

  Since the shape of the selected subset from the second source matches the overall shape, the chance for an observation to match is the product

$$\begin{gathered}P(\text{observation is highlighted})\ P(\text{observation between 4 and 4.5}) \\ = P_1(4 \le S \le 4.5)\ P_2(4 \le S \le 4.5) = P_c(4 \le S \le 4.5)\end{gathered}$$

This product probability is just the product of the heights of the two histograms matched over common intervals. The histogram of the simulated matched sample is roughly this product (roughly since these are simulated values and there's a lot of sampling variation).

➢ What did we want from pooling anyway?

- Going way back to the start, we wanted to combine the information of the two sources, as represented in the two intervals.

- What does it mean to combine information?

  We want to learn the combined probability (denoted P_c) from the two sources that the unknown value falls into any range.

  This is exactly what we are learning from the matched samples!

➢ Problem with the simulation/subset method

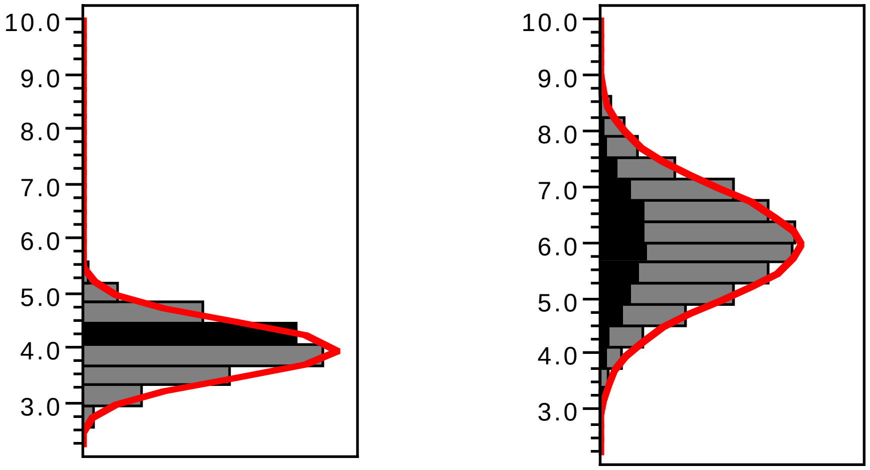
- Since we are computing the product of two probabilities, the value of P_c can get quite small when either of the sources assigns small probability to a range.

- This tiny probability becomes apparent when the sources are normal and contradictory. We just don't get the data to be able to estimate P_c.

➢ Remedy

- Why not just multiply the two histograms?

  Well, you have to match the interval endpoints, get the software to do it, then figure out the interval from the product of the two…



- Better idea is to go yet one step further…

  - We know the underlying "population histogram" is normal since that's how we have simulated these data.

  - Rather than multiply the histogram probabilities and have to deal with simulation variation and intervals, why not just multiply the two known population histograms?

Density Functions

➢ What's a density function

- The "right" name for a population histogram. It's the function that the histogram approaches as the sample size grows infinitely large.

- Interpretation

  A density function *f* describes how probabilities are allocated for different ranges of values. To find a probability for a random variable associated with this density, such as P(a ≤ X ≤ b) say, integrate the density function over that interval,

$$P(a \le X \le b) = \int_a^b f(x)dx$$

  Thus, you can think of a density function as a very fine histogram with very narrow bins of width *dx*. The probability of a value falling into one of these bins is *f(x)dx*.

- Two key properties of a density function

  (1) Positive, $f(x) \ge 0$

  (2) Area integrates to one, $\int f(x)\,dx = 1$. the total probability is one.

➢ Normal density functions

- Standard normal (with 2/3 probability between –1 and +1 and 19/20 between –2 and +2)

$$f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

(The exact range for 95% is –1.96 to 1.96, but ± 2 is easier to remember.)

- JMP does the standard normal for you in its calculator (prob. functions)

- Standard normal with mean μ and variance σ^2 is found by a simple change to this density function

$$f(x) = \frac{1}{\sigma} \times \frac{e^{-((x-\mu)/\sigma)^2/2}}{\sqrt{2\pi}}$$

This expression just replaces x in the standard normal by (x-μ)/ σ and divides the whole expression by σ.

➢ Cauchy density functions

- Standard Cauchy density (with exactly 1/2 probability between –1 and +1)

$$f(x) = \frac{1}{\pi(1+x^2)}$$

- Cauchy chosen to match a subjective interval with center *c* and *half-length w* (i.e., one that puts half of its probability into this region)

$$f(x) = \frac{1}{w} \times \frac{1}{\pi(1+((x-c)/w)^2)}$$

- Note that this construction parallels the one used to get the general normal from the standard normal. We have replaced *x* by *(x-c)/w* and divided the whole expression by the scaling term *w*.

Pooling via Density Functions

➢ Procedure

- Define a density function for each source, such as normal for those from data or Cauchy for a subjective interval, say $f_1(x)$ and $f_2(x)$.

- Match characteristics of the chosen type of density to the source interval.

- Multiply the two density functions to form the product,

$$f(x) = f_1(x)\, f_2(x)$$

➢ Hard parts of this method

- The product of two density functions is not a density. It's not negative, but it will not typically integrate to one.

- Hard task #1

  We have to normalize the product *f(x)* so that it integrates to one.

  Alternatively, we can just "look at the picture" and get a sense of the relative concentration of probabilities.

- Hard task #2

  Now find a 50% or 95% interval or the CEV that comes from the pooled density function.

  This can be very, very hard in cases in which *f(x)* is bimodal (below).

- The simulation method looks more and more attractive since it does not have these two problems.

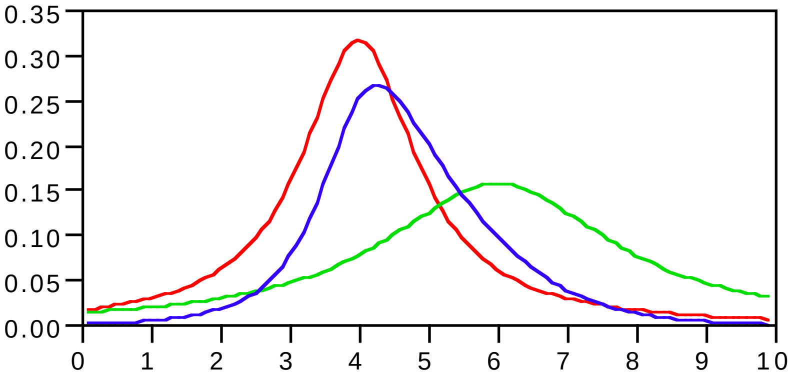
Examples of Pooling via Density Functions

➢ Shortcuts in these examples

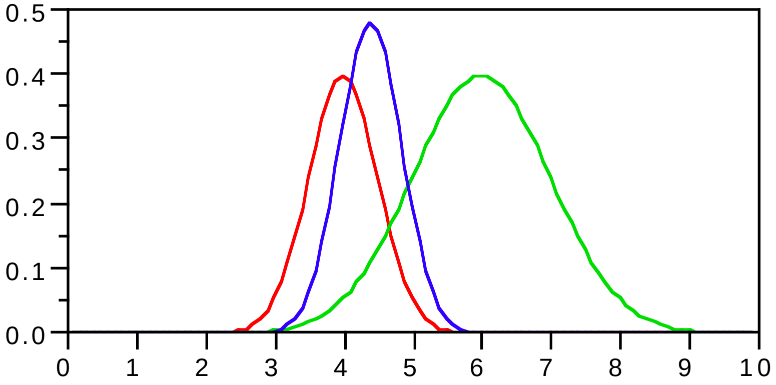
- Graph the product probabilities *k f(x)*, rather than figure out the exact normalizing constant for them.

➢ For two overlapping sources, [3,5]& [4,8]

- Cauchy: a somewhat asymmetric product of the two density functions.



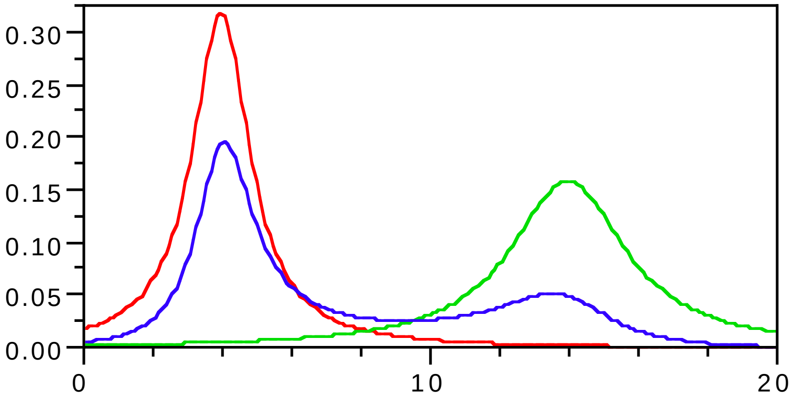
- Normal sources, we obtain another normal as the product. It is also a tighter distribution than for the Cauchy pair.



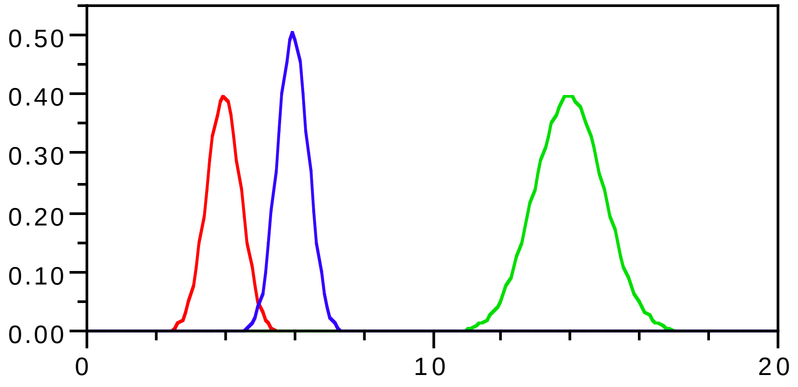
- For two contradictory sources, [3,5]& [12,16]

  - Very big difference this time between normal methods and Cauchy methods for pooling the information.

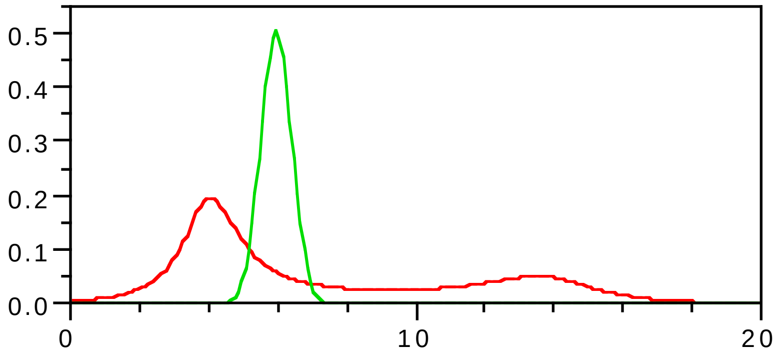
  - For Cauchy sources we obtain this plot of the two density functions and the product. With the scaling fixed, the product – or likelihood – is seen to be bimodal, with two peak areas. Thus, when we simulate the matched samples, we get values from the two peaks as well as the other values in between.



- When the sources are normal, we get a very different picture, with the product showing a third normal located between the other two. Rather than blurring as in the Cauchy case, we get one density with less variation.



- Compared to the normal, the Cauchy is basically flat, indicating the confusion associated with the two disjoint intervals. The normal is concentrated at about 6, just *above* the upper end of the narrow interval [3,5].



- Which is the right characterization for what you know? (Recall the gravity homework assignment.)

Combining More than Two Subjective Sources

➢ Need to use a mathematical method like the density products

- Simulation would take far too large a sample size.
- Have to keep normalizing the product to make product integrate to one. (The pictures shown above are rough estimates of product densities.)
- Need some better tools to make this procedure work more easily.

- Approximate pooling method for subjective intervals
  - Based on the Cauchy model and some other calculations
  - Works empirically with previous experimental data. YMMV.
  - Can use the simulation method for pairs, density method for more, if you have the time and want something more precise.
- Procedure (three steps)
  - Omit the extremes, the wild intervals that are clearly out of touch with the rest of the intervals.
  - Center the pooled interval at the average or median of the centers of the remaining intervals.
  - Set the width of the pooled interval at 2/3 of the typical width (such as the median interval width of those remaining after removing the extremes).
- Big difference from data-based normal intervals
  - The pooled interval is not that much more narrow than any of the constituent intervals. It may even be longer than some.
  - The pooled subjective interval allows for some serious bias.

Next Time

Game theory.