Notes on Subjective Probability Distributions*

Lucas Rosso

1 Overview

Eliciting agent expectations is an ongoing challenge for economic research. In that regard, several surveys have included questions on expectations over aggregate and individual variables (e.g. inflation, GDP growth, own prices, etc). Moreover, since Manski (2004), there has been increasing interest for estimating agents subjective distributions rather than self-reported expectations. Some benefits of this approach include estimating consistent moments across agents and being able to estimate subjective uncertainty measures. Conversely, these questions typically raise concerns on whether respondents are able to correctly report their probabilities, bias raised by bin definitions, among others.

These notes intend to broadly describe different options for computing individual subjective probability distributions from survey data. This will allow us to compute measures of central tendency for each individual (e.g density mean, median) and subjective uncertainty (variance).

Several expectation surveys (SCE, SPF) ask respondents on their expectations over specific variables (growth, interest rate, inflation, etc). I will refer to this response as point prediction. They are also asked to assign probabilities to the occurrence of different events. For example, the probability that output increase by 2-4% over the next 12 months, or that inflation is greater than 12%. These questions present the challenge of computing the respondents subjective distributions, as well moments from this distribution, with an "incomplete" probability distribution.

Why bother computing central tendency expectations? If households already report point estimates for their expectations, why not use that variable rather than computing central tendency expectations? One reason is that most surveys it is not totally clear to the respondents that they are asked for the mean of their subjective probability distribution, and thus can report other statistic such as median or mode. Thus, central tendency expectations provide a more con-

*lrosso@fen.uchile.cl.

sistent measure across individuals (Armantier et al., 2017). Another reason is that estimating a subjective probability distribution allows us to analyze respondents subjective uncertainty.

Table 1 presents an example from the Survey of Consumer Expectations (SCE), where individuals report their probability distribution over 12-month inflation expectations.[1] This table shows two of the main problems faced when computing respondents subjective mean and variance: (i) first and last unbounded; (ii) distribution within each bin is unknown.

Table 1: *Question 9 in the SCE*

And in your view, what would you say is the percent chance that, over the next twelve months, the average home price nationwide will ...

Increase by 12 percent or more	____ *percent chance*
Increase by 8 to 12 percent	____ *percent chance*
Increase by 4 to 8 percent	____ *percent chance*
Increase by 2 to 4 percent	____ *percent chance*
Increase by 0 to 2 percent	____ *percent chance*
Decrease by 0 to 2 percent	____ *percent chance*
Decrease by 2 to 4 percent	____ *percent chance*
Decrease by 4 to 8 percent	____ *percent chance*
Decrease by 8 to 12 percent	____ *percent chance*
Decrease by 12 percent or more	____ *percent chance*

2 Most Basic Approach: Histogram Mean and Variance

The most simple approach to compute measure of central tendency out of subjective distributions is to assume that the probability mass is uniform within intervals. Then, define x_s, as the median value of the bin s (e.g. 3% in the 2-4% bin).

With that, plus some additional assumption over unbounded intervals, one can simply compute central tendency expectations as the mean of the histogram. That is, compute

$$\mathbb{E}(x_{t+1}|\mathcal{I}_{i,t}) = \sum_{s \in S} \mathbb{P}(x_s) x_s, \tag{1}$$

where $\mathbb{E}(x_{t+1}|\mathcal{I}_{i,t})$ denotes expectation of x_{t+1} given information set $\mathcal{I}_{i,t}$, S the set of intervals of x and $\mathbb{P}(x_s)$ the subjective probability of x_s. Likewise, variance (often referred to as subjective

[1]For simplicity, I will often use inflation expectations as my default example, though this is valid for expectations on several other macro variables surveyed.

uncertainty) can be computed as

$$Var(\pi_{t+1}|\mathcal{I}_{i,t}) = \sum_{s \in S}(x_s - \mu)^2 \mathbb{P}(x_s), \tag{2}$$

where $\mu \equiv \mathbb{E}(x_{t+1}|\mathcal{I}_{i,t})$.

Pros. Easy to compute

Cons. Strong assumptions on probability mass within-bins and at the tails of the distribution. Zero uncertainty for individuals that only assign positive probability to one bin (more than 10% of inflation expectations respondents in SCE, see Figure 2).

3 Density Estimation

Another way to estimate respondents subjective distribution is to follow the parametric analysis presented in Engelberg et al. (2009). This method fits a generalized beta distribution to respondents probabilities, minimizing distance of estimated cumulative distribution function (CDF) and revealed CDF points in the data. The authors makes two assumptions. First, that individual probabilistic beliefs are unimodal. Second, that the subjective distribution follows a generalized beta distribution. With both assumptions, they estimate 4 parameters using the data, 2 to describe the shape of beliefs and 2 to give their support (due to unbounded intervals).

It is possible to fit a unique beta distribution to the data iff respondents assign positive probability to at least three bins. For the remaining respondents, Engelberg et al. (2009) propose to assume that the distribution has the shape of an isosceles triangle. In particular, they proceed as follows.

1. **Respondent uses 1 bin.** Subjective distribution has the shape of an isosceles triangle whose support is the interval. Then for example, if an individual places all probability in the interval 0-2%, the distribution will have support $[0 - 0.02]$, the center of the triangle will at 0.01 and the height of the triangle will be 100.
2. **If respondent uses 2 bins.** For respondents that use two *adjacent* intervals[2] the procedure is as follows. If respondent assigns the same probability to both intervals, the support is the union of the intervals and then fit an isosceles triangle as the 1-bin case.
   If the respondent places more probability to one interval than the other, assume that the

[2]Typically, almost all respondent selecting two bins use adjacent bins. The few observations with two *non-adjacent* intervals are dropped.

support of the subjective distribution contains the entirety of the more probable interval, and only a fraction of the least probable one (see Figure 1). Again, the distribution has the shape of an isosceles triangle, with mass $\alpha < 0.5$ on the smaller interval and mass $1-\alpha$ on the larger one.

We still need to identify the cutoff for the least probable interval and the height of the triangle. Using the same example of Figure 1, note that the height is given by $h = \frac{200}{t+1}$. Further, as the triangle with area α is similar to a triangle given by the height of the triangle and $(t+1)/2$, we have that:

$$\underbrace{\alpha}_{\substack{\text{mass small}\\ \text{interval}}} = \underbrace{\frac{1}{2}}_{\substack{\text{area half}\\ \text{isoceles}}} \times \underbrace{\left(\frac{2t}{t+1}\right)^2}_{\text{ratio areas}} \tag{3}$$

Solving for t yields

$$t = \frac{\sqrt{\frac{\alpha}{2}}}{1-\sqrt{\frac{\alpha}{2}}}. \tag{4}$$

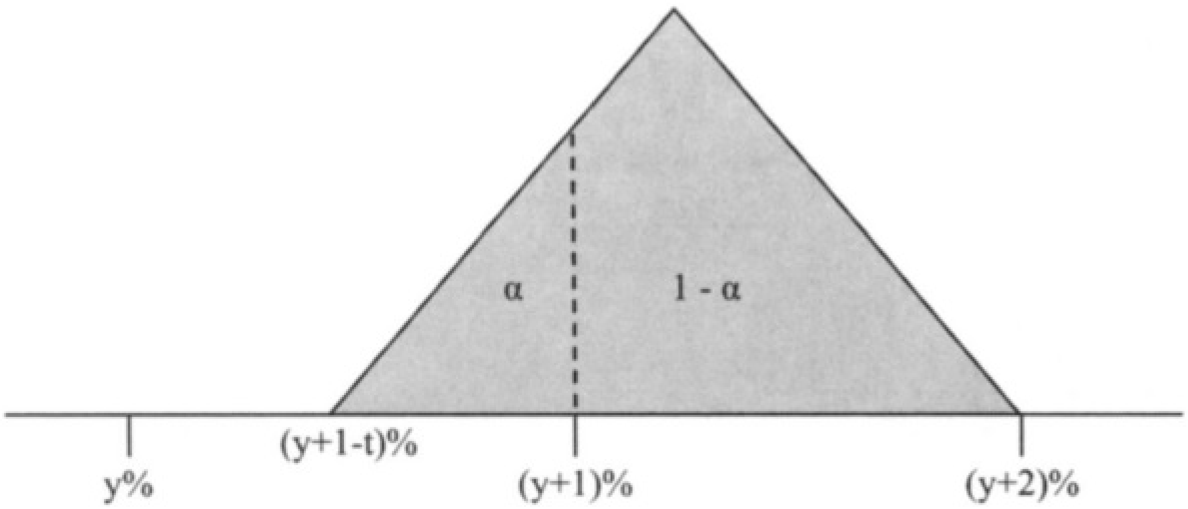
To get a functional form for the density, take the general case where we denote the left and right edges of the base as l and r respectively. Given the height $h = 2/(r-l)$, we can get both slopes using the points $((r-l)/2, h)$, $(l, 0)$ and $(r, 0)$. Then using any of the latter two points we get the intercept. With that, density $f(x)$ is given by

$$f(x) = \begin{cases} \frac{4}{(r-l)^2}(x-l) & l < x \leq \frac{l+r}{2} \\ \frac{4}{(r-l)^2}(r-x) & \frac{l+r}{2} < x \leq r \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Finally, some important notes for cases related to bins with unequal widths and open intervals (at both ends of the distribution). For the former, the support of the triangle is assumed to include the entire smaller-width interval if $\alpha > 0.4$ and include the entire larger-width interval otherwise. Then, proceed as the equal-width interval case to get t and h. For the latter, if one of the intervals include an open-ended interval, the base includes the entire inner interval, and part of the open-ended adjacent.

3. **If respondent uses more than 2 bins.** As noted by Engelberg et al. (2009), forecaster's probabilities reveal points in the cumulative distribution function (CDF) of his/her beliefs. For example, if a respondent assigns 20% chance to inflation to be on the range $1-2\%$, 70% chance to the range $2\%-3\%$ and 10% to the range $3\%-4\%$, then $F(0.01) = 0$, $F(0.02) = 0.2$, $F(0.03) = 0.9$ and $F(0.04) = 1$. In the more common case were the respondent use more than two intervals, fit a unimodal generalized Beta distribution to the data.

Figure 1: *Illustration of two-interval case in (Engelberg et al., 2009)*



The generalized Beta distribution CDF is given by

$$\hat{F}(x) = \begin{cases} 0 & x \leq l \\ \frac{1}{B(\alpha,\beta)} \int_l^t \frac{(x-l)^{\alpha-1}(r-x)^{\beta-1}}{(r-l)^{\alpha+\beta-1}} \mathrm{d}x & l < x \leq r \\ 1 & x > r, \end{cases} \tag{6}$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}\mathrm{d}x$. The parameters l and r are the outer bounds of the bottom/top intervals with positive probability, specifying the support of the distribution. We choose α and β.

Namely, to fit the distribution, assume the respondent observed points in the CDF $F(x_j)$, where $x_j \in \{x_1, \ldots, x_J\}$ denotes the right endpoints of the J intervals. Then, solve

$$\min_{\alpha>1,\beta>1} \sum_{j=1}^{J} \left(\hat{F}(x_j) - F(x_j)\right)^2. \tag{7}$$

If the forecaster assigns positive probability to the top and/or bottom interval, then we also need to choose l and r. For that, add the restriction that l and r cannot exceed (in magnitude) the largest observation seen in the aggregate data. For example, if the respondent assigns positive probability to the top interval, then solve

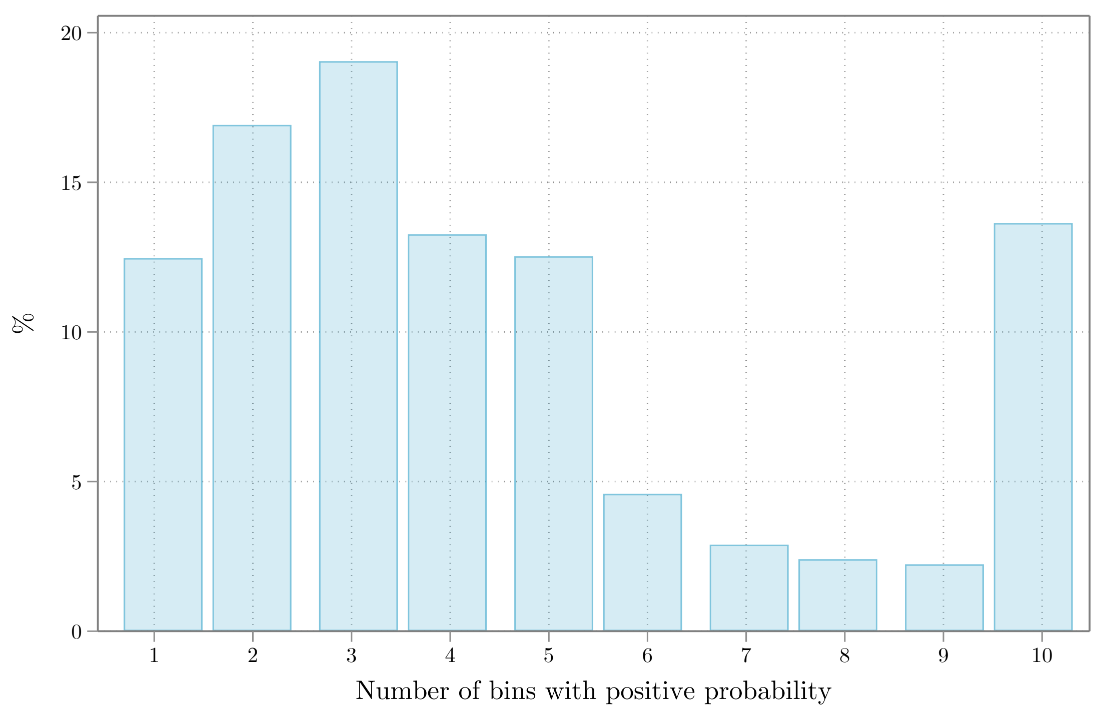
$$\min_{\alpha>1,\beta>1,r\leq r^{max}} \sum_{j=1}^{J} \left(\hat{F}(x_j) - F(x_j)\right)^2, \tag{8}$$

where r^{max} denotes the maximum value observed for the forecasted variable in the data. This method can be somewhat difficult to compute for individual-level forecasts. Also, surveys typically don't include moments estimated under this method. To my knowledge, there are two notable exceptions. First, the SCE that includes standard statistics (mean, variance, IQR, etc) estimated from households' subjective distribution. Second, replication files from Wang (2022) that provide individual moments from the SPF.

Pros. Uncertainty for respondents with one bin is not zero. Rather than assuming edges for the open-ended intervals, it includes to additional parameters for those edges on the optimization problem.

Cons. Difficult to code. Does not work well when respondents report less than 3 bins with positive probability, which is typically a large share. For example, in the SCE, roughly 30% of respondents only use one or two bins (see Figure 2).

Figure 2: *Distributions of Bins with Positive Probability in SCE*



Notes: This figure reports the distribution (in %)of number of bins with positive reported probability in the SCE over the period 2013-2021.

4 Comparing methods

Under complete rationality, point prediction of expectation should be equal to the central tendency expectation. However, for both the histogram mean and the expectations derived from the density estimation, I am making strong assumptions to compute the subjective distributions. Also, some respondents may be simply give a different statistic (median, mode).

Table 2 presents descriptive statistics on inflation expectations reported by individuals (point prediction), histogram mean and central tendency expectations in the SCE. Note that households report higher inflation expectations than the ones predicted from their subjective distribution. Also note from Panel B over half of the sample present deviations higher than 1%. This result is substantially higher the one found in Engelberg et al. (2009) for GDP growth (80% within 1 percent) using professional forecasters (SPF).

Table 2: *Descriptive Stats on Reported and Computed Inflation Expectations*

	Panel A: Descriptive statistics of inflation expectations								
Stat.	Mean	Std. Dev.	p25	p50	p90	p99	Min	Max	Obs.
Point Predicion (PP)	6.11	14.46	2.00	3.00	20.00	60.00	-75.00	100.00	121,311
Histogram Mean (HM)	4.05	4.80	1.28	3.00	10.00	19.67	-12.00	19.67	121,630
Central Tendency Exp. (CT)	3.80	5.55	1.16	2.93	9.88	25.00	-36.28	36.28	120,084
	Panel B: Differences between point predicions and central tendency expectation								
Share (%)	< 0.1	< 1	< 3	< 5	< 10	< 15	< 25	< 50	> 50
PP - HM	10.03	44.48	76.58	83.67	90.15	92.85	95.87	98.71	1.54
PP - CT	12.41	46.98	77.04	83.11	89.51	92.79	95.88	98.64	2.72

Notes: This table presents descriptive statistics of households inflation expectation using data from the SCE. Panal A presents stats for reported (point prediction) inflation expectation and computed (central tendency from specific event probabilities). Panel B presents the share of observations with difference between reported and computed inflation expectation within different thresholds.

References

Armantier, O., G. Topa, W. Van der Klaauw, and B. Zafar (2017): "An overview of the survey of consumer expectations," *Economic Policy Review*, 51–72.

Engelberg, J., C. F. Manski, and J. Williams (2009): "Comparing the point predictions and subjective probability distributions of professional forecasters," *Journal of Business & Economic Statistics*, 27, 30–41.

Manski, C. F. (2004): "Measuring expectations," *Econometrica*, 72, 1329–1376.

Wang, T. (2022): "How Do Agents Form Inflation Expectations? Evidence from the Forecast Uncertainty."

[2] Also, if respondents report subjective mean rather than other statistic.