EUROPEAN CENTRAL BANK

EUROSYSTEM

Working Paper Series

Jonas Dovern, Alexander Glas, Geoff Kenny

Testing for differences in survey-based density expectations: a compositional data approach

No 2791 / February 2023

Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Abstract

We propose to treat survey-based density expectations as compositional data when testing either for heterogeneity in density forecasts across different groups of agents or for changes over time. Monte Carlo simulations show that the proposed test has more power relative to both a bootstrap approach based on the KLIC and an approach which involves multiple testing for differences of individual parts of the density. In addition, the test is computaionally much faster than the KLIC-based one, which relies on simulations, and allows for comparisons across multiple groups. Using density expectations from the ECB Survey of Professional Forecasters and the U.S. Survey of Consumer Expectations, we show the usefulness of the test in detecting possible changes in density expectations over time and across different types of forecasters.

JEL classification: C12, D84, E27
Keywords: compositional data, density forecasts, survey forecasts, disagreement.

Non-Technical Summary

Being able to test for differences amongst economic variables or groups of economic variables is a central concern in empirical economics. For example, much recent work in macroeconomics and monetary economics has emphasized the importance of heterogeneity in the behavior and expectations of economic agents. However, a powerful test is needed to determine if any observed difference across variables or agents is genuinely robust. Equally, tests of difference are important for identifying changes in economic variables over time, i.e., their temporal stability. In a monetary policy context, a good example in this respect is ways to test for possible changes in agent' medium- to long-term inflation expectations in the current context of elevated inflation.

In this paper, we suggest a new method to test for differences in survey-based density expectations that recently have become popular. These density expectations offer a more complete picture of expectations relative to conventional point or mean expectations which indicate only the average outcome that is expected. Using such probabilistic expectations, it is possible to understand not just the average outcome that people expect but also how agents assess the uncertainty, skewness and tail risks related to the possible future evolution of the economy.

The central insight that we develop in this paper is that one should treat such density expectations that surveys elicit in the form of vectors of event probabilities as compositional data. Compositional data consist of vectors of proportions (in our context probabilities) that are subject to the constraint that the sum of all elements must equal a fixed value (in our context this value is set to one). We show how tests that have been developed for compositional data can be used to test for differences in probabilistic expectations across different groups of individuals or survey waves. Using a controlled simulation environment, we show that the proposed compositional approach is much more powerful than other commonly used methods. As a result, it may be able to detect important differences in probabilistic expectations that other methods may fail to detect. Moreover, our proposed test is much faster computationally because it does not require extensive computer simulations due to the fact that the underlying distribution of the test statistic is known under the null hypothesis that both sets of density expectations are from the same population, i.e., that there is no heterogeneity across groups or change over time. In addition, the test allows for a joint comparison across multiple groups or observations, whereas the alternative approaches that are currently available do not allow for such joint comparisons. With such joint comparisons it is possible to find out if the probabilistic expectations of one group of agents is different from the expectations of all other groups of agents that are available for analysis.

The paper then applies the proposed tests to four different research questions. Specifically, we test for heterogeneity of expectations and changes over time among different groups of professional macroeconomic forecasters (based on data from the European Cen-

tral Bank's Survey of Professional Forecasters) and private households (based on data from the Federal Reserve Bank of New York's Survey of Consumer Expectations). In these applications, we show that (i) professional forecasters quickly changed their short-term inflation density forecasts in response to the recent period of rising inflation rates, whereas long-term expectations reacted much more gradually, (ii) for most periods inflation and GDP growth expectations significantly differ between experts that round their probability statements and those that do not, (iii) comparing the inflation density forecasts of men and women there is strong evidence for gender differences in these expectations, and (iv) for a substantial fraction of periods in our sample households from different regions of the U.S. report significantly different density expectations for the future change of nationwide house prices.

Overall, our findings show that it is beneficial to treat survey-based density expectations as compositional data. This might be relevant also in other contexts where such survey data is used. Our results could be extended by using the panel structure of most expectation surveys. So far, we have analyzed each survey wave independently as separate samples but one could also jointly analyze the full sample of expectation data. For instance, by adopting a dynamic model for the compositional expectation data.

1 Introduction

Expectations are central both for microeconomic decision-making and for macroeconomic dynamics. Hence, it is not surprising that a large body of literature studies the properties of expectations in various economic contexts. In recent years, the use of survey-based expectation data has become increasingly more common. This process is particularly strong in macroeconomics where more and more surveys are set up to study the macroeconomic expectations of private households (D'Acunto et al., 2021; Conrad et al., 2022), firms (Coibion et al., 2018, 2020; Andrade et al., 2022), and professional forecasters (Rich and Tracy, 2021; Glas and Hartmann, 2022).

We observe two tendencies in this literature that we want to bring together in our paper. On the one hand, there is a growing focus on understanding the reasons for and the effects of heterogeneity of macroeconomic expectations at least since the seminal contribution by Mankiw et al. (2003). On the other hand, there is a tendency towards the analysis of probabilistic (density) expectations that offer a more complete picture of expectations relative to conventional point expectations (Manski, 2018). What is missing, so far, are major efforts to combine these two important aspects.

The contribution of this paper is to suggest a method that can be used to test for heterogeneity of probabilistic expectations. Such probabilistic expectations are usually elicited by asking agents to assign probabilities to intervals of potential outcomes. Our central insight is that these vectors of probabilities are compositional data. Compositional data consist of vectors of proportions (here: probabilities) that are subject to the constraint that the sum of all elements must equal a fixed value (here: one), see Aitchison (1982, 1986). We propose to use tests that have been developed for compositional data to test for differences in probabilistic expectations across different groups of individuals or survey waves.

Using Monte Carlo simulations, we show that the proposed compositional approach is more powerful than an alternative bootstrap-based approach that builds on the more traditional way of comparing (expectation) distributions using distance measures such as the Kullback Leibler Information Criterion (KLIC). Moreover, our proposed test is much faster because it does not require simulations due to the fact that the distribution of the test statistic is known under the null hypothesis. In addition, the test allows for a joint comparison of multiple groups, whereas the KLIC-based approach can only be used to compare two groups at a time. We then apply the method to four different research questions that have recently been discussed in the literature. Specifically, we test for heterogeneity of expectations and changes over time among different groups of professional macroeconomic forecasters (based on data from the European Central Bank's Survey of Professional Forecasters) and private households (based on data from the Federal Reserve Bank of New York's Survey of Consumer Expectations).

The results from the Monte Carlo simulations suggest that the tests designed for compositional data have high power against alternatives that imply only moderate differences in density expectations between two subpopulations, especially when the sample size is relatively small. In the applications, we show that (i) professional forecasters quickly changed their short-term inflation density forecasts in response to the recent period of rising inflation rates, whereas long-term expectations reacted more gradually, (ii) for most periods inflation and GDP growth expectations significantly differ between experts that round their probability statements and those that do not, (iii) there is strong evidence against the hypothesis that private households' inflation expectations reported by men and women are equal and (iv) for a substantial fraction of periods in our sample households from different regions report significantly different density expectations for the future change of nationwide house prices.

First and foremost, our work relates to other studies that analyze heterogeneity of macroeconomic expectations across individuals or firms. For example, Malmendier and Nagel (2011, 2016) show that U.S. households who experienced low stock market returns and/or high inflation rates during their lifetime tend to be more pessimistic with respect to future stock market developments and/or inflation than individuals with more moderate life-time experiences. Similarly, Kuchler and Zafar (2019) find that local house price experiences affect households' expectations about future house price changes. In particular, the experience of volatile house prices leads to a higher dispersion of house price expectations. With respect to firm expectations, Kumar et al. (2015) find that the inflation expectations of firm managers in New Zealand are heavily dispersed, at odds with the notion of anchored or fully rational expectations. A common feature shared by these studies is that they focus on point forecasts. Our contribution is to provide methods that allow us to analyze heterogeneity of density expectations and, thus, to move beyond the analysis of heterogeneity of point expectations.[1]

In terms of the methodology used, our work relates to—and borrows heavily from—the literature on compositional data. Aitchison (1986) and Filzmoser et al. (2018) offer comprehensive overviews of methodological aspects that are important when dealing with such data. The methods are widely applied in many disciplines, including geochemistry (e.g., Reimann et al., 2012; Buccianti, 2018), sedimentology (Weltje and von Eynatten, 2004), demography (Lloyd et al., 2012) and medicine (Kitano et al., 2020; Braga and Feingenbaun, 2020). In economics, methods for compositional data have been used, for instance, to analyze income or expenditure shares (Fry et al., 1996) and how time budgets

[1] A related paper that analyzes heterogeneity of density expectations is Mitchell and Hall (2005), who use the KLIC as a measure of heterogeneity. They propose a KLIC-based test of equal predictive accuracy of two density forecasts that is conceptually similar to the popular Diebold-Mariano test. Comparing the density forecasts ('fan charts') for inflation in the UK reported by the Bank of England and the National Institute of Economic and Social Research, Mitchell and Hall (2005) find that the former tend to be more accurate than the latter.

are shared for different activities (Gupta et al., 2020). Our contribution is to show that these methods are also relevant and helpful when dealing with probabilistic expectations.

The rest of this paper is structured as follows. Section 2 briefly summarizes the basics of compositional data and describes the tests that we propose to use for the analysis of heterogeneity and temporal stability in probabilistic expectations. Section 3 presents the results from the Monte Carlo simulations that we use to assess the properties of the tests. Section 4 describes the applications of the proposed method. Section 5 concludes.

2 Methodology

We consider probabilistic survey expectations reported by individuals $i = 1, \ldots, N$ at time $t = 1, \ldots, T$ for some future (macroeconomic) outcome in period $t+h$ so that h indicates the forecast horizon. In practice, such expectations are usually elicited by asking subjects to assign probabilities to a set of K different outcome intervals (or 'bins'). The assigned values indicate the probabilities by which subjects expect the outcome to fall into the corresponding intervals. Hence, each probabilistic expectation is characterized by a vector $\mathbf{p}_{i,t,h} = (p_{i,t,h,1}, \ldots, p_{i,t,h,K})'$ with non-negative elements $p_{i,t,h,k}$ for $k = 1, \ldots, K$.[2] Since the union of all intervals covers the entire outcome space, a natural constraint (which is usually enforced by the survey design) is that $p_{i,t,h,1} + p_{i,t,h,2} + \ldots + p_{i,t,h,K} = 1$.

We are interested in the following problem: given two sets of density forecasts, denoted as $g \in \{A, B\}$, we want to test the null hypothesis that individuals from both groups draw their probabilistic expectations from the same distribution. More formally, let $(\mu^g_{t,h,1}, \mu^g_{t,h,2}, \ldots, \mu^g_{t,h,K})' = \boldsymbol{\mu}^g_{t,h} = \mathbf{E}(\mathbf{p}^g_{i,t,h})$ denote the expected value of the vector of interval probabilities for any individual from group g. Our null hypothesis then is H_0: $\boldsymbol{\mu}^A_{t,h} = \boldsymbol{\mu}^B_{t,h}$ against the alternative hypothesis that H_1: $\boldsymbol{\mu}^A_{t,h} \neq \boldsymbol{\mu}^B_{t,h}$. We want to test this hypothesis about the two population moments using two samples of observed probabilistic expectations of size N_A and N_B (with $N_A + N_B = N$).

In the following subsections, we first propose a test for analyzing differences of histogram forecasts across groups of individuals that we borrow from the literature on analyses of compositional data.[3] We then describe two alternative approaches. The first alternative breaks down our null hypothesis into K interval-specific testable hypothesis and uses the Bonferroni correction to control for the size of the test of the primary null hypothesis. The second alternative is a bootstrap-based approach that is based on the traditional way of measuring the dissimilarity between two distributions using the KLIC.

[2] In the applications below, we require strictly positive probabilities to simplify the analysis. Hence, we replace all zero entries by very small numbers in the applications and adjust the other entries accordingly to ensure that the unit constraint is still met.

[3] We use the terms 'histogram' and 'density forecast' as synonyms when referring to these kind of expectation data.

2.1 Compositional Data Approach

Treating expectations of the form considered in this paper as compositional data starts from the insight that the sum of all probabilities must equal one. With respect to the statistical modeling of the distribution of the vectors of probabilities $\mathbf{p}_{i,t,h}$ this implies that the sample space is not simply the K-dimensional space of non-negative[4] real numbers $\mathbb{R}_+^K$ but the so-called $K-1$-dimensional *simplex* defined by

$$\mathbb{S}^{K-1} = \{(p_{i,t,h,1}, \ldots, p_{i,t,h,K}) : p_{i,t,h,1} \geq 0, \ldots, p_{i,t,h,K} \geq 0;\ p_{i,t,h,1} + \ldots + p_{i,t,h,K} = 1\}. \quad (1)$$

Failing to take account of this—by applying 'standard' statistical methods—will lead to various problems, including problematic interpretation of the covariance of the interval probabilities (see Aitchison, 1986, Chapter 3).

Instead, one needs to apply a proper one-to-one transformation that leads to a vector of random variables that one can handle more easily. Commonly, the *additive logratio transformation* is used and we adopt this choice in our paper, too. Choosing the Kth probability as the reference category (without loss of generality), the transformed expectation data is given by

$$\tilde{p}_{i,t,h,k} = \ln\left(\frac{p_{i,t,h,k}}{p_{i,t,h,K}}\right) \qquad \text{for } k = 1, \ldots, K-1. \quad (2)$$

This transformation makes the constraint that elements must add up to one obsolete. Instead, the sample space for the transformed object $\tilde{\mathbf{p}}_{i,t,h} = (\tilde{p}_{i,t,h,1}, \ldots, \tilde{p}_{i,t,h,K-1})'$ is $\mathbb{R}^{K-1}$. We will assume that $\tilde{\mathbf{p}}_{i,t,h}$ follows a multivariate normal distribution $\mathcal{N}(\tilde{\boldsymbol{\mu}}_{\boldsymbol{t,h}}, \boldsymbol{\Sigma}_{t,h})$ with $(K-1)$-dimensional mean vector $\tilde{\boldsymbol{\mu}}_{\boldsymbol{t,h}}$ and $(K-1)\times(K-1)$-dimensional covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{t,h}}$.[5] This implies that the original vector of probabilities $\mathbf{p}_{i,t,h}$ follows an *additive logistic normal distribution* according to the definition in Aitchison (1986, p. 113).

The above stated hypothesis test translates into a simple test of equality of population means in two subpopulations, i.e., H_0: $\tilde{\boldsymbol{\mu}}^A = \tilde{\boldsymbol{\mu}}^B$ versus H_1: $\tilde{\boldsymbol{\mu}}^A \neq \tilde{\boldsymbol{\mu}}^B$, leaving aside for a moment the indices for different time periods and horizons.[6] This can be implemented by a Hotelling test using the test statistic

$$\mathcal{Q} = \frac{N_A N_B (N_A + N_B - K)}{(N_A + N_B)(N_A + N_B - 2)(K-1)} (\bar{\tilde{\boldsymbol{p}}}_A - \bar{\tilde{\boldsymbol{p}}}_B)' \boldsymbol{S}^{-1} (\bar{\tilde{\boldsymbol{p}}}_A - \bar{\tilde{\boldsymbol{p}}}_B), \quad (3)$$

[4] Martín-Fernández et al. (2003) discuss various strategies of dealing with zeroes and missing values in compositional data.

[5] Using the suitable variant of the central limit theorem (Aitchison, 1986, p. 124), the normality assumption for the distribution of transformed probabilities can be relaxed when conducting tests based on sample sizes that are sufficiently large.

[6] We assume that $\boldsymbol{\Sigma}$ is the same in each subpopulation. However, this assumption can be easily relaxed in the used framework.

where

$$\bar{\tilde{\boldsymbol{p}}}_A = \frac{1}{N_A}\sum_{i=1}^{N_A}\tilde{\mathbf{p}}_i,$$
$$\bar{\tilde{\boldsymbol{p}}}_B = \frac{1}{N_B}\sum_{j=1}^{N_B}\tilde{\mathbf{p}}_j$$

and

$$\boldsymbol{S} = \frac{1}{N_A+N_B}\left(\sum_{i=1}^{N_A}(\tilde{\mathbf{p}}_i - \bar{\tilde{\boldsymbol{p}}}_A)(\tilde{\mathbf{p}}_i - \bar{\tilde{\boldsymbol{p}}}_A)' + \sum_{j=1}^{N_B}(\tilde{\mathbf{p}}_j - \bar{\tilde{\boldsymbol{p}}}_B)(\tilde{\mathbf{p}}_j - \bar{\tilde{\boldsymbol{p}}}_B)'\right)$$

denote the maximum likelihood estimates of the population parameters. The test statistic $\mathcal{Q}$ in Eqn. (3) follows an F distribution with $K-1$ and N_A+N_B-K degrees of freedom.[7]

One advantage of treating histogram expectations as compositional data when testing for mean differences across groups is that we can easily extend the approach to allow for a joint comparison of more than two groups. This can be done by applying an ANOVA-type analysis to test the null hypothesis $H_0 : \tilde{\boldsymbol{\mu}}^1 = \tilde{\boldsymbol{\mu}}^2 = \ldots = \tilde{\boldsymbol{\mu}}^G$ against the alternative that at least one mean is different from the others. Another benefit is that the computations necessary for this approach are very fast which is a key advantage over the commonly-used KLIC-based approach discussed below in Section 2.3.

2.2 Multiple Testing Bonferroni Approach

The second approach for testing the null hypothesis described above deconstructs the histograms and compares the probabilistic expectations interval by interval. The primary null hypothesis implies for all $k = 1, \ldots, K$ that $H_0^k : \mu^A_{t,h,k} = \mu^B_{t,h,k}$ is true. The alternative in each case is $H_1^k : \mu^A_{t,h,k} \neq \mu^B_{t,h,k}$. For each k, we can use a standard two-sample t-test to test this. The primary null hypothesis is rejected if we can reject the implied null hypothesis for at least one of the bins. To ensure good small sample properties, we apply the test to the log probabilities, i.e., $\ln(p_{i,t,h,k})$ for $k = 1, \ldots, K$.

Since this approach gets us into a multiple-testing setup, we have to apply a correction to the significance level used for the individual t-tests to control the overall size of our testing approach. A common approach to do so is the Bonferroni correction that implies using a significance level of α/K for each individual hypothesis H_0^k, where α is the overall size that should be achieved.

Similar to the Hotelling test from the previous subsection, the approach described here can deal with more than two groups and does not require much computing power. A drawback of this approach is that the Bonferroni correction is known to be conservative

[7] See Section 7.5 of Aitchison (1986) for more details about testing hypotheses about the population parameters of subsamples of vectors that follow an additive logistic normal distribution.

('undersized') when the individual test statistics are correlated which—due to the compositional nature of our data—is the case in our context. This reduces the power of the testing approach.

2.3 KLIC-Based Approach

The third approach for testing the primary null hypothesis starts from the fact that the KLIC is commonly used to compare probability distributions. The KLIC describes the expected value of the logarithmic difference between two sets of probability distributions (Mitchell and Hall, 2005). For discrete probability distributions, such as the histogram forecasts described above, the KLIC is defined as

$$\mathrm{KLIC}(\bar{\boldsymbol{p}}^A_{t,h}, \bar{\boldsymbol{p}}^B_{t,h}) = \sum_{k=1}^{K} \bar{p}^A_{t,h,k} \ln\left(\frac{\bar{p}^A_{t,h,k}}{\bar{p}^B_{t,h,k}}\right). \tag{4}$$

In Eqn. (4), the elements of the vectors $\bar{\boldsymbol{p}}^A_{t,h}$ and $\bar{\boldsymbol{p}}^B_{t,h}$ represent the average (non-transformed) probability mass assigned to bin k based on the individuals in a particular group.

Under the null hypothesis defined above, the aggregate distributions of both groups are very similar for finite group sizes and asymptotically identical. In this case, the KLIC from Eqn. (4) is close to zero. The more $\bar{p}^A_{t,h,k}$ and $\bar{p}^B_{t,h,k}$ deviate from each other, the larger the value of $\mathrm{KLIC}(\bar{\boldsymbol{p}}^A_{t,h}, \bar{\boldsymbol{p}}^B_{t,h})$. To test whether $\mathrm{KLIC}(\bar{\boldsymbol{p}}^A_{t,h}, \bar{\boldsymbol{p}}^B_{t,h})$ is significantly different from zero and, hence, the null hypothesis should be rejected, we use a bootstrap approach. Specifically, we draw Z random samples of size N with replacement from the available density expectation data. We then randomly assign N_A of the drawn histograms to group A and N_B drawn histograms to group B. For each bootstrap sample, we then calculate the KLIC as described in Eqn. (4). We conclude that $\mathrm{KLIC}(\bar{\boldsymbol{p}}^A_{t,h}, \bar{\boldsymbol{p}}^B_{t,h})$ is significantly different from zero whenever it exceeds the 95%-quantile of the Z bootstrapped KLIC values.[8]

The use of the KLIC for comparing histogram forecasts has several disadvantages though. First, the KLIC can only be used to compare the probability distributions of two groups. In case the number of groups exceeds two, only pairwise comparisons can be carried out. Second, calculation of the bootstrapped KLICs is computationally intensive. These are severe shortcomings relative to the previously discussed alternatives.

[8]This approach is similar to Clements (2022) who proposes a test for heterogeneity in the revisions of GDP growth expectations in the U.S. Survey of Professional Forecasters. In order to address potential issues due to small sample size, Clements (2022) simulates a set of imaginary SPF participants by randomly drawing from the set of forecast revisions reported in a given survey wave. While his bootstrap approach focuses on revisions of point forecasts, we randomly draw and reassign entire density forecasts.

3 Monte Carlo Simulations

We now assess the properties of the testing approaches discussed in the previous section by means of Monte Carlo simulations. In particular, we compare the rejection frequencies of the Hotelling test, the multiple testing Bonferroni approach and the KLIC-based test under the null hypothesis of equal subpopulation means and various altenatives.

3.1 Simulation Setup

For the Monte Carlo evaluation, and without any loss in generality, we calibrate our benchmark histograms to the one-year-ahead inflation histograms from the 2020Q1 wave of the SPF (see Section 4 for details on the survey). We first obtain an estimate of $\mathbf{\Sigma}$ by applying the additive logratio transformation in Eqn. (2) to the individual histograms and calculating the corresponding covariance matrix. To obtain $\tilde{\boldsymbol{\mu}}^A$, we fit a normal distribution to the aggregate SPF histogram. To do so, we first calculate the mean and standard deviation of the aggregate histogram by assuming that the probability mass in each bin is centered at the midpoint and use those parameters as starting values for the optimization.[9] We then calculate the probability mass for each bin using the fitted normal density and apply the additive logratio transformation to these probabilities.

For each scenario described below, we simulate $S = 2000$ artificial data sets of histograms. We then apply the Hotelling test, the Bonferroni-adjusted t-tests and the KLIC-based test as described in the previous section and calculate the rejection frequencies in each case. For the KLIC-based test we set the number of bootstrap replications to $Z = 250$. Finally, we choose a nominal level of $\alpha = 0.05$.

In practice, the SPF histograms are relatively coarse and many individual histograms do not closely resemble a normal distribution. For the one-year-ahead inflation expectations, almost two third of the SPF participants assign nonzero probability to at most five bins. Therefore, a possible concern could be that the choice of a Gaussian distribution for the simulations is not appropriate for individual survey responses and, thus, might yield a misleading impression of the tests' properties in real applications. To assess how deviations from normality affect the size and power of the tests, we conduct a second set of simulations with a data generating process (DGP) that mimics this data feature. In particular, we transform the simulated histograms into more coarse versions with nonzero probability assigned only to the (two to five) bins with the highest probabilities. For each individual histogram, we randomly draw the precise number of bins with nonzero probability with selection probabilities equal to the relative frequency of observations in

[9] Figure A.1 shows the aggregate SPF histogram (reporting densities instead of bin probabilities) based on the predictions reported by 46 survey participants. Mean and standard deviation based on the 'mass-at-midpoint'-approach are 1.26 percentage points and 0.60 percentage point, respectively. The black line shows the fitted normal distribution, which has a mean of 1.25 percentage points and a standard deviation of 0.54 percentage point.

the SPF with two to five bins (rescaled to sum to unity).[10] We refer to the two settings as the Gaussian setup and the truncated-probabilities setup below.

3.2 Results for Gaussian Setup

In a first step, we analyze whether the different tests are correctly sized for varying group size. For each group, we consider group sizes of 10, 25, 50, 75, 100, 200 and 500 individuals. While a group size of approximately 25 individuals seems to be a good description of surveys among professional forecasters, a group size of 500 individuals is more representative of typical household surveys.

Table 1: Monte Carlo simulation results: size analysis

Group size ($N_A = N_B$)	10	25	50	75	100	200	500
Panel A: Gaussian setup, baseline within-group heterogeneity							
Hotelling	0.055	0.052	0.056	0.048	0.046	0.046	0.048
Bonferroni	0.053	0.050	0.049	0.043	0.033	0.038	0.033
KLIC	0.062	0.050	0.063	0.049	0.054	0.047	0.049
Panel B: Gaussian setup, low within-group heterogeneity (rel. to baseline)							
Hotelling	0.055	0.048	0.048	0.042	0.051	0.057	0.045
Bonferroni	0.035	0.040	0.043	0.040	0.039	0.040	0.038
KLIC	0.047	0.049	0.042	0.051	0.047	0.066	0.050
Panel C: Gaussian setup, high within-group heterogeneity (rel. to baseline)							
Hotelling	0.052	0.049	0.051	0.050	0.058	0.054	0.050
Bonferroni	0.048	0.044	0.042	0.037	0.043	0.037	0.039
KLIC	0.038	0.053	0.047	0.052	0.047	0.052	0.055
Panel D: Truncated-prob. setup, baseline within-group heterogeneity							
Hotelling	0.050	0.047	0.044	0.038	0.049	0.046	0.045
Bonferroni	0.061	0.043	0.043	0.044	0.042	0.041	0.040
KLIC	0.048	0.046	0.052	0.046	0.053	0.045	0.049

Notes: The panels shows rejection frequencies for the Hotelling test, the multiple testing approach and the KLIC-based test under the null hypothesis of no expectation difference between two groups for varying group size. In the simulations, all tests are used with a nominal size of 0.05. Panel A presents results for the Gaussian setup with baseline within-group heterogeneity. Panels B and C show rejection rates when we multiply the baseline $\mathbf{\Sigma}$ by 0.5 and 5, respectively. Panel D presents our findings for the truncated-probabilities setup with baseline within-group heterogeneity.

[10] These frequencies are 9.9%, 20.0%, 16.4% and 15.1% for the one-year-ahead inflation expectations in the SPF.

Table 1 shows the rejection frequencies of all three tests under the null hypothesis. While panel A shows the results for our baseline calibration of the covariance matrix that determines the within-group heterogeneity, panels B and C present findings for lower/higher within-group heterogeneity. In these settings, we multiply $\mathbf{\Sigma}$ by a factor c, where c equals 0.5 (panel B) or 5 (panel C). For both the Hotelling test for compositional data and the KLIC-based approach (and for all sample sizes and levels of within-group heterogeneity) the empirical size is very close to the nominal size of 0.05. In contrast, the multiple testing Bonferroni approach is, as expected, undersized. With respect to the speed of the MC simulations, performing the KLIC-based test 2,000 times takes a little more than 14 hours on a standard desktop computer while 2,000 Hotelling tests take only nine seconds.

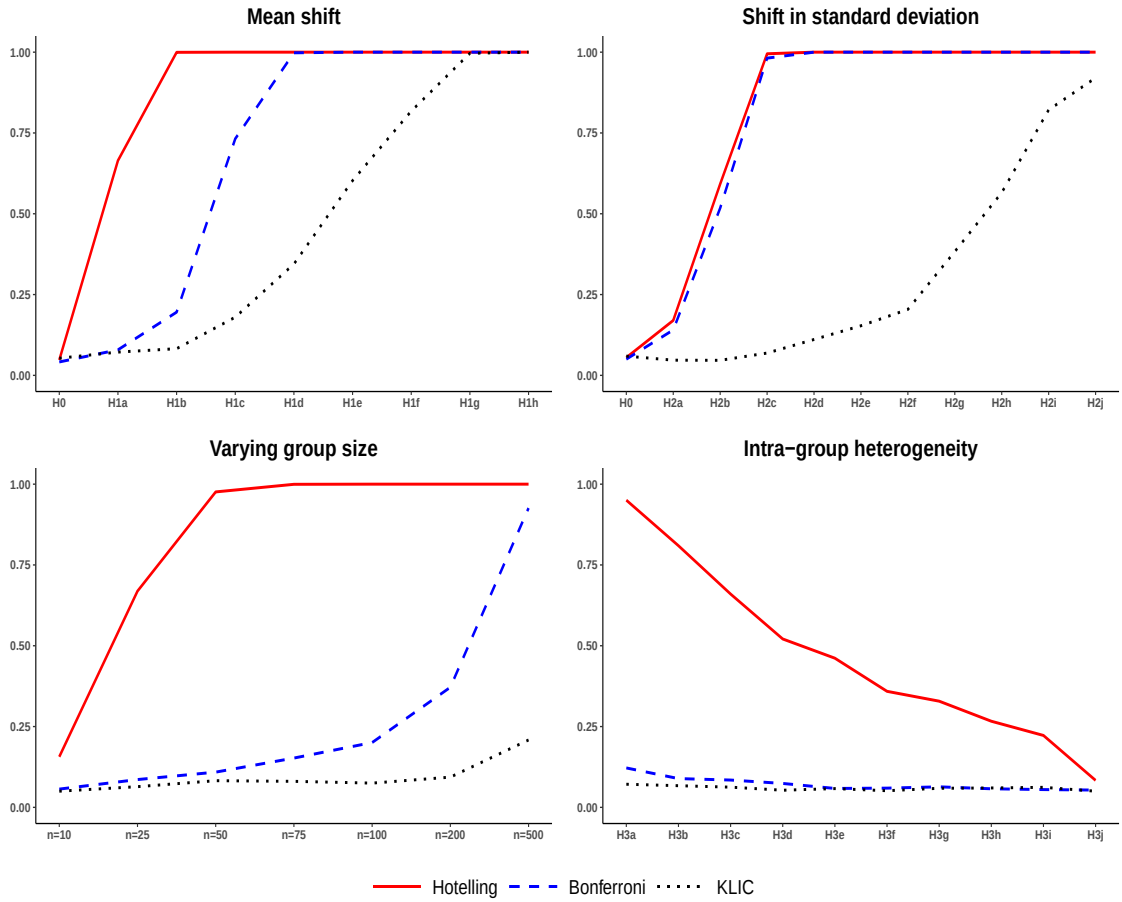
Next, we turn to an assessment of the power of the tests. Unless explicitly stated otherwise, we set the group sizes to $N_A = N_B = 25$. Since the Bonferroni approach is too conservative in the sense that it suffers from size distortions, we report size-adjusted power statistics for this approach.

We first consider shifts in the expected first moment of the histograms. Under H_0, all histograms have the same expected value. We then shift the expected value of the histograms for one group. We consider the following shifts: H_{1a}: $0.05, H_{1b}$: $0.1, H_{1c}$: $0.2, H_{1d}$: $0.3, H_{1e}$: $0.4, H_{1f}$: $0.5, H_{1g}$: 0.75 and H_{1h}: 1.00. Next, we change the population standard deviation of the histograms for one group by multiplying the standard deviation under H_0 by a factor unequal to one. We consider the following factors: H_{2a}: $1.05, H_{2b}$: $1.1, H_{2c}$: $1.2, H_{2d}$: $1.3, H_{2e}$: $1.4, H_{2f}$: $1.5, H_{2g}$: $1.75, H_{2h}$: $2.00, H_{2i}$: 2.50 and H_{2j}: 3.00. Next, we change the group sizes under the assumption of a moderate mean shift of 0.05 (i.e., under H_{1a}). Finally, we consider changes in the within-group heterogeneity by adjusting the covariance matrix $\mathbf{\Sigma}$ (again assuming a mean shift of 0.05 as in H_{1a}). In particular, we consider a range of settings for $c\mathbf{\Sigma}$, where c assumes the following values in the different alternative scenarios: H_{3a}: $0.5, H_{3b}$: $0.75, H_{3c}$: $1.0, H_{3d}$: $1.25, H_{3e}$: $1.5, H_{3f}$: $1.75, H_{3g}$: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, H_{3j}: 10.0.

The plot in the upper-left of Figure 1 shows the rejection frequencies for the set of alternatives for which we shift the mean expectations of one group. Evidently, the performances of the three approaches are very different. While the Hotelling test for compositional data rejects the null hypothesis in about 70% of cases already for a small shift of 0.05, the KLIC-based test produces much lower rejection frequencies for small deviations from the null hypothesis. It matches the performance of the Hotelling test only for very large mean shifts of 0.75 or more. The size-adjusted power of the Bonferroni approach lies somewhere in between, matching the power of the Hotelling test for mean shifts of 0.3 or more.

The upper-right plot of Figure 1 shows analogous rejection rates for the second set of simulations that analyze the test performance against alternatives that deviate from the null hypothesis due to differences in the population standard deviation of the density

Figure 1: Monte Carlo simulation results: power analysis for Gaussian setup



Notes: The plots show rejection frequencies based on the Gaussian setup for the Hotelling test (solid red lines), the multiple testing approach (dashed blue lines) and the KLIC-based test (dotted black lines) under various alternatives. The upper-left plot corresponds to alternative hypotheses with differences in means of density expectations (H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75 and H_{1h}: 1.00). In the upper-right plot corresponds to alternative hypotheses with differences in the standard deviation of density expectations (H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50 and H_{2j}: 3.00). The lower-left plot corresponds to simulations with varying group size, assuming mean differences as in H_{1a}. The lower-right plot corresponds to simulations with varying within-group heterogeneity for which we multiply our baseline calibration for the covariance matrix by a factor c, where c assumes the following values: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, H_{3j}: 10.0 (again assuming mean differences as in H_{1a}). In the simulations, all tests are used with a nominal size of 0.05.

expectations across groups. The alternatives range from moderate deviations (for which the ratio of the implied standard deviation of the two groups is 1.05) to extreme (ratio of 3). Again, the Hotelling test yields high rejection frequencies for all alternatives except H_{2a}. The Bonferroni approach here shows very similar (size adjusted) power to the compositional approach. In contrast, rejection frequencies of the KLIC-based test are low for the alternatives that do not differ much from the null hypothesis; we observe rejection frequencies above 25% only for alternatives that are based on a ratio of the standard deviations of 1.5 or larger.

Next, we assess how the group size affects the rejection frequencies. Again, we analyze this for the alternative H_{1a}, which implies a mean shift of 0.05 of mean expectations in one of the groups. The results in the lower-left plot show that increasing the sample size—even to numbers that would be common in household surveys—does not substantially increase the rejection frequency for the KLIC-based test. Rejection frequencies are much higher for the Hotelling test—for small sample sizes and increasingly so for larger sample sizes. Again, the Bonferroni approach is somewhere in-between, exhibiting very low (size adjusted) power for small to medium sample sizes but catching up with the Hotelling test for large sample sizes of $N_A = N_B = 500$.

Finally, the lower-right plot of Figure 1 shows how the level of within-group heterogeneity affects rejection frequencies. It is evident that the KLIC-based test and the Bonferroni approach have no power against H_{1a} independently of the level of within-group heterogeneity. For the Hotelling test, we observe that—not surprisingly—it has good power against a small difference in the expected value of the histograms implied by H_{1a} when the heterogeneity within groups is small, but less so when it is high. Rejection frequencies decline considerably from almost 100% to around 10% over the scenarios considered in our simulations. Still, for any level of within-group heterogeneity the rejection frequencies are substantially higher than those of the KLIC-based test and the Bonferroni approach.

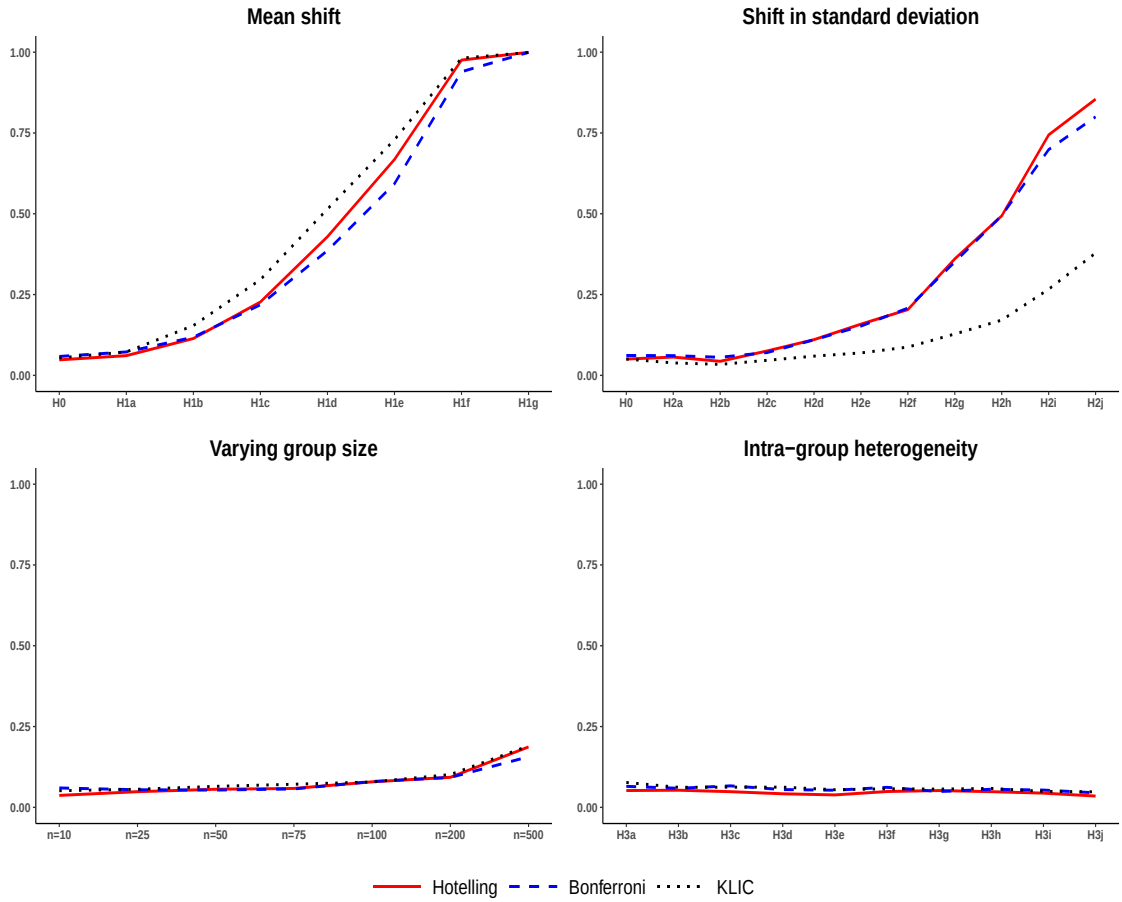
3.3 Results for Truncated-Probabilities Setup

We now turn to the Monte Carlo simulation for the alternative truncated-probabilities setup. Panel D of Table 1 shows that the tests are still appropriately sized for the alternative DGP.

Figure 2 presents the results for the power analysis. Clearly, the power of all tests is affected negatively when the data is not normally distributed. The upper-left plot shows that the rejection frequencies for the Hotelling and Bonferroni tests are much lower for small mean shifts and are now in a similar range as those for the KLIC-based test. In fact, rejection frequencies for the KLIC are slightly higher than those for the other tests for intermediate mean shifts, although the differences are relatively small. The upper-right plot shows that the power to detect shifts in the standard deviation is reduced for all three tests relative to the Gaussian setup, although the relative ranking remains the same. The lower-left plot shows that with the truncated-probabilities setup the low rejection frequencies for a small mean shift cannot be improved upon by increasing the group size to 500. Finally, the lower-right plot shows that the competitive edge of the Hotelling test in settings with low intra-group heterogeneity disappears for the truncated-probabilities setup.

In summary, the Hotelling test for compositional data and the KLIC-based test are appropriately sized while the Bonferroni approach is, as expected, undersized. For the Gaussian setup, the Hotelling test clearly outperforms the other two approaches in terms of power against all alternatives considered here. In particular, the Hotelling test detects

Figure 2: Monte Carlo simulation results: power analysis for truncated-probabilities setup



Notes: The plots show rejection frequencies based on the truncated-probabilities setupP for the Hotelling test (solid red lines), the multiple testing approach (dashed blue lines) and the KLIC-based test (dotted black lines) under various alternatives. The upper-left plot corresponds to alternative hypotheses with differences in means of density expectations (H_{1a}: 0.05, H_{1b}: 0.1, H_{1c}: 0.2, H_{1d}: 0.3, H_{1e}: 0.4, H_{1f}: 0.5, H_{1g}: 0.75 and H_{1h}: 1.00). In the upper-right plot corresponds to alternative hypotheses with differences in the standard deviation of density expectations (H_{2a}: 1.05, H_{2b}: 1.1, H_{2c}: 1.2, H_{2d}: 1.3, H_{2e}: 1.4, H_{2f}: 1.5, H_{2g}: 1.75, H_{2h}: 2.00, H_{2i}: 2.50 and H_{2j}: 3.00). The lower-left plot corresponds to simulations with varying group size, assuming mean differences as in H_{1a}. The lower-right plot corresponds to simulations with varying within-group heterogeneity for which we multiply our baseline calibration for the covariance matrix by a factor c, where c assumes the following values: H_{3a}: 0.5, H_{3b}: 0.75, H_{3c}: 1.0, H_{3d}: 1.25, H_{3e}: 1.5, H_{3f}: 1.75, H_{3g}: 2.0, H_{3h}: 2.5, H_{3i}: 3.0, H_{3j}: 10.0 (again assuming mean differences as in H_{1a}). In the simulations, all tests are used with a nominal size of 0.05.

significant group differences for much smaller differences in the expected value of density expectations relative to the KLIC-based approach and the Bonferroni approach, especially when group sizes and/or within-group heterogeneity are small. It also has much higher power compared to the KLIC-based approach against differences in the standard deviation of the density expectations across groups. While the rejection frequencies of the Hotelling test are considerably lower for the case of the truncated-probabilities setup, the power of the other tests are not substantially higher for any of the considered alternatives in this

setup. The results from the Monte Carlo simulations thus complement the conceptual advantages of our approach as described in Section 2.

4 Empirical Applications

In this section, we consider a range of applications for which the discussed tests can be useful. All applications deal with aspects of expectation heterogeneity that have recently been discussed in the literature. The data in the applications are either from the Survey of Professional Forecasters (SPF) conducted by the European Central Bank (ECB) for the euro area (as described in Bowles et al., 2007) or from the Survey of Consumer Expectations (SCE) conducted by the Federal Reserve Bank of New York among U.S. households (see Armantier et al., 2015). For all applications, we exclude those histograms from the sample that assign 100% probability to a single bin. Moreover, since the probability mass assigned to the exterior bin is zero in many cases, we choose the sixth bin as the reference category throughout.

4.1 Response to Rising Inflation Rates

After several years of low inflation rates, inflation in the euro area began to increase midway through 2021. In this section, we analyze whether the steady rise in inflation changed the inflation density expectations of SPF participants at different forecast horizons, i.e., we test for differences in the aggregate densities across time, thereby shedding light on their temporal stability. Intuitively, one might expect to observe an immediate adjustment of short-term expectations while long-term expectations would not change if they were firmly anchored.

The SPF asks experts from financial and non-financial institutions to report predictions for several macroeconomic outcomes in the euro area, including one- and five-year-ahead inflation expectations. It has been conducted by the ECB since 1999 at a quarterly frequency. We focus on the density expectations which are elicited by asking panelists to state probabilities for a range of bins (e.g., the likelihood that inflation turns out to be between 1.5% and 1.9%). An attractive feature of the SPF is that the bins have a constant width with the exception of the exterior bins, which are half-open. In particular, the intervals as defined in the SPF questionnaire have a width of 0.4 percentage point with a gap of 0.1 percentage point between bins.[11]

To analyze whether the increase in inflation had an effect on inflation expectations, we use 2021Q1 as a reference wave and compare the density expectations from each subsequent wave (up until 2022Q2) to those reported in this reference wave. For all of these waves, the SPF bin definitions for inflation have remained identical and comprise

[11] One exception is a recent change in the survey design for expectations of GDP growth. In 2020Q2, bins with a width of two percentage points have been introduced.

$K = 12$ bins. To provide some descriptive evidence, we compute the first four moments of the one- and five-year-ahead aggregate density expectations from each wave, i.e., $\bar{\boldsymbol{p}}_{t,h}$. Next, we formally test for each wave and horizon whether expectations have changed relative to the 2021Q1 wave by using the three testing approaches described in Section 2.

Panels A and C of Table 2 present the number of panelists in each wave along with the estimated moments (based on the 'mass-at-midpoint' approach) of the aggregate probability distributions for the one- and five-year-ahead inflation expectations. Panels B and D show the test statistics along with the corresponding p-values for the three tests. Note that to make results comparable we report the minimum of one and twelve times the smallest of the twelve p-value in case of the Bonferroni approach.[12]

For the one-year-ahead expectations, Panel A shows an upward shift in the histogram mean over time as well as an increase in the standard deviation for the 2022Q1 and 2022Q2 waves. As shown in Section 3, the Hotelling test should be able to detect such differences. For skewness and kurtosis, we do not observe any clear patterns. The upward shift in the mean is visible for virtually all of the underlying individual density expectations and suggests that SPF participants quickly reacted to the rising inflation rates. However, these changes are relatively small in magnitude from one period to the next relative to the heterogeneity of individual expectations. As a result, all tests do not reject the null hypothesis when comparing the 2021Q1 and 2021Q2 waves. In contrast, the tests detect significant differences in short-term density forecasts when comparing subsequent waves to the reference period.

We also observe an increase in the histogram mean of the five-year-ahead expectations, although the changes from one period to the next are clearly smaller than those for the one-year-ahead expectations. This likely reflects the fact that such expectations are more anchored and less impacted by price shocks that are perceived to have a large transitory component. In addition, we do not observe an increase in the standard deviation towards the end of the sample. As a result, the tests reject the null hypothesis only for the 2022Q2 wave relative to 2021Q1.

We conclude that the SPF participants quickly adapted their short-term inflation expectations in response to the recent inflation shock. Long-term expectations reacted less strongly and more gradually but also increased significantly relative to 2021Q1. This suggests that there was a deterioration in the degree to which medium term expectations were anchored.[13] This finding is consistent with the results in Binder et al. (2022) for U.S. forecasters. The low p-values for the KLIC-based test are in line with our Monte Carlo simulation results for the truncated-probabilities setup. As seen in Figure 2, the

[12]The displayed test statistics are the largest of the twelve bin-specific t-statistics.

[13]The ECB revised its inflation target in 2021. Comparing the five-year-ahead inflation expectations from the 2021Q3 wave (elicited just before the publication of the revised ECB strategy) with those from the 2021Q4 wave, we do not reject the null hypothesis. A possible explanation for this finding is that professional forecasters adjusted their density expectations to the new inflation target well in advance to the official announcement by the ECB.

Table 2: Differences in inflation expectations over time

	2021Q1	2021Q2	2021Q3	2021Q4	2022Q1	2022Q2
Panel A: Histogram moments (one-year-ahead expectations)						
Group size	39	39	34	38	38	31
Mean	1.24	1.36	1.51	1.71	1.94	2.73
Standard deviation	0.78	0.78	0.74	0.79	0.98	1.06
Skewness	0.12	0.16	−0.19	0.11	0.29	−0.44
Kurtosis	3.85	4.38	4.40	3.90	3.15	2.82
Panel B: Distance measures (one-year-ahead expectations)						
Hotelling	–	0.525	1.895	2.813	3.083	12.300
	–	(0.880)	(0.058)	(0.005)	(0.002)	(0.000)
Bonferroni	–	−1.685	2.536	3.759	3.599	−7.166
	–	(1.000)	(0.161)	(0.004)	(0.007)	(0.000)
KLIC	–	0.016	0.094	0.199	0.303	1.051
	–	(0.562)	(0.002)	(0.000)	(0.000)	(0.000)
Panel C: Histogram moments (five-year-ahead expectations)						
Group size	40	43	35	37	42	38
Mean	1.60	1.62	1.75	1.86	1.87	2.02
Standard deviation	0.83	0.80	0.88	0.89	0.89	0.85
Skewness	0.03	0.01	0.09	0.28	0.17	−0.05
Kurtosis	3.95	4.38	3.92	4.06	3.92	3.52
Panel D: Distance measures (five-year-ahead expectations)						
Hotelling	–	0.247	0.504	1.157	1.625	4.701
	–	(0.993)	(0.894)	(0.334)	(0.111)	(0.000)
Bonferroni	–	1.063	0.985	1.958	−1.946	−3.505
	–	(1.000)	(1.000)	(0.647)	(0.662)	(0.009)
KLIC	–	0.004	0.017	0.053	0.060	0.140
	–	(0.864)	(0.380)	(0.080)	(0.058)	(0.000)

Notes: Panel A presents moments (based on the 'mass-at-midpoint' approach) for the aggregate one-year-ahead inflation expectations from the 2021Q1 to 2022Q2 waves of the SPF. Panel B shows the test statistics relative to the 2021Q1 wave along with corresponding p-values in parentheses. For the multiple testing approach, we report the largest test statistic across the twelve distinct bins and twelve times the corresponding p-value. Panels C and D present the results for the five-year-ahead inflation expectations.

power of the KLIC-based test can exceed that of the Hotelling test for moderate mean shifts in non-normal settings.

4.2 Different Types of Forecasters

A number of studies have recently observed that one can distinguish two types of survey-based forecasts based on the rounding behavior of the panelists (Binder, 2017; Clements, 2021; Glas and Hartmann, 2022; Reiche and Meyler, 2022). For density forecasts, Glas and Hartmann (2022) show that one type of panelists ("rounders") state interval probabilities that are multiples of five or ten and tend to assign positive probabilities to only a relatively small subset of the surveyed bins while another type of panelists ("non-rounders") reports probabilities that do not share a common divisor and tend to consider a larger number of bins, often reporting probabilities with higher precision (i.e., to at least one decimal point) for most, or indeed, all of the surveyed bins.

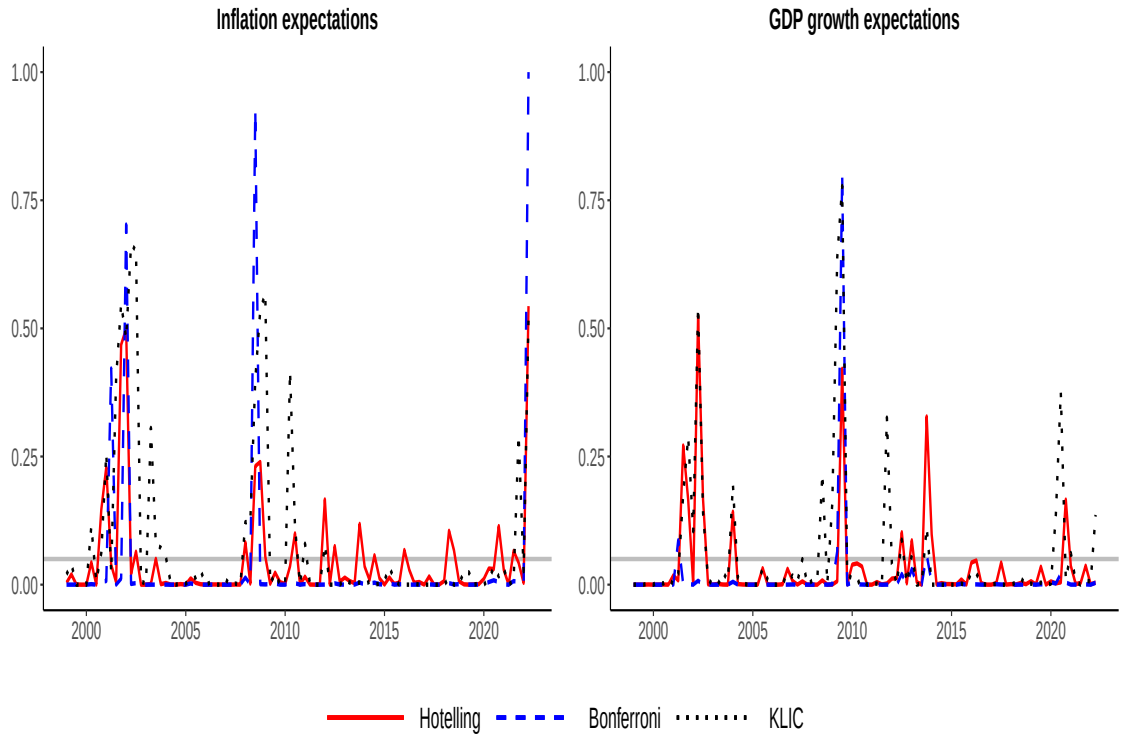
We test whether the reported expectations from rounders and non-rounders are indeed sampled from different populations. We define rounders as those panelists who report histograms containing probabilities of which more than half are multiples of five. Glas and Hartmann (2022) document that these two groups of forecasters differ in terms of the variances of their probabilistic forecasts. As shown in Section 3, the Hotelling test is able to detect such differences in second moments even if they are small.

For this analysis, we look at the SPF density forecasts for inflation and real GDP growth and focus on the one-year-ahead density forecasts from all available survey waves. Overall, the sample includes information from 108 panelists and $T = 94$ survey rounds, covering the period 1999Q1–2022Q2. The panel is unbalanced due to frequent dropouts and entries of new participants. The black lines in Figure A.2 in the appendix show that, on average, 45–55 panelists report density forecasts for inflation and GDP growth each quarter (with declining trend).

Figure 3 shows the p-value for all tests and each survey wave; to ensure comparability of results we again show the smallest p-value multiplied by the (time-varying) number of bins for the Bonferroni approach. The null hypothesis of equal expectations is rejected for most survey waves. We obtain the lowest rejection frequency of 75% of the survey waves for the KLIC-based test in the case of inflation expectations. In line with the evidence in Glas and Hartmann (2022), the rejections are driven primarily by the lower variances of the histograms reported by the rounders rather than differences in mean expectations (see Figure A.3 in the appendix). We observe a few large p-values in the first half of the sample. The red lines in Figure A.2 show that only a small number of non-rounders are included in these particular waves, leading to very low power of the tests. Another spike is visible in 2009Q1. This can be explained by a pile-up of probabilities in the lowest bin due to the Great Recession, which partially masks the differences in the second moments between both groups (see Figure A.3).

Referring back to the discussion of the Gaussian setup versus the truncated-probabilities setup in Section 3, the histograms reported by the non-rounders are more in line with the normality assumption than those of the rounders. As such, the SPF data can be thought of as a mixture of 'well-behaved' and relatively coarse histograms.

Figure 3: p-values for heterogeneity tests (SPF): rounders vs. non-rounders



Notes: The plot shows the p-values from the Hotelling test (solid red lines), the multiple testing approach (dashed blue lines) and the KLIC-based test (dotted black lines) for the analysis of differences in inflation expectations (left) and GDP growth expectations (right) between rounders and non-rounders. For the multiple testing approach, we report (the minimum of one and) the smallest p-value multiplied by the number of bins to make it comparable. The sample period is 1999Q1–2022Q2.

With that in mind, we briefly return to the previous application and now focus on the subsample of non-rounders only. Broadly speaking, we find that the non-rounders adjust their short-term expectations more slowly than the rounders (Table A.1). In particular, the standard deviation of the aggregate histogram is essentially constant. As before, the tests detect significant differences for the short-term expectations before such differences are evident for the long-term expectations. Interestingly, we observe that the KLIC-based test does not produce smaller p-values than the Hotelling test, unlike in Table 2. It is likely that this is because the histograms of the non-rounders are more in line with the normality assumption.

4.3 Gender Differences in Expectations

A drawback of the SPF data is the relatively small cross-section. Figure 1 shows that a small group size negatively affects the power of all tests including the Hotellig test. Therefore, we now turn to data on expectations of private households with a larger number of individual survey responses. A potential disadvantage is that within-group heterogeneity

may be larger for households than for experts. Moreover, it is likely that the data contains more frequent violations of the normality assumption than the SPF data.

First, we compare inflation expectations of men and women. Among others, D'Acunto et al. (2021) show that, on average, women expect higher inflation rates than men due to higher exposure to price changes for certain household items during grocery shopping. One may expect to find a similar divergence in the density forecasts reported by both genders. For U.S. households, Armantier et al. (2021) show that female survey participants assign more probability mass to both exterior bins, resulting in higher uncertainty. Similarly, using survey data from German households, Conrad et al. (2022) find that the inflation histograms of women tend to be more dispersed than those of men.

Here, we tackle the question of gender differences in inflation expectations using the full density forecasts reported in the SCE, which is a monthly and representative survey among U.S. households that asks questions about socioeconomic characteristics and macroeconomic expectations. The SCE has been conducted since June 2013. Each wave includes roughly $1,300$ households with a balanced relation between male and female household heads.[14]

We use density forecasts for the consumer price inflation rate over the next twelve months (*Q9* in the survey questionnaire).[15] Our sample includes responses from 18,066 households across $T = 103$ survey waves, covering the period from June 2013 to December 2021. The density forecasts in the SCE are conceptually similar to those in the SPF. The specific design differs, however, in the sense that the width of the intervals is larger and varies across bins.[16]

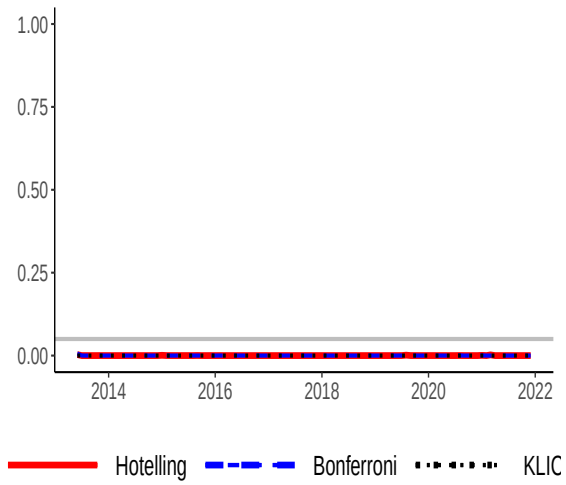
Figure 4 clearly shows that all tests reject the null hypothesis of no differences in the density forecasts of men and women for each survey wave. This is not surprising given the large differences in expectations across genders. The left plot in Figure A.5 in the appendix shows the aggregate histograms (pooled across households and survey waves) for men and women. In line with the studies discussed above, we observe that women assign more probabiliy mass to the exterior bins and have higher mean expectations and variances. The latter can be seen more clearly in Figure A.6 which presents the time series for the first four moments of the aggregate histograms of men and women. The figure also shows that the aggregate distribution of men has lower skewness and higher kurtosis. Given that the histograms reported by men and women strongly differ in terms of all four moments (and that a large sample size is available), it is not surprising that all tests reject the null hypothesis despite potentially large within-group heterogeneity.

[14] Figure A.4 in the appendix shows the sample size and the number of women per survey round.

[15] We obtain nearly identical results if we focus on long-term inflation expectations (*Q9c*). Results are available upon request by the authors.

[16] In particular, households are asked to assign probabilities to the following outcomes for future inflation (in percent): $(-\infty, -12], (-12, -8], (-8, -4], (-4, -2], (-2, 0], (0, 2], (2, 4], (4, 8], (8, 12], (12, +\infty)$

Figure 4: p-values for heterogeneity tests (SCE): gender differences in inflation expectations



Notes: The plot shows the p-values from the Hotelling test (solid red line), the multiple testing approach (dashed blue line) and the KLIC-based test (dotted black line) for the analysis of gender differences in inflation expectations. For the multiple testing approach, we report (the minimum of one and) the smallest p-value multiplied by the number of bins to make it comparable. The sample period is June 2013 to December 2021.

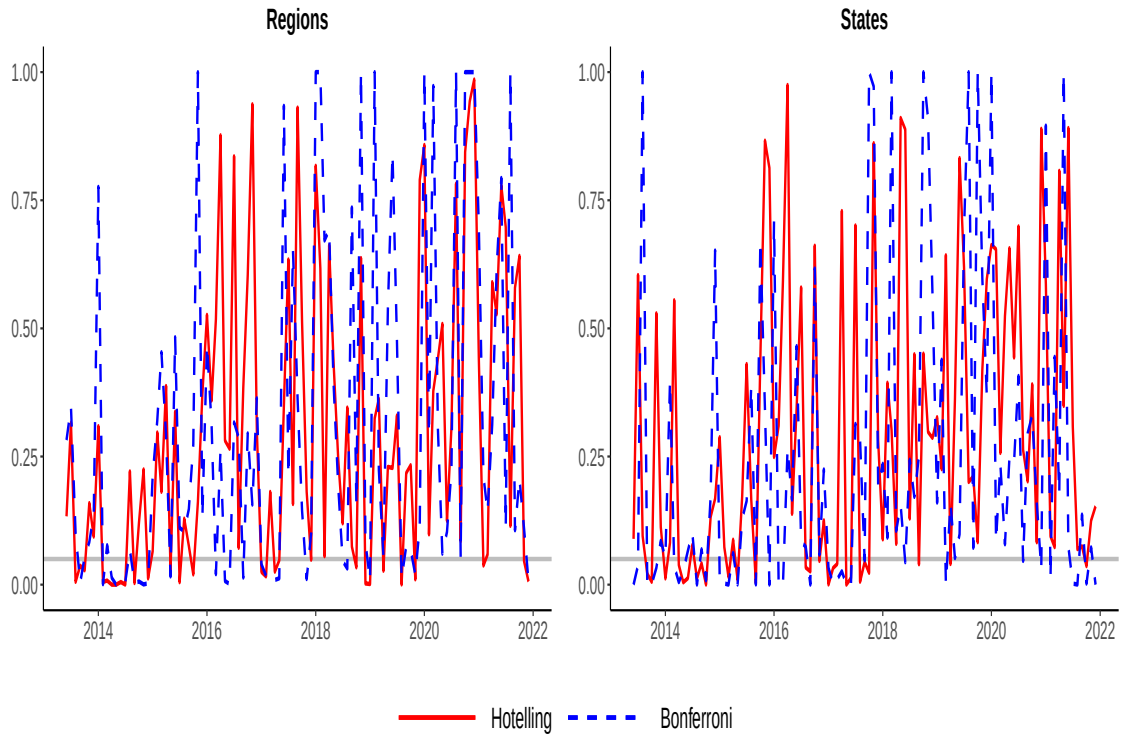
4.4 Regional Differences in Expectations for National House Prices

In this final application, we demonstrate that the KLIC-based approach ceases to be feasible in setups where expectations of more than two groups need to be compared. The choice is motivated by the finding of Kuchler and Zafar (2019) that differences in local house price dynamics tend to translate into dispersed forecasts of future nationwide house prices changes, a finding that appears inconsistent with full information rational expectations.

We test for differences of house price expectations across households from the 50 U.S. states and Washington D.C.—or, alternatively, from four broader regions ('West', 'Midwest', 'Northeast' and 'South'; see Figure A.4 for the number of households from each region). In particular, we test the hypothesis that the density expectations from all states (regions) are from the same population in an ANOVA framework. The data are again from the SCE and we focus on expectations for the change of average house prices nationwide (*C1*). The right plot in Figure A.5 in the appendix shows the aggregate histograms for the different regions. Figure A.7 shows the time series of the moments for the aggregate histograms. The figures do not reveal clear evidence of differences with one exception: the mean of the aggregate histogram for the 'West'-region is noticeably higher than those for the other regions in the first couple of survey waves.

The plots in Figure 5 present the results based on regions (left plot) and states (right plot). Evidently, there is more time variation in the p-values than for the gender differ-

Figure 5: p-values for heterogeneity tests (SCE): local differences in house price expectations



Notes: The plot shows the p-values from the Hotelling test (solid red line) and the multiple testing approach (dashed blue line) for the analysis of differences in house price expectations across regions (left plot) or states (right plot). For the multiple testing approach, we report (the minimum of one and) the smallest p-value multiplied by the number of bins to make it comparable. The sample period is June 2013 to December 2021.

ences in inflation expectations. For the Hotelling test we reject the null hypothesis of no differences in the house price expectations across regions (states) for 28 (26) of the 103 survey waves. The evidence for the multiple testing approach is similar. The periods with significantly different house price expectations are distributed without any obvious systematic pattern, although particularly for the state-level analysis we observe more differences in the beginning of the sample between 2013 and 2015. This is likely due to the higher mean expectations for the 'West'-region during this period.

5 Conclusion

We propose a new test for heterogeneity and differences in density expectations. This test builds on the insight that probabilistic survey forecasts are compositional data. For normally distributed data, our Monte Carlo simulations show the superior performance of this test relative to a more traditional bootstrap-based approach using the KLIC as a distance measure between two densities and an approach which involves multiple testing for differences of individual parts of the density. The novel test has high power especially

when intra-group heterogeneity is relatively low. For settings that mimic more closely the coarse density expectations observed in many surveys all tests have very similar power. However, the novel test is always much faster compared to the KLIC-based test because it does not rely on simulations and allows for comparisons across more than two groups.

In four applications we analyze survey-based density expectations of professional forecasters and households. First, we show that the short-term inflation expectations of experts adjusted rapidly in response to rising inflation rates in the euro area. Long-term expectations were not fully anchored but changed less strongly and more gradually. Second, we find that for most periods short-run inflation and growth expectations significantly differ between forecasters that round their probability statements and those that do not. Third, we find very strong evidence against the hypothesis that inflation expectations of men and women are equal, confirming earlier results in the literature based on point forecasts. Finally, consistent with a role for local developments and information sets influencing subjective expectations data for aggregate outcomes, we show that for a substantial fraction of periods in our sample households from different regions report significantly different density expectations for the future change of nationwide house prices.

Our findings show that it is beneficial to treat survey-based density expectations as compositional data. This might be relevant also in other contexts where such survey data is used. Our results could be extended by using the panel structure of most expectation surveys. So far, we have analyzed each survey wave as separate data samples but one could also jointly analyze the full sample of expectation data. For instance, by adopting a dynamic model for the compositional expectation data.

References

Aitchison, J., 1982. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. Series B (Methodological) 44 (2), 139–177.

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Springer Netherlands.

Andrade, P., Coibion, O., Gautier, E., Gorodnichenko, Y., 2022. No Firm Is an Island? How Industry Conditions Shape Firms' Expectations. Journal of Monetary Economics 125, 40–56.

Armantier, O., Bruine de Bruin, W., Topa, G., van der Klaauw, W., Zafar, B., 2015. Inflation Expectations and Behavior: Do Survey Respondents Act on Their Beliefs? International Economic Review 56 (2), 505–536.

Armantier, O., Kosar, G., Pomerantz, R., Skandalis, D., Smith, K., Topa, G., van der Klaauw, W., 2021. How Economic Crises Affect Inflation Beliefs: Evidence from the Covid-19 Pandemic. Journal of Economic Behavior & Organization 189, 443–469.

Binder, C., Janson, W., Verbrugge, R., 2022. Out of Bounds: Do SPF Respondents Have Anchored Inflation Expectations? Journal of Money, Credit and Banking forthcoming.

Binder, C. C., 2017. Measuring Uncertainty Based on Rounding: New Method and Application to Inflation Expectations. Journal of Monetary Economics 90, 1–12.

Bowles, C., Friz, R., Genre, V., Kenny, G., Meyler, A., Rautanen, T., Apr 2007. The ECB Survey of Professional Forecasters (SPF) - A Review After Eight Years' Experience. Occasional Paper Series 59, European Central Bank.

Braga, L., Feingenbaun, D., 2020. Assessing Global Covid-19 Cases Data through Compositional Data Analysis (CoDa). medRxiv.

Buccianti, A., 2018. Water Chemistry: Are New Challenges Possible from CoDA (Compositional Data Analysis) Point of View? In: Handbook of Mathematical Geosciences. Springer, Cham, pp. 299–311.

Clements, M. P., 2021. Rounding Behaviour of Professional Macro-Forecasters. International Journal of Forecasting 37 (4), 1614–1631.

Clements, M. P., 2022. Individual Forecaster Perceptions of the Persistence of Shocks to GDP. Journal of Applied Econometrics 36 (3), 640–656.

Coibion, O., Gorodnichenko, Y., Kumar, S., September 2018. How Do Firms Form Their Expectations? New Survey Evidence. American Economic Review 108 (9), 2671–2713.

Coibion, O., Gorodnichenko, Y., Ropele, T., 2020. Inflation Expectations and Firm Decisions: New Causal Evidence. The Quarterly Journal of Economics 135 (1), 165–219.

Conrad, C., Enders, Z., Glas, A., 2022. The Role of Information and Experience for Households' Inflation Expectations. European Economic Review 143, 104015.

D'Acunto, F., Malmendier, U., Ospina, J., Weber, M., 2021. Exposure to Grocery Prices and Inflation Expectations. Journal of Political Economy 129 (5), 1615–1639.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis. Cham: Springer.

Fry, J. M., Fry, T. R., McLaren, K. R., 1996. The Stochastic Specification of Demand Share Equations: Restricting Budget Shares to the Unit Simplex. Journal of Econometrics 73 (2), 377–385.

Glas, A., Hartmann, M., 2022. Uncertainty Measures from Partially Rounded Probabilistic Forecast Surveys. Quantitative Economics 13 (3), 979–1022.

Gupta, N., Rasmussen, C. L., Holtermann, A., Mathiassen, S. E., 2020. Time-Based Data in Occupational Studies: the Whys, the Hows, and Some Remaining Challenges in Compositional Data Analysis (CoDA). Annals of Work Exposures and Health 64 (8), 778–785.

Kitano, N., Kai, Y., Jindo, T., Tsunoda, K., Arao, T., 2020. Compositional Data Analysis of 24-Hour Movement Behaviors and Mental Health in Workers. Preventive medicine reports 20, 101213.

Kuchler, T., Zafar, B., 2019. Personal Experiences and Expectations about Aggregate Outcomes. The Journal of Finance 74 (5), 2491–2542.

Kumar, S., Afrouzi, H., Coibion, O., Gorodnichenko, Y., 2015. Inflation Targeting Does not Anchor Inflation Expectations: Evidence from Firms in New Zealand. Brookings Papers on Economic Activity 46, 151–225.

Lloyd, C., Pawlowsky-Glahn, V., Egozcue, J., 2012. Compositional Data Analysis in Population Studies. Annals of the Association of American Geographers 102 (6), 1251–1266.

Malmendier, U., Nagel, S., 2011. Depression Babies: Do Macroeconomic Experiences Affect Risk Taking? The Quarterly Journal of Economics 126 (1), 373–416.

Malmendier, U., Nagel, S., 2016. Learning from Inflation Experiences. The Quarterly Journal of Economics 131 (1), 53–87.

Mankiw, N. G., Reis, R., Wolfers, J., 2003. Disagreement about Inflation Expectations. NBER Macroeconomics Annual 18, 209–248.

Manski, C. F., 2018. Survey Measurement of Probabilistic Macroeconomic Expectations: Progress and Promise. NBER Macroeconomics Annual 32 (1), 411–471.

Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology 35, 253–278.

Mitchell, J., Hall, S. G., 2005. Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR Fan Charts of Inflation. Oxford Bulletin of Economics and Statistics 67, 995–1033.

Reiche, L., Meyler, A., Feb. 2022. Making Sense of Consumer Inflation Expectations: The Role of Uncertainty. Working Paper Series 2642, European Central Bank.

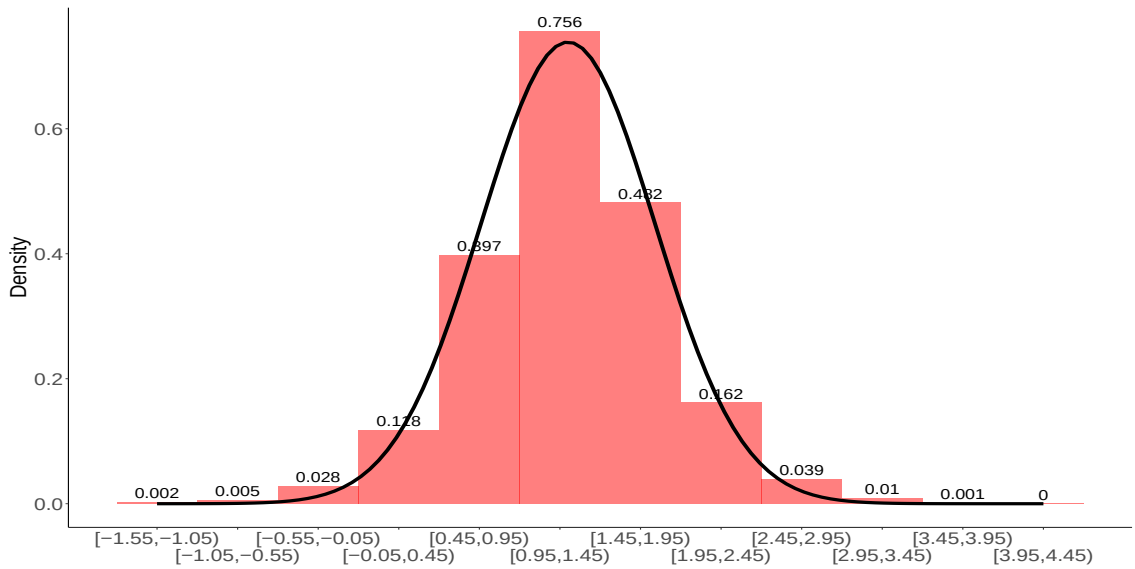
Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Dinelli, E., Ladenberger, A., Team, T., 2012. The Concept of Compositional Data Analysis in Practice. Total Major Element Concentrations in Agricultural and Grazing Land Soils of Europe. Science of the Total Environment 426, 196–210.

Rich, R., Tracy, J., 2021. A Closer Look at the Behavior of Uncertainty and Disagreement: Micro Evidence from the Euro Area. Journal of Money, Credit and Banking 53 (1), 233–253.

Weltje, G., von Eynatten, H., 2004. Quantitative Provenance Analysis of Sediments: Review and Outlook. Sedimentary Geology 171 (1-4), 1–11.

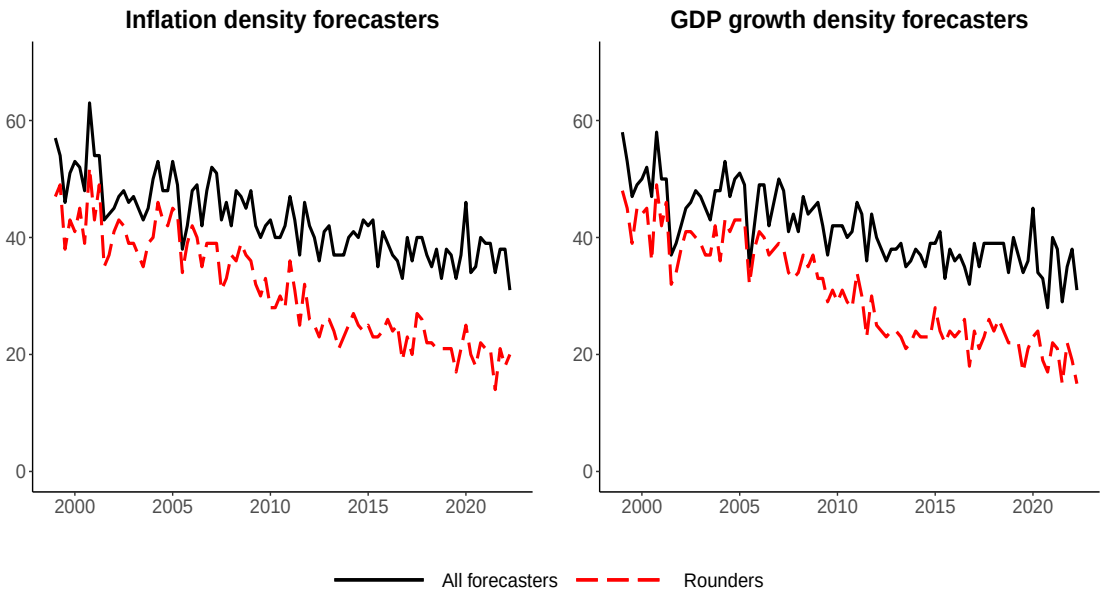
Appendix

Figure A.1: Aggregate one-year-ahead SPF inflation expectations (2020Q1 wave)



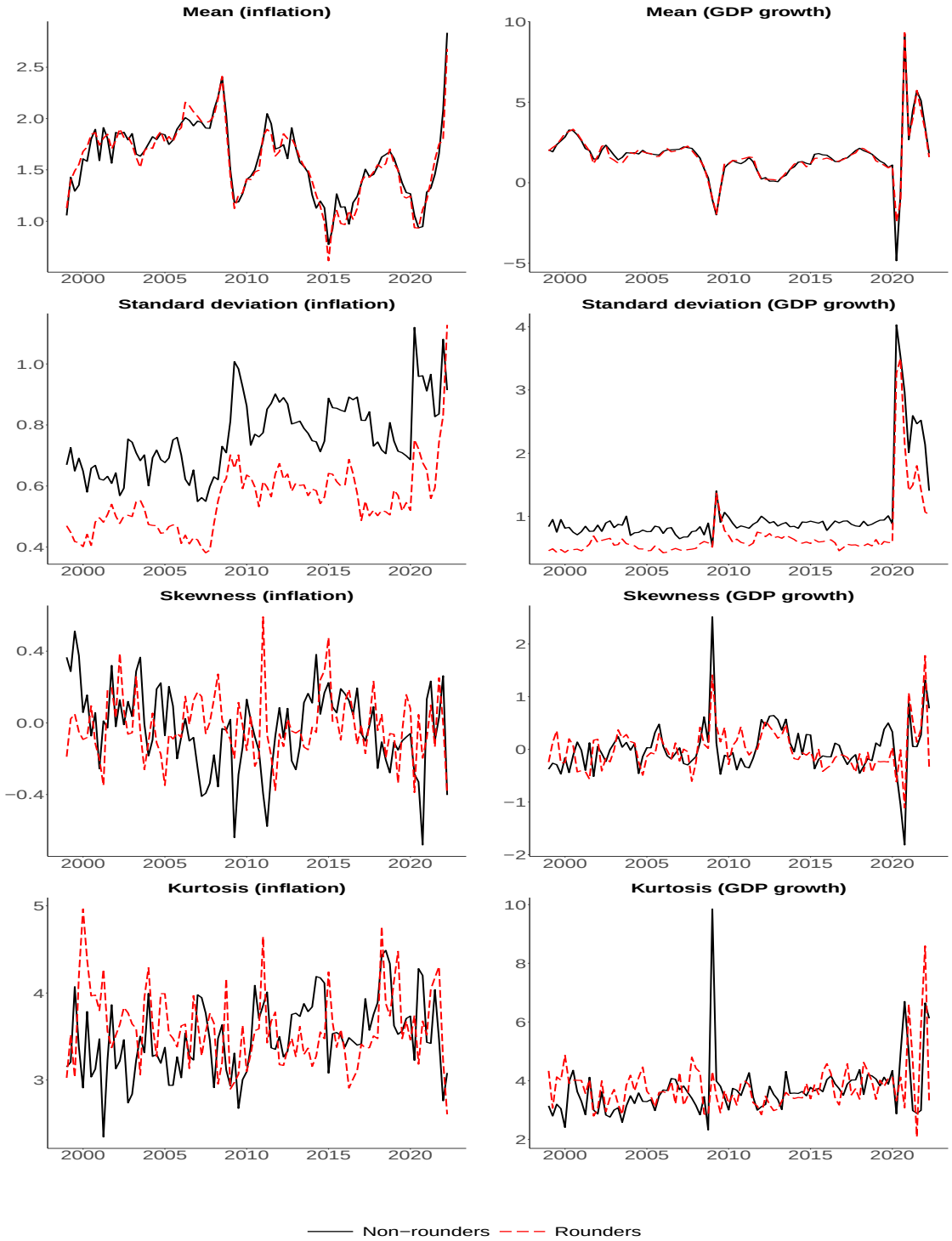
Notes: This plot shows the aggregate one-year-ahead inflation histogram (reporting densities instead of bin probabilities) from the 2020Q1 wave of the SPF. The black line shows the fitted normal distribution. The mean and standard deviation of this normal distribution are used to calibrate the baseline histogram in the Monte Carlo simulations.

Figure A.2: Number of density forecasts in the SPF data



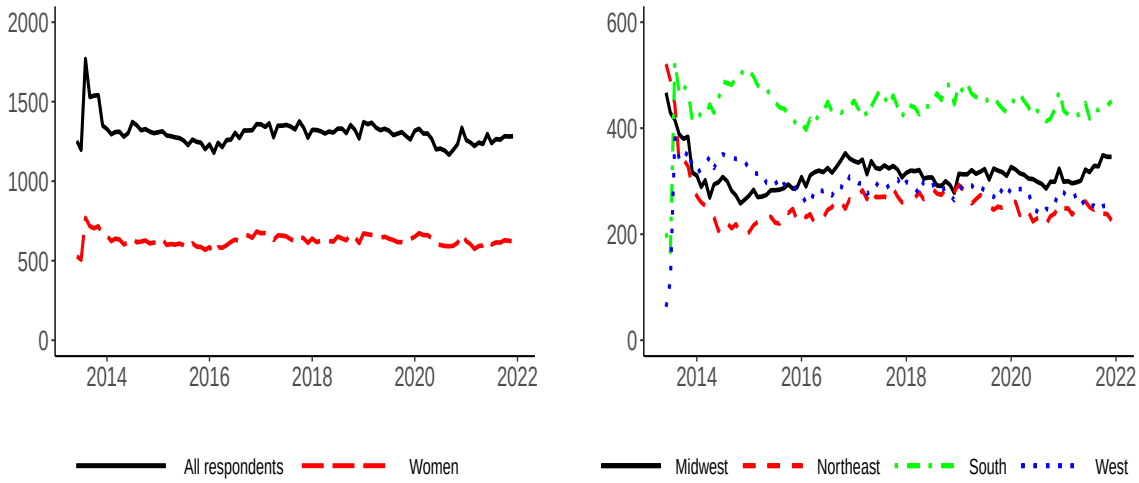
Notes: For each survey round, the plots depict the number of one-year-ahead density forecasts reported by SPF participants (black lines) as well as the number of forecasts reported by rounders (red lines). The sample period is 1999Q1–2022Q2.

Figure A.3: Histogram moments (SPF): rounders vs. non-rounders



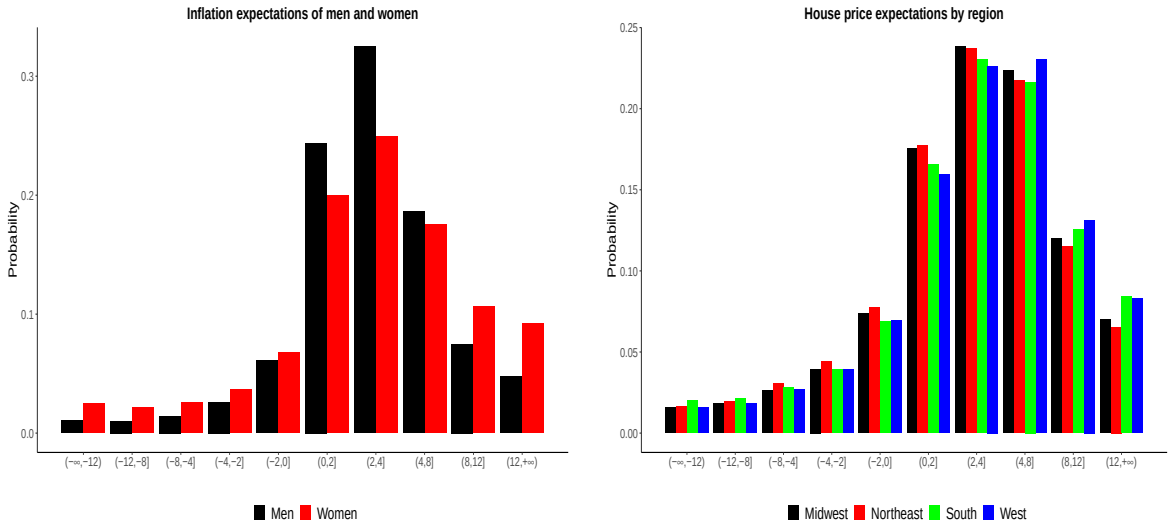
Notes: For each survey period, the plots depict the first four moments (derived under the 'mass-at-midpoint'-approach) of the aggregate SPF histograms reported by non-rounders (black lines) and rounders (red lines) for inflation (first column) or GDP growth (second column). The first row shows the means, the second row the standard deviations, the third row the skewness and the fourth row the kurtosis of the aggregate histograms. The sample period is 1999Q1–2022Q2.

Figure A.4: Number of density forecasts in the SCE data



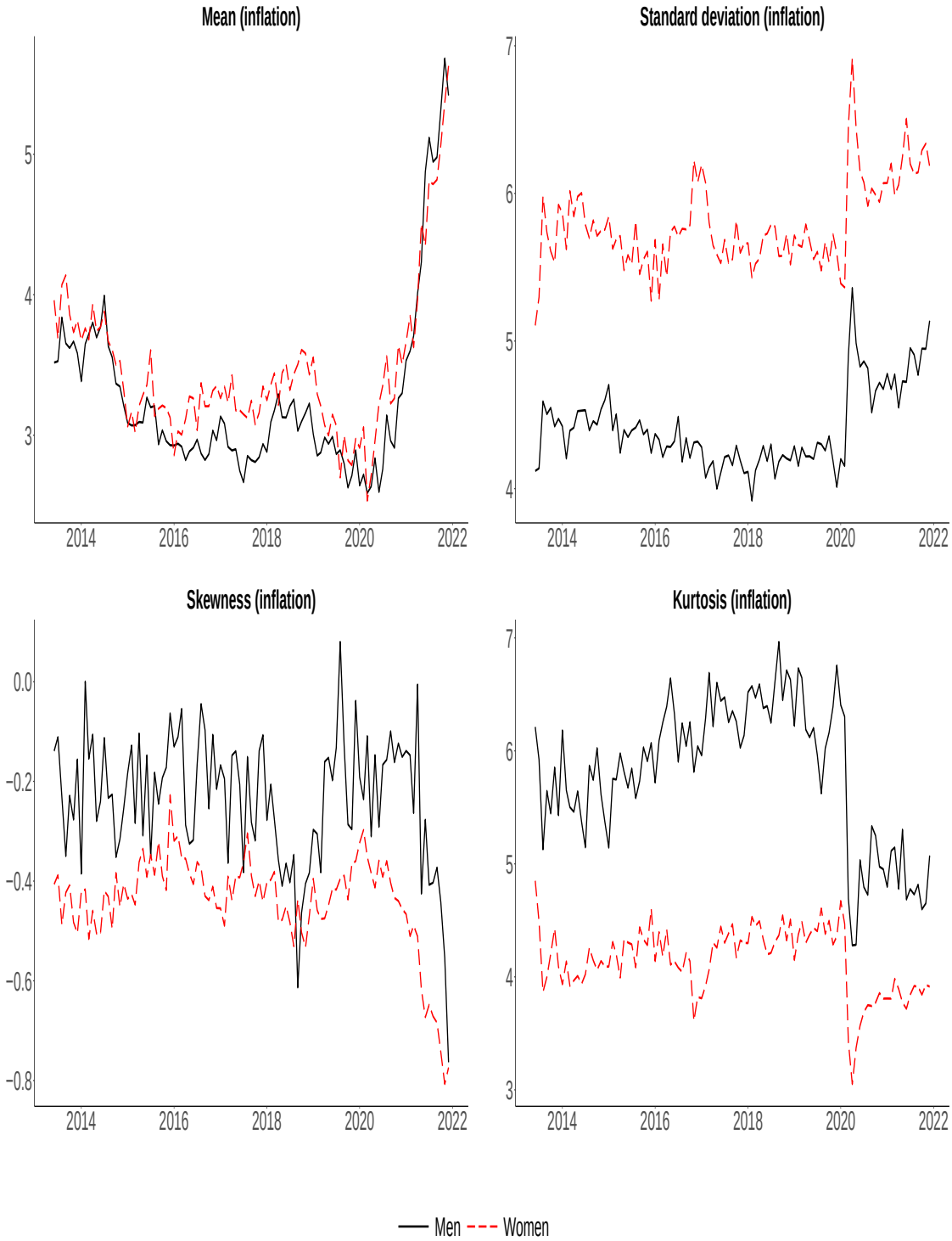
Notes: For each survey round, the left plot depicts the total number of SCE respondents (black line) as well as the number of women (red line). The right plots shows the number of households from each of the four U.S. regions. The sample period is June 2013 to December 2021.

Figure A.5: Aggregate one-year-ahead SCE inflation and national house price expectations



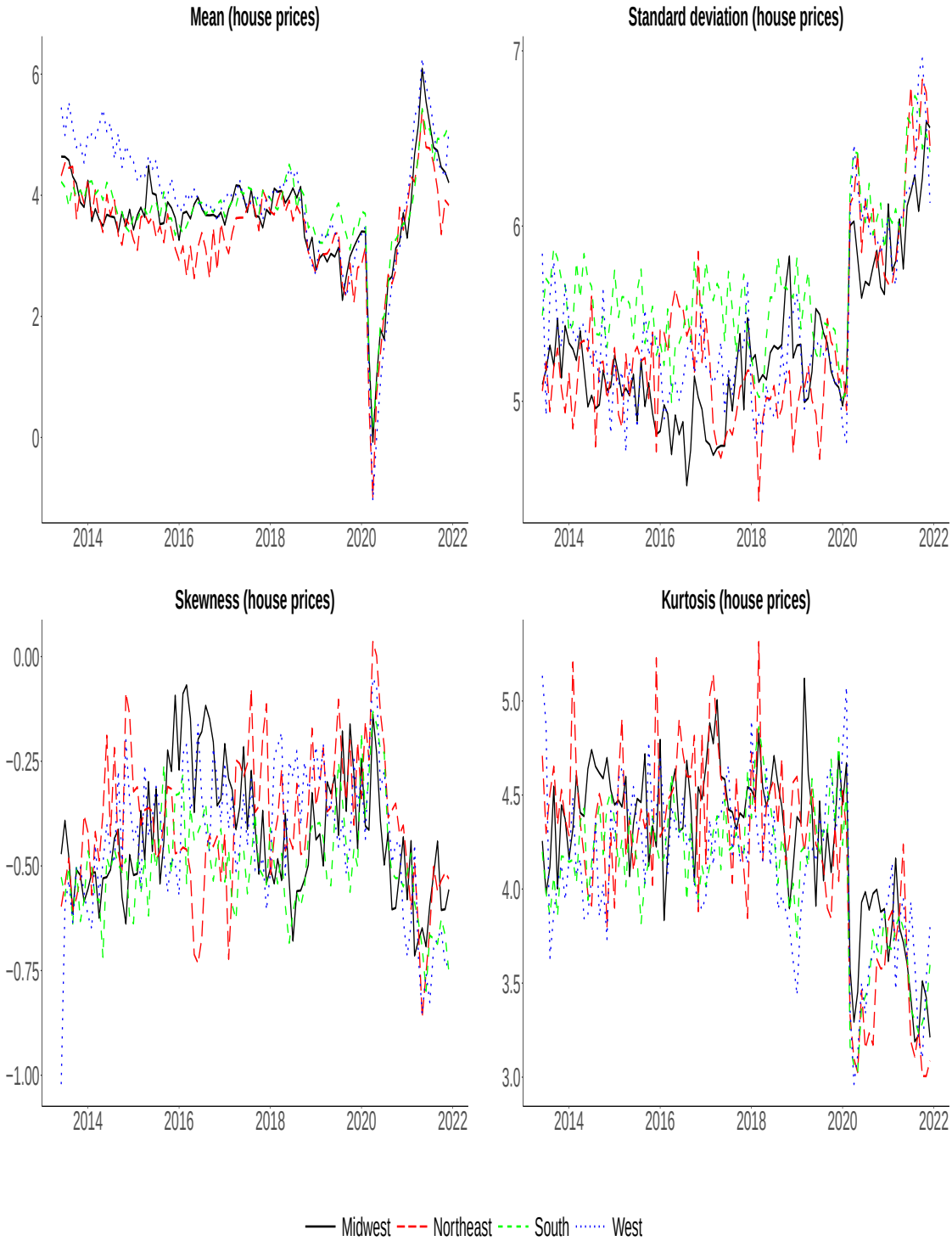
Notes: This plot shows the aggregate one-year-ahead inflation (left) and house price (right) histograms (pooled over households and survey waves) from the SCE based on gender (left) and regions (right).

Figure A.6: Histogram moments (SCE): women vs. men



Notes: For each survey period, the plots depict the first four moments (derived under the 'mass-at-midpoint'-approach) of the aggregate SCE histograms reported by men (black lines) and women (red lines) for one-year-ahead inflation. The upper left plot shows the means, the upper right plot the standard deviations, the lower left plot the skewness and the lower right plot the kurtosis of the aggregate histograms. The sample period is June 2013 to December 2021.

Figure A.7: Histogram moments (SCE): regional differences



Notes: For each survey period, the plots depict the first four moments (derived under the 'mass-at-midpoint'-approach) of the aggregate SCE histograms reported by households from Midwest (black lines), Northeast (red lines), South (green lines) or West (blue lines) for one-year-ahead nationwide house price. The upper left plot shows the means, the upper right plot the standard deviations, the lower left plot the skewness and the lower right plot the kurtosis of the aggregate histograms. The sample period is June 2013 to December 2021.

Table A.1: Differences in inflation expectations over time (nonrounders)

	2021Q1	2021Q2	2021Q3	2021Q4	2022Q1	2022Q2
Panel A: Histogram moments (one-year-ahead expectations)						
Group size	18	18	20	17	20	11
Mean	1.28	1.32	1.46	1.66	2.08	2.84
Standard deviation	0.91	0.97	0.83	0.84	1.08	0.91
Skewness	0.13	0.23	−0.10	0.03	0.26	−0.40
Kurtosis	3.43	3.42	4.04	3.47	2.76	3.08
Panel B: Distance measures (one-year-ahead expectations)						
Hotelling	–	0.503	1.375	2.105	3.260	5.705
	–	(0.882)	(0.242)	(0.064)	(0.006)	(0.001)
Bonferroni	–	−1.029	−2.370	2.021	−2.874	−4.734
	–	(1.000)	(0.279)	(0.617)	(0.081)	(0.001)
KLIC	–	0.008	0.049	0.116	0.309	1.304
	–	(0.892)	(0.158)	(0.018)	(0.008)	(0.000)
Panel C: Histogram moments (five-year-ahead expectations)						
Group size	19	18	13	13	18	17
Mean	1.66	1.62	1.78	1.97	1.99	2.08
Standard deviation	0.95	0.96	1.07	1.11	1.07	0.97
Skewness	−0.15	−0.17	−0.13	0.09	0.02	−0.26
Kurtosis	3.58	3.58	3.39	3.18	3.24	3.34
Panel D: Distance measures (five-year-ahead expectations)						
Hotelling	–	0.436	0.576	1.049	1.718	1.818
	–	(0.924)	(0.826)	(0.444)	(0.127)	(0.107)
Bonferroni	–	−0.594	1.581	1.499	1.418	−3.865
	–	(1.000)	(1.000)	(1.000)	(1.000)	(0.006)
KLIC	–	0.002	0.023	0.074	0.071	0.109
	–	(0.994)	(0.442)	(0.270)	(0.152)	(0.000)

Notes: Panel A presents moments (based on the 'mass-at-midpoint' approach) for the aggregate one-year-ahead inflation expectations from the 2021Q1 to 2022Q2 waves of the SPF. Panel B shows the test statistics relative to the 2021Q1 wave along with corresponding p-values in parentheses. For the multiple testing approach, we report the largest test statistic across the twelve distinct bins and twelve times the corresponding p-value. Panels C and D present the results for the five-year-ahead inflation expectations.

Acknowledgements

Jonas Dovern (Corresponding author)
Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, Germany; email: jonas.dovern@fau.de

Alexander Glas
Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, Germany; email: alexander.glas@fau.de

Geoff Kenny
European Central Bank, Frankfurt am Main, Germany; email: geoff.kenny@ecb.europa.eu

© European Central Bank, 2023

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

All rights reserved. Any reproduction, publication and reprint in the form of a different publication, whether printed or produced electronically, in whole or in part, is permitted only with the explicit written authorisation of the ECB or the authors.

This paper can be downloaded without charge from www.ecb.europa.eu, from the Social Science Research Network electronic library or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website.

PDF ISBN 978-92-899-5987-2 ISSN 1725-2806 doi:10.2866/36457 QB-AR-23-028-EN-N