The Wisdom of Multiple Guesses

Johan Ugander, Microsoft Research
Joint work with Ryan Drapeau and Carlos Guestrin, University of Washington

ACM EC’15
June 19, 2015
Wisdom of Crowds

Francis Galton at a country fair in 1907:

- 787 people guessing the weight of ox
- Median of guesses was 1207 lbs
- True weight was 1198 lbs
Heterogeneous Wisdom of Crowds

Francis Galton at a country fair in 1907:
- 787 people guessing the weight of ox
- Median of guesses was 1207 lbs
- True weight was 1198 lbs

This talk:
- Heterogeneously uncertain crowds
- How can/should we elicit uncertainty?
- How can/should use use uncertainty?

Related: [Jose et al. 2013, Budescu and Chen 2014, Goldstein et al. 2014, Davis-Stober et al. 2014]
Aggregation with uncertainty

California’s Drought

Californians would be better off on average if all final users in the state paid the same price for water — adjusted for quality, place and time — even if, as a result, some food prices rose sharply and some farms failed.
Individual uncertainty

Premise:

• Individuals have **belief distributions** [Wallsten et al. '97, Vul–Pashler '08]
• Possess different information/data [Frongillo et al. '15]
Individual uncertainty

Premise:

• Individuals have belief distributions
  [Wallsten et al. ’97, Vul–Pashler ’08]
• Possess different information/data
  [Frongillo et al. ’15]
Individual uncertainty

Premise:

- Individuals have belief distributions \[\text{[Wallsten et al. ’97, Vul–Pashler ‘08]}\]
- Possess different information/data \[\text{[Frongillo et al. ’15]}\]
Individual uncertainty

Premise:

- Individuals have belief distributions [Wallsten et al. ’97, Vul–Pashler ’08]
- Possess different information/data [Frongillo et al. ’15]
- Independent, no social interference [Lorenz et al. ’11, Das et al. ’13]
Measures of uncertainty

Possible approaches:

• Variance, standard deviation
• Interquantile ranges: [5%, 95%], [25%, 75%]
• Many others measures of dispersion (MAD, etc.)
Measures of uncertainty

Possible approaches:

- Variance, standard deviation
- Interquantile ranges: [5%, 95%], [25%, 75%]
- Many others measures of dispersion (MAD, etc.)

What’s “useful” for crowd aggregation?
Uncertainty for crowd aggregation

Best aggregation strategy depends on shape of belief distributions.

**Weighted mean:**
MLE if people’s guesses are drawn from $X_i \sim \text{Normal}(\mu,\sigma_i^2)$

$$
\hat{\mu}_1 = \frac{1}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}
$$

**Weighted median:**
MLE if people’s guesses are drawn from $X_i \sim \text{Laplace}(\mu,\sigma_i^2)$

$$
\hat{\mu}_2 = \arg\min_m \sum_{i=1}^{n} \frac{1}{\sigma_i} |x_i - m|
$$
Uncertainty for crowd aggregation

Best aggregation strategy depends on shape of belief distributions.

**Weighted mean:**
MLE if people’s guesses are drawn from $X_i \sim \text{Normal}(\mu, \sigma_i^2)$

$$\hat{\mu}_1 = \frac{1}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} \sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}$$

**Weighted median:**
MLE if people’s guesses are drawn from $X_i \sim \text{Laplace}(\mu, \sigma_i^2)$

$$\hat{\mu}_2 = \arg\min_m \sum_{i=1}^{n} \frac{1}{\sigma_i} |x_i - m|$$

**Galton:** means give “voting power to cranks in proportion to their crankiness”.
Uncertainty for crowd aggregation

Aggregators want $\text{var/std}$. What if we have confidence intervals?
Uncertainty for crowd aggregation

Aggregators want var/std. What if we have confidence intervals?

**Proposition.** For any $X$ belonging to a location-scale family $F$, any interquantile range between fixed quantiles $p$ and $q$ is proportional to the standard deviation,

$$IQR(X; p, q) = c_F(p, q) \sqrt{Var(X)}$$

with a constant that depends only on $F$ for all $X$. 
Uncertainty for crowd aggregation

Aggregators want var/std. What if we have confidence intervals?

**Proposition.** For any $X$ belonging to a location-scale family $F$, any interquantile range between fixed quantiles $p$ and $q$ is proportional to the standard deviation,

$$IQR(X; p, q) = c_F(p, q) \sqrt{Var(X)}$$

with a constant that depends only on $F$ for all $X$. 

![Graph showing height of the Space Needle, m vs. height in meters]
Uncertainty for crowd aggregation

Aggregators want var/std. What if we have confidence intervals?

**Proposition.** For any $X$ belonging to a location-scale family $F$, any interquantile range between fixed quantiles $p$ and $q$ is proportional to the standard deviation,

$$IQR(X; p, q) = c_F(p, q) \sqrt{Var(X)}$$

with a constant that depends only on $F$ for all $X$. 

![Graph showing interquantile ranges for different distributions.](graph.png)
Uncertainty for crowd aggregation

Aggregators want \( \text{var}/\text{std}. \) What if we have confidence intervals?

**Proposition.** For any \( X \) belonging to a location–scale family \( F \), any interquantile range between fixed quantiles \( p \) and \( q \) is proportional to the standard deviation,

\[
IQR(X; p, q) = c_F(p, q) \sqrt{Var(X)}
\]

with a constant that depends only on \( F \) for all \( X \).

**Result:** Can aggregate using interquantile ranges \( u_i \) instead of std \( \sigma_i \):

\[
\hat{\mu}_1 = \frac{1}{\sum_{j=1}^{n} \frac{1}{u^2_j}} \sum_{i=1}^{n} \frac{x_i}{u^2_i} \quad \hat{\mu}_2 = \arg\min_m \sum_{i=1}^{n} \frac{1}{u_i} |x_i - m|
\]
Uncertainty for crowd aggregation

Aggregators want var/std. What if we have confidence intervals?

**Proposition.** For any $X$ belonging to a location-scale family $F$, any interquantile range between fixed quantiles $p$ and $q$ is proportional to the standard deviation,

$$IQR(X; p, q) = c_F(p, q) \sqrt{Var(X)}$$

with a constant that depends only on $F$ for all $X$.

**Result:** Can aggregate using interquantile ranges $u_i$ instead of std $\sigma_i$:

$$\hat{\mu}_1 = \frac{1}{\sum_{j=1}^{n} \frac{1}{u_j^2}} \sum_{i=1}^{n} \frac{x_i}{u_i^2} \quad \hat{\mu}_2 = \arg\min_m \sum_{i=1}^{n} \frac{1}{u_i} |x_i - m|$$

$p=0.25, q=0.75$

Normal $c_F = 1.349$
Laplace $c_F = 1.386$
Eliciting what we can use

We can use std or interquantile range.

What can we elicit? Can we incentivize people to honestly state their uncertainty?

Yes, with scoring rules that incentivize honest responses from expected utility maximizers.

[Brier '50; Savage '71]
Eliciting what we can use

We can use std or interquantile range.

What can we elicit? Can we incentivize people to honestly state their uncertainty?

Yes, with scoring rules that incentivize honest responses from expected utility maximizers.

[Brier ’50; Savage ’71]

Other angles: competitive games, reputations, “Bayesian Truth Serum”
Eliciting uncertainty

Known scoring rule for first and second moments $m_1$, $m_2$:

$$S_{\text{Brier}}(m_1, m_2; X) = (2m_1 X - m_1^2) + (2m_2 X^2 - m_2^2)$$

Known scoring rule for [25%, 75%] confidence interval:

$$S_{\text{interval}}(\ell, u; X) = (u - \ell) + 4(\ell - X)1[X < \ell] + 4(X - u)1[X > u]$$
Eliciting uncertainty

Known scoring rule for first and second moments $m_1, m_2$:

$$S_{\text{Brier}}(m_1, m_2; X) = (2m_1 X - m_1^2) + (2m_2 X^2 - m_2^2)$$

Known scoring rule for [25%, 75%] confidence interval:

$$S_{\text{interval}}(\ell, u; X) = (u - \ell) + 4(\ell - X)1[X < \ell] + 4(X - u)1[X > u]$$

Just because a scoring rule makes people honest doesn’t make it accurate.
Multiple guesses scoring rule

We propose and analyze a multiple guesses scoring rule:

$$S_{MG,k}(\{r_1, \ldots, r_k\}; X) = \min\{|X - r_1|, \ldots, |X - r_k|\}$$

“Make multiple guesses, you’re rewarded based on closest guess”

Can think of as harnessing “dialectical crowds within” [Herzog–Hertwig ’09]
Multiple guesses scoring rule

We propose and analyze a multiple guesses scoring rule:

\[ S_{MG,k}(\{r_1, \ldots, r_k\}; X) = \min\{|X - r_1|, \ldots, |X - r_k|\} \]

“Make multiple guesses, you’re rewarded based on closest guess”

Can think of as harnessing “dialectical crowds within” [Herzog–Hertwig ’09]

Simplest case, two guesses scoring rule:

\[ S_{MG,2}(\{r_1, r_2\}; X) = \min\{|X - r_1|, |X - r_2|\} \]

Intuitively, spread out your guesses:

![Graph showing distribution functions]
Multiple guesses scoring rule

\[ S_{MG,2}(\{r_1, r_2\}; X) = \min\{|X - r_1|, |X - r_2|\} \]

Do guesses correspond to fixed quantiles \( p, q \) of belief distributions? If so, we can use the \textbf{inter-guess range} for weighted aggregation.
Multiple guesses scoring rule

\[ S_{MG,2}(\{r_1, r_2\}; X) = \min\{|X - r_1|, |X - r_2|\} \]

Do guesses correspond to fixed quantiles \( p, q \) of belief distributions? If so, we can use the **inter-guess range** for weighted aggregation.
Multiple guesses scoring rule

$$S_{MG,2}(\{r_1, r_2\}; X) = \min\{|X - r_1|, |X - r_2|\}$$

Do guesses correspond to fixed quantiles p, q of belief distributions? If so, we can use the **inter-guess range** for weighted aggregation.

For what belief distributions do multiple guesses “work”?
**Proposition.** For any log-concave $X$ the multiple guesses scoring rule is strictly proper for a set of quantiles $r_1, \ldots, r_k$.

**Proposition.** These quantiles are fixed for all symmetric $X$ within the same location–scale family.
**Proposition.** For any log-concave X the multiple guesses scoring rule is strictly proper for a set of quantiles \( r_1, \ldots, r_k \).

**Proof:** Corollary of log-concavity being a sufficient condition for uniqueness of k-medians for continuous 1D distributions.

Proven by the Mountain Pass Theorem: global min is the only local min!
Multiple guesses scoring rule

**Proposition.** For any log-concave $X$ the multiple guesses scoring rule is strictly proper for a set of quantiles $r_1,\ldots,r_k$.

**Proof:** Corollary of log-concavity being a sufficient condition for uniqueness of $k$-medians for continuous 1D distributions.

Proven by the Mountain Pass Theorem: global min is the only local min!

Gradient descent finds the global min. Not crazy to think that agents with bounded rationality can do well.
So far:

- Uncertainty-weighted aggregation:
  - \( \sigma_i^2 \)-weighted mean, \( \sigma_i \)-weighted median
  - Assume location-scale family: can replace with interquantile ranges
  - If symmetric log-concave: two guesses scoring rule elicits [25%, 75%]
What if uncertainties are wrong?

- Tukey contamination model: mixture of \( N(0, 1) \) and \( N(0, b) \) beliefs.
What if uncertainties are wrong?

- Tukey contamination model: mixture of $N(0,1)$ and $N(0,b)$ beliefs.

- Need better methods to handle “certainty-cranks”
Experiments

- Is weighted aggregation better than unweighted?
- Better to use weighted mean or weighted median?
- Better to ask for Interval or to use multiple guesses?
Mechanical Turk experiments

Experiments on Amazon Mechanical Turk using a “Dot Guessing Game”:

• Players saw 30 images with variable numbers of dots
Mechanical Turk experiments

Experiments on Amazon Mechanical Turk using a “Dot Guessing Game”:

• Players saw 30 images with variable numbers of dots
• Split in 3 rounds (random order): 1 guess, 2 guesses, 3 guesses
Mechanical Turk experiments

Experiments on Amazon Mechanical Turk using a “Dot Guessing Game”:

• Players saw 30 images with variable numbers of dots
• Split in 3 rounds (random order): 1 guess, 2 guesses, 3 guesses

• Pre-game tutorial, feedback about bonuses
Mechanical Turk experiments

- Dot counts ranged from 27 to 226.
- Very fewer dots (=very easy task): two guesses “gets in way”
- Rest: relative MSE was ~3x lower with 2-guess weighted aggregation

Weighted Median vs. Median

Weighted Mean vs. Mean
Mechanical Turk experiments

- **3 Guesses:** Symmetric?
  - Look at gap $g_3 - g_2$ vs. $g_2 - g_1$
  - 48% of triplets perfectly symmetric

- 3-guess aggregation statistically indistinguishable from 2-guesses aggregation.
Mechanical Turk experiments

- Calibration experiment: 2-guesses rule vs. Interval rule for [25%, 75%]

- Interval-weighted aggregation statistically indistinguishable from 2-guess weighted aggregation.
Concluding thoughts

- Eliciting and utilizing uncertainty: smarter use of (smaller) crowds
- Better ways to elicit/utilize? Ask questions that are easy for humans to answer accurately, make algorithms do the heavy lifting.
- “Conditionally strictly proper scoring rules”: strictly proper conditional on (hopefully reasonable) assumptions.
- Global min is only local min: interesting notion of efficiently computable.
- Shape of belief distribution family important.
- Methods for “certainty-cranks”
- Symmetric beliefs: not helpful to ask for more than 2 guesses.