Applied Data Analytics

Statistics — Dispersion & concentration

Measurement error and summary statistics

Hans-Martin von Gaudecker and Aapo Stenhammar

Goals

- Remember measurement issues are pervasive
- Show that unsystematic measurement errors impact only variance and higher
- Show that systematic measurement errors impact all statistics
- Realise that we need to understand where the data comes from

Model

- True value: x_i^*
- Actual measurement: $x_i \equiv x_i^* + \varepsilon_i$
- $\sum_{i=1}^{n} \varepsilon_i \equiv \bar{\varepsilon} = 0$
- $\frac{1}{n-1} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 > 0$

Mean

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\sum_{i=1}^{n} (x_i^* + \varepsilon_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^* + \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^* \\
&= \bar{x^*}
\end{aligned}
$$

Variance

$$
\begin{aligned}
s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^* + \varepsilon_i - \bar{x^*})^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^* - \bar{x^*})^2 + \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2
\end{aligned}
$$

Note: Third line needs more assumptions on "unsystematic".

Systematic Measurement Error

- Example: Censoring
- Real-world: German Employment Agency (BA) data
  - Reports of earnings are precise until a threshold
  - Only know that they are at least this threshold value beyond that

Model

$$x_i = \begin{cases} x_i^* & \text{if } \; x_i^* \leq 90,600\,€ \\ 90,600\,€ & \text{otherwise} \end{cases}$$

If I was to use this data directly:

- $\bar{x} < \bar{x^*}$
- $s_x < s_{x^*}$
- $x_{q_z} = x^*_{q_z}$ if $x^*_{q_z} \leq 90,600\,€$