The case of bounds in noisy protection methods: Selected risk and utility perspectives from official population statistics

2023 UNECE Expert Meeting on SDC, 26 – 28 September 2023
Risk assessment: Privacy, confidentiality, and disclosure vs utility

Fabian BACH
European Commission – Eurostat
Unit F2 – Population and migration

Outline

1. Intro: Noisy methods and bounds in a nutshell
2. Specific utility flaws of *unbounded* noise
3. Specific additional disclosure risks of *bounded* noise
4. Conclusions

Intro: noisy methods and bounds in a nutshell

- SDC ←→ protect individuals

SEX \\ POB*	Total	Country	Outside
Total	**42**	**35**	**7**
Male	**22**	17	5
Female	**20**	18	**2**

* Place of birth (POB)

Intro: noisy methods and bounds in a nutshell

- SDC ←→ protect individuals
- old-school suppression often inefficient and inconsistent

SEX \\ POB	Total	Country	Outside
Total	42	35	7
Male	22	C	C
Female	20	C	C

Intro: noisy methods and bounds in a nutshell

- SDC ←→ protect individuals
- old-school suppression often inefficient and inconsistent
- Noise in action: Is this better?

SEX \\ POB	Total	Country	Outside
Total	42	37	7
Male	23	15	4
Female	21	16	3

Intro: noisy methods and bounds in a nutshell

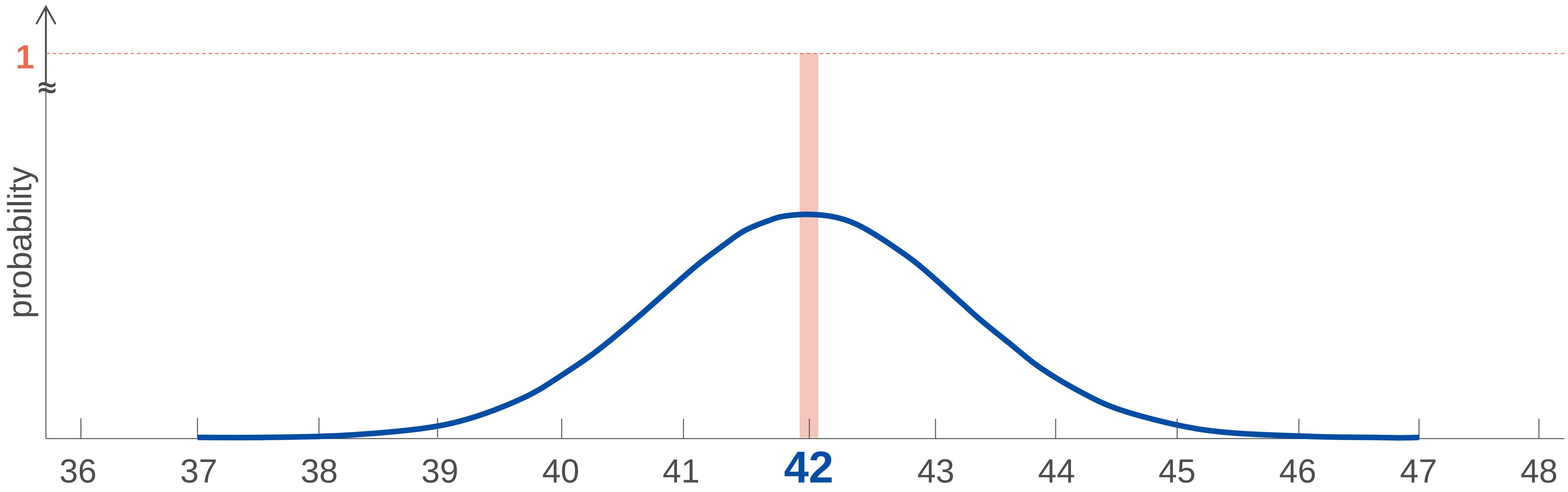
- SDC ←→ protect individuals
- old-school suppression often inefficient and inconsistent
- Noise in action: Is this better?

SEX \\ POB	Total	Country	Outside
Total	**42**	**37**	**7**
Male	**23**	15	4
Female	**21**	16	3

… a closer look at a single statistic …

Intro: noisy methods and bounds in a nutshell

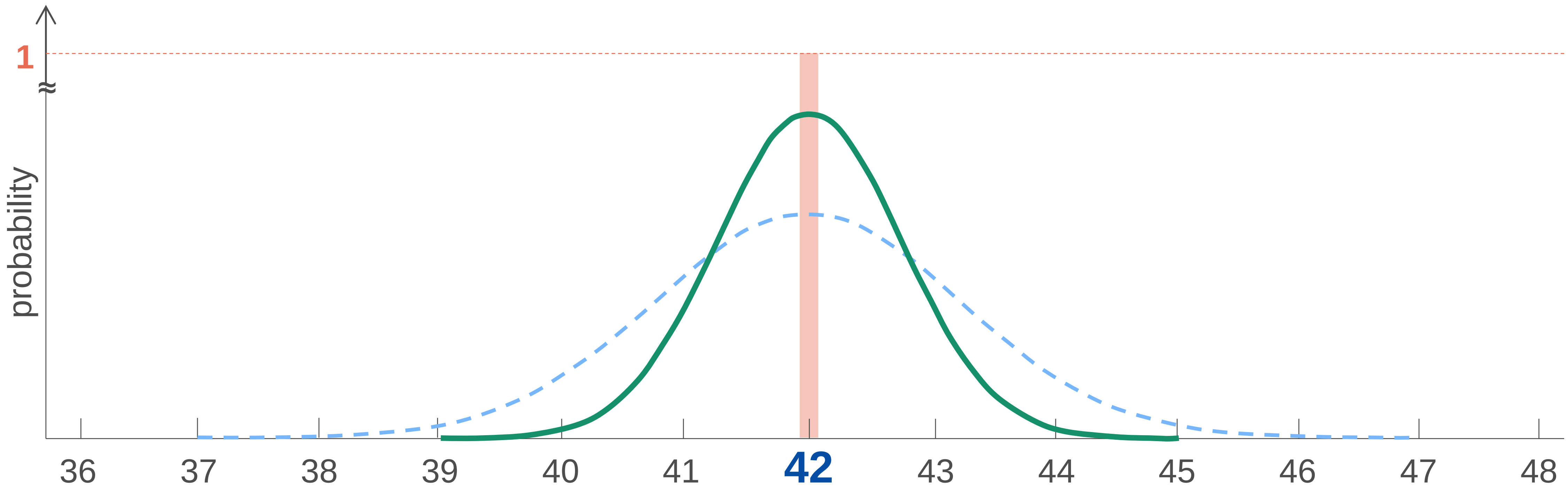
… a closer look at single statistic level: intrinsic uncertainty



intrinsic uncertainty: ± 2

Intro: noisy methods and bounds in a nutshell

… a closer look at single statistic level: intrinsic uncertainty vs. noise

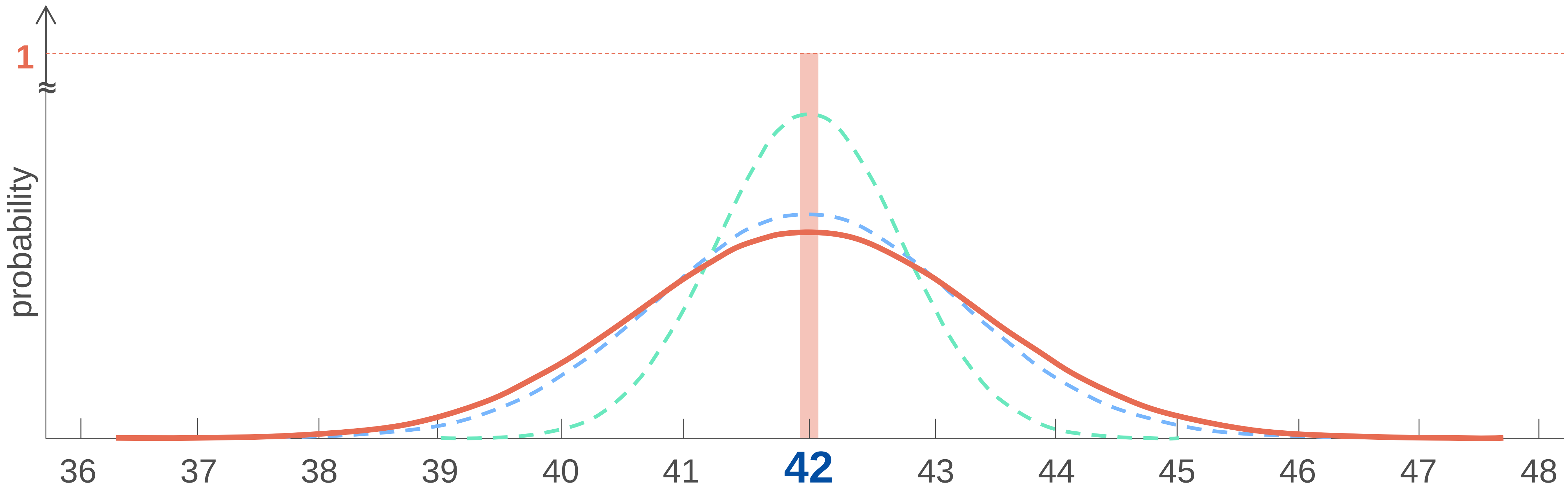


intrinsic uncertainty: ± 2

protective noise added: ± 1

Intro: noisy methods and bounds in a nutshell

… a closer look at single statistic level: intrinsic uncertainty vs. noise



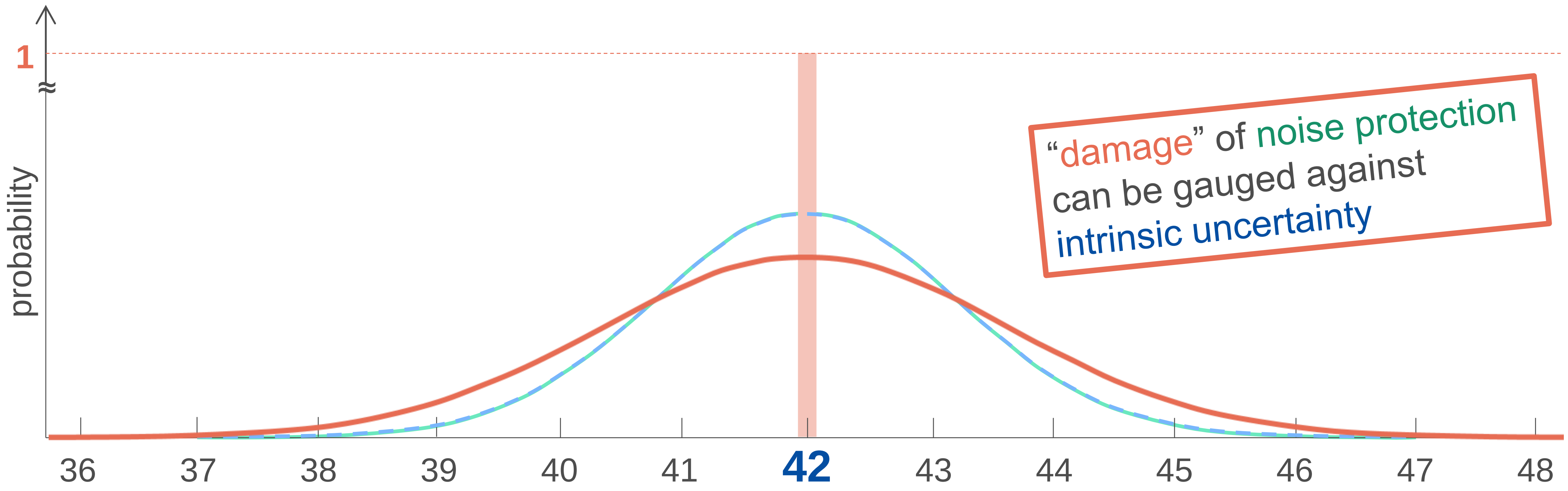
intrinsic uncertainty: ± 2

protective noise added: ± 1

total uncertainty: ± 2.2

Intro: noisy methods and bounds in a nutshell

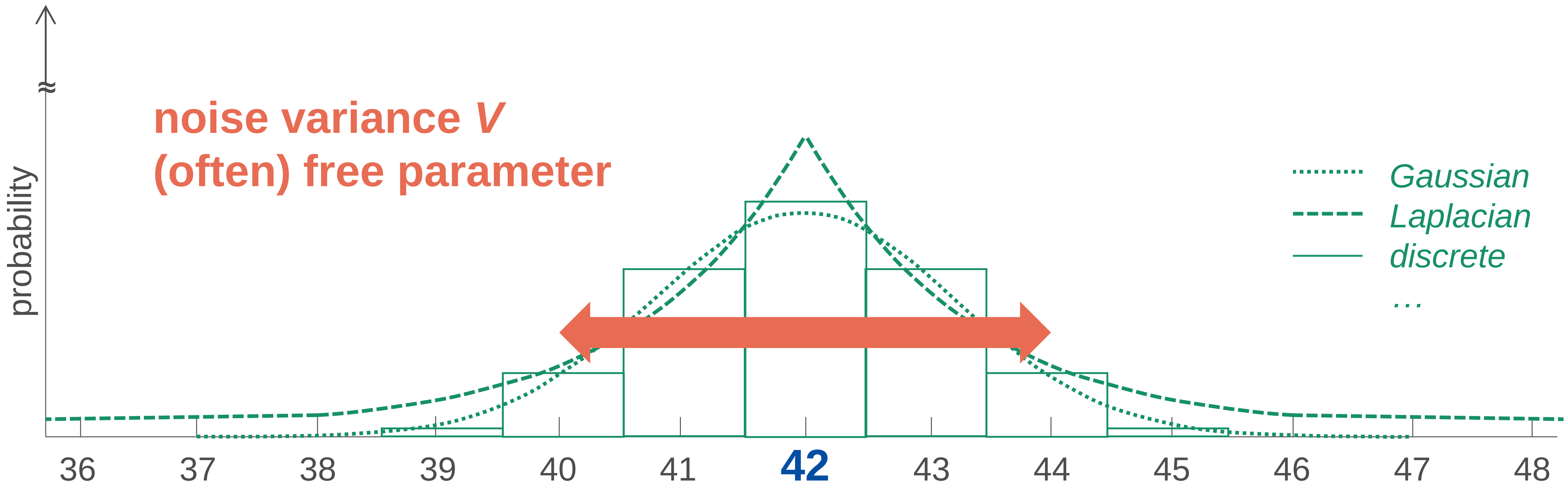
… a closer look at single statistic level: intrinsic uncertainty vs. noise



intrinsic uncertainty: ± 2
protective noise added: **± 2**
total uncertainty: ± 2.8

Intro: noisy methods and bounds in a nutshell

… a closer look at single statistic level: **noise distributions**

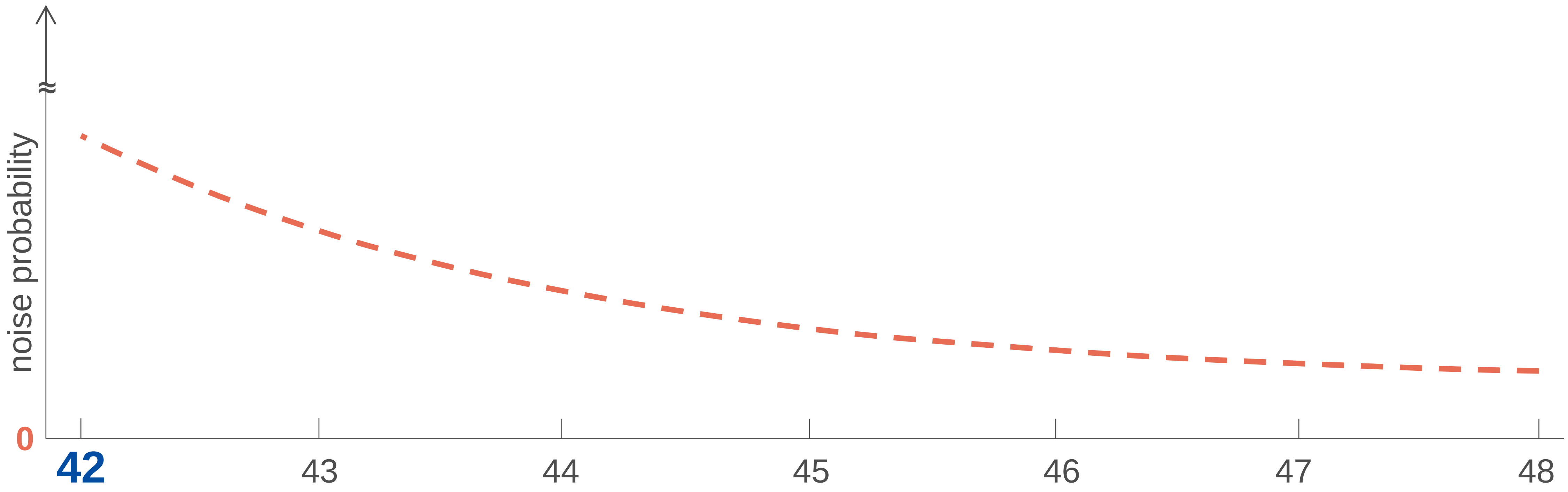


protective noise added: **± 2**

European Commission

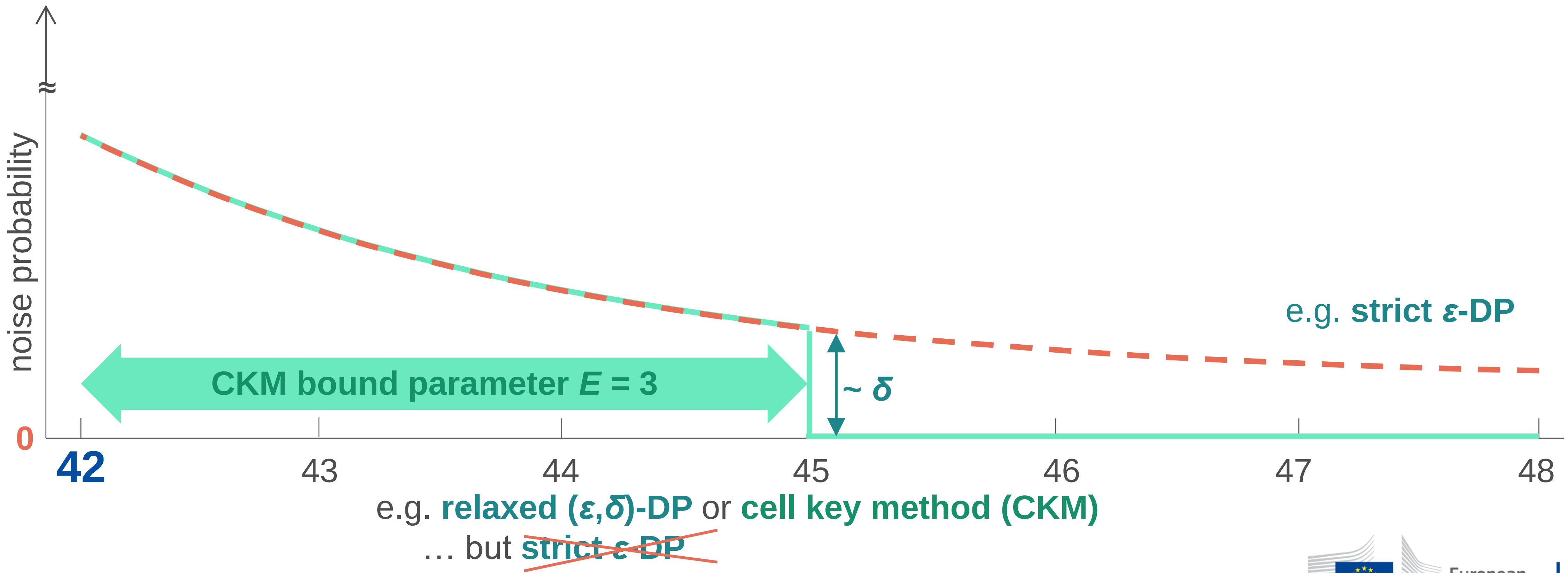
Intro: noisy methods and bounds in a nutshell

- **Noise distributions**: how long is the **tail**?



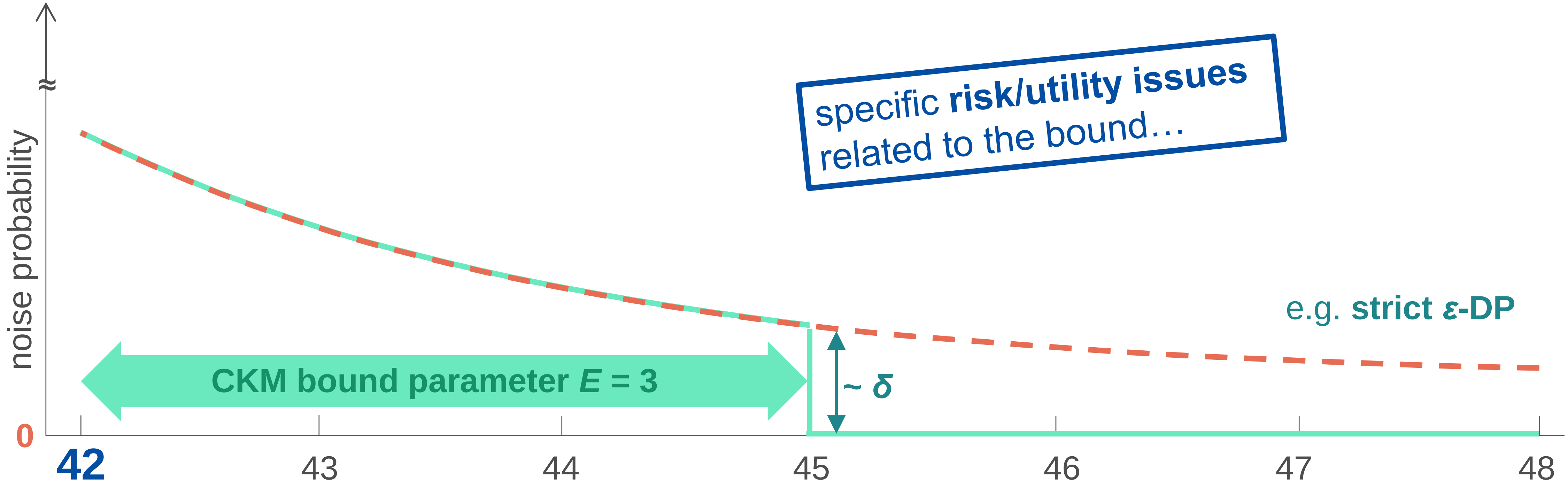
Intro: noisy methods and bounds in a nutshell

- **Noise distributions**: how long is the **tail**?



Intro: noisy methods and bounds in a nutshell

- **Noise distributions**: how long is the **tail**?



e.g. **relaxed (ε,δ)-DP** or **cell key method (CKM)**

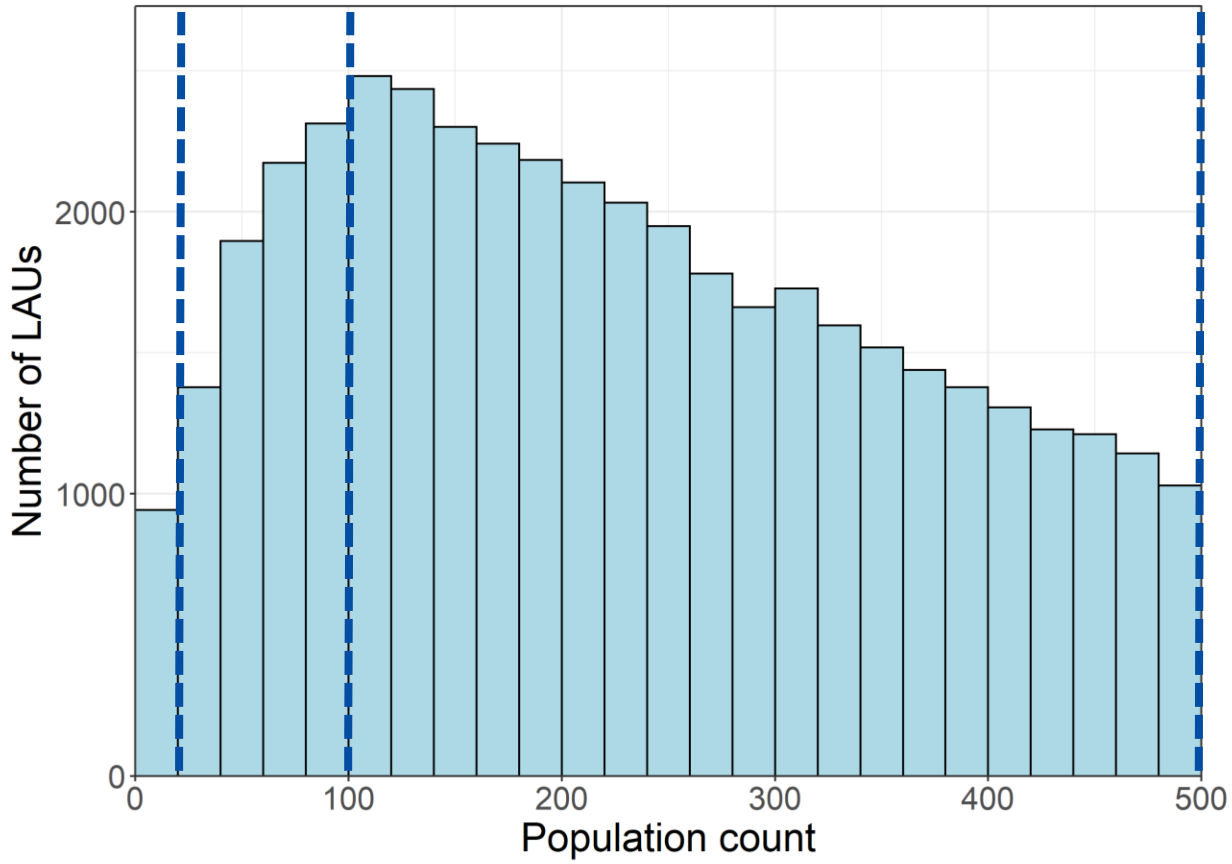
… but ~~**strict ε-DP**~~

European Commission

Utility flaws of *unbounded* noise

- 2021 EU census: ca. 110 000 **L**ocal **A**dministrative **U**nits (~ municipalities), of which
  - 43 395 with <500 people
  - 8 502 with <100 people
  - 866 with <20 people
- Could we accept here e.g. Pr(|noise|>100) = 0.1% or more?

❑ **Yes** ❑ **No**

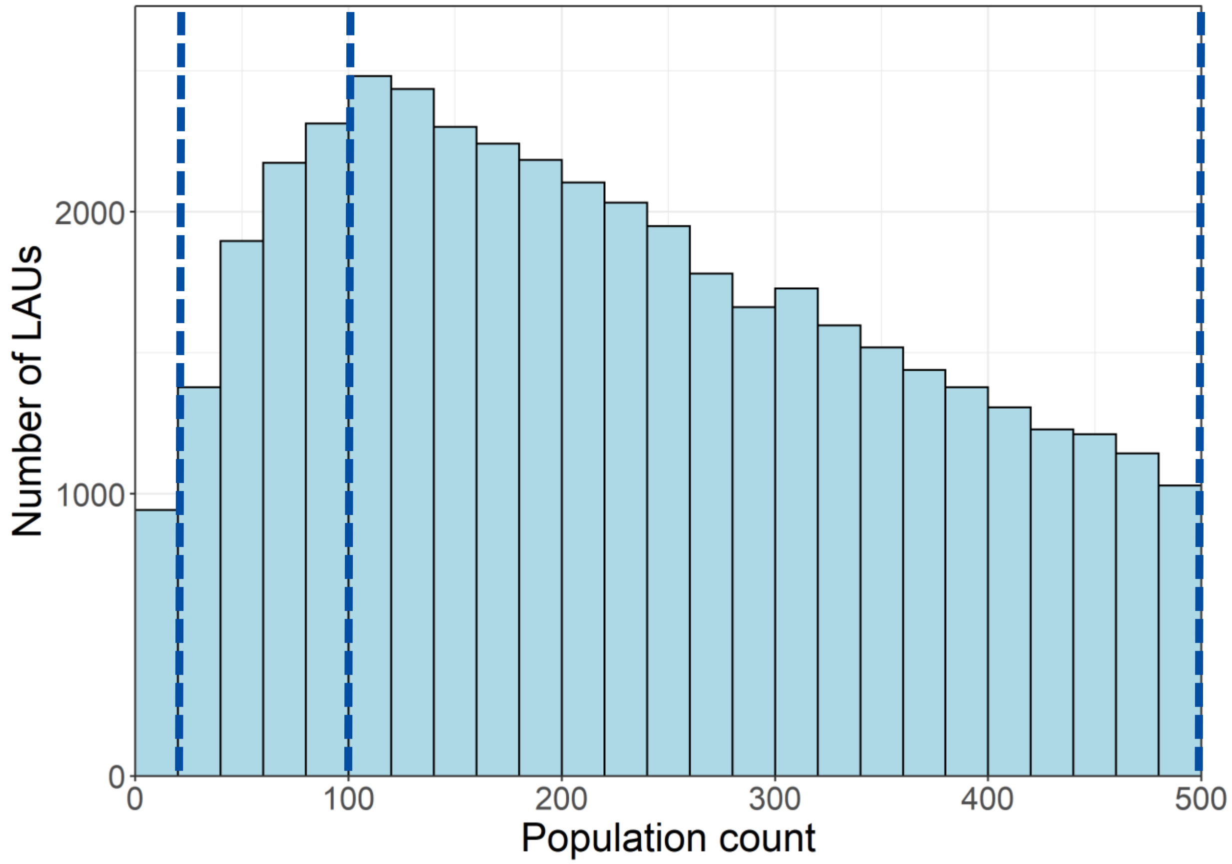


European Commission

Utility flaws of *unbounded* noise

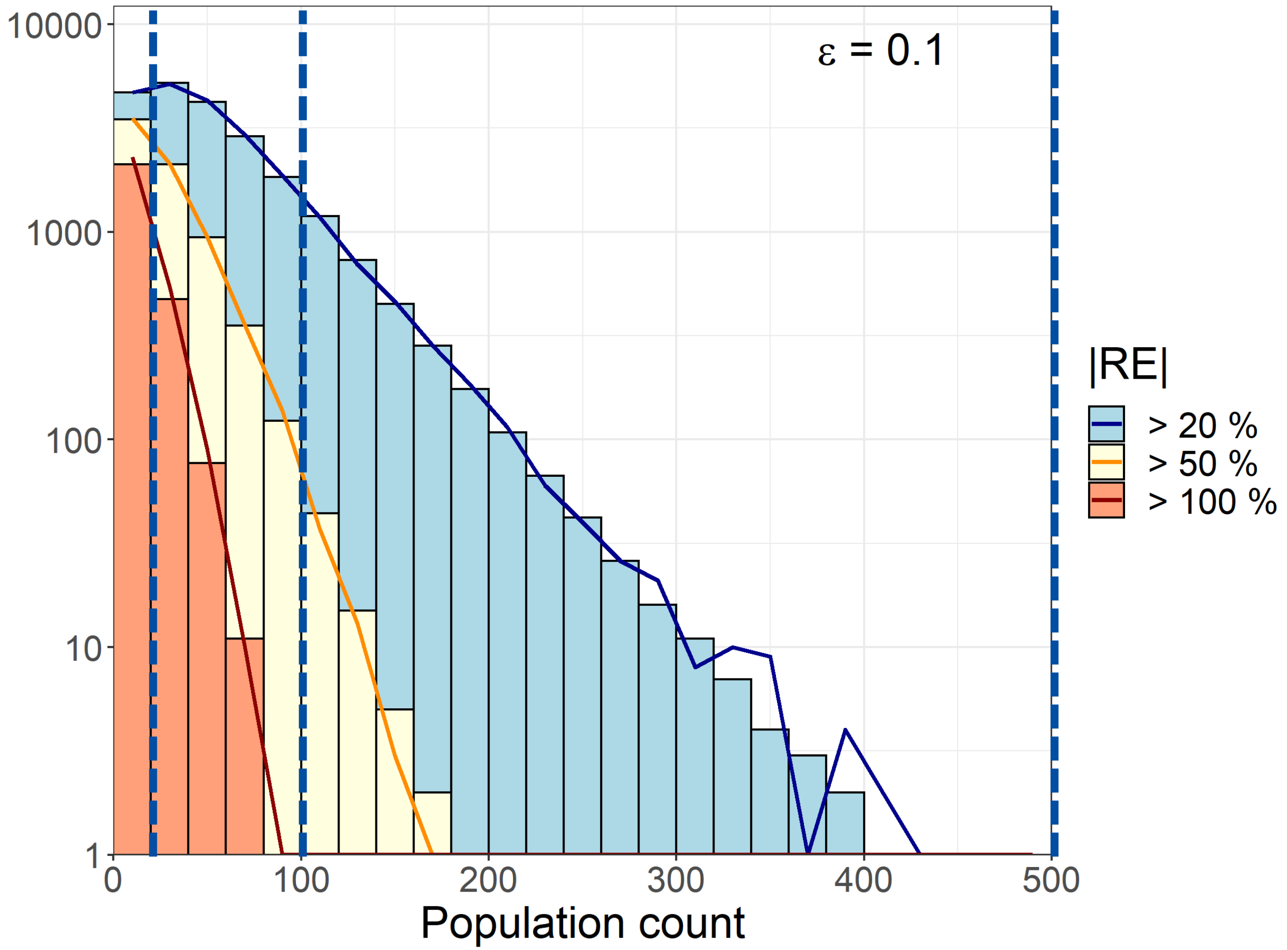
- 2021 EU census: ca. 110 000 **L**ocal **A**dministrative **U**nits (~ municipalities), of which
  - 43 395 with <500 people
  - 8 502 with <100 people
  - 866 with <20 people
- Could we accept here e.g. Pr(|noise|>100) = 0.1% or more?

❑ **Yes** ☑ **No**



Utility flaws of *unbounded* noise: counts

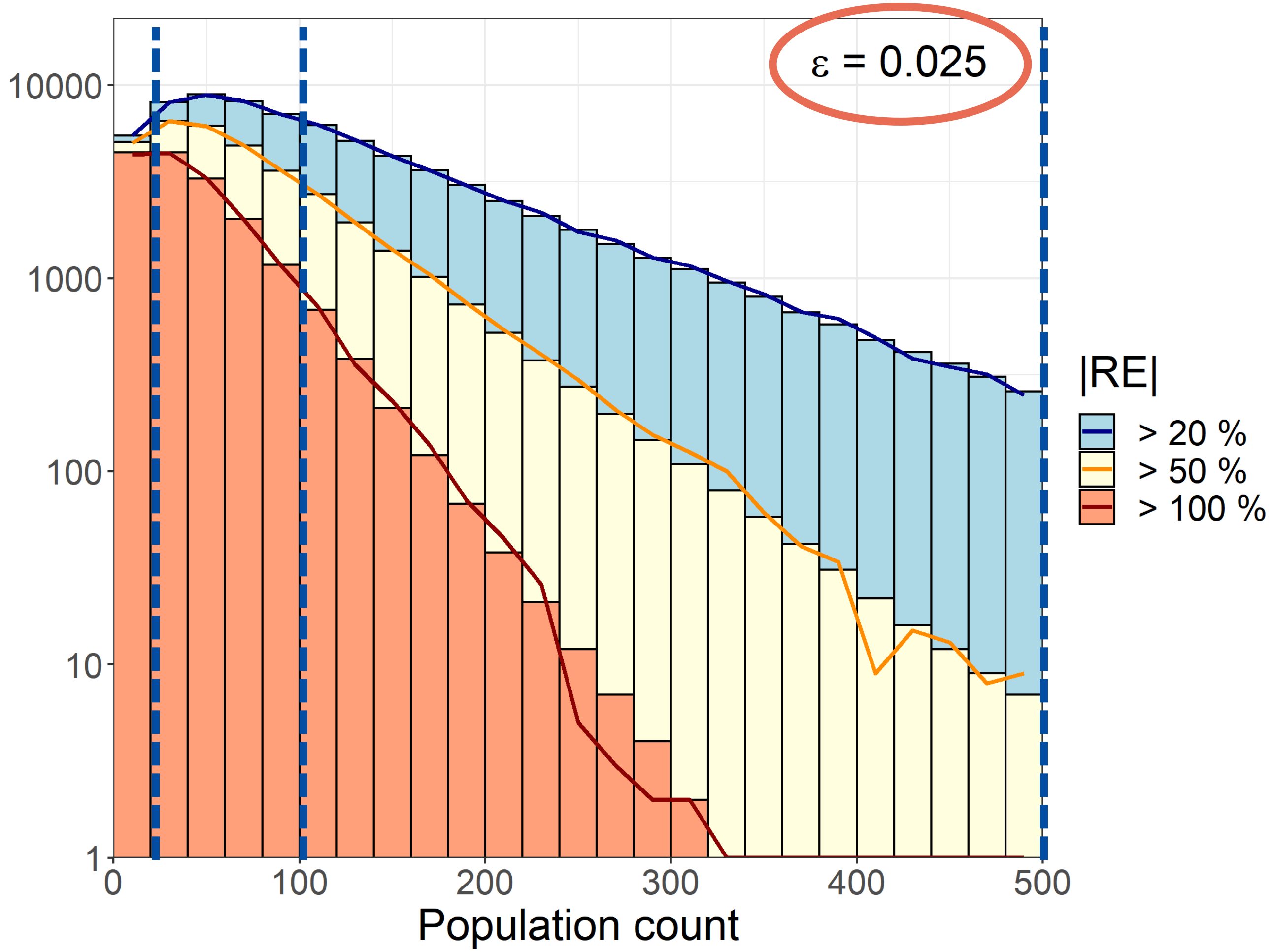
- E.g. 2020 U.S. census test setup with moderate tabular $\varepsilon = 0.1$
- expectation for individual LAU counts to obtain noise of relative size ±20, ±50 and ±100%
- analytical estimation (bins) and numerical simulation (lines)



European Commission

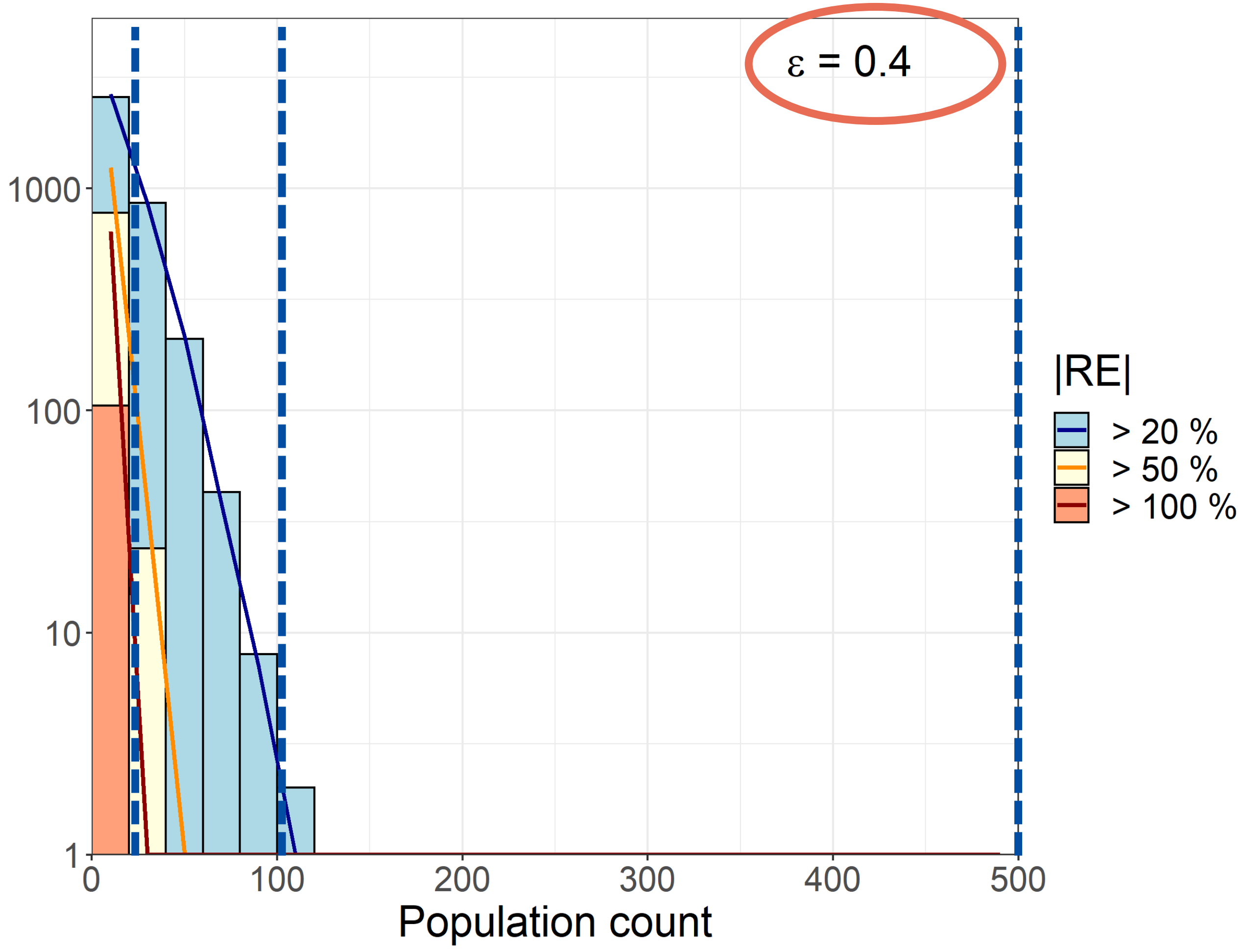
Utility flaws of *unbounded* noise: counts

- E.g. 2020 U.S. census test setup with restrictive tabular $\varepsilon = 0.025$
- expectation for individual LAU counts to obtain noise of relative size ±20, ±50 and ±100%
- analytical estimation (bins) and numerical simulation (lines)



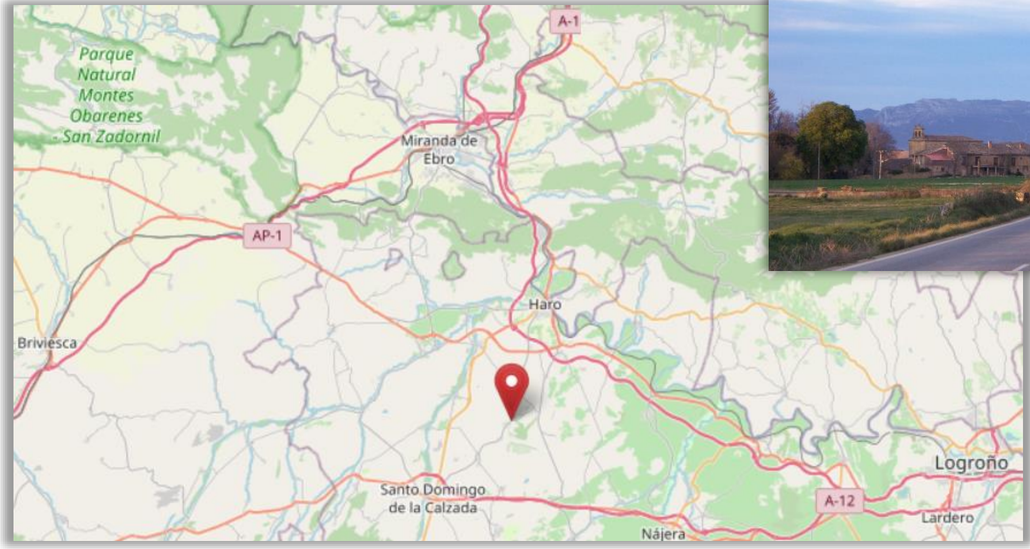
Utility flaws of *unbounded* noise: counts

- E.g. 2020 U.S. census test setup with generous tabular ε = 0.4
- expectation for individual LAU counts to obtain noise of relative size ±20, ±50 and ±100%
- analytical estimation (bins) and numerical simulation (lines)



Utility flaws of *unbounded* noise: counts

- Even worse: several counts (e.g. **T**otal, **M**ales, **F**emales) are distorted consistently
- E.g. 2020 U.S. census test setup with with moderate tabular $\varepsilon = 0.1$



source: OpenStreetMap

source: Wikipedia

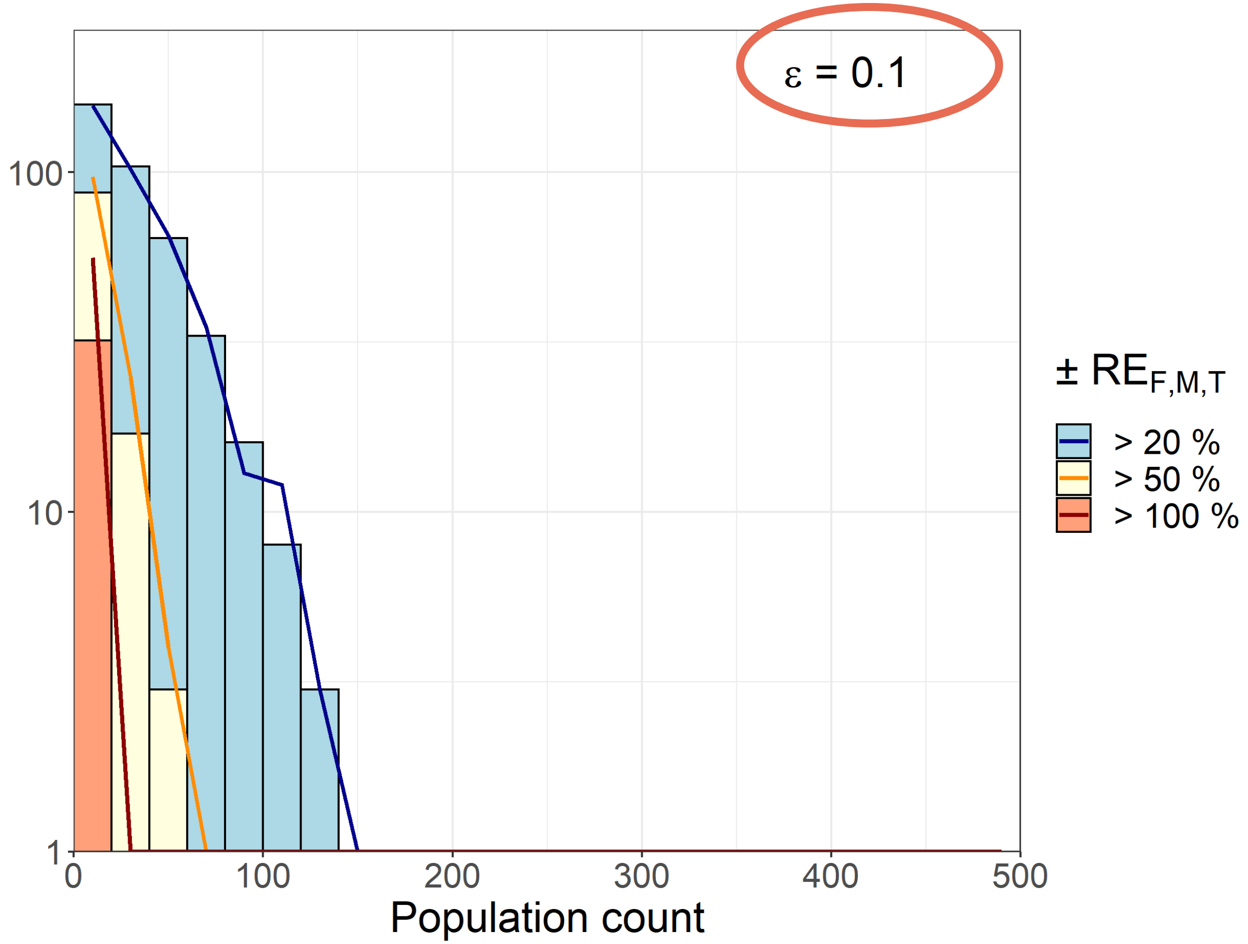
Cidamón, La Rioja, Spain
ES230_26048

	2011 census	U.S. setup (ε = 0.1)
Total	**30**	**-17**
Male	20	-1
Female	15	-9

European Commission

Utility flaws of *unbounded* noise: counts

- Even worse: several counts (e.g. **T**otal, **M**ales, **F**emales) are distorted consistently up or down
- E.g. 2020 U.S. census test setup with with moderate tabular $\varepsilon = 0.1$
- still ~20 small LAUs where ±100% would happen (~100 LAUs with ±50%)



Utility flaws of *unbounded* noise: ratios

- take very simple ratio indicator e.g. share of females: $r := F/T$

➔ standard deviation of *r* as a function of generic noise variance *V*:

$$\mathrm{sd}_r(V) = \frac{1}{T}\sqrt{V\left(1 + r^2\right)}$$

Utility flaws of *unbounded* noise: ratios

- take very simple ratio indicator e.g. share of females: $r := F/T$

  ➔ standard deviation of *r* as a function of generic noise variance *V*:

$$\mathrm{sd}_r(V) = \frac{1}{T}\sqrt{V\left(1+r^2\right)}$$

- to quantify bound effects, approximate noise effects $i = i_0 + x_i$ ($i = F, T$) as

$$r - r_0 = r\left(\xi_F - \xi_T\right) + O\left(\xi^2\right) \quad \text{with} \quad \xi_i \equiv x_i/i \ll 1$$

  ➔ in the presence of a bound *E*: $\max|r - r_0| \simeq \frac{E}{T}(1+r)$

Utility flaws of *unbounded* noise: ratios

- take very simple ratio indicator e.g. share of females: $r := F/T$

  ➔ standard deviation of *r* as a function of generic noise variance *V*:

$$\mathrm{sd}_r\,(V) = \frac{1}{T}\sqrt{V\,(1+r^2)}$$

- to quantify bound effects, approximate noise effects $i = i_0 + x_i$ ($i = F, T$) as

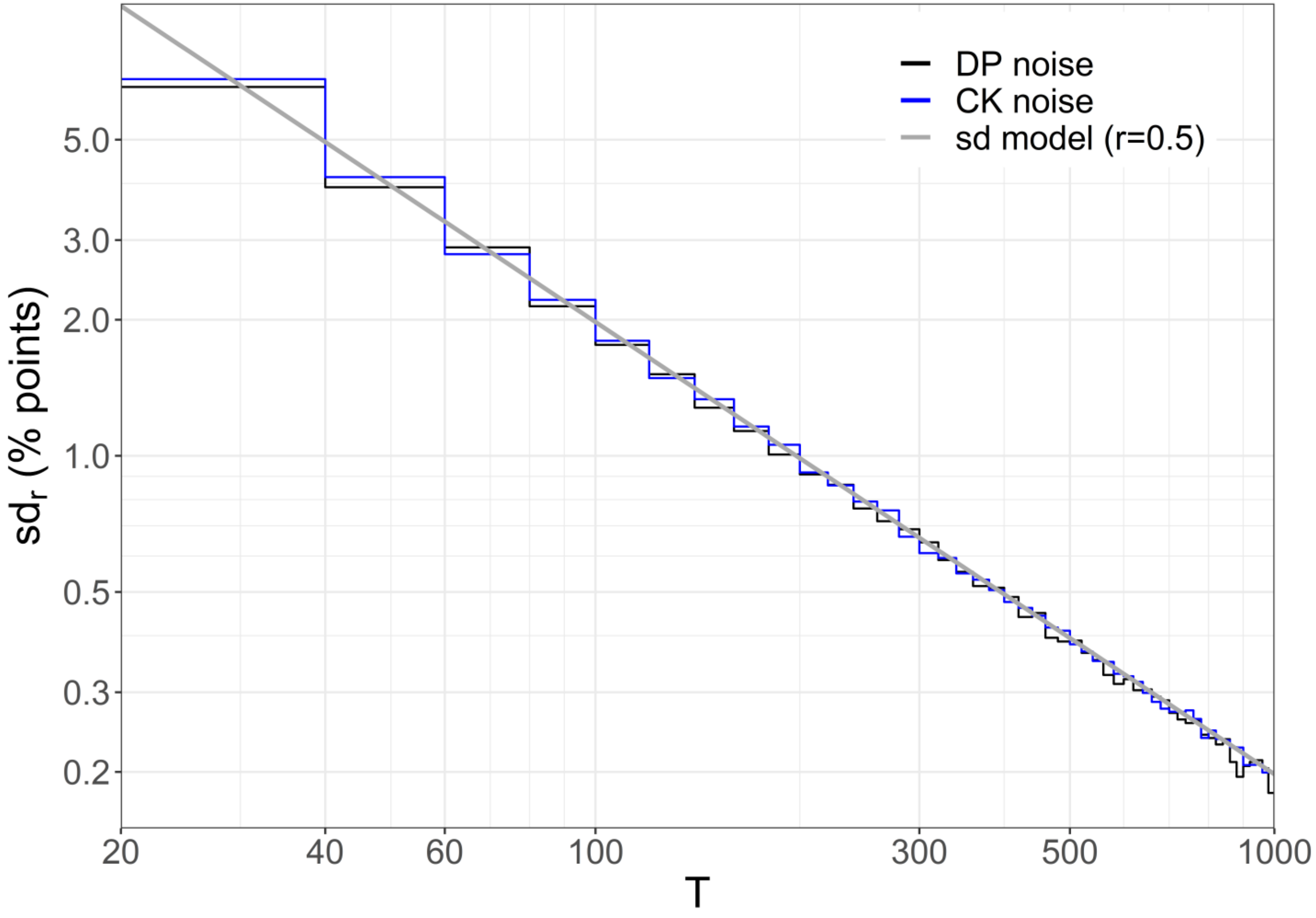
$$r - r_0 = r\,(\xi_F - \xi_T) + O\left(\xi^2\right) \quad \text{with} \quad \xi_i \equiv x_i/i \ll 1$$

  ➔ in the presence of a bound *E*: $\max|r - r_0| \simeq \frac{E}{T}(1+r)$

- this can be tested numerically with noise samples from CKM (e.g. *V*=3, *E*=6) and for comparison from unbounded *ε*-DP setup (*ε*=0.8)

Utility flaws of *unbounded* noise: ratios

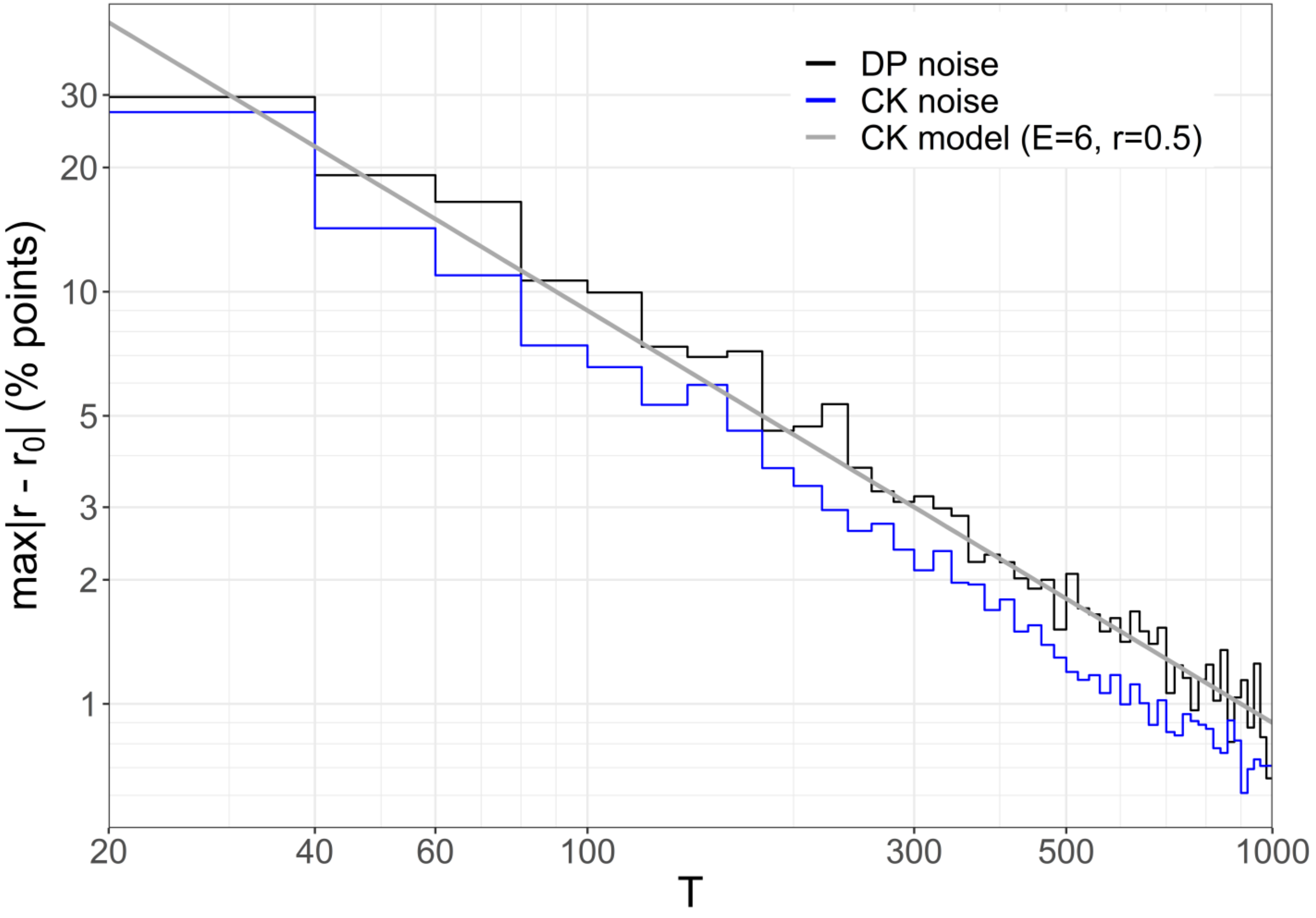
sanity check on sd_r



Utility flaws of *unbounded* noise: ratios

bound effects in max|r-r_0|

- *bounded* noise (CK) consistently below model
- *unbounded* noise (DP) consistently above model
- typical size of difference: ~5 % points across bins
- i.e. huge relative diff. for small r < 0.1 (e.g. minorities)



Additional disclosure risks of *bounded* noise

- Now would you bet all your money on a guess for the true count of the …
  - ❑ … total population?
  - ❑ … country-born males?
  - ❑ … total females?
  - ❑ … total foreign-born?

SEX \\ POB	Total	Country	Outside
Total	**42**	**37**	**7**
Male	**23**	15	4
Female	**21**	16	3

each count with noise variance *V* = 1
and noise bound *E* = 2

Additional disclosure risks of *bounded* noise

- Now would you bet all your money on a guess for the true count of the …
  - ❑ … total population?
  - ✓ … country-born males (= 17)
  - ❑ … total females?
  - ❑ … total foreign-born?

SEX \\ POB	Total	Country	Outside
Total	42	37 = 35+2	7
Male	23	15 = 17-2	4
Female	21	16 = 18-2	3

each count with noise variance $V = 1$
and noise bound $E = 2$

- But **how often** does this happen?

Additional disclosure risks of *bounded* noise

- linear constraints in breakdowns – e.g. dichotomous SEX = {*F*,*M*,*T*}:

  expectation $(F + M - T) = 0$ ➔ bound estimator $\widehat{E} = \left\lceil \left| \frac{F + M - T}{3} \right| \right\rceil$

Additional disclosure risks of *bounded* noise

- linear constraints in breakdowns – e.g. dichotomous SEX = {*F*,*M*,*T*}:

  expectation $(F + M - T) = 0$ ➔ bound estimator $\widehat{E} = \left\lceil \left| \frac{F + M - T}{3} \right| \right\rceil$

  prob. to reveal *E* from a single 3-tuple: $p_1 := \Pr[|F + M - T| > 3(E - 1)]$

  ➔ p_1 fixed by noise distribution (e.g. CKM pars. *V* and *E*)

Additional disclosure risks of *bounded* noise

- linear constraints in breakdowns – e.g. dichotomous SEX = {*F*,*M*,*T*}:

  expectation $(F + M - T) = 0$ ➔ bound estimator $\widehat{E} = \lceil \left| \frac{F+M-T}{3} \right| \rceil$

  prob. to reveal *E* from a single 3-tuple: $p_1 := \Pr[|F + M - T| > 3(E-1)]$

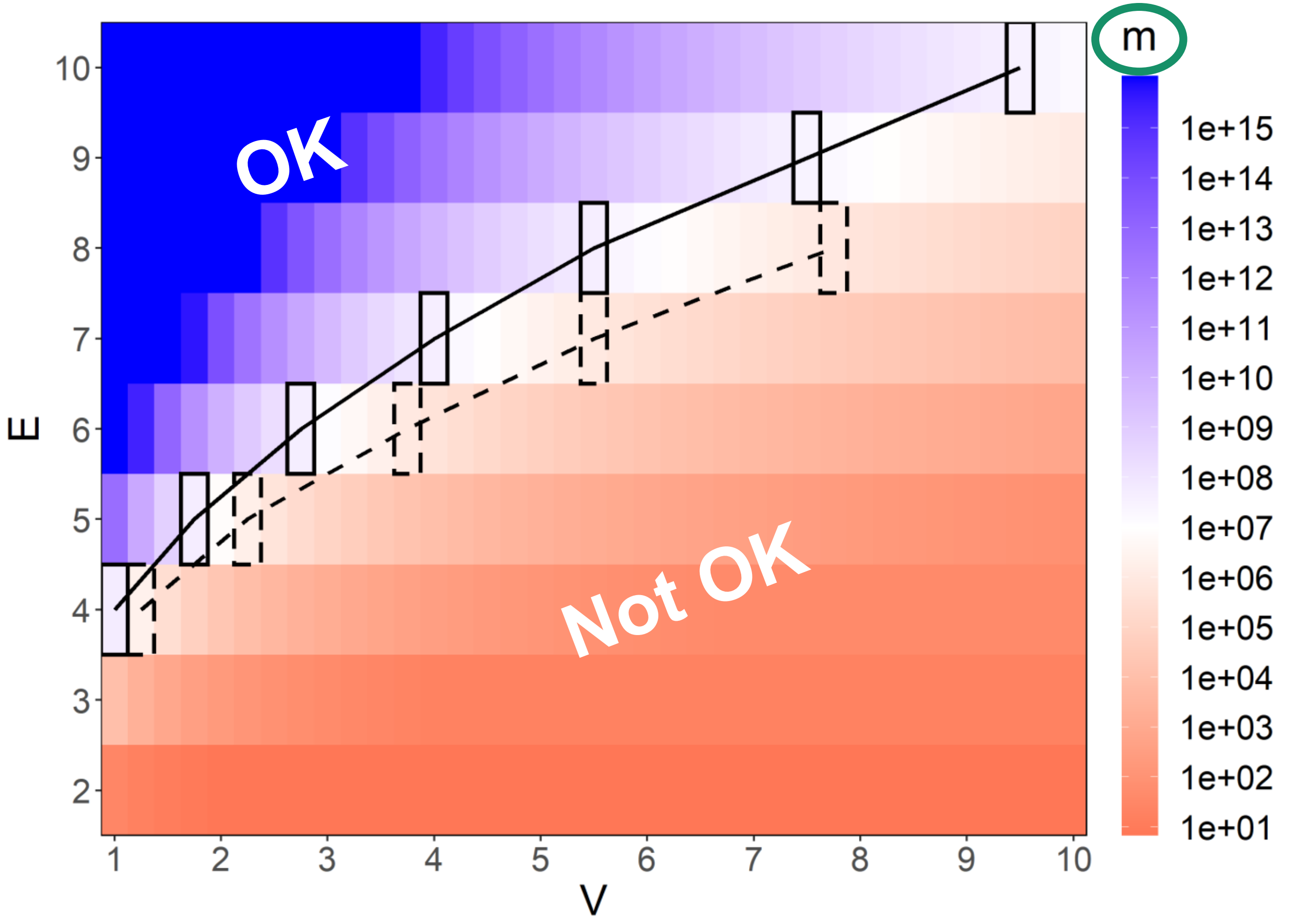
  ➔ p_1 fixed by noise distribution (e.g. CKM pars. *V* and *E*)

- number of 3-tuples needed to disclose *E* at c.l. α: $m = \lceil \frac{\log(1-\alpha)}{\log(1-p_1)} \rceil$

  ➔ available *m* fixed by table output

Additional disclosure risks of *bounded* noise

➔ Knowing the full output, the risk can be quantified systematically – e.g. for the 2021 EU census output:



m: number of 3-tuples needed in output to get ca. one *E*-disclosive noise pattern

black boxes showing where *m* exceeds the number of available 3-tuples for Malta (dashed) and Germany (solid)

Vanilla CKM from SDCTools on GitHub

European Commission

Conclusions

- in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both utility and disclosure risks

Conclusions

- in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both utility and disclosure risks – shown today:

- utility – *unbounded* noise cannot guarantee useful outputs on *all* small areas in a large output programme (e.g. EU census LAU data)

  ➔ holds for raw counts and more pronounced for shares/ratios, even with moderate noise variance (e.g. $V \sim 3$)

Conclusions

- in noisy approaches to confidentiality, whether the noise is *bounded* or *unbounded* is a key question with consequences for both utility and disclosure risks – shown today:

- utility – *unbounded* noise cannot guarantee useful outputs on *all* small areas in a large output programme (e.g. EU census LAU data)

  ➔ holds for raw counts and more pronounced for shares/ratios, even with moderate noise variance (e.g. $V \sim 3$)

- risks – *bounded* noise is additionally vulnerable to constraint exploits

  ➔ risk can be controlled by tuning noise to output complexity, with moderate noise parameters ($V \sim 2$, $E \sim 5$)

Thank you

Slide XX: map section, source: screenshot from OpenStreetMap; Slide XX: view of Cidamón, source: photo by Bigsus from Wikipedia