Why projects go over time?

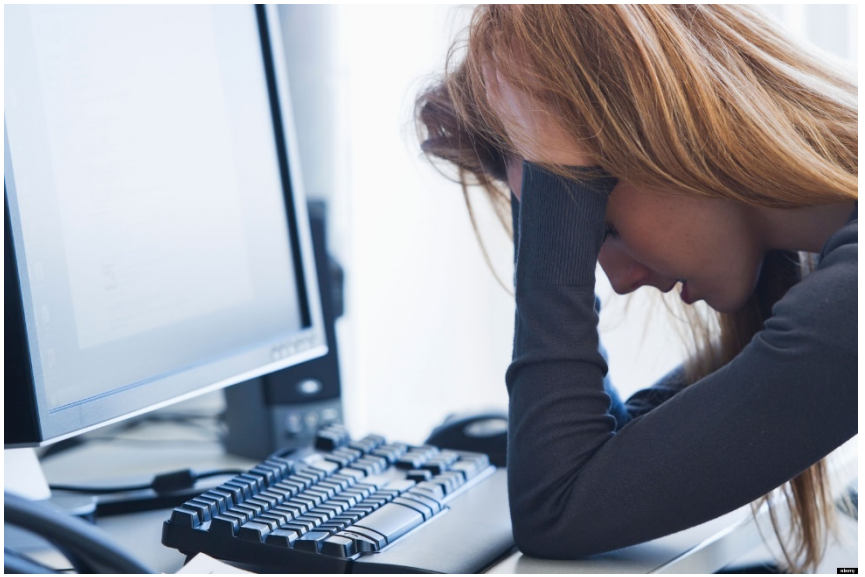
DAVID GUSZEJNOV

Underestimating project length

We have all ran out of time before deadlines or had projects „blow up”

It is common in both business and academia

Somehow we are really bad at estimating the length of projects



Are we bad at statistics?

Hypothesis from the data science blog of Erik Bernhardsson:

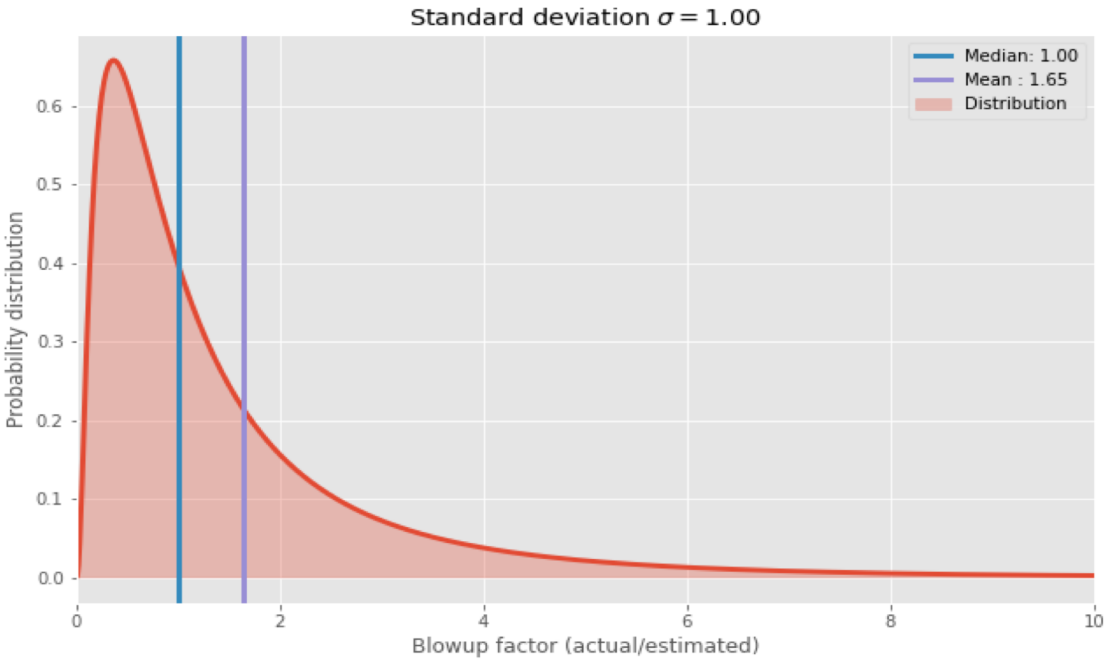
Humans can only estimate the median, not the mean

Let's define the **blowup factor** as

$$f = \frac{\Delta t_{actual}}{\Delta t_{estimate}}$$

Let's assume f is **lognormal**:

$$\log f \sim N(0, \sigma)$$



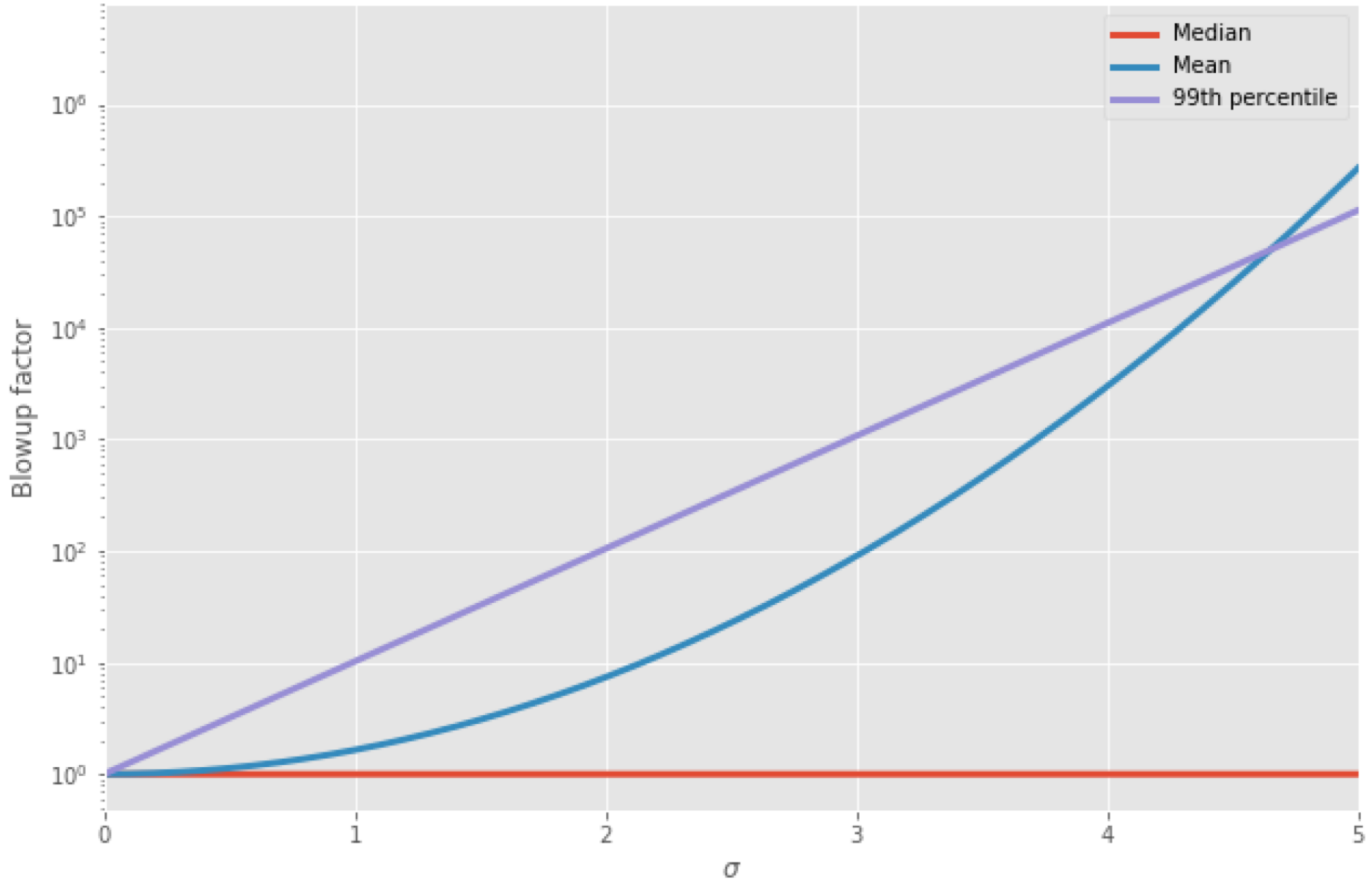
How multiple tasks add-up

	Median	Mean	99%
Task A	1.00	1.65	10.24
Task B	1.00	1.65	10.24
Task C	1.00	1.65	10.24
SUM	3.98	4.95	18.85

Even worse if projects are allowed to have different variances for the log f

	Median	Mean	99%
Task A ($\sigma = 0.5$)	1.00	1.13	3.20
Task B ($\sigma = 1$)	1.00	1.65	10.24
Task C ($\sigma = 2$)	1.00	**7.39**	**104.87**
SUM	**4.00**	**10.18**	**107.99**

Mean vs Median for high σ



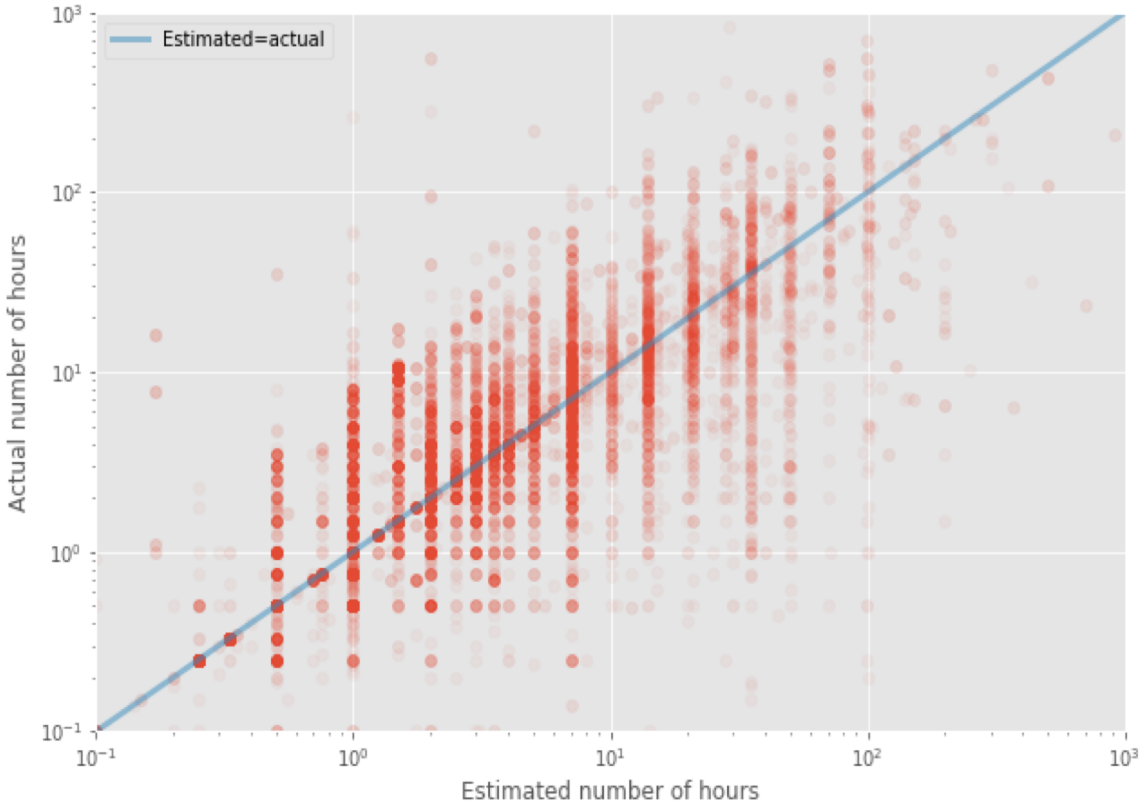
Real data on blowup factor

So far this was only a hypothesis

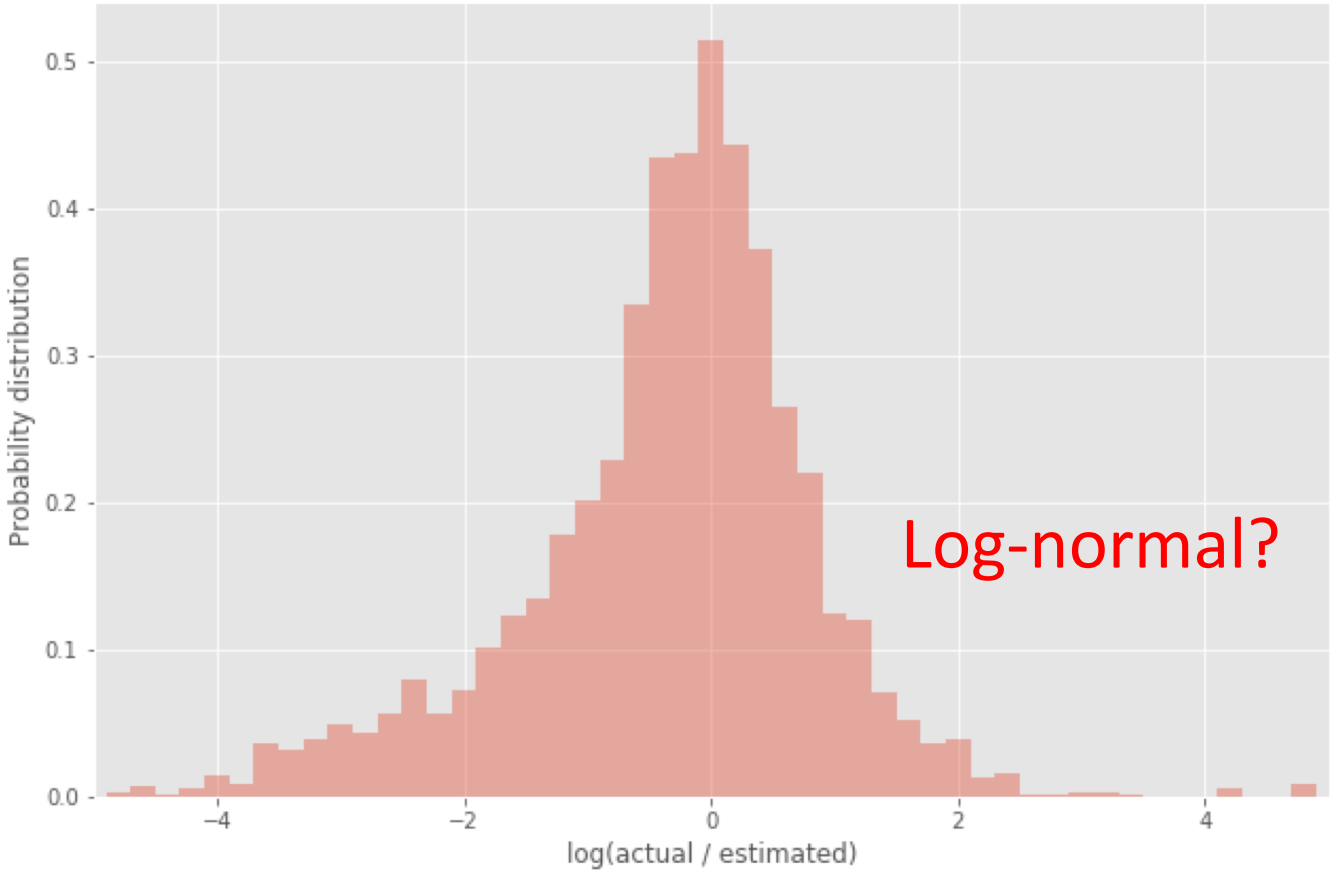
Actual data for blowup factor in software development projects:

- median = 1
- mean = 1.8

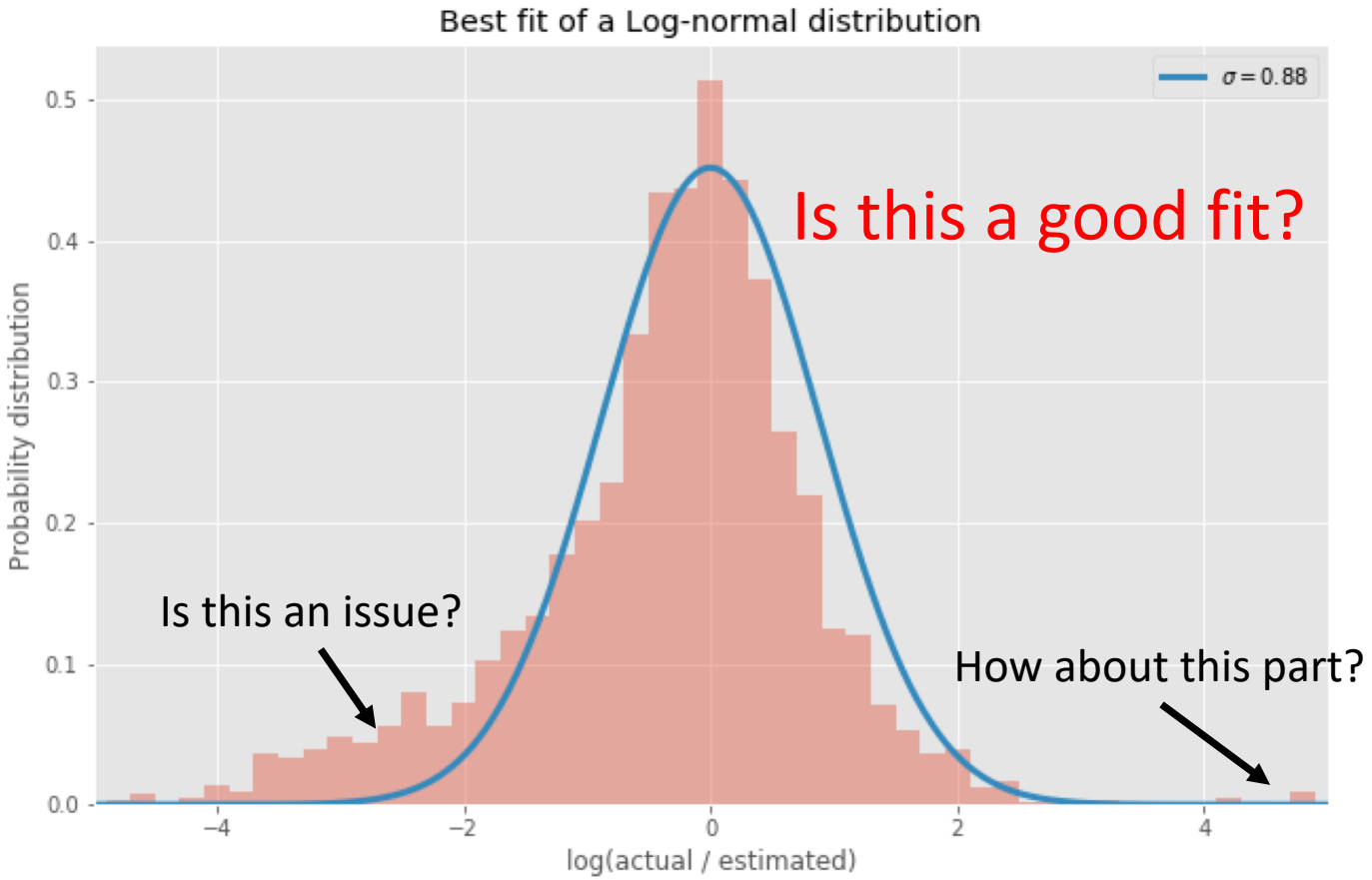
Maybe our hypothesis is not that bad?



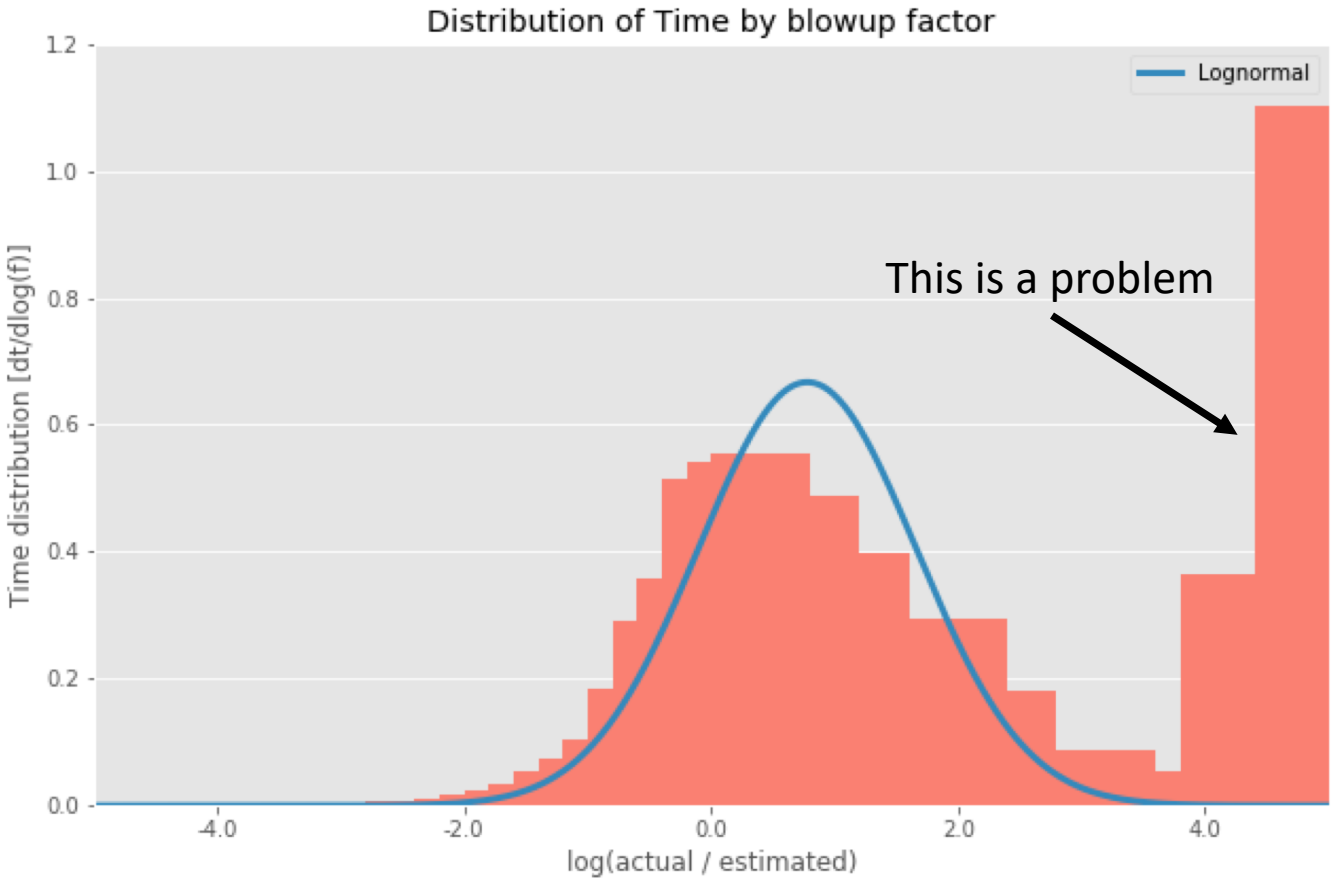
Distribution of log blowup factor



Is log-normal a good fit?



Is log-normal a good fit?



Fitting an analytic PDF

Let's fit an analytic PDF to the data!

But which one?

Hypothesis: Each project has an *f* blowup value that is lognormal with unit median and σ variance

Data: Mixture of different projects, with some distribution over σ

Data distribution = compound distribution of f_σ and σ

PDF of data $p(f)$ = $\int$ $p(f|\sigma)$ $p_\sigma(\sigma)d\sigma$ **PDF of σ among projects**

Blowup factor for each project (lognormal)

Fitting an analytic PDF

We have the freedom to choose $p_\sigma(\sigma)$

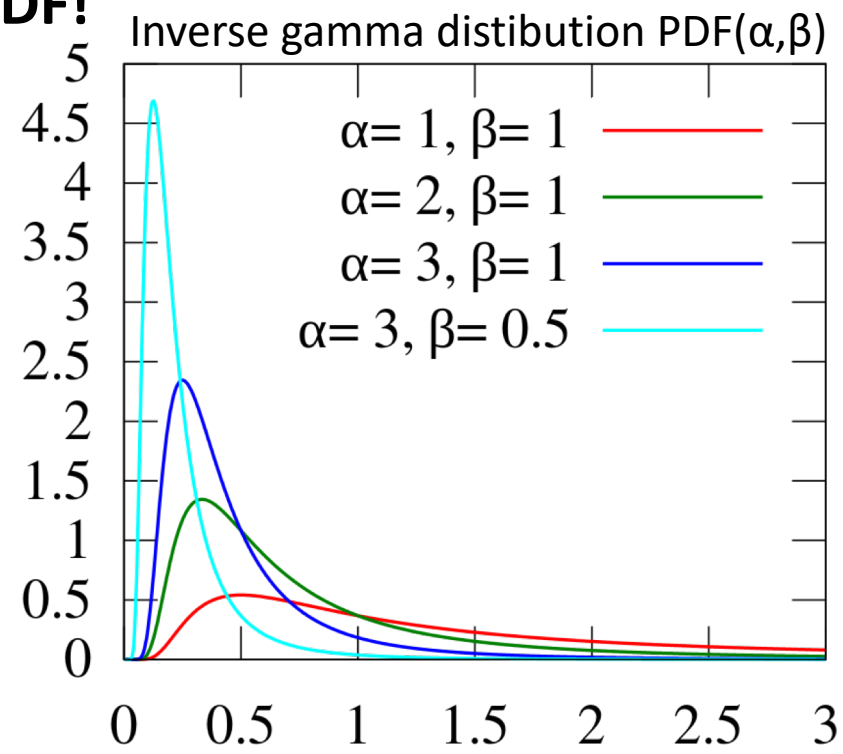
Let's choose the **inverse gamma PDF!**

Advantages:

- Positive values
- Flexible shape (2 parameters)
- **Gives an analytical PDF if compounded with a Gaussian**

⇩

Generalized Student t-distribution

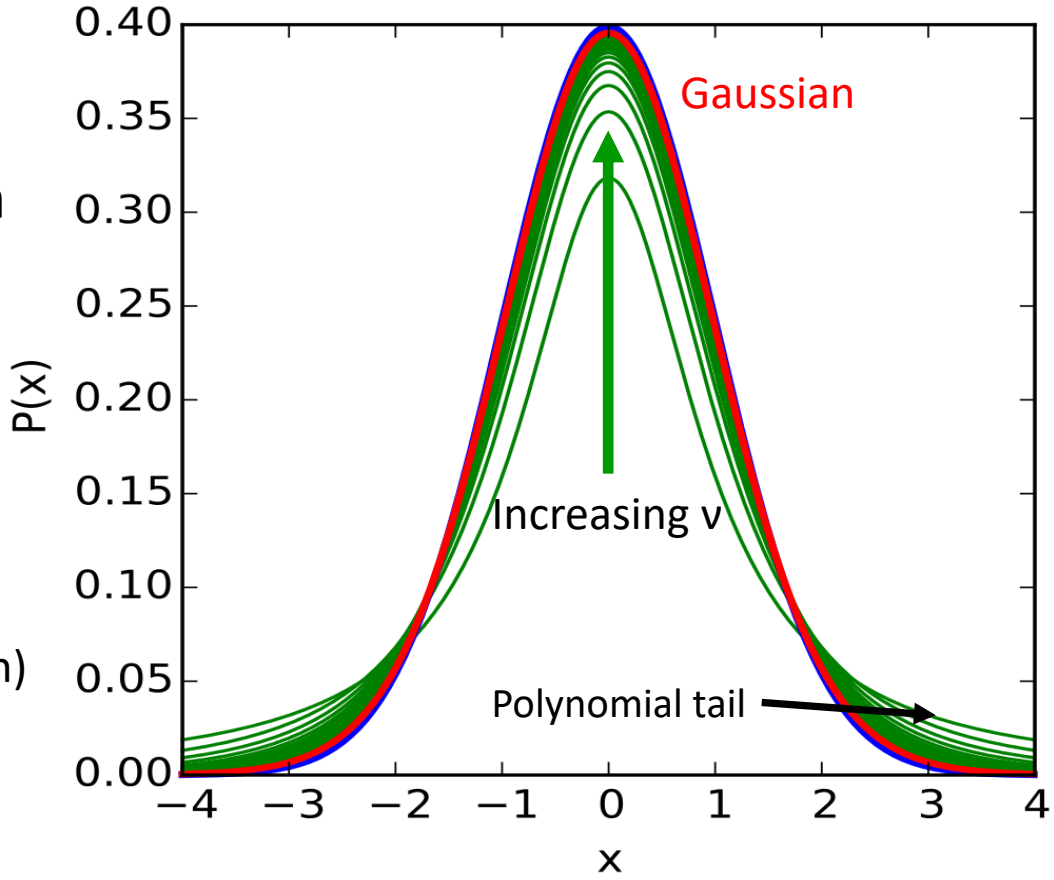


Generalized Student-t distribution

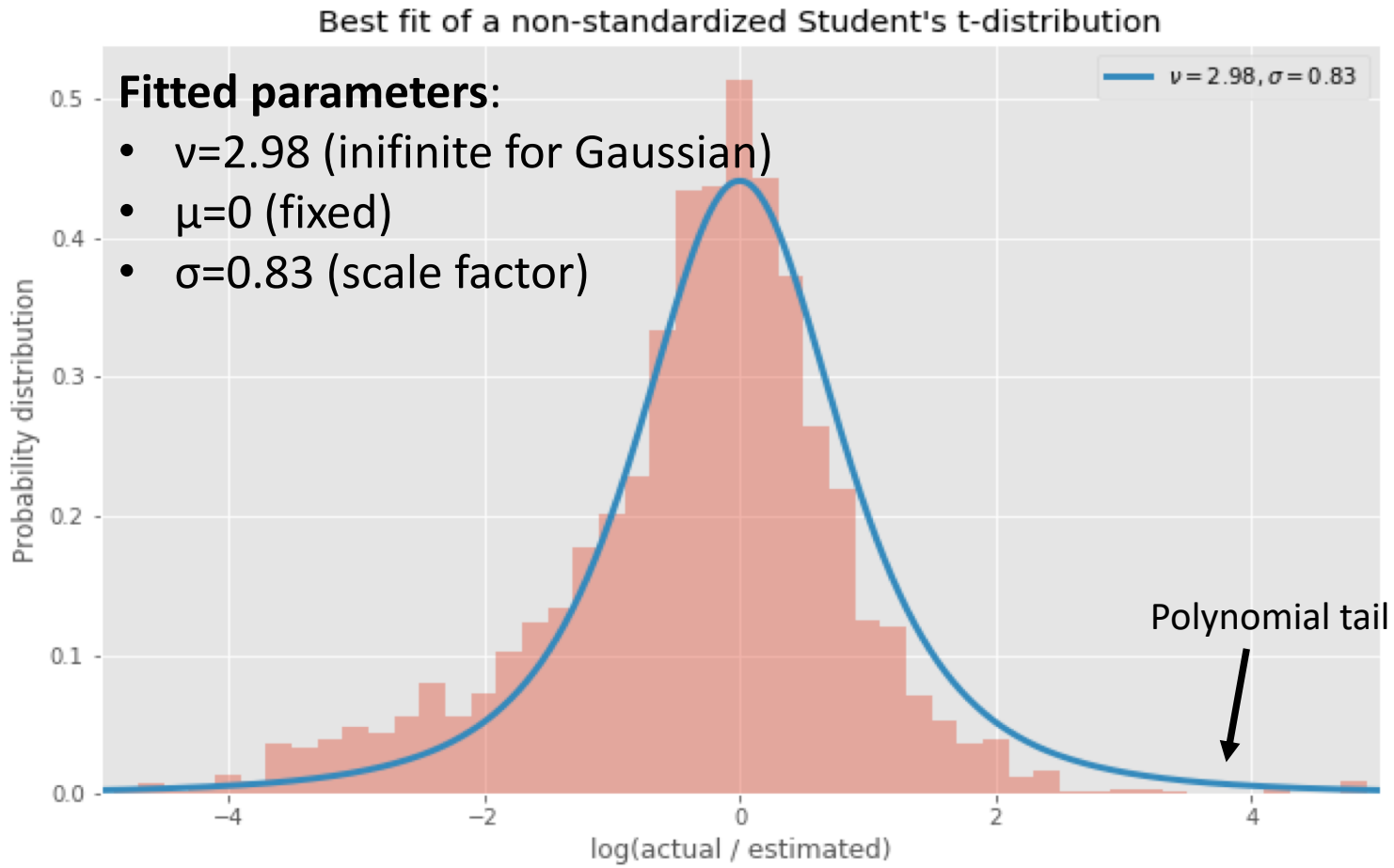
Arises when calculating the mean of a finite samples from a Gaussian distribution

Parameters:

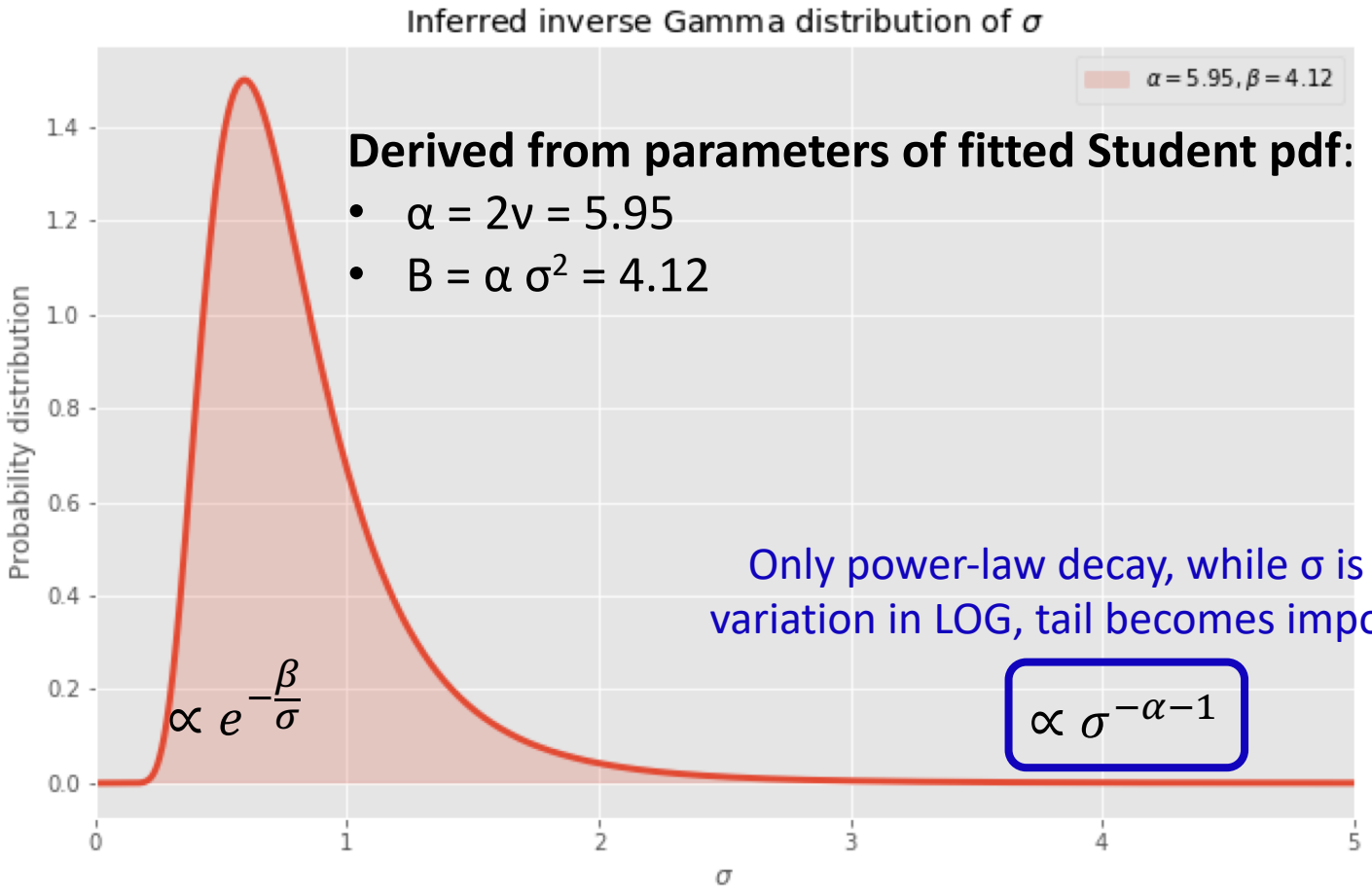
- ν = degrees of freedom
- μ = location parameter
- σ = scale factor (not variation)



Fitting an analytic PDF



Distribution of σ for data

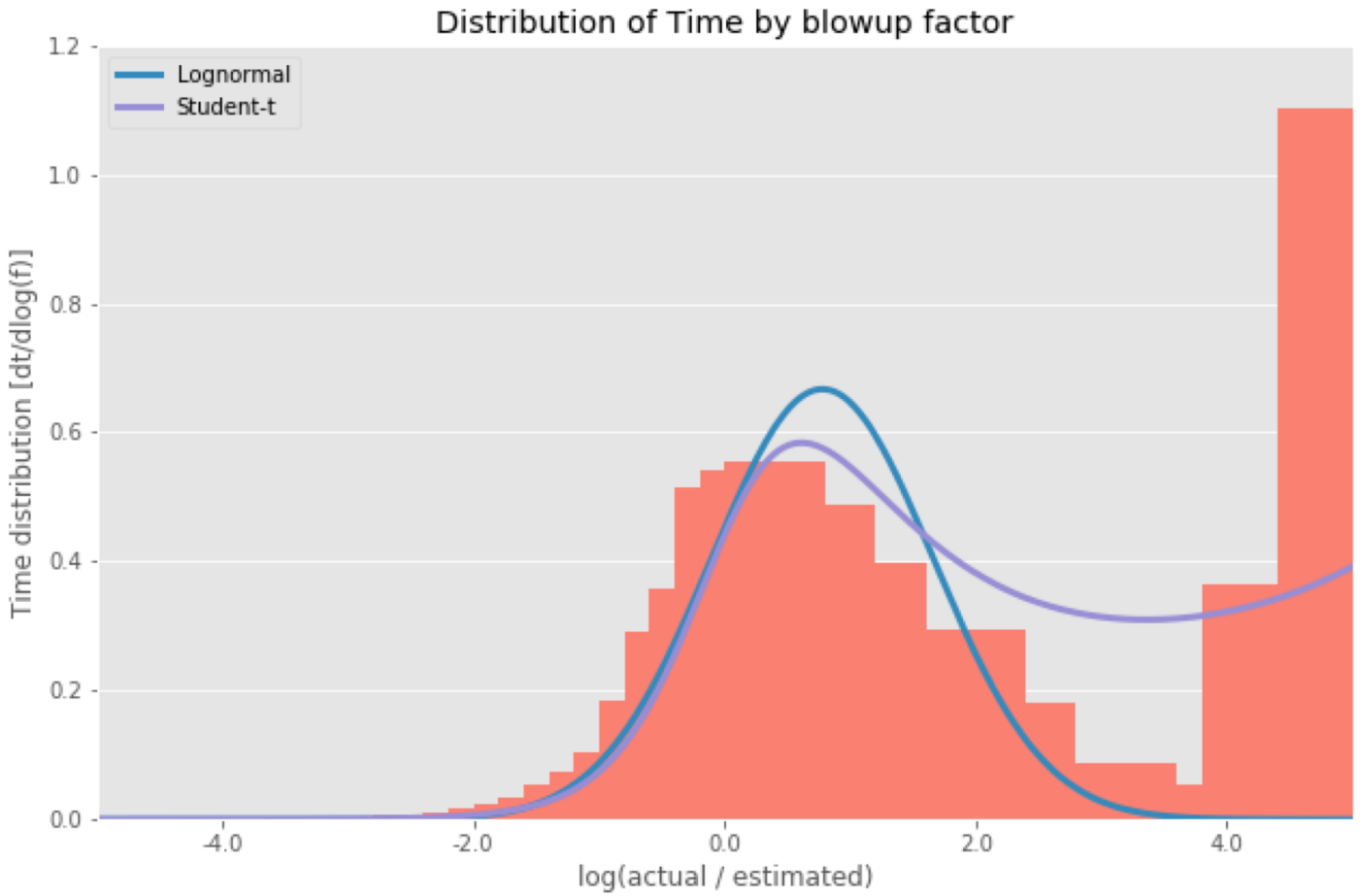


Derived from parameters of fitted Student pdf:

- $\alpha = 2\nu = 5.95$
- $B = \alpha \sigma^2 = 4.12$

Only power-law decay, while σ is the variation in LOG, tail becomes important

Distribution of project lengths



Derived quantities from fit to data

Median value	**1**
90th percentile	3.92
99th percentile	44
Mean value	**Infinite!**

Polynomially rare events require exponentially longer time, which leads to an infinite expected value

⇩

The mean time to solve a completely unknown project is infinite.
No wonder we can't keep the deadlines...

Summary

Robust results from data

- People actually estimate the **median** time a task requires, not the mean
- For skewed distributions the **medians don't add up**, unlike the means, leading to people underestimating the total time needed
- High blowup-factor projects dominate the project time length ditribution

Modeling results

- Hypothesis: Blowup factor of projects is well approximated by a lognormal distribution
- Blowup-factor well characterzed by a Student-t distribution
- Probability of high f blowup-factor events decay as polynomial of $\log f$, so these dominate the mean

Best practice: to estimate a project's length always look for the part that can potentially take the longest