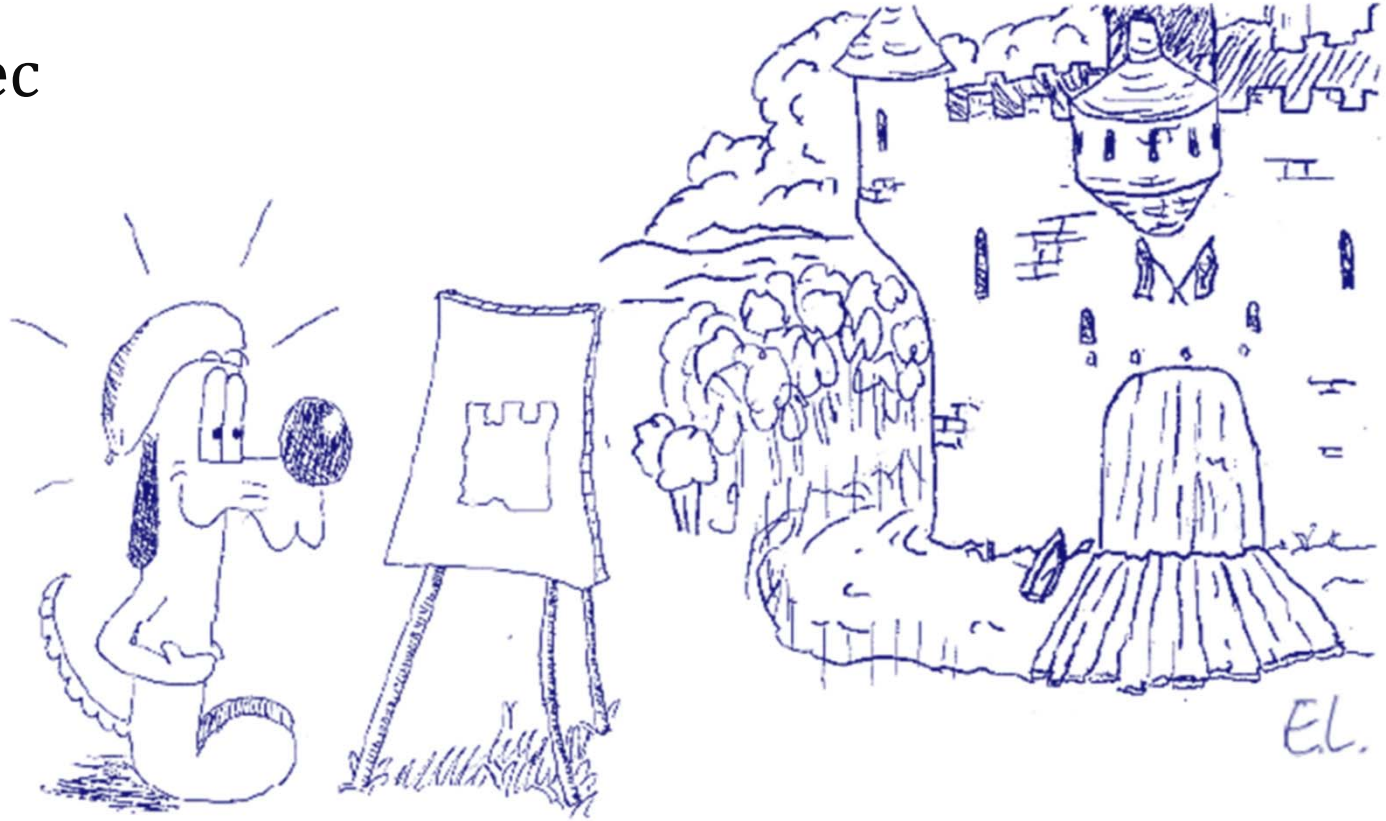
Model Fitting Bonus

JY Le Boudec

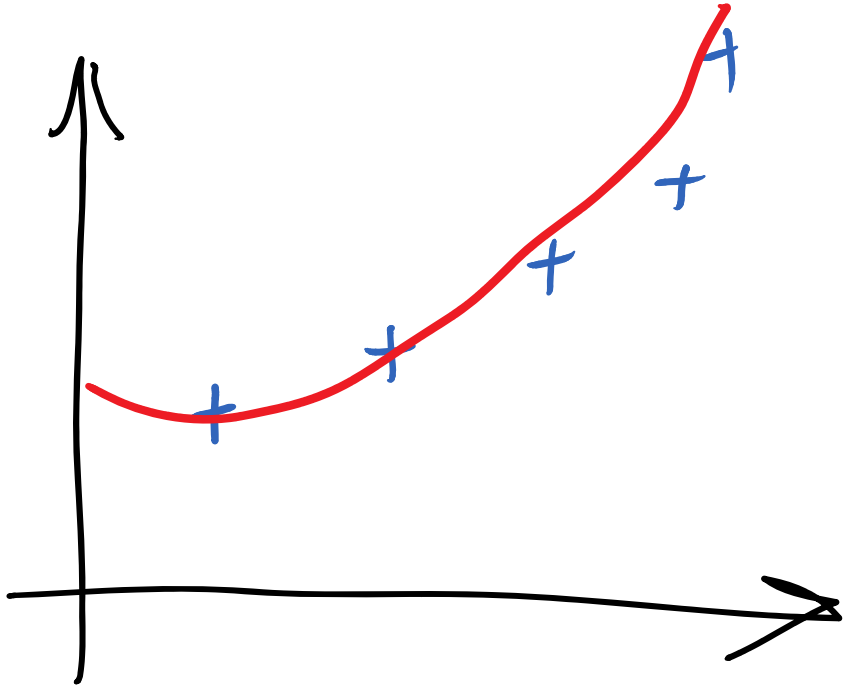
May 2015



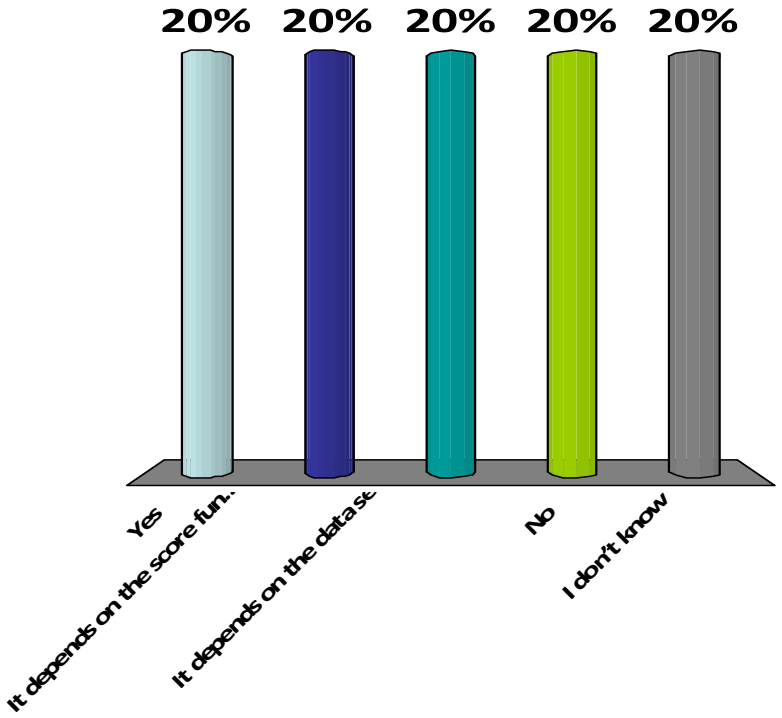
We want to fit a data set y_i to a polynomial of degree 2:

$$y_i = at_i^2 + bt_i + c.$$

Is this a linear regression model ?



A. Yes
B. It depends on the score function
C. It depends on the data set
D. No
E. I don't know

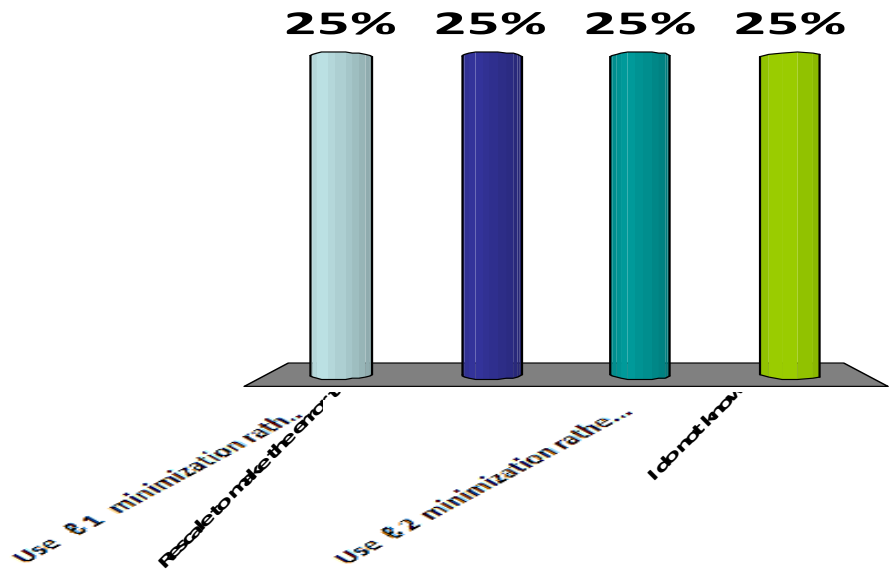


Solution

Linear regression means linear with respect to the parameters of the model other than the noise terms. Here the parameters are a, b, c and the model depends linearly on them. Answer A.

If the error terms in a fitting model are not homoscedastic, it is better to …

A. Use ℓ^1 minimization rather than ℓ^2
B. Rescale to make the error term homoscedastic
C. Use ℓ^2 minimization rather than ℓ^1
D. I do not know



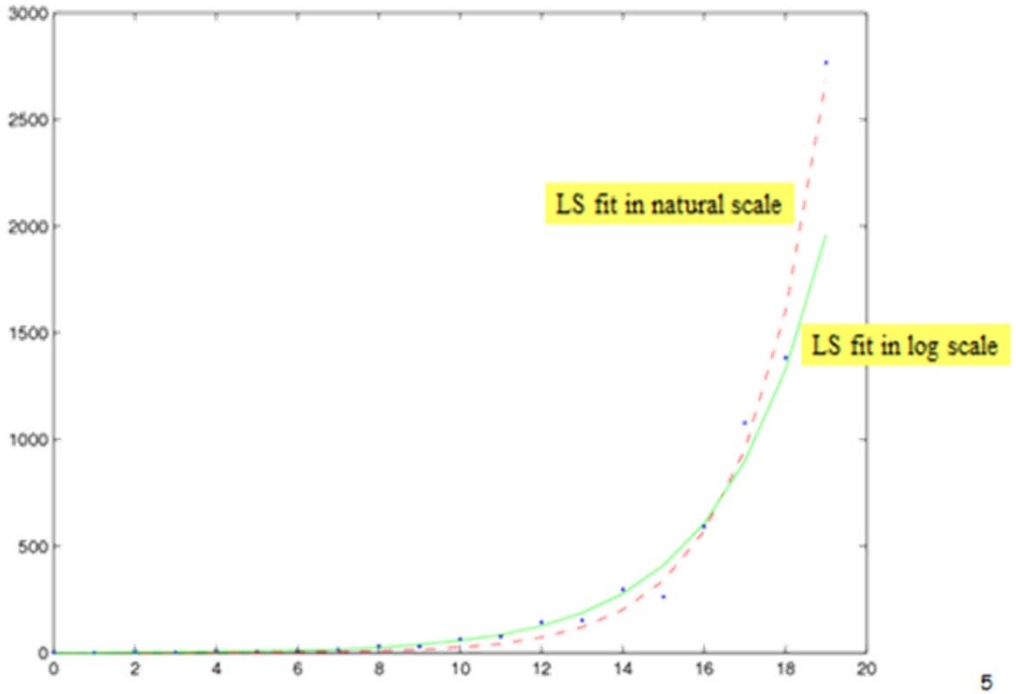
Solution

If the error terms are not homoscedastic, it is better to change the model, for example by rescaling, or to use weights in the ℓ^1 or ℓ^2 scores. Changing from ℓ^1 to ℓ^2 or vice-versa does not solve the problem.

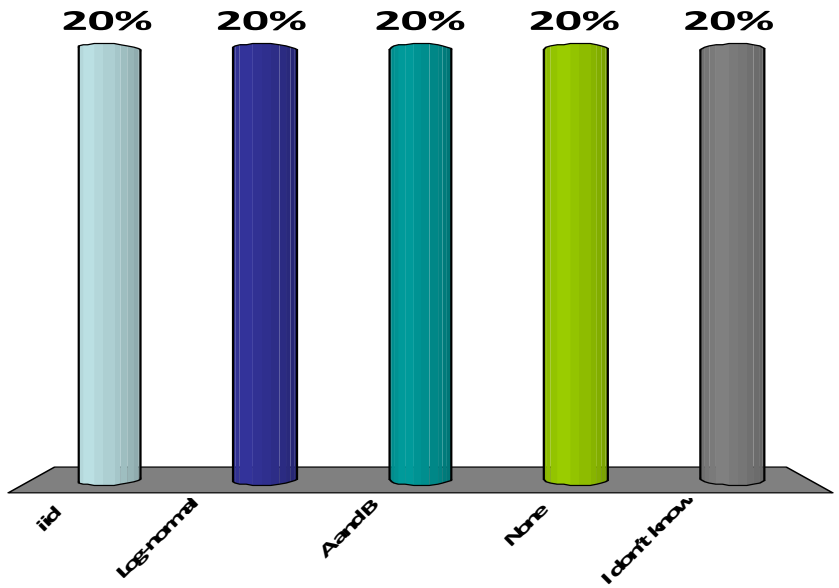
Answer B.

The green estimation corresponds to assuming that the error terms (blue dot – green curve) are …

A. iid
B. Log-normal
C. A and B
D. None
E. I don't know



5



Solution → $\log a + \alpha t_i$

$$\log y_i = \log f_i(\beta) + \varepsilon_i$$

$$\varepsilon_i \sim \text{iid } N(0, \sigma^2)$$

$$y_i = f_i(\beta)\, e^{\varepsilon_i}$$

error terms: $y_i - f_i(\beta) =$

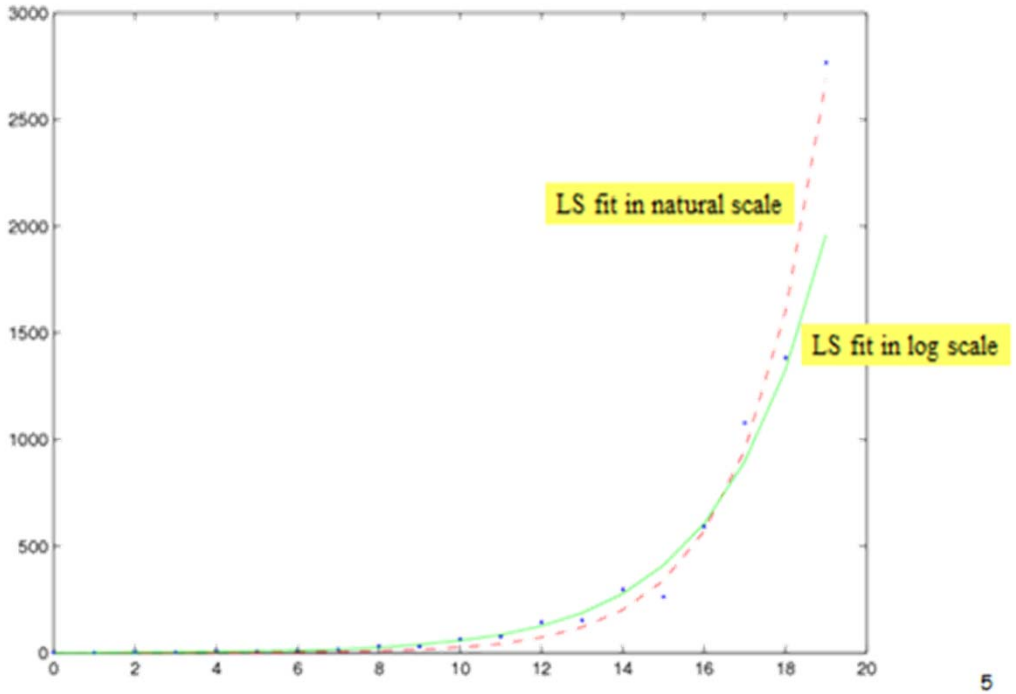
$f_i(\beta)(e^{\varepsilon_i} - 1)$ log normal (shifted)

not iid

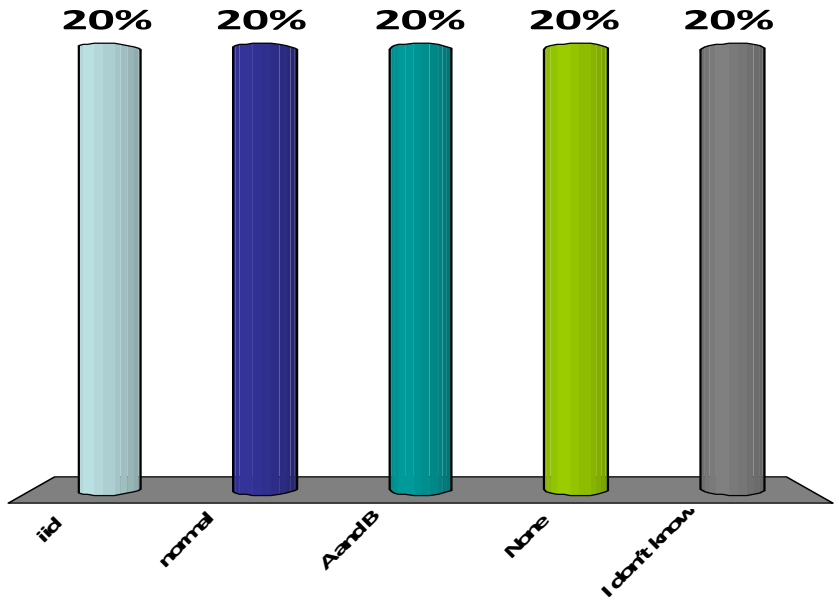
Answer B

The green estimation corresponds to assuming that the *relative* error terms (blue dot – green curve) are …

A. iid
B. normal
C. A and B
D. None
E. I don't know



5



Solution relative errors: $\frac{y_i - f_i(\beta)}{f_i(\beta)} \overset{\text{def}}{=} r_i$

$$r_i = \frac{f_i(\beta)\, e^{\varepsilon_i} - f_i(\beta)}{f_i(\beta)} = e^{\varepsilon_i} - 1$$

if σ^2 small, ε_i small and

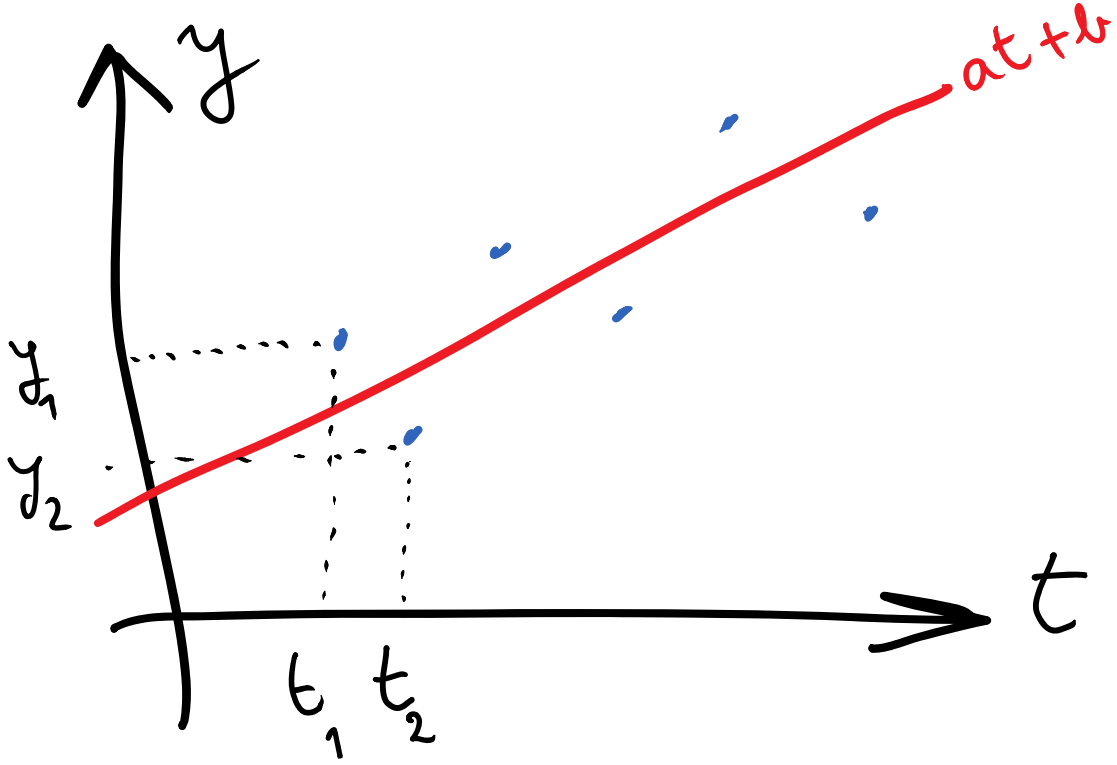
$$e^{\varepsilon_i} - 1 \simeq 1 + \varepsilon_i - 1 = \varepsilon_i$$

iid, almost normal

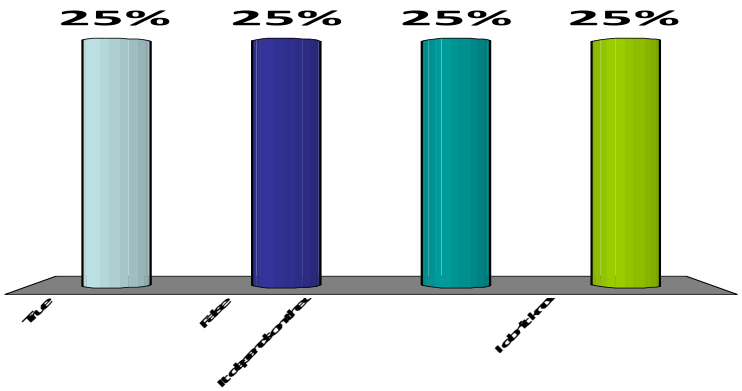
Answer C is (approximately) correct

We fit the model $y_i = at_i + b$ using least squares.

The obtained line is such that the average distance from the points to the line is 0.



A. True
B. False
C. It depends on the data
D. I don't know



Solution $(\hat{a}, \hat{b})$ = fitted parameters.

$(\hat{a}, \hat{b})$ minimizes $\sum_i [y_i - (at_i + b)]^2$

Let $y_i - \hat{a}t_i \overset{\text{def}}{=} x_i$

$\hat{b}$ minimizes $\sum_i [y_i - (\hat{a}t_i + b)]^2$

$\hat{b}$ minimizes $\sum_i (x_i - b)^2$

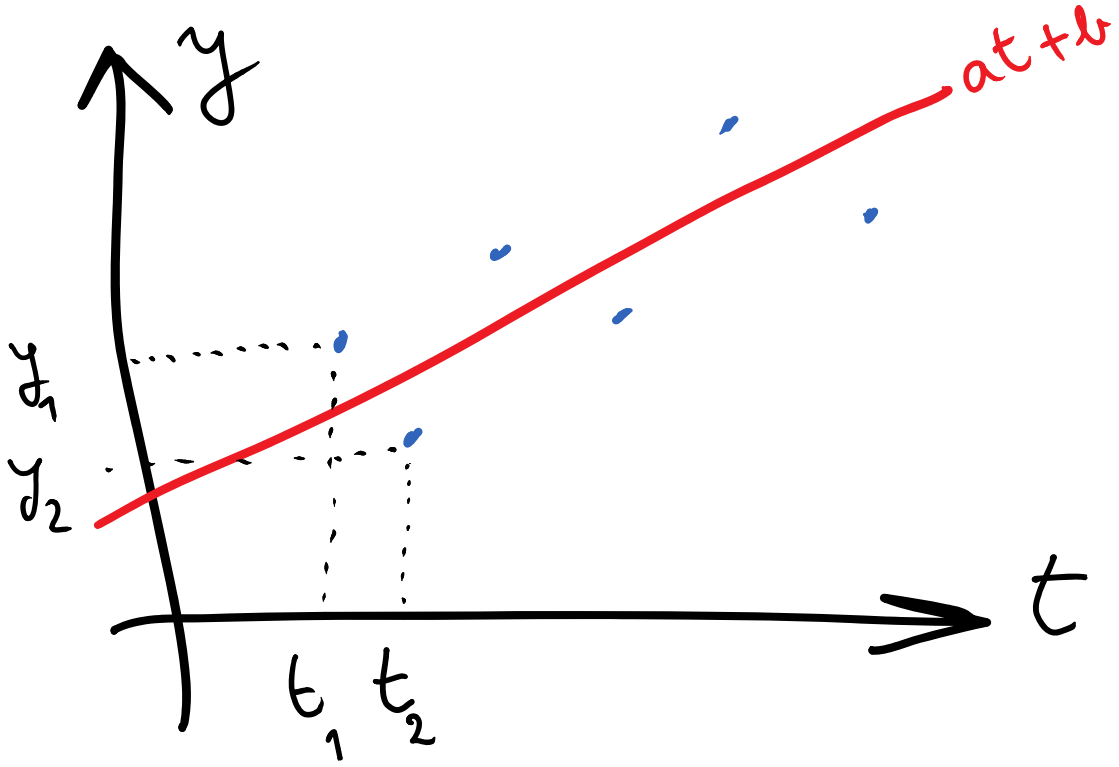
thus $\hat{b}$ = mean of $x_i = \bar{x}$

thus average of $y_i - at_i - b$ is 0

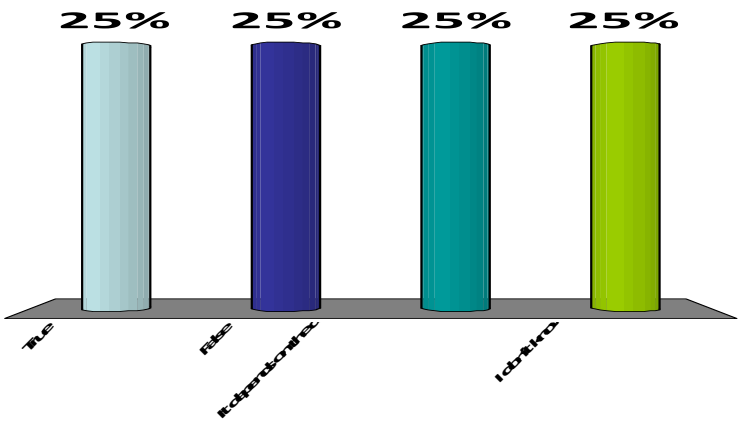
Answer A

We fit the model $y_i = at_i + b$ using ℓ^1 norm minimization.

The obtained line leaves an equal number of points on each side



A. True
B. False
C. It depends on the data
D. I don't know



Solution $(\hat{a}, \hat{b})$ = fitted parameters.

$(\hat{a}, \hat{b})$ minimizes $\sum_i |y_i - (at_i + b)|$

Let $y_i - \hat{a}t_i \stackrel{\text{def}}{=} x_i$

$\hat{b}$ minimizes $\sum_i |y_i - (\hat{a}t_i + b)|$

b minimizes $\sum_i |x_i - b|$

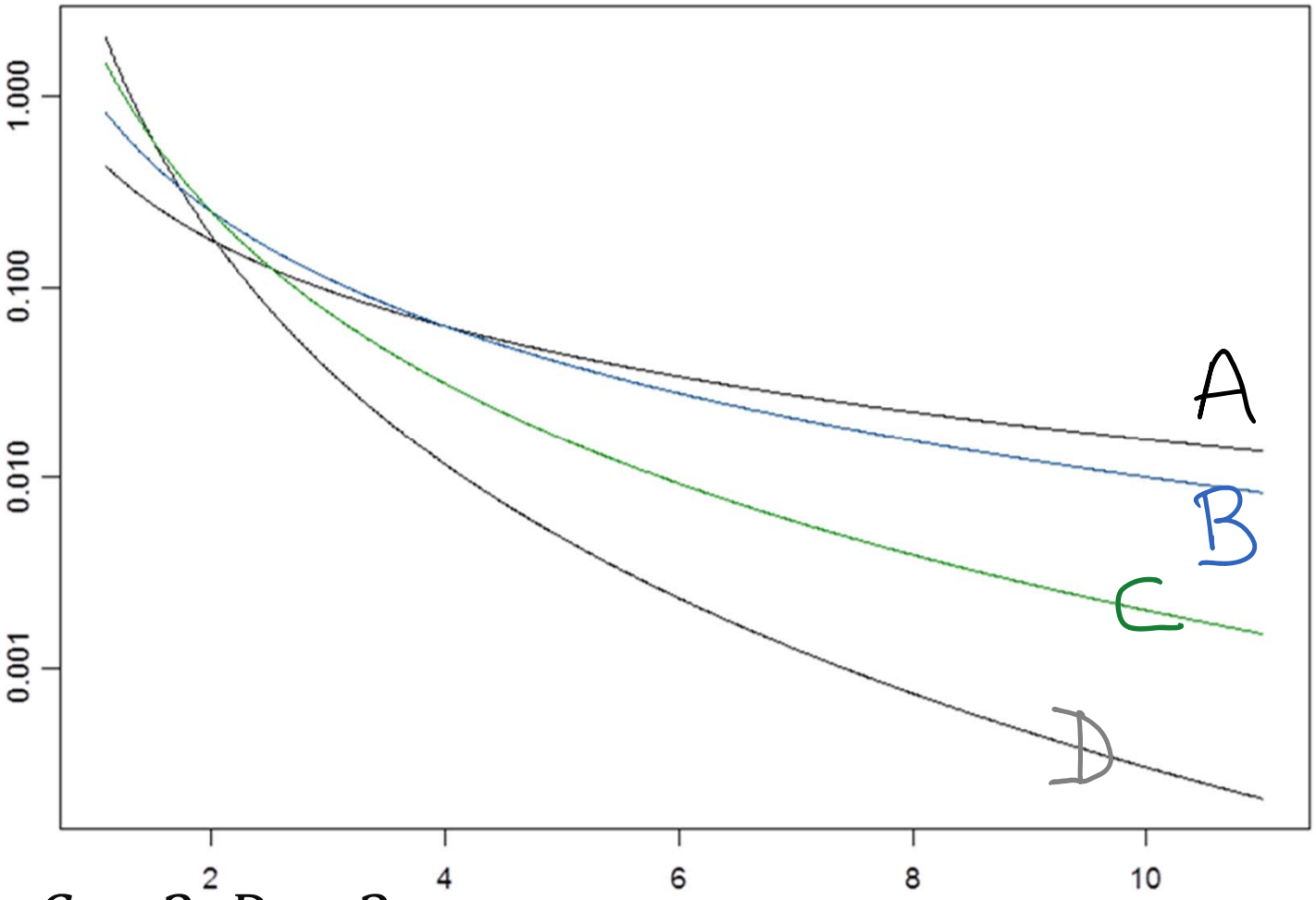
thus $\hat{b}$ = median of x_i

equal number of points on each side

Answer A

Find the parameter p for each of these standard Pareto PDFs

$$f(x) = \frac{p}{x^{p+1}} \mathbf{1}_{x>1}$$



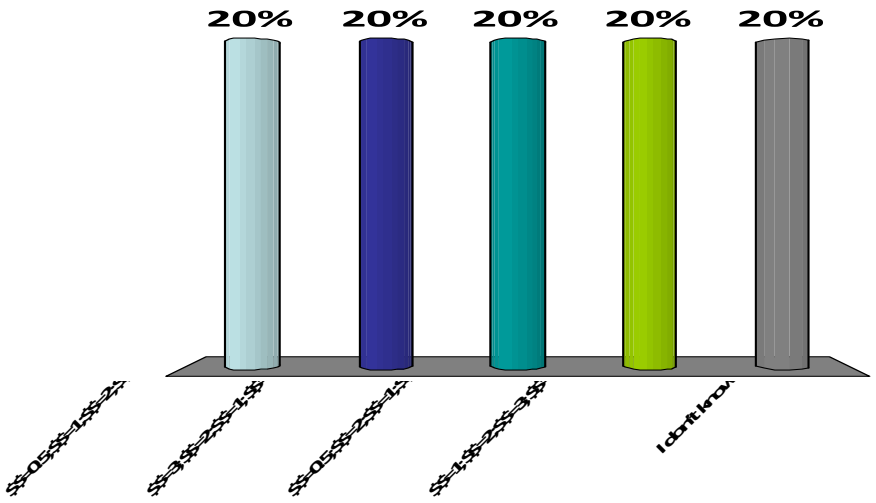
A. $A = 0.5; B = 1; C = 2; D = 3$

B. $A = 3; B = 2; C = 1; D = 0.5$

C. $A = 0.5; B = 2; C = 1; D = 3$

D. $A = 1; B = 2; C = 3; D = 0.5$

E. I don't know



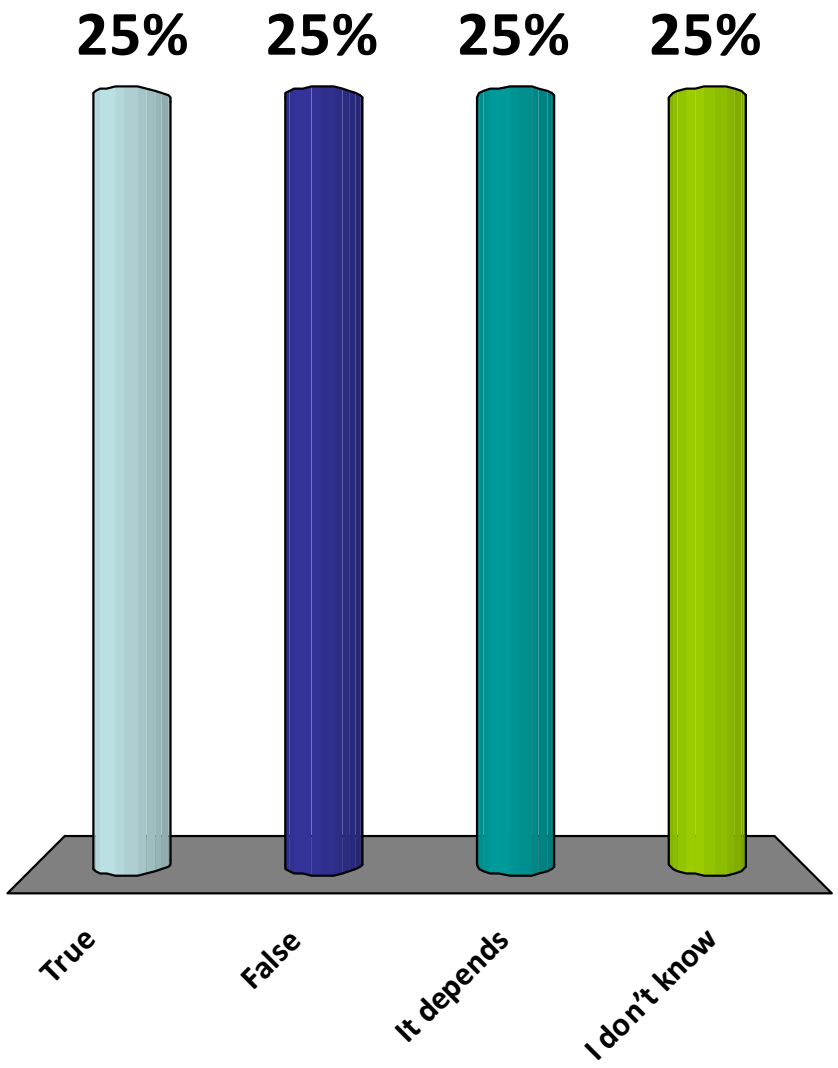
Solution

The decay is faster when the exponent is larger. Answer A.

Note that we have a semi-log plot. In a log-log plot we would have straight lines.

If a positive random variable has a finite mean and is heavy tailed, its variance is infinite

A. True
B. False
C. It depends
D. I don't know

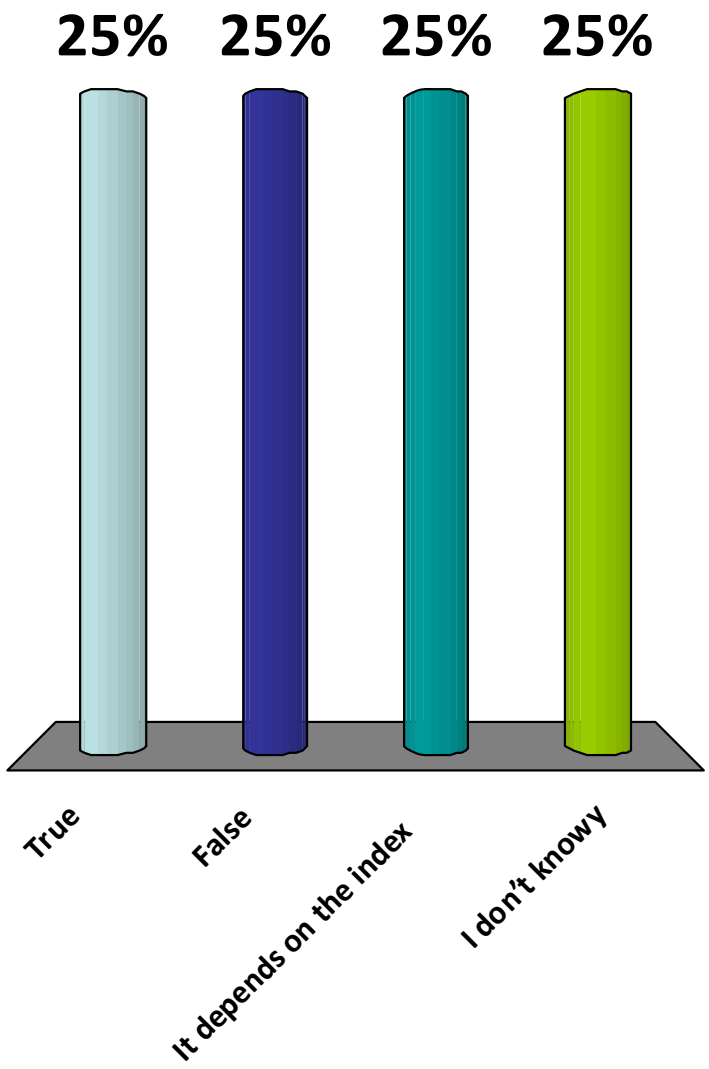


Solution

Answer A. Heavy tail means 1-CDF decays like $\frac{1}{x^p}$ with $p < 2$. The pdf is in $1/x^{p+1}$ therefore the second moment, which is the integral of $x^2 f(x)$, gives an infinite integral in $+\infty$ since $x^2 f(x) \sim \frac{1}{x^{p-1}}$ and $p - 1 < 1$

The Complementray CDF of a Pareto distribution follows a power law…

A. True
B. False
C. It depends on the index p
D. I don't knowy



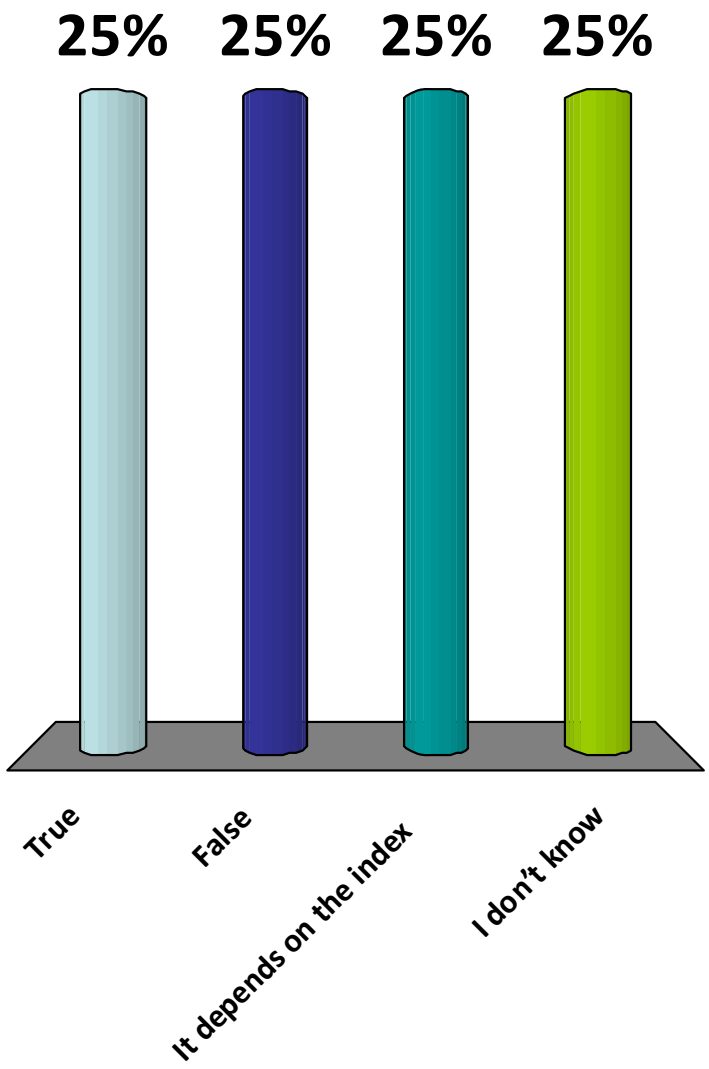
Solution

The complementary CDF is

1-CDF= $\frac{1}{x^p}$ for $x > 1$ and is a power law. Answer A

A Pareto distribution is heavy tailed …

A. True
B. False
C. It depends on the index p
D. I don't know

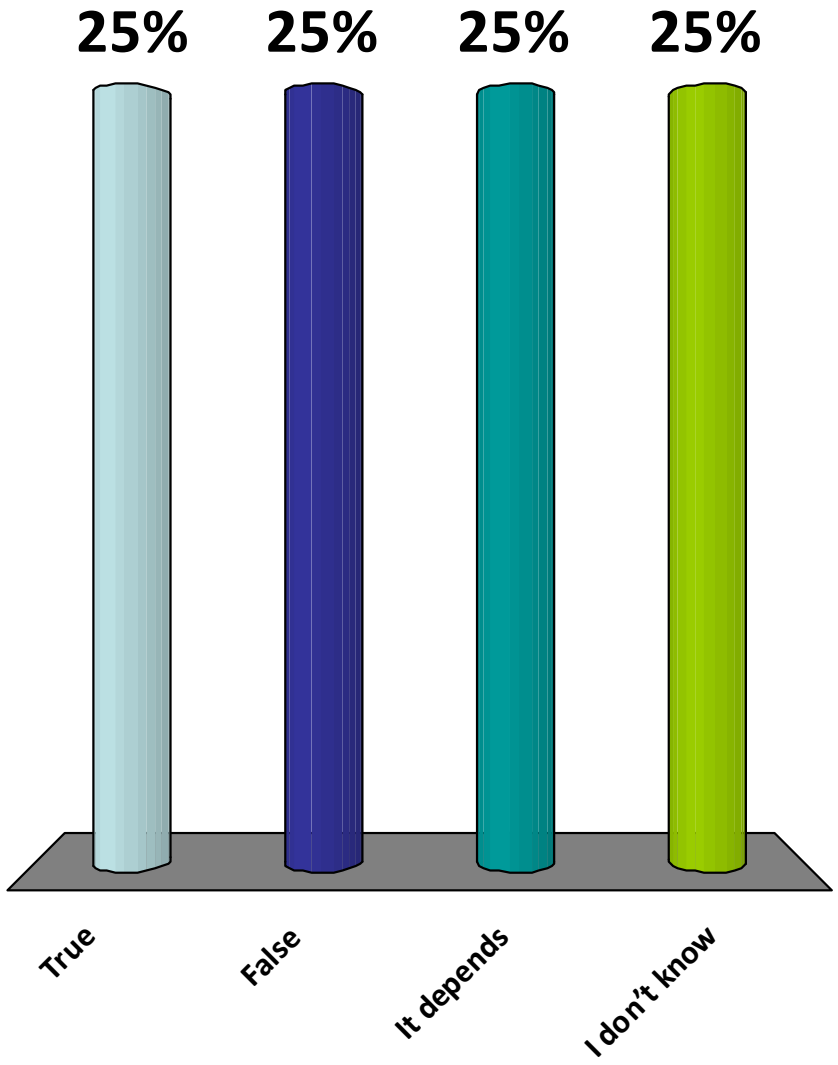


Solution

Answer C. The CCDF is $\frac{1}{x^p}$ and is heavy tailed when $p < 2$. For $p \geq 2$ the Pareto distribution is not heavy tailed.

For a Pareto distribution, the hazard rate $\lambda(t)$ is such that $\lim_{t \to \infty} \lambda(t) = 0$

A. True
B. False
C. It depends
D. I don't know



Solution

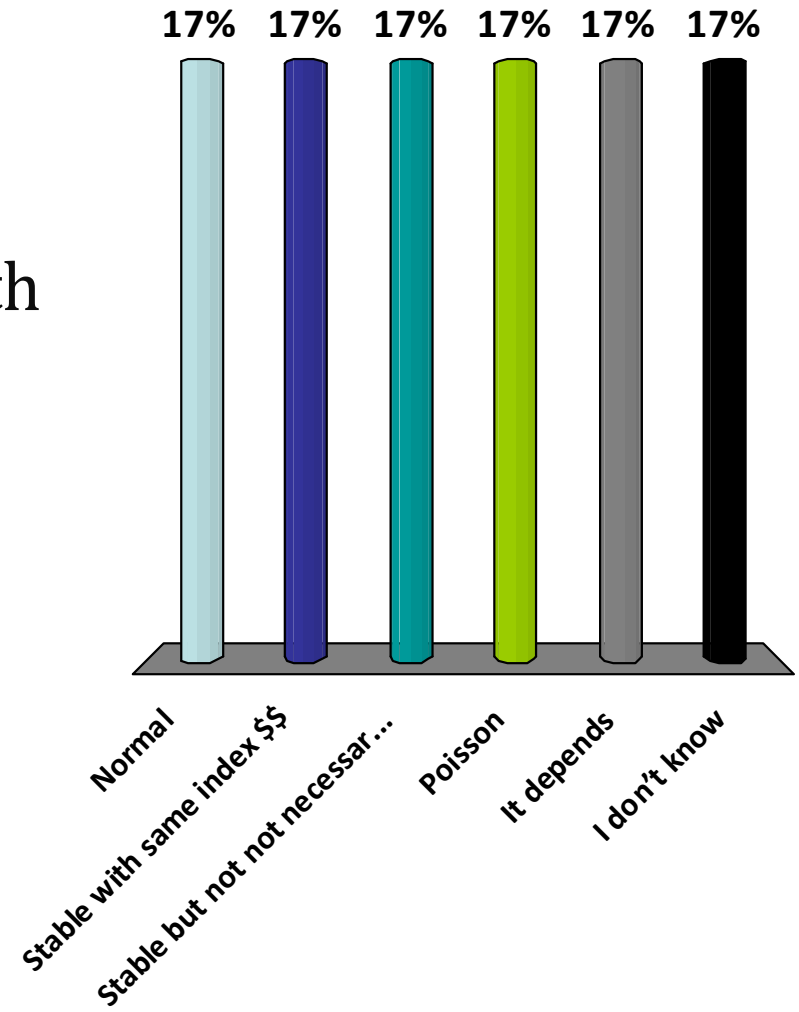
Answer A.

The hazard rate is $\lambda(x) = \frac{f(x)}{1-F(x)} = \frac{p}{x}$ and this holds for all p

The hazard rate vanishes for large x, i.e. Pareto(p) is fat-tailed for any value of p (but is heavy tailed only for $p < 2$)

The distribution of the sum of n iid random variables with heavy tail and index $p < 2$, for large n, is approximately…

A. Normal
B. Stable with same index p
C. Stable but not not necessarily with same index p
D. Poisson
E. It depends
F. I don't know

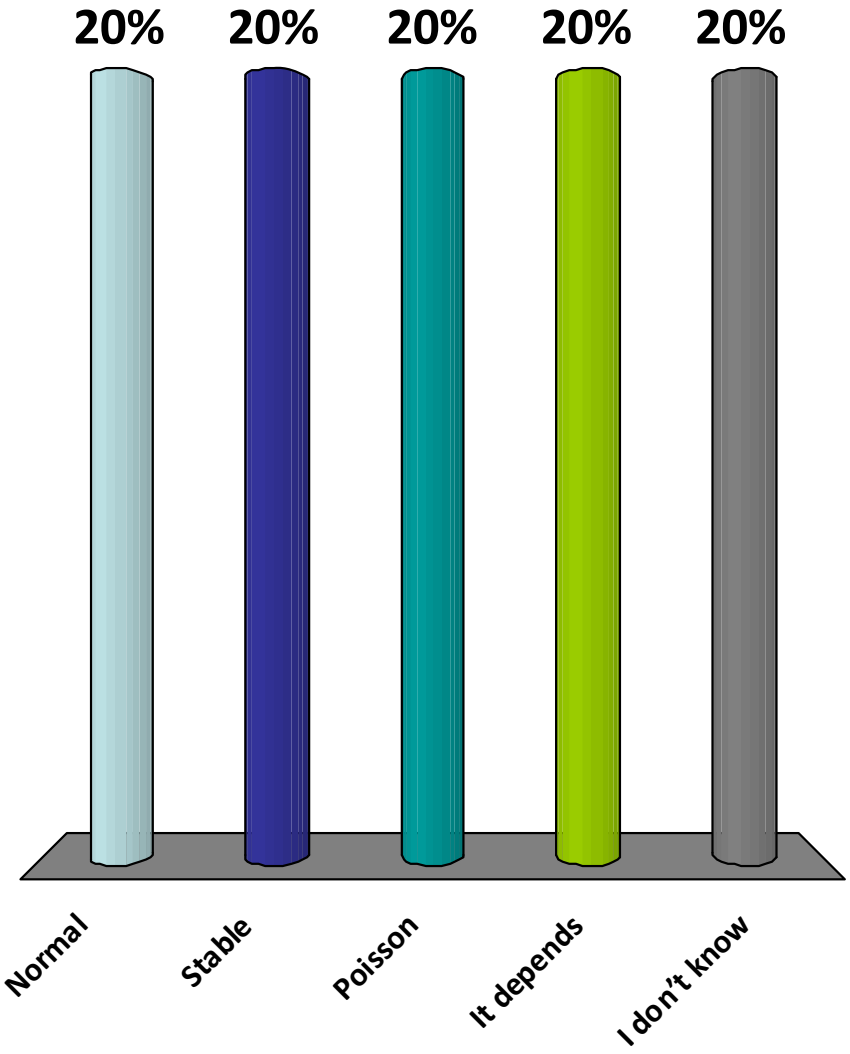


Solution

Answer B. The limit is not gaussian but is stable with same index.

The distribution of the sum of n iid random variables with finite variance, for large n, is approximately…

A. Normal
B. Stable
C. Poisson
D. It depends
E. I don't know



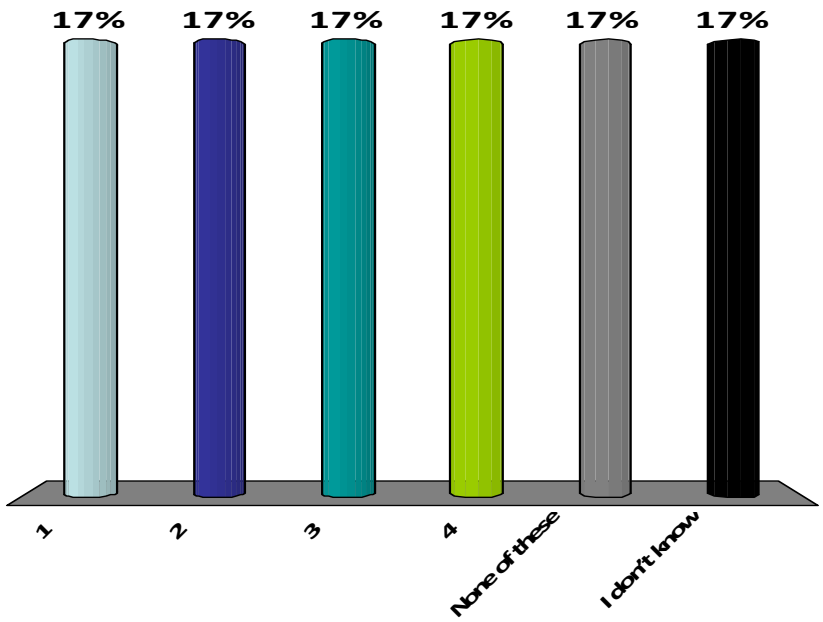
Solution

Answer A. The limit is gaussian (= normal).

We want to estimate some quantity μ. We have $m+n$ independent measurements $X_1, \dots, X_m \sim iid\ N(\mu, \sigma^2)$ and $Y_1, \dots, Y_n \sim iid\ N(\mu, \lambda^2\sigma^2)$ λ is known. We do weighted least square estimation of μ. What do we obtain ?

1. $\hat{\mu}_1 = \frac{X_1+\cdots+X_m+Y_1+\cdots+Y_n}{m+n}$
2. $\hat{\mu}_2 = \frac{X_1+\cdots+X_m+\lambda(Y_1+\cdots+Y_n)}{m+\lambda n}$
3. $\hat{\mu}_3 = \frac{X_1+\cdots+X_m+\frac{Y_1+\cdots+Y_n}{\lambda}}{m+\frac{n}{\lambda}}$
4. $\hat{\mu}_4 = \frac{X_1+\cdots+X_m+\frac{Y_1+\cdots+Y_n}{\lambda^2}}{m+\frac{n}{\lambda^2}}$

A. 1
B. 2
C. 3
D. 4
E. None of these
F. I don't know



Solution

The weighted least square fitting should minimize the score defined by

$$\sum_i (x_i - \mu)^2 + \sum_j \frac{(y_j - \mu)^2}{\lambda^2}$$

The derivative w.r. μ is

$$-2\sum_i (x_i - \mu) - 2\sum_j \frac{y_j - \mu}{\lambda^2}$$

The optimal μ is

$$\hat{\mu}_4 = \frac{\sum_i x_i + \sum_j \frac{y_j}{\lambda^2}}{m + \frac{n}{\lambda^2}} = \frac{m\bar{x} + \frac{n\bar{y}}{\lambda^2}}{m + \frac{n}{\lambda^2}}$$

Answer D

LS score functions

- Given a model to be fitted

$$Y_i = f_i(\theta) + \text{noise}$$

we have the equivalences:

Score	Statistical model
Least Square: score $= \sum_i (\text{noise}_i)^2$	noise $\sim$ iid $N(0, \sigma^2)$
Weighted Least Square: score $= \sum_i (w_i \times \text{noise}_i)^2$	$\text{noise}_i \sim$ ind $N\left(0, \frac{\sigma^2}{w_i^2}\right)$

i.e. WLS gives a weight inversely proportional to σ_i of noise

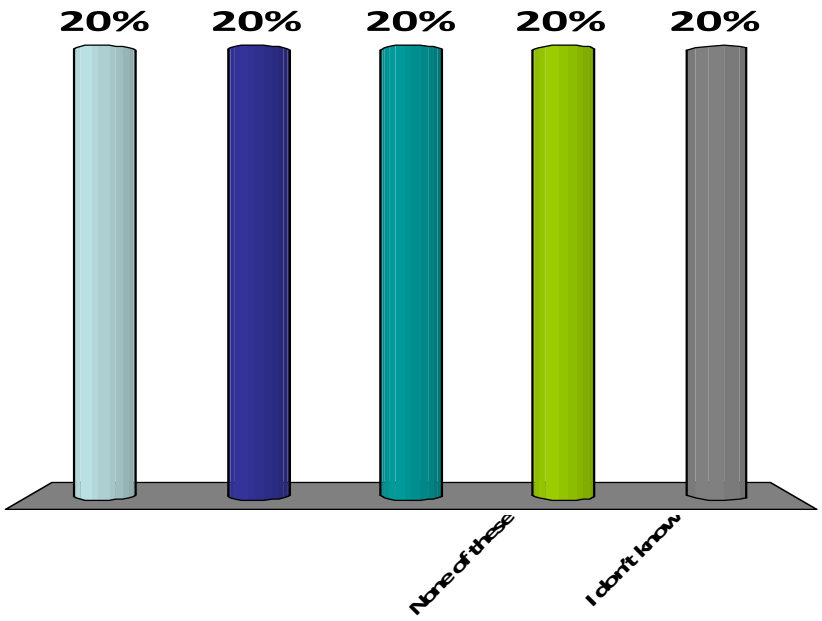
We want to estimate some quantity μ. We have $m + n$ independent measurements $X_1, \dots, X_m \sim iid\ N(\mu, \sigma^2)$ and $Y_1, \dots, Y_n \sim iid\ N(\mu, \lambda^2\sigma^2)$

σ and λ are unknown but we think that $\lambda >> 1$. i.e. Y is very noisy; $m \approx n$

To estimate μ, say which formula you prefer:

1. $\hat{\mu}_1 = \frac{X_1 + \cdots + X_m}{m}$
2. $\hat{\mu}_2 = \frac{Y_1 + \cdots + Y_n}{n}$
3. $\hat{\mu}_3 = \frac{X_1 + \cdots + X_m + Y_1 + \cdots + Y_n}{m+n}$

1. 1
2. 2
3. 3
4. None of these
5. I don't know



Solution

None of these estimates is the maximum likelihood estimator, seen in the previous question. Since we can only choose from these 3, let us see how good they are.

First note that all 3 estimators have expectation equal to μ, so we can compare them by comparing their variances.

$$\mathrm{var}(\hat{\mu}_1) = \frac{\sigma^2}{m}\ \mathrm{var}(\hat{\mu}_3)$$

$$\mathrm{var}(\hat{\mu}_2) = \frac{\lambda^2\sigma^2}{n} \gg \mathrm{var}(\hat{\mu}_1)$$

$$\mathrm{var}(\hat{\mu}_3) = \frac{\sigma^2(m+\lambda^2 n)}{(m+n)^2} \approx \frac{\sigma^2}{m}\frac{\lambda^2}{4} \gg \mathrm{var}(\hat{\mu}_1)$$

$\hat{\mu}_1$ is the best: it is better to ignore the measurements y_j than to use formula $\hat{\mu}_3$ which does the average. More is less !
Answer 1

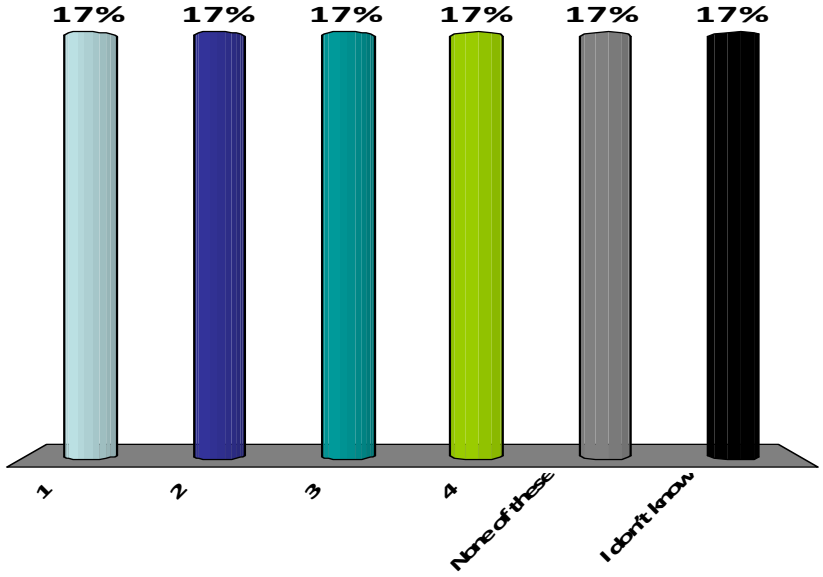
We want to estimate some quantity μ. We have $m + n$ independent measurements $X_1, \ldots, X_m \sim iid\ N(\mu, \sigma^2)$ and $Y_1, \ldots, Y_n \sim iid\ N(\mu, \lambda^2\sigma^2)$

σ and λ are unknown but we think that $\lambda >> 1$.

To estimate μ, say which formula you prefer:

1. $\hat{\mu}_1 = \frac{X_1 + \cdots + X_m}{m}$
2. $\hat{\mu}_2 = \frac{Y_1 + \cdots + Y_n}{n}$
3. $\hat{\mu}_3 = \frac{X_1 + \cdots + X_m + Y_1 + \cdots + Y_n}{m+n}$
4. $\hat{\mu}_4 = \frac{X_1 + \cdots + X_m + \frac{Y_1 + \cdots + Y_n}{\lambda^2}}{m + \frac{n}{\lambda^2}}$

A. 1
B. 2
C. 3
D. 4
E. None of these
F. I don't know



Solution

The only question is whether 4 is better than 1.

The expectation of $\hat{\mu}_4$ is μ like for the other 3 and its variance is

$$\mathrm{var}(\hat{\mu}_4) = \mathrm{var}(\hat{\mu}_4) = \frac{\sigma^2}{m + \frac{n}{\lambda^2}} < \mathrm{var}(\hat{\mu}_1)$$

Answer D : $\hat{\mu}_4$ is the best estimator

This is not surprising as it is the maximum likelihood estimator and such estimators are asymptotically optimal.

Take-home message: adding some very noisy measurements may lead to worst estimation, unless you compensate for the increased noise by appropriately weighting the estimator !