What does the distribution of change intervals look like?

How many pages….

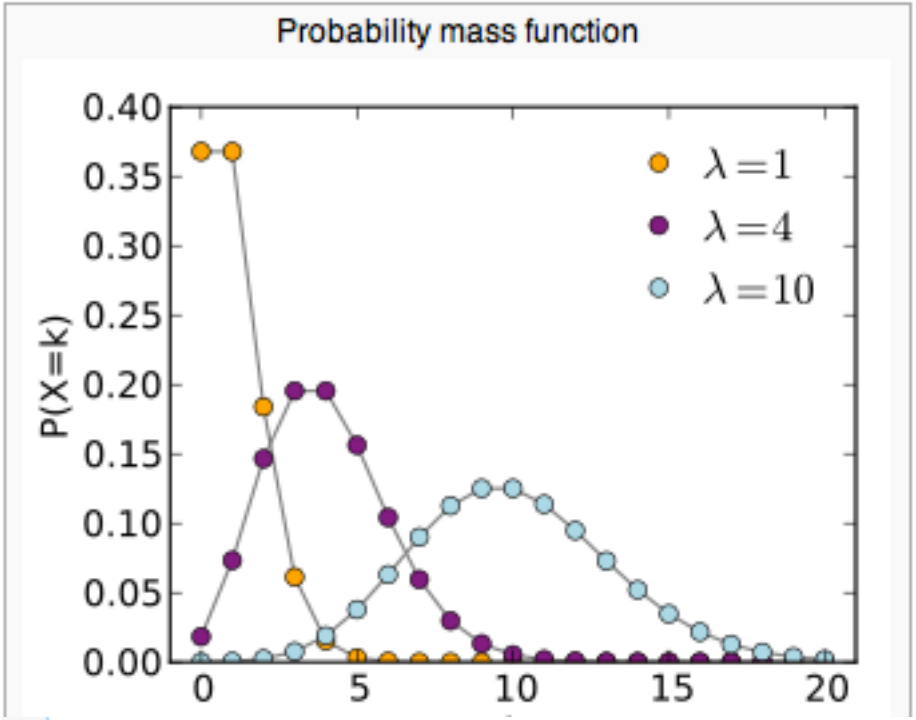
- Change regularly? Every day, every week…
- Don’t change at all?
- Change “randomly”?

Make a graph of the distribution of “changes” …

How do we compare “fast” vs “slow” pages?

Break to look at data

Poisson Process



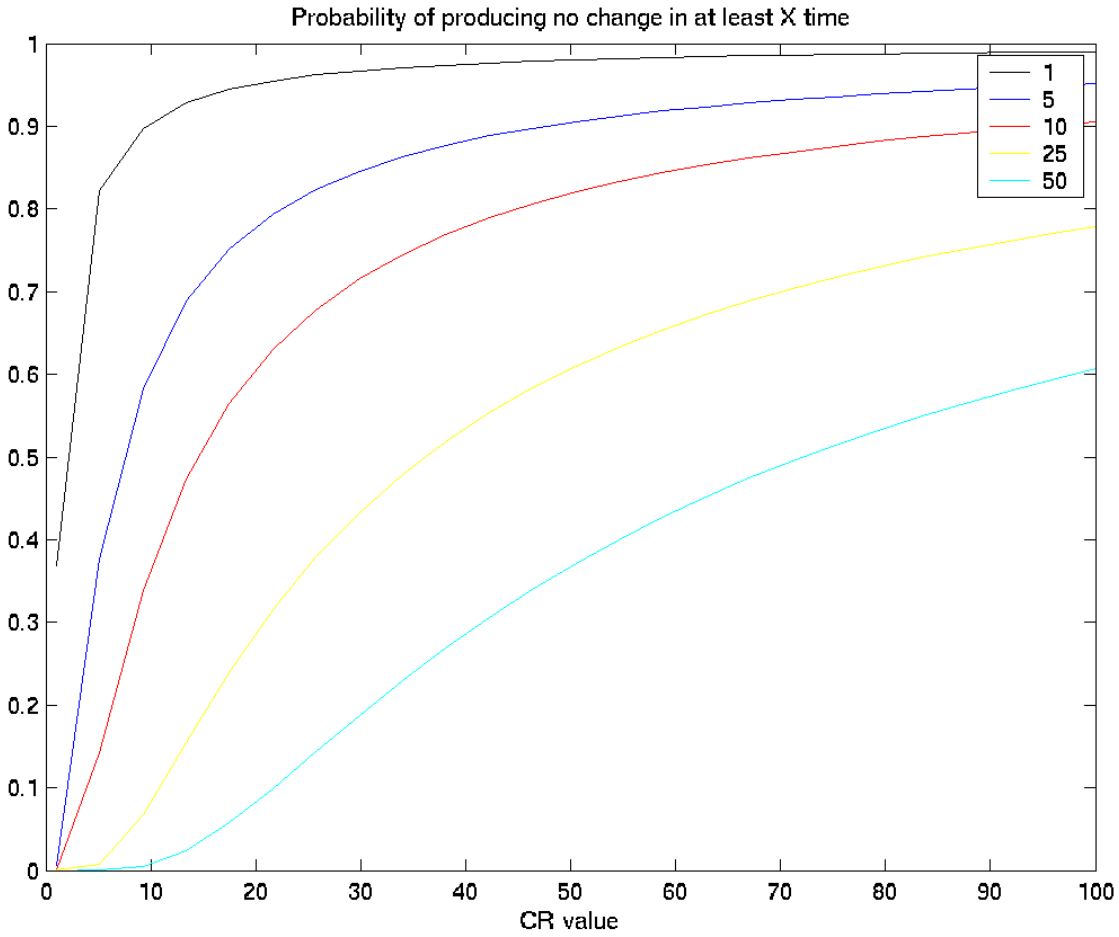
If the expected number of occurrences in this interval is λ, then the probability that there are exactly *n* occurrences (*n* being a non-negative integer, *n* = 0, 1, 2, ...) is equal to

$$f(n;\lambda) = \frac{\lambda^n e^{-\lambda}}{n!},$$

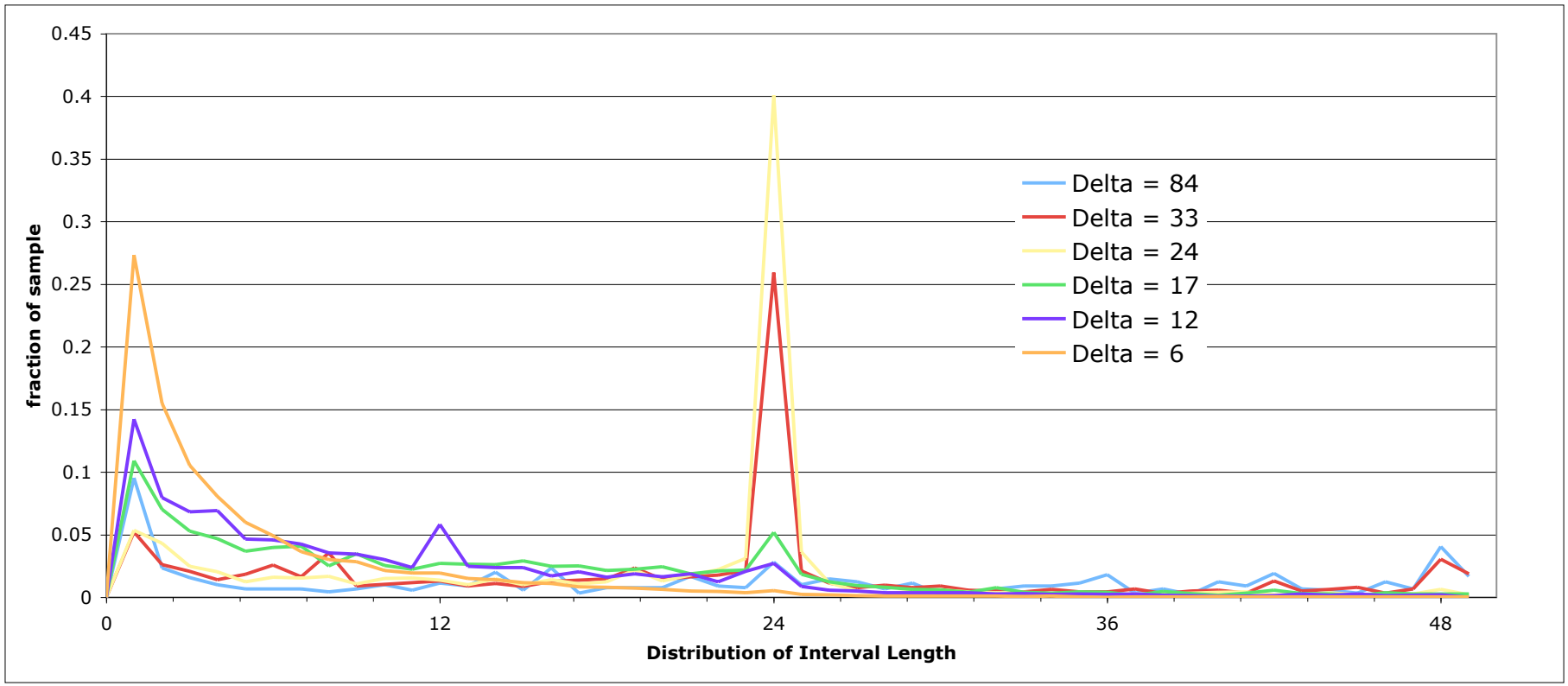
where

- *e* is the base of the natural logarithm (*e* = 2.71828...)
- *n* is the number of occurrences of an event - the probability of which is given by the function
- *n*! is the factorial of *n*
- λ is a positive real number, equal to the expected number of occurrences that occur during the given interval. For instance, if the events occur on average 4 times per minute, and you are interested in probability for n times of events occurring in a 10 minute interval, you would use as your model a Poisson distribution with λ = 10×4 = 40.

Implications of a Poisson

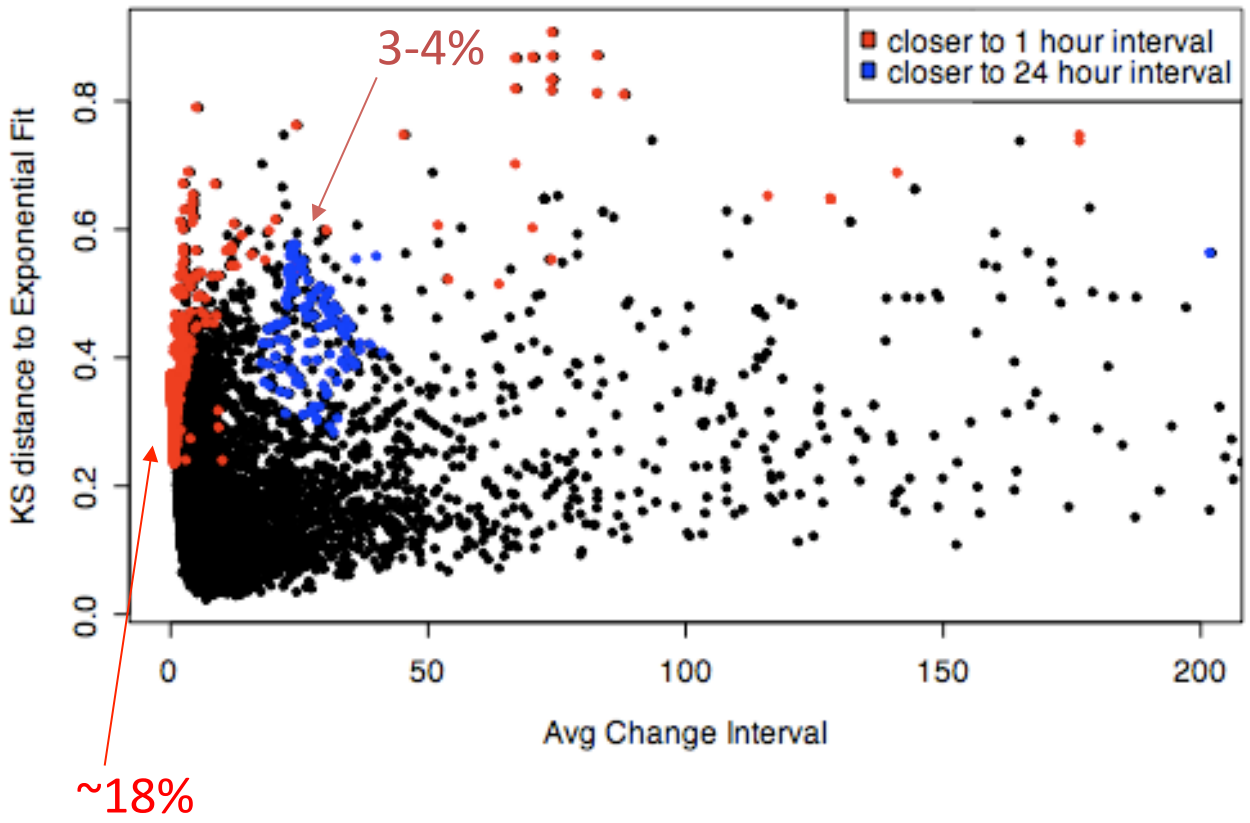


How many pages really change regularly?

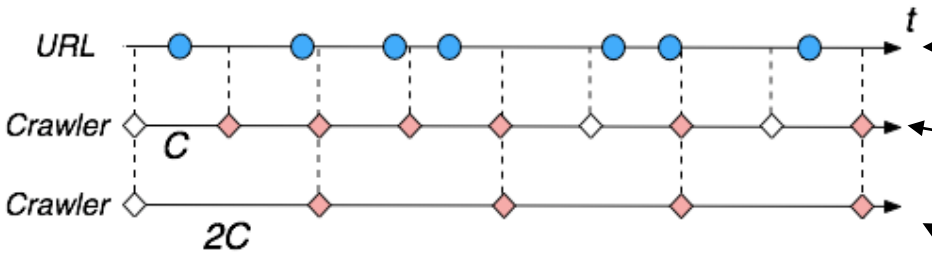


Google™

Prevalence of Discrepancies



‘Naïve’ Estimators



[a] 7 actual changes

[b] 6 changes observed out of 8 intervals

[c] 4 changes observed out of 4 intervals

For each interval, C_j, observe $x_j = 1_{\{change\ occurred\ on\ jth\ interval\}}$

then the simple estimator is:

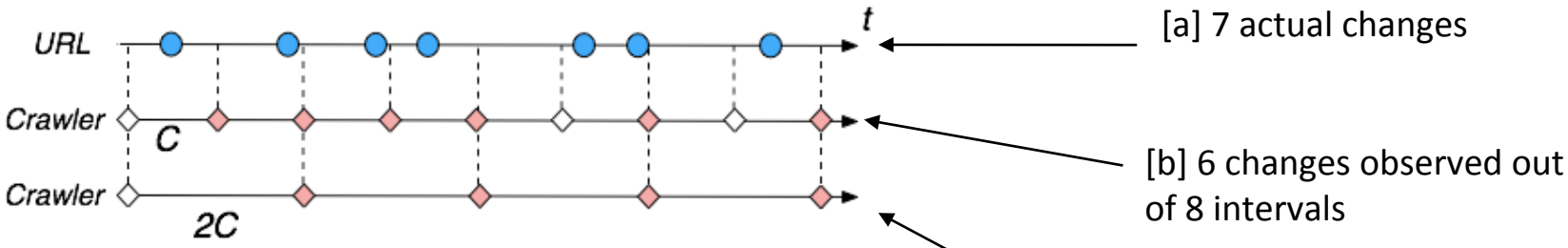
where

$$\hat{\lambda} = \frac{X}{T} \qquad X = \sum_j x_j \qquad T = \sum_j C_j$$

Taylor and Karlin 1998

Google

How to estimate with censored data?



$$\hat{\lambda} = \frac{X}{T}$$

[a] True frequency = 7/T

[b] 6/(8*C) = 6/T

[c] 4/(4*2C) = 4/T

“Better” estimators

- Cho, Garcia-Molina 2002 derive an MLE for the **regular crawl interval case**:

$$C/\Delta = -log\left(\frac{\#unchanged + 0.5)}{n + 0.5}\right)$$

C = length of your crawl interval (12 hours)
Δ = time between changes
n = number of intervals sampled

Questions:

- For what pages does the censoring matter most?
- For what pages does it make very little difference?
- How does it change the overall picture?

Break to look at data

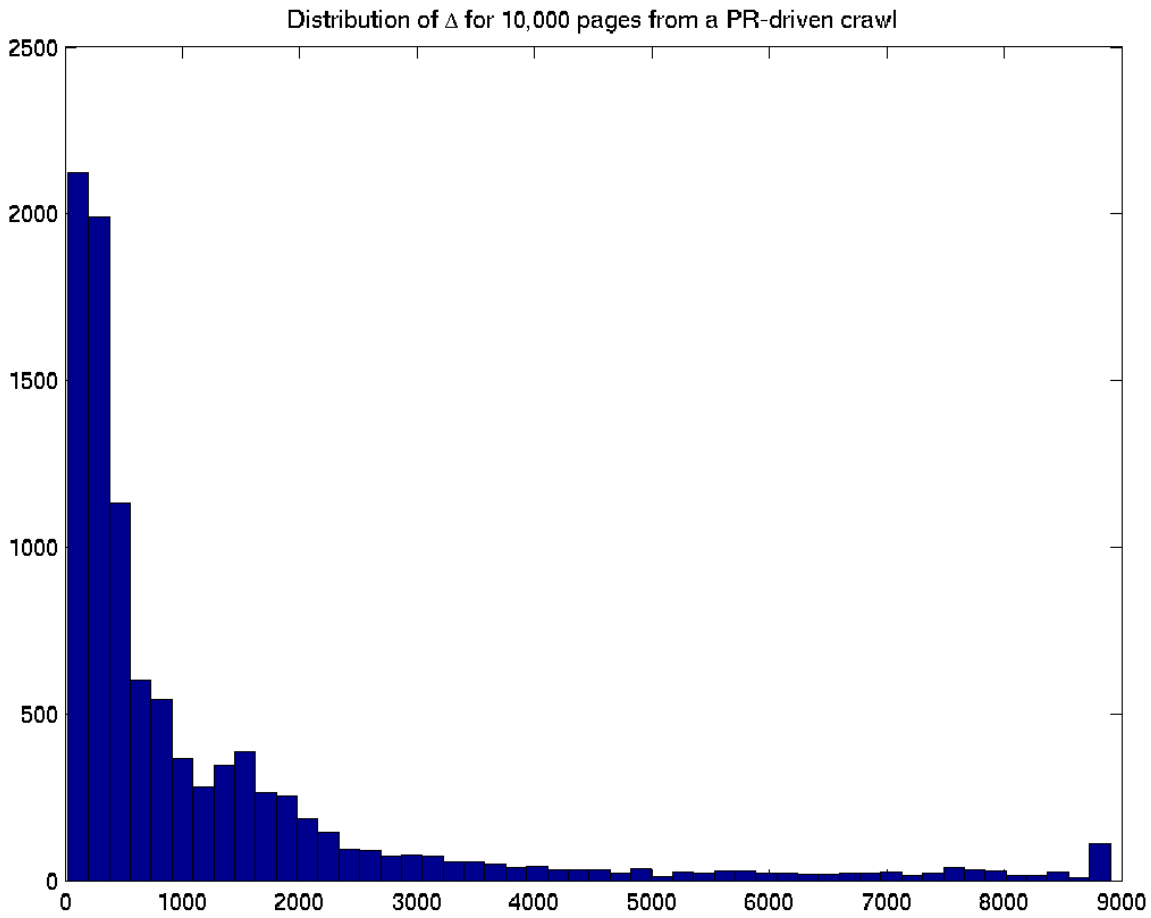
How does it alter the picture…

- This estimator has significantly smaller bias than the naïve estimator for larger ratios

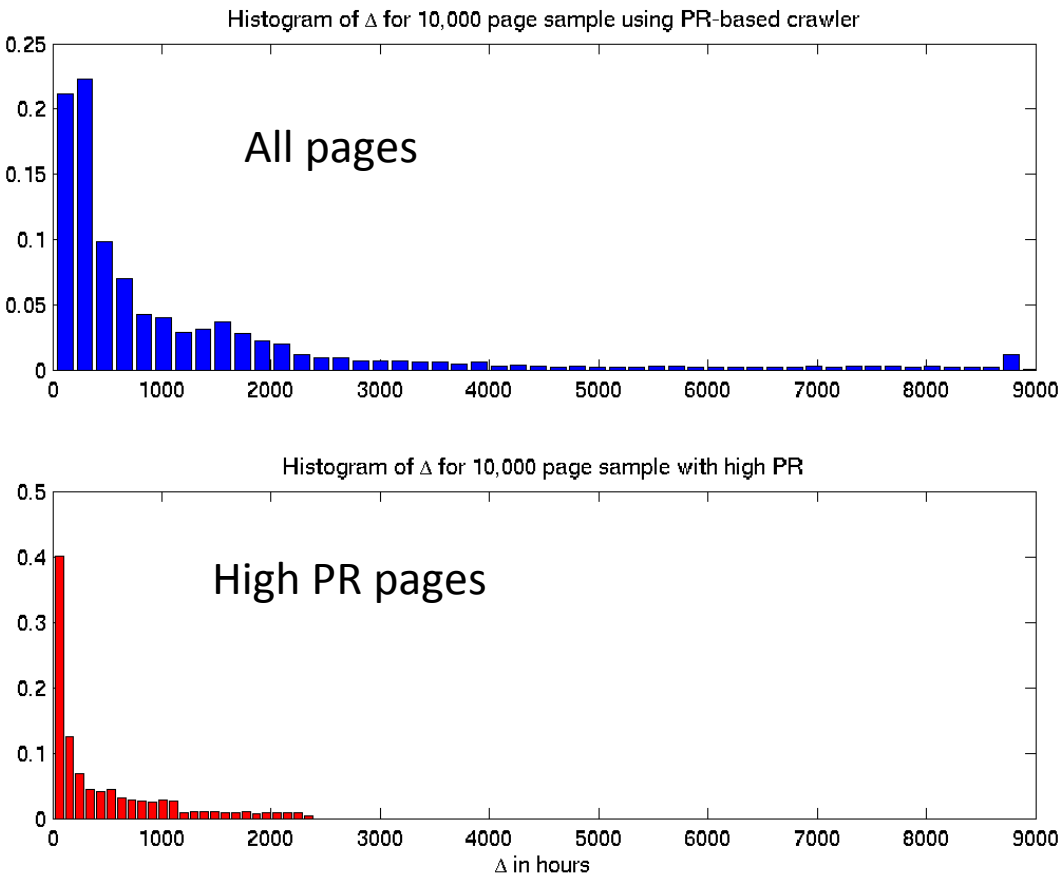
Let’s say you start from scratch

- Given the data you have, I give you a new page… what rate of change do you think it has?

Prior Distributions…



Different types of pages



Other options…

- Using the name of the host to “smear” expected rates of change
- Using other characteristics of the page: how much text, what types of content, etc.