Brief Intro to Statistics for use in Emprical Software Engineering

- By Abram Hindle <abram.hindle@ualberta.ca>

- Version 1.0/20120126

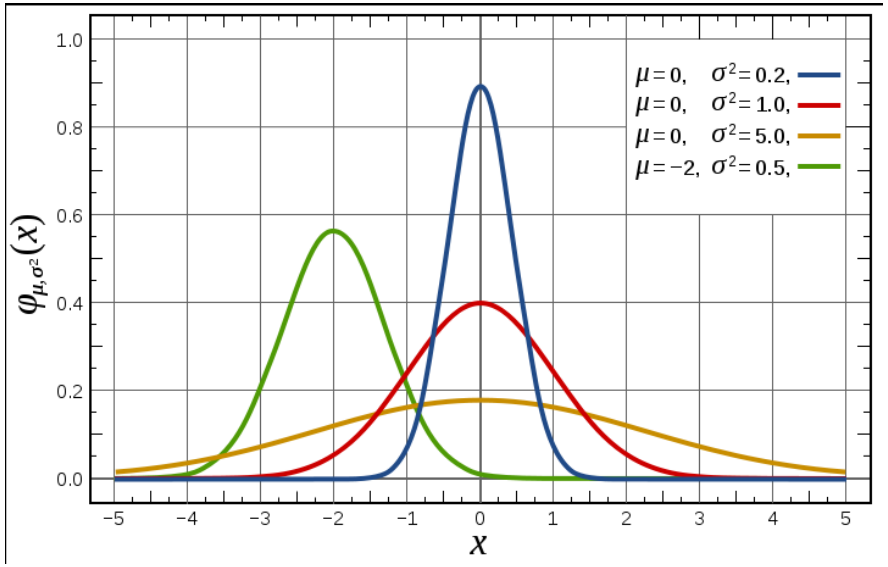
Brief Intro to Statistics

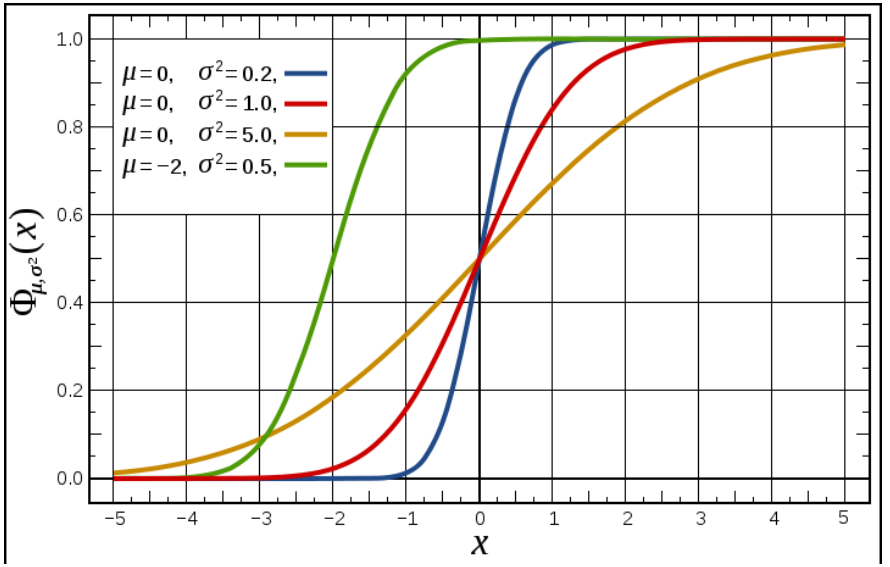
- Statistics allow us to measure and reason about data
- Statistics are well established and empirically validated, hence why we use them over other made up measures.
- Many statistical techniques were made before the widespread use of computers, thus they tend to be overconfident with large data-sets.
- Of course there are some new techniques that work well with modern computers and datasets

Distributions

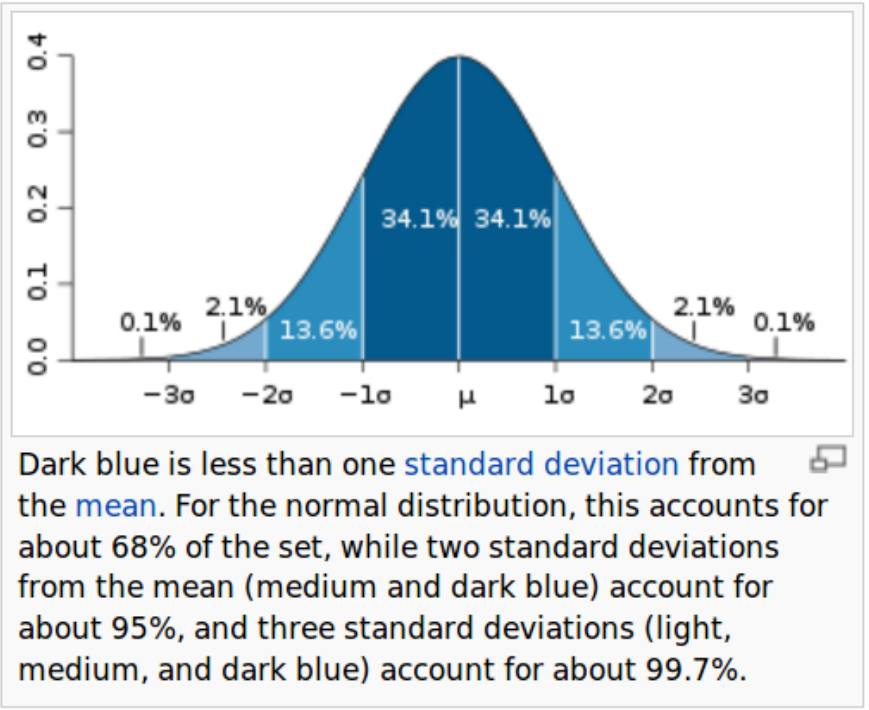
- How are values distributed over a space
- Frequency distribution
  - How often a value appears
- Probability distribution
  - The probability of a value occuring
    - Discrete [1,2,3,4,..]
    - Continuous [0.0, 1.0)
    - http://en.wikipedia.org/wiki/Probability_distribution

Normal Distribution





$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set, while two standard deviations from the mean (medium and dark blue) account for about 95%, and three standard deviations (light, medium, and dark blue) account for about 99.7%.

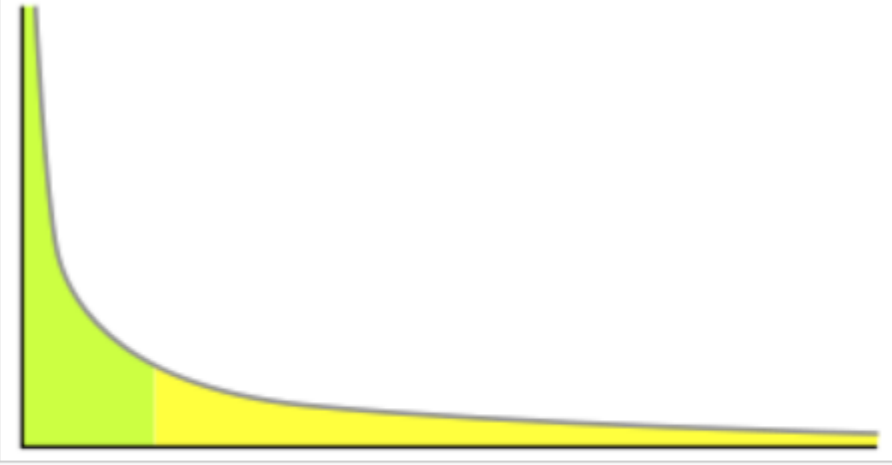
- http://en.wikipedia.org/wiki/Normal_distribution
- CC-BY-SA Wikipedia

Normal Distribution

- Normal or Gaussian Distribution
  - Natural distribution of errors
  - Well studied
  - Follow central limit theorm
    - Means of a random variable tend to produce normal distributions
    - E.g. We measure the average number of commits per day in a month, the mean of those monthly measurements likely converges to a normal distribution
  - Has nice parameters like mean and standard deviation

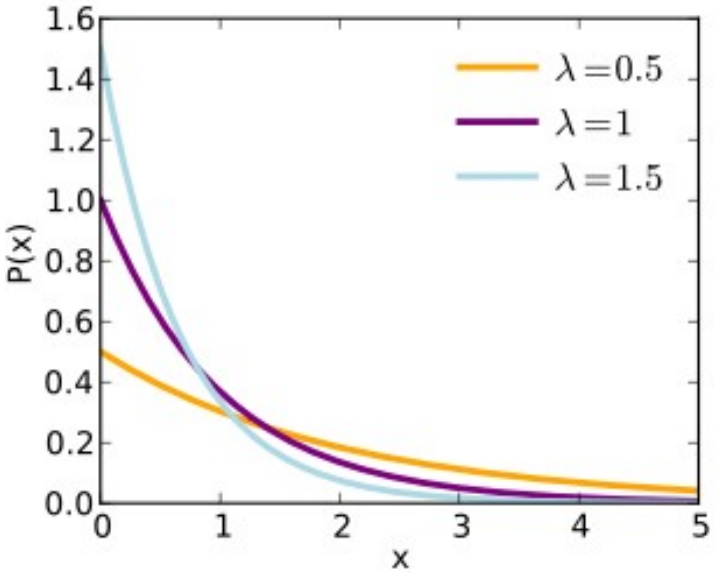
Power Law

- Common kind of relationship
  - In Software Engineering Data
- $y = ax^k + \varepsilon.$
- Powerlaws are fractals!
- http://softwareprocess.es/static/M
- Scale Free
  - They look the same at different scales.
- http://en.wikipedia.org/wiki/Power_law

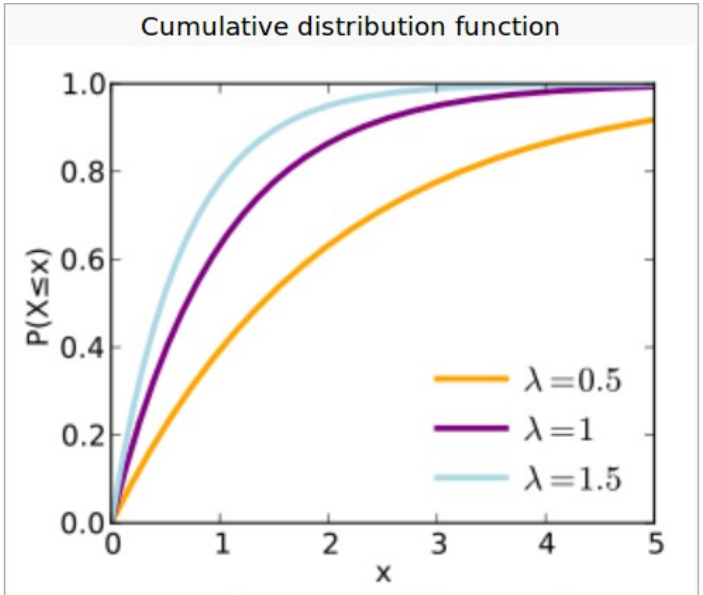


An example power-law graph, being used to demonstrate ranking of popularity. To the right is the long tail, and to the left are the few that dominate (also known as the 80-20 rule).

Exponential Distribution



- http://en.wikipedia.org/wiki/Exponential_distribution
- Common in Software engineering data
- Has a long tail.
- Easy to estimate and calculate
- Not fun to work with
- Bad for machine learners :(



Parameters	$\lambda > 0$ rate, or inverse scale
Support	$x \in [0, \infty)$
PDF	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$
Mean	λ^{-1}
Median	$\lambda^{-1} \ln 2$
Mode	0
Variance	λ^{-2}
Skewness	2
Ex. kurtosis	6
Entropy	$1 - \ln(\lambda)$
MGF	$\left(1 - \frac{t}{\lambda}\right)^{-1}$
CF	$\left(1 - \frac{it}{\lambda}\right)^{-1}$

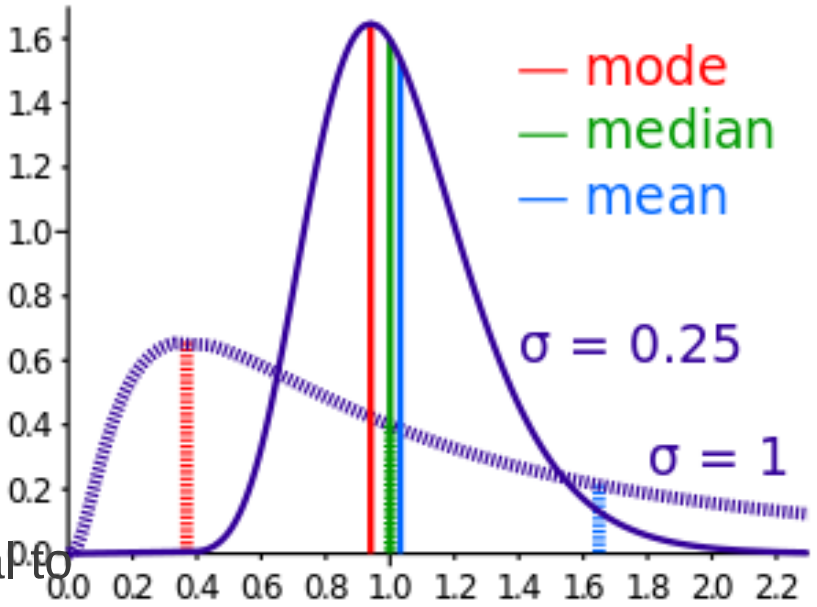
Summary Statistics

- Statistics used to describe data
  - Descriptive Statistics
- Location: mean, median, mode
- Spread: standard deviation, variance, quartiles
  - Gini coeffecient, exponents for skews (powerlaw)
- Shape: skew, kurtosis
- Percentiles: quartiles
- Dependence: correlation

Summary Statistics

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}.$$

- Average/Mean
  - What's the value of a random element
  - Works well on normal distributions
    - Doesn't work well to summarize skewed distribution
  - http://en.wikipedia.org/wiki/Average
- Median
  - The middle element
    - Half the population is greater than or equal to the median
    - Half the population is less than or equal to the median
  - http://en.wikipedia.org/wiki/Median



Summary Statistics

- Standard Deviation and Variance
  - Spread
  - Continuous
  - Discrete
  - Describes normal distribution well
  - Very wacky for exponential distributions (usually large too)

$$\mathrm{Var}(X) = \int (x-\mu)^2 f(x)\,dx \qquad \mu = \int x\, f(x)\,dx$$

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2 \qquad \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

Quartiles and Boxplots

- Quartiles partition values into 4 equal groups
  - [1,2,3][4,5,6][7,8,9][10,11,12] – quartiles of natural numbers between 1 and 12
  - Often seen in boxplots (median + 2 quartiles)
- Boxplots show median, Quartile 2 and Quartile 3 and sometimes the other 2 quartiles or and then last 2 quartiles of 98% of the data and then some outliers (2% of the data). They generally indicate max and min, which might be represented by outliers (circles)