UiO : Faculty of Mathematics and Natural Sciences
University of Oslo

Department of Informatics
Networks and Distributed Systems (ND) group

INF 5060/9060

Quantitative Performance Analysis

Özgü Alay
Carsten Griwodz

Why do we need statistics?

1. Noise, noise, noise, noise, noise!

445 446 397 226
388 3445 188 1002
47762 432 54 12
98 345 2245 8839
77492 472 565 999
1 34 882 545 4022
827 572 597 364

$$\bar{x} = \ldots$$

2. Aggregate data into meaningful information.

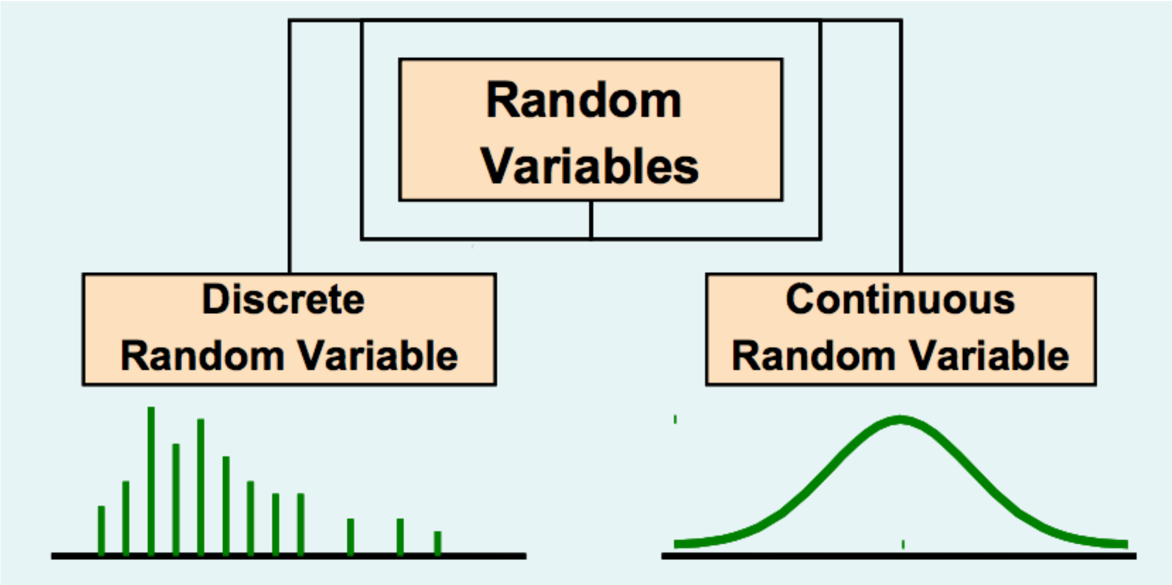
"Impossible things usually don't happen."

- Sam Treiman, Princeton University

Statistics helps us quantify "**usually**."

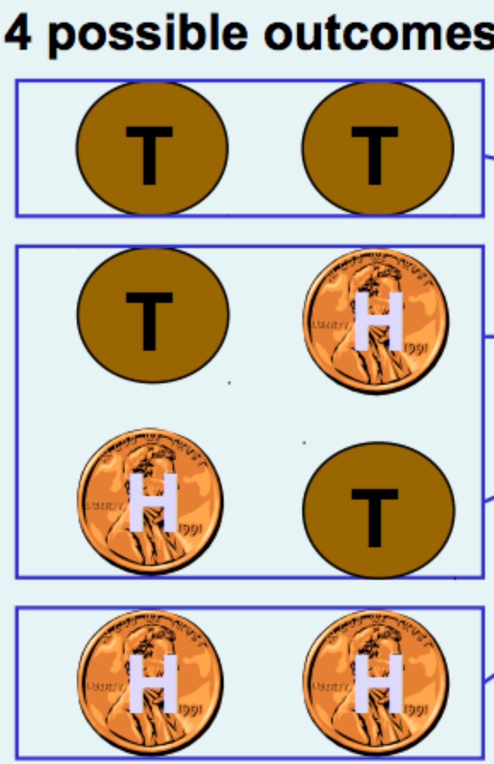
Basic Probability and Statistics Concepts

- Independent Events:
  - One event does not affect the other
  - Knowing probability of one event does not change estimate of another
- Random Variable:
  - A variable is called a random variable if it takes one of a specified set of values with a specified probability



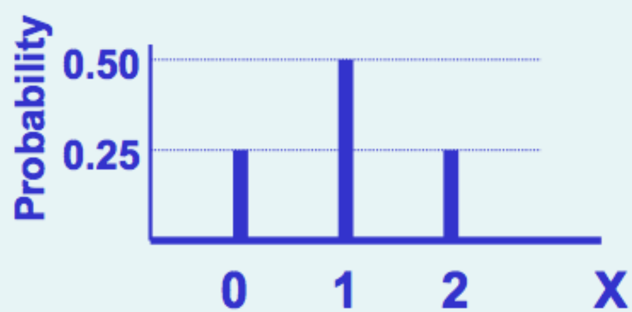
Example of a Discrete Random Variable Probability Distribution

Experiment: Toss 2 Coins. Let X = # heads



Probability Distribution

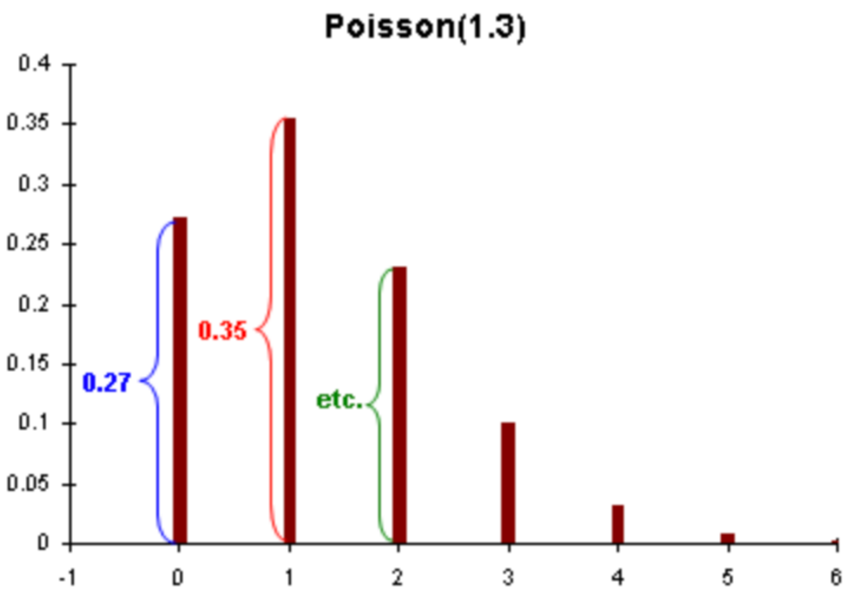
X Value	Probability
0	1/4 = 0.25
1	2/4 = 0.50
2	1/4 = 0.25



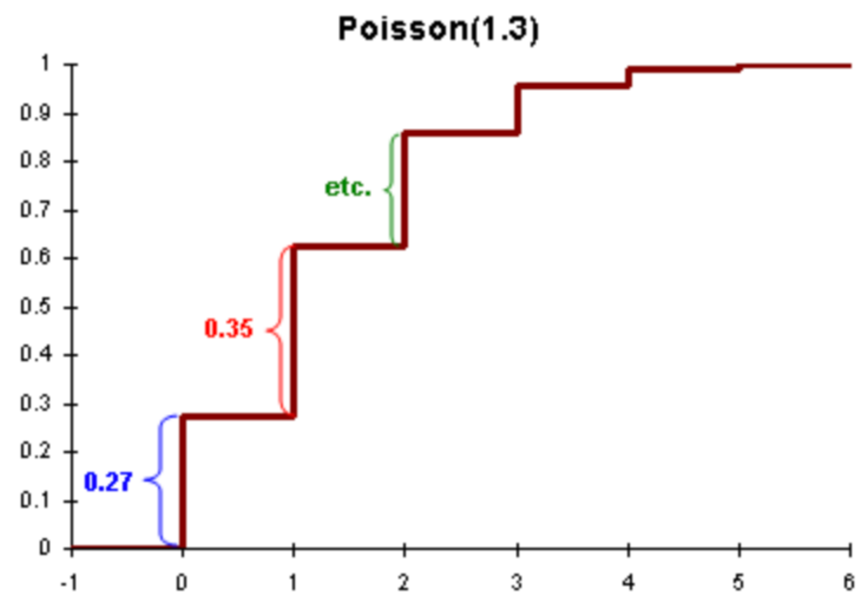
Histogram and Cumulative Distribition

Histogram: $f(x_i) = p_i$

Cumulative Distribution Function: $F_x(a) = P(x \leq a)$



Histogram



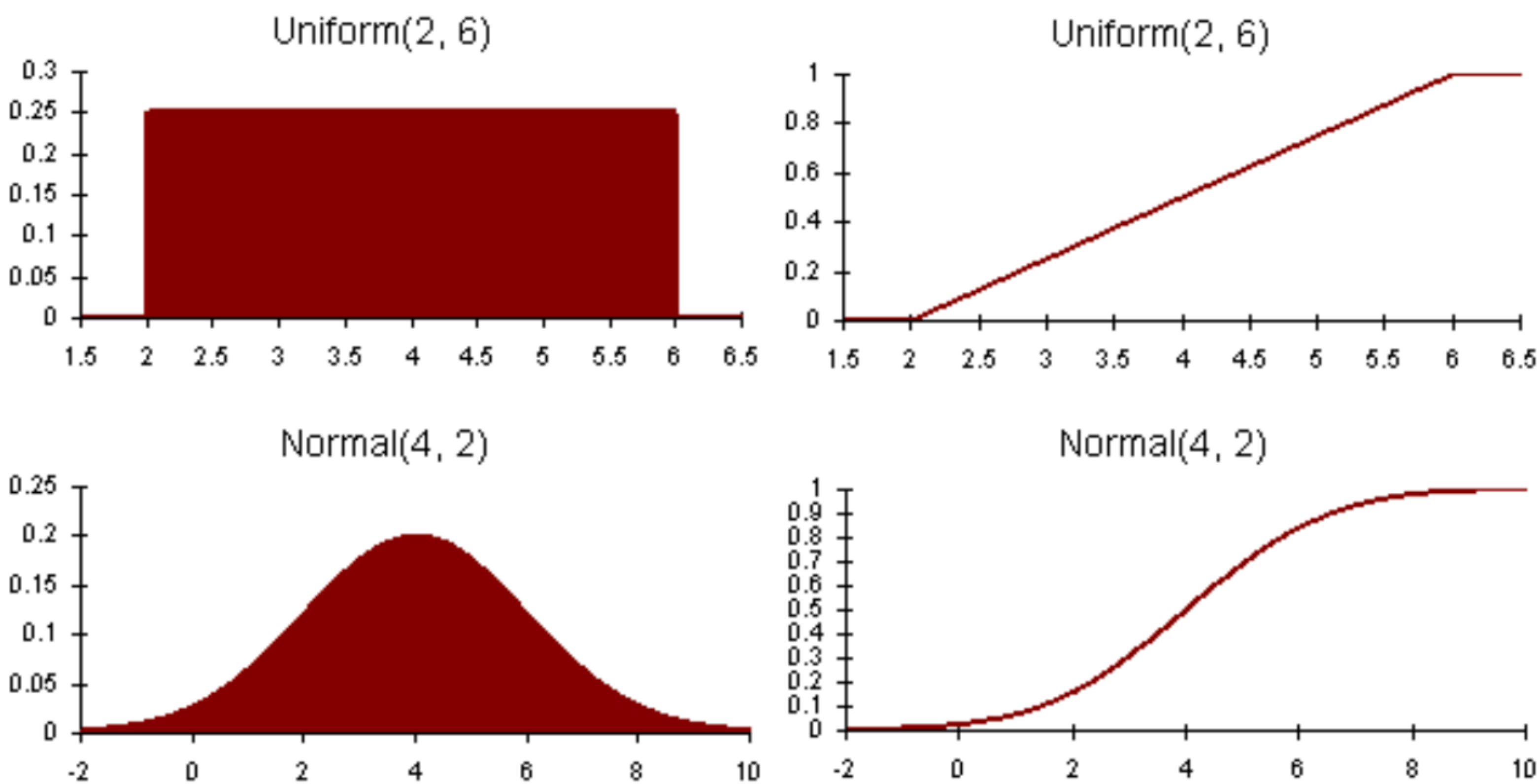
Cumulative Distribution Function (CDF)

Continuous Random Variable Probability Density Function

The probability density function, *pdf*, as $f(x)$.

The cumulative distribution function, *cdf*, as $F(x)$.

$$F_x(a) = P(x \leq a)$$

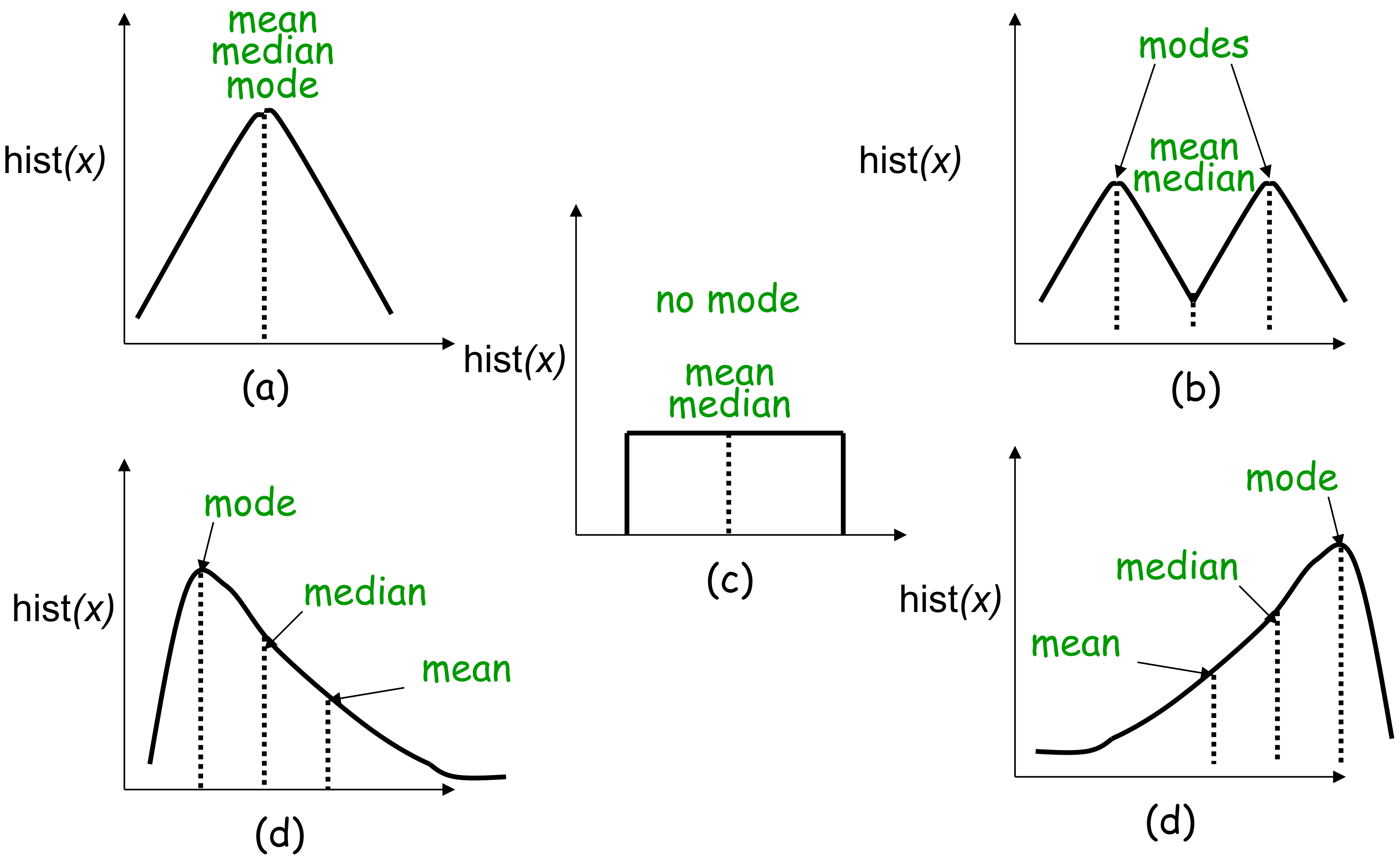


Indices of central tendency

Summarizing Data by a Single Number

- Mean – sum all observations, divide by number
- Median – sort in increasing order, take middle
- Mode – plot histogram and take largest bucket
- Mean can be affected by outliers, while median or mode ignore lots of info
- Mean has additive properties (mean of a sum is the sum of the means), but not median or mode

Relationship Between Mean, Median, Mode

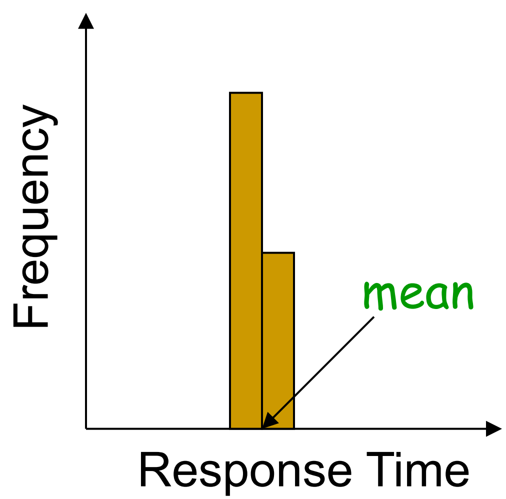


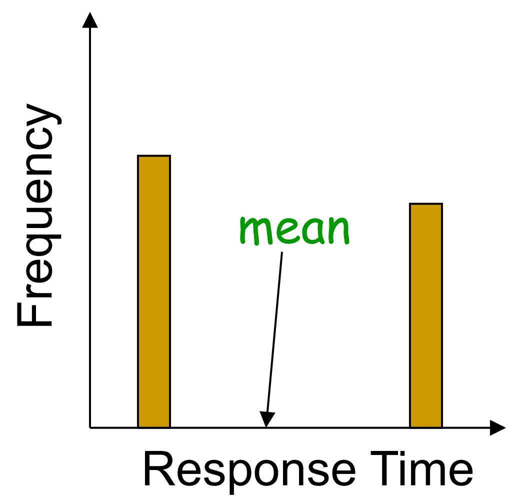
Summarizing Variability

"Then there is the man who drowned crossing a stream with an average depth of six inches." – W.I.E. Gates

- Summarizing by a single number is rarely enough

→ need statement about *variability*





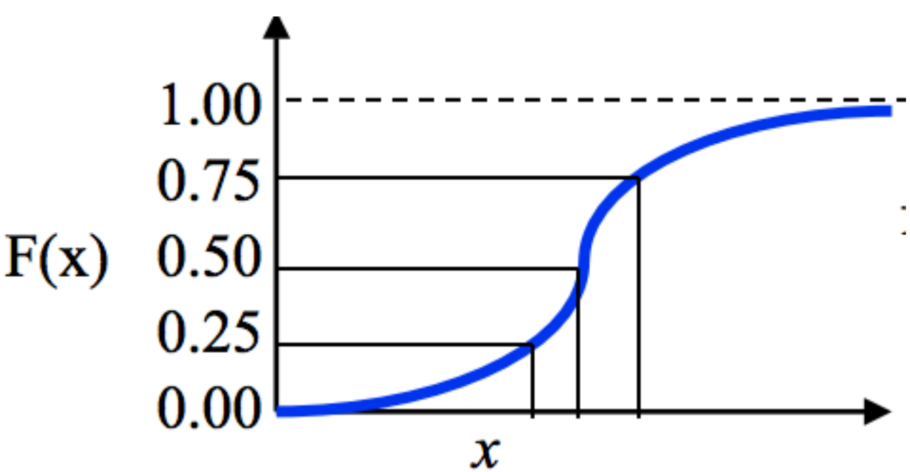
If two systems have same mean, tend to prefer one with <u>less variability</u>

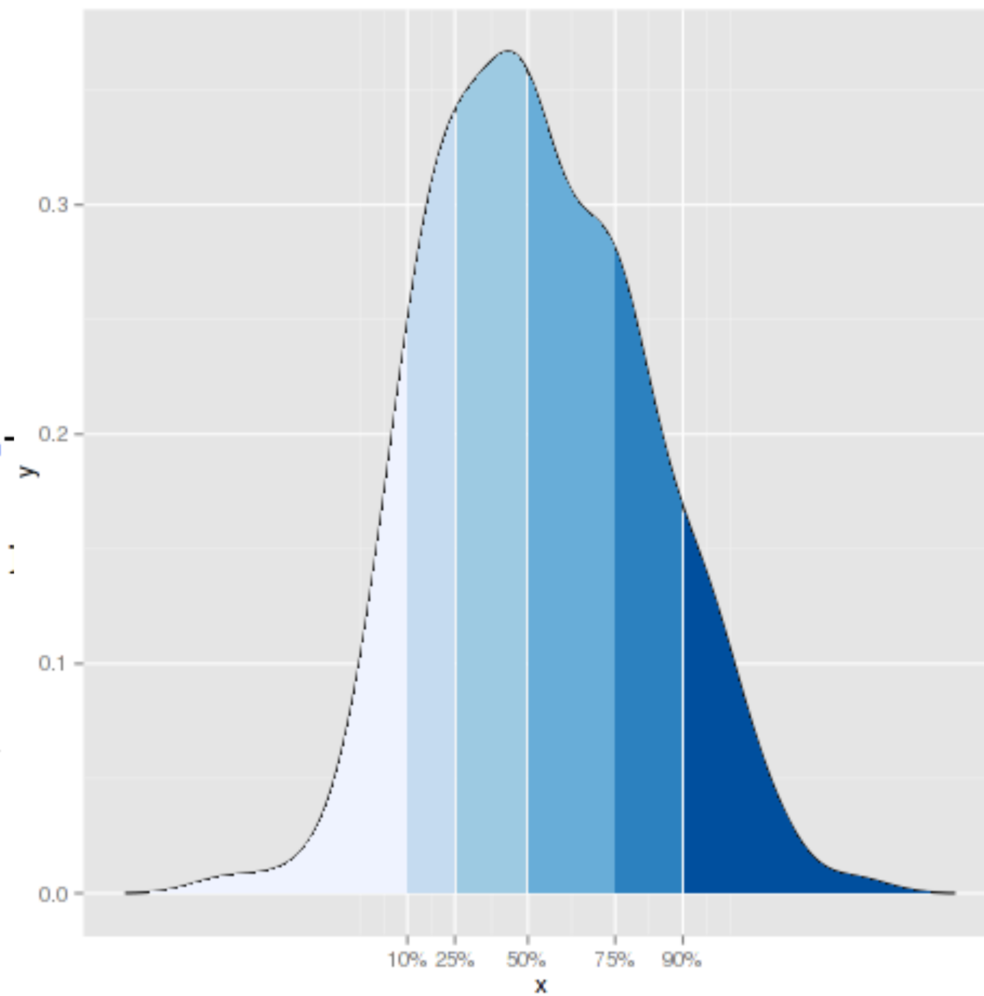
Indices of Dispersion

- *Range* – min and max values observed
- *Variance* or *standard deviation* or *CoV*
  - Variance: Square of the distance between a set of values x_i with relative frequency p_i and the mean μ
    - $\sigma^2 = E[(x-\mu)^2] = \sum_{i=1}^{n} p_i (x_i - \mu)^2$
  - or, if you have exactly n samples x_1 ... x_n
    - $\sigma^2 = E[(x-\mu)^2] = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$
  - Standard deviation, σ, is square root of variance
  - Coefficient of Variation (C.O.V.): Ratio of standard deviation to mean: = σ / μ
- *Percentiles*
  - The x value at which the *cdf* takes a value α is called the α-percentile and denoted x_α, so $F(x_\alpha) = \alpha$

Indices of Dispersion

- 10- and 90-*percentiles*
- (*Semi-)interquartile* range (SIQR)
  - Q1, Q2 and Q3





Determining Distribution of Data

- Additional summary information could be the *distribution* of the data
  - Ex: Disk I/O mean 13, variance 48. Ok. Perhaps more useful to say data is *uniformly distributed* between 1 and 25.
  - Plus, distribution useful for later simulation or analytic modeling

- How do determine distribution?
  - Plot histogram

For more formal testing: statistical comparison of
CDF (*Komolgorov-Smirnov test*) or PDF (*Chi-square test*)
The Art of Computer Systems Performance Analysis, pp. 460-465

Comparing Systems Using Sample Data

"Statistics are like alienists – they will testify for either side." – Fiorello La Guardia

- The word "sample" comes from the same root word as "example"
- Similarly, one sample does not prove a theory, but rather is an example
- Basically, a definite statement cannot be made about characteristics of all systems
- Instead, make probabilistic statement about range of most systems
  - *Confidence intervals*

Sample versus Population

- Say we generate 1-million random numbers
  - mean μ and stddev σ.
  - μ is *population mean*
- Put them in an urn draw sample of *n*
  - Sample $\{x_1, x_2, \ldots, x_n\}$ has mean $\bar{x}$, stddev s
- $\bar{x}$ is likely different than μ!
  - With many samples, $\overline{x_1} \neq \overline{x_2} \neq \cdots$
- Typically, μ is not known and may be impossible to know
  - Instead, get estimate of μ from $\overline{x_1}, \overline{x_2}, \ldots$

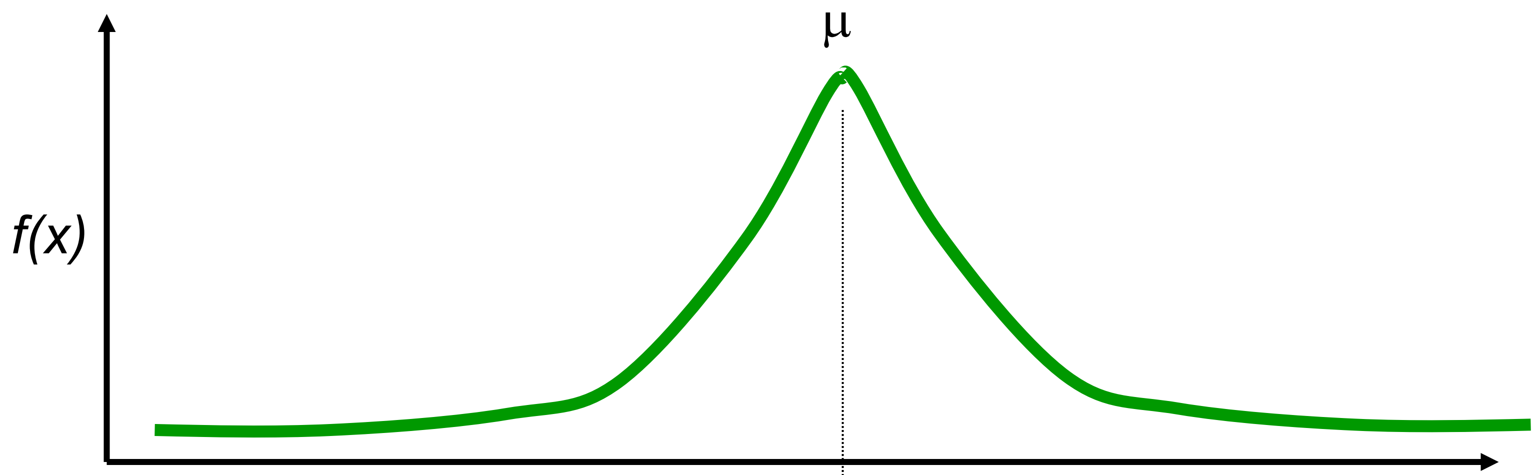
Confidence Interval for the Mean

- Obtain probability of μ in interval $[c_{1,} c_2]$
  - $Prob(c_1 \leq \mu \leq c_2) = 1 - \alpha$
    - $[c_{1,} c_2]$ is the *confidence interval*
    - α is the *significance level*
    - $100(1 - \alpha)$ is the *confidence level*

- Typically want α small so confidence level 90%, 95% or 99% (more later)
- Use 5-percentile and 95-percentile of the sample means to get 90% confidence interval

Meaning of Confidence Interval

- For a 90% confidence level, if we take 100 samples and construct the confidence interval for each sample, the interval would include the population mean in 90 cases.



Sample	Includes μ?
1	yes
2	yes
3	No
…	…
Total	yes $\geq 100(1-\alpha)$

What if *n* not large?

- Above only applies for large samples, 30+
- For smaller *n*, can only construct confidence intervals if observations come from normally distributed population: t-variate
  - $\left[\bar{x} - t_{\left[\frac{1-\alpha}{2};n-1\right]} \frac{s}{\sqrt{n}}; \bar{x} + t_{\left[\frac{1-\alpha}{2};n-1\right]} \frac{s}{\sqrt{n}}\right]$

$\bar{x}$: sampled mean

s: sampled standard deviation

n: number of samples

$t_{\left[\frac{1-\alpha}{2};n-1\right]}$: tabulated value of the t distribution

- Table A.4 of Jain's book

What Confidence Level to Use?

- Often see 90% or 95% (or even 99%), but…
- Example:
  - Lottery ticket $1, pays $5 million
  - Chance of winning is 10^{-7} (1 in 10 million)
  - To win with 90% confidence, need 9 million tickets
    - No one would buy that many tickets!
  - So, most people happy with 0.01% confidence