Technical Definitions

STATISTICS, PROBABILITY AND RANDOM VARIABLES

Statistics is a broad mathematical discipline which studies ways to collect, summarize and draw conclusions from data. Mathematical quantification of uncertainty (mean, Std. Dev., Var., …).

Probability: Mathematicians think of probabilities as numbers in the interval from 0 to 1 assigned to "events" whose occurrence or failure to occur is random. Uses the information from set theory and statistics to compute the likelihood of events.

A **random variable** is a variable whose value depends upon the outcome of a random experiment: a function (rule) that assigns a real number ($X(\varpi)$, $-\infty < X(\varpi) < \infty$) to each outcome ($\varpi$) of a random experiment in a sample space.

REAL WORLD EXPERIMENT (data) $\Leftrightarrow$ **MATHEMATICAL MODEL** (a Probability System)

(1) A set of possible experimental outcomes	(1) A sample space (subset of population)
(2) A grouping of outcomes into classes (results)	(2) An event (or set of events)
(3) The relative frequency of occurrence of the results in many independent trials	(3) A probability measure

An **experiment** is a process whose outcome is not known with certainty.

A **sample space** corresponds to the set of all possible (mutually exclusive) outcomes of an experiment. Each individual, possible outcome (ϖ) is a called a **sample point**.

An **event** is a subspace of the sample space containing a set of sample points $\{\varpi\}$. An event containing only one sample point is called an **elementary (simple) event**. An event matching the sample space is a **certain (sure, exhaustive) event**. The null set is called the **impossible event**. An event corresponds to a result in the real world.

A **probability measure** is an assignment (mapping) of the events defined on the sample space into the set of real numbers. Pr[A] denotes the real number associated with the event A.

(a) For any event A, $0 \leq Pr[A] \leq 1$.

(b) $Pr[S] = 1$ where S is the collection of all possible events.

(c) If A and B are mutually exclusive events, $Pr[A \cup B] = Pr[A] + Pr[B]$.

Statistical independence : Two events A and B are independent iff $Pr[AB] = Pr[A]Pr[B]$ => knowledge of the occurrence of the event B in no way affects the probability of occurrence of the event A.

Probability (Cumulative) Distribution Function (PDF): a way to express the probabilities associated with a random variable. The probability that the random variable X takes on a value less than or equal to x is expressed

mannCreek.com

Technical Definitions

$$F_X(x) \equiv Pr[X \le x] \qquad \text{for } -\infty < x < \infty$$

and has the following properties:

1) $0 \le F(x) \le 1 \quad \forall x$

2) F(x) is non-decreasing [if $x_1 < x_2$ then $F(x_1) < F(x_2)$]

3) $\lim x \to \infty\ F(x) = 1$ and $\lim x \to -\infty\ F(x) = 0$

A random variable that takes on a finite number of values $x_1, x_2, \ldots x_n$ is **discrete** and is defined:

$$F_X(x) = \sum p(x_i) \qquad \text{for} \quad x_i \le x \quad \text{and} \quad -\infty < x < \infty$$

A random variable is said to be **continuous** if there exists a non-negative function f(x) (**probability density function (pdf)**) which when integrated over an interval gives the probability that the random variable X takes on a value that lies in that interval:

$$Pr[X \in B] = \int_B f_X(x)\,dx \qquad \text{and} \qquad \int f_X(x)\,dx = 1 \qquad \text{(integrated over x from } -\infty \text{ to } \infty)$$

Relationships between PDF and pdf:

$$f_X(x) \equiv dF_X(x) / dx \qquad \text{or} \qquad F_X(x) \equiv \int f_X(x)\,dx$$

$$Pr[X \in [a,b]] = \int f_Y(y)\,dy = F_X(b) - F_X(a)$$

Joint (Probability) Distribution Function (joint PDF): for two random variables X and Y defined on the outcome(s) in the same sample space, define the probability that X takes on a value less than or equal to x at the same time Y takes on a value less than or equal to y. This is the sum of the probabilities associated with all sample points in the intersection of the two events $\{\varpi: X(\varpi) \le x\}$, $\{\varpi: Y(\varpi) \le y\}$.

$$Pr[X \le x, Y \le y] \equiv F_{XY}(x,y)$$

Joint (probability) density function (joint pdf):

$$f_{XY}(x,y) \equiv d^2F_{XY}(x,y) / dx\,dy$$

$$\Rightarrow Pr[X \in A, Y \in B] = \int_B \int_A f_{XY}(x,y)\,dx\,dy$$

"Marginal" (probability) density function (marginal pdf):

$$f_X(x) = \int f_{XY}(x,y)\,dy \text{ (integrated over y from } -\infty \text{ to } \infty)$$

Technical Definitions

Random variable independence: The random variables X and Y are independent if knowing the value that one random variable takes on tells us nothing about the distribution of the other. Two random variable are said to be independent if and only if their joint pdf factors into the product of their one-dimensional pdf factors.

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) \;\forall x,y$$

For three or more random variables,

$$f_z(\zeta) = f_{X_1}(x_1)\, f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

where $_z$ and ζ are vectors of $X_1, X_2, \ldots, X_n$ and $x_1, x_2, \ldots x_n$ respectively.

The **expected value** of a random variable X gives a single value which acts as a representative of the possible values of X. For a discrete random variable X having the possible values $x_1, x_2, \ldots x_n$, the expected value of X is

$$E(X) = x_1 Pr(X=x_1) + \ldots + x_n Pr(X=x_n) = \sum x_j Pr(X=x_j)$$

$$= x_1 f(x_1) + \ldots + x_n f(x_n) = \sum x_j f(x_j)$$

$$= x_1 + x_2 + \ldots + x_n / n \quad = \mu \text{ (mean) if all probabilities are equal}$$

For a continuous random variable having density function f(x), the expected value is

$$E(X) = \int xf(x)dx \quad \text{(integrated from } -\infty \text{ to } \infty)$$

The **variance** is a measure of the scatter of the values of the random variable about its mean.

$$Var(X) = E[(X - \mu)^2]$$

For a discrete random variable, with probability density function f(x),

$$Var(X) = E(x_j - \mu)^2 f(x_j)$$

$$= [(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2] / n \quad \text{if all probabilities are equal}$$

For a continuous random variable with probability density function f(x),

$$Var(X) = \int (x - \mu)^2 f(x)\, dx \quad \text{(integrated from } -\infty \text{ to } \infty)$$

A measure of dependence between two random variables is the **covariance**. If Cov() = 0, the random variables X_i and X_j are **uncorrelated**. If X_i and X_j are independent, then Cov() = 0 (the reverse is not usually

Technical Definitions

true). If X_i and X_j are jointly normally distributed with Cov() = 0, then they are independent. Cov() > 0 => positive correlation. Cov() < 0 => negative correlation.

$$Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_iX_j) - \mu_i\mu_j$$

Another measure of dependence is **correlation**. Cor() > 0 => positive correlation. Cor() < 0 => negative correlation.

$$Cor(X,Y) = Cov(X,Y) / [Var(X)\ Var(Y)]^{1/2}$$

Observations in most simulations are not independent (are correlated). E.g. for queuing simulations, if the waiting time for the i^{th} job is large, the waiting time for the $(i+1)^{th}$ job would be large too. The successive waiting times are highly correlated. This correlation means the waiting time variance will be inaccurate (computed value will be too narrow). To correctly compute the variance of the mean of correlated observations, use **independent replications**. (1) For m replications of size $n+n_0$ observations each, discard the n_0 transient observations and compute a mean for each replication. $\bar{x}_i = \frac{1}{n}\sum_{j=n_0+1}^{n_0+n} x_{ij}$ i=1,2,…,m (2) Compute the overall mean for all replications. $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} x_i$ (3) Calculate the variance of the replicate means.

$Var(\bar{x}) = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_i - \bar{\bar{x}})^2$ The **confidence interval** for the mean response is $[\bar{\bar{x}} \mp z_{1-\alpha/2} Var(\bar{x})]$ if m ≥ 30 or $[\bar{\bar{x}} \mp z_{t_{[1-\alpha/2;m-1]}} Var(\bar{x})]$ if m < 30. (Ref. Jain, pp. 430-431)

Any quantity obtained from a sample for the purpose of estimating a population parameter is called a **sample statistic** (**statistic**). When the mean or expected value of a statistic equals the corresponding population parameter, the statistic is called an **unbiased estimator** of the parameter.

For $X_1, X_2, \ldots X_n$ IID (independent identically distributed) random variables (observations) with finite population mean μ and finite population variance σ^2 we have the following unbiased estimator statistics. The **sample mean**

$$\bar{X}(n) = \sum X_i / n \qquad i = 1 \rightarrow n$$

is an unbiased estimator of μ. The **sample variance**

$$S^2(n) = \sum([X_i - \bar{X}(n)]^2) / n - 1$$

is an unbiased estimator of σ^2 and $S^2(n)/n$ is an unbiased estimator of **Var[** $\bar{X}(n)$ **]**.

An unbiased estimator of the **covariance** is

$$n_a^*(\beta) = \min\left\{ i \geq \pi : t_{i-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{S^2(n)}} \leq \beta \right\}.$$

$$n_r^*(\gamma) = \min\left\{ i \geq \pi : \frac{|\bar{X}(n)|}{|\bar{X}(n) - \gamma|} \leq 1 - \gamma \right\}$$

Cov(X_i, X_{i+j}) = $\sum$([X_i - $\bar{X}(n)$] [X_{i+j} - $\bar{X}(n)$]) / n - j i = 1 → n - j.

An unbiased estimator of the **correlation** is

Cor(X_i, X_{i+j}) = Cov(X_i, X_{i+j}) / S^2(n).

Because $\bar{X}(n)$ is a random variable, there is no way of assessing how close it is to μ. The usual way to assess the precision of $\bar{X}(n)$ is to construct a **confidence interval** for μ. If the X_j's are **normal random variables**, an exact (for any n ≥ 2) 100(1 - α) percent confidence interval for μ is given by the **t confidence interval**

$$\bar{X}(n) \ (\pm)\ t_{n-1,1-\alpha/2}[S^2(n) / n]^{1/2}.$$

In practice, the X_i's are rarely normally distributed and the above coverage will be approximate. Ref. Jain, p. 204.

However, in arriving at this confidence interval equation, the **central limit theorem** is used (the most important probability theory). It predicts that the above interval will get closer to the desired coverage of μ as n gets larger. The central limit theorem basically states that (1) if n (sample size) is "sufficiently large" and (2) if the observations in the sample are independent and come from the same population that has a mean μ and a standard deviation σ, then the random variable Z_n = [$\bar{X}(n)$ – μ]/ [σ^2 / n] ½ will be approximately distributed as a standard normal random variable regardless of the distributions of the X_i's. Reference p. 290, Law & Kelton. Also, "The sum of a large number of independent observations from any distribution tends to have a normal distribution". Reference p. 182, Jain. And "If observations in a sample {x_1,x_2,…,x_n} are independent and come from the same population that has a mean μ and a standard deviation σ, then the sample mean for large samples is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$", p. 205, Jain.

The **strong law of large numbers** (the second most important probability theory): If one performs an infinite number of experiments each resulting in an $\bar{X}(n)$ and n is sufficiently large, then $\bar{X}(n)$ will be arbitrarily close to μ for almost all of the experiments.

$$\bar{X}(n) \to \mu \quad \text{w.p. 1} \quad \text{as } n \to \infty$$

Independent replications: Basic assumption here is that the means of independent replications are independent even though observations in a single replication are correlated.

Find n such that $|\bar{X}(n) - \mu| \leq \beta.$

Find n such that

Technical Definitions

Hypothesis test for $\mu = \mu_0$ (hypothesized value of μ):

If $| t_n | > t_{n-1,1-\alpha/2}$, this implies μ_0 not in confidence interval for μ

$\leq t_{n-1,1-\alpha/2}$, this implies μ_0 may be in confidence interval

where

$$t_n = [\overline{X}(n) - \mu_0] / [S^2(n) / n]^{1/2} .$$

Indices of central tendency: ref. Jain, p. 183 for when to use which.

Mean: the sum of all observations divided by the number of observations; gives equal weight to all observations.

Median: the observation that is in the middle of an increasing sort of observations; resistant to outlying observations.

Mode: the midpoint of the bucket where a histogram of all observations peaks; resistant to outlying observations.

manncreek.com

Technical Definitions

STOCASTIC PROCESSES

A **stochastic process** is a collection of random variables ordered over time, which are all defined on a common sample space. Alternatively, for each sample point ϖ assign a time function $X(t,\varpi)$. This family of functions forms a stochastic process. (Also, a stochastic process is a function $X(t,\varpi)$ whose values are random variables.)

The set of all possible values that the collection of random variables of a stochastic process can take on is called the **state space**.

A stochastic process $X(t,\varpi)$ is said to be **stationary** if all $F_X(x,t) \equiv Pr[X(t_1,\varpi) \le x_1, X(t_2,\varpi) \le x_2, \ldots X(t_n,\varpi) \le x_n]$ are invariant to shifts in time, where X, x and t are vectors of $X_1, X_2, \ldots, X_n$, $x_1, x_2, \ldots x_n$ and $t_1, t_2, \ldots t_n$.

A discrete-time stochastic process $X_1, X_2, \ldots$ is said to be **covariance stationary** if

$\mu_i = \mu$ for $i = 1, 2, \ldots$ and $-\infty < \mu < \infty$

$\sigma_i^2 = \sigma^2$ for $i = 1, 2, \ldots$ and $\sigma^2 < \infty$

and $Cov(X_i, X_{i+j})$ is independent of i for $j = 1, 2, \ldots$.

$$\Rightarrow COR(X_i, X_{i+j}) = COV(X_i, X_{i+j}) / \sigma^2.$$

In **second order theory**, stochastic processes are classified and distinguished only on the basis of their **mean** $\overline{X}(t,\varpi) = E[X(t,\varpi)] = \int x f_X(x,t)dx$ (integrated from $-\infty$ to ∞) and autocorrelation $R_{XX}(t_1,t_2) = E[X(t_1)X(t_2)]$. For a **stationary stochastic process**, $\overline{X}(t,\varpi) = \overline{X}$ and $R_{XX}(t_1,t_2) = R_{XX}(t_2 - t_1)$ (R_{XX} is only a function of the time difference). Such a process is said to be **wide-sense stationary**.

A stochastic process $\{X(t), t \ge 0\}$ is said to be a **Poisson process** if:

(1) Customers arrive one at a time.

(2) $X(t+s) - X(t)$ (the number of arrivals in the time interval $(t,t+s]$) is independent of the number of arrivals in the earlier time interval $(t-s,t)$ and the times at which those arrivals occur.

(3) The distribution of $X(t+s) - X(t)$ is independent of t for all $t, s \ge 0$ (soften this to "arrival rate is relatively constant over the interval").

Let $X_1, X_2, \ldots, X_n$ be a stochastic process of length n observations, generated during a single simulation run with IID input probability distributions (probability density functions). Then define

manncreek.com

Technical Definitions

$$
\begin{matrix}
{}_iX_{jk} = {}_iX_{11}, {}_iX_{12}, \ldots, {}_iX_{1n} \\
{}_iX_{21}, {}_iX_{22}, \ldots, {}_iX_{2n} \\
\vdots \quad \vdots \quad \vdots \\
{}_iX_{m1}, {}_iX_{m2}, \ldots, {}_iX_{mn}
\end{matrix}
$$

where X = a random variable
- ${}_iX_{jk}$ = a random variate generated from X for model i
- j = replication number
- k = simulation stage (observation)
- m = number of replications of each model
- n = number of stages (observations) in simulation run
- n_0 = length of transient stage

X may be an input random variable (e.g. arrival or service) or an output random variable (e.g. queue delay, number in queue, throughput, time in system, proportion of items delayed < t or > t, utilization).

Output processes of most simulations are **non-stationary** and **autocorrelated**, therefore not IID.

If an input random variable, the ${}_iX_{jk}$ are generated as IID **variates** and (1) the sample mean is an unbiased estimator of the finite population means ($E({}_iX_j) = {}_i\mu_j$) , (2) the sample variance is an unbiased estimator of the finite population variance ($Var({}_iX_j) = ({}_i\sigma_j)^2$) .

If an output random variable from a covariant-stationary process, the ${}_iX_{jk}$ are then not generally IID and (1) holds while (2) does not. However, the assumption that the means of independent replications are independent even though observations in a single replication are correlated leads to the method of Independent Replications (e.g. Jain, pp. 431, 432).

1. Compute the mean of each replication $\bar{x}_i = \frac{1}{n}\sum_{j=n_0+1}^{n_0+n} x_{ij}$ i = 1,2,…,m
2. Compute the overall mean for all replications $\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i$
3. Calculate the variance of replicate means $Var(\bar{x}) = \frac{1}{m-1}\sum_{i=1}^{m} (\bar{x}_i - \bar{\bar{x}})^2$
4. Calculate the confidence interval for the mean response $\left[\bar{\bar{x}} \mp z_{1-\alpha/2} Var(\bar{x})\right]$

Independence: the random variables X_i and X_j are independent if knowing the value that one random variable takes on tells us nothing about the distribution of the other.

Identically distributed: the variates have the same probability distributions.

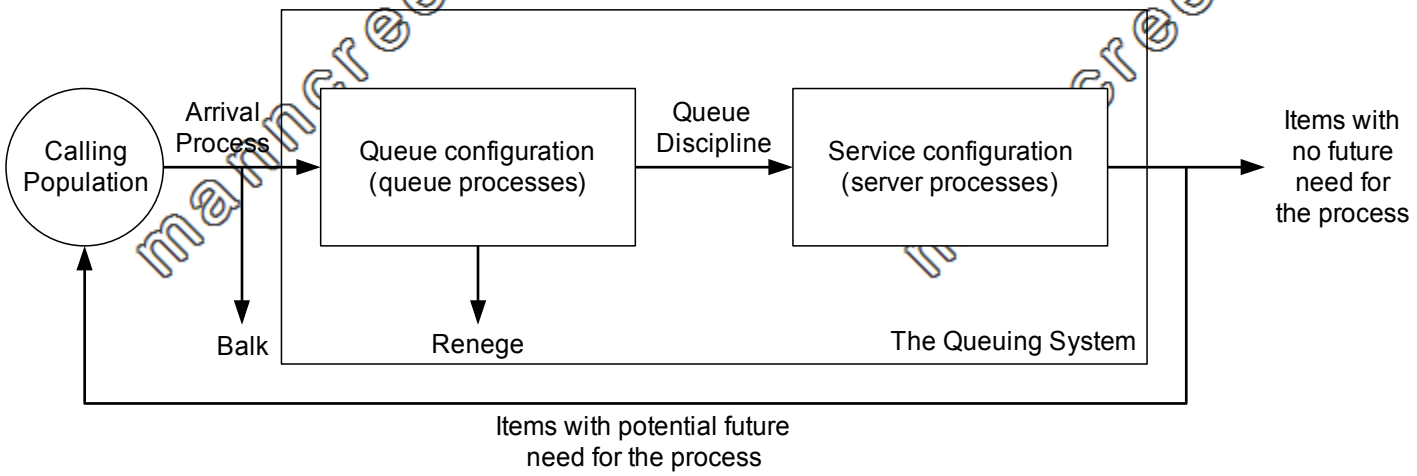
Random variate: random value (observation, realization) generated from a desired probability distribution.

Technical Definitions

QUEUEING SYSTEMS

Queuing theory: the mathematical study of waiting lines (queues).

Queuing Process and queuing system[1]:



Calling population: homogeneous / heterogeneous, finite / infinite, sample / population.

Arrival process: temporal and spatial distribution of queuing system demand.

Queue configuration: number of and location relative to other process components.

Queue discipline: queue policies (FIFO, LIFO, priority …).

Service configuration: number of, location, timings, distributions…

Balk: choice to not join queue.

Renege: leave queue before being served.

Interarrival time = A_i = time between arrivals of $(i - 1)^{st}$ and i^{th} item. If $A_1, A_2, \ldots$ are assumed to be IID random variables, then denote the **mean (expected) interarrival time** by E(A) and call $\lambda = 1/E(A)$ the **arrival rate** of items.

Service time = S_i = service time of i^{th} arriving item. If $S_1, S_2, \ldots$ are assumed to be IID random variables, then denote the **mean service time** of server by E(S) and call $\omega = 1/E(S)$ the **service rate** of a server.

Queuing systems notation (Kendall notation):

A/S/m/b/K/SD denotes an m-server queuing system with storage capacity b, item population size K and service discipline SD (e.g. FIFO), where:

1 Ref. Manual Laguna, Marklund, Johan: Business Process Modeling, Simulation and Design

Technical Definitions

A = interarrival time distribution

S = service time distribution

and the distributions take on one of the following values:

M (exponential or a variant (e.g. m-Erlang))

E_r (r-stage Erlangian)

H_R (R-stage hyperexponential)

D (deterministic)

G (general)

Measures of steady-state queuing system performance for M/M/b queues: (could try to give complete probabilistic description ((probability) distribution function), but usually settle for mean, variance, etc.)

average number of visits to queue j (V_j)

average throughput of queue j (X_j) = average number of transactions that complete from queue j per time unit. The **Flow Equilibrium Assumption** states that the number of arrivals and departures to a queue are *almost* the same, which implies $\lambda_j = X_j$.

average system throughput (X_0) = average number of transactions that complete per time unit

average number of transactions waiting for queue j (N^w_j).

average number of transactions receiving service from queue j (N^s_j)

average number of transactions at queue j waiting or receiving service ($N_j = N^w_j + N^s_j$)

time-in-queue (delay) for item i (D_i)

service time for i^{th} arriving item (S_i)

time in system for i^{th} item ($W_i = D_i + S_i$)

number of items in queue j at time t ($Q_j(t)$)

number of items in system at time t (L(t))

Then:

$d = \lim (\sum D_i / n)$ as $n \to \infty$ w.p. 1 and $w = \lim(\sum W_i / n)$ as $n \to \infty$ w.p. 1, summed from i=1 to n

are called the queue **steady-state average delay** and **steady-state average waiting time**, and

$Q = \lim (\int Q(t)\,dt / T)$ as $T \to \infty$ w.p. 1 and $L = \lim (\int L(t)\,dt / T$ as $T \to \infty$ w.p.1, integrated from 0 to T

are called the **steady-state time-average number of items in queue** and **steady-state time-average number of items in system**.

Utilization factor = $\rho = \lambda / (b\omega)$ where $(b\omega)$ is the service rate of the system when all servers are busy.

queue size depends on the arrival rate (λ) and the service rate ($b\omega$), and their statistical fluctuations.

$\lambda < (b\omega) \Rightarrow \rho < 1 \Rightarrow$ queues form due to statistical fluctuations. As $\rho \to 1$, effects of statistical fluctuations become more pronounced.

$\lambda > (b\omega) \Rightarrow \rho > 1 \Rightarrow$ average overload dominates growth of queues and unbounded queues begin to form.

Utilization Law: For server i,

$$U_i = X_i * S_i = \lambda_i * S_i \text{ (per Flow Equilibrium Assumption)}$$

Forced Flow Law: The average throughput of queue j,

$$X_i = V_i * X_0$$

Little's Law: The average arrival rate (λ) of items to a "system" times the average time spent by items in that "system" (T) is equal to the average number of items in the "system" (N) where "system" may be queue(s) + server(s), queue(s) only or server(s) only.

$$N = \lambda T$$

Conservation equations: For queuing systems for which d and w exist,

$$Q = \lambda d \quad \text{and} \quad L = \lambda w$$

Also, $w = d + E(S)$

manncreek.com

Technical Definitions

Key Simulation Variables and Relationships

Queue-related

Random Variables (a variable that takes on one of a specified set of values with a specified probability).

τ = interarrival time. Can be a function of the state of the system (e.g. number of jobs already in the system).

s = service time per job.

n = number of jobs in the system (a.k.a. queue length). Includes jobs currently receiving service as well as those waiting in the queue.

n_q = number of jobs waiting to receive service.

n_s = number of jobs receiving service.

r = response time or time in the system. Includes time waiting for service and time receiving service.

w = waiting time = time interval between arrival time and the instant the service begins.

Variables

λ = mean arrival rate = 1/E[τ]. E[τ]= $\sum_{i=1}^{n} p_i \tau_i = \int_{-\infty}^{+\infty} \tau \ f(\tau) d\tau$ where f(τ) = probability density function = dF(τ)/dτ and F(τ) = Cumulative Distribution Function of a random variable.

μ = mean service time per server = 1/E[s].

Relationships

Stability condition: $\lambda < m\mu$ where m = number of servers. Applies only to infinite buffer systems.

Number of jobs in system: $n = n_q + n_s$ => E[n]= E[n_q]+E[n_s]

If job service rate of each server is independent of the number of jobs in it's queue,

Cov(n_q, n_s)=0 and Var[n]=Var[n_q]+Var[n_s]

Job time in system: $r = w + s$ => E[r]=E[w]+E[s]

If job service rate is independent of the number of jobs in the queue,

Cov(w,s)=0 and Var[r]=Var[w]+Var[s]

Operational-related

Variables

λ_i = arrival rate for device i = number of arrivals / time = A/T.

X_i = throughput for device i = number of completions / time = C/T.

U_i = utilization for device i = busy time / total time = B/T.

S_i = mean service time for device i = total time served / number jobs served = B/C_i.

R_i = response time at ith device.

V_i = visit ratio of ith device = C_i/C_0 = device i jobs completed / total jobs completed.

Laws

Utilization law: $U_i = X_i S_i$

Forced flow law: $X_i = X V_i$ where X = jobs completed (system throughput) / total time = C_0/T, applies when there is job flow balance (A_i - C_i small wrt C_i)

Little's law: $Q_i = \lambda_i R_i$ for balanced flow $\lambda_i = X_i$

General response time law: $R = \sum_{i=1}^{M} R_i V_i$ where R is the system response time.

Interactive response time law: $R = (N/X) - Z$ where N = # users, Z = think time.

Bottleneck analysis: $X(N) \le min\{1/D_{max}, N/(D+Z)\}$ $R(N) \ge max\{D, ND_{max}-Z\}$ where $D = \Sigma D_i$ is the sum of total service demands on all devices except terminals. Ref Jain p. 563.

Conservation equations: for queues in which d and w exist $Q = \lambda d$ and $L = \lambda w$

manncreek.com