All You Need to Know About Queueing – The Most Useful Results of Queueing Analysis Applied in Telecommunications

Bob Warfield

May 2018

Purpose of This Presentation

- The purpose of this presentation is to summarise some of the most applicable results from queueing theory
- Familiarity with basic Probability Theory is assumed – Random Variable, Expectation, Variance, Standard Deviation, Normal (Gaussian) Distribution are all taken as known concepts
- In most cases no rigorous derivation of the results will be given – the conclusions are simply presented for you to learn, and practice applying to the solution of practical problems in telecommunication networks
- Often, intuitive explanations are offered, and in a few cases a derivation of sorts is laid out
- The derivations that are included are not intended to provide rigorous proof – rather they are included as mnemonic aids

The Beginning – Number of Arrivals in Time T

We will assume that the arrival process is "Random" or "Pure Chance". If the Random Variable $\mathbf{N}$ represents the number of arrivals in a deterministic time interval, T, then $\mathbf{N}$ has a Poisson Distribution, with "rate" λ:

$$P\{\mathbf{N} = n\} = e^{-\lambda T} \frac{(\lambda T)^n}{n!} \tag{1}$$

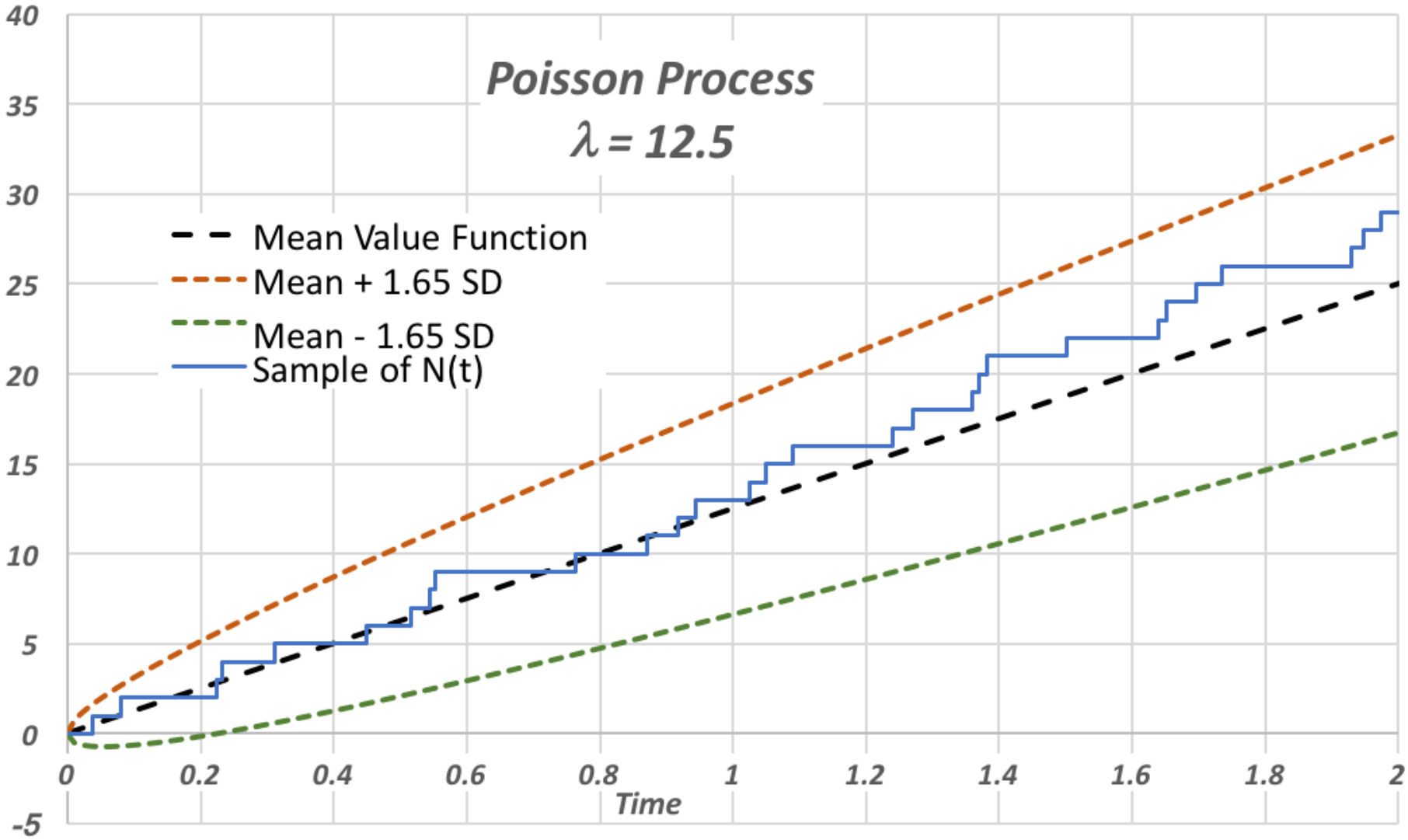
$$\text{for } n = 0, 1, 2, \ldots$$

$$\mathsf{E}\{\mathbf{N}\} = \lambda T \tag{2}$$

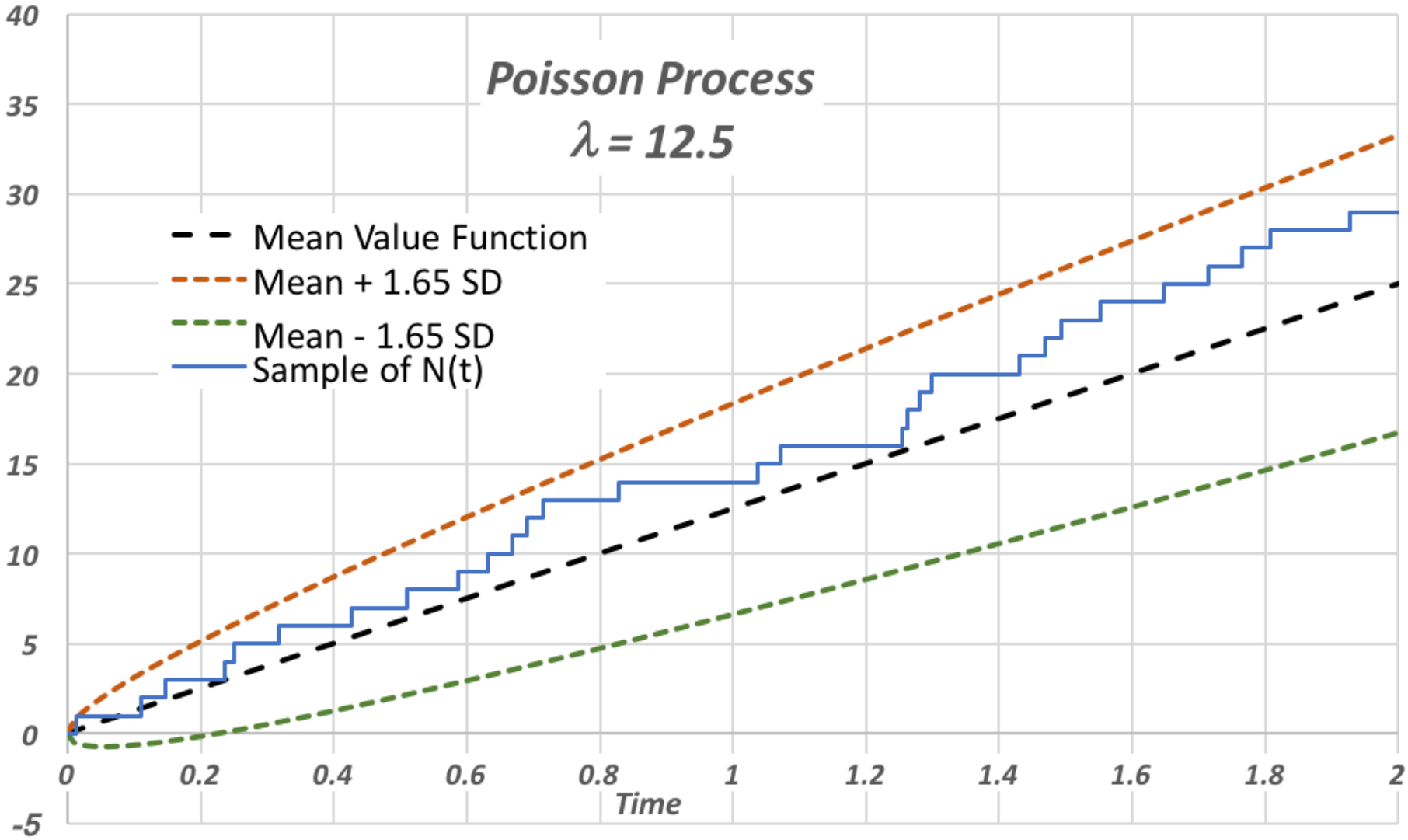
$$\mathsf{Var}\{\mathbf{N}\} = \lambda T \tag{3}$$

The fact that the Variance equals the Mean is a very special property of Poisson random variables

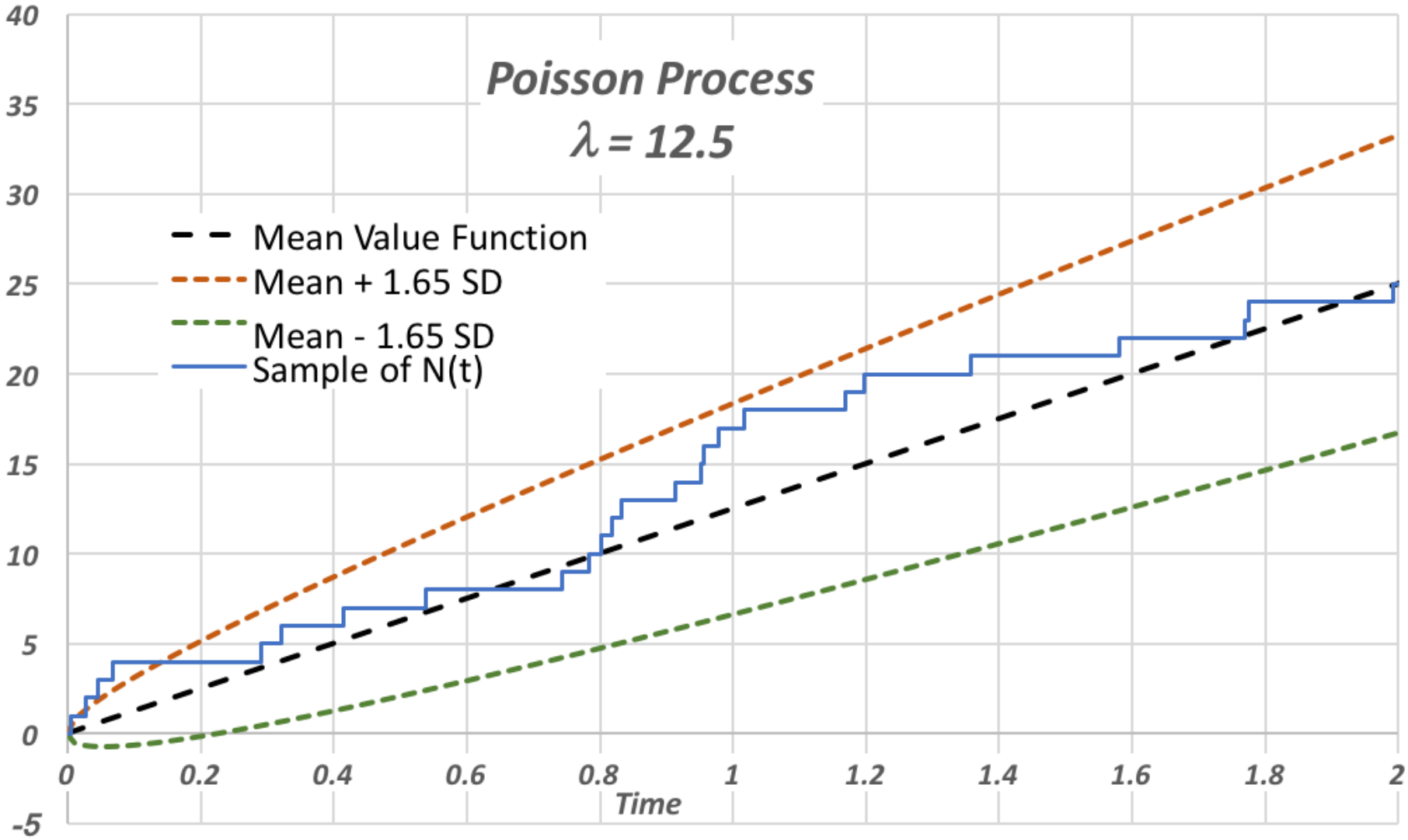
Samples of a Poisson Counting Process



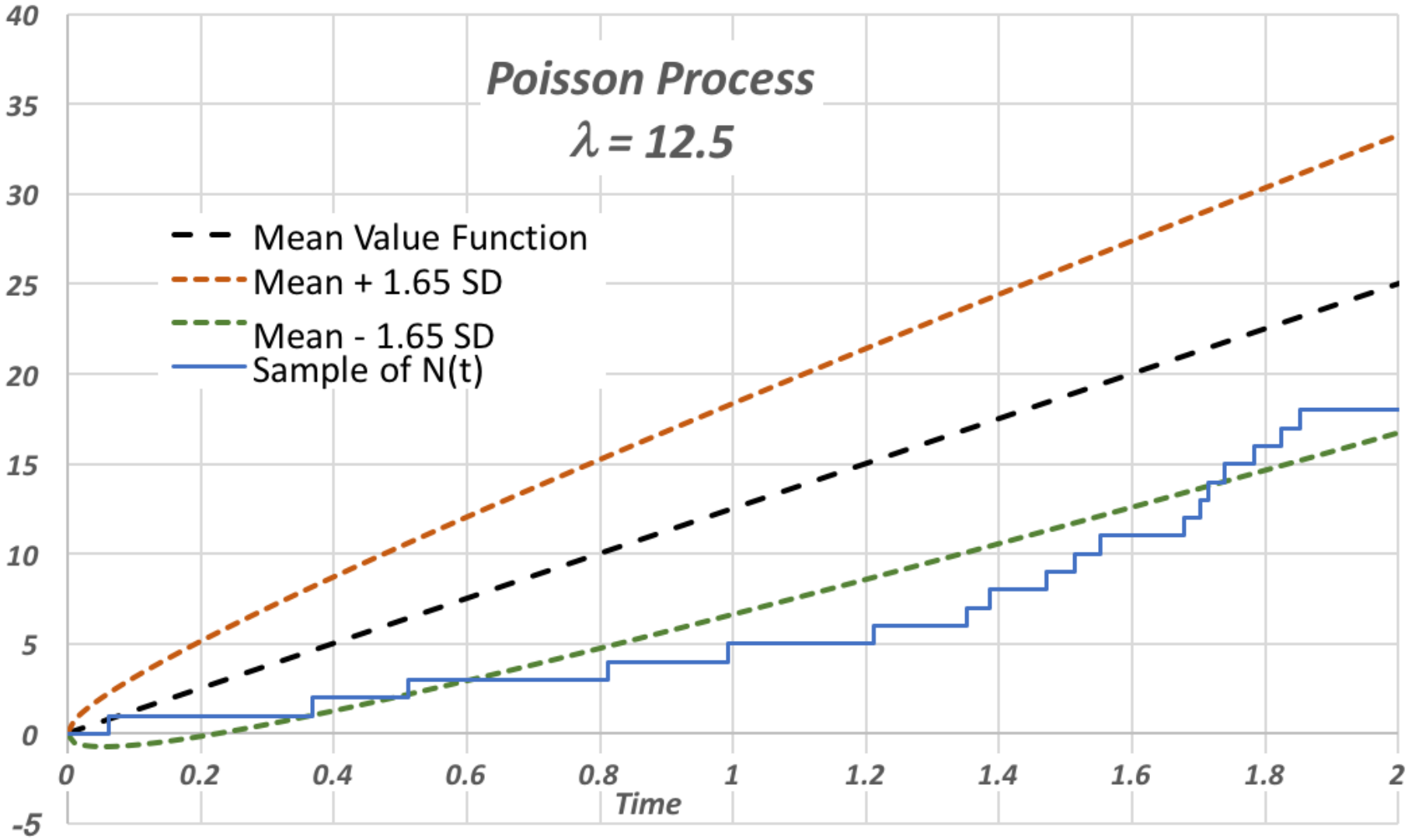
Samples of a Poisson Counting Process



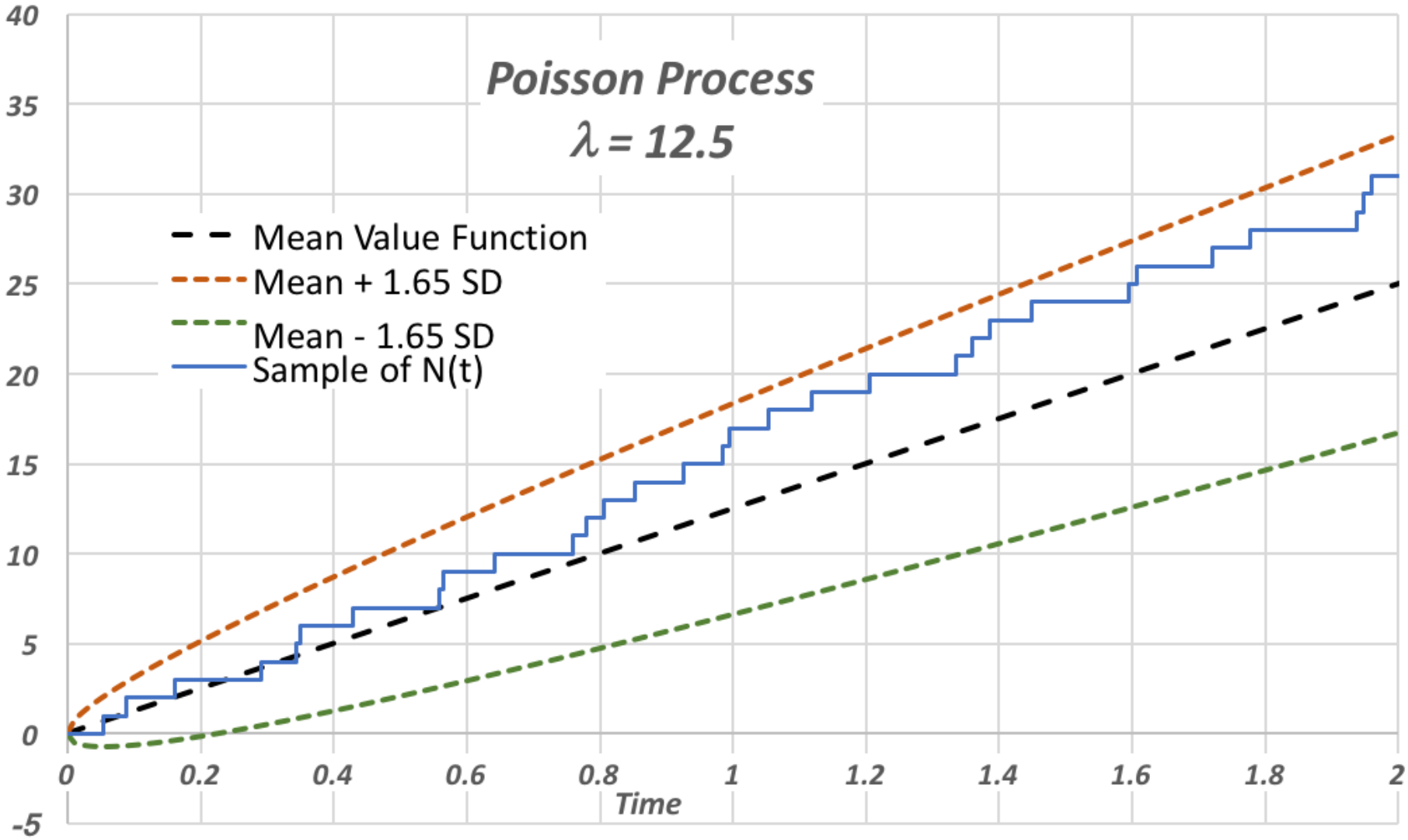
Samples of a Poisson Counting Process



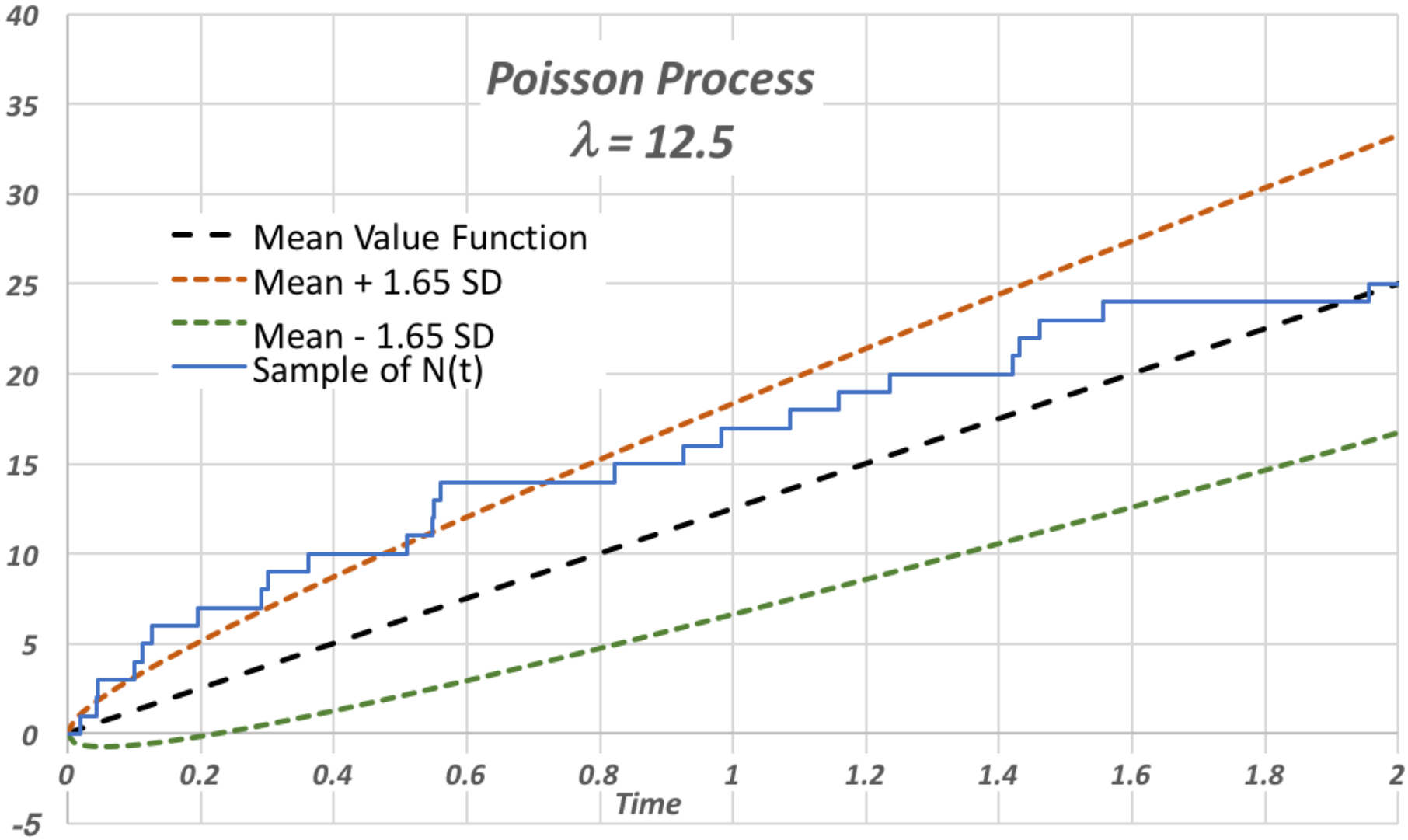
Samples of a Poisson Counting Process



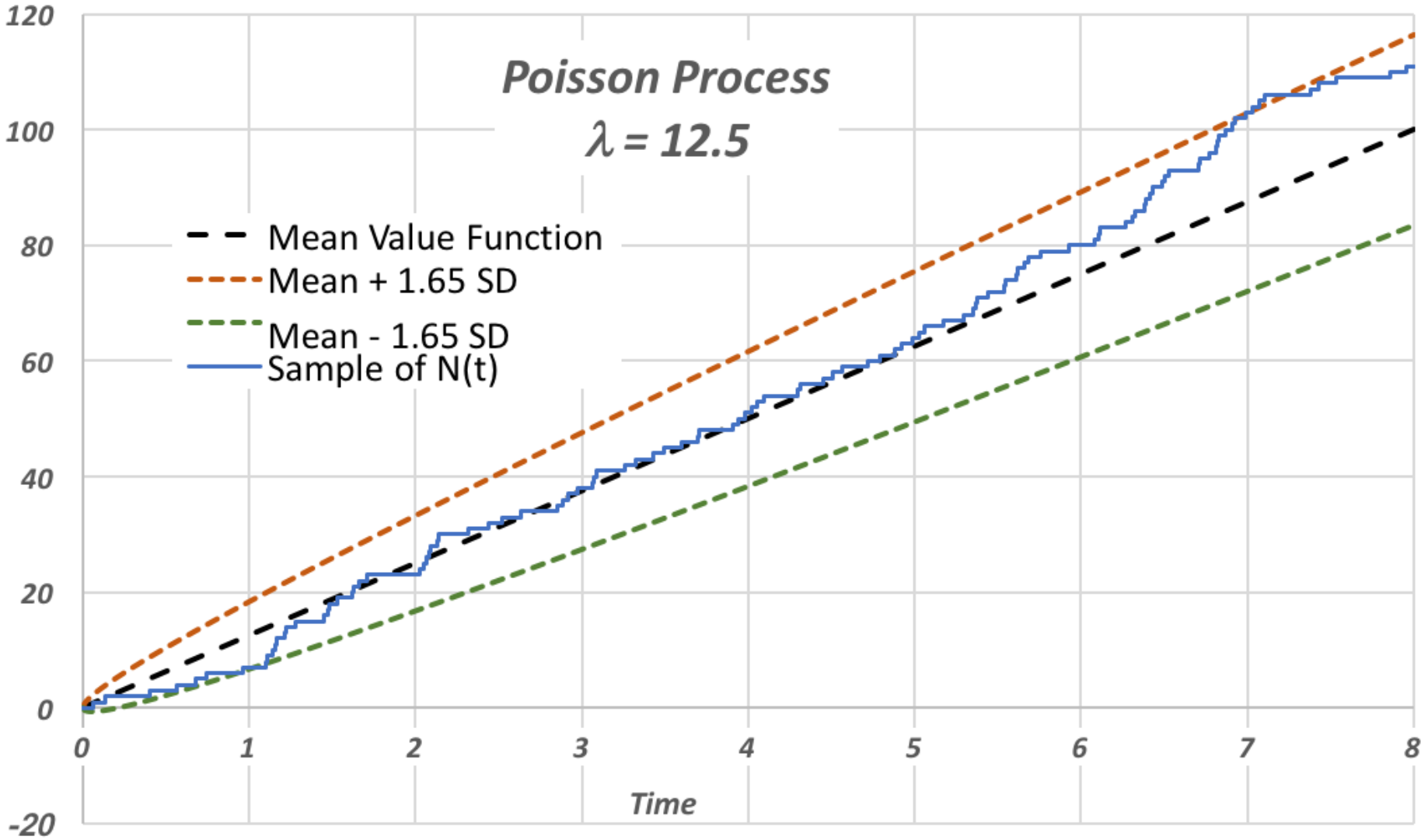
Samples of a Poisson Counting Process



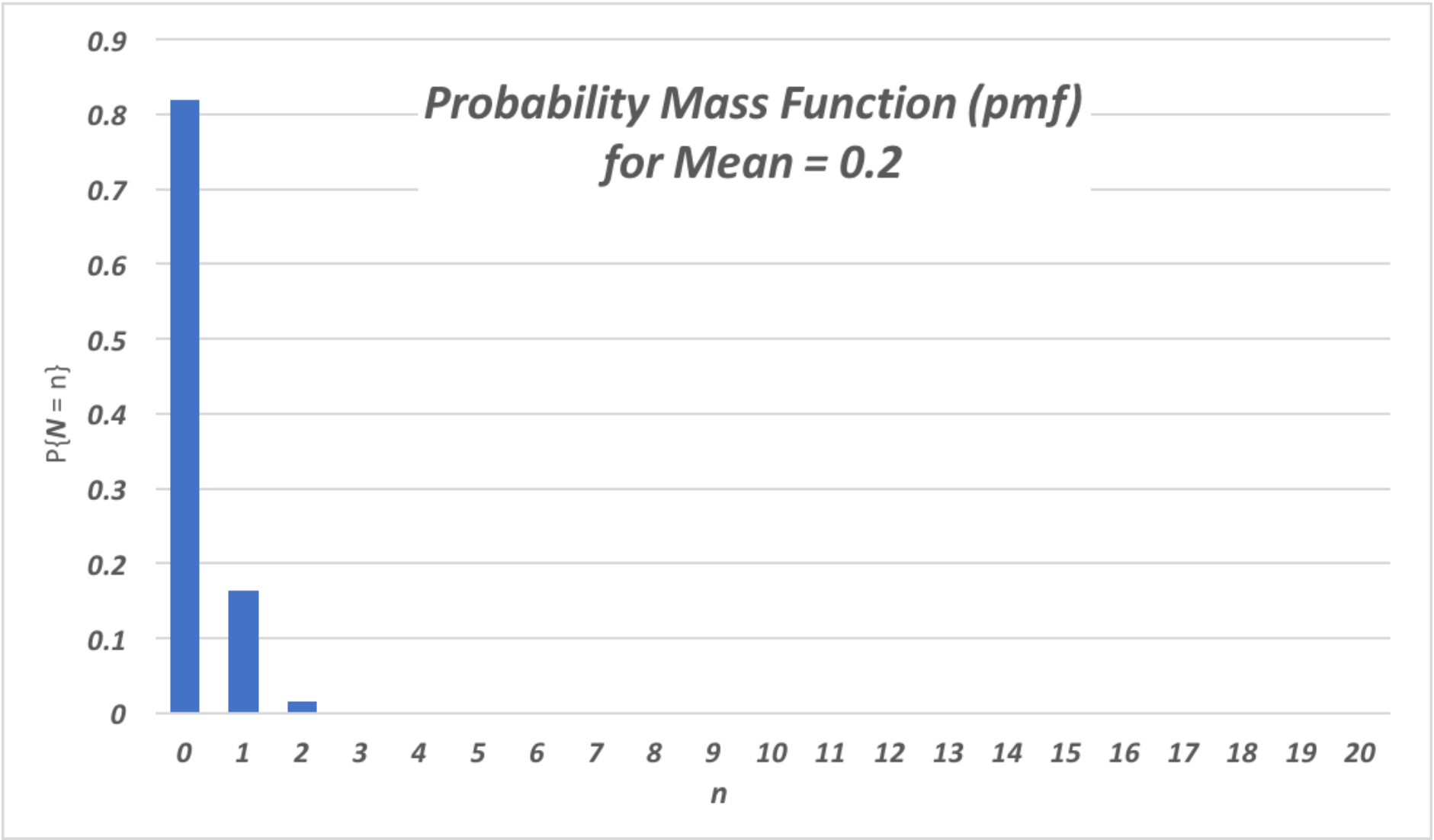
Samples of a Poisson Counting Process



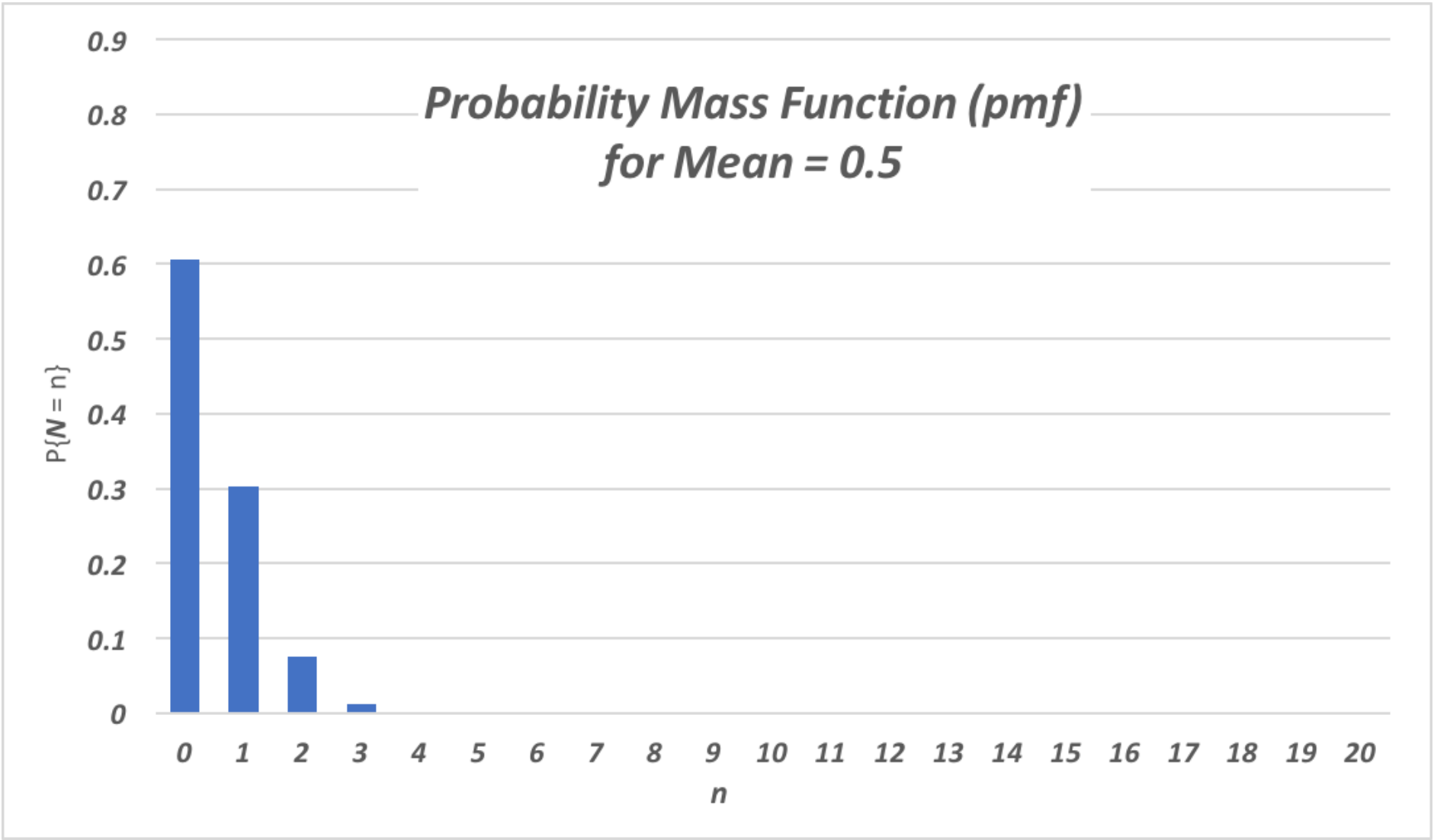
Observe Standard Deviation in Proportion to the Mean



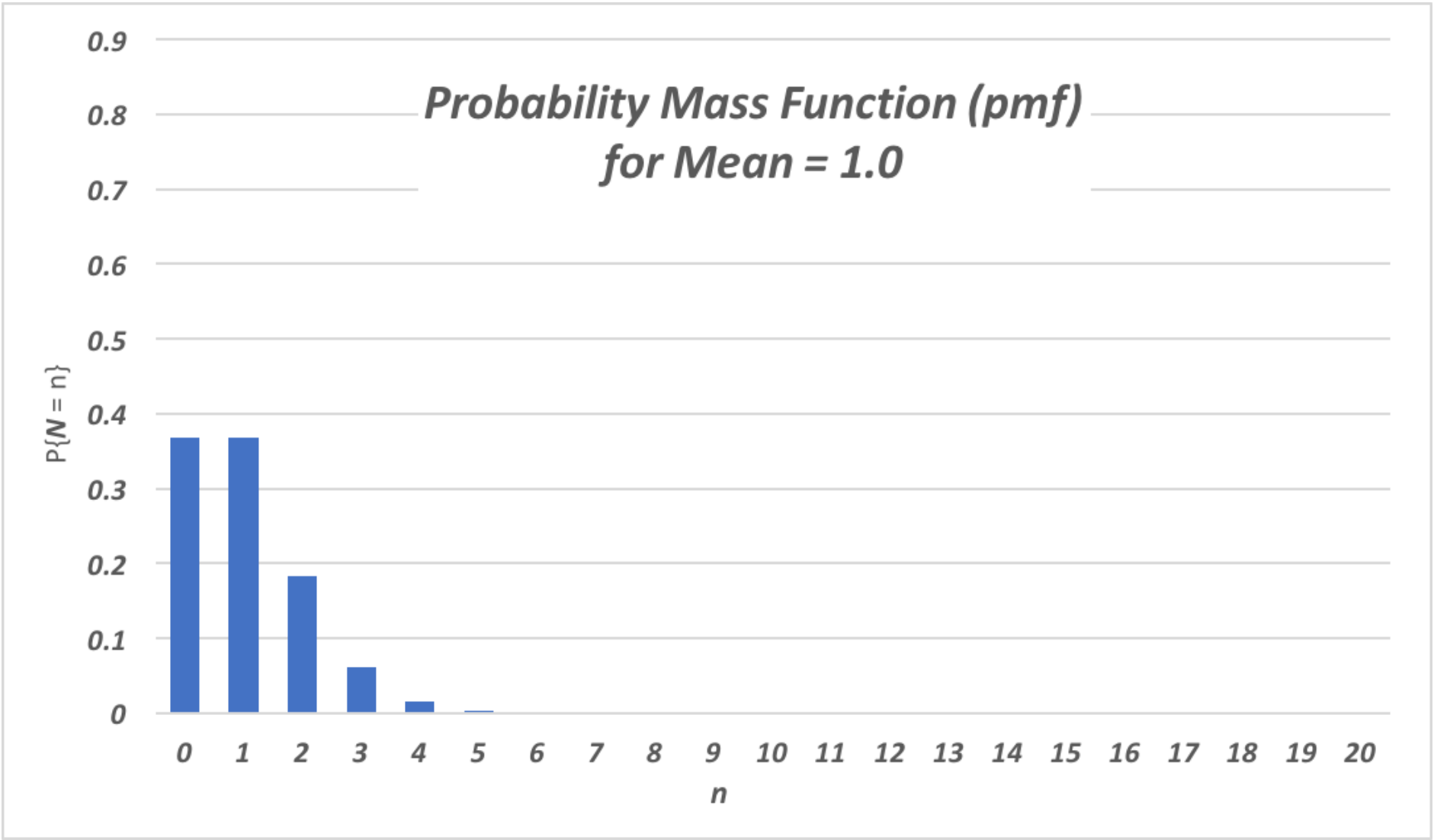
Probability Mass Function of $\mathbf{N}$



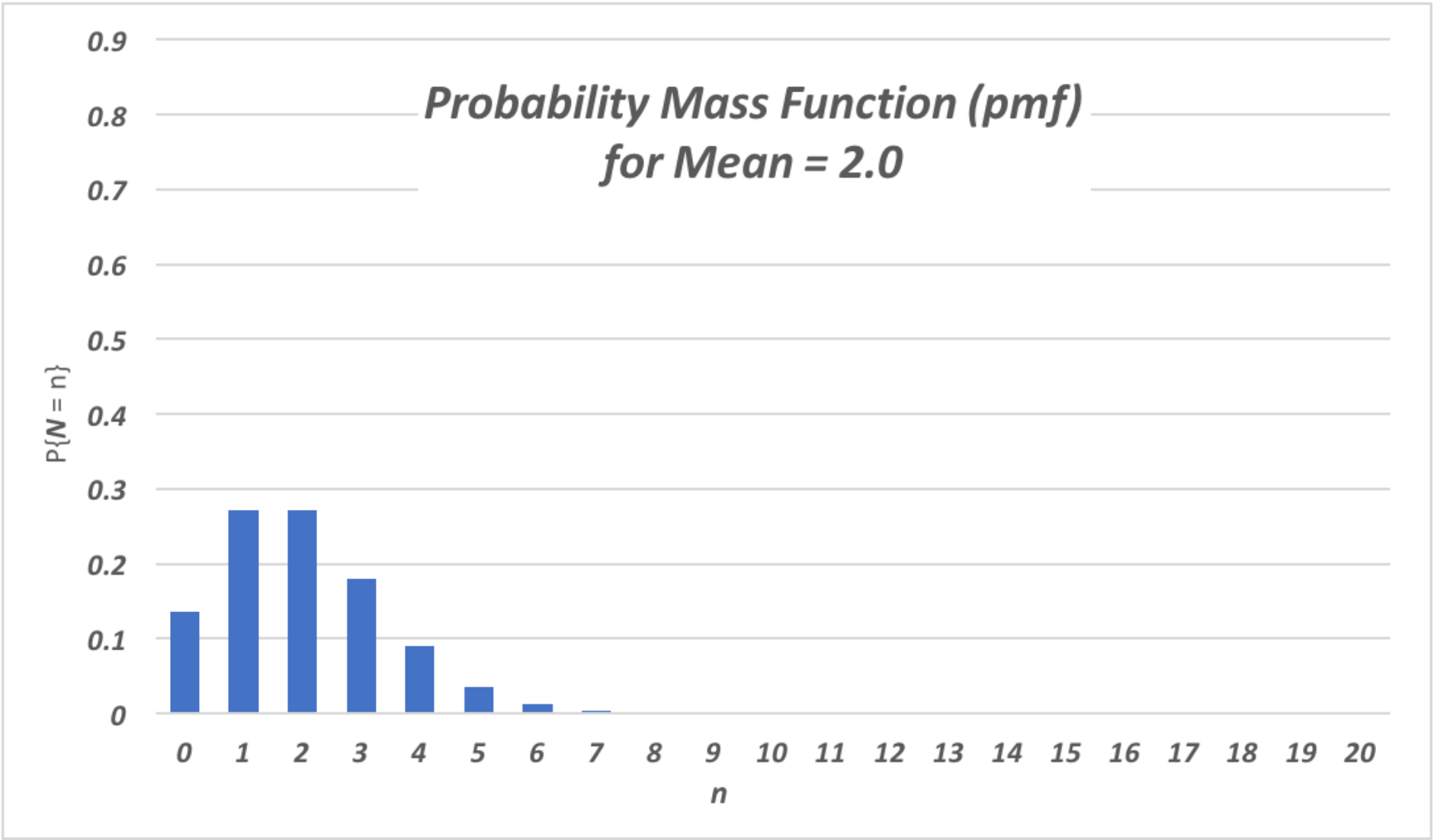
Probability Mass Function of $\mathbf{N}$



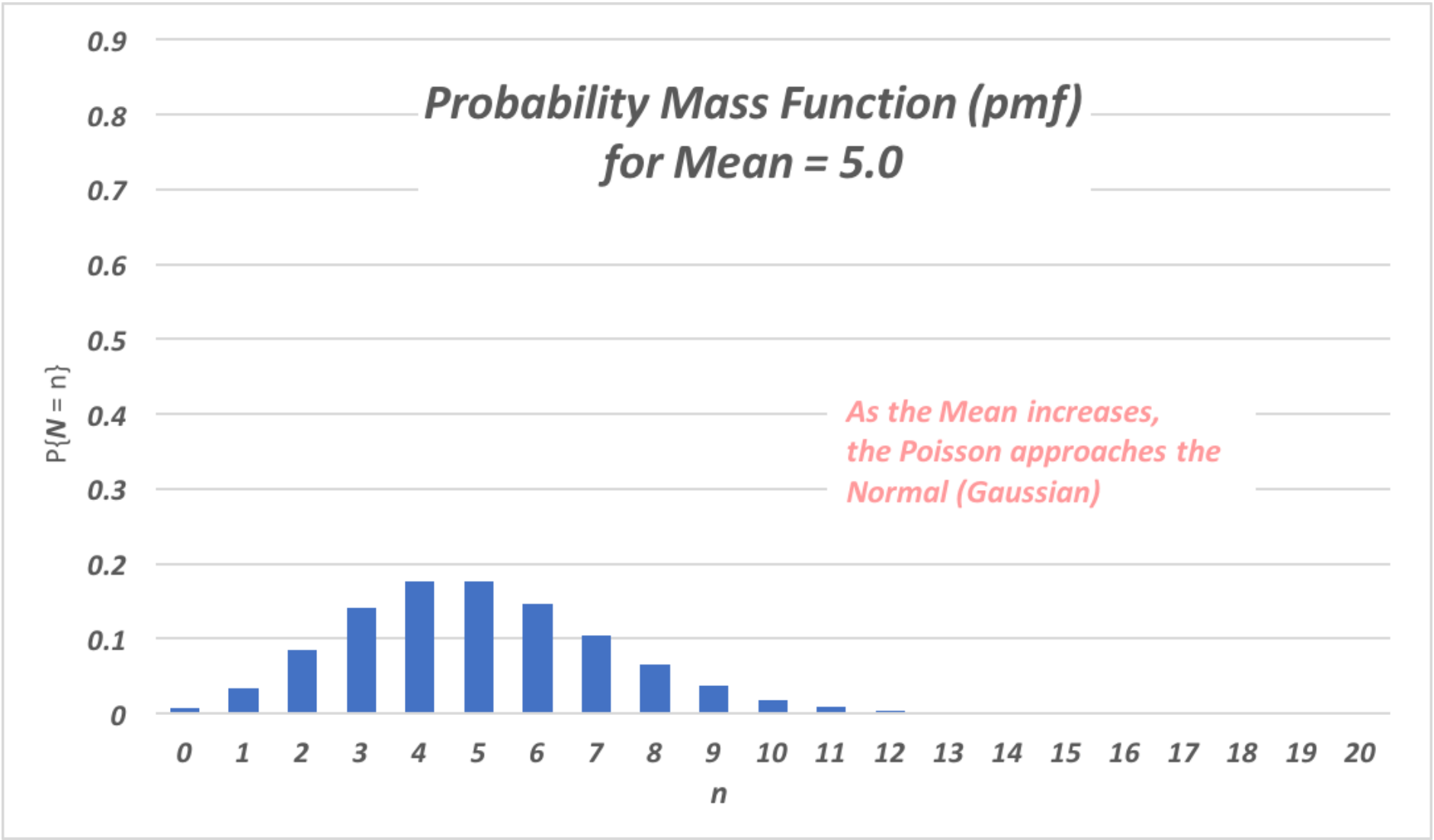
Probability Mass Function of $\mathbf{N}$



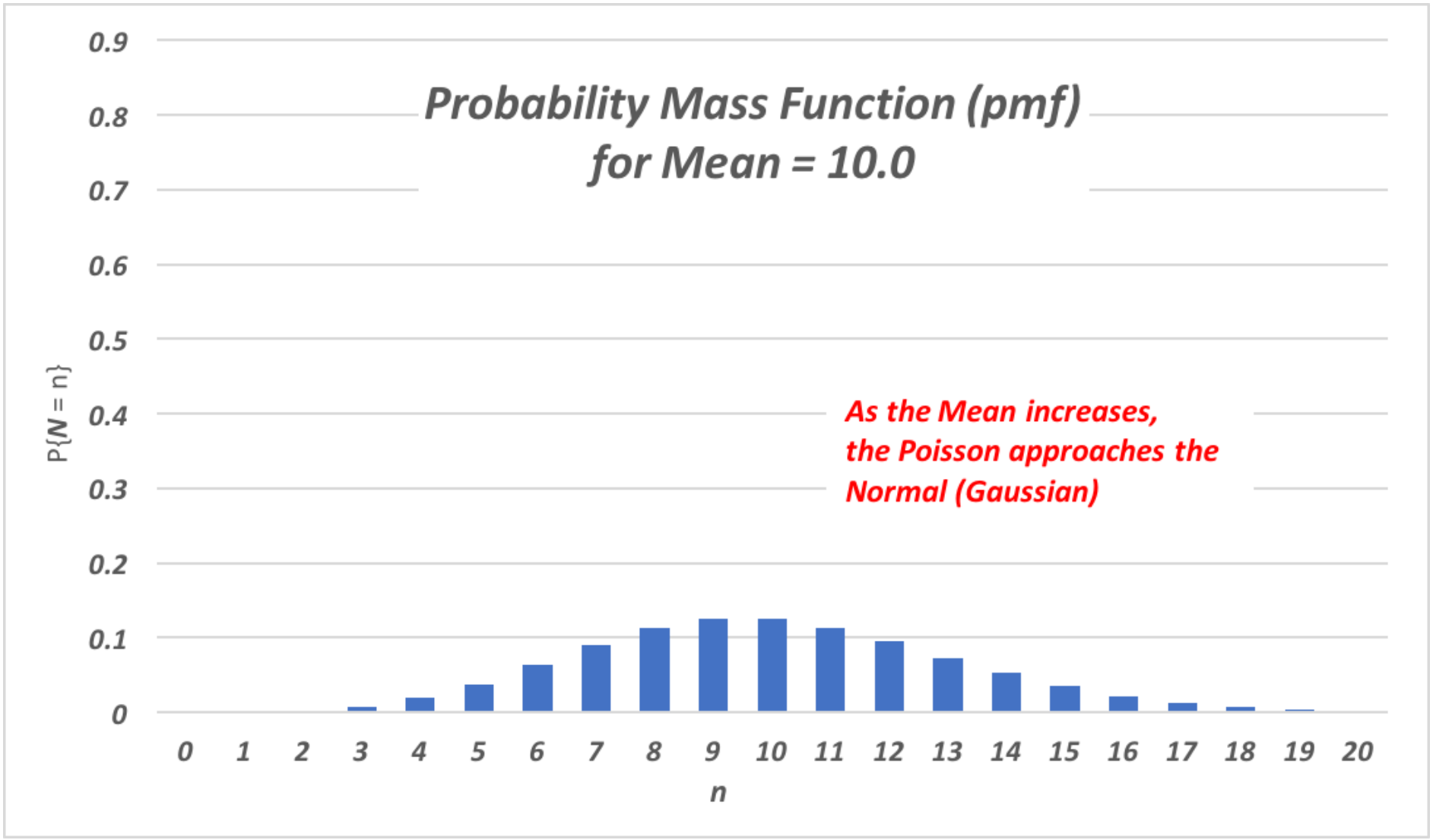
Probability Mass Function of $\mathbf{N}$



Probability Mass Function of $\mathbf{N}$



Probability Mass Function of $\mathbf{N}$



Time to the First Arrival

Starting at time zero, the time elapsed till the first arrival can also be thought of as the time for which $\mathbf{N} = 0$:

$$\begin{aligned} P\{\mathbf{T_a} > t\} &= P\{\mathbf{N}(t) = 0\} \\ &= e^{-\lambda t}\frac{(\lambda t)^0}{0!} \\ &= e^{-\lambda t} \end{aligned} \tag{4}$$

Hence the random time to the first arrival is a Negative Exponential Random Variable, and hence "memoryless".

As explained on the following slide, the time from any randomly selected instant to the next arrival has the same distribution as $\mathbf{T_a}$.

Memoryless Property of Negative Exponential

Suppose you have been waiting up to time t, with no arrival in sight. What is the probability that you will have to wait an additional time τ? Using Bayes' Rule:

$$\begin{aligned} P\{\mathbf{T_a} > t+\tau | \mathbf{T_a} > t\} &= \frac{P\{\{\mathbf{T_a} > t+\tau\} \cap \{\mathbf{T_a} > t\}\}}{P\{\mathbf{T_a} > t\}} \\ &= \frac{P\{\mathbf{T_a} > t+\tau\}}{P\{\mathbf{T_a} > t\}} \\ &= \frac{e^{-\lambda(t+\tau)}}{e^{-\lambda t}} = e^{-\lambda\tau} \end{aligned} \tag{5}$$

Hence the time to the next arrival, having waited up to time t, is simply a Negative Exponential, with the same distribution as $\mathbf{T_a}$.

This is termed the "memoryless" property
of Negative Exponential Random Variables.

"Rate," Mean, and Variance of Arrival Time

Suppose you have been waiting up to time t without any arrival. The probability of an arrival in the next (infinitesmial) interval dt is:

$$1 - e^{-\lambda dt} \rightarrow \lambda dt \tag{6}$$

This explains why λ is termed the "rate" of the Poisson Process – it is the probability per unit time of seeing an arrival, independently of how long you have already been waiting.
The Mean and Variance of $\mathbf{T_a}$ are given by:

$$\begin{aligned} \mathsf{E}\{\mathbf{T_a}\} &= \frac{1}{\lambda} \\ \mathsf{Var}\{\mathbf{T_a}\} &= \left(\frac{1}{\lambda}\right)^2 \end{aligned} \tag{7}$$

Note that the standard deviation of $\mathbf{T_a}$ equals the mean.

Percentiles of $\mathbf{T_a}$

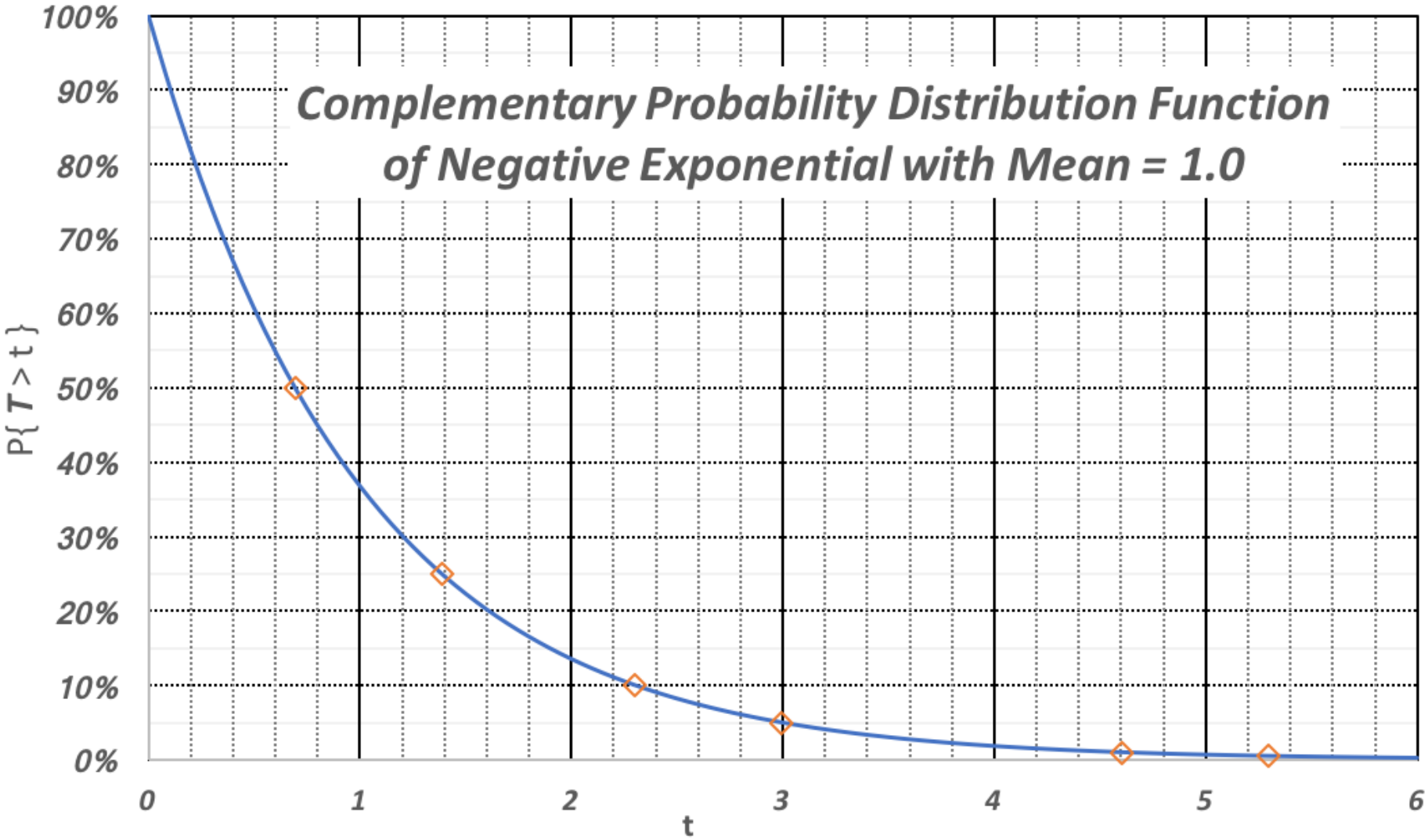
One common example of exponentially distributed lifetimes is provided by radioactive decay. If you were observing a single radioactive isotope with rate of decay λ (per sec), what would be the half-life of the isotope?

$$\begin{aligned} P\{\mathbf{T_a} > T_h\} &= e^{-\lambda T_h} = 0.5 \\ \Rightarrow T_h &= \frac{\ln(2)}{\lambda} \approx \mathbf{0.7} \times \mathrm{E}\{\mathbf{T_a}\} \end{aligned} \tag{8}$$

Similarly the 90th percentile for T_a is approximately $\mathbf{2.3} \times \mathrm{E}\{\mathbf{T_a}\}$.

t	$P\{\mathbf{T_a} > t\}$	Note:
$0.7\times$Mean	50%	Remember this,
$2.3\times$Mean	10%	and this.
$1.4\times$Mean	25%	2×0.7
$3.0\times$Mean	5%	$2.3 + 0.7$
$4.6\times$Mean	1%	2×2.3

Complementary PDF



Packet Length

- The Negative Exponential is such a wonderful distribution, we make the simplifying assumption that packet length takes that distribution – whether it is close to the truth or not. We even ignore the fact that packet length is an integer and the Negative Exponential distribution applies to Reals.
- Before considering the "rate" of serving a packet, we use $\mathbf{L_p}$ as the random packet length, in bits, and we denote its mean value by L_p (the only difference is the font).

$$P\{\mathbf{L_p} > l\} = e^{-(l/L_p)} \tag{9}$$

$$\mathsf{E}\{\mathbf{L_p}\} = L_p \tag{10}$$

$$\mathsf{Var}\{\mathbf{L_p}\} = L_p^2 \tag{11}$$

Time Taken to Serve a Packet

The time to serve a packet depends on the bitrate, B, at which it is served. Transmission Engineers use the term "Emission Time". We denote the random quantity $\mathbf{T_e}$ and the Mean Value T_e.

$$T_e = \frac{L_p}{B}$$
$$P\{\mathbf{T_e} > t\} = e^{-t/T_e} \tag{12}$$

Teletraffic Engineers talk in terms of "Holding Time," and often use the symbol h in place of T_e. Mathematicians who are experts in Queueing Theory do not use T_e. Instead they refer to the "Service Rate," μ, where

$$\mu = \frac{1}{T_e} \tag{13}$$

For "convenience" we will use all three terminologies in different contexts.

M/M/1 Queueing Model

- Kendall's Notation for queueing models describes the input, or arrival process, the service process, the number of servers, and the capacity of the system overall (servers plus waiting places).
- "M" stands for memoryless (or Markovian). This is the model we have already used for arrivals and also service times, hence: "M/M".
- The "/1", with no symbol or number following, indicates a single server and infinite capacity queue.
- Of course there are many things wrong with this model as a description of a router in the Internet - that is obvious.
- Nonetheless the model gives such simple results we will persevere with it in the hopes of getting at least some insight, even if very approximate, using simple "back of the envelope" formulas.
- After that, we will consider the problem of actually measuring parameters such as λ and μ from real-world measurements

Occupancy

The occupancy of the server is defined as the probability it is busy at a random instant of time. Denoting the mean occupancy by ρ we have

$$\rho = \frac{\lambda}{\mu} = \lambda h = \frac{\lambda L_p}{B} \tag{14}$$

The term λL_p is the number of packets arriving per sec times the length of the packets in bits. Hence it is the total bitrate coming in. The server has a capacity of B. Dividing the incoming bitrate by the bitrate of the server gives its average occupancy.

We must have $\rho < 1$ for the simple equations for a steady-state solution of the M/M/1 model to be valid. That is, the average incoming bitrate (λL_p) must be less than B, the bitrate capacity of the server.

Probability Distribution of The State

The probability the server is free at a random instant of time is $(1-\rho)$. Suppose we observe the M/M/1 system for a long period of time, T_t, and observe that the total time the server is free is T_0. The limit of $\frac{T_0}{T_t}$ is $(1-\rho)$. If we denote the number of packets in the system at a random instant of time as $\mathbf{J}$, the number of transitions from $\{\mathbf{J}=0\}$ to $\{\mathbf{J}=1\}$ is just the mean arrival rate times the duration $(=\lambda T_0)$.
Similarly, the number of transitions from $\{\mathbf{J}=1\}$ to $\{\mathbf{J}=0\}$ is just service rate times duration $=\mu T_1$, where T_1 is the total time spent in the state $\{\mathbf{J}=1\}$. Therefore

$$\frac{\mu T_1}{T_t} = \frac{\lambda T_0}{T_t} \tag{15}$$

$$P\{\mathbf{J}=1\} = \rho P\{\mathbf{J}=0\} \tag{16}$$

And by a similar process:

$$P\{\mathbf{J}=j\} = \rho P\{\mathbf{J}=(j-1)\} \tag{17}$$

Moments of the State

The variable $\mathbf{J}$ is termed the "state" of the system. Denoting $P\{\mathbf{J} = j\}$ by P_j,

$$P_j = (1 - \rho)\rho^j \tag{18}$$

Which is the Geometric Distribution – a discrete version of the Negative Exponential. The moments are given by:

$$\begin{aligned} \mathsf{E}\{\mathbf{J}\} &= \frac{\rho}{1 - \rho} \\ \mathsf{Var}\{\mathbf{J}\} &= \frac{\rho}{(1 - \rho)^2} \end{aligned} \tag{19}$$

Time in the System – Little's Law

The time a packet spends in the system, comprising any delay plus one service time, is denoted $\mathbf{T_s}$. We can find the mean value of $\mathbf{T_s}$ by taking the point of view of a single "typical" packet. Suppose a packet arrives at time zero, and departs at time T_s. Just at the moment of its departure, we pose the question: "how many packets have arrived since this one?" The answer is λT_s. But this also answers the question: "how many packets remain in the system after the packet departs?" which equals $\mathsf{E}\{\mathbf{J}\}$. Hence

$$\begin{aligned} \lambda T_s &= \mathsf{E}\{\mathbf{J}\} \\ &= \frac{\rho}{1-\rho} \\ T_s &= \frac{1/\mu}{1-\rho} = \frac{h}{1-\rho} \end{aligned} \tag{20}$$

This is a very important result for Time in the System (also called "Response Time" or "Sojourn Time").

Distribution of Time in the System

For any packet arriving while the system is in state j, the time in the system is dependent on the value of j. By the law of Total Probability, the probability density function of $\mathbf{T_s}$ is given by

$$f(t_s) = \sum_{j=0}^{j=\infty} P_j f(t_s|j) \tag{21}$$

You may (or may not) be surprised to learn that $\mathbf{T_s}$ turns out to be Negative Exponential! Hence

$$\begin{aligned} P\{\mathbf{T_s} > t\} &= e^{-t/T_s} \\ Var(\mathbf{T_s}) = T_s^2 &= \left[\frac{h}{1-\rho}\right]^2 \end{aligned} \tag{22}$$

Time Spent in the Queue

The time spent in the queue is denoted $\mathbf{T_q}$. Only some of the arriving packets go into the queue – for all the rest $\mathbf{T_q} = 0$. The probability of being delayed is given by

$$P\{\mathbf{J} > 0\} = P\{\mathbf{T_q} > 0\} = \rho \tag{23}$$

So the average arrival rate into the queue (delayed packets) is $\rho\lambda$. The average number of packets in the system is $\mathrm{E}\{\mathbf{J}\}$, but that includes an average of ρ packets in service (not in the queue). From Little's Law, denoting $\mathrm{E}\{\mathbf{T_q}|\mathbf{T_q} > 0\}$ by T_q:

$$\begin{aligned} \rho\lambda T_q &= \frac{\rho}{1-\rho} - \rho \\ \lambda T_q &= \frac{1}{1-\rho} - 1 \\ T_q &= \frac{h}{1-\rho} \end{aligned} \tag{24}$$

Verification from Mean Sojourn Time

We can verify that

$$\mathsf{E}\{\mathbf{T_s}\} = \mathsf{E}\{\mathbf{T_q}|\mathbf{T_q} > 0\} \tag{25}$$

by the following derivation:

$$\begin{aligned}\mathsf{E}\{\mathbf{T_s}\} &= h + P\{\mathbf{T_q} > 0\}\mathsf{E}\{\mathbf{T_q}|\mathbf{T_q} > 0\} \\ &= h + \rho\frac{h}{1-\rho} \\ &= \frac{h}{1-\rho}\end{aligned} \tag{26}$$

Hence we enjoy the very convenient result that

$$T_s = T_q = \frac{h}{1-\rho} \tag{27}$$

Variance of Delays

For the Negative Exponential Random Variable, $\mathbf{T_s}$,

$$\begin{aligned} \mathrm{E}(\mathbf{T_s}) &= \frac{h}{1-\rho} \\ \mathrm{Var}(\mathbf{T_s}) &= [\mathrm{E}(\mathbf{T_s})]^2 \\ &= \left[\frac{h}{1-\rho}\right]^2 \end{aligned} \tag{28}$$

This variance is the sum of the variance of the service time (which is h^2) and the Variance of the Queueing Time, $\mathbf{T_q}$, is therefore

$$\mathrm{Var}(\mathbf{T_q}) = \left[\frac{h}{1-\rho}\right]^2 - h^2 \tag{29}$$

Variance of $\mathbf{T_q}$ as a Mixture

It can be shown that, conditional on $\mathbf{T_q} > 0, \mathbf{T_q}$ is Negative Exponential with mean T_q and variance T_q^2. That means that the RV $\mathbf{T_q}$ is a mixture of a Negative Exponential with probability ρ, and the fixed value 0 with probability $(1-\rho)$. The variance of this mixture is given by:

$$\begin{aligned} \text{Var}\left(\mathbf{T_q}\right) &= \text{E}\left[\text{Var}(\mathbf{T_q})\right] + \text{Var}\left[\text{E}(\mathbf{T_q})\right] \\ &= \rho T_q^2 + \rho\left[T_q(1-\rho)\right]^2 + (1-\rho)\left[\rho T_q\right]^2 \end{aligned} \tag{30}$$

Substituting $T_q = \frac{h}{1-\rho}$ and simplifying yields, as before:

$$\text{Var}\left(\mathbf{T_q}\right) = \left[\frac{h}{1-\rho}\right]^2 - h^2 \tag{31}$$

The method of derivation given on the previous slide seems much more intuitive. This derivation just serves to confirm it.

Free Capacity

Returning to our result for mean time in the system,

$$\begin{aligned} T_s = T_q &= \frac{h}{1-\rho} \\ &= \frac{L_p}{B(1-\lambda(L_p/B))} \\ &= \frac{L_p}{B-\lambda L_p} \end{aligned} \tag{32}$$

Therefore we define Free Capacity, F by:

$$\begin{aligned} F &\triangleq B - \lambda L_p \\ &= B(1-\rho) \end{aligned} \tag{33}$$

Relationship to Circuit Switching

If a circuit with dedicated bitrate B_c carries packets with mean length L_p, the mean time to serve each packet is simply

$$\frac{L_p}{B_c} \tag{34}$$

If we set the bitrate of a dedicated circuit equal to the Free Capacity for a M/M/1 system, $B_c = F$, then the average delay through the two systems will be the same.

Delay Jitter

- Using the results for the variance and distributions of delays, we can quantify delay jitter as well as mean delay. We know the variance of the time in the system, and if there are several such systems in series we simply add the means and add the variances (assuming independence). Most packet delays will be less than mean plus 3 (for example) standard deviations.
- In the case where a particular stream of traffic has constant packet length, we wish to use only the variance of the Queueing Delays. For this we use:

$$\text{Var}\left(\mathbf{T_q}\right) = \left[\frac{h}{1-\rho}\right]^2 - h^2 \tag{35}$$

Resources

- Wikipedia gives a useful summary of the properties of all common Probability Distributions, also descriptions of the terms used
- The classic textbook on Queueing Theory by Cooper is available for free online