What we know so far: response time in M/G/1/FCFS

$$\mathbf{E}[T_Q] \;=\; \frac{\rho}{1-\rho} \cdot \frac{\mathbf{E}\left[S^2\right]}{2\mathbf{E}[S]}$$

Here's another formula (see p. 404 in your book):

$$\mathbf{Var}(T_Q) \;=\; \mathbf{E}[T_Q]^2 + \frac{\lambda \mathbf{E}\left[S^3\right]}{3(1-\rho)}$$

In fact, we can derive any moment of T_Q, e.g., $\mathbf{E}\left[T_Q^k\right]$, for any integer k.

(We can do this by deriving the Laplace Transform of T_Q and then differentiating that transform. See Chpts 25, 26 of your book, or come see me.)

What we know so far: response times with scheduling

We have covered a bunch of scheduling policies for the M/G/1 queue:

- P-Prio
- NP-Prio
- SJF
- SRPT
- LCFS
- P-LCFS

For each, we've seen a formula for $\mathbf{E}[T]$.

Turns out that we can also derive a formula for $\mathbf{Var}(T)$ for each of these.

In fact, we can derive any moment of T, e.g., $\mathbf{E}\left[T^k\right]$, for any integer k.

(We can do this by deriving the Laplace Transform of T and then differentiating that transform. See some of the exercises in Chpts 29-33 or come see me.)

But what if we want the tail of response time?

GOAL:

$$\mathbf{P}\{T > t\}$$

Here t is the SLO, e.g. 0.5s (below 0.5s is not noticeable).

Question: Can we convert knowledge of $\mathbf{E}[T]$ and $\mathbf{Var}(T)$ to $\mathbf{P}\{T > t\}$?

INTUITIONS:

- Why is $\mathbf{E}[T]$ not really enough?
- Why is knowing $\mathbf{Var}(T)$ a lot better?

OUTLINE FOR TODAY:

- Markov's Inequality: From mean to tail
- Chebyshev's Inequality: From mean & variance to tail
- Central Limit Theorem: From mean & variance to aggregate tail
- Beyond Tails: percentiles, like T_{99}.

Markov's Inequality

Thm: Let X be a non-negative r.v. with finite mean. Then $\forall a > 0$,

$$\mathbf{P}\{X \geq a\} \leq \frac{\mathbf{E}[X]}{a}$$

Question: The mean class grade is 40%. What fraction of the class has a grade > 80%?

Question: Why does the above make obvious sense?

Chebyshev's Inequality

Thm: Let X be an r.v. with finite mean and variance. Then $\forall a > 0$,

$$\mathbf{P}\left\{\Big|X - \mathbf{E}[X]\Big| \geq a\right\} \leq \frac{\mathbf{Var}(X)}{a^2}$$

Question: The mean class grade is 40%, with a std of 10%. What fraction of the class has a grade > 80%?

Central Limit Theorem

I've modified the actual theorem to make it easier. In practice, this is all your need.
The formal statement is more complex (see p. 61 of your book if you want the gory details).

Thm: Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.'s $\sim X$.
All have mean $\mu = \mathbf{E}[X]$ and variance $\sigma^2 = \mathbf{Var}(X)$.

We're not assuming anything about the distribution X.

Let

$$S_n = \sum_{i=1}^{n} X_i \ .$$

As n gets high, we have the following *approximation*:

$$S_n \sim \text{Normal}(\underline{\hspace{3cm}}, \underline{\hspace{3cm}}) \ .$$

$$\frac{S_n}{n} \sim \text{Normal}(\underline{\hspace{3cm}}, \underline{\hspace{3cm}}) \ .$$

I will give you 3 examples of where this comes up!

Central Limit Theorem, cont.

<u>Thm</u>: Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.'s $\sim X$.
All have mean $\mu = \mathbf{E}[X]$ and variance $\sigma^2 = \mathbf{Var}(X)$.

Let

$$S_n = \sum_{i=1}^{n} X_i \ .$$

As n gets high, we have the following *approximation*:

$$S_n \sim \text{Normal}(\underline{\hspace{2cm}}, \underline{\hspace{2cm}}) \ .$$

$$\frac{S_n}{n} \sim \text{Normal}(\underline{\hspace{2cm}}, \underline{\hspace{2cm}}) \ .$$

Question: There are 25 students in my class. Their scores are independent. Each student is well-modeled by a mean of 40% and a std of 10%. What is the chance that the class average is $> 50\%$?

Central Limit Theorem, cont.

Thm: Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.'s $\sim X$.
All have mean $\mu = \mathbf{E}[X]$ and variance $\sigma^2 = \mathbf{Var}(X)$.

Let

$$S_n = \sum_{i=1}^{n} X_i \ .$$

As n gets high, we have the following *approximation*:

$$S_n \sim \text{Normal}(\underline{\hspace{2cm}}, \underline{\hspace{2cm}}) \ .$$

$$\frac{S_n}{n} \sim \text{Normal}(\underline{\hspace{2cm}}, \underline{\hspace{2cm}}) \ .$$

Question: We are trying to transmit a signal. There are 100 independent sources of noise, each making a low amount of noise distributed Uniformly between -1 and 1. If the absolute value of the total amount of noise > 10, then the signal gets corrupted. What is the probability that the signal is corrupted?

Central Limit Theorem, cont.

Thm: Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.'s $\sim X$.
All have mean $\mu = \mathbf{E}[X]$ and variance $\sigma^2 = \mathbf{Var}(X)$.

Let

$$S_n = \sum_{i=1}^{n} X_i \ .$$

As n gets high, we have the following *approximation*:

$$S_n \sim \text{Normal}(\underline{\hspace{3cm}}, \underline{\hspace{3cm}}) \ .$$

$$\frac{S_n}{n} \sim \text{Normal}(\underline{\hspace{3cm}}, \underline{\hspace{3cm}}) \ .$$

Question: Google capacity provisioning problem. Describe Mor's double-estimator solution!

Suppose I want to know T_{99}

Question: Assume I know how to estimate $\mathbf{P}\{T > t\}$ for any t. Can I get T_{99}?

Hint: What does T_{99} mean?

Question: Suppose what we have is only an *upper bound* on $\mathbf{P}\{T > t\}$? Does this lead to an upper bound on T_{99}? Or a lower bound on T_{99}?

Yet another queueing theory tool: Setup times

The photocopier queue ...

- Always shuts off when not in use.
- There's a setup time when first person comes in.
- Setup affects others who follow, but not clear how much it affect overall mean

Let's draw a picture:

- M/M/1.
- $\rho = 0.5$.
- $S \sim \text{Exp}(1)$ minute.
- Setup time $I = 10$ minutes.

Question: Should setup time affect queue with high load more?

Question: Should setup time affect queue with high C_S^2 more?

Yet another queueing theory tool: Setup times

Thm: (Chpt 27 of your book)

$$\mathbf{E}[T_Q]^{M/G/1/setup} = \mathbf{E}[T_Q]^{M/G/1} + \frac{2\mathbf{E}[I] + \lambda \mathbf{E}[I^2]}{2(1 + \lambda \mathbf{E}[I])}$$

Let's compute for above example:

- M/M/1.
- $\rho = 0.5$.
- $S \sim \text{Exp}(1)$ minute.
- Setup time $I = 10$ minutes.

More complex models:

1. Delayed-Off model: Wait some time before shutting off
2. M/G/k, rather than M/G/1.

Setup times in multiserver systems are a new area in Queueing theory. Recent work:

- Gandhi et al. "Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward" SIGMETRICS 2013.
- Jalani Williams et al. "The M/M/k with Deterministic Setup Times" – SIGMETRICS 2023.