

Lecture Notes on
Probabilistic Task Durations in Projects

Yakov Ben-Haim
Yitzhak Moda'i Chair in
Technology and Economics
Faculty of Mechanical Engineering
Technion — Israel Institute of Technology
Haifa 32000 Israel
<http://tx.technion.ac.il/~yakov>
yakov@technion.ac.il

A Note to the Student: These lecture notes are not a substitute for the thorough study of books. These notes are no more than an aid in following the lectures.

Contents

1	Basic Problem	2
2	Project Reliability with a Global Time Buffer: Theory	3
3	Probabilistic Approaches	6
4	Central Limit Theorem	8
5	T is Normal	10
6	Weibull Distribution	16
7	T is an Extreme Value	18
8	c_m is Normal	20
9	c_m is a Convolution	22

1 Basic Problem

¶ A project is characterized by:

- A flow-chart of tasks.
- Uncertainty in the duration of each task.
(Alternatively: cost uncertainty.)
- Global requirement: complete project on time.

¶ Questions:

- How risky is the project given random variation of task times?
- How can the risk be reduced?
- How robust is the project to uncertainty in the probabilistic models?
- How can the robustness be increased (and the risk reduced)?
 - Re-structuring the project.
 - On-line monitoring.
 - Gathering information.
- How opportune is the project?
Can windfalls be exploited?

2 Project Reliability with a Global Time Buffer: Theory

¶ Consider a project whose task flow chart is:

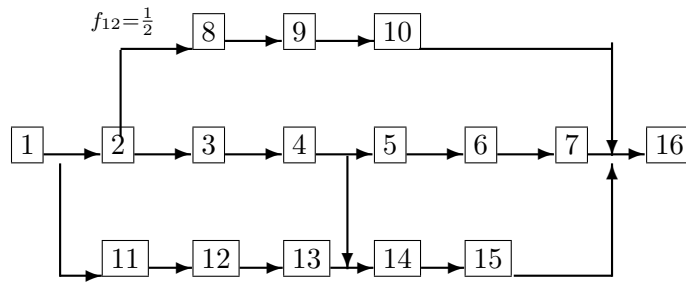


Figure 1: A 16-activity project schedule.

This project has 4 task paths:

Path 1: $1 \rightarrow 2 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 16$.

Path 2: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 16$.

Path 3: $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 14 \rightarrow 15 \rightarrow 16$.

Path 4: $1 \rightarrow 11 \rightarrow 12 \rightarrow 13 \rightarrow 14 \rightarrow 15 \rightarrow 16$.

¶ In order to answer the questions in section 1 on page 2 we need:

- Dynamic model: describing the task-path structure
and its relation to total project duration.
- Failure criterion.
- Uncertainty model.

¶ We first consider the **dynamic model**.

t_n = unknown duration of n th task, $n = 1, \dots, N$.

$t = (t_1, \dots, t_N)^T$

There are M paths.

f_{mn} = fractional participation of task n in path m .

m : path.

n : task.

In path m , the task following task n

begins when task n is fraction f_{mn} complete.

¶ E.g., in path 1 of fig. 1:

task 8 begins when task 2 is 1/2 complete:

$f_{12} = 0.5$.

¶ The duration of the m th path, c_m ,

equals the sum of the durations of **all tasks**

weighted by their fractional participations in path m :

$$c_m = \sum_{n=1}^N f_{mn} t_n, \quad m = 1, \dots, M \quad (1)$$

For instance, the duration of the 1st path is:

$$c_1 = 1 \cdot t_1 + \frac{1}{2} \cdot t_2 + 1 \cdot t_8 + 1 \cdot t_9 + 1 \cdot t_{10} + 1 \cdot t_{16} \quad (2)$$

Define F = matrix of participation factors $f_{mn} \in \mathfrak{R}^{M \times N}$.

For instance, for fig. 1:

$$F = \begin{pmatrix} 1 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad (3)$$

¶ Now the relation between task- and path-durations is:

$$c = Ft \tag{4}$$

The **dynamic model** is the duration of the longest path:

$$T = \|c\| = \max_{1 \leq m \leq M} |c_m| = \max_{1 \leq m \leq M} \sum_{n=1}^N f_{mn} t_n \tag{5}$$

Note that $\|c\|$ is in fact a vector norm, sometimes called the “zero norm”.

¶ The **failure criterion**:

the project fails if the duration of the longest path exceeds a critical value:

$$T > T_c \tag{6}$$

3 Probabilistic Approaches

¶ In this section we briefly outline four probabilistic approaches to the analysis of risk. We study these four approaches in subsequent sections.

¶ The analysis depends on three components:

- System model.
- Failure criterion.
- Probability (or other uncertainty) model.

¶ **System model**, which is T in eq.(5), p.5, the total duration of the project, which we re-write without absolute values:

$$T = \max_{1 \leq m \leq M} c_m = \max_{1 \leq m \leq M} \sum_{n=1}^N f_{mn} t_n \quad (7)$$

¶ **Failure criterion:** The project fails if the total project duration exceeds a critical value:

$$T > T_c \quad (8)$$

The probabilistic risk is the probability of failure:

$$P_f = \text{Prob}(T > T_c) \quad (9)$$

From the point of view of probabilistic design, we require that the probability of failure be no greater than a critical value:

$$P_f \leq P_c \quad (10)$$

or equivalently:

$$\text{Prob}(T > T_c) \leq P_c \quad (11)$$

¶ **Probabilistic models.** We will consider several different probability models:

1. The total project time T is the linear superposition of a large number of independent identically distributed (iid) random variables. Thus T is normally distributed, as implied by the central limit theorem.
2. T is the maximum of a large number of iid random variables. Thus T has an extreme-value distribution.
3. The duration of the n th task, t_n , is the linear superposition of a large number of iid random variables. Thus t_n is normally distributed, as implied by the central limit theorem. The duration of the m th task-path, c_m , is a linear combination of task times. Thus c_m is also normally distributed.
4. c_m is the linear superposition of independent random variables t_1, \dots, t_N . Thus the probability density function (pdf) of c_m is the convolution of the pdfs of t_1, \dots, t_N .

¶ Note the following **assumptions** inherent in some of these probability models:

- Asymptotics: large number of random variables.
- Independence of random variables: no common-mode failures.
- Identical distributions: the underlying random variables are derived from a single common factor of randomness.
- We know pdfs.
- Linearity.

4 Central Limit Theorem

¶ Source material:

- DeGroot, Morris H., *Probability and Statistics*, 2nd ed., Addison-Wesley, Reading, MA, 1986, pp.274–281.

- Kendall, Maurice, Alan Stuart and J. Keith Ord, *The Advanced Theory of Statistics*. Vol. 1: *Distribution Theory*, 5th ed., Oxford University Press, New York, 1987, § 7.35–7.40,

¶ **Central limit theorem: iid variables.** If x_1, x_2, \dots are iid random variables with mean μ and finite variance σ^2 , then:

$$\bar{x}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^N x_i \quad (12)$$

tends to a normal distribution as $N \rightarrow \infty$.

Note: this does not depend on how the x_i are distributed, or even if they are discrete or continuous random variables.

Specifically:

$$\lim_{N \rightarrow \infty} \text{Prob} \left[\frac{\sqrt{N}(\bar{x}_N - \mu)}{\sigma} \leq x \right] = \Phi(x) \quad (13)$$

where $\Phi(x)$ is the cdf of the standard normal distribution defined below.

¶ Definition: t has a **normal distribution** if its pdf is:

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(t - \mu)^2}{2\sigma^2} \right), \quad -\infty < t < \infty \quad (14)$$

We write:

$$t \sim \mathcal{N}(\mu, \sigma^2) \quad (15)$$

- The **standard normal distribution** is $t \sim \mathcal{N}(0, 1)$. Its pdf is:

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{t^2}{2} \right), \quad -\infty < t < \infty \quad (16)$$

The cdf is denoted $\Phi(t)$.

• **Standardization.** If $T \sim \mathcal{N}(\mu, \sigma^2)$, then:

$$F_T(t) = \text{Prob}(T \leq t) = \text{Prob}\left(\frac{T - \mu}{\sigma} \leq \frac{t - \mu}{\sigma}\right) \quad (17)$$

But:

$$\frac{T - \mu}{\sigma} \sim \mathcal{N}(0, 1) \quad (18)$$

Thus:

$$F_T(t) = \Phi\left(\frac{t - \mu}{\sigma}\right) \quad (19)$$

¶ **Central limit theorem: independent but not necessarily identically distributed variables.** Consider a sequence of independent random variables x_i with means μ_i and finite variances σ_i^2 . Define the random variables:

$$y_n = \frac{\sum_{i=1}^n x_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \quad (20)$$

Note that $E(y_n) = 0$ and $\text{var}(y_n) = 1$. Under certain conditions, y_n is asymptotically normal.

• Specifically:

If:

$$E(|x_i - \mu_i|^3) < \infty \quad \text{for all } i \quad (21)$$

and **if:**

$$\lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E(|x_i - \mu_i|^3)}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0 \quad (22)$$

then:

$$\lim_{n \rightarrow \infty} \text{Prob}(y_n \leq x) = \Phi(x) \quad (23)$$

5 T is Normal

¶ We consider the first probabilistic model described on p.7:

The total project time T is the linear superposition of a large number of iid random variables. Thus T is normally distributed, as implied by the central limit theorem:

$$T \sim \mathcal{N}(\mu, \sigma^2) \quad (24)$$

We assume that:

- We know the mean μ and variance σ^2 .
- We know how μ and σ depend on the project design (participation matrix F , resource allocations, etc.).

¶ The probability of failure, eq.(9), p.6, is:

$$P_f = \text{Prob}(T > T_c) = \text{Prob}\left(\underbrace{\frac{T - \mu}{\sigma}}_z \geq \underbrace{\frac{T_c - \mu}{\sigma}}_{z_c}\right) = 1 - \Phi(z_c) \quad (25)$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. z_c is a known value and z has a standard normal distribution:

$$z \sim \mathcal{N}(0, 1) \quad (26)$$

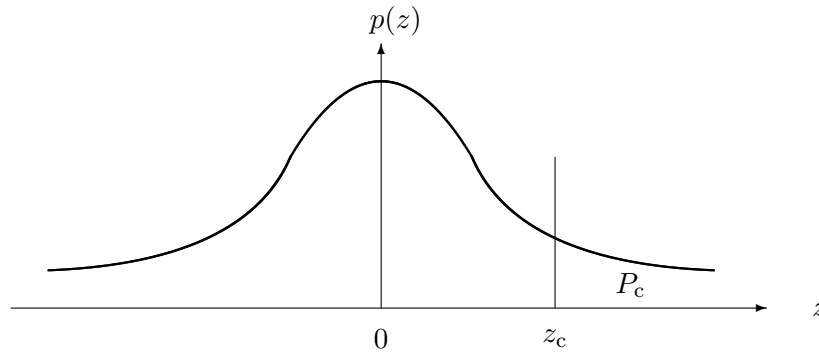


Figure 2: Sketch of probability density illustrating failure quantile z_c in eq.(28).

¶ **Failure quantile.** The probabilistic performance requirement for the project, eq.(11) p.6, is that the probability of failure be less than a critical value:

$$\text{Prob}(T > T_c) \leq P_c \quad (27)$$

With eq.(25) this becomes:

$$\Phi(z_c) \geq 1 - P_c \quad (28)$$

Thus z_c must be no less than the $(1 - P_c)$ quantile of the standard normal distribution, as illustrated in fig. 2, where z_c is:

$$z_c = \frac{T_c - \mu}{\sigma} \quad (29)$$

¶ Design implications.

- Let q denote the various project parameters which the manager can influence: topology of flow chart, resource allocation, managerial attention, etc.

- We have assumed that we know how the project design influences the mean and variance of T . That is we know the functions $\mu(q)$ and $\sigma^2(q)$. Thus we know how the design influences z_c , so we know the function $z_c(q)$:

$$z_c(q) = \frac{T_c - \mu(q)}{\sigma(q)} \quad (30)$$

- In light of eq.(28) we see that we prefer q over q' if the quantile is greater with q than with q' :

$$q \succ q' \quad \text{if} \quad z_c(q) > z_c(q') \quad (31)$$

- Recall that $P_f = 1 - \Phi(z_c)$, eq.(25), p.10. We can assess how much we prefer q over q' in eq.(31) in terms of how much smaller the probability of failure is with q than with q' . That is, we greatly prefer q over q' if:

$$\Phi[z_c(q)] - \Phi[z_c(q')] \gg 0 \quad (32)$$

Conversely, we only slightly prefer q over q' if:

$$1 \gg \Phi[z_c(q)] - \Phi[z_c(q')] > 0 \quad (33)$$

- Carrying eq.(31) to the limit, the “best design” maximizes z_c :

$$q^* = \arg \max_q z_c(q) \quad (34)$$

This is equivalent to minimizing the probability of failure:

$$q^* = \arg \min_q P_f(q) \quad (35)$$

- Later we will consider the question: how **robust**, to uncertainty in the underlying models, is this optimization? We will be especially interested in uncertainty in the function $z_c(q)$. We will find that the robustness is **zero**. This will give special meaning to the concept of “best design”.

¶ Cost vs. Performance.

- The quantile function $z_c(q)$ can be used to explore the trade-off and/or trade-on between cost and performance. Let $\$(q)$ denote the monetary cost of a project design specified by parameters q . Assume that we know this function.

- Recall that the quantile z_c is a monotonic function of the allowed probability of time-overrun P_c : P_c gets smaller as z_c gets larger, as in fig. 2 and eq.(28), p.11.

- We assume that we know:

- The cost function $\$(q)$.
- The quantile function $z_c(q)$.
- The probability distribution of this quantile, $\Phi(z_c)$
- The relation between z_c and the critical probability: $\Phi(z_c) = 1 - P_c$.

Thus we can plot cost versus critical time-overrun probability P_c . This curve can appear as any of the possibilities shown in figs. 3–5 on p.14. The project design q is varying along each of the curves. These curves give the manager a quantitative tool for choosing between designs.

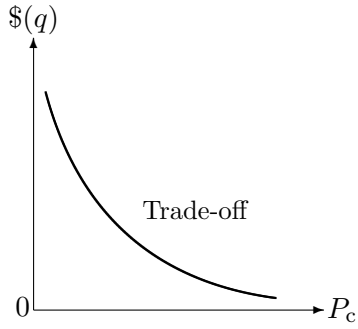


Figure 3: Project cost vs. critical probability of time overrun: trade-off. q varies along the curve.

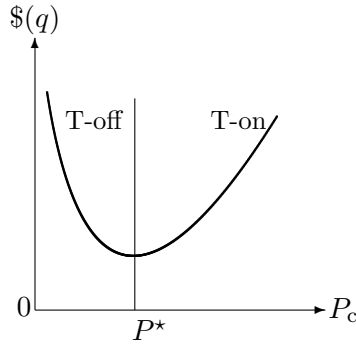


Figure 4: Project cost vs. critical probability of time overrun: trade-off and trade-on. q varies along the curve.

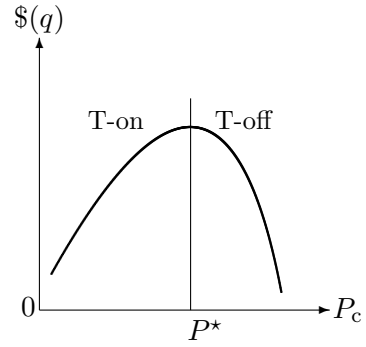


Figure 5: Project cost vs. critical probability of time overrun: trade-off and trade-on. q varies along the curve.

- Fig. 3 shows pure trade-off between money and performance. The critical probability P_c can be reduced by investing more resources.
- Fig. 4 shows both trade-off and trade-on:
 - Trade-off to the left of the vertical line: P_c is reduced by investing more money.
 - Trade-on to the right of the vertical line: P_c is reduced by investing less money.
 - If P^* is low enough critical probability of time overrun, then the design at this point on the curve should probably be adopted.
- Fig. 5 shows both trade-off and trade-on:
 - Trade-off to the right of the vertical line: P_c is reduced by investing more money.
 - Trade-on to the left of the vertical line: P_c is reduced by investing less money.
 - Even if P^* is acceptably low critical probability of time overrun, the design at this point on the curve is maximally expensive and should probably not be adopted. Designs to the left are cheaper and have lower critical probability.

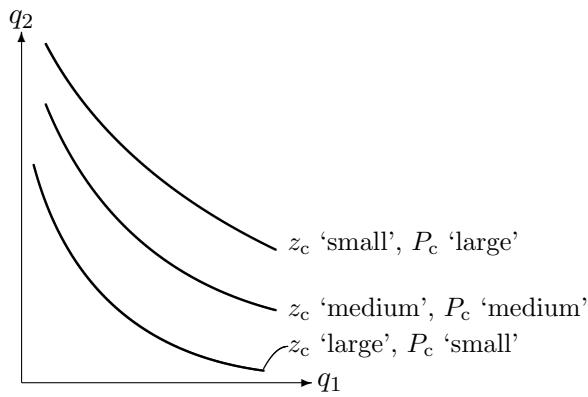


Figure 6: Design curves of constant $z_c(q)$ and $P_c(q)$.

¶ **Trade-off between design variables.** The quantile function $z_c(q)$ can be used to explore the trade-off between design variables.

- Consider two design variables, q_1 and q_2 , which may be hours of managerial attention and quality of equipment respectively.
- Fig. 6 shows schematic plots of constant failure quantile $z_c(q)$ as q_1 and q_2 vary.
- These curves show how resources can be re-allocated between different functions while keeping the overall project reliability constant.

6 Weibull Distribution

¶ Source material:

- Lawless, J.F., *Statistical Models and Methods for Lifetime Data*, 1982, John Wiley, New York, chapter 4.

- Høyland, A. and M. Rausand, 1994, *System Reliability Theory: Models and Statistical Methods*, Wiley, New York. 1994, pp37–40.

¶ Definition: a random variable T has a Weibull distribution if its cdf is:

$$F_T(t) = \text{Prob}(T \leq t) = \begin{cases} 1 - e^{-(\lambda t)^\alpha}, & t > 0 \\ 0 & t \leq 0 \end{cases} \quad (36)$$

- The pdf of T is:

$$f_T(t) = \frac{dF_T(t)}{dt} = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha}, \quad t > 0 \quad (37)$$

¶ The Weibull distribution arises as a **limit distribution**.

- Weibull describes the distribution of the **least** or **greatest** of a large number of identically distributed non-negative random variables.

- The Weibull distribution is thus sometimes called the **weakest link** distribution.

¶ **Example.** Consider a system with N components, whose lifetimes are identically distributed.

The system fails as soon as one component fails.

Let T_1, \dots, T_N be the TTFs of the components.

The TTF of the system is:

$$T = \min_{1 \leq n \leq N} T_n \tag{38}$$

The distribution of T will tend to a Weibull distribution for large N . ■

7 T is an Extreme Value

¶ We consider the second probabilistic model described on p.7:

T is the maximum of a large number of iid random variables. Thus T has an extreme-value distribution.

¶ Recall the system model, eq.(7), p.6:

$$T = \max_{1 \leq m \leq M} c_m \quad (39)$$

- **If:**

- The number of paths, M , is very large.
- The path durations c_m are independent random variables.

Then T will have an extreme-value distribution such as a Weibull distribution.

- The first assumption will often hold. The second assumption is problematic since the task-paths usually overlap. We adopt both assumptions.

¶ **Probability of failure.** We assume that we know the parameters of the distribution of T . E.g., λ and α in eq.(36), p.16, if T has a Weibull distribution.

- This allows us to evaluate the probability of failure as in eq.(9), p.6:

$$P_f = \text{Prob}(T > T_c) \quad (40)$$

¶ **Project analyses.** We assume that we know how the parameters λ and α depend on project-design variables q . That is, we know the functions $\lambda(q)$ and $\alpha(q)$.

- This allows us to perform the:
 - Design analysis as before, p.12.
 - Cost vs. performance analysis as before, p.13 and figs. 3–5, p.14.
 - Design-variable trade-off analysis as before, fig. 6, p.15.

8 c_m is Normal

¶ We consider the third probabilistic model described on p.7:

The duration of the n th task, t_n , is the linear superposition of a large number of iid random variables. Thus t_n is normally distributed, as implied by the central limit theorem.

The duration of the m th task-path, c_m , is a linear combination of task times, eq.(1), p.4:

$$c_m = \sum_{n=1}^N f_{mn} t_n, \quad m = 1, \dots, M \quad (41)$$

Thus c_m is also normally distributed.

¶ We focus on:

- The nominally critical path: the path whose average duration is maximum.
- Paths with large variance in their durations, even if on average they are not critical paths.

¶ **Assumptions:**

- The number of tasks in path m is very large. That is, $\sum_{n=1}^N f_{mn}$ is large.
- The tasks in path m are independent random variables.
- We know the mean and variance, μ and σ^2 , of the normal distribution of c_m .
- Given design variables of the project, q , we know the functional dependence on q of the mean and variance: $\mu(q)$ and $\sigma^2(q)$.

¶ **Probability of failure.** Since we know the cdf of c_m we can evaluate the probability that the project will fail due to time-overrun of the m th path, analogous to eq.(9), p.6:

$$P_{f,m} = \text{Prob}(c_m > T_c) \quad (42)$$

¶ **Project analyses.** Since we know the functions $\mu(q)$ and $\sigma^2(q)$, we can perform, for path m , the:

- Design analysis as before, p.12.
- Cost vs. performance analysis as before, p.13 and figs. 3–5, p.14.
- Design-variable trade-off analysis as before, fig. 6, p.15.

9 c_m is a Convolution

¶ We consider the fourth probabilistic model described on p.7:

- c_m is a linear superposition of independent random variables t_1, \dots, t_N .

Thus the pdf of c_m is the convolution of the pdfs of t_1, \dots, t_N .

- Specifically, the duration of the m th task-path, c_m , is the following linear combination of task times, eq.(1), p.4:

$$c_m = \sum_{n=1}^N f_{mn} t_n, \quad m = 1, \dots, M \quad (43)$$

¶ Assumptions:

- The pdf and cdf of task-time t_n are known: $p_n(t_n)$ and $P_n(t_n)$, respectively.
- We know how the task-time distributions depend on the project design variables q .
- The task durations are statistically independent.

¶ With these assumptions we can show that the distribution of the duration of the m th task path is related to the task-time distributions by convolution. We now explain this.

¶ Suppose $N = 2$ or that there are only two tasks in path m , tasks 1 and 2.

Then:

$$c_m = f_{m1}t_1 + f_{m2}t_2 \quad (44)$$

Thus:

$$\text{Prob}(c_m \leq x) = \text{Prob}(f_{m1}t_1 + f_{m2}t_2 \leq x) \quad (45)$$

$$= \int_0^x \text{Prob}(f_{m1}t_1 \leq y) \text{Prob}\left(f_{m2}t_2 \in (x - y) \pm \frac{dy}{2}\right) dy \quad (46)$$

$$= \int_0^x P_1\left(\frac{y}{f_{m1}}\right) p_2\left(\frac{x - y}{f_{m2}}\right) dy \quad (47)$$

which is the convolution of $P_1(\cdot)$ with $p_2(\cdot)$.

¶ Now consider the general case of N tasks in path m . Then, from eq.(43):

$$\text{Prob}(c_m \leq x) = \text{Prob}\left(\sum_{n=1}^N f_{mn}t_n \leq x\right) \quad (48)$$

$$= \int_0^x \text{Prob}\left(\sum_{n=1}^{N-1} f_{mn}t_n \leq y\right) \text{Prob}\left(f_{mN}t_N \in (x - y) \pm \frac{dy}{2}\right) dy \quad (49)$$

$$= \int_0^x P_{1\dots N-1}(y) p_N\left(\frac{x - y}{f_{mN}}\right) dy \quad (50)$$

where $P_{1\dots N-1}(y)$ is the cdf of the first $N - 1$ tasks. Thus eq.(50) is a recursive convolution between the distributions for N , $N - 1$ and 1 tasks:

$$\text{Prob}(c_m \leq x) = P_{1\dots N}(y) = \int_0^x P_{1\dots N-1}(y) p_N\left(\frac{x - y}{f_{mN}}\right) dy \quad (51)$$

¶ These convolutions may be difficult to evaluate, but in principle they can be determined, so we know the probability distribution of the duration of the m th path. We can now proceed as before.

¶ **Probability of failure.** Since we know the cdf of c_m we can evaluate the probability that the project will fail due to time-overrun of the m th path, analogous to eq.(9), p.6:

$$P_{f,m} = \text{Prob}(c_m > T_c) \quad (52)$$

¶ **Project analyses.** We know how the task-time distributions depend on the design variables q . Thus we know how the path-duration distribution depends on q . Thus we can perform, for path m , the:

- Design analysis as before, p.12.
- Cost vs. performance analysis as before, p.13 and figs. 3–5, p.14.
- Design-variable trade-off analysis as before, fig. 6, p.15.