Project Risk Management

Topic 10

Quantitative Risk Analysis - Probability Distributions

Contents

Topic 10: Quantitative Risk Analysis - Probability Distributions

Introduction to topic

A key concept in quantitative risk analysis is the probability distribution. Topic 10 describes the characteristics of common probability distributions. This is needed in order to understand and apply Monte Carlo simulation.

Learning outcomes from this topic

On completion of this topic, you should:

- Describe the characteristics of common probability distributions

Copyright Notice

1.0 PROBABILITY DISTRIBUTIONS – INTRODUCTION

1.1. Random Variable

One of the basic concepts in probability theory is that of the random variable. When a characteristic is observed to assume different values in different situations, that characteristic is called a *variable*. By contrast if a characteristic retains the same value from situation to situation, it is called a constant. Examples of variables are heights of adult males, cost or duration of a project activity, the spots showing on tossing a die. A *random* variable has a value that changes from situation to situation. A random variable can be discrete or continuous:

- **Discrete Random Variable -** A discrete random variable has numerical values are limited to *specific values* within a range. *Counting* will always result in discrete numerical data. Discrete data have values that are limited to specific points within a range of values. For example, the number of workers to be employed on a project will be 0, 1, 2 etc., but it cannot be say 2.7 or 3.2. Discrete data need not be just whole numbers. For example, shoe sizes are discrete data because they come in whole and half sizes, such as 8, 8.5, 9, but there cannot be values *between* shoes sizes 8 and 8.5 Hence a discrete variable is characterised by 'gaps' between the values that the variable can assume.
- **Continuous Random Variable -** A continuous random variable can take any value over a continuous range of value. Therefore the possible values of a continuous random variable are not isolated numbers but an entire span of numbers. Continuous data is usually based on the *measure* of a quantity. So continuous data are measurements that can include *any* value within a range. e.g., the length of someone's foot. This can be measured in centimetres to any number of digits to the right of the decimal place, depending on the accuracy of the measuring device. In reality data cannot be really continuous as there are limitations on the ability to obtain accurate measurements. So the continuous data for foot lengths may be limited to, say, a hundredth of a centimetre. So whilst the data are conceptually continuous, and dealt with as such, the recorded data are discrete. In projects, the random variables of cost and time are most commonly continuous e.g. a cost element is estimated to be between $100 and $150 and the actual will be any value within the range

1.2. Probability Distribution

The range of possible values that a random variable can take and the probability of occurrence for each of these values can be modelled as a *probability distribution,* which is used to represent aleatoric (natural variability) and/or epistemic (knowledge) uncertainty. The probability distribution sets out the probabilities for a set of value i.e. not the actual outcome but a theoretical outcome. The total of these probabilities in the probability distribution is 1. The probabilities can be structural, objective or subjective (see paper 9). A probability distribution is very similar in nature to a frequency distribution. In particular a probability distribution is analogous to the *relative frequency distribution*, that is, they both show the *proportion* (not the number) of times a value is expected to occur. Therefore, whilst a frequency distribution is based on the outcomes that *actually occurred*, a probability distribution is based on probabilities of all the possible values that *could result* for a variable. A probability distribution has the same format as a frequency distribution, whereby the horizontal axis (x-axis) covers a range of possible values that the variable could take and the vertical axis (y-axis) gives a probability for value a within that range. A discrete variable is represented by a discrete probability distribution and a continuous variable by a continuous probability distribution.

1.3. Expected Value, Variance, Standard Deviation

In Paper 8 we discovered how to describe the characteristics of frequency distributions by numerical measures such as the mean and standard deviation. Equally, we can describe probability distributions in similar terms:

Expected Value (Mean)

It is important to understand the difference between the various measures of central tendency in probability distributions. I.e. expected value (mean), median, mode. For a symmetrical distribution, these have the same value. For non-symmetrical distributions, such as skewed, these values are:

- Mode – peak of the distribution, and often referred to as the 'most likely' or 'most probable' value
- Expected value –the mean value for the distribution
- Median –the 50% percentile i.e. 50% probability of being exceeded or not being exceeded

Variance and Standard Deviation

The variability or spread of a probability distribution is an important and informative factor, of which a main measurement is standard deviation, which is the square root of the variance. Variance is a measure of how much the distribution is spread from the expected value (mean). The standard deviation of a random variable indicates the extent of the dispersion or variability among values that the random variable may assume. The standard deviation of a probability distribution is analogous to the standard deviation of a frequency distribution, which is explained in Paper 8. Whilst the standard deviation for a frequency distribution measures average deviation from the *mean*, the standard deviation for a probability distribution measures average deviation from the *EV*.

2.0. PROBABILITY DISTRIBUTIONS – Discrete and Continuous

2.1 Discrete Probability Distribution

Every possible value of a *discrete* random variable *can be assigned a probability* to form a discrete probability distribution, which is a list of all the possible values of the random variable and the associated probabilities. The list can be in the form of a formula, table or chart (Figure 1).

Value (days)	Probability
10	15
14	60
20	25

Table

Chart

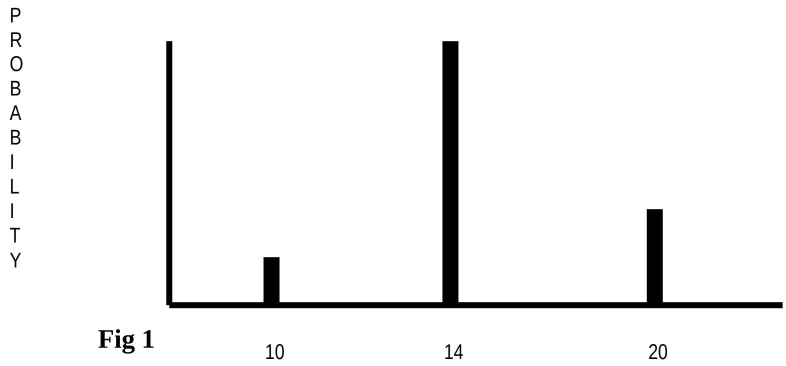


Fig 1

For a discrete probability distribution, the expected value and variance are

Expected Value = Σ x. p(x)	**Variance = Σ (x- EV)^2 * P(x)**
Where: x a value of the random variable P(x) the probability of x occurring	Where: x a value of the random variable EV expected value of the random variable. (x - EV) deviation from the EV for each value P(x) the probability of x occurring

Expected Value

The EV of a discrete random variable is the *weighted average* of that variable's possible values, where the respective probabilities are used as weightings. The EV of a *discrete* random variable is arrived at by multiplying each possible value of the random variable by the probability of its occurrence, and then sum these products.

Example: discrete probability distribution for a project activity is:

Value (days)	Probability (%)
10	15
14	60
20	25

EV for the duration of this project activity = **Σ x. p(x) =** 10(0.15) + 14(0.60) + 20(0.25) = 1.5 + 8.4 + 5.0 = 14.90

Variance and Standard Deviation

SD of a discrete probability distribution = **√ variance** = **√ Σ (x- EV)^2 * P(x)**

Example – Using previous example, where the EV for a project activity is 14.90 days, so:

Variance = Σ (x - EV)^2 * P(x) = Σ (x – 14.9)^2 * P(x)

= (10-14.9)^2 (.15) + (14-14.9)^2 (.60) + (20-14.9)^2 (.25) = 3.60 + 0.48 + 6.50 = <u>10.58</u>.

Standard Deviation = √ variance = √ 10.58 <u>= 3.25</u>

One useful type of discrete probability distribution is the ***binomial***, which is a yes/no distribution. It has only two values, each which its associated probability. For example, the *probability* of a risk event occurring; which is either yes it will occur (say 30% probability) or no it will not (say 70% probability).

2.2 Continuous Probability Distributions

A continuous random variable, unlike discrete random variables, can take any value over a continuous range of values. Therefore the possible values of a continuous random variable are not a discrete number of values but an entire span of values within a range. So it is not possible to assign probabilities to a particular possible value within the range as in a discrete probability distribution. A continuous probability distribution represent the probability for a continuous random variable. A continuous probability distribution assigns probabilities by means of *areas under a curve* known as the *probability density function* i.e., area is used to assign probabilities. A continuous probability distributions could be a convenient representation of a discrete distribution that have many possible outcomes, all very close to each other.

As with discrete random variables, the probabilities of all the possible values of a continuous random variable must total 1. Therefore the area between the curve and the *x*-axis (horizontal) must equal 1, because the curve represents all possible values and one of them will occur. This curve is useful in determining the probability that the random variable will take a value between two values and this probability corresponds to the *area under the curve between these two values*. Hence one of the key differences between discrete and continuous random variables is:

- *discrete variable* - the probability of a single value occurring can be defined
- *continuous variable* - probability is interpreted by area and only considered in terms of intervals between values rather than the probability of individual values.

Risk analysts can specify a probability distribution for any project variable (e.g., duration of an activity; cost of a project item). Several types of continuous probability distribution exist, common ones including: Uniform; Triangular; Normal; Poisson; Binomial; Lognormal; Exponential; Beta. This paper will only deal with some commonly used continuous probability distributions

2.3. Uniform Distribution

In a continuous uniform distribution, all equal intervals in the range of the distribution have the same probability. The minimum and maximum values are fixed but all values between minimum and maximum are equally likely to occur (Figure 2). So a uniform distribution is useful when we can identify a range of possible values but are unable to decide which values are most likely to occur than others. So this distribution is useful where we only have a minimum and maximum value to go on and used as an approximation of uncertainty where there is little available data. In practice, it has limited applications, although it is useful as it brings attention that the variable is very poorly known. The expected value (mean) will be the average of the two extreme values and there is no mode.

Expected value = $\dfrac{\text{min + max}}{2}$	**Variance =** $\dfrac{(\text{max} - \text{min})^2}{12}$

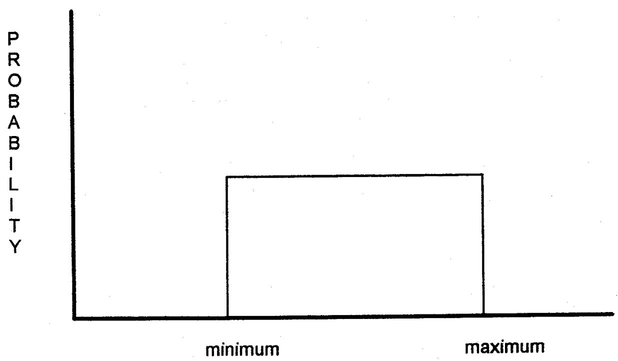


Fig 2

2.4. Triangular Distribution

A triangular distribution is applied to variables where the minimum, maximum and most likely (i.e. mode) values can be estimated (Figure 3). (It is interesting to note that the 'most likely' is in effect the mode or peak of the probability distribution, although those stipulating the 'most likely' may be assuming it represents the mean or median. Mathematically the mean would be more appropriate but this is requires knowledge of all data within a distribution which is not practical so the mode/peak is used). Where values near the minimum and maximum are less likely to occur than those near the most likely, then the triangular distribution provides a convenient means of representing this uncertainty. The triangular distribution is the most commonly used distribution for modelling expert opinion and is sufficient for most practical situations. It is favoured because it is relatively easy to specify the minimum, maximum, and most likely values, it does not require dealing with standard deviations, and it can also be made exhibit the skewness often associated with the expected range of outcomes. In project applications to represent the cost or duration of a variable, the triangular distribution is typically skewed to the right because there is a practical restriction on how quick or cheap a variable will be but there is greater potential for a longer duration or higher cost i.e. a higher probability of exceeding the most likely value compare to being lower than the most likely. For example it is broadly suggested that for project activities, durations tend to have a 1:2 skew to the right (Chapman & Ward, 2011), e.g. if the most likely duration is 10 days and the minimum is 8 days, then the maximum would be 14 days. A problem with the triangular distribution is that is that setting an absolute maximum value may be unreliable where the maximum is difficult to determine. The triangular distribution may also overestimate the probability of values towards the tails (minimum and maximum) of the distribution and underestimate the shoulders in comparison with other distributions.

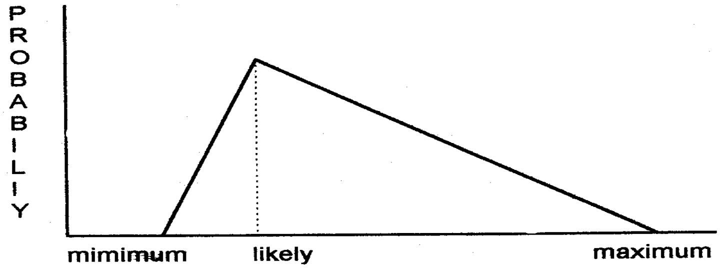


Fig 3A

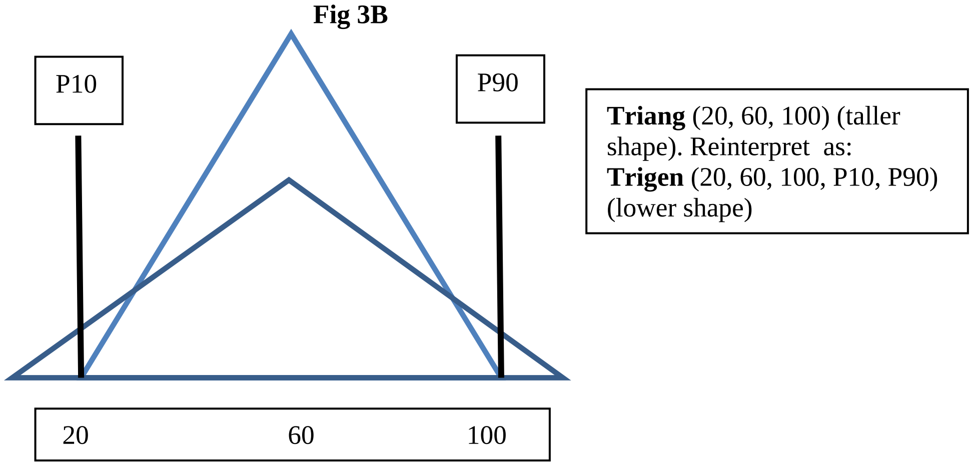
For a triangular distribution, If a = min, b = most likely, c = max, then:

Expected Value = $\dfrac{a + b + c}{3}$	**Variance** = $\dfrac{a^2 + b^2 + c^2 - ab - ac - bc}{18}$

Trigen (triangular generation) (Fig 3B)

Trigen is a variation of the triangular distribution. It has three points: lowest, most likely, highest. The lowest and highest are set at selected percentile levels, typically P10-P90. For example, a Trigen of (20,60,100; 10,90) specifies a *lowest* (not absolute minimum) value of 20 (with 10% probability of being lower) , most likely 60 and highest (not absolute maximum) of 100 (with 90% probability of being lower). Typically, people are subconsciously overconfident so they understate the minimum and maximum values for a variable. Furthermore, the minimum and maximum values in a triangular distribution are the *absolute* extreme values, which are often difficult to estimate. To counteract this, we may interpret the minimum and maximum values as being other possible extreme values. So, people may state the minimum and maximum is 20 and 100, but we reinterpret that these as simply 'practical' extremes and decide how likely they will be exceeded (e.g. P10-P90). So the values at the P10 and P90 are extended to arrive at the extreme values at P0 and P100, thereby widening the overall range of the distribution. It may not be appreciated that the final range of the Trigen distribution has been extended so it is wise to plot the distribution and agree before using it.

Fig 3B



2.5. Normal Distribution

The importance of the normal distribution is that it has been found that quantitative data gathered for a wide variety of phenomena form distributions close to that of a normal distribution. So it is a reasonably good approximation for many situations. That is, values cluster around some central value with deviations above and below that value being equally likely and have decreasing frequency as the deviations increase (figure 4). However there is no way to be sure that a particular distribution closely follows a normal curve without collecting data and testing to see if the normal distribution provides a good fit.

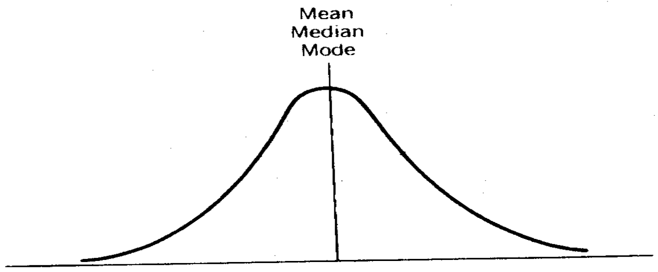


Fig 4

A normal distribution has the following characteristics (Figure 4):

- it has one mean, median, and mode. i.e., they all have the same value and lie at the centre of the curve
- it has a single peak and is symmetrical around the mean
- it is a bell-shaped curve, and spreads outwards and downwards
- it is composed of an infinite number of cases. So the tails of the curve never touch the horizontal axis.

- the probability that a variable will lie within (Fig 5):
  plus and minus one standard deviation of its mean is 68.3%
  plus and minus two standard deviations of its mean is 95.4%
  plus and minus three standard deviations of its mean is 99.7%

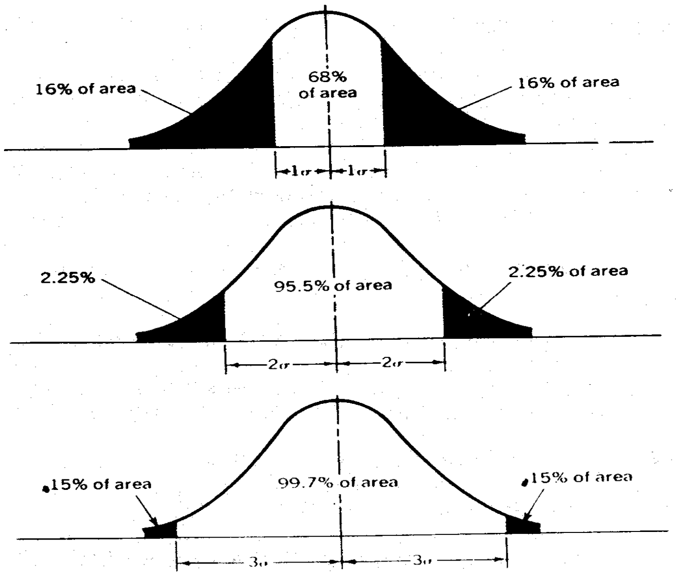
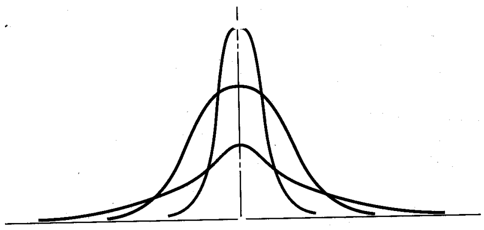


Fig 5

The shape of the curve is determined by two factors - the mean and the standard deviation (see Figure 6). The mean determines the height and location whilst the standard deviation determines the spread.

Normal distributions
Identical mean
Different standard deviation



Normal distributions
Different mean
Same standard deviation

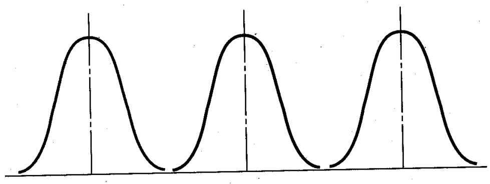


Fig 6

Using z-score to calculate probabilities (see appendix for table)

A project activity has been considered to be a normally distributed random variable with a mean of 120 hours and a standard deviation of 3 hours. *What is the probability of the project activity will take between 120 and 125 hours?*

1. *Find the location on the normal distribution curve corresponding to these two values (i.e., 120 and 125 hours).*

   The location of these two values are each expressed in terms of the position from the expected value, by the amount of standard deviation from the expected value. This is known as the Z score, which is the number of standard deviations from x to the mean, and the formula is

Z = (x-mean) / SD
x is any value on the normal distribution curve
z is the number of standard deviations from x to the mean

So:

The Z score for 120 = (120-120)/3 = <u>0 standard deviations</u>

The Z score for 125 = (125-120)/3 = <u>+1.67 standard deviations</u>

2.. Determine the area under the Normal curve between the z scores of the two values

-*W*e must determine the area between 0 and +1.67, as this area = probability. The area between any two points under a normal distribution curve can be calculated using a Normal Probability Distribution Table (see Appendix).

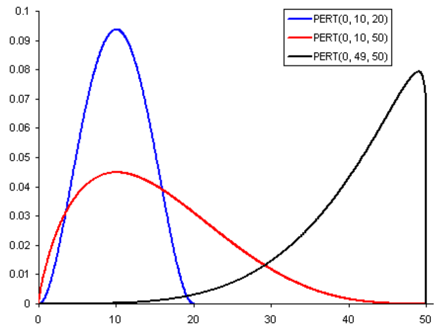
This table shows that a z score 1.67 of *results in a 0.4525 probability that task will take 120-125 hours*

2.6. BetaPERT Distribution

Beta distributions have a smooth curve rising from zero likelihood to a rounded peak, before falling away again to zero, and can be skewed– see figure 7 below. However, it requires some rather obscure parameters to describe it. A special version of a beta distribution is the PERT (Program Evaluation and Review Technique) distribution, which is defined by *minimum, most likely maximum* values, so in that regard it is like a triangular distribution. The PERT distribution assumes that mean = (minimum + 4* most-likely + maximum)/6. This allows the shape to be derived from three values – minimum, most-likely, maximum – so it can be used as a pragmatic and readily understandable distribution that can be based on expert opinion. It may be considered as superior to the triangular distribution when the parameters result in a skewed distribution, as the smooth shape of the curve places less emphasis in the direction of skew. The mean is four times more sensitive to the most likely value that minimum and maximum values, compared to the triangular where the mean is sensitive equally to all three values. Therefore the mean and standard deviation in a BetaPERT distribution is not overly biased by highly skewed distributions caused by an extreme maximum value, compared to the triangular distribution. So where it is considered that a cost variable has a quiet high pessimistic/maximum value, relative to its most likely and optimistic/low value, and this pessimistic value has a low likelihood, then a BetaPERT may be appropriate. The BetaPERT is useful if we are confident that the most likely value is based on good information and/or opinion, and we have lesser confidence in the extreme values.

In contrast, the triangular may be preferred because in distributions skewed to the right (i.e. skewed to pessimistic maximum value), the triangular will have a 'fat tail" (ie greater probability of higher pessimistic values), which offsets people's unwillingness to state their true pessimistic estimate. So for a set of minimum, most likely and maximum values for a variable, the triangular will have a mean and median further away from the most likely than a BetaPERT, and a larger standard deviation. So a triangular is often used as it conveys a greater degree of uncertainty for a set of minimum, most likely and maximum values

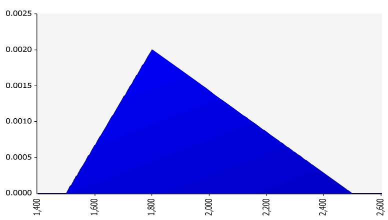
Fig 7 – Examples of BetaPERT distributions

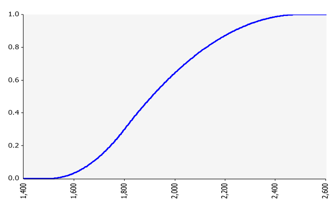


Expected value = $\dfrac{\textbf{Min + 4 Most Likely + max}}{\textbf{6}}$	**Variance =** $\dfrac{\textbf{(EV-min)(max-EV)}}{\textbf{7}}$

2.7 Cumulative Distribution Function

A *probability density function* can be displayed into a *cumulative distribution function* (figure 7). The CDF is simply the arithmetic summation of the probability density function. A continuous probability distribution helps to answer the question "what is the probability that a continuous random variable will be less or equal to a stated value?" Many *cumulative distribution functions* have the appearance of a stretched S shape and makes it difficult to identify its probability density function shape fromm which it is derived.





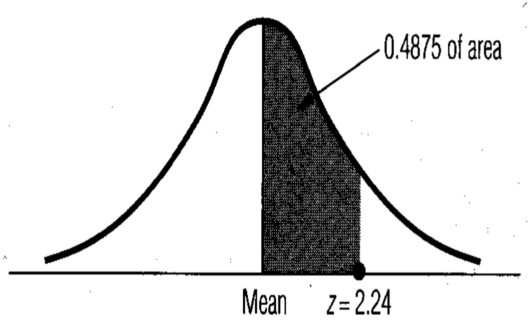
Probability Density Function	**Cumulative Distribution Function**

Fig 7: Probability density function displayed into a cumulative distribution function

3.0. REFERENCES

CHAPMAN, C.B. & WARD, S.C. (2011). *How to manage project opportunity and risk,* 3rd ed., Wiley: Chichester, England,

APPENDIX - Z Score



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990