Statistics for the Performance Analyst

Tom Wilson

1 Introduction

Many engineers come into the performance world from different backgrounds and lack the appropriate education. For example, a software developer may be challenged to look at the performance problems in some code and might transition into software testing as a result. He learns what he needs to know *on-the-fly*. Many times, he does not know that he needs to know something and does what he can. If there is any review, he learns from his mistakes.

Statistics is one of those fields that is necessary for performance analysis, but may not be given proper attention in advance. It might just be that it has been years since you encountered the topic in high school or college. This issue has been addressed in the literature before ([Mac89], [Kal02]). For some, basic statistics is sufficient. For others, more advanced knowledge is necessary.

This paper is an introduction to statistics from the performance analyst's perspective. It will focus on understanding the common concepts in order to *use* them rather than *implement* them. Many tools (e.g., Excel, R) implement the concepts and the analyst just needs to understand how to use them. So, we will not be writing out equations nor deriving them. For more details in any particular area, numerous materials exist ([Sto98], [NIS06], [Lan07], and [wik10, "Statistics"]). Strangely, only a few performance books provide a chapter or section on statistics ([Jai91], [Lil05], [SW07]).

Statistics refers to a range of techniques for analyzing data, interpreting data, displaying data, and making decisions based on data. A *statistic* is a numerical quantity (e.g., mean) calculated from data. Statistics is a vast, well-studied science. Not all aspects of statistics are applicable to performance, so only relevant terms and concepts are discussed here. Nonetheless, the absence of something does not imply that it cannot be used in our performance world.

We will divide statistics into two basic areas: descriptive statistics and predictive statistics. The former is for anyone looking at measurements; the latter is for anyone doing modeling. We will forgo the foundational discussions of populations, samples, and random variables, and take the viewpoint that you have a bunch of numbers that you have to summarize and/or understand. Of course, understanding those foundational concepts will improve your analyses.

2 Descriptive Statistics

Descriptive statistics summarize a large amount of data using a small amount of data, often using only one number. Descriptive statistics come in two general forms: numerical and graphical. Numerical forms use numbers to convey information. Graphical forms use images containing various graphics primitives (e.g., points, symbols, text, lines, polygons, colors). Sometimes a categorical description, instead of a numerical statistic, is sufficient. Descriptive statistics introduce error through the loss of information, but this is a necessary trade. This error is different than measurement error or statistical error.

Descriptive statistics can be divided into (at least) four categories: measures of (1) location, (2) dispersion, (3) shape, and (4) association. The first three measures are applicable to one data set. The fourth is applicable when there are two data sets. Any descriptive statistic can be misleading when used in isolation. *Summary statistics* group multiple descriptive statistics together to describe data more thoroughly.

2.1 Measures of Location

A *measure of location* (or *measure of central tendency*) is a value that attempts to define where most of the data are clustered. Common measures of location are the mean, the median, and the mode. They all have the same intent, but are applied in different situations. The measure of location is meant to be the one value that best represents the entire set of values. The measure need not be one of the values.

There are a few definitions for *mean* (or *average*) that exist. The *arithmetic mean* is the sum of all values divided by the number of values. The arithmetic mean is a good measure of location when the data are symmetrically distributed (a term defined in Section 2.3), but can be misleading when they are not. Each value contributes to the mean, so it is sensitive to extreme values. Table 1a shows three example data sets. Table 1b shows the arithmetic mean for each data set. Each data set was intentionally created so that the arithmetic mean would be the same. This provides interesting contrast when looking at the other descriptive statistics.

Table 1: Numerical Descriptive Statistics

(a) Example Data Sets

Set 1		Set 2		Set 3	
5.000	5.000	1.045	4.301	1.094	2.099
5.000	5.000	1.194	5.662	1.221	2.984
5.000	5.000	1.337	6.277	1.354	3.634
5.000	5.000	1.394	6.535	1.379	3.740
5.000	5.000	1.728	7.905	1.482	7.309
5.000	5.000	2.058	9.064	1.484	7.981
5.000	5.000	2.552	9.271	1.612	9.437
5.000	5.000	2.945	9.865	1.642	12.957
5.000	5.000	3.082	10.247	1.686	15.588
5.000	5.000	3.201	10.337	1.835	19.482

(b) Various Statistics (Rounded)

Statistic	Set 1	Set 2	Set 3
Arithmetic Mean	5.000	5.000	5.000
Harmonic Mean	5.000	2.835	2.266
Geometric Mean	5.000	3.814	3.133
Median	5.000	3.751	1.967
Standard Deviation	0.000	3.409	5.426
25^{th} Percentile (Q_1)	5.000	1.976	1.484
50^{th} Percentile (Q_2)	5.000	3.751	1.967
75^{th} Percentile (Q_3)	5.000	8.195	7.477
Range	0.000	9.292	18.388

The *harmonic mean* is a companion to the arithmetic mean, but is applicable to rates rather than measurements. A rate is a ratio of measurements (e.g., jobs per second, kilobytes per user). A simple scenario is averaging throughput rates (e.g., transactions per second). Refer to [Lil05] or [wik10, "Harmonic mean"] for brief discussions and examples. So, we should now emphasize that the arithmetic mean is appropriate for measurements that are not rates. Table 1b shows the harmonic mean for the three example data sets. We should not compute the arithmetic mean and harmonic mean for the same data; we are presenting the data without associating units of measure, and demonstrating each statistic as if it were appropriate to apply it.

The *geometric mean* is applicable to a series of proportions rather than to measurements or rates. This statistic is not that common. Refer to [Jai91], [Cie01], or [wik10, "Geometric mean"] for brief discussions and examples. Table 1b shows the geometric mean for the three example data sets.

The *median* is the middle value in an ordered set of values when there is an odd number of values, and the average of the middle two values when there is an even number of values. The median is less sensitive to extreme values than the mean and is usually a better measure of location than the mean for asymmetric distributions (a term defined in Section 2.3). Table 1b shows the median for the three data sets. The examples demonstrate the disparity that can exist between the arithmetic mean and the median.

The *mode* is the most frequent value in the set of values, but it is not often that we will want to know the mode. However, we are interested in a property related to the mode: the modality of a distribution. That will be discussed in Section 2.3.

The arithmetic mean and the median are contrasted in several successive figures. Their relative variations are affected by the shape of data rather than the dispersion of data.

2.2 Measures of Dispersion

A *measure of dispersion* is a value that relates how spread out a set of values is. The term *spread* or *variability* is sometimes used instead of *dispersion*. Common measures of dispersion are the standard deviation, percentiles, quartiles, and range. The boxplot is a graphical technique for expressing dispersion.

The *standard deviation* bases its value on how each of the data values relates to the mean. If the values are clustered around the mean, then the standard deviation is small. If the values are dispersed away from the mean, then the standard deviation is large. This is what we want the measure to tell us. As with the mean, each value contributes to the standard deviation. Therefore, it is sensitive to extreme values. Also, the measure is more appropriate for symmetric distributions. Table 1b shows the standard deviation for the three data sets. Each set is progressively more dispersed.

So, if there are three kinds of means, should there be three kinds of standard deviations? [wik10, "Geometric standard deviation"] discusses the geometric standard deviation in relation to the geometric mean. [Cie01] discusses

the application of the geometric mean and standard deviation in analyzing web response times. We found no references to a harmonic standard deviation.

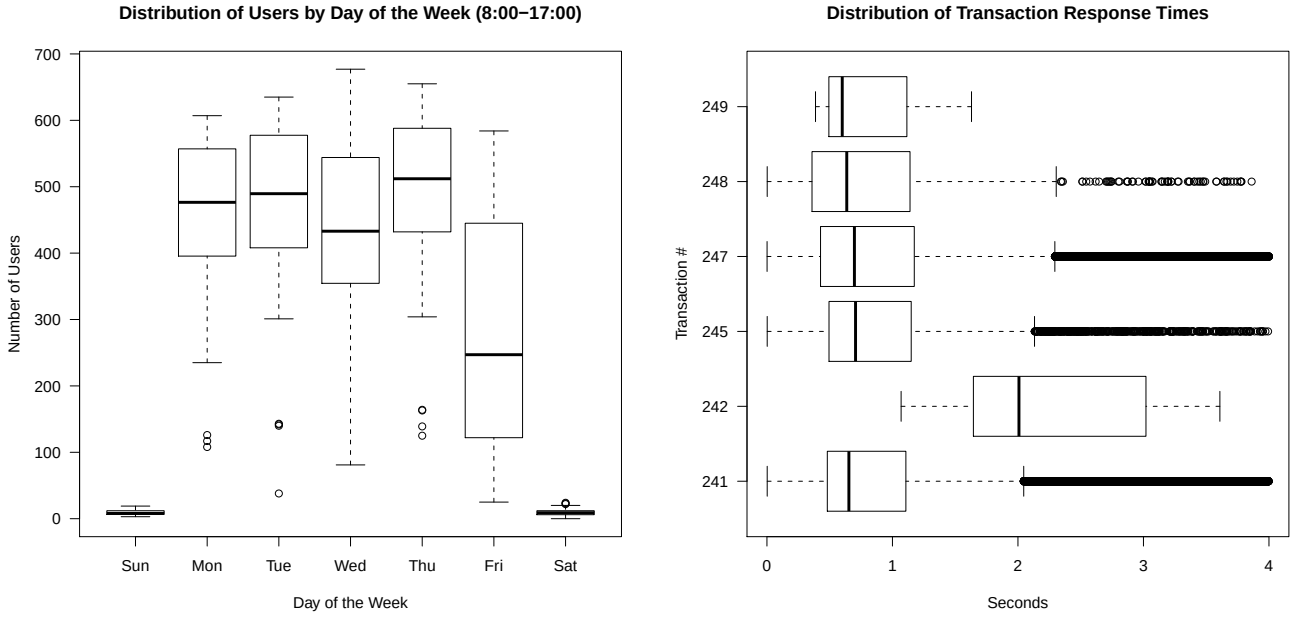
A *percentile* is a value in an ordered set of values below which a certain percentage of values fall. So if x is the p^{th} percentile, p% of the values are less than x. The 50^{th} percentile is the median, and divides the values into two halves. Other common percentiles are: 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, and 98^{th}. Percentiles are less sensitive to extreme values and provide a better measure of dispersion for asymmetric distributions.

Percentiles often appear in *service level agreements* (SLAs) and requirements. For example, an SLA might specify that 98% of the response times be less than or equal to 5 seconds. If we compute the 98^{th} percentile for all of the response times for the measurement period, that value must meet the 5 second goal. The other values are irrelevant beyond impacting the ordering of the values. Performance is usually viewed from the average case, worst case, or some percentile. The average case involves using one of the aforementioned means. The worst case uses the maximum value in a data set. Both of these choices have strong disadvantages when applied to performance data. So, percentiles may be a better choice.

A *quartile* is any of three values which divide the ordered values into four equal sets. The first (Q_1), second (Q_2), and third (Q_3) quartiles are equivalent to the 25^{th}, 50^{th}, and 75^{th} percentiles. Quartiles are used in boxplots (discussed below). Table 1b shows the quartiles and associated percentiles for the three data sets.

The *range* is simply the maximum value minus the minimum value. We are often interested in noting the minimum and maximum values explicitly without computing the range itself. Range is a poor measure since it only references the extremes. Table 1b shows the range for the three data sets.

The *boxplot* (or *box-and-whisker plot*) is a graph that shows the quartiles and some other descriptive statistics. The other statistics can vary, depending on the implementation or parameter choices. Figure 1 shows two boxplot examples, one vertical and one horizontal. Each box reflects Q_1 (box bottom or left) and Q_3 (box top or right), with Q_2 shown as a horizontal or vertical line within the box. Whiskers (solid lines perpendicular to the dashed lines emanating from the box) span a distance that can vary by implementation. A common choice is upto 1.5 times the *interquartile range*, which is the size of the box (i.e., $Q_3 - Q_1$). Points beyond the whiskers, if any, are plotted as circles, which may be either empty and/or filled. Such points are outliers. An *outlier* is a value that is statistically distant from most of the other values. Such values can be removed if they can be judged to be errors or anomalies.



(a) Vertical Boxplot **(b)** Horizontal Boxplot

Figure 1: These graphs show example (a) vertical and (b) horizontal boxplots. In both cases, multiple categorical data sets are plotted together for comparison. The x-axis in (b) has been limited so that the boxes are not compressed (i.e., there are many points beyond the drawing area).

Boxplots allow multiple data sets to be plotted together for comparison. This allows visual comparison of the statistics for all data sets. A good example is the separating of response times by transaction as shown in Figure 1b.

2.3 Measures of Shape

A *measure of shape* is a value that relates characteristics of the shape of the distribution of the data. In performance analysis, we are probably more interested in a categorical representation of shape rather than a quantitative one. So, we will dispense with defining some measures that you will not likely use. The shape of a distribution is often best conveyed with a graphical representation. This is typically done by drawing a histogram or a density plot. Other summary statistics can be added to enhance the analysis.

A *histogram* is used to plot the probability distribution, which expresses the probability of a certain value occurring (we will discuss probability distributions in Section 3.2). A histogram has an associated set of values, called *bins*. They are usually equally-spaced, but need not be. Each bin has a count associated with it. Each value in the data is assigned to the largest bin that is less than or equal to the value. Assigning a value to a bin results in the bin's count being incremented. When the histogram is drawn, a bar is drawn for each bin with a height equal to the bin's count. Alternatively, a frequency can be computed and substituted for the count (when we are only interested in the shape, this choice is not important).

Bin size and location are important factors in determining the histogram's appearance. Figure 2 shows two histograms of the same data. The difference between the two histograms is only the starting bin value. Figure 2a has bins that fall on integer values, while Figure 2b has bins that fall halfway between integer values. A similar disparity in shape can be created by varying the bin size. [Gan98] states that one way to lessen the side effect of the histogram's sensitivity is to create several histograms with different bin sizes; the features that are prominent in all of them are part of the data's true structure. He also says that histograms are more sensitive to the starting location than to the bin size when the bins are equal in size.

(a) **(b)**

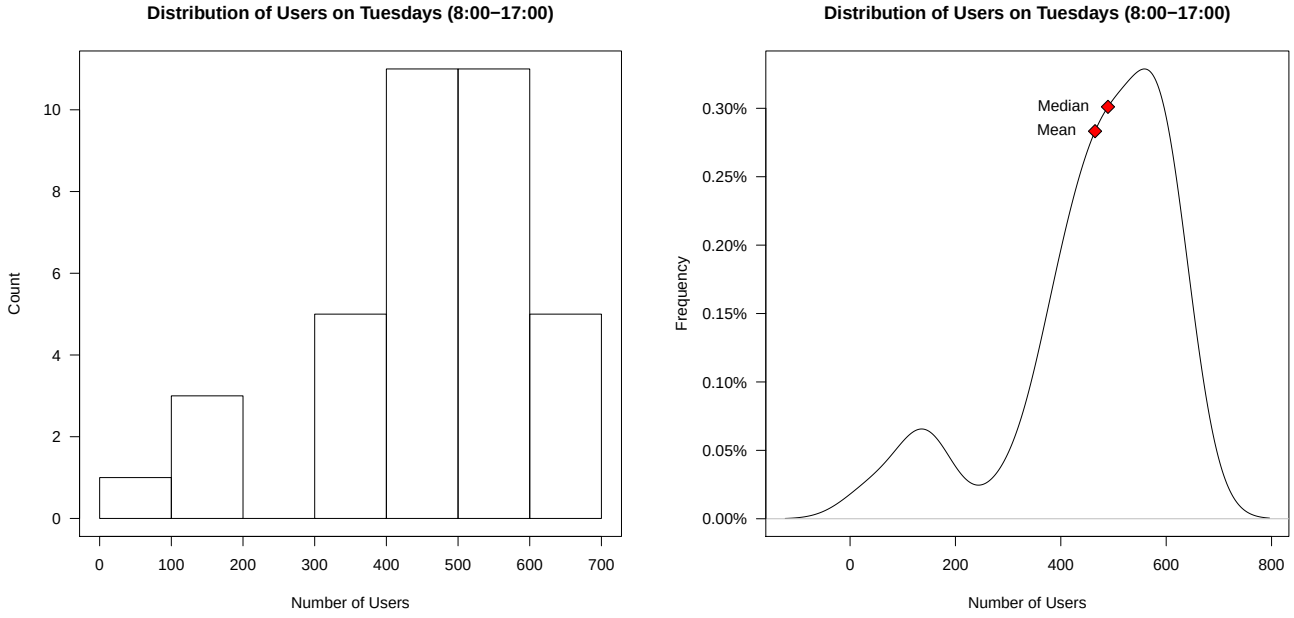
Figure 2: A comparison of two histograms for the same data.

A *density plot* is an alternative to the histogram, and is drawn with a line. It has a smoother appearance because a smoothing function is applied to the data before they are plotted. The density plot communicates similar shape information as the histogram, but is less sensitive to variation in the data. Multiple data sets can be drawn together and are easier to comprehend than histograms if they do not overlap very much. Different density functions may exist and may have parameters to control their behavior. Some subsequent figures will illustrate the density plot.

We are mostly interested in whether a distribution is symmetric or asymmetric. A *symmetric* distribution has a similar, reflective shape on either side of some midpoint. For such distributions, the mean and median are close in value. An *asymmetric* distribution often has a lopsided shape with the bulk of the values on one side and a tail on the other. The term *skewed* is often used instead of *asymmetric*. A skewed distribution is further classified as either *right-skewed* (or *positively-skewed*) or *left-skewed* (or *negatively-skewed*), indicating which side the tail is on.

The mean is often closer to the tail than the median, but this need not always be true.

Figure 3a shows a histogram of the number of users of a system. In this case, the shape is left-skewed. Figure 3b shows the same data using a density plot. Notice that there appears to be some subtle differences when the density plot is compared to the histogram. This is due to the smoothing operation provided by the density function.



(a) Histogram **(b)** Density Plot

Figure 3: A comparison of (a) a histogram and (b) a density plot for the same data.

Figure 4a shows a histogram of some response times. Figure 4b shows the corresponding density plot. Both communicate the right-skewed shape effectively.

The boxplot is a less effective indicator of shape because the boxplot hides most count information: Compare either graph in Figure 3 with the "Tuesday" column of Figure 1a (each summarizes the same data). A boxplot with many outliers on one side of the box is likely to be skewed in that direction.

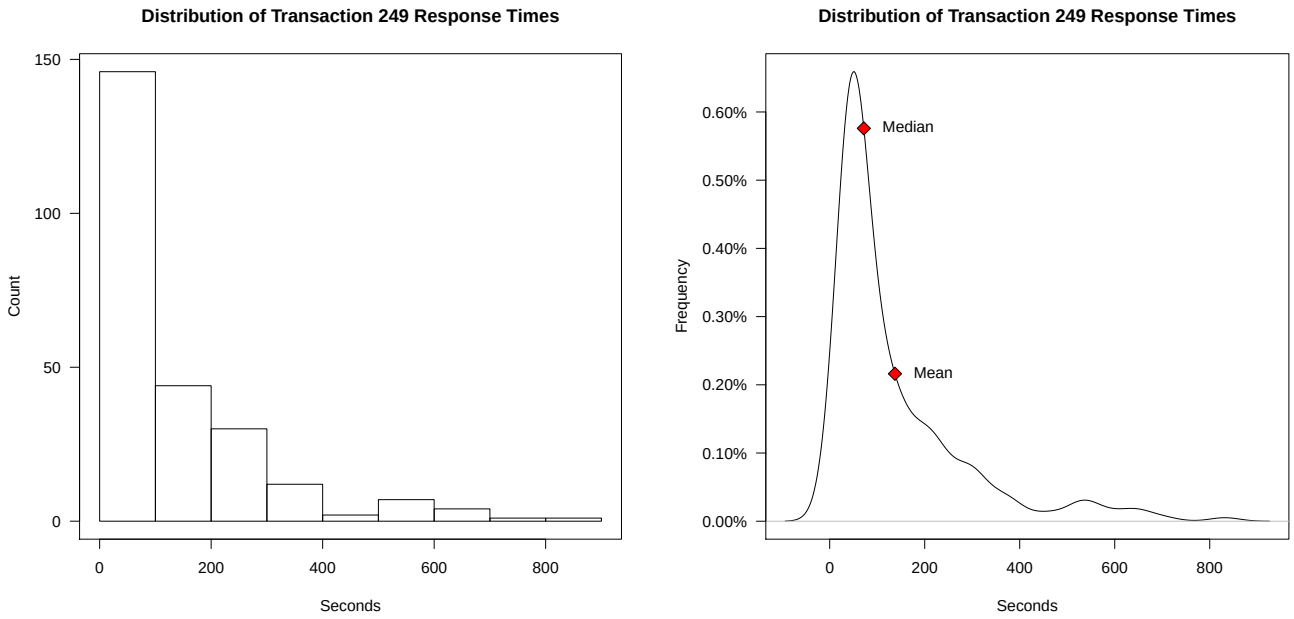
The simplest symmetric distributions are the normal and uniform distributions. Figure 5a shows a normal distribution; Figure 5b shows a uniform distribution. These distributions will be discussed further in Section 3.2. In most cases, the mean and median will not be equal, but will be close in value.

Figure 6a shows a notional left-skewed distribution. The curve need not be so well-formed. The mean and median show their possible locations. Figure 3b shows a more realistic example of a left-skewed distribution. Figure 6b shows a notional right-skewed distribution. This example is simply a reflection of the left-skewed distribution. Figure 4b shows a more realistic example. Skewed distributions are common in performance data because of the left-bound at 0 ([Gan98]).

As previously mentioned, the modality of a distribution is an important characteristic of shape. A distribution is *unimodal* if it has one mode. A distribution is *multimodal* if it has more than one mode. This definition is often relaxed to mean that the distribution has more than one local maximum rather than multiple values with the same maximum frequency. A *bimodal* distribution is a specific type of multimodal distribution where there are two modes (or local maximums). When a distribution is multimodal, descriptive statistics may not describe the data very well. Sometimes, the data are a combination of multiple groups of data. If the data can be separated, each can be described separately [Mac89]. Figure 5a, Figure 6a, and Figure 6b are unimodal. Figure 3b is multimodal. Figure 4b is multimodal, but the second local maximum is very small. Figure 5b has no mode.

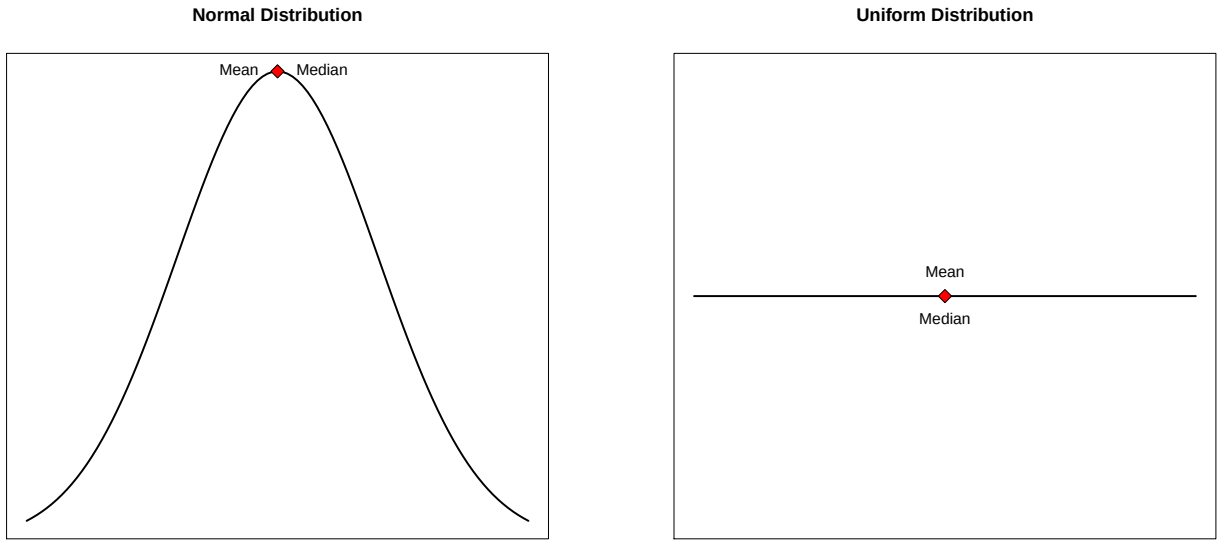
2.4 Measures of Association

A *measure of association* (or *measure of correlation*) is a value that expresses the relationship between two data sets. If a relationship exists, then one data set may be a function of the other. This is not truly evidence of causality; it may be mere coincidence that the two data sets appear to be related. Measures of association are more useful



(a) Histogram **(b)** Density Plot

Figure 4: A comparison of (a) a histogram and (b) a density plot for the same data.



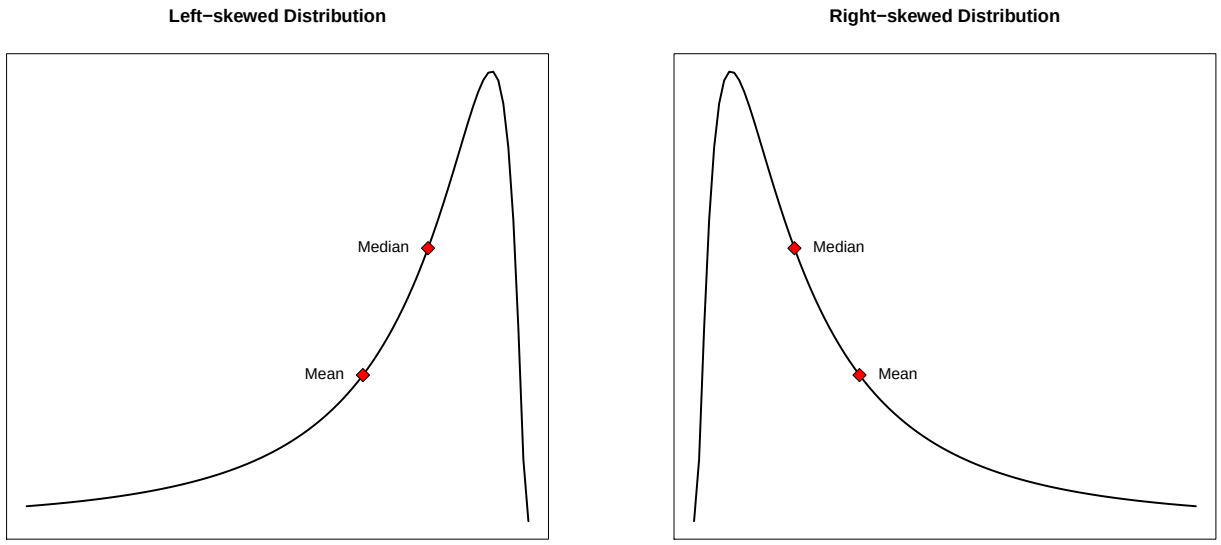
(a) **(b)**

Figure 5: These graphs show notional symmetric distributions: (a) a normal distribution and (b) a uniform distribution. In these examples, the median and mean are equal, but, in general, they need not be.

when used with predictive statistics, but they can be used on their own to study measurements. Having many data points increases the reliability of the measure.

The *simple linear correlation coefficient*[1] is one such measure. It computes the linear relationship between one data

[1]This is more formally known as the *Pearson product-moment correlation coefficient*.



(a) **(b)**

Figure 6: These graphs show example (a) left-skewed and (b) right-skewed distributions. In both cases, the tail is on the indicated side (left vs. right). The mean is typically closer to the tail than the median.

set and the other. The result is a value between -1 and +1 (inclusive). Values near +1 imply a direct relationship. Values near -1 imply an inverse relationship. Values near 0 imply no relationship. The simple linear correlation coefficient is sensitive to extreme values since its equation references all values in each data set, as well as each data set's mean and standard deviation. Therefore, a strong relationship can be weakened by a few extreme values. It is also not too hard to create data sets where a correlation coefficient can indicate a relationship where none exists. Correlation coefficients exist besides the one already mentioned.

Measures of association should be accompanied by a visualization of the data. A *scatterplot* is a two-dimensional graph of points constructed from the two data sets. Values from one data set (the dependent set) are the x-values and values from the other data set (the dependent set) are the y-values. A scatterplot communicates the data sets' relationship. It is common to draw a regression line (discussed in Section 3.1) on the scatterplot so the linear relationship can be confirmed visually.

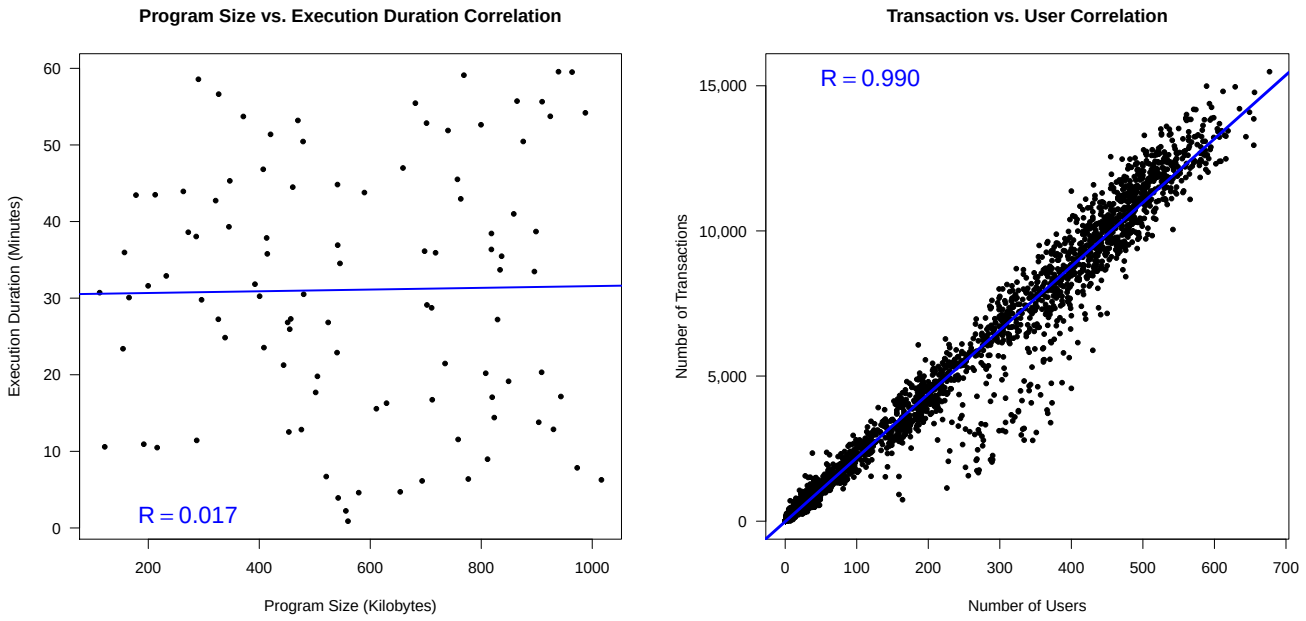
Figure 7a shows a scatterplot of the relationship between a program's size in kilobytes and its corresponding execution duration. The points appear rather random because they are: All values were generated randomly and are not real measurements. The correlation coefficient is 0.017 and reflects the lack of a linear relationship. It is unlikely that real data would indicate any relationship either (programs of any size can execute for any duration).

Figure 7b shows a scatterplot of some real data. This is a comparison between the number of users of a system and the number of transactions generated by those users. Most values cluster along the increasing diagonal. The correlation coefficient is 0.990, indicating a very strong linear relationship. This does not prove causality, but it supports the claim. Transactions are generated by the users, and we would expect a relationship to exist.

It is good to view the scatterplot and the regression line to judge if the fit is good. A line can always be generated for a set of points, but may not be appropriate (like Figure 7a). Related data sets will have a similar distribution.

2.5 Other Useful Statistics

A *weighted statistic* is a product of the statistic for a set of values and a weight assigned to it. The *weighted average* is an example of a weighted statistic. The use can be illustrated with an example. The response times in a transaction system can be grouped according to the transaction type. The average can be computed for each type. The weight assigned to each type is computed from the number of that type divided by the total number of transactions (i.e., this is the frequency of the transaction type). Such a value is useful for assessing contribution to the overall response time. It can be an indicator of which transaction should be targeted for improvement, should



(a) No Correlation

(b) Very High Correlation

Figure 7: Two correlation examples: (a) one with no correlation and (b) one with very high correlation. The points are accompanied by a blue regression line. The correlation coefficient (R) is also shown.

the overall response time be deemed too high (refer to [Wil10a]).

A *count* is a statistic that is not often defined as such. It has the property that most other descriptive statistics have: It summarizes a large amount of data using a small amount of data. They are akin to a histogram without the graph. In evaluating system performance, counts are used to summarize events (e.g., the number of file reads during a time interval). Counts often get converted to rates (e.g., files reads per second) when the counts are sampled over some interval. Counts are often the data we are describing with a statistic (e.g., mean). They are viewed as measurements, but should not be ([Wil10b]).

2.6 General Guidance for Describing Data

Some general guidance for describing performance data follows.

- Check the shape with a histogram or density plot
  - Is the shape unimodal or multimodal? If it is multimodal, consider separating the data. Otherwise, use care in summarizing the data (perhaps using some or all of the modes)
  - Is the shape symmetrical or asymmetrical? The answer dictates how location and dispersion are represented
- Summarize the location and dispersion
  - If the shape is symmetrical, use the (appropriate) mean and standard deviation
  - If the shape is asymmetrical, use the median and the other quartiles or a few percentiles (a boxplot is a graphical alternative)
- When relating data sets...
  - Create a scatterplot and check for linearity
  - Compare the distributions of the data sets
  - Compute the correlation coefficient

It is important to keep in mind your objective in describing data. Some aspects in the guidance may not be pertinent. Additional analysis beyond what is suggested may also be necessary.

3 Predictive Statistics

Predictive statistics also summarize a large amount of data using a small amount of data, but this summary often comes in the form of an equation (also called a model). Such statistics attempt to predict future or missing values based on past or existing values. Predictive statistics is a large and complex area, so only a few concepts are presented here. Advanced topics may not be available in the simpler tools.

Predictive statistics are useful to the performance modeler for estimation, prediction, trending, forecasting, and resource planning. In performance testing, we often want to model user arrival rates, session durations, and think times. If we have real data available, we want to turn such data into a model that can be implemented in the test. A *measure of fit* is a value that expresses how well the model fits the data. This is closely related to the amount of error in the model's output.

Two common forms of modeling are linear regression and distribution fitting. In order to discuss distribution fitting, we will need to define what probability distributions are.

3.1 Linear Regression

Linear regression is a modeling technique that derives a line equation, relating a dependent data set to an independent data set. The two main applications of linear regression are forecasting and correlation. Linear regression often uses "least squares" (i.e., minimizing the sum of the squares of the errors) to determine the best line. This value is used in a calculation that results in a number between 0 and 1 (inclusive), and is the measure of fit, called the *coefficient of determination*. This measure is equal to the square of the previously defined correlation coefficient. The coefficient of determination is the fraction of variation in the dependent variable that is a result of the variation in the independent variable. Linear regression uses all values in a data set when creating a model. Therefore, it is sensitive to extreme values. It is inappropriate when the data indicate a non-linear relationship. That may be determined manually by viewing the scatterplot of the data. Figure 7 shows examples that we already encountered during the discussion on measures of association.

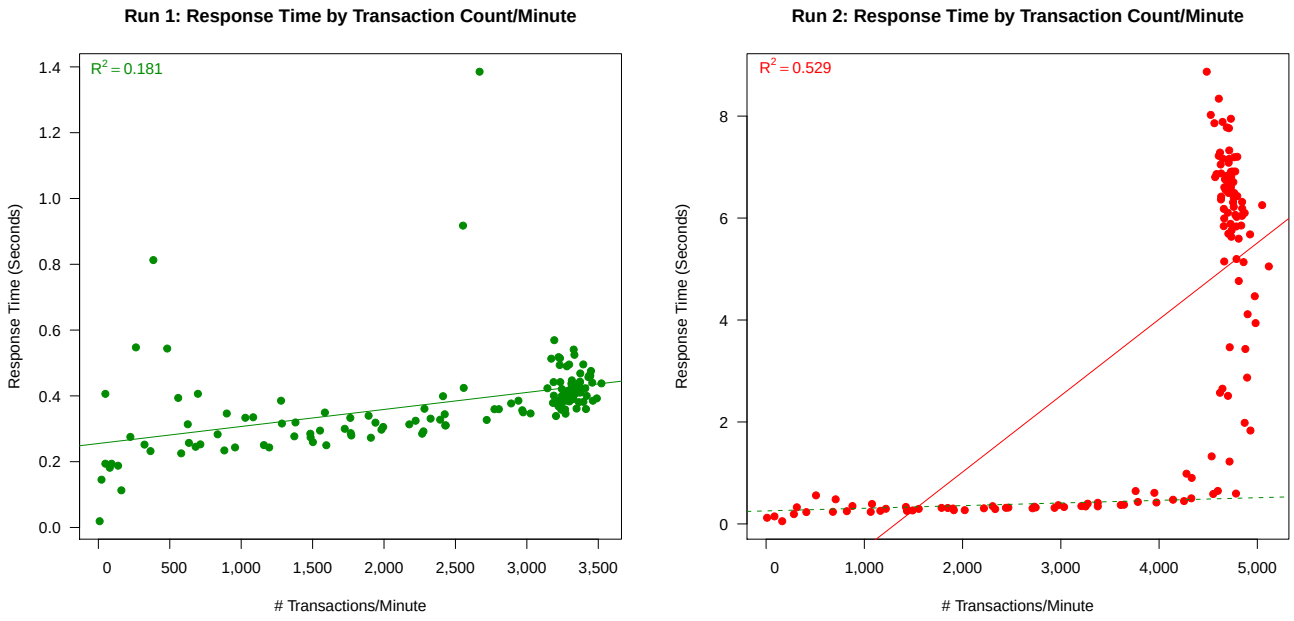
Interpolation infers values from within the range of values used to build a model. This has error that can be computed. *Extrapolation* infers values from outside the range of values used to build a model. The error cannot be computed. Extrapolation carries considerable risk, especially when the values stray greatly from the range. However, such estimates are better than no estimates. Although we have only mentioned "errors" in this brief discussion, it is an important topic.

Figure 8 is an interesting linear regression example. Data were collected from a performance test and are plotted as Run 1 in Figure 8a. The regression line looks like an acceptable fit. However, the coefficient of determination (R^2) is 0.181! This basically says that the first data set does not determine the second data set very well. If we used this model to predict response times for higher transaction loads, we might think that the model would provide an acceptable estimate (note that we are extrapolating here). However, Run 2 in Figure 8b shows how performance becomes non-linear. Creating a regression line for these data results in a line that does not fit the data very well. However, the higher coefficient of determination of 0.529 might lead us to believe that we have a better model. In each case, the visual plot and the measure of fit are an important aid to understanding the value of the model. The fact that the two models are so different is also helpful in concluding that they might not be very accurate.

3.2 Probability Distributions

In order to discussion probability distributions, additional terminology is necessary. The term "probability distribution" is often used instead of *probability mass function* (PMF) or *probability density function* (PDF). The word "probability" is often omitted when discussing distributions. A distribution is *discrete* if the values that can exist are countable (finite or infinite). In this case, the distribution is called a PMF. A distribution is *continuous* if the values that can exist are uncountable. In this case, the distribution is called a PDF. The *cumulative distribution function* (CDF) is a companion to the PMF or PDF that accumulates probability. For a given value, the CDF states the probability that any value less than or equal to that value would occur. We will encounter this function in Section 3.3.

Continuous distributions are more prevalent in performance analysis. Most of the distributions discussed in this paper are continuous. A distribution may be bounded (on both sides), bounded on the left, or unbounded. Performance data are usually bounded on the left and do not fit unbounded distributions very well. Distributions area really families of distributions with individual instances that are defined by the location, scale (i.e., dispersion), and/or shape parameters.



(a) **(b)**

Figure 8: Linear regression examples for two data sets. Neither are very good models. (a) This model *looks* good, but the coefficient of determination tells us that it does not describe the data very well. (b) This model does not *look* good, but has a better coefficient of determination (although, it is not good enough).

The *exponential* distribution is used when very small values are common while very large values are rare. It is bounded on the left at 0 and is infinite to the right. It is commonly used for modeling interarrival times. Figure 9a shows an example. Distributions within the family are defined by the scale parameter. For more examples, refer to [wik10, "Exponential distribution"]).

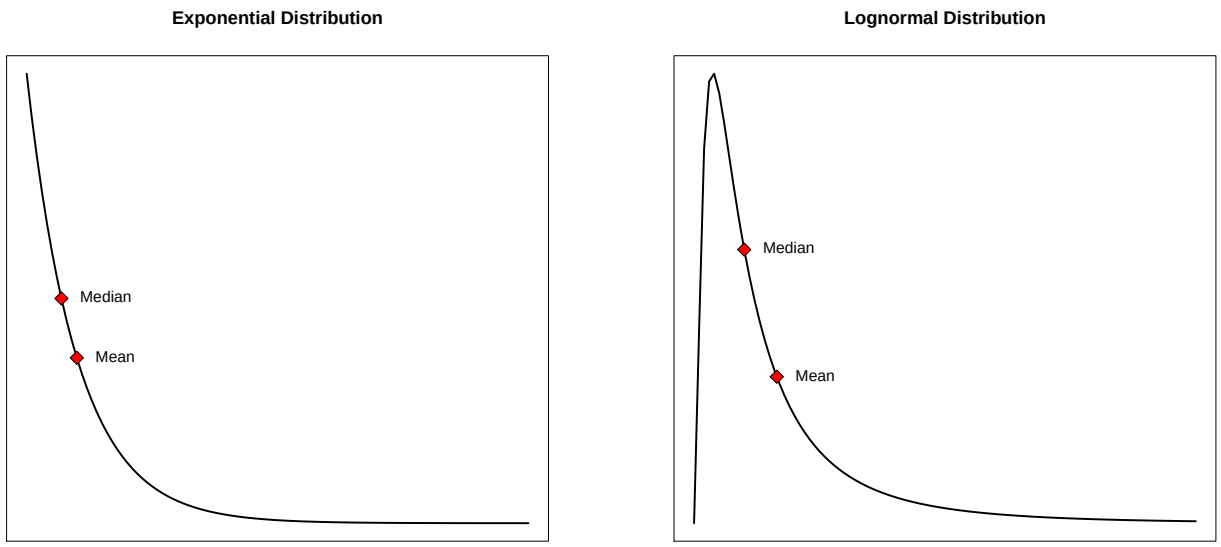
The *lognormal* distribution is used when small values are common, very small values are uncommon, and very large values are rare. The distribution is bounded on the left at 0 and is infinite to the right. It is commonly used for modeling certain types of durations, such as repair times or think times. Figure 9b shows a notional example. Distributions within the family are defined by the location and scale parameters. The shape of this distribution can be grossly affected by the scale parameter. It can take on shapes similar to the exponential and normal distributions. Figure 9b shows the more typical shape. For more examples, refer to [wik10, "Lognormal distribution"]). [Cie01] discusses the application of the lognormal distribution in analyzing web response times.

The *normal*, or *Gaussian*, distribution is used when a certain value is common with values to either side of it becoming less common. In many disciplines (e.g., psychology or sociology), it is common to default to this distribution. However, because the distribution is unbounded, it usually does not describe performance data. Figure 5a shows an example. Distributions within the family are defined by the location and scale parameters. For more examples, refer to [wik10, "Normal distribution"]).

[Szi96] provides an interesting discussion of why the normal distribution has the name "normal". Normal means independent. Flipping a coin numerous times is a series of independent experiments. Plotting the distribution of the number of heads results in a normal distribution. A series of response time measurements is probably not normally distributed because the measurements may not be independent.

The *uniform* distribution is used when no information is available besides the minimum and maximum values. It has the property that all values are equally likely. It is commonly encountered in random number generation. Figure 5b shows a notional example. Distributions within the family are defined by the location parameters. There are both continuous and discrete forms for this distribution. For more examples, refer to [wik10, "Uniform distribution (continuous)"]).

The *Poisson* distribution is a discrete distribution that is useful for modeling the number of events in an interval. It has a shape like the normal distribution, but it is not continuous. It is commonly used in queueing models. A few other distributions can take on a variety of shapes by means of a shape parameter: beta, Erlang, gamma, and Weibull. Numerous other distributions exist, some of which may be applicable in performance situations.



(a) **(b)**

Figure 9: These graphs show the (a) exponential and (b) lognormal distributions. Each is right-skewed with the mean usually being greater than the median.

Table 2 summarizes the distributions that have been previously mentioned. Their bounds provide interesting contrast. The number of distribution parameters for each distribution is also shown.

Table 2: Summary of Common Probability Distributions

Name	Type	Bound	Parameters		
			Loc.	Scale	Shape
Beta	Continuous	$[0, 1]$			2
Erlang	Continuous	$[0, \infty)$		1	1
Exponential	Continuous	$[0, \infty)$		1	
Gamma	Continuous	$[0, \infty)$		1	1
Lognormal	Continuous	$[0, \infty)$	1	1	
Normal	Continuous	$(-\infty, \infty)$	1	1	
Poisson	Discrete	$[0, \infty)$		1	
Uniform	Both	$[a, b]$	2		
Weibull	Continuous	$[0, \infty)$		1	1

Statistics tools may provide different functions to interact with a distribution: (1) a density function, (2) a CDF, (3) a quantile[2] function, and (4) a function for generating random numbers according to the distribution. Care should be taken in understanding and using them.

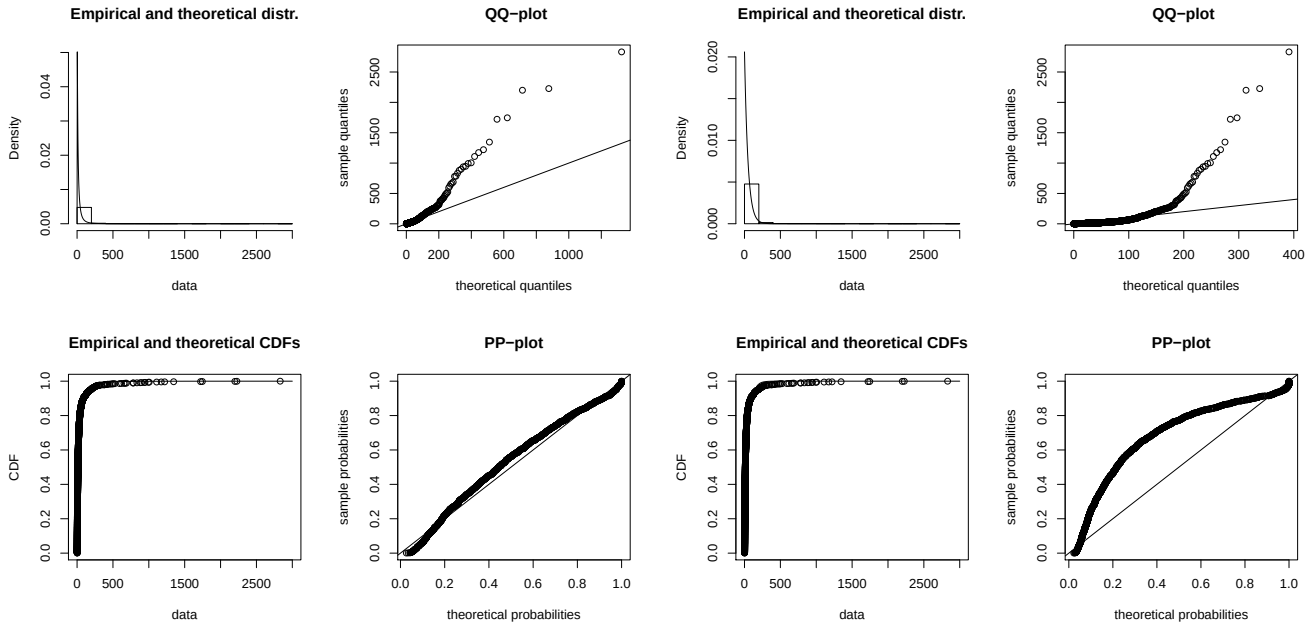
3.3 Distribution Fitting

Distribution fitting is a process of selecting the best distribution model from those constructed. The goal of distribution fitting is to identify the distribution family and the parameters associated with the distribution (i.e., location, scale, and/or shape). A histogram is a manual starting point for defining the data's distribution. For each model constructed, a measure of fit is computed, and the best measure of fit determines the most accurate model.

[2]A *quantile* is a value in an ordered set of values below which a certain number of values fall. Percentiles and quartiles are specific instances of quantiles.

Some graphical approaches allow visual comparisons to be made, rather than using a measure-of-fit value. A *probability-probability plot* (or *P-P plot*) graphs the CDF of the fitted probability distribution against the cumulative frequency of the data. A *quantile-quantile plot* (or *Q-Q plot*) graphs the quantiles of the fitted probability distribution against the quantiles of the data. For both the P-P plot and the Q-Q plot, a perfectly fitted distribution will plot as a straight line from the lower left to the upper right edge of the chart. A *CDF difference plot* graphs the difference, at each point, between the CDF of the fitted distribution and the cumulative frequency of the sample data. A perfectly fitted distribution will plot as a horizontal straight line coinciding with the x-axis.

Some tools, such as R ([R D09]), may have facilities to do most of the distribution fitting work. The R package "fitdistrplus" ([DMPDD10]) contains such functions. To demonstrate, a large collection of think times were collected from a real system. Lognormal and exponential distributions were fitted to the data. One command is sufficient to do most of the work. A second command plots the graphs shown in Figure 10.



(a) Lognormal Distribution Results **(b)** Exponential Distribution Results

Figure 10: Results of distribution fitting: (a) a lognormal distribution and (b) an exponential distribution.

Figure 10a shows the graphs for the lognormal distribution fit, while Figure 10b shows the graphs for the exponential distribution fit. We will focus on the Q-Q plot and P-P plot. For the Q-Q plot, the exponential distribution looks closer to the line. This is deceiving because the scale of exponential distribution graph has a smaller x-axis limit. Nonetheless, both Q-Q plots appear to differ in the tails. But we should note that the 95^{th} percentile of the data is 186. So the differences in the tails exist at the extreme of the distribution.

For the P-P plot, it is clear that the lognormal distribution is much better than the exponential distribution. The fitdistrplus package allows various measure-of-fit values to be computed. Also, a simple command can output the distribution's parameters (location, scale, and/or shape). Without this package, all of the same results can be obtained, but it requires more work by the analyst. [Ric05] describes how to do the fitting work above, plus much more, using the "MASS" package ([VR02]). Much of this work can be done in a spreadsheet tool like Excel, but many of the functions must be implemented by the analyst.

4 Conclusions

This paper provided an introduction to statistics from the performance analyst's perspective. Statistics is generally about summarizing a lot of data with only a little data, and this satisfies the performance analyst's need. The performance analyst's goal is to understand a system's performance, not fumble around with thousands or millions of numbers. Statistics is a communication mechanism—a means, not an end. The next step for the reader is to learn how to use the concepts in a tool. In using a tool, experience will be gained as well as exposure to new frontiers that

the robustness of a tool contains.

While we defined many statistics terms, we did not define others: population, sample, moment, variance, kurtosis, skewness, coefficient of variation, analysis of variance (ANOVA), degree of freedom, confidence interval, residual, time series, and numerous distributions. Some of these terms are foundational and become important for understanding more complex topics. Other terms are infrequent in performance analysis or occur in more advanced predictive statistics concepts.

References

[Cie01] David M. Ciemiewicz. "What Do You 'Mean'? Revisiting Statistics for Web Response Time Measurements". *Proceedings of the CMG 2001 International Conference*, 2001.

[DMPDD10] Marie Laure Delignette-Muller, Regis Pouillot, Jean-Baptiste Denis, and Christophe Dutang. *fitdistrplus*, 2010. R package version 0.1-3.

[Gan98] Donald T. Gantz. "Tutorial on Statistical Methods for Analysis and Reporting for Capacity Planning". *Proceedings of the CMG 1998 International Conference*, pages 871–880, 1998.

[Jai91] Raj Jain. *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* John Wiley & Sons, Inc., 1991.

[Kal02] Denise Kalm. "The Stork Correlation: Use and Abuse of Statistics in Performance and Capacity Planning". *Journal of Computer Resource Management*, 107, Summer 2002.

[Lan07] David M. Lane. *HyperStat Online Statistics Textbook.* `http://davidmlane.com/hyperstat/index.html`, 2007.

[Lil05] David J. Lilja. *Measuring Computer Performance: A Practitioner's Guide.* Cambridge University Press, 2005.

[Mac89] Douglas R. MacKinnon. "Lies, Damned Lies and Statistics". *Proceedings of the CMG 1989 International Conference*, pages 392–403, 1989.

[NIS06] NIST/SEMATECH. *e-Handbook of Statistical Methods.* `http://www.itl.nist.gov/div898/handbook`, 2006.

[R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009.

[Ric05] Vito Ricci. "Fitting Distributions with R". Release 0.4-21, February 2005.

[Sto98] David W. Stockburger. *Introductory Statistics: Concepts, Models, and Applications.* `http://www.psychstat.missouristate.edu/introbook/sbk00.htm`, 1998.

[SW07] Keith Smith and Bob Wescott. *Fundamentals of Performance Engineering: You Can't Spell Firefighter Without IT.* Hyperformix, 2007.

[Szi96] Dick Sziede. "Statistics for the Algebraically Challenged". *Proceedings of the CMG 1996 International Conference*, 1996.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S.* Springer, New York, fourth edition, 2002.

[wik10] Wikipedia. `http://en.wikipedia.org`, 2010. Page specifically referenced is noted in each citation. *Caveat lector: Because of the number of contributors to Wikipedia and the ease with which bad information can be introduced, readers should always validate what they are reading by other established sources. The references are given because of their simple availability.

[Wil10a] Tom Wilson. "Developing Toward an SLA: Understanding Transaction Performance". *CMG MeasureIT*, July 2010.

[Wil10b] Tom Wilson. "Principles of Performance Measurement". *CMG MeasureIT*, June 2010.