1 Introduction to Probability Theory and Statistics

This part introduces the basic concepts of probability theory and statistics. An understanding of these concepts will help you analyze and interpret data properly. After reading this material, you will be able to answer questions such as the ones that follow.

Application - Measuring the performance of a computer system

1. How should you report the performance as a single number?
2. Is specifying the mean the correct way to summarize a sequence of measurements?
3. How should you report the variability of measured quantities?
4. What are the alternatives to variance and when are they appropriate?
5. How should you interpret the variability?
6. How much confidence can you put on data with a large variability?
7. How many measurements are required to get a desired level of statistical confidence?
8. How should you summarize the results of several different workloads on a single computer system?
9. How should you compare two or more computer systems using several different workloads?
10. Is comparing the mean performance sufficient?
11. What model best describes the relationship between two variables? Also, how good is the model?

1.0.1 4.1 Statistical Concepts

1. **Independent Events**: Two events are called independent if the occurrence of one event does not in any way affect the probability of the other event. Thus, knowing that one event has occurred does not in any way change our estimate of the probability of the other event.
2. **Random Variable**: A variable is called a random variable if it takes one of a specified set of values with a specified probability.
3. **Cumulative Distribution Function**: The Cumulative Distribution Function (CDF) of a random variable maps a given value a to the probability of the variable taking a value less than or equal to a:
$$F_x(a) = P(x \leq a) \tag{1}$$
4. **Probability Density Function**: The derivative
$$f(x) = \frac{dF(x)}{dx} \tag{2}$$
of the CDF $F(x)$ is called the probability density function (pdf) of x. Given a pdf $f(x)$, the probability of x being in the interval (x_1, x_2) can also be computed by integration:
$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)dx \tag{3}$$
5. **Probability Mass Function**: For discrete random variable, the CDF is not continuous and, therefore, not differentiable. In such cases, the probability mass function (pmf) is used in place of pdf. Consider a discrete random variable x that can take n distinct values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ such that the probability of the ith value x_i is p_i. The pmf maps x_i to p_i:
$$f(x_i) = p_i \tag{4}$$
. The probability of x being in the interval (x_1, x_2) can also be computed by summation:
$$P(x_1 < x \leq x_2) = F(x_2) - F(x_1) = \sum_{\substack{i \\ x_1 < x_i \leq x_2}} p_i \tag{5}$$
6. **Mean or Expected Value**:
$$\text{Mean } \mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{+\infty} x f(x)dx \tag{6}$$

Summation is used for discrete and integration for continuous variables, respectively.

7. **Variance**: The quantity $(x - \mu)^2$ represents the square of distance between x and its mean. The expected value of this quantity is called the variance x:
$$\text{Var}(x) = E\left[(x-\mu)^2\right] = \sum_{i=1}^{n} p_i (x_i - \mu)^2 = \int_{-\infty}^{+\infty} (x_i - \mu)^2 f(x)dx \tag{7}$$

The variance is traditionally denoted by σ^2. The square root of the variance is called the standard deviation and is denoted by σ.

8. **Coefficient of Variation**: The ratio of the standard deviation to the mean is called the Coefficient of Variation (C.O.V.):
$$\text{C.O.V.} = \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma}{\mu} \tag{8}$$
9. **Covariance**: Given two random variables x and y with means μ_x and μ_y, their covariance is
$$\text{Cov}(x, y) = \sigma_{xy}^2 = E\left[(x - \mu_x)\left(y - \mu_y\right)\right] = E(xy) - E(x)E(y) \tag{9}$$

For independent variables, the covariance is zero since
$$E(xy) = E(x)E(y) \tag{10}$$

1.0.1.1 Expected Value of a Random Variable

Let X be a random variable with a finite number of finite outcomes

$$x_1, x_2, \ldots, x_k \tag{16}$$

occurring with probabilities

$$p_1, p_2, \ldots, p_k, \tag{17}$$

respectively. The expectation of X is defined as

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k \tag{18}$$

Since all probabilities p_i add up to 1 $(p_1 + p_2 + \cdots + p_k = 1)$, the expected value is the weighted average, with p_i's being the weights.

If all outcomes x_i are equiprobable (that is, $p_1 = p_2 = \cdots = p_k$), then the weighted average turns into the simple average. If the outcomes x_i are not equiprobable, then the simple average must be replaced with the weighted average, which takes into account the fact that some outcomes are more likely than the others.

1.0.1.2 Variation and Standard Deviation of a Random Variable

The formula for the sample standard deviation is

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{19}$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items, $\bar{x}$ is the mean value of these observations, and N is the number of observations in the sample.

There are two parameters μ and σ, which are also the mean and standard deviations of x. A normal variate is denoted by $N(\mu, \sigma)$. A normal distribution with zero mean and unit variance is called a unit normal or standard normal distribution and is denoted as $N(O, 1)$. In statistical modeling, you will frequently need to use quantiles of the unit normal distribution. An α-quantile of a unit normal variate z $N(0, 1)$ is denoted by z_α. If a random variable x has a $N(\mu, \sigma)$ distribution, then $(x-\mu)/\sigma$ has a $N(0, 1)$ distribution. Thus,

$$P\left(\frac{x-\mu}{\sigma} \leq z_\alpha\right) = \alpha \tag{20}$$

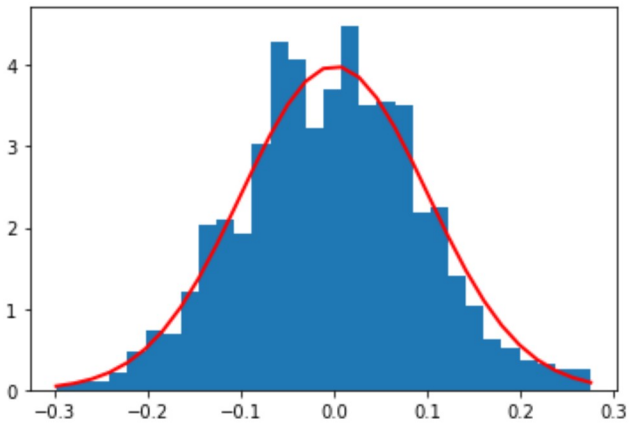
or

$$P(x \leq \mu + z_\alpha \sigma) = \alpha \tag{21}$$

The areas under the unit normal pdf between 0 and z for various values of z can be found from appropriate tables or computed by appropriate functions.

In [23]:

```
#Examples

#Draw samples from the distribution:


mu, sigma = 0, 0.1 # mean and standard deviation
s = np.random.normal(mu, sigma, 1000)
#Verify the mean and the variance:


abs(mu - np.mean(s)) < 0.01


abs(sigma - np.std(s, ddof=1)) < 0.01

#Display the histogram of the samples, along with the probability density fur


import matplotlib.pyplot as plt
count, bins, ignored = plt.hist(s, 30, density=True)
plt.plot(bins, 1/(sigma * np.sqrt(2 * np.pi)) *
                np.exp( - (bins - mu)**2 / (2 * sigma**2) ),
          linewidth=2, color='r')
plt.show()
```



There are two main reasons for the popularity of the normal distribution:

(a) The sum of n independent normal variates is a normal variate. If $x_i \approx N(\mu_i, \sigma_i)$, then $x = \sum_{i=1}^{n} a_i x_i$ has a normal distribution with mean $\mu = \sum_{i=1}^{n} a_i \mu_i$ and variance $\sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$. As a result of this linearity property, normal processes remain normal after passing through linear systems, which are popular in electrical engineering.
(b) The sum of a large number of independent observations from any distribution tends to have a normal distribution. This result, which is called the **central limit theorem**, is true for observations from all distributions. As a result of this property, experimental errors, which are contributed by many factors, are modeled with a normal distribution.

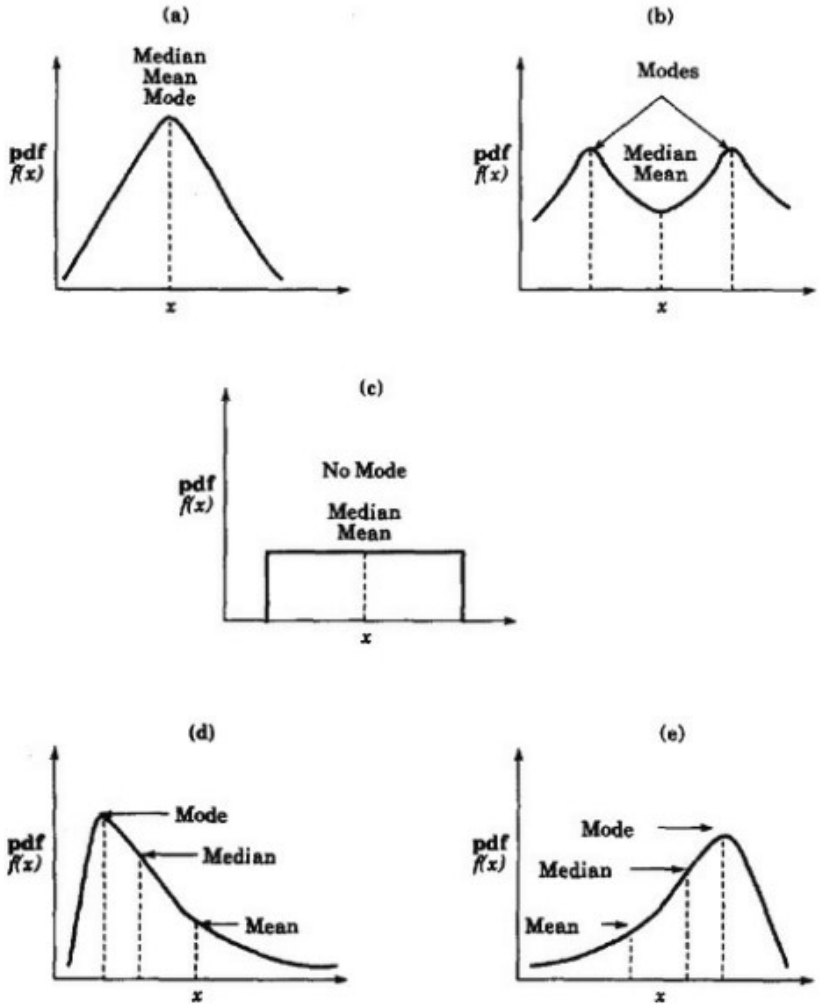
1.0.2 SUMMARIZING DATA BY A SINGLE NUMBER

In the most condensed form, a single number may be presented that gives the key characteristic of the data set. This single number is usually called an **average** of the data. To be meaningful, this average should be representative of a major part of the data set. Three popular alternatives to summarize a sample are to specify its **mean, median, or mode**. These measures are what statisticians call **indices of central tendencies**. The name is based on the fact that these measures specify the center of location of the distribution of the observations in the sample.

Sample mean is obtained by taking the sum of all observations and dividing this sum by the number of observations in the sample. **Sample median** is obtained by sorting the observations in an increasing order and taking the observation that is in the middle of the series. If the number of observations is even, the mean of the middle two values is used as a median. **Sample mode** is obtained by plotting a histogram and specifying the midpoint of the bucket where the histogram peaks. For categorical variables, mode is given by the category that occurs most frequently.

The word **sample** in the names of these indices signifies the fact that the values obtained are based on just one sample. However, if it is clear from the context that the discussion is about a single sample, and there is no ambiguity, the shorter names **mean, median, and mode** can be used.

Mean and median always exist and are unique. Given any set of observations, the mean and median can be determined. Mode, on the other hand, may not exist. An example of this would be if all observations were equal. In addition, even if modes exist, they may not be unique. There may be more than one mode, that is, there may be more than one local peak in the histogram.



1.0.3 SELECTING AMONG THE MEAN, MEDIAN, AND MODE

A common mistake inexperienced analysts make is to specify the wrong index of central tendency. For example, it is common to specify the mean regardless of its validity in a particular situation.

The flow chart of Figure 12.2 shows a set of guidelines to select a proper index of central tendency. The first consideration is the type of variable. If the variable is categorical, the mode is the proper single measure that best describes that data. An example of categorical data is the type of microprocessor in various workstations. A statement such as “the most frequent microprocessor used in workstations is the 68000” makes sense. The mean or median of the type of processor is meaningless.

The second consideration in selecting the index is to ask whether the total of all observations is of any interest. If yes, then the mean is a proper index of central tendency. For example, total CPU time for five queries is a meaningful number. On the other hand, if we count number of windows on the screen during each query, the total number of windows during five queries does not seem to be meaningful. If the total is of interest, specify the mean.

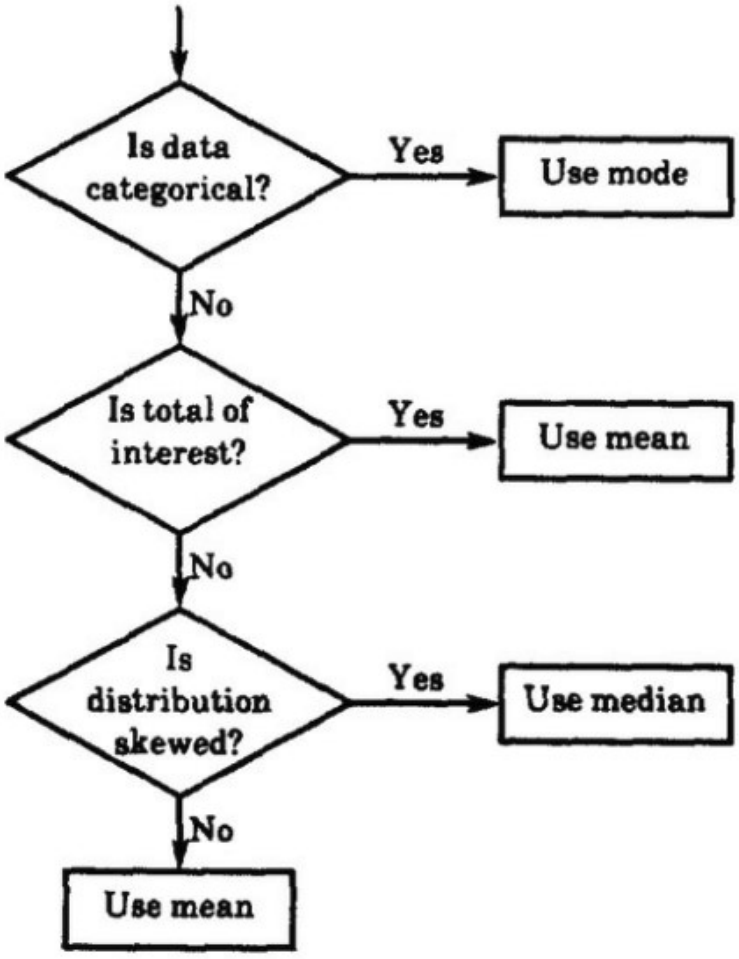


FIGURE 2 Selecting among the mean, meadian, and mode.

If the total is of no interest, one has to choose between median and mode. If the histogram is symmetrical and unimodal, the mean, median, and mode are all equal and it does not really matter which one is specified.

If the histogram is skewed, the median is more representative of a typical observation than the mean. For example, the number of disk drives on engineering workstations is expected to have skewed distribution, and therefore, it is appropriate to specify the median number. One simple to way to determine skewness for small samples is to examine the ratio of the maximum and minimum, ymax/ymin, of the observations. If the ratio is large, the data is skewed.

1.0.4 SUMMARIZING VARIABILITY

Then there is the man who drowned crossing a stream with an average depth of six inches.

— W. I. E. Gates

Given a data set, summarizing it by a single number is rarely enough. It is important to include a statement about its variability in any summary of the data. This is because given two systems with the same mean performance, one would generally prefer one whose performance does not vary much from the mean. For example, Figure 3 shows histograms of the response times of two systems. Both have the same mean response time of 2 seconds. In case (a), the response time is always close to its mean value, while in case (b), the response time can be 1 millisecond sometimes and 1 minute at other times. Which system would you prefer? Most people would prefer the system with low variability.

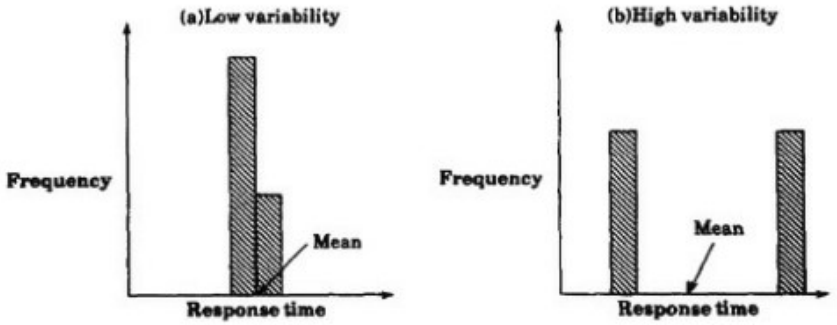


FIGURE 12.3 Histograms of response times of two systems.

Variability is specified using one of the following measures, which are called indices of dispersion:

- Range — minimum and maximum of the values observed
- Variance or standard deviation
- 10- and 90-percentiles
- Semi-interquantile range
- Mean absolute deviation

The range of a stream of values can be easily calculated by keeping track of the minimum and the maximum. The variability is measured by the difference between the maximum and the minimum. The larger the difference, the higher the variability. In most cases, the range is not very useful. The minimum often comes out to be zero and the maximum comes out to be an “outlier” far from typical values. Unless there is a reason for the variable to be bounded between two values, the maximum goes on increasing with the number of observations, the minimum goes on decreasing with the number of observations, and there is no “stable” point that gives a good indication of the actual range. The conclusion is that the range is useful if and only if there is a reason to believe that the variable is bounded. The range gives the best estimate of these bounds.

The variance of a sample of n observations $x_1, x_2, \ldots, x_n$ is calculated as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{22}$$

The quantity s^2 is called the **sample variance** and its square root s is called the **sample standard deviation**. The word sample can be dropped if there is no ambiguity and it it is clear from the context that the quantities refer to just one sample. Notice that in computing the variance, the sum of squares is divided by $n-1$ and not n. This is because only $n-1$ of the n differences are independent. Given $n-1$ differences, the nth difference can be computed since the sum of all n differences must be zero. The number of independent terms in a sum is also called its *degrees of freedom*.

In practice, the main problem with variance is that it is expressed in units that are the square of the units of the observations. For example, the variance of response time could be 4 seconds squared or 4,000,000 milliseconds squared. Changing the unit of measurement has a squared effect on the numerical magnitude of the variance. For this reason, it is preferable to use the standard deviation. It is in the same unit as the mean, which allows us to compare it with the mean. Thus, if the mean response time is 2 seconds and the standard deviation is 2 seconds, there is considerable variability. On the other hand, a standard deviation of 0.2 second for the same mean would be considered small. In fact, the ratio of standard deviation to the mean, or the coefficient of variation (C.O.V.), is even better because it takes the scale of measurement (unit of measurement) out of variability consideration. A C.O.V. of 5 is large, and a C.O.V. of 0.2 (or 20%) is small no matter what the unit is.

Percentiles are also a popular means of specifying dispersion. Specifying the 5-percentile and the 95-percentile of a variable has the same impact as specifying its minimum and maximum. However, it can be done for any variable, even for variables without bounds. When expressed as a fraction between 0 and 1 (instead of a percentage), the percentiles are also called quantiles. Thus 0.9-quantile is the same as 90-percentile.

Another term used is fractile, which is synonymous with quantile. The percentiles at multiples of 10% are called deciles. Thus, the first decile is 10-percentile, the second decile is 20-percentile, and so on. Quartiles divide the data into four parts at 25, 50, and 75%. Thus, 25% of the observations are less than or equal to the first quartile Q_1, 50% of the observations are less than or equal to the second quartile Q_2, and 75% are less than or equal to the third quartile Q_3. Notice that the second quartile Q_2 is also the median. The α-quantiles can be estimated by sorting the observations and taking the $[(n-1)\alpha + 1]$th element in the ordered set. Here, $[.]$ is used to denote rounding to the nearest integer. For quantities exactly halfway between two integers, use the lower integer.

The range between Q_3 and Q_1 is called the interquartile range of the data. One half of this range is called Semi-Interquartile Range (SIQR), that is,

$$\text{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{x_{0.75} - x_{0.25}}{2} \tag{23}$$

Another measure of dispersion is the mean absolute deviation, which is calculated as follows:

$$\text{Mean absolute deviation } = \frac{1}{n}\sum_{i=1}^{n} |x_i - x| \tag{24}$$

The key advantage of the mean absolute deviation over the standard deviation is that no multiplication or square root is required.

Among the preceding indices of dispersion, the range is affected considerably by outliers. The sample variance is also affected by outliers, but the effect is less than that on the range. The mean absolute deviation is next in resistance to outliers. The semi-interquantile range is very resistant to outliers. It is preferred to the standard deviation for the same reasons that the median is preferred to the mean. Thus, if the distribution is highly skewed, outliers are highly likely and the SIQR is more representative of the spread in the data than the standard deviation. In general, the SIQR is used as an index of dispersion whenever the median is used as an index of central tendency.

Finally, it should be mentioned that all of the preceding indices of dispersion apply only for quantitative data. For qualitative (categorical) data, the dispersion can be specified by giving the number of most frequent categories that comprise the given percentile, for instance, the top 90%.

Example 1 In an experiment, which was repeated 32 times, the measured CPU time was found to be {3.1, 4.2, 2.8, 5.1, 2.8, 4.4, 5.6, 3.9, 3.9, 2.7, 4.1, 3.6, 3.1, 4.5, 3.8, 2.9, 3.4, 3.3, 2.8, 4.5, 4.9, 5.3, 1.9, 3.7, 3.2, 4.1, 5.1, 3.2, 3.9, 4.8, 5.9, 4.2}. The sorted set is {1.9, 2.7, 2.8, 2.8, 2.8, 2.9, 3.1, 3.1, 3.2 3.2, 3.3, 3.4, 3.6, 3.7, 3.8, 3.9, 3.9, 3.9, 4.1, 4.1, 4.2, 4.2, 4.4, 4.5, 4.5, 4.8, 4.9, 5.1, 5.1, 5.3, 5.6, 5.9}. Then

```
The 10-percentile is given by [1 + (31)(0.10)] = 4th element = 2.8.

The 90-percentile is given by [1 + (31)(0.90)] = 29th element = 5.1.
```

1.0.5 Summarizing Observations

Given: A sample $x_1, x_2, \ldots, x_n$ of n observations.

1. Sample arithmetic mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
2. Sample geometric mean: $\dot{x} = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$
3. Sample harmonic mean: $x = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$
4. Sample median:

$$\begin{cases} x_{((n-1)/2)} & \text{if } n \text{ is odd} \\ 0.5 \left(x_{(n/2)} + x_{((1+n)/2)} \right) & \text{otherwise} \end{cases} \tag{26}$$

   Here x(i) is the ith observation in the sorted set.
5. Sample mode = observation with the highest frequency (for categorical data).
6. Sample variance: $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
7. Sample standard deviation: $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$
8. Coefficient of variation = $\frac{s}{\bar{x}}$
9. Coefficient of skewness = = $\frac{1}{ns^3} \sum_{i=1}^{n} (x_i - \bar{x})^3$
10. Range: Specify the minimum and maximum.
11. Percentiles: 100p-percentile
12. Semi-interquartile range $\mathrm{SIQR} = \frac{Q_3 - Q_1}{2} = \frac{x_{0.75} - x_{025}}{2}$
13. Mean absolute deviation = $\frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$

1.0.6 DETERMINING DISTRIBUTION OF DATA

In the last two cells we discussed how a measured data set could be summarized by stating its average and variability. The next step in presenting a summary could be to state the type of distribution the data follows. For example, a statement that the number of disk I/O's are uniformly distributed between 1 and 25 is a more meaningful summary than to specify only that the mean is 13 and the variance is 48. The distribution information is also required if the summary has to be used later in simulation or analytical modeling.

The simplest way to determine the distribution is to plot a histogram of the observations. This requires determining the maximum and minimum of the values observed and dividing the range into a number of subranges called cells or buckets. The count of observations that fall into each cell is determined. The counts are normalized to cell frequencies by dividing by the total number of observations. The cell frequencies are plotted as a column chart.

The key problem in plotting histograms is determining the cell size. Small cells lead to very few observations per cell and a large variation in the number of observations per cell. Large cells result in less variation but the details of the distribution are completely lost. Given a data set, it is possible to reach very different conclusions about the distribution shape depending upon the cell size used. One guideline is that if any cell has less than five observations, the cell size should be increased or a variable cell histogram should be used.

A better technique for small samples is to plot the observed quantiles versus the theoretical quantile in a quantile-quantile plot. Suppose, $y_(i)$ is the observed q_ith quantile. Using the theoretical distribution, the q_ith quantile x_i is computed and a point is plotted at $(x_i, y_(i))$. If the observations do come from the given theoretical distribution, the quantile-quantile plot would be linear.

To determine the q_ith quantile x_i, we need to invert the cumulative distribution function. For example, if $F(x)$ is the CDF for the assumed distribution,

$$q_i = F(x_i) \tag{27}$$

or

$$x_i = F^{-1}(q_i) \tag{28}$$

For those distributions whose CDF can be inverted, determining the x-coordinate of points on a quantile-quantile plot is straightforward.

For other distributions one can use tables and interpolate the values if necessary. For the unit normal distribution $N(0, 1)$, the following approximation is often used:

$$x_i = 4.91\left[q_i^{0.14} - (1 - q_i)^{0.14}\right] \tag{29}$$

For $N(\mu, \sigma)$, the x_i values computed by the above Equation are scaled to $\mu + \sigma x_i$ before plotting.

One advantage of a quantile-quantile plot is that often it is sufficient to know the name of the possible distribution. The parameter values are not required. This happens if the effect of the parameters is simply to scale the quantile. For example, in a normal quantile-quantile plot, x-coordinates can be obtained using the unit normal $N(0, 1)$ distribution. The intercept and the slope of the resulting line give the values of location and shape parameters μ and σ.

Example 2 The difference between the values measured on a system and those predicted by a model is called modeling error. The modeling error for eight predictions of a model were found to be -0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, and 0.09.

In [25]:

```
# Setup
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats

np.random.seed(0)

mu = 0  # mean
sigma = 1  # standard deviation

points = np.random.normal(mu, sigma, 1000)

print("First 10 points (of 1000):", points[:10])
```

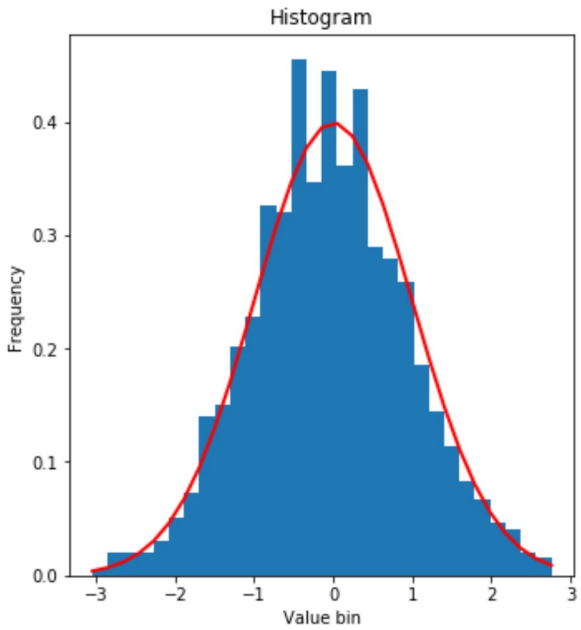
```
First 10 points (of 1000): [ 1.76405235  0.40015721  0.97873798  2.2408932   1.8
6755799 -0.97727788
  0.95008842 -0.15135721 -0.10321885  0.4105985 ]
count    1000.000000
mean       -0.045257
std         0.987527
min        -3.046143
25%        -0.698420
50%        -0.058028
75%         0.606951
max         2.759355
dtype: float64
```

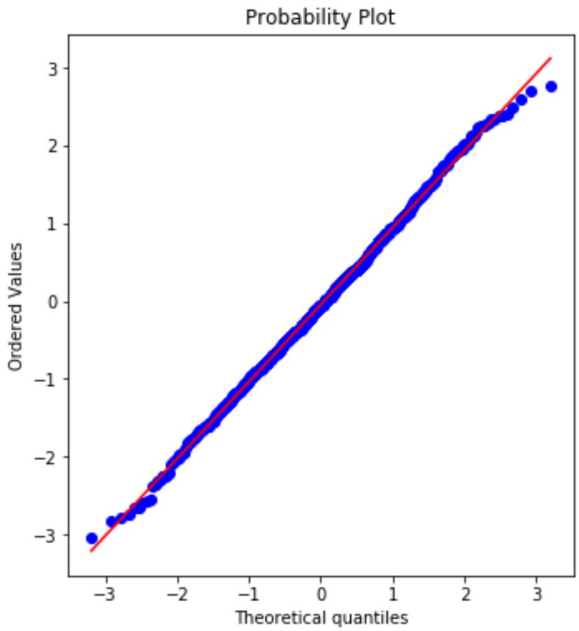
In [24]:

```
def plot_histogram_and_qq(points, mu, sigma, distribution_type="norm"):
  # Plot histogram of the 1000 points
  plt.figure(figsize=(12,6))
  ax = plt.subplot(1,2,1)
  count, bins, ignored = plt.hist(points, 30, normed=True)
  ax.set_title('Histogram')
  ax.set_xlabel('Value bin')
  ax.set_ylabel('Frequency')

  # Overlay the bell curve (normal distribution) on the bins data
  bell_curve = 1/(sigma * np.sqrt(2 * np.pi)) * np.exp( - (bins - mu)**2 / (2
  plt.plot(bins, bell_curve, linewidth=2, color='r')

  # Q-Q plot
  plt.subplot(1,2,2)
  res = stats.probplot(points, dist=distribution_type, plot=plt)
  (osm, osr) = res[0]
  (slope, intercept, r) = res[1]
  # For details see: https://docs.scipy.org/doc/scipy-0.14.0/reference/genera
  print("slope, intercept, r:", slope, intercept, r)
  print("r is the square root of the coefficient of determination")

```

In [26]:
```
1  # Run on the initial normally distributed data
```

```
C:\Users\IliasAlexis\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: Matplo
tlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in 3.1.
Use 'density' instead.
  """

slope, intercept, r: 0.9892713568120576 -0.045256707490195364 0.9994824641317025
r is the square root of the coefficient of determination
```

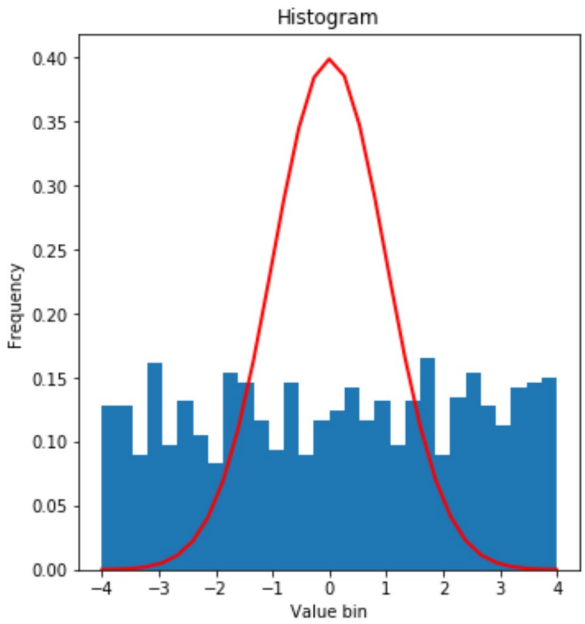


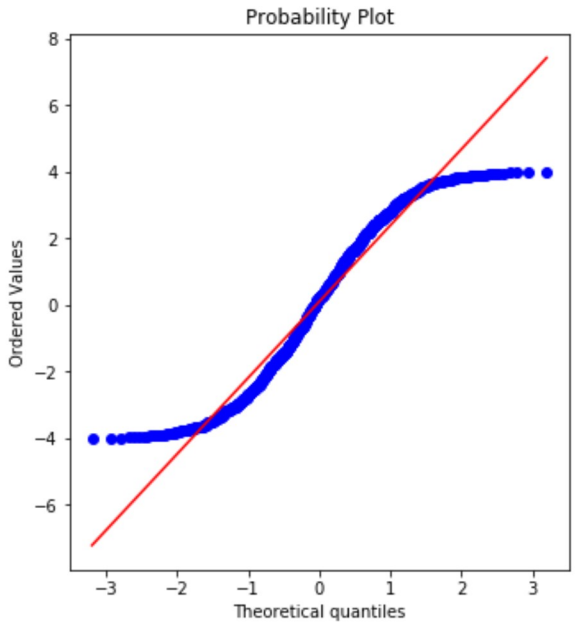


In [27]:

```
1  points = np.random.uniform(low=-4, high=4, size=1000)
2
3  print("First 10 points (of 1000):", points[:10])
4  print(pd.Series(points).describe())
5
6  # Run on the initial setup
```

```
First 10 points (of 1000): [ 2.57523127  1.60422898  3.06462078  3.73260086  2.1
9798091  3.95386467
  0.91815909 -3.70296317 -3.88598788 -1.263169  ]
count    1000.000000
mean        0.102926
std         2.337933
min        -3.999410
25%        -1.851285
50%         0.225588
75%         2.152776
max         3.988213
dtype: float64

C:\Users\IliasAlexis\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: Matplo
tlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in 3.1.
Use 'density' instead.
  """

slope, intercept, r: 2.288755908442107 0.10292591860780509 0.9767335657066317
r is the square root of the coefficient of determination
```

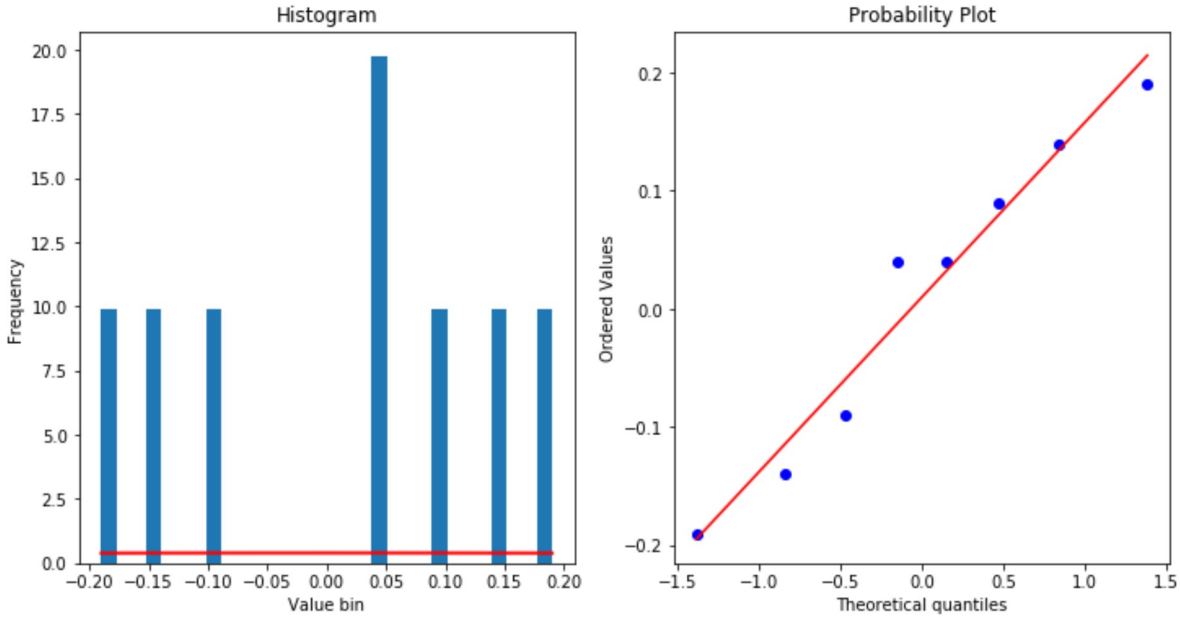




In [40]:

```
1
2
3  points = (0.04, -0.19, 0.14, -0.09, -0.14, 0.19, 0.04, 0.09)
4  n = 8
5  q=np.zeros(n)
6  print(q)
7  x = (0.157,-1.535,0.885,-0.487,-0.885 ,1.535 ,0.157 ,0.487 )
8  for i in range(8):
9      q[i]= (i-0.5)/n
10 mu=0
11 sigma = 1
12 plot_histogram_and_qq(points, mu, sigma)
```

```
C:\Users\IliasAlexis\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: Matplo
tlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in 3.1.
Use 'density' instead.
  """

slope, intercept, r: 0.14768759695342298 0.01000000000000001 0.9796285130684926
r is the square root of the coefficient of determination
```



1.1 Q Q Plots: Simple Definition & Example

1.1.1 Descriptive Statistics > Q Q plots

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

1.1.2 How to Make a Q Q Plot

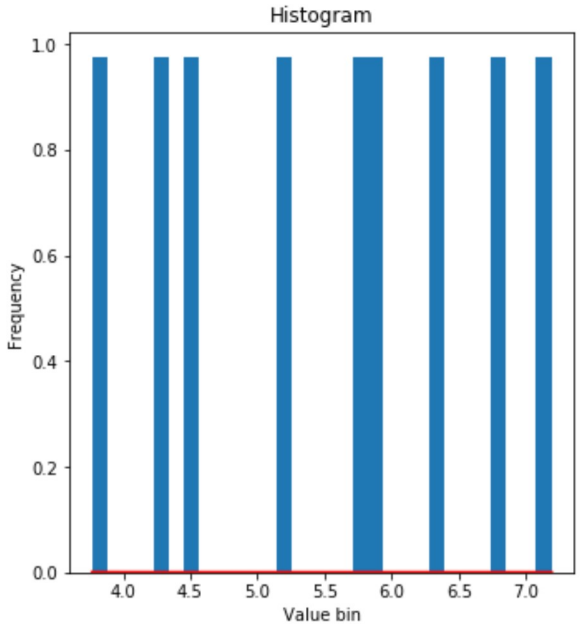
Sample question: Do the following values come from a normal distribution? 7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79.

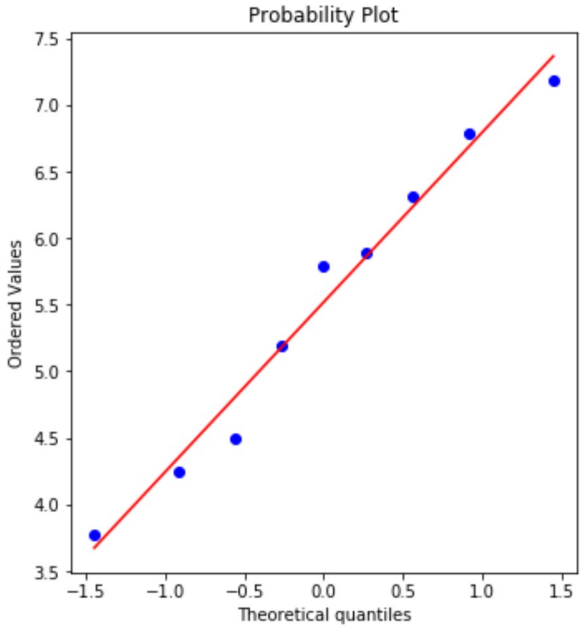
```
Step 1: Order the items from smallest to largest.
Step 2: Draw a normal distribution curve. Divide the curve into n+1 segments.
We have 9 values, so divide the curve into 10 equally-sized areas. For this e
xample, each segment is 10% of the area (because 100% / 10 = 10%).
Step 3: Find the z-value (cut-off point) for each segment in Step 3. These se
gments are areas, so refer to a z-table (or use software) to get a z-value fo
r each segment.
```

In [41]:

```
points = ( 7.19, 6.31, 5.89, 4.5, 3.77, 4.25, 5.19, 5.79, 6.79)
n = 9
q=np.zeros(n)
print(q)
for i in range(8):
    q[i]= (i-0.5)/n
mu=0
sigma = 1
```

```
[0. 0. 0. 0. 0. 0. 0. 0. 0.]
slope, intercept, r: 1.2768810949841627 5.52 0.9896069959517892
r is the square root of the coefficient of determination

C:\Users\IliasAlexis\Anaconda3\lib\site-packages\ipykernel_launcher.py:5: Matplo
tlibDeprecationWarning:
The 'normed' kwarg was deprecated in Matplotlib 2.1 and will be removed in 3.1.
Use 'density' instead.
  """
```





Exercise 1 A distributed system has three file servers, which are chosen independently and with equal probabilities whenever a new file is created. The servers are named A, B, and C. Determine the probabilities of the following events: a. Server A is selected b. Server A or B is selected c. Servers A and B are selected d. Server A is not selected e. Server A is selected twice in a row f. Server selection sequence ABCABCABC is observed (in nine successive file creations)

Exercise 2

The traffic arriving at a network gateway is bursty. The burst size x is geometrically distributed with the following pmf.

$$f(x) = (1-p)^{x-1}p, x = 1, 2, \ldots, \infty \tag{30}$$

Compute the mean, variance, standard deviation, and coefficient of variation of the burst size. Plot the pmf and CDF for P = 0.2.

Exercise 3 The number of I/O requests received at a disk during a unit interval follows a Poisson distribution with the following mass function:

$$f(x) = \lambda^x \frac{e^{-\lambda x}}{x!}, x = 0, 1, 2 \ldots, \infty \tag{31}$$

Here, λ is a parameter. Determine the mean, variance, and coefficient of variation of the number. Plot the pmf and CDF for λ = 8.

Exercise 4 Two Poisson streams merge at a disk. The pmf for the two streams are as follows:

$$f(x) = \lambda^x \frac{e^{-\lambda x}}{x!}, x = 0, 1, 2, \ldots, \infty \tag{32}$$

$$f(y) = \lambda^y \frac{e^{-\lambda y}}{y!}, y = 0, 1, 2, \ldots, \infty \tag{33}$$

Determine the following: a. Mean of $x + y$ b. Variance of $x + y$ c. Mean of $x - y$ d. Variance of $x - y$ e. Mean of $3x - 4y$ f. Coefficient of variation of $3x - 4y$

Exercise 5 The response time of a computer system has an Erlang distribution with the following CDF:

$$F(x) = 1 - e^{-x/a}\left(\sum_{i=0}^{m-1} \frac{(x/a)^i}{i!}\right) \tag{34}$$

Find expressions for the pdf, mean, variance, mode, and coefficient of variation of the response time.

Exercise 6 The execution times of queries on a database is normally distributed with a mean of 5 seconds and a standard deviation of 1 second. Determine the following:

a. What is the probability of the execution time being more than 8 seconds?

b. What is the probability of the execution time being less than 6 seconds?

c. What percentage of responses will take between 4 and 7 seconds?

d. What is the 95-percentile execution time?

Exercise 7

Plot a normal quantile-quantile plot for the following sample of errors: -0.04444 -0.04439 -0.04165 -0.03268 -0.03235 -0.03182 0.02771 0.02650
-0.02569 -0.02358 0.02330 0.02305 0.02213 0.02128 0.01793 0.01668
-0.01565 -0.01509 0.01432 0.00978 0.00889 0.00687 0.00543 0.00084

In [ ]:

```
1
```