A STATE OF STATISTICAL CONTROL IS NOT A NATURAL STATE FOR A MANUFACTURING PROCESS. IT IS INSTEAD AN ACHIEVEMENT, ARRIVED AT BY ELIMINATING ONE BY ONE, BY DETERMINED EFFORT, THE SPECIAL CAUSES OF EXCESSIVE VARIATION.

W. EDWARDS DEMING

Introduction

Basic Statistics is presented in two major category areas:

- **General Concepts**
- **Calculations**

General Concepts is reviewed in the following topic areas:

- **Terminology**
- **Frequency Distributions**

Terminology

To follow are a number of basic quality and statistical terms. Most of these are included in the CQT BOK. Sources of the definitions are identified in the references at the end of this Primer Section.

Parameter

The true population value, often unknown, estimated by a statistic.

(Omdahl, 2010)[7]

Population

All possible observations of similar items from which a sample is drawn.

(Omdahl, 2010)[7]

Quality

The degree to which a set of inherent characteristics fulfill requirements. Also, the totality of features and characteristics of a product, activity, or system that bears on its ability to satisfy stated or implied needs. (ASQ, 1992)[1], (ISO 8402, 1994)[5]

Terminology (Continued)

Quality Assurance

All those planned and systematic actions necessary to provide adequate confidence that a product or service will satisfy given quality requirements.
(ANSI/ASQ ISO 9000:2005)[1]

Quality Control

The operational techniques and activities that are used to fulfill requirements for quality. **(ANSI/ASQ ISO 9000:2005)[1]**

Quality System

The organizational structure, responsibilities, procedures, processes, and resources for implementing quality management. **(ISO 8402, 1992)[5]**

Sample

A group of units or observations taken from a larger collection of units or observations that serves to provide an information basis for making a decision concerning the larger quantity. **(Omdahl, 2010)[7]**

Statistic

A numerical data measurement taken from a sample that may be used to make an inference about a population. **(Omdahl, 2010)[7]**

Random Variable

A random variable is any observation that can vary. It can represent either discrete or continuous data. **(Grant, 1988)[4]**

Continuous Frequency Distributions

A continuous distribution contains infinite (variable) data points that may be displayed on a continuous measurement scale. Examples: include normal, exponential and Weibull distributions.

The exponential and Weibull distributions will not be on the CQT exam. They are widely used in reliability areas. Only statisticians work with the normal distribution formula. Most professionals use Z values to determine failure rates. A discussion of Z values occurs later in this Primer Section.

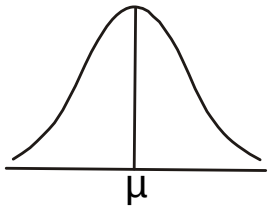
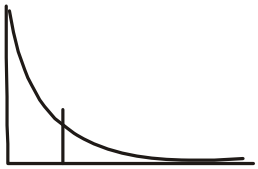
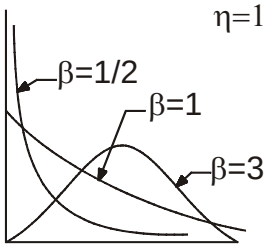
	NORMAL (GAUSSIAN)	EXPONENTIAL	WEIBULL
SHAPE			
FORMULAS	$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ μ = Mean σ = Standard deviation e = 2.718	$P(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$ or $P(x) = \lambda e^{-\lambda x}$ μ = θ = Mean X = X axis reading λ = failure rate	$P(x) = \frac{\beta}{\eta}(x-\gamma)^{\beta-1} e^{-\frac{(x-\gamma)^{\beta}}{\eta}}$ η = Scale parameter β = Shape parameter γ = Location parameter
APPLICATIONS	Numerous applications. Useful when it is equally likely that readings will fall above or below the average.	Describes constant failure rate conditions. Applies for the useful life cycle of many products. Often, time(t) is used for X.	Used for many reliability applications. Can test for the end of the infant mortality period. Can also describe the normal and exponential distributions.

Table 4.1 A Comparison of Continuous Distributions

Discrete Frequency Distributions

A discrete distribution results from countable (attribute) data that has a finite number of possible values. Examples include binomial, Poisson, and hypergeometric distributions.

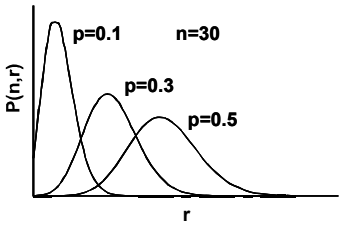
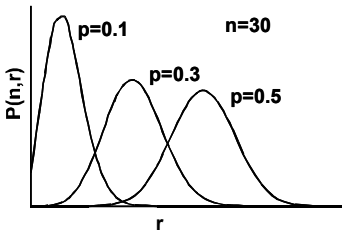
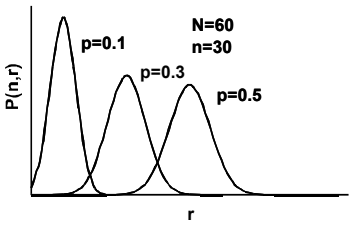
	POISSON	BINOMIAL	HYPERGEOMETRIC
SHAPE			
FORMULAS	$P(r) = \frac{(np)^r e^{-np}}{r!}$ n = Sample size r = Number of occurrences p = Probability np = μ = Average	$P(r) = \frac{n!}{r!(n-r)!} p^r q^{n-r}$ n = Sample size r = Number of occurrences p = Probability q = 1 - p	$P(r) = \frac{\binom{d}{r}\binom{N-d}{n-r}}{\binom{N}{n}}$ n = Sample size r = Number of occurrences d = Occurrences in population N = Population size
APPLICATIONS	The Poisson is used as a distribution for defect counts and can be used as an approximation to the binomial. For np < 5 the binomial is better approximated by the Poisson than the Normal.	The binomial is an approximation to the hypergeometric. Sampling is with replacement. The sample size is less than 10 % of N (n < 10 % of N). The normal distribution approximates the binomial when np ≥ 5.	Used when the number of defects (d) is known. Sampling is without replacement. The population size (N) is frequently small. Applied when the sample (n) is a relatively large proportion of the population (n > 10 % of N).

Table 4.2 A Comparison of Discrete Distributions

Hypergeometric Frequency Distribution

The hypergeometric distribution will not be on the CQT exam. It probably should be reviewed from a comprehension standpoint. ANSI/ASQ Z1.4 (2008)[3] for small populations is based on the hypergeometric distribution.

The hypergeometric distribution applies when the population is small compared to the sample size. Sampling is done without replacement. The hypergeometric distribution is a complex combination calculation.

The number of occurrences (r)* in the sample follows the hypergeometric function:

$$P(r) = \frac{C_r^d C_{n-r}^{N-d}}{C_n^N}$$

Where:
- **N = Population size**
- **n = Sample size**
- **d = Number of occurrences in the population**
- **N - d = Number of non occurrences in the population**
- **r = Number of occurrences* in the sample**

***r can also equal the number of defectives or successes in a sample. The term X is used instead of r in many texts.**

Example 4.1: From a group of 20 products, containing 5 defectives, 10 are selected at random. What is the probability that these 10 contain the 5 defectives?

N = 20, n = 10, d = 5, (N-d) = 15 and r = 5

$$P(r) = \frac{C_5^5 \, C_5^{15}}{C_{10}^{20}} \qquad \text{note that } C_r^n = \frac{n!}{r!(n-r)!}$$

$$P(r) = \frac{\left(\frac{5!}{5!0!}\right)\left(\frac{15!}{5!10!}\right)}{\left(\frac{20!}{10!10!}\right)} = \left(\frac{15!}{5!10!}\right)\left(\frac{10!10!}{20!}\right)$$

Answer $P(r) = 0.0163 = 1.63\ \%$

Binomial Frequency Distribution

The binomial distribution applies when the population is large (N > 50) and the sample size is small compared to the population. Generally, n is less than 10 % of N. It is most appropriate to use when the proportion defective is equal to or greater than (0.1).

$$P(r) = C_r^n p^r (1 - p)^{n - r} = \frac{n!}{r!(n - r)!} p^r (1 - p)^{n - r}$$

Where: n = Sample size
r = Number of defectives
p = Proportion defective

Note: 4! = 1x 2 x 3 x 4 = 24

Example 4.2: A random sample of 10 units is taken from a steady stream of product from a press. Past experience has shown 10 % defective parts. Find the probability of exactly one bad part.

$$n = 10 \quad r = 1 \quad p = 0.1$$

$$P(r) = C_r^n p^r (1 - p)^{n - r}$$

$$P(r) = \frac{10!}{1!9!}(0.1)^1(0.9)^9$$

$$P(r) = (10)(0.1)(0.3874)$$

$$\text{Answer } P(r) = 0.3874 = 38.74\ \%$$

Solve for 2 bad parts (answer = 19.37 %). Solve for 0 bad parts (answer = 34.87 %)

Note: There is a limited binomial probability table in the Appendix. This table will save considerable calculation time if the sample size is 10 or less.

Binomial Frequency Distribution (Continued)

The binomial distribution average and sigma can be obtained from the following calculations:

$$\text{The binomial average} = \mu = n\bar{p}$$

$$\text{The binomial sigma} = \sigma = \sqrt{np(1 - p)}$$

Example 4.3: If one tosses an honest coin 100 times, what is expected to be the average number of heads? What will be the 3 sigma variation?

$n = 100 \quad \bar{p} = 0.5$

Answer: $\mu = n\bar{p} = (100)(0.5) = 50$ heads

$\text{Sigma} = \sqrt{n\bar{p}(1 - \bar{p})} = \sqrt{(50)(1 - 0.5)} = \sqrt{25} = 5$

Answer: $\mu \pm 3s = 50 \pm 15$ heads

Poisson Frequency Distribution

The Poisson is another discrete distribution that has numerous applications in industry. The Poisson is also an approximation to the binomial distribution when p is equal to or less than 0.1, and the sample size n is fairly large.

$$P(r) = \frac{\mu^r e^{-\mu}}{r!}$$

Where: $\mu = n\bar{p}$ **= the population mean**

r = number of defectives

e = 2.71828 the base of natural logarithms

Poisson Frequency Distribution (Continued)

Example 4.4: A continuous process is running a 2 % defective rate. What is the probability that a 100 piece sample will contain exactly 2 defectives?

$$\mu = n\bar{p} = (100)(0.02) = 2 \qquad r = 2$$

$$P(r) = \frac{\mu^r e^{-\mu}}{r!} = \frac{2^2 e^{-2}}{2!} = \frac{2^2}{2!e^2}$$

$$\text{Answer } P(r) = \frac{4}{(2)(7.389)} = 0.27 = 27\ \%$$

Solve for r = 0 Answer 0.135 (13.5 %)
Solve for r = 1 Answer 0.27 (27 %)
Solve for r = 3 Answer 0.18 (18 %)

The very good news is that Poisson tables exist which allow an easy determination of probabilities. To use the Poisson table, a quick calculation of $n\bar{p}$ is needed. In a previous problem, $n\bar{p} = \mu = 2$ From the Poisson table in the Appendix: The probability of stated or fewer defectives equals:

np \ r	0	1	2	3	4	5	6
2	0.135	0.406	0.677	0.857	0.947	0.983	0.995

Thus, for np = 2.0 the probability of stated defects only equals:

np \ r	0	1	2	3	4	5	6
2	0.135	0.271	0.271	0.180	0.090	0.036	0.012

The Poisson distribution average and sigma can be obtained from the following calculations:

$$\text{The Poisson average} = \mu = n\bar{p} = \bar{c}\ *$$

$$\text{The Poisson sigma} = \sigma = \sqrt{\mu} = \sqrt{n\bar{p}} = \sqrt{\bar{c}}\ *$$

*** From the attribute c chart.**

Calculations

Calculations are presented in the following topic areas:

- **Measures of Central Tendency**
- **Measures of Dispersion**
- **Statistical Inference**
- **Confidence Limits**
- **Probability**

Measures of Central Tendency

Measures of central tendency represent different ways of characterizing the central value of a collection of data. Three of these measures will be addressed here: mean, mode and median.

The Mean (X-bar, $\bar{X}$)

The mean is the sum total of all data values divided by the number of data points.

Formula: $\bar{X} = \frac{\sum X}{n}$

$\bar{X}$ is the mean

X represents each number

$\sum$ means summation

n is the sample size

Example 4.5: (9 Numbers) 5 3 7 9 8 5 4 5 8

Find $\bar{X}$: **Answer: 6**

The arithmetic mean is the most widely used measure of central tendency.

<u>Advantages of using the mean:</u>

- **It is the center of gravity of the data**
- **It uses all data**
- **No sorting is needed**

Measures of Central Tendency (Continued)

The Mean (X-bar, $\bar{X}$) (Continued)

Disadvantages of using the mean:

- Extreme data values may distort the picture
- It can be time consuming
- The mean may not be the actual value of any data points

The Mode

The mode is the most frequently occurring number in a data set.

Example 4.6: (9 Numbers) 5 3 7 9 8 5 4 5 8
Find the mode: Answer: 5

Note: It is possible for groups of data to have more than one mode.

Advantages of using the mode:

- No calculations or sorting is necessary
- It is not influenced by extreme values
- It is an actual value
- It can be detected visually in distribution plots

Disadvantage of using the mode:

- The data may not have a mode

Measures of Central Tendency (Continued)

The Median (Midpoint)

The median is the middle value when the data is arranged in ascending or descending order. For an even set of data, the median is the average of the middle two values.

Examples 4.7: (10 Numbers)	**2 2 2 3 4 6 7 7 8 9**
(9 Numbers)	**2 2 3 4 5 7 8 8 9**
Find the median:	**Answer: 5 for both examples**

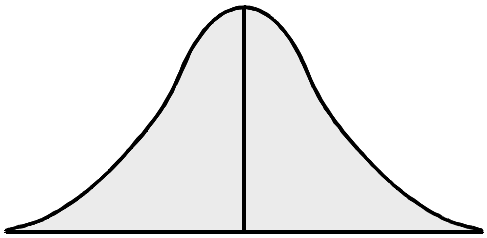
<u>Advantages of using the median:</u>

- **Provides an idea where most data are located**
- **Little calculation required**
- **Insensitivity to extreme values**

<u>Disadvantages of using the median:</u>

- **The data must be sorted and arranged**
- **Extreme values may be important**
- **Two medians cannot be averaged to obtain a combined median**
- **The median will have more variation (between samples) than the average ($\bar{X}$)**

For a Normal Distribution



For a Right Skewed Distribution

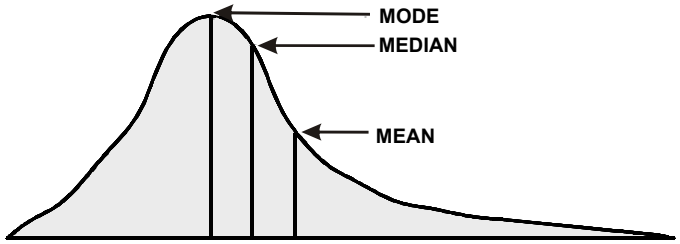


Figure 4.3 A Comparison of Central Tendency in Normal and Skewed Distributions

Measures of Dispersion

Other than central tendency, the other important parameter to describe a set of data is spread or dispersion. Three main measures of dispersion will be reviewed: range, variance, and standard deviation.

Range (R)

The range of a set of data is the difference between the largest and smallest values.

Example 4.8: (9 Numbers) 5 3 7 9 8 5 4 5 8

Find R: Answer: 6

Variance (σ^2, S^2)

The variance, σ^2 or S^2, equal to the sum of the squared deviations from the mean, divided by the sample size. The formulas for variance are:

$$\text{Population, } \sigma^2 = \frac{\sum(X - \mu)^2}{N} \qquad \text{Sample, } S^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

The variance is equal to the standard deviation squared.

Standard Deviation (σ, s)

The standard deviation is the square root of the variance.

$$\text{Population, } \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \qquad \text{Sample, } S = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Note: N is used for a population, and n - 1 for a sample (to remove potential bias in relatively small samples - less than 30)

Coefficient of Variation (COV)

The coefficient of variation equals the standard deviation divided by the mean and is expressed as a percentage.

$$COV = \frac{S}{\bar{X}} \bullet 100 \quad \text{or} \quad COV = \frac{\sigma}{\mu} \bullet 100$$

