7. Product and Process Comparisons

This chapter presents the background and specific analysis techniques needed to compare the performance of one or more processes against known standards or one another.

1. Introduction

1. Scope
2. Assumptions
3. Statistical Tests
4. Confidence Intervals
5. Equivalence of Tests and Intervals
6. Outliers
7. Trends

2. Comparisons: One Process

1. Comparing to a Distribution
2. Comparing to a Nominal Mean
3. Comparing to Nominal Variability
4. Fraction Defective
5. Defect Density
6. Location of Population Values

3. Comparisons: Two Processes

1. Means: Normal Data
2. Variability: Normal Data
3. Fraction Defective
4. Failure Rates
5. Means: General Case

4. Comparisons: Three + Processes

1. Comparing Populations
2. Comparing Variances
3. Comparing Means
4. Variance Components
5. Comparing Categorical Datasets
6. Comparing Fraction Defectives
7. Multiple Comparisons

Detailed table of contents
References for Chapter 7

7. Product and Process Comparisons - Detailed Table of Contents [7.]

7.1. Introduction

Goals of this section

The primary goal of this section is to lay a foundation for understanding statistical tests and confidence intervals that are useful for making decisions about processes and comparisons among processes. The materials covered are:

- Scope
- Assumptions
- Introduction to hypothesis testing
- Introduction to confidence intervals
- Relationship between hypothesis testing and confidence intervals
- Outlier detection
- Detection of sequential trends in data or processes

Hypothesis testing and confidence intervals

This chapter explores the types of comparisons which can be made from data and explains hypothesis testing, confidence intervals, and the interpretation of each.

7.1.1. What is the scope?

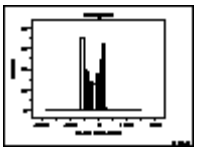
Data from one process	This section deals with introductory material related to comparisons that can be made on data from one process for cases where the process standard deviation may be known or unknown.

7.1.2. What assumptions are typically made?

Validity of tests

The validity of the tests described in this chapter depend on the following assumptions:

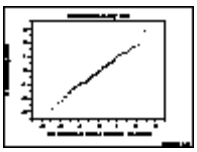
1. The data come from a single process that can be represented by a single statistical distribution.
2. The distribution is a normal distribution.
3. The data are uncorrelated over time.



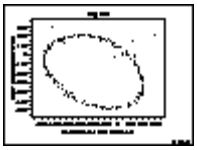
An easy method for checking the assumption of a single normal distribution is to construct a histogram of the data.

Clarification

The tests described in this chapter depend on the assumption of normality, and the data should be examined for departures from normality before the tests are applied. However, the tests are robust to small departures from normality; i.e., they work fairly well as long as the data are bell-shaped and the tails are not heavy. Quantitative methods for checking the normality assumption are discussed in the next section.



Another graphical method for testing the normality assumption is the normal probability plot.



A graphical method for testing for correlation among measurements is a time-lag plot. Correlation may not be a problem if measurements are properly structured over time. Correlation problems often occur when measurements are made close together in time.

7.1.3. What are statistical tests?

What is meant by a statistical test?

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process. The conjecture is called the null hypothesis. Not rejecting may be a good result if we want to continue to act as if we "believe" the null hypothesis is true. Or it may be a disappointing result, possibly indicating we may not yet have enough data to "prove" something by rejecting the null hypothesis.

For more discussion about the meaning of a statistical hypothesis test, see Chapter 1.

Concept of null hypothesis

A classic use of a statistical test occurs in process control studies. For example, suppose that we are interested in ensuring that photomasks in a production process have mean linewidths of 500 micrometers. The null hypothesis, in this case, is that the mean linewidth is 500 micrometers. Implicit in this statement is the need to flag photomasks which have mean linewidths that are either much greater or much less than 500 micrometers. This translates into the alternative hypothesis that the mean linewidths are not equal to 500 micrometers. This is a two-sided alternative because it guards against alternatives in opposite directions; namely, that the linewidths are too small or too large.

The testing procedure works this way. Linewidths at random positions on the photomask are measured using a scanning electron microscope. A test statistic is computed from the data and tested against pre-determined upper and lower critical values. If the test statistic is greater than the upper critical value or less than the lower critical value, the null hypothesis is rejected because there is evidence that the mean linewidth is not 500 micrometers.

One-sided tests of hypothesis

Null and alternative hypotheses can also be one-sided. For example, to ensure that a lot of light bulbs has a mean lifetime of at least 500 hours, a testing program is implemented. The null hypothesis, in this case, is that the mean lifetime is greater than or equal to 500 hours. The complement or alternative hypothesis that is being guarded against is that the mean lifetime is less than 500 hours. The test statistic is compared with a lower critical value, and if it

is less than this limit, the null hypothesis is rejected.

Thus, a statistical test requires a pair of hypotheses; namely,

- H_0: a null hypothesis
- H_a: an alternative hypothesis.

Significance levels

The null hypothesis is a statement about a belief. We may doubt that the null hypothesis is true, which might be why we are "testing" it. The alternative hypothesis might, in fact, be what we believe to be true. The test procedure is constructed so that the risk of rejecting the null hypothesis, when it is in fact true, is small. This risk, α, is often referred to as the *significance level* of the test. By having a test with a small value of α, we feel that we have actually "proved" something when we reject the null hypothesis.

Errors of the second kind

The risk of failing to reject the null hypothesis when it is in fact false is not chosen by the user but is determined, as one might expect, by the magnitude of the real discrepancy. This risk, β, is usually referred to as the *error of the second kind*. Large discrepancies between reality and the null hypothesis are easier to detect and lead to small errors of the second kind; while small discrepancies are more difficult to detect and lead to large errors of the second kind. Also the risk β increases as the risk α decreases. The risks of errors of the second kind are usually summarized by an *operating characteristic curve (OC)* for the test. OC curves for several types of tests are shown in (Natrella, 1962).

Guidance in this chapter

This chapter gives methods for constructing test statistics and their corresponding critical values for both one-sided and two-sided tests for the specific situations outlined under the scope. It also provides guidance on the sample sizes required for these tests.

Further guidance on statistical hypothesis testing, significance levels and critical regions, is given in Chapter 1.

7.1.3.1. Critical values and p values

Determination of critical values

Critical values for a test of hypothesis depend upon a test statistic, which is specific to the type of test, and the significance level, α, which defines the sensitivity of the test. A value of α = 0.05 implies that the null hypothesis is rejected 5% of the time when it is in fact true. The choice of α is somewhat arbitrary, although in practice values of 0.1, 0.05, and 0.01 are common. Critical values are essentially cut-off values that define regions where the test statistic is unlikely to lie; for example, a region where the critical value is exceeded with probability α if the null hypothesis is true. The null hypothesis is rejected if the test statistic lies within this region which is often referred to as the rejection region(s). Critical values for specific tests of hypothesis are tabled in chapter 1.

Information in this chapter

This chapter gives formulas for the test statistics and points to the appropriate tables of critical values for tests of hypothesis regarding means, standard deviations, and proportion defectives.

P values

Another quantitative measure for reporting the result of a test of hypothesis is the *p*-value. The *p*-value is the probability of the test statistic being at least as extreme as the one observed given that the null hypothesis is true. A small *p*-value is an indication that the null hypothesis is false.

Good practice

It is good practice to decide in advance of the test how small a *p*-value is required to reject the test. This is exactly analagous to choosing a significance level, α for test. For example, we decide either to reject the null hypothesis if the test statistic exceeds the critical value (for α = 0.05) or analagously to reject the null hypothesis if the *p*-value is smaller than 0.05. It is important to understand the relationship between the two concepts because some statistical software packages report *p*-values rather than critical values.

7.1.4. What are confidence intervals?

How do we form a confidence interval?

The purpose of taking a random sample from a lot or population and computing a statistic, such as the mean from the data, is to approximate the mean of the population. How well the sample statistic estimates the underlying population value is always an issue. A confidence interval addresses this issue because it provides a range of values which is likely to contain the population parameter of interest.

Confidence levels

Confidence intervals are constructed at a *confidence level*, such as 95%, selected by the user. What does this mean? It means that if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95% of the cases. A confidence stated at a $1 - \alpha$ level can be thought of as the inverse of a significance level, α .

One and two-sided confidence intervals

In the same way that statistical tests can be one or two-sided, confidence intervals can be one or two-sided. A two-sided confidence interval brackets the population parameter from above and below. A one-sided confidence interval brackets the population parameter either from above or below and furnishes an upper or lower bound to its magnitude.

Example of a two-sided confidence interval

For example, a 100($1 - \alpha$)% confidence interval for the mean of a normal population is;

$$\bar{Y} \pm \frac{z_{1-\alpha/2}\sigma}{\sqrt{N}}$$

where $\bar{Y}$ is the sample mean, $z_{1-\alpha/2}$ is the 1-α/2 critical value of the standard normal distribution which is found in the table of the standard normal distribution, σ is the known population standard deviation, and N is the sample size.

Guidance in this chapter

This chapter provides methods for estimating the population parameters and confidence intervals for the situations described under the scope.

Problem with unknown standard deviation

In the normal course of events, population standard deviations are not known, and must be estimated from the data. Confidence intervals, given the same confidence level, are by necessity wider if the standard deviation is estimated from limited data because of the uncertainty in this estimate. Procedures for creating confidence intervals in this situation are described fully in this chapter.

More information on confidence intervals can also be found in Chapter 1.

7.1.5. What is the relationship between a test and a confidence interval?

There is a correspondence between hypothesis testing and confidence intervals

In general, for every test of hypothesis there is an equivalent statement about whether the hypothesized parameter value is included in a confidence interval. For example, consider the previous example of linewidths where photomasks are tested to ensure that their linewidths have a mean of 500 micrometers. The null and alternative hypotheses are:

H_0: mean linewidth = 500 micrometers

H_a: mean linewidth $\neq$ 500 micrometers

Hypothesis test for the mean

For the test, the sample mean, $\bar{Y}$, is calculated from N linewidths chosen at random positions on each photomask. For the purpose of the test, it is assumed that the standard deviation, σ, is known from a long history of this process. A test statistic is calculated from these sample statistics, and the null hypothesis is rejected if:

$$\frac{\bar{Y} - 500}{\sigma/\sqrt{N}} \leq z_{\alpha/2} \quad \text{or} \quad \frac{\bar{Y} - 500}{\sigma/\sqrt{N}} \geq z_{1-\alpha/2}$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are tabled values from the normal distribution.

Equivalent confidence interval

With some algebra, it can be seen that the null hypothesis is rejected if and only if the value 500 micrometers is not in the confidence interval

$$\bar{Y} \pm \frac{z_{1-\alpha/2}\,\sigma}{\sqrt{N}}$$

Equivalent confidence interval

In fact, all values bracketed by this interval would be accepted as null values for a given set of test data.

7.1.6. What are outliers in the data?

Definition of outliers

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Ways to describe data

Two activities are essential for characterizing a set of data:

1. Examination of the overall shape of the graphed data for important features, including symmetry and departures from assumptions. The chapter on Exploratory Data Analysis (EDA) discusses assumptions and summarization of data in detail.

2. Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. Two graphical techniques for identifying outliers, scatter plots and box plots, along with an analytic procedure for detecting outliers when the distribution is normal (Grubbs' Test), are also discussed in detail in the EDA chapter.

Box plot construction

The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q2, then the difference (Q2 - Q1) is called the interquartile range or IQ.

Box plots with fences

A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called *fences*) are needed for identifying extreme values in the tails of the distribution:

1. lower inner fence: Q1 - 1.5*IQ
2. upper inner fence: Q2 + 1.5*IQ
3. lower outer fence: Q1 - 3*IQ
4. upper outer fence: Q2 + 3*IQ

Outlier detection criteria

A point beyond an inner fence on either side is considered a **mild outlier**. A point beyond an outer fence is considered an **extreme outlier**.

Example of an outlier

The data set of $N = 90$ ordered observations as shown below is examined for outliers:

box plot

30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441

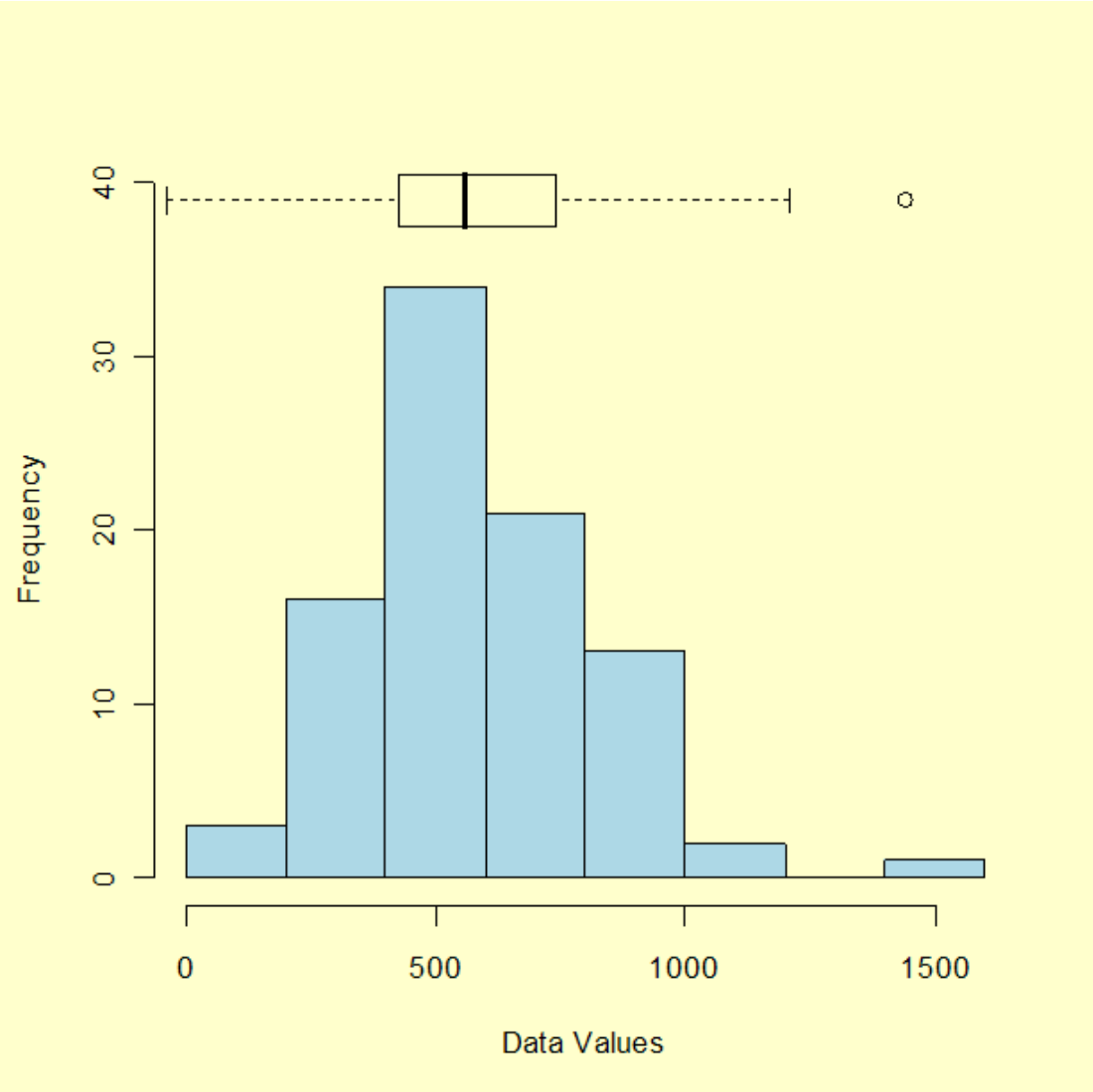
The computations are as follows:

- Median = (n+1)/2 largest data point = the average of the 45th and 46th ordered points = (559 + 560)/2 = 559.5
- Lower quartile = .25(N+1)th ordered point = 22.75th ordered point = 411 + .75(436-411) = 429.75
- Upper quartile = .75(N+1)th ordered point = 68.25th ordered point = 739 +.25(752-739) = 742.25
- Interquartile range = 742.25 - 429.75 = 312.5
- Lower inner fence = 429.75 - 1.5 (312.5) = -39.0
- Upper inner fence = 742.25 + 1.5 (312.5) = 1211.0
- Lower outer fence = 429.75 - 3.0 (312.5) = -507.75
- Upper outer fence = 742.25 + 3.0 (312.5) = 1679.75

From an examination of the fence points and the data, one point (1441) exceeds the upper inner fence and stands out as a mild outlier; there are no extreme outliers.

Histogram with box plot

A histogram with an overlaid box plot are shown below.



The outlier is identified as the largest value in the data set, 1441, and appears as the circle to the right of the box plot.

Outliers may contain important information

Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points.

7.1.7. What are trends in sequential process or product data?

Detecting trends by plotting the data points to see if a line with an obviously non-zero slope fits the points

Detecting trends is equivalent to comparing the process values to what we would expect a series of numbers to look like if there were no trends. If we see a significant departure from a model where the next observation is equally likely to go up or down, then we would reject the hypothesis of "no trend".

A common way of investigating for trends is to fit a straight line to the data and observe the line's direction (or slope). If the line looks horizontal, then there is no evidence of a trend; otherwise there is. Formally, this is done by testing whether the slope of the line is significantly different from zero. The methodology for this is covered in Chapter 4.

Other trend tests

A non-parametric approach for detecting significant trends known as the Reverse Arrangement Test is described in Chapter 8.

7.2. Comparisons based on data from one process

Questions answered in this section

For a single process, the current state of the process can be compared with a nominal or hypothesized state. This section outlines techniques for answering the following questions from data gathered from a single process:

1. Do the observations come from a particular distribution?
    1. Chi-Square Goodness-of-Fit test for a continuous or discrete distribution
    2. Kolmogorov- Smirnov test for a continuous distribution
    3. Anderson-Darling and Shapiro-Wilk tests for a continuous distribution
2. Are the data consistent with the assumed process mean?
    1. Confidence interval approach
    2. Sample sizes required
3. Are the data consistent with a nominal standard deviation?
    1. Confidence interval approach
    2. Sample sizes required
4. Does the proportion of defectives meet requirements?
    1. Confidence intervals
    2. Sample sizes required
5. Does the defect density meet requirements?
6. What intervals contain a fixed percentage of the data?
    1. Approximate intervals that contain most of the population values
    2. Percentiles
    3. Tolerance intervals
    4. Tolerance intervals based on the smallest and largest observations

General forms of testing

These questions are addressed either by an hypothesis test or by a confidence interval.

Parametric vs. non-parametric

All hypothesis-testing procedures can be broadly described as either parametric or non-parametric/distribution-free. Parametric test procedures are those that:

testing

1. Involve hypothesis testing of specified parameters (such as "the population mean=50 grams"...).
2. Require a stringent set of assumptions about the underlying sampling distributions.

When to use nonparametric methods?

When do we require non-parametric or distribution-free methods? Here are a few circumstances that may be candidates:

1. The measurements are only categorical; i.e., they are nominally scaled, or ordinally (in ranks) scaled.
2. The assumptions underlying the use of parametric methods cannot be met.
3. The situation at hand requires an investigation of such features as randomness, independence, symmetry, or goodness of fit rather than the testing of hypotheses about specific values of particular population parameters.

Difference between non-parametric and distribution-free

Some authors distinguish between non-parametric and distribution-free procedures.

Distribution-free test procedures are broadly defined as:

1. Those whose test statistic does not depend on the form of the underlying population distribution from which the sample data were drawn, or
2. Those for which the data are nominally or ordinally scaled.

Nonparametric test procedures are defined as those that are not concerned with the parameters of a distribution.

Advantages of nonparametric methods.

Distribution-free or nonparametric methods have several advantages, or benefits:

1. They may be used on all types of data-categorical data, which are nominally scaled or are in rank form, called ordinally scaled, as well as interval or ratio-scaled data.
2. For small sample sizes they are easy to apply.
3. They make fewer and less stringent assumptions than their parametric counterparts.
4. Depending on the particular procedure they may be *almost* as powerful as the corresponding parametric procedure when the assumptions of the latter are

met, and when this is not the case, they are generally more powerful.

Disadvantages of nonparametric methods

Of course there are also disadvantages:

1. If the assumptions of the parametric methods can be met, it is generally more efficient to use them.

2. For large sample sizes, data manipulations tend to become more laborious, unless computer software is available.

3. Often special tables of critical values are needed for the test statistic, and these values cannot always be generated by computer software. On the other hand, the critical values for the parametric tests are readily available and generally easy to incorporate in computer programs.

7.2.1. Do the observations come from a particular distribution?

Data are often assumed to come from a particular distribution.

Goodness-of-fit tests indicate whether or not it is reasonable to assume that a random sample comes from a specific distribution. Statistical techniques often rely on observations having come from a population that has a distribution of a specific form (e.g., normal, lognormal, Poisson, etc.). Standard control charts for continuous measurements, for instance, require that the data come from a normal distribution. Accurate lifetime modeling requires specifying the correct distributional model. There may be historical or theoretical reasons to assume that a sample comes from a particular population, as well. Past data may have consistently fit a known distribution, for example, or theory may predict that the underlying population should be of a specific form.

Hypothesis Test model for Goodness-of-fit

Goodness-of-fit tests are a form of hypothesis testing where the null and alternative hypotheses are

H_0: Sample data come from the stated distribution.
H_A: Sample data **do not** come from the stated distribution.

Parameters may be assumed or estimated from the data

One needs to consider whether a simple or composite hypothesis is being tested. For a simple hypothesis, values of the distribution's parameters are specified prior to drawing the sample. For a composite hypothesis, one or more of the parameters is unknown. Often, these parameters are estimated using the sample observations.

A simple hypothesis would be:

H_0: Data are from a normal distribution, $\mu = 0$ and $\sigma = 1$.

A composite hypothesis would be:

H_0: Data are from a normal distribution, unknown μ and σ.

Composite hypotheses are more common because they allow us to decide whether a sample comes from any distribution of a specific type. In this situation, the form of the distribution is of interest, regardless of the values of the parameters. Unfortunately, composite hypotheses are more difficult to

work with because the critical values are often hard to compute.

Problems with censored data

A second issue that affects a test is whether the data are censored. When data are censored, sample values are in some way restricted. Censoring occurs if the range of potential values are limited such that values from one or both tails of the distribution are unavailable (e.g., right and/or left censoring - where high and/or low values are missing). Censoring frequently occurs in reliability testing, when either the testing time or the number of failures to be observed is fixed in advance. A thorough treatment of goodness-of-fit testing under censoring is beyond the scope of this document. See D'Agostino & Stephens (1986) for more details.

Three types of tests will be covered

Three goodness-of-fit tests are examined in detail:

1. Chi-square test for continuous and discrete distributions;
2. Kolmogorov-Smirnov test for continuous distributions based on the empirical distribution function (EDF);
3. Anderson-Darling test for continuous distributions.

A more extensive treatment of goodness-of-fit techniques is presented in D'Agostino & Stephens (1986). Along with the tests mentioned above, other general and specific tests are examined, including tests based on regression and graphical techniques.

7.2.1.1. Chi-square goodness-of-fit test

Choice of number of groups for "Goodness of Fit" tests is important - but only useful rules of thumb can be given	The test requires that the data first be grouped. The actual number of observations in each group is compared to the expected number of observations and the test statistic is calculated as a function of this difference. The number of groups and how group membership is defined will affect the power of the test (i.e., how sensitive it is to detecting departures from the null hypothesis). Power will not only be affected by the number of groups and how they are defined, but by the sample size and shape of the null and underlying (true) distributions. Despite the lack of a clear "best method", some useful rules of thumb can be given.
Group Membership	When data are discrete, group membership is unambiguous. Tabulation or cross tabulation can be used to categorize the data. Continuous data present a more difficult challenge. One defines groups by segmenting the range of possible values into non-overlapping intervals. Group membership can then be defined by the endpoints of the intervals. In general, power is maximized by choosing endpoints such that group membership is equiprobable (i.e., the probabilities associated with an observation falling into a given group are divided as evenly as possible across the intervals). Many commercial software packages follow this procedure.
Rule-of-thumb for number of groups	One rule-of-thumb suggests using the value $2n^{2/5}$ as a good starting point for choosing the number of groups. Another well known rule-of-thumb requires every group to have at least 5 data points.
Computation of the chi-square goodness-of-fit test	The formulas for the computation of the chi-square goodnes-of-fit test are given in the EDA chapter.

7.2.1.2. Kolmogorov- Smirnov test

The K-S test is a good alternative to the chi-square test.

The Kolmogorov-Smirnov (K-S) test was originally proposed in the 1930's in papers by Kolmogorov (1933) and Smirnov (1936). Unlike the Chi-Square test, which can be used for testing against both continuous and discrete distributions, the K-S test is only appropriate for testing data against a continuous distribution, such as the normal or Weibull distribution. It is one of a number of tests that are based on the empirical cumulative distribution function (ECDF).

K-S procedure

Details on the construction and interpretation of the K-S test statistic, *D*, and examples for several distributions are outlined in Chapter 1.

The probability associated with the test statistic is difficult to compute.

Critical values associated with the test statistic, *D*, are difficult to compute for finite sample sizes, often requiring Monte Carlo simulation. However, some general purpose statistical software programs support the Kolmogorov-Smirnov test at least for some of the more common distributions. Tabled values can be found in Birnbaum (1952). A correction factor can be applied if the parameters of the distribution are estimated with the same data that are being tested. See D'Agostino and Stephens (1986) for details.

7.2.1.3. Anderson-Darling and Shapiro-Wilk tests

Purpose: Test for distributional adequacy

The Anderson-Darling Test

The Anderson-Darling test (Stephens, 1974) is used to test if a sample of data comes from a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails of the distribution than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested.

Requires critical values for each distribution

The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. Tables of critical values are not given in this handbook (see Stephens 1974, 1976, 1977, and 1979) because this test is usually applied with a statistical software program that produces the relevant critical values. Currently, Dataplot computes critical values for the Anderson-Darling test for the following distributions:

- normal
- lognormal
- Weibull
- extreme value type I.

Anderson-Darling procedure

Details on the construction and interpretation of the Anderson-Darling test statistic, A^2, and examples for several distributions are outlined in Chapter 1.

Shapiro-Wilk test for normality

The Shapiro-Wilk Test For Normality

The Shapiro-Wilk test, proposed in 1965, calculates a W statistic that tests whether a random sample, $x_1, x_2, ..., x_n$ comes from (specifically) a normal distribution . Small values of W are evidence of departure from normality and percentage points for the W statistic, obtained via Monte Carlo simulations, were reproduced by Pearson and Hartley (1972, Table 16). This test has done very well in

comparison studies with other goodness of fit tests.

The *W* statistic is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

where the $x_{(i)}$ are the ordered sample values ($x_{(1)}$ is the smallest) and the a_i are constants generated from the means, variances and covariances of the order statistics of a sample of size *n* from a normal distribution (see Pearson and Hartley (1972, Table 15)).

For more information about the Shapiro-Wilk test the reader is referred to the original Shapiro and Wilk (1965) paper and the tables in Pearson and Hartley (1972).

7.2.2. Are the data consistent with the assumed process mean?

The testing of H_0 for a single population mean

Given a random sample of measurements, Y_1, ..., Y_N, there are three types of questions regarding the true mean of the population that can be addressed with the sample data. They are:

1. Does the true mean agree with a known standard or assumed mean?
2. Is the true mean of the population less than a given standard?
3. Is the true mean of the population at least as large as a given standard?

Typical null hypotheses

The corresponding null hypotheses that test the true mean, μ, against the standard or assumed mean, μ_0 are:

1. $H0 : \mu = \mu0$
2. $H0 : \mu \leq \mu0$
3. $H0 : \mu \geq \mu0$

Test statistic where the standard deviation is not known

The basic statistics for the test are the sample mean and the standard deviation. The form of the test statistic depends on whether the poulation standard deviation, σ, is known or is estimated from the data at hand. The more typical case is where the standard deviation must be estimated from the data, and the test statistic is

$$t = \frac{\bar{Y} - \mu_0}{s / \sqrt{N}}$$

where the sample mean is

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

and the sample standard deviation is

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

with N - 1 degrees of freedom.

Comparison with critical values

For a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10, the hypothesis associated with each case enumerated above is rejected if:

1. $|t| \geq t_{1-\alpha/2,\, N-1}$
2. $t \geq t_{1-\alpha,\, N-1}$
3. $t \leq t_{\alpha,\, N-1}$

where $t_{1-\alpha/2,\, N-1}$ is the 1-α/2 critical value from the t distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the *t*-table in Chapter 1.

Test statistic where the standard deviation is known

If the standard deviation is known, the form of the test statistic is

$$z = \frac{\bar{Y} - \mu_0}{\sigma / \sqrt{N}}$$

For case (1), the test statistic is compared with $z_{1-\alpha/2}$, which is the 1-α/2 critical value from the standard normal distribution, and similarly for cases (2) and (3).

Caution

If the standard deviation is assumed known for the purpose of this test, this assumption should be checked by a test of hypothesis for the standard deviation.

An illustrative example of the t-test

The following numbers are particle (contamination) counts for a sample of 10 semiconductor silicon wafers:

```
50   48   44   56   61   52   53   55   67   51
```

The mean = 53.7 counts and the standard deviation = 6.567 counts.

The test is two-sided

Over a long run the process average for wafer particle counts has been 50 counts per wafer, and on the basis of the sample, we want to test whether a change has occurred. The null hypothesis that the process mean is 50 counts is tested against the alternative hypothesis that the process mean is not equal to 50 counts. The purpose of the two-sided alternative is to rule out a possible process change in either direction.

Critical values

For a significance level of α = 0.05, the chances of erroneously rejecting the null hypothesis when it is true are 5

% or less. (For a review of hypothesis testing basics, see Chapter 1).

Even though there is a history on this process, it has not been stable enough to justify the assumption that the standard deviation is known. Therefore, the appropriate test statistic is the *t*-statistic. Substituting the sample mean, sample standard deviation, and sample size into the formula for the test statistic gives a value of

$$t = 1.782$$

with degrees of freedom $N - 1 = 9$. This value is tested against the critical value

$$t_{1-0.025;9} = 2.262$$

from the *t*-table where the critical value is found under the column labeled 0.975 for the probability of exceeding the critical value and in the row for 9 degrees of freedom. The critical value is based on $\alpha/2$ instead of α because of the two-sided alternative (two-tailed test) which requires equal probabilities in each tail of the distribution that add to α.

Conclusion Because the value of the test statistic falls in the interval (-2.262, 2.262), we cannot reject the null hypothesis and, therefore, we may continue to assume the process mean is 50 counts.

7.2.2.1. Confidence interval approach

Testing using a confidence interval

The hypothesis test results in a "yes" or "no" answer. The null hypothesis is either rejected or not rejected. There is another way of testing a mean and that is by constructing a confidence interval about the true but unknown mean.

General form of confidence intervals where the standard deviation is unknown

Tests of hypotheses that can be made from a single sample of data were discussed on the foregoing page. As with null hypotheses, confidence intervals can be two-sided or one-sided, depending on the question at hand. The general form of confidence intervals, for the three cases discussed earlier, where the standard deviation is unknown are:

1. Two-sided confidence interval for μ:

$$\bar{Y} + \frac{s}{\sqrt{N}} t_{\alpha/2,N-1} \leq \mu \leq \bar{Y} + \frac{s}{\sqrt{N}} t_{1-\alpha/2,N-1}$$

2. Lower one-sided confidence interval for μ:

$$\mu \geq \bar{Y} + \frac{s}{\sqrt{N}} t_{\alpha,N-1}$$

3. Upper one-sided confidence interval for μ:

$$\mu \leq \bar{Y} + \frac{s}{\sqrt{N}} t_{1-\alpha,N-1}$$

where $t_{\alpha/2,\, N-1}$ is the $\alpha/2$ critical value from the t distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the t table in Chapter 1.

Confidence level

The confidence intervals are constructed so that the probability of the interval containing the mean is 1 - α. Such intervals are referred to as 100(1-α)% confidence intervals.

A 95% confidence interval for

The corresponding confidence interval for the test of hypothesis example on the foregoing page is shown below. A 95 % confidence interval for the population mean of particle counts per wafer is given by

the example

$$\bar{Y} + \frac{s}{\sqrt{N}}(t_{0.025,9}) \leq \mu \leq \bar{Y} + \frac{s}{\sqrt{N}}(t_{0.975,9})$$

$$53.7 + \frac{6.567}{\sqrt{10}}(-2.262) \leq \mu \leq 53.7 + \frac{6.567}{\sqrt{10}}(2.262)$$

$$49.0 \leq \mu \leq 58.4$$

Interpretation The 95 % confidence interval includes the null hypothesis if, and only if, it would be accepted at the 5 % level. This interval includes the null hypothesis of 50 counts so we cannot reject the hypothesis that the process mean for particle counts is 50. The confidence interval includes all null hypothesis values for the population mean that would be accepted by an hypothesis test at the 5 % significance level. This assumes, of course, a two-sided alternative.

7.2.2.2. Sample sizes required

The computation of sample sizes depends on many things, some of which have to be assumed in advance

Perhaps one of the most frequent questions asked of a statistician is,

> "How many measurements should be included in the sample? "

Unfortunately, there is no correct answer without additional information (or assumptions). The sample size required for an experiment designed to investigate the behavior of an unknown population mean will be influenced by the following:

- value selected for α, the risk of rejecting a true hypothesis
- value of β, the risk of accepting a false null hypothesis when a particular value of the alternative hypothesis is true.
- value of the population standard deviation.

Application - estimating a minimum sample size, N, for limiting the error in the estimate of the mean

For example, suppose that we wish to estimate the average daily yield, μ, of a chemical process by the mean of a sample, Y_1, ..., Y_N, such that the error of estimation is less than δ with a probability of 95%. This means that a 95% confidence interval centered at the sample mean should be

$$\bar{Y} - \delta \le \mu \le \bar{Y} + \delta$$

and if the standard deviation is known,

$$\delta = \frac{\sigma}{\sqrt{N}} z_{1-0.025}$$

The critical value from the normal distribution for 1-α/2 = 0.975 is 1.96. Therefore,

$$N \ge \left(\frac{1.96}{\delta}\right)^2 \sigma^2$$

Limitation and interpretation

A restriction is that the standard deviation must be known. Lacking an exact value for the standard deviation requires some accommodation, perhaps the best estimate available from a previous experiment.

Controlling the risk of accepting a false hypothesis

To control the risk of accepting a false hypothesis, we set not only α, the probability of rejecting the null hypothesis when it is true, but also β, the probability of accepting the null hypothesis when in fact the population mean is $\mu+\delta$ where δ is the difference or shift we want to detect.

Standard deviation assumed to be known

The minimum sample size, N, is shown below for two- and one-sided tests of hypotheses with σ assumed to be known.

$$N = (z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{\sigma}{\delta}\right)^2 \rightarrow two-sided\ test$$

$$N = (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma}{\delta}\right)^2 \rightarrow one-sided\ test$$

The quantities $z_{1-\alpha/2}$ and $z_{1-\beta}$ are critical values from the normal distribution.

Note that it is usual to state the shift, δ, in units of the standard deviation, thereby simplifying the calculation.

Example where the shift is stated in terms of the standard deviation

For a one-sided hypothesis test where we wish to detect an increase in the population mean of one standard deviation, the following information is required: α, the significance level of the test, and β, the probability of failing to detect a shift of one standard deviation. For a test with α = 0.05 and β = 0.10, the minimum sample size required for the test is

$$N = (1.645 + 1.282)^2 = 8.567 \sim 9.$$

More often we must compute the sample size with the population standard deviation being unknown

The procedures for computing sample sizes when the standard deviation is not known are similar to, but more complex, than when the standard deviation is known. The formulation depends on the *t* distribution where the minimum sample size is given by

$$N = (t_{1-\alpha/2} + t_{1-\beta})^2 \left(\frac{s}{\delta}\right)^2 \rightarrow two-sided\ test$$

$$N = (t_{1-\alpha} + t_{1-\beta})^2 \left(\frac{s}{\delta}\right)^2 \rightarrow one-sided\ test$$

The drawback is that critical values of the *t* distribution depend on known degrees of freedom, which in turn depend upon the sample size which we are trying to estimate.

Iterate on the initial estimate using critical values from

Therefore, the best procedure is to start with an intial estimate based on a sample standard deviation and iterate. Take the example discussed above where the the minimum sample size is computed to be N = 9. This estimate is low. Now use the formula above with degrees of freedom N - 1 = 8 which gives a second estimate of

the t table

$$N = (1.860 + 1.397)^2 = 10.6 \sim 11.$$

It is possible to apply another iteration using degrees of freedom 10, but in practice one iteration is usually sufficient. For the purpose of this example, results have been rounded to the closest integer; however, computer programs for finding critical values from the *t* distribution allow non-integer degrees of freedom.

Table showing minimum sample sizes for a two-sided test

The table below gives sample sizes for a two-sided test of hypothesis that the mean is a given value, with the shift to be detected a multiple of the standard deviation. For a one-sided test at significance level α, look under the value of 2α in column 1. Note that this table is based on the normal approximation (i.e., the standard deviation is known).

Sample Size Table for Two-Sided Tests

α	β	$\delta = .5\sigma$	$\delta = 1.0\sigma$	$\delta = 1.5\sigma$
.01	.01	98	25	11
.01	.05	73	18	8
.01	.10	61	15	7
.01	.20	47	12	6
.01	.50	27	7	3
.05	.01	75	19	9
.05	.05	53	13	6
.05	.10	43	11	5
.05	.20	33	8	4
.05	.50	16	4	3
.10	.01	65	16	8
.10	.05	45	11	5
.10	.10	35	9	4
.10	.20	25	7	3
.10	.50	11	3	3
.20	.01	53	14	6
.20	.05	35	9	4
.20	.10	27	7	3
.20	.20	19	5	3
.20	.50	7	3	3

7.2.3. Are the data consistent with a nominal standard deviation?

The testing of H_0 for a single population mean

Given a random sample of measurements, Y_1, ..., Y_N, there are three types of questions regarding the true standard deviation of the population that can be addressed with the sample data. They are:

1. Does the true standard deviation agree with a nominal value?
2. Is the true standard deviation of the population less than or equal to a nominal value?
3. Is the true stanard deviation of the population at least as large as a nominal value?

Corresponding null hypotheses

The corresponding null hypotheses that test the true standard deviation, σ, against the nominal value, σ_0 are:

1. H_0: $\sigma = \sigma_0$
2. H_0: $\sigma <= \sigma_0$
3. H_0: $\sigma >= \sigma_0$

Test statistic

The basic test statistic is the chi-square statistic

$$\chi^2 = \frac{(N-1)s^2}{\sigma_0^2}$$

with N - 1 degrees of freedom where s is the sample standard deviation; i.e.,

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

.

Comparison with critical values

For a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10, the hypothesis associated with each case enumerated above is rejected if:

1. $\chi^2 \geq \chi^2_{1-\alpha/2}$ or $\chi^2 \leq \chi^2_{\alpha/2}$
2. $\chi^2 \geq \chi^2_{1-\alpha}$
3. $\chi^2 \leq \chi^2_{\alpha}$

where $X^2_{\alpha/2}$ is the $\alpha/2$ critical value from the chi-square distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the chi-square table in Chapter 1.

Warning

Because the chi-square distribution is a non-negative, asymmetrical distribution, care must be taken in looking up critical values from tables. For two-sided tests, critical values are required for both tails of the distribution.

Example

A supplier of 100 ohm·cm silicon wafers claims that his fabrication process can produce wafers with sufficient consistency so that the standard deviation of resistivity for the lot does not exceed 10 ohm·cm. A sample of N = 10 wafers taken from the lot has a standard deviation of 13.97 ohm.cm. Is the suppliers claim reasonable? This question falls under null hypothesis (2) above. For a test at significance level, α = 0.05, the test statistic,

$$\chi^2 = \frac{(N-1)s^2}{\sigma_0^2} = \frac{9(13.97)^2}{100} = 17.56$$

is compared with the critical value, $X^2_{0.95,\,9}$ = 16.92.

Since the test statistic (17.56) exceeds the critical value (16.92) of the chi-square distribution with 9 degrees of freedom, the manufacturer's claim is rejected.

7.2.3.1. Confidence interval approach

Confidence intervals for the standard deviation

Confidence intervals for the true standard deviation can be constructed using the chi-square distribution. The 100(1- α)% confidence intervals that correspond to the tests of hypothesis on the previous page are given by

1. Two-sided confidence interval for σ

$$\frac{s\sqrt{N-1}}{\sqrt{\chi^2_{1-\alpha/2,N-1}}} \leq \sigma \leq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{\alpha/2,N-1}}}$$

2. Lower one-sided confidence interval for σ

$$\sigma \geq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{1-\alpha,N-1}}}$$

3. Upper one-sided confidence interval for σ

$$0 \leq \sigma \leq \frac{s\sqrt{N-1}}{\sqrt{\chi^2_{\alpha,N-1}}}$$

where for case (1), $X^2_{\alpha/2}$ is the $\alpha/2$ critical value from the chi-square distribution with N - 1 degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the chi-square table in Chapter 1.

Choice of risk level α can change the conclusion

Confidence interval (1) is equivalent to a two-sided test for the standard deviation. That is, if the hypothesized or nominal value, σ_0, is not contained within these limits, then the hypothesis that the standard deviation is equal to the nominal value is rejected.

A dilemma of hypothesis testing

A change in α can lead to a change in the conclusion. This poses a dilemma. What should α be? Unfortunately, there is no clear-cut answer that will work in all situations. The usual strategy is to set α small so as to guarantee that the null hypothesis is *wrongly* rejected in only a small number of

cases. The risk, β, of failing to reject the null hypothesis when it is false depends on the size of the discrepancy, and also depends on α. The discussion on the next page shows how to choose the sample size so that this risk is kept small for specific discrepancies.

7.2.3.2. Sample sizes required

Sample sizes to minimize risk of false acceptance

The following procedure for computing sample sizes for tests involving standard deviations follows W. Diamond (1989). The idea is to find a sample size that is large enough to guarantee that the risk, β, of accepting a false hypothesis is small.

Alternatives are specific departures from the null hypothesis

This procedure is stated in terms of changes in the variance, not the standard deviation, which makes it somewhat difficult to interpret. Tests that are generally of interest are stated in terms of δ, a discrepancy from the hypothesized variance. For example:

1. Is the true variance larger than its hypothesized value by δ ?
2. Is the true variance smaller than its hypothesized value by δ ?

That is, the tests of interest are:

1. $H_0: \sigma^2 \geq \sigma_0^2 + \delta; \quad \delta \geq 0$
2. $H_0: \sigma^2 \leq \sigma_0^2 - \delta; \quad \delta \geq 0$

Interpretation

The experimenter wants to assure that the probability of erroneously accepting the null hypothesis of unchanged variance is at most β. The sample size, N, required for this type of detection depends on the factor, δ; the significance level, α; and the risk, β.

First choose the level of significance and beta risk

The sample size is determined by first choosing appropriate values of α and β and then following the directions below to find the degrees of freedom, ν, from the chi-square distribution.

The calculations should be done by creating a table or

First compute

$$R = 1 + \frac{\delta}{\sigma_0^2}$$

Then generate a table of degrees of freedom, ν, say

spreadsheet

between 1 and 200. For case (1) or (2) above, calculate β_ν and the corresponding value of C_ν for each value of degrees of freedom in the table where

$$\begin{aligned} &1. \quad \beta_\nu = \chi^2_{1-\alpha,\nu}/R \\ &\qquad C_\nu = \Pr(\chi^2_\nu < \beta_\nu) \\ &2. \quad \beta_\nu = \chi^2_{\alpha,\nu}/R \\ &\qquad C_\nu = \Pr(\chi^2_\nu > \beta_\nu) \end{aligned}$$

The value of ν where C_ν is closest to β is the correct degrees of freedom and

$$N = \nu + 1$$

Hints on using software packages to do the calculations

The quantity $X^2_{1-\alpha,\,\nu}$ is the critical value from the chi-square distribution with ν degrees of freedom which is exceeded with probability α. It is sometimes referred to as the percent point function (PPF) or the inverse chi-square function. The probability that is evaluated to get C_ν is called the cumulative density function (CDF).

Example

Consider the case where the variance for resistivity measurements on a lot of silicon wafers is claimed to be 100 $(\text{ohm}\cdot\text{cm})^2$. A buyer is unwilling to accept a shipment if δ is greater than 55 ohm·cm for a particular lot. This problem falls under case (1) above. How many samples are needed to assure risks of $\alpha = 0.05$ and $\beta = 0.01$?

Calculations

If software is available to compute the roots (or zero values) of a univariate function, then we can determine the sample size by finding the roots of a function that calculates C_ν for a given value of ν. The procedure is:

1. Define constants.
   $\alpha = 0.05$
   $\beta = 0.01$
   $\delta = 55$
   $\sigma_0^2 = 100$
   $R = 1 + \delta/\sigma_0^2$

2. Create a function, Cnu.
   $Cnu = F(\ F^{-1}(\alpha,\ \nu)/R,\ \nu\ ) - \beta$

   $F(x,\ \nu)$ returns the probability of a chi-square random variable with ν degrees of freedom that is less than or equal to x and $F^{-1}(\alpha,\ \nu)$ returns x such that $F(x,\ \nu) = \alpha$.

3. Find the value of ν for which the function, Cnu, is zero.

Using this procedure, Cnu is zero when ν is 169.3. Therefore, the minimum sample size needed to guarantee the risk level is N = 170.

Alternatively, we can determine the sample size by simply printing computed values of Cnu for various values of ν.

1. Define constants.
   α = 0.05
   δ = 55
   σ_0^2 = 100
   R = 1 + δ/σ_0^2

2. Generate Cnu for values of ν from 1 to 200.
   Bnu = $F^{-1}(\alpha, \nu)$ / R
   Cnu = F(Bnu, ν)

The values of Cnu generated for ν between 165 and 175 degrees of freedom are shown below.

ν	Bnu	Cnu
165	126.4344	0.0114
166	127.1380	0.0110
167	127.8414	0.0107
168	128.5446	0.0104
169	129.2477	0.0101
170	129.9506	0.0098
171	130.6533	0.0095
172	131.3558	0.0092
173	132.0582	0.0090
174	132.7604	0.0087
175	133.4625	0.0085

The value of Cnu closest to 0.01 is 0.0101, which is associated with ν = 169 degrees of freedom. Therefore, the minimum sample size needed to guarantee the risk level is N = 170.

The calculations used in this section can be performed using both Dataplot code and R code.

7.2.4. Does the proportion of defectives meet requirements?

Testing proportion defective is based on the binomial distribution

The proportion of defective items in a manufacturing process can be monitored using statistics based on the observed number of defectives in a random sample of size N from a continuous manufacturing process, or from a large population or lot. The proportion defective in a sample follows the binomial distribution where p is the probability of an individual item being found defective. Questions of interest for quality control are:

1. Is the proportion of defective items within prescribed limits?
2. Is the proportion of defective items less than a prescribed limit?
3. Is the proportion of defective items greater than a prescribed limit?

Hypotheses regarding proportion defective

The corresponding hypotheses that can be tested are:

1. $p \neq p_0$
2. $p \leq p_0$
3. $p \geq p_0$

where p_0 is the prescribed proportion defective.

Test statistic based on a normal approximation

Given a random sample of measurements Y_1, ..., Y_N from a population, the proportion of items that are judged defective from these N measurements is denoted $\hat{p}$. The test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$$

depends on a normal approximation to the binomial distribution that is valid for large N, ($N > 30$). This approximation simplifies the calculations using critical values from the table of the normal distribution as shown below.

Restriction on sample size

Because the test is approximate, N needs to be large for the test to be valid. One criterion is that N should be chosen so that

$$\min\{Np_0, N(1 - p_0)\} \geq 5$$

For example, if $p_0 = 0.1$, then N should be at least 50 and if $p_0 = 0.01$, then N should be at least 500. Criteria for choosing a sample size in order to guarantee detecting a change of size δ are discussed on another page.

One and two-sided tests for proportion defective

Tests at the 1 - α confidence level corresponding to hypotheses (1), (2), and (3) are shown below. For hypothesis (1), the test statistic, z, is compared with $z_{1-\alpha/2}$, the critical value from the normal distribution that is exceeded with probability $\alpha/2$ and similarly for (2) and (3). If

1. $|z| \geq z_{1-\alpha/2}$
2. $z \leq z_{\alpha}$
3. $z \geq z_{1-\alpha}$

the null hypothesis is rejected.

Example of a one-sided test for proportion defective

After a new method of processing wafers was introduced into a fabrication process, two hundred wafers were tested, and twenty-six showed some type of defect. Thus, for N= 200, the proportion defective is estimated to be $\hat{p}$ = 26/200 = 0.13. In the past, the fabrication process was capable of producing wafers with a proportion defective of at most 0.10. The issue is whether the new process has degraded the quality of the wafers. The relevant test is the one-sided test (3) which guards against an increase in proportion defective from its historical level.

Calculations for a one-sided test of proportion defective

For a test at significance level α = 0.05, the hypothesis of no degradation is validated if the test statistic z is less than the critical value, $z_{0.95}$ = 1.645. The test statistic is computed to be

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}} = \frac{0.13 - 0.10}{\sqrt{\frac{0.10(0.90)}{200}}} = 1.414$$

Interpretation

Because the test statistic is less than the critical value (1.645), we cannot reject hypothesis (3) and, therefore, we cannot conclude that the new fabrication method is degrading the quality of the wafers. The new process may, indeed, be worse, but more evidence would be needed to reach that conclusion at the 95% confidence level.

7.2.4.1. Confidence intervals

Confidence intervals using the method of Agresti and Coull

The method recommended by Agresti and Coull (1998) and also by Brown, Cai and DasGupta (2001) (the methodology was originally developed by Wilson in 1927) is to use the form of the confidence interval that corresponds to the hypothesis test given in Section 7.2.4. That is, solve for the two values of p_0 (say, p_{upper} and p_{lower}) that result from setting $z = z_{1-\alpha/2}$ and solving for $p_0 = p_{upper}$, and then setting $z = z_{\alpha/2}$ and solving for $p_0 = p_{lower}$. (Here, as in Section 7.2.4, $z_{\alpha/2}$ denotes the variate value from the standard normal distribution such that the area to the left of the value is $\alpha/2$.) Although solving for the two values of p_0 might sound complicated, the appropriate expressions can be obtained by straightforward but slightly tedious algebra. Such algebraic manipulation isn't necessary, however, as the appropriate expressions are given in various sources. Specifically, we have

Formulas for the confidence intervals

$$\text{U.L.} = \frac{\hat{p} + \frac{z^2_{1-\alpha/2}}{2n} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{1-\alpha/2}}{4n^2}}}{1 + \frac{z^2_{1-\alpha/2}}{n}}$$

$$\text{L.L.} = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + \frac{z^2_{\alpha/2}}{n}}$$

Procedure does not strongly depend on values of p and n

This approach can be substantiated on the grounds that it is the exact algebraic counterpart to the (large-sample) hypothesis test given in section 7.2.4 and is also supported by the research of Agresti and Coull. One advantage of this procedure is that its worth does not strongly depend upon the value of *n* and/or *p*, and indeed was recommended by Agresti and Coull for virtually all combinations of *n* and *p*.

Another advantage is that the lower limit cannot be negative

Another advantage is that the lower limit cannot be negative. That is not true for the confidence expression most frequently used:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

A confidence limit approach that produces a lower limit which is an impossible value for the parameter for which the interval is constructed is an inferior approach. This also applies to limits for the control charts that are discussed in Chapter 6.

One-sided confidence intervals

A one-sided confidence interval can also be constructed simply by replacing each $z_{\alpha/2}$ by z_{α} in the expression for the lower or upper limit, whichever is desired. The 95% one-sided interval for *p* for the example in the preceding section is:

Example

$$p \geq \text{lower limit}$$

$$p \geq \frac{\hat{p} + \frac{z_\alpha^2}{2n} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_\alpha^2}{4n^2}}}{1 + \frac{z_\alpha^2}{n}}$$

$$p \geq \frac{0.013 + \frac{(-1.645)^2}{2(200)} - 1.645\sqrt{\frac{0.013(1-0.013)}{200} + \frac{(-1.645)^2}{4(200)^2}}}{1 + \frac{(-1.645)^2}{200}}$$

$$p \geq 0.09577$$

Conclusion from the example

Since the lower bound does not exceed 0.10, in which case it would exceed the hypothesized value, the null hypothesis that the proportion defective is at most 0.10, which was given in the preceding section, would not be rejected if we used the confidence interval to test the hypothesis. Of course a confidence interval has value in its own right and does not have to be used for hypothesis testing.

Exact Intervals for Small Numbers of Failures and/or Small Sample Sizes

Construction of exact two-sided confidence intervals based on the binomial distribution

If the number of failures is very small or if the sample size *N* is very small, symmetical confidence limits that are approximated using the normal distribution may not be accurate enough for some applications. An *exact method* based on the binomial distribution is shown next. To construct a two-sided confidence interval at the 100(1-α)% confidence level for the true proportion defective *p* where N_d defects are found in a sample of size *N* follow the steps below.

1. Solve the equation

$$\sum_{k=0}^{N_d}\binom{N}{k}p_U^k(1-p_U)^{N-k} = \alpha/2$$

   for p_U to obtain the upper 100(1-α)% limit for *p*.

2. Next solve the equation

$$\sum_{k=0}^{N_d-1} \binom{N}{k} p_L^k (1-p_L)^{N-k} = 1-\alpha/2$$

for p_L to obtain the lower 100(1-α)% limit for p.

Note

The interval (p_L, p_U) is an exact 100(1-α)% confidence interval for p. However, it is not symmetric about the observed proportion defective, $\hat{p}=N_d/N$.

Binomial confidence interval example

The equations above that determine p_L and p_U can be solved using readily available functions. Take as an example the situation where twenty units are sampled from a continuous production line and four items are found to be defective. The proportion defective is estimated to be $\hat{p}$ = 4/20 = 0.20. The steps for calculating a 90 % confidence interval for the true proportion defective, p follow.

1. Initalize constants.
   ```
   alpha = 0.10
   Nd = 4
   N = 20
   ```
2. Define a function for upper limit (fu) and a function for the lower limit (fl).
   ```
   fu = F(Nd,pu,20) - alpha/2
   fl = F(Nd-1,pl,20) - (1-alpha/2)
   ```
   F is the cumulative density function for the binominal distribution.
3. Find the value of pu that corresponds to fu = 0 and the value of pl that corresponds to fl = 0 using software to find the roots of a function.

The values of pu and pl for our example are:

```
pu = 0.401029
pl = 0.071354
```

Thus, a 90 % confidence interval for the proportion defective, p, is (0.071, 0.400). Whether or not the interval is truly "exact" depends on the software.

The calculations used in this example can be performed using both Dataplot code and R code.

7.2.4.2. Sample sizes required

Derivation of formula for required sample size when testing proportions

The method of determining sample sizes for testing proportions is similar to the method for determining sample sizes for testing the mean. Although the sampling distribution for proportions actually follows a binomial distribution, the normal approximation is used for this derivation.

Minimum sample size

If we are interested in detecting a change in the proportion defective of size δ in either direction, the minimum sample size is

1. For a two-sided test

$$N \geq \frac{p(1-p)}{\delta^2} z^2_{1-\alpha/2}$$

2. For a one-sided test

$$N \geq \frac{p(1-p)}{\delta^2} z^2_{1-\alpha}$$

Interpretation and sample size for high probability of detecting a change

This requirement on the sample size only guarantees that a change of size δ is detected with 50% probability. The derivation of the sample size when we are interested in protecting against a change δ with probability 1 -β (where β is small) is

1. For a two-sided test

$$N \geq (z_{1-\alpha/2} + z_{1-\beta})^2 \left(\frac{p(1-p)}{\delta^2}\right)$$

2. For a one-sided test

$$N \geq (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{p(1-p)}{\delta^2}\right)$$

where $z_{1-\beta}$ is the critical value from the normal distribution that is

exceeded with probability β.

Value for the true proportion defective

The equations above require that p be known. Usually, this is not the case. If we are interested in detecting a change relative to an historical or hypothesized value, this value is taken as the value of p for this purpose. Note that taking the value of the proportion defective to be 0.5 leads to the largest possible sample size.

Example of calculating sample size for testing proportion defective

Suppose that a department manager needs to be able to detect any change above 0.10 in the current proportion defective of his product line, which is running at approximately 10% defective. He is interested in a one-sided test and does not want to stop the line except when the process has clearly degraded and, therefore, he chooses a significance level for the test of 5%. Suppose, also, that he is willing to take a risk of 10% of failing to detect a change of this magnitude. With these criteria:

1. $z_{0.95}$ = 1.645; $z_{0.90}$=1.282
2. δ = 0.10
3. p = 0.10

and the minimum sample size for a one-sided test procedure is

$$N \geq \frac{p(1-p)}{\delta^2}(z_{0.95} + z_{0.90})^2 = \frac{(0.10)(0.90)(2.927)^2}{(0.10)^2} \cong 77$$

7. Product and Process Comparisons
7.2. Comparisons based on data from one process

7.2.5. Does the defect density meet requirements?

Testing defect densities is based on the Poisson distribution

The number of defects observed in an area of size A units is often assumed to have a Poisson distribution with parameter A x D, where D is the actual process defect density (D is defects per unit area). In other words:

$$P\{\#\text{Defects} = n\} = \frac{(AD)^n}{n!}e^{-AD}.$$

The questions of primary interest for quality control are:

1. Is the defect density within prescribed limits?
2. Is the defect density less than a prescribed limit?
3. Is the defect density greater than a prescribed limit?

Normal approximation to the Poisson

We assume that AD is large enough so that the normal approximation to the Poisson applies (in other words, $AD > 10$ for a reasonable approximation and $AD > 20$ for a good one). That translates to

$$P\{\#\text{Defects} < n\} = \Phi\left(\frac{n - AD}{\sqrt{AD}}\right)$$

where Φ is the standard normal distribution function.

Test statistic based on a normal approximation

If, for a sample of area A with a defect density target of D_0, a defect count of C is observed, then the test statistic

$$Z = \frac{C - AD_0}{\sqrt{AD_0}}$$

can be used exactly as shown in the discussion of the test statistic for fraction defectives in the preceding section.

Testing the hypothesis that the process defect density is less than or equal to D_0

For example, after choosing a sample size of area A (see below for sample size calculation) we can reject that the process defect density is less than or equal to the target D_0 if the number of defects C in the sample is greater than C_A, where

$$C_A = z_{1-\alpha}\sqrt{AD_0} + AD_0$$

and $z_{1-\alpha}$ is the 100(1-α) percentile of the standard normal distribution. The test significance level is 100(1-α). For a 90% significance level use $z_{0.90}$ = 1.282 and for a 95% test use $z_{0.95}$ = 1.645. α is the maximum risk that an acceptable process with a defect density at least as low as D_0 "fails" the test.

Choice of sample size (or area) to examine for defects

In order to determine a suitable area A to examine for defects, you first need to choose an unacceptable defect density level. Call this unacceptable defect density $D_1 = kD_0$, where $k > 1$.

We want to have a probability of less than or equal to β is of "passing" the test (and not rejecting the hypothesis that the true level is D_0 or better) when, in fact, the true defect level is D_1 or worse. Typically β will be 0.2, 0.1 or 0.05. Then we need to count defects in a sample size of area A, where A is equal to

$$A = \frac{k}{D_0}\left(\frac{\frac{z_{1-\alpha}}{\sqrt{k}} - z_\beta}{k-1}\right)^2$$

Example

Suppose the target is D_0 = 4 defects per wafer and we want to verify a new process meets that target. We choose α = 0.1 to be the chance of failing the test if the new process is as good as D_0 (α = the Type I error probability or the "producer's risk") and we choose β = 0.1 for the chance of passing the test if the new process is as bad as 6 defects per wafer (β = the Type II error probability or the "consumer's risk"). That means $z_{1-\alpha}$ = 1.282 and z_β = -1.282.

The sample size needed is A wafers, where

$$A = \frac{1.5}{4}\left(\frac{\frac{1.282}{\sqrt{1.5}} - (-1.282)}{k-1}\right)^2 = 8.1$$

which we round up to 9.

The test criteria is to "accept" that the new process meets target unless the number of defects in the sample of 9 wafers exceeds

$$C_A = z_{1-\alpha}\sqrt{AD_0} + AD_0 = 1.282\sqrt{36} + 36 = 43.7$$

In other words, the reject criteria for the test of the new process is 44 or more defects in the sample of 9 wafers.

Note: Technically, all we can say if we run this test and end up *not rejecting* is that we do not have statistically significant evidence that the new process exceeds target. However, the way we chose the sample size for this test assures us we most likely would have had statistically significant evidence for rejection if

the process had been as bad as 1.5 times the target.

7.2.6. What intervals contain a fixed percentage of the population values?

Observations tend to cluster around the median or mean

Empirical studies have demonstrated that it is typical for a large number of the observations in any study to cluster near the median. In right-skewed data this clustering takes place to the left of (i.e., below) the median and in left-skewed data the observations tend to cluster to the right (i.e., above) the median. In symmetrical data, where the median and the mean are the same, the observations tend to distribute equally around these measures of central tendency.

Various methods

Several types of intervals about the mean that contain a large percentage of the population values are discussed in this section.

- Approximate intervals that contain most of the population values
- Percentiles
- Tolerance intervals for a normal distribution
- Tolerance intervals based on the smallest and largest observations

7.2.6.1. Approximate intervals that contain most of the population values

Empirical intervals

A rule of thumb is that where there is no evidence of significant skewness or clustering, two out of every three observations (67%) should be contained within a distance of one standard deviation of the mean; 90% to 95% of the observations should be contained within a distance of two standard deviations of the mean; 99-100% should be contained within a distance of three standard deviations. This rule can help identify outliers in the data.

Intervals that apply to any distribution

The **Bienayme-Chebyshev** rule states that regardless of how the data are distributed, the percentage of observations that are contained within a distance of k tandard deviations of the mean is at least $(1 - 1/k^2)100\%$.

Exact intervals for the normal distribution

The Bienayme-Chebyshev rule is conservative because it applies to any distribution. For a *normal* distribution, a higher percentage of the observations are contained within k standard deviations of the mean as shown in the following table.

Percentage of observations contained between the mean and k standard deviations

k, No. of Standard Deviations	Empircal Rule	Bienayme-Chebychev	Normal Distribution
1	67%	N/A	68.26%
2	90-95%	at least 75%	95.44%
3	99-100%	at least 88.89%	99.73%
4	N/A	at least 93.75%	99.99%

7.2.6.2. Percentiles

Definitions of order statistics and ranks

For a series of measurements Y_1, ..., Y_N, denote the data ordered in increasing order of magnitude by $Y_{[1]}$, ..., $Y_{[N]}$. These ordered data are called order statistics. If $Y_{[j]}$ is the order statistic that corresponds to the measurement Y_i, then the rank for Y_i is j; i.e.,

$$Y_{[j]} \sim Y_i \Rightarrow r_i = j$$

Definition of percentiles

Order statistics provide a way of estimating proportions of the data that should fall above and below a given value, called a *percentile*. The pth percentile is a value, $Y_{(p)}$, such that at most $(100p)$ % of the measurements are less than this value and at most $100(1-p)$ % are greater. The 50th percentile is called the *median*.

Percentiles split a set of ordered data into hundredths. (Deciles split ordered data into tenths). For example, 70 % of the data should fall below the 70th percentile.

Estimation of percentiles

Percentiles can be estimated from N measurements as follows: for the pth percentile, set $p(N+1)$ equal to $k + d$ for k an integer, and d, a fraction greater than or equal to 0 and less than 1.

1. For $0 < k < N$, $Y(p) = Y_{[k]} + d\left(Y_{[k+1]} - Y_{[k]}\right)$
2. For $k = 0$, $Y(p) = Y_{[1]}$
3. For $k = N$, $Y(p) = Y_{[N]}$

Example and interpretation

For the purpose of illustration, twelve measurements from a gage study are shown below. The measurements are resistivities of silicon wafers measured in ohm·cm.

i	Measurements	Order stats	Ranks
1	95.1772	95.0610	9
2	95.1567	95.0925	6
3	95.1937	95.1065	10

4	95.1959	95.1195	11
5	95.1442	95.1442	5
6	95.0610	95.1567	1
7	95.1591	95.1591	7
8	95.1195	95.1682	4
9	95.1065	95.1772	3
10	95.0925	95.1937	2
11	95.1990	95.1959	12
12	95.1682	95.1990	8

To find the 90th percentile, $p(N+1) = 0.9(13) = 11.7$; $k = 11$, and $d = 0.7$. From condition (1) above, $Y(0.90)$ is estimated to be 95.1981 ohm·cm. This percentile, although it is an estimate from a small sample of resistivities measurements, gives an indication of the percentile for a population of resistivity measurements.

Note that there are other ways of calculating percentiles in common use

Some software packages set $1+p(N-1)$ equal to $k + d$, then proceed as above. The two methods give fairly similar results.

A third way of calculating percentiles (given in some elementary textbooks) starts by calculating pN. If that is not an integer, round up to the next highest integer k and use $Y_{[k]}$ as the percentile estimate. If pN is an integer k, use $0.5(Y_{[k]} + Y_{[k+1]})$.

Definition of Tolerance Interval

An interval covering population percentiles can be interpreted as "covering a proportion p of the population with a level of confidence, say, 90 %." This is known as a tolerance interval.

7.2.6.3. Tolerance intervals for a normal distribution

Definition of a tolerance interval

A confidence interval covers a population parameter with a stated confidence, that is, a certain proportion of the time. There is also a way to cover a fixed proportion of the population with a stated confidence. Such an interval is called a *tolerance interval*. The endpoints of a tolerance interval are called *tolerance limits*. An application of tolerance intervals to manufacturing involves comparing specification limits prescribed by the client with tolerance limits that cover a specified proportion of the population.

Difference between confidence and tolerance intervals

Confidence limits are limits within which we expect a given population parameter, such as the mean, to lie. Statistical tolerance limits are limits within which we expect a stated proportion of the population to lie.

Not related to engineering tolerances

Statistical tolerance intervals have a probabilistic interpretation. *Engineering tolerances* are specified outer limits of acceptability which are usually prescribed by a design engineer and do not necessarily reflect a characteristic of the actual measurements.

Three types of tolerance intervals

Three types of questions can be addressed by tolerance intervals. Question (1) leads to a two-sided interval; questions (2) and (3) lead to one-sided intervals.

1. What interval will contain p percent of the population measurements?
2. What interval guarantees that p percent of population measurements will not fall below a lower limit?
3. What interval guarantees that p percent of population measurements will not exceed an upper limit?

Tolerance intervals for measurements from a normal distribution

For the questions above, the corresponding tolerance intervals are defined by lower (L) and upper (U) tolerance limits which are computed from a series of measurements $Y_1, ..., Y_N$:

1. $Y_L = \bar{Y} - k_2 s; \quad Y_U = \bar{Y} + k_2 s$
2. $Y_L = \bar{Y} - k_1 s$
3. $Y_U = \bar{Y} + k_1 s$

where the k factors are determined so that the intervals cover at least a proportion p of the population with confidence, γ.

Calculation

If the data are from a normally distributed population, an approximate value for the

of k factor for a two-sided tolerance limit for a normal distribution

factor as a function of p and γ for a two-sided tolerance interval (Howe, 1969) is

$$k_2 = \sqrt{\frac{(N-1)\left(1+\frac{1}{N}\right) z^2_{1-(1-p)/2}}{\chi^2_{1-\gamma,N-1}}}$$

where $X^2_{1-\gamma,\, N-1}$, is the critical value of the chi-square distribution with degrees of freedom, N - 1, that is exceeded with probability γ and $z_{1-(1-p)/2}$ is the critical value of the normal distribution which is exceeded with probability (1-p)/2.

Example of calculation

For example, suppose that we take a sample of N = 43 silicon wafers from a lot and measure their thicknesses in order to find tolerance limits within which a proportion p = 0.90 of the wafers in the lot fall with probability γ = 0.99.

Use of tables in calculating two-sided tolerance intervals

Values of the k factor as a function of p and γ are tabulated in some textbooks, such as Dixon and Massey (1969). To use the tables in this handbook, follow the steps outlined below:

1. Calculate $\alpha = (1 - p)/2 = 0.05$
2. Go to the page describing critical values of the normal distribution and in the summary table under the column labeled 0.95 find $z_{1-(1-p)/2} = z_{0.95} = 1.645$.
3. Go to the table of lower critical values of the chi-square distribution and under the column labeled 0.01 in the row labeled degrees of freedom = 42, find
   $X^2_{1-\gamma,\, N-1} = X^2_{0.01,\, 42} = 23.650$.
4. Calculate

$$k_2 = \sqrt{\frac{(N-1)\left(1+\frac{1}{N}\right) z^2_{1-(1-p)/2}}{\chi^2_{1-\gamma,N-1}}} = \sqrt{\frac{42\left(\frac{44}{43}\right)(1.645)^2}{23.650}} = 2.217$$

The tolerance limits are then computed from the sample mean, $\bar{Y}$, and standard deviation, s, according to case(1).

Important notes

The notation for the critical value of the chi-square distribution can be confusing. Values as tabulated are, in a sense, already squared; whereas the critical value for the normal distribution must be squared in the formula above.

Some software is capable of computing a tolerance intervals for a given set of data so that the user does not need to perform all the calculations. All the tolerance intervals shown in this section can be computed using both Dataplot code and R code. R and Dataplot examples include the case where a tolerance interval is computed automatically from a data set.

Calculation of a one-sided tolerance interval for a normal

The calculation of an approximate k factor for one-sided tolerance intervals comes directly from the following set of formulas (Natrella, 1963):

distribution

$$k_1 = \frac{z_p + \sqrt{z_p^2 - ab}}{a}$$

$$a = 1 - \frac{z_\gamma^2}{2(N-1)}$$

$$b = z_p^2 - \frac{z_\gamma^2}{N}$$

A one-sided tolerance interval example

For the example above, it may also be of interest to guarantee with 0.99 probability (or 99% confidence) that 90% of the wafers have thicknesses less than an upper tolerance limit. This problem falls under case (3). The calculations for the k_1 factor for a one-sided tolerance interval are:

$$a = 1 - \frac{1}{2(43-1)}(2.3263)^2 = 0.9356$$

$$b = (1.2816)^2 - \frac{1}{43}(2.3263)^2 = 1.5165$$

$$k_1 = \frac{1.2816 + \sqrt{(1.2816)^2 - (0.9356)(1.5165)}}{0.9356} = 1.8752$$

Tolerance factor based on the non-central t distribution

The value of k_1 can also be computed using the inverse cumulative distribution function for the non-central *t* distribution. This method may give more accurate results for small values of *N*. The value of k_1 using the non-central *t* distribution (using the same example as above) is:

$$\delta = z_p\sqrt{N} = 1.2816\sqrt{43} = 8.4037$$

$$k_1 = \frac{t_{\gamma, N-1, \delta}}{\sqrt{N}} = \frac{12.28834}{\sqrt{43}} = 1.8740$$

where δ is the non-centrality parameter.

In this case, the difference between the two computations is negligble (1.8752 versus 1.8740). However, the difference becomes more pronounced as the value of *N* gets smaller (in particular, for $N \leq 10$). For example, if $N = 43$ is replaced with $N = 6$, the non-central *t* method returns a value of 4.4111 for k_1 while the method based on the Natrella formuals returns a value of 5.2808.

The disadvantage of the non-central *t* method is that it depends on the inverse cumulative distribution function for the non-central *t* distribution. This function is not available in many statistical and spreadsheet software programs, but it is available in Dataplot and R (see Dataplot code and R code). The Natrella formulas only depend on the inverse cumulative distribution function for the normal distribution (which is available in just about all statistical and spreadsheet software programs). Unless you have small samples (say $N \leq 10$), the difference in the

methods should not have much practical effect.

7.2.6.4. Tolerance intervals based on the largest and smallest observations

Tolerance intervals can be constructed for a distribution of any form

The methods on the previous pages for computing tolerance limits are based on the assumption that the measurements come from a normal distribution. If the distribution is not normal, tolerance intervals based on this assumption will not provide coverage for the intended proportion *p* of the population. However, there are methods for achieving the intended coverage if the form of the distribution is not known, but these methods may produce substantially wider tolerance intervals.

Risks associated with making assumptions about the distribution

There are situations where it would be particularly dangerous to make unwarranted assumptions about the exact shape of the distribution, for example, when testing the strength of glass for airplane windshields where it is imperative that a very large proportion of the population fall within acceptable limits.

Tolerance intervals based on largest and smallest observations

One obvious choice for a two-sided tolerance interval for an unknown distribution is the interval between the smallest and largest observations from a sample of Y_1, ..., Y_N measurements. Given the sample size *N* and coverage *p*, an equation from Hahn and Meeker (p. 91),

$$\gamma = 1 - Np^{N-1} + (N-1)p^N$$

allows us to calculate the confidence γ of the tolerance interval. For example, the confidence levels for selected coverages between 0.5 and 0.9999 are shown below for *N* = 25.

Confidence	Coverage
1.000	0.5000
0.993	0.7500
0.729	0.9000
0.358	0.9500
0.129	0.9750
0.026	0.9900
0.007	0.9950
0.0	0.9990
0.0	0.9995
0.0	0.9999

Note that if 99 % confidence is required, the interval that covers the entire sample data set is guaranteed to achieve a coverage of only 75 % of the population values.

What is the optimal sample size?

Another question of interest is, "How large should a sample be so that one can be assured with probability γ that the tolerance interval will contain at least a proportion *p* of the population?"

Approximation for N

A rather good approximation for the required sample size is given by

$$N \cong \frac{1}{4}\frac{(1+p)}{(1-p)}\chi^2_{\gamma,4} + \frac{1}{2}$$

where $X^2_{\gamma,\,4}$ is the critical value of the chi-square distribution with 4 degrees of freedom that is exceeded with probability γ.

Example of the effect of p on the sample size

Suppose we want to know how many measurements to make in order to guarantee that the interval between the smallest and largest observations covers a proportion p of the population with probability γ = 0.95. From the table for the upper critical value of the chi-square distribution, look under the column labeled 0.95 in the row for 4 degrees of freedom. The value is found to be $X^2_{0.95,\,4}$ = 9.488 and calculations are shown below for p equal to 0.90 and 0.99.

For p = 0.90, γ = 0.95:

$$N \cong \frac{1}{4}\frac{(1+0.90)}{(1-0.90)}\chi^2_{0.95,4} + \frac{1}{2} = 0.25(19)(9.488) + 0.5 = 45.57 = 46$$

For p = 0.99, γ = 0.95:

$$N \cong \frac{1}{4}\frac{(1+0.99)}{(1-0.99)}\chi^2_{0.95,4} + \frac{1}{2} = 0.25(199)(9.488) + 0.5 = 472.5 = 473$$

These calculations demonstrate that requiring the tolerance interval to cover a very large proportion of the population may lead to an unacceptably large sample size.

7.3. Comparisons based on data from two processes

Outline for this section

In many manufacturing environments it is common to have two or more processes performing the same task or generating similar products. The following pages describe tests covering several of the most common and useful cases for two processes.

1. Do two processes have the same mean?
    1. Tests when the standard deviations are equal
    2. Tests when the standard deviations are unequal
    3. Tests for paired data
2. Do two processes have the same standard deviation?
3. Do two processes produce the same proportion of defectives?
4. If the observations are failure times, are the failure rates (or mean times to failure) the same?
5. Do two arbitrary processes have the same central tendency?

Example of a dual track process

For example, in an automobile manufacturing plant, there may exist several assembly lines producing the same part. If one line goes down for some reason, parts can still be produced and production will not be stopped. For example, if the parts are piston rings for a particular model car, the rings produced by either line should conform to a given set of specifications.

How does one confirm that the two processes are in fact producing rings that are similar? That is, how does one determine if the two processes are similar?

The goal is to determine if the two processes are similar

In order to answer this question, data on piston rings are collected for each process. For example, on a particular day, data on the diameters of ten piston rings from each process are measured over a one-hour time frame.

To determine if the two processes are similar, we are interested in answering the following questions:

1. Do the two processes produce piston rings with the same diameter?
2. Do the two processes have similar variability in the

diameters of the rings produced?

Unknown standard deviation

The second question assumes that one does not know the standard deviation of either process and therefore it must be estimated from the data. This is usually the case, and the tests in this section assume that the population standard deviations are unknown.

Assumption of a normal distribution

The statistical methodology used (i.e., the specific test to be used) to answer these two questions depends on the underlying distribution of the measurements. The tests in this section assume that the data are normally distributed.

7.3.1. Do two processes have the same mean?

Testing hypotheses related to the means of two processes

Given two random samples of measurements,

$$\boldsymbol{Y_1}, ..., \boldsymbol{Y_N} \text{ and } \boldsymbol{Z_1}, ..., \boldsymbol{Z_N}$$

from two independent processes (the Y's are sampled from process 1 and the Z's are sampled from process 2), there are three types of questions regarding the true means of the processes that are often asked. They are:

1. Are the means from the two processes the same?
2. Is the mean of process 1 less than or equal to the mean of process 2?
3. Is the mean of process 1 greater than or equal to the mean of process 2?

Typical null hypotheses

The corresponding null hypotheses that test the true mean of the first process, μ_1, against the true mean of the second process, μ_2 are:

1. H_0: $\mu_1 = \mu_2$
2. H_0: μ_1 < or equal to μ_2
3. H_0: μ_1 > or equal to μ_2

Note that as previously discussed, our choice of which null hypothesis to use is typically made based on one of the following considerations:

1. When we are hoping to prove something new with the sample data, we make that the alternative hypothesis, whenever possible.
2. When we want to continue to assume a reasonable or traditional hypothesis still applies, unless very strong contradictory evidence is present, we make that the null hypothesis, whenever possible.

Basic statistics from the two processes

The basic statistics for the test are the sample means

$$\overline{Y} = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i \; ; \; \overline{Z} = \frac{1}{N_2}\sum_{i=1}^{N_2} Z_i$$

and the sample standard deviations

$$s_1 = \sqrt{\frac{\sum_{i=1}^{N_1}(Y_i - \bar{Y})^2}{N_1 - 1}}$$

$$s_2 = \sqrt{\frac{\sum_{i=1}^{N_2}(Z_i - \bar{Z})^2}{N_2 - 1}}$$

with degrees of freedom $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$ respectively.

Form of the test statistic where the two processes have equivalent standard deviations

If the standard deviations from the two processes are equivalent, and this should be tested before this assumption is made, the test statistic is

$$t = \frac{\bar{Y} - \bar{Z}}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

where the pooled standard deviation is estimated as

$$s = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 - 1) + (N_2 - 1)}}$$

with degrees of freedom $\nu = N_1 + N_2 - 2$.

Form of the test statistic where the two processes do NOT have equivalent standard deviations

If it cannot be assumed that the standard deviations from the two processes are equivalent, the test statistic is

$$t = \frac{\bar{Y} - \bar{Z}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

The degrees of freedom are not known exactly but can be estimated using the Welch-Satterthwaite approximation

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1 - 1)} + \frac{s_2^4}{N_2^2(N_2 - 1)}}$$

Test strategies

The strategy for testing the hypotheses under (1), (2) or (3) above is to calculate the appropriate *t* statistic from one of the formulas above, and then perform a test at significance level α, where α is chosen to be small, typically .01, .05 or .10. The hypothesis associated with each case enumerated above is rejected if:

1. $|t| \geq t_{1-\alpha/2,\,\nu}$
2. $t \geq t_{1-\alpha,\,\nu}$
3. $t \leq t_{\alpha,\,\nu}$

Explanation of critical values

The critical values from the *t* table depend on the significance level and the degrees of freedom in the standard deviation. For hypothesis (1) $t_{1-\alpha/2,\,\nu}$ is the 1-α/2 critical value from the *t* table with ν degrees of freedom and similarly for hypotheses (2) and (3).

Example of unequal number of data points

A new procedure (process 2) to assemble a device is introduced and tested for possible improvement in time of assembly. The question being addressed is whether the mean, μ_2, of the new assembly process is smaller than the mean, μ_1, for the old assembly process (process 1). We choose to test hypothesis (2) in the hope that we will reject this null hypothesis and thereby feel we have a strong degree of confidence that the new process is an improvement worth implementing. Data (in minutes required to assemble a device) for both the new and old processes are listed below along with their relevant statistics.

Device	Process 1 (Old)	Process 2 (New)
1	32	36
2	37	31
3	35	30
4	28	31
5	41	34
6	44	36
7	35	29
8	31	32
9	34	31
10	38	
11	42	
Mean	36.0909	32.2222
Standard deviation	4.9082	2.5386
No. measurements	11	9
Degrees freedom	10	8

Computation of the test statistic

From this table we generate the test statistic

$$t = \frac{\bar{Y} - \bar{Z}}{\sqrt{s_1^2/N_1 + s_2^2/N_2}} = \frac{36.0909 - 32.2222}{\sqrt{4.9082^2/11 + 2.5386^2/9}} = 2.2694$$

with the degrees of freedom approximated by

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}} = \frac{\left(\frac{4.9082^2}{11} + \frac{2.5386^2}{9}\right)^2}{\frac{4.9082^4}{1210} + \frac{2.5386^4}{648}} = 15.5$$

Decision process

For a one-sided test at the 5% significance level, go to the *t* table for 0.95 signficance level, and look up the critical value for degrees of

freedom v = 16. The critical value is 1.746. Thus, hypothesis (2) is rejected because the test statistic (t = 2.269) is greater than 1.746 and, therefore, we conclude that process 2 has improved assembly time (smaller mean) over process 1.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

7.3.1.1. Analysis of paired observations

Definition of paired comparisons

Given two random samples,

$$\boldsymbol{Y_1}, ..., \boldsymbol{Y_N} \quad \text{and} \quad \boldsymbol{Z_1}, ..., \boldsymbol{Z_N}$$

from two populations, the data are said to be paired if the *i*th measurement on the first sample is naturally paired with the *i*th measurement on the second sample. For example, if *N* supposedly identical products are chosen from a production line, and each one, in turn, is tested with first one measuring device and then with a second measuring device, it is possible to decide whether the measuring devices are compatible; i.e., whether there is a difference between the two measurement systems. Similarly, if "before" and "after" measurements are made with the same device on *N* objects, it is possible to decide if there is a difference between "before" and "after"; for example, whether a cleaning process changes an important characteristic of an object. Each "before" measurement is paired with the corresponding "after" measurement, and the differences

$$d_i = Y_i - X_i \quad (i = 1, \cdots, N)$$

are calculated.

Basic statistics for the test

The mean and standard deviation for the differences are calculated as

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

and

$$s_d = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i - \bar{d})^2}$$

with $v = N - 1$ degrees of freedom.

Test statistic based on the t

The paired-sample *t* test is used to test for the difference of two means before and after a treatment. The test statistic is:

distribution

$$t = \frac{\bar{d}}{s_d / \sqrt{N}}$$

The hypotheses described on the foregoing page are rejected if:

1. $|t| \geq t_{1-\alpha/2, \nu}$
2. $t \geq t_{1-\alpha, \nu}$
3. $t \leq t_{\alpha, \nu}$

where for hypothesis (1) $t_{1-\alpha/2, \nu}$ is the 1-α/2 critical value from the t distribution with ν degrees of freedom and similarly for cases (2) and (3). Critical values can be found in the t table in Chapter 1.

7.3.1.2. Confidence intervals for differences between means

Definition of confidence interval for difference between population means

Given two random samples,

$$Y_1, ..., Y_N \quad \text{and} \quad Z_1, ..., Z_N$$

from two populations, two-sided confidence intervals with 100 (1- α)% coverage for the difference between the unknown population means, μ_1 and μ_2, are shown in the table below. Relevant statistics for paired observations and for unpaired observations are shown elsewhere.

Two-sided confidence intervals with 100(1- α)% coverage for μ_1 - μ_2:

Paired observations

$\mu_1 - \mu_2$ (where $\sigma_1 = \sigma_2$)	$\bar{d} \pm t_{1-\alpha/2,N-1} \frac{s_d}{\sqrt{N}}$

Unpaired observations

$\mu_1 - \mu_2$ (where $\sigma_1 = \sigma_2$)	$\bar{Y} - \bar{Z} \pm t_{1-\alpha/2,N_1+N_2-2}\ s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$
$\mu_1 - \mu_2$ (where $\sigma_1 \neq \sigma_2$)	$\bar{Y} - \bar{Z} \pm t_{1-\alpha/2,effective\ df}\ s\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$

Interpretation of confidence interval

One interpretation of the confidence interval for means is that if zero is contained within the confidence interval, the two population means are equivalent.

7.3.2. Do two processes have the same standard deviation?

Testing hypotheses related to standard deviations from two processes

Given two random samples of measurements,

$$Y_1, ..., Y_N \quad \text{and} \quad Z_1, ..., Z_N$$

from two independent processes, there are three types of questions regarding the true standard deviations of the processes that can be addressed with the sample data. They are:

1. Are the standard deviations from the two processes the same?
2. Is the standard deviation of one process less than the standard deviation of the other process?
3. Is the standard deviation of one process greater than the standard deviation of the other process?

Typical null hypotheses

The corresponding null hypotheses that test the true standard deviation of the first process, σ_1, against the true standard deviation of the second process, σ_2 are:

1. H_0: $\sigma_1 = \sigma_2$
2. H_0: $\sigma_1 \leq \sigma_2$
3. H_0: $\sigma_1 \geq \sigma_2$

Basic statistics from the two processes

The basic statistics for the test are the sample variances

$$s_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N1} (Y_i - \bar{Y})^2$$

$$s_2^2 = \frac{1}{N_2 - 1} \sum_{i=1}^{N2} (Z_i - \bar{Z})^2$$

and degrees of freedom $\nu_1 = N_1 - 1$ and $\nu_2 = N_2 - 1$, respectively.

Form of the test statistic

The test statistic is

$$F = \frac{s_1^2}{s_2^2}$$

Test strategies

The strategy for testing the hypotheses under (1), (2) or (3) above is to calculate the *F* statistic from the formula above, and then perform a test at significance level α, where α is chosen to be small, typically 0.01, 0.05 or 0.10. The hypothesis associated with each case enumerated above is rejected if:

1. $F \leq \frac{1}{F_{\alpha/2;\nu 2;\nu 1}}$ or $F \geq F_{\alpha/2;\nu 1;\nu 2}$
2. $F \geq F_{\alpha;\nu 1;\nu 2}$
3. $F \leq \frac{1}{F_{\alpha;\nu 2;\nu 1}}$

Explanation of critical values

The critical values from the *F* table depend on the significance level and the degrees of freedom in the standard deviations from the two processes. For hypothesis (1):

- $F_{\alpha/2;\nu_2;\nu_1}$ is the upper critical value from the F table with
- $\nu_2 = N_2 - 1$ degrees of freedom for the numerator and
- $\nu_1 = N_1 - 1$ degrees of freedom for the denominator

and

- $F_{\alpha/2;\nu_1;\nu_2}$ is the upper critical value from the F table with
- $\nu_1 = N_1 - 1$ degrees of freedom for the numerator and
- $\nu_2 = N_2 - 1$ degrees of freedom for the denominator.

***Caution** on looking up critical values*

The *F* distribution has the property that

$$F_{1-\alpha/2;\nu_1;\nu_2} = \frac{1}{F_{\alpha/2;\nu_2;\nu_1}}$$

which means that only upper critical values are required for two-sided tests. However, note that the degrees of freedom are interchanged in the ratio. For example, for a two-sided test at significance level 0.05, go to the *F* table labeled "2.5% significance level".

- For $F_{\alpha/2;\nu_2;\nu_1}$, reverse the order of the degrees of freedom; i.e., look across the top of the table for and down the table for .

$\nu_2 = N_2 - 1$ $\nu_1 = N_1 - 1$

- For $F_{\alpha/2;\nu_1;\nu_2}$, look across the top of the table for $\nu_1 = N_1 - 1$ and down the table for $\nu_2 = N_2 - 1$.

Critical values for cases (2) and (3) are defined similarly, except that the critical values for the one-sided tests are based on α rather than on $\alpha/2$.

Two-sided confidence interval

The two-sided confidence interval for the ratio of the two unknown variances (squares of the standard deviations) is shown below.

Two-sided confidence interval with 100(1- α)% coverage for:

$$\frac{\sigma_1^2}{\sigma_2^2} \quad \left[\frac{1}{F_{\alpha/2;N1-1;N2-1}} \left(\frac{s_1^2}{s_2^2} \right), F_{\alpha/2;N2-1;N1-1} \left(\frac{s_1^2}{s_2^2} \right) \right]$$

One interpretation of the confidence interval is that if the quantity "one" is contained within the interval, the standard deviations are equivalent.

Example of unequal number of data points

A new procedure to assemble a device is introduced and tested for possible improvement in time of assembly. The question being addressed is whether the standard deviation, σ_2, of the new assembly process is better (i.e., smaller) than the standard deviation, σ_1, for the old assembly process. Therefore, we test the null hypothesis that $\sigma_1 \leq \sigma_2$. We form the hypothesis in this way because we hope to reject it, and therefore accept the alternative that σ_2 is less than σ_1. This is hypothesis (2). Data (in minutes required to assemble a device) for both the old and new processes are listed on an earlier page. Relevant statistics are shown below:

	Process 1	Process 2
Mean	36.0909	32.2222
Standard deviation	4.9082	2.5874
No. measurements	11	9
Degrees freedom	10	8

Computation of the test statistic

From this table we generate the test statistic

$$F = \frac{s_1^2}{s_2^2} = \left(\frac{4.9082}{2.5874} \right)^2 = 3.60$$

Decision process

For a test at the 5% significance level, go to the *F* table for 5% signficance level, and look up the critical value for numerator degrees of freedom $\nu_1 = N_1 - 1$ = 10 and denominator degrees of freedom $\nu_2 = N_2 - 1$ = 8. The critical value is 3.35. Thus, hypothesis (2) can be rejected because

the test statistic (F = 3.60) is greater than 3.35. Therefore, we accept the alternative hypothesis that process 2 has better precision (smaller standard deviation) than process 1.

7.3.3. How can we determine whether two processes produce the same proportion of defectives?

Case 1: Large Samples (Normal Approximation to Binomial)

The hypothesis of equal proportions can be tested using a z statistic

If the samples are reasonably large we can use the normal approximation to the binomial to develop a test similar to testing whether two normal means are equal.

Let sample 1 have x_1 defects out of n_1 and sample 2 have x_2 defects out of n_2. Calculate the proportion of defects for each sample and the z statistic below:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$$

where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

Compare $|z|$ to the normal $z_{1-\alpha/2}$ table value for a two-sided test. For a one-sided test, assuming the alternative hypothesis is $p_1 > p_2$, compare z to the normal $z_{1-\alpha}$ table value. If the alternative hypothesis is $p_1 < p_2$, compare z to z_{α}.

Case 2: An Exact Test for Small Samples

The Fisher Exact Probability test is an excellent choice for small samples

The **Fisher Exact Probability Test** is an excellent nonparametric technique for analyzing discrete data (either nominal or ordinal), when the two independent samples are small in size. It is used when the results from two independent random samples fall into one or the other of two mutually exclusive classes (i.e., defect versus good, or successes vs failures).

Example of a 2x2

In other words, every subject in each group has one of two possible scores. These scores are represented by frequencies

contingency table

in a 2x2 contingency table. The following discussion, using a 2x2 contingency table, illustrates how the test operates.

We are working with two independent groups, such as experiments and controls, males and females, the Chicago Bulls and the New York Knicks, etc.

	-	+	Total
Group I	A	B	A+B
Group II	C	D	C+D
Total	A+C	B+D	N

The column headings, here arbitrarily indicated as plus and minus, may be of any two classifications, such as: above and below the median, passed and failed, Democrat and Republican, agree and disagree, etc.

Determine whether two groups differ in the proportion with which they fall into two classifications

Fisher's test determines whether the two groups differ in the proportion with which they fall into the two classifications. For the table above, the test would determine whether Group I and Group II differ significantly in the proportion of plusses and minuses attributed to them.

The method proceeds as follows:

The exact probability of observing a particular set of frequencies in a 2 × 2 table, when the marginal totals are regarded as fixed, is given by the hypergeometric distribution

$$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{N}{A+B}}$$

$$= \frac{\frac{(A+C)!}{A!C!}\frac{(B+D)!}{B!D!}}{\frac{N!}{(A+B)!(C+D)!}}$$

$$p = \frac{(A+B)!(C+D)!(A+C)!(B+D)!}{N!A!B!C!D!}$$

But the test does not just look at the observed case. If needed, it also computes the probability of more extreme outcomes, *with the same marginal totals.* By "more extreme", we mean relative to the null hypothesis of equal proportions.

Example of Fisher's test

This will become clear in the next illustrative example. Consider the following set of 2 x 2 contingency tables:

Observed Data			More extreme outcomes with same marginals					
(a)			(b)			(c)		
2	5	7	1	6	7	0	7	7
3	2	5	4	1	5	5	0	5
5	7	12	5	7	12	5	7	12

Table (a) shows the observed frequencies and tables (b) and (c) show the two more extreme distributions of frequencies that could occur with the same marginal totals 7, 5. Given the observed data in table (a) , we wish to test the null hypothesis at, say, α = 0.05.

Applying the previous formula to tables (a), (b), and (c), we obtain

$$p_a = \frac{7!5!5!7!}{12!2!5!3!2!} = .26515$$

$$p_b = \frac{7!5!5!7!}{12!1!6!4!1!} = .04419$$

$$p_c = \frac{7!5!5!7!}{12!0!7!5!0!} = .00126$$

The probability associated with the occurrence of values as extreme as the observed results under H_0 is given by adding these three p's:

$$.26515 + .04419 + .00126 = .31060$$

So p = 0.31060 is the probability that we get from Fisher's test. Since 0.31060 is larger than α, we cannot reject the null hypothesis.

Tocher's Modification

Tocher's modification makes Fisher's test less conservative

Tocher (1950) showed that a slight modification of the Fisher test makes it a more useful test. Tocher starts by isolating the probability of all cases more extreme than the observed one. In this example that is

$$p_b + p_c = .04419 + .00126 = .04545$$

Now, if this probability is larger than α, we cannot reject H_0. But if this probability is less than α, while the probability that we got from Fisher's test is greater than α

(as is the case in our example) then Tocher advises to compute the following ratio:

$$\frac{\alpha - p_{more\ extreme\ cases}}{p_{observed\ alone}}$$

For the data in the example, that would be

$$\frac{\alpha - (p_b + p_c)}{p_a} = \frac{.05 - .04545}{.2615} = .0172$$

Now we go to a table of random numbers and at random draw a number between 0 and 1. If this random number is *smaller* than the ratio above of 0.0172, we reject H_0. If it is larger we cannot reject H_0. This added small probability of rejecting H_0 brings the test procedure Type I error (i.e., α value) to exactly 0.05 and makes the Fisher test less conservative.

The test is a one-tailed test. For a two-tailed test, the value of *p* obtained from the formula must be doubled.

A difficulty with the Tocher procedure is that someone else analyzing the same data would draw a different random number and possibly make a different decision about the validity of H_0.

7.3.4. Assuming the observations are failure times, are the failure rates (or Mean Times To Failure) for two distributions the same?

Comparing two exponential distributions is to compare the means or hazard rates

The comparison of two (or more) life distributions is a common objective when performing statistical analyses of lifetime data. Here we look at the one-parameter exponential distribution case.

In this case, comparing two exponential distributions is equivalent to comparing their means (or the reciprocal of their means, known as their hazard rates).

Type II Censored data

Definition of Type II censored data

Definition: Type II censored data occur when a life test is terminated exactly when a pre-specified number of failures have occurred. The remaining units have not yet failed. If n units were on test, and the pre-specified number of failures is r (where r is less than or equal to n), then the test ends at t_r = the time of the r-th failure.

Two exponential samples oredered by time

Suppose we have Type II censored data from two exponential distributions with means θ_1 and θ_2. We have two samples from these distributions, of sizes n_1 on test with r_1 failures and n_2 on test with r_2 failures, respectively. The observations are time to failure and are therefore ordered by time.

$$t_{1(1)} < \cdots < t_{1(r_1)} \quad (r_1 \le n_1)$$
$$t_{2(1)} < \cdots < t_{2(r_2)} \quad (r_2 \le n_2)$$

Test of equality of θ_1 and θ_2 and confidence

Letting

$$T_i = \sum_{j=1}^{r_i} t_{i(j)} + (n_i - r_i)t_{i(r_i)} \qquad i = 1, 2$$

interval for θ_1/θ_2

Then

$$2T_1/\theta_1 \approx \chi^2_{2r1}$$

and

$$2T_2/\theta_2 \approx \chi^2_{2r2}$$

with T_1 and T_2 independent. Thus

$$U = \frac{2T_1/(2r_1\theta_1)}{2T_2/(2r_2\theta_2)} = \frac{\hat{\theta}_1\theta_2}{\hat{\theta}_2\theta_1},$$

where

$$\hat{\theta}_1 = \frac{T_1}{r_1} \text{ and } \hat{\theta}_2 = \frac{T_2}{r_2}$$

has an F distribution with $(2r_1, 2r_2)$ degrees of freedom. Tests of equality of θ_1 and θ_2 can be performed using tables of the F distribution or computer programs. Confidence intervals for θ_1 / θ_2, which is the ratio of the means or the hazard rates for the two distributions, are also readily obtained.

Numerical example

A numerical application will illustrate the concepts outlined above.

For this example,

H_0: $\theta_1/\theta_2 = 1$

H_a: $\theta_1/\theta_2 \neq 1$

Two samples of size 10 from exponential distributions were put on life test. The first sample was censored after 7 failures and the second sample was censored after 5 failures. The times to failure were:

Sample 1: 125 189 210 356 468 550 610
Sample 2: 170 234 280 350 467

So $r_1 = 7$, $r_2 = 5$ and $t_{1,(r1)} = 610$, $t_{2,(r2)}=467$.

Then $T_1 = 4338$ and $T_2 = 3836$.

The estimator for θ_1 is 4338 / 7 = 619.71 and the estimator for θ_2 is 3836 / 5 = 767.20.

The ratio of the estimators = U = 619.71 / 767.20 = .808.

If the means are the same, the ratio of the estimators, U, follows an F distribution with $2r_1$, $2r_2$ degrees of freedom. The $P(F < .808) = .348$. The associated p-value is $2(.348) = .696$. Based on this p-value, we find no evidence to reject the null hypothesis (that the true but unknown ratio = 1). Note that this is a two-sided test, and we would reject the null hyposthesis if the p-value is either too small (i.e., less or equal to .025) or too large (i.e., greater than or equal to .975) for a 95% significance level test.

We can also put a 95% confidence interval around the ratio of the two means. Since the .025 and .975 quantiles of $F_{(14,10)}$ are 0.3178 and 3.5504, respectively, we have

$$\Pr(U/3.5504 < \theta_1/\theta_2 < U/.3178) = .95$$

and (.228, 2.542) is a 95% confidence interval for the ratio of the unknown means. The value of 1 is within this range, which is another way of showing that we cannot reject the null hypothesis at the 95% significance level.

7. Product and Process Comparisons
7.3. Comparisons based on data from two processes

7.3.5. Do two arbitrary processes have the same central tendency?

The nonparametric equivalent of the t test is due to Mann and Whitney, called the U test

By "arbitrary" we mean that we make no underlying assumptions about normality or any other distribution. The test is called the **Mann-Whitney U Test**, which is the nonparametric equivalent of the *t* test for means.

The *U*-test (as the majority of nonparametric tests) uses the rank sums of the two samples.

Procedure

The test is implemented as follows.

1. Rank all (n_1 + n_2) observations in ascending order. Ties receive the average of their observations.
2. Calculate the sum of the ranks, call these T_a and T_b
3. Calculate the U statistic,

$$U_a = n_1(n_2) + 0.5(n_1)(n_1 + 1) - T_a$$

or

$$U_b = n_1(n_2) + 0.5(n_2)(n_2 + 1) - T_b$$

where $U_a + U_b = n_1(n_2)$.

Null Hypothesis

The null hypothesis is: the two populations have the same central tendency. The alternative hypothesis is: The central tendencies are **NOT** the same.

Test statistic

The test statistic, *U*, is the smaller of U_a and U_b. For sample sizes larger than 20, we can use the normal *z* as follows:

$$z = [\ U - E(U)] / \sigma$$

where

$$E(U) = .5(n_1)(n_2) \text{ and } \sigma^2 = \left[n_1(n_2)(n_1 + n_2 + 1)\right] / 12$$

The critical value is the normal tabled *z* for $\alpha/2$ for a two-tailed test or *z* at α

level, for a one-tail test.

For small samples, tables are readily available in most textbooks on nonparametric statistics.

Example

An illustrative example of the U test

Two processing systems were used to clean wafers. The following data represent the (coded) particle counts. The null hypothesis is that there is no difference between the central tendencies of the particle counts; the alternative hypothesis is that there is a difference. The solution shows the typical kind of output software for this procedure would generate, based on the large sample approximation.

Group A	Rank	Group B	Rank
.55	8	.49	5
.67	15.5	.68	17
.43	1	.59	9.5
.51	6	.72	19
.48	3.5	.67	15.5
.60	11	.75	20.5
.71	18	.65	13.5
.53	7	.77	22
.44	2	.62	12
.65	13.5	.48	3.5
.75	20.5	.59	9.5

	N	Sum of Ranks	U	Std. Dev of U	Median
A	11	106.000	81.000	15.229	0.540
B	11	147.000	40.000	15.229	0.635

For U = 40.0 and E[U] = 0.5(n_1)(n_2) = 60.5, the test statistic is

$$z = \frac{U - \mathrm{E}(U)}{\sigma} = \frac{40.0 - 60.5}{15.23} = -1.346$$

where

$$\sigma = \sqrt{\frac{n_1(n_2)(n_1 + n_2 + 1)}{12}} = \sqrt{\frac{11(11)(11 + 11 + 1)}{12}} = 15.23$$

For a two-sided test with significance level α = 0.05, the critical value is $z_{1-\alpha/2}$ = 1.96. Since |z| is less than the critical value, we do not reject the null

hypothesis and conclude that there is not enough evidence to claim that two groups have different central tendencies.

7. Product and Process Comparisons

7.4. Comparisons based on data from more than two processes

Introduction This section begins with a nonparametric procedure for comparing several populations with unknown distributions. Then the following topics are discussed:

- Comparing variances
- Comparing means (ANOVA technique)
- Estimating variance components
- Comparing categorical data
- Comparing population proportion defectives
- Making multiple comparisons

7.4.1. How can we compare several populations with unknown distributions (the Kruskal-Wallis test)?

The Kruskal-Wallis (KW) Test for Comparing Populations with Unknown Distributions

A nonparametric test for comparing population medians by Kruskal and Wallis

The KW procedure tests the null hypothesis that k samples from possibly different populations actually originate from similar populations, at least as far as their central tendencies, or medians, are concerned. The test assumes that the variables under consideration have underlying continuous distributions.

In what follows assume we have k samples, and the sample size of the i-th sample is n_i, $i = 1, 2, \ldots, k$.

Test based on ranks of combined data

In the computation of the KW statistic, each observation is replaced by its rank in an ordered combination of all the k samples. By this we mean that the data from the k samples combined are ranked in a single series. The minimum observation is replaced by a rank of 1, the next-to-the-smallest by a rank of 2, and the largest or maximum observation is replaced by the rank of N, where N is the total number of observations in all the samples (N is the sum of the n_i).

Compute the sum of the ranks for each sample

The next step is to compute the sum of the ranks for each of the original samples. The KW test determines whether these sums of ranks are so different by sample that they are not likely to have all come from the same population.

Test statistic follows a χ^2 distribution

It can be shown that if the k samples come from the same population, that is, if the null hypothesis is true, then the test statistic, H, used in the KW procedure is distributed approximately as a chi-square statistic with $df = k - 1$, provided that the sample sizes of the k samples are not too small (say, $n_i > 4$, for all i). H is defined as follows:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

where

- k = number of samples (groups)
- n_i = number of observations for the i-th sample or group
- N = total number of observations (sum of all the n_i)
- R_i = sum of ranks for group i

Example

An illustrative example

The following data are from a comparison of four investment firms. The observations represent percentage of growth during a three month period.for recommended funds.

A	B	C	D
4.2	3.3	1.9	3.5
4.6	2.4	2.4	3.1
3.9	2.6	2.1	3.7
4.0	3.8	2.7	4.1
	2.8	1.8	4.4

Step 1: Express the data in terms of their ranks

	A	B	C	D
	17	10	2	11
	19	4.5	4.5	9
	14	6	3	12
	15	13	7	16
		8	1	18
SUM	65	41.5	17.5	66

Compute the test statistic

The corresponding H test statistic is

$$H = \frac{12}{(19)(20)}\left[\frac{65^2}{4} + \frac{41.5^2}{5} + \frac{17.5^2}{5} + \frac{66^2}{5}\right] - 3(20) = 13.678$$

From the chi-square table in Chapter 1, the critical value for $1-\alpha = 0.95$ with $df = k-1 = 3$ is 7.812. Since 13.678 > 7.812, we reject the null hypothesis.

Note that the rejection region for the KW procedure is one-sided, since we only reject the null hypothesis when the H statistic is too large.

7. Product and Process Comparisons
7.4. Comparisons based on data from more than two processes

7.4.2. Assuming the observations are normal, do the processes have the same variance?

Before comparing means, test whether the variances are equal

Techniques for comparing means of normal populations generally assume the populations have the same variance. Before using these ANOVA techniques, it is advisable to test whether this assumption of homogeneity of variance is reasonable. The following procedure is widely used for this purpose.

Bartlett's Test for Homogeneity of Variances

Null hypothesis

Bartlett's test is a commonly used test for equal variances. Let's examine the null and alternative hypotheses.

$$H_0 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

against

$$H_a = \text{the } \sigma_i^2 \text{ are not all equal}$$

Test statistic

Assume we have samples of size n_i from the i-th population, $i = 1, 2, \ldots, k$, and the usual variance estimates from each sample:

$$s_1^2, s_2^2, \ldots, s_k^2$$

where

$$s_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2/(n_j - 1)$$

Now introduce the following notation: $\nu_j = n_j - 1$ (the ν_j are the degrees of freedom) and

$$\nu = \sum_{i=1}^{k} \nu_i$$

$$s^2 = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{\nu}$$

The Bartlett's test statistic M is defined by

$$M = \nu \log s^2 - \sum_{i=1}^{k} \nu_i \log s_i^2$$

Distribution of the test statistic

When none of the degrees of freedom is small, Bartlett showed that M is distributed approximately as χ^2_{k-1}. The chi-square approximation is generally acceptable if all the n_i are at least 5.

Bias correction

This is a slightly biased test, according to Bartlett. It can be improved by dividing M by the factor

$$C = 1 + \frac{1}{3(k-1)}\left(\left[\sum_{i=1}^{k} \frac{1}{\nu_i}\right] - \frac{1}{\nu}\right)$$

Instead of M, it is suggested to use M/C for the test statistic.

Bartlett's test is not robust

This test is not robust, it is very sensitive to departures from normality.

An alternative description of Bartlett's test appears in Chapter 1.

Gear Data Example (from Chapter 1):

An illustrative example of Bartlett's test

Gear diameter measurements were made on 10 batches of product. The complete set of measurements appears in Chapter 1. Bartlett's test was applied to this dataset leading to a rejection of the assumption of equal batch variances at the .05 critical value level. applied to this dataset

The Levene Test for Homogeneity of Variances

The Levene test for equality of variances

Levene's test offers a more robust alternative to Bartlett's procedure. That means it will be less likely to reject a true hypothesis of equality of variances just because the distributions of the sampled populations are not normal. When non-normality is suspected, Levene's procedure is a better choice than Bartlett's.

Levene's test is described in Chapter 1. This description also includes an example where the test is applied to the gear data. Levene's test does not reject the assumption of equality of batch variances for these data. This differs from the conclusion drawn from Bartlett's test and is a better answer if, indeed, the batch population distributions are non-normal.

7.4.3. Are the means equal?

Test equality of means

The procedure known as the *Analysis of Variance* or *ANOVA* is used to test hypotheses concerning means when we have several populations.

The Analysis of Variance (ANOVA)

The ANOVA procedure is one of the most powerful statistical techniques

ANOVA is a general technique that can be used to test the hypothesis that the means among two or more groups are equal, under the assumption that the sampled populations are normally distributed.

A couple of questions come immediately to mind: **what** means? and why analyze **variances** in order to derive conclusions about the **means**?

Both questions will be answered as we delve further into the subject.

Introduction to ANOVA

To begin, let us study the effect of temperature on a passive component such as a resistor. We select three different temperatures and observe their effect on the resistors. This experiment can be conducted by measuring all the participating resistors before placing *n* resistors each in three different ovens.

Each oven is heated to a selected temperature. Then we measure the resistors again after, say, 24 hours and analyze the responses, which are the differences between before and after being subjected to the temperatures. The temperature is called a *factor*. The different temperature settings are called *levels*. In this example there are three levels or settings of the factor Temperature.

What is a factor?

A factor is an independent treatment variable whose settings (values) are controlled and varied by the experimenter. The intensity setting of a factor is the level.

- **Levels may be quantitative numbers or, in many cases, simply "present" or "not present" ("0" or "1").**

The 1-way

In the experiment above, there is only one factor,

ANOVA

temperature, and the analysis of variance that we will be using to analyze the effect of temperature is called a *one-way* or *one-factor ANOVA*.

The 2-way or 3-way ANOVA

We could have opted to also study the effect of positions in the oven. In this case there would be two factors, temperature and oven position. Here we speak of a *two-way* or *two-factor ANOVA*. Furthermore, we may be interested in a third factor, the effect of time. Now we deal with a *three-way* or *three-factor* ANOVA. In each of these ANOVA's we test a variety of hypotheses of equality of means (or average responses when the factors are varied).

Hypotheses that can be tested in an ANOVA

First consider the one-way ANOVA. The null hypothesis is: there is no difference in the population means of the different levels of factor A (the only factor).

The alternative hypothesis is: the means are not the same.

For the 2-way ANOVA, the possible null hypotheses are:

1. There is no difference in the means of factor A
2. There is no difference in means of factor B
3. There is no interaction between factors A and B

The alternative hypothesis for cases 1 and 2 is: the means are not equal.

The alternative hypothesis for case 3 is: there is an interaction between A and B.

For the 3-way ANOVA: The main effects are factors A, B and C. The 2-factor interactions are: AB, AC, and BC. There is also a three-factor interaction: ABC.

For each of the seven cases the null hypothesis is the same: there is no difference in means, and the alternative hypothesis is the means are not equal.

The n-way ANOVA

In general, the number of main effects and interactions can be found by the following expression:

$$N = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n}$$

The first term is for the overall mean, and is always 1. The second term is for the number of main effects. The third term is for the number of 2-factor interactions, and so on. The last term is for the *n*-factor interaction and is always 1.

In what follows, we will discuss only the 1-way and 2-way ANOVA.

7.4.3.1. 1-Way ANOVA overview

Overview and principles

This section gives an overview of the one-way ANOVA. First we explain the principles involved in the 1-way ANOVA.

Partition response into components

In an analysis of variance the variation in the response measurements is partitoned into components that correspond to different sources of variation.

The goal in this procedure is to split the total variation in the data into a portion due to random error and portions due to changes in the values of the independent variable(s).

Variance of n measurements

The variance of *n* measurements is given by

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

where $\bar{y}$ is the mean of the *n* measurements.

Sums of squares and degrees of freedom

The numerator part is called the *sum of squares* of deviations from the mean, and the denominator is called the *degrees of freedom*.

The variance, after some algebra, can be rewritten as:

$$s^2 = \frac{\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 / n}{n-1}$$

The first term in the numerator is called the "*raw sum of squares*" and the second term is called the "*correction term for the mean*". Another name for the numerator is the "*corrected sum of squares*", and this is usually abbreviated by *Total SS or SS(Total)*.

The SS in a 1-way ANOVA can be split into two components, called the *"sum of squares of treatments*" and *"sum of squares of error*", abbreviated as SST and SSE, respectively.

The guiding principle behind ANOVA is the decomposition of the sums of squares, or Total SS

Algebraically, this is expressed by

$$\text{Total SS} = \text{SST} + \text{SSE}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{..}\right)^2 = \sum_{i=1}^{k} n_i\left(\bar{y}_{i.}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_{i.}\right)^2$$

where k is the number of *treatments* and the bar over the $y_{..}$ denotes the "grand" or "overall" mean. Each n_i is the number of observations for treatment i. The total number of observations is N (the sum of the n_i).

Note on subscripting

Don't be alarmed by the double subscripting. The total SS can be written single or double subscripted. The double subscript stems from the way the data are arranged in the data table. The table is usually a rectangular array with k columns and each column consists of n_i rows (however, the lengths of the rows, or the n_i, may be unequal).

Definition of "Treatment"

We introduced the concept of treatment. The definition is:
A treatment is a specific combination of factor levels whose effect is to be compared with other treatments.

7.4.3.2. The 1-way ANOVA model and assumptions

A model that describes the relationship between the response and the treatment (between the dependent and independent variables)

The mathematical model that describes the relationship between the response and treatment for the one-way ANOVA is given by

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where Y_{ij} represents the j-th observation (j = 1, 2, ...n_i) on the i-th treatment (i = 1, 2, ..., k levels). So, Y_{23} represents the third observation using level 2 of the factor. μ is the common effect for the whole experiment, τ_i represents the i-th treatment effect and ϵ_{ij} represents the random error present in the j-th observation on the i-th treatment.

Fixed effects model

The errors ϵ_{ij} are assumed to be normally and independently (NID) distributed, with mean zero and variance σ_ϵ^2. μ is always a fixed parameter and $\tau_1, \tau_2, ..., \tau_k$ are considered to be fixed parameters *if the levels of the treatment are fixed*, and not a random sample from a population of possible levels. It is also assumed that μ is chosen so that

$$\sum \tau_i = 0 \qquad i = 1, ..., k$$

holds. This is the *fixed effects model*.

Random effects model

If the k levels of treatment are chosen at random, the model equation remains the same. However, now the τ_i's are random variables assumed to be NID(0, σ_τ). This is the *random effects model*.

Whether the levels are fixed or random depends on how these levels are chosen in a given experiment.

7.4.3.3. The ANOVA table and tests of hypotheses about means

Sums of Squares help us compute the variance estimates displayed in ANOVA Tables

The sums of squares **SST** and **SSE** previously computed for the one-way ANOVA are used to form two mean squares, one for *treatments* and the second for *error*. These mean squares are denoted by *MST* and *MSE*, respectively. These are typically displayed in a tabular form, known as an *ANOVA Table*. The ANOVA table also shows the statistics used to test hypotheses about the population means.

Ratio of MST and MSE

When the null hypothesis of equal means is true, the two mean squares estimate the same quantity (error variance), and should be of approximately equal magnitude. In other words, their ratio should be close to 1. If the null hypothesis is false, MST should be larger than MSE.

Divide sum of squares by degrees of freedom to obtain mean squares

The mean squares are formed by dividing the sum of squares by the associated degrees of freedom.

Let $N = \sum n_i$. Then, the degrees of freedom for treatment, DFT = k - 1, and the degrees of freedom for error, DFE = N - k.

The corresponding *mean squares* are:

$$MST = SST / DFT$$
$$MSE = SSE / DFE$$

The F-test

The test statistic, used in testing the equality of treatment means is: F = MST / MSE.

The critical value is the tabular value of the F distribution, based on the chosen α level and the degrees of freedom DFT and DFE.

The calculations are displayed in an ANOVA table, as follows:

ANOVA table

Source	SS	DF	MS	F

Treatments	SST	k-1	SST / (k-1)	MST/MSE
Error	SSE	N-k	SSE / (N-k)	
Total (corrected)	SS	N-1		

The word "source" stands for source of variation. Some authors prefer to use "between" and "within" instead of "treatments" and "error", respectively.

ANOVA Table Example

A numerical example

The data below resulted from measuring the difference in resistance resulting from subjecting identical resistors to three different temperatures for a period of 24 hours. The sample size of each group was 5. In the language of Design of Experiments, we have an experiment in which each of three treatments was replicated 5 times.

	Level 1	Level 2	Level 3
	6.9	8.3	8.0
	5.4	6.8	10.5
	5.8	7.8	8.1
	4.6	9.2	6.9
	4.0	6.5	9.3
means	5.34	7.72	8.56

The resulting ANOVA table is

Example ANOVA table

Source	SS	DF	MS	F
Treatments	27.897	2	13.949	9.59
Error	17.452	12	1.454	
Total (corrected)	45.349	14		
Correction Factor	779.041	1		

Interpretation of the ANOVA table

The test statistic is the F value of 9.59. Using an α of .05, we have that $F_{.05;\,2,\,12} = 3.89$ (see the F distribution table in Chapter 1). Since the test statistic is much larger than the critical value, we reject the null hypothesis of equal population means and conclude that there is a (statistically)

significant difference among the population means. The *p*-value for 9.59 is .00325, so the test statistic is significant at that level.

Techniques for further analysis

The populations here are resistor readings while operating under the three different temperatures. What we do **not** know at this point is whether the three means are all different or which of the three means is different from the other two, and by how much.

There are several techniques we might use to further analyze the differences. These are:

- constructing confidence intervals around the difference of two means,
- estimating combinations of factor levels with confidence bounds
- multiple comparisons of combinations of factor levels tested simultaneously.

7.4.3.4. 1-Way ANOVA calculations

Formulas for 1-way ANOVA hand calculations

Although computer programs that do ANOVA calculations now are common, for reference purposes this page describes how to calculate the various entries in an ANOVA table. Remember, the goal is to produce two variances (of treatments and error) and their ratio. The various computational formulas will be shown and applied to the data from the previous example.

Step 1: compute CM

STEP 1 Compute CM, the correction for the mean.

$$CM = \frac{\left(\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}\right)^2}{N_{total}} = \frac{(\text{Total of all observations})^2}{N_{total}}$$

$$= \frac{(108.1)^2}{15} = 779.041$$

Step 2: compute total SS

STEP 2 Compute the total SS.

The total SS = sum of squares of all observations - CM

$$SS_{total} = \sum_{i=1}^{3}\sum_{j=1}^{5} y_{ij}^2 - CM$$

$$= (6.9)^2 + (5.4)^2 + \cdots + (6.9)^2 + 9.3^2 - CM$$

$$= 829.390 - 779.041 = 45.349$$

The 829.390 SS is called the "raw" or "uncorrected " sum of squares.

Step 3: compute SST

STEP 3 Compute SST, the treatment sum of squares.

First we compute the total (sum) for each treatment.

T_1 = (6.9) + (5.4) + ... + (4.0) = 26.7

$T_2 = (8.3) + (6.8) + ... + (6.5) = 38.6$
$T_1 = (8.0) + (10.5) + ... + (9.3) = 42.8$

Then

$$SST = \sum_{i=1}^{3} \frac{T_i^2}{n_i} - CM$$
$$= \frac{(26.7)^2}{5} + \frac{(38.6)^2}{5} + \frac{(42.8)^2}{5} - 779.041 = 27.897$$

Step 4: compute SSE

STEP 4 Compute SSE, the error sum of squares.

Here we utilize the property that the treatment sum of squares plus the error sum of squares equals the total sum of squares.

Hence, SSE = SS Total - SST = 45.349 - 27.897 = 17.45.

Step 5: Compute MST, MSE, and F

STEP 5 Compute MST, MSE and their ratio, *F*.

MST is the mean square of treatments, MSE is the mean square of error (MSE is also frequently denoted by $\hat{\sigma}_\varepsilon^2$).

MST = SST / (k-1) = 27.897 / 2 = 13.949
MSE = SSE / (N-k) = 17.452/ 12 = 1.454

where *N* is the total number of observations and *k* is the number of treatments. Finally, compute *F* as

F = MST / MSE = 9.59

That is it. These numbers are the quantities that are assembled in the ANOVA table that was shown previously.

7.4.3.5. Confidence intervals for the difference of treatment means

Confidence intervals for the difference between two means

This page shows how to construct a confidence interval around (μ_i- μ_j) for the one-way ANOVA by continuing the example shown on a previous page.

Formula for the confidence interval

The formula for a (1- α) 100% confidence interval for the difference between two treatment means is:

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm t_{1-\alpha/2,N-k} \sqrt{\hat{\sigma}_\epsilon^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\hat{\sigma}_\epsilon^2$ = *MSE*.

Computation of the confidence interval for μ_3 - μ_1

For the example, we have the following quantities for the formula:

- $\bar{y}_3$ = 8.56
- $\bar{y}_1$ = 5.34
- $\sqrt{1.454(1/5 + 1/5)} = 0.763$
- $t_{0.975,\ 12}$ = 2.179

Substituting these values yields (8.56 - 5.34) $\pm$ 2.179(0.763) or 3.22 $\pm$ 1.616.

That is, the confidence interval is from 1.604 to 4.836.

Additional 95% confidence intervals

A 95% confidence interval for μ_3 - μ_2 is: from -1.787 to 3.467.

A 95% confidence interval for μ_2 - μ_1 is: from -0.247 to 5.007.

Contrasts

Later on the topic of estimating more general linear

discussed later	combinations of means (primarily contrasts) will be discussed, including how to put confidence bounds around contrasts.

7.4.3.6. Assessing the response from any factor combination

Contrasts

This page treats how to estimate and put confidence bounds around the response to different combinations of factors. Primary focus is on the combinations that are known as *contrasts*. We begin, however, with the simple case of a single factor-level mean.

Estimation of a Factor Level Mean With Confidence Bounds

Estimating factor level means

An unbiased estimator of the factor level mean μ_i in the 1-way ANOVA model is given by:

$$\hat{\mu}_i = \bar{Y}_{i.}$$

where

$$\bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} = \frac{Y_{i.}}{n_i}$$

Variance of the factor level means

The variance of this sample mean estimator is

$$s^2_{\bar{Y}_{i.}} = \frac{MSE}{n_i} = \frac{\hat{\sigma}^2_e}{n_i}$$

Confidence intervals for the factor level means

It can be shown that:

$$t = \frac{\bar{Y}_{i.} - \mu_i}{s_{\bar{Y}_{i.}}}$$

has a t distribution with (N - k) degrees of freedom for the ANOVA model under consideration, where N is the total number of observations and k is the number of factor levels or groups. The degrees of freedom are the same as were used to calculate the MSE in the ANOVA table. That is: *dfe* (degrees of freedom for error) = N - k. From this we can calculate (1- α)100% confidence limits for each μ_i. These are given by:

$$\bar{Y}_{i\cdot} \pm t_{1-\alpha/2,N-k}\sqrt{\frac{\hat{\sigma}_\epsilon^2}{n_i}}$$

Example 1

Example for a 4-level treatment (or 4 different treatments)

The data in the accompanying table resulted from an experiment run in a completely randomized design in which each of four treatments was replicated five times.

						Total	Mean
Group 1	6.9	5.4	5.8	4.6	4.0	26.70	5.34
Group 2	8.3	6.8	7.8	9.2	6.5	38.60	7.72
Group 3	8.0	10.5	8.1	6.9	9.3	42.80	8.56
Group 4	5.8	3.8	6.1	5.6	6.2	27.50	5.50
All Groups						135.60	6.78

1-Way ANOVA table layout

This experiment can be illustrated by the table layout for this 1-way ANOVA experiment shown below:

Level	Sample j						
i	1	2	...	5	Sum	Mean	N
1	Y_{11}	Y_{12}	...	Y_{15}	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$	n_1
2	Y_{21}	Y_{22}	...	Y_{25}	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$	n_2
3	Y_{31}	Y_{32}	...	Y_{35}	$Y_{3\cdot}$	$\bar{Y}_{3\cdot}$	n_3
4	Y_{41}	Y_{42}	...	Y_{45}	$Y_{4\cdot}$	$\bar{Y}_{4\cdot}$	n_4
All					$Y_{\cdot}$	$\bar{Y}_{\cdot\cdot}$	n_t

ANOVA table

The resulting ANOVA table is

Source	SS	DF	MS	F
Treatments	38.820	3	12.940	9.724
Error	21.292	16	1.331	
Total (Corrected)	60.112	19		
Mean	919.368	1		
Total (Raw)	979.480	20		

The estimate for the mean of group 1 is 5.34, and the sample size is $n_1 = 5$.

Computing the confidence interval

Since the confidence interval is two-sided, the entry (1 - $\alpha/2$) value for the t table is (1 - 0.05/2) = 0.975, and the associated degrees of freedom is N - 4, or 20 - 4 = 16.

From the t table in Chapter 1, we obtain $t_{0.975;16} = 2.120$.

Next we need the standard error of the mean for group 1:

$$s^2_{Y_{1.}} = \frac{\text{MSE}}{n_1} = \frac{1.331}{5} = 0.2662$$

$$s_{Y_{1.}} = \sqrt{0.2662} = 0.5159$$

Hence, we obtain confidence limits 5.34 ± 2.120 (0.5159) and the confidence interval is

$$4.246 \leq \mu_1 \leq 6.434$$

Definition and Estimation of Contrasts

Definition of contrasts and orthogonal contrasts

Definitions

A contrast is a linear combination of 2 or more factor level means with coefficients that sum to zero.

Two contrasts are orthogonal if the sum of the products of corresponding coefficients (i.e., coefficients for the same means) adds to zero.

Formally, the definition of a contrast is expressed below, using the notation μ_i for the i-th treatment mean:

$$C = c_1\mu_1 + c_2\mu_2 + \ldots + c_j\mu_j + \ldots + c_k\mu_k$$

where

$$c_1 + c_2 + \ldots + c_j + \ldots + c_k = \sum_{j=1}^{k} c_j = 0$$

Simple contrasts include the case of the difference between two factor means, such as $\mu_1 - \mu_2$. If one wishes to compare treatments 1 and 2 with treatment 3, one way of expressing this is by: $\mu_1 + \mu_2 - 2\mu_3$. Note that

$\mu_1 - \mu_2$ has coefficients +1, -1
$\mu_1 + \mu_2 - 2\mu_3$ has coefficients +1, +1, -2.
These coefficients sum to zero.

An example of orthogonal contrasts

As an example of *orthogonal contrasts*, note the three contrasts defined by the table below, where the rows denote coefficients for the column treatment means.

	μ_1	μ_2	μ_3	μ_4
c_1	+1	0	0	-1
c_2	0	+1	-1	0
c_3	+1	-1	-1	+1

Some properties of orthogonal contrasts

The following is true:

1. The sum of the coefficients for each contrast is zero.
2. The sum of the products of coefficients of each pair of contrasts is also 0 (orthogonality property).
3. The first two contrasts are simply pairwise comparisons, the third one involves all the treatments.

Estimation of contrasts

As might be expected, *contrasts are estimated by taking the same linear combination of treatment mean estimators*. In other words:

$$\hat{C} = \sum_{i=1}^{r} c_i \overline{Y}_{i\cdot}$$

and

$$Var\left(\hat{C}\right) = \sum_{i=1}^{r} c_i^2 Var\left(\overline{Y}_{i\cdot}\right) = \sum_{i=1}^{r} c_i^2 \left(\frac{\sigma^2}{n_i}\right) = \sigma^2 \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

Note: These formulas hold for any linear combination of treatment means, not just for contrasts.

Confidence Interval for a Contrast

Confidence intervals for contrasts

An unbiased estimator for a contrast *C* is given by

$$\hat{C} = \sum_{i=1}^{r} c_i \overline{Y}_{i\cdot}$$

The estimator of $Var(\hat{C})$ is

$$s_{\hat{C}} = \hat{\sigma}_e \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

The estimator $\hat{C}$ is normally distributed because it is a linear combination of independent normal random variables. It can be shown that:

$$\frac{\hat{C} - C}{s_{\hat{C}}}$$

is distributed as t_{N-r} for the one-way ANOVA model under discussion.

Therefore, the 1- α confidence limits for *C* are:

$$\hat{C} \pm t_{1-\alpha/2,N-r}\ s_{\hat{C}}$$

Example 2 (estimating contrast)

Contrast to estimate

We wish to estimate, in our previous example, the following contrast:

$$C = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

and construct a 95 % confidence interval for *C*.

Computing the point estimate and standard error

The point estimate is:

$$\hat{C} = \frac{\bar{Y}_1 + \bar{Y}_2}{2} - \frac{\bar{Y}_3 + \bar{Y}_4}{2} = -0.5$$

Applying the formulas above we obtain

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = 0.2$$

and

$$s_{\hat{C}}^2 = MSE \sum_{i=1}^{4} \frac{c_i^2}{n_i} = 1.331(0.2) = 0.2662$$

and the standard error is $\sqrt{0.2661}$ = 0.5159.

Confidence interval

For a confidence coefficient of 95 % and *df* = 20 - 4 = 16, $t_{0.975,16}$ = 2.12. Therefore, the desired 95 % confidence interval is -0.5 ± 2.12(0.5159) or

(-1.594, 0.594).

Estimation of Linear Combinations

Estimating linear combinations

Sometimes we are interested in a linear combination of the factor-level means that is not a contrast. Assume that in our sample experiment certain costs are associated with each group. For example, there might be costs associated with each factor as follows:

Factor	Cost in $
1	3
2	5
3	2
4	1

The following linear combination might then be of interest:

$$C = 3\mu_1 + 5\mu_2 + 2\mu_3 + 1\mu_4$$

Coefficients do not have to sum to zero for linear combinations

This resembles a contrast, **but the coefficients c_i do not sum to zero**. A linear combination is given by the definition:

$$C = \sum_{i=1}^{r} c_i \mu_i$$

with no restrictions on the coefficients c_i.

Confidence interval identical to contrast

Confidence limits for a linear combination C are obtained in precisely the same way as those for a contrast, using the same calculation for the point estimator and estimated variance.

7.4.3.7. The two-way ANOVA

Definition of a factorial experiment

The 2-way ANOVA is probably the most popular layout in the Design of Experiments. To begin with, let us define a *factorial experiment*:

An experiment that utilizes every combination of factor levels as treatments is called a factorial experiment.

Model for the two-way factorial experiment

In a factorial experiment with factor A at *a* levels and factor B at *b* levels, the model for the general layout can be written as

$$Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$
$$i = 1, 2, \ldots, a;\ j = 1, 2, \ldots, b;\ k = 1, 2, \ldots, r$$

where μ is the overall mean response, τ_i is the effect due to the *i*-th level of factor A, β_j is the effect due to the *j*-th level of factor B and γ_{ij} is the effect due to any interaction between the *i*-th level of A and the *j*-th level of B.

Fixed factors and fixed effects models

At this point, consider the levels of factor A and of factor B chosen for the experiment to be the only levels of interest to the experimenter such as predetermined levels for temperature settings or the length of time for process step. The factors A and B are said to be *fixed factors* and the model is a *fixed-effects model*. Random actors will be discussed later.

When an *a* x *b* factorial experiment is conducted with an equal number of observations per treatment combination, the total (corrected) sum of squares is partitioned as:

SS(total) = SS(A) + SS(B) + SS(AB) + SSE

where AB represents the interaction between A and B.

For reference, the formulas for the sums of squares are:

$$SS(A) = rb\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SS(B) = ra\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SS(AB) = r\sum_{j=1}^{b}\sum_{i=1}^{a}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

$$SSE = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{ij.})^2$$

$$SS(Total) = \sum_{k=1}^{r}\sum_{j=1}^{b}\sum_{i=1}^{a}(y_{ijk} - \bar{y}_{...})^2$$

The breakdown of the total (corrected for the mean) sums of squares

The resulting ANOVA table for an *a* x *b* factorial experiment is

Source	*SS*	*df*	*MS*
Factor A	*SS*(A)	(*a* - 1)	*MS*(A) = *SS*(A)/(*a*-1)
Factor B	*SS*(B)	(*b* - 1)	*MS*(B) = *SS*(B)/(*b*-1)
Interaction AB	*SS*(AB)	(*a*-1)(*b*-1)	*MS*(AB)= *SS*(AB)/(*a*-1)(*b*-1)
Error	*SSE*	(*N* - *ab*)	*SSE*/(*N* - *ab*)
Total (Corrected)	SS(Total)	(*N* - 1)	

The ANOVA table can be used to test hypotheses about the effects and interactions

The various hypotheses that can be tested using this ANOVA table concern whether the different levels of Factor A, or Factor B, really make a difference in the response, and whether the AB interaction is significant (see previous discussion of ANOVA hypotheses).

7.4.3.8. Models and calculations for the two-way ANOVA

Basic Layout

The balanced 2-way factorial layout

Factor A has 1, 2, ..., *a* levels. Factor B has 1, 2, ..., *b* levels. There are *ab* treatment combinations (or cells) in a complete factorial layout. Assume that each treatment cell has *r* independent obsevations (known as replications). When each cell has the same number of replications, the design is a *balanced factorial*. In this case, the *abr* data points $\{y_{ijk}\}$ can be shown pictorially as follows:

		Factor B			
		1	2	...	b
Factor A	1	$y_{111}, y_{112}, ..., y_{11r}$	$y_{121}, y_{122}, ..., y_{12r}$	...	$y_{1b1}, y_{1b2}, ..., y_{1br}$
	2	$y_{211}, y_{212}, ..., y_{21r}$	$y_{221}, y_{222}, ..., y_{22r}$	...	$y_{2b1}, y_{2b2}, ..., y_{2br}$
	.	...	....		...
	.				
	a	$y_{a11}, y_{a12}, ..., y_{a1r}$	$y_{a21}, y_{a22}, ..., y_{a2r}$	...	$y_{ab1}, y_{ab2}, ..., y_{abr}$

How to obtain sums of squares for the balanced factorial layout

Next, we will calculate the sums of squares needed for the ANOVA table.

- Let A_i be the sum of all observations of level *i* of factor A, *i* = 1, ... ,*a*. The A_i are the row sums.
- Let B_j be the sum of all observations of level *j* of factor B, *j* = 1, ...,*b*. The B_j are the column sums.
- Let $(AB)_{ij}$ be the sum of all observations of level *i* of A and level *j* of B. These are cell sums.
- Let *r* be the number of replicates in the experiment; that is: the number of times each factorial treatment combination appears in the experiment.

Then the total number of observations for each level of factor A is *rb* and the total number of observations for each level of factor B is *ra* and the total number of observations for each interaction is *r*.

Finally, the total number of observations *n* in the experiment is *abr*.

With the help of these expressions we arrive (omitting derivations) at

$$CM = \frac{(\text{Sum of all observations})^2}{rab}$$

$$SS_{total} = \sum(\text{each observation})^2 - CM$$

$$SS(A) = \frac{\sum_{i=1}^{a} A_i^2}{rb} - CM$$

$$SS(B) = \frac{\sum_{i=1}^{b} B_i^2}{ra} - CM$$

$$SS(AB) = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(AB)_{ij}^2}{r} - CM - SS(A) - SS(B)$$

$$SSE = SS_{total} - SS(A) - S(B) - SS(AB)$$

These expressions are used to calculate the ANOVA table entries for the (fixed effects) 2-way ANOVA.

Two-Way ANOVA Example:

Data

An evaluation of a new coating applied to 3 different materials was conducted at 2 different laboratories. Each laboratory tested 3 samples from each of the treated materials. The results are given in the next table:

	Materials (B)		
LABS (A)	1	2	3
1	4.1	3.1	3.5
	3.9	2.8	3.2
	4.3	3.3	3.6
2	2.7	1.9	2.7
	3.1	2.2	2.3
	2.6	2.3	2.5

Row and column sums

The preliminary part of the analysis yields a table of row and column sums.

Material (B)

Lab (A)	1	2	3	Total (A_i)
1	12.3	9.2	10.3	31.8
2	8.4	6.4	7.5	22.3
Total (B_j)	20.7	15.6	17.8	54.1

ANOVA table From this table we generate the ANOVA table.

Source	*SS*	*df*	*MS*	*F*	*p*-value
A	5.0139	1	5.0139	100.28	0
B	2.1811	2	1.0906	21.81	.0001
AB	0.1344	2	0.0672	1.34	.298
Error	0.6000	12	0.0500		
Total (Corr)	7.9294	17			

7.4.4. What are variance components?

Fixed and Random Factors and Components of Variance

A fixed level of a factor or variable means that the levels in the experiment are the only ones we are interested in

In the previous example, the levels of the factor temperature were considered as *fixed*; that is, the three temperatures were the only ones that we were interested in (this may sound somewhat unlikely, but let us accept it without opposition). The model employed for fixed levels is called a *fixed model*. When the levels of a factor are random, such as operators, days, lots or batches, where the levels in the experiment might have been chosen at *random* from a large number of possible levels, the model is called a *random model*, and inferences are to be extended to all levels of the population.

Random levels are chosen at random from a large or infinite set of levels

In a random model the experimenter is often interested in estimating *components of variance*. Let us run an example that analyzes and interprets a component of variance or random model.

Components of Variance Example for Random Factors

Data for the example

A company supplies a customer with a larger number of batches of raw materials. The customer makes three sample determinations from each of 5 randomly selected batches to control the quality of the incoming material. The model is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

and the k levels (e.g., the batches) are chosen at random from a population with variance σ_τ. The data are shown below

Batch				
1	2	3	4	5
74	68	75	72	79

76	71	77	74	81
75	72	77	73	79

ANOVA table for example

A 1-way ANOVA is performed on the data with the following results:

ANOVA

Source	*SS*	*df*	*MS*	*EMS*
Treatment (batches)	147.74	4	36.935	$\sigma_\epsilon^2 + 3\sigma_\tau^2$
Error	17.99	10	1.799	σ_ϵ^2
Total (corrected)	165.73	14		

Interpretation of the ANOVA table

The computations that produce the SS are the same for both the fixed and the random effects model. For the random model, however, the treatment sum of squares, SST, is an estimate of $\{\sigma_\epsilon^2 + 3\sigma_\tau^2\}$. This is shown in the EMS (Expected Mean Squares) column of the ANOVA table.

The test statistic from the ANOVA table is F = 36.94 / 1.80 = 20.5.

If we had chosen an α value of .01, then the F value from the table in Chapter 1 for a *df* of 4 in the numerator and 10 in the denominator is 5.99.

Method of moments

Since the test statistic is larger than the critical value, we reject the hypothesis of equal means. Since these batches were chosen via a random selection process, it may be of interest to find out how much of the variance in the experiment might be attributed to batch diferences and how much to random error. In order to answer these questions, we can use the EMS column. The estimate of σ_ϵ^2 is 1.80 and the computed treatment mean square of 36.94 is an estimate of $\sigma_\epsilon^2 + 3\sigma_\tau^2$. Setting the MS values equal to the EMS values (this is called the *Method of Moments*), we obtain

$$s_\epsilon^2 = 1.80 \quad \text{and} \quad s_\epsilon^2 + 3s_\tau^2 = 36.94$$

where we use s^2 since these are estimators of the corresponding σ^2's.

Computation of the

Solving these expressions

components of variance

$$s_\tau^2 = \frac{36.94 - 1.80}{3} = 11.71$$

The total variance can be estimated as

$$s_{total} = s_\tau^2 + s_\epsilon^2 = 11.71 + 1.80 = 13.51$$

Interpretation

In terms of percentages, we see that 11.71/13.51 = 86.7 percent of the total variance is attributable to batch differences and 13.3 percent to error variability within the batches.

NIST SEMATECH HOME TOOLS & AIDS SEARCH BACK NEXT

7.4.5. How can we compare the results of classifying according to several categories?

Contingency Table approach

When items are classified according to two or more criteria, it is often of interest to decide whether these criteria act independently of one another.

For example, suppose we wish to classify defects found in wafers produced in a manufacturing plant, first according to the type of defect and, second, according to the production shift during which the wafers were produced. If the proportions of the various types of defects are constant from shift to shift, then classification by defects is independent of the classification by production shift. On the other hand, if the proportions of the various defects vary from shift to shift, then the classification by defects depends upon or is *contingent* upon the shift classification and the classifications are dependent.

In the process of investigating whether one method of classification is contingent upon another, it is customary to display the data by using a cross classification in an array consisting of r rows and c columns called a ***contingency table***. A contingency table consists of r x c cells representing the r x c possible outcomes in the classification process. Let us construct an industrial case:

Industrial example

A total of 309 wafer defects were recorded and the defects were classified as being one of four types, *A*, *B*, *C*, or *D*. At the same time each wafer was identified according to the production shift in which it was manufactured, 1, 2, or 3.

Contingency table classifying defects in wafers according to type and production shift

These counts are presented in the following table.

	Type of Defects				
Shift	A	B	C	D	Total
1	15(22.51)	21(20.99)	45(38.94)	13(11.56)	94
2	26(22.9)	31(21.44)	34(39.77)	5(11.81)	96
3	33(28.50)	17(26.57)	49(49.29)	20(14.63)	119
Total	74	69	128	38	309

(Note: the numbers in parentheses are the expected cell frequencies).

Column probabilities

Let p_A be the probability that a defect will be of type A. Likewise, define p_B, p_C, and p_D as the probabilities of observing the other three types of defects. These probabilities, which are called the ***column probabilities***, will satisfy the requirement

$$p_A + p_B + p_C + p_D = 1$$

Row probabilities

By the same token, let p_i (i=1, 2, or 3) be the **row probability** that a defect will have occurred during shift i, where

$$p_1 + p_2 + p_3 = 1$$

Multiplicative Law of Probability

Then if the two classifications are independent of each other, a cell probability will equal the product of its respective row and column probabilities in accordance with the Multiplicative Law of Probability.

Example of obtaining column and row probabilities

For example, the probability that a particular defect will occur in shift 1 and is of type A is (p_1) (p_A). While the numerical values of the cell probabilities are unspecified, the null hypothesis states that each cell probability will equal the product of its respective row and column probabilities. This condition implies independence of the two classifications. The alternative hypothesis is that this equality does not hold for at least one cell.

In other words, we state the null hypothesis as H_0: the two classifications are independent, while the alternative hypothesis is H_a: the classifications are dependent.

To obtain the observed column probability, divide the column total by the grand total, n. Denoting the total of column j as c_j, we get

$$\hat{p}_A = \frac{c_1}{n} = \frac{74}{309} \qquad \hat{p}_C = \frac{c_3}{n} = \frac{128}{309}$$
$$\hat{p}_B = \frac{c_2}{n} = \frac{69}{309} \qquad \hat{p}_D = \frac{c_4}{n} = \frac{38}{309}$$

Similarly, the row probabilities p_1, p_2, and p_3 are estimated by dividing the row totals r_1, r_2, and r_3 by the grand total n, respectively

$$\hat{p}_1 = \frac{r_1}{n} = \frac{94}{309} \quad \hat{p}_2 = \frac{r_2}{n} = \frac{96}{309} \quad \hat{p}_3 = \frac{r_3}{n} = \frac{119}{309}$$

Expected cell frequencies

Denote the observed frequency of the cell in row i and column $j$of the contingency table by n_{ij}. Then we have

$$\hat{E}(n_{ij}) = n(\hat{p}_i \hat{p}_j) = n\left(\frac{r_i}{n}\right)\left(\frac{c_j}{n}\right) = \frac{r_i c_j}{n}$$

Estimated expected cell frequency when H_0 is true.

In other words, when the row and column classifications are independent, the estimated expected value of the observed cell frequency n_{ij} in an r x c contingency table is equal to its respective row and column totals divided by the total frequency.

$$\hat{E}(n_{ij}) = \frac{r_i c_j}{n}$$

The estimated cell frequencies are shown in parentheses in the contingency table above.

Test statistic

From here we use the expected and observed frequencies shown in the table to calculate the value of the test statistic

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{4} \frac{[n_{ij} - \hat{E}(n_{ij})]^2}{\hat{E}(n_{ij})}$$

$$\chi^2 = \frac{(15 - 22.51)^2}{22.51} + \frac{(26 - 22.99)^2}{22.99} + \cdots + \frac{(20 - 14.63)^2}{14.63} = 19.18$$

df = (r-1)(c-1)

The next step is to find the appropriate number of degrees of freedom associated with the test statistic. Leaving out the details of the derivation, we state the result:

> *The number of degrees of freedom associated with a contingency table consisting of r rows and c columns is (r-1) (c-1).*

So for our example we have (3-1) (4-1) = 6 d.f.

Testing the null hypothesis

In order to test the null hypothesis, we compare the test statistic with the critical value of $X^2_{1-\alpha/2}$ at a selected value of α. Let us use α = 0.05. Then the critical value is $X^2_{0.95,6}$ = 12.5916 (see the chi square table in Chapter 1). Since the test statistic of 19.18 exceeds the critical value, we reject the null hypothesis and conclude that there is significant evidence that the proportions of the different defect types vary from shift to shift. In this case, the *p*-value of the test statistic is 0.00387.

7.4.6. Do all the processes have the same proportion of defects?

The contingency table approach

Testing for homogeneity of proportions using the chi-square distribution via contingency tables

When we have samples from n populations (i.e., lots, vendors, production runs, etc.), we can test whether there are significant differences in the proportion defectives for these populations using a contingency table approach. The contingency table we construct has two rows and n columns.

To test the null hypothesis of no difference in the proportions among the n populations

$$H_0: \ p_1 = p_2 = ... = p_n$$

against the alternative that not all n population proportions are equal

$$H_1: \text{Not all } p_i \text{ are equal } (i = 1, 2, ..., n)$$

The chi-square test statistic

we use the following test statistic:

$$\chi^2 = \sum_{\text{all cells}} \frac{(f_o - f_c)^2}{f_c}$$

where f_o is the observed frequency in a given cell of a 2 x n contingency table, and f_c is the theoretical count or expected frequency in a given cell if the null hypothesis were true.

The critical value

The critical value is obtained from the χ^2 distribution table with degrees of freedom (2-1)(n-1) = n-1, at a given level of significance.

An illustrative example

Data for the example

Diodes used on a printed circuit board are produced in lots of size 4000. To study the homogeneity of lots with

respect to a demanding specification, we take random samples of size 300 from 5 consecutive lots and test the diodes. The results are:

	Lot					
Results	1	2	3	4	5	Totals
Nonconforming	36	46	42	63	38	225
Conforming	264	254	258	237	262	1275
Totals	300	300	300	300	300	1500

Computation of the overall proportion of nonconforming units

Assuming the null hypothesis is true, we can estimate the single overall proportion of nonconforming diodes by pooling the results of all the samples as

$$\bar{p} = \frac{(36+46+42+63+38)}{(5 \text{ x } 300)} = 225/1500 = .15$$

Computation of the overall proportion of conforming units

We estimate the proportion of conforming ("good") diodes by the complement 1 - 0.15 = 0.85. Multiplying these two proportions by the sample sizes used for each lot results in the expected frequencies of nonconforming and conforming diodes. These are presented below:

Table of expected frequencies

	Lot					
Results	1	2	3	4	5	Totals
Nonconforming	45	45	45	45	45	225
Conforming	255	255	255	255	255	1275
Totals	300	300	300	300	300	1500

Null and alternate hypotheses

To test the null hypothesis of homogeneity or equality of proportions

H_0: $p_1 = p_2 = ... = p_5$

against the alternative that not all 5 population proportions are equal

H_1: Not all p_i are equal (i = 1, 2, ...,5)

Table for computing the test statistic

we use the observed and expected values from the tables above to compute the χ^2 test statistic. The calculations are presented below:

2 2

f_o	f_c	$(f_o - f_c)$	$(f_o - f_c)$	$(f_o - f_c)\ /\ f_c$
36	45	-9	81	1.800
46	45	1	1	0.022
42	45	-3	9	0.200
63	45	18	324	7.200
38	45	-7	49	1.089
264	225	9	81	0.318
254	255	-1	1	0.004
258	255	3	9	0.035
237	255	-18	324	1.271
262	255	7	49	0.192
				12.131

Conclusions If we choose a .05 level of significance, the critical value of χ^2 with 4 degrees of freedom is 9.488 (see the chi square distribution table in Chapter 1). Since the test statistic (12.131) exceeds this critical value, we reject the null hypothesis.

7.4.7. How can we make multiple comparisons?

What to do after equality of means is rejected

When processes are compared and the null hypothesis of equality (or homogeneity) is rejected, all we know at that point is that there is no equality amongst them. But we do not know the form of the inequality.

Typical questions

Questions concerning the reason for the rejection of the null hypothesis arise in the form of:

- "Which mean(s) or proportion (s) differ from a standard or from each other?"
- "Does the mean of treatment 1 differ from that of treatment 2?"
- "Does the average of treatments 1 and 2 differ from the average of treatments 3 and 4?"

Multiple Comparison test procedures are needed

One popular way to investigate the cause of rejection of the null hypothesis is a *Multiple Comparison Procedure.* These are methods which examine or compare more than one pair of means or proportions at the same time.

Note: Doing pairwise comparison procedures over and over again for all possible pairs will not, in general, work. This is because the overall significance level is not as specified for a single pair comparison.

ANOVA F test is a preliminary test

The ANOVA uses the F test to determine whether there exists a significant difference among treatment means or interactions. In this sense it is a preliminary test that informs us if we should continue the investigation of the data at hand.

If the null hypothesis (no difference among treatments or interactions) is accepted, there is an implication that no relation exists between the factor levels and the response. There is not much we can learn, and we are finished with the analysis.

When the F test rejects the null hypothesis, we usually want to undertake a thorough analysis of the nature of the factor-

level effects.

Procedures for examining factor-level effects

Previously, we discussed several procedures for examining particular factor-level effects. These were

- Estimation of the Difference Between Two Factor Means
- Estimation of Factor Level Effects
- Confidence Intervals For A Contrast

Determine contrasts in advance of observing the experimental results

These types of investigations should be done on combinations of factors that were determined in advance of observing the experimental results, or else the confidence levels are not as specified by the procedure. Also, doing several comparisons might change the overall confidence level (see note above). This can be avoided by carefully selecting contrasts to investigate in advance and making sure that:

- the number of such contrasts does not exceed the number of degrees of freedom between the treatments
- only orthogonal contrasts are chosen.

However, there are also several powerful multiple comparison procedures we can use after observing the experimental results.

Tests on Means after Experimentation

Procedures for performing multiple comparisons

If the decision on what comparisons to make is withheld until after the data are examined, the following procedures can be used:

- Tukey's Method *to test all possible pairwise differences of means to determine if at least one difference is significantly different from 0.*
- Scheffé's Method *to test all possible contrasts at the same time, to see if at least one is significantly different from 0.*
- Bonferroni Method *to test, or put simultaneous confidence intervals around, a pre-selected group of contrasts*

Multiple Comparisons Between Proportions

Procedure for proportion defective data

When we are dealing with population proportion defective data, the Marascuilo procedure can be used to simultaneously examine comparisons between all groups after the data have been collected.

7.4.7.1. Tukey's method

Tukey's method considers all possible pairwise differences of means at the same time

The Tukey method applies simultaneously to the set of all pairwise comparisons

$$\{\mu_i - \mu_j\}$$

The confidence coefficient for the set, when all sample sizes are equal, is exactly 1-α. For unequal sample sizes, the confidence coefficient is greater than 1-α. In other words, the Tukey method is conservative when there are unequal sample sizes.

Studentized Range Distribution

The studentized range q

The Tukey method uses the *studentized range distribution*. Suppose we have r independent observations $y_1, ..., y_r$ from a normal distribution with mean μ and variance σ^2. Let w be the range for this set , i.e., the maximum minus the minimum. Now suppose that we have an estimate s^2 of the variance σ^2 which is based on ν degrees of freedom and is independent of the y_i. The studentized range is defined as

$$q_{r,\nu} = w/s$$

The distribution of q is tabulated in many textbooks and can be calculated using Dataplot

The distribution of q has been tabulated and appears in many textbooks on statistics. In addition, Dataplot has a CDF function (SRACDF) and a percentile function (SRAPPF) for q.

As an example, let r = 5 and ν = 10. The 95th percentile is $q_{.05;5,10}$ = 4.65. This means:

$$P\left\{\frac{w}{s} \le 4.65\right\} = .95$$

So, if we have five observations from a normal distribution, the probability is .95 that their range is not more than 4.65 times as great as an independent sample standard deviation estimate for which the estimator has 10 degrees of freedom.

Tukey's Method

Confidence limits for Tukey's method

The Tukey confidence limits for all pairwise comparisons with confidence coefficient of at least 1- α are:

$$\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm \frac{1}{\sqrt{2}} q_{\alpha;r,N-r} \hat{\sigma}_\epsilon \sqrt{\frac{2}{n}} \qquad i,j = 1,\ldots,r; i \neq j$$

Notice that the point estimator and the estimated variance are the same as those for a single pairwise comparison that was illustrated previously. The only difference between the confidence limits for simultaneous comparisons and those for a single comparison is the multiple of the estimated standard deviation.

Also note that the sample sizes must be equal when using the studentized range approach.

Example

Data

We use the data from a previous example.

Set of all pairwise comparisons

The set of all pairwise comparisons consists of:

μ_2 - μ_1, μ_3 - μ_1, μ_1 - μ_4,
μ_2 - μ_3, μ_2 - μ_4, μ_3 - μ_4

Confidence intervals for each pair

Assume we want a confidence coefficient of 95 percent, or .95. Since r = 4 and n_t = 20, the required percentile of the studentized range distribution is $q_{.05;\,4,16}$. Using the Tukey method for each of the six comparisons yields:

$$0.29 \leq \mu_2 - \mu_1 \leq 4.47$$

$$1.13 \leq \mu_3 - \mu_1 \leq 5.31$$

$$-2.25 \leq \mu_1 - \mu_4 \leq 1.93$$

$$-2.93 \leq \mu_2 - \mu_3 \leq 1.25$$

$$0.13 \leq \mu_2 - \mu_4 \leq 4.31$$

$$0.97 \leq \mu_3 - \mu_4 \leq 5.15$$

Conclusions

The simultaneous pairwise comparisons indicate that the differences μ_1 - μ_4 and μ_2 - μ_3 are not significantly different from 0 (their confidence intervals include 0), and all the other pairs are significantly different.

Unequal sample sizes

It is possible to work with unequal sample sizes. In this case, one has to calculate the estimated standard deviation for each

pairwise comparison. The Tukey procedure for unequal sample sizes is sometimes referred to as the *Tukey-Kramer Method*.

7.4.7.2. Scheffe's method

Scheffe's method tests all possible contrasts at the same time

Scheffé's method applies to the set of estimates of all possible contrasts among the factor level means, not just the pairwise differences considered by Tukey's method.

Definition of contrast

An arbitrary contrast is defined by

$$C = \sum_{i=1}^{r} c_i \mu_i$$

where

$$\sum_{i=1}^{r} c_i = 0$$

Infinite number of contrasts

Technically there is an infinite number of contrasts. The simultaneous confidence coefficient is exactly 1- α, whether the factor level sample sizes are equal or unequal.

Estimate and variance for C

As was described earlier, we estimate C by:

$$\hat{C} = \sum_{i=1}^{r} c_i \bar{Y}_{i\cdot}$$

for which the estimated variance is:

$$s^2_{\hat{C}} = \hat{\sigma}^2_e \sum_{i=1}^{r} \frac{c_i^2}{n_i}$$

Simultaneous confidence interval

It can be shown that the probability is 1 - α that all confidence limits of the type

$$\hat{C} \pm \sqrt{(r-1) F_{\alpha;r-1,N-r}}\ s_{\hat{C}}$$

are correct simultaneously.

Scheffe method example

Contrasts to estimate

We wish to estimate, in our previous experiment, the following contrasts

$$C_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$C_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

and construct 95 percent confidence intervals for them.

Compute the point estimates of the individual contrasts

The point estimates are:

$$\hat{C}_1 = \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} = -0.5$$
$$\hat{C}_2 = \frac{\overline{Y}_1 + \overline{Y}_3}{2} - \frac{\overline{Y}_2 + \overline{Y}_4}{2} = .34$$

Compute the point estimate and variance of C

Applying the formulas above we obtain in both cases:

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = .2$$

and

$$s^2_{\hat{C}} = \hat{\sigma}^2_\epsilon \sum_{i=1}^{4} \frac{c_i^2}{4} = 1.331(.2) = .2661$$

where σ^2_ϵ = 1.331 was computed in our previous example. The standard error = .5158 (square root of .2661).

Scheffe confidence interval

For a confidence coefficient of 95 percent and degrees of freedom in the numerator of r - 1 = 4 - 1 = 3, and in the denominator of 20 - 4 = 16, we have:

$$\sqrt{(r-1)F_{\alpha;r-1;N-r}} = \sqrt{3F_{.05;3;16}} = 3.12$$

The confidence limits for C_1 are -.5 ± 3.12(.5158) = -.5 ± 1.608, and for C_2 they are .34 ± 1.608.

The desired simultaneous 95 percent confidence intervals are

$$-2.108 \leq C_1 \leq 1.108$$

$$-1.268 \leq C_2 \leq 1.948$$

Comparison to confidence interval for a single contrast

Recall that when we constructed a confidence interval for a single contrast, we found the 95 percent confidence interval:

$$-1.594 \leq C \leq 0.594$$

As expected, the Scheffé confidence interval procedure that generates simultaneous intervals for all contrasts is considerabley wider.

Comparison of Scheffé's Method with Tukey's Method

Tukey preferred when only pairwise comparisons are of interest

If only pairwise comparisons are to be made, the Tukey method will result in a narrower confidence limit, which is preferable.

Consider for example the comparison between μ_3 and μ_1.

Tukey: $1.13 < \mu_3 - \mu_1 < 5.31$

Scheffé: $0.95 < \mu_3 - \mu_1 < 5.49$

which gives Tukey's method the edge.

The normalized contrast, using sums, for the Scheffé method is 4.413, which is close to the maximum contrast.

Scheffe preferred when many contrasts are of interest

In the general case when many or all contrasts might be of interest, the Scheffé method tends to give narrower confidence limits and is therefore the preferred method.

7.4.7.3. Bonferroni's method

Simple method

The Bonferroni method is a simple method that allows many comparison statements to be made (or confidence intervals to be constructed) while still assuring an overall confidence coefficient is maintained.

Applies for a finite number of contrasts

This method applies to an ANOVA situation when the analyst has picked out a particular set of pairwise comparisons or contrasts or linear combinations in advance. This set is not infinite, as in the Scheffé case, but may exceed the set of pairwise comparisons specified in the Tukey procedure.

Valid for both equal and unequal sample sizes

The Bonferroni method is valid for equal and unequal sample sizes. We restrict ourselves to only linear combinations or comparisons of treatment level means (pairwise comparisons and contrasts are special cases of linear combinations). We denote the number of statements or comparisons in the finite set by *g*.

Bonferroni general inequality

Formally, the Bonferroni general inequality is presented by:

$$P\left(\bigcap_{i=1}^{g} A_i\right) \geq 1 - \sum_{i=1}^{g} P[\bar{A}_i]$$

where A_i and its complement $\bar{A}_i$ are any events.

Interpretation of Bonferroni inequality

In particular, if each A_i is the event that a calculated confidence interval for a particular linear combination of treatments includes the true value of that combination, then the left-hand side of the inequality is the probability that all the confidence intervals simultaneously cover their respective true values. The right-hand side is one minus the sum of the probabilities of each of the intervals missing their true values. Therefore, if simultaneous multiple interval estimates are desired with an overall confidence coefficient 1-α, *one can construct each interval with confidence coefficient (1-*α*/g), and the Bonferroni inequality insures that the overall confidence coefficient is at least 1-*α.

Formula for Bonferroni confidence interval

In summary, the Bonferroni method states that the confidence coefficient is at least 1- α that simultaneously all the following confidence limits for the *g* linear combinations C_i are "correct" (or capture their respective true values):

$$\hat{C}_i \pm t_{1-\alpha/(2g),N-r}\ s_{\hat{C}_i}$$

where

$$s_{\hat{C}_i} = \hat{\sigma}_\epsilon \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}}$$

Example using Bonferroni method

Contrasts to estimate

We wish to estimate, as we did using the Scheffe method, the following linear combinations (contrasts):

$$C_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$
$$C_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2}$$

and construct 95 % confidence intervals around the estimates.

Compute the point estimates of the individual contrasts

The point estimates are:

$$\hat{C}_1 = \frac{\overline{Y}_1 + \overline{Y}_2}{2} - \frac{\overline{Y}_3 + \overline{Y}_4}{2} = -0.5$$
$$\hat{C}_2 = \frac{\overline{Y}_1 + \overline{Y}_3}{2} - \frac{\overline{Y}_2 + \overline{Y}_4}{2} = .34$$

Compute the point estimate and variance of C

As before, for both contrasts, we have

$$\sum_{i=1}^{4} \frac{c_i^2}{n_i} = \frac{4(1/2)^2}{5} = .2$$

and

$$s_{\hat{C}}^2 = \hat{\sigma}_\epsilon^2 \sum_{i=1}^{4} \frac{c_i^2}{4} = 1.331(.2) = .2661$$

where σ_ϵ^2 = 1.331 was computed in our previous example. The standard error is .5158 (the square root of .2661).

Compute the Bonferroni simultaneous confidence interval

For a 95 % overall confidence coefficient using the Bonferroni method, the t value is $t_{1-0.05/(2*2),16} = t_{0.9875,16}$ = 2.473 (from the t table in Chapter 1). Now we can calculate the confidence intervals for the two contrasts. For C_1 we have confidence limits -0.5 ± 2.473 (.5158) and for C_2 we have confidence limits 0.34 ± 2.473 (0.5158).

Thus, the confidence intervals are:

$$-1.776 \leq C_1 \leq 0.776$$
$$-0.936 \leq C_2 \leq 1.616$$

Comparison to Scheffe interval

Notice that the Scheffé interval for C_1 is:

$$-2.108 \leq C_1 \leq 1.108$$

which is wider and therefore less attractive.

Comparison of Bonferroni Method with Scheffé and Tukey Methods

No one comparison method is uniformly best - each has its uses

1. If all pairwise comparisons are of interest, Tukey has the edge. If only a subset of pairwise comparisons are required, Bonferroni may sometimes be better.
2. When the number of contrasts to be estimated is small, (about as many as there are factors) Bonferroni is better than Scheffé. Actually, unless the number of desired contrasts is at least twice the number of factors, Scheffé will always show wider confidence bands than Bonferroni.
3. Many computer packages include all three methods. So, study the output and select the method with the smallest confidence band.
4. No single method of multiple comparisons is uniformly best among all the methods.

7.4.7.4. Comparing multiple proportions: The Marascuillo procedure

Testing for equal proportions of defects

Earlier, we discussed how to test whether several populations have the same proportion of defects. The example given there led to rejection of the null hypothesis of equality.

Marascuilo procedure allows comparison of all possible pairs of proportions

Rejecting the null hypothesis only allows us to conclude that not (in this case) all lots are equal with respect to the proportion of defectives. However, it does not tell us which lot or lots caused the rejection.

The Marascuilo procedure enables us to simultaneously test the differences of all pairs of proportions when there are several populations under investigation.

The Marascuillo Procedure

Step 1: compute differences p_i - p_j

Assume we have samples of size n_i (i = 1, 2, ..., k) from k populations. The first step of this procedure is to compute the differences p_i - p_j, (where i is not equal to j) among all $k(k\text{-}1)/2$ pairs of proportions.

The absolute values of these differences are the test-statistics.

Step 2: compute test statistics

Step 2 is to pick a significance level and compute the corresponding critical values for the Marascuilo procedure from

$$r_{ij} = \sqrt{\chi^2_{1-\alpha,k-1}} \sqrt{\frac{p_i(1-p_i)}{n_i} + \frac{p_j(1-p_j)}{n_j}}$$

Step 3: compare test statistics against corresponding critical values

The third and last step is to compare each of the $k(k\text{-}1)/2$ test statistics against its corresponding critical r_{ij} value. Those pairs that have a test statistic that exceeds the critical value are significant at the α level.

Example

Sample proportions

To illustrate the Marascuillo procedure, we use the data from the previous example. Since there were 5 lots, there are (5 x 4)/2 = 10 possible pairwise comparisons to be made and ten critical ranges to compute. The five sample proportions are:

$$p_1 = 36/300 = .120$$
$$p_2 = 46/300 = .153$$
$$p_3 = 42/300 = .140$$
$$p_4 = 63/300 = .210$$
$$p_5 = 38/300 = .127$$

Table of critical values

For an overall level of significance of 0.05, the critical value of the chi-square distribution having four degrees of freedom is $X^2_{0.95,4} = 9.488$ and the square root of 9.488 is 3.080. Calculating the 10 absolute differences and the 10 critical values leads to the following summary table.

contrast	value	critical range	significant
$\|p_1 - p_2\|$	.033	0.086	no
$\|p_1 - p_3\|$	.020	0.085	no
$\|p_1 - p_4\|$	.090	0.093	no
$\|p_1 - p_5\|$	.007	0.083	no
$\|p_2 - p_3\|$	.013	0.089	no
$\|p_2 - p_4\|$	.057	0.097	no
$\|p_2 - p_5\|$	.026	0.087	no
$\|p_3 - p_4\|$	.070	0.095	no
$\|p_3 - p_5\|$	.013	0.086	no
$\|p_4 - p_5\|$	.083	0.094	no

The table of critical values can be generated using both Dataplot code and R code.

No individual contrast is statistically significant

A difference is statistically significant if its value exceeds the critical range value. In this example, even though the null hypothesis of equality was rejected earlier, there is not enough data to conclude any particular difference is significant. Note, however, that all the comparisons involving population 4 come the closest to significance - leading us to suspect that more data might actually show that population 4 does have a significantly higher proportion of defects.

7.5. References

Primary References

Agresti, A. and Coull, B. A. (1998). *Approximate is better than "exact" for interval estimation of binomial proportions*", The American Statistician, 52(2), 119-126.

Berenson M.L. and Levine D.M. (1996) *Basic Business Statistics*, Prentice-Hall, Englewood Cliffs, New Jersey.

Bhattacharyya, G. K., and R. A. Johnson, (1997). *Statistical Concepts and Methods*, John Wiley and Sons, New York.

Birnbaum, Z. W. (1952). "*Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size*", Journal of the American Statistical Association, 47, page 425.

Brown, L. D. Cai, T. T. and DasGupta, A. (2001). *Interval estimation for a binomial proportion*", Statistical Science, 16(2), 101-133.

Diamond, W. J. (1989). *Practical Experiment Designs*, Van-Nostrand Reinhold, New York.

Dixon, W. J. and Massey, F.J. (1969). *Introduction to Statistical Analysis*, McGraw-Hill, New York.

Draper, N. and Smith, H., (1981). *Applied Regression Analysis*, John Wiley & Sons, New York.

Fliess, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, Third Edition, John Wiley & Sons, New York.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners*, John Wiley & Sons, New York.

Hicks, C. R. (1973). *Fundamental Concepts in the Design of Experiments*, Holt, Rinehart and Winston, New York.

Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*, John Wiley & Sons, New York.

Howe, W. G. (1969). *"Two-sided Tolerance Limits for Normal Populations - Some Improvements"*, Journal of the Americal Statistical Association, 64 , pages 610-620.

Kendall, M. and Stuart, A. (1979). *The Advanced Theory of*

Statistics, Volume 2: Inference and Relationship. Charles Griffin & Co. Limited, London.

Mendenhall, W., Reinmuth, J. E. and Beaver, R. J. *Statistics for Management and Economics*, Duxbury Press, Belmont, CA.

Montgomery, D. C. (1991). *Design and Analysis of Experiments*, John Wiley & Sons, New York.

Moore, D. S. (1986). "*Tests of Chi-Square Type*". From Goodness-of-Fit Techniques (D'Agostino & Stephens eds.).

Myers, R. H., (1990). *Classical and Modern Regression with Applications*, PWS-Kent, Boston, MA.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models*, 3rd Edition, Irwin, Boston, MA.

Lawless, J. F., (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Pearson, A. V., and Hartley, H. O. (1972). *Biometrica Tables for Statisticians, Vol 2*, Cambridge, England, Cambridge University Press.

Sarhan, A. E. and Greenberg, B. G. (1956). "*Estimation of location and scale parameters by order statistics from singly and double censored samples*," Part I, Annals of Mathematical Statistics, 27, 427-451.

Searle, S. S., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, John Wiley & Sons, New York.

Siegel, S. (1956). *Nonparametric Statistics*, McGraw-Hill, New York.

Shapiro, S. S. and Wilk, M. B. (1965). "*An analysis of variance test for normality (complete samples)*", Biometrika, 52, 3 and 4, pages 591-611.

Some Additional References and Bibliography

Books

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-FitTechniques*, Marcel Dekker, Inc., New York.

Hicks, C. R. 1973. *Fundamental Concepts in the Design of Experiments*. Holt, Rhinehart and Winston,New-York

Miller, R. G., Jr. (1981). *Simultaneous Statistical Inference*, Springer-Verlag, New York.

Neter, Wasserman, and Whitmore (1993). *Applied Statistics*, 4th Edition, Allyn and Bacon, Boston, MA.

Neter, J., Wasserman, W. and Kutner, M. H. (1990). *Applied Linear Statistical Models,* 3rd Edition, Irwin, Boston, MA.

Scheffe, H. (1959). *The Analysis of Variance*, John Wiley, New-York.

Articles

Begun, J. M. and Gabriel, K. R. (1981). "*Closure of the Newman-Keuls Multiple Comparisons Procedure*", Journal of the American Statistical Association, 76, page 374.

Carmer, S. G. and Swanson, M. R. (1973. *"Evaluation of Ten Pairwise Multiple Comparison Procedures by Monte-Carlo Methods"*, Journal of the American Statistical Association, 68, pages 66-74.

Duncan, D. B. (1975). "*t-Tests and Intervals for Comparisons suggested by the Data*" Biometrics, 31, pages 339-359.

Dunnett, C. W. (1980). "*Pairwise Multiple Comparisons in the Homogeneous Variance for Unequal Sample Size Case*", Journal of the American Statistical Association, 75, page 789.

Einot, I. and Gabriel, K. R. (1975). "*A Study of the Powers of Several Methods of Multiple Comparison*", Journal of the American Statistical Association, 70, page 351.

Gabriel, K. R. (1978). "*A Simple Method of Multiple Comparisons of Means*", Journal of the American Statistical Association, 73, page 364.

Hochburg, Y. (1974). "*Some Conservative Generalizations of the T-Method in Simultaneous Inference*", Journal of Multivariate Analysis, 4, pages 224-234.

Kramer, C. Y. (1956). "*Extension of Multiple Range Tests to Group Means with Unequal Sample Sizes*", Biometrics, 12, pages 307-310.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). "*On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance*", Biometrics, 63, pages 655-660.

Ryan, T. A. (1959). "*Multiple Comparisons in Psychological Research*", Psychological Bulletin, 56, pages 26-47.

Ryan, T. A. (1960). "*Significance Tests for Multiple Comparisons of Proportions, Variances, and Other Statistics*", Psychological Bulletin, 57, pages 318-328.

Scheffe, H. (1953). "*A Method for Judging All Contrasts in the Analysis of Variance*", Biometrika,40, pages 87-104.

Sidak, Z., (1967). "*Rectangular Confidence Regions for the Means of Multivariate Normal Distributions*", Journal of the American Statistical Association, 62, pages 626-633.

Tukey, J. W. (1953). *The Problem of Multiple Comparisons*, Unpublished Manuscript.

Waller, R. A. and Duncan, D. B. (1969). "*A Bayes Rule for the Symmetric Multiple Comparison Problem*", Journal of the American Statistical Association 64, pages 1484-1504.

Waller, R. A. and Kemp, K. E. (1976). "*Computations of Bayesian t-Values for Multiple Comparisons*", Journal of Statistical Computation and Simulation, 75, pages 169-172.

Welsch, R. E. (1977). "*Stepwise Multiple Comparison Procedure*", Journal of the American Statistical Association, 72, page 359.