

Sample Biases Vol 1, pg 442**Data-Mining Bias**

Data mining is the practice of extensively searching through a data set for statistically significant relationships till a pattern ‘that works’ is discovered. In the process of data mining, large numbers of hypotheses about a single data set are tested in a very short time by searching for combinations of variables that might show a correlation.

Warning signs that data mining bias might exist are:

- Too much digging warning sign.
- No story, no future warning sign.

The best way to avoid data-mining bias is to test the ‘apparently statistically significant relationships’ on ‘out-of-sample’ data to check whether they hold.

Sample Selection Bias

Sample selection bias results from the exclusion of certain assets from a study due to the unavailability of data.

Some databases use historical information and may suffer from a type of sample selection bias known as **survivorship bias**. Databases that only list companies or funds currently in existence suffer from this bias.

Look-Ahead Bias

An analyst may not have complete information at the time of testing. Look-ahead bias arises if the analyst uses an assumed figure instead. The actual figure may be different from the one used in the study.

Time-Period Bias

Time-period bias arises if the sample data is drawn from a certain time period. The results obtained from the study of such a data set will be time-specific.

HYPOTHESIS TESTING

 Cross-Reference to CFA Institute Assigned Reading #11

Steps in Hypothesis Testing Vol 1, pg 457

- State the hypothesis.
- Select the appropriate test-statistic.
- Specify the level of significance.
- State the decision rules.
- Calculate the sample statistic.
- Make a decision regarding the validity of the hypothesis.

Null versus Alternative Hypotheses Vol 1, pg 458

- The **null hypothesis** (H_0) generally represents the status quo, and is the hypothesis that we are *interested in rejecting*. This hypothesis will not be rejected unless the sample data provides sufficient evidence to suggest otherwise.
- The **alternate hypothesis** (H_A) is the statement that will only be accepted if the sample data provides convincing evidence of its truth. It is the conclusion of the test if the null hypothesis is rejected. *The alternate hypothesis is essentially the statement whose validity we are trying to evaluate.*

One-Tailed versus Two-Tailed Tests

Under **one-tailed tests**, we assess whether the value of the population parameter is either *greater than*, or *less than* a given hypothesized value. Hypotheses for one-tailed hypotheses tests are structured as:

- $H_0: \mu \leq \mu_0$ versus $H_a: \mu > \mu_0$; when determining whether the population mean is *greater* than a hypothesized value.
- $H_0: \mu \geq \mu_0$ versus $H_a: \mu < \mu_0$; when determining whether the population mean is *less* than a hypothesized value.

The following rejection rules apply when we are trying to determine whether the population mean is *greater* than the hypothesized value.

- Reject H_0 when:
Test statistic $>$ positive critical value
- Fail to reject H_0 when:
Test statistic \leq positive critical value

When we want to ascertain whether the population mean is *greater* than the hypothesized mean we compare the test statistic to the *positive* critical value.

The following rejection rules apply when we are trying to determine whether the population mean is *lower* than the hypothesized value.

- Reject H_0 when:
Test statistic $<$ negative critical value
- Fail to reject H_0 when:
Test statistic \geq negative critical value

When we want to ascertain whether the population mean is *less* than the hypothesized mean we compare the test statistic to the *negative* critical value.

Under **two-tailed tests**, we assess whether the value of the population parameter is *simply different from* a given hypothesized value. Hypotheses for two-tailed hypotheses tests are structured as:

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

The following rejection rules apply for two-tailed hypothesis test

- Reject H_0 when:
 Test statistic < lower critical value
 Test statistic > upper critical value
- Fail to reject H_0 when:
 Lower critical value \leq test statistic \leq Upper critical value

Type I versus Type II Errors Vol 1, pg 461

- **Type I error:** Rejecting the null when it is actually true.
- **Type II error:** Not rejecting the null when it is actually false.

The significance level (α) represents the probability of making a Type I error. A significance level of 5% means that there is a 5% chance of rejecting the null when it is actually true.

If we were to fail to reject the null hypothesis given the lack of overwhelming evidence in favor of the alternate, we risk a Type II error- *failing to reject the null hypothesis when it is false.*

Sample size and the choice of significance level (probability of Type I error) together determine the probability of a Type II error.

The power of a test is the probability of *correctly* rejecting the null hypothesis when it is false.

$$\text{Power of a test} = 1 - P(\text{Type II error})$$

Errors in Hypothesis Testing

Decision	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Incorrect decision Type II error
Reject H_0	Incorrect decision Type I error Significance level (α)	Correct decision Power of the test = $1 - P(\text{Type II error})$

- The *higher* the power of the test, the *better* is it for purposes of hypothesis testing.
- An *increase* in the power of a test comes at the cost of *increasing* the probability of a Type I error.
- The only way to *decrease* the probability of a Type II error given the significance level is to *increase the sample size.*

Confidence Intervals versus Hypothesis Tests Vol 1, pg 463

- In a confidence interval we state that the population parameter lies within the interval, which represents the 'fail-to-reject-the-null region' with a $(1 - \alpha)$ level of confidence.
- In a hypothesis test, we examine whether the population parameter lies in the rejection region, or outside the interval, at the α level of significance.

P-Values and Hypothesis Tests Vol 1, pg 465

The p-value is the smallest level of significance at which the null hypothesis can be rejected. It is the probability of obtaining a critical value that would lead to the rejection of the null hypothesis.

- If the p-value is *less* than the required level of significance, we can *reject* the null hypothesis.
- If the p-value is *greater* than the required level of significance, we *fail to reject* the null.

Summary of Hypothesis Tests on the Mean of a Single Population

Type of test	Null hypothesis	Alternate hypothesis	Reject null if	Fail to reject null if	P-value represents
One-tailed (upper tail)	$H_0: \mu \leq \mu_0$	$H_a: \mu > \mu_0$	Test statistic $>$ critical value	Test statistic \leq critical value	Probability that lies above the computed test statistic.
One-tailed (lower tail)	$H_0: \mu \geq \mu_0$	$H_a: \mu < \mu_0$	Test statistic $<$ critical value	Test statistic \geq critical value	Probability that lies below the computed test statistic.
Two-tailed	$H_0: \mu = \mu_0$	$H_a: \mu \neq \mu_0$	Test statistic $<$ Lower critical value Test statistic $>$ Upper critical value	Lower critical value \leq test statistic \leq Upper critical value	Probability that lies above the positive value of the computed test statistic <i>plus</i> the probability that lies below the negative value of the computed test statistic.

Hypothesis Tests Concerning a Single Mean Vol 1, pg 466

The **t-test** is used when the variance of the population is unknown *and* either of the conditions below hold:

- The sample size is large.
- The sample is small, but the underlying population is normally distributed or approximately normally distributed.

The t-statistic for hypothesis test concerning the mean of a single population is:

$$t\text{-stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In a t-test, the sample's t-statistic is compared to the critical t-value with degrees of freedom (n-1) at the desired level of significance.

The **z-test** is used to conduct hypothesis tests of the population mean when the population is normally distributed and its variance is known.

$$z\text{-stat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

The z-test can also be used when the population's variance is unknown, but the sample size is large.

$$z\text{-stat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

In a z-test, the z-statistic is compared to the critical z-value at the given level of significance.

Hypothesis Tests Concerning the Mean of Two Populations Vol 1, pg 474

Population distribution	Relationship between samples	Assumption regarding variance	Type of test
Normal	Independent	Equal	t-test with pooled variance
Normal	Independent	Unequal	t-test with variance not pooled
Normal	Dependent	N/A	t-test with paired comparisons

Hypothesis Tests Concerning the Variance Vol 1, pg 482

Hypothesis tests for the variance of a normally distributed population involve the use of the chi-square distribution where the test statistic is denoted as χ^2 . Three important features of the chi-square distribution are:

- It is asymmetrical.
- It is bounded by zero. Chi-square values cannot be negative.
- It approaches the normal distribution in shape as degrees of freedom increase.

The chi-square test statistic with n-1 degrees of freedom is calculated as:

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

Hypotheses related to the equality of the variance of two populations are tested with an F-test. This test is used under the assumptions that:

- The populations from which samples are drawn are normally distributed.
- The samples are independent.

The test statistic for the F-test is given by:

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

Features of the F-distribution:

- It is skewed to the right.
- It is bounded by zero on the left.
- It is defined by two separate degrees of freedom

The rejection region for any F-test, whether it be one-tailed or two-tailed, always lies in the right tail. This unique feature makes the F-test different from other hypothesis tests.

Hypothesis Test Concerning	Appropriate test statistic
Variance of a single, normally distributed population	Chi-square stat
Equality of variance of two independent, normally distributed populations	F-stat

Parametric versus Nonparametric Tests Vol 1, pg 488

A parametric test has at least one of the following two characteristics:

- It is concerned with parameters, or defining features, of a distribution.
- It makes a definite set of assumptions.

A non-parametric test is not concerned with a parameter, and makes only a minimal set of assumptions regarding the population. Non-parametric tests are used when:

- The researcher is concerned about quantities other than the parameters of the distribution.
- The assumptions made by parametric tests cannot be supported.
- When the data available is ranked. For example, non-parametric methods are widely used for studying populations such as movie reviews that receive one to five stars based on people’s preferences.

Statistically versus Economically Meaningful Results

Even though a trading strategy that is being studied provides a statistically significant return of greater than zero (based on the hypothesis test) it does not guarantee that trading on this strategy would result in economically meaningful returns. The returns may not be economically significant after accounting for taxes, transaction costs and risks inherent in the strategy.