

Statistics basics.

Vladimir Lobaskin

Oktober 22, 2005

The main concepts

Population - as the entire collection of items that is the focus of concern.

Descriptive Statistics describes characteristics of a population

Inferential Statistics makes educated inferences about the characteristics of a population by drawing a random sample.

Random sample is a set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample.

Location is the typical or central value that best describes the data.

Scale is the measure of the dispersion of the data.

The probability distribution

Discrete

1. The probability that x can take a specific value is $p(x)$: $P[X = x] = p(x) = p_x$
2. $0 \leq p(x) \leq 1$.
3. $\sum_j p_j = 1$.

Continuous

1. The probability that x is between two points a and b is

$$p[a \leq x \leq b] = \int_a^b f(x)dx \quad (1)$$

2. It is non-negative for all real x .
- 3.

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (2)$$

Location

The most common definitions:

1. **mean (average)**

$$\bar{Y} = \sum_{i=1}^N Y_i / N \quad (3)$$

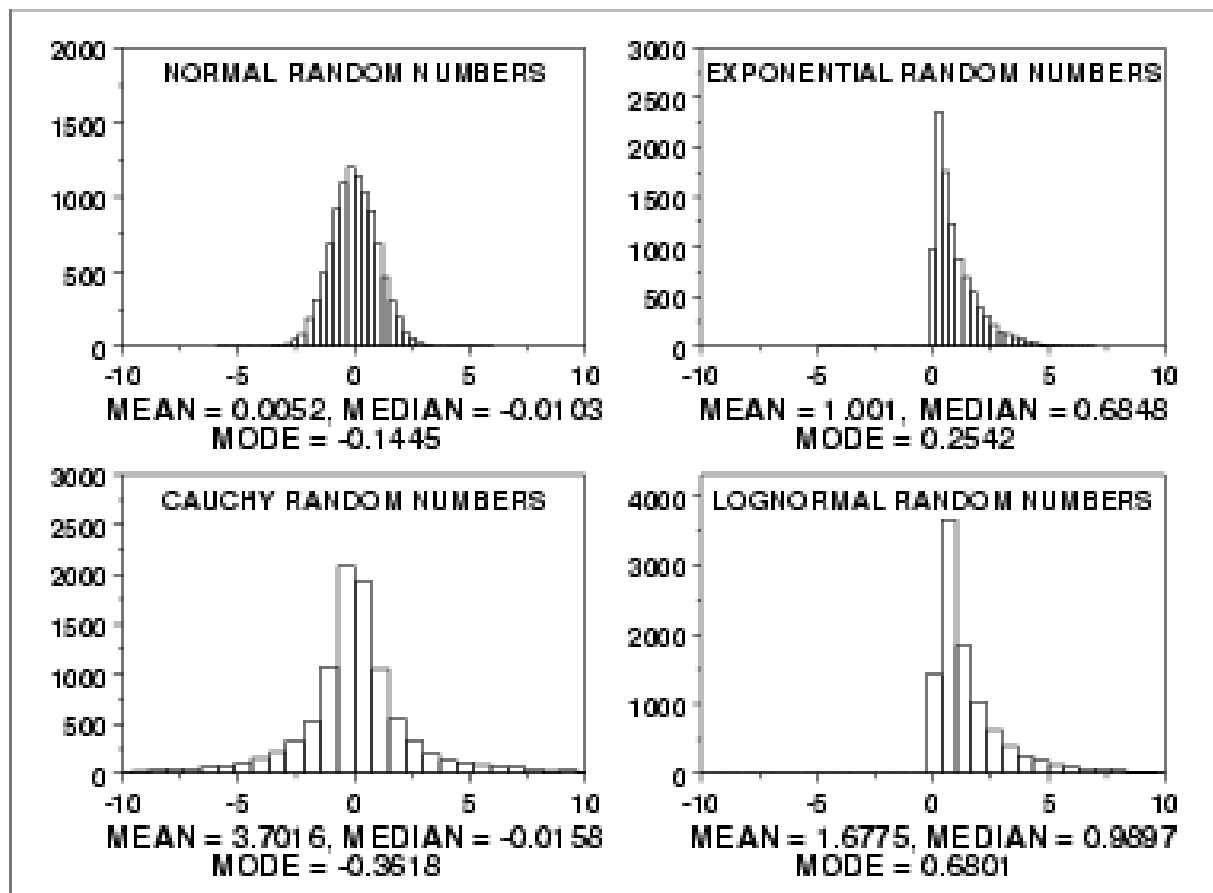
2. **median** - the median is the value of the point which has half the data smaller than that point and half the data larger than that point. If X_1, X_2, \dots, X_N is a random sample sorted from smallest value to largest value,

$$\tilde{Y} = Y_{(N+1)/2} \text{ if } N \text{ is odd}$$

$$\tilde{Y} = (Y_{N/2} + Y_{(N/2)+1}) / 2 \text{ if } N \text{ is even}$$

3. **mode** - the mode is the value of the random sample that occurs with the greatest frequency.

Location



Scale / Variability / Spread

The common numerical measures of the spread:

1. variance

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

2. standard deviation

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

standard error (of the mean)

$$SEM = \sqrt{\frac{s^2}{N}} = \sqrt{\frac{s^2}{\bar{N}}}$$

The standard deviation is expressed in the units of the location.

3. **range** - the range is the largest value minus the smallest value in a data set.

4. average absolute deviation

$$AAD = \frac{\sum_{i=1}^N (|Y_i - \bar{Y}|)}{N}$$

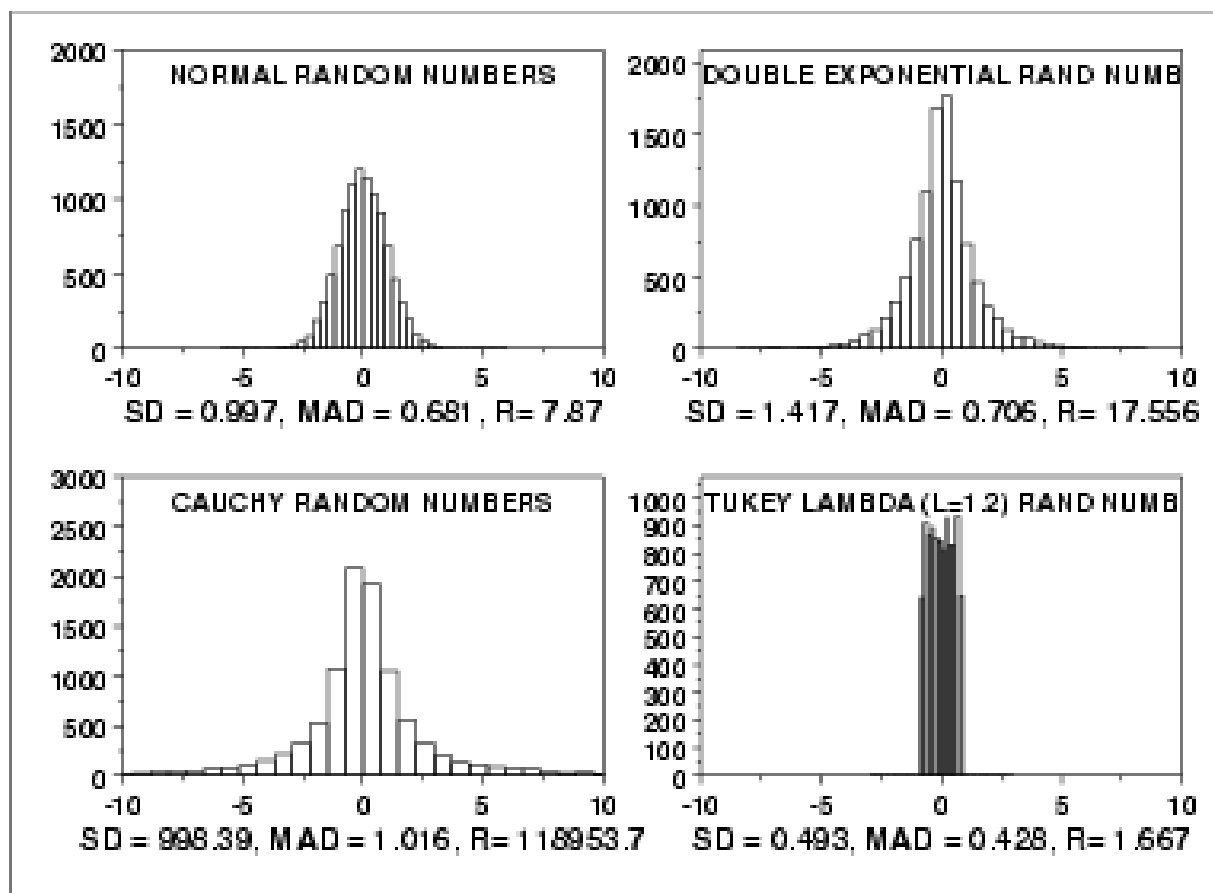
5. median absolute deviation

$$MAD = MEDIAN(|Y_i - \tilde{Y}|)$$

where \tilde{Y} is the median of the data.

6. interquartile range - this is the value of the 75th percentile minus the value of the 25th percentile.

Scale



46 64 54 77 67 68 62 56 38 Population
N = 9

Random
Sample
n = 4 38 62 67 62

$$\bar{X} = \frac{\sum x}{n} = \frac{229}{4} = 57.25$$

The mean of this Random Sample
equals 57.25 (i.e. $\bar{X} = 57.25$)

$$\mu_x = \frac{\sum x}{N} = \frac{532}{9} = 59.11$$

The Mean of this Population (μ_x)
equals 59.11 (i.e. $\mu_x = 59.11$)

The Central Limit Theorem tells us
that \bar{X} is an unbiased estimate
of μ_x . (i.e. $\bar{X} \rightarrow \mu_x$)

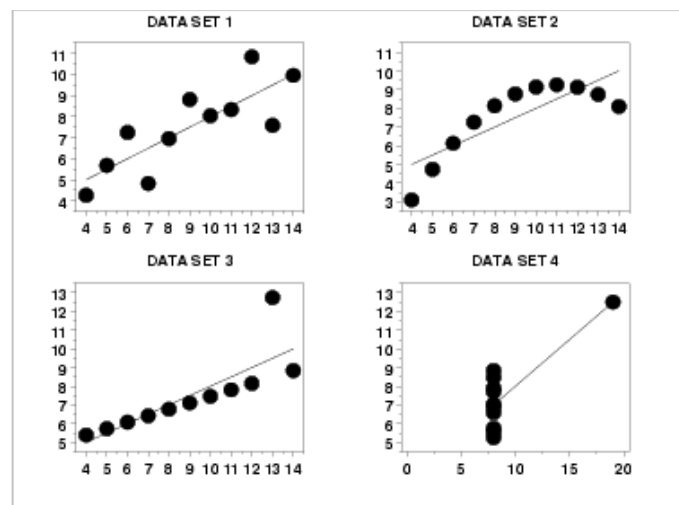
In short, with only one random sample to go on, the mean of the
sample ($\bar{X} = 57.25$) is our best estimate of the population mean (μ_x)

Exploratory data analysis: questions

1. What is a typical value?
2. What is the uncertainty for a typical value?
3. What is a good distributional fit for a set of numbers?
4. Does a factor have an effect?
5. What are the most important factors?
6. Are measurements coming from different sets equivalent?
7. What is the best function for relating a response variable to a set of factor variables?
8. What are the best settings for factors?
9. Can we separate signal from noise in time dependent data?
10. Can we extract any structure from multivariate data?
11. Does the data have outliers?

Exploratory data analysis: methods

Graphical analysis



These four sets produce equal parameters when fitted with a straight line.

$$N = 11$$

$$\text{Mean of } X = 9.0$$

$$\text{Mean of } Y = 7.5$$

$$\text{Intercept} = 3$$

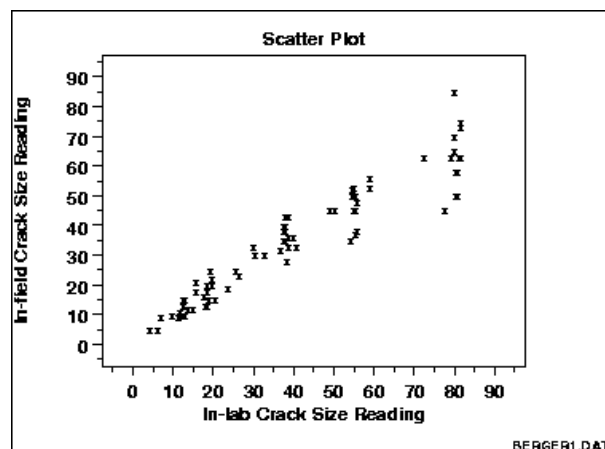
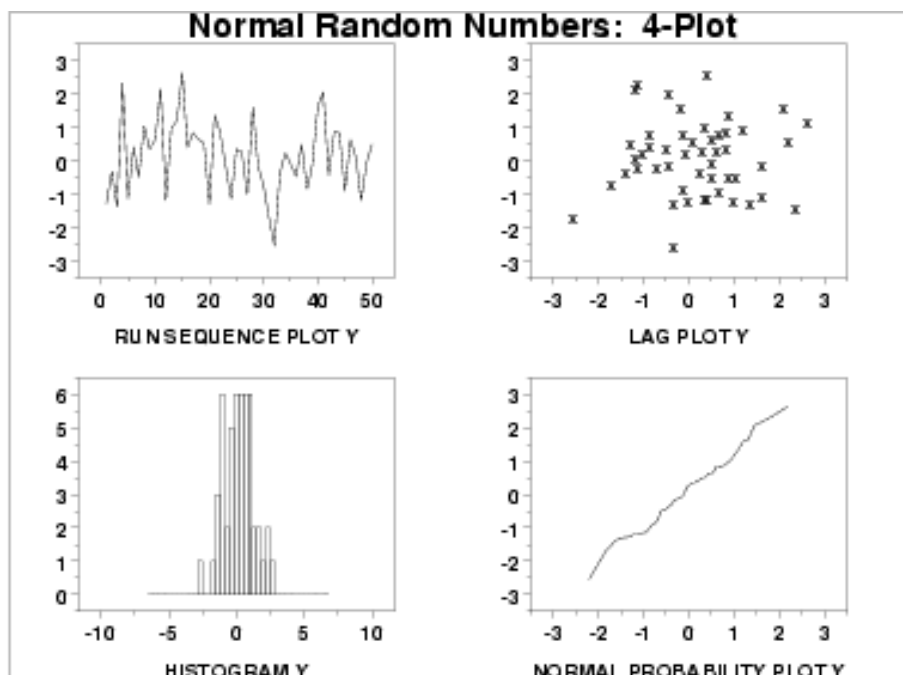
$$\text{Slope} = 0.5$$

$$\text{Residual standard deviation} = 1.237$$

$$\text{Correlation} = 0.816$$

Exploratory data analysis: methods

Graphical analysis



Exploratory data analysis: methods

Quantitative analysis

1. Interval estimation
2. Hypothesis tests

Hypothesis test

A common format for a hypothesis test is:

H_0 : A statement of the **null hypothesis**, e.g., two population means are equal.

H_a : A statement of the **alternative hypothesis**, e.g., two population means are not equal.

Test Statistic: The test statistic is based on the specific hypothesis test.

Significance Level: α defines the sensitivity of the test. A value of $\alpha = 0.05$ means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true (the type I error).

Critical Region: The critical region encompasses those values of the test statistic that lead to a rejection of the null hypothesis.

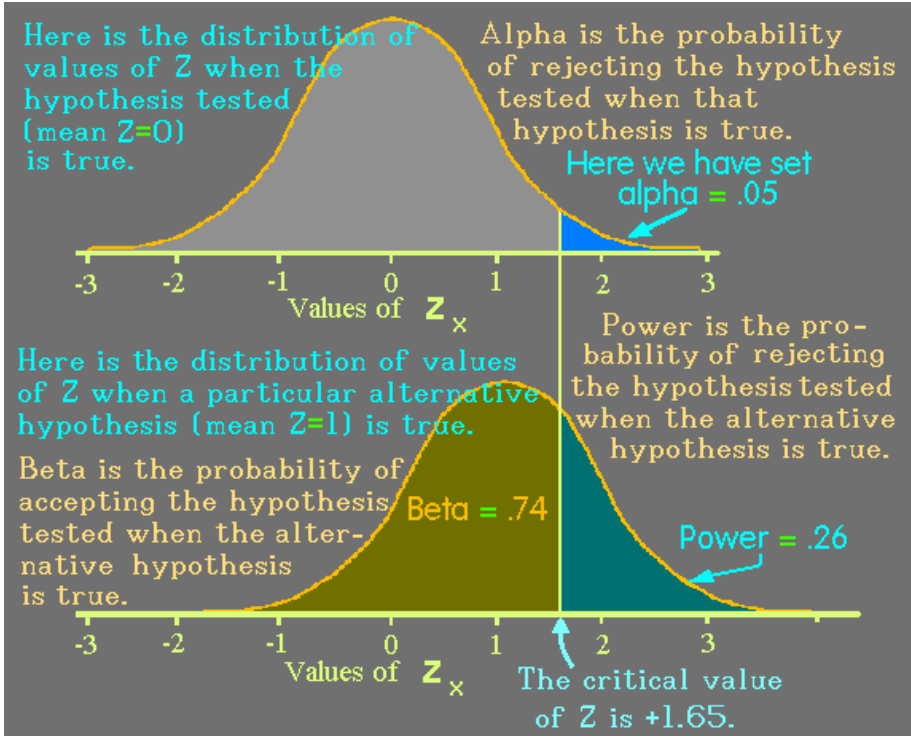
Significance level: α

Statisticians use α to indicate the probability of rejecting the statistical hypothesis tested when in fact, that hypothesis is true. Before conducting any statistical test, it is important to establish a value for alpha. For most psychologists, and for many other scientists, it is customary to set alpha at 0.05.

This is the equivalent of asserting that you will reject the hypothesis tested if the obtained statistic is among those that would occur only 5 out of 100 times that random samples are drawn from a population in which the hypothesis is true. If your obtained statistic leads you to reject the hypothesis tested, it's not because you believe that the obtained statistic could not have occurred by chance.

It's that you are asserting that the odds of obtaining that statistic by chance only are sufficiently low (one out of twenty) that it is reasonable to conclude that your results are not due to chance. Could you be in error? Of course you could, but at least you know the probability of such an error. It is exactly equal to the value you have previously established for alpha.

The deviation from the hypothesis is called **statistically significant** if the test statistic is outside the expected dispersion interval for the given α and the number of degrees of freedom (measurements).



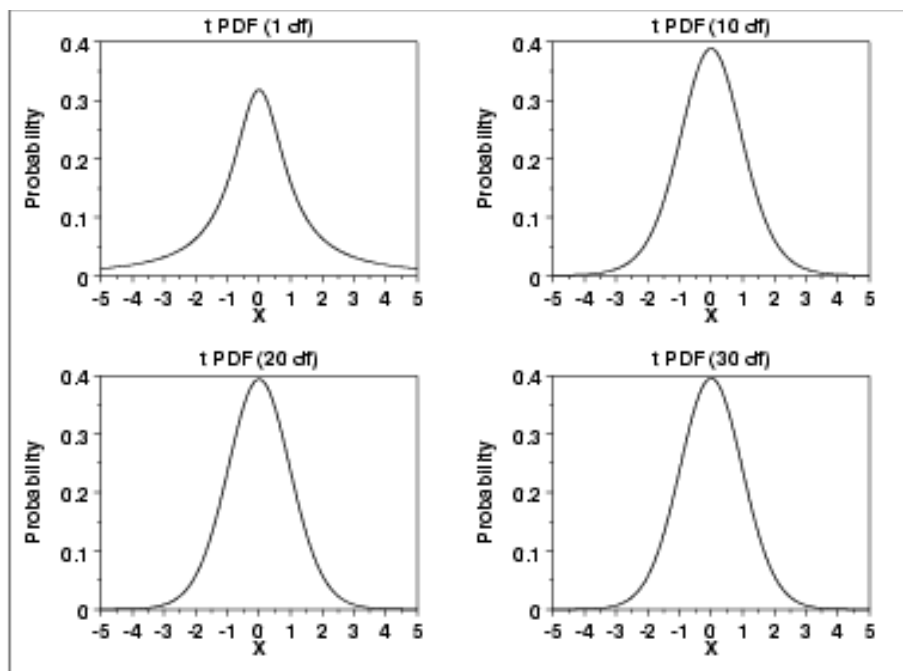
t distribution

W.S. Gosset (Student)

$$f(x) = \frac{(1+x^2/\nu)^{-(\nu+1)/2}}{\gamma(0.5,0.5\nu)\sqrt{\nu}} \quad (4)$$

ν is a positive integer shape parameter.

$$\gamma(\alpha, \beta) = \int_0^1 (t^{\alpha-1}(1-t)^{\beta-1}) dt \quad (5)$$



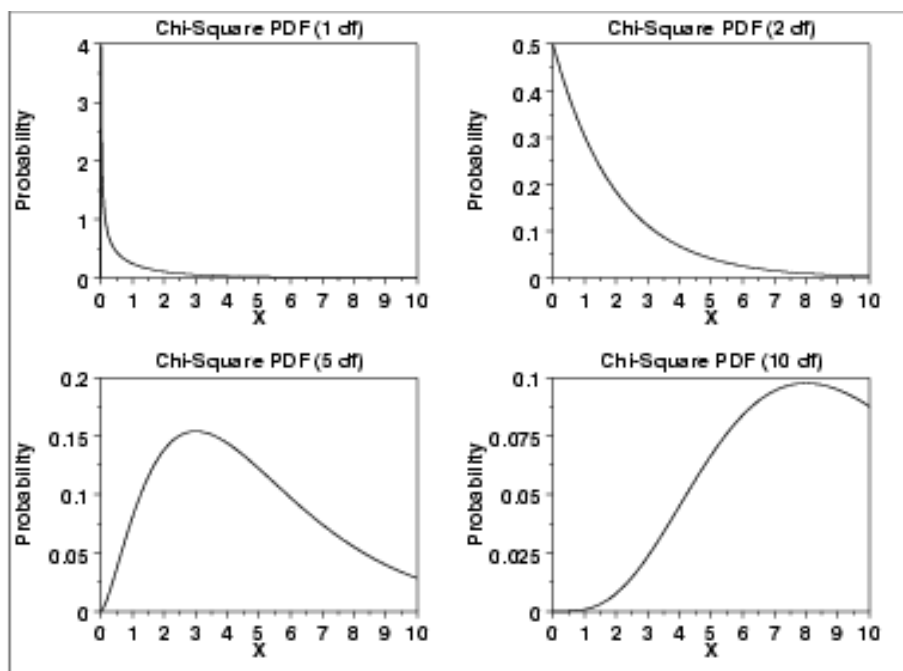
χ^2 distribution

The χ^2 distribution results when ν independent variables with standard normal distributions are squared and summed. The probability density

$$f(x) = \frac{\exp(-x/2)x^{\nu/2-1}}{(2^{\nu/2}\Gamma(\nu/2))}, x \geq 0 \quad (6)$$

where ν is the shape parameter. Gamma function:

$$\Gamma(a) = \int_0^{\infty} t^{a-1}e^{-t}dt \quad (7)$$



Hypothesis test

The **null hypothesis** is a term that statisticians often use to indicate the statistical hypothesis tested. The purpose of most statistical tests, is to determine if the obtained results provide a reason to reject the hypothesis that they are merely a product of chance factors. For example, in an experiment in which two groups of randomly selected subjects have received different treatments and have yielded different means, it is always necessary to ask if the difference between the obtained means is among the differences that would be expected to occur by chance whenever two groups are randomly selected. In this example, the hypothesis tested is that the two samples are from populations with the same mean. Another way to say this is to assert that the investigator tests the null hypothesis that the difference between the means of the populations from which the samples were drawn, is zero. If the difference between the means of the samples is among those that would occur rarely by chance when the null hypothesis is true, the null hypothesis is rejected and the investigator describes the results as **statistically significant**.

χ^2 test says whether the data are like the results of drawing at random from the population whose *contents* are given

z or t test says whether the data are like the results of drawing at random from the population whose *average* is given

Confidence intervals: t-test

Confidence limits are commonly defined as:

$$Y = \bar{Y} \pm t_{(\alpha/2, N-1)} s / \sqrt{N} \quad (8)$$

\bar{Y} is the sample mean

s the sample standard deviation

N is the sample size

α is the desired significance level

$t_{(\alpha/2, N-1)}$ is the upper critical value of the t distribution with $N - 1$ degrees of freedom.

The confidence coefficient is $1 - \alpha$.

Confidence intervals: one sample t-test

$H_0: \mu = \mu_0$ the population mean has a specific value, μ_0

$H_a: \mu \neq \mu_0$

Test Statistic: $T = (\bar{Y} - \mu_0)/(s/\sqrt{N})$

Significance Level: α . The most commonly used value for α is 0.05.

Critical Region: Reject the null hypothesis that the mean is a specified value, μ_0 , if

$$T < -t_{(\alpha/2, N-1)} \quad (9)$$

or

$$T > t_{(\alpha/2, N-1)} \quad (10)$$

Two-sample t-test

$H_0: \mu_1 = \mu_2$ two populations have the same mean

$H_a: \mu_1 \neq \mu_2$

Test Statistic:

$$T = (\bar{Y}_1 - \bar{Y}_2) / \sqrt{s_1^2/N_1 + s_2^2/N_2} \quad (11)$$

where N_1 and N_2 are the sample sizes, \bar{Y}_1 and \bar{Y}_2 are the sample means, and s_1^2 and s_2^2 are the sample variances.

If equal variances are assumed, then

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2)}{[s_p \sqrt{(1/N_1) + (1/N_2)}]} \quad (12)$$

where

$$s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{(N_1 + N_2 - 2)} \quad (13)$$

Significance Level: α .

Critical Region: Reject the null hypothesis that the two means are equal if

$$T < -t_{(\alpha/2, \nu)} \quad (14)$$

or

$$T > t_{(\alpha/2,\nu)} \quad (15)$$

where $t_{(\alpha/2,\nu)}$ is the critical value of the t distribution with ν degrees of freedom where

$$\nu = \frac{(s_1^2/N_1 + s_2^2/N_2)^2}{(s_1^2/N_1)^2/(N_1 - 1) + (s_2^2/N_2)^2/(N_2 - 1)} \quad (16)$$

If equal variances are assumed, then the formula reduces to:

$$\nu = N_1 + N_2 - 2 \quad (17)$$

Do the observations come from a particular distribution?

Three goodness-of-fit tests are often used:

1. χ^2 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.

Other general and specific tests might be used, including tests based on regression and graphical techniques.

χ^2 test

H_0 (Null-hypothesis): The data follow a specified distribution.

H_a : The data do not follow the specified distribution.

Test Statistic:

$$\chi^2 = \sum_{i=1}^k (O_i - E_i)^2 / E_i \quad (18)$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin i . The expected frequency is calculated by

$$E_i = N(F(Y_u) - F(Y_l)) \quad (19)$$

where F is the cumulative distribution function for the distribution being tested, Y_u is the upper limit for class i , Y_l is the lower limit for class i , and N is the sample size.

Significance Level: α .

Critical Region: The hypothesis is rejected if

$$\chi^2 > \chi^2_{(\alpha, k-c)} \quad (20)$$

The Kolmogorov-Smirnov test

H_0 : The data follow a specified distribution.

H_a : The data do not follow the specified distribution

Test Statistic:

$$D = \max_{1 \leq i \leq N} [F(Y_i) - (i-1)/N, (i/N) - F(Y_i)] \quad (21)$$

F - theoretical cumulative distribution of the tested continuous distribution, which must be fully specified.

Significance Level: α .

Critical Values: The hypothesis regarding the distributional form is rejected if the test statistic, D , is greater than the critical value obtained from a table.

Linear Least Squares Regression

used to fit the data of the form

$$f(\vec{x}; \vec{\beta}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 \dots \quad (22)$$

or a more general form.

1. each explanatory variable in the function is multiplied by an unknown parameter,
2. there is at most one unknown parameter with no corresponding explanatory variable,
3. all of the individual terms are summed to produce the final function value.

The function is *linear* with respect to the parameters.

Then, one minimizes the value

$$Q = \sum_{i=1}^n [y_i - f(\vec{x}; \vec{\beta}')]^2 \quad (23)$$

χ^2 -test of the scale

Purpose: Test if standard deviation is equal to a specified value

Questions: The χ^2 test can be used to answer the following questions:

1. Is the standard deviation equal to some pre-determined threshold value?
2. Is the standard deviation greater than some pre-determined threshold value?
3. Is the standard deviation less than some pre-determined threshold value?

$H_0: \sigma = \sigma_0$

$H_a: \sigma < \sigma_0$ for a lower one-tailed test

$\sigma > \sigma_0$ for an upper one-tailed test

$\sigma \neq \sigma_0$ for a two-tailed test

Test Statistic:

$$T = (N - 1)(s/\sigma_0)^2 \quad (24)$$

Significance Level: α

Critical Region: Reject the null hypothesis that the standard deviation is a specified value, σ_0 , if

$T > \chi^2_{(\alpha, N-1)}$ for an upper one-tailed alternative

$T < \chi^2_{(1-\alpha, N-1)}$ for a lower one-tailed alternative

$T < \chi^2_{(1-\alpha/2, N-1)}$ for a two-tailed test
or

$T > \chi^2_{(1-\alpha/2, N-1)}$

where $\chi^2_{(\dots, N-1)}$ is the critical value of the χ^2 distribution with $N - 1$ degrees of freedom.

$\chi^2_{(\alpha)}$ is the upper critical value and $\chi^2_{(1-\alpha)}$ is the lower critical value from the χ^2 distribution.

The formula for the hypothesis test can be converted to form an interval estimate for the standard deviation:

$$\sqrt{\frac{(N-1)s^2}{\chi^2_{(\alpha/2, N-1)}}} \leq \sigma \leq \sqrt{\frac{(N-1)s^2}{\chi^2_{(1-\alpha/2, N-1)}}} \quad (25)$$