

Basic statistical concepts

1. Variable types.

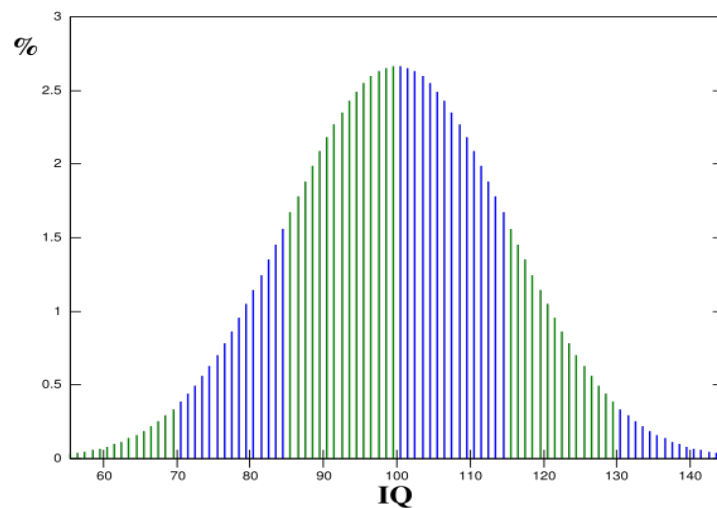
1. independent variable (IV)
2. dependent variable (DV)
3. moderator variable
4. covariate
5. random variable (subjects)

types of variables, according to type of data:

1. nominal (categorical)
2. ordinal (categorical)
3. interval (numerical, continuous)

2. Distribution

Distribution: The distribution of probabilities for various values. For example, look at the distribution of IQ scores, which follow the standard bell curve shape.



Many things measured numerically in the natural world or in social sciences follow this distribution, known as the normal, standard normal, or Gaussian distribution. People's scores will more often fall into the 85-115 range, followed by the 70-85 and 115-130 ranges, and so on. These ranges represent certain distances from the mean ($\bar{x}=100$) that are defined by standard deviations.

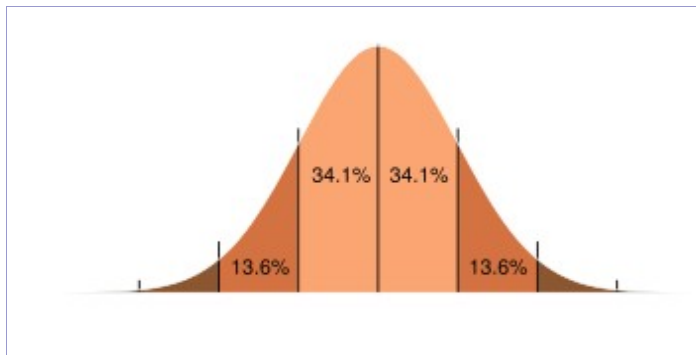
Since many things follow the normal distribution, we assume this in our statistics, and try to fit data to the normal distribution – or a variant form of it, adapted to particular statistical tests, such as the Student distribution for t-tests (for simple comparisons of two groups of a similar and “normal” size) or the F-distribution for ANOVA tests.

3. Variance.

An important part of the calculations in statistics tests involves variance. Let's look at the IQ test. Each individual's score varies from the overall group mean of 100. We subtract the individual score from the group mean ($|100-x|$) for the individual variance. Then we average all the individual variances for the average amount of deviation from the mean – the standard deviation¹.

score	variance	=
71	100-71	= 29
87	100-87	= 13
92	100-92	= 8
98	100-98	= 2
102	100-102	= 2
107	100-107	= 7
116	100-116	= 16
133	100-133	= 33
...

If we do this for a relatively large sample of the population – for a random sample of subjects – we can take the average of all the individual variances, and find that the $sd=15$ (for the Wechsler IQ test) or $sd=16$ (Stanford-Binet IQ test). This is true for all adult humans who have taken an IQ test. So they will all fall into the same distribution with the same standard deviation.



1 SD = 68.27% of the set
2 SDs = 95.45%
3 SDs = 99.73%
4 SDs = 99.994%

Standard deviation units.

Once the mean and SD are known for a set of data, the scores (such as subjects' test scores or responses) can be standardized by expressing the score in terms of SD units. This is known as a z-score. The z-score is computed as follows, for a score x and a mean \bar{x} :

$$z = \frac{\bar{x} - x}{sd}$$

4. Comparison of groups.

Stats tests essentially compare two or more groups to determine if they fall within the same distribution, and are thus the same thing, not different from each other – or fall into two separate normal distributions, and are thus separate entities. For example, in comparing two pedagogical techniques, we collect data such as student's scores on a post-test. If one is better than the other, they are different, and the students' scores are actually meaningfully different. Thus, scores from the two groups belong to two different normal distributions, and you can claim that the two are different, or that one is better. But if student scores are not meaningfully different, the scores follow the same distribution, and student performance is the same for both groups.

This meaningfulness in group difference is referred to as *significant difference*. Part of the stats results includes a p value, or probability value. This represents the probability that the two groups are the same, and the probability that the stats test results are due to random error.

In the social sciences, we rely on a p -value of $p<.05$ as a standard cut-off for significance, $p<.01$ is

¹ To avoid negative signs (like -7), since t-tests and ANOVAs were invented in the early 20th century before calculators existed, this is done by squaring the terms - computing the sum of squares (SS), or ordinary least squares (OLS). For example, in the IQ score example: $SS = 29^2 + 13^2 + 8^2 + 2^2 + 2^2 + 7^2 + 16^2 + 33^2 + \dots$; this is then fed into the t-test or ANOVA equation.

common in medical research.

Degrees of freedom: # groups & N-size.

Calculation of probability from the statistical test (t-test, ANOVA, correlation, etc.) depends on degrees of freedom, the dimensionality of the data.

df = number of “dimensions”; = # of possible comparisons of data points

$$\text{df for subjects} = N - 1$$

$$\text{df for conditions or groups} = k - 1 \quad (N = \text{\#subjects}, k = \text{\#groups/cond.})$$

Some tests use an overall df also: $\text{df}_{\text{total}} = N - k - 1$

5. t-test

For a simple comparison of two groups (groups X_1 and X_2 ; S here = standard deviation, N = # subjects – actually, $N-1$). Similar assumptions as for ANOVAs.

$$t = \frac{X_1 - X_2}{S - \sqrt{N}}$$

6. ANOVA

= Analysis of variance

Assumptions:

1. Independence of cases (one measurement doesn't affect a subsequent measurement)
2. Normality – normal distributions
3. Equal variances across groups
4. Interval data

For non-interval data, use a chi-square test or other non-parametric test.

Theoretical basis of ANOVAs:

$$SS_{\text{total}} = SS_{\text{treatment}} + SS_{\text{error}}$$

Computation:

$$F = \frac{\text{variance of group means}}{\text{mean of within – group variances}} \quad \text{or} \quad F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$

The results of the F-test are compared to the F-distribution values on a chart to determine significance (p). If an overall significant difference is found, then come follow-up or post hoc tests such as Tukey's, Scheffe's, or Huyn-Feldt (or Sidak's, Duncan's, Bonferroni's, LSD, Neuman-Keuls...).

If standard ANOVA assumptions don't apply, alternatives to a regular ANOVA would be less powerful tests like the Mann-Whitney or Kruskal-Wallis tests.

ANCOVA: analysis of covariance

A control variable or covariate is a second factor that you want to control for and factor out first, to separate its influence from that of the main IV.

Repeated measures ANOVA.

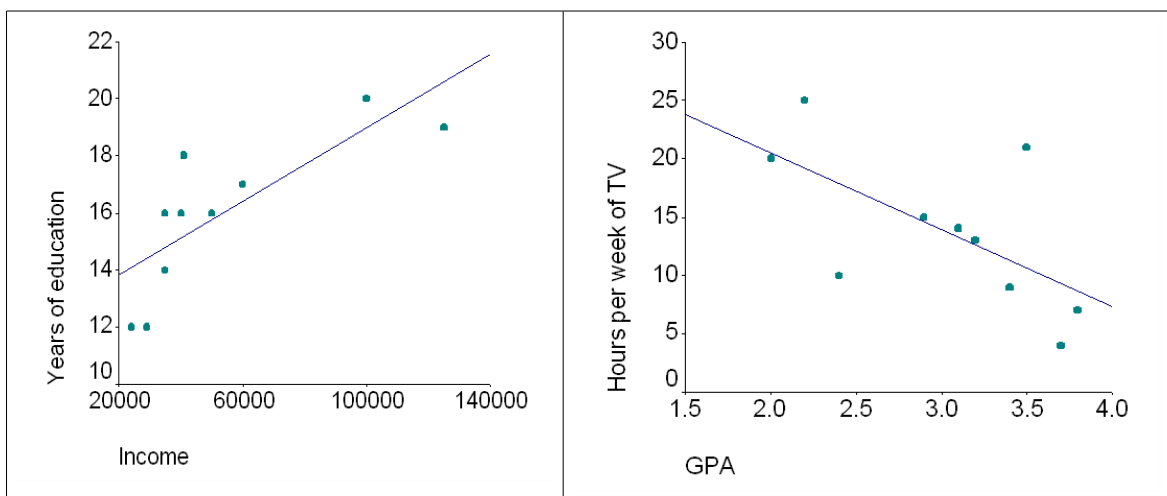
The subject responds twice in one session on one type of item – i.e., two or more responses on the same or similar item are collected from the same subject, at a time interval close enough that the second response could be influenced by the first response. For example, reading times on the same word or grammar structure that is encountered more than once would be a repeated measures. Special adjustments are used in the ANOVA computations for the variance between time intervals.

7. Correlation (regression).

$$y = a + bx + e$$

Are two groups correlated? Does one IV affect the DV in a regular, predicable manner? ANOVAs tell you if two groups are different or similar. Correlation tells you whether they are similar, and how similar. It is expressed as an r value (Pearson's r).

$$r = \frac{cov(X, Y)}{s_X s_Y}$$



Note: Correlation does not necessarily imply causation.

More complex cases:

multivariate regression

$$y = a + b_1x_1 + b_2x_2$$

Two or more IVs are entered at once.

hierarchical regression

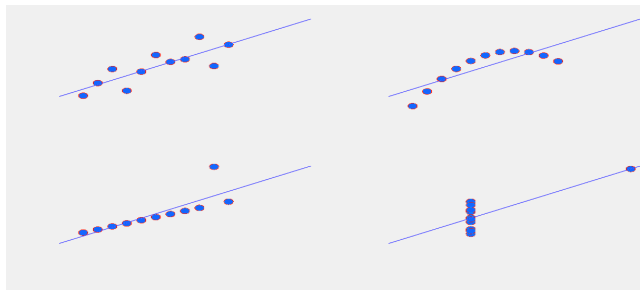
$$y = a + b_1x_1 + b_2x_2$$

Each IV is entered one at a time, to factor out other effects.

non-linear regression

Some non-linear cases might be described by a special form of the standard regression model with a quadratic term (x^2) or even a cubic term (x^3).

$$y = a + bx^2 + bx$$



Non-parametric equivalents also exist, when standard assumptions don't work, such as chi-square tests (proportions) or Spearman's rho (regression with categorical variables). Sometimes the regression might follow a logarithmic pattern, where logistic regression can be used.

8. Other tests.

8.1. Factor analysis, principal components analysis: For complex analyses of many factors, to see how some variables pattern together, or to find patterns in complex data. For example, one might subject survey data to FA / PCA to see what kinds of responses or factors are correlated with each other, or what kinds of other factors underly the responses.

8.2. MANOVA: Multiple analysis of variants – to test two DVs at once. This is not often used, and not having the same number of responses or data for each condition (e.g., if some subjects drop out or don't respond to some items) are more problematic for MANOVAs than for other tests.

8.3. Time series analysis: Somewhat like an ANOVA or regression model, for measuring events at different times and correlations between times and events (e.g., measuring performance on a particular language skill over time, and including the times as a factor in the equation). More complex analyses use what's known as a generalized liner model, such as:

$$y = t_1(b_{1x_1} \dots) + t_2 (b_{1x_1} \dots) \dots \quad \text{where } t = \text{measurement times}$$

8.4. Hierarchical linear model (HLM, or mixed level models): Somewhat like an ANOVA or regression model, but for a variety of different kinds individual and group level variables in a more complex statistical model. For example, we might have a complex study with two numerical IVs (like scores on two different tests as IVs), plus the students' classroom as a group variable (g_1) and students' school as another group variable (g_2), while explicitly factoring out each student's individual variability (r , for subject as random variable):

$$y = a + g_1 + g_2 + b_1x_1 + b_2x_2 + r$$

8.5. Logistic regression: correlations based on logarithmic functions; see below.

8.6. Cronbach's alpha, Kramer's kappa. When two raters code or rate data (e.g., evaluating ESL students' essays), their scores are compared with such a test to determine how consistent the two raters' scores are with each other (agreement, reliability) – hence, they are called tests of inter-rater reliability. For example, the essay ratings are a variable in the study. Ratings are a somewhat subjective variable, which have to be done by at least two persons (since one person's judgment is subjective at best); an average of the two is taken and used as a variable in the main analysis. But to make sure that the two raters are fairly consistent, a Cronbach's α or Kramer's κ is first performed.

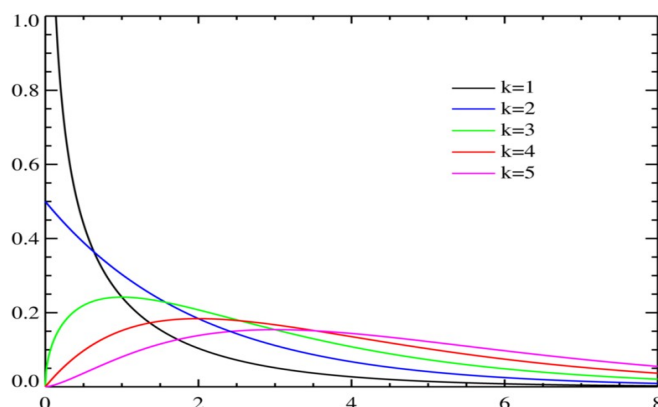
9. Other distributions.

In most language research, the normal distribution will suffice. The following are used in specialized areas, where certain kinds of data do not follow the normal distribution. In these cases, the shape of each distribution varies according to the number of subjects or n-size.

9.1. Chi-square distribution (χ^2)

A chi-square test is used to compare proportions of two groups, e.g., proportions of patients who improve in a drug treatment group cf. those in a control group. Thus, it is often used for categorical dependent variables, or for group comparisons where standard ANOVA assumptions do not hold.

Most often this is used for categorical dependent variables, i.e., if the DV or main IV pertains to identifying which category an item belongs to; for example, in predicting whether the subject is classified into certain categories:



1) successful L2 attainment:

{successful / moderate / average / poor} \Leftrightarrow score on aptitude test

2) successful L2 attainment:

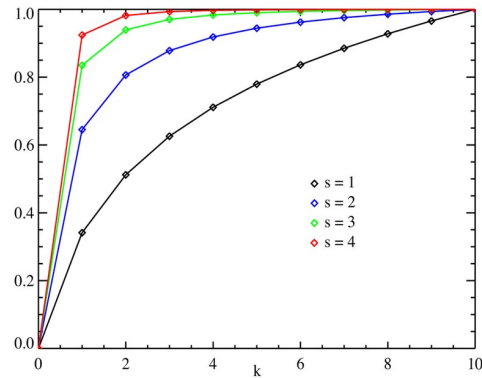
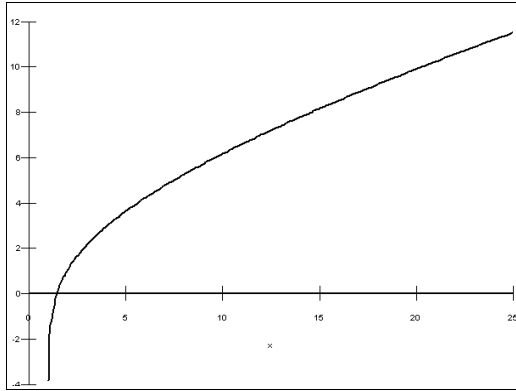
{successful / moderate / avg. / poor} \Leftrightarrow type of motivation {extrinsic / pragmatic / social}

Chi-square tests are also used when a variable doesn't follow the standard distribution, e.g., due to too much variability or range in scores, or when scores aren't true numerical variables (e.g., Likert scale surveys, where the numbers represent categories, or otherwise don't follow a full, boundless numerical distribution).

This includes random variables, too. Thus, the chi-square is also used in some more advanced procedures when particular subjects (random effect) are an important variable to be measured and controlled for.

Other statistical tests based on some other distributions below are more complicated, in that the tests use a chi-square test to test for functions of logarithmic, Poisson, and other distributions.

9.2. logarithmic distribution (similar to exponential distribution)



This is used, for example, for word frequencies in corpus data, since word frequencies follow a logarithmic distribution. If you rank words in frequency of occurrence (1st, 2nd, 3rd, ...), e.g., based on counts in a 10 million word corpus, the most frequent words will be very high in frequency, and the lowest frequency words may occur only once or a couple of times. If you plot rank by frequency count, you'll get a graph similar to those above. Their distribution is based on logarithmic values (logarithms of frequency counts), and are ideally computed in logistic regression models. The logarithms used here are not base-10 logs, but natural logs (ln), based on $e=2.71828183$.

The correlation is defined by the following equation, which provides more statistical power than standard regression or standard statistical models.

$$\log(y) = a + b_1x_1 + b_2x_2 \dots$$

9.3. Poisson distribution (simple counts)

The Poisson is for simple counts of events, e.g., for how many car accidents happen along a certain road, or how many students fail to show up for an exam. It is not likely that you will encounter this, except perhaps in logistic regression with corpus data or event data.

