Statistics Formulas

This web page lists statistics formulas used in the Stat Trek tutorials. Each formula links to a web page that explains how to use the formula.

Parameters

- Population mean = $\mu = (\Sigma X_i) / N$
- Population standard deviation = $\sigma = \mathrm{sqrt} [\Sigma (X_i - \mu)^2 / N]$
- Population variance = $\sigma^2 = \Sigma (X_i - \mu)^2 / N$
- Variance of population proportion = $\sigma_P^2 = PQ / n$
- Standardized score = $Z = (X - \mu) / \sigma$
- Population correlation coefficient = $\rho = [1 / N] * \Sigma \{ [(X_i - \mu_X) / \sigma_x] * [(Y_i - \mu_Y) / \sigma_y] \}$

Statistics

Unless otherwise noted, these formulas assume simple random sampling.

- Sample mean = $\bar{x} = (\Sigma x_i) / n$
- Sample standard deviation = $s = \mathrm{sqrt} [\Sigma (x_i - \bar{x})^2 / (n - 1)]$
- Sample variance = $s^2 = \Sigma (x_i - \bar{x})^2 / (n - 1)$
- Variance of sample proportion = $s_p^2 = pq / (n - 1)$
- Pooled sample proportion = $p = (p_1 * n_1 + p_2 * n_2) / (n_1 + n_2)$
- Pooled sample standard deviation = $s_p = \mathrm{sqrt} [(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2] / (n_1 + n_2 - 2)]$
- Sample correlation coefficient = $r = [1 / (n - 1)] * \Sigma \{ [(x_i - \bar{x}) / s_x] * [(y_i - \bar{y}) / s_y] \}$

Correlation

- Pearson product-moment correlation = $r = \Sigma (xy) / \mathrm{sqrt} [(\Sigma x^2) * (\Sigma y^2)]$
- Linear correlation (sample data) = $r = [1 / (n - 1)] * \Sigma \{ [(x_i - \bar{x}) / s_x] * [(y_i - \bar{y}) / s_y] \}$
- Linear correlation (population data) = $\rho = [1 / N] * \Sigma \{ [(X_i - \mu_X) / \sigma_x] * [(Y_i - \mu_Y) / \sigma_y] \}$

Simple Linear Regression

- Simple linear regression line: $\hat{y} = b_0 + b_1x$
- Regression coefficient = $b_1 = \Sigma [(x_i - \bar{x}) (y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$
- Regression slope intercept = $b_0 = \bar{y} - b_1 * \bar{x}$
- Regression coefficient = $b_1 = r * (s_y / s_x)$
- Standard error of regression slope = $s_{b_1} = \mathrm{sqrt} [\Sigma(y_i - \hat{y}_i)^2 / (n - 2)] / \mathrm{sqrt} [\Sigma(x_i - \bar{x})^2]$

Counting

- n factorial: $n! = n * (n-1) * (n - 2) * \ldots * 3 * 2 * 1$. By convention, $0! = 1$.
- Permutations of *n* things, taken *r* at a time: ${}_nP_r = n! / (n - r)!$
- Combinations of *n* things, taken *r* at a time: ${}_nC_r = n! / r!(n - r)! = {}_nP_r / r!$

Probability

- Rule of addition: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Rule of multiplication: $P(A \cap B) = P(A) P(B|A)$
- Rule of subtraction: $P(A') = 1 - P(A)$

Random Variables

In the following formulas, *X* and *Y* are random variables, and *a* and *b* are constants.

- Expected value of X = $E(X) = \mu_x = \Sigma [x_i * P(x_i)]$
- Variance of X = $Var(X) = \sigma^2 = \Sigma [x_i - E(x)]^2 * P(x_i) = \Sigma [x_i - \mu_x]^2 * P(x_i)$
- Normal random variable = z-score = $z = (X - \mu)/\sigma$
- Chi-square statistic = $X^2 = [(n - 1) * s^2] / \sigma^2$
- f statistic = $f = [s_1^2/\sigma_1^2] / [s_2^2/\sigma_2^2]$
- Expected value of sum of random variables = $E(X + Y) = E(X) + E(Y)$
- Expected value of difference between random variables = $E(X - Y) = E(X) - E(Y)$
- Variance of the sum of *independent* random variables = $Var(X + Y) = Var(X) + Var(Y)$
- Variance of the difference between *independent* random variables = $Var(X - Y) = Var(X) + Var(Y)$

Sampling Distributions

- Mean of sampling distribution of the mean = $\mu_{\bar{x}} = \mu$
- Mean of sampling distribution of the proportion = $\mu_p = P$
- Standard deviation of proportion = $\sigma_p = \mathrm{sqrt}[P * (1 - P)/n] = \mathrm{sqrt}(PQ / n)$
- Standard deviation of the mean = $\sigma_{\bar{x}} = \sigma/\mathrm{sqrt}(n)$
- Standard deviation of difference of sample means = $\sigma_d = \mathrm{sqrt}[(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)]$
- Standard deviation of difference of sample proportions = $\sigma_d = \mathrm{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$

Standard Error

- Standard error of proportion = $SE_p = s_p = \mathrm{sqrt}[p * (1 - p)/n] = \mathrm{sqrt}(pq / n)$
- Standard error of difference for proportions = $SE_p = s_p = \mathrm{sqrt}\{ p * (1 - p) * [(1/n_1) + (1/n_2)] \}$
- Standard error of the mean = $SE_{\bar{x}} = s_{\bar{x}} = s/\mathrm{sqrt}(n)$
- Standard error of difference of sample means = $SE_d = s_d = \mathrm{sqrt}[(s_1^2 / n_1) + (s_2^2 / n_2)]$
- Standard error of difference of paired sample means = $SE_d = s_d = \{ \mathrm{sqrt} [(\Sigma(d_i - \bar{d})^2 / (n - 1)] \} / \mathrm{sqrt}(n)$
- Pooled sample standard error = $s_{pooled} = \mathrm{sqrt} [(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2] / (n_1 + n_2 - 2)]$
- Standard error of difference of sample proportions = $s_d = \mathrm{sqrt}\{ [p_1(1 - p_1) / n_1] + [p_2(1 - p_2) / n_2] \}$

Discrete Probability Distributions

- Binomial formula: $P(X = x) = b(x; n, P) = {}_nC_x * P^x * (1 - P)^{n - x} = {}_nC_x * P^x * Q^{n - x}$
- Mean of binomial distribution = $\mu_x = n * P$
- Variance of binomial distribution = $\sigma_x^2 = n * P * (1 - P)$
- Negative Binomial formula: $P(X = x) = b^*(x; r, P) = {}_{x-1}C_{r-1} * P^r * (1 - P)^{x - r}$
- Mean of negative binomial distribution = $\mu_x = rQ / P$
- Variance of negative binomial distribution = $\sigma_x^2 = r * Q / P^2$
- Geometric formula: $P(X = x) = g(x; P) = P * Q^{x - 1}$
- Mean of geometric distribution = $\mu_x = Q / P$
- Variance of geometric distribution = $\sigma_x^2 = Q / P^2$
- Hypergeometric formula: $P(X = x) = h(x; N, n, k) = [{}_kC_x] [{}_{N-k}C_{n-x}] / [{}_NC_n]$
- Mean of hypergeometric distribution = $\mu_x = n * k / N$
- Variance of hypergeometric distribution = $\sigma_x^2 = n * k * (N - k) * (N - n) / [N^2 * (N - 1)]$
- Poisson formula: $P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$
- Mean of Poisson distribution = $\mu_x = \mu$
- Variance of Poisson distribution = $\sigma_x^2 = \mu$
- Multinomial formula: $P = [n! / (n_1! * n_2! * \ldots n_k!)] * (p_1^{n_1} * p_2^{n_2} * \ldots * p_k^{n_k})$

Linear Transformations

For the following formulas, assume that Y is a linear transformation of the random variable X, defined by the equation: Y = aX + b.

- Mean of a linear transformation = $E(Y) = \bar{Y} = a\bar{X} + b$.
- Variance of a linear transformation = $Var(Y) = a^2 * Var(X)$.
- Standardized score = $z = (x - \mu_x) / \sigma_x$.
- t statistic = $t = (x - \mu_x) / [s/\mathrm{sqrt}(n)]$.

Estimation

- Confidence interval: Sample statistic $\pm$ Critical value * Standard error of statistic
- Margin of error = (Critical value) * (Standard deviation of statistic)
- Margin of error = (Critical value) * (Standard error of statistic)

Hypothesis Testing

- Standardized test statistic = (Statistic - Parameter) / (Standard deviation of statistic)
- One-sample z-test for proportions: z-score = $z = (p - P_0) / \mathrm{sqrt}(p * q / n)$
- Two-sample z-test for proportions: z-score = $z = z = [(p_1 - p_2) - d] / SE$
- One-sample t-test for means: t statistic = $t = (\bar{x} - \mu) / SE$
- Two-sample t-test for means: t statistic = $t = [(\bar{x}_1 - \bar{x}_2) - d] / SE$
- Matched-sample t-test for means: t statistic = $t = [(\bar{x}_1 - \bar{x}_2) - D] / SE = (\bar{d} - D) / SE$
- Chi-square test statistic = $X^2 = \Sigma[(Observed - Expected)^2 / Expected]$

Degrees of Freedom

The correct formula for degrees of freedom (DF) depends on the situation (the nature of the test statistic, the number of samples, underlying assumptions, etc.).

- One-sample t-test: DF = n - 1
- Two-sample t-test: $DF = (s_1^2/n_1 + s_2^2/n_2)^2 / \{ [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)] \}$
- Two-sample t-test, pooled standard error: $DF = n_1 + n_2 - 2$
- Simple linear regression, test slope: DF = n - 2
- Chi-square goodness of fit test: DF = k - 1
- Chi-square test for homogeneity: DF = (r - 1) * (c - 1)
- Chi-square test for independence: DF = (r - 1) * (c - 1)

Sample Size

Below, the first two formulas find the smallest sample sizes required to achieve a fixed margin of error, using simple random sampling. The third formula assigns sample to strata, based on a proportionate design. The fourth formula, Neyman allocation, uses stratified sampling to minimize variance, given a fixed sample size. And the last formula, optimum allocation, uses stratified sampling to minimize variance, given a fixed budget.

- Mean (simple random sampling): $n = \{ z^2 * \sigma^2 * [N / (N - 1)] \} / \{ ME^2 + [z^2 * \sigma^2 / (N - 1)] \}$
- Proportion (simple random sampling): $n = [(z^2 * p * q) + ME^2] / [ME^2 + z^2 * p * q / N]$
- Proportionate stratified sampling: $n_h = (N_h / N) * n$
- Neyman allocation (stratified sampling): $n_h = n * (N_h * \sigma_h) / [\Sigma (N_i * \sigma_i)]$
- Optimum allocation (stratified sampling):
  $n_h = n * [(N_h * \sigma_h) / \mathrm{sqrt}(c_h)] / [\Sigma (N_i * \sigma_i) / \mathrm{sqrt}(c_i)]$