

# STATISTICAL FORMULAE

## DESCRIPTIVE STATISTICS

Create a dataset and store it in a variable:

$x = c(x_1, x_2, x_3, \dots)$

### Sample Statistics

Sample mean:  $\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k$  mean(x)

Sample variance:  $s^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2$  var(x)

Sample standard deviation:  $s = \sqrt{\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x})^2}$  sd(x)

Pearson Correlation Coefficient:  $r = \frac{1}{n-1} \sum_{k=1}^n \left( \frac{x_k - \bar{x}}{s_x} \right) \left( \frac{y_k - \bar{y}}{s_y} \right)$  cor(x, y)

### Simple Linear Regression Line

Slope:  $b_1 = r \frac{s_y}{s_x}$       y-intercept:  $b_0 = \bar{y} - b_1 \bar{x}$  lm(y ~ x)

### Measures of Position

z-score:  $z = \frac{x - \bar{x}}{s}$

Sample quantiles: quantile(x, q)

Empirical cdf: ecdf(x)(d)

Lower and Upper Fences:  $Q_1 - \frac{3}{2}IQR$     and     $Q_3 + \frac{3}{2}IQR$

### Empirical Rule

If a distribution is approximately normal, then approximately 68% of all data are within 1 standard deviation of the mean, approximately 95% of all data are within 2 standard deviations of the mean, and almost all of the data are within 3 standard deviations of the mean. This can also be remembered as 34% - 13.5% - 2.5% rule.

### Chebyshev's Inequality

For any distribution with finite non-zero variance and any real  $k > 1$ , at least  $\left(1 - \frac{1}{k^2}\right) 100\%$  of all observations fall within  $k$  standard deviations of the mean.

## PROBABILITY AND COUNTING

### Set Notation

$A'$  is the event “not  $A$ ”

$A \cap B$  is the event “ $A$  and  $B$ ”

$A \cup B$  is the event “ $A$  or  $B$  or both”

### Kolmogorov Axioms

(1) If  $E$  is any event, then  $P(E) \geq 0$

(2) If  $S$  is a sample space, then  $P(S) = 1$

(3) If  $A$  and  $B$  are disjoint, then  $P(A \cup B) = P(A) + P(B)$

### Rule of complements

$$P(A') = 1 - P(A)$$

### General Addition Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

### Independence Criterion

$$P(A \cap B) = P(A)P(B)$$

### Conditional Probability

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

### General Multiplication Rule

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

### Counting without Replacement

The number of Permutations:  ${}_N P_n = \frac{N!}{(N-n)!}$

The number of Combinations:  ${}_N C_n = \frac{N!}{n!(N-n)!}$

choose(N, n)

## DISCRETE RANDOM VARIABLES

### Mean (Expected Value)

$$\mu_X = EX = \sum_k x_k P(X = x_k)$$

### Variance

$$\sigma_X^2 = \text{Var}(X) = E(X^2) - (EX)^2$$

## BINOMIAL DISTRIBUTION

$$X \sim \text{Bin}(n, p)$$

$$\mu_X = np$$

$$\sigma_X = \sqrt{np(1-p)}$$

Binomial pmf:

`dbinom(x, n, p)`

Binomial cdf:

`pbinom(x, n, p)`

Binomial inverse cdf:

`qbinom(q, n, p)`

## NORMAL DISTRIBUTION

$$X \sim N(\mu, \sigma)$$

Normal cdf:

`pnorm(x, mu, sigma)`

Normal inverse cdf:

`qnorm(q, mu, sigma)`

### Sampling Distribution of the Sample Mean

$$\bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}})$$

$$\mu_{\bar{X}} = \mu_X$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

## CONFIDENCE INTERVALS

### Population Proportion

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

Clopper-Pearson 'exact' binomial interval: `binom.test(x, n, conf.level=)`

Minimal sample size for obtaining a confidence interval for the population proportion, given the desired level of significance  $\alpha$  and margin of error  $E$  (use  $\hat{p} = 0.5$  if the estimate is unavailable):

$$n = \left\lceil \hat{p}(1-\hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2 \right\rceil$$

### Population Mean

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \quad \text{t.test(x, conf.level=)}$$

Minimal sample size for obtaining a confidence interval for the population mean, given the standard deviation estimate  $s$ , the desired level of significance  $\alpha$ , and the margin of error  $E$ :

$$n = \left\lceil \left( \frac{z_{\alpha/2} \cdot s}{E} \right)^2 \right\rceil$$

### Population Standard Deviation

Given a normal population or a large sample size:

$$\left( \sqrt{\frac{(n-1)s^2}{\chi_{\alpha/2}^2}}, \sqrt{\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}} \right)$$

### Difference Between Proportions

Normal approximation:  $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Chi-squared test: `prop.test(c(x1, x2), c(n1, n2), conf.level=)`

### Difference Between Means

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \text{t.test(x, y, conf.level=)}$$

## HYPOTHESIS TESTING

### Proportion

Normal approximation test statistic:

$$z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Clopper-Pearson 'exact' binomial test:

`binom.test(x, n, p=, conf.level=)`

### Mean

Test statistic:

$$t_0 = \frac{\bar{x} - \mu_0}{s} \cdot \sqrt{n} \quad \text{where df} = n - 1$$

*t*-test:

`t.test(x, mu=, conf.level=)`

### Standard Deviation

Test statistic:

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \quad \text{where df} = n - 1$$

### Independence

Test statistic:

$$\chi_0^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \quad \text{where df} = (r-1)(c-1)$$

Chi-squared test:

`chisq.test(m)`

Where *m* is the matrix with data:

`m = matrix(c(x1, x2, ...), ncol=)`

### Correlation

Test statistic:

$$t_0 = r \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \quad \text{where df} = n - 2$$

Pearson *t*-test:

`cor.test(x, y, conf.level=)`

## Difference Between Proportions

Normal approximation test statistic:  $z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$  where  $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$

Chi-squared test: `prop.test(c(x1, x2), c(n1, n2), conf.level=)`

## Difference Between Means

Test statistic:  $t_0 = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$  where  $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}$

*t*-test: `t.test(x, y, mu=, conf.level=)`

## R DISPLAYS

Barplot:	<code>barplot(x)</code>
Histogram:	<code>hist(x)</code>
Stem-and-leaf plot:	<code>stem(x)</code>
Box-and-whiskers plot:	<code>boxplot(x)</code>
Scatter plot:	<code>plot(x, y)</code>
Linear regression line:	<code>abline(lm(y ~ x))</code>
QQ plot:	<code>qqplot(x, y)</code> <code>qqline(x, y)</code>

## R ARITHMETIC

$a + b$	<code>a+b</code>
$a - b$	<code>a-b</code>
$ab$	<code>a*b</code>
$a/b$	<code>a/b</code>
$a^b$	<code>a^b</code>
$\sqrt{a}$	<code>sqrt(a)</code>