# Statistical Formulae

## Descriptive Statistics

Create a dataset and store it in a variable: `x = c(x1, x2, x3, ...)`

### Sample Statistics

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

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

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

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

### Simple Linear Regression Line

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

### Measures of Position

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

Sample quantiles: `quantile(x, q)`

Empirical cdf: `ecdf(x)(d)`

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

### 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**

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

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

$\sigma_{\overline{X}} = \dfrac{\sigma_X}{\sqrt{n}}$

## 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

$\overline{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$ `t.test(data, 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:

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

### 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

$\overline{x}_1 - \overline{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ `t.test(data1, data2, conf.level=)`

---

$z_{\alpha/2} =$ `qnorm(1-α/2)` $t_{\alpha/2,n-1} =$ `qt(1-α/2, n-1)` $\chi^2_{\alpha/2,n-1} =$ `qchisq(1-α/2, n-1)`

# 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=)`

**Mean**

Test statistic: $$t_0 = \frac{\overline{x} - \mu_0}{s} \cdot \sqrt{n} \qquad \text{where df} = n - 1$$

$t$-test: `t.test(data, mu=)`

**Standard Deviation**

Test statistic: $$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \qquad \text{where df} = n - 1$$

**Independence**

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

Chi-squared test: `chisq.test(m)`

Where `m` is the matrix with data: `m = matrix(c(x1, x2, ...), ncol=)`

**Correlation**

Test statistic: $$t_0 = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{where df} = n - 2$$

Pearson $t$-test: `cor.test(data1, data2)`

**Difference Between Proportions**

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

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

**Difference Between Means**

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

$t$-test: `t.test(data1, data2, mu=)`

## R Displays

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

## R Arithmetic

| | |
|---|---|
| $a + b$ | `a+b` |
| $a - b$ | `a-b` |
| $ab$ | `a*b` |
| $a/b$ | `a/b` |
| $a^b$ | `a^b` |
| $\sqrt{a}$ | `sqrt(a)` |