# KEY FORMULAS

## Prem S. Mann • Introductory Statistics, Fourth Edition

### CHAPTER 2 • ORGANIZING DATA

- Relative frequency of a class $= f/\Sigma f$
- Percentage of a class = (Relative frequency) × 100
- Class midpoint or mark = (Upper limit + Lower limit)/2
- Class width = Upper boundary − Lower boundary
- Cumulative relative frequency

$$= \frac{\text{Cumulative frequency}}{\text{Total observations in the data set}}$$

- Cumulative percentage

$$= (\text{Cumulative relative frequency}) \times 100$$

### CHAPTER 3 • NUMERICAL DESCRIPTIVE MEASURES

- Mean for ungrouped data: $\mu = \Sigma x/N$ and $\bar{x} = \Sigma x/n$
- Mean for grouped data: $\mu = \Sigma mf/N$ and $\bar{x} = \Sigma mf/n$ where $m$ is the midpoint and $f$ is the frequency of a class
- Median for ungrouped data

$$= \text{Value of the } \left(\frac{n+1}{2}\right)\text{th term in a ranked data set}$$

- Range = Largest value − Smallest value
- Variance for ungrouped data:

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

where $\sigma^2$ is the population variance and $s^2$ is the sample variance

- Standard deviation for ungrouped data:

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

where $\sigma$ and $s$ are the population and sample standard deviations, respectively

- Variance for grouped data:

$$\sigma^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n-1}$$

- Standard deviation for grouped data:

$$\sigma = \sqrt{\frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n-1}}$$

- Chebyshev's theorem:
  For any number $k$ greater than 1, at least $(1 - 1/k^2)$ of the values for any distribution lie within $k$ standard deviations of the mean.
- Empirical rule:
  For a specific bell-shaped distribution, about 68% of the observations fall in the interval $(\mu - \sigma)$ to $(\mu + \sigma)$, about 95% fall in the interval $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$, and about 99.7% fall in the interval $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$.
- Interquartile range: $\text{IQR} = Q_3 - Q_1$
  where $Q_3$ is the third quartile and $Q_1$ is the first quartile
- The $k$th percentile:

$$P_k = \text{Value of the } \left(\frac{kn}{100}\right)\text{th term in a ranked data set}$$

- Percentile rank of $x_i$

$$= \frac{\text{Number of values less than } x_i}{\text{Total number of values in the data set}} \times 100$$

### CHAPTER 4 • PROBABILITY

- Classical probability rule for a simple event:

$$P(E_i) = \frac{1}{\text{Total number of outcomes}}$$

- Classical probability rule for a compound event:

$$P(A) = \frac{\text{Number of outcomes in } A}{\text{Total number of outcomes}}$$

- Relative frequency as an approximation of probability:

$$P(A) = \frac{f}{n}$$

- Conditional probability of an event:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

- Condition for independence of events:

$$P(A) = P(A|B) \quad \text{and/or} \quad P(B) = P(B|A)$$

- For complementary events: $P(A) + P(\bar{A}) = 1$
- Multiplication rule for dependent events:

$$P(A \text{ and } B) = P(A)\, P(B|A)$$

- Multiplication rule for independent events:

$$P(A \text{ and } B) = P(A)\, P(B)$$

- Joint probability of two mutually exclusive events:

$$P(A \text{ and } B) = 0$$

- Addition rule for mutually nonexclusive events:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- Addition rule for mutually exclusive events:

$$P(A \text{ or } B) = P(A) + P(B)$$

### CHAPTER 5 • DISCRETE RANDOM VARIABLES AND THEIR PROBABILITY DISTRIBUTIONS

- Mean of a discrete random variable $x$: $\mu = \Sigma x P(x)$
- Standard deviation of a discrete random variable $x$:

$$\sigma = \sqrt{\Sigma x^2 P(x) - \mu^2}$$

- $n$ factorial: $n! = n(n-1)(n-2)\ldots 3 \cdot 2 \cdot 1$
- Number of combinations of $n$ items selected $x$ at a time:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- Binomial probability formula: $P(x) = \binom{n}{x} p^x q^{n-x}$
- Mean and standard deviation of the binomial distribution:

$$\mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

- Poisson probability formula: $P(x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$
- Mean, variance, and standard deviation of the Poisson probability distribution:

$$\mu = \lambda, \quad \sigma^2 = \lambda, \quad \text{and} \quad \sigma = \sqrt{\lambda}$$

### CHAPTER 6 • CONTINUOUS RANDOM VARIABLES AND THE NORMAL DISTRIBUTION

- $z$ value for an $x$ value: $z = \dfrac{x - \mu}{\sigma}$
- Value of $x$ when $\mu$, $\sigma$, and $z$ are known: $x = \mu + z\sigma$

### CHAPTER 7 • SAMPLING DISTRIBUTIONS

- Mean of $\bar{x}$: $\mu_{\bar{x}} = \mu$
- Standard deviation of $\bar{x}$ when $n/N \le .05$: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- $z$ value for $\bar{x}$: $z = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$
- Population proportion: $p = X/N$
- Sample proportion: $\hat{p} = x/n$
- Mean of $\hat{p}$: $\mu_{\hat{p}} = p$
- Standard deviation of $\hat{p}$ when $n/N \le .05$: $\sigma_{\hat{p}} = \sqrt{pq/n}$
- $z$ value for $\hat{p}$: $z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}}$

### CHAPTER 8 • ESTIMATION OF THE MEAN AND PROPORTION

- Margin of error for the point estimation of $\mu$:

$$\pm 1.96\, \sigma_{\bar{x}} \quad \text{or} \quad \pm 1.96\, s_{\bar{x}}$$

where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $s_{\bar{x}} = s/\sqrt{n}$

- Confidence interval for $\mu$ for a large sample:

$$\bar{x} \pm z\, \sigma_{\bar{x}} \quad \text{if } \sigma \text{ is known}$$

$$\bar{x} \pm z\, s_{\bar{x}} \quad \text{if } \sigma \text{ is not known}$$

where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $s_{\bar{x}} = s/\sqrt{n}$

- Confidence interval for $\mu$ for a small sample:

$$\bar{x} \pm t s_{\bar{x}} \quad \text{where} \quad s_{\bar{x}} = s/\sqrt{n}$$

- Margin of error for the point estimation of $p$:

$$\pm 1.96 s_{\hat{p}} \quad \text{where} \quad s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$$

- Confidence interval for $p$ for a large sample:

$$\hat{p} \pm z s_{\hat{p}} \quad \text{where} \quad s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$$

- Maximum error of the estimate for $\mu$:

$$E = z\sigma_{\bar{x}} \quad \text{or} \quad z s_{\bar{x}}$$

- Determining sample size for estimating $\mu$: $n = z^2\sigma^2/E^2$
- Maximum error of the estimate for $p$:

$$E = z s_{\hat{p}} \quad \text{where} \quad s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$$

- Determining sample size for estimating $p$: $n = z^2 pq/E^2$

### CHAPTER 9 • HYPOTHESIS TESTS ABOUT THE MEAN AND PROPORTION

- Test statistic $z$ for a test of hypothesis about $\mu$ for a large sample:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \text{if } \sigma \text{ is known, where } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

or

$$z = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{if } \sigma \text{ is not known, where } s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Test statistic for a test of hypothesis about $\mu$ for a small sample:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad \text{where} \quad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Test statistic for a test of hypothesis about $p$ for a large sample:

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} \quad \text{where} \quad \sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

### CHAPTER 10 • ESTIMATION AND HYPOTHESIS TESTING: TWO POPULATIONS

- Mean of the sampling distribution of $\bar{x}_1 - \bar{x}_2$:

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

- Confidence interval for $\mu_1 - \mu_2$ for two large and independent samples:

$$(\bar{x}_1 - \bar{x}_2) \pm z\sigma_{\bar{x}_1 - \bar{x}_2} \quad \text{if } \sigma_1 \text{ and } \sigma_2 \text{ are known}$$

or

$$(\bar{x}_1 - \bar{x}_2) \pm z s_{\bar{x}_1 - \bar{x}_2} \quad \text{if } \sigma_1 \text{ and } \sigma_2 \text{ are not known}$$

where

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{and} \quad s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- Test statistic for a test of hypothesis about $\mu_1 - \mu_2$ for two large and independent samples:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

If $\sigma_1$ and $\sigma_2$ are not known, then replace $\sigma_{\bar{x}_1 - \bar{x}_2}$ by its point estimator $s_{\bar{x}_1 - \bar{x}_2}$.

- For two small and independent samples taken from two populations with equal standard deviations:

  Pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

  Estimate of the standard deviation of $\bar{x}_1 - \bar{x}_2$:

$$s_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

  Confidence interval for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t s_{\bar{x}_1 - \bar{x}_2}$

  Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$

- For two small and independent samples selected from two populations with unequal standard deviations:

  Degrees of freedom: $df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

  Estimate of the standard deviation of $\bar{x}_1 - \bar{x}_2$:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

  Confidence interval for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t s_{\bar{x}_1 - \bar{x}_2}$

  Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$

- For two paired or matched samples:

  Sample mean for paired differences: $\bar{d} = \frac{\Sigma d}{n}$

  Sample standard deviation for paired differences:

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n - 1}}$$

  Mean and standard deviation of the sampling distribution of $\bar{d}$:

$$\mu_{\bar{d}} = \mu_d \quad \text{and} \quad s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

  Confidence interval for $\mu_d$:

$$\bar{d} \pm t\, s_{\bar{d}} \quad \text{where} \quad s_{\bar{d}} = \frac{s_d}{\sqrt{n}}$$

  Test statistic for a test of hypothesis about $\mu_d$:

$$t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$$

- For two large and independent samples, confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z s_{\hat{p}_1 - \hat{p}_2}$$

  where $s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

- For two large and independent samples, for a test of hypothesis about $p_1 - p_2$ with $H_0$: $p_1 - p_2 = 0$:

  Pooled sample proportion:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \text{or} \quad \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

  Estimate of the standard deviation of $\hat{p}_1 - \hat{p}_2$:

$$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  Test statistic: $z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{s_{\hat{p}_1 - \hat{p}_2}}$

## CHAPTER 11 • CHI-SQUARE TESTS

- Expected frequency for a category for a goodness-of-fit test:

$$E = np$$

- Degrees of freedom for a goodness-of-fit test:

$$df = k - 1 \quad \text{where } k \text{ is the number of categories}$$

- Expected frequency for a cell for an independence or homogeneity test:

$$E = \frac{\text{(Row total)(Column total)}}{n}$$

- Degrees of freedom for a test of independence or homogeneity:

$$df = (R - 1)(C - 1)$$

  where $R$ and $C$ are the total number of rows and columns, respectively, in the contingency table

- Test statistic for a goodness-of-fit test and a test of independence or homogeneity:

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

- Confidence interval for the population variance $\sigma^2$:

$$\frac{(n - 1)s^2}{\chi^2_{\alpha/2}} \quad \text{to} \quad \frac{(n - 1)s^2}{\chi^2_{1-\alpha/2}}$$

- Test statistic for a test of hypothesis about $\sigma^2$:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

## CHAPTER 12 • ANALYSIS OF VARIANCE

Let:

$k$ = the number of different samples (or treatments)
$n_i$ = the size of sample $i$
$T_i$ = the sum of the values in sample $i$
$n$ = the number of values in all samples $= n_1 + n_2 + n_3 + \cdots$
$\Sigma x$ = the sum of the values in all samples $= T_1 + T_2 + T_3 + \cdots$
$\Sigma x^2$ = the sum of the squares of values in all samples

- For the $F$ distribution:

  Degrees of freedom for the numerator $= k - 1$
  Degrees of freedom for the denominator $= n - k$

- Between-samples sum of squares:

$$\text{SSB} = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right) - \frac{(\Sigma x)^2}{n}$$

- Within-samples sum of squares:

$$\text{SSW} = \Sigma x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right)$$

- Total sum of squares: $\text{SST} = \text{SSB} + \text{SSW} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$
- Variance between samples: $\text{MSB} = \frac{\text{SSB}}{k - 1}$
- Variance within samples: $\text{MSW} = \frac{\text{SSW}}{n - k}$
- Test statistic for a one-way ANOVA test: $F = \frac{\text{MSB}}{\text{MSW}}$

## CHAPTER 13 • SIMPLE LINEAR REGRESSION

- Simple linear regression model: $y = A + Bx + \epsilon$
- Estimated simple linear regression model: $\hat{y} = a + bx$
- Sum of squares of $xy$, $xx$, and $yy$:

$$\text{SS}_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

$$\text{SS}_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \quad \text{and} \quad \text{SS}_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

- Least squares estimates of $A$ and $B$:

$$b = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

- Standard deviation of the sample errors:

$$s_e = \sqrt{\frac{\text{SS}_{yy} - b\,\text{SS}_{xy}}{n - 2}}$$

- Error sum of squares: $\text{SSE} = \Sigma e^2 = \Sigma(y - \hat{y})^2$
- Total sum of squares: $\text{SST} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$
- Regression sum of squares: $\text{SSR} = \text{SST} - \text{SSE}$
- Coefficient of determination: $r^2 = b\frac{\text{SS}_{xy}}{\text{SS}_{yy}}$
- Confidence interval for $B$:

$$b \pm t\, s_b \quad \text{where} \quad s_b = \frac{s_e}{\sqrt{\text{SS}_{xx}}}$$

- Test statistic for a test of hypothesis about $B$: $t = \frac{b - B}{s_b}$
- Linear correlation coefficient: $r = \frac{\text{SS}_{xy}}{\sqrt{\text{SS}_{xx}\,\text{SS}_{yy}}}$
- Test statistic for a test of hypothesis about $\rho$: $t = r\sqrt{\frac{n - 2}{1 - r^2}}$
- Confidence interval for $\mu_{y|x}$:

$$\hat{y} \pm t s_{\hat{y}_m} \quad \text{where} \quad s_{\hat{y}_m} = s_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\text{SS}_{xx}}}$$

- Prediction interval for $y_p$:

$$\hat{y} \pm t\, s_{\hat{y}_p} \quad \text{where} \quad s_{\hat{y}_p} = s_e\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\text{SS}_{xx}}}$$

## CHAPTER 14 • NONPARAMETRIC METHODS

- Test statistic for a sign test about the population proportion for a large sample:

$$z = \frac{(X \pm .5) - \mu}{\sigma} \quad \text{where} \quad \mu = np \quad \text{and} \quad \sigma = \sqrt{npq}$$

  Here, use $(X + .5)$ if $X \le n/2$ and $(X - .5)$ if $X > n/2$

- Test statistic for a sign test about the median for a large sample:

$$z = \frac{(X + .5) - \mu}{\sigma}$$

  where $\mu = np$, $p = .5$, and $\sigma = \sqrt{npq}$

  Let $X_1$ be the number of plus signs and $X_2$ the number of minus signs in a test about the median. Then, if the test is two-tailed, either of the two values can be assigned to $X$; if the test is left-tailed, $X$ = smaller of the values of $X_1$ and $X_2$; if the test is right-tailed, $X$ = larger of the values of $X_1$ and $X_2$.

- Test statistic for the Wilcoxon signed-rank test for a large sample:

$$z = \frac{T - \mu_T}{\sigma_T}$$

  where $\mu_T = \frac{n(n + 1)}{4}$ and $\sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$

- Test statistic for the Wilcoxon rank sum test for large and independent samples:

$$z = \frac{T - \mu_T}{\sigma_T}$$

  where $\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2}$ and $\sigma_T = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$

- Test statistic for the Kruskal–Wallis test:

$$H = \frac{12}{n(n + 1)}\left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \cdots + \frac{R_k^2}{n_k}\right] - 3(n + 1)$$

  where $n_i$ = size of $i$th sample, $n = n_1 + n_2 + \cdots + n_k$, $k$ = number of samples, and $R_i$ = sum of ranks for $i$th sample

- Test statistic for the Spearman rho rank correlation coefficient test:

$$r_s = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

  where $d_i$ = difference in ranks of $x_i$ and $y_i$

- Test statistic for the runs test for randomness for a large sample:

$$z = \frac{R - \mu_R}{\sigma_R}$$

  where $R$ is the number of runs in the sequence

$$\mu_R = \frac{2n_1 n_2}{n_1 + n_2} + 1 \quad \text{and} \quad \sigma_R = \sqrt{\frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$$