

FORMULAE FOR BASIC STATISTICS

A. Descriptive Statistics (for Samples)

A.1 Mean and variance are

$$\bar{y} = \sum y / n; \quad s^2 = TSS / (n - 1), \quad TSS = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 / n.$$

A.2 Quartiles: $R_\alpha = \alpha (1+n) / 4 = i + d$, $\alpha = 1, 2, 3$; $Q_\alpha = (1-d)y_{(i)} + dy_{(i+1)}$
where i is the largest integer not exceeding R_α .

A.3 Mean and the variance for grouped data:

$$\bar{y} = \sum yf / n; \quad s^2 = TSS / (n - 1), \quad TSS = \sum y^2 f - (\sum yf)^2 / n.$$

A.4 Coefficient of Variation : $CV = s / \bar{y}$.

A.5 Coefficient of Skewness : $CS = 3(\bar{y} - \tilde{y}) / s$.

B. Glossary of Probability of Set Events (Two Sets)

	Verbal Description of Event	Probability
B.1	A but not B (=Only A = A alone)	$P(A \cap \bar{B}) = P(A) - P(A \cap B)$
B.2	B but not A (=Only B = B alone)	$P(\bar{A} \cap B) = P(B) - P(A \cap B)$
B.3	None (=Neither A nor B)	$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$
B.4	Exactly one	$P(A \cap \bar{B}) + P(\bar{A} \cap B)$
B.5	Both (=Exactly two = Two)	$P(A \cap B) = 1 - P(A \cap \bar{B}) = 1 - P(\bar{A} \cup \bar{B})$ $P(A \cap B) = P(B)P(A B) = P(A)P(B A)$ $P(A \cap B) = P(A)P(B)$ iff A and B are independent
B.6	Not both	$P(\overline{A \cap B}) = 1 - P(A \cap B)$
B.7	A given B	$P(A B) = \frac{P(A \cap B)}{P(B)}$ if $P(B) \neq 0$ $P(A B) = P(A)$ iff A and B are independent
B.8	At least one of the two (= A or B)	$P(A \cup B) = P(\text{exactly one}) + P(\text{exactly two})$ $= P(A \cap \bar{B}) + P(\bar{A} \cap B) + P(A \cap B)$ $= P(A) + P(B) - P(A \cap B)$ $= 1 - P(\overline{A \cup B}) = 1 - P(\bar{A} \cap \bar{B})$

Independence: $P(A|\bar{B}) = P(A|B) = P(A)$, or, $P(B|\bar{A}) = P(B|A) = P(B)$

For intersection of three sets A, B and C , the following results are important:

$$B.9 \quad P(\text{none}) + P(\text{at least one}) = 1$$

$$B.10 \quad P(\text{none}) = P(\bar{A} \cap \bar{B} \cap \bar{C}) = (P(\bar{A})P(\bar{B})P(\bar{C})) \text{ by independence}$$

$$B.11 \quad P(\text{at least one}) = P(A \cup B \cup C) \\ = P(A) + P(B) + P(C) - [P(A \cap B) + P(A \cap C) + P(B \cap C)] + P(A \cap B \cap C)$$

C. Discrete Probability Distributions

$$C.1 \quad P(a \leq Y \leq b) = \sum_y f(y), \quad F(b) = \sum_{y \leq b} f(y)$$

$$C.2 \quad \mu = E(Y) = \sum_y y f(y), \quad (p89)$$

$$C.3 \quad E(Y^2) = \sum y^2 f(y), \quad \sigma^2 = E(Y - \mu)^2 = E(Y^2) - \mu^2, \quad (p96)$$

	Probability Density function $p(x)$	Mean (μ) and Variance(σ^2)
C.4	The Binomial Distribution: $B(n, p)$ (p119) $f(y) = \binom{n}{y} p^y (1-p)^{n-y}; \quad y = 0, 1, \dots, n$	$\mu = E(Y) = np$ $\sigma^2 = V(Y) = np(1-p).$
C.5	The Geometric Distribution $f(y) = q^{y-1} p; \quad y = 1, 2, \dots$	$\mu = E(Y) = 1/p$ $\sigma^2 = V(Y) = q/p^2$
C.6	The Hypergeometric Distribution (p128) $f(y) = \binom{K}{y} \binom{N-K}{n-y} \div \binom{N}{n},$ $\max\{0, n - (N - K)\} \leq y \leq \min\{n, K\}$	$\mu = E(Y) = np, \quad (p = K/N)$ $\sigma^2 = V(Y) = (1-c) npq, \quad (q = 1-p)$ where $(N-1)c = n-1$
C.7	The Poisson Distribution (p136) $f(y) = \frac{e^{-\lambda t} (\lambda t)^y}{y!}; \quad y = 0, 1, \dots$	$\mu = E(Y) = \lambda t$ $\sigma^2 = V(Y) = \lambda t$

D. Continuous Probability Distributions

For a continuous random variable Y with pdf $f(y)$

$$D.1 \quad P(a < Y < b) = \int_a^b f(y) dy, \quad F(u) = \int_{-\infty}^u f(y) dy$$

$$D.2 \quad \mu = E(Y) = \int_{-\infty}^{\infty} y f(y) dy, \quad (p89)$$

$$D.3 \quad E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y) dx, \quad \sigma^2 = E(Y - \mu)^2 = E(Y^2) - \mu^2, \quad (p96)$$

	<i>Probability Density function</i>	<i>Mean and Variance</i>
D.4	<i>The Normal Distribution: $N(\mu, \sigma^2)$</i>	<i>Mean = $E(Y) = \mu$ Variance = $V(Y) = \sigma^2$</i>
D.5	<i>The Exponential Distribution</i> $f(y) = \frac{1}{\beta} e^{-y/\beta}, 0 < y$	<i>Mean = $E(Y) = \beta$ Variance = $V(Y) = \beta^2$</i>
D.6	<i>Waiting Time Distribution</i> $f(t) = \lambda e^{-\lambda t}, 0 < t$	<i>Mean = $E(T) = 1/\lambda$ Variance = $V(T) = 1/\lambda^2$</i>
D.7	<i>The Gamma Distribution</i> $f(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, 0 < y, 0 < \alpha, 0 < \beta$	<i>Mean = $E(Y) = \alpha\beta$ Variance = $V(Y) = \alpha\beta^2$</i>

E. Sampling Distributions

E.1 Suppose that Y has a distribution with mean μ and variance σ^2 . Additionally if the distribution is normal then $\frac{\sum Y - n\mu}{\sqrt{n\sigma^2}} = \frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} = Z$.

E.2 Suppose that Y has a distribution with mean μ and variance σ^2 . However if the distribution is not normal but $n \geq 30$, then $\frac{\sum Y - n\mu}{\sqrt{nS^2}} = \frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \approx Z$ (p210). This is known as Central Limit Theorem. Note S^2 converges to σ^2 in probability (WLLN).

E.3 The Student T - statistic is defined by $T = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$, with $\nu = n - 1$ (p220)

E.4 The Sampling Distribution of the Proportion (p258)

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{Y/n - p}{\sqrt{p(1-p)/n}} \approx Z$$

F. Statistical Estimation (with Random Sample / Samples)

F.1 Confidence Interval Estimates of the Mean μ

F.1.0 CI for μ , (σ known, any n , normal): $\bar{y} \mp z_{\alpha/2} (\sigma/\sqrt{n})$ for (235)

F.1.1 ACI for μ , (σ known, nonnormal, large n): $\bar{y} \mp z_{\alpha/2} (\sigma/\sqrt{n})$ (cf. p235).

F.1.1(a) sample size for estimating μ : $n = z_{\alpha/2}^2 \sigma^2 / e^2$ (p237).

F.1.2 ACI for μ , (σ unknown, large n , nonnormal): $\bar{y} \mp z_{\alpha/2} (s/\sqrt{n})$ (cf. p235).

F.1.3 CI for μ , (σ unknown, $n \geq 2$, normal): $\bar{y} \mp t_{\alpha/2} (s/\sqrt{n})$ for any $n \geq 2$ (p239).

F.2 Confidence Interval for $\mu_1 - \mu_2$ (Based on Random and Independent Samples)

F.2.3 CI for $\mu_1 - \mu_2$, (small n_i , unknown $\sigma_1^2 = \sigma_2^2$ but unknown, normal):

$$(\bar{y}_1 - \bar{y}_2) \mp t_{\alpha/2} \sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}, \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}, \quad \nu = (n_1 - 1) + (n_2 - 1),$$

where s^2 is the combined variance ($s_1^2 \leq s^2 \leq s_2^2$), (p249)

F.3 Confidence Interval for Proportion p

F.3.1 CI for p when n large: $\hat{p} \mp z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$, (p258)

F.3.1 (a) Large sample size for estimating p : $n = z_{\alpha/2}^2 \hat{p}(1-\hat{p})/e^2$. (p 259)

F.4 Confidence Interval for Variance σ^2 : $[(n-1)s^2/u, (n-1)s^2/l]$

where $l = \chi_{1-\alpha/2}^2 \ll u = \chi_{\alpha/2}^2$

G. Testing of Hypotheses (with Random Sample/ Samples)

Reject H_0 for $p\text{-value} \leq \alpha$; Don't reject H_0 for $0 \leq \alpha < p\text{-value}$.

G.1 Testing of a Mean μ

G.1.0 σ known, normal: $z = \frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}$ (p300)

G.1.1 σ known, large n , nonnormal: $z \approx \frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}$

G.1.2 σ unknown, large sample: $z \approx \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$ (p300)

H_0	H_a	RR (for H_0)	p -value
$\mu = \mu_0$	$\mu < \mu_0$	$z < -z_\alpha$	$P(Z < z)$
$\mu = \mu_0$	$\mu > \mu_0$	$z > z_\alpha$	$P(Z > z)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$2 P(Z > z)$

G.1.3 σ unknown, normal population: $t = \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$, ($\nu = n - 1 \geq 1$) (p304)

H_0	H_a	RR (for H_0)	p -value
$\mu = \mu_0$	$\mu < \mu_0$	$t < -t_\alpha$	$P(T < t)$
$\mu = \mu_0$	$\mu > \mu_0$	$t > t_\alpha$	$P(T > t)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	$2 P(T > t)$

G.2 Testing $\mu_1 - \mu_2 = \delta$ (Random and Independent Samples)

G.2.3 Small n_i , unknown $\sigma_1^2 = \sigma_2^2$, normal)

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{\sqrt{(s^2/n_1) + (s^2/n_2)}}, \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \quad v = (n_1 - 1) + (n_2 - 1),$$

where s^2 is the combined variance ($s_1^2 \leq s^2 \leq s_2^2$), (p308).

H_0	H_a	RR (for H_0)	p-value
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 < \delta$	$t < -t_\alpha$	$P(T < t)$
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 > \delta$	$t > t_\alpha$	$P(T > t)$
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \neq \delta$	$t < -t_{\alpha/2}$ or $t > t_{\alpha/2}$	$2 P(T > t)$

G.3.1 Testing of a proportion (p)

Large sample: $z \approx \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

H_0	H_a	RR (for H_0)	p-value
$p = p_0$	$p < p_0$	$z < -z_\alpha$	$P(Z < z)$
$p = p_0$	$p > p_0$	$z > z_\alpha$	$P(Z > z)$
$p = p_0$	$p \neq p_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$2 P(Z > z)$

H. Linear Regression Analysis (Degrees of Freedom: $v = n - 2$)

H.1 Line of Best Fit

H.1.0 Model: $y = \beta_0 + \beta_1 x + \varepsilon$ for a given x , $Y \sim N(\mu, \sigma^2)$ where $\mu = \beta_0 + \beta_1 x$

Fitted Model: $\hat{y} = \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$ for a given x

H.1.1 $\hat{\beta}_1 = \frac{s_{xy}}{s_{xx}}$; $s_{xy} = \sum xy - (\sum x)(\sum y)/n$; $s_{xx} = \sum x^2 - (\sum x)^2/n$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. (p536)

H.1.2 Pearson product moment coefficient of correlation: $r = \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$ (p392)

H.1.3 Standard error of the estimate: $s = \sqrt{SSE/(n-2)} = \sqrt{MSE}$

H.1.4 $TSS = \sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2/n$; $SSR = \hat{\beta}_1 s_{xy}$; $SSE = TSS - SSR$

H.1.5 $R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{SSR}{TSS}$

H.2 Inference Regarding the Regression Coefficients

$$H.2.1 \ 100(1-\alpha)\% \text{ confidence interval for } \beta_0 : \quad \hat{\beta}_0 \mp t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right) MSE} .$$

$$H.2.2 \ \text{Testing the hypothesis } H_0 : \beta_0 = c : \quad t = \frac{\hat{\beta}_0 - c}{\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right) MSE}} \quad (p364)$$

$$H.2.3 \ 100(1-\alpha)\% \text{ confidence interval for } \beta_1 : \quad \hat{\beta}_1 \mp t_{\alpha/2} \sqrt{MSE / s_{xx}}$$

$$H.2.4 \ \text{Testing the hypothesis } H_0 : \beta_1 = c : \quad t = \frac{\hat{\beta}_1 - c}{\sqrt{MSE / s_{xx}}} \quad (p364)$$

Inference Regarding the Response Variable

H.2.5 Confidence Interval for Mean for a given x :

$$\hat{y} \pm t_{\alpha/2} \sqrt{\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}\right) MSE} ,$$

H.2.6 Prediction Interval for an Individual Y Given x :

$$\hat{y} \pm t_{\alpha/2} \sqrt{\left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}\right) MSE} \quad (p368)$$

Table for Estimation or Tests of Hypotheses on Mean or Difference in Means (In this table sample is always drawn randomly, and two samples considered are always independent)

σ_i^2	n_j	Y	CI	z/t	TEST	page
K N O W N	1	Any, Normal	F.1.0	z	G.1.0	235
	1	Large ($n \geq 30$), nonnormal	F.1.1	$\approx z$	G1.1	235
U N K N O W N	1	Large ($n \geq 30$), Nonnormal	F.1.2	$\approx z$	G.1.2	239
	1	At least 2, Normal	F.1.3	t	G.1.3	239
	2	Each at least 2, Normal ($\sigma_1^2 = \sigma_2^2$)	F.2.3	t	G.2.3	249