FORMULAE FOR STAT 319

(Edited by Anwar Joarder)

A. Descriptive Statistics (for Samples)

A.1 ***Mean*** *and* ***variance*** *are* $\bar{y} = \frac{1}{n} \sum y$ *and*

$s^2 = \frac{TSS}{n-1}$ where $TSS = \sum (y - \bar{y})^2 = \sum y^2 - \frac{1}{n}\left(\sum y\right)^2$.

A.2 ***Quartiles:*** $R_\alpha = \alpha \frac{1+n}{4} = i + d, \quad \alpha = 1,\ 2,\ 3;\ Q_\alpha = (1-d)y_{(i)} + dy_{(i+1)}$.

A.3 ***Mean*** *and the* ***variance*** *for* ***grouped data:***

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

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

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

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

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

Independence:

$$P(A\mid \overline{B}) = P(A\mid B) = P(A)$$
$$P(B\mid \overline{A}) = P(B\mid A) = P(B)$$

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

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

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

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

C. Discrete Probability Distributions

C.0a $\mu = E(Y) = \sum yp(y),\qquad (p89)$

C.0b $E\left(Y^2\right) = \sum y^2\, p(y),\ \ \sigma^2 = E(Y-\mu)^2 = E\left(Y^2\right)-\mu^2,\ (p96)$

	Probability Density function $p(x)$	*Mean* (μ) *and Variance* (σ^2)
C.1	*The Binomial Distribution:* $B(n,p)$ *(p119)* $f(y) = \binom{n}{y}\, p^y(1-p)^{n-y},\quad y = 0,\ 1,\ \ldots,\ n\,,$ *where* $\binom{n}{y} = \dfrac{n(n-1)\ldots(n-y+1)}{y!}.$	$\mu = E(Y) = np$ $\sigma^2 = V(Y) = np(1-p).$
C.2	*The Hypergeometric Distribution (p128)* $f(y) = \dfrac{\binom{K}{y}\binom{N-K}{n-y}}{\binom{N}{n}}\,,\ \ y = \ 0,\ 1,\ \ldots,$	$\mu = E(Y) = n\ \left(K/N\right)$
C.3	*The Poisson Distribution (p136)* $f(y) = \dfrac{e^{-\lambda t}\left(\lambda t\right)^y}{y!},\ y = \ 0,\ 1,\ \ldots$	$\mu = E(Y) = \lambda t$ $\sigma^2 = V(Y) = \lambda t$

D. Continuous Probability Distributions

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

$D.0 \quad P(a < Y < b) = \int_a^b f(y)dy, \;\; P(Y \le u) = \int_{-\infty}^{u} f(y)dy$

$D.0a \quad \mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy, \;\; (p89)$

$D.0b \quad E\left(Y^2\right) = \int_{-\infty}^{\infty} y^2 f(y)dx, \;\; \sigma^2 = E(Y-\mu)^2 = E\left(Y^2\right) - \mu^2, \;\; (p96)$

	Probability Density function	*Mean and Variance*
D.1	*The Normal Distribution:* $N(\mu, \sigma^2)$	*Mean* = $E(Y) = \mu$ *Variance* = $V(Y) = \sigma^2$
D.2	*The Exponential Distribution* $f(y) = \frac{1}{\beta} e^{-y/\beta}, \;\; 0 < y$	*Mean* = $E(Y) = \beta$ *Variance* = $V(Y) = \beta^2$
D.3	*Waiting Time Distribution* $f(t) = \lambda e^{-\lambda t}, \;\; 0 < t$	*Mean* = $E(T) = 1/\lambda$ *Variance* = $V(T) = 1/\lambda^2$
D.4	*The Gamma Distribution* $f(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, \;\; 0 < y, \; 0 < \alpha, \; 0 < \beta$	*Mean* = $E(Y) = \alpha\beta$ *Variance* = $V(Y) = \alpha\beta^2$

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* $n \ge 30$, *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 that* σ^2 *is estimated by* s^2 *(SLLN).*

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)*

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

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):** $\bar{y} \mp z_{\alpha/2}\left(\sigma/\sqrt{n}\right)$ for (235)*

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

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

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

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

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

***F.2.0 CI for $\mu_1 - \mu_2$, (σ_i^2 known, normal):** $(\bar{y}_1 - \bar{y}_2) \mp z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ (cf. p247)*

F.2.1 ACI for $\mu_1 - \mu_2$, (σ_i^2 known, large n_i, nonnormal):

$(\bar{y}_1 - \bar{y}_2) \mp z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$ *(cf. p247)*

***F.2.2 CI for $\mu_1 - \mu_2$, (σ_i^2 unknown, large n_i, nonnormal):** $(\bar{y}_1 - \bar{y}_2) \mp z_{\alpha/2}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ (p247)*

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

$$(\bar{y}_1 - \bar{y}_2) \mp t_{\alpha/2}\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}, \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}, \quad \nu = (n_1 - 1) + (n_2 - 1), \text{ (p249)}$$

F.2.4 CI for $\mu_1 - \mu_2$, (small $n_1 = n_2$, $\sigma_1^2 \ne \sigma_2^2$, normal):

$$(\bar{y}_1 - \bar{y}_2) \mp t_{\alpha/2}\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}, \quad \nu = (n - 1) + (n - 1) \text{ (p249)}$$

F.2.5 ACI for $\mu_1 - \mu_2$, (small $n_1 \neq n_2$, $\sigma_1^2 \neq \sigma_2^2$, normal):

$$(\bar{y}_1 - \bar{y}_2) \mp t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}, \qquad \text{(p251)}$$

F.2.6 CI for $\mu_1 - \mu_2$ using matched pairs): $\bar{d} \mp t_{\alpha/2}\sqrt{s_d^2/n}$, ($df = n-1$), (p254)

F.3 Confidence Interval for Proportion p

F.3.1 CI for π 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.3.2 CI for $\pi_1 - \pi_2$ with large sample sizes :

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

F.4 Confidence Interval for Variance σ^2: $[(n-1)s^2/l, \quad (n-1)s^2/u]$
where $l = \chi^2_{\alpha/2}$ and $u = \chi^2_{1-\alpha/2}$

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

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

G.1 Testing of a Mean μ

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

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

G.1.2 σ unknown, large sample: $z \approx \dfrac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$ *(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 : $t = \dfrac{\bar{y} - \mu_0}{\sqrt{s^2 / n}}$, $(\nu = n - 1 \ge 1)$ *(p304)*

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

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

G.2.0 known σ_i , any n_i , normal: $z = \dfrac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)}}$

G.2.1 known σ_i , large n_i , nonnormal: $z \approx \dfrac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)}}$

G.2.2 unknown σ_i , large n_i , nonnormal: $z \approx \dfrac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{\left(s^2 / n_1\right) + \left(s^2 / n_2\right)}}$,

$s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, $\nu = (n_1 - 1) + (n_2 - 1)$

Table for the above three tests are given below:

H_0	H_a	RR (for H_0)	p-value
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 < \delta$	$z < -z_\alpha$	$P(Z < z)$
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 > \delta$	$z > z_\alpha$	$P(Z > z)$
$\mu_1 - \mu_2 = \delta$	$\mu_1 - \mu_2 \ne \delta$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$2\ P(Z > \lvert z \rvert)$

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

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{\left(s^2 / n_1\right) + \left(s^2 / n_2\right)}} \ , \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \ , \quad \nu = (n_1 - 1) + (n_2 - 1) \ , \ \textit{(p308)}$$

G.2.4 Small $n_1 = n_2$, $\sigma_1^2 \ne \sigma_2^2$, normal)

$$t = \frac{\bar{y}_1 - \bar{y}_2 - \delta}{\sqrt{\left(s_1^2 / n\right) + \left(s_2^2 / n\right)}} \ , \quad \nu = (n - 1) + (n - 1)$$

G.2.5 Small $n_1 \neq n_2$, $\sigma_1^2 \neq \sigma_2^2$, normal)

$$t \approx \frac{\overline{y}_1 - \overline{y}_2 - \delta}{\sqrt{\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)}}\ ,\quad \nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}\text{, (p309).}$$

Tables for the above three tests are provided below:

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 $t > t_{\alpha/2}$	$2\ P(T > \vert t \vert)$

G.2.5 Testing $\mu_1 - \mu_2$ using matched pairs): $t = \overline{d}/\sqrt{s_d^2/n}$, $(df = n-1)$, *(p310)*

G.3.1 Testing of a proportion

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

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 $z > z_{\alpha/2}$	$2\ P(Z > \vert z \vert)$

G.3.2 Testing the difference of two proportions

Large n_i : $z \approx \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})/n_1 + \hat{p}(1-\hat{p})/n_2}}$ *where* $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \dfrac{y_1 + y_2}{n_1 + n_2}$

H_0	H_a	RR (for H_0)	p-value
$p_1 = p_2$	$p_1 < p_2$	$z < -z_\alpha$	$P(Z < z)$
$\pi_1 = \pi_2$	$p_1 > p_2$	$z > z_\alpha$	$P(Z > z)$
$\pi_1 = \pi_2$	$p_1 \neq p_2$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$2\ P(Z > \vert z \vert)$

G.4.1 Testing a variance $\chi^2 = (n-1)s^2/\sigma_0^2$

To test $H_0 : \sigma^2 = \sigma_0^2,\ H_0 : \sigma^2 > \sigma_0^2$ **use** $\chi^2 > \chi_\alpha^2$
To test $H_0 : \sigma^2 = \sigma_0^2,\ H_0 : \sigma^2 < \sigma_0^2$ **use** $\chi^2 < \chi_\alpha^2$
To test $H_0 : \sigma^2 = \sigma_0^2,\ H_0 : \sigma^2 \neq \sigma_0^2$ **use** $\chi^2 < \chi_{1-\alpha/2}^2$ or $\chi^2 < \chi_{\alpha/2}^2$

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

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 $\hat{\beta}_1 = \dfrac{s_{xy}}{s_{xx}}$, $s_{xy} = \sum xy - \left(\sum x\right)\left(\sum y\right)/n$, $s_{xx} = \sum x^2 - \left(\sum x\right)^2/n$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. *(p536)*

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

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

H.1.4 *TSS =* $\sum (y-\bar{y})^2 = \sum y^2 - \dfrac{\left(\sum y\right)^2}{n}$, *SSR =* $bs_{xy} = b^2\, s_{xx}$, *SSE =* $TSS - SSR$

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

H.2 Inference Regarding the Regression Coefficient

H.2.1 100(1-α)% confidence interval for β_0 ***:*** $\hat{\beta}_0 \mp \sqrt{\left\{\dfrac{1}{n} + \dfrac{\bar{x}^2}{s_{xx}}\right\} MSE}$.

H.2.2 100(1-α)% confidence interval for β_1 ***:*** $\hat{\beta}_1 \mp t_{\alpha/2}\sqrt{MSE/s_{xx}}$

H.2.3 Testing the hypothesis $H_0 : \beta_1 = c$ ***:*** $t = \dfrac{\hat{\beta}_1 - c}{\sqrt{MSE/s_{xx}}}$ *(p364)*

H.3 Confidence Interval

H.3.1 Confidence Interval for Mean for a given x : $\hat{y} \pm t_{\alpha/2} \sqrt{\left\{\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{s_{xx}}\right\} MSE}$,

H.3.2 Prediction Interval for an Individual Y ***Given*** x***:***

$\hat{y} \pm t_{\alpha/2} \sqrt{\left\{1 + \dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{s_{xx}}\right\} MSE}$ *(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_i		Y	CI	z/t	TEST	page
KNOWN		1	Normal	F.1.0	z	G.1.0	235
KNOWN	Large	1	nonnormal	F.1.1	$\approx z$	G1.1	235
KNOWN		2	Normal	F.2.0	z	G.2.0	247
	Large	2	nonnormal	F.2.1	$\approx z$	G.2.1	247
UNKNOWN		1	Normal	F.1.3	t	G.1.3	239
UNKNOWN	Large	1	Nonnormal	F.1.2	$\approx z$	G.1.2	239
UNKNOWN		2	Normal $(\sigma_1^2 = \sigma_2^2)$	F.2.3	t	G.2.3	249
UNKNOWN	Small $(n_1 = n_2)$	2	Normal	F.2.4	$\approx t$	G.2.4	251
UNKNOWN		2	Normal	F.2.5	$\approx t$	G.2.5	251
UNKNOWN	Large	2	Nonnormal	F.2.2	$\approx z$	G.2.2	251

Distribution Guide

n, p → **Binomial**

$(p = N_0 / N)$ → *SWR: Binomial*

N

(N_0, n) → *SWOR: Hypergeometric*

- ***First success in a trial: Geometric***
- ***Second/ third/.. success in a trail: Negative Binomial***
- ***Mean & discrete variable: Poisson***

(# accidents occurs in a given time interval)

- ***Time between accidents: Exponential***

Lifetime : Exponential

$$mean = \beta,\ \ parameter = \lambda = 1/\beta$$