Formulae for
Statistics for Science, Engineering and Technology

A. Descriptive Statistics (for Samples)

A.1 Percentiles: The position of the 100α-th percentile is $R_\alpha = \alpha\ (1+n) = i + d,$ so that the 100α-th percentile is $(1-d)y_i + dy_{i+1}$, where $\alpha = 0.01,\ 0.02,\ \cdots, 0.99,$ i is the largest integer not exceeding R_α and y_i is i-th value of the sorted sample.

A.2 Average and variance are

$$\overline{x} \equiv \frac{1}{n}\sum x; \quad CSS \equiv \sum (x-\overline{x})^2 = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2;\ s^2 \equiv \frac{CSS}{n-1}.$$

A.3 Mean and the variance for grouped data:

$$\overline{x} \equiv \frac{1}{n}\sum xf;\ CSS \equiv \sum x^2 f\ - \frac{1}{n}\left(\sum xf\right)^2;\ s^2 \equiv \frac{CSS}{n-1}.$$

where $x's$ are the mid values of each class and the sum is over the number of classes.

A.4 Standardized score of an observation x is $z(x) \equiv (x-\overline{x})/s.$

A.5 Coefficient of Variation : $CV \equiv s/\overline{x}.$

A.6 Coefficient of Skewness : $CS \equiv \frac{3(\overline{x}-\tilde{x})}{s}$, where $\tilde{x}$ is the sample median.

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

B.1 A but not B (=Only A): $P(AB') = P(A) - P(AB).$

B.2 $P(A'B') + P(AB') + P(A'B) + P(AB) = 1$

B.3 Law of Addition

B.3a $P(A \cup B) = P(AB') + P(A'B) + P(AB).$

B.3b $P(A \cup B) = P(A) + P(B) - P(AB).$

B.3c $P(A \cup B) = 1 - P(A \cup B)' = \ 1 - P(A'B').$ De Morgan's Law

B.4 $P(AB)' = P(A' \cup B').$ De Morgan's Law

B.5 Law of Multiplication

$$P(A \mid B) = \frac{P(AB)}{P(B)}, \quad P(B) \neq 0; \quad P(AB) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

B.6 Independence: $P(A \mid B') = P(A) = P(A \mid B),\ \ P(AB) = P(A)P(B).$.

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

B.7 None: $P(A'B'C') = P(A')P(B')P(C')$ iff A, B and C are independent.

B.8 $P(A \cup B \cup C) = P(A) + P(B) + P(C) - [P(AB) + P(AC) + P(BC)] + P(ABC).$

C. Discrete Probability Distributions

C.1 $P(a \le X \le b) = \sum_{x} f(x);\ P(X \le b) = \sum_{x \le b} f(x),$

C.2 $\mu \equiv E(X) = \sum_{x} x\, f(x).$

C.3 $E(X^2) = \sum x^2\, f(x),\ \ \sigma^2 \equiv E(X - \mu)^2 = E(X^2) - \mu^2.$

C.4 The Binomial Distribution $B(n, p)$: $f(x) = \binom{n}{x} p^x q^{n-x};\ \ x = 0,1,...,n;\ \ 0 < p < 1;$

$q = 1 - p;\ \mu = np,\ \sigma^2 = npq.$

C.5 The Geometric Distribution: $f(x) = q^x p,\ \ x = 0,1,2,\cdots;\ q = 1 - p;\ \ \mu = 1/p,\ \sigma^2 = q/p^2.$

C.7 The Hypergeometric Distribution

$$f(x) = \binom{K}{x}\binom{N-K}{n-x} \div \binom{N}{n}, \quad \max\{0, n-(N-K)\} \le y \le \min\{n, K\};\ \mu = np,$$

$\sigma^2 = (1-c)\ npq,\ \ (N-1)\ c = n-1,\ p = (K/N),\ q = 1-p.$

The mass function of hypergeometric distribution can also be written as

$$P(X = x) = \binom{n}{x} P(D_1 D_2 \cdots D_x D'_{x+1} \cdots D'_n) = \binom{n}{x}\left(P_x^K P_{n-x}^{N-K} \div P_n^N\right).$$

C.8 The Poisson Distribution $f(x) = \dfrac{(\lambda t)^x}{x!} e^{-\lambda t};\ x = 0,1,...;\ \mu = \lambda t,\ \sigma^2 = \lambda t.$

D. Continuous Probability Distributions

For a continuous random variable X with pdf $f(x)$,

D.1 $P(a < X < b) = \int_a^b f(x)dx; \quad F(k) = \int_{-\infty}^{k} f(x)dx$ where k is a particular value of x.

D.2 $\mu \equiv E(X) = \int_{-\infty}^{\infty} x\, f(x)dx.$

D.3 $E(X^2) = \int_{-\infty}^{\infty} x^2\, f(x)d\,x\; \sigma^2 \equiv V(X) = E(X^2) - \mu^2.$

D.4 The Normal Distribution $N(\mu, \sigma^2)$

D.5 The Exponential Distribution: $f(x) = \frac{1}{\beta}\, e^{-x/\beta}, \quad 0 \le x;\ \mu = \beta,\ \sigma^2 = \beta^2.$

D.6 Waiting Time Distribution: $f(t) = \lambda e^{-\lambda t}, \quad 0 \le t;\ \mu = 1/\lambda,\ \sigma^2 = 1/\lambda^2,$

D.7 The Gamma Distribution: $f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\, x^{\alpha-1}\, e^{-x/\beta}, \quad 0 < x;\ 0 < \alpha,\ 0 < \beta,\ \mu = \alpha\beta,$

$\sigma^2 = \alpha\beta^2.$

E. Sampling Distributions

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

E.2 Suppose that Y has a distribution with mean μ and variance σ^2. However if the distribution is not normal but $30 \le n$, then

$\frac{\sum X - n\mu}{\sqrt{n\sigma^2}} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx Z$ (weak) .

This is known as Central Limit Theorem (CLT). In case σ^2 is unknown,

$\frac{\sum X - n\mu}{\sqrt{nS^2}} = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx\approx Z$ (weaker)

by Slutsky's Theorem (CT). Note S^2 converges stochastically to σ^2 .

E.3 The Student T- statistic is defined by $T = \dfrac{\overline{X} - \mu}{\sqrt{S^2 / n}}$, with $\nu = n - 1$

E.4 The Sampling Distribution of the Proportion

$$\frac{X - np}{\sqrt{npq}} = \frac{(X/n) - p}{\sqrt{pq/n}} \approx Z.$$

E.5 The binomial event $X = x$ can be inflated to $x - 0.50 \le X \le x + 0.50$ if $4 \le np$ and $4 \le q$.

The $100(1-\alpha)\%$ percentiles of a random variable Z is denoted by ${}_{1-\alpha}z_{\alpha}$ or often by z_{α}. For example 95th percentile of the standard normal distribution is given by ${}_{0.95}z_{0.05} \approx 1.645$.

F. Statistical Estimation (with Random Sample / Samples)

F.1 Confidence Interval Estimates of the Mean μ

F.1.1 CI for μ, (σ known, any n, normal): $\overline{x} \mp z_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$.

F.1.2 ACI for μ, (σ known, large n, nonnorma): $\overline{x} \mp z_{\alpha/2}\left(\dfrac{\sigma}{\sqrt{n}}\right)$.

F.1.2a sample size for estimating μ : $n = \dfrac{z_{\alpha/2}^2}{e^2}\sigma^2$, where $P(|\overline{X} - \mu| \le e) = 1 - \alpha$.

F.1.3 ACI for μ, (σ unknown, large n, nonnormal): $\overline{x} \mp z_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right)$.

F.1.3a sample size for estimating μ : $n = \dfrac{z_{\alpha/2}^2}{e^2}s^2$, where $P(|\overline{X} - \mu| \le e) = 1 - \alpha$.

F.1.4 CI for μ, (σ unknown, $n \ge 2$, normal): $\overline{x} \mp t_{\alpha/2}\left(\dfrac{s}{\sqrt{n}}\right)$ for any $n \ge 2$.

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

F.2.1 CI for $\mu_1 - \mu_2$, (σ_i^2 known, any n_i, normal):

$$(\overline{x}_1 - \overline{x}_2) \mp z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

F.2.2 ACI for $\mu_1 - \mu_2$, (σ_i^2 known, large n_i, nonnormal):

$$(\bar{x}_1 - \bar{x}_2) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

F.2.3 CI for $\mu_1 - \mu_2$, (σ_i^2 unknown, large n_i, nonnormal):

$$(\bar{x}_1 - \bar{x}_2) \mp z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

F.2.4 CI for $\mu_1 - \mu_2$, (small n_i, unknown $\sigma_1^2 = \sigma_2^2$ but equal, normal):

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

F.3 Confidence Interval for Proportion p

F.3.1 CI for p when n large: $\hat{p} \mp z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}.$

F.3.1 a Large sample size for estimating p: $n = \frac{z_{\alpha/2}^2}{e^2} \hat{p}\hat{q}$ where e is the error in estimation.

F.3.2 CI for $p_1 - p_2$ with large sample sizes :

$$\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}.$$

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

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

G.1 Testing of a Mean μ

G.1.1 σ known, normal: $z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}}.$

G.1.2 σ known, large n, nonnormal: $z \approx \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2/n}}.$

G.1.3 σ unknown, large sample: $z \approx \frac{\bar{y} - \mu_0}{\sqrt{s^2/n}}$

G.1.4 σ unknown, normal population : $t = \frac{\bar{x} - \mu_0}{\sqrt{s^2/n}}$, $(\nu = n - 1 \ge 1).$

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

G.2.1 known σ_i, large n_i, normal: $z = \dfrac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)}}.$

G.2.2 known σ_i, large n_i, nonnormal: $z \approx \dfrac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)}}.$

G.2.3 unknown σ_i, large n_i, nonnormal: $z \approx \dfrac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\left(s_1^2 / n_1\right) + \left(s_2^2 / n_2\right)}}.$

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

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \delta_0}{\sqrt{\left(s_w^2 / n_1\right) + \left(s_w^2 / n_2\right)}} \quad , \quad s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \quad , \quad \nu = (n_1 - 1) + (n_2 - 1),$$

where s_w^2 is the weighted or pooled combined variance $\min(s_1^2, s_2^2) \le s_w^2 \le \max(s_1^2, s_2^2).$

G.3.1 Testing of a proportion

Large sample: $z \approx \dfrac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$, where $q_0 = 1 - p_0.$

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

H.1 Line of Best Fit

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

Estimated Model: $\hat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ for a given x. $\hat{\mu}(x)$ is often denoted by $\hat{y}$. Error $\varepsilon(x) = y - \mu(x)$ is estimated by $e = y - \hat{\mu}(x).$ Also $\Delta\mu(x) = \mu(x + h) - \mu(x) = \beta_1 h.$

H.1.1 $\hat{\beta}_1 = \dfrac{s_{xy}}{s_{xx}}$, $s_{xy} = \sum xy - \dfrac{1}{n}\left(\sum x\right)\left(\sum y\right),$ $s_{xx} = \sum x^2 - \dfrac{1}{n}\left(\sum x\right)^2$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

H.1.2 Coefficient of correlation: $r = \dfrac{s_{xy}}{\sqrt{s_{xx}\ s_{yy}}}.$ $\left(r\sqrt{s_{yy}} = \hat{\beta}_1\sqrt{s_{xx}}\right)$

H.1.3 $CSS = \sum y^2 - \dfrac{1}{n}\left(\sum y\right)^2$, $SSR = \hat{\beta}_1 s_{xy}$, $SSE = CSS - SSR.$

H.1.4 Estimate of σ^2: $s_e^2 = SSE / (n - 2)$ often denoted by $MSE.$

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

H.2 Inference Regarding the Regression Coefficients

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

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

H.2.3 $100(1-\alpha)\%$ confidence interval for β_1: $\hat{\beta}_1 \mp t_{\alpha/2}\sqrt{\dfrac{MSE}{s_{xx}}}$.

H.2.4 Testing the hypothesis $H_0:\beta_1=c$: $t=\dfrac{\hat{\beta}_1-c}{\sqrt{MSE/s_{xx}}}$.

H2.5 Testing the hypothesis $H_0:\rho=0$: $t=\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. $\quad(\rho\sigma_y=\beta_1\sigma_x)$

Inference Regarding the Response Variable

H.3.1 $100(1-\alpha)\%$ Confidence Interval of $\mu(x)$: $\hat{\mu}(x) \pm t_{\alpha/2}\sqrt{\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}\right)MSE}$.

H.3.2 $100(1-\alpha)\%$ Prediction Interval for an Individual Y for a given x:

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

GOLDEN TABLE (Sigma-Distribution)

	$X \sim N(\mu,\sigma^2)$	$X \not\sim N(\mu,\sigma^2)$
σ known (Moya moujud)	$\dfrac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}=Z$	Small n: Assume $X \sim N(\mu,\sigma^2)$; Exact $\dfrac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}=Z$. Large n: $\dfrac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \approx Z$ (CLT)
s known (Ma fiyy moya)	$\dfrac{\bar{X}-\mu}{\sqrt{S^2/n}}=T$	Large n: $\dfrac{\bar{X}-\mu}{\sqrt{S^2/n}} \approx\approx Z$ (ST)