Descriptive Statistics

Uni-modal: one peak; Bi-modal: two peaks; Multi-modal: more than 2 peaks

⋆ Left-skewed: A long *left* tail; Symmetric

Range = max - min

Sample variance: $s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

Standard deviation: $s = \sqrt{s^2}$

Coefficient of variation: $CV = 100 \times \frac{s}{\overline{x}}$

Only have meaning when all the numbers are positive.

Mean absolute deviation: $MAD = \frac{\sum_{i=1}^{n}|x_i-\overline{x}|}{n}$

Standardized variable: $z_i = \frac{x_i-\overline{x}}{s}$

A positive z means the observation is to the right of the mean, at the position of the z^{th} standard deviation.

In the boxplot, $Q3-Q1 = IQR$, if $N-Q_3 > 1.5\times IQR$ or $Q_1-N > 1.5\times IQR$ then N is an outlier.

Correlation coefficient: $r = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$

Covariance: $s_{XY} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{n-1}$

Skewness = $\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(\frac{x_i-\overline{x}}{s})^3$

Skewness < 0 : Skewed Left; Skewness > 0 : Skewed Right

Probability

denotation:$P(AB)$ means $P(A\cap B)$

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

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

Mutually Exclusive Events:$A\cup B=\emptyset, P(AB)=0$

Collectively exhaustive Events: $A\cup B = U$

Conditional Probability: $P(A|B) = \frac{P(AB)}{P(B)}$

Law of Total Probability:$P(B) = \sum_{i=1}^{n}P(A_i)P(B|A_i)$

Specially, $P(B) = P(B|A)P(A)+P(B|\overline{A})P(\overline{A})$

Odds in favor of A: $\frac{P(A)}{P(\overline{A})}$, Odds against A: $\frac{P(\overline{A})}{P(A)}$

Bayes Theorem:

$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\overline{B})p(\overline{B})}$

The second expression is generally more useful

General Form of Bayes' Theorem:

$P(B_i|A = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+\cdots+P(A|B_n)P(B_n)})$

Discrete probability distribution

The following $p(x)$ functions are the pmf

CDF: $F(x) = P(X\le x) = \sum_{y:y\le x}p(y)$

Expected value(mean value): $E(X) = \sum_{x\in D}x\cdot p(x)$

$E(X)$ can also be denoted as μ_X

Expected value of function: $E[h(X)] = \sum_D h(x)\cdot p(x)$

Variance: $V(X) = \sigma_X^2 = \sum_D (x-\mu)^2\cdot p(x) = E(X-\mu)^2$

μ is the Expected value of X

SD(Standard deviation): $\sigma_X = \sqrt{\sigma_X^2}$

Skewness: $\frac{E[(X-\mu)^3]}{\sigma^3} = E[\frac{X-\mu}{\sigma}^3]$

mgf(Moment Generating Function): $M_X(t) = E(e^{tX}) = \sum_{x\in D}e^{tx}p(x)$

Continuous probability distribution

The following f(x) are the pdf.

$P(a\le X\le b) = \int_a^b f(x)dx$

CDF: $F(x) = P(X\le x) = \int_{-\infty}^{x}f(y)dy$

Expected value: $\mu_X = E(X) = \int_{-\infty}^{\infty}x\cdot f(x)dx$

$E(X^2) = \int_{-\infty}^{\infty}x^2\cdot f(x)dx$

$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty}h(x)\cdot f(x)dx$

Variance: $V(X) = \sigma_X^2 = \int_{-\infty}^{\infty}(x-\mu)^2\cdot f(x)dx = E[(x-\mu)^2]$

SD(Standard deviation): $\sigma_X = \sqrt{V(x)}$

mgf(Moment Generating Function):

$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty}e^{tx}f(x)dx$

Percentile: $p = F[\eta(p)] = \int_{-\infty}^{\eta(p)}f(y)dy$

$p\in[0,1]$, *$\eta(p)$ means the p^{th} percentile, get the number by solving the equation above.*

$\tilde{\mu}$ denotes the median, where $F(\tilde{\mu}) = 0.5$

For n sample observations, the i^{th} smallest observation is the $[100(i-0.5)/n]^{th}$ sample percentile.

Useful mutual shortcut formula:

$E(aX+b) = a\cdot E(X)+b$

$V(X) = \sigma_X^2 = E(X^2)-[E(X)]^2$

$V(aX+b) = \sigma_{aX+b}^2 = a^2\cdot\sigma_X^2 = a^2\cdot V(X)$

$\sigma_{aX+b} = |a|\cdot\sigma_X$

$E(X^r) = M_X^{(r)}(0)$, the (r) means r-order derivation.

Let X have mgf $M_X(t)$, let $Y = aX+b$, then $M_Y(t) = e^{bt}M_X(at)$.

Joint probability distribution

Two discrete variables

pmf: $p(x,y) = P(X = x\ and\ Y = y)$

$P[(X,Y)\in A] = \sum\sum_{(x,y)\in A}p(x,y)$

marginal pmf: $p_X(x) = \sum_y p(x,y)$; $p_Y(y) = \sum_x p(x,y)$

independence: $p(x,y) = p_X(x)\cdot p_Y(y)$

$E[h(X,Y)] = \sum_x\sum_y h(x,y)\cdot p(x,y)$

Covariance: $Cov(X,Y) = \sum_x\sum_y(x-\mu_X)(y-\mu_Y)p(x,y)$

Correlation coefficient:

$Corr(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}$

Two continuous variables

The following f(x,y) are pdf.

cdf: $P[(X,Y)\in A] = \iint_A f(x,y)dxdy$

particularly, if A is a rectangle:

$P[(X,Y)\in A] = P(a\le X\le b, c\le Y\le d) = \int_a^b\int_c^d f(x,y)dxdy$

marginal pdf:

$f_X(x) = \int_{-\infty}^{\infty}f(x,y)dy$, for $-\infty < x < \infty$

$f_Y(y) = \int_{-\infty}^{\infty}f(x,y)dy$, for $-\infty < y < \infty$

independence: $f(x,y) = f_X(x)\cdot f_Y(y)$

$E[h(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(x,y)\cdot p(x,y)dxdy$

Covariance:

$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy$

Correlation coefficient:

$Corr(X,Y) = \rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X\cdot\sigma_Y}$

Mutual properties

$E[h_1(X_1)\cdot h_2(X_2)\cdots h_n(X_n)]$

$= E[h_1(X_1)]\cdot E[h_2(X_2)]\cdots E[h_n(X_n)]$

$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(X\cdot Y)-\mu_X\cdot\mu_Y$

$Cov(aX+bY,Z) = a\cdot Cov(X,Z)+b\cdot Cov(Y,Z)$

To compute σ_X, integrate continuous y from $-\infty$ to ∞ or sum all the discrete y up, so that x is the only variable.

*If X and Y are **independent**, $Cov = \rho = 0$, but $\rho = 0$ does not imply independence. $\rho = \pm 1$ means strictly $Y = aX+b$.*

$Corr(aX+b, cY+d) = Corr(X,Y)$

For any X,Y: $-1\le Corr(X,Y)\le 1$

Linear combination of random variables

Let $X_1,\cdots,X_n$ and $Y_1,\cdots,Y_m$ be random variables with $E(X_i) = \mu_i$ and $E(Y_j) = \xi_j$. Define $U_1 = \sum_{i=1}^{n}a_iX_i$ and $U_2 = \sum_{j=1}^{n}b_jY_j$ for constants a_i and b_i:

$E(U_1) = \sum_{i=1}^{n}a_i\mu_i$;

If all X_i are independent: $V(U_1) = a_1^2V(X_1)+\cdots+a_n^2V(X_n)$

For any X_i: $V(U_1) = \sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jcov(X_i,X_j)$

$Cov(U_1,U_2) = \sum_{i=1}^{n}\sum_{j=1}^{m}a_ib_jCov(X_i,Y_j)$;

Conditional distribution

The formula for two discrete and two continuous variables are given next to each other in the following text.

Conditional pmf/pdf of Y given X = x:

$p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}$; $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$

Conditional mean of Y given that X = x:

$\mu_{Y|X=x} = E(Y|X=x) = \sum_{y \in D_Y} y p_{Y|X}(y|x)$

$\mu_{Y|X=x} = E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$

Conditional mean of any function $g(Y)$:

$E(g(Y)|X=x) = \sum_{y \in D_Y} g(y) p_{Y|X}(y|x)$

$E(g(Y)|X=x) = \int_{-\infty}^{\infty} g(y) f_{Y|X}(y|x) dy$

Conditional variance of Y given X = x:

$\sigma^2_{Y|X=x} = V(Y|X=x) = E\left\{[Y - E(Y|X=x)]^2 | X=x\right\}$

$= E(Y^2|X=x) - \mu^2_{Y|X=x}$

⋆ Key theorems when solving problems:

$E(Y) = E[E(Y|X)]$

$V(Y) = V[E(Y|X)] + E[V(Y|X)]$

Transformations of a Random Variable

Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g has an inverse function $X = h(Y)$. Then $f_Y(y) = f_X(h(y))\,|h'(y)|$. $h'(y)$ is the derived function.

Transformations of Random Variables

Given two random variables X_1 and X_2, consider forming two new random variables $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$. Let $f(x_1, x_2)$ and $g(y_1, y_2)$ be their joint pdf.

$g(y_1, y_2) = f(x_1, x_2) \cdot \left| \frac{\partial(x_1, x_2)}{\partial(y_1, y_2)} \right|$

This can be extended to n variables.

Sampling distributions

$\overline{X}$: the sample mean regarded as a statistic, $\overline{x}$: the calculated value, S: the sample standard deviation regarded as a statistic, s: the computed value

Let $X_1, \cdots X_n$ be a random sample from a distribution with mean μ and standard deviation σ.

This will be called "in general cases" in the following texts.

$E(\overline{X}) = \mu_{\overline{X}} = \mu; V(\overline{X}) = \sigma^2_{\overline{X}} = \sigma^2/n; \sigma_{\overline{X}} = \sigma/\sqrt{n}$

$T_0 = X_1 + \cdots + X_n, E(T_0) = n\mu; V(T_0) = n\sigma^2, \sigma_{T_0} = \sqrt{n}\sigma$

The case of a normal distribution

If $X_1, \cdots X_n$ is a random sample from a normal distribution, then $\overline{X}$ and T_0 are normally distributed for any n. Their mean and variance can be seen above.

The $\overline{X}$ and S^2 are independent.

The central limit theorem (CLT)

In general cases, as $n \to \infty$, the standardized versions of $\overline{X}$ and T_0 have the standard normal distribution.

$\lim_{n\to\infty} P(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le z) = P(Z \le z) = \Phi(z)$

$\lim_{n\to\infty} P(\frac{T_0 - n\mu}{\sqrt{n}\sigma} \le z) = P(Z \le z) = \Phi(z)$

This can be used only when $n > 30$.

Law of large numbers

In general cases: $E[(\overline{X} - \mu)]^2 \to 0$ as $n \to \infty$

$P(|\overline{X} - \mu| \ge \epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$

Binomial Distribution

n: number of trials; π: probability of success of each trial. X: total number of success

pmf: $P(X=x) = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}$

Mean: $n\pi$; Standard deviation: $\sqrt{n\pi(1-\pi)}$

Skewed right: $\pi < 0.5$

Geometric Distribution

π: probability of success of each trial. X: this is the first successful trial

pmf: $P(X=x) = \pi(1-\pi)^{x-1}$; cdf: $P(X \le x) = 1 - (1-\pi)^x$

Mean: $\frac{1}{\pi}$; Standard deviation: $\sqrt{\frac{1-\pi}{\pi^2}}$

Poisson Distribution

λ: mean arrivals per unit of time. X: the arrivals per unit of time

pmf:$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$; Mean: λ; Standard deviation: $\sqrt{\lambda}$

Always skewed right.

mgf: $M_X(t) = e^{\lambda(e^t - 1)}$

Hypergeometric distribution

N: population, M: number of success in the population; n: the number in the sample. X: the number of success in the sample

pmf: $P(X=x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

Mean: $E(X) = n \cdot \frac{M}{N}$

Variance: $V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N}\left(1 - \frac{M}{N}\right)$

Negative binomial distribution

p: probability of success of each trial; r: number of successes observed. X: the number of failures before the $r^t h$ success

pmf: $nb(x; r, p) = \binom{x+r-1}{r-1} p^r (1-p)^x$

mgf: $M_X(t) = \frac{p^r}{[1 - e^t(1-p)]^r}$

Mean: $E(X) = \frac{r(1-p)}{p}$, Variance: $\frac{r(1-p)}{p^2}$

Uniform Distribution

a: lower limit b: upper limit

pdf: $\frac{1}{b-a}$; cdf: $P(X \le x) = \frac{x-a}{b-a}$; Mean: $\frac{a+b}{2}$; Sd: $\sqrt{\frac{(b-a)^2}{12}}$

Normal Distribution

μ: population mean, σ: population standard deviation

pdf: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$; Mean: μ; Standard deviation: σ

Standard Normal Distribution

Normalizing process: $Z = \frac{X-\mu}{\sigma}$

pdf: $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$; Mean: 0; Standard deviation: 1

Using $\Phi(z)$ to represent $cdf(Z \le z)$

Normal Approximations to the Binominal

X: a binomial distribution, n: number of trials, p: probability of success

$\mu = np, \sigma = \sqrt{np(1-p)}$

$P(X \le x) = B(x; n, p) \approx \Phi(\frac{x + 0.5 - np}{\sqrt{np(1-p)}})$

The **"+0.5"** is called Continuity Correction which **must** be applied. The approximation can be applied when $np \ge 10$ and $n(1-p) \ge 10$.

Normal Approximations to the Poisson

$\mu = \lambda, \sigma = \sqrt{\lambda}; P(X \le x) \approx \Phi(\frac{x + 0.5 - \lambda}{\sqrt{\lambda}})$

⋆ The pdf of the following distribution are all in such form:

$$f(x; *args) = \begin{cases} f_0(x; *args) & , x \in Domain \\ 0 & , otherwise \end{cases}$$

Only $f_0(x; *args)$ will be listed.

Gamma Distribution

gamma function: $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$

$\Gamma(x+1) = x\Gamma(x)$; $\Gamma(n) = (n-1)!$; $\Gamma(\frac{1}{2}) = \sqrt{\pi}$;

pdf: $f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0$

Mean: $E(X) = \mu = \alpha\beta$; Variance: $V(X) = \sigma^2 = \alpha\beta^2$

Standard gamma distribution: let $\beta = 1$ in the above pdf.

cdf - standard gamma distribution: $F(x; \alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)} dy$

cdf - any gamma distribution with parameter α and β:

$P(X \le x) = F(x; \alpha, \beta) = F(\frac{x}{\beta}; \alpha)$

Exponential Distribution

pdf: $f(x; \lambda) = \lambda e^{-\lambda x}, x \ge 0$

It is a specific gamma distribution where $\alpha = 1, \beta = 1/\lambda$.

cdf: $P(X \le x) = 1 - e^{-\lambda x}$;

Mean: $1/\lambda$; Standard deviation: $1/\lambda$

Exponential Distribution has property of "memoryless".

The exponential distribution with mean 2 is χ^2_2.

Chi-Squared Distribution

pdf:$f(x; v) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2}, x \ge 0$

It is a specific gamma distribution where $\alpha = v/2, \beta = 2$.

v is called degrees of freedom (df) of X.

The symbol "χ^2_v" represents chi-squared distribution with v df.

Constructing the chi-squared distribution:

If Z has a standard normal distribution and $X = Z^2$, then the pdf of X is chi-squared with 1 df, $X \sim \chi^2_1$.

If $X_1 \sim \chi^2_{v_1}, X_2 \sim \chi^2_{v_2}$ and they are independent, then $X_1 + X_2 \sim \chi^2_{v_1+v_2}$.

If $Z_1, \cdots, Z_n$ are independent and each has the standard normal distribution, then $Z_1^2 + \cdots + Z_n^2 \sim \chi^2_n$.

If $X_1, \cdots, X_n$ are a random sample from a normal distribution□ $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

The χ^2_2 is an exponential distribution with mean 2.

Weibull Distribution

pdf: $f(x;\alpha,\beta) = \frac{\alpha}{\beta}x^{\alpha-1}e^{-(x/\beta)^\alpha}, x \geq 0$

Mean: $\mu = \beta\Gamma(1+\frac{1}{\alpha})$

Variance: $\sigma^2 = \beta^2\left\{\Gamma\left(1+\frac{2}{\alpha}\right) - \left[\Gamma\left(1+\frac{1}{\alpha}\right)\right]^2\right\}$

cdf: $F(x;\alpha,\beta) = 1 - e^{-(x/\beta)^\alpha}, x \geq 0$

Weibull distribution can have a third parameter γ, this equals to $X-\gamma$ having the above-mentioned pdf. So $x-\gamma$ replaces s in the new cdf.

Lognormal Distribution

X is lognormal distributed when $ln(X)$ has a normal distribution, whose mean is μ and standard deviation is σ.

pdf: $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x}e^{-[ln(x)-\mu]^2/(2\sigma^2)}, x \geq 0$

Mean: $E(X) = e^{\mu+\sigma^2/2}$

Variance: $V(X) = e^{2\mu+\sigma^2}\cdot(e^{\sigma^2}-1)$

cdf: $F(x;\mu,\sigma) = P(X \leq x) = \Phi\left[\frac{ln(x)-\mu}{\sigma}\right]$

Beta Distribution

pdf: $f(x;\alpha,\beta,A,B) = \frac{1}{B-A}\cdot\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)}\left(\frac{x-A}{B-A}\right)^{\alpha-1}\left(\frac{B-x}{B-A}\right)^{\beta-1}$

, where $x \in [A,B]$

Mean: $\mu = A + (B-A)\cdot\frac{\alpha}{\alpha+\beta}$

Variance: $\sigma^2 = \frac{(B-A)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Standard Beta Distribution: $A=0, B=1$

cdf: Use the general integration method.

⋆ Remember that the lower bound is A.

Multinormal distribution

n: number of trials, r: number of possible outcomes, p_i*: possibility of outcome being i in a trial.* X_i*: the number of trials resulting the* i^{th} *outcome* $(i \in [0,r])$

⋆ *The binomial distribution is a specific case when n = n, r = 2,* X_1 *= failure(0),* X_2 *= success(1).*

$p(x_1,\cdots,x_r) = \frac{n!}{(x_1!)(x_2!)\cdots(x_r!)}p_1^{x_1}\cdots p_r^{x_r},\ x_i = 0,1,2\cdots$, with $x_1+\cdots+x_r = n$

Bivariate normal distribution

pdf: $f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^{\Delta}$, where $\Delta = -\left\{[(x-\mu_1)/\sigma_1]^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + [\frac{(y-\mu_2)}{\sigma_2}]^2\right\}/[2(1-\rho^2)]$

μ_1,σ_1 *and* μ_2,σ_2*: mean and standard deviation of X and Y;* ρ*: the correlation between X and Y.*

Conditional mean: $\mu_{Y|X=x} = E(Y|X=x) = \mu_2 + \rho\sigma_2\frac{x-\mu_1}{\sigma_1}$

Conditional variance: $\sigma^2_{Y|X=x} = V(Y|X=x) = \sigma_2^2(1-\rho^2)$

Chi-squared Distribution is also a part of this section.

t Distribution

Let Z be a standard normal rv and let X be a χ^2_v rv independent of Z. The t distribution with df v is: $T = \frac{Z}{\sqrt{X/v}}$

If $X_1,\cdots,X_n$ is a random sample from a normal distribution $N(\mu,\sigma^2)$, then $T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ has the t distribution with $(n-1)$ df.

pdf: $f(t) = \frac{1}{\sqrt{\pi v}}\frac{\Gamma[(v+1)/2]}{\Gamma(v/2)}\frac{1}{(1+t^2/v)^{(v+1)/2}}, t \in R$

Cauchy Distribution

t distribution with 1 df, pdf: $\frac{1}{\pi(1+t^2)}$

F distribution

Let X_1 and X_2 be independent chi-squared random variables with v_1 and v_2 df. The F distribution with v_1 numerator df and v_2 denominator df is: $F = \frac{X_1/v_1}{X_2/v_2}$

R Language

Basic calculation

```
(7 * 3) + 12 / 2 - 7 ^ 2 + sqrt(4)
log(3) + sqrt(2) * sin(pi) - exp(3)
```

Creating vector

`c(1,2,3,4,5)`, out: `1 2 3 4 5`

`24:20`, out: `24 23 22 21 20`

```
seq(from = 5, to = 25, by = 5)
seq(from = 5, by = 3, length.out = 6)
```

out: `5 8 11 14 17 20`

```
rep(x=5, times = 10)
rep(x=1:5, length.out = 15)
```

out: `1 2 3 4 5 1 2 3 4 5 1 2 3 4 5`

```
rep(x=1:3, times = 3:1)
```

out: `1 1 1 2 2 3`

```
vec <- c(1,2,3,14,15)
```

`vec[ c(1,4) ]`, out: `1 14`

More about vectors

Boolean calculation of vectors:

```
x <- 5; y <- c(3,5,7)
```

`y <= x`, out: `TRUE TRUE FALSE`

Assuming X and Y are two vectors with the same length:

`X & Y, X | Y, X == Y, X != Y` each produces a vector by comparing each of the elements in X and Y.

`X && Y, X || Y` produce a single answer by looking at the first element of X and Y.

Loop and If

```
#Ignoring indent so as to save space!
for(i in c(10,20,30)){if(i == 20){a <- i + a}}
while(i < 10){i <- i + 1}
```

R code in Homework

Stem-leaf graph:

```
library(aplpack)
stem.leaf.backback(vec1,vec2, m=1,depths = FALSE)
```

Frequency distribution table:

```
transform(table(cut(vec, breaks)))
```

Histogram: `hist(vec, breaks = breaks)`

The proportion of elements in the sample are less than 100:

```
length(vec[vec < 100]) / length(vec)
```

The sample median, 10% trimmed mean, and sample mean:

```
median(vec); mean(vec, trim = 0.1); mean(vec)
```

Sum of the elements, sum of the square of the elements, sample variance and the standard deviation:

```
sum(vec); sum(vec^2); var(vec); sd(vec)
```

Upper and lower fourth:

```
quantile(vec, 0.75); quantile(vec, 0.25)
```

Sorting: `sort(vec, decreasing = TRUE)`

Boxplot: `boxplot(vec1,vec2,names=c("Name1","Name2"))`

Normal probability plot + line: `qqnorm(vec); qqline(vec)`

R normal distribution

`dnorm(vec,mean_,sd)` returns the vector pdf value of the input vector.

`pnorm(vec,mean_,sd)` returns the cdf.

`qnorm(vec,mean_,sd)` returns the inverse function of cdf.

`rnorm(n,mean_,sd)` gives n random samples from normal distribution.

R binomial distribution

```
dbinorm(vec,size,prob); pbinorm(vec,size,prob)
qbinorm(vec,size,prob); rbinorm(n,size,prob)
```

Point Estimation

point estimator of θ : $\hat{\theta}$

Consistency: A consistent estimator converge toward the parameter being estimated as the sample size increases.

Bias of $\hat{\theta}$: $E(\hat{\theta})-\theta$; Unbiased: $E(\overline{X}) = \mu$

Mean square error (MSE) of an estimator $\hat{\theta}$: $E[(\hat{\theta}-\theta)^2]$

$MSE = V(\hat{\theta}) + [E(\hat{\theta})-\theta]^2$ = variance of estimator + bias^2

Unbiased estimator of θ: $E(\hat{\theta}) = \theta$

Efficiency: a more efficient estimator has smaller variance

Minimum variance unbiased estimator (MVUE): the one unbiased estimator of θ that has minimum variance

Let $X_1,\cdots,X_n$ be a random sample from a $\mathcal{N}(\mu,\sigma)$, the estimator $\hat{\mu} = \overline{X}$ is the MVUE for μ

Standard error of the estimator $\hat{\theta}$: $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$

Estimated standard error: $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$

The Bootstrap

The population pdf is $f(x;\theta)$, and data $x_1,\cdots,x_n$ gives $\hat{\theta} = \hat{\theta}_0$. Obtain "bootstrap samples" from $f(x;\hat{\theta}_0)$, and for each sample, calculate a "bootstrap estimate" $\hat{\theta}_1^*,\cdots,\hat{\theta}_B^*$. Let $\overline{\theta}^* = \sum\hat{\theta}_i^*/B$.

The Bootstrap estimate of $\hat{\theta}$'s standard error is:

$S_{\hat\theta} = \sqrt{\frac{1}{B-1}\sum(\hat\theta_i^* - \overline{\theta}^*)^2}$

The Method of Moments

Let $X_1, \cdots, X_n$ be a random sample from $f(x)$. The k^{th} population moment, or k^{th} moment of the distribution $f(x)$, is $E(X^k)$. The k^{th} sample moment is $\frac{1}{n}\sum_{i=1}^n X_i^k$.

Let $X_1, \cdots, X_n$ be random sample from distribution with $f(x; \theta_1, \cdots, \theta_m)$, where θ_i are unknown. Then the moment estimators $\theta_1, \cdots, \theta_m$ are obtained by equating the first *m* sample moments to the corresponding first *m* population moments and solving for $\theta_1, \cdots, \theta_m$.

Maximum Likelihood Estimation

Let $X_1, \cdots, X_n$ have joint pmf/pdf $f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)$, where the parameters $\theta_1, \cdots, \theta_m$ have unknown values. When x_i are the observed sample values and the equation is regarded as a function of θ_i, it is called the likelihood function.

The maximum likelihood estimates $\hat\theta_1, \cdots, \hat\theta_m$ are those values of the θ_is that maximize the likelihood function:
$f(x_1, \cdots, x_n; \hat\theta_1, \cdots, \hat{\theta_m}) \geq f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m)$ for any θ_i

When the X_i are substituted in place of the x_i, the maximum likelihood estimators (mle's) result.

A common method is to take "ln" against the original likelihood function and let its derivatives against the estimator be 0.

Some properties of MLEs

- The Invariance Principle

Let $\hat\theta_1, \cdots, \hat\theta_m$ be the mle's of the parameters $\theta_1, \cdots, \theta_m$. Then the mle of any function $h(\theta_1, \cdots, \theta_m)$ of these parameters is the function $h(\hat\theta_1, \cdots, \hat{\theta_m})$.

- Large Sample Behavior of the MLE

$\hat\theta$ is approximately the MVUE of θ.

- Large Sample Properties of the MLE:

Given a random sample $X_1, \cdots, X_n$ from $f(x;\theta)$, assume that the set of possible x values does not depend on θ. Then for large n the MLE $\hat\theta$ has approximately a normal distribution with mean θ and variance $1/[nI(\theta)]$. The limiting distribution of $\sqrt{n}(\hat\theta - \theta)$ is normal with mean 0 and variance $1/I(\theta)$.

Information and Efficiency

Fisher information in a single observation from $f(x;\theta)$:
$I(\theta) = V\left[\frac{\partial}{\partial\theta}ln(f(X;\theta))\right] = -E\left[\frac{\partial^2}{\partial\theta^2}ln(f(X;\theta))\right]$

Fisher information in a random sample $X_1, \cdots, X_n$ with $f(x;\theta)$:

$$I_n(\theta) = V\left[\frac{\partial}{\partial\theta}lnf(X_1;\theta) + \cdots + \frac{\partial}{\partial\theta}lnf(X_n;\theta)\right]$$

$$= nV\left[\frac{\partial}{\partial\theta}lnf(X_1;\theta)\right] = nI(\theta)$$

Cramér-Rao Inequality:

Assume a random sample $X_1, \cdots, X_n$ from $f(x;\theta)$ such that the set of possible values does not depend on θ. If the statistic $T = t(X_1, \cdots, X_n)$ is an unbiased estimator for the parameter θ, then:
$V(T) \geq \frac{1}{V\{\frac{\partial}{\partial\theta}[lnf(X_1, \cdots, X_n;\theta)]\}} = \frac{1}{nI(\theta)} = \frac{1}{I_n(\theta)}$

Let T be an unbiased estimator of θ.

Efficiency: the ratio of the lower bound of Cramér-Rao Inequality to the variance of T.

T is an efficient estimator if T achieves the Cramér-Rao lower bound (the efficiency is 1). And it is a MVUE.

Statistical Intervals Based on a Single Sample

The following CI means confident interval.

CI (Confidence Interval) for normal distribution

A 100(1-α)% CI for the mean μ of a normal population where σ is known is: $\left(\overline{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$

Sample size n necessary to ensure an interval width w is $n = (2z_{\alpha/2}\cdot\frac{\sigma}{w})^2$.

The above $z_{\alpha/2} = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1-\alpha/2)$, meaning the area under the curve to the right of $z_{\alpha/2}$ is $\alpha/2$

A large-sample asymptotic interval for μ

A large sample CI for μ with confidence level approximately $100(1-\alpha)\%$ is $\overline{x} \pm z_{\alpha/2}\cdot\frac{s}{\sqrt{n}}$, regardless of the shape of the population distribution. Can be applied when $n > 40$.

A general large-sample CI

Suppose that $\hat\theta$ is an estimator who has approximately a normal distribution, an available expression for $\sigma_{\hat\theta}$ and is unbiased.

Then $P\left(-z_{\alpha/2} < \frac{\hat\theta - \theta}{\sigma_{\hat\theta}} < z_{\alpha/2}\right) \approx 1-\alpha$.

Large-sample Score CI for Proportion [Recommended]

Let p denote the proportion of "successes" in a population. X is the number of successes in the sample. X can be regarded as a binomial rv with E(X) = np and $\sigma_x = \sqrt{np(1-p)}$.

Let $\tilde{p} = \frac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n}$. Then a CI for a population proportion p with confidence level approximately $100(1-\alpha)\%$ is
$\tilde{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}\hat{q}/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$ where $\hat{q} = 1 - \hat{p}$. This is often referred to as the "score CI" for p.

If an upper/lower bound is needed, replace all the $z_{\alpha/2}$ with z_α and choose the +/- sign.

Sample size n necessary to give interval - width w is
$n = \frac{2z^2\hat{p}\hat{q} - z^2w^2 \pm \sqrt{4z^4\hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2z^4}}{w^2}$, approximately $n = \frac{4z_{\alpha/2}^2\hat{p}\hat{q}}{w^2}$

Simpler Traditional CI for a proportion

If n is large, then the score CI is approximately $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$

One-sided Large-sample Confidence Bound

Confidence level of the following intervals is about $100(1-\alpha)\%$
Large-sample upper/lower confidence bound for μ: $\overline{x} \pm z_\alpha\cdot\frac{s}{\sqrt{n}}$

Intervals Based on a Normal Population Distribution

Let $t_{\alpha,v}$ = the number on the x-axis for which the area under the t curve with v df to the right of $t_{\alpha,v}$ is α. It is called "t critical value".

Let $\overline{x}$ and s be the sample mean and sd computed from the results of a random sample from a normal population with mean μ. Then a $100(1-\alpha)\%$ CI for μ, the one-sample t CI,is
$\left(\overline{x} - t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}},\ \overline{x} + t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}\right)$.

The upper/lower confidence bound for $\mu: \overline{x} \pm t_{\alpha,n-1}\cdot\frac{s}{\sqrt{n}}$

A $100(1-\alpha)\%$ CI for the variance σ^2 of a normal population has lower limit $(n-1)s^2/\chi^2_{\alpha/2,n-1}$ and upper limit $(n-1)s^2/\chi^2_{1-\alpha/2,n-1}$.

Prediction Interval for a Single Future Value

A prediction interval for a single observation to be selected from a normal population distribution is $\overline{x} \pm t_{\alpha/2,n-1}\cdot s\sqrt{1+\frac{1}{n}}$. The prediction level is $100(1-\alpha)\%$

Bootstrap Percentile Interval

The bootstrap percentile interval with a confidence level of $100(1-\alpha)\%$ for a specified parameter is obtained by:

-Generate B bootstrap samples, for each calculate particular statistics that estimates the parameter and sort them ascending.

-Compute $k = \alpha(B+1)/2$ and choose the k^{th} value from each end of the sorted list. The two values are the confidence limits.

One-sample Hypothesis tests

H_0: null hypothesis, initially assumed to be true.

H_a: alternative hypothesis, in contradiction to H_0.

The test procedure is specified by 1) test statistic 2) rejection region. H_0 will be rejected iff. 1) falls in 2).

Type I error: rejecting H_0 when it's true. Probability: α, the integration of the rejection region.

Type II error: accepting H_0 when it's false. Probability: β

Power: $1-\beta$

Normal Population with Known σ

H_0: $\mu = \mu_0$; Test statistic: $z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$. Let $\Delta = \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}$

H_a	Rejection Region	β
$\mu > \mu_0$	$z \geq z_\alpha$	$\Phi\left(z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
$\mu < \mu_0$	$z \leq -z_\alpha$	$1 - \Phi\left(-z_\alpha + \frac{\mu_0 - \mu'}{\sigma/\sqrt{n}}\right)$
$\mu \neq \mu_0$	$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$	$\Phi\left(z_{\alpha/2} + \Delta\right) - \Phi\left(-z_{\alpha/2} + \Delta\right)$

Sample size n for which a level α test has some β at alternative value μ' is $\left[\frac{\sigma(z_\alpha + z_\beta)}{\mu_0 - \mu'}\right]^2$ for one tail, $\left[\frac{\sigma(z_{\alpha/2} + z_\beta)}{\mu_0 - \mu'}\right]^2$ for two tails.

Large-Sample Tests

Replacing σ with s in normal population test when making z.

Normal Population with UNKNOWN σ

H_0: $\mu = \mu_0$; Test statistic: $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$; H_a and rej. region:
$\mu > \mu_0 : t \geq t_{\alpha,n-1}$; $\quad \mu < \mu_0 : t \leq -t_{\alpha,n-1}$
$\mu \neq \mu_0 : t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$

Large-Sample tests concerning population proportion

"Large" when $np_0 \geq 10$ and $n(1-p_0) \geq 10$.

H_0: $p = p_0$; Test statistic: $z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$, rejection region:
$p > p_0 : z \geq z_\alpha$; $\quad p < p_0 : z \leq -z_\alpha$
$p \neq p_0 : z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

The β value for two-tailed test:

$$\Phi\left[\frac{p_0-p'+z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right] - \Phi\left[\frac{p_0-p'-z_{\alpha/2}\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$$

for $H_a : p > p_0$ and $H_a : p < p_0$, accordingly:

$$\Phi\left[\frac{p_0-p'+z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right], \text{ and } 1-\Phi\left[\frac{p_0-p'-z_\alpha\sqrt{p_0(1-p_0)/n}}{\sqrt{p'(1-p')/n}}\right]$$

Sample size n for which the level α test satisfies some β is:

one tail: $\left[\frac{z_\alpha\sqrt{p_0(1-p_0)}+z_\beta\sqrt{p'(1-p')}}{p'-p_0}\right]^2$, two tails: replace α with $\alpha/2$

Small-Sample tests concerning population proportion

When the samples are based directly on the binomial distribution.

H_0: $p = p_0$; $H_a : p > p_0$.

Test statistic: X, number of successes in the sample.

The upper-tailed rejection region is $x \geq c$, where c is the largest number satisfying $B(c; n, p_0) \neq \alpha$.

$P(\text{type I error}) = 1 - B(c-1; n, p_0)$

$\beta(p') = P(\text{type II error when } p = p') = B(c-1; n, p')$

P-values

P-value is a probability calculated assuming that the null hypothesis is true.

To determine it, first decide which values of the test statistic are at least contradicting H_0.

P-value is the smallest significance level α where H_0 can be rejected. Also referred to as the observed significance level (OSL).

> Reject H_0 if $p \leq \alpha$, accept if $p > \alpha$.
> Test something, something is H_a

For z tests

$$p = \begin{cases} 1-\Phi(z) & \text{for an upper-tailed test} \\ \Phi(z) & \text{for a lower-tailed test} \\ 2[1-\Phi(|z|)] & \text{for a two-tailed test} \end{cases}$$

For t tests

Respectively, $t_{p,n-1} =$ calculated t; $t_{1-p,n-1} =$ calculated t; $t_{p/2,n-1} =$ calculated t and solve for p.

Two sample hypothesis tests

$X_1, \cdots, X_m$ is a random sample from a population with mean μ_1 and variance σ_1^2. Y_i is similar, with μ_2 and σ_2^2. X and Y are independent. $E(\overline{X}-\overline{Y}) = \mu_1 - \mu_2$, $\sigma_{\overline{X}-\overline{Y}} = \sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}$

Normal populations with known variances

$H_0 : \mu_1 - \mu_2 = \Delta_0$. Test statistic: $z = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n}}}$
$\mu_1 - \mu_2 > \Delta_0 : z \geq z_\alpha$; $\quad \mu_1 - \mu_2 < \Delta_0 : z \leq -z_\alpha$
$\mu_1 - \mu_2 \neq \Delta_0 : z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

p value is the same as the last part.

The β for two-tailed test:

$$\Phi\left(z_{\alpha/2} - \frac{\Delta'-\Delta_0}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta'-\Delta_0}{\sigma}\right)$$

for $H_a : \mu_1 - \mu_2 > \Delta_0$ and $H_a : \mu_1 - \mu_2 < \Delta_0$, correspondingly:

$$\Phi\left(z_\alpha - \frac{\Delta'-\Delta_0}{\sigma}\right) \text{ and } 1-\Phi\left(-z_\alpha - \frac{\Delta'-\Delta_0}{\sigma}\right)$$

where real $\mu_1 - \mu_2 = \Delta'$, $\sigma = \sigma_{\overline{X}-\overline{Y}} = \sqrt{(\sigma_1^2/m)+(\sigma_2^2/n)}$

Sample size m and n needed is the value satisfying:

$\frac{\sigma_1^2}{m}+\frac{\sigma_2^2}{n} = \frac{(\Delta'-\Delta_0)^2}{(z_\alpha+z_\beta)^2}$, when $m = n$, $m = n = \frac{(\sigma_1^2+\sigma_2^2)(z_\alpha+z_\beta)^2}{(\Delta'-\Delta_0)^2}$.

That's for one-tailed test, replacing α with $\alpha/2$ for two-tailed test.

Large sample tests with UNKNOWN variances

Replace σ_1, σ_2 with s_1, s_2 in the formula above when $m, n > 40$.

CI for $\mu_1 - \mu_2$ with a confidence level of about $100(1-\alpha)$%:

$$\overline{x} - \overline{y} \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}$$

This is for two-tailed. For one-tailed CI, replacing $z_{\alpha/2}$ with z_α and choose the appropriate sign. Sample size needed for a $100(1-\alpha)$% CI of width w is $n = \frac{4z_{\alpha/2}^2(\sigma_1^2+\sigma_2^2)}{w^2}$

Small sample tests

t CI for $\mu_1 - \mu_2$ with confidence level $100(1-\alpha)$% is:

$$\overline{x} - \overline{y} \pm t_{\alpha/2,v}\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}, \text{ where } v = \left\lfloor \frac{\left(\frac{s_1^2}{m}+\frac{s_2^2}{n}\right)^2}{\frac{(s_1^2/m)^2}{m-1}+\frac{(s_2^2/n)^2}{n-1}} \right\rfloor$$

two-sample t test:

H_0: $\mu_1 - \mu_2 = \Delta_0$, Test statistic: $t = \frac{\overline{x}-\overline{y}-\Delta_0}{\sqrt{\frac{s_1^2}{m}+\frac{s_2^2}{n}}}$
$\mu_1 - \mu_2 > \Delta_0 : t \geq t_{\alpha,v}$; $\quad \mu_1 - \mu_2 < \Delta_0 : t \leq -t_{\alpha,v}$
$\mu_1 - \mu_2 \neq \Delta_0 : t \geq t_{\alpha/2,v}$ or $t \leq -t_{\alpha/2,v}$

Paired t-Tests

Data: n independently selected pairs $(X_1, Y_1), \cdots, (X_n, Y_n)$ with $E(X_i) = \mu_1$ and $E(Y_i) = \mu_2$. Let $D_i = X_i - Y_i$. The D_i are normally distributed with mean μ_D and variance σ_D^2.

$H_0 : \mu_D = \Delta_0$, Test statistic: $t = \frac{\overline{d}-\Delta_0}{s_D/\sqrt{n}}$
$\mu_D > \Delta_0 : t \geq t_{\alpha,n-1}$; $\quad \mu_D < \Delta_0 : t \leq -t_{\alpha,n-1}$
$\mu_D \neq \Delta_0 : t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$

p-value can be calculated as was done for earlier t tests.

Two-side paired t CI for μ_D: $\overline{d} \pm t_{\alpha/2,n-1} \cdot s_D/\sqrt{n}$, for one-side, replace $t_{\alpha/2}$ with t_α

Difference of two sample proportions

Let $X \sim Bin(m, p_1)$ and $Y \sim Bin(n, p_2)$, independent.
$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$, $V(\hat{p}_1 - \hat{p}_2) = \frac{p_1q_1}{m} + \frac{p_2q_2}{n}$, where $q_i = 1 - p_i$.

Let $\hat{p} = \frac{X+Y}{m+n} = \frac{m}{m+n}\hat{p}_1 + \frac{n}{m+n}\hat{p}_2$, $\hat{q} = 1 - \hat{p}$

H_0: $p_1 - p_2 = 0$. Test statistic: $z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{m}+\frac{1}{n}\right)}}$
$p_1 - p_2 > p_0 : z \geq z_\alpha$; $\quad p_1 - p_2 < p_0 : z \leq -z_\alpha$
$p_1 - p_2 \neq p_0 : z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$

p-value is the same as previous z tests.

Calculating beta:

H_0	$\beta(p_1, p_2)$
$p_1 - p_2 > 0$	$\Phi\left[\Delta(z_\alpha)\right]$
$p_1 - p_2 < 0$	$1 - \Phi\left[\Delta(-z_\alpha)\right]$
$p_1 - p_2 \neq 0$	$\Phi\left[\Delta(z_{\alpha/2})\right] - \Phi\left[\Delta(-z_{\alpha/2})\right]$

where $\Delta(z) = \frac{z\sqrt{\overline{p}\overline{q}\left(\frac{1}{m}+\frac{1}{n}\right)}-(p_1-p_2)}{\sigma}$

$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1q_1}{m}+\frac{p_2q_2}{n}}$, $\overline{p} = \frac{mp_1+np_2}{m+n}$, $\overline{q} = \frac{mq_1+nq_2}{m+n}$

sample size needed:

$n = \frac{\left[z_\alpha\sqrt{(p_1+p_2)(q_1+q_2)/2}+z_\beta\sqrt{p_1q_1+p_2q_2}\right]^2}{d^2}$, where $p_1 - p_2 = d$.

That's for one tail, replace α with $\alpha/2$ for two tail.

The $100(1-\alpha)$% CI is $\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{m}+\frac{\hat{p}_2\hat{q}_2}{n}}$.

Inference - Two Population Variances, F-distribution

Let $X_1, \ldots, X_m$ be a random sample from a normal distribution with variance σ_1^2, and Y_i with σ_2^2, independently. Let S_1^2 and S_2^2 denote the two sample variances. Then $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ has an F distribution with $v_1 = m-1$ and $v_2 = n-1$.

H_0: $\sigma_1^2 = \sigma_2^2$, Test statistic: $f = s_1^2/s_2^2$
$\sigma_1^2 > \sigma_2^2 : f \geq F_{\alpha,m-1,n-1}$; $\quad \sigma_1^2 < \sigma_2^2 : f \leq F_{1-\alpha,m-1,n-1}$
$\sigma_1^2 \neq \sigma_2^2 f \geq F_{\alpha/2,m-1,n-1}$ or $f \leq F_{1-\alpha/2,m-1,n-1}$

p = area under the F curve to the right of the calculated f.

ANOVA

X_{ij}: the rv denoting the j^{th} measurement from the i^{th} population; x_{ij}: the observed value of X_{ij}.

H_0: $\mu_1 = \cdots = \mu_I$, the mean of all populations are equal.

Assume X_{ij} is normally distributed: $E(X_{ij}) = \mu_i$, $V(X_{ij}) = \sigma^2$.

$\overline{X}_{i.} = \frac{\sum_{j=1}^J X_{ij}}{J}$, $i = 1, \cdots, I$; $\overline{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^J X_{ij}}{IJ}$

$S_i^2 = \frac{\sum_{j=1}^J \left(X_{ij}-\overline{X}_i\right)^2}{J-1}$, $i = 1, \cdots, I$

Treatment sum of squares:

SSTr $= J\left[\left(\overline{X}_{1.} - \overline{X}_{..}\right)^2 + \cdots + \left(\overline{X}_{I.} - \overline{X}_{..}\right)^2\right]$

Error sum of squares:

SSE $= \sum_i \sum_j \left(X_{ij} - \overline{X}_{i.}\right)^2 = (J-1)\left[S_1^2 + S_2^2 + \cdots + S_I^2\right]$

Total sum of squares: SST = $\sum_i \sum_j \left(x_{ij} - \overline{x}_{..}\right)^2$
Mean square for treatments: MSTR = SSTr$/(I-1)$
Mean square for error: MSE = SSE$/[I(J-1)]$
SSE$/\sigma^2$ has a χ^2 distribution with $I(J-1)$ df.
When H_0 is true, SSTr$/\sigma^2$ has a χ^2 distribution with $I-1$ df.

F test

Computing Formula: $x_{i.}$*: **sum** of all* x_{ij} *for fixed i;* $x_{..}$*: **sum** of all* x_{ij}
SST = $\sum_i \sum_j x_{ij}^2 - x_{..}^2/IJ$, df = $IJ-1$.
SSTr = $\frac{\sum_i x_{i.}^2}{J} - \frac{x_{..}^2}{IJ}$, df = $I-1$
SSE = SST - SSTr, df = $I(J-1)$,
Test statistic: $F = \frac{\text{SSTr}/(I-1)}{\text{SSE}/I(J-1)} = \frac{\text{MSTr}}{\text{MSE}}$
Rejection region: $f \geq F_{\alpha,I-1,I(J-1)}$ for an upper-tailed test with the significance level α.

P-value for it is the area under the relevant F curve to the right of the calculated f.

ANOVA Table, the following "^{2}" means "Square"

Source	df	Sum of ^{2}s	Mean ^{2}	f
Treatments	I-1	SSTr	MSTr	MSTr/MSE
Error	I(J-1)	SSE	MSE	
Total	IJ-1	SST		

Tukey's Procedure

Let $Z_1, \cdots, Z_m$ be m independent standard normal rv's and W be a χ^2 rv with v df.
$Q = \frac{\max|Z_i - Z_j|}{\sqrt{W/v}} = \frac{\max(Z_1,\ldots,Z_m) - \min(Z_1,\ldots,Z_m)}{\sqrt{W/v}}$
is called the studentized range distribution with parameters: m: the number of Z_i, v: denominator df. Critical value $Q_{\alpha,m,v}$ captures upper-tail area α under the density curve of Q.

For each $i < j$, form the interval:
$\overline{x}_{i.} - \overline{x}_{j.} \pm Q_{\alpha,I,I(J-1)}\sqrt{MSE/J}$.

There are $I(I-1)/2$ such intervals, each for $\mu_1 - \mu_2, \cdots, \mu_{I-1} - \mu_I$. The simultaneous confidence level that every interval includes the corresponding $\mu_i - \mu_j$ is $100(1-\alpha)$%.

The procedure: Select α, extract $Q_{\alpha,I,I(J-1)}$ and calculate $w = Q_{\alpha,I,I(J-1)} \cdot \sqrt{\text{MSE}/J}$. List the sample means in increasing order and underline those pairs that differ by less than w. Any pair of sample means not underscored by the same line corresponds to a pair of population or treatment means that are judged significantly different. w is called Tukey's honestly significantly difference (HSD).

CI for other parametric functions

$100(1-\alpha)$% CI for $\sum c_i \mu_i$:
$\sum c_i \overline{x}_{i.} \pm t_{\alpha/2,I(J-1)} \sqrt{\left(\text{MSE} \sum c_i^2\right)/J}$

Alternative description for ANOVA

$H_0 : \alpha_1 = \cdots = \alpha_I = 0$, $E(\text{MSTr}) = \sigma^2 + \frac{J}{I-1}\sum \alpha_i^2$

Single-Factor ANOVA with unequal sample sizes

n: total number of observations

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i} X_{ij}^2 - \frac{1}{n}X_{..}^2$$

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{i.} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{I}\frac{1}{J_i}X_{i.}^2 - \frac{1}{n}X_{..}^2$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\left(X_{ij} - \overline{X}_{i}\right)^2 = \text{SST} - \text{SSTr}$$

Test statistic value: $f = \frac{\text{MSTr}}{\text{MSE}} = \frac{\text{SSTr}/(I-1)}{\text{SSE}/(n-I)}$
Rejection region: $f \geq F_{\alpha,I-1,n-I}$

Multiple comparisons with unequal sample sizes

Let $w_{ij} = Q_{\alpha,I,n-I} \cdot \sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)}$, the probability is about $1-\alpha$ that $\overline{X}_i. - \overline{X}_j. - w_{ij} \leq \mu_i - \mu_j \leq \overline{X}_i. - \overline{X}_j. + w_{ij}$ for every i and j with $i \neq j$

A Random Effects Model

H_0: $\sigma_A^2 = 0$, Test statistic: $F = \frac{\text{MSTr}}{\text{MSE}}$, reject H_0 if $f \geq F_{\alpha,I-1,n-I}$

Two-Factor ANOVA with $K_{ij} = 1$

The estimators: $\hat{\mu} = \overline{X}.$; $\hat{\alpha}_i = \overline{X}_{i.} - \overline{X}.$; $\hat{\beta}_j = \overline{X}_{.j} - \overline{X}$
SST = $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{..}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}^2 - \frac{1}{IJ}X_{..}^2$, df = $IJ-1$
SSA = $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{i.} - \overline{X}_{..}\right)^2 = \frac{1}{J}\sum_{i=1}^{I} X_{i.}^2 - \frac{1}{IJ}X_{..}^2$, df = $I-1$
SSB = $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{.j} - \overline{X}_{..}\right)^2 = \frac{1}{I}\sum_{j=1}^{J} X_{.j}^2 - \frac{1}{IJ}X_{..}^2$, df = $J-1$
SSE = $\sum_{i=1}^{I}\sum_{j=1}^{J}\left(X_{ij} - \overline{X}_{i.} - \overline{X}_{.j} + \overline{X}_{..}\right)^2$, df = $(I-1)(J-1)$
SST = SSA + SSB + SSE

H_{0A}: $\alpha_1 = \cdots = \alpha_I = 0$, H_{aA}: at least one $a_i \neq 0$; Test statistic: $f_A = \frac{\text{MSA}}{\text{MSE}} = \frac{\text{SSA}/(I-1)}{\text{SSE}/[(I-1)(J-1)]}$, rejection region: $f_A \geq F_{\alpha,I-1,(I-1)(J-1)}$

H_{0B}: $\beta_1 = \cdots = \beta_J = 0$, H_{aB}: at least one $\beta_j \neq 0$; Test statistic: $f_B = \frac{\text{MSB}}{\text{MSE}} = \frac{\text{SSB}/(J-1)}{\text{SSE}/[(I-1)(J-1)]}$, rejection region: $f_B \geq F_{\alpha,J-1,(I-1)(J-1)}$

Multiple Comparisons

For comparing A, $w = Q_{\alpha,I,(I-1)(J-1)} \cdot \sqrt{\text{MSE}/J}$; for comparing B, $w = Q_{\alpha,J,(I-1)(J-1)} \cdot \sqrt{\text{MSE}/I}$. Arrange the sample means in increasing order, underscore those pairs differing by less than w, identify pairs not underscored by the same line as corresponding to significantly different levels of the given factor.

Two-Factor ANOVA with Replications($K_{ij} > 1$)

$\mu = \frac{1}{IJ}\sum_i \sum_j \mu_{ij}, \overline{\mu}_{i.} = \frac{1}{J}\sum_j \mu_{ij}, \overline{\mu}_{.j} = \frac{1}{I}\sum_i \mu_{ij}$
$\alpha_i = \overline{\mu}_{i.} - \mu, \beta_j = \overline{\mu}_{.j} - \mu, \gamma_{ij} = \mu_{ij} - (\mu + \alpha_i + \beta_j)$
SST = $\sum\limits_i \sum\limits_j \sum\limits_k (X_{ijk} - \overline{X}_{...})^2 = \sum\limits_i \sum\limits_j \sum\limits_k X_{ijk}^2 - \frac{1}{IJK}X_{...}^2$, df = $IJK-1$
SSE = $\sum\limits_i \sum\limits_j \sum\limits_k (X_{ijk} - \overline{X}_{ij.})^2 = \sum\limits_i \sum\limits_j \sum\limits_k X_{ijk}^2 - \frac{1}{K}\sum\limits_i \sum\limits_j X_{ij.}^2$, df = $IJ(K-1)$
SSA = $\sum_i \sum_j \sum_k (\overline{X}_{i..} - \overline{X}_{...})^2 = \frac{1}{JK}\sum_i X_{i..}^2 - \frac{1}{IJK}X_{...}^2$, df = $I-1$
SSB = $\sum_i \sum_j \sum_k (\overline{X}_{.j.} - \overline{X}_{...})^2 = \frac{1}{IK}\sum_j X_{.j.}^2 - \frac{1}{IJK}X_{...}^2$, df = $J-1$
SSAB = $\sum_i \sum_j \sum_k (X_{ij.} - \overline{X}_{i..} - \overline{X}_{.j.} + \overline{X}_{...})^2$, df = $(I-1)(J-1)$
SST = SSA + SSB + SSAB + SSE

H_{0A}: $\alpha_1 = \cdots = \alpha_I = 0$; H_{aA}: at least one $\alpha_i \neq 0$.
Test statistic: $f_A = \frac{\text{MSA}}{\text{MSE}}$, Rej. region: $f_A \geq F_{\alpha,I-1,IJ(K-1)}$
H_{0B}: $\beta_1 = \cdots = \beta_J = 0$; H_{aB}: at least one $\beta_j \neq 0$.
Test statistic: $f_B = \frac{\text{MSB}}{\text{MSE}}$, Rej. region: $f_B \geq F_{\alpha,J-1,IJ(K-1)}$
H_{0AB}: $\gamma_{ij} = 0$ for all i, j; H_{aAB}: at least one $\gamma_{ij} \neq 0$.
Test statistic: $f_{AB} = \frac{\text{MSAB}}{\text{MSE}}$, Rej. region: $f_{AB} \geq F_{\alpha,(I-1)(J-1),IJ(K-1)}$

Regression

Linear Regression Model: $Y = \beta_0 + \beta_1 x + \varepsilon$, the rv ε is assumed to be normally distributed with mean 0 and var σ^2

Logistic Regression Model: $p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$

Estimating Model Parameters

Vertical deviation of the point $(x_i), y_i$ from the line $y = b_0 + b_1 x$ is $y_i - (b_0 + b_1 x_i)$

The sum of squared vertical deviations from points $(x_1, y_1), \cdots, (x_n, y_n)$ to the line is $f(b_0, b_1) = \sum_{i=1}^{n}[y_i - (b_0 + b_1 x_i)]^2$.

The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ satisfies $f\left(\hat{\beta}_0, \hat{\beta}_1\right) \leq f(b_0, b_1)$ for any b_0 and b_1.

The estimated regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.
$\hat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$
$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$
$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}$

The fitted/predicted values $\hat{y}_i$ are obtained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The residuals are $y_i - \hat{y}_i$.

Error sum of squares: SSE $= \sum (y_i - \hat{y}_i)^2$
$= \sum \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2 = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$.

The last formula is sesitive, **use as many digits from the calculator as possible**.

Total sum of squares:
SST $= S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2/n$
SST = SSE + SSR
$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$, the least square estimate of σ^2.
Coefficient of determination: $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$

Inferences about β_1

Mean of $\hat{\beta}_1$ is $E\left(\hat{\beta}_1\right) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is unbiased.

$\hat{\beta}_1$ has a normal distribution with
$V\left(\hat{\beta}_1\right) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}}, \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$
where $S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \left(\sum x_i\right)^2/n$, replacing σ by s gives an estimate: $s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$

Variable $T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$ has a t distribution with n-2 df, called T ratio.

A $100(1-\alpha)\%$ CI for $\hat{\beta}_1$ is $\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$

Hypothesis test: H_0: $\beta_1 = \beta_{10}$, test statistic: $t = \frac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

$\beta_1 > \beta_{10}$: $t \geq t_{\alpha,n-2}$; $\beta_1 < \beta_{10}, t \leq -t_{\alpha,n-2}$

$\beta_1 \neq \beta_{10}, t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$.

The **model utility test** is the test of H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$, test statistic: $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

Simple linear regression ANOVA:

Source	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	n-2	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	n-1	SST		

Inferences concerning $\mu_{\mathbf{Y}\cdot\mathbf{x}^*}$ and predicting future Y

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is fixed:

$E(\hat{Y}) = E\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$, so $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e. $\mu_{Y \cdot x^*}$)

$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum x_i^2 - (\sum x_i)^2/n}\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]$

and the sd $\sigma_{\hat{Y}}$ is its root, the estimated sd of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is:

$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$. And $\hat{Y}$ has a normal distribution.

The variable $T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}}$ has a t distribution with n-2 df. Conduct t-test using it as the test statistic, rejection region is $t_{\alpha,n-2}$ or $t_{\alpha/2,n-2}$

A $100(1-\alpha)\%$ CI for $\mu_{Y \cdot x^*}$ is:

$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2,n-2} \cdot s_{\hat{Y}}$

A $100(1-\alpha)\%$ PI for a future Y to be made when $x = x^*$ is:

$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$

$= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}}$

$= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}$

Sample correlation coefficient is:

$r = \frac{S_{xy}}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$

Properties of r:

- Independent of which of the two rv is labeled x and which is labeled y.
- Independent of the units.
- $-1 \leq r \leq 1$. $r = 1$ iff all (x_i, y_i) lies on a straight line with positive slope, and $r = -1$ when negative slope.
- $(r)^2 = r^2$

Population correlation coefficient ρ:

$\hat{\rho} = R = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2}\sqrt{\sum\left(Y_i - \overline{Y}\right)^2}}$

Testing for the absence of correlation:

H_0: $\rho = 0$; Test statistic: $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$

$\rho > 0$: $t \geq t_{\alpha,n-2}$; $\rho < 0$: $t \leq -t_{\alpha,n-2}$;

$\rho \neq 0$: $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$

Assessing Model Adequacy

Standardized residuals: $e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}}}, i = 1, \cdots, n$

Multiple Regression Analysis

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$, $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$

Let x_{ij} denote the value of the jth predictor x_j in the ith observation. $i \in [1, n]; j \in [1, k]$

Estimating $\hat{\beta}_i$: solving

$nb_0 + \left(\sum x_{i1}\right) b_1 + \left(\sum x_{i2}\right) b_2 + \cdots + \left(\sum x_{ik}\right) b_k = \sum y_i$

$\left(\sum x_{i1}\right) b_0 + \left(\sum x_{i1}^2\right) b_1 + \left(\sum x_{i1} x_{i2}\right) b_2 + \cdots + \left(\sum x_{i1} x_{ik}\right) b_k = \sum x_{i1} y_i$

... ...

$\left(\sum x_{ik}\right) b_0 + \left(\sum x_{i1} x_{ik}\right) b_1 + \cdots + \left(\sum x_{i,k-1} x_{ik}\right) b_{k-1} + \left(\sum x_{ik}^2\right) b_k = \sum x_{ik} y_i$

$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-(k+1)} = \text{MSE}$, $\hat{\sigma} = s = \sqrt{s^2}$

Coefficient of multiple determination $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$

Adjusted coefficient of multiple determination:

$R_a^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{\text{SSE}/[n-(k+1)]}{\text{SST}/(n-1)} = 1 - \frac{n-1}{n-(k+1)}\frac{\text{SSE}}{\text{SST}}$

A Model Utility Test

H_0: All $\beta_i = 0$, H_a: at least one $\beta_i \neq 0$

Test statistic: $f = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{\text{SSR}/k}{\text{SSE}/[n-(k+1)]} = \frac{\text{MSR}}{\text{MSE}}$

SSR = regression sum of squares = SST - SSE

Rejection region for a level α test: $f \geq F_{\alpha,k,n-(k+1)}$

Inferences in Multiple Regression

All for level $100(1-\alpha)\%$ test:

CI for β_i, the coefficient of x_i is $\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} \cdot S_{\hat{\beta}_i}$

A test for H_0: $\beta_i = \beta_{i0}$

Test statistic: $t = \left(\hat{\beta}_i - \beta_{i0}\right)/s_{\hat{\beta}_i}$, df: $n - (k+1)$

CI for $\mu_{Y \cdot x_1^*, \ldots, x_k^*}$: $\hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot S_{\hat{Y}}$, $\hat{y}$: estimate y by x^*

PI for future y: $\hat{y} \pm t_{\alpha/2,n-(k+1)} \cdot \sqrt{s^2 + s^2_{\hat{Y}}}$

Goodness-of-fit Tests

Situation:

Category	$i=1$	$i=2$	...	$i=k$	Row Total
Observed	n_1	n_2	...	n_k	n
Expected	np_{10}	np_{20}	...	np_{k0}	n

H_0: All $p_i = p_{i0}, i \in [1, k]$. H_a: at least one $p_i \neq p_{i0}$

Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{\text{all cells}} \frac{(\text{observed-expected})^2}{\text{expected}}$,

Rejection region: $\chi^2 \geq \chi^2_{\alpha,k-1}$

When parameters are estimated

k denotes the number of categories or cells and p_i denotes the probability of an observation falling in the ith cell. Each π_i is a function.

H_0: All $p_i = \pi_i(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$, H_a: H_0 is not true.

Test statistic: $\chi^2 = \sum_{i=1}^{k} \frac{\left[N_i - n\pi_i(\hat{\boldsymbol{\theta}})\right]^2}{n\pi_i(\hat{\boldsymbol{\theta}})}$,

Rejection region: $\chi^2 \geq \chi^2_{\alpha,k-1-m}$

This test can be used if $n\pi_i(\hat{\boldsymbol{\theta}}) \geq 5$ for any i

χ^2 Test for Independence

p_{ij} = the proportion of individuals in the population who belong in category i of factor 1 and category j of factor 2

then, $p_{i.} = \sum_j p_{ij}$, $p_{.j} = \sum_i p_{ij}$

The mle are $\hat{p}_{i.} = \frac{n_{i.}}{n}$, $\hat{p}_{.j} = \frac{n_{.j}}{n}$

$\hat{e}_{ij} = n \cdot \hat{p}_{i.} \cdot \hat{p}_{.j} = n \cdot \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} = \frac{n_{i.} \cdot n_{.j}}{n} = \frac{(\text{ith row total})(\text{jth column total})}{n}$

H_0: $p_{ij} = p_{i.} \cdot p_{.j}$ for every pair (i, j), H_a: null hypothesis is wrong.

Test statistic:

$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} = \sum_{\text{all cells}} \frac{(\text{observed - estimated expected})^2}{\text{estimated expected}}$

Rejection region: $\chi^2 \geq \chi^2_{\alpha,(I-1)(J-1)}$, applicable: all $\hat{e}_{ij} \geq 5$.

Example of problems

Two-Factor ANOVA with $K_{ij} > 1$

Three different varieties of tomato and 4 different plant densities are being considered for planting.

Variety	Planting Density 10000			20000			30000		
H	10.5	9.2	7.9	12.8	11.2	13.3	12.1	12.6	14.0
Ife	8.1	8.6	10.1	12.7	13.7	11.5	14.4	15.4	13.7
P	16.1	15.3	17.5	16.6	19.2	18.5	20.8	18.0	21.0
$x_{.j.}$	103.3			129.5			142.0		
$\bar{x}_{.j.}$	11.48			14.39			15.78		

Here, $I = 3$, $J = 4$ and $K = 3$, for a total of $IJK = 36$ observations.

For the given data, $x_{\ldots}^2 = 500^2 = 250000$

$\sum_i \sum_j \sum_k x_{ijk}^2 = 10.5^2 + 9.2^2 + \cdots + 18.9^2 + 17.2^2 = 7404.80$

$\sum_i x_{i..}^2 = 136.0^2 + 146.5^2 + 217.5^2 = 87,264.50$

$\sum_j x_{.j.}^2 = 63280.18$

The cell totals $(x_{ij.})$ are

	10000	20000	30000	4000
H	27.6	37.3	38.7	32.4
Ife	26.8	37.9	43.5	38.3
P	48.9	54.3	59.8	54.5

From that we get $\sum_i \sum_j x_{ij.}^2 = 27.6^2 + \cdots + 54.5^2 = 22,100.28$

Then:

$\text{SST} = 7404.80 - \frac{1}{36}(250000) = 7404.80 - 6944.44 = 460.36$

$\text{SSA} = \frac{1}{12}(87264.50) - 6944.44 = 327.60$

SSB $= \frac{1}{9}(63280.18) - 6944.44 = 86.69$
SSE $= 7404.80 - \frac{1}{3}(22100.28) = 38.04$
SSAB $= 460.36 - 327.60 - 86.69 - 38.04 = 8.03$
The resulting ANOVA Table:

Source	df	Sum of ^{2}	Mean ^{2}	f
Varieties	2	327.6	163.8	$f_A = 103.02$
Density	3	86.69	28.9	$f_B = 18.18$
Interaction	6	8.03	1.34	$f_{AB} = 0.84$
Error	24	38.04	1.59	
Total	35	460.36		

Fitting the Logistic Regression Model

The dependent variable Y is 1 if the observation is a success and 0 otherwise. The probability of success is related to x by the logit function: $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$, (It can be shown that $\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$.) Fitting the model requires β_0 and β_1 be estimated.

Suppose $n = 5$ and the observations made at x_2, x_4 and x_5 are success whereas the other two are failures. The likelihood function is thus:

$$[1-p(x_1)][p(x_2)][1-p(x_3)][p(x_4)][p(x_5)]$$
$$= \left[\frac{1}{1+e^{\beta_0+\beta_1 x_1}}\right]\left[\frac{e^{\beta_0+\beta_1 x_2}}{1+e^{\beta_0+\beta_1 x_2}}\right]\left[\frac{1}{1+e^{\beta_0+\beta_1 x_3}}\right]\left[\frac{e^{\beta_0+\beta_1 x_4}}{1+e^{\beta_0+\beta_1 x_4}}\right]\left[\frac{e^{\beta_0+\beta_1 x_5}}{1+e^{\beta_0+\beta_1 x_5}}\right]$$

No straightforward formula can be derived. Use iterative numerical methods to maximize it.

Explanation of MiniTab Output

In the following table, "##" marks irrelevant items. All the items' relative position in the table are identical to that on the example of textbook.

The "p" is the p-value for model utility test.

```
The regression equation is
(The Equation)

Predictor        Coef        SE Coef   T            P
Constant         β̂0          ##        ##           ##
(Variable)       β̂1          s_β̂1      t = β̂1/s_β̂1  p
S = ##           R-Sq = r^2  R-Sq(adj)  = ##

Analysis of Variance
Source           DF          SS
Regression       ##          ##
Residual Error   ##          SSE
Total            ##          SST
```