conditional probability

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{P(B/A_i)\,P(A_i)}{\sum_{i=1}^{k} P(B/A_i) \cdot P(A_i)} \rightarrow \text{Baye's theorem}$$

Chebyshev's theorem :-

$$P(\,|X-\mu| \geq k\sigma) \leq 1/k^2$$

Discrete Random Variables

Bernoulli r.v = either 0 or 1

$E(X) = \mu_x = \Sigma x \cdot p(x)$ $\qquad \mu_{aX+b} = a\mu_x + b$

$E(h(X)) = \Sigma h(x) \cdot p(x)$ $\qquad V(aX+b) = a^2 V(X)$

Binomial distribution : each trail is bernoulli trail

$b(x;n,p) = {}^nC_x \, p^x (1-p)^{n-x} \rightarrow P(X=x)$

$B(x;n,p) = \sum_{y=0}^{x} b(y;n,p) \rightarrow P(X \leq x)$

$\mu_x = np \qquad \sigma_x^2 = npq$

$m_x = (q+pe^t)^n \quad , q = 1-p$

Hypergeometric distribution : a defective in N, x defective out of n. choosing

$$h(x;n,a,N) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

finite population correction value

$\mu_x = \frac{n \cdot a}{N} \qquad \sigma_x^2 = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{a}{N}\left(1-\frac{a}{N}\right)$

$= np \qquad \sigma_x^2 = \left(\frac{N-n}{N-1}\right) np(1-p)$

Negative Binomial distribution: no. of failures before r successes.

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

$$\mu_x = \frac{r(1-p)}{p} \qquad \sigma_x^2 = \frac{r(1-p)}{p^2}$$

special case when $r=1$ (trails stop after single success)

Geometric distribution

* memory loss property $P(x > s+t / x > s) = P(x > t)$

$$m_x = \frac{pe^t}{1-qe^t}, \quad q = 1-p$$

Poisson distribution:

$$f(x) = \frac{e^{-\lambda} \lambda^x}{x!} \qquad (\lambda > 0, x \geq 0)$$

$$m_x(t) = e^{\lambda(e^{t}-1)}$$

$$\mu_x = \lambda = \sigma_x^2$$

$\lambda \to$ mean per unit (t)

* poisson process, $f(x; \lambda, t) = \dfrac{e^{-\lambda t} (\lambda t)^x}{x!}$

Continuous distribution:

$$f(x) = 1/n$$

$$m_x(t) = \frac{1}{n} \sum_{i=1}^{n} e^{tx_i} \qquad \mu_x = \frac{\Sigma x}{n} \qquad \sigma_x^2 = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

Continous Random variables

PDF

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

CDF

$$P(-\infty \le X \le x) = \int_{-\infty}^{x} f(x)\,dx \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1$$

k^{th} moment about mean:

$$E\left((X-\mu)^k\right) = \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,dx$$

↳ k^{th} moment about mean

Normal k^{th} moment → $\mu = 0$

percentile:

* $\eta \rightarrow (100p)^{th}$ percentile , $\tilde{\mu} = 50^{th}$ percentile (median)

$$p = \int_{-\infty}^{\eta(p)} f(y)\,dy = F(\eta(p))$$

$$P(x \le X \le x + \Delta x) \simeq f(x)\Delta x \quad ; \quad f(x) = \frac{dF(x)}{dx}$$

Uniform Distribution: → Rectangular distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{other} \end{cases} \qquad F(x) = \begin{cases} 0 & x \le a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x \ge b \end{cases}$$

$$m_x(t) = \frac{e^{bt} - e^{at}}{(b-a)t} \qquad \mu_x = \frac{b+a}{2} \qquad \sigma_x^2 = \frac{(b-a)^2}{12}$$

Gamma distribution:

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

$$\Gamma(\alpha) = (\alpha-1)\,\Gamma(\alpha-1)$$

$$\Gamma(n) = (n-1)! \qquad (n \text{ is +ve integer})$$

$$\Gamma(1/2) = \sqrt{\pi}$$

$$f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \ge 0 \\ 0 & \text{other} \end{cases}$$

$$\mu_x = \alpha\beta \qquad \sigma_x^2 = \alpha\beta^2$$

$$m_x = (1-\beta t)^{-\alpha} \qquad (t < 1/\beta)$$

* incomplete gamma function $\longrightarrow$ $y = \frac{x}{\beta}$

$$F(y;\alpha) = \int_0^x \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy$$

Exponential distribution: Gamma with $\alpha = 1$

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{other} \end{cases}$$

* distribution of elapsed $dt \leftarrow$ time between successive events in poisson process $d \leftarrow$ is exponential.

$$\mu = 1/\lambda \qquad \sigma^2 = \frac{1}{\lambda^2}$$

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

Chi squared distribution: $\alpha = \frac{\nu}{2}$ $\beta = 2$

$$\chi^2(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} \cdot e^{-x/2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$\nu \rightarrow$ degrees of freedom (df)

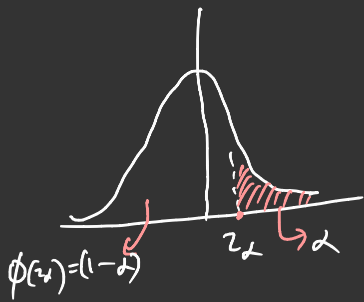
Normal Distribution:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

$$\sim N(\mu, \sigma^2)$$

$$m_x = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

standard normal distribution



$$z = \frac{x-\mu}{\sigma} \rightarrow f(z;0,1)$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

$$P(Z \le z) = \phi(z) \rightarrow \text{table}$$

$$z_\alpha \Rightarrow \phi(z_\alpha) = 1-\alpha \quad (\text{critical value})$$

$$z_\alpha = -\phi^{-1}(\alpha) \longrightarrow \text{look for } -\alpha \text{ in table}$$

Binomial approximation

np and $nq > 10$

$$B(x;n,p) = F(x+0.5;\, np, npq) = \phi\left(\frac{x+.5-np}{\sqrt{npq}}\right)$$

Correction factor

* lognormal $\Rightarrow$ $\ln(x)$ follows normal distribution

Joint Distributions

pmf

$$P(x,y) = P(X=x \text{ and } Y=y) \leftrightarrow f_{XY}(X=x \text{ and } Y=y)$$

$$\sum_x \sum_y P(x,y) = 1 \iff \iint_{\mathbb{R}} f(x,y)\,dx\,dy = 1$$

cdf

$$F(x,y) = \sum_{X \le x} \sum_{Y \le y} P(x,y) \leftrightarrow \int_a^b \int_c^d f(x,y)\,dx\,dy$$

Marginal density function

$$f_X(x) = \sum_y f(x,y) \iff \int_{-\infty}^{\infty} f_{XY}\,dy$$

$\downarrow$ for independent random variables

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y) \quad \forall (x,y)$$

conditional probability

$$f_{Y/X}(y/x) = \frac{f(x,y)}{f_X(x)}$$

* Expectation

can be done from marginal

$$E(X) = \sum_x \sum_y x f_{xy} = \mu_x \leftrightarrow \iint_{\mathbb{R}} x \cdot f_{xy}\, dy\, dx$$

$$E(H(X,Y)) = \sum_x \sum_y H(x,y) f(x,y) \leftrightarrow \iint_{\mathbb{R}} H(x,y) f_{xy}\, dy\, dx$$

* Covariance

$$E((X-\mu_x)(Y-\mu_y)) = E(XY) - E(X)E(Y)$$

$$= E(XY) - \mu_x \mu_y$$

* Coefficient of correlation

$$\rho_{xy} = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \qquad [-1 \le \rho \le 1]$$

$|\rho| = 1 \Rightarrow Y = aX + b$

If X, Y are independant $\Rightarrow \rho = 0$

$\rho \to$ degree of linear relationship

$\rho = 0 \Rightarrow X, Y$ are uncorrelated (need not be independant)

Statistics

Random samples (X)

rv's $x_1, \dots, x_n$ form a random sample if each X is independant and follows same probability distribution.

$\bar{x} = \frac{\sum x_i}{n}$ → sample mean

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \rightarrow \text{sample variance}$$

$$= \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

Distributions

* Sample Mean ; $\mu_{\bar{x}} = \mu$ $\quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

* Total ; $\mu_{T_0} = n\mu$ $\quad \sigma_{T_0} = \sqrt{n}\,\sigma$

independent

* Difference ; $\mu_Y = \mu_1 - \mu_2$ $\quad \sigma_Y^2 = \sigma_1^2 + \sigma_2^2$

* $Y = a_1 X_1 + a_2 X_2 + \dots + a_n X_n$

$$\mu_Y = a_1 \mu_1 + a_2 \mu_2 + \dots + a_n \mu_n$$

$$\sigma_Y^2 = \sum_i \sum_j a_i a_j \operatorname{Cov}(X_i, X_j)$$

↓ for independant random variables

$$\sigma_Y^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \dots + a_n^2 \sigma_n^2$$

* Central limit theorem : for $n > 30$ ~ Y follows normal distribution, even if X not normal

Point estimation

$\hat{\theta} \rightarrow$ point estimate of θ (a sensible value)

unbiased estimator :

$\hat{\theta}$ is an unbiased estimator of θ if $E(\hat{\theta}) = \theta$

$E(\hat{\theta}) - \theta$ is called the bias of θ.

some unbiased estimators

* $\hat{p} = \frac{X}{n}$

* $\hat{\sigma}^2 = S^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1}$

* $\hat{\mu} = \bar{X}$; for continous and symmetric $\tilde{X}$ and trimmed mean are also unbiased estimators

MVUE $\rightarrow$ Minimum variance unbiased estimator

Method of moments:

Find $E(X), \dots E(X^k)$ from distribution and equate with sample data.

$E(X) = \bar{X}$, $E(X^2) = S^2 + \bar{X}^2$

k

equate first k moments to find $\theta_1, \dots, \theta_k$

Method of Maximum Likelyhood:

likelyhood function → $f(x_1, x_2, \ldots, x_n ; \theta_1, \ldots \theta_m)$

maximum likelyhood estimates (mle's) are the $\hat{\theta}_1, \ldots, \hat{\theta}_m$ which maximize the likelyhood function

most probable estimate → MODE

apply log to f, differentiate wrt $\hat{\theta}_i$ and equate to 0.

When n is large, mle's are the MVUE's.

Confidence Intervals

$100(1-\alpha)\%$ confidence interval for

Mean

→ two tailed : $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

→ upper tailed : $\left(\bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}},\ \infty\right)$ → lower bound

→ lower tailed : $\left(-\infty,\ \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}\right)$ → upper bound

$$n = \left(2z_{\alpha/2}\cdot\frac{\sigma}{w}\right)^2$$

Population proportion

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \simeq 1-\alpha$$

$$\Rightarrow \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}$$ → Score CI

for one tailed, $+z_\alpha$ (or) $-z_\alpha$

CI for no σ → t - distribution

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ → n-1 degrees of freedom

same as z formula, replace z_α by $t_{\alpha, n-1}$

Prediction interval

$$\bar{x} \pm t_{\alpha/2, n-1} \times s\sqrt{1 + \frac{1}{n}}$$

Tolerance interval

$\bar{x} \pm$ (tolerance critical value) $\cdot s$

CI for σ^2 → χ^2 distribution

$$\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}} , \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}} \right)$$ → for sigma take squareroot of upper and lower bound

Hypothesis

lower tailed — upper tailed — two-tailed

$H_0 : \hat{\theta} = \theta_0$; $H_a : \hat{\theta} < \theta_0$ — $H_a : \hat{\theta} > \theta_0$ — $H_a : \hat{\theta} \neq \theta_0$

H_0 → Null Hypothesis

H_a → Alternate Hypothesis → $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$

for some θ in H_a, we reject H_0 ⇒ called rejection region

otherwise, we fail to reject H_0 (always assume H_0 correct and try to reject it)

Type 1 error (α) → rejecting H_0 when its true

Type 2 error (β) → not rejecting H_0 when its false

→ Power of test $= 1 - \beta$
(more power → better test)

General test statistic

$$z / t = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

	Rejection region for α-test	$\beta(\theta')$
$H_a : \hat{\theta} > \theta_0$	$Z \geq Z_\alpha$	$P\left(z < \dfrac{\theta_u - \theta'}{\sigma_{\theta'}}\right)$
$H_a : \hat{\theta} < \theta_0$	$Z \leq -Z_\alpha$	$P\left(z > \dfrac{\theta_l - \theta'}{\sigma_{\theta'}}\right)$
$H_a : \hat{\theta} \neq \theta_0$	$Z \geq Z_{\alpha/2}$ or $Z \leq -Z_{\alpha/2}$	→ $P\left(z < \dfrac{\theta_u - \theta'}{\sigma_{\hat{\theta}}}\right) + P\left(z > \dfrac{\theta_l - \theta'}{\sigma_{\hat{\theta}}}\right)$

$$n = \begin{cases} \left[\dfrac{\sigma_{\hat{\theta}}(Z_\alpha + Z_\beta)}{\theta_0 - \theta'}\right]^2 & \text{→ one tailed} \\[2ex] \left[\dfrac{\sigma_{\hat{\theta}}(Z_{\alpha/2} + Z_\beta)}{\theta_0 - \theta'}\right]^2 & \text{→ two tailed} \end{cases}$$

θ is $\hat{\theta}$ value for $\dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}} = Z_\alpha$

$$\left.\begin{aligned} \theta_u &= \theta_0 + \sigma_{\hat{\theta}} Z_\alpha \\ \theta_l &= \theta_0 - \sigma_{\hat{\theta}} Z_\alpha \end{aligned}\right\} \text{→ one tailed}$$

(Z_α replaced by $Z_{\alpha/2}$ → two tailed) for large sample → Z - critical value → $\sigma = s$

for small sample → t - critical value → $\sigma = s$

for Mean,

$\hat{\theta} = \bar{X}$ $\quad \theta_0 = \mu_0$ $\quad \sigma_{\hat{\theta}} = \dfrac{\sigma}{\sqrt{n}}$

for population proportion,

no unknowns

$\hat{\theta} = \hat{p} = \dfrac{X}{n}$ $\quad \theta_0 = p_0$ $\quad \sigma_{\hat{\theta}} = \dfrac{\sqrt{npq}}{n} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$

$\sigma_{\theta'} = \sqrt{\dfrac{p'(1-p')}{n}}$

p value for General test statistic

$z/t = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$

$\left.\begin{array}{l}\text{reject } H_0 \text{ if } P\text{-value} \le \alpha \\ \text{do not reject if } P\text{-value} > \alpha\end{array}\right\}$ only condition

$$\left.\begin{array}{lll} H_a: \hat{\theta} > \theta_0 & & 1 - \phi(z) \\ H_a: \hat{\theta} < \theta_0 & & \phi(z) \\ H_a: \hat{\theta} \ne \theta_0 & & 2[1 - \phi(|z|)] \end{array}\right\} \text{p-value}$$

if (or z) t, find α value for t value by reverse, and use in place of $[1 - \phi(z)] \rightarrow \alpha$

Regression

$$E(\varepsilon) = 0$$

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad V(\varepsilon) = \sigma^2$$

↳ Simple linear regression model

least square fit method

$$f(b_0, b_1) = \sum \left[y_i - (b_0 - b_1 x_i) \right]^2$$

$\hat{\beta}_0, \hat{\beta}_1 \rightarrow$ values which maximize f (most no. of points almost passing through)

↳ estimated regression line / least square line

$$b_1 = \text{slope}(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{\sum x_i^2}{n}}$$

$$b_0 = \text{intercept}(\hat{\beta}_0) = \bar{Y} - \hat{\beta}_1 \bar{X}$$

↳ $= \dfrac{S_{xy}}{S_{xx}}$

Normal equations :-

$$n b_0 + (\Sigma x) b_1 = \Sigma y$$

$$(\Sigma x) b_0 + (\Sigma x^2) b_1 = \Sigma xy$$

Best fit values :

find $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ for each $x_1, \dots, x_n$ by substituting in least square line. $(y_i - \hat{y}_i)$ is residual for x_i (i^{th} sample)

↳ more value of residual $\Rightarrow$ more σ^2 (deviation) error of sum of squares:

$$SSE = \sum \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1) x\right)^2$$

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum [y - \hat{y}_i]^2}{n-2} \rightarrow \text{(residuals)}$$

$$SSE = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

$$SST = S_{yy} = \sum y_i^2 - (\sum y_i)^2/n$$

Coefficient of Determination:

$$r^2 = 1 - \frac{SSE}{SST}$$

Menu, statistics

$y = ax+b$

regression calc

$\sum x$, $\sum x^2 \rightarrow$ 2 variable calc

* use as many digits as possible in intermediate calculations.

Simulation

General method:

find cdf of any distribution / density, then assign 2/3 digit numbers limits corresponding to each different outcome, then check in which region random number lies in and equate to cdf to the random number and find corresponding sample value.

Transformation of random variable:

$$g(y) = f(x) \left| \frac{dx}{dy} \right|$$

→ Discrete

Eg: $p = .05$ → geometric RV

unacceptable → 0-4

$1 \to 0.05$ → 5-9

$2 \to 0.0975$ → 10-14

$3 \to 0.143$ → 15-18

$4 \to 0.185$ → 19-22

$5 \to 0.226$ → 23-26

$6 \to 0.264$ → 27-30

$7 \to 0.302$ → 30-33

$8 \to 0.336$

$9 \to 0.369$

⋮

and thus take the distribution

→ Continuous

Eg: $$f(x) = \begin{cases} \frac{1}{8}|x-1| & -1 \le x < 3 \\ 1/4 & 3 \le x < 5 \\ 0 & \text{other} \end{cases} \rightarrow \begin{cases} \frac{1-x}{8} & -1 \le x \le 1 \\ \frac{x-1}{8} & 1 < x < 3 \\ 1/4 & 3 \le x < 5 \\ 0 & \text{other} \end{cases}$$

(a) $$F(x) = \begin{cases} \frac{x}{8} - \frac{x^2}{16} + \frac{3}{16} & -1 \le x \le 1 \longleftrightarrow 0 - .25 \\ \frac{5}{16} + \frac{x^2}{16} - \frac{x}{8} & 1 < x < 3 \longleftrightarrow .25 - .5 \\ \frac{x}{4} - \frac{1}{4} & 3 \le x \le 5 \longleftrightarrow .5 - .99 \\ 1 & \text{other} \end{cases}$$

(b) 65, 09, 26, 80, 21, 34

$.65 = \frac{x}{4} - 0.25$

$x_1 = 3.6$

$.09 = -\frac{x^2}{16} + \frac{x}{8} + \frac{3}{16}$

$x_2 = -0.6$

same way

$x_3 = 1.4$

$x_4 = 4.2$

$x_5 = 0.2$

$x_6 = 2.2$

$u = \frac{1}{4}(x-1)$

$4u + 1 = x$

$u = 0.65$

x