

1 Distributions

1.1 Binomial

Sum of n bernoulli trials with probability of success p .

$$X \sim B(n, p)$$

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad k = 0, 1, \dots, n$$

$$\text{mgf: } M_X(t) = (pe^t + 1 - p)^n$$

$$E X = np, \quad \text{Var } X = np(1-p)$$

1.2 Poisson

$$X \sim P(\lambda)$$

$$p(k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k = 0, 1, \dots$$

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

$$E X = \lambda, \quad \text{Var } X = \lambda$$

1.3 Multivariate Normal

$$X \sim N(\mu, \Sigma), \Sigma \text{ positive definite}$$

$$f(x) = \frac{\exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}}{(2\pi)^{\frac{k}{2}} \sqrt{\det(\Sigma)}}$$

1.4 Beta

$$X \sim \text{Beta}(\alpha, \beta)$$

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad 0 \leq x \leq 1$$

using the beta function:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

$$E X = \frac{\alpha}{\alpha+\beta}, \quad \text{Var } X = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

1.5 Gamma

$$X \sim \text{Gamma}(\alpha, \lambda)$$

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} \quad x > 0$$

$$\text{mgf: } M_X(t) = (1 - \frac{t}{\lambda})^{-\alpha}, t < \lambda$$

$$X \sim \text{Gamma}(\alpha, \lambda) \iff \lambda X \sim \text{Gamma}(\alpha, 1)$$

$$X_i \text{ iid } \text{Gamma}(\alpha_i, \lambda), \text{ then}$$

$$\sum X_i \sim \text{Gamma}(\sum \alpha_i, \lambda)$$

$$\text{Gamma function: } \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt.$$

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

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

$$\Gamma(k) = (k-1)! \text{ for } k \text{ positive integer.}$$

1.6 Exponential

$$\text{Special case: } \text{Exp}(\lambda) \equiv \text{Gamma}(1, \lambda)$$

1.7 Chi square

$$\text{Special case: } \chi_n^2 \equiv \text{Gamma}(\frac{n}{2}, \frac{1}{2})$$

$$\text{Let } Z_i \text{ be iid } N(0, 1).$$

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

$$\text{Noncentral } \chi^2. \text{ Let } Y \sim N(\mu, I) \text{ be an } n \text{ vector.}$$

Then

$$\|Y\|^2 \sim \chi_n^2(\|\mu\|^2)$$

1.8 F

$$F(m, n) \equiv \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}}$$

Where numerator and denominator are independent χ^2 .

1.9 T

$$t(n) = \frac{N(0, 1)}{\sqrt{\frac{\chi_n^2}{n}}}$$

Where numerator and denominator are independent.

1.10 Studentized range distribution

Let z_1, \dots, z_K be i.i.d $N(0, 1)$ and X be an independent χ^2_ν . Then

$$\frac{\max_i z_i - \min_i z_i}{\sqrt{(X/\nu)}} \sim q_{K,\nu}$$

2 Math

Moment generating functions determine distribution

$$M_X(t) \equiv E(e^{tX}), \quad M'_X(0) = E(X)$$

For X_i independently distributed:

$$M_{\sum X_i}(t) = \prod M_{X_i}(t)$$

Use this to find distributions of sums of RV's.

2.1 Inequalities

Jensen's Inequality if $S \subset R^k$ convex and closed, g convex on S , $P[X \in S] = 1$, and $E X$ is finite, then $E X \in S$, $E g(X)$ exists, and

$$E g(X) \geq g(E X)$$

Holder's Inequality if $r, s > 1$ and $\frac{1}{r} + \frac{1}{s} = 1$ then

$$E |XY| \leq (E |X|^r)^{\frac{1}{r}} (E |X|^s)^{\frac{1}{s}}$$

Order statistics for sorted sample $X_{(1)}, \dots, X_{(n)}$ has pdf:

$$n! \prod_{i=1}^n f(X_{(i)}) \quad I(X_{(1)} < \dots < X_{(n)})$$

$T(X)$ Sufficient means the distribution of $X|T(X)$ does not depend on θ .

Factorization theorem: $T(x)$ is sufficient \iff

$$f_\theta(x) = h(x)g(\theta, T(x))$$

$L_x(\theta) = p_\theta(x) = p(x, \theta)$ likelihood is function of θ , density is function of x .

The likelihood ratio

$$\lambda_x(\theta) = \frac{L_x(\theta)}{L_x(\theta_0)}$$

is minimal sufficient. To show $T(x)$ is minimal sufficient show that it is sufficient and a function of the likelihood $\lambda_x(\theta)$.

Fisher information

$$I(\theta) = E_\theta \left[\frac{\partial}{\partial \theta} \log L_X(\theta) \right]^2 = E_\theta \left[-\frac{\partial^2}{\partial \theta^2} \log L_X(\theta) \right]$$

2.2 Bias, Variance

bias $\hat{v} \equiv E(\hat{v}) - v$

$$MSE(\hat{v}) \equiv E(\hat{v} - v)^2 = \text{Var}(\hat{v}) + (\text{bias } \hat{v})^2$$

Rao-Blackwell Let $S(X)$ be an unbiased point estimator for $g(\theta)$. Conditioning on a sufficient statistic $T(X)$ reduces variance.

$$\text{Var}_\theta(S(X)) \geq \text{Var}_\theta(E(S(X)|T(X)))$$

Also holds for more general convex loss function L :

$$R(\theta, S) \equiv E_\theta L(\theta, S(X)) \geq E_\theta L(\theta, E(S(X)|T(X)))$$

Completeness $T(X)$ is complete if $E g(T(X)) = 0$ implies $g = 0$ almost surely for all θ .

Cramer Rao Inequality Let $g : \Theta \rightarrow R$. Suppose there exists an unbiased estimator $U(X)$, $E U(X) = g(\theta)$. Then

$$\text{Var}_\theta U(X) \geq \left(\frac{\partial g(\theta)}{\partial \theta} \right)^T I(\theta)^{-1} \left(\frac{\partial g(\theta)}{\partial \theta} \right)$$

Basu's Theorem - If $T(X)$ complete sufficient statistic and $A(X)$ is ancillary then $A(X)$ and $T(X)$ are independent.

Lehmann - Scheffe Suppose $T(X)$ is complete sufficient. Then there exists unique unbiased estimator $E h(T(x))$ of $g(\theta) \in R$ with smallest variance (MVUE)

2.3 Exponential Families

One parameter model indexed by θ .

$$p(x, \theta) = h(x) \exp\{\eta(\theta)T(x) - B(\theta)\}$$

Canonical form model indexed by η .

$$q(x, \eta) = h(x) \exp\{\eta T(x) - A(\eta)\}$$

Then moment generating function for $T(X)$ is

$$M_{T(X)}(t) = \exp\{A(t + \eta) - A(\eta)\}$$

2.4 Multivariate Normal

Stein's formula: $X \sim N(\mu, \sigma)$

$$E(g(X)(X - \mu)) = \sigma^2 E(g'(X))$$

assuming these expectations are finite.

$X \sim N(\mu, \Sigma)$, A an $m \times n$ matrix, then

$$AX \sim N(A\mu, A\Sigma A^t)$$

For Σ full rank it's possible to transform between $z \sim N(0, I)$ and X :

$$X = \Sigma^{1/2}z + \mu \quad z = \Sigma^{-1/2}(X - \mu)$$

Conditional distribution:

2.5 Conditional Distributions

Conditional pdf:

$$f_{x|y}(x, y) \equiv \frac{f_{x,y}(x, y)}{f_y(y)}$$

Iterated expectation:

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

Conditional variance formula:

$$\text{Var}(Y) = \text{Var}(E(Y|X)) + E(\text{Var}(Y|X))$$

3 Applied

Matrix / Vector differentiation

$\frac{\partial A^T \beta}{\partial \beta} = A$, $\frac{\partial \beta^T A \beta}{\partial \beta} = (A + A^T)\beta = 2A\beta$ for A symmetric.

3.1 Linear Models

Least Squares Principle

$$\arg \min_{\beta} \|Y - X\beta\|^2$$

Normal Model

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

Normal Equations - Any b satisfying this solves the least squares

$$X^T X b = X^T y$$

Gauss Markov Theorem - $\hat{\beta}$ is Best Linear Unbiased Estimator (BLUE) of β .

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim N(\beta, \sigma(X^T X)^{-1})$$

Estimating the variance: $\frac{\|y - X\hat{\beta}\|^2}{\sigma^2} \sim \chi_{n-p}^2$.

$$\hat{\sigma}^2 = \frac{\|y - X\hat{\beta}\|^2}{n - p}$$

Use t test for hypothesis testing and confidence intervals for the value of a particular β_j coefficient. Let w_{ii} be the i th diagonal entry of $(X^T X)^{-1}$.

$$\frac{\beta_j - \beta_j^*}{\hat{\sigma} \sqrt{w_{ii}}} \sim t_{n-p}$$

$1 - \alpha$ Confidence intervals for new observation Y_h at x_h and $E[Y_h]$:

$$E[y_h] \approx \hat{y}_h \pm t(n - p, 1 - \frac{\alpha}{2}) \hat{\sigma} \sqrt{x_h^T (X^T X)^{-1} x_h}$$

$$y_h \approx \hat{y}_h \pm t(n - p, 1 - \frac{\alpha}{2}) \hat{\sigma} \sqrt{1 + x_h^T (X^T X)^{-1} x_h}$$

General linear tests. Partition $\beta = (\beta_1, \beta_2)$ where β_1 is an r vector and β_2 is $p - r$. Null hypothesis $H_0 : \beta_2 = \beta_2^*$ (often 0), and $H_a : \beta_2 \neq \beta_2^*$. Then $SSE_r = \|y - X_2 \beta_2^* - X_1 \tilde{\beta}_1\|^2$ is the sum of squared error for the reduced model and $SSE_f = \|y - X\hat{\beta}\|^2$

is the squared sum of error for the full model. Under H_0 :

$$\frac{\frac{SSE_r - SSE_f}{p - r}}{\frac{SSE_f}{n - p}} \sim F_{p-r, n-p}$$

Alternate forms of linear test, and testing a linear combination if $R\beta = r$.

$$\frac{(R\hat{\beta} - r)^T (R(X^T X)^{-1} R^T)^{-1} (R\hat{\beta} - r) / s^2}{\hat{\sigma}} \sim F_{s, n-p}$$

3.2 Model selection and diagnostics

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \|y - \hat{y}\|^2 = \|(I - H)y\|^2$$

If the model contains the intercept in the column space of X then $SSTO = SSR + SSE$.

$$R^2 = 1 - \frac{SSE}{SSTO}$$

$$\text{Adjusted } R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

$$AIC = n \log SSE + 2p$$

$$BIC = n \log SSE + p \log n$$

$$Cp = \frac{SSE}{MSE} - (n - 2p)$$

$$\text{Residuals: } \hat{\epsilon}_i = y_i - \hat{y}_i$$

Studentized residuals (`rstandard` in R):

$$\gamma_i = \frac{\hat{\epsilon}_i}{s\{\hat{\epsilon}_i\}} = \frac{\hat{\epsilon}_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

Prediction sum of squares (PRESS) is the same as leave one out cross validation (LOOCV). Prediction error on i th observation is called deleted residuals:

$$y_i - \hat{y}_{i(-i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}}$$

Works for ridge regression also, letting $H = X(X^T X + \lambda I)^{-1} X^T$.

Studentized deleted residuals:

$$t_i = \frac{\hat{\epsilon}_i}{\sqrt{MSE_{(-i)}(1 - h_{ii})}} \sim t_{n-p-1}$$

TODO- Don't think this is correct:

Where $MSE_{(-i)} = SSE_{(-i)} / (n - 1 - p)$ and $SSE_{(-i)} = SSE - \frac{\hat{\epsilon}_i^2}{1 - h_{ii}}$ can be used to calculate without refitting model.

3.3 ANOVA

Three principles of experimental design: 1) Replication 2) Randomization 3) Blocking

One way ANOVA with n total observations, K groups:

| SS | | DF |
|------|--|---------|
| SSTR | $\sum_{j=1}^K n_j (\bar{y}_{j\cdot} - \bar{y}_{\cdot\cdot})^2$ | $K - 1$ |
| SSE | $\sum_{i=1}^n (y_{ij} - \bar{y}_{j\cdot})^2$ | $n - K$ |
| SSTO | $\sum_{i=1}^n (y_{ij} - \bar{y}_{\cdot\cdot})^2$ | $n - 1$ |

Contrasts are sums of the form $\Phi = \sum_{i=1}^K c_i \mu_i$ with $\sum_{i=1}^K c_i = 0$. Tukey's works for all pairwise contrasts. Scheffe's and extended Tukey works for all contrasts. Bonferroni's is for a limited number of pre specified contrasts.

Ridge Regression for $\lambda > 0$ solves

$$\min_{\beta} \|Y - X\beta\|^2 + \lambda \|\beta\|^2$$

3.4 General Techniques

Singular Value Decomposition (SVD) Any matrix X can be written

$$X = UDV^T$$

with U, V orthogonal, and D diagonal.

Moore Penrose Pseudoinverse A^+ exists uniquely for every matrix A .

Projection matrix P are symmetric and idempotent. They have eigenvalues either 0 or 1.

$$P = P^T \quad P^2 = P$$

Covariance of linear transformations

$$\text{Cov}(Ay, Bx) = A \text{Cov}(y, x) B^T$$

4 Miscellaneous

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Integration by parts:

$$\int uv' = uv - \int u'v$$

5 Course Specific

5.1 Applied Lectures

30 Nov - Outliers in x and y , leverage, DFFITS, cook's distance, influence plot, add variable plot, robust regression

25 Nov - Model selection criteria, deleted residuals, forward and backward selection in lab

23 Nov - BIC, AIC derivations, Mallows' cp , stepwise selection algorithms, outliers and studentized residuals

18 Nov - F test with orthogonalized X , model selection criteria, bootstrap t method in lab

16 Nov - Multicollinearity, Variance Inflation Factor, ridge regression, bias variance tradeoff, AIC, BIC, cross validation, proof leave one out cross validation formula for OLS

11 Nov - Holiday

9 Nov - Linear model with random X , transformations of y and X , box cox procedure, bootstrap with percentile- t and fixed X sampling, weighted least squares

4 Nov - Interaction plots for two way ANOVA with balanced design, Linear models with random X

2 Nov - Midterm

28 Oct - Kronecker product formulae for two way ANOVA

26 Oct - Kronecker product 1 way ANOVA, decomposition of two way ANOVA, noncentral χ^2 distributions for ANOVA table SSA, SSB, SSAB

21 Oct - Tukey's method for pairwise contrasts, Bonferroni's method, definition and properties of Kronecker product

5.2 Math Lectures

30 Nov - Bayesian inference, risk, decision rules, conjugate families, Binomial, normal examples

25 Nov - Exponential families, GLM's, full rank exp families, decision theory, Bayes risk

23 Nov - Fisher information, Cramer Rao inequality, Exponential families and properties

18 Nov - Rao-Blackwell theorem, Lehmann-Scheffe theorem, UMVUE examples for normal, uniform, Poisson, Fisher information

16 Nov - Minimal sufficiency, likelihood ratio, ancillary statistics, completeness, Basu's theorem, loss functions

11 Nov - Holiday

9 Nov - Distribution of order statistics, factorization theorem, sufficient statistics for Exponential families and uniform dist

4 Nov - Midterm

2 Nov - Location-scale families, invariance, ancillary and sufficient statistics, order statistics, multinomial distribution

28 Oct - Transformation of discrete and continuous random variables, Jacobian, examples with beta distributions, Dirichlet distribution

26 Oct - Jensen's and Holder's inequality, convex functions and sets, products of normal random variables

21 Oct - Convolution formula, examples with Uniform, Gamma, Poisson, marginal and conditional distributions for multivariate normal

Table 1: Problems in past 231 exams - Came from a brief glance at the question statements. TODO-make second updated table after solving questions that shows which techniques are used.

| | |
|-------------------------|-------|
| Binomial | ***** |
| Poisson | ***** |
| Uniform | ***** |
| Normal | ***** |
| Gamma | *** |
| Exponential | ** |
| Negative binomial | * |
| Beta | * |
| Geometric | * |
| MLE | ***** |
| asymptotic distribution | ***** |
| Bayes estimator / risk | ***** |
| UMVUE / Cramer-Rao | ***** |
| minimax | ***** |
| UMP test | ***** |
| linear regression | ***** |
| likelihood ratio | **** |
| Wald's test | *** |
| sufficient statistic | ** |
| Hierarchical model | * |
| method of moments | * |
| hypothesis testing | * |
| order statistics | * |

5.3 Problem Solving Strategies

Read the whole question

Read carefully and do the right problem! If there's a hint, it should probably be used. Early parts of a question can help for later parts, and later parts occasionally provide insight for earlier parts.

First principles

When in doubt, work from definitions

Look for Distributions

Can the question be solved by knowing the distribution of some quantity? Example: $\sum (x_i - \bar{x})^2$ is χ^2_{n-1} for normal x_i .

Fast and correct algebra

Better to write more than to make a simple algebra mistake. Practice common manipulations so don't have to think about them when testing.

5.4 Math questions

1. State and prove the Lehmann Scheffe Theorem.

5.5 Applied questions

1. Derive ridge regression estimates, along with their bias and variances.