Probability & Distributions

Multivariate Distributions & Moments

if $X_1, \dots, X_n$ are **i.i.d.** with *CDF* $F(x)$ then:

$$F_{X_{\min}}(x) = 1 - [1 - F(x)]^n$$
$$F_{X_{\max}}(x) = [F(x)]^n$$

For random vector $\boldsymbol{X} \in \mathbb{R}^p$, covariance matrix:

$$\Sigma = \mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \in \mathbb{R}^{p\times p}$$

Properties:

- if $\boldsymbol{a}$ is constant vector then: $\mathrm{Var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\,\boldsymbol{a}$
- if $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}$ then $\mathrm{Var}(\boldsymbol{X}+\boldsymbol{Y}) = \mathrm{Var}(\boldsymbol{X}) + \mathrm{Var}(\boldsymbol{Y})$

Correlation coefficient:

$$\mathrm{Corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}}$$

Multivariate Normal (MVN):

$\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ means $\boldsymbol{X}$ has density

$$f_X(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}.$$

Properties:

- Any linear combination of components of $\boldsymbol{X}$ is normal.
- If $A \in \mathbb{R}^{k\times p}$, then $\boldsymbol{Y} = A\boldsymbol{X} \sim N_k(A\boldsymbol{\mu}, A\Sigma A^T)$.
- Marginals of a MVN are normal: any subset X_I is $N_{|I|}(\mu_I, \Sigma_{II})$.
- If $\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$, then the conditional distribution $X_1 \mid X_2 = x_2$ is:

$$N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right),$$
$$E[X_1|X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2).$$

- The **MGF** of $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ is:

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^T\boldsymbol{\mu} + \tfrac{1}{2}\boldsymbol{t}^T\Sigma\,\boldsymbol{t}).$$

Independence:
Components of $\boldsymbol{X}$ are independent $\iff$ Σ is diagonal (for MVN, uncorrelated $\Rightarrow$ independent).
Mahalanobis: If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then

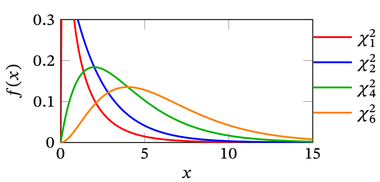
$$(\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \sim \chi^2_p \quad \text{and} \quad \Sigma^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu}) \sim N_p(0, I_p).$$

Multivariate Chi-squared Distribution:

If $\boldsymbol{X} \sim N_p(0, I_p)$, then $\chi^2 = \boldsymbol{X}^T\boldsymbol{X} = \sum_{i=1}^p X_i^2 \sim \chi^2_p$. For non-central case, if $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, I_p)$, then $\boldsymbol{X}^T\boldsymbol{X} \sim \chi^2_p(\lambda)$ with non-centrality $\lambda = \boldsymbol{\mu}^T\boldsymbol{\mu}$.
Properties:

- $E[\chi^2_p] = p$, $\mathrm{Var}(\chi^2_p) = 2p$
- For independent $\chi^2_{p_1}, \chi^2_{p_2}$: $\chi^2_{p_1} + \chi^2_{p_2} \sim \chi^2_{p_1+p_2}$
- MGF: $M_{\chi^2_p}(t) = (1-2t)^{-p/2}$ for $t < 1/2$

Connection to quadratic forms: If $\boldsymbol{Z} \sim N_p(0, I_p)$ and A is a symmetric idempotent matrix of rank r, then $\boldsymbol{Z}^TA\boldsymbol{Z} \sim \chi^2_r$.



Principal Component Analysis (PCA):

For data with covariance matrix Σ (or correlation R), find eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ and eigenvectors $e_1, \dots, e_p$ (orthonormal) solving $\Sigma e_i = \lambda_i e_i$.
Properties:

- The ith PC is $Z_i = e_i^T(\boldsymbol{X} - \bar{\boldsymbol{X}})$ with $\mathrm{Var}(Z_i) = \lambda_i$
- PCs are uncorrelated (orthogonal directions of maximal variance)
- Proportion of variance explained by first k PCs: $\frac{\lambda_1+\dots+\lambda_k}{\lambda_1+\dots+\lambda_p}$

Quadratic Forms & Idempotent Matrices:

Idempotent Matrices: A matrix H is *idempotent* if $H^2 = H$.

- Idempotent H has eigenvalues 0 or 1 only
- $\mathrm{rank}(H) = \mathrm{tr}(H)$ (number of eigenvalues = 1)

Quadratic Forms: If $Z \sim N(0, I_n)$ and P is symmetric idempotent of rank r, then:

$$Q = Z^TPZ \sim \chi^2_r$$
$$Z \sim N(0, \sigma^2 I_n) \implies Q/\sigma^2 \sim \chi^2_r$$

Independence: If P_1 and P_2 are symmetric idempotent with $P_1P_2 = 0$ (projections onto orthogonal subspaces), then Z^TP_1Z and Z^TP_2Z are independent.

Linear Model Setup:

Assume $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$ where $\boldsymbol{Y}$ is $n\times 1$, $\boldsymbol{X}$ is $n \times p$ (full rank p), β is $p\times 1$ of unknown parameters, and $\varepsilon \sim N(0, \sigma^2 I_n)$ (errors independent, homoscedastic, normal). *Assumptions:* linear relationship, $E[\varepsilon] = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$, independent errors (normality for inference). Under these assumptions,

$$\boldsymbol{Y} \sim N(\boldsymbol{X}\beta,\ \sigma^2 I).$$

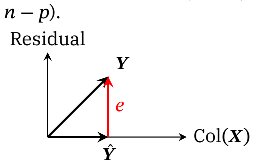
The ordinary least squares (OLS) estimator is:

$$\hat\beta = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \quad \begin{cases} E[\hat\beta] = \beta & \text{(unbiased)} \\ \mathrm{Var}(\hat\beta) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} & \\ \mathrm{SE}(\hat\beta) = \sqrt{\mathrm{diag}(\mathrm{Var}(\hat\beta))} = \sigma\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}} & \text{(std. error)} \end{cases}$$

solving the normal equations $\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat\beta) = 0$. The *fitted values* are,

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat\beta = H\boldsymbol{Y}.$$

Hat matrix $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is symmetric and idempotent (rank p). The *residuals* are $e = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I - H)\boldsymbol{Y}$, where $(I-H)$ is symmetric idempotent (rank $n - p$).



Gauss–Markov theorem: $\hat\beta$ is the best linear unbiased estimator (minimal variance). If ε is normal, then $\hat\beta \sim N(\beta,\ \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

$$\hat{\boldsymbol{Y}} = H\boldsymbol{Y} \sim N(\boldsymbol{X}\beta,\ \sigma^2 H), \quad e = (I-H)\boldsymbol{Y} \sim N(0,\ \sigma^2(I-H)).$$

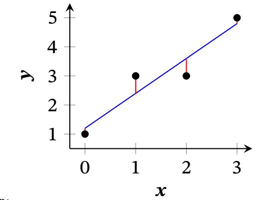
Moreover, $\hat{\boldsymbol{Y}}$ and e are independent (since $H(I-H) = 0$).
If an *intercept* (β_0) is included,

$$\mathrm{SST} = \sum \overbrace{(\hat Y_i - \bar Y)^2}^{\mathrm{SSR}} + \sum \overbrace{(Y_i - \hat Y_i)^2}^{\mathrm{SSE}} = \sum (Y_i - \bar Y)^2 \tag{1}$$

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}, \quad \mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p-1}(1 - \mathrm{R}^2) \tag{2}$$

$$\mathrm{df}_{\mathrm{SSR}} = p - 1, \quad \mathrm{df}_{\mathrm{SSE}} = n - p, \quad \mathrm{df}_{\mathrm{SST}} = n - 1 \qquad \text{(DoF)}$$

Under normal errors, $\mathrm{SS_{err}}/\sigma^2 \sim \chi^2_{n-p}$ then $\mathrm{SS_{err}}$ is independent of $\hat\beta$. An unbiased estimator of σ^2 is $\hat\sigma^2 = \frac{1}{n-p}\mathrm{SS_{err}}$.



Inference in Linear Model:

For each parameter β_j, under $H_0 : \beta_j = 0$,

$$t = \frac{\hat\beta_j}{\mathrm{SE}(\hat\beta_j)} \sim t_{n-p}.$$

Thus $(1-\alpha)$ CI for β_j is:

$$\hat\beta_j \pm t_{n-p,\,\alpha/2}\,\mathrm{SE}(\hat\beta_j)$$

For a new predictor value x_0, the predicted response $\hat y_0 = x_0\hat\beta$ has $\mathrm{Var}(\hat y_0) = \sigma^2 x_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_0^T$. A $(1-\alpha)$ CI for the mean at x_0 is $\hat y_0 \pm t_{n-p,\,\alpha/2}\ \hat\sigma\sqrt{x_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_0^T}$, and a prediction interval for a new $\boldsymbol{Y}$ at x_0 is $\hat y_0 \pm t_{n-p,\,\alpha/2}\ \hat\sigma\sqrt{x_0(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_0^T + 1}$.

Hypothesis Testing:

F-test:

For general linear hypothesis $H_0 : L\beta = 0$ where L is a $d \times p$ matrix of rank d:

$$F = \frac{(L\hat\beta)^T[L(\boldsymbol{X}^T\boldsymbol{X})^{-1}L^T]^{-1}(L\hat\beta)}{\hat\sigma^2} \sim F_{d,n-p}$$

- Testing $H_0 : \beta_j = 0$: $t^2 = F_{1,n-p}$
- Global test $H_0 : \beta_1 = \dots = \beta_{p-1} = 0$: $F = \frac{\mathrm{SSR}/(p-1)}{\mathrm{SSE}/(n-p)} \sim F_{p-1,n-p}$

Reject H_0 if $F > F_{d,n-p,\alpha}$.

Model Selection:

Common criteria: $\mathrm{AIC} = n\ln(\mathrm{SSE}/n) + 2p$, $\quad \mathrm{BIC} = n\ln(\mathrm{SSE}/n) + p\ln n$ (smaller is better). Mallows' $C_p \approx \frac{\mathrm{SSE_{model}}}{\hat\sigma^2_{\mathrm{(full)}}} - (n - 2p)$, targeting $C_p \approx p$.

Diagnostics:

Residual standard deviation: $\hat\sigma = \sqrt{\mathrm{SSE}/(n-p)}$. Check residual plots for constant variance (no patterns in residuals vs fitted) and for normality (e.g. Q–Q plot). High-leverage points: $h_{ii} = H_{ii}$; large h_{ii} (relative to p/n) indicates an outlier in predictor space. Standardized residual $r_i = \frac{e_i}{\hat\sigma\sqrt{1-h_{ii}}}$ (should be $\approx N(0,1)$ under the model). Outliers may have $|r_i| > 2$ or 3. Influence can be assessed by Cook's distance: $D_i = \frac{e_i^2}{p\,\hat\sigma^2}\frac{h_{ii}}{(1-h_{ii})^2}$ (values $D_i > 1$ are often considered large). If assumptions are violated (nonlinearity, heteroscedasticity, non-normal errors), consider remedies such as transforming variables or using a different model. For variance stabilization, choose $g(\boldsymbol{y})$ such that $\mathrm{Var}(g(\boldsymbol{Y}))$ is roughly constant. E.g. for Poisson $\boldsymbol{Y}$ (Var $\boldsymbol{\mu}$), use $g(\boldsymbol{Y}) = \sqrt{\boldsymbol{Y}}$; for Binomial proportion (Var $\approx \boldsymbol{\mu}(1-\boldsymbol{\mu})$), use $g(\boldsymbol{Y}) = \arcsin\sqrt{\boldsymbol{Y}}$; for Var $\propto \boldsymbol{\mu}^2$, use $\log \boldsymbol{Y}$ (Box–Cox power transform can find an optimal λ).

Multiple Testing:

When performing m hypothesis tests, the family-wise error rate (FWER) at level α is Pr(any false rejection) $\leq \alpha$.
Bonferroni correction: $\alpha = \frac{\alpha}{m}$ for each test (controls FWER $\leq \alpha$).
Šidák correction: $\alpha = 1 - (1-\alpha)^{1/m}$.

Variance–Stabilising Transformations: Step-by-Step Guide

Given a random variable Y with mean $\mu = \mathbb{E}[Y]$ and variance $\mathrm{Var}(Y) = v(\mu)$, a variance–stabilising transformation g is obtained as follows:

$$\mu = \mathbb{E}[Y], \quad v(\mu) = \mathrm{Var}(Y) \tag{3}$$

$$\mathrm{Var}\big[g(Y)\big] \approx \big[g'(\mu)\big]^2 v(\mu) = C \tag{4}$$

$$g'(\mu) = \frac{\sqrt{C}}{\sqrt{v(\mu)}} \tag{5}$$

$$g(y) = \sqrt{C} \int \frac{dy}{\sqrt{v(y)}} \tag{6}$$

Typically choose $C = 1$ and drop the additive constant C_0.

ANOVA (Analysis of Variance):

One-way ANOVA with a groups (levels) and total N observations: model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, for $i = 1, \dots, a$ and $j = 1, \dots, n_i$, where τ_i is the effect of group i (with $\sum_i \tau_i = 0$) and $\varepsilon_{ij} \sim N(0, \sigma^2)$. Null hypothesis $H_0 : \tau_1 = \tau_2 = \cdots = \tau_a = 0$ (all group means equal) is tested by

$$F = \frac{\mathrm{SSB}/(a-1)}{\mathrm{SSW}/(N-a)} \sim F_{a-1, N-a},$$

where $\mathrm{SSB} = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$ (between-group SS) and $\mathrm{SSW} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ (within-group SS). Total $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}$ with df $N - 1 = (a-1) + (N-a)$. If H_0 is rejected, follow-up with multiple-comparison tests (adjusting for multiple testing). Two-way ANOVA (two factors) and factorial designs partition variability into main effects and interaction effects similarly, each tested with an F-ratio of mean squares.

Design of Experiments (DOE):

In a 2^k full factorial design, k factors each have 2 levels (often coded -1 and $+1$). All 2^k combinations are run (possibly with replication). The *main effect* of factor A is the difference in average response between high and low levels of A; an *interaction* effect (e.g. AB) is the difference in A's effect between the two levels of B. For example, A effect $= \bar{Y}(A^+) - \bar{Y}(A^-)$, and AB interaction $= [(\bar{Y}_{A+B+} - \bar{Y}_{A-B+}) - (\bar{Y}_{A+B-} - \bar{Y}_{A-B-})]$. Effects can be estimated via a regression model $Y = \beta_0 + \sum \beta_i x_i + \sum \beta_{ij} x_i x_j + \cdots$ with $x_i = \pm 1$. (In this coding, β_i equals half the main effect for factor i, etc.) Orthogonal designs: in a full factorial, the design matrix columns for each effect are orthogonal, simplifying estimation and interpretation (no confounding among effects). Blocking: to account for nuisance variables, experiments may be divided into blocks. In a 2^k design with 2 blocks, one effect (usually a high-order interaction) is *confounded* with the block factor (indistinguishable from a block effect). E.g. in a 2^3 design, to run in 2 blocks we can confound the ABC interaction with blocks by assigning all runs with $ABC = +1$ to Block 1 and $ABC = -1$ to Block 2. Then any systematic block difference will appear as an ABC effect (and vice versa). Fractional factorial 2^{k-p} designs run a fraction of the 2^k runs. Specified by p *generators* (defining relations). E.g. a 2^{3-1} half-fraction with generator $I = ABC$ means we run only combinations where $ABC = +1$. This yields aliasing: A is aliased with BC, B with AC, C with AB. (Defining relation $I = ABC$; resolution III design since smallest alias involves 3 factors.) Design resolution R: no effect involving $< R$ factors is aliased with any other effect with $< R$ factors. Higher resolution designs reduce confounding (e.g. resolution IV: no main effect aliased with any other main or 2-factor effect; resolution V: no main or 2-factor aliased with any other up to 2-factor effect, etc.). To de-alias key effects, one can run fold-over (the complementary fraction) or combine fractional designs to a higher resolution.