10/26/2023

Outline

1) χ^2, t and F distributions
2) Canonical linear model
3) General linear model

"Gaussian-adjacent" distributions

If $Z_1, \dots, Z_d \overset{iid}{\sim} N(0,1)$ then

$$V = \sum Z_i^2 \sim \chi_d^2 = \text{Gamma}(\overset{\text{shape}}{d/2}, \overset{\text{scale}}{2})$$

$$\mathbb{E} V = d, \quad \text{Var}(V) = 2d$$

CLT: $\dfrac{V-d}{\sqrt{d}} \Rightarrow N(0,2)$

(informal) $\dfrac{V}{d} \approx N(1, 2/\sqrt{d}) \to 1$

If $Z \sim N(0,1)$ and $V \sim \chi_d^2$, $Z \perp\!\!\!\perp V$ then

$$\frac{Z}{\sqrt{V/d}} \sim t_d \Rightarrow N(0,1) \quad \text{as} \quad d \to \infty$$

If $V_1 \sim \chi_{d_1}^2$ and $V_2 \sim \chi_{d_2}^2$, $V_1 \perp\!\!\!\perp V_2$ then

$$\frac{V_1/d_1}{V_2/d_2} \sim F_{d_1,d_2} \Rightarrow \frac{1}{d_1}\chi_{d_1}^2 \quad \text{as} \quad d_2 \to \infty$$

Note if $T \sim t_d$ then $T^2 \sim F_{1,d}$

Recall: $Z \sim N_d(\mu, \Sigma)$, $A \in \mathbb{R}^{k \times d}$, $b \in \mathbb{R}^k$

$$\Rightarrow AZ + b \sim N_k(A\mu + b, A\Sigma A')$$

Change of basis (1-sample t-test)

Let $Q = \begin{pmatrix} | & | & & | \\ q_1 & q_2 & \cdots & q_n \\ | & | & & | \end{pmatrix} = \begin{array}{c c} & \begin{array}{cc} 1 & n-1 \end{array} \\ n & \begin{pmatrix} | & \\ q_1 & Q_r \\ | & \end{pmatrix} \end{array}$

where $q_1 = \frac{1}{\sqrt{n}} \cdot 1_n$,

$q_2, \ldots, q_n$ complete orthonormal basis

(e.g. via Gram-Schmidt)

$X_i \overset{iid}{\sim} N(\mu, \sigma^2) \iff X \sim N_n(\mu \cdot 1_n, \sigma^2 I_n)$

New basis:

$$Z = Q'X = \begin{array}{c c} & 1 \\ \begin{array}{c} 1 \\ n-1 \end{array} & \begin{pmatrix} q_1' X \\ Q_r' X \end{pmatrix} \end{array} = \begin{pmatrix} \sqrt{n}\, \bar{X} \\ Q_r' X \end{pmatrix}$$

$$\|Q_r' X\|^2 = \|Q'X\|^2 - \|q_1' X\|^2$$

$$= \|X\|^2 - n\bar{X}^2 \qquad (Q'Q = I_n)$$

$$= (n-1) s^2$$

$$Q'X \sim N_n\left(\begin{pmatrix} \sqrt{n}\,\mu \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ \sigma^2 I_n \right)$$

$$Z_1 \sim N(\sqrt{n}\,\mu, \sigma^2)$$

$$Z_r = Q_r' X \sim N(0, \sigma^2 I_{n-1})$$

$$\Rightarrow s^2 = \frac{1}{n-1} \|Z_r\|^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$$

and $s^2 \perp\!\!\!\perp Z_1$ (we already knew, from Basu)

Geometric interp.

$$T^2 = \frac{n\bar{X}^2}{S^2} = \frac{\|\mathrm{Proj}_{1_n} X\|^2}{\frac{1}{n-1}\|\mathrm{Proj}_{1_n}^{\perp} X\|^2} \sim F_{1,n-1}$$

$$= \frac{(\text{magnitude of } X \text{ in special dir.})^2}{(\text{average magn. of } X \text{ in resid. dir.s})^2}$$

Independent of total magnitude (under H_0):

$$n\bar{X}^2 \overset{H_0}{\sim} \sigma^2 \chi_1^2 = \mathrm{Gamma}\left(\tfrac{1}{2}, 2\sigma^2\right)$$

$$(n-1)S^2 \sim \sigma^2 \chi_{n-1}^2 = \mathrm{Gamma}\left(\tfrac{n-1}{2}, 2\sigma^2\right)$$

$$\|X\|^2 = n\bar{X}^2 + (n-1)S^2 \overset{H_0}{\sim} \sigma^2 \chi_n^2 = \mathrm{Gamma}\left(\tfrac{n}{2}, 2\sigma^2\right)$$

$$\Rightarrow \quad \frac{n\bar{X}^2}{\|X\|^2} \sim \mathrm{Beta}\left(\tfrac{1}{2}, \tfrac{n-1}{2}\right), \quad \text{indep. of } \|X\|^2$$

$$= \frac{n\bar{X}^2}{n\bar{X}^2 + (n-1)S^2}$$

F_{d_1,d_2} related to $\mathrm{Beta}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$: If $U \sim \mathrm{Beta}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$

Then $\dfrac{U/d_1}{(1-U)/d_2} \sim F_{d_1,d_2}$

Canonical Linear Model

Assume
$$Z = \begin{array}{r} d_0 \\ d_1 = d - d_0 \\ d_r = n - d \end{array} \begin{pmatrix} Z_0 \\ Z_1 \\ Z_r \end{pmatrix} \sim N_n\left(\begin{pmatrix} \mu_0 \\ \mu_1 \\ 0 \end{pmatrix}, \sigma^2 I_n \right)$$

$$\mu_0 \in \mathbb{R}^{d_0}, \quad \mu_1 \in \mathbb{R}^{d_1}, \quad \sigma^2 > 0$$

Test $H_0 : \mu_1 = 0$ vs. $H_1 : \mu_1 \neq 0$

(or possibly one-sided, if $d_1 = 1$).

Exp. Fam.:
$$p(z) = e^{\frac{\mu_1}{\sigma^2}' z_1 + \frac{\mu_0}{\sigma^2}' z_0 - \frac{1}{2\sigma^2} \|z\|^2}$$

σ^2 known, $d_1 = 1$:

"Cond. on Z_0", reject for large(/small/extreme) Z_1

$Z_0 \perp\!\!\!\perp Z_1$, test stat is $Z_1 \sim N(\mu_1, \sigma^2)$

$$\frac{Z_1}{\sigma} \overset{H_0}{\sim} N(0,1) \qquad (\underline{z\text{-test}})$$

σ^2 known, $d_1 \geq 1$: reject for large $\|Z_1\|$

unless we have anisotropic prior on μ_1

$$\|Z_1\|^2 / \sigma^2 \sim \chi^2_{d_1} \qquad (\underline{\chi^2\text{-test}})$$

σ^2 unknown, $d_1 = 1$:

Cond. on Z_0, $\|Z\|^2 = \|Z_1\|^2 + \|Z_0\|^2 + \|Z_r\|^2$

Reject for large(/small/extreme) Z_1

$\iff$ Reject for large $Z_1 / \|Z\|$

$\iff$ Reject for large $\dfrac{Z_1}{\sqrt{\|Z_r\|^2/(n-d)}} \overset{H_0}{\sim} t_{d_r}$

(t - test)

σ^2, $d_1 \geq 1$: Reject for (conditionally) large $\|Z_1\|^2$

$\iff$ Reject for large $\dfrac{\|Z_1\|^2/d_1}{\|Z_r\|^2/(n-d)} \overset{H_0}{\sim} F_{d_1, n-d}$

(F - test)

Here $\|Z_r\|^2/d_r \sim \dfrac{\sigma^2}{d_r} \chi^2_{n-d}$

functioning as estimator of σ^2

$\mathbb{E}\,\hat{\sigma}^2 = \sigma^2$, $\mathrm{Var}(\hat{\sigma}^2) = 2\sigma^2/(n-d)$

Compare:

z: Z_1/σ $\qquad$ t: $Z_1/\hat{\sigma}$

χ^2: $\|Z_1\|^2/\sigma^2$ $\qquad$ F: $\dfrac{\|Z_1\|^2/d_1}{\hat{\sigma}^2}$

Intervals for Canonical Model

How to test $H_0 : \mu_1 = \mu_1^0 \in \mathbb{R}^d$?

<u>Problem:</u> μ_1 is not a natural parameter.

Translate problem:

$$\begin{pmatrix} Z_0 \\ Z_1 - \mu_1^0 \\ Z_r \end{pmatrix} \sim N_d\left(\begin{pmatrix} \mu_0 \\ \mu_1 - \mu_1^0 \\ 0 \end{pmatrix}, \ \sigma^2 I_n \right)$$

Can do same tests with $Z_1 - \mu_1^0$ replacing Z_1

Invert:

<u>$d_1 = 1$, σ^2 kn</u> $\quad \frac{Z_1 - \mu_1}{\sigma} \sim N(0,1) \rightsquigarrow$ CI $\quad Z_1 \pm \sigma z_{\alpha/2}$

$$= \left[Z_1 - \sigma z_{\alpha/2},\ Z_1 + \sigma z_{\alpha/2} \right]$$

<u>$d_1 = 1$, $\hat{\sigma}$ unkn</u> $\quad \frac{Z_1 - \mu_1}{\hat{\sigma}} \sim t_{n-d} \rightsquigarrow \quad Z_1 \pm \hat{\sigma}\, t_{n-d}^{(\alpha/2)}$

$\{x : \|x\| \le 1\}$

<u>$d_1 \ge 1$, σ^2 kn :</u> $\quad \frac{\|Z_1 - \mu_1\|}{\sigma} \sim \chi^2_{d_1} \rightsquigarrow \quad Z_1 + \sigma \sqrt{c_{\chi^2}(\alpha)}\, B_1(0)$

upper-α quantile

<u>$d_1 \ge 1$, σ^2 unkn :</u> $\quad \frac{\|Z_1 - \mu_1\|}{\hat{\sigma}} \sim F_{d_1, n-d} \rightsquigarrow \quad Z_1 + \hat{\sigma} \sqrt{c_F(\alpha)}\, B_1(0)$

General Linear Model:

Many problems can be put into canonical linear model after change of basis.

Basic setup:

Observe $Y \sim N_n(\theta, \sigma^2 I_n)$, $\sigma^2 > 0$

(known or unknown)

Test $\theta \in \Theta_0$ vs. $\theta \in \Theta \setminus \Theta_0$

where $\Theta_0 \subseteq \Theta$ are subspaces of $\mathbb{R}^n$

$\dim(\Theta_0) = d_0$, $\dim(\Theta) = d = d_0 + d_1$

Idea: rotate into canonical form

$$Q = \begin{array}{c c} & \begin{array}{ccc} d_0 & d_1 & n-d \end{array} \\ n & \begin{bmatrix} Q_0 & Q_1 & Q_r \end{bmatrix} \end{array}$$

orthonormal basis for Θ_0

o.b. for $\Theta \cap \Theta_0^\perp$

o.b. for $\mathbb{R}^n \cap \Theta^\perp$

$$Z = Q'Y \sim N_n\left(\begin{pmatrix} Q_0'\theta \\ Q_1'\theta \\ 0 \end{pmatrix}, \sigma^2 I_n \right)$$

$H_0: Q_1'\theta = 0$

Do z, χ^2, t, or F-test as appropriate

Ex. Linear Regression $\quad x_i \in \mathbb{R}^d$ fixed

$$Y_i = x_i' \beta + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$Y \sim N_d(X\beta, \sigma^2 I_n) \qquad X = \begin{pmatrix} -x_1'- \\ \vdots \\ -x_n'- \end{pmatrix} \in \mathbb{R}^{d \times n}$$

$$= \begin{pmatrix} | & & | \\ X_1 & \cdots & X_d \\ | & & | \end{pmatrix}$$ capital letters

(Assume X has full column rank)

$$\theta = X\beta \in \Theta = \text{Span}(X_1, \dots, X_d)$$

$$H_0: \beta_1 = \cdots = \beta_{d_1} = 0, \qquad (1 \le d_1 \le d)$$

$$\iff \theta \in \text{Span}(X_{d_1+1}, \dots, X_d)$$

$$(\text{or } \{0\} \text{ if } d_1 = d)$$

$$\|Z_r\|^2 = \|Y - \text{Proj}_{\Theta}(Y)\|^2$$

$$= \|Y - X\hat{\beta}_{OLS}\|^2 \qquad \hat{\beta}_{OLS} = \text{argmin} \|Y - X\beta\|^2$$

$$= (X'X)^{-1}X'Y$$

$$= \sum (Y_i - x_i'\hat{\beta})^2$$

$= \underline{\text{Residual sum of squares}}$ (RSS)

$$\|Z_1\|^2 + \|Z_r\|^2 = \|Y - \text{Proj}_{\Theta_0}(Y)\|^2 = RSS_0 \quad (\text{null RSS})$$

F-statistic is $\dfrac{\|z_1\|^2/(d-d_0)}{\|z_r\|^2/(n-d)} = \dfrac{(RSS_0 - RSS)/(d-d_0)}{RSS/(n-d)}$

$n-d$ called <u>residual degrees of freedom</u>

$d_1 = 1$: Let $X_0 = (X_2 \cdots X_d) \in \mathbb{R}^{d_0 \times n}$

Let $X_{1\perp} = X_1 - \mathrm{Proj}_{\omega_0}(X_1)$

$= X_1 - X_0 (X_0' X_0)^{-1} X_0' X_1$

$= X_1 - X_0 \gamma$

<u>Reparametrize</u>:

$$\theta = X\beta \iff \theta = X_{1\perp}\beta_1 + X_0 \overbrace{(\beta_{-1} + \gamma)}^{\delta}$$

$$\begin{array}{c} \\ 1 \\ d-1 \end{array} \begin{pmatrix} \overset{1}{\hat\beta_1} \\ \hat\delta \end{pmatrix} = \begin{array}{c} \\ 1 \\ d-1 \end{array} \begin{array}{c} \begin{array}{cc} 1 & d-1 \end{array} \\ \begin{bmatrix} X_{1\perp}' X_{1\perp} & 0 \\ 0 & X_0' X_0 \end{bmatrix}^{-1} \end{array} \begin{pmatrix} X_{1\perp}' Y \\ X_0' Y \end{pmatrix} = \begin{pmatrix} X_{1\perp}'Y / \|X_{1\perp}\|^2 \\ (X_0'X_0)^{-1} X_0' Y \end{pmatrix}$$

$\hat\beta_1 = X_{1\perp}' Y / \|X_{1\perp}\|^2$, $s.e.(\hat\beta_1) = \sigma / \|X_{1\perp}\|$

$q_1 = X_{1\perp} / \|X_{1\perp}\|$, $Q_1 = (q_1)$, $Q_0 = X_0 (X_0'X_0)^{-1} X_0'$

t-statistic: $\dfrac{q_1' Y}{\sqrt{RSS/(n-d)}} = \dfrac{\hat\beta_1}{\hat\sigma / \|X_{1\perp}\|} = \dfrac{\hat\beta_1}{\widehat{s.e.}(\hat\beta_1)}$

Ex: Two-sample t-test (equal variance)

$$Y_1, \dots, Y_m \overset{iid}{\sim} N(\mu, \sigma^2) \qquad Y_{m+1}, \dots, Y_{n+m} \overset{iid}{\sim} N(\nu, \sigma^2)$$

Model: $\theta = \mathbb{E}Y = \begin{pmatrix} \mu 1_m \\ \nu 1_n \end{pmatrix} \iff \theta \in \operatorname{Span}\left(\begin{pmatrix} 1_m \\ -1_n \end{pmatrix}, 1_{n+m} \right)$

$H_0: \mu_1 = \mu_2 \iff \theta \in \operatorname{Span}(1_{n+m})$

$d_0 = 1, \quad d = 2, \quad d_r = n + m - 2$

Orthogonalize $\begin{pmatrix} 1_m \\ -1_n \end{pmatrix} \rightsquigarrow \begin{array}{c} \\ m\left\{ \vphantom{\begin{matrix} a \\ a \\ a \end{matrix}} \right. \\ n\left\{ \vphantom{\begin{matrix} a \\ a \\ a \end{matrix}} \right. \end{array} \begin{pmatrix} 1/m \\ \vdots \\ 1/m \\ -1/n \\ \vdots \\ -1/n \end{pmatrix}$

$\Rightarrow$ Reject for large

$$\frac{\frac{1}{m} \sum_{i \le m} Y_i - \frac{1}{n} \sum_{i > m} Y_i}{\sqrt{\frac{1}{m} + \frac{1}{n}} \cdot \sqrt{RSS/(n+m-2)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{\sigma} \cdot \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

Ex. One-way ANOVA: (fixed effects)

$$Y_{k,i} \overset{\text{ind.}}{\sim} \mu_k + \varepsilon_{k,i} \qquad \varepsilon_{k,i} \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

$$k = 1, \dots, m \qquad i = 1, \dots, n$$

$$H_0 : \mu_1 = \dots = \mu_m = \mu$$

$$\bar{Y}_k = \frac{1}{n} \sum_i Y_{k,i} \qquad S_k^2 = \frac{1}{n-1} \sum_i (Y_{k,i} - \bar{Y}_k)^2$$

$$\bar{Y} = \frac{1}{mn} \sum_k \sum_i Y_{k,i} \qquad S_0^2 = \frac{1}{mn-1} \sum_k \sum_i (Y_{k,i} - \bar{Y})^2$$

$$d_0 = 1, \quad d = m, \quad d_r = m(n-1)$$

$$RSS = \sum_{k,i} (Y_{k,i} - \bar{Y}_k)^2 = \|Y\|^2 - n \sum_k \bar{Y}_k^2$$

$$RSS_0 = \sum_{k,i} (Y_{k,i} - \bar{Y})^2 = \|Y\|^2 - mn \bar{Y}^2$$

$$RSS_0 - RSS = n \left(\sum_k \bar{Y}_k^2 - m\bar{Y}^2 \right)$$

$$= n \sum_k (\bar{Y}_k - \bar{Y})^2$$

$$F\text{-stat} = \frac{\frac{n}{m-1} \sum_k (\bar{Y}_k - \bar{Y})^2}{\frac{1}{m(n-1)} \sum_k \sum_i (Y_{k,i} - \bar{Y}_k)^2} \qquad \begin{matrix} \leftarrow \text{"between" variance} \\ \leftarrow \text{"within" variance} \end{matrix}$$