Inferences about multivariate means

March 7, 2024

Distributions used in statistical inference

- There is a group of continuous variables and corresponding distributions that are important not for modeling real data but because they have been found useful in statistical procedures.
- We discuss three classes:
    - Chi-square distribution with k degrees of freedom
    - Student's t-distribution with k degrees of freedom
    - F-distribution with k and l degrees of freedom

Student's t-distribution

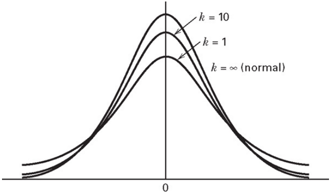
Defintion of Student's t-distribution

Student's t-distribution or simply t-distributions with k degrees of freedom is the distribution of

$$\sqrt{k}\frac{Z}{\sqrt{Y}},$$

where Z is a standard normal $N(0,1)$ independent of the chi-squared distributed random variable Y with k degrees of freedom.

Density plot:



Historical remark

William S. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name “Student.” His main contribution was published in 1908.

Student's *t*-distribution, further properties

- The following variable has also t distribution with $k = n - 1$ df

$$X = \sqrt{n}\frac{\overline{Y} - \mu}{S},$$

where as before $\overline{Y}$ is the average of Y_i's and S^2 is the sample variance of independent Y_i's having the $N(\mu, \sigma^2)$ distribution.

- For large k the distribution resembles that of the standard normal distribution– essentially for $k > 100$ they are identical.
- For $k > 1$ the expected value is well defined and is equal to zero. The variance is well defined for $k > 2$, where it is given as $Var(X) = k/(k-2)$. We note that it is larger than 1.

F distribution

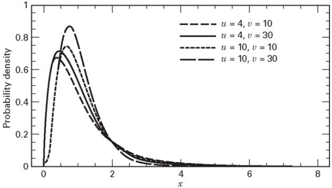
Definition of F-distribution

F distribution with k and l degrees of freedom is the distribution of

$$X = \frac{U^2/k}{V^2/l},$$

where U^2 and V^2 are independently distributed chi-square random variables with k and l degrees of freedom, respectively, i.e. $U^2 \sim \chi^2(k)$ and $V^2 \sim \chi^2(l)$. We write it shortly $X \sim F(k, l)$.

Densities:



F distribution, further properties

- Consider two sets of variables $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$.
- Let all these variables be independent of each other.
- Let $Y_i \sim N(\mu_Y, \sigma_Y^2)$ and $Z_i \sim N(\mu_Z, \sigma_Z^2)$
- Let S_Y^2 and S_Z^2 be sample variances computed for each set of variables.
- Then
$$\frac{S_Y^2}{\sigma_Y^2} \Big/ \frac{S_Z^2}{\sigma_Z^2} \sim F(n-1, m-1).$$
- Exercise: Let T be t distributed with n degrees of freedom. Is T^2 somehow related to some F distribution?

The Student t-statistic

- Let $X_1, \ldots, X_n$ be independent and identically distributed as $\mathcal{N}(\mu, \sigma^2)$.
- To test a hypothesis $H_0 : \mu = \mu_0$ we use the test-statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

- If the null hypothesis is true, then t is distributed as Student's t-distribution with $n - 1$ degrees of freedom.
- Then also

$$t^2 = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0)$$

  is distributed as the F-distribution with 1 and $n - 1$ degrees of freedom.
- The two tests are equivalent if the alternative is two-tailed.

Estimates of mean and covariance matrix

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ independent p-dimensional observations from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Let

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$$

and

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$$

Test of $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ multivariate.

- Use the statistic (Hotelling's T^2)

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$$

  to test the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$
- If the null hypothesis is true, then

$$\frac{n-p}{(n-1)p}T^2$$

  is distributed as F with p and $n-p$ degrees of freedom.
- Large values of T^2 lead to rejection of the hypothesis.

Confidence region based on T^2

- By the duality between tests and confidence intervals, a $100(1-\alpha)$ confidence region for $\boldsymbol{\mu}$ is determined by the set of all μ such that

$$n(\bar{\mathbf{X}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

- The region is an ellipsoid centered in $\bar{\mathbf{X}}$ and with orientation and size determined by the eigenvectors and eigenvalues of **S**.

Relation to the eigenvalues/eigenvectors

- The length of the main axes are given by

$$\sqrt{\lambda_i}\sqrt{\frac{(n-1)p}{n(n-p)}F_{p,n-p}(\alpha)}$$

  with direction $\mathbf{e}_i$ where $\lambda_1, \ldots, \lambda_n$ and $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are the eigenvalues and eigenvectors of **S**.
- As a confidence region this will cover the true value of $\boldsymbol{\mu}$ with probability $1-\alpha$.
- Projections of $\boldsymbol{\mu}$ will be covered by the corresponding projections of the ellipsoid with simultaneous probability $1-\alpha$.

Simultaneous intervals

Probability that for *all* $\mathbf{a} \in R^p$ *simultaneously*, $\mathbf{a}'\boldsymbol{\mu}$ is covered by the interval with endpoints

$$\mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\frac{p(n-1)}{n(n-p)} F_{p,n-p}(\alpha)\mathbf{a}'\mathbf{S}\mathbf{a}}$$

is equal to $1-\alpha$.

Theorem (Maximization Lemma)

B - positive definite matrix of dimension $p \times p$
d - any vector of dimension $p \times 1$

$$max_{a\neq 0} \frac{(a'd)^2}{a'Ba} = d'B^{-1}d$$

Thus

$$T^2 = \max_{\mathbf{a}} n \frac{(a'(\bar{\mathbf{X}} - \boldsymbol{\mu}))^2}{a'Sa}$$

With probability $1 - \alpha$

$$T^2 = \max_{\mathbf{a}} t_{\mathbf{a}}^2 \leq \frac{p(n-1)}{n-p)} F_{p,n-p}(\alpha) = c^2 \quad \Leftrightarrow$$

$$t_{\mathbf{a}}^2 \leq c^2 \quad \forall \mathbf{a} \quad \Leftrightarrow$$

$$-c \leq t_{\mathbf{a}} \leq c \quad \forall \mathbf{a} \quad \Leftrightarrow$$

$$-c \leq \frac{\mathbf{a}'(\bar{\mathbf{X}} - \boldsymbol{\mu})}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}/n}} \leq c \quad \forall \mathbf{a} \quad \Leftrightarrow$$

$$\mathbf{a}'\boldsymbol{\mu} \in \mathbf{a}'\bar{\mathbf{X}} \pm \sqrt{\frac{p(n-1)}{(n-p)} F_{p,n-p}(\alpha) \mathbf{a}'\mathbf{S}\mathbf{a}/n}$$

Simultaneous confidence intervals

Define E_i as the event that

$$\bar{X}_i \pm t_{n-1}(\alpha/2)\sqrt{s_{ii}/n}$$

cover μ_i (the t-interval for μ_i cover μ_i.) Then

$$\begin{aligned} P(\cap_{i=1}^{p} E_i) &= 1 - P(\cup_{i=1}^{p} E_i^c) \\ &\geq 1 - \sum_{i=1}^{p} P(E_i^c) \qquad \text{(Bonferroni)} \\ &= 1 - \sum_{i=1}^{p} \alpha = 1 - p\alpha \end{aligned}$$

Choosing α

- Simultaneous $100(1-p\alpha)$ confidence intervals can be obtained from p, t-intervals each of confidence $100(1-\alpha)$. So if we want to have an overall level of $100(1-\alpha)$ this suggest setting each t-interval to confidence $100(1-\alpha/p)$.
- **Special case:** If the events E_i are independent, then

$$P(\cap_{i=1}^p E_i) = (1-\alpha)^p < 1-\alpha$$

  when $p > 1$.
- **EXAMPLE** With $\alpha = 0.01$, $p = 10$, and confidence intervals represented by E_i are independent, then $(1-\alpha)^p = 0.99^{10} = 0.9044$.
- **Bonferroni interval** Distribute α over the t-intervals:

$$\bar{X}_i \pm t_{n-1}(\alpha_i/2)\sqrt{s_i^2/n}$$

  where $\sum_{i=1}^p \alpha_i = \alpha$.

Comments

- In practice we have two methods for constructing confidence intervals.
- Hotellings T^2 have exact confidence for all linear conbinations.
- Bonferronis method concerns a finite number of such and can be more liberal than Hotellings method.

Multiple testing

March 7, 2024

Identifying genes associated with cancer

$X_{n_1 \times p}$ - expressions of p genes for n_1 healthy individuals

$Y_{n_2 \times p}$ - expressions of p genes for n_2 cancer patients

Assumption: X_{ij} for $i = 1, \dots, n_1$ are iid with $E(X_{ij}) = \mu_{1j}$ and $Var(X_{ij}) = \sigma^2_{1j} < \infty$

Y_{ij} for $i = 1, \dots, n_2$ are iid with $E(Y_{ij}) = \mu_{2j}$ and $Var(Y_{ij}) = \sigma^2_{2j} < \infty$

Gene j is associated with cancer if $\mu_{1j} \neq \mu_{2j}$

We test $H_{0j} : \mu_{1j} = \mu_{2j}$ with a t-test $t_j = \frac{\bar{X}_{.j} - \bar{Y}_{.j}}{S(\bar{X}_{.j} - \bar{Y}_{.j})}$, where $S(\bar{X}_{.j} - \bar{Y}_{.j})$ is the estimate of the standard deviation of $\bar{X}_{.j} - \bar{Y}_{.j}$

If n_1 and n_2 are large enough then $t_j \sim N(\mu_j, 1)$ with $\mu_j = \frac{\mu_{1j} - \mu_{2j}}{\sigma_{1j}/\sqrt{n_1} + \sigma_{2j}/\sqrt{n_2}}$ and $H_{0j} : \mu_j = 0$

Multiple testing (1)

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \;\; \text{vs} \;\; \mu_i \neq 0$

Reject H_{0i} when $|X_i| > c$

Multiple comparison problem: if all μ_is are equal to zero than $max(|X_1|, \ldots, |X_p|) = \sqrt{2 \log p}(1 + o_p)$

Thus to separate signal from noise we need $c = c(p) \to \infty$ as $p \to \infty$.

Testing for global null, Bonferroni procedure

$X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, p$

$H_{0i} : \mu_i = 0 \quad \text{vs} \quad \mu_i \neq 0$

$$H_0 : \bigcap_{i=1}^{p} H_{0i}$$

Bonferroni procedure: Reject H_0 when

$\max(|X_1|, \ldots, |X_p|) \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = c_{Bon}$

Probability of type I error:

$$P_{H_0}\left(\bigcup_{j=1}^{p}\{|X_j| > c_{Bon}\}\right) \leq \sum_{j=1}^{p} P(\{|X_j| > c_{Bon}\} = \alpha$$

Exact type I error of Bonferroni

Due to independence

$$\begin{aligned} P(Type\ I\ Error) &= 1 - P_{H_0}\left(\bigcap_{j=1}^{p}\{|X_j| < c_{Bon}\}\right) \\ &= 1 - \left(1 - \frac{\alpha}{p}\right)^{p} \to 1 - e^{-\alpha} = \alpha + o(\alpha) \end{aligned}$$

$\alpha = 0.05$, $n = 30000, P(Type\ I\ Error) \approx 0.0488$

Multiple testing

We now separately test each of hypotheses $H_{0i} : \mu_i = 0$

	H_0 accepted	H_0 rejected	
H_0 true	U	V	p_0
H_0 false	T	S	p_1
	W	R	p

$$FWER = P(V > 0), \qquad FDR = E\left(\frac{V}{R \vee 1}\right)$$

$$E(V) = \alpha p_0$$

$$\alpha = 0.05, p_0 = 5000 \rightarrow E(V) = 250$$

Multiple testing procedures

Bonferroni correction: Use significance level $\frac{\alpha}{p}$.

Reject H_{0i} if $|X_i| \geq \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) = \sqrt{2 \log p}(1 + o(1))$

Benjamini-Hochberg (1995) procedure:

(1) $|X|_{(1)} \geq |X|_{(2)} \geq \ldots \geq |X|_{(p)}$

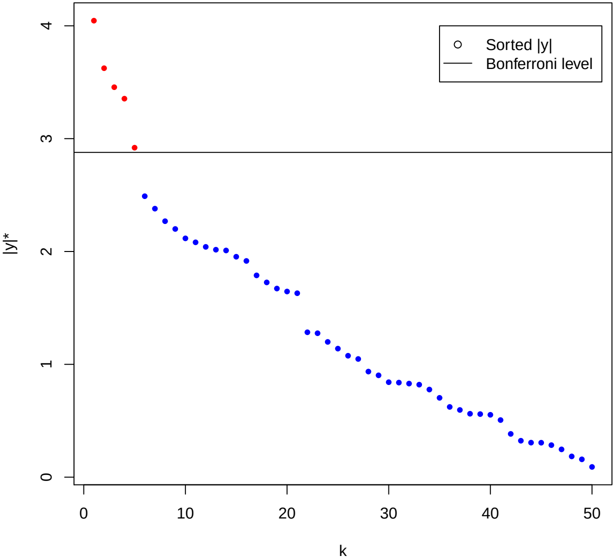
(2) Find the largest index i such that

$$|X|_{(i)} \geq \Phi^{-1}(1 - \alpha_i), \quad \alpha_i = \alpha \frac{i}{2p}, \tag{1}$$

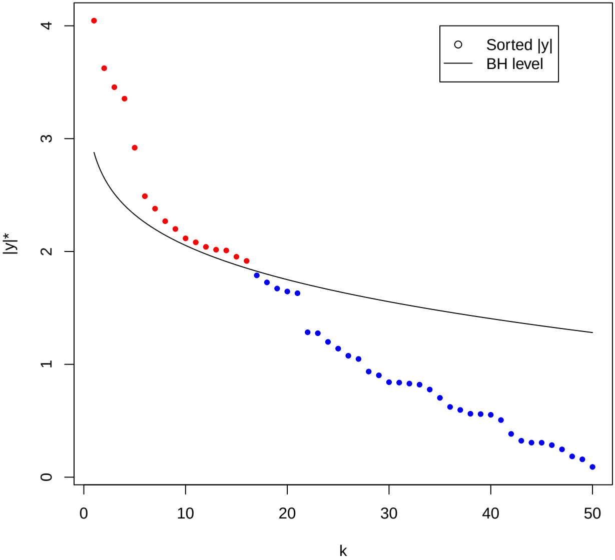
Call this index i_{SU}.

(3) Reject all $H_{(i)}$'s for which $i \leq i_{\mathrm{SU}}$

Bonferroni correction



Benjamini and Hochberg correction



FWER and FDR control

For Bonferroni correction $FWER \leq \alpha$

(Benjamini,Hochberg, 1995) If $X_1, \ldots, X_p$ are independent then BH controls FDR at:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{R \vee 1}\right] = \alpha\frac{p_0}{p}, \tag{2}$$

where p_0 is the number of true null hypotheses, $p_0 = |\{i : \mu_i = 0\}|$

(Benjamini, Yekutieli, 2001) When test statistics are ”positively correlated” then BH controls FDR at or below the level $\alpha\frac{p_0}{p}$. Independently of the correlation structure FDR is controlled at or below the level $\alpha\frac{p_0}{p}$ if $|X|_{(j)}$ is compared to $\Phi^{-1}\left(1 - \frac{j\alpha}{2p\sum_{i=1}^{p}\frac{1}{i}}\right)$.