Tutorial - Week 4

Please read the related material and attempt these questions before attending your allocated tutorial. Solutions are released on Friday 4pm.

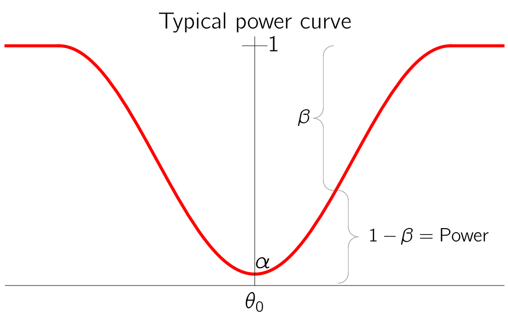
A large part of this course is concerned with trying to argue that various new and improved techniques (that use random matrix theory) are better than the well-known "classic" multivariate methods used in statistics **[A]** and machine learning **[B]**. At the core of most methods is the presence of hypothesis testing problem and we are trying to compare various methods/tests in regards to their *power*, the probability under repeated sampling of correctly rejecting an incorrect hypothesis, subject to some pre-specified *size*[1], the probability of incorrectly rejecting a true hypothesis.

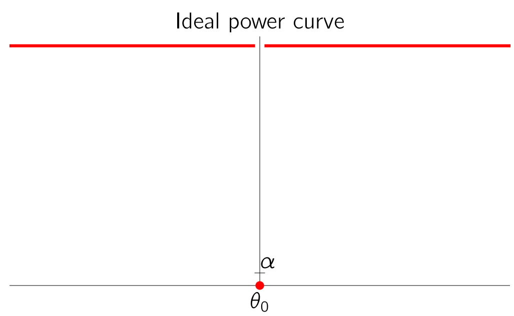
More precisely, suppose we have a parameter space Θ and consider hypotheses of the form

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are two disjoint subsets of Θ, possibly, but not necessarily, satisfying $\Theta_0 \cup \Theta_1 = \Theta$. For example, the parameter space could consist of the mean μ with $\Theta_0 = \{\mu : \mu = \mu_0\}$ and $\Theta_1 = \{\mu : \mu \neq \mu_0\}$. We call H_0 the *null hypothesis* and H_1 the *alternate hypothesis* (sometimes written H_a). A *Type I error* is made if H_0 is rejected when H_0 is true. The *probability of a Type I error* is denoted by α and is called the *size* of the test. A *Type II error* is made if H_0 is accepted when H_1 is true. The *probability of a type II error* is denoted by β. If θ_1 is the value of θ in the alternative hypothesis H_1, then the *power* of the test is $1 - \beta$.

If $\Theta_0 = \{\theta : \theta = \theta_0\}$ and we plot power for varying $\theta \in \Theta$, we get the following typical power curve that we might observe compared to an ideal (i.e., best possible) power curve (1 everywhere except for $\theta = \theta_0$ where it is zero).





A test with a high statistical power means we have a small risk of committing Type II errors (false negatives) so, ideally, we want choose the method/test that has the highest power. In other words, the faster the power $1 - \beta$ increases as you move away from θ_0 the better your method is, with the extreme example being the ideal power curve in the figure.

[1] size is aka. significance level, level, probability of a Type I error, or α.

A crucial ingredient to graphing such a plot for a given method is to be able to compute the value of $\beta(\theta)$ for every $\theta \in \Theta$. In some special cases[2] a closed-form solution can be found for simple distributions and simple parameters θ; see **[C]** for a refresher. Unfortunately, in this course, closed-form solutions are typically unavailable and we need to resort to a *simulation* approach; see **[D]** and **[E]**. For the case $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, the essence of the simulation approach is:

- To evaluate whether the *size* of a test achieves advertised α, sample data under $\theta = \theta_0$ and calculate the proportion of rejections of H_0. This approximates the true probability of rejecting H_0 when it is true. The proportion of rejections should be $\approx \alpha$.
- To evaluate *power*, sample data under some alternative $\theta \neq \theta_0$ and calculate the proportion of rejections of H_0. The approximates the true probability of rejecting H_0 when the alternative is true. The proportion of rejections ρ should be $\approx \beta$, then power is $\approx 1 - \rho$.

Careful, if the actual size of the test is $> \alpha$, then the evaluation of power is flawed. This is because a good test is one that makes $1 - \beta(\theta)$ as large as possible for θ on Θ_1 while satisfying the constraint $1 - \beta(\theta) \leq \alpha$ for all $\theta \in \Theta_0$; see ideal figure above for the case $\Theta_0 = \{\theta : \theta = \theta_0\}$ and $\Theta_1 = \{\theta : \theta \neq \theta_0\}$.

Question 1

We consider a classic multivariate dataset of skull sizes. The data was collected in southeastern and eastern Tibet. Skulls 1-17 were found in one place (Sample 1) and skulls 18-32 were found in another place (Sample 2). On each skull, we have $p = 5$ measurements, all in millimeters. They are (1) `Length`: greatest length of the skull, (2) `Breadth`: greatest horizontal breadth of skull, (3) `Height`: height of the skull, (4) `Fheight`: upper face height, (5) `Fbreadth`: face breadth, between outermost points of cheek bones. An anthropologist would be interested to know if these skulls come from the same population or two different populations.

A standard way to answer this question is to use a multivariate generalisation of the Student's t-statistic called *Hotelling's T^2 statistic* which is given by

$$T^2 = n(\bar{\mathbb{x}} - \mu)\mathbb{S}^{-1}(\bar{\mathbb{x}} - \mu),$$

where $\mathbb{S}$ is the sample covariance, $\bar{\mathbb{x}}$ is the sample mean, and μ is the population mean; see **[F]**. Under the assumption that $\mathbb{x}_1, \ldots, \mathbb{x}_n$ is a sample from $N_p(\mu_0, \Sigma)$, then the distribution of

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

is the non-central F-distribution with parameter p, $n - p$ degrees of freedom, and non-centrality parameter $\Delta = n(\mu - \mu_0)'\Sigma^{-1}(\mu - \mu_0)$. If $\mu = \mu_0$ then the F-distribution is central.

(a) Load the `tibetskull.dat` into R and plot the data in various ways so you get familiar with the data.

[2] Remember this from STAT2001?

(b) To illustrate the Hotelling method, consider the (artificial) example whereby we perform a one-sample test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ where $\mu_0 = (180, 140, 135, 74, 135)'$. Do you reject or accept H_0?

(c) Perform the two-sample test

$$H_0 : \mu_1 = \mu_2 \qquad \text{vs.} \qquad H_1 : \mu_1 \neq \mu_2,$$

where μ_k is the population mean of Sample k. You can consult **[F]** for the derivation of the two-sample test. Do you reject or accept H_0? What could an anthropologist conclude?

Question 2

Let's consider the performance of the Hotelling T^2 test.

(a) Write a function to compute the power of the Hotelling T^2 test using the "closed-form" functions `qf` and `pf` in R. Plot the power function in the case where $\alpha = 0.05$, $p = 10$, $\Sigma = I_p$, $n_1 = 50$, $n_2 = 50$, over the alternative $\mu_1 - \mu_2 = (\delta, \delta, \ldots, \delta)'$ with $-0.5 \leq \delta \leq 0.5$. Add a horizontal line at height α on the same plot.

(b) Write a function to compute the power of the Hotelling T^2 test using a simulation approach under the same assumptions as (a). Using $m = 5000$ simulations for each δ, plot the closed-form and simulation power curves for $0 \leq \delta \leq 0.5$ on the same figure. Do they match? If not, explain why not.

(c) Redo (b) with $m = 10000$. What does it change?

References

[A] Anderson (2003). An Introduction to Multivariate Statistical Analysis. Wiley.

[B] James, Witten, Hastie, Tibshirani (2013). An Introduction to Statistical Learning. Springer.

[C] https://online.stat.psu.edu/stat415/lesson/25/25.2

[D] http://www4.stat.ncsu.edu/~davidian/st810a/simulation_handout.pdf

[E] https://stats.stackexchange.com/a/40874

[F] https://en.wikipedia.org/wiki/Hotelling%27s_T-squared_distribution