# Problem Set 4–STAT 7200

Unless otherwise stated, the significance level is $\alpha = 0.05$.

1. Let $(\mathbf{Y}_1, \mathbf{X}_1), \ldots, (\mathbf{Y}_n, \mathbf{X}_n)$ be a paired random sample. Construct a test statistic for testing the equality of the means, i.e. $H_0 : E(\mathbf{Y}_i) = E(\mathbf{X}_i)$. Give the distribution of your test statistic under the null hypothesis, and describe the distributional assumptions you are making.

2. Using the result of Problem **??** on the `gapminder` dataset, test the null hypothesis of no change in infant mortality, life expectancy and fertility from 2012 to 2013 (only keep countries with complete observations for both years). Construct the simultaneous confidence intervals and the Bonferroni-corrected univariate confidence intervals. Discuss the results.

3. Investigate the multivariate normality assumption for the dataset in Problem **??**. Perform a permutation test and compare the results with what you obtained earlier.

4. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n \sim N_p(\mu, \Sigma)$, with $\Sigma$ positive definite, and let $\bar{\mathbf{Y}}$ and $S_n$ be the sample mean and covariance, respectively. Let $C$ be a $(p-1) \times p$ matrix of rank $p-1$ such that $C\mathbf{1} = 0$, i.e. the rows of $C$ sum to zero. Prove the following:
   (a) The following null hypotheses are equivalent:
   $$H_0 : C\mu = 0 \Leftrightarrow H_0 : \mu_1 = \cdots = \mu_p.$$
   (b) Any $p-1$ columns of $C$ are linearly independent, which implies that we can write
   $$C = A(I_{p-1}, -\mathbf{1}),$$
   for some invertible $(p-1) \times (p-1)$ matrix $A$.
   (c) Using part (b), show that the test statistic
   $$T_2 = n(C\bar{\mathbf{Y}})^T (C S_n C^T)^{-1} (C\bar{\mathbf{Y}})$$
   does not depend on $C$.

(d) Under the null hypothesis of part (a), the distribution of $T^2$ is

$$T^2 \sim \frac{(n-1)(p-1)}{n-p+1} F_\alpha(p-1, n-p+1).$$

5. Using the result of Problem **??** on the Ramus dataset, perform a hypothesis test that the ramus lengths are constant over time. Construct the simultaneous confidence intervals and the Bonferroni-corrected univariate confidence intervals. Discuss the results.

6. Let $GL_p$ be the set of invertible $p \times p$ matrices.[1] Any matrix $A \in GL_p$ induces a transformation on the data $\mathbf{Y}_i \mapsto A\mathbf{Y}_i$, which also induces a transformation on the sample statistics
$$(\bar{\mathbf{Y}}, S_n) \mapsto (A\bar{\mathbf{Y}}, AS_nA^T),$$
and on the population parameter
$$(\bar{\mu}, \Sigma) \mapsto (A\mu, A\Sigma A^T).$$
We say a test statistic $f(\bar{\mathbf{Y}}, S_n)$ is *invariant* if it yields the same value on the transformed data:
$$f(\bar{\mathbf{Y}}, S_n) = f(A\bar{\mathbf{Y}}, AS_nA^T), \qquad \text{for all } A \in GL_p.$$
By choosing an appropriate $A$, show that an invariant test statistic satisfies
$$f(\bar{\mathbf{Y}}, S_n) = f((\bar{\mathbf{Y}}^T S_n^{-1} \bar{\mathbf{Y}})^{1/2}, I).$$
Conclude that an invariant test statistic depends on the data only through the statistic $T^2 = n\bar{\mathbf{Y}}^T S_n^{-1} \bar{\mathbf{Y}}$.

[1] This set is actually called the *general linear group*, and it is ubiquitous in mathematics.