STATG006: Exercise Sheet #2

The exercises in this sheet focus on the basics of hypothesis testing. Some questions, marked with the indication "(COMPUTER IMPLEMENTATION)*" may require programming, with* R *being the suggested language. Even if* R *programming is not necessary for the STATG006 exams, it is strongly encouraged you attempt these exercises even if your knowledge of* R *is currently incipient (it is assumed that you already know or are currently learning fundamentals of programming, a much more general skill that knowing a particular programming language). Part of being a Data Scientist is teaching yourself how to use different tools. Use your stay at UCL to have the chance of discussing your problem solving skills during class time and office hours.*

1. If you toss a coin 1,000 times and observe 570 heads, how would you assess the claim that the coin is fair?

2. Now say you toss 1,000 different coins once each, which you assume are identically distributed. Would you change your test for fairness? Would you think of ways in which the test could fail to behave as expected?

3. *(Adapted from Wasserman, Chapter 10)* Let your data be $X_1, \ldots X_n \sim Poisson(\lambda)$. Find a size[1] α Wald test for

   $$H_0 : \lambda = \lambda_0 \quad \text{versus} \quad H_1 : \lambda \neq \lambda_0.$$

   (COMPUTER IMPLEMENTATION) Now, let us experiment with the precision of the Wald test. Let $\lambda_0 = 1, n = 20$ and $\alpha = 0.05$. Simulate[2] $X_1, \ldots, X_n \sim Poisson(\lambda_0)$. Do the Wald test. Repeat this many times and compute the frequency in which the null is rejected. Compare that to the advertised error rate of 0.05.

4. *(Adapted from Wasserman, Chapter 10)* There is a theory that people can postpone their death until after an important event. Here are the numbers, in a particular year, of elderly Jewish and Chinese women who died just before and after the Chinese Harvest Moon Festival.

[1] The **size** of a test is just the worst-case scenario of a Type I error, where the worst-case is the defined as the highest Type I error probability among all values of the parameter of interest compatible with the null. The notion of **level** of a test, as we saw in the slides, is just a number that upper bounds the size of the test. If the null is a single point, as it is here and in most of our examples, then size and the level of a test are the same.

[2] R programmers: do you want to know how to sample from a Poisson? Go to `http://bfy.tw/82L5`. (Notice: this is a tongue-in-cheek link, a friendly nudge that you should try to be as independent as possible.)

Week	Chinese	Jewish
-2	55	141
-1	33	145
1	70	139
2	49	161

Compare the two mortality patterns using a hypothesis test, explaining your reasoning.

5. *(Adapted from Wasserman, Chapter 10)* The following table summarizes data from a double-blind experiment that aims at comparing particular drugs for nausea reduction against a placebo. Assume each patient is independently assigned one of the treatment groups, or the placebo.

	Number of patients	Incidence of Nausea
Placebo	80	45
Chlorpromazine	75	26
Dimenhydrinate	85	52
Pentobarbital (100 mg)	67	35
Pentobarbital (150 mg)	85	37

Test each drug versus the placebo at the 5 per cent level. Report what the result would be under a Bonferroni adjustment, and how the interpretation changes.

(Computer implementation) Let us simulate what happens if we ignore multiple adjustment. Pretend that each of the p-values on the four tests above are independent random variables[3]. Perform a simulation of the distribution of the minimum of the these four p-values. Which lessons can you conclude out of it? (Notice: it is not that hard to derive analytically what the distribution of this minimum should be. Do you want to give it a shot?)

6. For a sample size $n = 10, 50, 100$, consider the test of $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ for the mean parameter μ of a sample following a Gaussian distribution $N(\mu, \sigma^2)$, where σ^2 is known and the level of the test is $\alpha = 0.05$. Write down the power of this test as a function of the true mean μ, whatever that is. How would this change if H_0 was $\mu \leq 0$?

(Computer implementation) Suppose you could not write down an intelligible formula for this power curve, and instead a numerical procedure would needed to be called for every value of μ. Write a computer program to plot the curve as a function of the true μ_0. Say, for $-3 \leq \mu_0 \leq 3$.

7. This is a classic example using Gregor Mendel's tests of his theory of heredity. Mendel bred four different types of peas, starting with round yellow seeds and wrinkled green seeds. Each pea could result in one of four categories: round yellow, wrinkled yellow,

[3] In reality, they are not as they share the common placebo group. This illustrates why methods like Bonferroni are used, as they do not depend on how different p-values are associated.

round green and wrinkled green. Mendel's theory dictates that these categories follow a discrete distribution with respective probabilities

$$p_0 = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right).$$

His experiment had a sample size $n = 556$, where the observed counts were[4] $X = (315, 101, 108, 32)$. We want to test whether this data falsify or not the theory.

For that, first calculate Pearson's χ^2 statistic

$$T \equiv \sum_{j=1}^{4} \frac{(X_j - E_j)^2}{E_j},$$

where X_j is the count data for category j, and E_j is the expected count under the null $H_0 : p = p_0$, with p being the distribution parameter vector of the multinomial[5].

If we have k categories, T will have a chi-squared distribution with $k-1$ degrees of freedom ("degrees of freedom" is just the fancy name given to the parameter of the chi-squared). Describe how you would use this chi-squared statistic to test Mendel's theory.

8. Now consider the twin data from the slides of Chapter 2. Write down the explicit form of Pearson's chi-squared statistic under the null that states that A_1 and D_2 are independent[6].

9. More on the chi-squared. Let us create a simple goodness-of-fit test on whether the data follows a particular $N(\mu, \sigma^2)$. To do that, let us partition the real line in a set of disjoint intervals $I_1, I_2, \ldots, I_k$. Within each interval $[I_j, I_{j+1}]$ we observe N_j data points. Describe how you would calculate the chi-squared statistic.

10. (Computer implementation) Implement the test above[7], knowing that the distribution of the statistic is χ^2_{k-1-s}, where s is the number of assumed parameters[8] in the model (2, in our Gaussian case – we are assuming a particular mean and a particular variance are given). Define the intervals according to quantiles: for instance, we can divide the real line so that the first interval can be $(-\infty, q_{0.25}]$, where $q_{0.25}$ is the 0.25 quantile of the target Gaussian; the consecutive intervals can be $[q_{0.25}, q_{0.50}], [q_{0.50}, q_{0.75}], [q_{0.75}, +\infty)$.

11. (Computer implementation) This exercise has nothing to do with statistical inference, but can give you some insights about the claim, given in class, of the link between the pdf and a histogram. Using your code from the previous question, write a program

[4] Notice these counts follow a **multinomial distribution**, where n and p_0 are the fixed parameters. If the sample size n was considered to be random, then this would not be a multinomial random vector anymore!

[5] If you have difficulties to figure out what E_j is, think of the Bernoulli case: if I toss a coin n times, what is the expected value of the sum $Y^{(1)} + \cdots + Y^{(n)}$, if each $Y^{(i)}$ is i.i.d. Bernoulli with parameter θ?

[6] Recall: $P(A_1 = a, D_2 = d) = P(A_1 = a)P(D_2 = d)$ under the null.

[7] R hint: function `pnorm` gives you the Gaussian cdf.

[8] If we had *estimated* parameters instead of assumed ones, the distribution would not be χ^2_{k-1-s} anymore, but the resulting p-value would be an approximate lower bound of the true p-value. See Wasserman, p. 169 for instance.

that draws a piecewise constant approximation of a distribution (use the Gaussian of the previous exercise) that corresponds the probability of a point falling in pre-specified interval $[I_j, I_{j+1}]$ for $j = 1, 2, \ldots, k - 1$. It is up to you to define these intervals, but make them in a way that it is easy to see what happens when we increase the number of intervals.

12. The `lead.dat` data in the Moodle page is a study D. Morton and collaborators (1982, American Journal of Epidemiology, No. 115, 549–555) on the concentration of lead in the blood of children whose parents worked in a factory where lead was used in making batteries. The Treatment column contains the measurement of 33 of such children, being the concentration of lead in $\mu g/dl$ of whole blood samples. The Control column contains the measurement of a corresponding "matched" control child that is similar to the treated child in ways that we will omit for simplicity. How would you set up a hypothesis test here? The goal is to assess evidence that children of the two types do have different lead concentrations in their blood. Would you choose a t-test, a Wald test, a paired t-test, a Wilcoxon rank sum test? Under which conditions? Also, would you test whether a particular summary of the distribution of the treated is different from the summary of the control, or whether one is larger than the other?

    (Computer implementation) Perform the test of your choice using a programming language such as R.

13. This example is by C. N. Morris, published in the March 1987 issue of the Journal of the American Statistical Association (vol. 82, No. 397, 131–133). Two candidates, Mr. Allen and Mr Baker, are running for a particular public post. Mr Allen would like to know, for $\theta \equiv$ the proportion of voters favoring him today, whether $\theta < 0.5$ or $H_1 : \theta > 0.5$. Assume we have a sample of size n, and the random number of votes Y for Mr Allen follow approximately a Binomial (n, θ), and that the sample average $\hat{\theta} \equiv Y/n$ is approximately such that $\hat{\theta} \sim N(\theta, 0.25/n)$ under the null. Which of the three outcomes would be the most encouraging to Mr Allen and why?

    (a) $Y = 15, n = 20, \hat{\theta} = 0.75$;

    (b) $Y = 115, n = 200, \hat{\theta} = 0.575$;

    (c) $Y = 1046, n = 2000, \hat{\theta} = 0.523$;