STATG006: Solutions to Exercise Sheet #2

The exercises in this sheet focus on the basics of hypothesis testing. As before, we provide solutions, sometimes detailed, sometimes a sketch that should point you to the complete solution.

1. "Fairness" is typically interpreted as a coin that has equal probability of tails or heads. In what follows, we are implicitly assuming that the coin tosses were independent and identically distributed!

   Let θ denote the probability of heads, we can set $H_0 : \theta = 0.5$ with the alternative being the usual "H_0 is false", that is, $H_1 : \theta \neq 0$. If X_i is a random variable encoding as 1 the event of a toss i showing up heads, 0 otherwise, our statistic is set to $Y = \sum_{i=1}^{1000} X_i$, which is known to follow a Binomial with parameters $n = 1000$ and $\theta = 0.5$ under the null.

   Now, if we want a **two-tailed test**, as implied by the null hypothesis, I will look at the $\alpha/2$ quantile and $1 - \alpha/2$ quantile of the distribution of Y under H_0. Using a standard package to calculate the corresponding quantiles at $\alpha = 0.05$, we get that the critical region is $[0, 469] \cup [531, 1000]$. 570 is in the critical region, so by this logic we reject that the coin is fair. How would you do it with a CLT approximation instead?

2. Using the same notation as before, the distribution of my test statistic still is $Bin(1000, 0.5)$ under H_0, again assuming the i.i.d. condition. Notice how different physical processes, when summarized by probability and statistics, end up being equivalent. Concerning failure of the test, we *will* get the distribution of Y wrong if the coin tosses are not independent. For instance, if you think of an unusual coin tossing process that first glues all coins to a large transparent disk, tossing the whole dis at once in a "fair" way. What happens in this case?

3. Let us use the sample average $\bar{X}$ as an estimate $\hat{\lambda}$ of the parameter of this Poisson. We need to look up the standard deviation of this sample average by looking up the variance of a $Poisson(\lambda)$. This variance is also λ. Hence, we know that the variance of the sample average will be λ/n, so we can plug in $\sqrt{\hat{\lambda}/n}$ as the estimated standard deviation. The Wald statistic

$$W \equiv \frac{\bar{X} - \lambda_0}{\sqrt{\hat{\lambda}/n}}$$

   approximately follows a $N(0, 1)$. So for this two-sided test I would repeat the same reasoning as in Question 1, where the quantiles now come from a standard Gaussian.

What if we ignored the fact that X came from a Poisson and plugged-in the empirical standard deviation without this assumption?

4. There are different ways of thinking about this. One is to split the time period into two bins, "before" (weeks -2 and -1) and "after" (weeks 1 and 2) the festival, treat the number of deaths fixed, and assume as random the period in which the deaths took place. This would give the following 2 x 2 contingency table

Period	Chinese	Jewish
1	88	286
2	110	300

We can now think of X and Y as "deaths in period 1" for each of the two groups. Under assumptions of independence of the samples, each follow a binomial distribution: $X \sim Binomial(198, p_1)$ and $Y \sim Binomial(586, p_2)$. The null hypothesis here is not that the two periods have a similar probability of death, but that this period should not have a particular meaningful impact across groups. So we set $H_0 : p_1 - p_2 = 0$ (with $H_1 : p_1 - p_2 \neq 0$). As it seems safe to assume the two populations are independent, we can use the difference of sample averages $\hat{p}_1 - \hat{p}_2$ as our test statistic. From the known result that the variance of the sum of two independent variables is the sum of the variances, its standard deviation is

$$\hat{se} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{198} + \frac{\hat{p}_2(1-\hat{p}_2)}{586}}.$$

Now you have the pieces to do a Wald test here. For a totally different approach, see the original paper by Phillips and Smith (1980, "Postponement of death until symbolically meaningful occasions", Journal of the American Medical Association, vol. 263, 1947–1951).

5. As the patients in each group are independent, the number of patients with nausea incidence can be assumed to follow independent binomial distributions. So to test the comparison of the placebo group against the chlorpromazine group, the reasoning stays the same as in the previous question.

What changes under the Bonferroni adjustment is that the level of the test is adjusted to be 0.05/4. This changes the critical regions and possibly the respective conclusions (notice that the *size* of the test is probably lower than 0.05, as they the tests are not independent- why not?).

What if we wanted to control for at least one test in this group being a false positive? In this case it is enough to characterize the distribution of the minimum of the four p-values. This might be hard if the tests are dependent, but the questions asks for the simpler case where they are independent. If we have four independent random variables X_1, X_2, X_3, X_4, what is the distribution of $Y = \min\{X_1, X_2, X_3, X_4\}$? The easiest way of finding this is by looking at the cdf of the corresponding events as encoded by Y and by X:

$$P(Y \geq y) = P(X_1 \geq y \ and \ X_2 \geq y \ and \ X_3 \geq y \ and \ X_4 \geq y) = \prod_{i=1}^{4} P(X_i \geq y)$$

and as such $P(Y \leq y) = 1 - P(Y \geq y) = 1 - \prod_{i=1}^{4}(1 - P(X_i \leq y))$. Now it is a matter of taking its derivative with respect to y. For the particular question asked in the exercise sheet, recall that the distribution of a p-value under H_0 is $U(0, 1)$.

6. First we need to find the critical region R. For $n = 10$, $\alpha = 0.05$, and a two-tailed test, the region is the quantile corresponding to a test statistic T (the sample average) following a $N(0, \sigma^2/10)$, which is approximately the region $(-\infty, -1.96\sqrt{\sigma^2/10}] \cup [1.96\sqrt{\sigma^2/10}, \infty)$. For an alternative $\mu \neq 0$, the power function boils down to $\beta(\mu) = P(T \in R) = P(T \leq -1.96\sqrt{\sigma^2/10}) + P(T \geq 1.96\sqrt{\sigma^2/10})$, where $T \sim N(\mu, \sigma^2/10)$. I leave the formula of this function to you, which should be expressed as the sum of two integrals (which cannot be solved analytically). It is common to express functions like this as functions of "$\Phi(\cdot)$", the cdf of a standard Gaussian. This means $P(T \leq -1.96\sqrt{\sigma^2/10}) = \Phi(\mu - 1.96\sqrt{\sigma^2/10})$ for $T \sim N(\mu, \sigma^2/10)$ etc.

   If H_0 is now $\mu \leq 0$, then we have that the region changes to $[1.65\sqrt{\sigma^2/10}, \infty)$, with the corresponding change in the power function.

7. If $X \sim Bin(n, p)$, then $E[X] = np$ (why?). So T can be written as

$$T = \frac{(315 - 556 \times 9/16)^2}{556 \times 9/16} + \text{ etc.} = 0.47$$

   We then need to find the critical region R for a χ^2_3 so that for each t in R we have $P(\chi^2_3 > t) \leq 0.05$. Using R or any other statistical package we find that this region is $[7.815, \infty)$, so that the model is *not* rejected. The p-value for the null hypothesis is actually very high, find it yourself.

8. We need to count the data for all combinations of A_1 and D_2. For instance, for $A_1 = 0$ and $D_2 = 0$, we see that the table give us $288 + 15 + 92 + 7$ (make sure you know how to read a contingency table). Let's call this n_{00} etc. Under independence, we have that $P(A_1 = a, D_2 = d) = P(A_1 = a)P(D_2 = d)$. This means that under H_0, the expected counts are $E[N_{ad}] = n\hat{p}_{A_1}(a)\hat{p}_{D_2}(d)$, where n is the total sample size and $\hat{p}_{A_1}(a)$ is the corresponding empirical frequency of the event $A_1 = a$. With this, you have the information to write the desired statistic.

9. If we partition the real line, then the probability of a point X falling into the interval $[I_j, I_{j+1}]$, according to the model, is

$$P(X \in [I_j, I_{j+1}]) = \int_{I_j}^{I_{j+1}} p(x; \mu, \sigma^2)\ dx = F(I_{j+1}; \mu, \sigma^2) - F(I_j; \mu, \sigma^2),$$

where $p(\cdot)$, $F(\cdot)$ are the corresponding Gaussian pdf/cdf. So if we have a sample size of n, the expected count of points falling in that particular bin is

$$E_j = n \times (F(I_{j+1}; \mu, \sigma^2) - F(I_j; \mu, \sigma^2)).$$

The rest is a matter of putting together the actual counts and expected counts in the χ^2 statistic, contrasting N_j to E_j.

10. See comments in EX2.R in the Moodle page.

11. See comments in EX2.R in the Moodle page.

12. It boils down to whether assumptions we can realistically use. A qqplot against normality reveals the columns are not quite Gaussian, but might be close enough that a Wald test can work well enough without the need to assume normality and the use of a t-test (a test of Gaussianity, the Shapiro-Wilk test, does not reject normality at 0.05). However, the choice of controls can be assumed to come from having observed the treatment children, so they are not independent. A **paired t-test** could work here. The rationale boils down as explained in class, that if we take the difference of the columns we have again some (Gaussian) random variable with a mean and standard deviation estimated from the data. The rest proceeds as before.

    If the Gaussianity assumption is not reasonable, we can use a nonparametric test. In class, the example we used as the Wilcoxon test. Notice however that in many implementations, this test does not deal properly with ties.

13. The p-values in all three situations are approximately the same, using as variance the variance of the null hypothesis $H_0 : \theta = 0.5$. Let the test statistic be $\hat{\theta}$, which is $N(0.5, 0.25/n)$, following the assumptions. The p-value is for the one-sided test $P(T > t; H_0)$, which for case (a) is for example 0.01. All three cases have very similar p-values.

    What about power? The critical regions for the tests under size 0.05 are approximately ≥ 0.68, ≥ 0.56 and ≥ 0.52 for (a), (b) and (c) respectively. If we plot the power curves, though, we see we get situation (c) presenting a much better control of Type II than (b), and (b) much better than (a).