

COMPLEMENTARY EXERCISES WITH STATISTICAL INFERENCE AND ANALYSIS OF ASSOCIATIONS

EX 1

In a sample of 40 patients, the mean of White Blood Count was $2.9 (\times 1000/\text{ml}^3)$ and the variance was 0.336. Provide a point estimate and an interval estimate by 90% Confidence Interval for the expected value of WBC (*i.e. mean of the population of all patients affected by that disease*).

EX 2

A group of 60 people is tested for the presence of a certain genetic abnormality; this is detected in 14 people. What is the estimate of the prevalence of this genetic abnormality? Provide an interval estimate with 95% confidence level.

EX 3

The distribution of Body Mass Index ($\text{BMI} = \text{weight (kg)} / [\text{height (m)}]^2$) in a certain population is approximately Normal, with standard deviation 2.2. In a group of 26 people chosen at random from that population, the mean BMI was 24.2. Compute a point and an interval (95%CI) estimate of mean BMI μ , and give an interpretation.

Afterwards, assuming that the unknown mean of the population is exactly equal to the point estimate, provide a range of values for BMI such that 95% of the population is included in that range.

EX 4

Among 80 patients treated with a new drug for a certain disease, it was observed a low average Count of Red Blood (equal to $3.4 \times 10^3/\text{ml}$), which is inferior to the reference value for that population, which is $5 \times 10^3/\text{ml}$. Knowing that the standard deviation in the sample was $4.5 \times 10^3/\text{ml}$, we want to verify whether the treatment can induce low RBC. Make a statistical hypothesis test, using both the method of computing the p-value, both the method of rejection regions.

EX 5

Among the same 80 patients of EX 4, we observed also the onset of dermatitis in 25 of them. In fact dermatitis is a side effect of the disease, which usually occurs in 10% of patients. Should we conclude that the treatment modifies the incidence of dermatitis? Make a two-sided test both computing the p-value and using the rejection regions at 5% level. Afterward, return a measure of the effect of treatment on the probability of onset of dermatitis by a Risk Ratio.

EX 6

A physician specialized in nutrition devises a new kind of diet (A) for losing weight. He wishes to compare the efficacy of this diet versus the kind of diet he used to prescribe (B). He has the results for two groups of patients, 39 treated with the old diet B, who lost on average 2.9 kg (dev. std.=1.2kg), and 34 patients treated with the new diet A, who had an average weight loss equal to 3.5 kg (std=1.1kg). Assuming that weight loss follows a Normal distribution and that the variances in the populations are the same, do a 2-sided hypothesis test at 5% level. Is it correct that diet A is better than diet B? what should be considered regarding the two samples in order to be sure of the conclusion?

EX 7

A group of 7 subjects is treated for hypertension. The diastolic blood pressure (minimum) is measured before and after a certain number of days of treatment: results are reported in the table below. The objective of the study is to assess whether the treatment has a beneficial effect. Can we solve the problem with the methods we know?

Subject	1	2	3	4	5	6	7
Before	118	100	108	97	120	102	104
After	100	105	89	78	96	97	88

EX 8

186 patients were treated for a certain disease; 122 had treatment A, the other 64 had treatment B. In group A, 37 subjects responded to therapy, in group B, there were 32 responders. The data are also reported in the contingency table below. Compute the response percentages and compare them using the Risk Ratio with group A as baseline.

To assess the significance of the difference we should use appropriate tests for frequencies, but we can also compute a measure of association for the contingency table and make a test for the presence of a relation. Use $\alpha=5\%$.

Treatment	Response		Tot
	no	yes	
A	85	37	122
B	32	32	64
tot	117	69	186

EX 9

Given the following series of data on Height and Weight of 8 patients, compute a measure of association, then identify the straight line that describes the dependence of Weight on Height, and predict the weight for a patient 165 cm tall. Observing that patient id=1 is 165 cm tall and weighs 55.2 kg, make remarks about the consistency between the model and the data.

id	Height, cm	Weight, kg
1	165	55.2
2	157	48.1
3	168	65.6
4	178	72.0
5	171	73.5
6	182	77.9
7	182	80.3
8	153	49.4

EX 10

The following table reports the results of a logistic regression model (a kind of linear model for dichotomous outcomes), i.e. the estimated effect of a series of factors on the risk of short-term progression in a specific kind of cancer. The effect is estimated in terms of odds ratio, with 95% Confidence Interval, and the last column reports the p-value for its significance. For each of the following questions, say if the answer given is True or False, and if False make the appropriate correction.

Factor	OR	95% CI		p-value
Treatment B vs. A	0.47	0.31	0.72	0.004280
Prior treatment	1.44	0.88	2.36	0.142867
High. Beta2 Microglob	2.36	1.35	4.12	0.002690
DEL13 abnormality	2.03	1.29	3.17	0.000518

- Does treatment have an impact of the risk of progression?
- What is the effect of a prior treatment?

- c) Which is the most important risk factor?
- d) What is the effect of high Beta2 microglobuline?

SOLUTIONS

EX 1

The point estimate of the mean of the population is the arithmetic mean computed in the sample, thus $\hat{\mu} = 2.9$; in fact the sample mean $\bar{x}$ is an unbiased estimator of μ with several other properties (efficiency, consistency etc). The requested interval estimate is a 90% Confidence Interval, we can use the formula we know from the course replacing the known variance σ^2 (parameter of the population) with the estimated variance from the sample s^2 – in fact the sample size is “large”, $n=40$.

$s^2=0.336$ (and thus $s=0.579655\approx 0.58$).

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(2.9 - 1.64 \cdot \frac{0.58}{\sqrt{40}}, 2.9 + 1.64 \cdot \frac{0.58}{\sqrt{40}} \right) = (2.75, 3.05)$$

(Notice that we use the quantile 1.64, which leaves in the tail a 5% probability and thus leaves externally to the interval a 10% probability; usually the confidence interval is requested at 95% confidence level, thus usually the quantile used is 1.96, which leaves 2.5% in the tail and overall 5% externally)

Interpretation: for this kind of patients we expect on average a WBC value between 2.75 and 3.05. Pay attention: this is NOT the range of WBC values, the range is larger; this is the range for the average WBC value. *See EX 3 and use this reasoning in that exercise.*

EX 2

The point estimate of the prevalence in the population (or in other terms of the frequency of the “disease” in the population, or of the probability of having the “disease” – here “disease” stays for a genetic abnormality) is just the proportion (percentage) observed in the sample, which is $\hat{\pi} = 14/60=23\%$. Since we have a large sample ($n=60$) for inference purposes we can assume that $\hat{\pi}$ behaves like a sample mean for a quantitative continuous random variable X . Thus:

$$\left(\hat{\pi} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}, \hat{\pi} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right) = \left(0.23 - 1.96 \sqrt{\frac{0.23 \cdot 0.77}{60}}, 0.23 + 1.96 \sqrt{\frac{0.23 \cdot 0.77}{60}} \right) = (0.13, 0.34)$$

Thus the prevalence of the genetic abnormality in the population ranges between 13% and 34%.

EX 3

According to the text of the exercise we can assume that BMI is distributed as a Normal variable and that we know the variance of the population (the text provides the standard deviation, which is its squared root). Thus even if the sample is small ($n=26$) we can use our formula for the 95% CI for the interval estimate requested – being the point estimate of μ simply equal to 24.2.

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = \left(24.2 - 1.96 \cdot \frac{2.2}{\sqrt{26}}, 24.2 + 1.96 \cdot \frac{2.2}{\sqrt{26}} \right) = (23.35, 24.05)$$

Thus for a person chosen at random from the same population used to extract the observed sample, we expect on average a BMI value equal to 24.2, or going from 23.35 and 24.05.

To be more specific, this is the predicted value for an “average individual”, but we know that there is some variation around the mean of a variable (BMI)– variability expressed by the standard deviation of the population, 2.2. To compute a range of possible values for BMI we must thus use

the properties of a Normal distribution, assuming (as the last question indicates) that it has mean equal to 24.2 and st. dev. equal to 2.2.

With $N(\mu=24.2, \sigma=2.2)$, and knowing that for each Normal curve 95% of values are included between $\mu - 1.96 \cdot \sigma$ and $\mu + 1.96 \cdot \sigma$, we conclude that 95% of possible BMI values of individuals of that population are included between 19.89 and 28.51. (A larger range and the one that we usually consider as the largest, i.e. from the minimum to the maximum, is the one that has a radius equal to 3σ around the mean, it provides almost all values – 99.7% – of the population: 17.6,30.8)

EX 4

First step is writing the hypotheses of the test. The hypotheses regard a parameter of a population, in this case the mean μ . The null hypothesis assumes a “neutral” knowledge, in this case it assumes that μ is equal to the reference value (known from the literature, from prior knowledge, etc), which is 5 (being the unit of measurement $10^3/\text{ml}$). The alternative hypothesis is our experimental hypothesis, a deviation from the neutral knowledge; usually it is a two-sided hypothesis ($\mu \neq 5$) but in this case we are interested in proving that the treatment reduces μ , and thus we will consider $H_1: \mu < 5$.

Notice: we make this choice here to illustrate the procedure for a one-sided test, but in real life it is usually NOT appropriate to do one-sided tests, or it is appropriate as long as the significance level is halved, for example either the test is 2-sided with $\alpha=0.05$ or the test is one-sided with $\alpha=2.5\%$: both provide the same tail as rejection region – and give the same result, provided that the test statistics falls on the rejection side – see below for further explanations.

So our hypotheses are:

$$H_0: \mu=5 \quad \text{vs} \quad H_1: \mu<5$$

Notice that when we write the hypotheses of a test we NEVER use the values observed in the sample of observations.

If the test for the mean of one population is a t-test, we can use our formula since we are in one of the situations when it is allowed: we have a large sample ($n=80$); the test statistic is:

$$t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{3.4 - 5}{4.5/\sqrt{80}} = -3.18$$

This value falls far away from the average value we expect under the null hypothesis (the value 0), still we cannot draw a conclusion since whether “it is really far” or not depends on what is the variability of the population (under the null hypothesis). That’s why we construct rejection regions or compute a p-value: because “far” means “odd, unexpected, wired” in the context of statistical tests.

The rejection region for $\alpha=0.025$ includes all values of t below the threshold -1.96, thus our test rejects the null hypothesis at 2.5% level (as said, for a one-sided test of level α we use the same rejection tail of a two-sided test of level 2α ; if we wanted to keep $\alpha=5\%$ and a one-sided test, we should have considered the threshold leaving a tail with probability 5%, which is -1.64, the same of a two-sided test with $\alpha=10\%$).

The approach of computing the p-value requires to compute the probability that the test statistics falls below the observed value t under the null hypothesis. In other terms, in this exercise we must compute the area under a standard Normal in the lower tail delimited by $z = -3.18$. Since the table has only positive z , we compute the area as:

$$p = 1 - \Phi(3.18) = 1 - 0.999 = 0.001$$

(Notice: if the test was two-sided, this area should have been multiplied by 2 to find the p-value, as we would have had 2 tails: $p=0.002$)

This is our p-value. It is very small; it says that IF the null hypothesis was true and thus the population of patients treated with the drug had “normal” RBC with expected value 5, in the sample we would have observed an average equal to 3.4 with 0.001 probability, i.e. in 1 out of 1,000 samples. We believe this is “too wired” to happen under the null hypothesis, and we reject it.

EX 5

This exercise is a replicate of the previous one, since being interested in a parameter which is a probability (of onset of dermatitis, π) and having a big sample we can use the t-test as if the parameter of interest was the mean of a continuous variable.

The hypotheses are based on the reference value, the usual probability of onset of dermatitis, which is 10%, the alternative is 2-sided as requested by the text:

$$H_0: \pi = 0.10 \quad \text{vs} \quad H_1: \pi \neq 0.10$$

(again: notice that we don't use the percentage observed in the sample to write the test hypotheses) From the sample we obtain the point estimate $\hat{\pi} = 0.31$, and from this its standard error

$$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.052; \text{ the test statistic is } 4.10.$$

This value is clearly very far from 0 which is expected under H_0 , in particular it is below 1.96 (threshold for $\alpha=5\%$). If we try to compute the p-value, we get $p=0$ (*in publication the notation usually used for $p=0$ is $p<0.0001$*).

Thus the test rejects H_0 and we say that the data show a significant difference of the observed incidence from the reference value. According to this study, there is a significant increase of the incidence of dermatitis among patients treated with the drug. The increase is measured by $RR=0.31/0.1 \approx 3$: the treatment induces an increase by 3 times of the incidence.

EX 6

Let us set as first the hypotheses of the test:

$$H_0: \delta = \mu_A - \mu_B = 0 \quad \text{vs} \quad H_1: \delta = \mu_A - \mu_B \neq 0$$

The two groups are independent since the subjects of group A are not the same as group B and they don't share any feature (cases when the two groups might share relevant features would be for example when comparing individuals of the same family like siblings – who might not live together and still share genetic factors or environmental elements during infancy – or partners – who might share environmental factors, diet, habits etc).

The comparison of means of continuous factors with ample sample sizes and same population variance (stated in the text) can be done by the t-test:

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(34-1)1.1^2 + (39-1)1.2^2}{34 + 39 - 2}} = 1.156$$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{3.5 - 2.9}{1.156 \sqrt{\frac{1}{34} + \frac{1}{39}}} = 2.22$$

Using the approach of rejection regions, we can reject the null hypothesis of no difference between the mean of A and the mean of B at 5% significance level (in fact the threshold is 1.96).

Using the calculation of the p-value for assessing the significance of the observed difference, we look up in the table of the standard Normal the value corresponding to $z = 2.22$: the area is 0.987; thus the area in one tail is $1 - 0.987 = 0.013$, and the total probability of the tails (2-sided test) is $p = 0.026$.

Thus the observed difference is pretty significant, the data allow to reject H_0 and support the hypothesis that diet A gives larger weight loss than diet B. This is a “statistical” association which anyway might be not casual; there can be a causal relationship between changed diet and larger weight loss IF the two groups of patients were similar with respect to all characteristics potentially influential on weight loss, like starting weight, gender, age, physical activity and so on, as well as if the methods and tools for measuring weight loss were the same in the two groups. In other terms any relevant difference with respect to one of these other causal factors might have determined the observed difference: the relationship between diet and weight loss might be confounded by other factors. To conclude, there is a statistically significant difference in weight loss between group A and group B, an analysis of potential confounders is necessary to claim that diet A is more effective than diet B.

EX 7

A general idea might be to compare the mean of the values “before” with the mean of the values “after”. The first problem we encounter is that the two samples of data are not independent: these data are “paired”, in fact they are measured on the same individual. Thus we cannot make a t-test like in EX 6. There is a specific t-test for paired data. Basically, it does the following:

We consider the variation of blood pressure, computed as Value After-Value Before. We get a sample of 7 values.

Subject	1	2	3	4	5	6	7
Reduction	18	-5	19	19	24	5	16

The test regards the mean of the population from which we extracted the 7 values, δ : in case of no effect, this mean is 0, otherwise it is different than 0:

$$H_0: \delta = 0 \quad \text{vs.} \quad H_1: \delta \neq 0$$

We can solve this problem using a t-test for one sample. Anyway we have a further problem, since we have a small sample and we don’t know the variance of the population. Thus we cannot solve the problem with the tools we know. The test can be applied using the tables of another distribution called “Student’s T”.

EX 8

The response percentages are computed as row-profiles, i.e. considering the conditional distributions of Response given the two different levels of the variable Treatment: $p_A = 37/122 = 0.3$, $p_B = 32/64 = 0.5$. The Risk Ratio is $0.5/0.3 = 1.65$: the probability of response with B is increased by 65% with respect to A.

The measure of association and the required test are the Chi-squared. The null hypothesis H_0 is that there is no association between Treatment and Response, i.e. that in a population the percentage of response with A and B are equal. The alternative hypothesis is that there is an association, i.e. that the two percentages of response in the population are different according to which treatment is given A or B.

The Chi-squared is based on a measure of the distance between the observed frequencies and the frequencies we expect under the null hypothesis, i.e. assuming that H_0 is true:

expected freq.:

Treatment	Response	
	no	yes
A	76.74	45.26
B	40.26	23.74

differences (sum to obtain the test statistic):

	Response	
Treatment	no	yes
A	0.889	1.507
B	1.694	2.872

Chi-squared=6.962, it is higher than the threshold for significance level $\alpha=5\%$ (1 degree of freedom, since the table is a 2x2) which is 3.841. Thus the test rejects the null hypothesis at 5% level. Notice that also with a lower significance level like $\alpha=0.01$ – and thus with a more cautious test, which requires a stronger evidence to reject H_0 – we could reject H_0 (threshold would be 6.635).

EX 9

The measure of association we can compute is the linear correlation coefficient: (using the “fast formulas” for covariance and standard deviations)

	Height, X	Weight, Y	X*Y	X^2	Y^2
	165	55.2	9108	27225	3047.04
	157	48.1	7551.7	24649	2313.61
	168	65.6	11020.8	28224	4303.36
	178	72	12816	31684	5184
	171	73.5	12568.5	29241	5402.25
	182	77.9	14177.8	33124	6068.41
	182	80.3	14614.6	33124	6448.09
	153	49.4	7558.2	23409	2440.36
sum	1356	522	89415.6	230680	35207.12
sum/n	169.5	65.25	11176.95	28835	4400.89
			- mean^2 * n/(n-1) sqrt	104.75 119.7143 10.9414	143.3275 163.8029 12.79855

Covariance = $11176.95 - 169.5 \cdot 65.25 = 117.065$

Correlation Coeff. $r = 117.065 / (10.94 \cdot 12.80) = 0.836$

The value r indicates a strong positive association between the two variables.

Let us now compute the slope and intercept for the regression line:

$\text{beta} = 11176.95 / 119.7143 = 0.978$

$\text{intercept} = 65.25 - 0.978 \cdot 169.5 = -100.513$

The weight value predicted by this regression model for a subject with height=165 is thus:

$y^* = -100.513 + 0.978 \cdot 165 = 60.84$

The model over-estimated the weight of subject id=1 (anyway the order of magnitude is rather consistent)

EX 10

- a) Does treatment have an impact of the risk of progression?
 - Yes, it has a very significant effect ($p < 0.001$), B being less at risk of progression than treatment A. In fact the OR is less than 1, which means that B compared to A (baseline, denominator) is a protective factor. → **Needs correction, underlined**

- b) What is the effect of a prior treatment?
- The average increase of the odds of progression is estimated to be equal to 44%, but the estimation has a high variability, according to this model it could even be a protective factor; this is a non-significant factor → **True**
- c) Which is the most important risk factor?
- It is the presence of DEL13 abnormality ($p=0.0005$) → **False**. The most relevant is the increase of Beta 2 microglobulin (since it has the larger OR)
- d) What is the effect of high Beta2 microglobuline?
- The odds of progression is increased by about 36% (the effect is highly significant) → **True**