Biostatistics VHM 8010 Course
Fall 2024, Atlantic Veterinary College, PEI
Henrik Stryhn

Solution to home assignment II

The solution is more detailed than required for a 100% mark, by including detailed discussions of multiple ways of answering the different questions and their interpretations. The 10% version of the home assignment only included Questions $1-4$.

1. Statistical comparisons with population mean values

We first describe the modelling options generally, and then present the results for the separate analyses of the data for women and men. We consider a single SRS (simple random sample, or i.i.d. observations) from a population, in this case the population represented by the Framingham study participants. The interest is in comparing the (Framingham) population mean with the mean from the national survey. As we don't have access to the data behind the national survey, we cannot do a two-sample comparison. Instead we consider the national survey mean value as fixed (instead of an estimate), with the justification that it is presumably based on a very large sample. The options for 1-sample inference covered in Sessions $1-6$ of the course include z-based and t-based inferences when the population standard deviation is known and unknown, respectively. Although the IPS exercises referred to in Home asssignment I mentioned a known population standard deviation (and we explored the origin of those values in Question 4 of that assignment), it seems unnatural to assume that the national population standard deviation would be valid for the Framingham study population. Therefore, for a comparison with the population mean we should not assume the standard deviation to be known.

Denote by $X_1, \ldots, X_n$ the observations in the sample, and assume these to be i.i.d. from $\mathrm{N}(\mu, \sigma)$. We estimate the two unknown parameters by their sample values: $\hat{\mu} = \bar{X}$ and $\hat{\sigma} = s_X$. Then we use a t-procedure for a 95% CI for μ, and we test the hypothesis $\mathrm{H}_0 : \mu = \mu_0$ (where μ_0 is the relevant population value, for either women or men) by a t-test. In formulae,

$$95\% \text{ for } \mu: \ \bar{X} \pm t^* s_X/\sqrt{n} \quad \text{and} \quad t = \frac{\bar{X} - \mu_0}{s_X/\sqrt{n}}$$

where the t-value should be compared to a t-distribution with $\mathrm{df} = n-1$. The table below gives statistics and conclusions for the analyses for women and men. Alternative hypotheses were two-sided (in absence of information indicating otherwise).

Statistic		Women	Men
sample size	n	46	40
sample mean	$\bar{X}$	222.33	208.38
sample std.dev.	s_X	40.81	34.94
95% CI for μ		$(210.21, 234.45)$	$(197.20, 219.55)$
hypothesis	H_0	$\mu = 214$	$\mu = 227$
test	t	1.38	-3.37
P-value	P	0.17	0.002
Conclusion		cannot reject H_0	reject H_0

The data for women showed no evidence whatsoever against a normal distribution, so we should consider the statistical results as exact. We are 95% confident that the (Framingham) population mean is in the range $210-234$, and although the data mostly point towards a higher population

mean than 214, there is no (convincing) evidence to say that it differs from 214, the value from the national survey. We can have high confidence in this conclusion because our assumptions were met.

The data for men showed a somewhat right-skewed and non-normal distribution. Therefore the statistical results to some extent are approximate, but according to our discussion of the robustness of t-procedures (Lecture 6) with moderately large sample size (< 40) and no strong skewness or outliers the approximation should be quite good and reliable. Our case here is actually $n \geq 40$, so we could have used the large sample size guideline, but it will not change the conclusion. There was strong evidence against the hypothesized mean, and the sample mean was much lower than the hypothesized value. We can have high confidence in our conclusion that the Framingham population mean is lower than the value from the national survey, because even some small uncertainty related to the assumptions not being totally met is unlikely to affect the results enough to change the conclusion. We are 95% confident that the population mean lies within the range $197-220$.

2. Statistical comparison of values for women and men

The samples for men and women should be considered as two independent samples; even if a few of the study subjects were part of the same household or family, such a dependence is likely to have only minimal impact (because most of the study subjects would be truly independent). The natural model is therefore that the data for women $(X_1, \ldots, X_{n_1})$ constitute a SRS from $\mathrm{N}(\mu_0, \sigma_0)$ and that the data for men $(Y_1, \ldots, Y_{n_2})$ constitute a SRS from $\mathrm{N}(\mu_1, \sigma_1)$. Based on this model we can use t-procedures to compute a 95% CI for the mean difference $(\mu_0 - \mu_1)$ and to test the null hypothesis of equal means, $\mathrm{H}_0 : \mu_0 = \mu_1$. Again, we have no information to justify focus on a particular direction, so we use a two-sided alternative, $\mathrm{H}_a : \mu_0 \neq \mu_1$. The table below gives the calculations for the two versions of the t-procedures, depending on whether one assumes equal standard deviations, i.e. $\sigma_0 = \sigma_1$, or makes no such assumption. The descriptive statistics for the two samples were already given in Question 1.

Statistic		no assumption about σ's	$\sigma_1 = \sigma_2$ assumed
stand.dev's	s	separate	(pooled) 38.20
degrees of freedom	df	83	84
95% CI for $\mu_0 - \mu_1$		$13.95 \pm 16.25 \ = (-2.30, 30.20)$	$13.95 \pm 16.42 \ = (-2.47, 30.37)$
test	t	1.71	1.69
P-value	P	0.091	0.095
Conclusion		cannot reject H_0	cannot reject H_0

The two versions of t-procedures give almost identical results and conclusions. This is not surprising because the sample standard deviations are quite close and the sample sizes are large and similar. Even if it may not be (biologically) obvious why there would be a different spread in cholesterol values for men and women, there is also no particular reason to assume them to be the same. That is, the analysis without the assumption is preferable. With both t-procedures, the P-value for the t-test is around 0.09, thus slightly above the standard significance level of 0.05. We therefore have no formal evidence against H_0, and although the cholesterol mean was higher for women than men in the sample that could well have happened by chance. With the P-value so close to 0.05 we could perhaps talk about an "indication", "trend" or "tendency" of a higher mean for women, and we could call the P-value "close to significant". One good way of quantifying the information in the data is simply by the 95% CI for $\mu_0 - \mu_1$, which does include zero but on the other hand extends far longer on the positive side, corresponding to the mean being highest for women.

As to the model assumptions, we already discussed the normality assumption for each of the samples. Taken together, and considering the added robustness when having two samples (of approximately

the same size) it seems fair to say that the statistical results are probably pretty exact. We can have high confidence in our P-value somewhat above the 5% significance level as well as our conclusion.

3. Restricting analysis to subjects with a complete series

The most obvious implication of restricting the analysis to subjects with a complete series is a change in reference populations for the analysis. With this restriction, the sample could be considered representative for subjects that participated in (or could have participated in) the Framingham study for at least 10 years. Those subjects could differ in gender or age distributions (or in their health status) from the subjects that had the first cholesterol sample taken. In principle it might be possible to assess such differences between the subjects with a complete series and those that *dropped out* of the study, but in practice our sample size becomes quite small for such comparisons. It is a bit cumbersome to identify the 67 subjects from the original sample that dropped out during the 10-year period, so for simplicity we could instead compare the distributions in the full and reduced samples (for proper statistical inference, one would need to identify the two subsamples). It is seen that the reduced sample has more women than men (73 out of 133) but the full sample had most men (103 out of 200); thus the gender distribution might be affected by the restriction. The average age at the time of the first cholesterol measurement is slightly higher in the full than restricted sample (42.6 and 42.0 years), and although it would make sense that some older subjects would drop out of the study over a 10-year period, this could perhaps also be a coincidence.

We could also hypothesize potential biases by excluding the incomplete series: it could lead to higher or lower mean cholesterol values. A potential confounding effect of the restriction to a complete series could occur if presence of a complete series was linked to both a predictor of interest and to the outcome (cholesterol levels). The only cholesterol levels we can compare based on the data at hand are those at onset, and working again with the full and reduced datasets we get mean (median) cholesterol values for those two datasets of 220.5 (214) and 220.2 (215), respectively. These values do not indicate major differences in cholesterol values at onset, but one should be cautious because these overall values do not account for differences in the composition of the two data sets, e.g. in their gender distribution.

One final, and rather obvious, implication of restricting the analysis to a subset is the loss of power and precision associated with the reduced sample size.

4. Comparison of cholesterol values at baseline and year 10

If we compare cholesterol values at years 0 and 10 in these data, we must consider the two sets of values as paired samples: they are from the same subjects. As our general approach to dealing with paired samples is to work with the differences, our analysis for this part will focus on the changes in cholesterol values from year 0 to year 10. For one gender group, the natural model is then a single SRS (of differences) assumed to follow $\mathrm{N}(\mu_D, \sigma_D)$, and the focus of our interest will be the mean difference μ_D and the hypothesis $\mathrm{H}_0 : \mu_D = 0$. In absence of specific information about an expected change in cholesterol values over the years, we will use a two-sided alternative hypothesis $\mathrm{H}_a : \mu_D \neq 0$. We already gave the formulae for this situation in Question 1, and therefore proceed to present the results in tabular form (where D refers to the differences year 10 minus year 0). The table also include statistics to aid in the assessment of the model assumptions: skewness and P-value for the A-D test of normality for the differences. Note that assessing normality for each sample (year 0 and year 10) is not helpful because the analysis is based entirely on the differences.

Statistic		Women	Men
sample size	n	73	60
sample mean	$\bar{D}$	23.03	35.78
sample std.dev.	s_D	32.84	36.67
skewness		-0.73	0.19
normality test	P_n	0.11	0.25
95% CI for μ		$(15.36, 30.69)$	$(26.31, 45.26)$
test	t	5.99	7.56
P-value	P	< 0.0005	< 0.0005
Conclusion		reject H_0	reject H_0

The conclusion from these analyses seems quite clear: cholesterol levels increased over the 10 years for both women and men. This is seen both in the estimates and in the strongly significant t-tests. The skewness is moderate in both samples, and the normality test is non-significant; thus, no serious concerns exist with the assumed normal distributions, and we can have high confidence in the results. The confidence intervals show that the increase in cholesterol (population mean) value is likely to be substantial for both men and women.

5. Comparison of cholesterol values for years 2,. . . ,10 to baseline

As in Question 4, if we compare cholesterol values at years 0 and 2 (say) in these data, we must consider the two sets of values as paired samples: they are from the same subjects. Also here we will work with the differences, specifically the changes in cholesterol values relative to baseline, that is, the value at year 2 minus the value at year 0, and so on for the subsequent years. For one gender group, the natural model is then a single SRS (of differences) assumed to follow $N(\mu_D, \sigma_D)$, and the focus of our interest will be the mean difference μ_D and the hypothesis $H_0 : \mu_D = 0$, again with a two-sided alternative hypothesis $H_a : \mu_D \neq 0$. We will essentially repeat the analyses from Question 4 for the different time steps.

		Women: year compared to 0				
Statistic		2	4	6	8	10
sample mean	$\bar{D}$	7.19	10.32	15.95	14.93	23.03
sample std.dev.	s_D	38.67	30.51	33.95	36.07	32.84
skewness		-0.04	-0.27	-0.95	-0.28	-0.73
normality test	P_n	0.019	0.46	0.005	0.58	0.11
95% CI for μ_D		$(-1.83, 16.22)$	$(3.20, 17.43)$	$(8.03, 23.87)$	$(6.52, 23.35)$	$(15.36, 30.69)$
test	t	1.59	2.89	4.01	3.54	5.99
P-value	P	0.12	0.005	$<$0.0005	0.001	$<$0.0005
Conclusion		cannot reject H_0	reject H_0	reject H_0	reject H_0	reject H_0

The conclusion from the analyses for women seems quite clear: cholesterol levels tend to increase over the years. Apart from changes from baseline to years 6 and 8 being similar, the cholesterol levels consistently increase over time. There is clear statistical evidence of this trend in the differences to year 0 from year 4 onward. The differences for years 2 and 6 do not look too nice for a normal distribution, in both cases due to some extreme values (for year 6 only in the left tail). The statistical results may be considered approximate, but probably still pretty good, for those years. As the conclusion for the statistical test in both cases was very clear, we can have high confidence in these, and all other, conclusions.

Statistic		Men: year compared to 0				
		2	4	6	8	10
sample mean	$\bar{D}$	2.17	10.75	22.82	24.25	35.78
sample std.dev.	s_D	28.54	26.55	31.15	30.23	36.67
skewness		-0.47	0.51	1.05	0.39	0.19
normality test	P_n	0.69	0.33	0.077	0.25	0.25
95% CI for μ_D		$(-5.21, 9.54)$	$(3.89, 17.61)$	$(14.77, 30.86)$	$(16.44, 32.06)$	$(26.31, 45.26)$
test	t	0.59	3.14	5.67	6.21	7.56
P-value	P	0.56	0.003	<0.0005	<0.0005	<0.0005
Conclusion		cannot reject H_0	reject H_0	reject H_0	reject H_0	reject H_0

Also for the men, the cholesterol levels tend to increase over the years, and there is clear statistical evidence of this trend in the differences to year 0 from year 4 onward. This increasing trend is consistent across all years, even if the increase from year 6 to year 8 is small compared to other two-year increases. The differences for year 6 are a bit problematic for a normal distribution, due to a couple of extreme values in the right tail, but the normality test is only close to significant. With the large sample size, we can probably consider the statistical results for this year's differences as pretty exact. Therefore, we can have high confidence in all our conclusions.

6. Comparison of changes in cholesterol levels between women and men

Our analyses of the changes to baseline in the previous question showed that for both women and men the cholesterol levels tended to increase over time. Here we continue working with the differences to baseline and compare these changes between women and men. The statistical design is two independent samples (the 73 women and the 60 men), for each of the years 2,...,10, and we will assume these samples drawn from two normal distributions: $N(\mu_0, \sigma_0)$ for the women, and $N(\mu_1, \sigma_1)$ for the men. The two-sample t-procedures were already described in Question 2, and we therefore proceed to present the results in tabular form for each of the comparisons to year 0. As for Question 2 there is a choice between using the methods assuming $\sigma_0 = \sigma_1$ and those without the assumption. As the estimated standard deviations in the data for women and men were similar (Question 5), both approaches can be justified, but the default would be to *not* make the extra assumption. For completeness we include separate tables for both approaches. In order to avoid negative signs on estimates and test statistics, we give estimates for $\mu_1 - \mu_0$ (men − women) and use the corresponding estimate in the numerator of the t-statistics.

Analyses without any assumption on σ_0 and σ_1

Statistic/Year		2	4	6	8	10
mean difference	$\hat{\mu}_1 - \hat{\mu}_0$	−5.03	0.43	6.87	9.32	12.76
degrees of freedom	df	129	130	129	130	119
margin of error for 95% CI		±11.55	±9.79	±11.18	±11.38	±12.12
test	t	−0.86	0.09	1.22	1.62	2.09
P-value	P	0.39	0.93	0.23	0.11	0.039
Conclusion		cannot reject H_0				reject H_0

Analyses assuming $\sigma_0 = \sigma_1$

Statistic/Year		2	4	6	8	10
mean difference	$\hat{\mu}_1 - \hat{\mu}_0$	-5.03	0.43	6.87	9.32	12.76
pooled s (df$=131$)		34.48	28.79	32.71	33.57	34.62
margin of error for 95% CI		± 11.89	± 9.92	± 11.28	± 11.57	± 11.94
test	t	-0.84	0.09	1.21	1.59	2.11
P-value	P	0.40	0.93	0.23	0.11	0.036
Conclusion		cannot reject H_0				reject H_0

The two sets of analyses are seen to produce similar results throughout, and they both point to the conclusion that changes in cholesterol levels tend to be somewhat more pronounced in men than women as time passes (from 6 years onward), but only at 10 years a (weakly) significant difference between the changes for women and men can be demonstrated. Considering the added robustness of the two-sample t-procedures referred to in Question 2, the only analysis where we should consider the results as approximate is for year 6. This is because problems with the normality appear for both the men and women, and the distributions in the two gender groups show different features, in particular opposite skewness. As the test outcome for year 6 was clearly non-significant we can probably still have high confidence in it. All other analyses should be reasonably exact, giving us high confidence in conclusions.

One important issue ignored in the analyses for Questions 5 and 6 is whether we can allow us to investigate differences for the years 2,...,10 separately, thus in a sense giving us 5 chances to obtain significance. In the context of 1-way ANOVA we will discuss methods to adjust the overall significance level for such multiple analyses. The most conservative method will require us to multiply all P-values for the number of analyses carried out, in our case 5 (within each of Questions 5 and 6). It is seen that the conclusions of Question 5 are unaffected by such a procedure, whereas the significant difference in change of cholesterol values after 10 years for women and men would be affected. In such a situation it is probably better to analyse the cholesterol values for all years (0,...,10) simultaneously, but such an analysis of *repeated measures data* is beyond the scope of the course. A proper repeated measures analysis would have the additional advantage that it could include incomplete series as well, thus avoiding to restrict the data to complete series as we did for Questions $4-6$.