Solution to home assignment 3

The solution is more detailed than required for a 100% mark, by presenting and discussing several choices for the different analyses. Nevertheless, a full mark requires a discussion and justification of the chosen analytical approaches.

1. Comparison of quantitative variables across grade categories

Analysis for means
For the purpose of this analysis, we regard the data as three independent samples (from the populations of animals with the respective grades). Among the quantitative variables, whose distributions we explored overall (across grades) in the first home assignment, `carc_wt` was closest to normal, in particular when considered at transformed scales. From this we cannot be certain the same will be true when considering the three samples separately, and the AAA grade has the A-D normality test significant (or close to) at all scales; the P-values are 0.025, 0.047 and 0.058 on original, square-root and log-transformed scales, respectively. Also, the A grade has a clearly skewed distribution (skewness 0.85) on original scale, and in view of the guidelines for the robustness of t-distribution procedures for the mean (Session 6), this is probably of largest concern (smallest sample size and skewness, whereas the other, and larger, groups are not clearly skewed or have strong outliers, so pose not real concern). In order to use the same analysis method for all grades (clearly desirable, even if not always feasible), it seems sensible to work on original scale, and note that there may be a small reservation about the confidence interval for grade A. If we computed the CIs on transformed scale and back-transformed them, they would be for the median, *not* the mean.

We therefore assume the observations in sample i, denoted by $X_{ij}, j = 1, \ldots, n_i$, to be i.i.d. from $\mathrm{N}(\mu_i\, \sigma_i)$, for $i = 1, 2, 3$ (corresponding to grades A, AA, and AAA). For each sample, the table below gives estimated means and standard deviations (SDs) together with a 95% confidence interval (CI) for the mean which can also be found in Minitab's Graphical Summary.

grade (i)	nobs (n_i)	mean ($\bar{X}_{i\cdot}$)	SD (s_i)	95% CI: $\bar{X}_{i\cdot} \pm t^*(n_i-1) \cdot s_i/\sqrt{n_i}$
A	42	296.67	45.24	$(282.57, 310.76)$
AA	257	326.58	43.35	$(321.25, 331.90)$
AAA	144	344.17	42.12	$(337.23, 351.10)$

The CIs are clearly non-overlapping, indicating differences in mean carcass weights for animals scored with grades A, AA, and AAA. As the standard deviations are quite similar in the three samples, we could have computed a pooled standard deviation ($s_p = 43.13$) from the one-way ANOVA and used this value and its larger degrees of freedom (DFE = 440) for the intervals. However, as all groups are fairly large, the gain in precision from using a t^* with larger DF would be minimal. One way to explore how sensitive the CI for grade A is to the assumed normal distribution is to compute a bootstrap CI (available in Minitab's `Calc` menu); one such interval for 10,000 resamples was $(283.58, 310.29)$, and thus pretty close to the interval above.

For a test to compare the carcass weights among the grades, the most obvious choice is the F-test of the one-way ANOVA which tests the hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$. Analysis by statistical software gives $F = 21.08$, which in a $F(2, 440)$-distribution is very clearly significant ($P < 0.001$ from PSLS Table F; $P < 0.0005$ from Minitab). We conclude that differences in carcass weights between the grades do exist. The minor concern with the assumed normal distribution for grade A should not affect our conclusion.

We note also that the textbook guideline for equality of standard deviations, i.e. that their ratios do not exceed 2, is easily met.

For a two-sample approach, pairwise t-tests between any of the two grades are also clearly significant, e.g. t-values of 3.97, 4.00 and 6.08 without any pooling of standard deviations. Combining two categories into a single category is less attractive here because the three distributions appear separated; thus, the combined category will have a somewhat bimodal distribution.

Analysis for medians
In the first home assignment, the other quantitative variables all had clearly non-normal distributions overall, so for each of those variables it would be natural to use non-parametric inference for the median to describe the distribution center. The basic assumption for these methods is that the observations in each sample are i.i.d. from a continuous distribution, but Wilcoxon signed rank test methods additionally assume the distribution to be symmetrical. The CI for the median listed in the Graphical Summary is based on the sign test and does not have this assumption. It is perhaps most natural to use this CI in order to avoid the extra assumption; also, some of the distributions were previously shown to be rather skewed. The table below gives estimated medians and CIs (from the sign test) for the four other quantitative variables. To save space in the table a measure of the spread in each distribution, such as the interquartile range, is not included. Additionally, the P-value for the Kruskal-Wallis test ($\sim$ a non-parametric one-way ANOVA) to compare the distributions for the three grades is given as well.

Variable		`backfat`		`ribeye`		`imfat`		`days`	
grade	nobs	median	95% CI	median	95% CI	median	95% CI	median	95% CI
A	46	2.71	(2.48, 3.10)	8.01	(7.74, 8.70)	3.91	(3.47, 4.05)	186	(166, 214)
AA	277	2.73	(2.69, 2.92)	8.74	(8.52, 8.97)	3.86	(3.73, 3.97)	218	(194, 233)
AAA	164	2.73	(2.59, 2.99)	8.43	(8.17, 8.77)	4.28	(4.10, 4.51)	217	(190, 232)
K-W test P		0.66		0.15		< 0.005		0.13	

The findings differ between variables. For `backfat`, no substantial differences between grades are seen in the medians and the CIs, and the overall P-value is large (clearly above 0.05); no evidence of differences. For `imfat`, the AAA grade has the largest median with a CI separated from the other two grades (which appear quite similar), and the overall P-value is strongly significant; differences do exist. The remaining two variables show no overall significance with P-values in the $0.10-0.15$ range, and only grade A stands out mildly with a somewhat lower median than the higher grades. The results for these two variables will at best be termed as inconclusive of any (major) differences between grades. In order to interpret the evidence from the Kruskal-Wallis test for `imfat` as indicating differences between the medians, one will need to make the additional assumption that the three grade distributions are of similar shape, but they have in fact different skewness, so this does not seem valid.

A two-sample non-parametric approach could compare two of the groups by the Mann-Whitney(-Wilcoxon) test. For `imfat`, such comparisons show grade AAA to have a clearly different distribution than both grades A and AA ($P < 0.0005$ in Minitab) whereas the comparison between grade A and AA is not significant ($P = 0.15$). With the medians in the two groups being very close, it is clear that the rank sum test picks up another difference between the distributions than in their medians.

2. Analysis for categorical versions of quantitative variables

The smallest grade category (A) has 42 observations, and if we want to avoid cells in the two-way table with expected counts less than 5, the categories should comprise well above $5/42 = 0.12$ or 12% of the carcasses. This shows we can accommodate maybe up till 6 categories of about the same size without violating the conditions for the test. We may want to make the categories a bit larger in order to not get too small observed counts, and a simple and quite sensible choice for a total size of about 40 observations is four categories, defined by the quartiles of the distribution. In order to keep the solution reasonably short, only the variables `carc_tw` and `imfat` will be included. The 5-number summaries (across grades) for the two variables are: $(209.09, 297.27, 325.91, 360.91, 474.55)$ and $(1.85, 3.50, 3.97, 4.65, 8.36)$, respectively. We create categories according to these cutpoints, with the cutpoint value included in the lower interval (the Minitab default, and as good as any other rule). Note that the intervals need not be of equal sizes, and it is more useful for analysis to have not too unequal counts across the intervals.

With the quantitative variable categorised into quartiles, the question becomes one of studying the association between this categorical variable and `grade`, which is also categorical. That is, a two-way table analysis is called for. We have a choice between two statistical models, termed Model I and II in the lecture handouts. In accordance with the study objective we should take `grade` as a response variable. Both `carc_wt` and `imfat` were measured variables in the study with random variation, and would therefore most naturally be considered as response variables. This would correspond to model II, or the model for assessing independence between two response variables. As `imfat` was measured upon entry to the feedlot and well before slaughter, one could argue that these measurements effectively divided the cattle into groups which are followed to slaughter and grade assessment. In this way a model I, or comparing several independent populations, could be motivated as well for `imfat`, but model I seems unnatural for `carc_wt`. For both variables the descriptive statistics of primary interest is the grade distributions in each of the four quartiles (under model II assumptions, the conditional distribution of grade conditional on the other response variable). The cross-tabulations (from Minitab) below show: actual counts, grade distributions, expected counts under the null hypothesis of equal grade distributions (for model II: independence), and contributions to the Pearson chi-square statistics with $(3-1)\cdot(4-1)=6$ degrees of freedom.

A1.MTW

Tabulated Statistics: grade, Recoded carc_wt

Rows: grade Columns: Recoded carc_wt

	Q1	Q2	Q3	Q4	All
A	23	9	6	4	42
	21.30	7.89	5.41	3.64	9.48
	10.24	10.81	10.52	10.43	
	15.9031	0.3025	1.9446	3.9631	
AA	65	66	72	54	257
	60.19	57.89	64.86	49.09	58.01
	62.65	66.14	64.40	63.81	
	0.0878	0.0003	0.8981	1.5096	
AAA	20	39	33	52	144
	18.52	34.21	29.73	47.27	32.51
	35.11	37.06	36.08	35.76	
	6.5001	0.1019	0.2631	7.3794	
All	108	114	111	110	443
	100.00	100.00	100.00	100.00	100.00

Cell Contents
Count
% of Column
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	38.854	6	0.000
Likelihood Ratio	36.607	6	0.000

A1.MTW

Tabulated Statistics: grade, Recoded imfat

Rows: grade Columns: Recoded imfat

	Q1	Q2	Q3	Q4	All
A	15	8	16	3	42
	13.76	7.21	14.29	2.70	9.48
	10.33	10.52	10.62	10.52	
	2.107	0.605	2.727	5.379	
AA	73	71	63	50	257
	66.97	63.96	56.25	45.05	58.01
	63.23	64.40	64.98	64.40	
	1.508	0.677	0.060	3.218	
AAA	21	32	33	58	144
	19.27	28.83	29.46	52.25	32.51
	35.43	36.08	36.41	36.08	
	5.878	0.462	0.319	13.315	
All	109	111	112	111	443
	100.00	100.00	100.00	100.00	100.00

Cell Contents
Count
% of Column
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	36.255	6	0.000
Likelihood Ratio	36.948	6	0.000

The table for carc_wt shows substantial differences between the four conditional distributions which are also highly significant ($P < 0.0005$ by the Minitab listing). The validity of the chi-square test is ensured by all expected counts being well above 5 (all above 10). The highest, and indeed only substantial, chi-square contributions come from the corner cells of the table. The cattle with highest weights are more frequently graded AAA and less frequently graded A than would be expected under the null hypothesis. This pattern is also demonstrated by the requested confidence intervals for the grade AAA proportions, see the table below. Conversely, the lowest weight category shows an over-representation of A grades and an under-representation of AAA grades. The two middle weight categories have distributions that are roughly, with some variability, similar and in-between these two extremes. In conclusion, it is clear that higher weight is associated with higher carcass grade, and carcasses weighting more than 360.91 kg have almost half of them graded as AAA.

The cross-tabulation of categorised imfat values and grades shows a similar picture. The chi-square test is also highly significant ($P < 0.0005$), and the three largest chi-square contributions are again from the four corner cells. The carcasses with highest imfat values are more frequently graded AAA and less frequently graded A than would be expected under the null hypothesis. Conversely, the lowest imfat category shows an over-representation of A grades and an under-representation of AAA grades, and again the two middle imfat categories are somewhat in-between. In conclusion, it is clear that higher imfat percentage is associated with higher carcass grade, and imfat values above 4.65 are particularly attractive with slightly more than half of all carcasses graded as AAA.

The counts across AAA categories are so large that we can use the classical, normal approximation method for confidence intervals for the proportions in the respective $\mathrm{B}(n, p)$ distributions. All observations are assumed independent, and without using any further information about the carcasses we would assume the grade AAA probabilities to be equal within each category for carc_wt and imfat.

Variable	Category	Count X	Total n	Proportion $\hat{p} = X/n$	95% CI $\hat{p} \pm 1.96 \cdot \sqrt{\hat{p}(1-\hat{p})/n}$
carc_wt	Q1	20	108	0.185	$(0.112, 0.258)$
	Q2	39	114	0.342	$(0.255, 0.429)$
	Q3	33	111	0.297	$(0.212, 0.382)$
	Q4	52	110	0.473	$(0.379, 0.566)$
imfat	Q1	21	109	0.193	$(0.119, 0.267)$
	Q2	32	111	0.288	$(0.204, 0.372)$
	Q3	33	112	0.295	$(0.210, 0.379)$
	Q4	58	111	0.523	$(0.430, 0.615)$

The confidence intervals confirm our impressions from the two-way table analysis. There are large differences between Q4 and Q1 proportions for grade AAA for both variables, and the confidence intervals are not even close to overlapping. Both of the intermediate quartiles also have non-overlapping confidence intervals with the highest quartile for imfat, demonstrating the significant improvement in carcass grade with for the highest imfat values.

3. Relations involved in association between implant and grade

This solution focuses on farm as the potential lurking variable. As explained and illustrated in the solution to home assignment I, the causal diagram involves three associations:

- between implant and grade,
- between farm and implant,

- between `farm` and `grade`.

Because all these variables are categorical, two-way table analysis is the relevant tool to study their associations. As before we consider `grade` as a response variable, and when collecting information about a sample of cattle at the feedlot we neither know or have predetermined the origin of the cattle (`farm`) or whether these were implanted or not. Therefore, in this context both `farm` and `implant` may also be considered as response variables, or we may consider the farms as explanatory and the use of implants within farms as random. In any case, the distributions of interest are conditional on farms (for both `implant` or `grade`). With the former interpretation of all variables as responses, the hypothesis of interest for each two-way table analysis is H_0: independence between classification by the two variables involved, and the results are summarised in the table below (with Minitab listings given at the end).

Association	`implant` vs. `grade`	`farm` vs. `implant`	`farm` vs. `grade`
chi-sq. statistic (X^2)	0.77	352.0	91.5
degrees of freedom (df)	2	5	10
expected values (e_{ij})	all > 10	all > 10	1 (8.3%) within $1 - 5$
P-value	0.68	$<$0.0005	$<$0.0005
conclusion	no association	strong association	strong association

In order for `farm` to be a lurking (confounding) variable for the association between `implant` and `grade`, its associations with these two variables must both exist (that is, be non-negligible). The table shows that these associations are indeed very strong. The one expected value less than 5 in the `farm` vs. `grade` table does not violate the guidelines (at least 80% of $e_{ij} > 5$). The strong association between `farm` and `implant` is due to very different uses of implants on the farms, ranging between both extremes of the proportions (0 and 1). The strong association between `farm` and `grade` is primarily due to a couple of farms with different grade distributions, in particular farm 4 with very few high grades and farms 2 and 7 with many high grades. The association between `implant` and `grade` in the data may therefore be strongly affected by farms. That overall association is totally non-significant: the distribution on the grades is virtually identical regardless of whether the animal was implanted or not. A more sophisticated analysis (beyond this course) is required to determine whether this association is still non-significant when the farms are accounted for. The absence of an effect of implant in the final model in the journal article would lead one to think that the association is also non-significant when farms are accounted for, and this is indeed the case.

Minitab listings for two-way table analyses on the next page:

A1.MTW

Tabulated Statistics: grade, implant

Rows: grade Columns: implant

	0	1	All
A	29	13	42
	31.29	10.71	
	0.16713	0.48808	
AA	192	65	257
	191.44	65.56	
	0.00161	0.00470	
AAA	109	35	144
	107.27	36.73	
	0.02795	0.08161	
All	330	113	443

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	0.771	2	0.680
Likelihood Ratio	0.745	2	0.689

A1.MTW

Tabulated Statistics: grade, farm

Rows: grade Columns: farm

	1	2	3	4	5	7	All
A	1	0	12	19	10	0	42
	5.31	5.40	12.70	7.68	6.16	4.74	
	3.498	5.404	0.039	16.688	2.390	4.740	
AA	28	23	88	59	37	22	257
	32.49	33.07	77.74	46.99	37.71	29.01	
	0.620	3.065	1.355	3.069	0.013	1.693	
AAA	27	34	34	3	18	28	144
	18.20	18.53	43.56	26.33	21.13	16.25	
	4.251	12.920	2.097	20.671	0.463	8.491	
All	56	57	134	81	65	50	443

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	91.466	10	0.000
Likelihood Ratio	106.863	10	0.000

1 cell(s) with expected counts less than 5.

A1.MTW

Tabulated Statistics: implant, farm

Rows: implant Columns: farm

	1	2	3	4	5	7	All
0	56	0	134	25	65	50	330
	41.72	42.46	99.82	60.34	48.42	37.25	
	4.89	42.46	11.70	20.70	5.68	4.37	
1	0	57	0	56	0	0	113
	14.28	14.54	34.18	20.66	16.58	12.75	
	14.28	124.00	34.18	60.44	16.58	12.75	
All	56	57	134	81	65	50	443

Cell Contents
Count
Expected count
Contribution to Chi-square

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	352.038	5	0.000
Likelihood Ratio	402.994	5	0.000