Slaughter Data Analysis Report

1. Problem Statement & Data Description

Data collected before and after pig slaughtering are important indices of carcass values. For this particular problem, we want to compare the carcass values of two groups of pigs with different genotypes by using data collected during slaughter. For each pig slaughtered, we record its genotype as **Sire Line**, the **index** of pig, its **HCW** (Hot Carcass Weight) and **BF** (Backfat thickness, at 6-7 ribs) along with its carcass **Grade.** For the following analysis I will assume that such Grade is made by some established criterions related with only HCW and BF.

2. Data Inspection

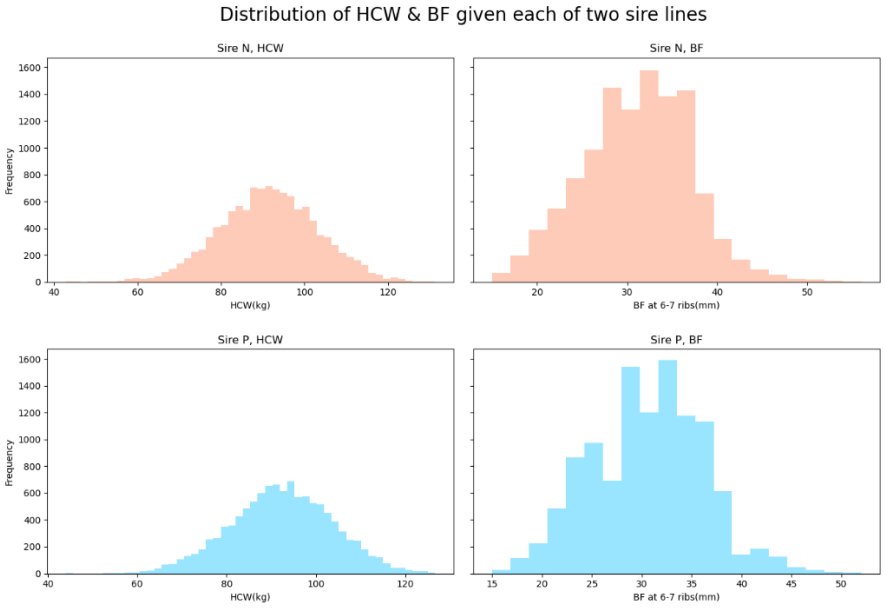
There is no blank cell in this dataset and there are 11419 samples from Sire N and 11196 samples from Sire P, which is very close. And as HCW, BF and Grade are all numeric variables, I calculated some basic statistics for them and recorded as follows:

	HCW	BF	Grade
count	22614.00000	22614.00000	22614.00000
mean	91.98045	30.88777	2.53060
std	11.56494	5.76043	1.18859
min	42.80000	15.00000	1.00000
25%	84.22500	27.00000	1.00000
50%	92.00000	31.00000	3.00000
75%	99.80000	35.00000	3.00000
max	131.20000	56.00000	5.00000

We can see that all the recordings from these varaibles lie in resonable range given their mean, min, max and standard deviation. Thus it seems like we don't need to do furthur wrangling or cleansing.

3. Data Visualization

Now, we want to visualize the distribution of HCW and BF for each of these two Sire Lines. Although we can still use some statistical testing such as t-test for large samples like this dataset despite its distribution, doing data visualization will always provide us extra information.



I think these two Sire Lines have very similar distributions both in HCW and BF, so I think now we can only use statistical testing to distinguish them.

4. Detailed Analysis and Results

More detailed statistics are calculated and store in the file 'Sire_Line_statistics.xlsx'. From there, we can Sire N and P have very close performance indeed, with P did slightly better in mean of HCW and N did slightly better in mean BF and Final Grade.

By implementing Kolmogorov–Smirnov tests for all three variables between N and P, all of the p-values are less than 0.05 (1.90 * 10^-7, 4.45 * 10^-18, 0.0087 for HCW, BF and Grade respectively), so we can accept the conclusion that these variables between N and P are from different distributions. Then we implement three t-test to compare their means, and p-values are 1.55 * 10^-8, 1.16 * 10^-16 and 0.206 for HCW, BF and Grade respectively. This indicates that the **mean of HCW from P is significantly higher than that of N**, and the **mean of BF from N is significantly higher than that of P**. However, **there is no significant difference between mean of Grade for N and P.**

5. Conclusions and Discussions

So how is HCW and BF related to carcass values specifically? Basically higher HCW means more saleable and edible meat, and also according to Harsh B N, Arkfeld E K et al, Heavier carcass weight produces thicker and firmer bellies [1]. Thus a high HCW value should be one the criterions for a good breed of meat pig.

As for BF, it can also be related with meat quality [2], but it's more often used as a significant parameter to consider when selecting female pigs into breeding herds since it dominates a number of reproductive performances [3].

In conclusion, there is no significant difference of the carcass value w.r.t its grade between Sire N and Sire P, but Sire P has more potential to be selected as meat pig while Sire N could be more suited to perform reproductive tasks.

For future work, maybe we can consider looking at the correlation between HCW and BF either integrally or sectionally.

One of the main problems of this dataset is absence of the cost for raising, for example, time from birth or average daily gain. This reduces the meaning behind such comparison of carcass values.

Finally, as a bonus part, I build a naïve Bayes classifier for Grade based on HCW and BF and this 5-class classification model gives a 57% accuracy rate when tested by 5-folder cross validation.

[1] Harsh B N, Arkfeld E K, Mohrhauser D A, et al. Effect of hot carcass weight on loin, ham, and belly quality from pigs sourced from a commercial processing facility[J]. Journal of animal science, 2017, 95(11): 4958-4970.

[2] Kim G W, Kim H Y. Effects of carcass weight and back-fat thickness on carcass properties of Korean native pigs[J]. Korean journal for food science of animal resources, 2017, 37(3): 385.

[3] Roongsitthichai A, Tummaruk P. Importance of backfat thickness to reproductive performance in female pigs[J]. The Thai Journal of Veterinary Medicine, 2014, 44(2): 171-178.