Chapter 8: Comparing Two Independent Populations
Part 1: The Two-Sample T-Test (Normal with Equal but Unknown Variances)

In the previous section, our data was a sample drawn from one population, and we saw several procedures for testing measures of central location about that population based on the sample. Here we will concern ourselves with comparing measures of central location of two independent populations, based on samples from each.

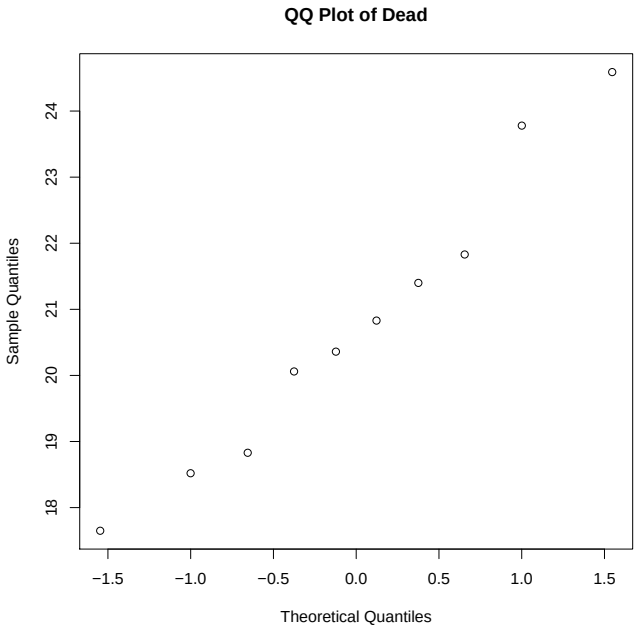
Example: The horned lizard *Phrynosoma mcalli* is named for the fringe of spikes around the back of the head. It was thought that the spikes may provide the lizard protection from its primary predator, the loggerhead shrike, *Lanius ludovicanus*, though there was not much existing quantitative evidence to support this. Researchers were interested in comparing two populations: the population of dead lizards known to be killed by shrikes, and the population of live lizards from the same geographic location. Random samples were taken from each population. The longest spike was measured on each sampled lizard, in mm. **The primary research question was, "Is there any difference in the size of the spikes between the two populations?" A difference in mean spike length between the two groups would indicate whether spike length was associated with survival.**

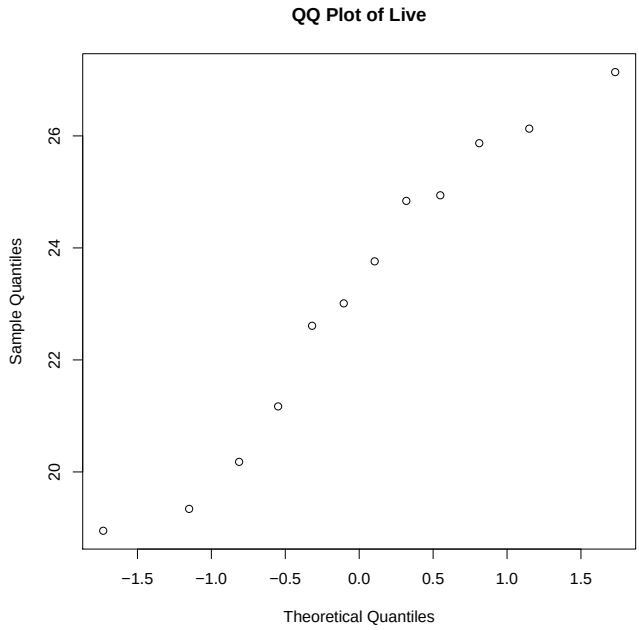
The data are as follows:

- Dead Group: 17.65, 20.83, 24.59, 18.52, 21.40, 23.78, 20.36, 18.83, 21.83, 20.06
- Live Group: 23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87, 18.95, 22.61

We find each data point is independent from the context.

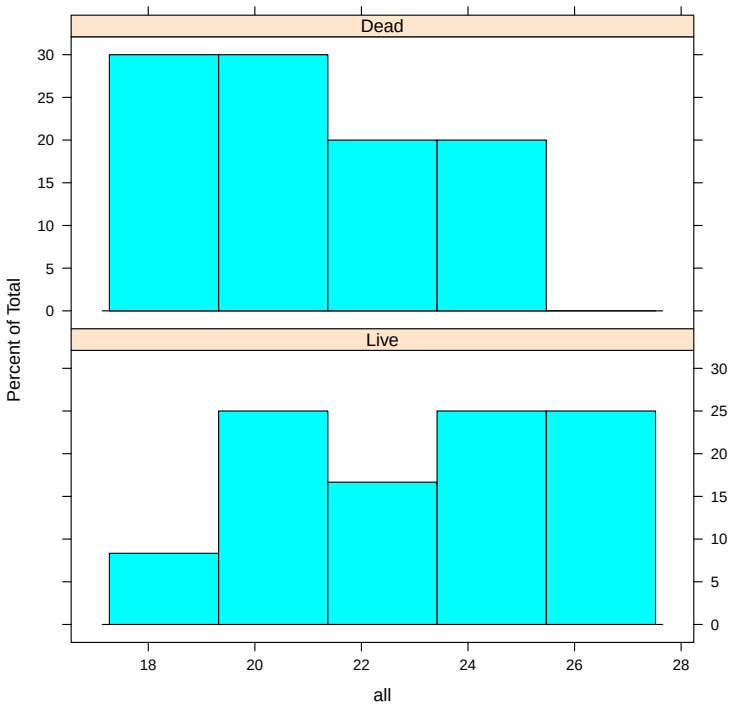
Let's start by graphing the data. We could first do separate QQ plots for each sample:





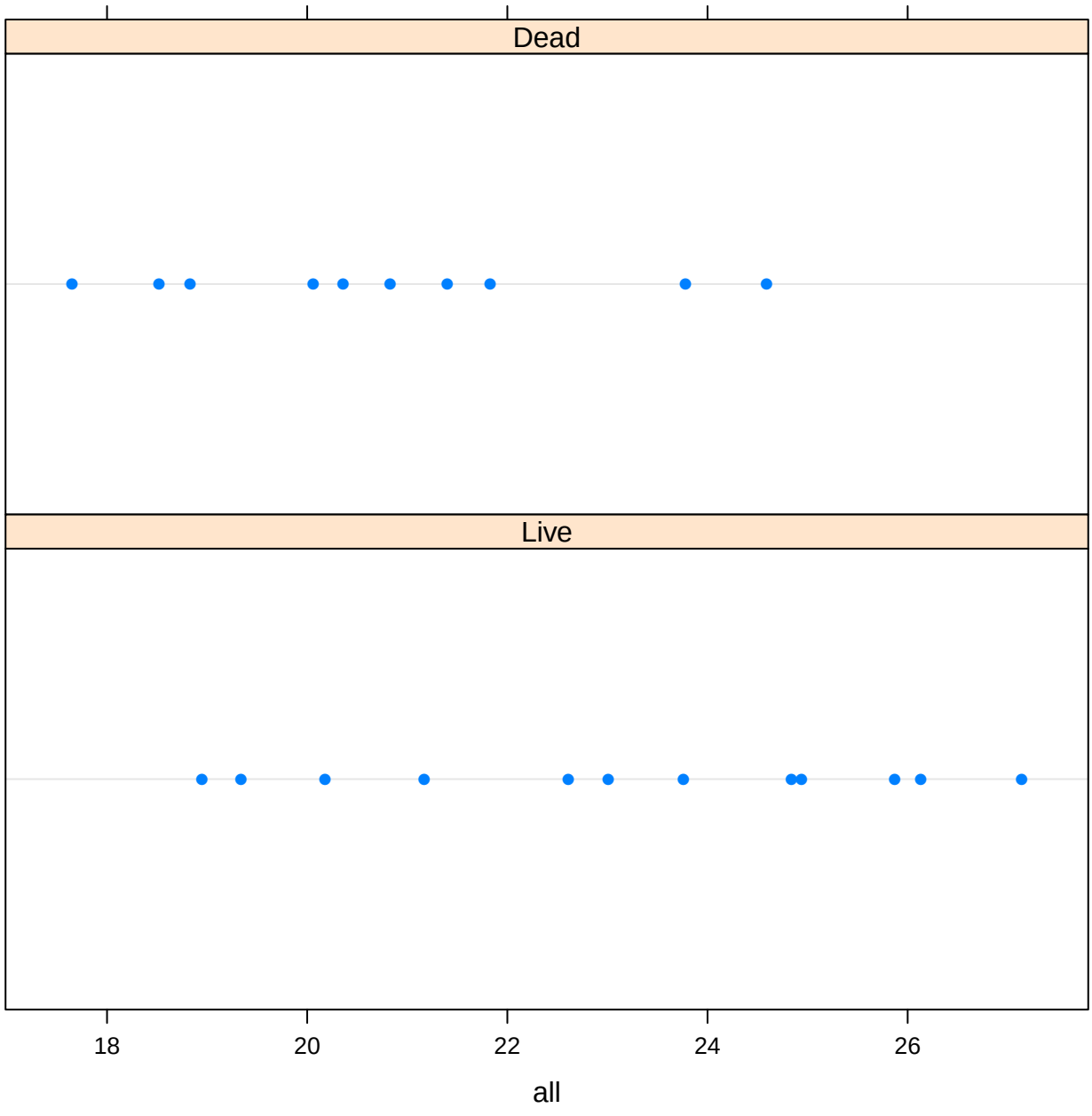
What do you find from the QQ plot? each sample comes from a normal population.

Next, we draw the histograms from each group, top to bottom:



These are pretty chunky because of the small sample size and may not be the best thing in this case. But often they will be quite helpful. Even with the small sample size you can see that the live group seems to be shifted to the right a bit.

Another helpful plot just plots the raw data, we call this a **dotplot**:



Again we can see the shift. **Dotplot is good when there isn't much data, but when there's a lot, sometimes it's hard to see the important aspects of the data, and we'd be better off summarizing somehow.** One choice is called a **boxplot**.

To make a boxplot:

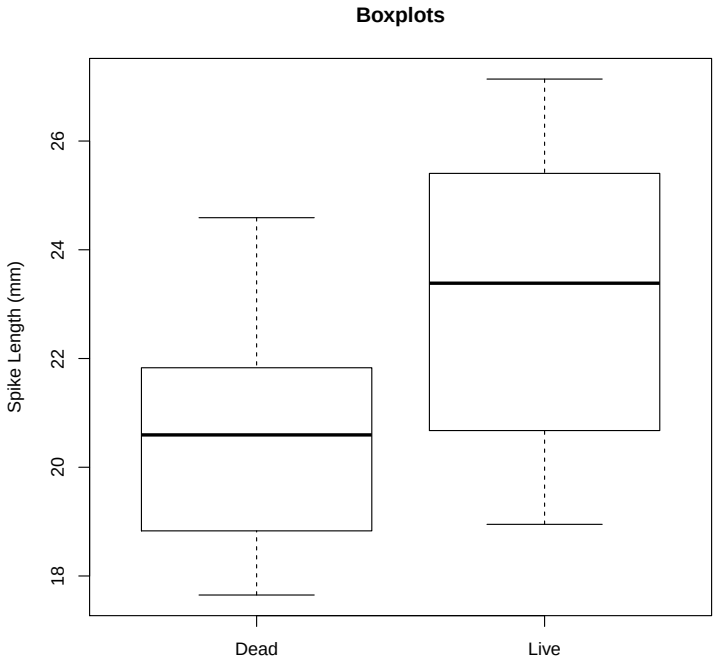
- Plot a bar at the median, and at the first and third quartiles.
- Connect the ends of the bars to make a box with a line in it.

- Extend whiskers out to a maximum of 1.5*IQR up from the third quartile and down from the first quartile, but only go to the largest or smallest actual data point within that range.
- Any other data point gets a dot.

Here is some numerical summary information that will be helpful:

Group	Sample Size, n	Sample Mean	Sample SD	1st Quartile	Sample Median	3rd Quartile
Dead	10	20.79	2.22	19.14	20.59	21.72
Live	12	23.16	2.76	20.92	23.16	25.17

Here are the boxplots side by side:



Boxplots are good for showing the rough location and spread of data, especially when there is a lot of data so that dotplots are too 'busy.' We can see in this boxplot, as we did in the histogram and dotplot, that live seems generally higher. The spread also seems about the same. Again, with a small number of points, boxplots may hide important features of the individual data points, and a dotplot might be better.

Next let's talk about the population standard deviations for each sample. In this example, we don't know the population standard deviation. **When the sample sizes are similar in the two groups, if** $0.5 \leq \frac{s_1}{s_2} \leq 2.0$**, we can assume the population standard deviations are equal.** In this case, $\frac{s_1}{s_2} = \frac{2.22}{2.76} = 0.8$, so we should be safe.

Now that we have explored the data, what kind of test might be appropriate? The fundamental test would be whether the means were different. If we introduce the notation that μ_{dead} = Mean of Dead Population, and μ_{live} = Mean of Live Population, we could define our hypotheses as:

$$H_0 : \mu_{dead} = \mu_{live} \text{ vs. } H_A : \mu_{dead} \neq \mu_{live}$$

Or, equivalently:

$$H_0 : \mu_{dead} - \mu_{live} = 0 \text{ vs. } H_A : \mu_{dead} - \mu_{live} \neq 0$$

Next we need to choose a test statistic. In the one-sample situation, we used a T-statistic, which in general has the form:

$$statistic = \frac{(estimate\ from\ data) - (hypothesized\ value\ under\ null)}{estimated\ standard\ error\ of\ estimator}$$

In this two-sample case with normal populations, equal but unknown variances, we use the following statistic:

$$T = \frac{\bar{X}_{dead} - \bar{X}_{live} - 0}{S_p \sqrt{\frac{1}{n_{dead}} + \frac{1}{n_{live}}}}$$

where

$$S_p^2 = \frac{(n_{dead}-1)S_{dead}^2 + (n_{live}-1)S_{live}^2}{n_{dead}+n_{live}-2}$$

Then the distribution of the statistic is t-distribution with $n_{dead} + n_{live} - 2$ degrees of freedom.

Now let's compute the statistic using the real data. The summary stats that we computed initially will be very helpful:

- $s_{dead}^2 = 2.22^2 = 4.93$
- $s_{live}^2 = 2.76^2 = 7.62$
- $s_p^2 = \frac{(10-1)4.93 + (12-1)7.62}{10+12-2} = 6.41$
- $t = \frac{20.79 - 23.16 - 0}{\sqrt{6.41}\sqrt{\frac{1}{10} + \frac{1}{12}}} = -2.195$

Compare this to a T distribution on $n_{dead} + n_{live} - 2 = 10 + 12 - 2 = 20$ df, and find $p = 0.040$. Given $\alpha = 0.05$, since the p-value is smaller than α, we reject the null hypothesis. Since the live group has longer spikes, we can tentatively conclude that longer spikes are associated with greater survival.