Chapter 8: Comparing two independent populations

Duzhe Wang

Part 1
https://dzwang91.github.io/stat371/

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

What do we study

- In the previous chapter, 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.
- **Key concepts in Part 1**: Two-sample T-test

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.

Example continued

- 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.

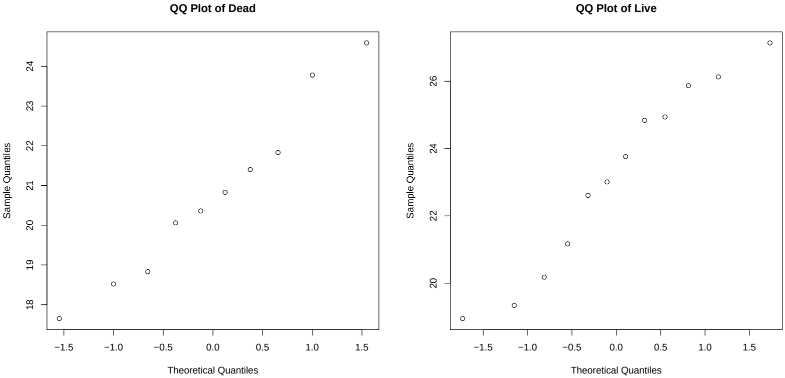
- The data are as follows:
  - Dead Group: 17.65, 20.83, 24.59, 18.52, 21.40, 23.78, 20.36, 18.83, 21.83, 20.06
  - Live Group: 23.76, 21.17, 26.13, 20.18, 23.01, 24.84, 19.34, 24.94, 27.14, 25.87, 18.95, 22.61

Example continued

- Each data point is independent from the context.

Example continued

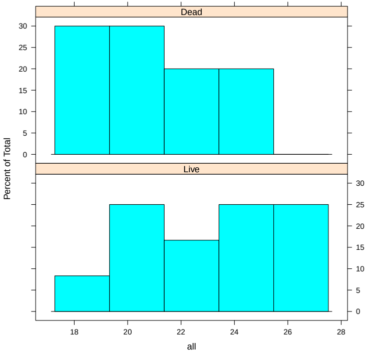
- Each data point is independent from the context.
- Separate QQ plots for each sample:



Each sample comes from a normal population.

Example continued

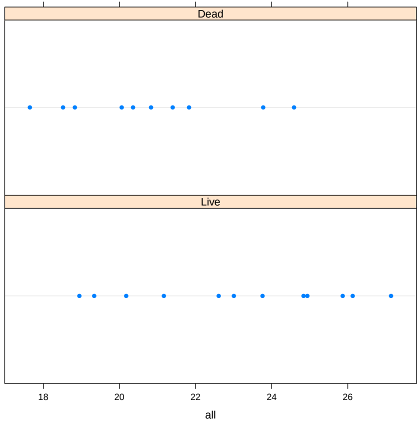
- Histograms from each group, top to bottom:



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.

Example continued

- 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.

Example continued

- When there is a lot of data, a good choice for showing the rough location and spread of data 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.
    - Any other data point outside of that range gets a dot.

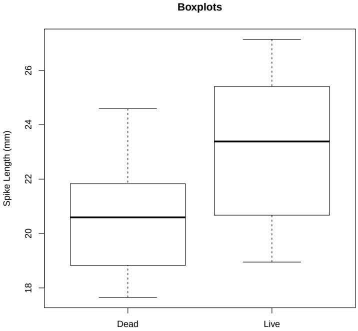
Example continued

- When there is a lot of data, a good choice for showing the rough location and spread of data 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.
  - Any other data point outside of that range gets a dot.
- Here is some numerical summary information that will be helpful:

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

Example continued

- Boxplots side by side:



- In this boxplot, as we did in the histogram and dotplot, live seems generally higher. The spread also seems about the same.

Example continued

- 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.

Example continued

- 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.
- State H_0 and H_A: The fundamental test in this example 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:

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

Example continued

- 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.
- State H_0 and H_A: The fundamental test in this example 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:

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

Example continued

- Choose a test statistic: In this two-sample case with normal populations, equal but unknown variances, we use the following test 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 test statistic is T-distribution with $n_{dead} + n_{live} - 2$ degrees of freedom.

- In this example, $s_{dead}^2 = 2.22^2 = 4.93$, $s_{live}^2 = 2.76^2 = 7.62$, $s_p^2 = \frac{(10-1)4.93+(12-1)7.62}{10+12-2} = 6.41$, $T_{obs} = \frac{20.79-23.16-0}{\sqrt{6.41}\sqrt{\frac{1}{10}+\frac{1}{12}}} = -2.195$, and $n_{dead} + n_{live} - 2 = 10 + 12 - 2 = 20$.

Example continued

- p-value: $2 \times P(T_{20} > 2.195) = 0.04$.
- Conclusion: 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.

Recap of two-sample T-test

- The data consists of separate simple random samples from two independent populations, label them 1 and 2. Let
  - $\mu_1 =$ true mean of population 1
  - $\mu_2 =$ true mean of population 2
  - $n_1 =$ sample size taken from population 1
  - $n_2 =$ sample size taken from population 2
  - $\sigma_1^2 =$ true variance of population 1
  - $\sigma_2^2 =$ true variance of population 2

Recap of two-sample T-test

- The data consists of separate simple random samples from two independent populations, label them 1 and 2. Let
  - μ_1 = true mean of population 1
  - μ_2 = true mean of population 2
  - n_1 = sample size taken from population 1
  - n_2 = sample size taken from population 2
  - σ_1^2 = true variance of population 1
  - σ_2^2 = true variance of population 2
- We wish to test: $H_0 : \mu_1 - \mu_2 = \delta$ vs. $H_A : \mu_1 - \mu_2 \neq \delta$(or other one-tailed alternative hypothesis).

Recap of two-sample T-test

- The data consists of separate simple random samples from two independent populations, label them 1 and 2. Let
  - $\mu_1 =$ true mean of population 1
  - $\mu_2 =$ true mean of population 2
  - $n_1 =$ sample size taken from population 1
  - $n_2 =$ sample size taken from population 2
  - $\sigma_1^2 =$ true variance of population 1
  - $\sigma_2^2 =$ true variance of population 2
- We wish to test: $H_0 : \mu_1 - \mu_2 = \delta$ vs. $H_A : \mu_1 - \mu_2 \neq \delta$(or other one-tailed alternative hypothesis).
- Good numerical and graphical summaries to explore the data might include means, medians, standard deviations, QQ plot, side-by-side boxplots, side-by-side dotplots, histograms.

Recap of two-sample T-test

- After exploring the data, if we are willing to assume:
  - All of the data points are independent, both within and between populations
  - The two populations each follow normal distributions
  - The variances of the two populations are equal so that $\sigma_1^2 = \sigma_2^2 = \sigma^2$

  then the test statistic is:

$$t = \frac{\bar{X}_1 - \bar{X}_2 - \delta}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  Where:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- Compute the p-value using T distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom and then make a conclusion using given α.

What's the next?

We'll discuss how to run hypothesis tests in the setting of unequal variance in the next lecture.