Chapter 5

Inference for numerical data

Chapter 4 introduced a framework for statistical inference based on confidence intervals and hypotheses. In this chapter, we encounter several new point estimates and scenarios. In each case, the inference ideas remain the same:

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Apply the ideas from Chapter 4 using the distribution from step 2.

Each section in Chapter 5 explores a new situation: the difference of two means (5.1, 5.2); a single mean or difference of means where we relax the minimum sample size condition (5.3, 5.4); and the comparison of means across multiple groups (5.5). Chapter 6 will introduce scenarios that highlight categorical data.

5.1 Paired data

Are textbooks actually cheaper online? Here we compare the price of textbooks at UCLA's bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010, representing less than 10% of all UCLA courses.[1] A portion of this data set is shown in Table 5.1.

	dept	course	ucla	amazon	diff
1	Am Ind	C170	27.67	27.95	-0.28
2	Anthro	9	40.59	31.14	9.45
3	Anthro	135T	31.68	32.00	-0.32
4	Anthro	191HB	16.00	11.52	4.48
⋮	⋮	⋮	⋮	⋮	⋮
72	Wom Std	M144	23.76	18.72	5.04
73	Wom Std	285	27.70	18.22	9.48

Table 5.1: Six cases of the `textbooks` data set.

[1] When a class had multiple books, only the most expensive text was considered.

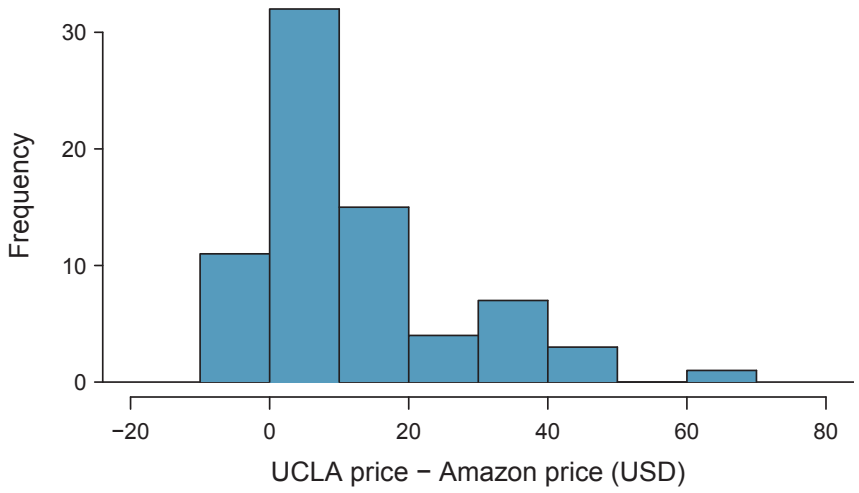

Figure 5.2: Histogram of the difference in price for each book sampled. These data are strongly skewed.

5.1.1 Paired observations and samples

Each textbook has two corresponding prices in the data set: one for the UCLA bookstore and one for Amazon. Therefore, each textbook price from the UCLA bookstore has a natural correspondence with a textbook price from Amazon. When two sets of observations have this special correspondence, they are said to be **paired**.

Paired data
Two sets of observations are *paired* if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations. In the `textbook` data set, we look at the difference in prices, which is represented as the `diff` variable in the `textbooks` data. Here the differences are taken as

$$\text{UCLA price} - \text{Amazon price}$$

for each book. It is important that we always subtract using a consistent order; here Amazon prices are always subtracted from UCLA prices. A histogram of these differences is shown in Figure 5.2. Using differences between paired observations is a common and useful way to analyze paired data.

⊙ **Exercise 5.1** The first difference shown in Table 5.1 is computed as $27.67 - 27.95 = -0.28$. Verify the differences are calculated correctly for observations 2 and 3.[2]

5.1.2 Inference for paired data

To analyze a paired data set, we use the exact same tools that we developed in Chapter 4. Now we apply them to the differences in the paired observations.

[2]Observation 2: $40.59 - 31.14 = 9.45$. Observation 3: $31.68 - 32.00 = -0.32$.

n_{diff}	$\bar{x}_{diff}$	s_{diff}
73	12.76	14.26

Table 5.3: Summary statistics for the price differences. There were 73 books, so there are 73 differences.

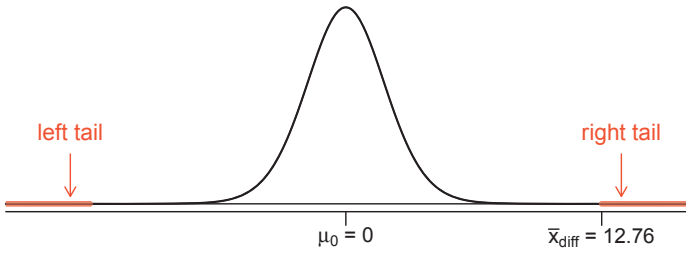

Figure 5.4: Sampling distribution for the mean difference in book prices, if the true average difference is zero.

Example 5.2 Set up and implement a hypothesis test to determine whether, on average, there is a difference between Amazon's price for a book and the UCLA bookstore's price.

There are two scenarios: there is no difference or there is some difference in average prices. The *no difference* scenario is always the null hypothesis:

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.
H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Can the normal model be used to describe the sampling distribution of $\bar{x}_{diff}$? We must check that the differences meet the conditions established in Chapter 4. The observations are based on a simple random sample from less than 10% of all books sold at the bookstore, so independence is reasonable; there are more than 30 differences; and the distribution of differences, shown in Figure 5.2, is strongly skewed, but this amount of skew is reasonable for this sized data set ($n = 73$). Because all three conditions are reasonably satisfied, we can conclude the sampling distribution of $\bar{x}_{diff}$ is nearly normal and our estimate of the standard error will be reasonable.

We compute the standard error associated with $\bar{x}_{diff}$ using the standard deviation of the differences ($s_{diff} = 14.26$) and the number of differences ($n_{diff} = 73$):

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{14.26}{\sqrt{73}} = 1.67$$

To visualize the p-value, the sampling distribution of $\bar{x}_{diff}$ is drawn as though H_0 is true, which is shown in Figure 5.4. The p-value is represented by the two (very) small tails.

To find the tail areas, we compute the test statistic, which is the Z score of $\bar{x}_{diff}$ under the null condition that the actual mean difference is 0:

$$Z = \frac{\bar{x}_{diff} - 0}{SE_{x_{diff}}} = \frac{12.76 - 0}{1.67} = 7.59$$

This Z score is so large it isn't even in the table, which ensures the single tail area will be 0.0002 or smaller. Since the p-value corresponds to both tails in this case and the normal distribution is symmetric, the p-value can be estimated as twice the one-tail area:

$$\text{p-value} = 2 \times (\text{one tail area}) \approx 2 \times 0.0002 = 0.0004$$

Because the p-value is less than 0.05, we reject the null hypothesis. We have found convincing evidence that Amazon is, on average, cheaper than the UCLA bookstore for UCLA course textbooks.

⊙ **Exercise 5.3** Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[3]

5.2 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. The methods are similar in theory but different in the details. Just as with a single sample, we identify conditions to ensure a point estimate of the difference $\bar{x}_1 - \bar{x}_2$ is nearly normal. Next we introduce a formula for the standard error, which allows us to apply our general tools from Section 4.5.

We apply these methods to two examples: participants in the 2012 Cherry Blossom Run and newborn infants. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

5.2.1 Point estimates and standard errors for differences of means

We would like to estimate the average difference in run times for men and women using the `run10Samp` data set, which was a simple random sample of 45 men and 55 women from all runners in the 2012 Cherry Blossom Run. Table 5.5 presents relevant summary statistics, and box plots of each sample are shown in Figure 5.6.

	men	women
$\bar{x}$	87.65	102.13
s	12.5	15.2
n	45	55

Table 5.5: Summary statistics for the run time of 100 participants in the 2009 Cherry Blossom Run.

The two samples are independent of one-another, so the data are not paired. Instead a point estimate of the difference in average 10 mile times for men and women, $\mu_w - \mu_m$, can be found using the two sample means:

$$\bar{x}_w - \bar{x}_m = 102.13 - 87.65 = 14.48$$

[3] Conditions have already verified and the standard error computed in Example 5.2. To find the interval, identify $z^\star$ (1.96 for 95% confidence) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 12.76 \pm 1.96 \times 1.67 \quad \rightarrow \quad (9.49, 16.03)$$

We are 95% confident that Amazon is, on average, between \$9.49 and \$16.03 cheaper than the UCLA bookstore for UCLA course books.