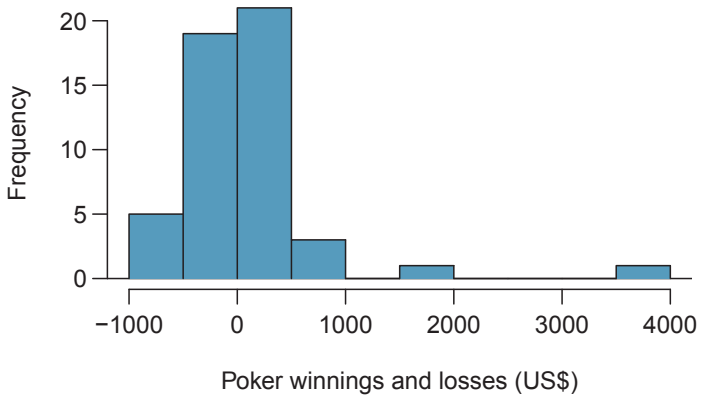

Figure 4.21: Sample distribution of poker winnings. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

Caution: Watch out for strong skew and outliers
Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$. There are no simple guidelines for what sample size is big enough for all situations, so proceed with caution when working in the presence of strong skew or more extreme outliers.

4.5 Inference for other estimators

The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample proportions closely resembles the normal distribution when the sample size is sufficiently large. In this section, we introduce a number of examples where the normal approximation is reasonable for the point estimate. Chapters 5 and 6 will revisit each of the point estimates you see in this section along with some other new statistics.

We make another important assumption about each point estimate encountered in this section: the estimate is unbiased. A point estimate is **unbiased** if the sampling distribution of the estimate is centered at the parameter it estimates. That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a "good" estimate. The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Finally, we will discuss the general case where a point estimate may follow some distribution other than the normal distribution. We also provide guidance about how to handle scenarios where the statistical techniques you are familiar with are insufficient for the problem at hand.

4.5.1 Confidence intervals for nearly normal point estimates

In Section 4.2, we used the point estimate $\bar{x}$ with a standard error $SE_{\bar{x}}$ to create a 95% confidence interval for the population mean:

$$\bar{x} \pm 1.96 \times SE_{\bar{x}} \tag{4.44}$$

We constructed this interval by noting that the sample mean is within 1.96 standard errors of the actual mean about 95% of the time. This same logic generalizes to any unbiased point estimate that is nearly normal. We may also generalize the confidence level by using a place-holder $z^\star$.

General confidence interval for the normal sampling distribution case
A confidence interval based on an unbiased and nearly normal point estimate is

$$\text{point estimate} \pm z^\star SE \tag{4.45}$$

where $z^\star$ is selected to correspond to the confidence level, and SE represents the standard error. The value $z^\star SE$ is called the *margin of error*.

Generally the standard error for a point estimate is estimated from the data and computed using a formula. For example, the standard error for the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

In this section, we provide the computed standard error for each example and exercise without detailing where the values came from. In future chapters, you will learn to fill in these and other details for each situation.

Example 4.46 In Exercise 4.1 on page 161, we computed a point estimate for the average difference in run times between men and women: $\bar{x}_{women} - \bar{x}_{men} = 14.48$ minutes. This point estimate is associated with a nearly normal distribution with standard error $SE = 2.78$ minutes. What is a reasonable 95% confidence interval for the difference in average run times?

The normal approximation is said to be valid, so we apply Equation (4.45):

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 14.48 \pm 1.96 \times 2.78 \quad \rightarrow \quad (9.03, 19.93)$$

Thus, we are 95% confident that the men were, on average, between 9.03 and 19.93 minutes faster than women in the 2012 Cherry Blossom Run. That is, the actual average difference is plausibly between 9.03 and 19.93 minutes with 95% confidence.

Example 4.47 Does Example 4.46 guarantee that if a husband and wife both ran in the race, the husband would run between 9.03 and 19.93 minutes faster than the wife?

Our confidence interval says absolutely nothing about individual observations. It only makes a statement about a plausible range of values for the *average* difference between all men and women who participated in the run.

⊙ **Exercise 4.48** What $z^\star$ would be appropriate for a 99% confidence level? For help, see Figure 4.10 on page 169.[34]

⊙ **Exercise 4.49** The proportion of men in the `run10Samp` sample is $\hat{p} = 0.45$. This sample meets certain conditions that ensure $\hat{p}$ will be nearly normal, and the standard error of the estimate is $SE_{\hat{p}} = 0.05$. Create a 90% confidence interval for the proportion of participants in the 2012 Cherry Blossom Run who are men.[35]

4.5.2 Hypothesis testing for nearly normal point estimates

Just as the confidence interval method works with many other point estimates, we can generalize our hypothesis testing methods to new point estimates. Here we only consider the p-value approach, introduced in Section 4.3.4, since it is the most commonly used technique and also extends to non-normal cases.

Hypothesis testing using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation.
2. Identify an appropriate point estimate of the parameter of interest.
3. Verify conditions to ensure the standard error estimate is reasonable and the point estimate is nearly normal and unbiased.
4. Compute the standard error. Draw a picture depicting the distribution of the estimate under the idea that H_0 is true. Shade areas representing the p-value.
5. Using the picture and normal model, compute the *test statistic* (Z score) and identify the p-value to evaluate the hypotheses. Write a conclusion in plain language.

⊙ **Exercise 4.50** A drug called sulphinpyrazone was under consideration for use in reducing the death rate in heart attack patients. To determine whether the drug was effective, a set of 1,475 patients were recruited into an experiment and randomly split into two groups: a control group that received a placebo and a treatment group that received the new drug. What would be an appropriate null hypothesis? And the alternative?[36]

We can formalize the hypotheses from Exercise 4.50 by letting $p_{control}$ and $p_{treatment}$ represent the proportion of patients who died in the control and treatment groups, respec-

[34] We seek $z^\star$ such that 99% of the area under the normal curve will be between the Z scores -$z^\star$ and $z^\star$. Because the remaining 1% is found in the tails, each tail has area 0.5%, and we can identify -$z^\star$ by looking up 0.0050 in the normal probability table: $z^\star = 2.58$. See also Figure 4.10 on page 169.

[35] We use $z^\star = 1.65$ (see Exercise 4.17 on page 170), and apply the general confidence interval formula:

$$\hat{p} \pm z^\star SE_{\hat{p}} \quad \rightarrow \quad 0.45 \pm 1.65 \times 0.05 \quad \rightarrow \quad (0.3675, 0.5325)$$

Thus, we are 90% confident that between 37% and 53% of the participants were men.

[36] The skeptic's perspective is that the drug does not work at reducing deaths in heart attack patients (H_0), while the alternative is that the drug does work (H_A).

tively. Then the hypotheses can be written as

$$H_0 : p_{control} = p_{treatment} \quad \text{(the drug doesn't work)}$$
$$H_A : p_{control} > p_{treatment} \quad \text{(the drug works)}$$

or equivalently,

$$H_0 : p_{control} - p_{treatment} = 0 \quad \text{(the drug doesn't work)}$$
$$H_A : p_{control} - p_{treatment} > 0 \quad \text{(the drug works)}$$

Strong evidence against the null hypothesis and in favor of the alternative would correspond to an observed difference in death rates,

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment}$$

being larger than we would expect from chance alone. This difference in sample proportions represents a point estimate that is useful in evaluating the hypotheses.

Example 4.51 We want to evaluate the hypothesis setup from Exericse 4.50 using data from the actual study.[37] In the control group, 60 of 742 patients died. In the treatment group, 41 of 733 patients died. The sample difference in death rates can be summarized as

$$\text{point estimate} = \hat{p}_{control} - \hat{p}_{treatment} = \frac{60}{742} - \frac{41}{733} = 0.025$$

This point estimate is nearly normal and is an unbiased estimate of the actual difference in death rates. The standard error of this sample difference is $SE = 0.013$. Evaluate the hypothesis test at a 5% significance level: $\alpha = 0.05$.

We would like to identify the p-value to evaluate the hypotheses. If the null hypothesis is true, then the point estimate would have come from a nearly normal distribution, like the one shown in Figure 4.22. The distribution is centered at zero since $p_{control} - p_{treatment} = 0$ under the null hypothesis. Because a large positive difference provides evidence against the null hypothesis and in favor of the alternative, the upper tail has been shaded to represent the p-value. We need not shade the lower tail since this is a one-sided test: an observation in the lower tail does not support the alternative hypothesis.

The p-value can be computed by using the Z score of the point estimate and the normal probability table.

$$Z = \frac{\text{point estimate} - \text{null value}}{SE_{\text{point estimate}}} = \frac{0.025 - 0}{0.013} = 1.92 \tag{4.52}$$

Examining Z in the normal probability table, we find that the lower unshaded tail is about 0.973. Thus, the upper shaded tail representing the p-value is

$$\text{p-value} = 1 - 0.973 = 0.027$$

Because the p-value is less than the significance level ($\alpha = 0.05$), we say the null hypothesis is implausible. That is, we reject the null hypothesis in favor of the alternative and conclude that the drug is effective at reducing deaths in heart attack patients.

[37] Anturane Reinfarction Trial Research Group. 1980. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. New England Journal of Medicine 302(5):250-256.

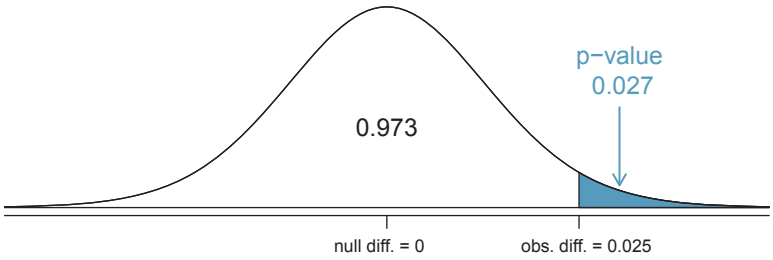

Figure 4.22: The distribution of the sample difference if the null hypothesis is true.

The Z score in Equation (4.52) is called a **test statistic**. In most hypothesis tests, a test statistic is a particular data summary that is especially useful for computing the p-value and evaluating the hypothesis test. In the case of point estimates that are nearly normal, the test statistic is the Z score.

Test statistic
A *test statistic* is a special summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. When a point estimate is nearly normal, we use the Z score of the point estimate as the test statistic. In later chapters we encounter situations where other test statistics are helpful.

4.5.3 Non-normal point estimates

We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:

- the sample size is too small for the normal approximation to be valid;
- the standard error estimate may be poor; or
- the point estimate tends towards some distribution that is not the normal distribution.

For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

4.5.4 When to retreat

Statistical tools rely on conditions. When the conditions are not met, these tools are unreliable and drawing conclusions from them is treacherous. The conditions for these tools typically come in two forms.

- **The individual observations must be independent.** A random sample from less than 10% of the population ensures the observations are independent. In experiments, we generally require that subjects are randomized into groups. If independence fails,

then advanced techniques must be used, and in some such cases, inference may not be possible.

- **Other conditions focus on sample size and skew.** For example, if the sample size is too small, the skew too strong, or extreme outliers are present, then the normal model for the sample mean will fail.

Verification of conditions for statistical tools is always necessary. Whenever conditions are not satisfied for a statistical technique, there are three options. The first is to learn new methods that are appropriate for the data. The second route is to consult a statistician.[38] The third route is to ignore the failure of conditions. This last option effectively invalidates any analysis and may discredit novel and interesting findings.

Finally, we caution that there may be no inference tools helpful when considering data that include unknown biases, such as convenience samples. For this reason, there are books, courses, and researchers devoted to the techniques of sampling and experimental design. See Sections 1.3-1.5 for basic principles of data collection.

4.6 Sample size and power (special topic)

The Type 2 Error rate and the magnitude of the error for a point estimate are controlled by the sample size. Real differences from the null value, even large ones, may be difficult to detect with small samples. If we take a very large sample, we might find a statistically significant difference but the magnitude might be so small that it is of no practical value. In this section we describe techniques for selecting an appropriate sample size based on these considerations.

4.6.1 Finding a sample size for a certain margin of error

Many companies are concerned about rising healthcare costs. A company may estimate certain health characteristics of its employees, such as blood pressure, to project its future cost obligations. However, it might be too expensive to measure the blood pressure of every employee at a large company, and the company may choose to take a sample instead.

Example 4.53 Blood pressure oscillates with the beating of the heart, and the systolic pressure is defined as the peak pressure when a person is at rest. The average systolic blood pressure for people in the U.S. is about 130 mmHg with a standard deviation of about 25 mmHg. How large of a sample is necessary to estimate the average systolic blood pressure with a margin of error of 4 mmHg using a 95% confidence level?

First, we frame the problem carefully. Recall that the margin of error is the part we add and subtract from the point estimate when computing a confidence interval. The margin of error for a 95% confidence interval estimating a mean can be written as

$$ME_{95\%} = 1.96 \times SE = 1.96 \times \frac{\sigma_{employee}}{\sqrt{n}}$$

[38] If you work at a university, then there may be campus consulting services to assist you. Alternatively, there are many private consulting firms that are also available for hire.