



CONTENTS

1	Statistical Inferences for Proportions	5
1.1	Interval Estimation and the Confidence Interval	5
1.2	Constructing a Confidence Interval for a Proportion	6
1.3	Inference: Tests of Significance	8
1.3.1	The Logic of a Significance Test	8
1.3.2	The Test Statistic and the p-value	8
1.3.3	Making a Decision	9
1.3.4	Type I and Type II Errors	9
2	Statistical Inferences for Means	11
2.1	The t-Distribution and Conditions for Inference about a Mean	11
2.1.1	Meet the t-Distribution	11
2.1.2	Conditions for Inference about a Mean	12
2.1.3	When the Data Comes in Pairs	13
2.2	Constructing and Justifying a Confidence Interval for a Population Mean	14
2.2.1	The Confidence Interval for a Population Mean	14
2.2.2	Interpreting and Justifying the Interval	15
2.3	Setting Up and Carrying Out a Test for a Population Mean	17
2.3.1	Setting Up the Test	17
2.3.2	Carrying Out the Test	18
3	Inference for Two Means	21
3.1	Confidence Intervals for a Difference in Two Means	21
3.1.1	Two Independent Samples	21

3.2	Significance Tests for a Difference in Two Means, and Selecting the Right Procedure	23
3.2.1	Setting Up and Carrying Out the Test	23
3.2.2	Selecting the Right Procedure	24
3.2.3	Common Errors	25
4	Inference for Two Proportions	27
4.1	The Difference Between Two Proportions	27
4.1.1	Variances Add, Even When You Subtract	27
4.2	Conditions	28
4.3	A Confidence Interval for $p_1 - p_2$	29
4.4	A Significance Test for $p_1 - p_2$	30
4.4.1	Pooling	30
4.5	Computing	31
5	Regression Basics for the Slope	35
5.1	The Population Regression Model	35
5.2	Estimating the Slope and Intercept	35
5.3	Computing in Excel	38
6	Statistical Inferences for Slopes	41
6.1	The Sampling Distribution of b_1 and the Confidence Interval for the Slope	41
6.2	Conditions for Inference on the Slope	44
6.3	Testing a Claim About the Slope	46
7	Evaluating the Fit of a Linear Regression Model	51
7.1	Look at the Residuals First	51
7.2	Splitting Up the Variation	52
7.3	The Coefficient of Determination	53
7.4	Root Mean Squared Error	54
7.5	The ANOVA Table and the F-Test	54
7.6	Grading a Line for Penguins	55
7.7	More Features, Higher R^2 , Worse Predictions	57
7.8	Limits of These Measures	58
8	Statistical Inferences: Chi-Square Distribution	61
8.1	The Chi-Square Distribution	61
8.1.1	Properties of the Chi-Square Distribution	62
8.1.2	Reading a Chi-Square Table	62
8.2	The Chi-Square Test for Goodness of Fit	63
8.2.1	Conditions	63

8.2.2	Stating the Hypotheses	64
8.3	The Chi-Square Test for Independence	65
8.3.1	Contingency Tables	65
8.3.2	Conditions	66
8.3.3	Hypotheses and Expected Counts	66
8.4	The Chi-Square Test for Homogeneity of Proportions	68
8.4.1	Conditions and Hypotheses	68
A	Answers to Exercises	71
Index		81

Statistical Inferences for Proportions

1.1 Interval Estimation and the Confidence Interval

Suppose you had to guess your own weight, and then guess a classmate's weight. You would probably feel more confident about your own guess. You might say 140 pounds for both, but you would trust your own number a lot more than the one for your classmate. A point estimate only gives you a single number. It does not tell you how much to trust that number.

A point estimate is a single value that is calculated from a sample. It serves as our best guess for a population parameter. Although it can be useful, it is incomplete. Two point estimates can look identical on paper while carrying very different amounts of certainty behind them, like the way your weight guess and your classmate's guess might both be 140 pounds while being backed by very different amounts of information.

This is where the *confidence interval* comes in. A confidence interval gives a range of plausible values for the population parameter that is built around the point estimate. The *margin of error* is the "plus or minus" amount that sets the width of that range. It is driven by how confident we want to be and how much the estimate varies from sample to sample.

A wider interval is more likely to capture the true value, but it also tells you less. This tradeoff between confidence and precision runs through everything in this chapter.

Before we build an interval, we check three conditions:

- **Random:** the sample comes from a random sample or a randomized experiment.
- **10% condition:** the sample is no more than 10% of the population, so the individuals are roughly independent.
- **Large counts:** $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$, so the sampling distribution is close to normal.

1.2 Constructing a Confidence Interval for a Proportion

When we discussed confidence intervals, we saw that a wider interval gives us more certainty at the cost of being precise. That tradeoff is captured by a single number called the *critical value*, written z^* .

Suppose a city council wants to estimate the proportion of residents who support building a new network of bike lanes. Before collecting any data, they need to decide how confident they want to be in their eventual estimate, say 95%. That confidence level determines z^* .

A 95% confidence level leaves 5% split evenly between the two tails of the standard normal distribution, so 2.5% in each tail. We need the value z^* such that the area to its right is 0.025, which means the area to its left is 0.975. Using a standard normal table or a calculator's inverse normal function, $z^* \approx 1.96$.

If the city council instead wanted 99% confidence, would z^* be larger or smaller than 1.96? (Larger: to catch the true value 99% of the time, the interval must reach farther out. For 99%, $z^* \approx 2.576$.)

Once we have z^* , we still need to decide how precise we want our estimate to be. An interval can technically be true and still be worthless. If the council's confidence interval for bike lane support turned out to be "somewhere between 5% and 95%," that would be almost useless for making a decision, even if it is a perfectly valid 95% confidence interval. The width of a useful interval, called the margin of error, depends on how large a sample the council collects. Larger samples produce narrower, more informative intervals for the same confidence level.

A confidence interval for a population proportion p has the form

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where $\hat{p}$ is the sample proportion, n is the sample size, and the square root term is the standard error. The margin of error is z^* times the standard error.

Example

The council surveys $n = 200$ residents and finds that 120 support the bike lanes, so $\hat{p} = \frac{120}{200} = 0.60$. For 95% confidence, $z^* = 1.96$. The standard error is $\sqrt{\frac{0.6(0.4)}{200}} \approx 0.0346$, and the margin of error is $1.96(0.0346) \approx 0.068$. The 95% confidence interval is 0.60 ± 0.068 , or $(0.532, 0.668)$. We are 95% confident that between 53.2% and 66.8% of residents support the bike lanes.

To plan a sample size, we solve the margin-of-error formula for n . Using the conservative guess $p^* = 0.5$ (which makes the standard error as large as possible), the sample size

needed for a margin of error E is

$$n = \left(\frac{z^*}{E} \right)^2 p^*(1 - p^*)$$

For example, to get a margin of error of 0.03 at 95% confidence, the council needs $n = \left(\frac{1.96}{0.03} \right)^2 (0.5)(0.5) \approx 1067.1$, so it should survey at least 1068 residents.

This is why sample size planning matters just as much as the confidence level. A researcher who wants a useful interval has to think ahead about how much data collecting that precision will require.

Computing in Excel

Put the counts in cells: n in B1 (200) and the number supporting in B2 (120). Then use these formulas:

- $\hat{p}$: =B2/B1, which gives 0.60.
- z^* : =NORM.S.INV(0.975), which gives 1.96.
- Margin of error: =B4*SQRT(B3*(1-B3)/B1), where B3 holds $\hat{p}$ and B4 holds z^* . This gives about 0.068.
- Interval: =B3-B5 and =B3+B5, which give 0.532 and 0.668.

Computing in Python

The `scipy` library finds z^* , and the `statsmodels` library builds the interval from the counts:

```

1 from scipy.stats import norm
2 from statsmodels.stats.proportion import proportion_confint
3
4 z_star = norm.ppf(0.975) # 1.96
5 low, high = proportion_confint(120, 200, alpha=0.05, method="normal")
6 print(f"z* = {z_star:.2f}")
7 print(f"95% CI: ({low:.3f}, {high:.3f})" # about (0.532, 0.668)

```

To plan a sample size, a short function solves the formula above:

```

1 import math
2
3 def sample_size(margin, z_star=1.96, p_star=0.5):
4     return math.ceil((z_star / margin) ** 2 * p_star * (1 - p_star))
5

```

```
6 | print(sample_size(0.03)) # 1068
```

1.3 Inference: Tests of Significance

Building a confidence interval answers “what is the population proportion likely to be?” But sometimes the question you actually care about is different: has something changed? *Statistical inference* is the process of using information from a sample to answer a specific question or make a decision about a population parameter, and a *test of significance* is the tool built for exactly this kind of question.

1.3.1 The Logic of a Significance Test

A streaming service offers a free trial. Historically, 20% of trial users convert to paying subscribers. The company redesigns its sign-up page and wants to know: did the conversion rate go up?

A significance test starts by assuming nothing changed, then asks whether the data would be surprising under that assumption.

- The *null hypothesis*, H_0 , is the “nothing changed” claim: $H_0 : p = 0.20$.
- The *alternative hypothesis*, H_a , is what we are looking for evidence of: $H_a : p > 0.20$. (If we cared about a change in either direction, we would write $H_a : p \neq 0.20$, a *two-sided* test.)

After the redesign, a random sample of $n = 400$ trial users shows 98 conversions, so $\hat{p} = 0.245$. That is higher than 0.20, but samples vary. Could a jump this big happen by chance if the true rate were still 20%?

1.3.2 The Test Statistic and the p-value

If H_0 is true, $\hat{p}$ is approximately normal with mean $p_0 = 0.20$ and standard deviation $\sqrt{\frac{p_0(1-p_0)}{n}}$. (Notice that we use p_0 , the value from H_0 , not $\hat{p}$. The conditions are the same as before, but the large-counts check uses p_0 : $np_0 = 80 \geq 10$ and $n(1 - p_0) = 320 \geq 10$.) The *test statistic* measures how many standard deviations $\hat{p}$ is from p_0 :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.245 - 0.20}{\sqrt{\frac{(0.2)(0.8)}{400}}} = \frac{0.045}{0.02} = 2.25$$

The *p-value* is the probability, assuming H_0 is true, of getting a result at least as extreme as the one observed. Here that is $P(Z \geq 2.25) \approx 0.012$. If the redesign did nothing, only about 1.2% of samples would show a conversion rate this high.

1.3.3 Making a Decision

Before collecting data, we choose a *significance level* α , the cutoff for “too surprising to be chance.” The most common choice is $\alpha = 0.05$.

- If the *p-value* $< \alpha$, **reject** H_0 : the data give convincing evidence for H_a .
- If the *p-value* $\geq \alpha$, **fail to reject** H_0 : the data do not give convincing evidence for H_a . (This does *not* prove H_0 is true.)

Since $0.012 < 0.05$, we reject H_0 . There is convincing evidence that the redesign raised the conversion rate above 20%.

In Python:

```

1 from scipy.stats import norm
2 import math
3
4 p0, n, successes = 0.20, 400, 98
5 p_hat = successes / n
6 z = (p_hat - p0) / math.sqrt(p0 * (1 - p0) / n)
7 p_value = 1 - norm.cdf(z)           # one-sided, H_a: p > p0
8 print(round(z, 3), round(p_value, 4)) # 2.25 0.0122

```

In Excel, the *p-value* is $=1-\text{NORM.S.DIST}(2.25, \text{TRUE})$. For a two-sided test, double the one-sided tail area.

1.3.4 Type I and Type II Errors

Because tests rely on samples, they can be wrong in two ways:

- A *Type I error* is rejecting H_0 when it is actually true: concluding the redesign helped when it did not. The probability of a Type I error is α .
- A *Type II error* is failing to reject H_0 when H_a is actually true: missing a real improvement.

The *power* of a test is the probability of correctly rejecting H_0 when H_a is true. Power

goes up with a larger sample size, a bigger true effect, or a larger α . For example, if the redesign really raised the conversion rate to 25%, this test (with $n = 400$ and $\alpha = 0.05$) would detect it about 79% of the time; the other 21% of the time it would make a Type II error.

Exercise 1 **The Cafeteria Survey**

The cafeteria claims that 60% of students like the new menu. The student council surveys a random sample of 150 students, and 81 say they like it. At $\alpha = 0.05$, is there convincing evidence that the true proportion differs from 60%?

Working Space

Answer on Page 71

Exercise 2 **Name the Error**

In the streaming-service example, suppose the redesign actually had no effect, but the test rejected H_0 anyway. What kind of error is this, and what is its consequence for the company?

Working Space

Answer on Page 71

Statistical Inferences for Means

In Statistical Inferences for Proportions, every question came down to a percentage: what proportion of trial users converted to paying customers, and whether that proportion changed after a redesign. Not every question reduces to a percentage.

Sometimes what you want to know is a genuine measurement, how long something takes, how much something weighs, how far something travels, on average.

A quantitative variable like this calls for the same two tools you already trust, a confidence interval to estimate a mean and a significance test to check a claim about one, but neither tool behaves the way it did for proportions. This is because sometimes you are missing crucial information: you almost never know the population's true standard deviation, only what your own sample can tell you. That gap forces a new distribution into the picture, which looks like the normal curve.

2.1 The t -Distribution and Conditions for Inference about a Mean

Suppose the streaming service now wants to know something different from its conversion rate: how long, on average, a subscriber watches in a single sitting. A product analyst pulls a random sample of viewing sessions and calculates a sample mean watch time, $\bar{x}$. To turn that one number into a confidence interval or a significance test, you need the standard error of $\bar{x}$, and the standard error depends on the population standard deviation, σ . Here is the catch: nobody knows σ . The population's true spread in viewing habits is exactly the kind of thing you would need a census to pin down, and if you had a census, you would not need inference in the first place.

2.1.1 Meet the t -Distribution

The sample standard deviation s estimates σ , but substituting s changes $\frac{\bar{x} - \mu}{s/\sqrt{n}}$ from a normal distribution to a t -distribution.

The t -distribution is continuous, symmetric, and centered at 0, but shorter and thicker-tailed than the normal distribution. Its shape depends on *degrees of freedom*, abbreviated df . For a single mean, $df = n - 1$. As n grows, the t -distribution converges to the standard normal distribution.

The critical value t^* replaces z^* and is read from a t -table using both df and the tail area.

For $n = 18$ ($df = 17$), a 90 percent confidence level gives $t^* = 1.740$, versus $z^* = 1.645$ for a known- σ case. At $df = 100$, $t^* = 1.660$, much closer to $z^* = 1.645$.

Exercise 3 Reading the t -Table

A logistics company samples $n = 25$ delivery times and wants to construct a 95 percent confidence interval for the true mean delivery time. State the degrees of freedom, and find t^* .

Working Space

Answer on Page 71

2.1.2 Conditions for Inference about a Mean

Two conditions must hold for a t -procedure to be valid.

Independence: the data come from a random sample or randomized experiment, and if sampling without replacement, the sample size is no more than 10 percent of the population.

Normal or large sample: the sampling distribution of $\bar{x}$ is approximately normal. Check this using a dotplot, stemplot, or normal probability plot for strong skew or outliers, or rely on the population being known normal, or $n \geq 30$.

Use t , not z , unless σ is explicitly given, a case that rarely occurs in practice or on the AP exam.

- Independence: random sample or randomized experiment, and $n \leq 10\%$ of the population when sampling without replacement.
- Normal or large sample: no strong skew or outliers in the data, the population is known to be normal, or $n \geq 30$.
- Default to a t -procedure. Only switch to z if σ is explicitly given.

Exercise 4 **Spot the Error**

A student is asked to construct a confidence interval for a population mean using a sample of $n = 22$ measurements. The student writes, "Since $n = 22$ is less than 30, I cannot use a t-interval. I need to collect a bigger sample first." Explain what is wrong with this reasoning.

Working Space

Answer on Page 71

2.1.3 When the Data Comes in Pairs

Some data comes in pairs: two measurements per subject, such as watch time before and after a recommendation-algorithm update for the same subscribers. The two measurements within a pair are not independent.

Data like this is called *matched pairs*. Subtract to get one difference per subject, $d_i = x_i - y_i$, reducing the data to a single sample of differences. The t-distribution, degrees of freedom, and the independence and normality conditions all apply directly to that sample of differences. Independence now refers to the pairs being independent of one another, not the two measurements within a pair.

Matched pairs is not the same as comparing two independent means: two independent means needs two separate random samples, while matched pairs needs one sample of paired subjects.

Exercise 5 **Paired or Independent?**

For each scenario, state whether the data described would be analyzed as matched pairs or as two independent samples, and justify your answer in one sentence.

Working Space

- (a) Twenty subscribers each have their nightly watch time recorded for one week before a recommendation algo-

rithm update and one week after.

- (b) A researcher randomly splits 200 subscribers into two groups, shows one group only comedy titles and the other only drama titles for a month, and compares the two groups' average watch time.

Answer on Page 72

2.2 Constructing and Justifying a Confidence Interval for a Population Mean

You now have everything you need to actually build an interval. Every confidence interval takes the shape:

$$\text{point estimate} \pm \text{margin of error}$$

2.2.1 The Confidence Interval for a Population Mean

The point estimate for a population mean μ is the sample mean, $\bar{x}$. The margin of error is the critical value t^* times the standard error of $\bar{x}$, which is $\frac{s}{\sqrt{n}}$. Putting the two together,

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

is a $(1 - \alpha)100\%$ confidence interval for μ , using t^* from a t -distribution with $n - 1$ degrees of freedom.

Back to the streaming service. The analyst samples $n = 24$ viewing sessions and finds a sample mean watch time of $\bar{x} = 47.8$ minutes with a sample standard deviation of $s = 9.6$ minutes. A dotplot of the sample shows no strong skew or outliers, and the sessions were pulled at random from the full session log, well under 10 percent of all sessions logged that month. Both conditions are satisfied, so a t -interval is appropriate.

For a 95 percent confidence interval, $df = 23$ and $t^* = 2.069$. The standard error is $\frac{9.6}{\sqrt{24}} \approx 1.960$, so the margin of error is $2.069 \times 1.960 \approx 4.05$ minutes. The interval is

$$47.8 \pm 4.05 \Rightarrow (43.75, 51.85)$$

so the analyst estimates the true mean session watch time to be between 43.75 and 51.85 minutes.

Exercise 6 Building a t -Interval

A city transportation office samples $n = 16$ commute times and finds $\bar{x} = 28.4$ minutes and $s = 6.2$ minutes. Conditions for inference are satisfied. Construct a 98 percent confidence interval for the true mean commute time.

Working Space

Answer on Page 72

2.2.2 Interpreting and Justifying the Interval

An interpretation references two things: the confidence level and the population. For the streaming service example: the analyst is 95 percent confident that the true mean watch time per session, across all sessions, falls between 43.75 and 51.85 minutes.

Do not say there is a 95 percent probability that μ falls in this interval. Once calculated, μ either is or is not inside it. The 95 percent describes the method: if the sampling process were repeated many times, about 95 percent of the resulting intervals would capture the true μ .

A confidence interval can justify a claim by checking whether the claimed value falls inside or outside it. If the product team's target is an average watch time of at least 45 minutes, the interval (43.75, 51.85) dips below 45, so it does not provide convincing evidence that the true mean is at least 45 minutes.

For a fixed confidence level, interval width shrinks as n grows, roughly proportional to $\frac{1}{\sqrt{n}}$: quadrupling n from 24 to 96 shrinks the margin of error from about 4.05 to about 1.95 minutes, since $\sqrt{4} = 2$. For a fixed sample, raising the confidence level widens the interval.

Computing in Excel

Excel does not have a single built-in function that hands back a finished confidence interval, but two functions cover everything once you have n , $\bar{x}$, and s .

If you have the raw session lengths in a column, say A2:A25, get the three summary values first:

- =AVERAGE(A2:A25) for $\bar{x}$
- =STDEV.S(A2:A25) for s
- =COUNT(A2:A25) for n

Suppose those land in cells B1 ($n = 24$), B2 ($\bar{x} = 47.8$), and B3 ($s = 9.6$), with the confidence level in B4 (0.95).

The most direct route is CONFIDENCE.T, which returns the margin of error by itself:

- =CONFIDENCE.T(1-B4, B3, B1) returns 4.054, the margin of error.

Then the interval bounds are just =B2-C1 and =B2+C1, where C1 holds that margin of error.

If you would rather see each piece of the formula spelled out, build it manually instead:

- =B1-1 for df
- =T.INV.2T(1-B4, C1) for t^* , where C1 is the df cell (T.INV.2T takes the two-tailed probability, so use $1 - 0.95 = 0.05$ directly, not 0.025)
- =B3/SQRT(B1) for the standard error
- =C2*C3 for the margin of error, where C2 and C3 hold t^* and the standard error

Both routes return the same margin of error, 4.054, and the same interval, (43.75, 51.85). The manual version is worth building at least once, since it makes visible exactly where the t -distribution and the sample size are doing their work; CONFIDENCE.T is what you would reach for afterward, once you trust the formula.

Exercise 7 A Matched-Pairs Interval

A call center tests a new training program on $n = 12$ employees, recording each employee's average call-handling time before and after training. The sample of differences (after minus before) has $\bar{d} = -3.2$ minutes and $s_d = 4.5$ minutes. Conditions for inference on the differences are satisfied. Construct a 90 percent confidence interval for μ_d , and

Working Space

state what it suggests about the training program.

Answer on Page 72

Exercise 8 Spot the Error

A student reports the streaming-service interval (43.75, 51.85) and writes, "There is a 95 percent probability that the true mean watch time is between 43.75 and 51.85 minutes." Explain what is wrong with this interpretation, and rewrite it correctly.

Working Space

Answer on Page 72

2.3 Setting Up and Carrying Out a Test for a Population Mean

2.3.1 Setting Up the Test

A test for a population mean starts the same way a test for a proportion did: state hypotheses about the parameter, never about a sample statistic. The null hypothesis takes the form $H_0 : \mu = \mu_0$, where μ_0 is whatever value is being challenged. The alternative hypothesis is $H_a : \mu < \mu_0$, $H_a : \mu > \mu_0$, or $H_a : \mu \neq \mu_0$, depending on which direction the question points.

Choosing the method and verifying conditions works exactly as it did in the last section: check independence, check that the population is normal or the sample is large enough and free of strong skew or outliers, and default to a t-test unless σ is somehow known. If the data arrives as matched pairs, subtract first, then run everything that follows on the sample of differences, testing $H_0 : \mu_d = 0$ against whichever alternative fits the question.

2.3.2 Carrying Out the Test

The test statistic for a population mean is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

with $n - 1$ degrees of freedom. Find the p-value from the t-distribution, matching the tail(s) to H_a , and compare to α . Reject H_0 if p-value $< \alpha$.

A streaming service wants to know whether average session watch time has increased above the old baseline of 42 minutes after a redesign. Sample: $n = 30$, $\bar{x} = 44.6$ minutes, $s = 8.2$ minutes. Conditions are satisfied.

$$H_0 : \mu = 42 \quad H_a : \mu > 42$$

$df = 29$, standard error $= \frac{8.2}{\sqrt{30}} \approx 1.497$, so

$$t = \frac{44.6 - 42}{1.497} \approx 1.737$$

p-value ≈ 0.0465 . Since $0.0465 < 0.05$, reject H_0 : sufficient evidence that mean session watch time increased above 42 minutes. A smaller p-value is stronger evidence against H_0 than one just under α .

Matched-pairs example: call-handling time, $\bar{d} = -3.2$ minutes, $s_d = 4.5$ minutes, $n = 12$. Testing $H_0 : \mu_d = 0$ against $H_a : \mu_d < 0$ gives $df = 11$, $t \approx -2.463$, p-value ≈ 0.0157 . At $\alpha = 0.05$, reject H_0 : sufficient evidence the training program decreased average call-handling time. For a two-sided test at α equal to 1 minus the confidence level, the interval and the test always agree.

Exercise 9 Setting Up and Testing a Claim

A bakery's recipe is supposed to produce loaves with a mean weight of 500 grams. A random sample of $n = 16$ loaves has $\bar{x} = 498.2$ grams and $s = 6.4$ grams, with no strong skew or outliers in the sample. At $\alpha = 0.05$, is there sufficient evidence that the true mean loaf

Working Space

weight differs from 500 grams?

Answer on Page 73

Exercise 10 A Matched-Pairs Test

A wellness program measures resting heart rate for $n = 15$ participants before and after an eight-week exercise regimen. The sample of differences (before minus after) has $\bar{d} = 4.8$ beats per minute and $s_d = 6.1$ beats per minute. Conditions for inference on the differences are satisfied. At $\alpha = 0.05$, is there sufficient evidence that the regimen decreased mean resting heart rate?

Working Space

Answer on Page 73

Exercise 11 Spot the Error

A student sets up a test with the hypotheses $H_0 : \bar{x} = 500$ and $H_a : \bar{x} \neq 500$. Explain what is wrong with this setup.

Working Space

Answer on Page 73

So far, every procedure has looked at one population, or at pairs of measurements on the same subjects. The next chapter, Inference for Two Means, compares two independent groups.

Inference for Two Means

Do students who listen to music while studying score differently from those who study in silence? Do paid subscribers watch longer than free users? Most interesting questions in statistics are comparisons. In the Statistical Inferences for Means chapter, you built confidence intervals and tests for a single mean, and for paired data. This chapter compares the means of two independent groups.

3.1 Confidence Intervals for a Difference in Two Means

In the Statistical Inferences for Means chapter, every example looked at one population at a time. Just as often, the real question is a comparison: not “what is the mean watch time,” but “is the mean watch time different between two groups.”

3.1.1 Two Independent Samples

Comparing average watch time between free-tier and paid-tier subscribers, two disjoint groups, means the samples are independent. Let μ_1, μ_2 be the true means. The parameter of interest is $\mu_1 - \mu_2$, point estimate $\bar{x}_1 - \bar{x}_2$.

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Degrees of freedom come from the Welch-Satterthwaite approximation: between $\min(n_1 - 1, n_2 - 1)$ and $n_1 + n_2 - 2$.

Conditions: both samples independently and randomly selected, each $n \leq 10\%$ of its population if sampling without replacement, and each sample free of strong skew/outliers or $n \geq 30$.

Sample: free-tier $n_1 = 20$, $\bar{x}_1 = 35.2$, $s_1 = 8.4$; paid-tier $n_2 = 18$, $\bar{x}_2 = 44.6$, $s_2 = 10.1$.

Conditions satisfied. For a 95 percent CI,

$$SE = \sqrt{\frac{8.4^2}{20} + \frac{10.1^2}{18}} \approx 3.032, \quad df \approx 33, \quad t^* \approx 2.035$$

Margin of error ≈ 6.17 minutes, point estimate $= -9.4$ minutes:

$$-9.4 \pm 6.17 \Rightarrow (-15.57, -3.23)$$

95 percent confident the true difference (free-tier minus paid-tier) is between -15.57 and -3.23 minutes. Since the entire interval is negative, paid-tier subscribers watch more per session on average.

Exercise 12 Building a Two-Sample Interval

A tech reviewer compares battery life between two phone models. Model A: $n_1 = 15$, $\bar{x}_1 = 11.2$ hours, $s_1 = 1.8$ hours. Model B: $n_2 = 17$, $\bar{x}_2 = 9.8$ hours, $s_2 = 2.1$ hours. The two samples are independent, and conditions for inference are satisfied. Construct a 90 percent confidence interval for the true difference in mean battery life, Model A minus Model B. Technology gives $df \approx 29$.

Working Space

Answer on Page 73

Exercise 13 Justifying a Claim

Using the battery-life interval from the previous exercise, does the data provide convincing evidence that Model A's true mean battery life exceeds Model B's by more than one hour? Explain using the interval.

Working Space

Answer on Page 73

Exercise 14 Spot the Error

A student compares two independent samples with visibly different spreads, one sample's dotplot is tightly clustered, the other is spread out much wider, and writes, "Since both populations are approximately normal, I will pool the two sample standard deviations into one estimate of σ and use that in my confidence interval." Explain what is wrong with this reasoning for an AP Statistics response.

*Working Space**Answer on Page 74***3.2 Significance Tests for a Difference in Two Means, and Selecting the Right Procedure****3.2.1 Setting Up and Carrying Out the Test**

Testing a claim about $\mu_1 - \mu_2$ follows the same four-step process, usually $H_0 : \mu_1 - \mu_2 = 0$, with $H_a : \mu_1 - \mu_2 < 0, > 0$, or $\neq 0$.

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

using the same Welch-approximated df and conditions as the two-sample interval.

Free-tier versus paid-tier, same samples as before: $n_1 = 20$, $\bar{x}_1 = 35.2$, $s_1 = 8.4$; $n_2 = 18$, $\bar{x}_2 = 44.6$, $s_2 = 10.1$. Testing whether paid-tier watches more on average:

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 < 0$$

SE = 3.032 (as before), so

$$t = \frac{35.2 - 44.6}{3.032} \approx -3.100$$

df ≈ 33 , one-sided p-value ≈ 0.00196 . Reject H_0 : convincing evidence paid-tier subscribers

watch more per session on average. For the matching two-sided test, the test and the interval always agree.

Exercise 15 Running a Two-Sample Test

Two store locations track customer wait times. Store 1: $n_1 = 22$, $\bar{x}_1 = 8.3$ minutes, $s_1 = 2.1$ minutes. Store 2: $n_2 = 25$, $\bar{x}_2 = 9.6$ minutes, $s_2 = 2.8$ minutes. The samples are independent, and conditions for inference are satisfied. Test, at $\alpha = 0.05$, whether Store 1's mean wait time is shorter. Technology gives $df \approx 44$.

Working Space

Answer on Page 74

3.2.2 Selecting the Right Procedure

By this point you have several inference procedures for means and, from the last chapter, inference for a proportion. The hardest part of a real problem is often not the arithmetic; it is recognizing which procedure actually fits. Figure 3.1 lays out the same three questions, in order, that will get you there every time.

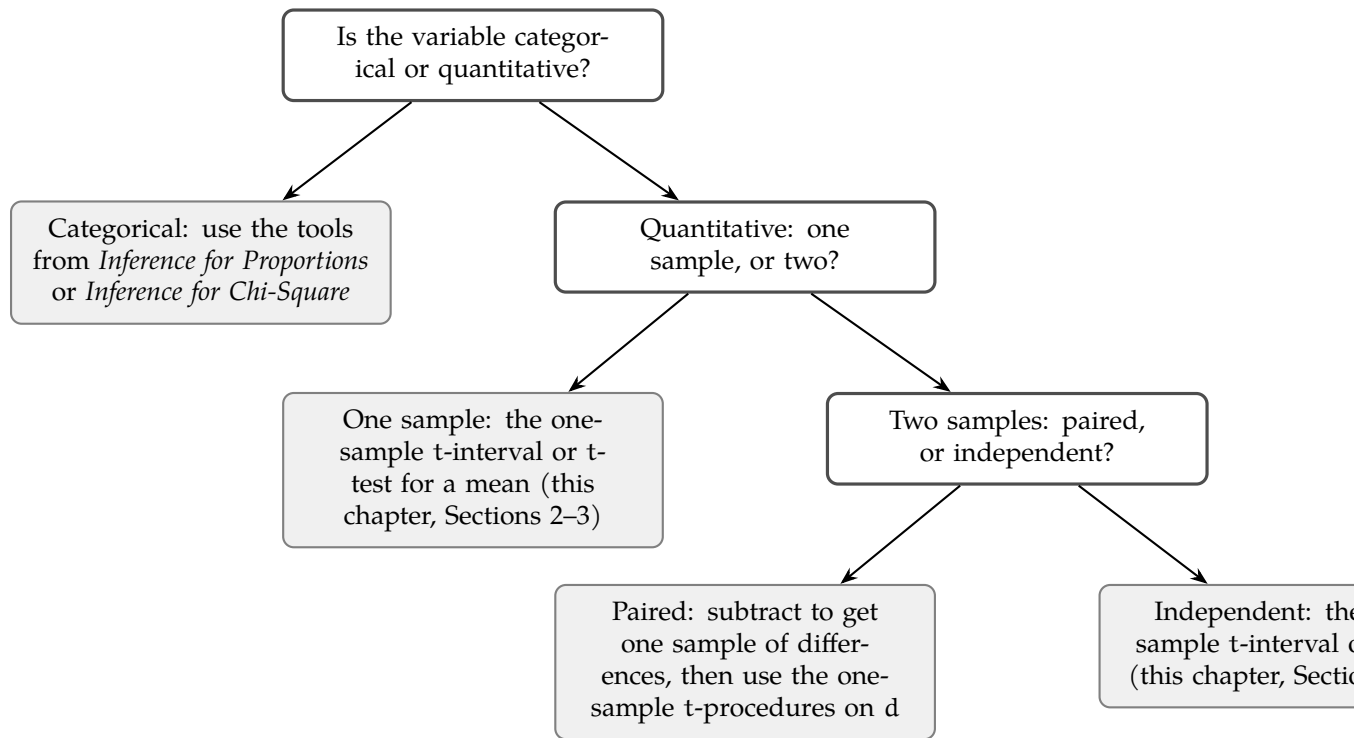


Figure 3.1: Selecting an inference procedure for a single quantitative or categorical variable

Categorical or quantitative is usually obvious from the variable itself. Paired or independent is worth slowing down for: does each measurement in one group have a natural partner in the other, or could the two groups' labels be shuffled without anything becoming meaningless? Same subscriber measured twice: shuffling is meaningless, so paired. Two disjoint groups: shuffling changes nothing, so independent.

3.2.3 Common Errors

Type I error, Type II error, and power all carry over from the last chapter without any changes to their definitions, just a change in what the parameter is. A Type I error here means rejecting $H_0 : \mu_1 = \mu_2$ when the two population means really are equal, concluding the streaming service's paid tier watches more when it actually does not. A Type II error means failing to reject H_0 when there really is a difference, missing a real effect. Power is still the probability of correctly detecting a real difference when one exists, and it still increases with a larger sample size, a larger true difference, or a higher significance level, exactly as it did for proportions.

Exercise 16 Naming the Error

In the store wait-time test from the pre-

Working Space

vious exercise, the data led to rejecting H_0 in favor of Store 1 having a shorter mean wait time. Suppose, unknown to the researcher, the two stores' true mean wait times are actually equal. What type of error, if any, occurred? Explain.

Answer on Page 74

Exercise 17 Choosing the Procedure

For each scenario, state which of the four procedures from Figure 3.1 applies: quantitative one-sample, quantitative paired, quantitative independent-samples, or categorical.

Working Space

- (a) A gym wants to know whether the average weight lifted by 30 randomly selected members changed after switching to a new equipment brand, measuring the same members' max lift before and after the switch.
- (b) A researcher wants to compare the proportion of two different email subject lines that get opened.
- (c) A nutritionist samples 40 adults at random and wants to estimate the true mean daily sugar intake in the population.

Answer on Page 74

That decision flowchart will keep earning its keep. Later chapters add inference for regression slopes and chi-square tests for categorical variables with more than two outcomes, and the first question will always be the same: which procedure fits the shape of the data in front of you?

Inference for Two Proportions

Does a text reminder make students more likely to return a permission slip? Do juniors get enough sleep more often than seniors? Does a new vaccine prevent infection better than a placebo? Each question compares two groups, and each answer is a proportion: a fraction of “yes” in one group against a fraction of “yes” in the other.

The Statistical Inferences for Proportions chapter studied one proportion at a time, and the Inference for Two Means chapter compared two means. This chapter puts the two ideas together: confidence intervals and significance tests for the *difference* between two population proportions. The reasoning is the same as before. The only new piece is figuring out how much a difference varies from sample to sample.

4.1 The Difference Between Two Proportions

Suppose a school counselor wants to compare how many juniors and seniors get at least 8 hours of sleep on school nights. Let p_1 be the true proportion of juniors and p_2 the true proportion of seniors. The parameter of interest is the difference

$$p_1 - p_2$$

The counselor surveys random samples: 84 of 240 juniors ($\hat{p}_1 = 0.35$) and 54 of 200 seniors ($\hat{p}_2 = 0.27$) say they sleep 8 hours or more. The point estimate is $\hat{p}_1 - \hat{p}_2 = 0.08$: juniors are 8 percentage points more likely to get enough sleep *in these samples*. Would another pair of samples show the same gap? To know how much to trust 0.08, we need the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

4.1.1 Variances Add, Even When You Subtract

From the Sampling Distributions and the Central Limit Theorem chapter, each sample proportion has mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. The Random Variables chapter showed that for independent random variables, variances add. That is true for differences as well as sums: $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$, because subtracting a random quantity adds just as much uncertainty as adding it. So:

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$

For independent random samples,

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2, \quad \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

The distribution is approximately normal when the large-counts condition holds in both groups.

It is tempting to subtract the two standard deviations instead. A simulation shows how wrong that would be. Here are 10,000 pairs of surveys with $p_1 = 0.35$ and $p_2 = 0.27$:

```

1 import numpy as np
2
3 rng = np.random.default_rng(3)
4 p1, n1 = 0.35, 240      # juniors
5 p2, n2 = 0.27, 200      # seniors
6
7 # 10,000 pairs of surveys
8 p_hat_1 = rng.binomial(n1, p1, size=10_000) / n1
9 p_hat_2 = rng.binomial(n2, p2, size=10_000) / n2
10 diff = p_hat_1 - p_hat_2
11
12 theory = np.sqrt(p1 * (1 - p1) / n1 + p2 * (1 - p2) / n2)
13 print(f"mean of differences: {diff.mean():.4f} (p1 - p2 = {p1 - p2:.2f})")
14 print(f"sd of differences: {diff.std():.4f} (formula: {theory:.4f})")
15 print(f"sd if you subtracted the sds: {abs(np.sqrt(p1*(1-p1)/n1) -
    ↪ np.sqrt(p2*(1-p2)/n2)):.4f}")

```

```

mean of differences: 0.0794 (p1 - p2 = 0.08)
sd of differences: 0.0439 (formula: 0.0440)
sd if you subtracted the sds: 0.0006

```

The differences really do spread out by about 0.044, just as the formula says. Subtracting the standard deviations would claim the difference is almost perfectly predictable, which is nonsense: two uncertain numbers give a *more* uncertain difference, not a less uncertain one.

4.2 Conditions

Before any two-proportion interval or test, check:

- **Random:** the data come from two independent random samples, or from two groups in a randomized experiment.

- **10%:** when sampling without replacement, each sample is no more than 10% of its population. (This is not needed for a randomized experiment.)
- **Large counts:** for a confidence interval, the four counts $n_1\hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1 - \hat{p}_2)$ are all at least 10. In other words, each group has at least 10 “yes” answers and at least 10 “no” answers.

4.3 A Confidence Interval for $p_1 - p_2$

We do not know p_1 and p_2 , so we estimate the standard deviation with the sample proportions. The result is the *standard error* of the difference, and the interval has the familiar shape: estimate $\pm$ critical value $\times$ standard error.

Two-Proportion z-Interval

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Example

For the sleep survey, the counts are 84 and 156 for juniors and 54 and 146 for seniors, all at least 10, and the samples are random and less than 10% of each class in a large district. The standard error is

$$SE = \sqrt{\frac{0.35(0.65)}{240} + \frac{0.27(0.73)}{200}} \approx 0.0440$$

For 95% confidence, $z^* = 1.96$, so the margin of error is $1.96(0.0440) \approx 0.086$, and the interval is

$$0.08 \pm 0.086, \quad \text{or} \quad (-0.006, 0.166)$$

We are 95% confident that the proportion of juniors who sleep at least 8 hours is between 0.6 percentage points *lower* and 16.6 points *higher* than the proportion of seniors.

Look at that interval closely. It contains 0. So although the samples showed an 8-point gap, the data are consistent with juniors and seniors sleeping equally well. An interval for a difference that contains 0 does not give convincing evidence of any difference at all; one that lies entirely above (or below) 0 does.

4.4 A Significance Test for $p_1 - p_2$

A school wants to know whether texting families a reminder raises the rate of returned permission slips. In a randomized experiment, 150 families get a reminder text and 150 do not. In the text group, 102 return the slip ($\hat{p}_1 = 0.68$); in the other group, 84 do ($\hat{p}_2 = 0.56$). Is a 12-point gap convincing?

The hypotheses are

$$H_0 : p_1 - p_2 = 0 \quad H_a : p_1 - p_2 > 0$$

where p_1 and p_2 are the true return rates with and without a reminder.

4.4.1 Pooling

A significance test always asks, “How surprising is this result *if H_0 is true?*” If H_0 is true, the two groups share a single return rate. Our best estimate of that common rate combines both samples into one *pooled proportion*:

$$\hat{p}_c = \frac{\text{total successes}}{\text{total sample size}} = \frac{102 + 84}{150 + 150} = \frac{186}{300} = 0.62$$

In the test, we use $\hat{p}_c$ in place of both $\hat{p}_1$ and $\hat{p}_2$ in the standard error, and in the large-counts check: $150(0.62) = 93$ and $150(0.38) = 57$, both at least 10.

Two-Proportion z-Test

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c(1 - \hat{p}_c) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Pool for a test (because H_0 says the proportions are equal), but not for an interval (because an interval makes no such assumption). This is the single most common mistake on two-proportion problems.

Example

$$SE = \sqrt{0.62(0.38) \left(\frac{1}{150} + \frac{1}{150} \right)} \approx 0.0560, \quad z = \frac{0.68 - 0.56}{0.0560} \approx 2.14$$

The p-value is $P(Z \geq 2.14) \approx 0.016$. Since $0.016 < 0.05$, we reject H_0 . There is convincing evidence that reminder texts raise the return rate.

Because families were *randomly assigned* to get a text or not, the two groups differed only

by chance and by the text itself. So we can conclude that the text *caused* the higher return rate. In the sleep survey, which was an observational study, even a significant difference would not show that being a junior causes better sleep. (The Experimental Design chapter explains why random assignment makes this difference.)

4.5 Computing

Excel. With $\hat{p}_1$, $\hat{p}_2$, n_1 , and n_2 in cells, the standard error and interval are ordinary formulas, and the p-value of a one-sided test is $=1-\text{NORM.S.DIST}(z, \text{TRUE})$. For the test above, $=1-\text{NORM.S.DIST}(2.14, \text{TRUE})$ gives about 0.016.

Python. The statsmodels library does both procedures from the counts:

```

1 from statsmodels.stats.proportion import proportions_ztest,
   ↪ confint_proportions_2indep
2
3 # Confidence interval: juniors vs. seniors who sleep 8+ hours
4 low, high = confint_proportions_2indep(84, 240, 54, 200, method="wald")
5 print(f"95% CI for p1 - p2: ({low:.3f}, {high:.3f})")
6
7 # Significance test: text reminder vs. no reminder
8 z, p_value = proportions_ztest([102, 84], [150, 150], alternative="larger")
9 print(f"z = {z:.2f}, p-value = {p_value:.4f}")

```

95% CI for p1 - p2: (-0.006, 0.166)
 z = 2.14, p-value = 0.0161

The option `method="wald"` gives the textbook interval from this chapter; statsmodels also offers more accurate variations.

Exercise 18 Biking to School

At School A, a random sample shows that 45 of 150 students bike to school. At School B, 30 of 160 do. Construct and interpret a 90% confidence interval for the difference in the proportions of students who bike. Is there convincing evidence of a difference?

Working Space

Answer on Page 75

Exercise 19 A New Math App

In a randomized experiment, 200 students used a new math app and 200 used the old one. On the unit test, 132 of the new-app students passed and 110 of the old-app students passed. Test, at $\alpha = 0.05$, whether the pass rates differ (a two-sided test).

Working Space

Answer on Page 75

Exercise 20 Reading an Interval

A 95% confidence interval for $p_1 - p_2$, the difference in the proportions of adults and teens who prefer a new soda, is $(-0.02, 0.10)$. A news site reports, "Adults prefer the new soda more than teens do." Is that justified?

Working Space

Answer on Page 75

Exercise 21 To Pool or Not to Pool?

Explain why a two-proportion significance test of $H_0 : p_1 = p_2$ uses the pooled proportion $\hat{p}_c$, but a confidence interval for $p_1 - p_2$ does not.

Working Space

Answer on Page 75

Exercise 22 **What Could Go Wrong?**

In the reminder-text experiment, the school rejected H_0 . Describe a Type I error and a Type II error in this context. Which one could the school have made?

Working Space

Answer on Page 76

Regression Basics for the Slope

A sleep researcher pulls data from a sample of 8 teenagers: daily screen time, in hours, and how many hours they slept that night. The researcher plots the points, fits a line, and finds a slope of $b_1 = -0.16$: for every extra hour of screen time, sleep drops by about ten minutes. Does this mean the effect is real across all teenagers, or could a sample this size show a slope like this by chance?

Recall from The Regression Line chapter that the least-squares line $\hat{y} = b_0 + b_1x$ describes your sample: descriptive statistics, not inference.

5.1 The Population Regression Model

A sample slope b_1 estimates an unknown population slope β_1 , the same way $\hat{p}$ estimates p and $\bar{x}$ estimates μ . Statisticians describe the true relationship between two quantitative variables using a *population regression model*:

$$Y = \beta_0 + \beta_1X + \varepsilon$$

β_0 is the population y-intercept, β_1 is the population slope, and ε is random error: the part of Y not explained by X . β_0 and β_1 are parameters, describing the entire population; b_0 and b_1 are statistics, calculated from your sample, used to estimate them.

This chapter sets up the model and the estimates. The next chapter, Statistical Inferences for Slopes, reasons from b_1 , the number your sample gives you, to β_1 , the value you actually care about, using the same confidence interval and significance test tools you have already built for proportions and means.

5.2 Estimating the Slope and Intercept

The formulas for b_0 and b_1 :

$$b_1 = r \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x}$$

where r is the sample correlation coefficient and s_x, s_y are the sample standard deviations of X and Y . If you are working from raw sums instead of r, s_x , and s_y , an algebraically

equivalent formula is

$$b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

Both formulas give the same b_1 . You can use whichever matches the summary statistics you have.

The residual for any data point is still $e = y - \hat{y}$, where $\hat{y} = b_0 + b_1x$ is the value your line predicts. The standard deviation of all the residuals in your sample, s_e , tells you the typical size of a prediction miss. You will use s_e heavily, because it is the key in measuring how much b_1 itself might vary from sample to sample.

For the study, the researcher samples 8 teenagers at random and records their daily screen time (X , in hours) and their sleep duration that night (Y , in hours):

Screen time (hr)	2	3	4	5	6	7	8	9
Sleep duration (hr)	7.85	8.05	7.40	7.55	6.90	7.35	6.70	6.95

For this sample, $\bar{x} = 5.5$, $\bar{y} = 7.34$, $s_x = 2.45$, $s_y = 0.473$, and $r = -0.854$. That gives

$$b_1 = (-0.854) \frac{0.473}{2.45} = -0.165 \quad b_0 = 7.34 - (-0.165)(5.5) = 8.25$$

so $\hat{y} = 8.25 - 0.165x$. The negative slope matches what the researcher suspected: more screen time is associated with less sleep, at least in this sample. The question this chapter answers is whether that association is strong enough and consistent enough, to conclude something about β_1 for teenagers in general, or whether a slope this far from zero is a believable result of sampling variation alone.

- $Y = \beta_0 + \beta_1 X + \varepsilon$ is the population regression model; β_0 and β_1 are unknown parameters, and ε is random error.
- b_0 and b_1 are sample statistics, calculated from your data, that estimate β_0 and β_1 .
- $b_1 = r \frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1 \bar{x}$, with an equivalent raw-sum formula available for b_1 .
- The residual $e = y - \hat{y}$ and its standard deviation s_e measure how far actual data points fall from the line.

Exercise 23 Identifying the Model

A horticulturist records the amount of fertilizer applied (in grams per plant) and the resulting plant height (in centimeters) for a random sample of tomato

Working Space

plants.

1. Identify X and Y in this study.
2. Write the population regression model for this relationship, using correct notation.
3. In context, what would it mean if $\beta_1 = 0$?
4. Explain why β_1 is not something the horticulturist can ever know exactly, even after collecting data from every plant they own.

Answer on Page 76

Exercise 24 Spot the Error

A student in the sleep study writes: "I calculated $\beta_1 = -0.165$ from my sample of 8 teenagers, so I can conclude that increasing screen time by one hour causes a decrease of 0.165 hours of sleep for every teenager." Identify two separate errors in this statement.

Working Space

Answer on Page 76

Exercise 25 Computing the Line

An engineer records the operating temperature (in degrees Celsius) and the battery life (in hours) for a random sample of 6 devices and finds $\bar{x} = 30$, $\bar{y} = 9.2$, $s_x = 8.5$, $s_y = 1.4$, and $r = -0.79$.

Working Space

1. Find b_1 and b_0 .
2. Write the least-squares regression equation.
3. Predict the battery life for a device operating at 35°C .

Answer on Page 76

5.3 Computing in Excel

Enter the sleep study data from earlier in this chapter into two columns: screen time in cells A2 through A9, and sleep duration in cells B2 through B9, with headers in row 1.

1. In an empty cell, enter `=SLOPE(B2:B9,A2:A9)`. Excel returns -0.1649 , matching $b_1 = -0.165$ from the hand calculation (Excel simply keeps more decimal places).
2. In another cell, enter `=INTERCEPT(B2:B9,A2:A9)`. Excel returns 8.2506 , matching $b_0 = 8.25$.
3. To check the correlation behind the calculation, enter `=CORREL(A2:A9,B2:B9)`. Excel returns -0.8536 , matching $r = -0.854$.
4. To find s_x and s_y directly instead of computing them by hand, use `=STDEV.S(A2:A9)` and `=STDEV.S(B2:B9)`. These return 2.4495 and 0.4732 .

Watch the argument order: for both `SLOPE` and `INTERCEPT`, the known y -values are listed first, then the known x -values. Reversing the two ranges is a common mistake, and it may hand you the wrong sign or a wrong intercept.

	A	B	C	D	
1	Screen time (hr)	Sleep (hr)			
2	2	7.85	Slope	-0.1649	-- =SLOPE(B2:B9,A2:A9)
3	3	8.05	Intercept	8.2506	-- =INTERCEPT(B2:B9,A2:A9)
4	4	7.40	Correlation	-0.8536	-- =CORREL(A2:A9,B2:B9)
5	5	7.55			
6	6	6.90			
7	7	7.35			
8	8	6.70			
9	9	6.95			

Figure 5.1: Spreadsheet layout for the sleep study: data in columns A and B, with each formula beside the result it returns in column D.

Statistical Inferences for Slopes

6.1 The Sampling Distribution of b_1 and the Confidence Interval for the Slope

Suppose the sleep researcher ran the study again with a brand new sample of 8 teenagers. A new sample of 8 teenagers would give a different fitted line and a different b_1 , the same way $\hat{p}$ and $\bar{x}$ vary from sample to sample. That variation is the *sampling distribution* of b_1 : the distribution of values you would get from repeated random samples of the same size. Under the conditions for inference, this distribution is approximately normal, centered at β_1 , with standard deviation σ_{b_1} .

σ_{b_1} depends on σ_ϵ , the population standard deviation of the errors, which is unknown, so you estimate it with s_{b_1} , the *standard error of the slope*. Substituting s_{b_1} for σ_{b_1} turns

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

from normal into a t-distribution with $df = n - 2$, the same trade as with means: estimated standard error in, precision out, t-distribution in place of normal.

Unlike proportions and means, you will almost never compute s_{b_1} by hand: it depends on s_e , sample size, and the spread of your x-values, and doing that from scratch is tedious enough that no exam or workplace expects it. Instead, s_{b_1} is read directly from a calculator's regression screen or software printout, in a column usually labeled "StDev" or "SE Coef." Recognizing and interpreting that column is the actual skill.

Constructing the Confidence Interval

With s_{b_1} in hand, a confidence interval for β_1 takes the same shape as every confidence interval you have built so far: statistic plus or minus a margin of error.

$$\text{Margin of Error} = t^* s_{b_1} \quad \text{Confidence Interval: } b_1 \pm t^* s_{b_1}$$

where t^* is the critical value from a t-distribution with $df = n - 2$, chosen to match your desired confidence level.

Let us return to the sleep study. A statistical package fits the regression line to the 8 data points and reports the following output:

Predictor	Coef	StDev	T	P
Constant	8.2506	0.2448	33.70	0.000
Screen time	-0.1649	0.0411	-4.01	0.007

with $n = 8$, so $df = n - 2 = 6$. To build a 90% confidence interval for β_1 , look up $t^* = t_{0.05}(6) = 1.943$. Then

$$\text{Margin of Error} = (1.943)(0.0411) = 0.0798$$

$$\text{Confidence Interval: } -0.1649 \pm 0.0798 = (-0.2447, -0.0850)$$

We are 90% confident that, for teenagers in general, each additional hour of daily screen time is associated with a change in average sleep duration of between -0.245 and -0.085 hours. Notice that this entire interval is negative. Since it does not contain 0, every plausible value of β_1 in this interval represents some negative relationship between screen time and sleep, which is a first hint that the effect the researcher found in the sample may reflect something real in the population, and not just sampling noise. The next two sections make that hint precise, first by checking whether this procedure was even valid to use, and then by testing the claim directly.

- b_1 varies from sample to sample; its sampling distribution is centered at β_1 under the conditions for inference.
- The standard error s_{b_1} is read directly from calculator or computer regression output; it is rarely computed by hand.
- $t = \frac{b_1 - \beta_1}{s_{b_1}}$ follows a t -distribution with $df = n - 2$.
- Confidence interval: $b_1 \pm t^* s_{b_1}$, with t^* taken from the t -distribution at $df = n - 2$.
- If a confidence interval for β_1 does not contain 0, that is an early signal, not yet a conclusion, that the population slope may be nonzero.

Exercise 26 Reading Computer Output

A researcher studies the relationship between weekly study hours (X) and the improvement in exam score, in points (Y), for a random sample of $n = 12$ students. Regression output gives $b_1 = 2.35$ and $s_{b_1} = 0.68$.

Working Space

1. Find the degrees of freedom for this problem.

2. Construct a 95% confidence interval for β_1 . ($t_{0.025}^*(10) = 2.228$.)
3. Interpret the interval in context.

Answer on Page 77

Computing in Excel

The regression output table above, including s_{b_1} , can be generated directly from the raw data rather than typed in by hand. Using the same screen time and sleep duration columns as in the previous chapter (A2:A9 and B2:B9):

1. Select a blank 2×2 range, say D2:E3. Type =LINEST(B2:B9,A2:A9,TRUE,TRUE) and confirm with Ctrl+Shift+Enter (in newer versions of Excel with dynamic arrays, a plain Enter is enough). This fills all four cells at once:

$b_1 = -0.1649$	$b_0 = 8.2506$
$s_{b_1} = 0.0411$	$s_{b_0} = 0.2448$

The top row gives the coefficients, in reverse order from how you might expect (slope first, then intercept); the row directly below gives each one's standard error.

2. For the critical value, enter =T.INV.2T(0.10,6) in an empty cell. The first argument is $1 - \text{confidence level}$, and the second is $df = n - 2$. Excel returns 1.9432, matching $t_{0.05}^*(6)$ from the table.
3. Margin of error: reference the two cells you already have, =D3*[t* cell], giving 0.0798.
4. Confidence interval bounds: =D2-[margin cell] and =D2+[margin cell], giving -0.2447 and -0.0850 .

LINEST is the same idea as SLOPE and INTERCEPT from the previous chapter, but it returns the standard errors in the same step, so you never have to hunt for s_{b_1} separately. The argument order is the same gotcha as before: known y's first, known x's second.

	A	B	C	D	E	F	G	
1	Screen time (hr)	Sleep (hr)		Slope	Intercept			
2	2	7.85		-0.1649	8.2506			--- =LINEST(B2:B9,A2:A9,TRUE,TRUE)
3	3	8.05		0.0411	0.2448			
4	4	7.40						
5	5	7.55				t^*	1.9432	--- =T.INV.2T(0.10,6)
6	6	6.90				Margin	0.0798	--- =D3*G5
7	7	7.35				Lower	-0.2447	--- =D2-G6
8	8	6.70				Upper	-0.0850	--- =D2+G6
9	9	6.95						

Figure 6.1: LINEST output in D2:E3 and the 90% confidence interval for the slope built from it in column G.

6.2 Conditions for Inference on the Slope

The confidence interval you just built assumed that it was valid to use a t -procedure at all. Every inference method in this book comes with conditions attached, and the slope is no exception.

You have already met four of these five conditions for proportions and for means.

- **Linearity:** the true relationship between X and Y is actually linear, not curved.
- **Independence:** individual observations do not influence one another; if sampling without replacement, the population is at least 10 times the sample size.
- **Normality:** for any fixed value of X , the residuals are approximately normally distributed.
- **Equal variance:** the spread of the residuals is roughly the same across all values of X .
- **Random:** the data come from a random sample or a randomized experiment.

These are remembered by the acronym LINER. Random and Independence are conditions you check from the design of the study itself, the same way you always have. Linearity, Normality, and Equal variance are different: you check all three using the same tool, a *residual plot*, which graphs each residual against its corresponding x -value.

Checking the Conditions: The Sleep Study

Return to the sleep researcher's 8 teenagers. Before trusting the confidence interval, check all five conditions.

Random. The researcher selected the 8 teenagers using a random sampling method, so this condition is satisfied by the design of the study.

Independence. Teenagers in general form a population far larger than 8, so sampling 8 of them without replacement does not meaningfully affect the independence of their screen time or sleep duration from one another. The 10% condition is easily satisfied.

Linearity, Normality, and Equal variance. These three come from the residuals themselves.

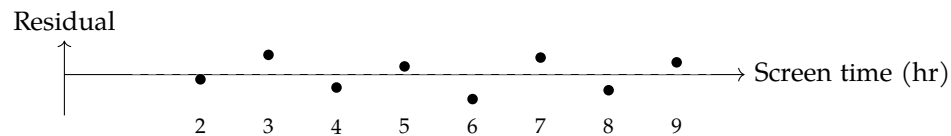


Figure 6.2: Residual plot for the sleep study regression.

The residuals in Figure 6.2 bounce above and below 0 with no visible curve, so linearity is reasonable: a straight line, rather than some curved model, is a sensible description of the relationship. The band of residuals also stays roughly the same width across the range of screen time values, from 2 hours to 9 hours, with no obvious funnel shape, so equal variance is reasonable as well.

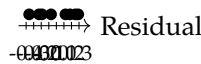


Figure 6.3: Dotplot of the eight residuals from the sleep study regression.

The dotplot in Figure 6.3 shows no strong skew and no outliers among the eight residuals, so it is reasonable to proceed as though normality holds. With only $n = 8$ observations, though, this check is genuinely limited; a small sample simply does not give you enough information to rule out non-normality with much confidence. The honest conclusion is not “the residuals are normal,” but “nothing in this small sample contradicts normality,” which is a weaker, more defensible claim, and one that is worth stating explicitly on an exam or in a report.

With all five conditions reasonably satisfied, the confidence interval, and the hypothesis test coming next, both rest on solid ground.

- The five conditions for slope inference: Linearity, Independence, Random, Equal variance, Normality (LINER).
- Random and Independence come from the design of the study.
- Linearity, Normality, and Equal variance are checked using a residual plot (and, for normality specifically, a dotplot, boxplot, or normal probability plot of the residuals).

- A small sample size limits how confidently you can assess normality; conclusions should be stated as “nothing contradicts” rather than “confirms.”

Exercise 27 Diagnosing a Residual Plot

A researcher studies the relationship between a car’s weight and its fuel efficiency. The residual plot shows points that fan out widely for heavy cars but cluster tightly near 0 for light cars.

Working Space

1. Which condition does this pattern call into question?
2. Explain, in your own words, why this pattern is a problem for the validity of a confidence interval or test built from this data.

Answer on Page 77

6.3 Testing a Claim About the Slope

The confidence interval hinted that the true slope β_1 is not 0, since $(-0.2447, -0.0850)$ never crosses it. A hint is not a conclusion. To make the claim precise, and to attach an actual probability to the strength of the evidence, you need a significance test.

Step 1: State the Hypotheses

The sleep researcher’s original question was directional: does more screen time predict *less* sleep, not just some unspecified difference. That points to a one-sided alternative.

$$H_0 : \beta_1 = 0 \quad H_a : \beta_1 < 0$$

H_0 says screen time has no linear relationship with sleep duration in the population of teenagers; H_a says the relationship is negative. Use $\alpha = 0.05$.

Step 2: Name the Procedure and Verify Conditions

Use a one-sample t-test for the slope of a regression line, with $df = n - 2$. The previous section already checked all five conditions, Random, Linearity, Independence, Normality, and Equal variance, against this exact data set, and found no serious violations. Since the same data set is being reused here, that check still applies; there is no need to redo it.

Step 3: Calculate the Test Statistic and P-value

Recall the regression output:

Predictor	Coef	StDev	T	P
Constant	8.2506	0.2448	33.70	0.000
Screen time	-0.1649	0.0411	-4.01	0.007

with $n = 8$, so $df = 6$. The test statistic is

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{-0.1649}{0.0411} = -4.01$$

matching the T column exactly, since that column is always $\frac{\text{Coef}}{\text{StDev}}$.

The P column, 0.007, is a two-sided p-value: it treats $H_a : \beta_1 \neq 0$ as the alternative, whether or not that is the alternative you actually care about. Since our alternative here is one-sided, that number needs one adjustment before you can use it.

Converting a two-sided p-value to a one-sided p-value:

1. Check whether the sign of t matches the direction claimed in H_a . Positive t matches $H_a : \beta_1 > 0$; negative t matches $H_a : \beta_1 < 0$.
2. If the sign matches, divide the reported two-sided p-value by 2.
3. If the sign does not match, the data point the opposite way from your claim. The correct one-sided p-value is actually greater than 0.5, no matter how small the two-sided number looks, and there is no evidence for H_a .

Here $t = -4.01$ is negative, matching the direction of $H_a : \beta_1 < 0$, so the sign check passes. The one-sided p-value is

$$\text{p-value} = \frac{0.007}{2} = 0.0035$$

Step 4: State a Conclusion

Because $p = 0.0035 < \alpha = 0.05$, reject H_0 . There is convincing statistical evidence that, in the population of teenagers this sample represents, more daily screen time is associated with a decrease in average nightly sleep duration. Because the screen time data were observed rather than randomly assigned, this conclusion is about association, not causation; some other factor could plausibly be driving both variables at once.

- Test statistic: $t = \frac{b_1 - 0}{s_{b_1}}$, with $df = n - 2$, using the value of β_1 specified in H_0 (usually 0).
- Standard computer output typically reports a two-sided p-value by default, regardless of which alternative you are actually testing.
- For a one-sided H_a , divide the reported two-sided p-value by 2 only if the sign of t matches the direction of H_a ; otherwise the true one-sided p-value exceeds 0.5.
- Reject H_0 when $p < \alpha$; state the conclusion in context, and be careful not to claim causation from an observational study.

Computing in Excel

Continuing from the LINEST output (cell D2 for b_1 , cell D3 for s_{b_1}):

1. Test statistic: in an empty cell, enter `=D2/D3`. Excel returns -4.0131 , matching the T column of the regression output.
2. One-sided p-value directly: enter `=T.DIST(D2/D3,6,TRUE)`. This gives the left-tail probability $P(T < t)$ for $df = 6$, which is exactly the one-sided p-value for $H_a : \beta_1 < 0$ when t is negative. Excel returns 0.0035, with no separate halving step needed.
3. If you only have the two-sided p-value from a printed output table, use `=T.DIST.2T(ABS(D2/D3),6)/2` instead, which reproduces the halving rule from Step 3 directly. Both approaches agree here: 0.0035.

`T.DIST` assumes a lower-tail test by default; for $H_a : \beta_1 > 0$, use `=1-T.DIST(t,df,TRUE)` or the equivalent `=T.DIST.RT(t,df)` instead, since you need the upper tail.

	A	B	C	D	E	F	G
1	Screen time (hr)	Sleep (hr)		Slope	Intercept		
2	2	7.85		-0.1649	8.2506		
3	3	8.05		0.0411	0.2448		
4	4	7.40					
5	5	7.55				<i>t</i>	-4.0131 --- =D2/D3
6	6	6.90				<i>P</i> -value	0.0035 --- =T.DIST(D2/D3,6,TRUE)
7	7	7.35				<i>P</i> -value	0.0035 --- =T.DIST.2T(ABS(D2/D3),6)/2
8	8	6.70					
9	9	6.95					

Figure 6.4: Test statistic and one-sided *P*-value for the sleep study, computed from the LINEST output in D2:E3.

Computing in Python

The full test, from raw data to conclusion, takes only a few lines with `scipy.stats`:

```

1 import numpy as np
2 from scipy import stats
3
4 x = np.array([2, 3, 4, 5, 6, 7, 8, 9])
5 y = np.array([7.85, 8.05, 7.40, 7.55, 6.90, 7.35, 6.70, 6.95])
6
7 result = stats.linregress(x, y)
8 df = len(x) - 2
9 t_stat = result.slope / result.stderr
10 p_one_sided = stats.t.cdf(t_stat, df) # left tail, for Ha: β1 < 0
11
12 print(f"b1 = {result.slope:.4f}, SE(b1) = {result.stderr:.4f}")
13 print(f"t = {t_stat:.4f}, df = {df}")
14 print(f"one-sided p-value = {p_one_sided:.4f}")

```

This prints $b_1 = -0.1649$, $SE(b_1) = 0.0411$, $t = -4.0131$, $df = 6$, and one-sided p -value = 0.0035 , matching the by-hand and Excel results exactly. `stats.t.cdf(t_stat, df)` gives the lower-tail probability directly, which is the correct one-sided p -value whenever H_a is "less than" and t is negative; for a "greater than" alternative, use $1 - \text{stats.t.cdf}(t_stat, df)$ instead.

Evaluating the Fit of a Linear Regression Model

Give a regression any two columns of numbers, say shoe size and math grade, and it will hand you back a line. It never refuses. So the real question is not “What is the line?” but “Is the line any good?” The last chapter answered one version of that question with a t-test for the slope. This chapter adds the other tools statisticians use to grade a fit: residual plots, R^2 , RMSE, and the ANOVA table with its F-test.

7.1 Look at the Residuals First

The residual for observation i is $e_i = y_i - \hat{y}_i$. A *residual plot* graphs e_i against x_i (or against $\hat{y}_i$). If the line is the right shape, the residuals scatter randomly around zero. If they form a pattern, the model has missed something.

Here is a driver’s-education example. Figure 7.1 shows the stopping distance of a car at different speeds. (The data is simulated from the physics of braking, plus some noise.) The line looks fine, and as you will see, it explains 97% of the variation. But the residuals make a clear U: the line overestimates in the middle and underestimates at both ends. Braking distance grows with the *square* of the speed, so a straight line is the wrong model, however good its score.

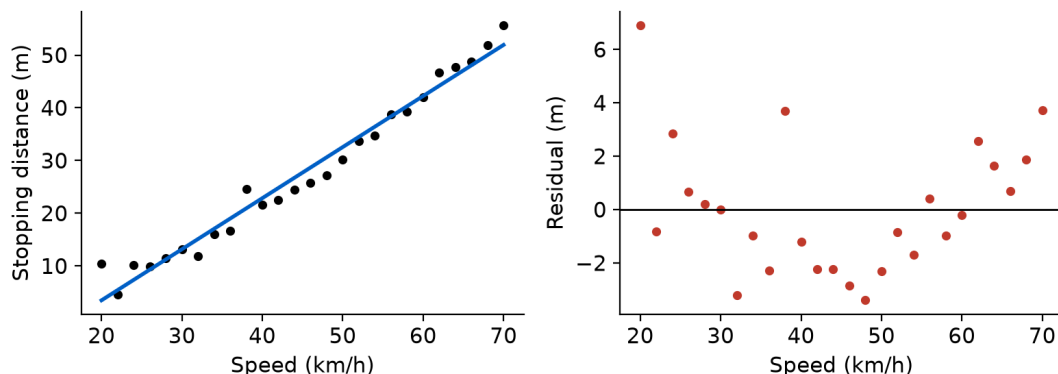


Figure 7.1: Left: stopping distance against speed, with the least squares line. Right: the residuals form a U, so the relationship is curved.

Residual pattern	What it suggests
random scatter around 0	the line is a reasonable model
a curve (U or upside-down U)	the relationship is not linear; try x^2 or a transformation
a fan (spread grows with x)	errors are bigger for big x ; predictions there are less precise
groups above and below 0	a missing category, such as grade level for students or species for animals
one point far from the rest	an outlier that may be pulling the line

7.2 Splitting Up the Variation

Why does a y value differ from the average $\bar{y}$? Part of the reason is that its x differs from average, which the line accounts for. The rest is the residual. For each point,

$$\underbrace{y_i - \bar{y}}_{\text{total}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{explained}} + \underbrace{y_i - \hat{y}_i}_{\text{residual}}$$

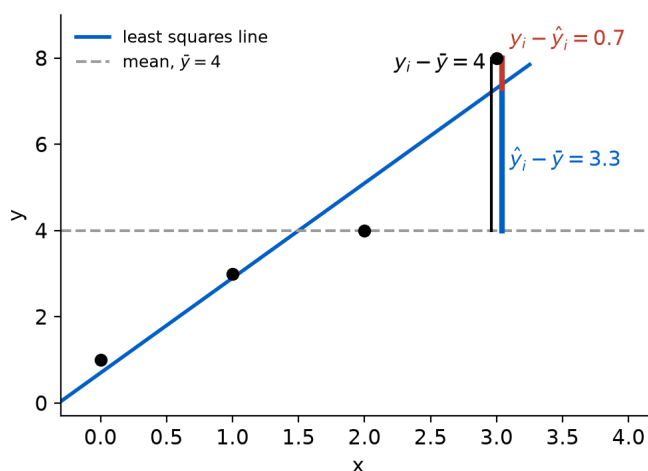


Figure 7.2: For the point (3, 8), the distance above the mean (4) splits into the part the line explains (3.3) and the residual (0.7).

Square these pieces and add them over all the points to get three *sums of squares*:

$$SS_{\text{tot}} = \sum (y_i - \bar{y})^2, \quad SS_{\text{reg}} = \sum (\hat{y}_i - \bar{y})^2, \quad SS_{\text{res}} = \sum (y_i - \hat{y}_i)^2$$

For a least squares fit with an intercept, they add up just like the pieces do:

$$SS_{\text{tot}} = SS_{\text{reg}} + SS_{\text{res}}$$

(It is not obvious that the squares should add up, too. It works because the least squares residuals are uncorrelated with the predictions: in the language of vectors, the residual vector is perpendicular to the vector of predictions, and this is the Pythagorean theorem in n dimensions.)

Example: For the points $(0, 1), (1, 3), (2, 4), (3, 8)$ with line $\hat{y} = 0.7 + 2.2x$, find the three sums of squares.

Solution: $\bar{y} = 4$. Work row by row:

x	y	$\hat{y}$	$(y - \bar{y})^2$	$(\hat{y} - \bar{y})^2$	$(y - \hat{y})^2$
0	1	0.7	9	10.89	0.09
1	3	2.9	1	1.21	0.01
2	4	5.1	0	1.21	1.21
3	8	7.3	16	10.89	0.49
Sum			$SS_{\text{tot}} = 26$	$SS_{\text{reg}} = 24.2$	$SS_{\text{res}} = 1.8$

And indeed $24.2 + 1.8 = 26$.

7.3 The Coefficient of Determination

The fraction of the total variation that the line explains is the *coefficient of determination*:

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- A perfect fit has $SS_{\text{res}} = 0$, so $R^2 = 1$.
- The laziest possible model ignores x and always predicts $\bar{y}$. It has $SS_{\text{res}} = SS_{\text{tot}}$, so $R^2 = 0$. You can read R^2 as “how much better than always guessing the average, on a scale from 0 to 1.”
- For simple linear regression, $R^2 = r^2$, the square of the correlation between x and y , as in the Regression Line chapter.

In the example above, $R^2 = 24.2/26 \approx 0.931$. Here is a second example. Five students studied for $x = 1, 2, 3, 4, 5$ hours and scored $y = 2.1, 3.9, 6.2, 7.8, 10.1$ on a 10-point quiz. The least squares line is $\hat{y} = 0.05 + 1.99x$, with $SS_{\text{res}} = 0.107$ and $SS_{\text{tot}} = 39.708$, so

$$R^2 = 1 - \frac{0.107}{39.708} \approx 0.997$$

Hours of study explain 99.7% of the variation in those five scores. Real data is rarely that tidy.

7.4 Root Mean Squared Error

R^2 has no units, which makes it easy to compare across problems but hard to feel. The *root mean squared error* (RMSE) is in the units of y :

$$\text{RMSE} = \sqrt{\frac{SS_{\text{res}}}{n}}$$

For the quiz data, $\text{RMSE} = \sqrt{0.107/5} \approx 0.146$ points: a typical prediction is off by about a seventh of a point. (The regression output in the last two chapters reports a close cousin, $s = \sqrt{SS_{\text{res}}/(n-2)}$, which divides by the degrees of freedom to estimate the spread of the errors in the population. RMSE divides by n , so it is exactly the root of the average squared error. For large n , the two are nearly equal.)

The two measures answer different questions. R^2 asks, “What fraction of the variation did we explain?” RMSE asks, “How far off is a typical prediction?” Report both.

7.5 The ANOVA Table and the F-Test

Could the pattern be a fluke? Even if y had nothing to do with x , random noise would give the line a slope that is not exactly zero and an R^2 that is not exactly zero. The *analysis of variance* (ANOVA) table tests whether the line explains more than noise would.

Each sum of squares is divided by its *degrees of freedom* to get a *mean square*, and the ratio of the two mean squares is the F statistic:

Source	SS	df	MS	F
Regression	SS_{reg}	1	$MS_{\text{reg}} = SS_{\text{reg}}/1$	$MS_{\text{reg}}/MS_{\text{res}}$
Residual	SS_{res}	$n - 2$	$MS_{\text{res}} = SS_{\text{res}}/(n - 2)$	
Total	SS_{tot}	$n - 1$		

The residual degrees of freedom are $n - 2$ because the line used up two numbers, b_0 and b_1 . With p features, the regression row gets p degrees of freedom and the residual row gets $n - p - 1$.

If x is useless, F tends to be around 1. If the line explains a lot, F is large. How large is large enough? If x really were useless, F would follow an *F distribution* with 1 and $n - 2$ degrees of freedom. The p -value is the chance that such a distribution produces a value at least as big as ours.

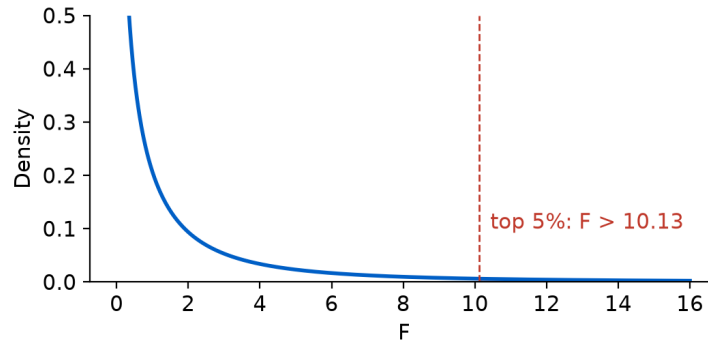


Figure 7.3: The F distribution with 1 and 3 degrees of freedom. If x were useless, F would land above 10.13 only 5% of the time.

Example: For the quiz data, $SS_{\text{reg}} = 39.601$, $SS_{\text{res}} = 0.107$, and $n = 5$. Find F .

Solution:

$$F = \frac{39.601/1}{0.107/3} \approx 1110$$

That is far beyond 10.13, off the right edge of Figure 7.3. The p-value is about 0.00006: strong evidence that the slope is not zero.

The F-test and the t-test. You already have a test for “is the slope zero?”: the t-test from the last chapter. With one predictor, the two tests always agree, because $F = t^2$ exactly. Take the sleep study, where 8 teenagers’ screen time and sleep gave $t = -4.01$ with a two-sided p-value of 0.007. Its sums of squares are $SS_{\text{tot}} = 1.567$, $SS_{\text{reg}} = 1.142$, and $SS_{\text{res}} = 0.425$, so

$$R^2 = \frac{1.142}{1.567} \approx 0.729, \quad F = \frac{1.142/1}{0.425/6} \approx 16.1 \approx (-4.01)^2$$

and the F-test’s p-value is 0.007, the same as the two-sided t-test. Why have both? The t-test can be one-sided and gives a confidence interval for the slope. The F-test extends to models with several predictors, where it asks whether *any* of them helps.

7.6 Grading a Line for Penguins

Now a bigger data set. From 2007 to 2009, researchers at Palmer Station, Antarctica, measured 344 penguins of three species; 333 have every measurement recorded. The data is free for anyone to use. Put `penguins.csv` and `penguins.py` in a folder; the function `load_penguins()` in `penguins.py` returns a dictionary of NumPy arrays, one per column, so `data["body_mass_g"]` is an array of 333 masses. (The machine-learning chapters later in the series use this data set throughout.)

Can a penguin's flipper length predict its body mass? SciPy knows the F distribution: `stats.f.sf(F, df1, df2)` gives the area to the right of F ("sf" stands for *survival function*). Here is `fit_quality.py`, which fits the line with `stats.linregress`, as in the last chapter, and grades it:

```

1 import numpy as np
2 from scipy import stats
3 from penguins import load_penguins
4
5 data = load_penguins()
6 x = data["flipper_length_mm"]
7 y = data["body_mass_g"]
8 n = len(y)
9
10 fit = stats.linregress(x, y)
11 y_hat = fit.intercept + fit.slope * x
12
13 ss_tot = np.sum((y - np.mean(y))**2)
14 ss_reg = np.sum((y_hat - np.mean(y))**2)
15 ss_res = np.sum((y - y_hat)**2)
16 print(f"SS_tot = {ss_tot:.4g}, SS_reg + SS_res = {ss_reg + ss_res:.4g}")
17
18 r_squared = 1 - ss_res / ss_tot
19 print(f"R^2 = {r_squared:.3f}, r^2 = {fit.rvalue**2:.3f}")
20 print(f"RMSE = {np.sqrt(ss_res / n):.0f} g")
21
22 # ANOVA: one predictor, n - 2 residual degrees of freedom
23 F = (ss_reg / 1) / (ss_res / (n - 2))
24 p_value = stats.f.sf(F, 1, n - 2)
25 print(f"F = {F:.0f}, p-value = {p_value:.1e}")
26
27 # The t statistic for the slope, from the previous chapter
28 t = fit.slope / fit.stderr
29 print(f"t = {t:.2f}, t^2 = {t**2:.0f}")

```

```

SS_tot = 2.153e+08, SS_reg + SS_res = 2.153e+08
R^2 = 0.762, r^2 = 0.762
RMSE = 392 g
F = 1060, p-value = 3.1e-105
t = 32.56, t^2 = 1060

```

The line is $\widehat{\text{mass}} = -5872 + 50.15 \times \text{flipper}$. Flipper length explains 76% of the variation in body mass, a typical prediction is off by 392 g, and $t^2 = F$ again. The p-value has 104 zeros after the decimal point; with 333 penguins, there is no doubt that the relationship is real. Notice that “real” is not the same as “strong”: with enough data, even a weak relationship gets a tiny p-value. R^2 and RMSE tell you how useful the line is.

7.7 More Features, Higher R^2 , Worse Predictions

A line uses one predictor. *Multiple regression* uses several: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p$, still choosing the b 's to make SS_{res} as small as possible. In NumPy, you stack the predictors as columns of a matrix X , with a first column of ones for b_0 , and `np.linalg.lstsq(X, y)` finds the least squares coefficients. (The Least Squares chapter in the machine-learning unit shows how.) The ANOVA table and R^2 work the same way, with p and $n - p - 1$ degrees of freedom.

There is a catch. Adding a predictor can never make SS_{res} go up on the data used to fit the model: least squares could always set the new coefficient to zero and do exactly as well as before. So R^2 never decreases as you add predictors, even useless ones.

The *adjusted* R^2 charges a fee for each of the p features:

$$R^2_{\text{adj}} = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

Let us test both. The program `noise.py` predicts mass from flipper length plus k columns of pure random numbers, which carry no information about penguins at all. It fits each model to 250 randomly chosen penguins, the *training* penguins, and then measures its RMSE on the other 83, which it never saw:

```

1 import numpy as np
2 from penguins import load_penguins
3
4 data = load_penguins()
5 y = data["body_mass_g"]
6 n = len(y)
7 order = np.random.default_rng(1).permutation(n)
8 train, test = order[:250], order[250:]
9
10 # 40 columns of pure random numbers: no information about penguins
11 junk = np.random.default_rng(0).normal(size=(n, 40))
12
13 print("junk    R^2    adj R^2    test RMSE")
14 for k in [0, 5, 10, 20, 40]:
15     X = np.column_stack([np.ones(n), data["flipper_length_mm"], junk[:, :k]])
16     b, *_ = np.linalg.lstsq(X[train], y[train])
17     res = y[train] - X[train] @ b
18     r2 = 1 - np.sum(res**2) / np.sum((y[train] - np.mean(y[train]))**2)
19     p = k + 1 # number of features
20     adj = 1 - (1 - r2) * (250 - 1) / (250 - p - 1)
21     rmse = np.sqrt(np.mean((y[test] - X[test] @ b)**2))
22     print(f"{k:4d}  {r2:.3f}  {adj:.3f}  {rmse:.0f}")

```

junk	R^2	adj R^2	test RMSE
0	0.747	0.746	343

5	0.750	0.743	344
10	0.764	0.753	354
20	0.783	0.763	378
40	0.807	0.768	414

R^2 climbs from 0.747 to 0.807 as we add garbage, while the RMSE on the unseen penguins gets worse by 71 g. The model is memorizing the random quirks of the training penguins, which is called *overfitting*. Adjusted R^2 grows more slowly, but even it is fooled here. The safest way to compare models is to measure them on data they did not see.

7.8 Limits of These Measures

- A high R^2 does not mean the model is the right shape; the stopping-distance line has $R^2 = 0.97$. Look at the residual plot.
- A low p-value means the relationship is probably not a fluke. It does not mean the relationship is strong, or causal.
- R^2 on the data used for fitting always rewards more predictors, so compare models on data they did not see.

Exercise 28 Compute R-squared

A regression has $SS_{\text{res}} = 25$ and $SS_{\text{tot}} = 100$. What is R^2 , and what does it mean?

Working Space

Answer on Page 77

Exercise 29 A model that predicts the mean

Explain why $R^2 = 0$ for a model that predicts $\bar{y}$ for every observation.

Working Space

Answer on Page 77

Exercise 30 **Read a residual plot**

In a residual plot, the residuals are positive for small $\hat{y}$, negative in the middle, and positive again for large $\hat{y}$. What does this suggest about the model?

*Working Space**Answer on Page 77***Exercise 31** **Fill in an ANOVA table**

A line is fit to 12 points. If $SS_{\text{tot}} = 80$ and $SS_{\text{res}} = 20$, find SS_{reg} , R^2 , and F .

*Working Space**Answer on Page 78***Exercise 32** **Correlation from R-squared**

A line of mass against flipper length has $R^2 = 0.762$. What is the correlation r ? How do you know its sign?

*Working Space**Answer on Page 78*

Exercise 33 **Fix the stopping distance**

For the stopping distances, the line has $R^2 = 0.973$. Adding a column of squared speeds, so that $\hat{d} = b_0 + b_1v + b_2v^2$, raises R^2 to 0.987. Since R^2 always rises when you add a column, why should you believe this gain, but not the gains from the junk columns?

Working Space

Answer on Page 78

Exercise 34 **The sleep study in code**

Change `fit_quality.py` to use the sleep-study data from the last chapter (it is in the `linregress` example there). Confirm that $R^2 \approx 0.729$, $F \approx 16.1$, and $F = t^2$.

Working Space

Answer on Page 78

Statistical Inferences: Chi-Square Distribution

8.1 The Chi-Square Distribution

A garden center might want to know whether the flower colors coming out of a seed mix really match the ratio printed on the packet. A hospital might want to know whether patient outcomes are spread across categories the way a published study says they should be. In cases like these you are comparing an entire distribution of counts to another distribution of counts, rather than comparing one number to another.

Suppose a garden center sells a wildflower seed mix advertised to bloom in a 3:2:5 ratio of red, yellow, and white flowers, that is, 30% red, 20% yellow, and 50% white. A gardener plants 200 seeds from this mix and, once everything blooms, counts 54 red flowers, 38 yellow flowers, and 108 white flowers. These three counts, 54, 38, and 108, are called *observed frequencies*, denoted $O_1, O_2, \dots, O_k$ for k categories. They always add up to the total sample size, here $54 + 38 + 108 = 200$.

If the packet's claim is true, how many flowers of each color should the gardener expect out of 200 seeds? Multiply the total by each claimed proportion: $200(0.30) = 60$ red, $200(0.20) = 40$ yellow, and $200(0.50) = 100$ white. These are the *expected frequencies*, denoted $E_1, E_2, \dots, E_k$. Like the observed frequencies, they sum to the total sample size.

Notice that the observed and expected counts are close, but not identical: 54 versus 60, 38 versus 40, 108 versus 100. Is that gap small enough to chalk up to ordinary sampling variability, or large enough to cast doubt on the packet's claim? To answer that, you need a single number that summarizes how far the observed counts are from the expected counts.

The natural first idea is to look at each difference, $O_i - E_i$. But differences can be positive or negative, and they would cancel out if you just added them up.

So, just as you did when building the variance of a data set, square each difference to keep everything positive: $(O_i - E_i)^2$.

A difference of 6 matters more when the expected count is 10 than when the expected count is 1,000. Dividing each squared difference by its own expected count, E_i , scales the discrepancy relative to how large that category was expected to be. Adding these scaled, squared discrepancies across all k categories gives the *chi-square statistic*:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

For the wildflower example, this comes out to

$$\chi^2 = \frac{(54 - 60)^2}{60} + \frac{(38 - 40)^2}{40} + \frac{(108 - 100)^2}{100} = 0.6 + 0.1 + 0.64 = 1.34$$

A small χ^2 value means the observed counts landed close to what was expected; a large value means they landed far away. What counts as “small” or “large” depends on comparing χ^2 to a reference distribution, which is exactly what the chi-square distribution gives you.

8.1.1 Properties of the Chi-Square Distribution

The χ^2 statistic follows an approximate *chi-square distribution*, a continuous probability distribution with the following properties:

- It only takes positive values, since it is built entirely out of squared terms.
- It is right-skewed, especially for small degrees of freedom.
- Its shape depends on its *degrees of freedom* (df). As df increases, the distribution stretches out and becomes more symmetric, shifting to the right.
- Its mean is equal to df, and its standard deviation is equal to $\sqrt{2 \cdot \text{df}}$.

You will see exactly how df is determined for each type of chi-square test in the sections that follow; it depends on the number of categories or the size of the table involved, not on the sample size itself.

8.1.2 Reading a Chi-Square Table

Because the chi-square distribution changes shape with df, a chi-square table has to account for both df and the tail area you care about. The value χ^2_{α} is defined as the value such that the area under the curve to its right equals α .

To find $\chi^2_{0.05}$ for df = 8, locate the row for df = 8 and the column for a right-tail area of 0.05, and read off the value where they intersect. That value is 15.507. To find $\chi^2_{0.01}$ for the same df = 8, you would instead go to the 0.01 column, giving 20.09. Notice that a smaller

tail area pushes you further out into the right tail of the distribution, so the critical value grows as α shrinks.

Table 8.1 gives critical values for the tail areas most commonly used in this course, for $df = 1$ through 10.

Table 8.1: Chi-square critical values, χ^2_α

df	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$
1	2.706	3.841	5.024	6.635
2	4.605	5.991	7.378	9.210
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.070	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.090
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209

Exercise 35 Reading the table

Using Table 8.1, find $\chi^2_{0.01}$ for $df = 3$.
Then find $\chi^2_{0.10}$ for $df = 6$.

Working Space

Answer on Page 78

8.2 The Chi-Square Test for Goodness of Fit

χ^2 measures how far a set of observed counts strays from what was expected, and the χ^2 distribution tells you how large that statistic can get just by chance. Putting the two together gives you a full hypothesis test: the *chi-square goodness-of-fit test*, used to check whether observed categorical data match a claimed distribution across categories.

8.2.1 Conditions

Before running the test, three conditions need to be met.

- The data are counts. They should not be percentages, averages, or any other summary.
- The sample was selected at random from the population of interest.
- The sample is large enough that the chi-square distribution is a reasonable approximation. This applies when every expected count is at least 5, that is, $E_i \geq 5$ for every category. If some expected counts fall below 5, combine that category with a neighboring one until every expected count clears 5.

8.2.2 Stating the Hypotheses

For k categories with claimed proportions $p_1, p_2, \dots, p_k$ (which must add to 1), the hypotheses are:

- H_0 : The population is distributed across the k categories according to $p_1, p_2, \dots, p_k$.
- H_a : The population is *not* distributed according to $p_1, p_2, \dots, p_k$; at least one category's true proportion differs from what was claimed.

Under H_0 , the test statistic $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ follows an approximate chi-square distribution with $df = k - 1$. You reject H_0 if χ^2 falls in the rejection region, that is, if $\chi^2 > \chi^2_{\alpha}(df)$, or equivalently, if the p -value is less than α .

Example

Return to the wildflower seed mix from the previous section. The packet claims a 30% red, 20% yellow, 50% white split. Out of 200 seeds planted, the gardener observed 54 red, 38 yellow, and 108 white flowers. Is there evidence that the true color distribution differs from the packet's claim?

State: Let p_1, p_2, p_3 be the true proportions of red, yellow, and white flowers produced by this seed mix.

$$H_0 : p_1 = 0.30, p_2 = 0.20, p_3 = 0.50 \quad H_a : \text{at least one } p_i \text{ differs from its claimed value}$$

Use $\alpha = 0.05$.

Plan: This is a chi-square goodness-of-fit test. Check conditions: the data are counts of flowers by color, the 200 seeds were planted as an independent random selection from the seed mix, and the expected counts (60, 40, and 100, calculated in the previous section) are all well above 5. All three conditions are satisfied, so a chi-square test is appropriate here. With $k = 3$ categories, $df = 3 - 1 = 2$.

Do: The test statistic was already computed: $\chi^2 = 1.34$. From Table 8.1, $\chi^2_{0.05}(2) = 5.991$. Because 1.34 is nowhere near 5.991, this is not going to be a close call. The exact p-value is $P(\chi^2(2) > 1.34) \approx 0.512$.

Conclude: Because $\chi^2 = 1.34 < 5.991$ (equivalently, because the p-value of 0.512 is far greater than 0.05), we fail to reject H_0 . There is not sufficient evidence that the true color distribution of this seed mix differs from the packet's claimed 30%/20%/50% split.

Once the observed and expected counts are in hand, the goodness-of-fit test is really just: find df, compare χ^2 to the table, and interpret.

Exercise 36 Goodness-of-Fit Test for a Bakery

A bakery believes weekday demand for its most popular pastry is the same on all five weekdays. Over 44 weeks, it records the number of times each weekday was the single best-selling day for that pastry: Monday 41, Tuesday 38, Wednesday 52, Thursday 45, Friday 44 (220 weekdays total). Test, at $\alpha = 0.05$, whether best-selling days are evenly spread across the five weekdays.

Working Space

Answer on Page 79

8.3 The Chi-Square Test for Independence

Suppose you have two categorical variables measured on the same individuals. How would you figure out whether they are related? Does a customer's preferred way of reaching support have anything to do with how satisfied they end up? Does a plant's soil type have anything to do with which color it blooms? These questions call for the *chi-square test of independence*.

8.3.1 Contingency Tables

Data for this test is organized into a *contingency table*: rows represent categories of one variable, columns represent categories of the other, and each cell holds the count of in-

dividuals falling into that combination. Every row has a row total, every column has a column total, and both sets of totals add up to the grand total, the overall sample size.

8.3.2 Conditions

The conditions are the same as for the goodness-of-fit test, adjusted for two variables instead of one.

- The data are counts of individuals falling into each combination of categories.
- Individuals were selected at random (or, if not literally random, in a way that can reasonably be treated as a random sample from the population of interest).
- Every expected count is at least 5. As before, if a cell's expected count is too small, adjacent categories should be combined.

8.3.3 Hypotheses and Expected Counts

For a test of independence, the hypotheses are always framed the same way, regardless of context:

- H_0 : the two categorical variables are independent.
- H_a : the two categorical variables are not independent (there is an association between them).

The tricky part is computing expected counts, since there is no claimed proportion handed to you the way there was for goodness-of-fit. Instead, expected counts are built directly from the table itself.

If the two variables really were independent, the proportion of individuals in any given column should be the same across every row. That leads to the expected count formula for the cell in row i , column j :

$$E_{ij} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{grand total}}$$

Once every E_{ij} is computed, the test statistic is the same $\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ you already know, and it follows an approximate chi-square distribution with

$$df = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns.

Example

A garden center wants to know whether the fertilizer its customers use is related to how well their plants grow. It surveys a random sample of 240 customers, recording each customer's fertilizer type and rating the customer's plant as thriving, average, or struggling. (It happens that exactly 80 customers used each type.) The results:

Table 8.2: Growth outcome by fertilizer type ($n = 240$)

	Thriving	Average	Struggling	Row total
Organic	30	31	19	80
Synthetic	27	29	24	80
None	27	30	23	80
Column total	84	90	66	240

The garden center's horticulturist runs a chi-square test for independence at $\alpha = 0.05$ and reports the critical value as $\chi^2_{0.05}(4) = 9.488$.

State: H_0 : fertilizer type and growth outcome are independent. H_a : fertilizer type and growth outcome are not independent.

Plan: This is a chi-square test of independence. The data are counts, the 240 customers were a random sample and were rated independently of one another, and every expected count (computed below) is well above 5. With $r = 3$ rows and $c = 3$ columns, $df = (3 - 1)(3 - 1) = 4$.

Do: Using $E_{ij} = \frac{(\text{row total})(\text{column total})}{240}$, every expected count in this table works out to 28, 30, or 22, since all three row totals equal 80:

Table 8.3: Expected counts under independence

	Thriving	Average	Struggling
Organic	28	30	22
Synthetic	28	30	22
None	28	30	22

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx 0.917$$

Conclude: Because $\chi^2 = 0.917 < 9.488$, we fail to reject H_0 . There is not sufficient evidence that fertilizer type and plant growth outcome are related at this garden center.

Exercise 37 Spotting a misinterpretation

Suppose the horticulturist looks at the same table and writes: "Since the observed counts differ from group to group, this proves fertilizer type affects growth, so we reject H_0 ." Is this conclusion correct? Justify your answer, and write the correct conclusion in context.

Working Space

Answer on Page 79

8.4 The Chi-Square Test for Homogeneity of Proportions

The test for independence and test for homogeneity for proportions are almost identical. They use the same test statistic, same $df = (r-1)(c-1)$, same table lookup. The differences include the question being asked and how the data were collected.

A test for independence starts with *one* sample and measures *two* categorical variables on each individual in it.

A *chi-square test for homogeneity*, by contrast, starts with *separate, independent samples*, one from each of several populations or groups, and measures a categorical variable on each individual.

8.4.1 Conditions and Hypotheses

The conditions are the same as for independence: counts, random samples (in this case, an independent random sample from each group), and every expected count at least 5.

The hypotheses:

- H_0 : the distribution of the categorical variable is the same across all of the populations or groups.

- H_a : the distribution of the categorical variable is not the same across all of the populations or groups.

Example

A garden center wants to compare three seed suppliers before choosing one to stock going forward. It plants exactly 100 seeds from each supplier, all of the same wildflower variety, and records the bloom color:

Table 8.4: Bloom color by seed supplier (100 seeds planted per supplier)

	Red	Yellow	White	Row total
Supplier A	50	20	30	100
Supplier B	30	40	30	100
Supplier C	20	45	35	100
Column total	100	105	95	300

Test, at $\alpha = 0.05$, whether bloom color distribution is the same across the three suppliers.

State: H_0 : bloom color distribution is the same across all three suppliers. H_a : bloom color distribution is not the same across all three suppliers.

Plan: This is a chi-square test for homogeneity: three independent samples, one fixed at 100 seeds per supplier by design, each measuring the same categorical variable (bloom color). Expected counts (below) are all well above 5. With $r = 3$ and $c = 3$, $df = 4$.

Do: Since all three row totals equal 100, every expected count is just the column total divided by 3:

Table 8.5: Expected counts under homogeneity

	Red	Yellow	White
Supplier A	33.33	35.00	31.67
Supplier B	33.33	35.00	31.67
Supplier C	33.33	35.00	31.67

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx 24.53$$

Conclude: From Table 8.1, $\chi^2_{0.05}(4) = 9.488$. Since $24.53 > 9.488$ ($p\text{-value} < 0.001$), we reject H_0 . There is sufficient evidence that bloom color distribution is not the same across the three seed suppliers. Looking back at the table, Supplier A stands out with far more red blooms than expected, which is likely driving most of that gap.

Exercise 38 Independence or homogeneity?

For each scenario, state whether a chi-square test for independence or a chi-square test for homogeneity is appropriate, and identify the categorical variable(s) involved.

1. A single random sample of 300 commuters is asked both their mode of transportation (car, bus, bike) and whether they arrived to work on time that day.
2. Independent random samples of 150 patients are drawn from three different hospitals, and each patient's recovery outcome (full recovery, partial recovery, no improvement) is recorded.

Working Space

Answer on Page 79

Answers to Exercises

Answer to Exercise 1 (on page 10)

$H_0 : p = 0.60$ and $H_a : p \neq 0.60$. The conditions hold: $np_0 = 90$ and $n(1 - p_0) = 60$ are both at least 10. $\hat{p} = \frac{81}{150} = 0.54$, and

$$z = \frac{0.54 - 0.60}{\sqrt{\frac{(0.6)(0.4)}{150}}} = \frac{-0.06}{0.04} = -1.5$$

The two-sided p-value is $2P(Z \leq -1.5) \approx 0.134$. Since $0.134 > 0.05$, fail to reject H_0 . The survey does not give convincing evidence that the true proportion differs from 60%.

Answer to Exercise 2 (on page 10)

A Type I error: rejecting a true H_0 . The company would believe the redesign works and might spend money rolling it out, even though it does not help.

Answer to Exercise 3 (on page 12)

$df = n - 1 = 24$. A 95 percent confidence interval splits the remaining 5 percent evenly between the two tails, leaving 2.5 percent, or 0.025, in the upper tail. Reading the t-table at $df = 24$ and upper-tail area 0.025 gives $t^* = 2.064$.

Answer to Exercise 4 (on page 13)

The $n \geq 30$ threshold is not a hard requirement for using a t-interval. It is one way to satisfy the normality condition when you cannot otherwise verify the shape of the population. If the population itself is known to be normally distributed, or if a dotplot or stemplot of the sample shows no strong skew or outliers, the normality condition is satisfied regardless of how small n is, and a t-interval is entirely appropriate. The student conflated a fallback rule with a strict cutoff.

Answer to Exercise 5 (on page 13)

(a) Matched pairs. The same twenty subscribers are measured twice, so the two sets of measurements are linked subscriber by subscriber, and the natural next step is to subtract to get one sample of twenty differences.

(b) Two independent samples. The comedy-watching group and the drama-watching group are made up of entirely different subscribers with no pairing between them, so there is no meaningful way to subtract one group's values from the other's.

Answer to Exercise 6 (on page 15)

$df = 15$. A 98 percent confidence interval leaves 1 percent, or 0.01, in each tail, giving $t^* = 2.602$. The standard error is $\frac{6.2}{\sqrt{16}} = 1.55$, so the margin of error is $2.602 \times 1.55 \approx 4.03$ minutes. The interval is $28.4 \pm 4.03 \Rightarrow (24.37, 32.43)$ minutes.

Answer to Exercise 7 (on page 16)

This is matched pairs, so the sample is the twelve differences, and everything proceeds exactly as it did for a single mean. $df = 11$, and a 90 percent confidence interval leaves 5 percent in each tail, giving $t^* = 1.796$. The standard error is $\frac{4.5}{\sqrt{12}} \approx 1.299$, so the margin of error is $1.796 \times 1.299 \approx 2.33$ minutes. The interval is $-3.2 \pm 2.33 \Rightarrow (-5.53, -0.87)$ minutes. Because the entire interval is negative, and the differences were defined as after minus before, this suggests call-handling time decreased after training, by somewhere between about 0.87 and 5.53 minutes on average.

Answer to Exercise 8 (on page 17)

The true mean μ is a fixed, though unknown, number. It is not a random variable, so it makes no sense to assign it a probability of falling anywhere. What is random is the interval itself, since a different sample would have produced a different $\bar{x}$ and a different interval. The corrected interpretation is that the analyst is 95 percent confident that the true mean watch time per session falls between 43.75 and 51.85 minutes, meaning that if this sampling process were repeated many times, about 95 percent of the resulting intervals would capture the true mean.

Answer to Exercise 9 (on page 18)

$H_0 : \mu = 500$, $H_a : \mu \neq 500$. Conditions are satisfied, so use a t-test with $df = 15$. The standard error is $\frac{6.4}{\sqrt{16}} = 1.6$, giving $t = \frac{498.2 - 500}{1.6} \approx -1.125$. For a two-sided test with $df = 15$, this gives a p-value of about 0.278. Because $0.278 > 0.05$, fail to reject H_0 . There is not sufficient evidence that the true mean loaf weight differs from 500 grams.

Answer to Exercise 10 (on page 19)

Since the differences are defined as before minus after, a decrease in heart rate shows up as a positive mean difference, so $H_0 : \mu_d = 0$, $H_a : \mu_d > 0$. This is matched pairs, so $df = 14$. The standard error is $\frac{6.1}{\sqrt{15}} \approx 1.575$, giving $t = \frac{4.8}{1.575} \approx 3.048$, with a one-sided p-value of about 0.0043. Because $0.0043 < 0.05$, reject H_0 . There is sufficient evidence that the regimen decreased mean resting heart rate.

Answer to Exercise 11 (on page 19)

Hypotheses are statements about the population parameter, μ , never about the sample statistic, $\bar{x}$. The sample mean is a single computed number from one sample; it is not something you test a hypothesis about, since it is already known exactly once the sample is collected. The correct hypotheses are $H_0 : \mu = 500$ and $H_a : \mu \neq 500$.

Answer to Exercise 12 (on page 22)

$SE = \sqrt{\frac{1.8^2}{15} + \frac{2.1^2}{17}} \approx 0.690$. With $df = 29$, $t^* \approx 1.699$, so the margin of error is $1.699 \times 0.690 \approx 1.17$ hours. The point estimate is $11.2 - 9.8 = 1.4$ hours, giving $1.4 \pm 1.17 \Rightarrow (0.23, 2.57)$ hours.

Answer to Exercise 13 (on page 22)

No. The interval $(0.23, 2.57)$ does contain values above 1, but it also contains values below 1, so it does not provide convincing evidence that the difference exceeds one full hour specifically. The interval does support the weaker claim that Model A's mean battery life is greater than Model B's, since the entire interval is positive, just not the stronger one-hour claim.

Answer to Exercise 14 (on page 23)

Pooling assumes the two population standard deviations are equal, $\sigma_1 = \sigma_2$. Here, the two samples show clearly different spreads, which is evidence against that assumption, not for it. Even when the spreads look similar, the unpooled two-sample t-interval using $\sqrt{s_1^2/n_1 + s_2^2/n_2}$, with technology-computed degrees of freedom, is the safe default, because it does not require equal spreads.

Answer to Exercise 15 (on page 24)

$H_0 : \mu_1 - \mu_2 = 0$, $H_a : \mu_1 - \mu_2 < 0$. The standard error is $\sqrt{\frac{2.1^2}{22} + \frac{2.8^2}{25}} \approx 0.717$, so $t = \frac{8.3 - 9.6}{0.717} \approx -1.813$, with $df \approx 44$. The one-sided p-value is about 0.0383. Since $0.0383 < 0.05$, reject H_0 . There is convincing evidence that Store 1's mean wait time is shorter than Store 2's.

Answer to Exercise 16 (on page 25)

This would be a Type I error. The researcher rejected H_0 (concluding Store 1's mean wait time is shorter), but if the true means are actually equal, H_0 was in fact true, so rejecting it was a mistake. Because the test used $\alpha = 0.05$, this kind of error would occur no more than about 5 percent of the time across many repeated studies like this one.

Answer to Exercise 17 (on page 26)

- (a) Quantitative, paired. The same 30 members are measured twice, before and after, so the natural approach is to subtract and work with the sample of differences.
- (b) Categorical. Proportions of an email being opened or not are categorical counts, not measurements, so this belongs to the tools from *Inference for Proportions*.
- (c) Quantitative, one sample. A single random sample is used to estimate one population mean, with no comparison to a second group or a second measurement per subject.

Answer to Exercise 18 (on page 31)

$\hat{p}_1 = 0.30$ and $\hat{p}_2 = 0.1875$. The counts 45, 105, 30, and 130 are all at least 10.

$$SE = \sqrt{\frac{0.30(0.70)}{150} + \frac{0.1875(0.8125)}{160}} \approx 0.0485$$

With $z^* = 1.645$, the margin of error is about 0.080, and the interval is 0.1125 ± 0.080 , or about (0.033, 0.192). We are 90% confident that the proportion of students who bike is between 3.3 and 19.2 percentage points higher at School A. The entire interval is above 0, so there is convincing evidence of a difference.

Answer to Exercise 19 (on page 32)

$H_0 : p_1 - p_2 = 0$ and $H_a : p_1 - p_2 \neq 0$. The pooled proportion is $\hat{p}_c = \frac{242}{400} = 0.605$, and the pooled counts $200(0.605) = 121$ and $200(0.395) = 79$ are at least 10.

$$SE = \sqrt{0.605(0.395) \left(\frac{1}{200} + \frac{1}{200} \right)} \approx 0.0489, \quad z = \frac{0.66 - 0.55}{0.0489} \approx 2.25$$

The two-sided p-value is $2P(Z \geq 2.25) \approx 0.024$. Since $0.024 < 0.05$, reject H_0 : there is convincing evidence that the pass rates differ, and because this was a randomized experiment, the new app caused the difference.

Answer to Exercise 20 (on page 32)

No. The interval includes 0, and even some negative values, so it is plausible that the two proportions are equal, or that teens prefer it slightly more. The data do not give convincing evidence of a difference.

Answer to Exercise 21 (on page 32)

A test computes how surprising the data would be if H_0 were true, and H_0 says the two proportions are the same, so the best estimate of that common value combines both samples. A confidence interval makes no assumption that the proportions are equal (in fact, it is estimating how different they are), so it uses each sample's own proportion.

Answer to Exercise 22 (on page 33)

A Type I error would be concluding that reminder texts raise the return rate when they actually make no difference. A Type II error would be failing to find that the texts help when they really do. Since the school rejected H_0 , the only error it could have made is a Type I error.

Answer to Exercise 23 (on page 36)

1. X is the amount of fertilizer applied (grams per plant); Y is the resulting plant height (centimeters).
2. $Y = \beta_0 + \beta_1 X + \varepsilon$, where β_0 is the population intercept, β_1 is the population slope relating fertilizer amount to height, and ε is random error.
3. If $\beta_1 = 0$, then fertilizer amount has no linear relationship with plant height in the population; the population regression line is flat, and knowing how much fertilizer a plant received tells you nothing about how tall it will grow.
4. β_1 describes the relationship for the entire population of plants the horticulturist could ever grow, under all possible conditions, not just the ones they happened to collect data from. Even collecting data from every plant they currently own would still leave β_1 unknown, since those plants are themselves a sample of all plants that could exist under those growing conditions. β_1 is a parameter; only a b_1 calculated from data is ever directly observed.

Answer to Exercise 24 (on page 37)

First, the value -0.165 was calculated from the sample, so it should be labeled b_1 , not β_1 . β_1 refers to the unknown population slope; b_1 is the sample statistic used to estimate it. Second, the statement claims causation from an observational study. Unless screen time was randomly assigned to teenagers in an experiment, a relationship in the data does not establish that screen time causes reduced sleep. Other variables, such as stress or extracurricular schedules, could be influencing both screen time and sleep.

Answer to Exercise 25 (on page 37)

1. $b_1 = (-0.79)\frac{1.4}{8.5} = -0.130$; $b_0 = 9.2 - (-0.130)(30) = 13.10$.

2. $\hat{y} = 13.10 - 0.130x$.
3. At $x = 35$: $\hat{y} = 13.10 - 0.130(35) = 8.55$ hours.

Answer to Exercise 26 (on page 42)

1. $df = n - 2 = 12 - 2 = 10$.
2. Margin of error $= (2.228)(0.68) = 1.515$. Confidence interval: $2.35 \pm 1.515 = (0.835, 3.865)$.
3. We are 95% confident that, in the population of students like those sampled, each additional hour of weekly study is associated with an average exam score improvement of between 0.835 and 3.865 points.

Answer to Exercise 27 (on page 46)

1. Equal variance. The spread of the residuals is clearly not constant across values of X ; it grows as car weight increases.
2. The standard error formula used to build s_{b_1} assumes the same amount of scatter around the line at every value of X . If the scatter actually changes with X , as it does here, the standard error is not accurately estimated, so the resulting confidence interval and p-value cannot be trusted, even if every calculation is performed correctly.

Answer to Exercise 28 (on page 58)

$R^2 = 1 - \frac{25}{100} = 0.75$. The line explains 75% of the variation in y .

Answer to Exercise 29 (on page 58)

Such a model has $\hat{y}_i = \bar{y}$, so $SS_{\text{res}} = \sum (y_i - \bar{y})^2 = SS_{\text{tot}}$. Then $R^2 = 1 - \frac{SS_{\text{tot}}}{SS_{\text{tot}}} = 0$.

Answer to Exercise 30 (on page 59)

The pattern is curved, not random, like the stopping-distance example. A straight line does not describe the relationship, and a model with a curved term, or a transformation

of the variables, would be a better choice.

Answer to Exercise 31 (on page 59)

$SS_{\text{reg}} = 80 - 20 = 60$ and $R^2 = 60/80 = 0.75$. The residual degrees of freedom are $12 - 2 = 10$, so $F = \frac{60/1}{20/10} = 30$.

Answer to Exercise 32 (on page 59)

$r = \pm\sqrt{0.762} \approx \pm 0.873$. The slope is positive (50 g per mm), and r always has the same sign as the slope, so $r \approx 0.873$.

Answer to Exercise 33 (on page 60)

The v^2 column comes from physics (braking distance grows with the square of speed), and it removes the U pattern from the residual plot. The junk columns had no reason to help and made the RMSE on unseen penguins worse. A column is believable when it fixes a visible pattern and improves performance on data the model did not see.

Answer to Exercise 34 (on page 60)

Replace the three lines that load the data with

```
1 x = np.arange(2, 10)
2 y = np.array([7.85, 8.05, 7.40, 7.55, 6.90, 7.35, 6.70, 6.95])
3 n = len(y)
```

and print the RMSE in hours with two decimal places (it is 0.23 hours). The program prints $SS_{\text{tot}} = 1.567$, $R^2 = 0.729$, $F = 16$ with $p \approx 7.0 \times 10^{-3}$, and $t = -4.01$ with $t^2 = 16$. (Use `:.2f` in the F line to see 16.10.)

Answer to Exercise 35 (on page 63)

$\chi^2_{0.01}(3) = 11.345$ and $\chi^2_{0.10}(6) = 10.645$.

Answer to Exercise 36 (on page 65)

H_0 : best-selling days are equally likely across the 5 weekdays ($p_i = 0.20$ for each). H_a : at least one weekday's true proportion differs from 0.20.

Conditions: counts, so assume the 220 weekdays behave as a random sample of best-selling-day outcomes; expected count per day is $220/5 = 44$, which is at least 5 for every category. The conditions are met.

With $k = 5$, $df = 4$. Expected count for every weekday is 44. The test statistic is

$$\chi^2 = \frac{(41 - 44)^2}{44} + \frac{(38 - 44)^2}{44} + \frac{(52 - 44)^2}{44} + \frac{(45 - 44)^2}{44} + \frac{(44 - 44)^2}{44} = 2.5$$

From Table 8.1, $\chi^2_{0.05}(4) = 9.488$. Since $2.5 < 9.488$ ($p\text{-value} \approx 0.645$), we fail to reject H_0 . There is not sufficient evidence that best-selling days are unevenly distributed across the weekdays.

Answer to Exercise 37 (on page 68)

No. The horticulturist is reasoning from the raw observed differences alone rather than from the test statistic and critical value. Even though the observed counts are not perfectly equal across groups, $\chi^2 = 0.917$ is far below the critical value of 9.488, so those differences are well within what random variation alone would produce. The correct conclusion is: there is not sufficient evidence that fertilizer type and growth outcome are associated; fail to reject H_0 . It is also worth noting that failing to reject H_0 does not *prove* the two variables are independent, so it only means this data set did not provide strong enough evidence of a relationship.

Answer to Exercise 38 (on page 70)

1. Test for independence. There is one sample, and two categorical variables (mode of transportation, on-time status) are measured on each commuter in it.
2. Test for homogeneity. There are three separate, independently drawn samples (one per hospital), and one categorical variable (recovery outcome) is compared across them.



INDEX

- F distribution, 54
- t-distribution, 11
- adjusted R^2 , 57
- adjusted R-squared, 57
- alternative hypothesis, 8
- analysis of variance, 54
- ANOVA
 - regression, 54
- chi-square distribution, 62
- chi-square goodness-of-fit test, 63
- chi-square statistic, 61
- chi-square test for homogeneity, 68
- chi-square test of independence, 65
- coefficient of determination, 53
- conditions for inference, 12
 - two proportions, 28
- confidence interval, 5
 - for a difference in proportions, 29
 - for a proportion, 6
 - for the slope, 41
 - interpretation, 15
- confidence level, 6
- contingency table, 65
- critical value, 6
- degrees of freedom, 11, 54, 62
- degrees of freedom (chi-square), 62
- expected frequencies, 61
- expected frequency, 61
- F-test, 54
- hypothesis test
 - two-sample, 23
- interval estimation, 5
- least-squares line, 35
- LINER conditions, 44
- LINEST, 43
- margin of error, 5
- matched pairs, 13
- mean square, 54
- Multiple regression, 57
- multiple regression, 57
- null hypothesis, 8
- observed frequencies, 61
- observed frequency, 61
- overfitting, 58
- p-value, 8, 9
- Palmer penguins, 55
- point estimate, 5
- pooled proportion, 30
- population regression model, 35
- power, 9, 25
- R-squared, 53
- regression slope
 - inference, 35

- residual, 36
 - plot, 51
- residual plot, 44, 51
- root mean squared error, 54
- sample size
 - planning, 7
- sampling distribution, 41
 - of the slope, 41
- significance level, 9
- significance test
 - for the slope, 46
 - two proportions, 30
- standard error, 11, 29
 - of the slope, 41
- standard error of the slope, 41
- Statistical inference, 8
- sums of squares, 52
- t-distribution, 11
- test of significance, 8
- test statistic, 8
- two-proportion inference, 27
- two-sample t-interval, 21
- Type I error, 9, 25
- Type II error, 9, 25
- variance
 - of a difference, 27
- Welch-Satterthwaite approximation, 21