Hypothesis Testing: statistical tests

Hypothesis Testing

Estimation vs Hypothesis testing

- Estimation puts bounds on the value of a population parameter.

- Hypothesis testing asks whether the parameter differs from a specific “null” expectation. “How large is the effect?”, “Is there any effect at all?”, so on.

Hypothesis testing compares data to what we would expect to see if a specific null hypothesis were true. If the data are too unusual, compared to what we would expect to see if the null hypothesis were true, then the null hypothesis is rejected.

Formal hypothesis testing begins with clear statements of two hypotheses - the null and alternative hypotheses - about a population.

The ***null hypothesis*** is a specific statement about a population parameter made for the purposes of argument. A good null hypothesis is a statement that would be interesting to reject.

The ***alternative hypothesis*** includes all other feasible values for the population parameter besides the value stated in the null hypothesis.

Hypothesis Testing

To reject or not to reject

Four basic steps are involved in hypothesis testing:

1. State the hypotheses.
2. Compute the test statistic.
3. Determine the *P*-value.
4. Draw the appropriate conclusions.

Hypothesis Testing

Hypothesis testing: an example

Do right-handed and left-handed toads occur with equal frequency in the toad population, or is one type more frequent than the other, as seen in human population?

Of the 18 toads tested, 14 were right-handed and four were left-handed. Are these results evidence of a predominance of one type of handedness in toads?

(1) State the hypotheses

- H0: Left- and right-handed toads are *equally frequent* in the population (i.e., $p = 0.5$).
- HA: Left- and right-handed toads are *not equally frequent* in the population (i.e., $p \neq 0.5$).

Note: The alternative hypothesis is **two-sided.** This just means that the alternative hypothesis allows for two possibilities: that p is greater than 0.5 (in which case right-handed toads outnumber left-handed toads in the population), or that p is less than 0.5 (i.e., left-handed toads predominate).

In a ***two-sided*** (or two-tailed) test, the alternative hypothesis includes parameter values on both sides of the parameter value specified by the null hypothesis.

Hypothesis Testing

2. Compute the test statistic

The ***test statistic*** is a number calculated from the data that is used to evaluate how compatible the data are with the result expected under the null hypothesis.

For the toad study, one can use the observed number of right-handed toads as the test statistic.

On average, if the null hypothesis were correct, one would expect to observe nine right-handed toads out of the 18 sampled (and nine left-handed toads, too). Instead, one observed 14 righthanded toads out of the 18 sampled.

Fourteen, then, is the value of our test statistic.

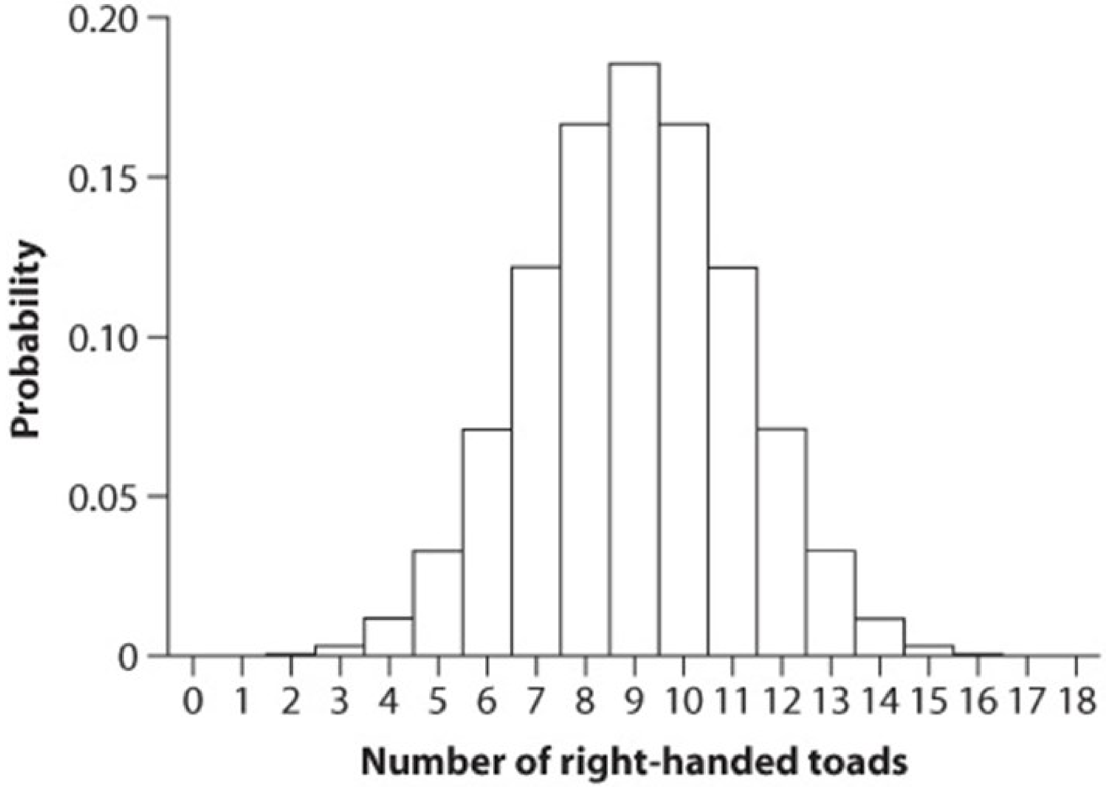
Hypothesis Testing

The null distribution

The ***null distribution*** is the sampling distribution of outcomes for a test statistic under the assumption that the null hypothesis is true.

X (no. of right-handed toads)	Probability
0	0.000004
1	0.00007
2	0.0006
3	0.0031
4	0.0117
5	0.0327
6	0.0708
7	0.1214
8	0.1669
9	0.1855

X (no. of right-handed toads)	Probability
10	0.1669
11	0.1214
12	0.0708
13	0.0327
14	0.0117
15	0.0031
16	0.0006
17	0.00007
18	0.000004
Total	1.0

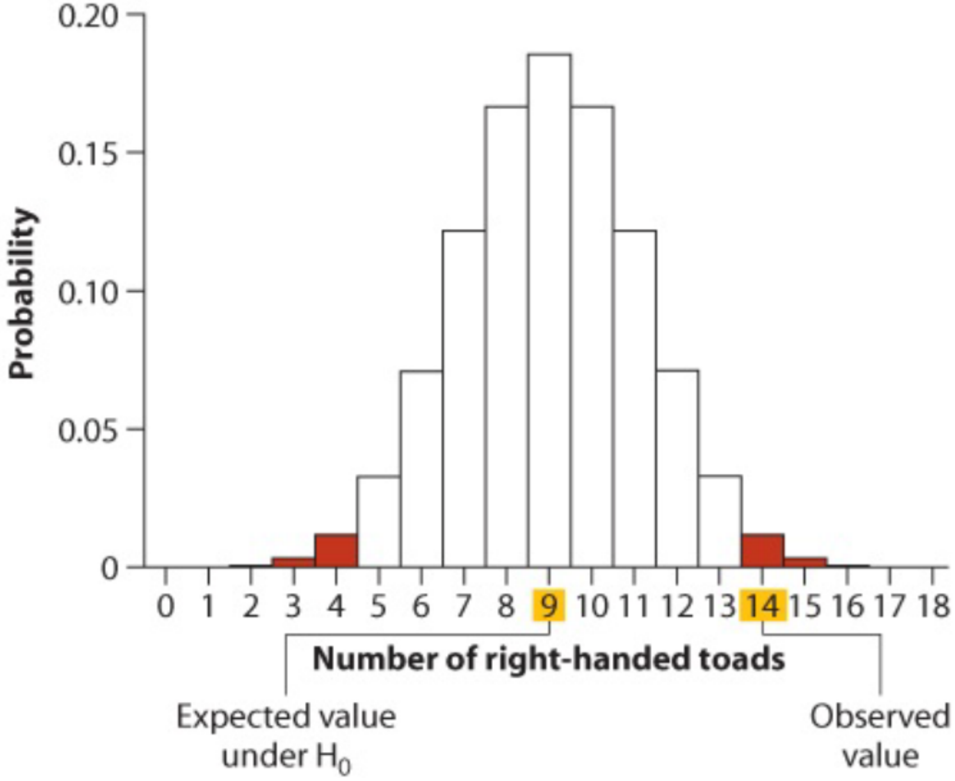


Hypothesis Testing

3. Quantifying uncertainty: the *P*-value

The ***P-value*** is the probability of obtaining the data (or data showing as great or greater difference from the null hypothesis) if the null hypothesis were true.

Note: The *P*-value is *not* the probability that the null hypothesis is true. In practice, one calculates the *P*-value from the null distribution for the test statistic.



Hypothesis Testing

3. Quantifying uncertainty: the *P*-value

According to null distribution, a total of 14 or more right-handed toads out of 18 is fairly unusual, assuming the null hypothesis.

These values lie at the right tail of the null distribution and have a low probability of occurring if H_0 is true. Equally unusual are 0, 1, 2, 3, or 4 righthanded toads, which are outcomes at the other tail of the null distribution.

Remember that our alternative hypothesis H_A is two-sided: it allows for the possibility that right-handed toads outnumber left-handed toads in the population, and also the possibility that left-handed toads outnumber right-handed toads.

Therefore, outcomes from both tails of the null distribution that are as unusual as the observed data, or even more unusual, must be accounted for in the calculation of the *P*-value.

Hypothesis Testing

3. Quantifying uncertainty: the *P*-value

Thus, let X: no. of right-handed toads

$$
\begin{aligned}
P(X \geq 14) &= P(X=14) + P(X=15) + P(X=16) + P(X=17) + P(X=18) \\
&= 0.0117 + 0.0031 + 0.0006 + 0.00007 + 0.000004 = 0.0155
\end{aligned}
$$

Note: this sum is not the *P*-value, because it does not yet include the equally extreme results at the left tail of the null distribution—that is, those outcomes involving a predominance of left-handed toads.

The quickest way to include the probabilities of the equally extreme results at the other tail is to take the above sum and multiply by two:

$$\text{P-value} = 2 \times P(X \geq 14) = 0.031$$

Hypothesis Testing

4. Draw the appropriate conclusions
Having calculated the *P*-value, what conclusion can we draw from it?

We also know that if *P-value* is “small,” we reject the null hypothesis; otherwise, we do not reject H0. But what value of *P-value* is small enough?

By convention in most areas of biological research, the boundary between small and not-small *P*-values is 0.05.

That is, if *P-value* is less than or equal to 0.05, then we reject the null hypothesis; if $P > 0.05$, we do not reject it.

This decision threshold for *P* (i.e., $P = 0.05$) is called the **significance level**, which is signified by *α*. In biology, the most widely used significance level is $\alpha = 0.05$, or $\alpha = 0.01$.

The ***significance level, α***, is a probability used as a criterion for rejecting the null hypothesis. If the *P*-value is less than or equal to *α*, then the null hypothesis is rejected. If the *P*-value is greater than *α*, then the null hypothesis is *not* rejected.

The *P*-value for the toad data, $P = 0.031$, is indeed less than 0.05, so we reject the null hypothesis that left-handed and right-handed toads are equally frequent in the toad population. We conclude from these data that most of the toads in the population are right-handed.

Hypothesis Testing

Reporting the results
When writing up your results in a research paper or laboratory report, always include the following information in the summary of the results of a statistical test:

1. The value of the test statistic.
2. The sample size.
3. The *P*-value.

In addition, the best practice is to provide confidence intervals, or at least the standard errors, for the parameters of interest.

This is because although the *P*-value indicates the weight of evidence against the null hypothesis (smaller *P* means stronger evidence), *P* does not measure the size of the effect. A very small *P*-value may result even when the size of the effect being measured is small.

The confidence interval puts bounds on the estimated magnitude of effect.

In the example of toads, $0.54 < P\text{-}value < 0.91$ with 95% confidence.

Hypothesis Testing

Errors in hypothesis testing

The most unsettling aspect of hypothesis testing is the possibility of errors. Rejecting H0 does not necessarily mean that the null hypothesis is false. Similarly, failing to reject H0 does not necessarily mean that the null hypothesis is true. This is because chance affects samples, sometimes with large impact. Some uncertainty can be quantified, though, if the data are a random sample, so making rational decisions is possible.

Type I and Type II errors

Type I error is rejecting a true null hypothesis. The significance level *α* sets the probability of committing a Type I error.

Type II error is failing to reject a false null hypothesis.

Decision	**Reality**	
	When H_0 is True	**When H_0 is False**
Reject	Type I Error (False Positive) Probability = α	Correct Decision (True Positive) Probability = 1 – β
Don’t Reject	Correct Decision (True negative) Probability = 1 – α	Type II Error (False negative) Probability = β

The ***power*** of a test is the probability that a random sample will lead to rejection of a false null hypothesis.

Hypothesis Testing

Hypothesis testing versus confidence intervals

The confidence interval puts bounds on the most-plausible values for a population parameter based on the data in a random sample.

Would confidence intervals and hypothesis tests on the same data, then, give the same answer?

In other words, if the parameter value stated in the null hypothesis fell outside the 95% confidence interval estimated for the parameter, would H0 be rejected by a hypothesis test at $\alpha = 0.05$?

And, if the parameter value stated in the null hypothesis fell inside the 95% confidence interval, would a test at $\alpha = 0.05$ fail to reject H0?

The answer is almost always “yes.” Why, then, don’t we just skip hypothesis testing altogether?

Hypothesis testing is mainly used to decide whether sufficient evidence has been presented to support a scientific claim.

The kinds of claims addressed by hypothesis testing are largely qualitative, such as “this new drug is effective” or “this pollutant harms fish.”

Hypothesis Testing

Why statistical significance is not the same as biological importance
In the early 20th century, the word “significant” had one dominant meaning.

If you said that something was “significant”, you meant that it *signified* or showed something—that it had or conveyed a meaning.

When R. A. Fisher said that a result was significant, he meant that the data showed some difference from the null hypothesis. In other words, one was able to learn something from those data.

This sense of the word “significant” has persisted in the scientific literature. When discussing data, a “statistically significant” result means that a null hypothesis has been rejected.

A statistically significant result is not the same as a biologically important result.

Full of sound and fury, signifying nothing. — Shakespeare, *Macbeth*

Confidence intervals

Confidence intervals

The **confidence interval** is another common way to quantify uncertainty about the value of a parameter. It is a range of numbers, calculated from the data, that is likely to contain within its span the unknown value of the target parameter.

A ***confidence interval*** is a range of values surrounding the sample estimate that is likely to contain the population parameter.

Confidence intervals can be calculated for means, proportions, correlations, differences between means, and other population parameters.

The ***95% confidence interval*** provides a most-plausible range for a parameter. Values lying within the interval are most plausible, whereas those outside are less plausible, based on the data.

Confidence intervals

The 2SE rule of thumb

A good “quick-and-dirty” approximation to the 95% confidence interval for the population mean is obtained by adding and subtracting two standard errors from the sample mean (the so called 2SE rule of thumb).

A rough approximation to the 95% confidence interval for a mean can be calculated as the sample mean plus and minus two standard errors.

The lower limit of the confidence interval is: $\bar{Y} - 2\ SE_{\bar{Y}}$.

Similarly, the uppper limit is: $\bar{Y} + 2\ SE_{\bar{Y}}$

One-sided or two-sides tests

Hypothesis Testing

One-sided tests

In a ***one-sided*** (or one-tailed) test, the alternative hypothesis includes parameter values on only one side of the value specified by the null hypothesis. H0 is rejected only if the data depart from it in the direction stated by HA.

For example, imagine a study designed to test whether daughters resemble their fathers. In each trial of the study a participant examines a photo of one girl and photos of two adult men, one of whom is the girl's father. The participant must guess which man is the father.

H0: Participants pick the father correctly half the time ($p = 1/2$).
HA: Participants pick the father correctly *more frequently than* half the time ($p > 1/2$).

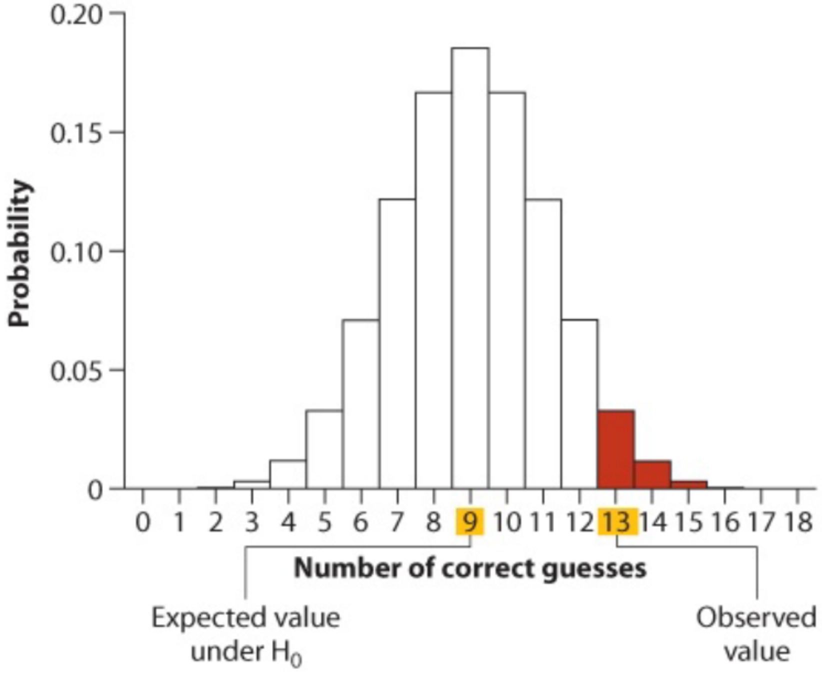
Now let's imagine that the study was carried out using 18 independent trials (which would require 18 different sets of photographs and participants) and that 13 out of 18 participants successfully guessed the father of the daughter.

Under the null hypothesis, we would expect only $18 \times 1/2 = 9$ correct guesses on average.

Hypothesis Testing

One-sided tests

The null distribution is



If X: no. of participants guessed the father correctly.

$$\textit{P-value} = P(X \geq 13) = P(X=13) + P(X=14) + P(X=15) + P(X=16) + P(X=17) + P(X=18)$$
$$= 0.0327 + 0.0117 + 0.0031 + 0.0006 + 0.00007 + 0.000004 = 0.048 < 0.05$$

(Thus, reject the null hypothesis).

Inference for a normal population

Inference for a normal population

The Z-test for sample means

As we know that the sampling distribution for the sample mean $\bar{Y}$ is a normal distribution if the variable *Y* is itself normally distributed. Thus,

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

However, this *Z*-standardization to $\bar{Y}$ can not be done for real data, as $\sigma_{\bar{Y}}$ is mostly unknown. However, the standard error can be estimated by

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

It is generally called **standard error of the mean**.

Inference for a normal population

The *t*-distribution for sample means

Student's *t*-distribution

As we know that the sampling distribution for the sample mean $\bar{Y}$ is a normal distribution if the variable *Y* is itself normally distributed. Thus,

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}}$$

However, this Z-standardization to $\bar{Y}$ can not be done for real data, as $\sigma_{\bar{Y}}$ is mostly unknown. However, the standard error can be estimated by

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

It is generally called **standard error of the mean**.

Substituting $SE_{\bar{Y}}$ for $\sigma_{\bar{Y}}$ in the Z-formula, we get

$$t = \frac{\bar{Y} - \mu}{SE_{\bar{Y}}}$$

This is called **Student's t-distribution** with n-1 degrees of freedom.

Inference for a normal population

Student's *t*-distribution

While the formula for *t* resembles that for *Z,* the important difference is that the sampling distribution for this statistic is not the normal distribution. Instead, *t* has a *t*-distribution.

$SE_{\bar{Y}}$ is not a constant like $\sigma_{\bar{Y}}$ but is a variable, varying by chance from sample to sample.

Therefore, the distribution of *t* is not the same as *Z*.

Substituting $SE_{\bar{Y}}$ for $\sigma_{\bar{Y}}$ adds sampling error to the quantity *t*.

As a result, the sampling distribution of *t* is wider than the standard normal distribution.

As the sample size increases, *t* becomes more like *Z*.

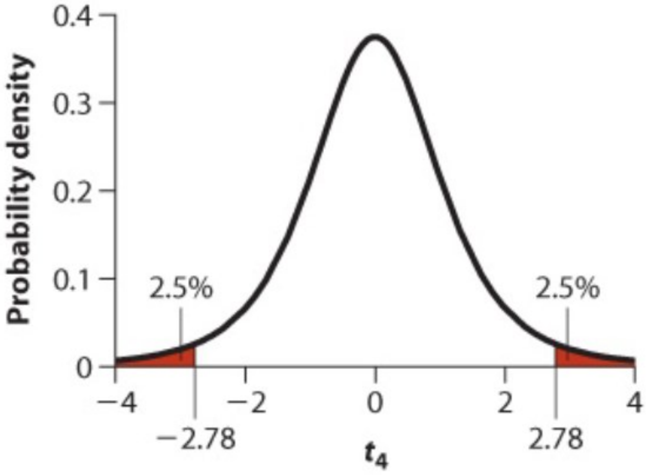
In most respects, the *t*-distribution is similar to the standard normal distribution. It is symmetric around a mean of zero, is roughly bell shaped, and it has tails that fall off toward plus infinity and minus infinity.

The *t*-distribution, however, is fatter in the tails than the standard normal distribution. The difference in the tails is crucial, because it is the tails that matter most when calculating confidence intervals and testing hypotheses.

Inference for a normal population

Finding critical values of the *t*-distribution

The value 2.78 is the “5% critical value” of the *t* distribution having *df* = 5 − 1 = 4. The 5% refers to the percentage of the area in the tails of the *t*-distribution.



We use the symbol $t_{0.05(2),\,df}$ to indicate the 5% critical *t*-value of a *t*-distribution having “*df*” degrees of freedom.

In this notation, the 0.05 stands for the fraction of the area under the curve lying in the tails of the distribution.

The “(2)” indicates that the 5% area is divided between the two tails of the *t*-distribution—that is, 2.5% of the area under the curve lies above $+t_{0.05(2),\,df}$ and 2.5% lies below $-t_{0.05(2),\,df}$.

Inference for a normal population

The confidence interval for the mean of a normal distribution

Generally, one approximates confidence interval for the mean using the two standard errors (2SE) rule of thumb. But it can be better estimated using *t*-distribution to calculate a more accurate confidence interval for the mean of a population having a normal distribution.

Example:

The span, in millimeters, from one eye to the other was measured in a random sample of nine male stalkeyed flies. The data are as follows:

8.69, 8.15, 9.25, 9.45, 8.96, 8.65, 8.43, 8.79, 8.63.

The 95% confidence interval for the mean

The mean and standard deviation of the eye-span sample are

$\bar{Y}$ =8.778 and s=0.398.

How precise is this estimate of the population mean?

Let's describe the precision by calculating a 95% confidence interval for the mean. For that, we will use the *t*-distribution.

Inference for a normal population

The confidence interval for the mean of a normal distribution

This means that in 95% of random samples from a normal distribution, the standardized difference will lie between

$$-t_{0\cdot\mathbf{05}(2),\mathbf{d}f} < \frac{\bar{Y}-\mu}{SE_{\bar{Y}}} < t_{0.05(2),\,\mathbf{d}f}$$

Re-arranging this equation gives,

$$\bar{Y} - t_{0.05(2),\mathrm{d}f} \times SE_{\bar{Y}} < \mu < \bar{Y} + t_{0\cdot05(2),\mathrm{d}f} \times SE_{\bar{Y}}$$

This is the exact 95% confidence interval for the mean.

Let’s calculate the 95% confidence interval for mean eye span in male stalk-eyed flies. To begin, we’ll need the standard error of the mean:

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0\cdot 398}{\sqrt{9}} = 0\cdot 133$$

And $t_{0.05(2),\,8}$ = 2.31.

Thus, the 95% confidence interval is

$$8.778 - (2.31 \times 0.133) < \mu < 8.778 + (2.31 \times 0.133) = 8.47 < \mu < 9.08 \text{ mm.}$$

Thus, the 95% confidence interval for the mean of the eye span in this species, calculated from this particular random sample, is from 8.47 mm to 9.08 mm.

The one-sample *t*-test

Inference for a normal population

Assumptions of the one-sample *t*-test

- The data are a random sample from the population.
- The variable is normally distributed in the population.

Inference for a normal population

The one-sample *t*-test

The **one-sample *t*-test,** is designed to compare the mean from a sample of individuals with a value for the population mean proposed in the null hypothesis.

H0: The true mean equals *μ*0.
HA: The true mean does not equal *μ*0.

Suppose, for example, we have a null hypothesis that the mean of a variable in a population is *μ*0 = 0.

We can then use the one-sample *t*-test to determine whether the $\bar{Y}$ that we calculate from a sample is sufficiently different from zero to warrant rejection of the null hypothesis.

The test statistic for the one-sample *t*-test is *t* (no surprise),

$$t = \frac{\bar{Y} - \mu_0}{SE_{\bar{Y}}}$$

The sampling distribution of this test statistic under H0 is the *t*-distribution having *n* − 1 degrees of freedom.

Inference for a normal population

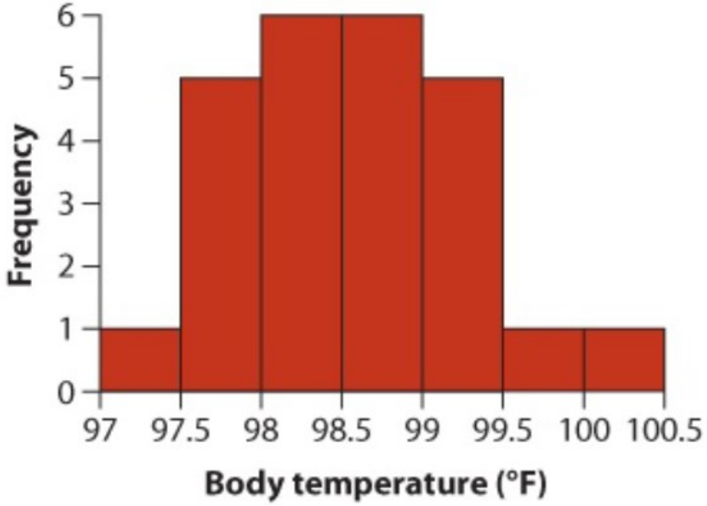
The one-sample *t*-test
Example

Normal human body temperature is 98.6°F. But how well is this supported by data? Researchers obtained body-temperature measurements on randomly chosen healthy people. The data for 25 of those people are as follows:

98.4, 98.6, 97.8, 98.8, 97.9, 99.0, 98.2, 98.8, 98.8, 99.0, 98.0, 99.2, 99.5, 99.4, 98.4, 99.1, 98.4, 97.6, 97.4, 97.5, 97.5, 98.8, 98.6, 100.0, 98.4.

H0: The mean human body temperature is 98.6°F.
HA: The mean human body temperature is different from 98.6°F.

The frequency distribution of body temperatures in a sample of 25 individuals.



The sample mean $\bar{Y}$ = 98.524
The sample standard deviation s = 0.678
The sample standard error $SE_{\bar{Y}}$ = √s = √0.678 = 0.136

The test statistic $t = \frac{\bar{Y}-\mu_0}{SE_{\bar{Y}}}$ = 98.524 – 98.6 / 0.136 = -0.5588

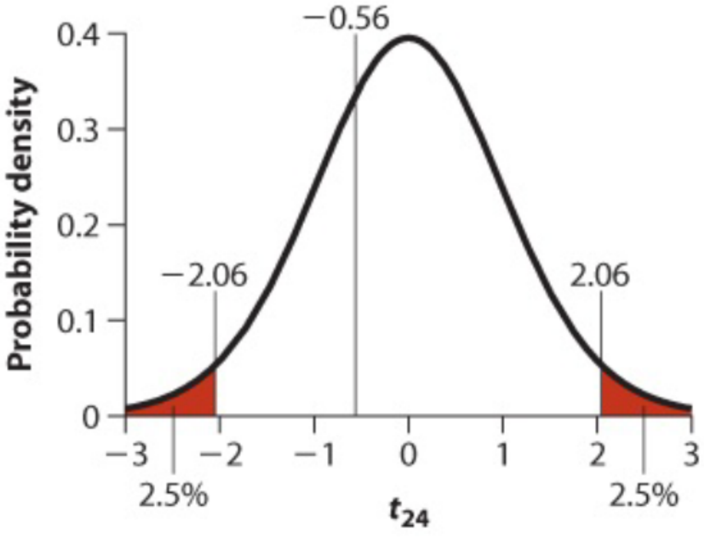
Inference for a normal population

The one-sample *t*-test
Example
Using t-distribution, $t_{0.05(2), 24}$=2.064.

The *t*-value of −0.56 that we calculated from the data occurs inside this range. The observed *t*-statistic does not fall within one of the tails. Therefore, *P* > 0.05, and we fail to reject the null hypothesis.

The 95% confidence interval for the mean

$$98.24 ◦F < \mu < 98.80 ◦F.$$



Comparing two means

Comparing two means

Here, we discuss procedures for comparing means of a numerical variable between two treatments or groups. All of the methods here assume that the measurements are normally distributed in the populations.

Paired sample versus two independent samples
Let’s use an example: does clear-cutting a forest affect the number of salamanders present?

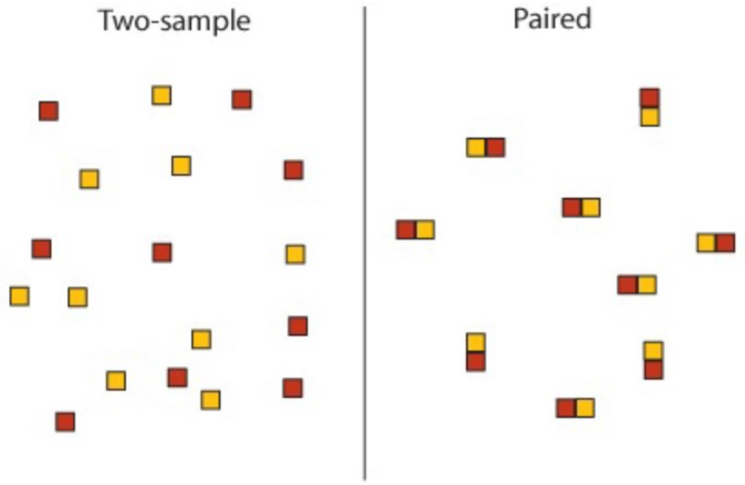
Here we have two treatments (clear-cutting/no clear-cutting), and we want to know if the mean of a numerical variable (the *number of salamanders*) differs between them.

“Clear-cut” is the treatment of interest, and “no clear-cut” is the control.

This is the same as asking whether these two variables, *treatment* (a categorical variable) and *salamander number* (a numerical variable), are associated.

We can design this kind of a study in either of two ways: a paired design or a two-sample design.

Comparing two means



In the *two sample* design, we take a random sample of forest plots from the population and then randomly assign either the clear-cut treatment or the no-clear-cut treatment to each plot. In this case, we end up with two independent samples, one from each treatment. The difference in the mean number of salamanders between the clear-cut and no-clear-cut areas estimates the effect of clear-cutting on salamander number.

In the *paired* design, we take a random sample of forest plots and clear-cut a randomly chosen half of each plot, leaving the other half untouched. Afterward, we count the number of salamanders in each half. The mean difference between the two sides estimates the effect of clear-cutting.

Comparing two means

Paired sample versus two independent samples

In the ***paired*** design, both treatments are applied to every sampled unit. In the *two-sample* design, each treatment group is composed of an independent, random sample of units.

Paired comparison of means

In paired samples, if 20 individuals are grouped into 10 pairs, there are 10 measurements of the difference between the two treatments. Ten would be the sample size. We can estimate and test the effect of treatment using the mean of the differences. Paired measurements are converted to a single measurement by taking the difference between them.

Estimating mean difference from paired data

This method assumes that we have a random sample of pairs and that the differences between members of each pair have a normal distribution.

Comparing two means

Example:

In many species, males are more likely to attract females if the males have high testosterone levels.

Are males with high testosterone paying a cost for this extra mating success in other ways?

One hypothesis is that males with high testosterone might be less able to fight off disease—that is, their high levels of testosterone might reduce their immunocompetent.

To test this idea, one experimentally increased the testosterone levels of 13 male red-winged blackbirds by surgically implanting a small permeable tube filled with testosterone.

They measured immunocompetence as the rate of antibody production in response to a nonpathogenic antigen in each bird's blood serum both before and after the implant.

The antibody production rates were measured optically, in units of log 10^{-3} optical density per minute (ln[mOD/min]).

Comparing two means

Example

What is the mean difference between the two treatments?

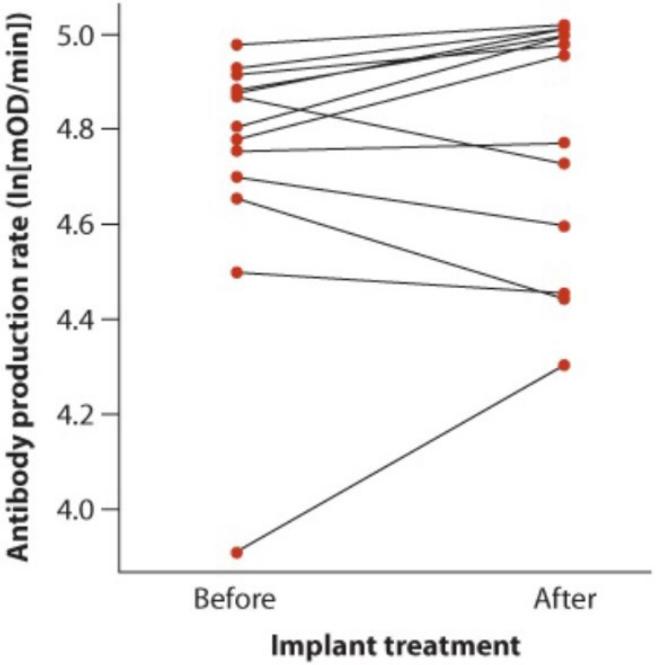
Male id	APBI	APAI	Difference (d)
1	4.65	4.44	-0.21
4	3.91	4.30	0.39
5	4.91	4.98	0.07
6	4.50	4.45	-0.05
9	4.80	5.00	0.20
10	4.88	5.00	0.12

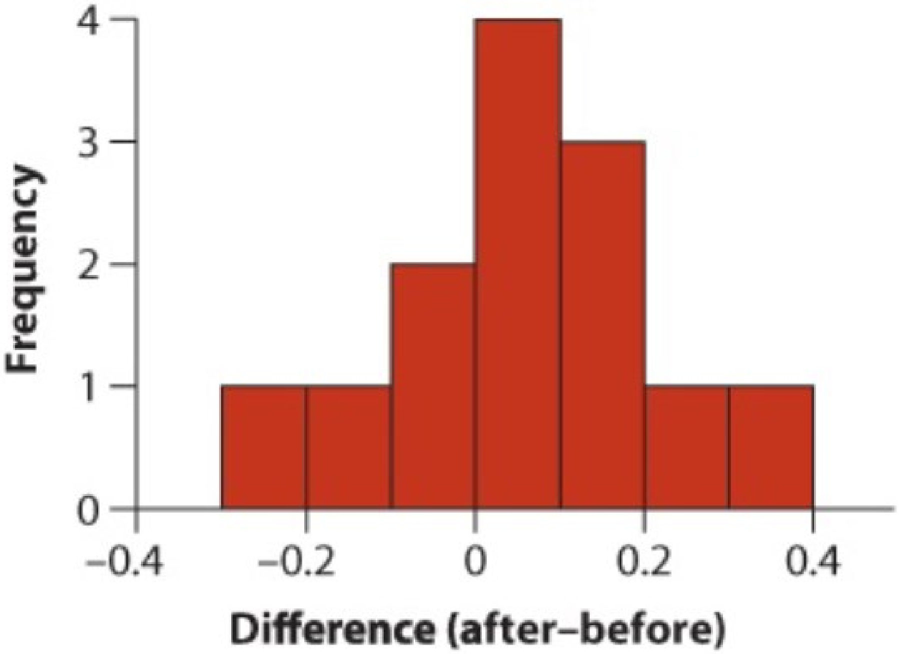
Male id	APBI	APAI	Difference (d)
15	4.88	5.01	0.13
16	4.78	4.96	0.18
17	4.98	5.02	0.04
19	4.87	4.73	-0.14
20	4.75	4.77	0.02
23	4.70	4.60	-0.10
24	4.93	5.01	0.08

Measurement unit: (ln[mOD/min]),
APBI: Antibody production before implant
APAI: Antibody production after implant

Comparing two means

Example





From the given data, we can calculate

$\bar{d}$=0.056, s_d = 0.159, and n=13.

Comparing two means

Example

The confidence interval for the mean difference (μ_d) is

$$\bar{d} - t_{\alpha(2),\mathrm{d}f} \times SE_{\bar{d}} < \mu_d < \bar{d} + t_{\alpha(2),df} \times SE_{\bar{d}}$$

Where

$$SE_{d^-} = \frac{s_d}{\sqrt{n}} = \frac{0.159}{\sqrt{13}} = 0.044 \text{ and } t_{0.05(2),\ 12} = 2.18.$$

Thus, the confidence interval is

$$0.056 - 2.18(0.044) < \mu_d < 0.056 + 2.18(0.044) \Rightarrow -0.040 < \mu_d < 0.152.$$

In other words, the most-plausible range for the true mean difference is between −0.040 and 0.152 ln[mOD/min].

Comparing two means

Paired t-test

The **paired *t*-test** is used to test a null hypothesis that the mean difference of paired measurements equals a specified value.

For the previous example of antibody production after implant,

H0: The mean change in antibody production after testosterone implants was zero. Or $H_0: \mu_d = 0$

HA: The mean change in antibody production after testosterone implants was not zero. Or $H_\mathrm{A}: \mu_d \neq 0$

where μ_d is the population mean difference between treatments.

We have already calculated

$\bar{d}$=0.056, s_d = 0.159, and n=13, $SE_{\bar{d}} = 0.044$,

From here on, the paired *t*-test is identical to a one-sample *t*-test on the differences. We can calculate the *t*-statistic as

$$t = \frac{\bar{d} - \mu_{d_0}}{SE_{\bar{d}}} = (0.056 – 0) / 0.044 = 1.27$$

Comparing two means

Paired t-test

And

$$t_{0.05(2),\ 12}=2.18.$$

Because *t* = 1.27 does not fall outside the critical limits of −2.18 and 2.18, at 0.05 significance level; we do not reject the null hypothesis.

Comparing two means

Two-sample comparison of means

Let us now discuss methods to analyze the difference between the means of two treatments or groups in the case of a two-sample design. In a two-sample design, the two treatments are applied to separate, independent samples from two populations.

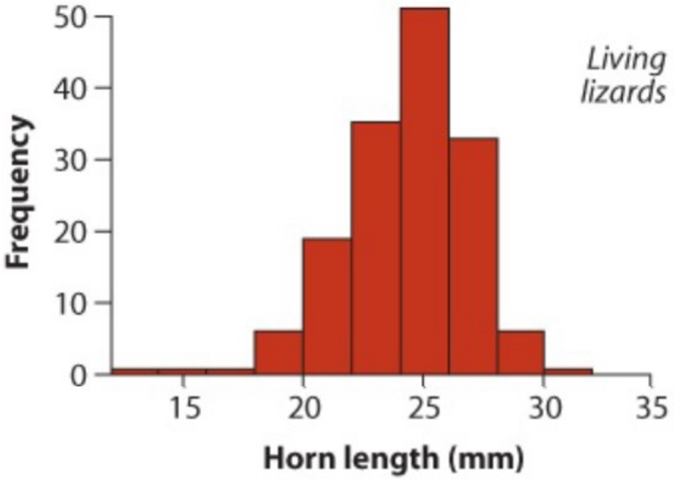
Example

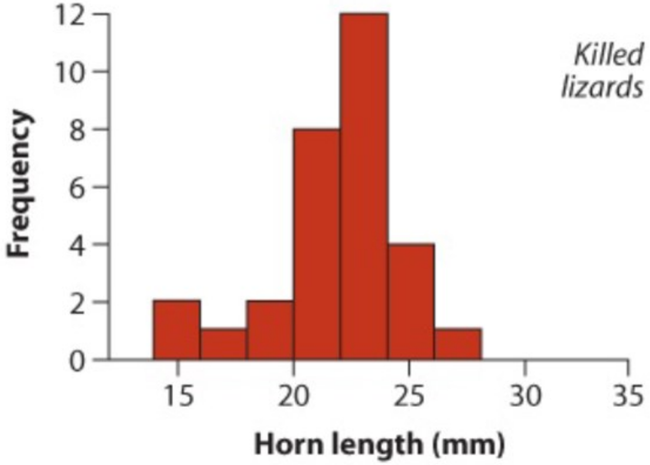
The horned lizard *Phrynosoma mcallii* has many unusual features, including the ability to squirt blood from its eyes. The species is named for the fringe of spikes surrounding the head. Herpetologists recently tested the idea that long spikes help protect horned lizards from being eaten, by taking advantage of the gruesome but convenient behavior of one of their main predators—the loggerhead shrike, *Lanius ludovicianus.* The loggerhead shrike is a small predatory bird that skewers its victims on thorns or barbed wire, to save for later eating.

The researchers identified the remains of 30 horned lizards that had been killed by shrikes and measured the lengths of their horns. As a comparison group, they measured the same trait on 154 horned lizards that were still alive and well. They compared the mean horn lengths of the dead lizards with those of the living lizards.

Comparing two means

Two-sample comparison of means
Example
Histograms of the horn lengths of the two groups are shown





Lizard group	Sample mean ($\bar{Y}$) in mm	Sample std. (s) in mm	Sample size
Living	24.28	2.63	154
Killed	21.88	2.71	30

Note: the lizards that were killed by shrikes are *different individuals* than the living lizards.

Comparing two means

Two-sample comparison of means

Example
Confidence interval for the difference between two means
How much longer on average are the horns of the surviving lizards?

The best estimate of the difference between two population means is the difference between the sample means, $\bar{Y}_1 - \bar{Y}_2$.

The method for confidence intervals makes use of the fact that, if the variable is normally distributed in both populations, then the sampling distribution for the *difference* between the sample means is also normal.

Thus, the Student's *t*-distribution will be very helpful in describing the sampling properties of the standardized difference.

Comparing two means

Two-sample comparison of means
Example
Confidence interval for the difference between two means

The standard error of $\overline{Y_1} - \bar{Y}_2$ is

$$SE_{\overline{Y_1}-\bar{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where s_p^2 is called the **pooled sample variance** and is given by

$$s_p^2 = \frac{\mathrm{d}f_1 \times s_1^2 + \mathrm{d}f_2 \times s_2^2}{\mathrm{d}f_1 + \mathrm{d}f_2}$$

Comparing two means

Two-sample comparison of means
Example
Confidence interval for the difference between two means

Because the sampling distribution of $\overline{Y_1} - \bar{Y}_2$, is normal, the sampling distribution of the following standardized difference has a Student's *t*-distribution:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE_{\overline{Y_1} - \bar{Y}_2}}$$

From these two formulas, we can calculate the confidence interval for the difference between two population means:

$$(\bar{Y}_1 - \bar{Y}_2) - t_{\alpha(2),\mathrm{d}f} \times SE_{\overline{Y_1} - \bar{Y}_2} < \mu_1 - \mu_2 < (\bar{Y}_1 - \bar{Y}_2) + t_{\alpha(2),\mathrm{d}f} \times SE_{\overline{Y_1} - \bar{Y}_2}$$

Comparing two means

Two-sample comparison of means
Example
Confidence interval for the difference between two means

Now let us use the given values,

$$\bar{Y}_1 - \bar{Y}_2 = 24.28 - 21.99 = 2.29 \text{ mm}$$

The pooled sample variance,

$$s_p^2 = \frac{\mathrm{d}f_1 \times s_1^2 + \mathrm{d}f_2 \times s_2^2}{\mathrm{d}f_1 + \mathrm{d}f_2}$$

$$= \frac{153(2.63^2)+29(2\cdot71^2)}{153+29} = 6.98$$

The standard error of the difference between the two means is then

$$SE_{\overline{Y_1}-\bar{Y}_2} = \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= \sqrt{6.98\left(\frac{1}{154}+\frac{1}{30}\right)} = 0.527$$

Comparing two means

Two-sample comparison of means
Example
Confidence interval for the difference between two means

The critical value of *t* for *df* = 182:

$$t_{0.05(2),\ 182}=1.97.$$

By plugging these quantities into the formula, we find the 95% confidence interval for the difference in mean horn length between the living and dead lizards is

$$2.29 - 1 \cdot 97(0 \cdot 527) < \mu_1 - \mu_2 < 2 \cdot 29 + 1.97(0 \cdot 527)$$
$$1.25 < \mu_1 - \mu_2 < 3.33$$

Thus, the 95% confidence interval for *μ*1 − *μ*2 is from 1.25 to 3.33 mm. We can be reasonably confident that surviving lizards have longer horns than lizards killed by shrikes, by an amount somewhere between 1.25 and 3.33 millimeters.

Comparing two means

Two-sample comparison of means
Example

Two-sample t-test
The *two-sample t-test* is the simplest method to compare the means of a numerical variable between two independent groups.

Its most common use is to test the null hypothesis that the means of two populations are equal.
H0: $\mu_1 = \mu_2$ and HA: $\mu_1 \neq \mu_2$

For example above (living and killed lizards) example,
H0: Lizards killed by shrikes and living lizards do not differ in mean horn length (i.e., $\mu 1 = \mu 2$).
HA: Lizards killed by shrikes and living lizards differ in mean horn length (i.e., $\mu 1 \neq \mu 2$).
So,

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{SE_{\overline{Y_1} - \bar{Y}_2}} = 2.29 / 0.527 = 4.35$$

Since $t_{0.05(2),\ 182} = 1.97$.

The t = 4.35 calculated from these data is much further into the tail of the distribution than this critical value, so we reject the null hypothesis. Based on these studies, there is a difference in the horn length of lizards eaten by shrikes, compared with live lizards.

Fitting probability models to frequency data : Chi-square test

Chi-square test

Goodness-of-fit test

A **goodness-of-fit test** is a method for comparing an observed frequency distribution with the frequency distribution that would be expected under a simple probability model governing the occurrence of different outcomes.

Goodness-of-fit test allows one to handle categorical and discrete numerical variables having more than two outcomes.

Chi-square test

The proportional model

The **proportional model** is a simple probability model in which the frequency of occurrence of events is proportional to the number of opportunities.

Example

Given the data below, are babies born at the same frequency on all seven days of the week?

Day	No. of births
Sunday	33
Monday	41
Tuesday	63
Wednesday	63
Thursday	47
Friday	56
Saturday	47
Total	350

Under the proportional model, which will be the null hypothesis, the number of births on Monday should be proportional to the numbers of Mondays in the year, except for chance differences.

The same should be true for the other days of the week.

Does the variation among days evident in the table represent only chance variation?

We can test the fit of the proportional model to the data with a χ2 goodness-of-fit test.

Chi-square test

***χ*2 goodness-of-fit test**

The ***χ*2 goodness-of-fit** test uses a test statistic called χ2 to measure the discrepancy between an observed frequency distribution and the frequencies expected under a simple probability model serving as the null hypothesis.

The simple model is rejected if the discrepancy, χ2, is too large. The χ2 ***goodness-of-fit test*** compares frequency data to a probability model stated by the null hypothesis.

For the previous example, under the proportional model, each day of the week should have the same probability of a birth, that is, 1/7 and

H0: The probability of birth is the same on every day of the week.
HA: The probability of birth is *not* the same on every day of the week.

Chi-square test

Observed and expected frequencies

Because the proportional model is the null hypothesis, we use it to generate the expected frequency of births on each day of the week.

Day	No. of days in year	Proportion of days in year	Expected freq. of births
Sunday	52	52/365	49.863
Monday	52	52/365	49.863
Tuesday	52	52/365	49.863
Wednesday	52	52/365	49.863
Thursday	52	52/365	49.863
Friday	53	53/365	50.822
Saturday	52	52/365	49.863
Total	365	1	350

Expected frequency = total no. of births x proportion = 350 x 52/365 = 49.863 and so on.

Chi-square test

The χ2 test statistic

The χ2 statistic measures the discrepancy between the observed and expected frequencies. It is calculated as

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Where Oi: ith observed and Ei: ith expected value. It's important to notice that the χ2 calculations use the ***absolute*** frequencies (i.e., counts) for the observed and expected frequencies, not proportions or *relative* frequencies.

Day	Observed no. of births	Expected no. of births	$\frac{(\mathbf{O_i} - \mathbf{E_i})^2}{\mathbf{E_i}}$
Sunday	33	49.863	5.70
Monday	41	49.863	1.58
Tuesday	63	49.863	3.46
Wednesday	63	49.863	3.46
Thursday	47	49.863	0.16
Friday	56	50.822	0.53
Saturday	47	49.863	0.16
Total	350	350	15.05

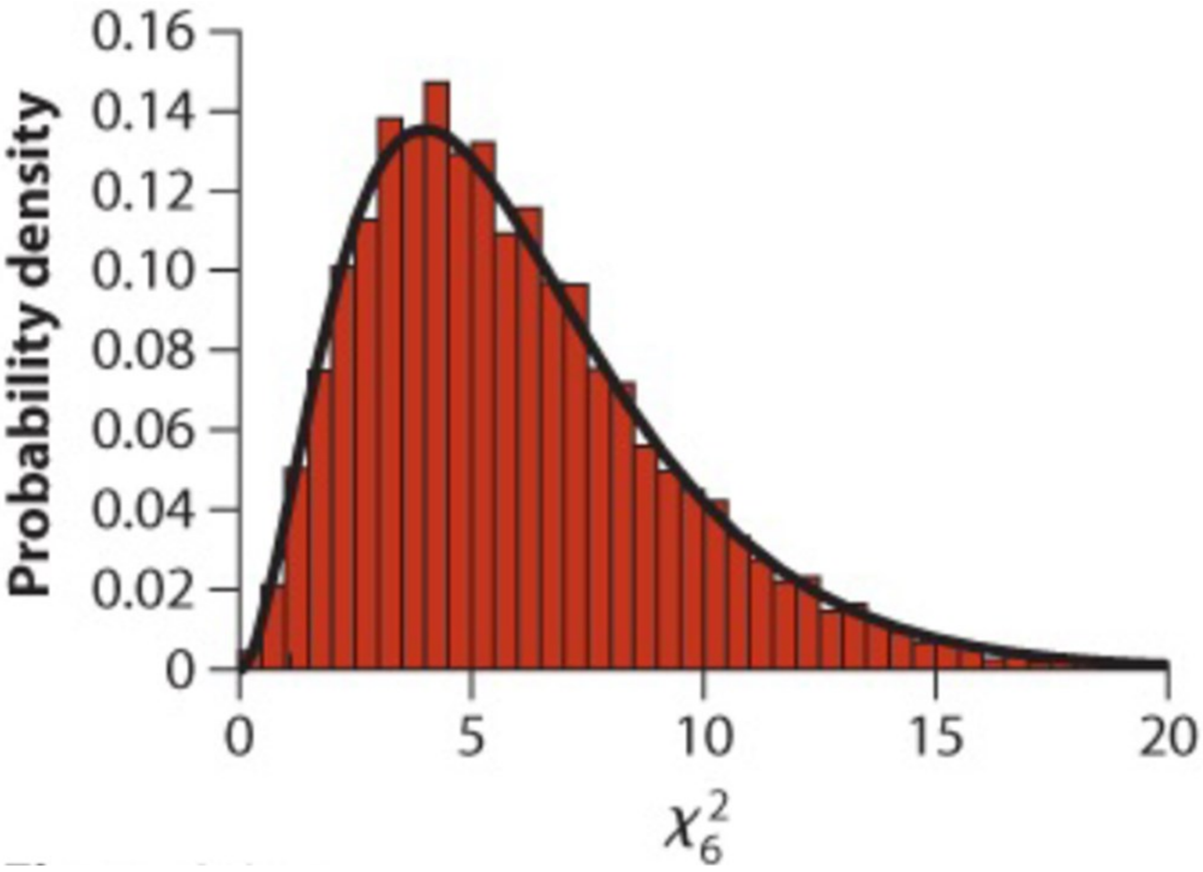
The χ2 statistic is the test statistic for the χ2 goodness-of-fit test, the quantity measuring the level of agreement between the data and the null hypothesis.

All we need know is the sampling distribution of the χ2 test statistic under H0.

This will allow us to decide whether χ2 = 15.05 is large enough to warrant rejection of the null hypothesis.

Chi-square test

The sampling distribution of χ2 under the null hypothesis
By simulation for these data, the approximate null distribution for the χ2 statistic looks like



The number of ***degrees of freedom*** of a χ2 statistic specifies which χ2 distribution to use as the null distribution.

df = (no. of categories) – 1 – (no. of parameters estimated from the data).

In the example above, no. of parameters estimated from the data = 0. Thus, for the above example,
df = 7 – 1 = 6.

This tells us that we need to compare our χ2 value calculated from the birth data (χ2 =15.05) to the χ^2_6 distribution with six degrees of freedom.

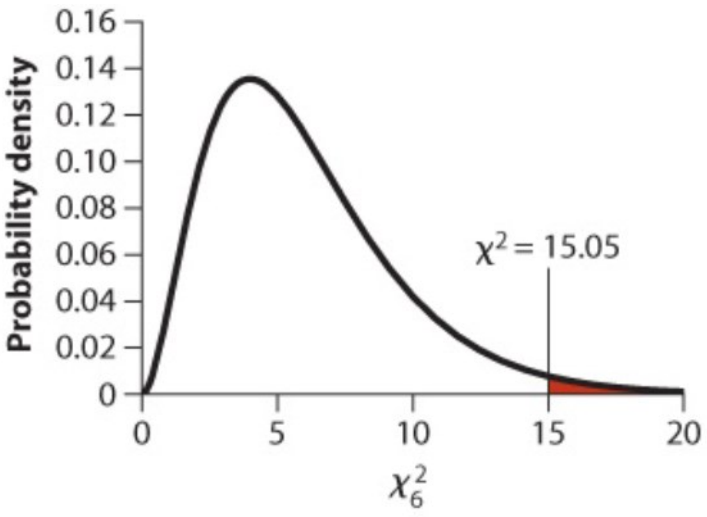
Chi-square test

Calculating the *P*-value

The *P*-value for a test is the probability of getting a result as extreme as, or more extreme than, the result observed if the null hypothesis were true.

Remember that if the data exactly matched the expectation of the null hypothesis, χ2 would be zero.

A deviation in either direction between an observed frequency and the expected frequency causes χ2 to be greater than zero. Greater deviations from the null expectation result in a higher value of χ2. As a result, we use only the right tail of the χ2 distribution to calculate *P*.



How do we go about finding this area beyond the measured χ2 value?

Chi-square test

Critical values for the χ2 distribution
A ***critical value*** is the value of a test statistic that marks the boundary of a specified area in the tail (or tails) of the sampling distribution under H0.

Using the χ2 table, we get

$$\chi^2_{0.05,6} = 12 \cdot 592$$

Under the null hypothesis, the probability of obtaining a χ2 value as extreme as, or more extreme than, 12.59 is 0.05:

Because our observed χ2 value (15.05) is greater than 12.59 (i.e., further out in the right tail of the distribution), χ2 values of 15.05 or greater occur more rarely under the null hypothesis than 5% of the time.

So we reject the null hypothesis.

Which test should I use?

Commonly used statistical tests for data on a single variable. These methods test whether a population parameter equals the value proposed in the null hypothesis or whether a specific probability model fits a frequency distribution.

Data Type	Goal	Test
Categorical	Use frequency data to test whether a population proportion equals a null hypothesized value	Binomial test, χ2 Goodness-of-fit test with two categories (when n is large)
	Use frequency data to test the fit of a specific population model	χ2 Goodness-of-fit test
Numerical	Test whether the mean equals a null hypothesized value when data are approximately normal	One-sample *t*-test
	Test whether the median equals a null hypothesized value when data are not normal	Sign test
	Use frequency data to test the fit of a discrete probability distribution	χ2 Goodness-of-fit test
	Use data to test the fit of the normal distribution	Shapiro-Wilk test

Which test should I use?

Commonly used tests of association between two variables.

		Type of explanatory variable	
Type of response variable		**Categorical**	**Numerical**
	Categorical	Contingency analysis	Logistic regression
	Numerical	*t*-tests, ANOVA, Mann-Whitney *U*-test, etc.	Linear and nonlinear regression Linear correlation (16) Spearman’s rank correlation (when data are not bivariate normal)

Which test should I use?

A comparison of methods to test differences between group means according to whether the tests assume normal distributions.

Number of Treatments	**Tests assuming normal distribution**	**Tests not assuming normal distributions**
Two treatments (independent samples)	Two-sample *t*-test Welch's *t*-test (used when variance is unequal in the two groups)	Mann-Whitney *U*-test
Two treatments (paired data)	Paired *t*-test	Sign test
More than two treatments	ANOVA	Kruskal-Wallis test

Thank You