The Significance of Evidence

7

Imagine you believe the mean human body weight is 72 kg. The mean human weight isn't a random number, but it's very hard to measure directly. You are forced to take a sample, and compute the sample mean. This sample mean is a random variable, and it will have different values for different samples. You need to know how to tell whether the difference between the observed value and 72 kg is just an effect of variance caused by sampling, or is because the mean weight actually isn't 72 kg. One strategy is to construct an interval around the sample mean within which the true value will lie for (say) 99% of possible samples. If 72 kg is outside that interval, then very few samples are consistent with the idea that 72 kg is the mean human body weight. If you want to believe the mean human body weight is 72 kg, you have to believe that you obtained a very odd sample.

You should think of the procedure I described as assessing the extent to which the evidence you have contradicts the original hypothesis. At first glance, this may seem strange to you—surely one wants to assess the extent to which the evidence *supports* the hypothesis—but in fact it's natural. You can't prove that a scientific hypothesis is true; you can only fail to show that it's false. Just one piece of evidence can destroy a scientific hypothesis, but no amount of evidence can remove all doubt.

There is an important, quite general, line of reasoning here. It is a bad idea to try and explain data using a hypothesis that makes the data you observed a rare event. We can use the reasoning of Sect. 6.2 to assess how rare the observed data is. In that section, we used the distribution that the sample mean would take to construct a confidence interval. This meant we could plot an interval in which the population mean would lie with (say) 95% confidence. To assess the rarity of the sample, we could ask how large a confidence interval we would have to draw around the hypothesized mean to cover the observed sample mean. If that interval is relatively small (say 50%), then it's quite possible that the population mean takes the value we hypothesized. A cleaner way to say this is we do not have enough evidence to reject the hypothesis. If that interval requires (say) 99.99% of possible samples, that's a strong suggestion that the sample is extremely unusual. Assessing the rarity of the sample using methods like this is usually talked about as testing the significance of evidence against the hypothesis.

Example 7.1 (Patriot Missiles) I got this example from "Dueling idiots", a nice book by P.J. Nahin, Princeton University Press. Apparently in 1992, the Boston Globe of Jan 24 reported on this controversy. The pentagon claimed that the patriot missile successfully engaged SCUD missiles in 80% of encounters. An MIT physicist, Theodore Postol, pointed out there was a problem. He viewed tapes of 14 patriot/SCUD encounters, with one hit and 13 misses. We can reasonably assume each encounter is independent. We can extract the probability of getting one hit and 13 misses if $P(\text{hit}) = 0.8$ from the binomial model, to get a number around 1e-8. Now you could look at this information and make several arguments: (a) the pentagon is right and the probability really is 0.8, but Postol looked at a really unlucky set of videotapes; (b) the probability is not 0.8, because you would need to fire 14 patriots at 14 SCUD missiles about 1e8 times to see this set of videotapes once; (c) for some reason, the videotapes are not independent—perhaps only unsuccessful encounters get filmed. If Postol viewed tapes at random (i.e. he didn't select only unsuccessful tapes, etc.), then argument (a) is easily dismissed, because the pentagon would have had to be unreasonably unlucky—it's a bad idea to try to explain data with a hypothesis that makes the data very unlikely.

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_7

This reasoning can be extended to compare populations. Imagine I want to know whether mice weigh more than rats. For practical reasons, I will estimate the weights using a sample. But this means that two different estimates will have different values purely because I used different samples. I am now in a difficult position—perhaps my observed sample mean for the weight of mice is smaller than that for rats because of random variation in the sample value. But now imagine drawing a (say) 95% confidence interval about each mean—if these don't overlap, then the population means are likely not the same. If you did the exercises for the last section, you will have noticed that I was signalling this idea. The principle here is that a very large difference may be hard to explain with random variation unless you are willing to believe in very odd samples. This leads to a procedure that can be used to decide whether (say) mice weigh more than rats.

7.1 Significance

Imagine we hypothesize that the average human body temperature is 95°. We collect temperature measurements x_i from a random sample of N people. The mean of this sample is unlikely to be 95°. The sample will likely have too many people who run too hot, or too cool, to get exactly the number we expect. We must now find what caused the difference between the sample mean and the value we hypothesized. We could be wrong about the average body temperature. Alternatively, we could be right, and the difference might just be because the sample is randomly chosen. We can assess the **significance** of the evidence against the hypothesis by finding out what fraction of samples would give us sample means like the one we observe *if* the hypothesis is true.

7.1.1 Evaluating Significance

We hypothesize that a population mean has some value $\mathsf{popmean}\,(\{X\})$ (a big letter, because we don't see the whole population). Write S for the random variable representing a possible sample mean. The mean of this random variable is the population mean, and the standard deviation of this random variable is estimated by the standard error, which we write $\mathsf{stderr}\,(\{x\})$ (small letters, because we got this from the sample) Now consider the random variable

$$T = \frac{(S - \mathsf{popmean}\,(\{X\}))}{\mathsf{stderr}\,(\{x\})}.$$

This random variable has a t-distribution with $N-1$ degrees of freedom (or, if N is big enough, we can regard it as a standard normal random variable). We now have a way to tell whether the evidence supports our hypothesis. We assess how strange the sample would have to be to yield the value that we actually see, *if* the hypothesis is true. We can do this, because we can compute the fraction of samples that would have a less extreme value. Write s for the value of S that we observe. This yields t (the observed value of T, which is usually known as the **test statistic**). Write $p_t(u; N-1)$ for the probability density function of a t-distribution with $N-1$ degrees of freedom. Now the fraction of samples that will have less extreme values of s *if* the population mean was, indeed, $\mathsf{popmean}\,(\{X\})$ is:

$$f = \frac{1}{\sqrt{2\pi}} \int_{-|s|}^{|s|} p_t(u; N-1) du.$$

Remember that, if N is sufficiently large, we can use a standard normal distribution in place of $p_t(u; N-1)$. Now assume we see a very large (or very small) value of v. The value of f will be close to one, which implies that most samples from the population will have a v value closer to zero if the hypothesis were true. Equivalently, this says that, if the hypothesis were true, our sample is highly unusual, which implies the data fails to support the hypothesis.

Worked example 7.1 (Samples of 44 Male Chow Eating Mice) Assume the mean weight of a male chow eating mouse is 35 gr. and the standard error of a sample of 44 such mice is 0.827 gr. What fraction of samples of 44 such mice will have a sample mean in the range 33–37 grams?

Solution You could use the data at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml, but you don't really need it to answer this question. Write S for the sample mean weight of a sample of 44 male chow eating mice. Because we assumed that the true mean weight is 35 gr, we have

$$T = \frac{S - 35}{0.827}$$

is a t-distributed random variable, with 43 degrees of freedom. The question is asking for the probability that T takes a value in the range $[(33 - 35)/0.827, (37 - 35)/0.827]$, which is $[-2.41, 2.41]$. There are enough degrees of freedom to regard S as normal, so this probability is

$$\int_{-2.41}^{2.41} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2) du \approx 0.984$$

a number I found in tables. In turn, this means about 98.4% of samples of 44 chow eating male mice will give a mean weight in this range, if the population mean is truly 35 gr.

Worked example 7.2 (Samples of 48 Chow-Eating Female Mice) Assume the population mean of the weight of a chow-eating female mouse is 27.8 gr. Use the data at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml to estimate the fraction of samples that will have mean weight greater than 29 gr.

Solution From Worked example 6.4, the standard error of a sample of 48 chow-eating female mice is 0.70 gr. Write S for a sample mean of a sample of 48 chow-eating female mice. Because we assumed that the true mean was 27.8 gr, we have

$$T = \frac{S - 27.8}{0.70}$$

is a t-distributed random variable, with 47 degrees of freedom. The question is asking for the probability that T takes a value greater than $(29 - 27.8)/0.7 = 1.7143$. There are enough degrees of freedom to regard T as normal, so this probability is

$$\int_{1.7143}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) dx \approx 0.043$$

a number I found in tables. In turn, this means about 4% of samples of 48 chow eating female mice will give a mean weight greater than 29 gr, if the population mean is truly 27.8 gr

7.1.2 P-Values

The procedure of the previous section computes the fraction of samples that would give a smaller absolute value of T than the one we observed *if* the hypothesis was true. I called this fraction f. It is easier (and more traditional) to think about $p = 1 - f$ than about f. You should think of p as representing the fraction of samples that would give a larger absolute value of T than the one observed, *if* the hypothesis was true. If this fraction is very small, then there is significant evidence against the hypothesis. The fraction is sometimes referred to as a **p-value**.

Definition 7.1 (p-Value) The p-value represents the fraction of samples that would give a more extreme value of the test statistic than that observed, *if* the hypothesis was true.

Here is one useful interpretation of a p-value. Assume that you are willing to believe the hypothesis when the p-value is α or less, and will reject it otherwise. Then the probability that you accept a hypothesis that is false (i.e. a false positive, or type I error) is the fraction of samples that would give you that p-value or less *even though* the hypothesis is true. But this is α—so you can interpret a p-value as the probability that you have accepted a hypothesis that is false. This view yields the definition of significance, which is delicate.

Definition 7.2 (Statistical Significance) Statistical significance is a term that is used in a variety of ways. One can refer to the significance of a set of measurements. In this context, the term means the p-value for the relevant test statistic for those measurements. One can refer to the significance of a study. In this context, the term refers to the value against which the p-value of the test statistic will be tested. This value is, ideally, chosen in advance. It should be interpreted as meaning the fraction of all possible samples that the study could encounter that would cause the chosen procedure to reject the null hypothesis, given that it was true.

The procedure I have described for evaluating evidence against a hypothetical population mean is known as a **T-test**. I have put the details in box 7.1. The procedure is worth knowing because it is useful and very widely used. It also sketches the general form for tests of significance. You determine a statistic which can be used to test the particular proposition you have in mind. This statistic needs to: (a) depend on your data; (b) depend on your hypothesis; and (c) have a known distribution under sampling variation. You compute the value of this statistic. You then look at the distribution to determine what fraction of samples would have a more extreme value. If this fraction is small, the evidence suggests your hypothesis isn't true.

Procedure 7.1 (The T-Test of Significance for a Hypothesized Mean) The initial hypothesis is that the population has a known mean, which we write μ. Write $\{x\}$ for the sample, and N for the sample size.

- Compute the sample mean, which we write $\mathsf{mean}\,(\{x\})$.
- Estimate the standard error $\mathsf{stderr}\,(\{x\})$ using

$$\mathsf{stderr}\,(\{x\}) = \frac{\mathsf{stdunbiased}\,(\{x\})}{\sqrt{N}}.$$

- Compute the test statistic using

$$v = \frac{(\mu - \mathsf{mean}\,(\{x\}))}{\mathsf{stderr}\,(\{x\})}.$$

- Compute the p-value, using one of the recipes below.
- The p-value summarizes the extent to which the data contradicts the hypothesis. A small p-value implies that, *if* the hypothesis is true, the sample is very unusual. The smaller the p-value, the more strongly the evidence contradicts the hypothesis.

It is common to think that a hypothesis can be rejected only if the p-value is less than 5% (or some number). You should not think this way; the p-value summarizes the extent to which the data contradicts the hypothesis, and your particular application logic affects how you interpret it.

There is more than one way to compute a p-value. In one approach, we compute the fraction of experiments that would give us a larger absolute value of t than the one we saw, computing

$$p = (1 - f) = 1 - \int_{-|s|}^{|s|} p_t(u; N-1)du$$

Here the probability distribution we use is either a t-distribution with $N-1$ degrees of freedom, or a normal distribution if N is sufficiently large. Recall I wrote S for the sample mean as a random variable (i.e. before we've actually drawn a sample) and s for the value of that random variable. You should interpret p using

$$p = P(\{S > |s|\}) \cup P(\{S < -|s|\}).$$

It's usual to look this number up in tables; alternatively, any reasonable math computing environment will produce a number. This is known as a **two-sided p-value** (because you are computing the probability that either $\{S > |s|\}$ or $\{S < -|s|\}$).

Procedure 7.2 (Computing a Two-Sided p-Value for a T-Test) Evaluate

$$p = (1 - f) = 1 - \int_{-t}^{t} p_t(u; N-1)du = P(\{S > |s|\}) \cup P(\{S < -|s|\})$$

where $p_t(u; N-1)$ is the probability density of a t-distribution. If $N > 30$, it is enough to replace p_t with the density of a standard normal distribution.

Under some circumstances, one might compute a **one-sided p-value**. Here one computes either

$$p = P(\{S > |s|\})$$

or

$$p = P(\{S < -|s|\}).$$

Generally, it is more conservative to use a two-sided test, and one should do so unless there is a good reason not to. Very often, authors use one-sided tests because they result in smaller p-values, and small p-values are often a requirement for publication. This is not sensible behavior.

Procedure 7.3 (Computing a One-Sided p-Value for a T-Test) First, don't do this unless you have a good reason (getting a value less than 0.05 doesn't count). Now determine which side is important to you—which of $P(\{S > |s|\})$ or $P(\{S < -|s|\})$ do you care about, and why? If this process of thought hasn't dissuaded you, compute

$$p = P(\{S > |s|\})$$

or

$$p = P(\{S < -|s|\})$$

using the probability density of a t-distribution, as above. If $N > 30$, it is enough to replace p_t with the density of a standard normal distribution.

Once we have the p-value, evaluating significance is straightforward. A small p-value means that very few samples would display more extreme behavior than what we saw, *if* the null hypothesis is true. In turn, a small p-value means that, to believe our null hypothesis, we are forced to believe we have an extremely odd sample. More formally, the p-value that we compute is described as an assessment of the significance of the evidence *against* the null hypothesis. The p-value is smaller when the evidence against the null hypothesis is stronger. We get to decide how small a p-value means we should reject the null hypothesis.

Worked example 7.3 (If the Mean Length of an Adult Male Mouse is 10 cm, How Unusual is the Sample in the Mouse Dataset?) The mouse dataset is the one at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml. The variable to look at is Length2 (which appears to be the length of the mouse at the point it is sacrificed). We need to compute the p value.

Solution The mean length of male mice in this data is 9.5 cm, and the standard error is 0.045 cm. Write S for the sample mean length of some sample of male mice. This (unknown) value is a random variable. Assuming that the mean length really is 10, we have that

$$T = \frac{S - 10}{0.045}$$

and there are enough mice to assume that this is a normal random variable. The value we observe is $t = (9.5 - 10)/0.045 = -11.1$ We are asking for the probability that $T \leq -|t|$ OR $T \geq |t|$. This is so close to 0 that the difference is of no interest to us. In turn, the sample in the mouse dataset is quite implausibly unlikely if the mean length of an adult mouse were 10 cm. We can interpret this as overwhelming evidence that the mean length isn't 10 cm.

It is conventional to reject the null hypothesis when the p-value is less than 0.05. This is sometimes called "a significance level of 5%". The phrase can mislead: the term "significance" seems to imply the result is important, or meaningful. Instead, you should interpret a p-value of 0.05 as meaning that you would see evidence this unusual in about one experiment in twenty if the null hypothesis was true. It's quite usual to feel that this means the hypothesis is unlikely to be true.

Sometimes, the p-value is even smaller, and this can be interpreted as very strong evidence the null hypothesis is wrong. A p-value of less than 0.01 allows one to reject the null hypothesis at "a significance level of 1%". Similarly, you should interpret a p-value of 0.01 as meaning that you would see evidence this unusual in about one experiment in a hundred if the null hypothesis was true.

Worked example 7.4 (Average Human Body Weight) Assess the significance of the evidence against the hypothesis that the average human body weight is 175 lb, using the height and weight data set of http://www2.stetson.edu/~jrasp/data.htm (called bodyfat.xls).

Solution The dataset contains 252 samples, so we can use a normal model. The average weight is 178.9 lb. This results in a two-sided p-value of 0.02. We can interpret this as quite strong evidence that the average human body weight is not, in fact, 175 lb. This p-value says that, if (a) the average human body weight is 175 lb and (b) we repeat the experiment (weigh 252 people and average their weights) 50 times, we would see a value as far from 175 lb about once.

Worked example 7.5 (Cholesterol Levels After Heart Attacks) At http://www.statsci.org/data/general/cholest.html, you can find data on 28 patients whose cholesterol level was measured at various days after a heart attack. The data is attributed to "a study conducted by a major northeastern medical center" (here northeastern refers to a location in the USA). Assess the significance of the evidence that, on day 2, the mean cholesterol level is 240 mg/dL.

Solution N is small enough to use a t-distribution. We have the sample mean is 253.9; the standard error is 9.02; and so the test statistic is 1.54. A two-sided test, using 27 degrees of freedom, gives a p-value of 0.135, too large to comfortably reject the hypothesis.

7.2 Comparing the Mean of Two Populations

We have two samples, and we need to know whether these samples come from populations that have the same mean. For example, we might observe people using two different interfaces, measure how quickly they perform a task, then ask are their performances different? As another example, we might run an application with no other applications running, and test how long it takes to run some standard tasks. Because we don't know what the operating system, cache, etc. are up to, this number behaves a bit like a random variable, so it is worthwhile to do several experiments, yielding one set of samples. We now do this with other applications running as well, yielding another set of samples—is this set different from the first set? For realistic datasets, the answer is always yes, because they're random samples. A better question is: could the differences be the result of chance, or are these datasets really samples of two different populations?

Worked example 7.6 (Male and Female Chow Eating Mice) Give a centered 95% confidence interval for the weight of a female mouse who ate chow and for the weight of a male mouse who ate chow, based on the dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml. Compare these intervals.

Solution We know from Worked example 6.4 that the interval we want runs from 26.38 to 29.18 gr. For male mice, the same procedure yields the interval 34.75–38.06 gr. Now these two ranges are quite distinct. This means that, to believe the two populations are the same, we'd have to believe we have a really strange sample of at least one population. Here is rather compelling evidence that male and female chow-eating mice do not have the same mean weight.

As Worked example 7.6 shows, we can use confidence intervals on the means to reason about populations. There's an alternative—we could look at the significance of the difference between means. We need some notation: write $\{X\}$ for the first population and $\{Y\}$ for the second population. Write $\{x\}$ for the first dataset, which has size k_x, and $\{y\}$ for the second, which has size k_y. These datasets need not be of the same size.

7.2.1 Assuming Known Population Standard Deviations

In the simplest case, assume the two populations each have *known* standard deviation, i.e. $\mathsf{popsd}(\{X\})$ and $\mathsf{popsd}(\{Y\})$ are *known*. In this case, the distribution of sample means is normal. We can use some simple facts about normal random variables to come up with a measure of significance.

Useful Facts 7.1 (Sums and Differences of Normal Random Variables)
Let X_1 be a normal random variable with mean μ_1 and standard deviation σ_1. Let X_2 be a normal random variable with mean μ_2 and standard deviation σ_2. Let X_1 and X_2 be independent. Then we have that:

- for any constant $c_1 \neq 0$, $c_1 X_1$ is a normal random variable with mean $c_1\mu_1$ and standard deviation $c_1\sigma_1$;
- for any constant c_2, $X_1 + c_2$ is a normal random variable with mean $\mu_1 + c_2$ and standard deviation σ_1;
- $X_1 + X_2$ is a normal random variable with mean $\mu_1 + \mu_2$ and standard deviation $\sqrt{\sigma_1^2 + \sigma_2^2}$.

I will not prove these facts; we already know the expressions for means and standard deviations from our results on expectations. The only open question is to show that the distributions are normal. This is easy for the first two results. The third requires a bit of integration that isn't worth our trouble; you could reconstruct the proof from section Worked example 14.13's notes on sums of random variables and some work with tables of integrals.

Now write $X^{(k_x)}$ for the random variable obtained by: drawing a random sample with replacement of k_x elements from the first population, then averaging this sample. Write $Y^{(k_y)}$ for the random variable obtained by: drawing a random sample with replacement of k_y elements from the first population, then averaging this sample. Each random variable is normal, because the population standard deviations are known. This means that $X^{(k_x)} - Y^{(k_y)}$ is a normal random variable.

Now write D for $X^{(k_x)} - Y^{(k_y)}$. If the two populations have the same mean, then

$$\mathbb{E}[D] = 0.$$

Furthermore,

$$\begin{aligned}\mathsf{std}(D) &= \sqrt{\mathsf{std}\left(X^{(k_x)}\right)^2 + \mathsf{std}\left(X^{(k_y)}\right)^2} \\ &= \sqrt{\frac{\mathsf{popsd}(\{X\})^2}{k_x} + \frac{\mathsf{popsd}(\{Y\})^2}{k_y}},\end{aligned}$$

which we can evaluate because we assumed $\mathsf{popsd}(\{X\})$ and $\mathsf{popsd}(\{Y\})$ were known. We can now use the same reasoning we used to test the significance of the evidence that a population had a particular, known mean. We have identified a number we can compute from the two samples. We know how this number would vary under random choices of sample. If the value we observe is too many standard deviations away from the mean, the evidence is against our hypothesis. If we wanted to believe the hypothesis, we would be forced to believe that the sample is extremely strange. I have summarized this reasoning in box 7.4.

Procedure 7.4 (Testing Whether Two Populations Have the Same Mean, for Known Population Standard Deviations) The initial hypothesis is that the populations have the same, unknown, mean. Write $\{x\}$ for the sample of the first population, $\{y\}$ for the sample of the second population, and k_x, k_y for the sample sizes.

- Compute the sample means for each population, $\mathsf{mean}(\{x\})$ and $\mathsf{mean}(\{y\})$.
- Compute the standard error for the difference between the means,

$$s_{ed} = \sqrt{\frac{\mathsf{popsd}(\{X\})}{k_x} + \frac{\mathsf{popsd}(\{Y\})}{k_y}}.$$

- Compute the value of the test statistic using

$$s = \frac{(\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\}))}{s_{ed}}.$$

- Compute the p-value, using

$$p = (1 - f) = (1 - \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp\left(\frac{-u^2}{2}\right) du)$$

- The p-value summarizes the extent to which the data contradicts the hypothesis. A small p-value implies that, *if* the hypothesis is true, the sample is very unusual. The smaller the p-value, the more strongly the evidence contradicts the hypothesis.

It is common to think that a hypothesis can be rejected only if the p-value is less than 5% (or some number). You should not think this way; the p-value summarizes the extent to which the data contradicts the hypothesis, and your particular application logic affects how you interpret it.

7.2.2 Assuming Same, Unknown Population Standard Deviation

Now assume the two populations each have the *same, unknown* standard deviation, i.e. $\mathsf{popsd}(\{X\}) = \mathsf{popsd}(\{Y\}) = \sigma$, with σ unknown. If $\mathsf{popmean}(\{X\}) = \mathsf{popmean}(\{Y\})$, then we have that $\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\})$ is the value of a random variable whose mean is 0, and whose variance is

$$\frac{\sigma^2}{k_x} + \frac{\sigma^2}{k_y} = \sigma^2 \frac{k_x k_y}{k_x + k_y}$$

We don't know this variance, but must estimate it. Because the variance is the same in each population, we can pool the samples when estimating the variance. This yields the following estimate of the standard error:

$$s_{ed}^2 = \left(\frac{\mathsf{std}(\{x\})^2 (k_x - 1) + \mathsf{std}(\{y\})^2 (k_y - 1)}{k_x + k_y - 2} \right) \left(\frac{k_x k_y}{k_x + k_y} \right).$$

Using our previous reasoning, we have that

$$\frac{\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\})}{s_{ed}}$$

is the value of a random variable with a t-distribution with $k_x + k_y - 2$ degrees of freedom. I have summarized this reasoning in box 7.5.

Procedure 7.5 (Testing Whether Two Populations Have the Same Mean, for Same But Unknown Population Standard Deviations) The initial hypothesis is that the populations have the same, unknown, mean. Write $\{x\}$ for the sample of the first population, $\{y\}$ for the sample of the second population, and k_x, k_y for the sample sizes.

- Compute the sample means for each population, $\mathsf{mean}(\{x\})$ and $\mathsf{mean}(\{y\})$.
- Compute the standard error for the difference between the means,

$$s_{ed}^2 = \left(\frac{\mathsf{std}(\{x\})^2 (k_x - 1) + \mathsf{std}(\{y\})^2 (k_y - 1)}{k_x + k_y - 2} \right) \left(\frac{k_x k_y}{k_x + k_y} \right).$$

- Compute the test statistic using

$$s = \frac{(\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\}))}{s_{ed}}.$$

- Compute the p-value, using the recipe of Procedure 7.2; the number of degrees of freedom is $k_x + k_y - 2$.
- The p-value summarizes the extent to which the data contradicts the hypothesis. A small p-value implies that, *if* the hypothesis is true, the sample is very unusual. The smaller the p-value, the more strongly the evidence contradicts the hypothesis.

It is common to think that a hypothesis can be rejected only if the p-value is less than 5% (or some number). You should not think this way; the p-value summarizes the extent to which the data contradicts the hypothesis, and your particular application logic affects how you interpret it.

7.2.3 Assuming Different, Unknown Population Standard Deviation

Now assume the two populations each have the *different, unknown* standard deviations. If $\mathsf{popmean}(\{X\}) = \mathsf{popmean}(\{Y\})$, then we have that $\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\})$ is the value of a random variable whose mean is 0, and whose variance is

$$\frac{\mathsf{popsd}(\{X\})^2}{k_x} + \frac{\mathsf{popsd}(\{Y\})^2}{k_y}$$

We don't know this variance, but must estimate it. Because the two populations have different standard deviations, we can't pool the estimate. An estimate is

$$s_{ed}^2 = \frac{\mathsf{stdunbiased}(\{x\})^2}{k_x} + \frac{\mathsf{stdunbiased}(\{y\})^2}{k_y}.$$

We can form the test statistic in the natural way, yielding

$$\frac{\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\})}{s_{ed}}.$$

But there is an issue here. This statistic does not have a t-distribution, and the form of its distribution is complicated. It can be approximated satisfactorily with a t-distribution. Write

$$W = \left(\frac{\left[\mathsf{stdunbiased}(\{x\})^2/k_x\right]^2}{k_x - 1} + \frac{\left[\mathsf{stdunbiased}(\{y\})^2/k_y\right]^2}{k_y - 1} \right).$$

Then the approximating t-distribution has

$$\frac{\left[\left(\mathsf{stdunbiased}(\{x\})^2/k_x\right) + \left(\mathsf{stdunbiased}(\{y\})^2/k_y\right)\right]^2}{W}$$

degrees of freedom. With this, everything proceeds as before.

Procedure 7.6 (Testing Whether Two Populations Have the Same Mean, for Different Population Standard Deviations) The initial hypothesis is that the populations have the same, unknown, mean. Write $\{x\}$ for the sample of the first population, $\{y\}$ for the sample of the second population, and k_x, k_y for the sample sizes.

- Compute the sample means for each population, $\mathsf{mean}(\{x\})$ and $\mathsf{mean}(\{y\})$.
- Compute the standard error for the difference between the means,

$$s_{ed}^2 = \frac{\mathsf{stdunbiased}(\{x\})^2}{k_x} + \frac{\mathsf{stdunbiased}(\{y\})^2}{k_y}.$$

- Compute the test statistic using

$$s = \frac{(\mathsf{mean}(\{x\}) - \mathsf{mean}(\{y\}))}{s_{ed}}.$$

(continued)

- Compute the p-value, using the recipe of Procedure 7.2; the number of degrees of freedom is

$$\frac{\left(\text{stdunbiased}\left(\{x\}\right)^2/k_x + \text{stdunbiased}\left(\{y\}\right)^2/k_y\right)^2}{\left(\begin{array}{l}\left[\text{stdunbiased}\left(\{x\}\right)^2/k_x\right]^2/\left(k_x-1\right)+\\ \left[\text{stdunbiased}\left(\{y\}\right)^2/k_y\right]^2/\left(k_y-1\right)\end{array}\right)}.$$

- The p-value summarizes the extent to which the data contradicts the hypothesis. A small p-value implies that, *if* the hypothesis is true, the sample is very unusual. The smaller the p-value, the more strongly the evidence contradicts the hypothesis.

It is common to think that a hypothesis can be rejected only if the p-value is less than 5% (or some number). You should not think this way; the p-value summarizes the extent to which the data contradicts the hypothesis, and your particular application logic affects how you interpret it.

Worked example 7.7 (Are US and Japanese Cars Different) At http://www.itl.nist.gov/div898/handbook/eda/section3/eda3531.htm you can find a dataset, published by NIST, giving measurements of miles per gallon for Japanese and US cars. Assess the evidence these two populations have the same mean MPG.

Solution There are 249 measurements for Japanese cars, and 79 for US cars. The mean for Japanese cars is 20.14, and for US cars is 30.48. The standard error is 0.798. The value of the test statistic is

$$\frac{\left(\text{mean}\left(\{x\}\right) - \text{mean}\left(\{y\}\right)\right)}{s_{ed}} = 12.95$$

and the number of degrees of freedom is about 214. There are enough degrees of freedom here that the t-distribution could be approximated with a normal distribution, so a reasonable approximate p-value is the probability of encountering a standard normal random variable of this value or greater. This is so close to zero I had some trouble getting sensible numbers; the evidence very strongly rejects this hypothesis. A version of this example is worked in the NIST/SEMATECH e-Handbook of Statistical Methods, at http://www.itl.nist.gov/div898/handbook/, as of 2017.

7.3 Other Useful Tests of Significance

There are many forms of significance test. Significance testing can get quite elaborate, because determining distributions under sampling variation can get tricky. Furthemore, there are still arguments about precisely what significance means (or should mean). Mostly, we can ignore these difficulties. There are two more of the very many available procedures that will be very useful for us.

7.3.1 F-Tests and Standard Deviations

Imagine we have two datasets. There are N_x items in $\{x\}$, and N_y items in $\{y\}$. We believe that each dataset is normally distributed (i.e. that the values are IID samples from a normal distribution). We wish to evaluate the significance of the evidence against the belief that the two datasets have the same variance. This test will prove useful in the discussion of experiments (Chap. 8).

The procedure we adopt follows that of the T-test. We assume the dataset are samples drawn from a population. We compute a statistic representing the data that we actually see. This statistic will have a known distribution under sampling. We then compute the probability of observing a value of the statistic even more unusual than the value we observe, *if* the

two dataset have the same variance. If that probability is small, then if we want to believe the two datasets have the same variance, we will have to believe we have an extremely strange sample.

The statistic: In the cases I will deal with, it is clear which of the two populations has largest variance if they are different. I will assume that the X population might have larger variance. If these two populations had the same variance, we would expect that

$$F = \frac{\text{stdunbiased}\left(\{x\}\right)^2}{\text{stdunbiased}\left(\{y\}\right)^2}$$

should be close to one. This statistic is known as the **F-statistic**, and we are concerned that it could be greater than or equal to one, justifying the use of a one sided p-value.

The distribution: The F-statistic has a known distribution (called the **F-distribution**), assuming that $\{x\}$ and $\{y\}$ are IID samples from normal distributions. The form of the distribution isn't particularly important to us. Appropriate values can be looked up in tables, or obtained from a computing environment. However, it is important to keep track of one detail. As the number of samples in either dataset goes up, the estimate of the variance obtained from the samples must get more accurate. This means that the distribution depends on the number of degrees of freedom for *each* dataset (i.e. $N_x - 1$ and $N_y - 1$). We write $p_f(u; N_x - 1, N_y - 1)$ for the probability density of the F-statistic. We chose the ratio that was greater than, or equal to, one. Write r for the ratio we observed. The probability of observing a value of the statistic that is even more unusual (i.e. bigger) than the one we observe is

$$\int_r^\infty p_f(u; N_x - 1, N_y - 1)du.$$

I write this integral for completeness, rather than because you'd ever need to actually work with it. In practice, one either works with tables or with a function from a software package.

Procedure 7.7 (The F-Test of Significance for Equality of Variance) Given two datasets $\{x\}$ of N_x items and $\{y\}$ of N_y items, we wish to assess the significance of evidence against the hypothesis that the populations represented by these two datasets have the same variance. We assume that the alternative possibility is that the population represented by $\{x\}$ has the larger variance. We compute

$$F = \frac{\text{stdunbiased}\left(\{x\}\right)^2}{\text{stdunbiased}\left(\{y\}\right)^2}$$

and obtain a p-value using tables or software to recover

$$\int_r^\infty p_f(u; N_x - 1, N_y - 1)du.$$

where p_f is the probability distribution for the F-statistic.

Worked example 7.8 (Yet More Male and Female Chow Eating Mice) Does the data support the idea that the variance of the weight of female mice who ate chow and the variance of the weight of male mice who ate chow are the same? Use the dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml, and an F-test.

Solution The value of the F-statistic is 1.686 (male mice have the larger variance). The p-value for this F-statistic is 0.035, which we can interpret as evidence that only 3.5% of samples would have a more unusual value of the F-statistic *if* the two populations actually had the same variance. In turn, this means the evidence quite strongly contradicts the hypothesis they have the same variance. One note for careful readers: F-tests have a reputation for being somewhat unreliable when the data isn't normally distributed. Good sense suggests that I should check that mouse weights are normally distributed before releasing the result we have here to the newspapers. Fortunately, Worked example 7.10 below suggests that mouse weights are normally distributed, so we're safe on this point.

7.3.2 χ^2 Tests of Model Fit

Sometimes we have a model, and we would like to know whether the data is consistent with that model. For example, imagine we have a six-sided die. We throw it many times, and record which number comes up each time. We would like to know if the die is fair (i.e. is the data consistent with the model that the die is fair). It is highly unlikely that each face comes up the same number of times, even if the die is fair. Instead, there will be some variation in the frequencies observed; with what probability is that variation, or bigger, the result of chance effects?

As another example, we decide that the number of calls by a telemarketer in each hour is distributed with a Poisson distribution. We don't know the intensity. We could collect call data, and use maximum likelihood to determine the intensity. Once we have the best estimate of the intensity, we still want to know whether the model is consistent with the data.

In each case, the model predicts the frequencies of events. For the six-sided die case, the model tells us how often we expect to see each side. For the call case, the model predicts how often we would see no calls, one call, two calls, three calls, etc. in each hour. To tell whether the model fits the data, we need to compare the frequencies we observed with theoretical frequencies.

We adopt the following procedure, which should now be familiar. We assume the dataset are samples drawn from a population. We compute a statistic representing the data that we actually see. This statistic will have a known distribution under sampling. We then compute the probability of observing a value of the statistic even more unusual than the value we observe, *if* the model correctly predicts the frequencies of events. If that probability is small, then if we want to believe the model correctly predicts the frequencies of events, we will have to be believe we have an extremely small sample.

The statistic: The appropriate statistic is computed as follows. Assume we have a set of k disjoint events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ which cover the space of outcomes (i.e. any outcome lies in one of these events). Assume we perform N experiments, and record the number of times each event occurs. We have a hypothesis regarding the probability of events. We can take the probability of each event and multiply by N to get a frequency under that hypothesis. Now write $f_o(\mathcal{E}_i)$ for the observed frequency of event i; $f_t(\mathcal{E}_i)$ for the theoretical frequency of the event under the null hypothesis. We form the statistic

$$C = \sum_i \frac{(f_o(\mathcal{E}_i) - f_t(\mathcal{E}_i))^2}{f_t(\mathcal{E}_i)}$$

which compares the observed and actual frequency of events. This statistic is known as the **χ^2-statistic** (say "khi-squared").

The distribution: It turns out that this statistic C has a distribution very close to a known form, called the **χ^2-distribution**, as long as each count is five or more. The distribution has two parameters; the statistic, and the number of degrees of freedom. The degrees of freedom refers to the dimension of the space of measurement values that you could have. We will need to fix some values. The number of values to fix has to do with the type of test you are doing. In the most common case, you want to inspect the counts in each of k bins to tell whether they are consistent with some distribution. We know the sum of counts is N. It turns out that we should compare what we observe with what we could have observed with the same N. In this case, the dimension of the space of measurement value is $k - 1$, because you have k numbers but they must add up to N. Now assume we have to estimate p parameters for the model. For example, rather than asking whether the data comes from a standard normal distribution, we might use the data to estimate the mean. We then test whether the data comes from a normal distribution with the estimated mean, but unit standard deviation. As another example, we could estimate both mean and standard deviation from the data. If we estimate p parameters from the data, then the number of degrees of freedom becomes $k - p - 1$ (because there are k counts, they must lead to p parameter values, and they must add to 1).

After this, things follow the usual recipe. We compute the statistic; we then look at tables, or use our programming environment, to find the probability that the statistic takes this value or greater under the null hypothesis. If this is small, then we reject the null hypothesis.

Procedure 7.8 (The χ^2-Test of Significance of Fit to a Model) The model consists of k disjoint events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ which cover the space of outcomes and the probability $P(\mathcal{E})$ of each event. The model has p unknown parameters. Perform N experiments and record the number of times each event occurs in the experiments. The theoretical frequency of the i'th event for this experiment is $NP(\mathcal{E})$. Write $f_o(\mathcal{E}_i)$ for the observed frequency of event i; $f_t(\mathcal{E}_i)$ for the theoretical frequency of the event under the null hypothesis. We form the statistic

(continued)

$$C = \sum_i \frac{(f_o(\mathcal{E}_i) - f_t(\mathcal{E}_i))^2}{f_t(\mathcal{E}_i)}$$

which has $k - p - 1$ degrees of freedom, and compute a p-value by using tables or software to evaluate

$$\int_C^\infty p_{\chi^2}(u; k - p - 1)du$$

where $p_{\chi^2}(u; k-p-1)$ is the χ^2 probability density function with $k-p-1$ degrees of freedom. This test is safest when there are at least five instances of each event. This test is extremely elastic and can be used in a variety of non-obvious ways as sketched in the worked examples.

Worked example 7.9 (χ^2 Test for Dice) I throw a die 100 times. I record the outcomes, in the table below. Is this a fair die?

Face	1	2	3	4	5	6
Count	46	13	12	11	9	9

Solution The expected frequency is 100/6 for each face. The χ^2 statistic has the value 62.7, and there are 5 degrees of freedom. We get a p-value of about 3e-12. You would have to run this experiment 3e11 times to see a table as skewed as this once, by chance. It's quite unreasonable to believe the die is fair—or, at least, if you wanted to do so, you would have to believe you did a quite extraordinary unusual experiment.

Worked example 7.10 (Are Mouse Weights Normally Distributed?) Assess the evidence against the hypothesis that the weights of all mice who ate chow are normally distributed, based on the dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml.

Solution This example takes a little thought. The way to check whether a set of data is (roughly) normally distributed, is to break the values into intervals, and count how many data items fall into each interval. This gives the observed frequencies. You can then also compute theoretical frequencies for those intervals with a normal distribution (or simulate). Then you use a χ^2 test to tell whether the two are consistent. The choice of intervals matters. It is natural to have intervals that are some fraction of a standard deviation wide, with the mean at the center. You should have one running to infinity, and one to minus infinity. You'd like to have enough intervals so that you can tell any big difference from normal, but it's important to have at least five data items in each interval. There are 92 mice who make it to whenever Weight2 is evaluated (sacrifice, I think). The mean of Weight2 is 31.91 and the standard deviation is 6.72. I broke the data into 10 intervals at breakpoints $[-\infty, -1.2, -0.9, -0.6, -0.3, 0, 0.3, 0.6, 0.9, 1.2, \infty] * 6.72 + 31.91$. This gave me a count vector $[10, 9, 12, 9, 7, 11, 7, 9, 8, 10]$. I simulated 2000 draws from a normal distribution with the given mean and standard deviation and sorted them into these intervals, getting $[250, 129, 193, 191, 255, 240, 192, 192, 137, 221]$ (if you are less idle, you'll evaluate the integrals, but this shouldn't make much difference). I found a statistic with value 5.6338. Using 7 degrees of freedom (10 counts, but there are two parameters estimated), I found a p-value of 0.5830979, meaning there is no reason to reject the idea that the weights are normally distributed.

Worked example 7.11 (Is Swearing Poisson?) A famously sweary politician gives a talk. You listen to the talk, and for each of 30 intervals 1 min long, you record the number of swearwords. You record this as a histogram (i.e. you count the number of intervals with zero swear words, with one, etc.), obtaining the table below.

No. of swear words	0	1	2	3	4
No. of intervals	13	9	8	5	5

The null hypothesis is that the politician's swearing is Poisson distributed, with intensity (λ) one. Can you reject this null hypothesis?

Solution If the null hypothesis is true, then the probability of getting n swear words in a fixed length interval would be $\frac{\lambda^n e^{-\lambda}}{n!}$. There are 10 intervals, so the theoretical frequencies are 10 times the following probabilities

No. of swear words	0	1	2	3	4
No. of intervals	0.368	0.368	0.184	0.061	0.015

so the χ^2 statistic takes the value 243.1 and there are 4 degrees of freedom. The significance is indistinguishable from zero by my programming environment, so you can firmly reject the null hypothesis. Of course, it may just be that the intensity is wrong (exercises).

Worked example 7.12 (Are Goals Independent of Gender?) Assess the evidence that student goals are independent of student gender in the dataset of Chase and Dunner, which you can find at http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html.

Solution This is an example of a common use of the χ^2 test. The table below shows the count of students in the study by gender and goal. I have inserted row and column totals for convenience. In total, there were 478 students.

	Boy	Girl	Total
Grades	117	130	247
Popular	50	91	141
Sports	60	30	90
Total	227	251	478

We will test whether the counts observed are different from those predicted if gender and goal are independent. If they are independent, then $P(\text{boy}) = 227/478 = 0.47$, and $P(\text{Grades}) = 247/478 = 0.52$ (and so on). This means that we can produce a table of theoretical counts under the model (below).

	Boy	Girl
Grades	117.29916	129.70084
Popular	66.96025	74.03975
Sports	42.74059	47.25941

There are 6 cells in our table. One degree of freedom is used up by requiring that there are 478 students. Two further degrees of freedom are used up by insisting that the Grades/Popular/Sports counts yield the distribution we observed. One further degree of freedom is used up by insisting that the gender counts yield the distribution we observe. This means that there are a total of two degrees of freedom. We compute a χ^2 value of 21.46. The p-value is 2e-5. In turn, if the two factors were independent, you'd see counts like these by chance about twice in a hundred thousand experiments. It's very hard to believe they are independent.

7.4 P-Value Hacking and Other Dangerous Behavior

Significance is a very useful way of telling whether an experimental observation might be the result of chance effects, but it is important to follow the logic of the method carefully. If you don't, you can fairly easily come to false conclusions.

There is an important danger here.

Removing data points and recomputing p-values is one way to have a serious problem. One context where this occurs in evaluating medical procedures. We would test the hypothesis that some sample mean of a treated population is the same as that of an untreated population. If this hypothesis fails, then the procedure did something. However, we might see outliers in the dataset. If we remove these outliers, then (of course) the p-value changes. This presents an temptation to remove outliers that shift the p-value in some desired direction. Of course, doing this consciously is fraud; but it's quite easy to simply fool yourself into removing data points whose absence is helpful.

Another way to fool yourself is to look at a lot of samples, take the one with the smallest p-value, then declare the evidence is against the hypothesis. The more samples you look at, the better your chance of seeing one which has a small p-value (that's what the p-value means). If you look at lots of samples or do lots of experiments, looking for one with a small p-value, then use that to argue the hypothesis is false, you are fooling yourself. Because fooling other people can be quite profitable, this practice is common enough to have a name: it's referred to as **p-value hacking**.

t's pretty clear that searching samples for one with a low p-value is bad behavior, but a subtle version of this mistake is to intermingle computing p-values with collecting data, then stop when you get the p-value you want. This is quite common behavior. A good example of how badly things can go wrong when you do this is described in the paper "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", by J.P Simmons, L.D. Nelson and U. Simonsohn (Psychological Science, 2011). The authors were able to use this strategy to collect data that showed that listening to a particular song would cause your chronological age to go down. I advise you to look at this paper, which is a good, easy, and highly informative read

Yet another way to fool yourself is to change hypotheses halfway through an experiment. Imagine we want to test the effect of a medical procedure. We decide to look at one particular sample mean, but halfway through collecting data it looks as though there won't be anything interesting to say about that mean. It's tempting to change measurements and focus on another statistic instead. If you do, the logic underlying the procedures we describe here fails. Essentially, you will bias the test to reject a hypothesis to a degree we can't go into.

Yet another way to fool yourself is to test multiple hypotheses at the same time, and reject the one with the smallest p-value. The problem here is that the test isn't taking into account that you're testing multiple hypotheses. If you repeatedly test different hypotheses using the same dataset, the chances go up that the data you observed are inconsistent with a hypothesis that is true, purely as a result of sampling. Special procedures are required for this case.

One solution is for these problems is really strict protocols, where you describe everything you will do and everything you will test before doing the experiment and then don't vary from that plan. But such protocols can be expensive and clumsy in practice.

You should not get confused about what a test means. The tests I have described are referred to as tests of statistical significance. I use this terminology because that's what everyone else uses, but personally I don't find it helpful, because there is an insidious suggestion that a statistically significant difference actually matters—i.e. is significant. What significance testing procedures tell you is what fraction of random samples of data have the mean that you observe, if your hypothesis is true, *and* if you have collected the data correctly, tested correctly, and so on. The procedures don't tell you that you've done important, or even interesting, science.

7.5 You Should

7.5.1 Remember These Definitions

7.5.2 Remember These Terms

7.5.3 Remember These Facts

7.5.4 Use These Procedures

7.5.5 Be Able to

- Compute the fraction of samples that would have a more extreme mean than some value, *if* the population mean had a given value.
- Evaluate the significance of the evidence against the hypothesis that a population mean has a given value using a normal distribution model (i.e. for a large sample).
- Evaluate the significance of the evidence against the hypothesis that a population mean has a given value using a t-distribution model (i.e. for a sample that isn't large).
- Evaluate the significance of the evidence against the hypothesis that two population means are the same using a t-distribution model (i.e. for a sample that isn't large).
- Evaluate the significance of the evidence against the hypothesis that two population standard deviations are the same using an F-test.
- Evaluate the significance of the evidence against a model using a χ^2 test.
- Avoid thinking that the significance of evidence is the same as the significance of a piece of science.
- Avoid "adjusting" your experimental data to improve the p-value, and avoid p-value hacking in general.

Problems

Fractions of Samples

7.1 In 1998, the average height of an adult male in South Africa was estimated to be 169 cm. Assume that this estimate is exact; assume also that the population standard deviation is 10 cm. What fraction of samples consisting of 50 adult males from South Africa (selected uniformly at random, and with replacement) will have average height greater than 200 cm?

7.2 Assume the average weight of an adult male short-hair house cat is 5 kg, and the standard deviation is 0.7 kg (these numbers are reasonable, but there's quite a lively fight between cat fanciers about the true numbers).

(a) What fraction of samples consisting of 30 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4 kg?
(b) What fraction of samples consisting of 300 adult male short-hair house cats (selected uniformly at random, and with replacement) will have average weight less than 4 kg?
(c) Why are these numbers different?

Significance

7.3 Yet more Mouse-weighing I claim the average weight of a mouse is 25 grams. You decide to evaluate the evidence in support of this claim. You obtain 10 mice, sampled uniformly at random and with replacement from the mouse population. Their weights are $21, 23, 27, 19, 17, 18, 20, 15, 17, 22$ grams respectively. Does the evidence support my claim? to what extent? Why?

7.4 How big are Parktown Prawns? The Parktown prawn is an impressively repellent large insect, common in Johannesburg (look them up on the Web). I claim that their average length is 10 cm. You collect 100 Parktown prawns (this will take about 10 mins, in the right places in Johannesburg; more difficult from the US). The mean length of these prawns is 7 cm. The standard deviation is 1 cm. Assess the evidence against my claim.

7.5 Two Populations of Rats Zucker rats are specially bred to have curious weight properties, related to their genetics (look them up on the Web). You measure 30 lean Zucker rats, obtaining an average weight of 500 grams with a standard deviation of 50 grams. You measure 20 fatty Zucker rats, obtaining an average weight of 1000 grams with a standard deviation of 100 grams. Assess the evidence against the claim that these populations have the same weight.

7.6 Male and Female pet Rats You measure 35 female pet rats, obtaining an average weight of 300 grams with a standard deviation of 30 grams. You measure 30 male pet rats, obtaining an average weight of 400 grams with a standard deviation of 100 grams. Assess the evidence against the claim that these populations have the same weight.

7.7 Lean and Fatty Zucker Rats Zucker rats are specially bred to have curious weight properties, related to their genetics (look them up on the Web). You measure 30 lean Zucker rats, obtaining an average weight of 500 grams with a standard deviation of 50 grams. You measure 35 fatty Zucker rats, obtaining an average weight of 1000 grams with a standard deviation of 100 grams. In steps, you will assess the evidence against the claim that a fatty Zucker rat has exactly twice the weight of a lean Zucker rat. You know that the product of a normal random variable and a constant is a normal random variable. You should assume (and accept, because I won't prove it) that the sum of two normal random variables is a normal random variable.

(a) Write $L^{(k)}$ for the random variable obtained by drawing a uniform sample of k lean rats and averaging their weights. You can assume that k is large enough that this is normal.

- What is $\mathbb{E}\left[L^{(k)}\right]$? (write an expression, no need to prove anything)
- What is $\mathsf{std}\left(L^{(k)}\right)$? (write an expression, no need to prove anything)

(b) Now write $F^{(s)}$ for the random variable obtained by drawing a uniform sample of s fatty rats and averaging their weights. You can assume that s is large enough that this is normal.

- What is $\mathbb{E}\left[F^{(s)}\right]$? (write an expression, no need to prove anything)
- What is $\mathsf{std}\left(F^{(s)}\right)$? (write an expression, no need to prove anything)

(c) Write $\mathsf{popmean}\left(\{L\}\right)$ for the population mean weight of lean rats, and $\mathsf{popmean}\left(\{F\}\right)$ for the population mean weight of fatty rats. Assume that $2\mathsf{popmean}\left(\{L\}\right) = \mathsf{popmean}\left(\{F\}\right)$.

- In this case, what is $\mathbb{E}\left[F^{(s)} - 2L^{(k)}\right]$?
- In this case, what is $\mathsf{std}\left(F^{(s)} - 2L^{(k)}\right)$?
- Your expression for $\mathsf{std}\left(F^{(s)} - 2L^{(k)}\right)$ will have contained terms in the population standard deviation of F and L. What is the standard error of $F^{(s)} - 2L^{(k)}$?

(d) Now assess the evidence against the hypothesis that a fatty Zucker rat weighs exactly twice as much as a lean Zucker rat.

7.8 Are boys and girls equiprobable? In Carcelle-le-Grignon at the end of the eighteenth century, there were 2009 births. There were 983 boys and 1026 girls. You can regard this as a fair random sample (with replacement, though try not to think too hard about what that means) of births. Assess the evidence against the hypothesis that a boy is born with probability exactly 0.5.

Chi-Squared Tests

7.9 You can find a dataset of the passenger list for the Titanic disaster at http://www.statsci.org/data/general/titanic.html.

(a) Assess the evidence that survival is independent of passenger ticket class.
(b) Assess the evidence that survival is independent of passenger gender.

7.10 You can find a dataset giving income data for US citizens at the UC Irvine Machine Learning data archive, at http://archive.ics.uci.edu/ml/datasets/Adult. Each item consists of a set of numeric and categorical features describing a person, together with whether their annual income is larger than or smaller than 50 K$.

(a) Assess the evidence that income category is independent of gender.
(b) Assess the evidence that income category is independent of education level.

7.11 Assess the evidence that the swearing behavior of the politician of Worked example 7.11 follows a Poisson distribution. *Hint:* Once you've estimated the intensity, the rest is like that example; be careful about the number of degrees of freedom.