Chapter 2

Confidence intervals and hypothesis tests

This chapter focuses on how to draw conclusions about populations from sample data. We'll start by looking at binary data (e.g., polling), and learn how to estimate the true ratio of 1s and 0s with confidence intervals, and then test whether that ratio is significantly different from some baseline value using hypothesis testing. Then, we'll extend what we've learned to continuous measurements.

■ 2.1 Binomial data

Suppose we're conducting a yes/no survey of a few randomly sampled people[1], and we want to use the results of our survey to determine the answers for the overall population.

■ 2.1.1 The estimator

The obvious first choice is just the fraction of people who said yes. Formally, suppose we have n samples $x_1, \dots, x_n$ that can each be 0 or 1, and the probability that each x_i is 1 is p (in frequentist style, we'll assume p is fixed but unknown: this is what we're interested in finding). We'll assume our samples are *indendepent and identically distributed (i.i.d.)*, meaning that each one has no dependence on any of the others, and they all have the same probability p of being 1. Then our estimate for p, which we'll call $\hat{p}$, or "p-hat" would be

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Notice that $\hat{p}$ is a *random* quantity, since it depends on the random quantities x_i. In statistical lingo, $\hat{p}$ is known as an **estimator** for p. Also notice that except for the factor of $1/n$ in front, $\hat{p}$ is almost a binomial random variable (that is, $(n\hat{p}) \sim B(n,p)$). We can compute its expectation and variance using the properties we reviewed:

$$\mathbb{E}[\hat{p}] = \frac{1}{n}np = p, \tag{2.1}$$

$$\mathrm{var}[\hat{p}] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}. \tag{2.2}$$

[1] We'll talk about how to choose and sample those people in Chapter 7.

Since the expectation of $\hat{p}$ is equal to the true value of what $\hat{p}$ is trying to estimate (namely p), we say that $\hat{p}$ is an **unbiased** estimator for p. Reassuringly, we can see that another good property of $\hat{p}$ is that its variance decreases as the number of samples increases.

■ 2.1.2 Central Limit Theorem

The *Central Limit Theorem*, one of the most fundamental results in probability theory, roughly tells us that if we add up a bunch of independent random variables that all have the same distribution, the result will be approximately Gaussian.

We can apply this to our case of a binomial random variable, which is really just the sum of a bunch of independent Bernoulli random variables. As a rough rule of thumb, if p is close to 0.5, the binomial distribution will look almost Gaussian with $n = 10$. If p is closer to 0.1 or 0.9 we'll need a value closer to $n = 50$, and if p is much closer to 1 or 0 than that, a Gaussian approximation might not work very well until we have much more data.

This is useful for a number of reasons. One is that Gaussian variables are completely specified by their mean and variance: that is, if we know those two things, we can figure out everything else about the distribution (probabilities, etc.). So, if we know a particular random variable is Gaussian (or approximately Gaussian), all we have to do is compute its mean and variance to know everything about it.

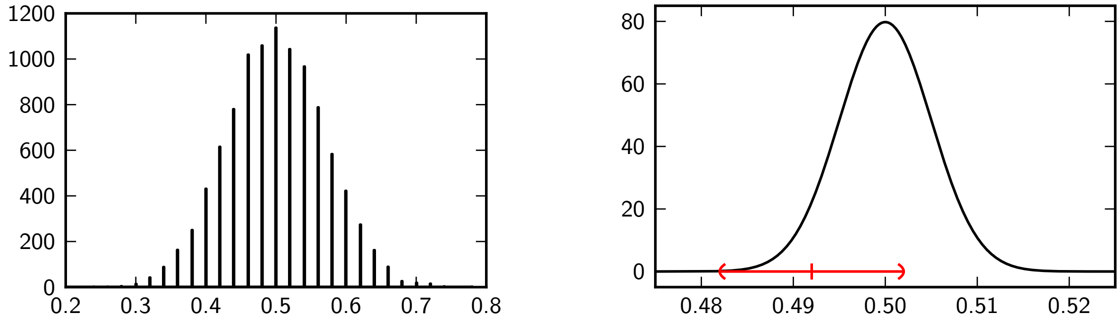
■ 2.1.3 Sampling Distributions

Going back to binomial variables, let's think about the distribution of $\hat{p}$ (remember that this is a random quantity since it depends on our observations, which are random). Figure 2.1a shows the **sampling distribution** of $\hat{p}$ for a case where we flip a coin that we hypothesize is fair (i.e. the true value p is 0.5). There are typically two ways we use such sampling distributions: to obtain **confidence intervals** and to perform **significance tests**.

■ 2.1.4 Confidence intervals

Suppose we observe a value $\hat{p}$ from our data, and want to express how certain we are that $\hat{p}$ is close to the true parameter p. We can think about how often the *random* quantity $\hat{p}$ will end up within some distance of the *fixed but unknown* p. In particular, we can ask for an interval around $\hat{p}$ for any sample so that *in 95% of samples, the true mean p will lie inside this interval.* Such an interval is called a **confidence interval**. Notice that we chose the number 95% arbitrarily: while this is a commonly used value, the methods we'll discuss can be used for any confidence level.

We've established that the random quantity $\hat{p}$ is approximately Gaussian with mean p and variance $p(1-p)/n$. We also know from last time that the probability of a Gaussian random variable being within about 2 standard deviations of its mean is about 95%. This means that there's a 95% chance of $\hat{p}$ being less than $2\sqrt{p(1-p)/n}$ away from p. So, we'll define



(a) The sampling distribution of the estimator $\hat{p}$: i.e. the distribution of values for $\hat{p}$ given a fixed true value $p = 0.5$.

(b) The 95% confidence interval for a particular observed $\hat{p}$ of 0.49 (with a true value of $p = 0.5$). Note that in this case, the interval contains the true value p. Whenever we draw a set of samples, there's a 95% chance that the interval that we get is good enough to contain the true value p.

Figure 2.1

the interval

$$\hat{p} \pm \underbrace{2}_{\text{coeff.}} \cdot \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{std. dev.}}. \tag{2.3}$$

With probability 95%, we'll get a $\hat{p}$ that gives us an interval containing p.

What if we wanted a 99% confidence interval? Since $\hat{p}$ is approximately Gaussian, its probability of being within 3 standard deviations from its mean is about 99%. So, the 99% confidence interval for this problem would be

$$\hat{p} \pm \underbrace{3}_{\text{coeff.}} \cdot \underbrace{\sqrt{\frac{p(1-p)}{n}}}_{\text{std. dev.}}. \tag{2.4}$$

We can define similar confidence intervals, where the standard deviation remains the same, but the coefficient depends on the desired confidence. While our variables being Gaussian makes this relationship easy for 95% and 99%, in general we'll have to look up or have our software compute these coefficients.

But, there's a problem with these formulas: they requires us to know p in order to compute confidence intervals! Since we don't actually know p (if we did, we wouldn't need a confidence interval), we'll approximate it with $\hat{p}$, so that (2.3) becomes

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{2.5}$$

This approximation is reasonable if $\hat{p}$ is close to p, which we expect to normally be the case. If the approximation is not as good, there are several more robust (but more complex) ways to compute the confidence interval.

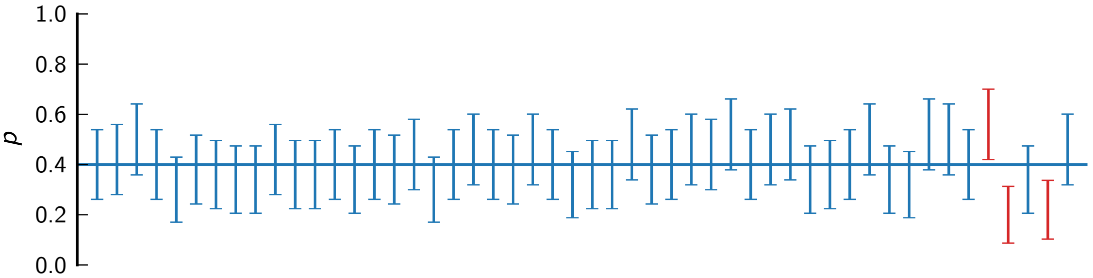


Figure 2.2: Multiple 95% confidence intervals computed from different sets of data, each with the same true parameter $p = 0.4$ (shown by the horizontal line). Each confidence interval represents what we might have gotten if we had collected new data and then computed a confidence interval from that new data. Across different datasets, about 95% of them contain the true interval. But, once we have a confidence interval, we can't draw any conclusions about where in the interval the true value is.

Interpretation

It's important not to misinterpret what a confidence interval is! This interval tells us nothing about the *distribution* of the true parameter p. In fact, p is a fixed (i.e., deterministic) unknown number! Imagine that we sampled n values for x_i and computed $\hat{p}$ along with a 95% confidence interval. Now imagine that we repeated this whole process a huge number of times (including sampling new values for x_i). Then about 5% of the confidence intervals constructed won't actually contain the true p. Furthermore, if p is in a confidence interval, we don't know where exactly within the interval p is.

Furthermore, adding an extra 4% to get from a 95% confidence interval to a 99% confidence interval doesn't mean that there's a 4% chance that it's in the extra little area that you added! The next example illustrates this.

In summary, a 95% confidence interval gives us a region where, had we redone the survey from scratch, then 95% of the time, the true value p will be contained in the interval. This is illustrated in Figure 2.2.

■ 2.1.5 Hypothesis testing

Suppose we have a hypothesized or baseline value p and obtain from our data a value $\hat{p}$ that's smaller than p. If we're interested in reasoning about whether $\hat{p}$ is "significantly" smaller than p, one way to quantify this would be to assume the true value were p and then compute the probability of getting a value smaller than or as small as the one we observed (we can do the same thing for the case where $\hat{p}$ is larger). If this probability is "very low", we might think the hypothesized value p is incorrect. This is the **hypothesis testing framework**.

We begin with a **null hypothesis**, which we call H_0 (in this example, this is the hypothesis that the true proportion is in fact p) and an **alternative hypothesis**, which we call H_1 or H_a (in this example, the hypothesis that the true mean is significantly smaller than p).

Usually (but not always), the null hypothesis corresponds to a baseline or boring finding, and the alternative hypothesis corresponds to some interesting finding. Once we have the two hypotheses, we'll use the data to test which hypothesis we should believe. "Significance" is usually defined in terms of a probability threshold α, such that we deem a particular result significant if the probability of obtaining that result under the null distribution is less than α. A common value for α is 0.05, corresponding to a 1/20 chance of error. Once we obtain a particular value and evaluate its probability under the null hypothesis, this probability is known as a ***p*-value**.

This framework is typically used when we want to disprove the null hypothesis and show the value we obtained is significantly different from the null value. In the case of polling, this may correspond to showing that a candidate has significantly more than 50% support. In the case of a drug trial, it may correspond to showing that the recovery rate for patients given a particular drug is significantly more than some baseline rate.

Here are some definitions:

- In a **one-tailed hypothesis test**, we choose one direction for our alternative hypothesis: we either hypothesize that the test statistic is "significantly big", or that the test statistic is "significantly small".
- In a **two-tailed hypothesis test**, our alternative hypothesis encompasses both directions: we hypothesize that the test statistic is simply *different* from the predicted value.
- A **false positive** or Type I error happens when the null hypothesis is true, but we reject it. Note that the probability of a Type I error is α.
- A **false negative** or Type II error happens when the null hypothesis is false, but we fail to reject it[2]
- The statistical **power** of a test is the probability of rejecting the null hypothesis when it's false (or equivalently, $1 -$ (probability of type II error).

  Power is usually computed based on a particular assumed value for the quantity being tested: "if the value is actually __, then the power of this test is __." It also depends on the threshold determined by α.

  It's often useful when deciding how many samples to acquire in an experiment, as we'll see later.

[2] Notice our careful choice of words here: if our result isn't significant, we can't say that we accept the null hypothesis. The hypothesis testing framework only lets us say that we fail to reject it.

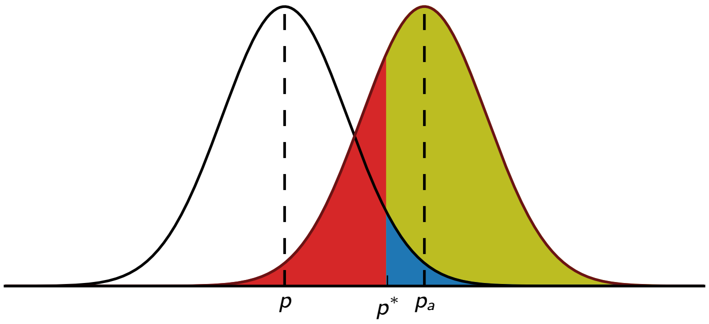


Figure 2.3: An illustration of statistical power in a one-sided hypothesis test on variable p.

Example

The concepts above are illustrated in Figure 2.3. Here, the null hypothesis H_0 is that $p = p_0$, and the alternative hypothesis H_a is that $p > p_0$: this is a one-sided test. In particular, we'll use the value p_a as the alternative value so that we can compute power. The null distribution is shown on the left, and an alternative distribution is shown on the right. The $\alpha = 0.05$ threshold for the alternative hypothesis is shown as p^*.

- When the null hypothesis is true, $\hat{p}$ is generated from the null (left) distribution, and we make the correct decision if $\hat{p} < p^*$ and make a Type I error (false positive) otherwise.
- When the alternative hypothesis is true, and if the true proportion p is actually p_a, $\hat{p}$ is generated from the right distribution, and we make the correct decision when $\hat{p} > p^*$ and make a Type II error (false negative) otherwise.

The power is the probability of making the correct decision when the alternative hypothesis is true. The probability of a Type I error (false positive) is shown in blue, the probability of a Type II error (false negative) is shown in red, and the power is shown in yellow and blue combined (it's the area under the right curve minus the red part).

Notice that a threshold usually balances between Type I and Type II errors: if we always reject the null hypothesis, then the probability of a Type I error is 1, and the probability of a Type II error is 0, and vice versa if we always fail to reject the null hypothesis.

Example: Drug therapy results: a warning about data collection

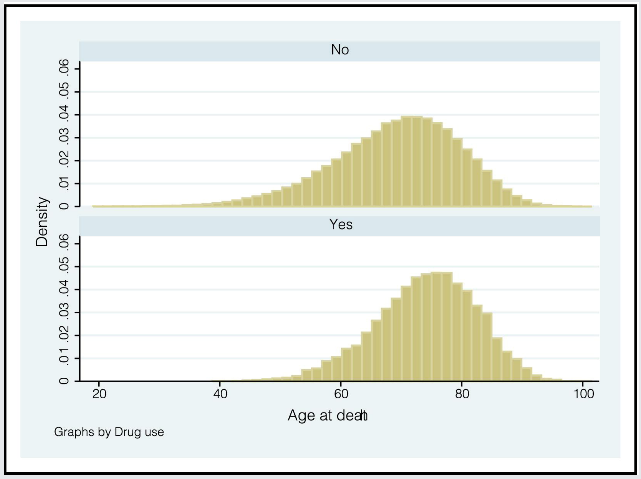


Figure 2.4: Results of a simulated drug trial measuring the effects of statin drugs on lifespan. The top figure shows the lifespan of subjects who *did not* receive treatment, and the bottom figure shows the lifespan of subjects who *did* receive it.

Figure 2.4 shows results from a simulated drug trial[a]. At first glance, it seems clear that people who received the drug (bottom) tended to have a higher lifespan than people who didn't (top), but it's important to look at hidden confounds! In this simulation, the drug actually had no effect, but the disease occurred more often in older people: these older people had a higher average lifespan simply because they had to live longer to get the drug.

Any statistical test we perform will say that the second distribution has a higher mean than the first one, but this is not because of the treatment, but instead because of how we sampled the data!

[a] Figure from: Støvring, et al. Statin Use and Age at Death: Evidence of a Flawed Analysis. The American Journal of Cardiology, 2007

2.2 Continuous random variables

So far we've only talked about binomial random variables, but what about continuous random variables? Let's focus on estimating the mean of a random variable given observations of it. As you can probably guess, our estimator will be $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

We'll start with the case where we know the true population standard deviation; call it σ. This is somewhat unrealistic, but it'll help us set up the more general case.

2.2.1 When σ is known

Consider random i.i.d. Gaussian samples $x_1, \ldots, x_n$, all with mean μ and variance σ^2. We'll compute the sample mean $\hat{\mu}$, and use it to draw conclusions about the true mean μ.

Just like $\widehat{p}$, $\hat{\mu}$ is a random quantity. Its expectation, which we computed in Chapter 1, is μ. Its variance is

$$\begin{aligned} \operatorname{var}[\hat{\mu}] &= \operatorname{var}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var}[x_i] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}. \end{aligned} \tag{2.6}$$

This quantity (or to be exact, the square root of this quantity) is known as the **standard error of the mean**. In general, the standard deviation of the sampling distribution of the a particular statistic is called the **standard error** of that statistic.

Since $\hat{\mu}$ is the sum of many independent random variables, it's approximately Gaussian. If we subtract its mean μ and divide by its standard deviation $\sigma/\sqrt{n}$ (both of which are deterministic), we'll get a standard normal random variable. This will be our test statistic:

$$z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}. \tag{2.7}$$

Hypothesis testing

In the case of hypothesis testing, we know μ (it's the mean of the null distribution), and we can compute the probability of getting z or something more extreme. Your software of choice will typically do this by using the fact that z has a standard normal distribution and report the probability to you. This is known as a ***z*-test**.

Confidence intervals

What about a confidence interval? Since z is a standard normal random variable, it has probability 0.95 of being within 2 standard deviations of its mean. We can compute the confidence interval by manipulating a bit of algebra:

$$\begin{aligned} P(-2 \le z \le 2) &\approx 0.95 \\ P(-2 \le \frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}} \le 2) &\approx 0.95 \\ P(-2\frac{\sigma}{\sqrt{n}} \le \hat{\mu} - \mu \le 2\frac{\sigma}{\sqrt{n}}) &\approx 0.95 \\ P(\hat{\mu} - \underbrace{2}_{\text{coeff.}} \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{std. dev.}} \le \mu \le \hat{\mu} + \underbrace{2}_{\text{coeff.}} \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{std. dev.}}) &\approx 0.95 \end{aligned}$$

This says that the probability that μ is within the interval $\hat{\mu} \pm 2\frac{\sigma}{\sqrt{n}}$ is 0.95. But remember: the only thing that's random in this story is $\hat{\mu}$! So when we use the word "probability" here, it's referring only to the randomness in $\hat{\mu}$. Don't forget that μ isn't random!

Also, remember that we chose the confidence level 0.95 (and therefore the threshold 2) somewhat arbitrarily, and we could just as easily compute a 99% confidence interval (which would correspond to a threshold of about 3) or an interval for any other level of confidence: we could compute the threshold by using the standard normal distribution.

Finally, note that for a two-tailed hypothesis test, the threshold at which we declare significance for some particular α is the same as the width of a confidence interval with confidence level $1 - \alpha$. Can you show why this is true?

Statistical power

If we get to choose the number of observations n, how do we pick it to ensure a certain level of statistical power in a hypothesis test? Suppose we choose α and a corresponding threshold x^*. How can we choose n, the number of samples, to achieve a desired statistical power? Since the width of the sampling distribution is controlled by n, by choosing n large enough, we can achieve enough power for particular values of the alternative mean.

The following example illustrates the effect that sample size has on significance thresholds.

Example: Fertility clinics

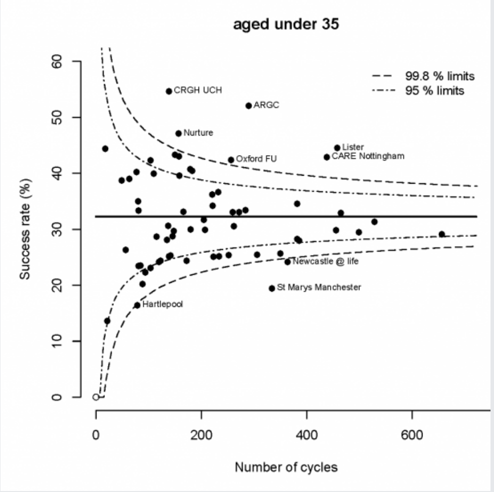


Figure 2.5: A funnel plot showing conception statistics from fertility clinics in the UK. The x-axis indicates the sample size; in this case that's the number of conception attempts (cycles). The y-axis indicates the quantity of interest; in this case that's the success rate for conceiving. The funnels (dashed lines) indicate thresholds for being significantly different from the null value of 32% (the national average). This figure comes from http://understandinguncertainty.org/fertility.

Figure 2.5 is an example of a *funnel plot*. We see that with a small number of samples, it's difficult to judge any of the clinics as significantly different from the baseline value, since exceptionally high/low values could just be due to chance. However, as the number of cycles increases, the probability of consistently obtaining large values by chance decreases, and we can declare clinics like Lister and CARE Nottingham significantly better than average: while other clinics have similar success rates over fewer cycles, these two have a high success rate over many cycles. So, we can be more certain that the higher success rates are not just due to chance and are in fact meaningful.

■ 2.2.2 When σ is unknown

In general, we won't know the true population standard deviation beforehand. We'll solve this problem by using the sample standard deviation. This means using $\hat{\sigma}^2/n$ instead of σ^2/n for $\text{var}(\hat{\mu})$. Throughout these notes, we'll refer to this quantity as the standard error of the mean (as opposed to the version given in Equation (2.6)).

But once we replace the fixed σ with the random $\hat{\sigma}$ (which we'll also write as s), our test statistic (Equation (2.7)) becomes

$$t = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}. \tag{2.8}$$

Since the numerator and denominator are both random, this is no longer Gaussian. The denominator is roughly χ^2-distributed quantity[3], and the overall statistic is t-distributed. In this case, our t distribution has $n - 1$ degrees of freedom.

Confidence intervals and hypothesis tests proceed just as in the known-σ case with only two changes: using $\hat{\sigma}$ instead of σ and using a t distribution with $n-1$ degrees of freedom instead of a Gaussian distribution. The confidence interval requires only $\hat{\mu}$ and the standard error s, while the hypothesis test also requires a hypothesis, in the form of a value for μ.

For example, a 95% confidence interval might look like

$$\hat{\mu} \pm t^* \frac{\hat{\sigma}}{\sqrt{n}} \tag{2.9}$$

To determine the coefficient t^*, we need to know the value where a t distribution has 95% of its probability. This depends on the degrees of freedom (the only parameter of the t distribution) and can easily be looked up in a table or computed from any software package. For example, if $n = 10$, then the t distribution has $n - 1 = 9$ degrees of freedom, and $k = 2.26$. Notice that this produces a wider interval than the corresponding Gaussian-based confidence interval from before. If we don't know the standard deviation and we estimate it, we're then less certain about our estimate $\hat{\mu}$.

To derive the t-test, we assumed that our data points were normally distributed. But, the t-test is fairly robust to violations of this assumption.

■ 2.3 Two-sample tests

So far, we've looked at the case of having one sample and determining whether it's significantly greater than some hypothesized amount. But what about the case where we're interested in the difference between two samples? We're usually interested in testing whether the difference is significantly different from zero. There are a few different ways of dealing with this, depending on the underlying data.

[3] In fact, the quantity $(n-1)\hat{\sigma}^2/\sigma^2$ is χ^2-distributed with $n-1$ degrees of freedom, and the test statistic $t = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma\sqrt{n-1}}{\hat{\sigma}\sqrt{n-1}}$ is therefore t-distributed.

- In the case of **matched pairs**, we have a "before" value and an "after" value for each data point (for example, the scores of students before and after a class). Matching the pairs helps control the variance due to other factors, so we can simply look at the differences for each data point, $x_i^{\text{post}} - x_i^{\text{pre}}$ and perform a one-sample test against a null mean of 0.

- In the case of two samples with **pooled variance**, the means of the two samples might be different (this is usually the hypothesis we test), but the variances of each sample are assumed to be the same. This assumption allows us to combine, or pool, all the data points when estimating the sample variance. So, when computing the standard error, we'll use this formula:

  $$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}.$$

  Our test statistic is then

  $$t = \frac{\hat{\mu}^{(1)} - \hat{\mu}^{(2)}}{s_p\sqrt{(1/n_1) + (1/n_2)}}.$$

  This test still provides reasonably good power, since we're using all the data to estimate s_p.

  In this setting, where the two groups have the same variance, we say the data are **homoskedastic**.

- In the general case of two samples with separate (not pooled) variance, the variances must be estimated separately. The result isn't quite a t distribution, and this variant is often known as **Welch's *t*-test**. It's important to keep in mind that this test will have lower statistical power since we are using less data to estimate each quantity. But, unless you have solid evidence that the variances are in fact equal, it's best to be conservative and stick with this test.

  In this setting, where the two groups have different variances, we say the data are **heteroskedastic**.

■ 2.4 Some important warnings for hypothesis testing

- **Correcting for multiple comparisons** (**very important**): suppose you conduct 20 tests at a significance level of 0.05. Then on average, just by chance, even if the null hypothesis is wrong, one of the tests will show a significant difference (see this relevant xkcd). There are a few standard ways of addressing this issue:

  - Bonferroni correction: If we're doing m tests, use a significance value of α/m instead of α. Note that this is very conservative, and will dramatically reduce the number of acceptances.

  - False discovery rate (Benjamini-Hochberg): this technique guarantees α overall error by using the very small significances to allow slightly larger ones through as well.

- **Rejecting the null hypothesis**: You can never be completely sure that the null hypothesis is false from using a hypothesis test! Any statement stronger than "the data do not support the null hypothesis" should be made with extreme caution.

- **Practical vs statistical significance**: with large enough n, any minutely small difference can be made statistically significant. The first example below demonstrates this point. Sometimes small differences like this matter (e.g., in close elections), but many times they don't.

- **Independent and identically distributed**: Many of our derivations and methods depend on samples being independent and identically distributed. There are ways of changing the methods to account for dependent samples, but it's important to be aware of the assumptions you need to use a particular method or test.

Example: Practical vs statistical significance

Suppose we are testing the fairness of a coin. Our null hypothesis might be $p = 0.5$. We collect 1000000 data points and observe a sample proportion $\hat{p} = 0.501$ and run a significance test. The large number of samples would lead to a p-value of 0.03. At a 5% significance level, we would declare this significant. But, for practical purposes, even if the true mean were in fact 0.501, the coin is almost as good as fair. In this case, the strong statistical significance we obtained does not correspond to a "practically" significant difference. Figure 2.6 illustrates the null sampling distribution and the sampling distribution assuming a proportion of $p = 0.501$.

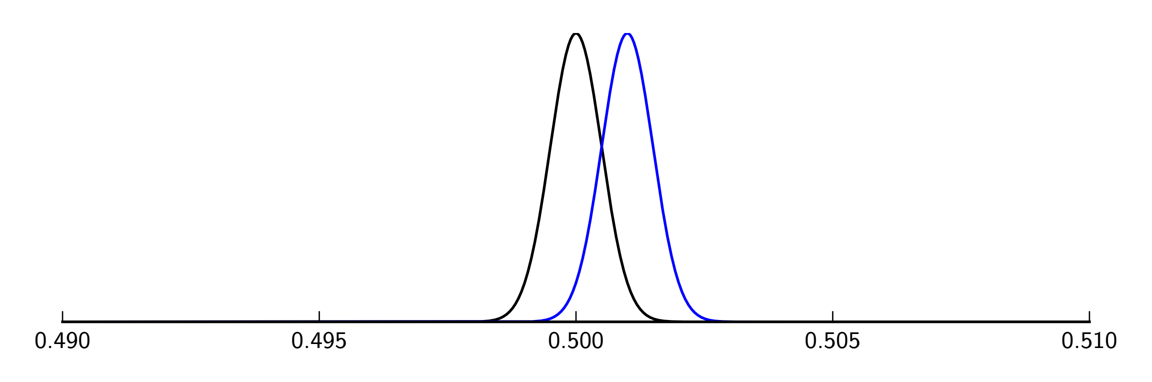


Figure 2.6: Sampling distributions for $p = 0.5$ (black) and $p = 0.501$ (blue) for $n = 1000000$. Note the scale of the x-axis: the large number of samples dramatically reduces the variance of each distribution.

Example: Pitfall of the day: Interpretation fallacies and Sally Clark

In the late 1990s, Sally Clark was convicted of murder after both her sons died suddenly within a few weeks of birth. The prosecutors made two main claims:

- The probability of two children independently dying suddenly from natural causes like Sudden Infant Death Syndrome (SIDS) is 1 in 73 million. Such an event would occur by chance only once every 100 years, which was evidence that the death was not natural.
- If the death was not due to two independent cases of SIDS (as asserted above), the only other possibility was that they were murdered.

The assumption of independence in the first item was later shown to be incorrect: the two children were not only genetically similar but also were raised in similar environments, causing dependence between the two events. This wrongful assumption of independence is a common error in statistical analysis. The probability then goes up dramatically[a].

Also, showing the unlikeliness of two chance deaths does *not* imply any particular alternative! Even if it were true, it doesn't make sense to consider the "1 in 73 million claim" by itself: it has to be compared to the probability of two murders (which was later estimated to be even lower). This second error is known as the *prosecutor's fallacy.* In fact, tests later showed bacterial infection in one of the children!

[a] See *Royal Statistical Society concerned by issues raised in Sally Clark Case*, October 2001.