MAT220 Lecture 13: More on Estimation

Contents

1 A Review of Estimation **2**

1.1 Review: estimating μ with known σ 2

2 Estimation with Unknown σ **3**

2.1 Where the known σ method breaks 3

2.2 Replacing σ with s 4

2.3 Student's t-distributions 4

2.4 Critical values for t 6

2.5 Confidence interval with unknown σ 7

2.6 Worked example: screws with unknown σ 8

3 Estimating a Binomial Proportion **8**

3.1 Defining a Sample Proportion 8

3.2 Defining a Sampling Distribution 9

3.3 Confidence intervals for p 10

3.4 The method 11

4 Worked Example: Basketball Scouting **11**

4.1 Example 1: 65 baskets in 95 attempts 12

4.2 Example 2: 773 baskets in 1000 attempts 12

In Lecture 12 we introduced the concept of *estimation*, and used it to guess the value of the population mean μ given some sample mean $\bar{x}$. In doing so, we deliberately made the assumption that σ, the true population standard deviation, was somehow already known. This helped simplify the concepts involved, which was very helpful at the time. However, we can imagine that in the real world, it is probably the case that if we don't know μ, then we don't know σ either. In this lecture we will explore the idea of estimation without assuming that σ is known. We will do this in two settings: estimating μ as in Lecture 12, and estimating the success probability p in a binomial experiment.

1 A Review of Estimation

Last lecture we introduced the concept of *confidence intervals* to estimate the true population mean μ. The main idea was to take a sample mean $\bar{x}$ and to use the Central Limit Theorem to construct the sampling distribution of $\bar{X}$, $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. This distribution has the true population mean μ at the very center, and therefore we can use it to relate $\bar{x}$ to μ. From here, we created small intervals around $\bar{x}$ which, by construction, would contain the true mean a certain amount of the time.

A common real-world use of this style of estimation appears in *quality control*. For the purposes of illustration, suppose that there is a factory that produces a large number of screws. In order to fit into whatever they are designed to fit into, these screws need to be about 20 mm long, with a 0.1 mm tolerance either way. This means that, for instance, a screw of 19.9 mm is acceptable, whereas a screw of length 19.89 mm is not.

In principle, we could guarantee a high quality of the screws by inspecting every single screw that the factory produces and measuring its length. That way, we could add up all of these numbers and divide by the total number of screws to get their true average length. However, in practice this would be far too time-consuming, and practically impossible. So, in a situation like this we can instead take some sample data and merely *estimate* the true average length of the screws. If this number were within the range between 19.9 and 20.1, then the screws are acceptable. This is the idea of *quality control*.

Quality control

Quality control uses sample data to decide whether a production process is behaving acceptably. Often it is unrealistic to test every single product, so it is easier and cheaper to use a statistical estimate instead.

1.1 Review: estimating μ with known σ

Keeping with the example of the factory, suppose that long-term studies tell us that the population standard deviation is known, with the exact value:

$$\sigma = 1.2 \text{ mm}.$$

Let's also suppose that quality control takes a simple random sample of $n = 50$ screws and finds a sample mean of $\bar{x} = 19.98$ mm.

We will now construct a 95% confidence interval for the true mean length μ. From Lecture 12, the confidence interval for μ with known σ can be calculated from the margin of error E. If we center the interval at an observed value of $\bar{x}$, the confidence interval will be:

$$\left[\bar{x} - z_c\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_c\frac{\sigma}{\sqrt{n}}\right].$$

For a 95% confidence interval, we know from the table in Lecture 12 that $z_c = 1.96$. So, the margin of error can be calculated:

$$E = z_c\frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{1.2}{\sqrt{50}} \approx 0.333.$$

Therefore, the 95% confidence interval centered around our observed sample mean $\bar{x} = 19.98$ will be:

$$\text{CI} = [19.98 - 0.333,\ 19.98 + 0.333] = [19.647,\ 20.313].$$

Interpretation

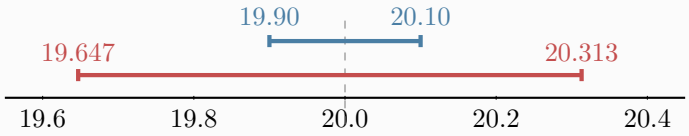
According to the setup, the allowed range for the average length is:

$$[19.90,\ 20.10].$$

But the 95% confidence interval that we found is

$$[19.647,\ 20.313].$$

This interval is much wider than the required engineering tolerance:

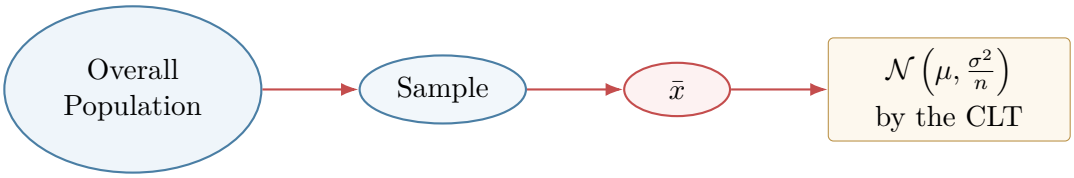


So, the sample does *not* provide strong evidence that the true mean is within the required tolerance. However, this shouldn't mean that we automatically shut down the factory. Instead, we have to accept that the sample data is not convincing enough either way: we don't have enough evidence to conclude that the true mean is outside of the desired 20 mm $\pm$ 0.1 mm, and we also don't have enough evidence to conclude that the true mean is inside those bounds either.

2 Estimation with Unknown σ

2.1 Where the known σ method breaks

In the previous example, we used the following workflow:



From there, we estimated μ via a margin of error calculated using:

$$E = z_c \frac{\sigma}{\sqrt{n}}.$$

However, in many real-world situations we simply *don't know* σ. This means that we cannot plug that number into the formula above, and therefore we can no longer estimate μ using this method. Instead, we need to replace σ with something that we do know, and then evaluate the consequences of this replacement.

2.2 Replacing σ with s

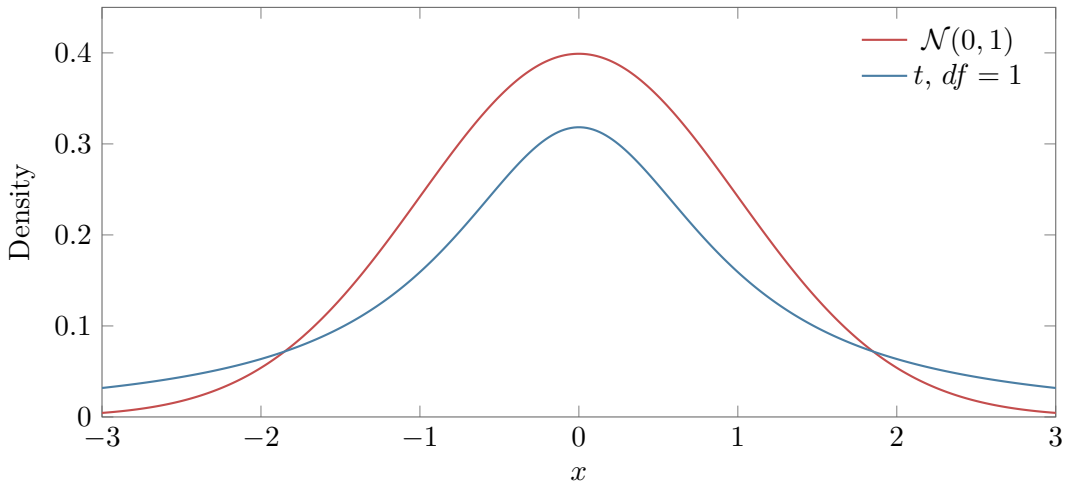
If we don't know what the true population standard deviation σ is, then it seems that the next best thing is to replace it with the sample standard deviation:

$$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}.$$

However, if we make this replacement, then we are introducing more uncertainty than we had before. Therefore, we have to be careful about which probability distribution we use to derive our estimates. As a matter of fact, it is no longer appropriate to use the standard normal distribution. Instead, we need to use a modification known as Student's t-distribution. We will now explore exactly what these t-distributions look like.

2.3 Student's t-distributions

Student's t-distributions are related to the standard normal distribution. The only major difference is that more of the area is distributed in the tails. A simple comparison is depicted below:



Here the standard normal distribution is in red, and the t-distribution is in blue. Notice that the blue curve is still symmetric and still centered at 0. In fact, the only clear difference is in the tails – they are *fatter*.[1] This is for a good reason: when σ is unknown we have more uncertainty, and therefore confidence intervals built off of this distribution should be wider than their normal-based relatives.

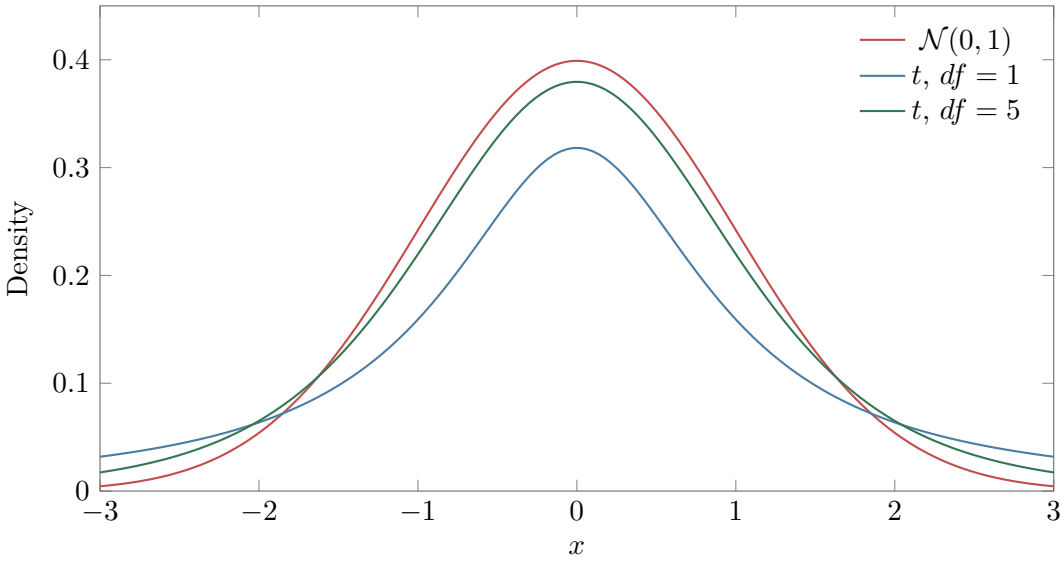
Notice in the graph above that there is an extra term "df" in the label of the t-distribution. In fact, this is an extra input that is required for the t-distribution to be fully described. This extra piece of information is known as the *degrees of freedom* (hence the letters df), and it is defined simply as:

$$df = n - 1.$$

The idea here is that this number will increase as the sample size does. As a sample size increases, we would generally expect that the sample standard deviation s becomes less influenced by random

[1] The technical term for this is *kurtosis*, which is a measure of how "heavy" a distribution's tails are.

noise, and therefore it becomes a better estimate of the true population parameter σ. If s gets closer to σ as n increases, then we would expect that the uncertainty built into the tails of the t-distribution should decrease. As n gets very large, we should have a sample standard deviation s very close to σ, and therefore we expect a t-distribution that looks very similar to the standard normal distribution $\mathcal{N}(0,1)$. In fact, that is exactly what happens: as df increases, a t-distribution looks more like the standard normal distribution:



As df gets bigger and bigger, the resulting t-distribution will look more and more like a normal distribution. This gives us another way to describe the standard normal distribution: it is a limit of standard t-distributions as $df \to \infty$.

Properties of Student's t-distributions

Student's t-distributions have the following properties.

1. They are symmetric about 0.

2. The total area under the curve is 1.

3. They have larger tails than $\mathcal{N}(0,1)$.

4. They require one extra input:

$$df = n - 1.$$

5. As df gets larger, the t-distribution gets closer to $\mathcal{N}(0,1)$.

2.4 Critical values for t

Student's t-statistic

Assume that the original population is normally distributed with mean μ. For samples of size n, with sample mean $\bar{X}$ and sample standard deviation S,[a] the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a Student's t-distribution with

$$df = n - 1$$

degrees of freedom.

[a] Here we have upgraded the sample standard deviation into a random variable in its own right.

The expression looks almost exactly like the z-score formula, except that we have replaced σ with s. Once have an observed sample, the formula changes:

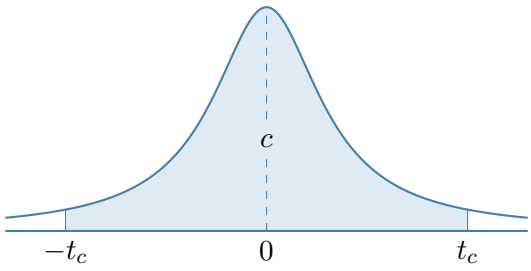
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \quad \xrightarrow{\text{Unknown } \sigma} \quad t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

Recall that when σ is known, we used a critical value z_c from the standard normal distribution. Geometrically, this corresponded to the area in the standard normal distribution that was exactly c. We then constructed confidence intervals to estimate μ by mapping this region back into the sampling distribution $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ using the z-score formula.

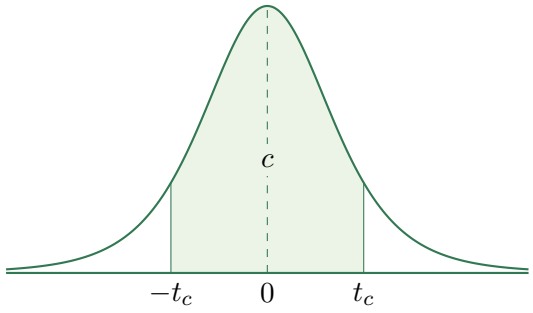
In contrast, when σ is unknown, we instead use a critical value from Student's t-distribution. In analogy to the critical value z_c, we will write the critical values in the t-distribution as t_c. Geometrically, these critical values play exactly the same role: they provide the endpoints of a region in the t-distribution that bounds an area of exactly c, centered at the origin. Since t-distributions depend on the degrees of freedom df, the exact value of t_c will depend on two pieces of information:

1. the confidence level c, and
2. the degrees of freedom $df = n - 1$.

For instance, consider the t-distributions for $df = 1$ and $df = 5$ drawn previously. Here, since the two curves have different shapes, the critical values t_c will change:



The area c when $df = 1$



The area c when $df = 5$

As you can see, the critical values for t_c will be larger when df is smaller. A table for common values of t_c is written below. Notice that it is much more complicated than the table for its counterpart z_c:

df	80%	90%	95%	98%	99%	99.5%
1	3.078	6.314	12.706	31.821	63.657	127.321
2	1.886	2.920	4.303	6.965	9.925	14.089
3	1.638	2.353	3.182	4.541	5.841	7.453
4	1.533	2.132	2.776	3.747	4.604	5.598
5	1.476	2.015	2.571	3.365	4.032	4.773
6	1.440	1.943	2.447	3.143	3.707	4.317
7	1.415	1.895	2.365	2.998	3.499	4.029
8	1.397	1.860	2.306	2.896	3.355	3.833
9	1.383	1.833	2.262	2.821	3.250	3.690
10	1.372	1.812	2.228	2.764	3.169	3.581
12	1.356	1.782	2.179	2.681	3.055	3.428
15	1.341	1.753	2.131	2.602	2.947	3.286
20	1.325	1.725	2.086	2.528	2.845	3.153
25	1.316	1.708	2.060	2.485	2.787	3.078
30	1.310	1.697	2.042	2.457	2.750	3.030
40	1.303	1.684	2.021	2.423	2.704	2.971
60	1.296	1.671	2.000	2.390	2.660	2.915
120	1.289	1.658	1.980	2.358	2.617	2.860
∞	1.282	1.645	1.960	2.326	2.576	2.807

Here, each column corresponds to a different common value of the confidence level c. Each row gives the value of t_c depending on some common degrees of freedom. The bottom row represents the case in which df gets arbitrarily large. In this case, the value of t_c will get closer and closer to the value of z_c, eventually approaching it in the limit.

2.5 Confidence interval with unknown σ

Now that we have access to critical values t_c, we follow the usual derivation of the margin of error to describe confidence intervals around observed values $\bar{x}$. In this case we have to use the t-statistic:

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

instead of the usual z-score formula. Using the critical values $\pm t_c$ and solving gives the margin of error:

$$E = t_c \frac{s}{\sqrt{n}}.$$

Notice the clear similarity to the case of known σ – here we have simply replaced σ with s and have used a t-distribution's critical value instead of the usual z_c. We can describe confidence intervals around observed sample means $\bar{x}$ as follows.

Confidence interval for μ with unknown σ

If the population standard deviation σ is unknown, we use the sample standard deviation s. The confidence interval for μ becomes:

$$[\bar{x} - E, \bar{x} + E] = \left[\bar{x} - t_c \frac{s}{\sqrt{n}},\ \bar{x} + t_c \frac{s}{\sqrt{n}}\right],$$

where $df = n - 1$.

The key differences between the known and unknown cases are summarized below.

	σ **known**	σ **unknown**
Score	$z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$
Critical value	z_c	t_c
Margin of error	$E = z_c \dfrac{\sigma}{\sqrt{n}}$	$E = t_c \dfrac{s}{\sqrt{n}}$
Confidence interval	$[\bar{x} - E, \bar{x} + E]$	$[\bar{x} - E, \bar{x} + E]$

2.6 Worked example: screws with unknown σ

Let's return to the example of the factory from Section 1. Suppose now that the factory does *not* know the true population standard deviation σ. Instead, a sample of 50 screws gives:

$$n = 50, \qquad \bar{x} = 19.98, \qquad s = 1.1 \text{ mm}.$$

If we now want to construct a 95% confidence interval for the true mean length μ, we must use s instead of σ, and we must use the t-statistic instead of the z-score. Since $n = 50$, the degrees of freedom are $df = n - 1 = 49$. From a t-table, for a 95% confidence interval with $df = 49$, the critical value is approximately:

$$t_c \approx 2.010.$$

Using $s = 1.1$ and the above, we compute the margin of error to be:

$$E = t_c \frac{s}{\sqrt{n}} = 2.010 \cdot \frac{1.1}{\sqrt{50}} \approx 0.313.$$

Therefore, the 95% confidence interval centered around our observed value of $\bar{x}$ will be:

$$\text{CI} = [19.98 - 0.313,\ 19.98 + 0.313] = [19.667,\ 20.293].$$

Here, we are met with the same interpretation as the previous example.

3 Estimating a Binomial Proportion

3.1 Defining a Sample Proportion

Suppose that we witness part of a binomial experiment. We happen to observe n trials, and we count r successes. Because of random variation in the sample, this is not enough information to deduce the true value of p, the probability of success. However, we can still use this information to estimate the value of p. As with the previous section and Lecture 12, we will create a confidence interval around sample data, which will give us an appropriate error bound around our estimate. Our first goal is to obtain a meaningful sample statistic.

Sample proportion

If we observe r successes in n binomial trials, then the sample proportion is defined as:

$$\hat{p} = \frac{r}{n}.$$

From this, we also define $\hat{q} = 1 - \hat{p}$.

The quantity $\hat{p}$ is going to be our point estimate for the true success probability p.

3.2 Defining a Sampling Distribution

When discussing the normal approximation to a binomial distribution in Lecture 11, we mentioned that the usual binomial random variable X that counts the number of successes within a binomial experiment can be broken down into a sum of Y_i, where Y_i is a random variable defined for each individual trial of the experiment:

$$Y_i = \begin{cases} 1 & \text{if trial } i \text{ is a success,} \\ 0 & \text{if trial } i \text{ is a failure.} \end{cases}$$

In symbols, we saw in Lecture 11 that:

$$X = Y_1 + Y_2 + \cdots + Y_n$$

and therefore dividing X by n gave a quantity that acts like a sample mean. We will now put this together with our sample proportion: we can introduce a random variable, written $\hat{P}$, if we define it as:

$$\hat{P} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}.$$

In fact, this is exactly what we did in Lecture 11 without giving it a name. Written this way, we can see that $\hat{P}$ is a random variable whose observed values $\hat{p}$ act like sample means taken from a sample of size n – the observed value r above is the sum of the outcomes of the n individual random variables Y_i, and then we divide by n to get the mean.

We saw in Lecture 11 that each one of these individual random variables Y_i has:

$$E(Y_i) = p \qquad \text{and} \qquad \text{Var}(Y_i) = pq.$$

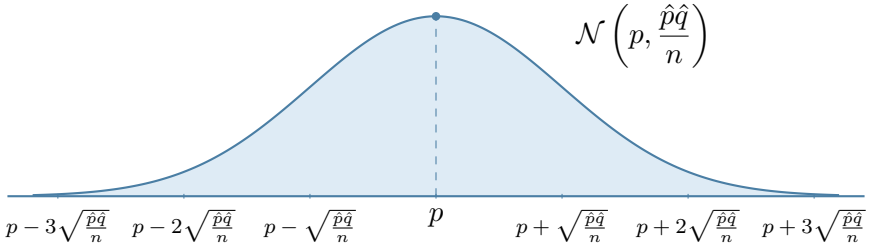
The Central Limit Theorem tells us that the collection of all sample means follows an approximate normal distribution in which the center is the mean of the original variable from which we form the sample, and the new variance is the variance of the original variable divided by n, the sample size. In our case, the Central Limit Theorem tells us that $\hat{P}$ follows an approximate normal distribution with parameters:

$$\mathcal{N}\left(p, \frac{pq}{n}\right).$$

This appears to give us precisely what we need: we have just confirmed that our sample proportion has an approximate normal distribution, and the true mean is at the center. From here, it seems that we can start to create some confidence intervals. However, there is one fundamental issue: we are trying to estimate p, the true probability of success. If we don't know what the exact value

of p is, then we *also* don't know what the value of $\frac{pq}{n}$ is either! In other words, if we were to try and continue, and create some sort of margin of error, how would we actually calculate it, without already knowing the value of p?

Here, we have to accept that the true sampling distribution for $\hat{P}$ cannot serve our needs. Instead, we have to do the next-best thing, which is to replace the variance term with point estimates. In practice, we replace p and q by the observed values $\hat{p}$ and $\hat{q}$, giving the following approximation:



This will be the setting of our estimation.

Sampling distribution for $\hat{P}$

For a binomial experiment, the sample proportion is approximately normally distributed:

$$\hat{P} \approx \mathcal{N}\left(p, \frac{\hat{p}\hat{q}}{n}\right).$$

This approximation is reasonable when

$$n\hat{p} > 5 \qquad \text{and} \qquad n\hat{q} > 5.$$

3.3 Confidence intervals for p

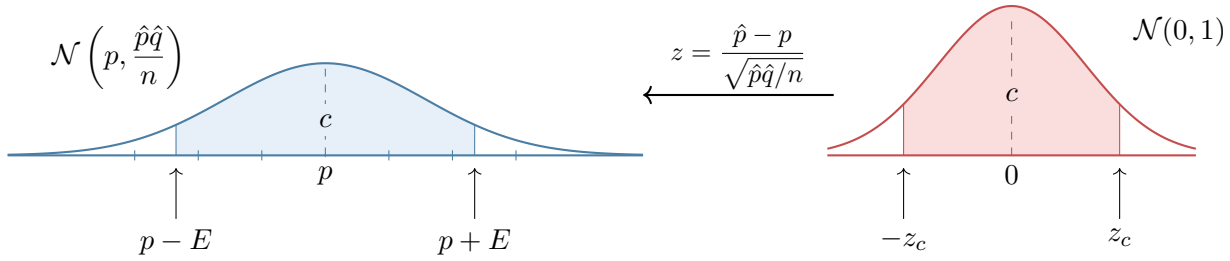
Now we construct confidence intervals exactly as before. The estimated standard error is

$$\sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

As such, the formula for the z-score becomes:

$$z = \frac{\hat{p} - p}{\sqrt{\hat{p}\hat{q}/n}}.$$

To derive a confidence interval with confidence level c, we find the critical value z_c in $\mathcal{N}(0,1)$ that gives $P(-z_c < Z < z_c) = c$. Using the z-score formula in reverse then maps this region back into the sampling distribution:



The margin of error E is given by z_c multiplied by the standard error. In this case:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Therefore, confidence intervals around observed values $\hat{p}$ can be defined as follows.

Confidence interval for a binomial proportion

Given r successes in n trials,

$$\hat{p} = \frac{r}{n}, \qquad \hat{q} = 1 - \hat{p}.$$

The confidence interval for the true success probability p is

$$\left[\hat{p} - z_c \sqrt{\frac{\hat{p}\hat{q}}{n}},\ \hat{p} + z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}\right].$$

3.4 The method

To estimate a binomial success probability p:

Step 1: Calculate

$$\hat{p} = \frac{r}{n}, \qquad \hat{q} = 1 - \hat{p}.$$

Step 2: Check that the normal approximation is reasonable:

$$n\hat{p} > 5 \qquad \text{and} \qquad n\hat{q} > 5.$$

Step 3: Find the critical value z_c for the chosen confidence level.

Step 4: Calculate the margin of error:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Step 5: Write the confidence interval:

$$[\hat{p} - E, \hat{p} + E].$$

4 Worked Example: Basketball Scouting

Suppose that a basketball team only wants to hire players whose true free-throw success rate is at least 70%. In symbols, the team wants evidence that $p \geq 0.70$.

You scout a player and watch him shoot some free throws. You cannot observe him all day, so you take a sample of his performance and decide to compute a 95% confidence interval to estimate his true shooting success rate p.

4.1 Example 1: 65 baskets in 95 attempts

Suppose the player takes $n = 95$ shots, and scores $r = 65$ of them. Then our point estimate of the success probability will be

$$\hat{p} = \frac{65}{95} \approx 0.684,$$

and therefore $\hat{q} = 1 - \hat{p} \approx 0.316$. Before going any further, we should first confirm the approximation conditions. We have:

$$n\hat{p} = 95(0.684) \approx 65, \qquad \text{and} \qquad n\hat{q} = 95(0.316) \approx 30.$$

These are both greater than 5, so the normal approximation is reasonable. For a 95% confidence interval, we have that $z_c = 1.96$. Therefore, we calculate the margin of error to be:

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{(0.684)(0.316)}{95}} \approx 0.093.$$

The 95% confidence interval around our observed value $\hat{p} = 0.684$ is therefore:

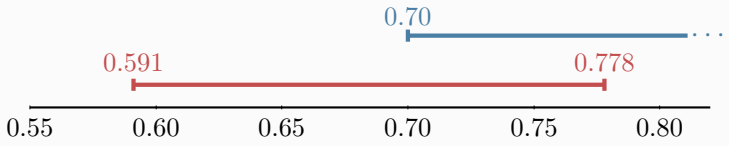
$$\text{CI} \approx [0.684 - 0.093,\ 0.684 + 0.093] = [0.591,\ 0.778].$$

Interpretation

According to the setup, we are looking for a desired threshold of $p = 0.70$. However, our confidence interval is

$$[0.591,\ 0.778].$$

This interval straddles the threshold:



Therefore, we have a similar situation to that of Section 1.1 – we cannot conclusively say that the player's true shooting rate is above or below 70%. Simply put, this sample does not provide enough convincing evidence either way.

4.2 Example 2: 773 baskets in 1000 attempts

After finding that your first observation was inconclusive, you go back to observe the player for a second time, this time for much longer. Suppose that now, you see the player take $n = 1000$ shots, and he scores $r = 773$ of them. Now, our point estimate of the success probability will be

$$\hat{p} = \frac{773}{1000} = 0.773,$$

and therefore $\hat{q} = 1 - \hat{p} = 0.227$. Again, we should first double-check that the approximation conditions hold. We have:

$$n\hat{p} = 1000(0.773) = 773, \qquad \text{and} \qquad n\hat{q} = 1000(0.227) = 227.$$

These are both greater than 5, so the normal approximation is reasonable. Again using the critical value $z_c = 1.96$, we calculate the margin of error:

$$E = z_c\sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96\sqrt{\frac{(0.773)(0.227)}{1000}} \approx 0.026.$$

The 95% confidence interval around our observed value $\hat{p} = 0.773$ is then:

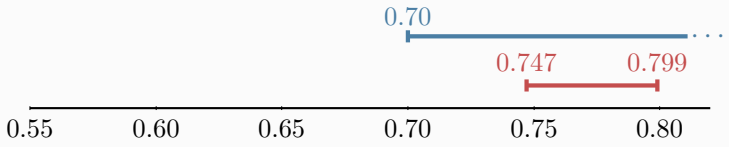
$$\text{CI} = [0.773 - 0.026,\ 0.773 + 0.026] = [0.747,\ 0.799].$$

Interpretation

In this case, our confidence interval is

$$[0.747,\ 0.799].$$

This entire interval now sits well above the desired threshold of 0.70:



Since this method captures the true value of p about 95% of the time, we have convincing evidence that the true value of p is above 0.70. So, this sample provides convincing evidence that the player meets the team's standard.