STATISTICS TOPIC NOTES

Predicting Individuals vs. Estimating Groups

Understanding the difference between predicting individual values and estimating group values is one of the most puzzling distinctions for students of statistics. The difference between individuals and groups is easy. The problem is that the uncertainty of individual prediction is different from the uncertainty of group estimation in ways that seem mysterious. "Which standard deviation do I use?" and "Do I divide by the square root of *n* or not?" are common student questions. It is extremely important to get this right. Uncertainty about individual values is usually orders of magnitude larger than uncertainty about group values. In this Topic Note, I try to clarify this very important topic. My objective is not to give you a set of rules to use. My objective is to explain the principle so well that you will inerrantly make the correct choice.

In this Topic Note, I consistently use the term *prediction* to refer to guessing the value of an individual; I consistently use the term *estimation* to refer to guessing the value of a group mean. However, you should be warned not to rely upon the appearance of the terms "prediction" or "estimation" to clue you which standard deviation to use. No one enforces a code of consistency of vocabulary upon writers of statistics, who often use these and other terms interchangeably. Moreover, when you are deciding how to approach a data analysis problem of your own, you will not have a written guide, so will have to understand the concepts in order to select the correct approach.

Let me begin with a simple example, stripped of the interesting but distracting detail of real life. I do this in order to expose the principle unambiguously. When you have understood the principle clearly, I will move up to a more practical example.

Example 1. Before writing this Topic Note, I drew a random sample of four numbers from an uncertainty distribution that I know, but I will not tell you what it is – yet. The distribution from which I drew is Population 1. So there is a population of all possible outcomes that I know, but you do not. There is also a set of probabilities – one probability for each possible outcome – that tells how likely I am to get each outcome on each draw. I also know the probabilities, and you do not. The set of outcomes and their probabilities remained the same every time I drew, and I drew independently, so that the outcome that I selected on one draw did not affect the outcome that I selected on any other draw. This method of selection means that my sample is a Random Sample and is likely to be representative of Pop 1 (at least, as representative as you can expect a sample of four draws to be.)

Here are the four outcomes that I drew, in order: 1, 3, 1, 5.
I now pose two tasks for you:

- T-ind: Predict the next number that I will draw from the distribution.
- T-mean: Estimate the mean of the distribution from which I am drawing.

Do you understand T-ind and T-mean are different tasks?

Task T-ind ("Task individual") requires the prediction of an *individual* outcome in Population 1. This is the value of one outcome – the outcome that I will select on the 5th draw. The other

.

members of Pop 1 contribute nothing to this value. The value is a property of the one outcome to be selected and is not shared with the group of other possible outcomes.

Task T-mean (“Task mean”) requires estimation of a *group* property – the mean of the entire Population 1. This is a number to which all possible outcomes contribute. The value does not belong to any one individual outcome, but is shared by all.

You should know now that the best solution for T-mean is the sample mean $\bar{x}$ = (1 + 3 + 1 + 5)/4 = 2.5. The reasons were discussed in detail in the Estimation section of the Topic Note on Estimation and Sampling Distributions. A brief summary of that discussion goes as follows: A randomly selected sample is representative of the population from which it is drawn. The larger the sample, the more the sample looks like its population. Therefore, the sample mean is representative of the population mean. The larger the sample, the closer the sample mean gets to the population mean.

Also, you may have guessed that the solution for T-ind is also the sample mean $\bar{x}$ = (1 + 3 + 1 + 5)/4 = 2.5. That is correct. But the reasons have not been stated explicitly before now. Here is the intuition: You do not know if the 5th draw will come from the high side of its population or from the low side. If you guess a value on the high side, then you will have a big potential error if the actual 5th draw is any of the low-side possibilities – although you will have a small potential error if the 5th draw is any of the high-side possibilities. If you guess a value on the low side, then you will have a big potential error if the actual 5th draw is any of the high-side possibilities – although you will have a small potential error if the 5th draw is any of the low-side possibilities. But in *neither* of these cases, high guess or low guess, do the potential high-side errors balance the potential low-side errors. If you guess high, the potential negative errors predominate; if you guess low, the potential positive errors predominate. The potential high-side errors balance the potential low-side errors at only one value: the true mean of the population. (This was demonstrated in the interpretation of the mean on pages 5-6 in the Topic Note on Mean and Standard Deviation.) So if you want to minimize the average magnitude of potential error in guessing the value of the 5th draw, you should guess the true mean of the population. However, you do not know the value of the true mean. Instead, as an approximation, you can use your best guess at the true mean, which is the sample mean. So you should predict that the 5th draw will be the sample mean = 2.5.

You now have solutions for the two tasks that I posed for you:

- T-ind: Predict the next number that I will draw to be the sample mean of the numbers that I have already drawn = 2.5.
- T-mean: Estimate the mean of the distribution from which I am drawing to be the sample mean of the numbers that I have already drawn = 2.5.

But you should not expect either the actual 5th draw or the actual population mean to be 2.5. You should expect there to be some deviation between your prediction and the actual 5th draw, and between your estimate and the true population mean. How much deviation should you expect in each case? This is the point at which you should really start paying attention in this Topic Note. For the confusion that students experience is not about the prediction or estimate, but about their expected error deviations. Let us examine the error deviation for each task, starting with T-ind.

Deviation for T-ind: The expected error in predicting the 5th draw

Let us call the as-yet-unknown value of the 5th draw x_5. Your estimate of draw 5 is the sample mean $\bar{x}$ = 2.5. The deviation of the 5th draw from your estimate is $x_5 - \bar{x} = x_5 - 2.5$. This is the error you make when you predict that the 5th draw will be 2.5. It is called the **prediction error**. You know the value of your estimate, 2.5. But you do not know the value of the 5th draw, x_5. So you do not yet know the value of the prediction error, x_5 - 2.5. You would like to guess the magnitude of the prediction error for the 5th draw. To this end, it would be helpful if you had some examples of prediction errors so that you could get an idea of what the prediction errors typically turn out to be. Even better than examples would be a representative sample of prediction errors. With a representative sample, or a Random Sample, you could guess the typical magnitude of the prediction error for the 5th draw and reasonably expect that the actual draw 5 would deviate from 2.5 by about that much.

But you do have a random sample of prediction errors! Consider the four draws (1, 3, 1, 5) that I have already made. They are a Random Sample from Pop 1. That is how I selected them. So their deviations from 2.5 – namely, 1 - 2.5, 3 - 2.5, 1 - 2.5, 5 - 2.5 – are a Random Sample of deviations from 2.5. Before I drew them, each of the first four draws had the same set of possible outcomes (Pop 1) to choose from as x_5 does, and the same selection probabilities. So before I drew them, each of the first four draws had the same set of possible *deviations* from 2.5 as x_5 does, and the same selection probabilities. That is, the uncertainty distribution of the deviation of x_5 from 2.5 is the same as the uncertainty distribution of the deviations of each of the first four draws from 2.5.[1] Therefore, you are entitled to use the first four deviations from 2.5 to guess the magnitude of the prediction error, x_5 - 2.5, the fifth deviation from 2.5.

How to do that? The first four deviations from 2.5 are 1 - 2.5, 3 - 2.5, 1 - 2.5, 5 - 2.5. In other words, -1.5, 0.5, -1.5, 2.5. You could calculate the mean of their absolute values: (1.5 + 0.5 + 1.5 + 2.5) / 4 = 1.5. But statisticians don't generally do that. Instead, for somewhat mysterious reasons, statisticians square the deviations and then average the squared deviations. But, again for somewhat mysterious reasons, when averaging, statisticians divide by 3 instead of 4.[2] Finally, statisticians take the square root. The result is $\sqrt{(1.5^2 + 0.5^2 + 1.5^2 + 2.5^2)/3}$ = 1.9149. If you have a sense of déjà vu in reading this paragraph, it is because you have seen this sort of thinking before – in the motivation for the standard deviation in the Topic Note on Mean and Standard Deviation. The ordinary standard deviation assesses the average magnitude by which the data in a sample differ from their mean. The ordinary standard deviation of the four sample draws 1, 3, 1, 5 is STDEV(1,3,1,5) = 1.9149.

The takeaway from the discussion of T-ind: The best prediction of the next draw from Pop 1 is the sample mean of a Random Sample already drawn from Pop 1. And the best guess of

[1] There is a little subtlety here. You may have noticed it: The prediction error itself has an uncertainty distribution that is related to, but different from, the uncertainty distribution of the five draws. The possible outcomes for the prediction error are the possible outcomes for the five draws, minus 2.5. So shift Pop1 down by 2.5, and carry along the probabilities with the shift to get the uncertainty distribution of the prediction errors.

[2] The reasons for squaring and dividing by 3 are justifiable, but take time to explain, and add no discernible value to the course.

.

the prediction error for the next draw from Pop 1 is the ordinary standard deviation of that Random Sample. It is vital to understand that prediction error estimates the deviation of an *individual* outcome from its prediction. It is vital to understand the conceptual difference between T-ind and T-mean. I next discuss T-mean, which is about the deviation of a *group* outcome.

Deviation for T-mean: The expected error in estimating the mean of Pop 1

Let us call μ the as-yet-unknown mean of Pop 1. Your estimate of μ is $\bar{x}$ = 2.5. The deviation of your estimate from μ is $\bar{x}$ - μ = 2.5 - μ. This is the error you make when you estimate that μ will be 2.5. You do not know the value of the error because you do not know the value of μ. Moreover, you are not likely ever to know it, in practice.[3] This is because you would need to exhaustively collect all of Pop 1 and average all of the possible outcomes in Pop 1 to calculate μ. By taking a sample, you have implicitly announced your intention to avoid that exhaustive analysis and settle for a statistical approximation.

Still, you would like to guess the magnitude of the estimation error for μ. To this end and paralleling the preceding discussion of T-ind, it would be helpful if you had some examples of estimation errors for μ so that you could get an idea of what the estimation errors typically turn out to be. Even better than examples would be a representative sample of estimation errors. With a representative sample, or a Random Sample, you could guess the typical magnitude of the estimation error for μ and reasonably expect that the estimate 2.5 would deviate from μ by about that much.

Alas! You will never get a sample of estimation errors! If you did have a sample of estimation errors, what would it look like? It would be $\bar{x}_1 - \mu, \bar{x}_2 - \mu, ..., \bar{x}_n - \mu,$ where each of the values $\bar{x}_1, \bar{x}_2, ..., \bar{x}_n$ is the mean of a sample of 4 draws (like our one sample of 4 draws: 1, 3, 1, 5). There are two reasons why you will never have such a sample: (1) You will never have μ, as noted in the preceding paragraph. (2) You will never have multiple means $\bar{x}_1, \bar{x}_2, ..., \bar{x}_n$. On point (2): If you were to try to get *n* more means, by making 4*n* additional draws, it would make sense to combine all of them to make one big sample and use its mean to estimate μ. Then you would be wondering about its expected deviation from μ! You would have to ponder the deviations not of means of samples of size 4, but of means of samples of size 4*n*. In practice, you will always have only one sample and you will not know the true mean.[4]

So what can be done? I suggest recalling appropriate material from the Topic Note on Estimation and Sampling Distributions. There, you learned three important properties about the uncertainty distribution (the sampling distribution, AKA Population 3) of the sample mean. You learned that the probability curve of the sampling distribution is approximately normal if the sample size is sufficiently large; that the true mean of the sampling distribution (Pop 3) is the same as the true mean of Pop 1; and that *the standard deviation of the sampling distribution (Pop 3) is the standard deviation of Pop 1, divided by the square root of the sample size.*

[3] This is unlike task T-ind. In predicting the 5th draw, you will learn the magnitude of your prediction error as soon as I make the 5th draw and tell you what it is. In this example, I happen to know all possible outcomes and probabilities for Pop 1. If I tell you, then you could calculate and know μ. But in a real-world problem, I would not know Pop 1 and could not tell you.

[4] If you did know μ, then you would know the answer! So why would you then be taking a sample to estimate μ?

In Example 1, the sample size (n = 4) is too small to invoke the first property. But the second and third properties still hold. These properties are important for task T-mean. In T-mean, you are estimating μ by the sample mean $\bar{x}$ = 2.5 and you want to guess the expected magnitude of your error. Your error is $\bar{x} - \mu$ = 2.5 - μ. Now 2.5 is one of the sample means in Pop 3. And the second property tells you that the true mean of Pop 3 is the same as the true mean of Pop 1 = μ. So the deviation of 2.5 from μ (i.e., 2.5 - μ) is the same as the deviation of 2.5 from the mean of Pop 3. Now, the average deviation of members of Pop 3 from the mean of Pop 3 is the standard deviation of Pop 3. (That is the meaning of standard deviation.[5]) So you can expect 2.5 to differ from μ by approximately the standard deviation of Pop 3. By the third property just referenced in the preceding paragraph, the standard deviation of Pop 3 is the standard deviation (call it σ) of Pop 1, divided by the square root of the sample size.

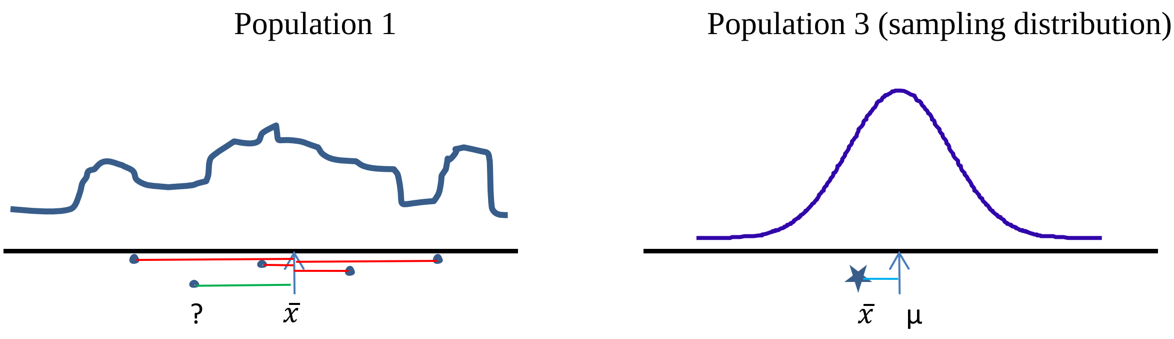
So the solution to the expected deviation for T-mean is $\sigma / \sqrt{n}$. To get an actual number, you need values for σ and n. The sample size is n = 4, but the true standard deviation σ of Pop 1 is unknown. However, you can estimate σ. In the Topic Note on Estimation and Sampling Distributions, you learned that the ordinary standard deviation of the sample is an excellent estimate of σ. The sample standard deviation = STDEV(1,3,1,5) = 1.9149. So the best estimate of the deviation for T-mean is 1.9149 / $\sqrt{4}$ = 0.9575.

The takeaway from the discussion of T-mean: The best estimate of the true mean of Pop 1 is the sample mean of a Random Sample already drawn from Pop 1. And the best estimate of the estimation error is the ordinary standard deviation of that Random Sample, divided by the square root of the sample size. It is vital to understand that estimation error is about the deviation of a *group* outcome from the true mean of group outcomes. It is vital to understand the conceptual difference between T-ind and T-mean.

Since understanding the distinction between T-ind and T-mean is so vital, I will come at the explanation again from another angle – graphical.

Figure 1 illustrates prediction error and estimation error in the general case.

Figure 1. Prediction error and estimation error



A. Prediction error

B. Estimation error

[5] You can review this interpretation of standard deviation in the Topic Note on Mean and Standard Deviation, especially at page 7.

On the left, Figure 1A shows a generic Population 1, consisting of individuals. Every member of Population 1 is an individual value. Four individuals from Pop 1 have been randomly sampled. They are the four dots below the distribution with horizontal red lines leading to their sample mean. The values of the four sampled individuals are known. Therefore, their sample mean $\bar{x}$ is known. Therefore, the deviations (the red lines) of all four individuals from the sample mean are known. In prediction, you must guess the value of the next individual to be randomly sampled from Pop 1. That 5th individual is tentatively shown by the question mark – indicating that its location in Pop 1 is not known until it is drawn and that the location shown is only one of a vast number of possibilities for the 5th draw. Your job in task T-ind is to predict the location of the 5th draw and estimate your prediction error. Your prediction is the sample mean, indicated by the arrow. That is where the positive and negative red deviations cancel out. If the 5th draw turns out to be the value shown at the question mark, then your prediction error will be the green line. The four known red deviations provide great clues as to what your as-yet unknown green deviation will be because the 5th draw will be done just like the first four were. So the prediction error of the 5th draw will be just like the first four deviations. The average magnitude of the four red deviations is measured by the standard deviation of the four draws. The standard deviation of the four draws is an excellent estimate of your prediction error.

Figure 1B shows a sampling distribution – Population 3 – the uncertainty distribution of the sample mean. Every member of Pop 3 is a sample mean, and all possible sample means are in Pop 3.[6] The sample mean of the four draws from Pop 1 is shown as the star in Figure 1B. The value of the star is known. The true mean of Pop 3 is indicated by the μ, which is also the true mean of Pop 1. μ is not known. So the position that Figure 1B shows for μ is really only one of a vast number of possibilities. In estimation, you must guess the value of the true mean μ. Your job in task T-mean is to estimate the location of the true mean μ and to guess your estimation error. Your estimate is the sample mean, indicated by the star. If the true mean happens to be at the value shown at μ, then your estimation error will be the light blue line from $\bar{x}$ to μ. Unlike the prediction error problem illustrated in Figure 1A, you have no sample of mean deviations to guide you in Figure 1B. However, you do not need them because you have the statistical theory that relates deviations of means in Figure 1B to deviations of individuals in Figure 1A. That theory says that the average deviation from the true mean for all of the sample means in Pop 3 is the average deviation from the true mean for all of the individuals in Pop 1, divided by $\sqrt{n}$. So the best estimate of the estimation error in Figure 1B is the sample standard deviation of the four individual draws, divided by $\sqrt{4}$.

You now have solutions for the expected error for the two tasks that I posed for you:

- T-ind: The error deviation for predicting the next number that I will draw is estimated by the standard deviation of the numbers that I have already drawn = 1.9149.
- T-mean: The error deviation for estimating the mean of the distribution from which I am drawing is estimated by the standard deviation of the numbers that I have already drawn, divided by the square root of the sample size = 0.9575.

[6] The shape of Pop 3 is shown as normal, which would be reasonable if the sample size were sufficiently large. Four is not sufficiently large, unless Pop 1 is normal, and Figure 1A does not look normal. But the shape of Pop 3 is not important for the discussion of this paragraph.

.

At the risk of over-simplification …

- Task T-ind is about individuals in Figure 1A.
- Task T-mean is about group means in Figure 1B.

{By the way, in Example 1, the actual fifth draw was 4, so the prediction error was 4 - 2.5 = 1.5. The actual uncertainty distribution (Pop 1) was the throw of a fair die, with possible outcomes 1,2,3,4,5,6, with equal probabilities of 1/6. So you can calculate $\mu = 3.5$ and the magnitude of the estimation error was |2.5 - 3.5| = 1.0.}

Now that you understand (hopefully) the principle of the distinction between prediction error and estimation error, I will apply it to a more practical example.

Example 2. Prediction and estimation for Austin apartment rent.

A Random Sample of 60 apartments has been drawn from the population of all Austin apartments. Their rents have been compiled and are shown in Table 1.

Table 1. The monthly rents of 60 randomly selected Austin apartments

519	530	450	425	470	415	505	470	625	470	659	605
765	580	520	770	700	399	445	470	745	480	650	929
475	995	495	445	450	585	565	700	540	460	750	695
575	565	420	510	785	525	650	455	650	600	455	455
415	620	575	635	485	495	515	550	595	575	430	1050

You have two tasks:

- T-ind. Predict the rent of 123 Lotus Avenue, an Austin apartment that is not among the sample of 60.
- T-mean. Estimate the mean rent of all Austin apartments.

The best solution for both tasks is the sample mean $\bar{x}$ = \$572.27, although for somewhat different reasons. For T-ind, nothing is known about 123 Lotus Avenue, except that it is an Austin apartment. You do not know how big it is or where it is located (unless you know where Lotus Avenue is). It could be any Austin apartment. You do not want to guess too big or too small and thereby expose yourself to the risk of a big prediction error. Ideally, your guess should be the true mean rent, at which value the average prediction error for all Austin apartments is a minimum. Since you do not know the true mean rent, you use the sample mean rent as the next best thing.

For T-mean, you use the sample mean to estimate the true mean because the sample should be representative of the population of rents. Therefore, the sample mean should be representative of the population mean.

Although the solutions for individual prediction and group estimation are the same number, the error estimates are different. For T-ind, the picture is like Figure 1A, but with 60 dots and 60 red deviation lines below the figure instead of four. The average magnitude of the 60 individual deviations from $572.27 is an excellent guess at the magnitude of the deviation from the $572.27 prediction for 123 Lotus Avenue. The average magnitude of the 60 deviations is the sample standard deviation of the 60 individual apartment rents = $140.52.

For T-mean, the picture is like Figure 1B. The statistical theory tells you that the average deviation of all possible sample means from the true mean rent is the same as the average deviation of all individual apartment rents from the true mean rent, divided by $\sqrt{60}$. That is, standard deviation of Figure 1B = standard deviation of Figure 1A ÷ $\sqrt{60}$. The best guess at this is $140.52 ÷ $\sqrt{60}$ = $18.14.

You predict that the rent of 123 Lotus Avenue is $572.27 and you expect your prediction to differ from the actual rent by ±$140.52.

You estimate that the mean rent of all Austin apartments is $572.27 and you expect your estimate to differ from the true mean rent by ±$18.14.

Here are two scenarios that illustrate practical business applications of T-ind and T-mean:

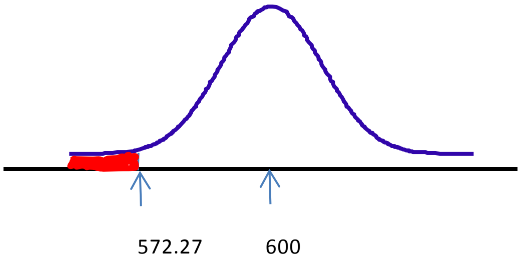
Business Scenario 1. You will be moving to Austin to pursue your MBA fulltime at UT. How much should you budget for rent? You have not yet rented an apartment, but you have obtained the Random Sample of 60 rents shown in Table 1. Even with such a rudimentary dataset, you can infer useful facts. You will "draw" your apartment from Population 1. Thus your question is a T-ind question – How much should you pay for one individual apartment? You can expect to pay $572.27, more or less, for your apartment. This gives you a starting point for budgeting. How much more or less? The typical apartment could easily rent for $140.52 more or less than the $572.27 prediction. The ±$140.52 reflects the cost of more or fewer amenities and attractions than the average apartment. So if you were hoping to rent a typical apartment on the high side of average with typical high-side size and amenities, but paying more than about $710 (= $572.27 + $140.52) would crimp your budget, you may need to revise your plans. On the other hand, if you know that you must really economize on rent, then the average plus-minus variation of $140.52 should give you hope that you can find a typical apartment on the low side of average for about $430 (= $572.27 - $140.52). {*If you want to do a more sophisticated analysis of how apartment attributes affect rent, see Example 3 and Business Scenarios 3 and 4, yet to come.*}

Business Scenario 2. You are an apartment landlord. You are considering entering the Austin apartment market. You acquire apartments that reflect the range of apartments in each of your cities – a mix of high-end, low-end, and a lot of middle-market properties. From experience, you know that your apartments need to average $600 or more rent for the financial numbers to work for you. You have the data in Table 1 available. Do you enter the Austin market? To be sure, you will no doubt do a quite detailed analysis before deciding. But even the simple data in Table 1 can help inform your business judgment. You will not acquire a single apartment but a group. So you do not have a T-ind question. Your apartment rents will vary. Some will be high, others low. You will get more than $600 on some, less on others. You will lose money on some and make money on others. As long as the average rent exceeds $600, you will do OK. Your question is a T-mean (Population 3) question. What is the mean rent of all Austin apartments? Since your apartments will reflect the population of Austin apartments, if the true mean of the Austin population exceeds $600, yours will probably also. The best estimate of the true mean of Austin apartment rents is $572.27, which is less than your $600 requirement. That number does not look good. On that alone, you could justify canceling plans to enter Austin. But you might reason that the 60 apartments in Table 1 are only a sample. They are not the whole population. The population could still have a true mean that exceeds $600. You are wondering: "What if the sample was a fluke? The sample is close to my cutoff. Maybe I should take a larger sample and/or do a more refined analysis. I don't want to miss out on a good business opportunity because of unrepresentative data."

The following is one approach to addressing your doubts: Hypothetically, what if the true mean rent were $600 – how likely is it that you would get a sample like Table 1 with a mean of $572.27 or less? If you knew that, you could judge whether Table 1 could be a fluky sample of unrepresentatively low rents. So the question is, What is the probability of getting a sample mean of $572.27 or less if the true mean is $600?[7] The question is now posed in a familiar statistical format that you can solve.

Figure 2 shows how to solve it. Figure 2 shows the population of all possible sample means as a normal curve. You know that the shape is normal because of the Central Limit Theorem (Property 1 of sampling distributions) since $n = 60$ is sufficiently large. You also know that the true mean of the sampling distribution in Figure 2 is $600 *if* (as we are hypothesizing) the true mean rent of Population 1 is $600. That is Property 2 of sampling distributions. The red shaded region shows all of the sample means that are less than $572.27. How many are there? That is an easy question to answer. All you need to finish the calculation is the standard deviation of the sampling distribution. You know that the standard deviation of the sampling distribution equals the standard deviation of Pop 1 ÷ $\sqrt{n}$ (Property 3 of sampling distributions), which is the estimation error, previously estimated to be $18.14. The Z-score of 572.27 is therefore (572.27 - 600)/18.14 = -1.5287. Thus, 572.27 is 1.5287 Pop 3 standard deviations below the mean. The red shaded region therefore contains NORMSDIST(-1.5287) = 0.0632 fraction of all sample means.

Figure 2. The probability of getting a sample mean less than $572.27



If the true mean is $600 or more, it is therefore rather unlikely that our sample, shown in Table 1, could be a sample that just flukily had unrepresentatively too low rents. The probability is only a little more than 6% that this sample could have been produced by a Pop 1 having a true mean of $600. This analysis adds further confirmation that entering the Austin market would likely be a mistake.

It is important to note that the business decision depends crucially upon using the correct standard deviation. If the individual apartment variability had been used in the calculation

[7] Assume that the data were honestly and competently drawn as a true Random Sample.

instead of the group mean variability, then we would mistakenly compute that $572.27 is only $\left(\frac{572.27-600}{140.52}\right)$ = -0.1973 standard deviations below the mean, instead of -1.5287. Being off by 0.1973 standard deviations is much more likely than being off by 1.5287. Furthermore, the red region in Figure 2 mistakenly would be computed to be NORMSDIST(-0.1973) = 0.4218. If the red shaded region contains 42% of all possible sample means, then it is quite conceivable that the true mean could be $600 or more. You may well be tempted to spend more money to continue your analysis of the Austin market under those circumstances.

Sidebar. The above discussion illustrates how statistical thinking can help test the plausibility of business hypotheses. The landlord hypothesizes that Austin apartments average $600 or more in monthly rent. This hypothesis may not be what the landlord actually believes, but it is what he needs in order to justify entering the Austin market. To test the hypothesis, a random sample of data is analyzed. The sample mean is 1.5287 standard deviations below the hypothesized mean of $600. That is rather far and rather unlikely (0.0632) to be consistent with the hypothesis. The hypothesis can be rejected if it is too far away in terms of standard deviations or too unlikely in terms of probability. The landlord gets to decide how far is “too far” and how unlikely is “too unlikely”. Using the correct standard deviation is critical. An hypothesis that is implausible with an estimation (group mean) standard deviation may be plausible with a prediction (individual) standard deviation.

End sidebar.

The preceding material should give you a basic understanding of the differences between prediction and estimation. It is but a short step further to apply these ideas in regression. I will now discuss prediction and estimation in regression. The essence of the distinction between prediction and estimation remains the difference between predicting an *individual* value and estimating a *group* value. In regression, once again, the same number is used as both prediction and estimate, but the error measures continue to differ. Example 3 provides an illustrative setting. I have structured my discussion of Example 3 to parallel my discussion of Example 2 so that you can see clearly the essential sameness of the ideas.

Example 3. Prediction and estimation of Austin apartment rent for a given area.

A Random Sample of 60 apartments has been drawn from the population of all Austin apartments. Their rents and areas have been compiled and are shown in Table 2.

Table 2. Rent and Area of 60 Randomly Selected Austin Apartments

Rent	Area	Rent	Area	Rent	Area	Rent	Area
519	725	425	620	505	672	470	751
765	995	770	1040	445	660	480	608
475	481	445	520	565	755	460	900
575	925	510	880	650	810	600	860
415	600	635	832	515	611	575	925
530	668	470	545	470	705	659	944
580	725	700	921	470	564	650	940

995	1421	450	577	700	1250	750	1048
565	672	785	1080	455	512	455	474
620	1025	485	710	550	630	430	700
450	781	415	605	625	850	605	921
520	800	399	680	745	1156	929	1229
495	870	585	730	540	932	695	896
420	700	525	687	650	755	455	630
575	800	495	703	595	1093	1050	1864

You have two tasks:

- T-ind. Predict the rent of 123 Lotus Avenue, an Austin apartment that has 1000 square feet of area and is not among the sample of 60.
- T-mean. Estimate the mean rent of all Austin apartments that have 1000 square feet of area.

Since you now know one of the attributes of the apartments at issue, you should take that information into account in your tasks. 1000 square feet is larger than average (the mean area is 816.05 square feet). So the apartments at issue should command more than average rent. You run the regression using Table 2 data and find that the estimated regression equation is *Estimated mean rent* = 160.1871 + 0.5050 *Area*. You plug in the value 1000 sq ft for *Area* and calculate 160.1871 + 0.5050 x 1000 = $665.19.

The best solution for both tasks is the plug-in regression estimate, $665.19, although for somewhat different reasons.

For T-ind, nothing further is known about 123 Lotus Avenue, other than its area. It could be any Austin apartment with 1000 square feet of area. You do not want to guess too big or too small and thereby expose yourself to the risk of a big prediction error. Ideally, your guess should be the true mean rent of all Austin apartments with 1000 square feet because the average prediction error is a minimum there. Since you do not know this true mean rent for 1000 sq ft apartments, you use the regression estimate of that mean, $665.19, as the next best thing.

For T-mean, you use the regression estimate, $665.19, because the purpose of regression is to estimate the mean *Y* for given values of the *X*-variables. If the four assumptions of the regression model are correct, the regression estimate is the best such estimate there is.

Although the regression solutions for individual prediction (T-ind) and group estimation (T-mean) are the same number ($665.19), the error estimates are different. The intuition for why the error estimates are different is still the distinction between prediction error for *individual* apartments and estimation error for the mean of a *group* of apartments. So if you had a sample of 1000-square foot apartments, you could apply the same pre-regression procedure for the two tasks that I developed earlier in this Topic Note in Examples 1 and 2.

Unfortunately, there is not even a single apartment among the 60 sample apartments that has exactly 1000 square feet of area. However, L,H,I,N impose strong assumptions on the statistical model that let us use all 60 apartments anyway to estimate both the mean rent of all 1000 square foot apartments and also estimate individual and group uncertainty. The key to estimating the uncertainty is the 60 residual deviations from the regression line. By L,H,I,N, the 60 residuals are approximately a random sample from a normal distribution with mean 0 and standard deviation σ_e. This is a sample and this is a distribution of regression deviations for *individual* apartments. Therefore, the average magnitude of the residuals is a solution for T-ind. Their average magnitude can be assessed by calculating their standard deviation.[8]

For T-ind, you predict that the rent of 123 Lotus Ave, an individual 1000-square foot apartment, is the plug-in regression estimate of the mean rent of all 1000-square foot apartments, namely $665.19. You expect that the actual rent of 123 Lotus Ave will deviate from this prediction by the average magnitude of the regression deviations (residuals). The best estimate of this average individual deviation is the sample standard deviation of the 60 residuals, namely $68.86.[9,10] This number is part of the regression output shown in the Topic Note on Simple Linear Regression.

For T-mean, you estimate that the mean rent of all 1000-square foot apartments is the plug-in regression estimate of the mean rent of all 1000-square foot apartments, namely $665.19. You expect that the actual true mean rent of all 1000-square foot apartments will deviate from this estimate by the average deviation of all possible sample means from the true mean rent. The statistical theory (Property 3) tells you that this is the same as the average deviation of all individual 1000-square foot apartment rents from their true mean rent, divided by $\sqrt{60}$. That is, the expected deviation of your estimate $665.19 from the true mean rent of all 1000-square foot apartments is best estimated by $68.86 ÷ $\sqrt{60}$ = $8.89. [11]

Here are two scenarios that illustrate practical business applications of T-ind and T-mean in regression:

Business Scenario 3. You will be moving to Austin to pursue your MBA fulltime at UT. How much should you budget for rent? You have not yet rented an apartment, but you have decided that you want to rent a 1000-square foot apartment. You have obtained the Random Sample of 60 rents and areas shown in Table 2. You will "draw" your apartment from Population 1, consisting of all Austin apartments having 1000 square feet. Thus your question is a T-ind

[8] Provided you divide by *n* – 2, instead of *n* – 1, in the standard deviation.

[9] Provided you divide by *n* – 2, instead of *n* – 1, in the standard deviation.

[10] This is an approximation that is a little too small. There is an exact formula. Its value gets larger the farther *x* is from $\bar{x}$. For about 95% of *x* values, the formula is less than 250/n % larger. So with n = 60, the estimate given here is less than 4% too small.

[11] As noted in a footnote in the Topic Note on Simple Linear Regression, this formula is correct for an apartment of average size ($x = \bar{x}$). But the formula should be increased the farther *x* is from $\bar{x}$. For a better estimate, increase the formula by $(Z_x)^2/2$ times, where Z_x is the Z-score of *x*, that is, $Z_x = (x-\bar{x})/stdev(x)$. For *x* = 1000 sq ft, the formula should be increased by about $(Z_x)^2/2$ = [(1000-816.05)/243.24]^2/2 = 28.6% times. So $8.89*(1+.286) = $11.43 is a better estimate.

question – for one individual apartment. You run the regression of rent (Y) on area (X). You obtain the regression equation *Estimated mean rent* = 160.19 + 0.505* *Area*. You plug your intended apartment area of 1000 into this equation and calculate $665.69. Since the estimated mean is the best prediction of an individual apartment rent, you can expect to pay $665.19, more or less, for your apartment. This gives you a starting point for budgeting. How much more or less? The typical 1000-square foot apartment could easily rent for $68.86 more or less than the $665.19 prediction. The ±$68.86 reflects the cost of more or fewer amenities and attractions than the average 1000-square foot apartment. So if you were hoping to rent a 1000-square foot apartment with better than average amenities for an apartment of that size, you should expect to pay more than $665.19. A 1000-square foot apartment with average high-side amenities could be expected to cost about $70 more. So if paying more than about $735 (= $665.19 + $68.86) would crimp your budget, you may need to revise your plans. On the other hand, if you know that you must really economize on rent, then you might settle for a 1000-square foot apartment with average low-side amenities. You could expect to save about $70 on the rent of such an apartment and pay about $595 (= $665.19 - $68.86).

Business Scenario 4. You are an apartment landlord. You are considering entering the Austin apartment market. You specialize in acquiring apartments with 1000 square feet of space. From experience, you know that your apartments need to average $700 or more rent for the financial numbers to work for you. You have the data in Table 2 available. Do you enter the Austin market? To be sure, you will no doubt do a quite detailed analysis before deciding. But even the simple data in Table 2 can help inform your business judgment. You will not acquire a single apartment but a group. So you do not have a T-ind question. Your apartment rents will vary, although your apartments will all have 1000 square feet. Some will have high rent, others low. You will get more than $700 on some, less on others. You will lose money on some and make money on others. As long as the average rent exceeds $700, you will do OK. Your question is a T-mean (Population 3) question. What is the mean rent of all Austin apartments having 1000 square feet? Since your apartments will reflect the population of Austin apartments having 1000 square feet, if the true mean of that Austin sub-population exceeds $700, yours will probably also. The best estimate of the true mean of 1000-square foot Austin apartment rents is $665.19, which is less than your $700 requirement. That shortfall does not look good. On that alone, you could justify canceling plans to enter Austin. But you might reason that the 60 apartments in Table 2 are only a sample. They are not the whole population. The population could still have a true mean that exceeds $700. You are wondering: "What if the sample was a fluke? The sample is somewhat close to my cutoff. Maybe I should take a larger sample and/or do a more refined analysis. I don't want to miss out on a good business opportunity because of unrepresentative data."

The following is one approach to addressing your doubts: Hypothetically, what if the true mean rent of 1000-square foot Austin apartments were $700 – how likely is it that you would get a sample like Table 2 with a regression estimated mean for 1000-square foot apartments of $665.19 or less? If you knew that, you could judge whether Table 2 could be a fluky sample of unrepresentatively low rents. So the question is, What is the probability of getting a sample regression estimate of the mean of 1000-square foot apartments of $665.19 or less if the true mean is $700?[12] The question is now posed in a familiar statistical format that you can solve.

[12] Assume that the data were honestly and competently drawn as a true Random Sample.

Figure 7. The probability of getting a sample regression mean for 1000-sq ft apartments that is less than $665.19

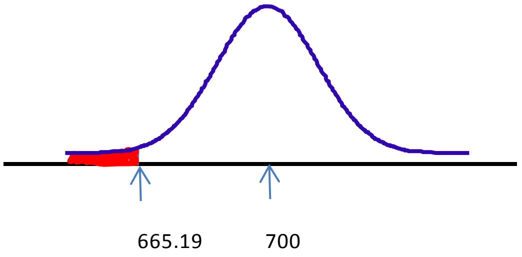


Figure 7 shows how to solve it. Figure 7 shows the population of all possible regression estimates of the mean of 1000-square foot apartments as a normal curve. You know that the shape is normal because of the Central Limit Theorem (Property 1 of sampling distributions) since $n = 60$ is sufficiently large. You also know that the true mean of the sampling distribution in Figure 7 is $700 *if* (as we are hypothesizing) the true mean rent of apartments with 1000 square feet is $700. That is Property 2 of sampling distributions. The red shaded region shows all of the regression mean estimates that are less than $665.19. How many are there? That is an easy question to answer. All you need to finish the calculation is the standard deviation of the sampling distribution. You know that the standard deviation of the sampling distribution equals the standard deviation of Pop 1 ÷ $\sqrt{n}$ (Property 3 of sampling distributions), which is the estimation error, previously estimated to be $68.86 ÷ $\sqrt{60}$ = $8.89. The Z-score of 665.19 is therefore (665.19 - 700)/8.89 = -3.9156. Thus, 665.19 is 3.9156 Pop 3 standard deviations below the mean. The red shaded region therefore contains NORMSDIST(-3.9156) = 0.000045 fraction of all sample means.

If the true mean is $700 or more, it is therefore highly unlikely that our sample in Table 2 just flukily has unrepresentatively too low rents. The probability is only a little more than 0.0045% that this sample could have been produced by a Pop 1 with a true mean of $700 for all 1000-square foot apartments. This analysis adds strong confirmation that entering the Austin market would likely be a mistake.

It is important to note that the business decision depends crucially upon using the correct standard deviation. If the individual variability of 1000-square foot apartments had been used in the calculation instead of the group variability, then we would mistakenly compute that $665.19 is only $\left(\dfrac{665.19-700}{68.86}\right)$ = -0.5055 standard deviations below the mean, instead of -3.9156. Being off by 0.5055 standard deviations is much more likely than being off by 3.9156. Furthermore, the red region in Figure 2 mistakenly would be computed to be NORMSDIST(-0.5055) = 0.3066. If the red shaded region contains 30% of all possible sample means, then it is quite conceivable that the true mean could be $700 or more. You may well be tempted to spend more money to continue your analysis of the Austin market under those circumstances.

Comment: Just as noted in the Sidebar at the end of Example 2, the preceding sort of discussion illustrates how statistical thinking can help test the plausibility of a business hypothesis. The landlord wants to know if the mean rent of 1000-square foot apartments is $700. The hypothesis can be rejected if it is too far away in terms of standard deviations or too unlikely in terms of probability.

SUMMARY

It is crucially important from a business standpoint to distinguish between the *prediction* of individual values and the *estimation* of group means. Although the numerical value of a parallel prediction and estimation problem are the same, the sizes of their uncertainties are often an order of magnitude different.

Suppose a Random Sample has already been drawn from Population 1.

The **prediction problem** is to predict the next individual value to be drawn from Pop 1.

- The best prediction is the sample mean of the Random Sample already drawn from Pop 1.
- The uncertainty of the prediction can be summarized by reporting the ±deviation by which the actual next individual value can be expected to differ from the sample mean estimate. This ±deviation is the sample standard deviation of the Random Sample already drawn from Pop 1. If Pop 1 is normally distributed, there is about a 68% chance that the actual next individual value will differ from the sample mean estimate by less than the ±sample standard deviation.
- The prediction problem is a Pop 1 issue.

The **estimation problem** is to estimate the true mean of Pop 1.

- The best estimate is the sample mean of the Random Sample already drawn from Pop 1.
- The uncertainty of the estimation can be summarized by reporting the ±deviation by which the true mean can be expected to differ from the sample mean estimate. This ±deviation is the standard deviation of Pop 3 (the sampling distribution of the sample mean), best estimated by the standard error, which is the sample standard deviation of the Random Sample already drawn from Pop 1, divided by $\sqrt{n}$. If n is sufficiently large or if Pop 1 is normally distributed, there is about a 68% chance that the true mean will differ from the sample mean estimate by less than the ±standard error.
- The estimation problem is a Pop 3 issue.

Application of prediction and estimation to regression:

Suppose that a dataset of Y's and X's satisfy the four regression assumptions (L,H,I,N).

The **prediction problem in regression** is to predict the next individual Y to be drawn at a given $X = x$. In this problem, Pop 1 becomes the sub-population of all Y's at the given x.

- The best prediction is the plug-in regression estimate *Estimated mean* $Y = a + b*x$.
- The uncertainty of the prediction can be summarized by reporting the ±deviation by which the actual next individual Y at $X = x$ can be expected to differ from $a + b*x$. This ±deviation is the root-mean-square-error of the regression, called the "standard error of the estimate" by StatTools. It is the standard deviation of the residuals. If the "N" assumption of regression is correct, then there is about a 68% chance that the actual next individual Y will differ from $a + b*x$ by less than the ±root-mean-square-error.

- The prediction problem in regression is a Pop 1 issue, where Pop 1 is the sub-population consisting of all *Y*'s at the given *X* = *x*.

The **estimation problem in regression** is to estimate the true mean *Y* at a given *X* = *x*. In this problem, Pop 1 becomes the sub-population of all *Y*'s at the given *x*. and Pop 3 becomes the set of all possible regression estimates *a* + *b***x*, one for each possible sample of *Y*'s and *X*'s.

- The best estimate is the plug-in regression estimate *Estimated mean Y* = *a* + *b***x*.
- The uncertainty of the estimation can be summarized by reporting the ±deviation by which the true mean *Y* at *X* = *x* can be expected to differ from *a* + *b***x*. This ±deviation is the root-mean-square-error of the regression, divided by $\sqrt{n}$.[13] The root-mean-square-error is called the "standard error of the estimate" by StatTools and just "standard error" by Excel. If *n* is sufficiently large or if the "N" assumption of regression is correct, then there is about a 68% chance that the true mean *Y* at *X* = *x* will differ from *a* + *b***x* by less than the ±root-mean-square-error, divided by $\sqrt{n}$.
- The estimation problem in regression is a Pop 3 issue, where Pop 3 is all possible regression estimates *a* + *b***x*, one for each possible sample of *Y*'s and *X*'s.

[13] As noted in footnote 11, this formula for the error deviation may be too small.