The t Distribution

As we've seen, in interval estimation for a population mean **when we do not know the standard deviation of the population** (which is all cases for us and almost all cases in real practice), we must use the appropriate t distribution to obtain the correct multiplier. Why is this; that is, how does the t distribution arise? To see the answer, we must look at two related problems involving the sampling distribution of sample means.

■ The Sampling Distribution of Sample Means

- **Here is the first type of problem.** Recall that IQ's ~ N(100, 15) (the ~ is read "are distributed"). What is the probability that, if we randomly select 36 people, **their average IQ** will be greater than 115?
- We know by the empirical rule that the probability of getting 1 random person with an IQ greater than 115 is approximately 0.16. What about the probability of getting 36 random people with an average IQ greater than 115? To find a probability about a sampling situation, we need to know the ***sampling distribution of sample means*** for the sampling situation.
- In this problem, **since the population distribution of IQ's is itself normal**, it can be shown that **the sampling distribution we must use is also normal.** We can show this by showing first that **the sums** of independent draws from a normal distribution have a normal distribution. Though this seems to be somewhat obvious, it actually requires some fairly sophisticated mathematics to show. Then, once we show that the sums of our 36 draws are normally distributed, it follows that **the averages** are as well.

- It can also be shown using some basic random variable theory that the sampling distribution we are looking for will have mean of 100 and its standard deviation will be ($\sigma/\sqrt{n}$) (15/6 = 2.5, in our problem). That is, **in general, if a population is distributed N(μ, σ) and we take random samples of size n from it, the means of these samples will be distributed N(μ, $\sigma/\sqrt{n}$). See Chapter 3, Appendix 2 – From Population Distribution to Sampling Distribution.** ***Notice that this is true no matter what the sample size n is.***
- To come back to our problem, if we randomly select 36 people from N(100, 15), the probability that their average IQ is greater than 115 comes from N(100, 2.5). Using the empirical rule and our basic knowledge of the normal distribution, we see that this probability is essentially 0. **(Can you see how we get this probability? Hint: Consider the z-score we would use if solving the problem using a z-score table.)**

- **Now, consider a second problem**: Suppose we are studying the IQ's of the Purdue undergrad population. We believe from years of past data that the average Purdue undergrad IQ is 115 and that the standard deviation is 15, the same standard deviation as the population at large. However, **we know that the distribution of Purdue undergrad IQ's is not normal** since it does not spread out symmetrically below 115 as a normal distribution would. We write Purdue **IQ's ~ ?(115, 15)** to indicate that we do not know the shape of the distribution.

--- Now, what is the probability that, if we randomly select 36 Purdue undergrads, their average IQ will be greater than or equal to 120?

--- It is clear that, again, we need to have the sampling distribution of sample means in this sampling situation. But this time **we do not have a normal population distribution.** In the last problem, having a normal population distribution ***seemed*** necessary for our solution. It turns out that because of a classic mathematical result known as the ***Central Limit Theorem*** we can calculate the sampling distribution for means in this situation **even though the population distribution is not normal.**

- **The Central Limit Theorem (CLT)** says:

--- ***Suppose a population has mean = μ and standard deviation = σ; that is, the population is distributed ?(μ, σ). For random samples*** **with fixed size n = 30 or more,** ***the sampling distribution of sample means is N(μ, (σ/√n))*** **– no matter what the shape of the population distribution is.**

> **Remember: The sampling distribution gives the distribution of the results of** ***many repeated random samplings*** **of the same size** ***from the same population. See the handout: From Population Distribution to Sampling Distribution – Sample Means, at the end of this chapter.***

- Now, applying the CLT to our problem, we see that the sampling distribution is N(115, 2.5) and, using **the empirical rule,** the probability that a sample of 36 students will have an average IQ greater than 120 is approximately 0.025.
- You see that the CLT is quite a remarkable result. ***If we take a sample of size 30 or more from any population, no matter how oddly shaped, the sampling distribution of sample means in that situation will be normal.*** The importance of the CLT in statistics cannot be overstated.

QUESTION 1: A population has mean = 26 and standard deviation of 8. Determine if the CLT can be applied and, if so, give the sampling distribution of sample means for samples of size:

a. **36**
b. **64**
c. **25**

QUESTION 2: We are operating a large catfish farm. We believe our catfish have an average weight of 8.8 lbs. Universally accepted scientific publications indicate that the population standard deviation of farmed catfish weights is 1.58 lbs. To test whether our assumption about our catfish's average weight is correct, we randomly sample 36 catfish from our ponds. If our assumption is about their average weight is correct, what is the probability that the sample mean of these 36 fish is 8.4 lbs or less? (Hint: Use the CLT to get the sampling distribution for means for this sampling situation.)

■ The t Distribution

- In the problems above, we used the normal distribution to reach our solution. So how does the t distribution come into the picture?
- The answer is that the t distribution arises as a sampling distribution in this type of problem **when we do not know the standard deviation of the population from which we are sampling.** Below, you will see how this happens.

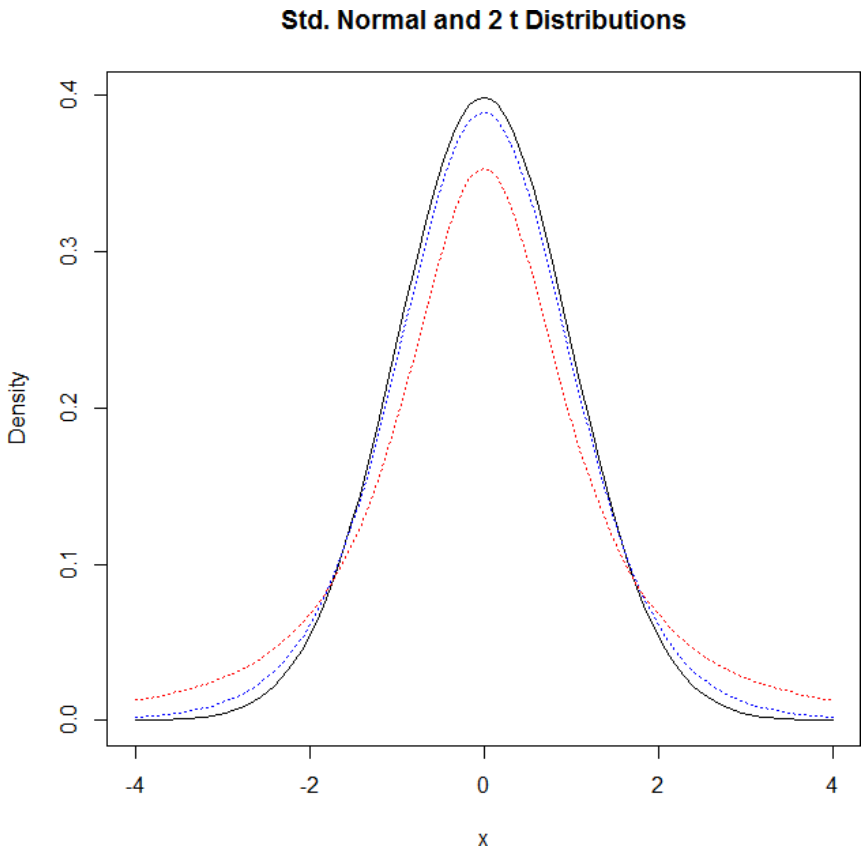
- The t distribution was discovered in the early 20th century by a statistician named Gosset who worked for the Guinness Brewing Co. in Dublin, Ireland. He noticed that when he worked problems like Question 2 above in the real world – that is, without assuming that he knew the population standard deviation - his probabilities were a little off.
- Let's go through the process using our catfish farm in Question 2, but this time without assuming that we know the population standard deviation of our catfish weights.
- By the CLT, the sampling distribution of sample means is N(8.8, σ/6). **But, we do not know σ!** What should we do? A reasonable approach would be to use the ***sample* standard deviation** to estimate σ. Suppose our sample standard deviation is 2.4. We want to use this value for σ. Before Gosset's derivation of the t distribution, that is what was done. The sampling distribution to use would then be N(8.8, 0.4) and we could get an answer – **an incorrect answer.** Because ***it is the use of the sample standard deviation to estimate σ that gives rise to the t distribution as the sampling distribution for this sampling situation.*** This is what Gosset discovered by closely considering the following type of thought experiment:
- Compare two sampling situations. In the first, we take random samples of size 36 from ?(μ, σ). By the CLT, we know that if we take many such samples (all of size 36), the sample means will be distributed N(μ, σ/6). Suppose we take many such samples and record the sample means for each sample. We have many sample means. If we standardize these (by subtracting μ and dividing by σ/6), they will have a standard normal distribution.
- Next, suppose we again take many random samples of size 36, but this time from ?(μ, ?). Notice the difference. Here **we do not know σ,** the population standard deviation. We

want to standardize our sample means again. So, we have to record our sample means **and our sample standard deviations** each time. By the CLT, we know that the sample means are normally distributed with mean μ and **some** standard deviation. But the standardized means in this case will not have the standard normal distribution. **Why not?**

- Because, when we standardize we will subtract μ, **but we will *not* divide by σ/6.** ***Remember, we do not know σ.*** So, to standardize **we will divide by: (sample standard deviation/6).** Thus, because each sample has a different ***sample*** standard deviation, **we are dividing by a different standard deviation each time we standardize.** The resulting means will therefore **not have the standard normal distribution.** Their distribution will be slightly different - **they will be t distributed.**
- This is what Gosset discovered and he was able to work out the mathematics that this estimation of the population standard deviation led to – **the formula for the family of t distributions**. (**Note therefore that the t distributions we will work with are standardized distributions.)** Guinness's policy required Gosset to publish anonymously, so he published his paper under the name "Student." So, the t distribution is sometimes called **"Student's t."**

The family of t distributions (revisited)

- The t distributions are **a family of distributions,** similar to the standard normal.
- They are all symmetric and all have mean 0 with a shape determined by what is known as **degrees of freedom (df).** These degrees of freedom determine the standard deviation and precise shape of the particular t distribution.



- The graph above shows the standard normal (on top) with 2 t distributions. The lowest curve is the t with 2 df; the middle curve is the t with 10 df.

- Notice that the higher the df's, the closer the t is to the standard normal (which, mathematically, can be shown to be the t with ∞ df).

- Notice that the t distributions have what are called ***heavy tails.*** There is more probability in the tails of t distributions; more in the bell of the standard normal. **However, the t distributions are symmetric and mound-shaped, so the empirical rule does apply fairly well to them as long as the degrees of freedom are more than 5.**

- We must use software to compute probabilities from the t distributions. We can use Minitab or, as with the normal, an online calculator such as Stat Trek.

- **Remember, we must always standardize based on the CLT estimate of the sampling distribution when computing t distribution probabilities.**

- Here is the process:

--- First, we obtain the standardized score, called a **t score** in this type of problem. It uses the same general formula as the z-score:

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

--- x-bar is the sample mean

--- μ-zero is the assumed population mean

--- s is the sample standard deviation

--- n is the sample size

- **Notice, we use the standard error (s/√n) to get our t score since we are standardizing based on the CLT estimate of the sampling distribution.**

- The t-score is then evaluated using the t distribution with **n - 1 df.** (Remember, n is the sample size.) This determination of degrees of freedom is beyond the scope of our course. You need only to remember the basic rule: in these types of problems, **df = n – 1.**

- **Now, let's consider our catfish farm from in Question 2 from this real-world point of view.**

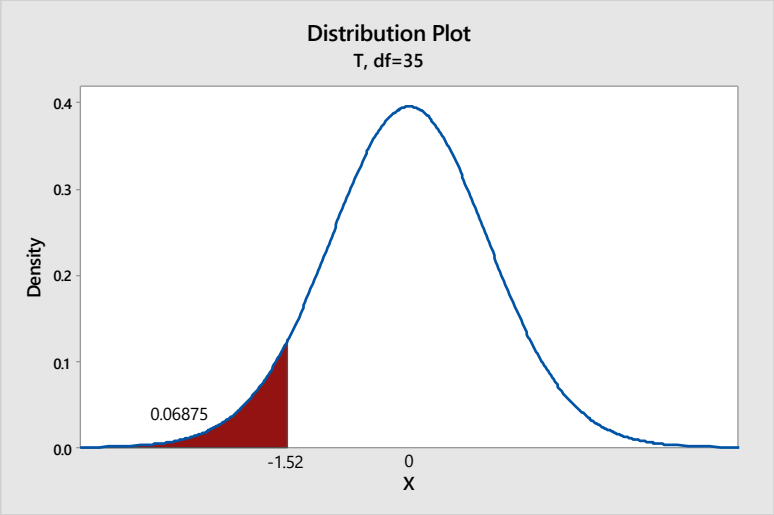
--- We want to know: If the average weight of all of our catfish (that is, the population mean) is 8.8 lbs,, what is the probability of getting a sample of 36 with a sample mean of 8.4 lbs or less? Notice: ***we do not claim to know and are not assuming that we know σ, the population standard deviation.***

We proceed as follows:

--- We will compute the **sample standard deviation** – denoted **s** – and use it to estimate σ. For our sample, suppose s = 1.58.

- Then, for our problem, t = (-0.4)/(1.58/6) = -1.52. It is evaluated on the t with 35 df – often written as: t_{35}. Using EXCEL:

  = t.dist(-1.52, 35, TRUE) = 0.068747



- A question that suggests itself is, **"What would the answer be if we just used the standard normal?"** It would be:

  = norm.s.dist(-1.52, TRUE) = 0.064355

- This is a probability only approximately 0.0045 lower than that obtained using the t (showing us, as stated above, that tail probabilities of the t distributions are greater than in the standard normal.) Notice that in general (that is, when degrees of freedom are around 20 or more), the t distributions are very close to the standard normal. It is quite remarkable that Gosset, in an era before computers or even calculators, detected them. The reason that he could is important to us. **In business applications which often involve estimations based on mass production processes, the extra precision obtained by using the t distributions gives far better estimates. Not using it can throw off the bottom line.**

QUESTION 3: Give an interpretation of the probability just computed: 0.0687. (Hint: it is a conditional probability.) Does our assumption that the population mean of our catfish is 8.8 lbs. appear to be sound?

SAMPLING DISTRIBUTIONS OF SAMPLE MEANS – A QUICK SUMMARY FOR THE SITUATIONS WE WILL ENCOUNTER

- The key steps are to realize that a) the sampling distribution of sample means, like proportions, **always depends on sample size** and b) to always **assess the population**

distribution.

- **Case 1:** If we know:

  --- the population is normally distributed;

  --- the mean of the population (μ); and,

  --- the standard deviation of the population (σ),

  then, the sampling distribution of sample means is **N(μ, (σ/√n)) no matter what the sample size is.**

- **Case 2:** If we know:

  --- the mean of the population (μ);

  --- the standard deviation of the population (σ);

  but, we do not know the shape of the population distribution (or we know that it is not normal), then the sampling distribution of sample means is **N(μ, (σ/√n)), as long as the sample size, n ≥ 30. If n < 30, we cannot reliably compute the sampling distribution.**

- **Case 3: If we do not know the shape of the population distribution nor the population standard deviation, the sampling distribution of sample means is t_{n-1}, as long as n ≥ 30. (**If n < 30, we cannot reliably compute the sampling distribution.**)**

- **Our interval estimation problems will always involve Case 3.** This is because, we will be estimating sample means in situations in which we will not know the population mean. It generally follows from this that we will not know the population standard deviation as well. Remember, it is because of this that we use the t distribution as the sampling distribution. **Thus, our multipliers for CI's for means must come from the appropriate t distribution.**