Problem 1. A population has a mean of 50 and a standard deviation of 6. What are the mean and standard deviation of the sampling distribution of the mean for $n = 16$?

Answer 1. By population mean and standard deviation, we are talking about the mean and standard deviation of a single draw, X_i, where $\mu = 50$ and $\sigma = 6$. The sampling distribution refers to $\bar{X}$. The mean of $\bar{X}$ is also $\mu = 50$, but the standard deviation of $\bar{X}$ is $6/\sqrt{16} = 1.5$. The latter number is called the *standard error*, denoted $\text{se}(\bar{X})$

If we know σ and if $n > 30$, then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ approximately. For known σ, if X_i is normally distributed, or if $n \to \infty$, then $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ exactly.

Problem 2. Given a test that is normally distributed with mean $\mu = 100$ and standard deviation of $\sigma = 12$. Find the following:

(a) the probability that a single score drawn at random will be less than 120
(b) the probability that a single score drawn at random will be greater than 123
(c) the probability that a sample of 25 scores will have a mean less than 106
(d) the probability that the mean of a sample of 36 scores will be either less than 95 or greater than 105
(e) the test score such that the probability of scoring above it is 5%

Answer 2.

(a) Let X denote a random test score. We want to find $P(X < 120)$. We first need to standardize the test score so that it has mean 0 and standard deviation 1, and accordingly we instead find

$$P\left(\frac{X - 100}{12} < \frac{120 - 100}{12}\right).$$

Let $Z \equiv (X - 100)/12$. Since the test is normally distributed, we also know that Z is normally distributed; and since we've standardized it, it is standard normally distributed. Hence we are to find $P(Z < 1.67)$ for $Z \sim \mathcal{N}(0, 1)$.

To solve this, we need to either appeal to a normal distribution table, or use R. To solve it with R, use the command `pnorm(1.67)`, which gives approximately 0.953. Using the normal table we are provided with, 1.67 is closest to 1.645, so we would use approximately 0.95.

(b) We set the problem up analogously and arrive at standardized probability $P(Z > 1.92)$. The problem is, `pnorm(1.92)` tells us the probability of Z being *below* 1.92, whereas

we are now trying to find the probability of Z being *above* 1.92. We can exploit the symmetry of the normal distribution to solve this: the probability of being above 1.92 is the same as the probability of being below -1.92.

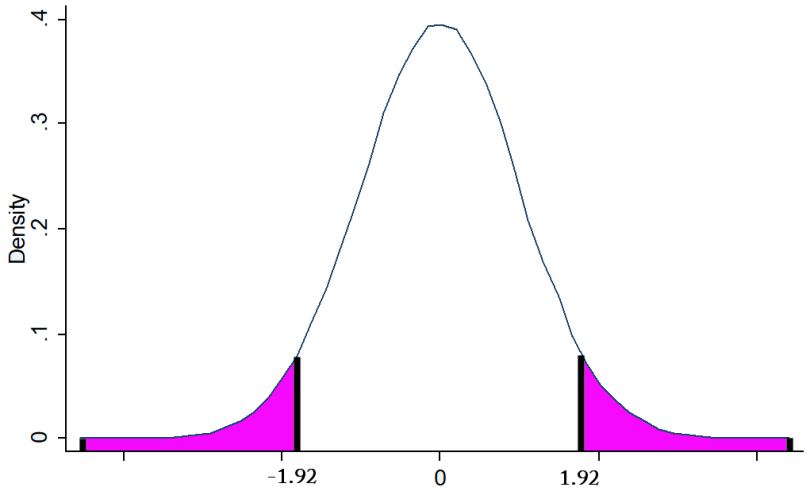

FIGURE 1: The probability of being above 1.92 is the same as the probability of being below -1.92.

Hence the problem can be solved with `pnorm(-1.92)`, which gives about 0.027.

Another problem: -1.92 is not a number that appears on the normal table. What we can do instead is recognize that the probability of Z being above 1.92 is the complementary probability of Z being below 1.92. That is, $P(Z > 1.92) = 1 - P(Z < 1.92)$.

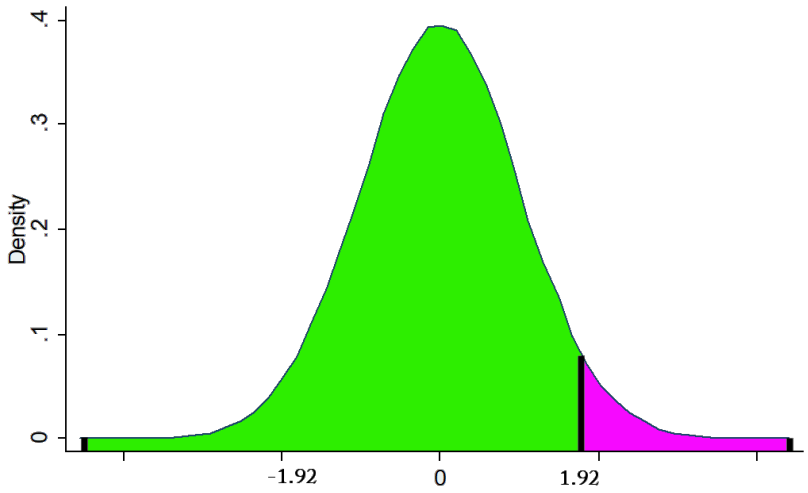

FIGURE 2: The area under the entire curve is 1. Hence, 1 minus the green area gives us the purple area. The green area is $P(Z < 1.92)$. Hence $P(Z > 1.92) = 1 - P(Z < 1.92)$.

Using the normal table, 1.92 is reasonably close to 1.96, so the answer is approximately $1 - 0.9750 = 0.025$.

(c) Now we are dealing with a sampling distribution, so we appeal to the central limit theorem, which tells us that

$$Z \equiv \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

We want to solve $P(\bar{X} < 106)$. We conform it to central limit theorem form by using

$$\begin{aligned} P(\bar{X} < 106) &= P\left(\frac{\bar{X} - 100}{12/\sqrt{25}} < \frac{106 - 100}{12/\sqrt{25}}\right) \\ &= P(Z < 2.50). \end{aligned}$$

Using R, this gives `pnorm(2.50)` ≈ 0.994. Using the normal table, the closest we have is 2.576, which gives probability 0.995.

(Note that if we did not know the population standard deviation σ, then we'd have to use estimate s instead of σ, and the $T(n-1)$ distribution instead of the standard normal. Also note that since $n \leq 30$, this problem only has a "good" answer because the underlying distribution is normal.)

(d) We want to find $P(\bar{X} < 95) + P(\bar{X} > 105)$. We first standardize each with respect to the central limit theorem, which gives

$$\begin{aligned} P\left(\bar{X} < 95\right) + P\left(\bar{X} > 105\right) &= P\left(\frac{\bar{X} - 100}{12/\sqrt{36}} < \frac{95 - 100}{12/\sqrt{36}}\right) + P\left(\frac{\bar{X} - 100}{12/\sqrt{36}} > \frac{105 - 100}{12/\sqrt{36}}\right) \\ &= P(Z < -2.50) + P(Z > 2.50). \end{aligned}$$

Since the normal distribution is symmetric, we know $P(Z < -2.50) = P(Z > 2.50)$. Hence we can instead find either $2 \times P(Z > 2.50)$ or $2 \times P(Z < -2.50)$. The latter we can do in R with command `2*pnorm(-2.50)`, which gives about 0.012.

But -2.50 isn't on the normal table. From the argument used in part (b), we know that $P(Z > 2.50) = 1 - P(Z < 2.50)$. From the normal table, 2.50 is close to 2.576, so we can conclude approximately that

$$P(Z > 2.50) = 1 - P(Z < 2.50) = 1 - 0.995 = 0.005.$$

Hence the answer is approximately $2 \times 0.005 = 0.01$.

(e) First standardize the test score X into Z in the usual way. We are looking for the value of Z that makes the right tail consist of 5% of the area under the curve. Which is another way of saying, we want the value of Z such that the area below that number is 0.95. According to our normal table, that number is 1.645. Using R, we find the number by using command `qnorm(0.95)`, which gives the same number.

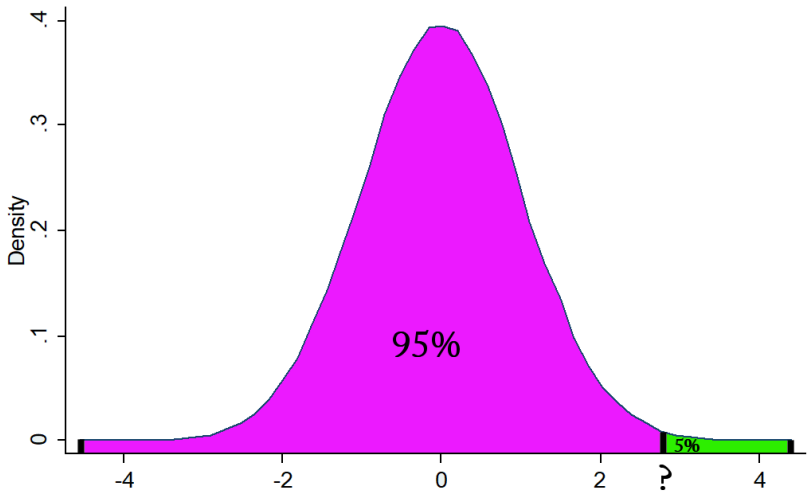

FIGURE 3: We want to find the value of Z such that the area underneath the curve above the value is 0.05.

But this is not a test score. To convert it back into a test score, we have to unstandardize it. So multiply it by the standard deviation and then add the mean back, and you get

$$1.645 \times 12 + 100 \approx 120.$$

Thus we conclude there is a 5% chance that someone receives a score above 120. Note that this is completely consistent with part (a), where we found the probability of being below a score of 120 is 0.95.

Problem 3. In Wisconsin, the mean donut consumption in a week is 48 donuts per person, and the standard deviation of weekly donut consumption is 12 donuts.[1] This week, Jiminy Glick has a weekly donut Z-score of 1.5. How many donuts did Jiminy Glick eat this week?

Answer 3. The Z-score tells us how many standard deviations from the mean. Since Jiminy's Z-score is 1.5, that means he ate $1.5 \times 12 = 18$ donuts more than the mean. Hence, he ate $48 + 18 = 66$ donuts this week.

[1] I miss Wisconsin.

Problem 4. On average, I eat 7 pizzas per week, with a standard deviation of 1 pizza, and my pizza consumption is normally distributed. What is the probability that I eat less than 5 pizzas in a given week? Don't use R or a normal table.

Answer 4. For any normal distribution, it is approximately true that

- 68% of the data lies within one standard deviation of the mean
- 95% of the data lies within two standard deviations of the mean
- 99.7% of the data lies within three standard deviations of the mean.

5 pizzas is two standard deviations less than the mean of 7, so let's consider the second bullet more closely. Since 95% of the data is within two standard deviations of the mean, that means the two tails outside of that must comprise the remaining 5% of the data. Since the normal distribution is symmetric, that means each tail gets 2.5%. Thus the probability of me consuming fewer than 5 pizzas in a week is approximately 2.5%.

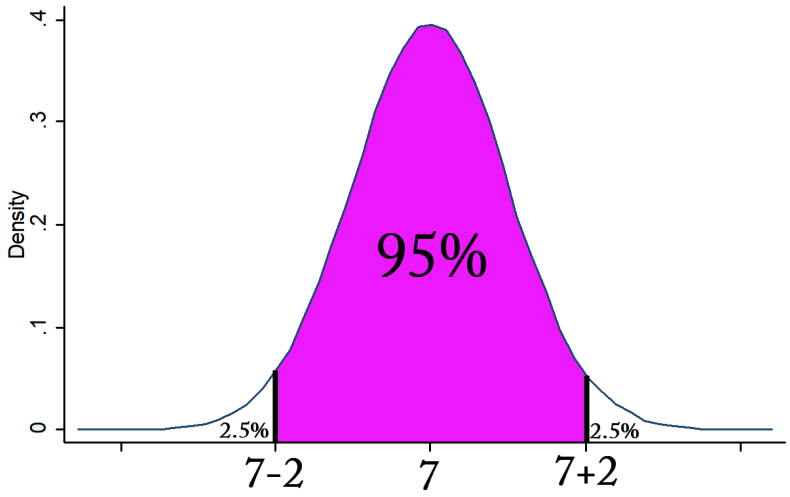

FIGURE 4: 95% of the data is found within ±2 standard deviations of the mean.

Problem 5. Explain when you should use the standard normal distribution and when you should use the $T(n-1)$ distribution for analyzing sampling means.

Answer 5. There are a lot of cases to consider. The most important cases are:

(a) If σ is known and $n > 30$, then use the standard normal distribution. (Approximation)
(b) If σ is known and the underlying distribution is normal, then use the standard normal distribution for any n. (Exact)
(c) If σ is not known and $n > 30$, then use the $T(n-1)$ distribution. (Approximation)

(d) If σ is not known and the underlying distribution is normal, then use the $T(n-1)$ distribution regardless of n. (Exact)

The approximate cases become exact as $n \to \infty$. On paper, if $n > 30$, then you can usually use the normal distribution instead of $T(n-1)$ because they will be very similar. (In fact, $T(\infty)$ is exactly standard normal.) If you're using R, then just use $T(n-1)$ anyway. In practice, we will usually have unknown σ and $n > 30$, so the $T(n-1)$ statistic is used heavily.

Note that if $n \leq 30$, then we can only do "reliable" inference if we have reason to believe that the underlying data is normally distributed. Accordingly, you should be skeptical of inference on small sample sizes.

Problem 6. Explain the difference between time series data, panel data, and cross-sectional data.

Answer 6. Cross-sectional data looks at observations of a number of individuals (which could be people, firms, countries, etc, usually denoted n) at a given point in time. For example, we might observe the GPA of each person in class at the beginning of the quarter.

Time series data looks at one piece of data over time, usually denoted t. For example, you might track your own GPA each quarter over your college career. You're looking at one piece of data, your own GPA, and you're looking at it for several terms.

Panel data is both. We look at a number of different individuals, n, and we see how observations about them change over time, t. For example, we observe each person's GPA at the end of the Fall quarter (there are n people, time $t = 1$). Then we look at the same people's GPAs at the end of the Winter quarter (now $t = 2$). Then we look at the same people's GPAs again at the end of the Spring quarter (now $t = 3$). So we've tracked the same n people's GPAs over $t = 3$ different periods.

You can think of time series as being panel data where $n = 1$, because we only look at one piece of data over time. Similarly, you can think of cross-sectional data as being panel data where $t = 1$, because we look at those n individuals in only one period. In practice, panel data tends to contain smaller time horizons than time series.

Student	F2018	W2019	S2019
Zarnold's GPA	3.00	3.10	3.88
Engelbert's GPA	2.85	3.35	2.95
Schtolteheim's GPA	3.78	2.50	1.90

FIGURE 5: The entire spreadsheet is panel data because we observe several individuals ($n > 1$) over time ($t > 1$). The green row is the time series of Zarnold's GPA. The red column is panel data for GPA of Fall 2018.