Sampling Distributions and Simulation

OPRE 6301

Basic Concepts...

We are often interested in calculating some properties, i.e., parameters, of a population. For very large populations, the exact calculation of a parameter is typically prohibitive.

A more economical/sensible approach is to take a random sample from the population of interest, calculate a **statistic** related to the **parameter** of interest, and then make an inference about the parameter based on the value of the statistic. This is called **statistical inference**.

Any statistic is a random variable. The distribution of a statistic is a **sampling distribution**. The sampling distribution helps us understand how close is a statistic to its corresponding population parameter.

Typical parameters of interest include:

- Mean
- Proportion
- Variance

Sample Mean...

The standard statistic that is used to infer about the population mean is the **sample mean**.

Rolling a Die...

Suppose a fair die is rolled an infinite number of times. "Imagine" a population that is consisted of such sequences of outcomes.

Let X be the outcome of a single roll. Then, the probability mass function of X is:

$X = x$	$P(x)$
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

Rolling the die once can be viewed as one "draw" from an infinite population with mean $\mu = E(X) = 3.5$ and variance $\sigma^2 = V(X) = 2.92$. Let us pretend that we do **not** know the population mean and would like to estimate that from a sample.

Each draw can be viewed as a sample of size 1. Suppose we simply used the outcome of a *single* roll, i.e., a realized value of X, as an estimate of μ. How good is this estimate?

Our estimate could assume any of the values 1, 2, ..., 6 with probability 1/6. If we used this approach **repeatedly**, then the average deviation of our estimates from μ would be $E(X - \mu) = 0$. This certainly is desirable, but note that the variance of the deviation $X - \mu$ is σ^2 and, in particular, that this "gap" between our estimate and μ is *never* less than 0.5!

Can we improve this situation? The **key concept** is that we need to increase the sample size.

Consider now a sample of size 2. Denote the outcomes of two independent rolls by X_1 and X_2. Let the average of the two outcomes be $\bar{X}$ ("X bar"), i.e.,

$$\bar{X} = \frac{X_1 + X_2}{2}.$$

How good is $\bar{X}$ as an estimate of μ?

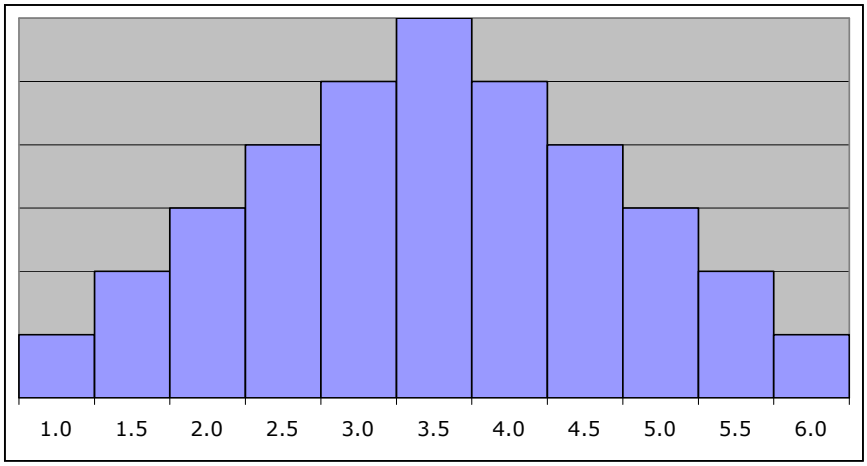
Terminology: A *method* for estimating a parameter of a population is called an **estimator**. Here, $\bar{X}$ is an estimator for μ and, for this reason, it is often denoted as $\hat{\mu}$ ("μ hat"). Note that $\hat{\mu}$ is a function of the observations X_1 and X_2.

Let us look at the *sampling* distribution of $\bar{X}$. There are 36 possible pairs of (X_1, X_2): (1, 1), (1, 2), (1, 3), ..., (6, 4), (6, 5), (6, 6). The corresponding values of $\bar{X}$ for these pairs are: 1, 1.5, 2, ..., 5, 5.5, 6. Using the fact that each pair has probability 1/36, we obtain (there are only 11 possible values for $\bar{X}$, since some values, 3.5 for example, occur in more than one way):

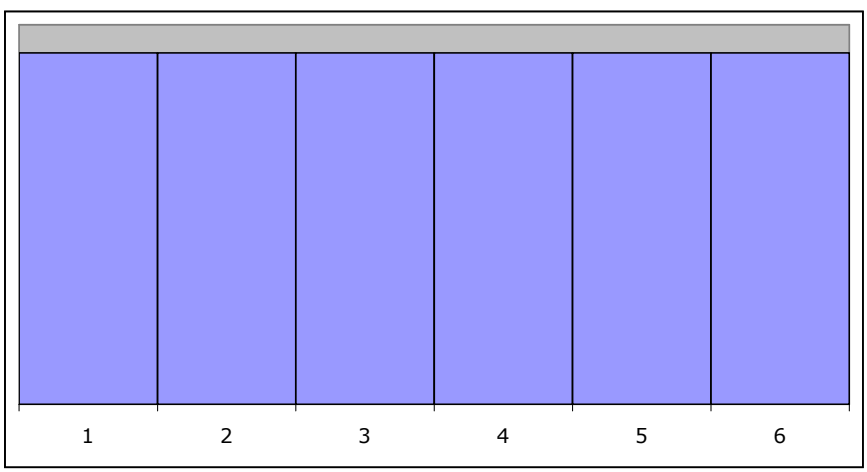
$\bar{X} = x$	$P(x)$	Pairs
1.0	1/36	(1,1)
1.5	2/36	(1,2), (2,1)
2.0	3/36	(1,3), (2,2), (3,1)
2.5	4/36	(1,4), (2,3), (3,2), (4,1)
3.0	5/36	(1,5), (2,4), (3,3), (4,2), (5,1)
3.5	6/36	(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)
4.0	5/36	(2,6), (3,5), (4,4), (5,3), (6,2)
4.5	4/36	(3,6), (4,5), (5,4), (6,3)
5.0	3/36	(4,6), (5,5), (6,4)
5.5	2/36	(5,6), (6,5)
6.0	1/36	(6,6)

As you can see, the sampling distribution of $\bar{X}$ is more complicated. (Imagine doing this for $n = 3, 4, \ldots$.)

The sampling distribution of $\bar{X}$ is charted below:



Notice that this distribution is symmetric around 3.5 and that it has *less* variability than that of X, which is also charted below:



Formally, we have

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2}{2}\right) \\ &= \frac{1}{2}\left(E(X_1) + E(X_2)\right) \\ &= \mu \end{aligned}$$

and

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{X_1 + X_2}{2}\right) \\ &= \frac{1}{4}\left(V(X_1) + V(X_2)\right) \\ &\qquad\qquad (X_1 \text{ and } X_2 \text{ are independent}) \\ &= \frac{\sigma^2}{2}. \end{aligned}$$

Thus, the variance of our estimator has been reduced in half, while the mean remained "centered."

Standard Estimator for Mean...

More generally, consider a sample of size n with outcomes X_1, X_2, ..., X_n from a population, where the X_is have a common **arbitrary** distribution, and are independent. (In our previous example, X happens to be discrete uniform over the values 1, 2, ..., 6.)

Define again

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,. \tag{1}$$

Then,

$$\mu_{\bar{X}} \equiv E(\bar{X}) = \mu \tag{2}$$

and

$$\sigma^2_{\bar{X}} \equiv V(\bar{X}) = \frac{1}{n} \sigma^2 \,. \tag{3}$$

The standard (best, in a certain sense) estimator for the population mean μ is $\mu_{\bar{X}}$. The standard deviation of $\bar{X}$, $\sigma_{\bar{X}}$, is called the **standard error** of the estimator.

Observe that

— The sample mean $\mu_{\bar{X}}$ does *not* depend on the sample size n.

— The variance $V(\bar{X})$ *does* depend on n, and it shrinks to zero as $n \to \infty$. (Recall the effect of increasing n in our example on discrimination.)

Moreover, it is most important to realize that the calculations in (2) and (3) do **not** depend on the assumed distribution of X, i.e., of the population.

Can we say the same about the sampling **distribution** of $\bar{X}$? We may be asking too much, but it turns out ...

Central Limit Theorem...

An interesting property of the normal distribution is that if X_1, X_2, ... are normally distributed, i.e., if the population from which successive samples are taken, has a normal distribution, then $\bar{X}$ is normally distributed (with parameters given in (2) and (3) above) for *all* n.

What if the population (i.e., X) is not normallly distributed?

Central Limit Theorem: For *any* infinite population with mean μ and variance σ^2, the sampling distribution of $\bar{X}$ is well approximated by the normal distribution with mean μ and variance σ^2/n, provided that n is sufficiently large.

This is the most fundamental result in statistics, because it applies to any infinite population. To facilitate understanding, we will look at several simulation examples in a separate Excel file (C9-01-Central_Limit_Theorem.xls).

The definition of "sufficiently large" depends on the extent of nonnormality of X (e.g., heavily skewed, multimodal, ...). In general, the larger the sample size, the more closely the sampling distribution of $\bar{X}$ will resemble a normal distribution. For most applications, a sample size of 30 is considered large enough for using the normal approximation.

For *finite* population, the standard error of $\bar{X}$ should be corrected to

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}},$$

where N is the size of the population. The term

$$\sqrt{\frac{N-n}{N-1}}$$

is called the *finite population correction factor*. For large N, this factor, of course, approaches 1 and hence can be ignored. The usual rule of thumb is to consider N large enough if it is at least 20 times larger than n.

Example 1: Soda in a Bottle

Suppose the amount of soda in each "32-ounce" bottle is normally distributed with a mean of 32.2 ounces and a standard deviation of 0.3 ounce.

If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces of soda? Answer:

This can be viewed as a question on the distribution of the population itself, i.e., of X (or of a sample of size one). We wish to find $P(X > 32)$, where X is normally distributed with $\mu = 32.2$ and $\sigma = 0.3$:

$$\begin{aligned} P(X > 32) &= P\left(\frac{X-\mu}{\sigma} > \frac{32-32.2}{0.3}\right) \\ &= P(Z > -0.67) \\ &= 1 - P(Z \leq -0.67) \\ &= 0.7468\,, \end{aligned}$$

where the last equality comes from the Excel function NORMSDIST().

Hence, there is about 75% chance for a single bottle to contain more than 32 ounces of soda.

Suppose now that a customer buys a box of *four* bottles, what is the probability for the mean amount of soda in these four bottles to be greater than 32 ounces? Answer:

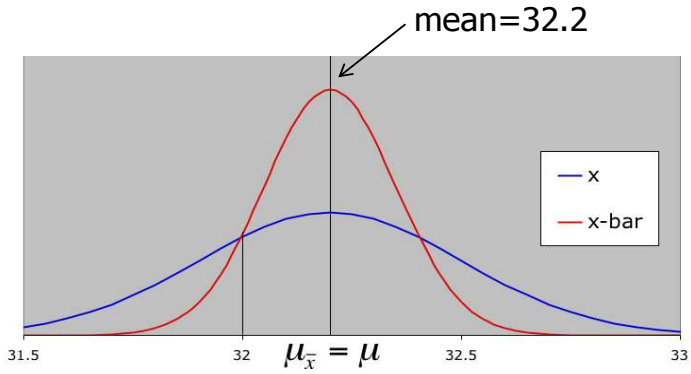
We are now interested in $\bar{X}$ for a sample of size 4. Since X is normally distributed, $\bar{X}$ also is. We also have $\mu_{\bar{X}} = \mu = 32.2$ and $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 0.3/\sqrt{4} = 0.15$. It follows that

$$\begin{aligned} P(\bar{X} > 32) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} > \frac{32 - 32.2}{0.15}\right) \\ &= P(Z > -1.33) \\ &= 1 - P(Z \leq -1.33) \\ &= 0.9082\,. \end{aligned}$$

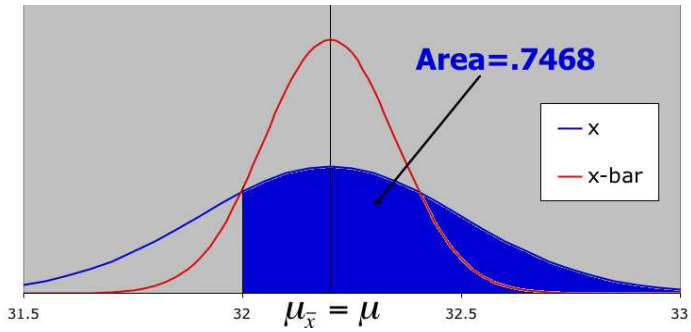
Thus, there is about 91% chance for the mean amount of soda in four bottles to exceed 32 ounces.

The answer here, 91%, is greater than that in a sample of size one. Is this expected? Note that the standard deviation of the sample mean ($\sigma_{\bar{X}}$) is *smaller* than the standard deviation of the population (σ), as highlighted in yellow above. Pictorially, ...

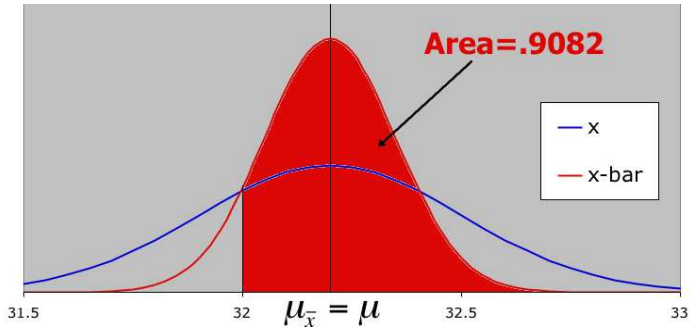
Reducing Variability:



Sample of Size One:



Sample of Size Four:



Example 2: Average Salary

The Dean of a Business school claims that the average salary of the school's graduates one year after graduation is \$800 per week with a standard deviation of \$100.

A second-year student would like to check whether the claim about the mean is correct. He carries out a survey of 25 people who graduated one year ago. He discovers the sample mean to be \$750. Based on this information, what can we say about the Dean's claim?

Analysis: To answer this question, we will calculate the probability for a sample of 25 graduates to have a mean of \$750 *or less* when the population mean is \$800 and the population standard deviation is \$100, i.e., the ***p*-value** for the given sample mean.

Although X is likely to be skewed, it seems reasonable to assume that $\bar{X}$ is normally distributed. The mean of $\bar{X}$ is $\mu_{\bar{X}} = 800$ and the standard deviation of $\bar{X}$ is $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 100/\sqrt{25} = 20$.

Therefore,

$$\begin{aligned} P(\bar{X} \leq 750) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \leq \frac{750 - 800}{20}\right) \\ &= P(Z \leq -2.5) \\ &= 0.0062\,. \end{aligned}$$

Since this p-value is extremely small (compared to 0.05 or 0.01), we conclude that the Dean's claim is not supported by data.

Standardizing the Sample Mean...

A common practice is to convert calculations regarding $\bar{X}$ into one regarding the standardized variable Z:

$$\begin{aligned} Z &= \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} \\ &= \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \,. \end{aligned} \tag{4}$$

Indeed, we had done this in the previous two examples; and this unifies discussion and notation.

From the central limit theorem, Z as defined in (4) is, for large n, normally distributed with mean 0 and variance 1. That is, the distribution of Z can be well approximated by the standard normal distribution.

Define $z_{\alpha/2}$ as the value such that

$$P(-z_{\alpha/2} < Z \leq z_{\alpha/2}) = 1 - \alpha \,. \tag{5}$$

The choice of α, the so-called **significance level**, is usually taken as 0.05 or 0.01.

The value $z_{\alpha/2}$ is called the *two*-tailed critical value at level α, in contrast with the *one*-tailed z_A, or z_α in our current notation, discussed earlier.

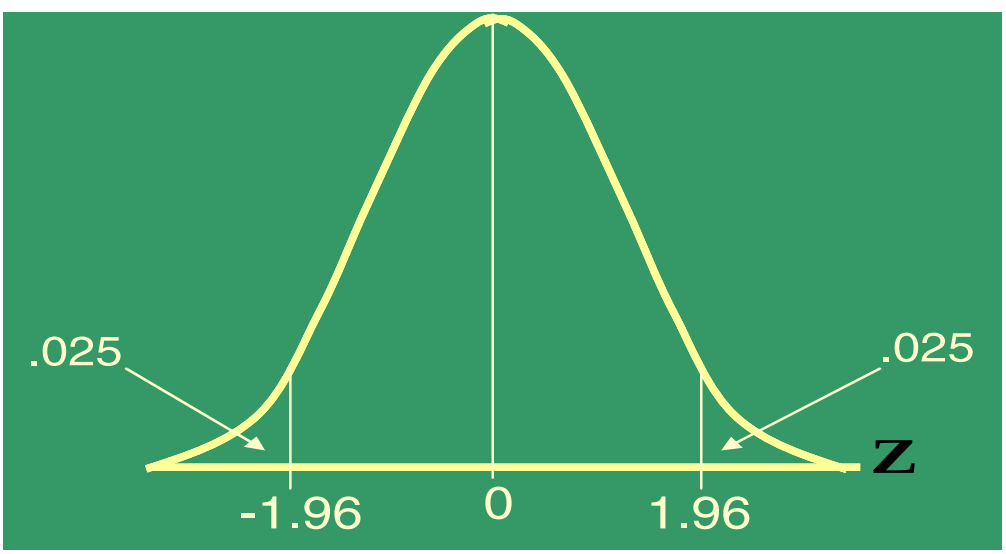
Using NORMSINV() (see C7-07-Normal.xls; "Heights" example, last question), it is easily found that

— For $\alpha = 0.1$, $z_{\alpha/2} = 1.645$.

— For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$.

— For $\alpha = 0.01$, $z_{\alpha/2} = 2.576$.

For $\alpha = 0.05$, the most-common choice, this means:



Upon substitution of (4), (5) becomes

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha\,,$$

which is equivalent to:

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha\,. \quad (6)$$

This means that for a given α, the probability for the sample mean $\bar{X}$ to fall in the interval

$$\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\ \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad (7)$$

is $100(1-\alpha)\%$. We will often use this form for statistical inference, since it is easy to check if a given $\bar{X}$ is contained in the above interval. The end points

$$\mu \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

are often called *control limits*; they can be computed without even taking any sample.

Example 2: Average Salary — Continued

We can also check the Dean's claim using a slightly different approach.

Let $\alpha = 0.05$ (say) and define an observed $\bar{X}$ as **rare** if it falls *outside* the interval (7).

Analysis: With $\mu = 800$, $\sigma = 100$, $n = 25$, our control limits are

$$800 \pm 1.96 \frac{100}{\sqrt{25}};$$

or, from 760.8 to 839.2. Since the observed $\bar{X} = 750$ is outside these limits, the Dean's claim is not justified.

Note that if we had observed an $\bar{X}$ of 850 (50 above the mean, as opposed to below), we would also have rejected the Dean's claim. This suggests that we could also define a "two-sided" p-value as:

$$P(\bar{X} \leq 750 \text{ or } \bar{X} \geq 850).$$

This is just twice the original (one-sided) p-value, and hence equals $2 \cdot 0.0062 = 0.0124$. Since this is less than the given $\alpha = 0.05$, we arrive, as expected, the same conclusion.

Sampling Distribution of a Proportion. . .

The central limit theorem also applies to "sample proportions." Let X be a binomial random variable with parameters n and p. Since each trial results in either a "success" or a "failure," we can define for trial i a variable X_i that equals 1 if we have a success and 0 otherwise. Then, the proportion of trials that resulted in a success is given by:

$$\hat{p} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{X}{n} \,. \tag{8}$$

In the "Discrimination" example (see C7-05-Binomial.xls), n is the number of available positions and p is the probability for hiring a female for a position. Let X_i "indicate" whether or not the ith hire is a female. Then, $\hat{p}$ as defined in (8) is the (random) proportion of hires that are female.

Observe that the middle term in (8) is an average, or a sample mean. Therefore, the central limit theorem implies that the sampling distribution of $\hat{p}$ is approximately normal with mean

$$E(\hat{p}) = p \tag{9}$$

(note that $E(X_i) = p$) and variance

$$V(\hat{p}) = \frac{p(1-p)}{n} \tag{10}$$

(note that $V(X_i) = p(1-p)$).

This discussion, in fact, explains why, under suitable conditions, the normal distribution can serve as a good approximation to the binomial; see C7-07-Normal.xls.

From (9) and (10), we see that $\hat{p}$ can be standardized to a Z variable:

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}\,.$$

Example: Discrimination — Continued

For $n = 50$, $p = 0.3$ and $\hat{p} = 0.1$ (10% of 50), we have

$$\begin{aligned} Z &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ &= \frac{0.1 - 0.3}{\sqrt{0.3(1-0.3)/50}} \\ &= -3.086\,. \end{aligned}$$

For $\alpha = 0.01$, we have $z_\alpha = 2.326$ (NORMSINV(0.99); note that this is a one-tailed critical value). Since -3.086 is less than -2.326 (the normal density is symmetric), we conclude that it is likely that discrimination exists. Note that our calculations here are based on a normal approximation to the *exact* binomial probability in C7-05-Binomial.xls. Since the approximation is good, the conclusions are consistent.

Sampling Distribution of a Difference. . .

We are frequently interested in comparing two populations.

One possible scenario is that we have two *independent* samples from each of two *normal* populations. In such a case, the sampling distribution of the **difference** between the two sample means, denoted by $\bar{X}_1 - \bar{X}_2$, will be normally distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \tag{11}$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\,. \tag{12}$$

If the two populations are *not* both normally distributed, then the above still applies, provided that the sample sizes n_1 and n_2 are "large" (e.g., greater than 30).

As usual, we can standardize the difference between the two means:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\,. \tag{13}$$

Example: MBA Salaries

The starting salaries of MBA graduates from two universities are \$62,000 and \$60,000, with respective standard deviations \$14,500 and \$18,300. Assume that the two populations of salaries are normally distributed.

Suppose $n_1 = 50$ and $n_2 = 60$ samples are taken from these two universities. What is the probability for the first sample mean $\bar{X}_1$ to exceed the second sample mean $\bar{X}_2$?

Analysis: We wish to find $P(\bar{X}_1 - \bar{X}_2 > 0)$, which can be computed via Z as:

$$
\begin{aligned}
P\left(Z > \frac{0 - (62000 - 60000)}{\sqrt{14500^2/50 + 18300^2/60}}\right) \\
&= P(Z > -0.64) \\
&= 1 - P(Z \leq -0.64) \\
&= 1 - 0.261 \\
&= 0.739\,.
\end{aligned}
$$

Thus, there is about a 74% chance for the sample mean starting salary of the first university to exceed that of the second university.