Chapter 1

Sampling Distributions

> Statistics mean never having to say you are certain
>
> unknown

Sample Statistics

In previous chapters you were introduced the notions of sample mean $\bar{X}$, sample proportion $\hat{p}$, sample standard deviation s. All of them are examples of sample statistics. Sample statistic is a *formula* producing numerical values calculated on sample datasets. Sometimes the values themselves are also called sample statistics. The resulting values are as usual some *characteristics* of a sample. Let's consider them in more detail.

Sampling Error

In chapter 6 we reached the conclusion that sampling error is inevitable. We carefully collect data and calculate characteristics we are interested in, however, different samples provide *different* values of calculated statistic. So, sample statistic is a *random variable*! As a result, every conclusion made about population contains some level of uncertainty.

We want to know some important truth which in practice is unachievable!

What to do?!

Fortunately Statistics allows to *quantify* sampling error, and predict behavior of sample statistics. For instance we can predict that the sample mean weight must almost surely vary from 171 to 181 pounds meaning that it is within ± 5 pounds

about 176; or that the proportion of broken details must almost surely be between 4.7 and 5.3 percent, that is within the margin of ±0.3% away from 5%. We can understand the *nature* of sampling error, and still provide valid statements about population.

More formally, as for any other random variable, the behavior of a sample statistic is analyzed with its *probability distribution*. Recall that probability distribution is the collection of possible values random variable takes with associated probabilities.

How sampling distribution is created?

Probability distribution of a *sample statistic* is called a **sampling distribution**. To understand exactly what it is and from where it appears consider a clarifying example.

Example "TEAM OF WORKERS". At the company where Masha does her internship one department consists of six workers. Their work experiences are correspondingly 2, 4, 6, 6, 7 and 8 years. A manager wants to form a small team of 4 employees to work on a new project. Work experience is a crucial characteristic of a worker but since the manager is busy he will choose workers randomly. What will be the *mean* work experience in a team?

Solution.

The total number of workers in the department is six. The average work experience of workers in the department is therefore: $\frac{2+4+6+6+7+8}{6} = 5.5$ years.

The team consisting of four employees will be a random sample of size 4 chosen from a population of size 6. There exist fifteen different samples that can be selected in this way (15 is the number of combinations C_6^4). The table below shows all of them with the corresponding mean work experience calculated for each sample.

	Samples	Mean work experience		Samples	Mean work experience
1.	2,4,6,6	4.5	**9.**	2,6,7,8	5.75
2.	2,4,6,7	4.75	**10.**	2,6,7,8	5.75
3.	2,4,6,8	5.0	**11.**	4,6,6,7	5.75
4.	2,4,6,7	4.75	**12.**	4,6,6,8	6
5.	2,4,6,8	5	**13.**	4,6,7,8	6.25
6.	2,4,7,8	5.25	**14.**	4,6,7,8	6.25
7.	2,6,6,7	5.25	**15.**	6,6,7,8	6.75
8.	2,6,6,8	5.5			

Table 1.1: Fifteen different samples.

Since procedure was random sampling, each out of fifteen sample has equal chance of being selected, namely $\frac{1}{15}$. Using this we can determine the probability of every possible value of mean. For instance, from the table we can see that three out of fifteen possible samples resulted in mean experience 5.75 years. Thus, the probability that the team for the project would have an average work experience of 5.75 years is $\frac{3}{15}$. In the same way we can calculate probability for all other values of means.

The resulting collection of values with corresponding probabilities constitutes the probability distribution of the sample mean or, equivalently, its sampling distribution. It is provided below in the table. Denote X as an employee's work experience. Then, the mean work experience can be denoted as $\bar{X}$. Its distribution is the following:

X	4.5	4.75	5	5.25	5.5	5.75	6	6.25	6.75
$P(X)$	1/15	2/15	2/15	2/15	1/15	3/15	1/15	2/15	1/15

We can also show it on a histogram.

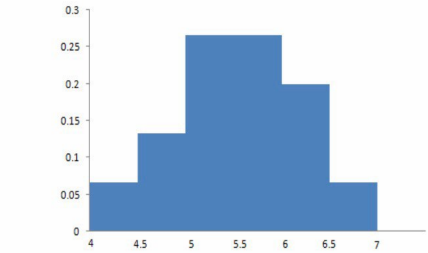


Figure 1.1: Distribution of mean in sample.

Note that while experience X varies from 2 to 8 years, *mean* experience $\bar{X}$ varies from 4.5 to 6.75 years. The values of mean have smaller range and are more concentrated in the center than the underling variable. This will be shown analytically on page 8.

Recall: Probability distribution is the collection of values with associated probabilities. So, distinguish among three notions:

- Sampling distribution – distribution not of X, but of some sample statistic among different samples
- Population distribution – the true distribution of all possible values of X in the world. Remember, we take samples from populations.
- Sample distribution (rarely used term). It is the way a particular sample is distributed. Sample distribution is often shown using a histogram. We did it a lot in Graph chapter. This you usually draw with histogram to show the range of values in the sample with associated relative frequencies (not probabilities).

Sample mean $\bar{X}$ is a variable, but in problems you will meet the sentences like "sample mean equals 50". In this context 50 is the *realization* of this variable – a particular value it took on a particular calculated sample.

Most common sample statistics

Note that the term statistic can be used both for the function (the formula) and for the values of the function (calculated on a given sample).

Sample Mean $\bar{X}$ is the average of sample observations.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n}$$

Imagine now that the procedure of obtaining data is repeated. This will result in different datasets. How sample mean $\bar{X}$ could vary among different samples? Read further to learn it.

Sample proportion $\hat{p}$ In fact initially sample proportion was introduced in Chapter 1. It is the same as relative frequency of some event estimated in limited number of experiments – a sample. It is then called a sample proportion.

It equals $\hat{p} = \frac{n_a}{n}$ – the number of successes n_a out of total number of trials (sample size) n. By success we mean some event of interest, like number of broken details or positive votes. We are interested in the proportion of n_a out of n. Check closer this expression. Sample size n is a constant while n_a is a random variable, distributed

binomially. Indeed, n_a is the number of people or objects of interest in a given sample. Each of these objects has a chance p to be in the required category of interest. Thus, $n_a \sim \text{Bi}(n, p)$.

Recall that binomial random variable is the number of successes in n independent trials given equal probability of success p. Expectation of binomial random variable equals np, while variance equals $np(1-p)$. For example, the number of yellow M&M's candies in a randomly chosen pack is binomial. Sometimes π is used instead of p.

Central Limit Theorem

We discussed that sample statistics have their own probability distribution since they are random variables, which is called sampling distribution. Which type of distribution do they have? It happens that in practice in many cases the distribution of sample statistic is approximately Normal! This surprising fact explains why normal distribution plays such a central role in Statistics – it is used to describe the results of sampling procedures.

The Central Limit Theorem (CLT):
The *sum* of independent identically distributed random variables approaches normal distribution as the number of summands approaches infinity.

That means that if the number of these variables is big enough, their *sum* is approximately normally distributed. When sample size is large, these variables reach approximately normal distribution. It does not happen without SRS.

Do we have many random variables? Don't we have only one? Random sampling assumes that we work with a set of random variables. Go back to the "TEAM OF WORKERS" example. Let X be the experience of a worker. Assume you are randomly taking 4 workers to then measure their experiences. *After* the workers are chosen and their experience is measured, we get a set of numbers x_1, x_2, ... x_n. But *before* workers are chosen, experience of a conventionally first worker in the sample is a random variable. Denote it X_1 as worker is called first for convenience. Experience of a second randomly chosen worker is also a random variable X_2, etc. So, generally speaking we are working with the set of random variables X_1, X_2, ..., X_n which are all the same. They are the same or, more formally, identically distributed since they follow the same population of X: $X_1 = X_2 = X_n = X$. The number of random variables in a set reflects the number of objects in the sample.

Also since workers are chosen independently, all X_is are assumed to be independent from each other. This is a very classic assumption in Statistics state X_i are iid – independently identically distributed.

CLT can be applied to sample mean $\bar{X}$ and to sample proportion $\hat{p}$, which are the most frequently used sample statistics, as well as to a binomial random variable. Let's prove why CLT is applicable to each of those statistics (assuming sample taken is SRS):

For sample mean $\bar{X}$:

- $\bar{X}$ is the sum of X_i divided by a constant n
- X_i are independent from each other since the sample is assumed to be randomly taken
- X_i-s are identically distributed because each of X_i is taken from the same underlying population

Thus $\bar{X}$ is distributed normally if sample taken is large.

For Binomial random variable X (which is n_a in proportion $\frac{n_a}{n}$):

Binomial random variable is the sum of Bernoulli random variables $n_a = \sum Y_i$, $Y_i \sim \text{Bernoulli}(p)$ (recall the chapter on discrete random variable)

- $X = \sum_i^n Y_i$ so it is the sum of independent random variables
- By definition trials are independent, thus Y_i-s are independent
- $X = \sum Y_i$, where $Y_i \sim \text{Bernoulli}(p)$ so are identically distributed

Thus X is distributed normally if sample taken is large.

For sample proportion $\hat{p}$:

$\hat{p} = \frac{n_a}{n}$ where n_a has binomial distribution. So, proportion is a binomial random variable divided over a constant. For binomial random variable we've just proven that it is distributed normally.

Thus $\hat{\boldsymbol{p}}$ is distributed normally if sample taken is large.

The idea of CLT can be described shortly in one phrase: *if sample is big, everything is going to be normal!*

Limiting distribution

What does the word "approaches" mean? The answer is the larger is n, the closer distribution is to normal. To understand how it works consider it on example of binomial random variable X. Note that any other discrete random variable which is the *sum* of iid variables will be applicable with the same way.

Example "YOUR ORDER HAS BEEN SHIPPED" Let X be the amount of ordered goods in some online shop that reached you. Then X is distributed binomially since this final amount is the sum of all "successes". Under success we understand successful receipt of one order. Assume that every time you make an order, the probability it will reach you is 0.7 so probability of success in each trial equals 0.7.

$$X \sim \text{Bi}(n, 0.7)$$

Let us draw the resulting distributions of X when total number of orders is 4, 10 and 25. This is done via histograms below. As the number of trials n increases, the distribution of a discrete random variable X becomes more and more smooth and close to the bell-shaped normal form! That means the way a binomial random variable is distributed becomes very similar to normal. In extreme case, when n is

infinity, X becomes a normally distributed variable. Note that in this extreme case, discrete random variable becomes a continuous one. This is what we call approaches normal distribution.

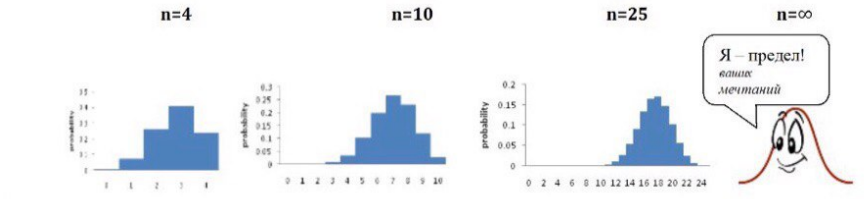


Figure 1.2: Distribution of number of orders recieved.

So we can view normal distribution as a *limiting distribution* for many others. Under conditions which are often met other distributions approach normal. Opportunity to use normal distribution to approximate other distributions is the main feature that distinguishes it from other distributions. It is a surprising fact of normality of our life.

How large is "large enough"?

Starting from what number can we say that variable is normal? In other words how to understand if a particular sample is already big enough to state that variable is approximately normally distributed. For sample mean we agreed to consider sample size to be large enough when n is at least **30**. For proportion and binomial distribution the conditions are:

$$\begin{cases} np > 5 \\ np(1-p) > 5 \end{cases}$$

Sometimes you can also meet "> 10" as a requirement for normal approximation. This is Ok, different sources use different thresholds.

The more closely the sampling distribution needs to resemble a normal distribution, the more sample observations are required. Also the more close the original population is to normal distribution, i.e. is smooth, bell-shaped, symmetric, the fewer sample points will be required for approximation to start working.

Under which conditions does CLT start working?

- as n approaches infinity
- the closer is the underlying population to normal, the earlier CLT starts working.

Normal approximation to Binomial Distribution

As we said, binomial random variable will be approximately normal under conditions often met. That means that, when the number of trials n increases, the way a binomial random variable is distributed becomes more and more similar to the normal. Thus we can use normal distribution to work with binomial variables. This is shown below.

Recall **"YOUR ORDER HAS BEEN SHIPPED"** Example. If the probability to receive one ordered good is 0.7, what is the probability that increasing number of orders to 100 at most 73 will be reached?

Solution. Look, we are considering probability on an interval. Let X be the number of orders received, then

$$X \sim \text{Bi}(100, 0.7)$$

Let us first use the exact binomial formula:

$$P(X \leqslant 73) = P(X = 0) + P(X = 1) + \ldots + P(X = 73) = C_0^{100}(0.7)^0(0.3)^{100} + \ldots = \mathbf{0.7756}$$

Now the normal approximation:
Since $np > 5$, $n(1-p) > 5$, by CLT $X \approx N(\mu, \sigma^2)$.

$$\mu = np = 100 \cdot 0.7 = \mathbf{70}$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0.7 \cdot 0.3} = \mathbf{4.5826}$$

$$P(X \leqslant 73) \approx P(Z \leqslant \frac{73 - 70}{4.5826}) = P(Z \leqslant 0.645) = \mathbf{0.7405}$$

As you see the resulting probability differs a bit. This happens because of approximation and also because normal distribution is continuous and binomial is discrete.

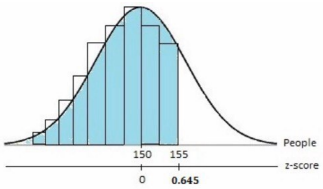


Figure 1.3: Normal approximation to binomial.

Which way is better? Of course the exact formula is the most precise. However sometimes it involves really long calculations, and for simplicity or to save time approximation is used. If the number of summands is too huge, even calculator can not count it or may stuck. In general the higher is the number of trials, the more precise will be your approximated probability.

Commonly used Sampling Distributions

We said that these variables have normal distribution. As we already know normal distribution is described by two parameters, mean (μ) and standard deviation (σ). Let us find the parameters of $\bar{X}$ and $\hat{p}$.

Sample mean $\bar{X}$

Sample mean $\bar{X}$ has the following sampling distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

As we've shown earlier, this normality holds approximately when sample is SRS and $n \geq 30$. Alternatively, $\bar{X}$ can be distributed normally even for small n, if X itself is normal (more on that in Chapter 9).

Proof Let's calculate the parameters of the random variable $\bar{X}$, namely $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}}$

- Mean of $\bar{X}$
  The following proof is based on properties of expectation: $\mu_{\bar{X}} = E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right)$ One over n is just a constant, so we put it out of brackets and from the chapter of random variables we know that the expectation of sum is the sum of expectations, so:

  $$E(\frac{\sum X_i}{n}) = \frac{1}{n}E(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}(E(X_1) + E(X_2) + \ldots + E(X_n))$$

  Since each random variable X_i has the same distribution and mean μ, we know each expectation: $\frac{1}{n}(E(X_1)+E(X_2)+\ldots+E(X_n)) = \frac{1}{n}(\mu+\mu+\ldots+\mu) = \frac{n\mu}{n} = \mu$

- Variance of $\bar{X}$:
  $Var(\bar{X}) = Var(\sum_i \frac{X_i}{n}) = \frac{1}{n^2}Var(X_1+X_2+\ldots+X_n) = \frac{1}{n^2}(Var(X_1)+Var(X_2)+\ldots+Var(X_n)) = \frac{1}{n^2}(\sigma^2+\sigma^2+\ldots+\sigma^2) = \frac{1}{n^2}n\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} < \sigma^2$

When I should divide by $\sqrt{n}$? The only way to remember this point is to understand that you should divide over the *standard deviation* of a variable. The standard, calculating probability of variable X like $P(X > 5)$ to get the standard z you should divide over σ since the standard deviation of X is σ. Calculating the probability of variable $\bar{X}$ converting to standard z you should divide over $\frac{\sigma}{\sqrt{n}}$ since the standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Further note that the variance of X is larger than that of $\bar{X}$. This is natural because X is a single characteristic and $\bar{X}$ is an *averaged* characteristic and part of variation of X is neutralized. We saw that in the "TEAM OF WORKERS" example.

So, finally what you should remember about the distribution of sample mean $\bar{X}$:

- is approximately normally distributed
- has mean μ
- has variance $\frac{\sigma^2}{n}$

Sample proportion $\hat{p}$ We've shown it most often is normally distributed. Thus $\hat{p}$ has the following distribution:

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}$$

This normality holds approximately when sample is SRS and $np > 5$, $n(1-p) > 5$.

Now let us calculate the characteristics of the random variable $\hat{p}$, namely its mean and variance.

- Mean of $\hat{p}$:
  $E(\hat{p}) = E(\frac{m}{n}) = \frac{1}{n}E(m) = \frac{1}{n}np = \frac{\text{np}}{n} = p$
- Variance of $\hat{p}$:
  $Var(\hat{p}) = Var(\frac{m}{n}) = \frac{1}{n^2}Var(m) = \frac{1}{n^2}np(1-p) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$
  Standard deviation is, as usually, the square root of variance: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

So, finally what you should remember about the distribution of sample proportion $\hat{p}$:

- is approximately normally distributed
- has mean p
- has variance $\frac{(1-p)}{n}$

Note. The variability of a sampling distribution is measured by its variance or its standard deviation. The standard deviation of sample statistic is also called the **standard error**.

Comparison of Parameters

Statistics allows us to compare several parameters. For instance, we might be interested in the questions like which of two financial firms pays a higher mean starting salary? Whether the proportion of defective smartphones is higher for Samsung than for Apple, or is the same for both brands? What can be said about the difference in the mean welfare of single-parent families versus two-parent families? To answer such questions we should consider the *difference* in two parameters as *one* parameter. For it, we should introduce a new sample statistic – the difference as one sample statistic, find its distribution and further analyse it.

Sample Means difference $\bar{X}_1 - \bar{X}_2$
The difference of sample means has the following distribution:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}\right)$$

This holds when samples of both X_1, X_2 are SRS, $n_1 \geqslant 30$, $n_2 \geqslant 30$ and X_1 and X_2 are independent from each other. The latter is necessary to ensure that difference

of two normal variables $\bar{X}_1$ and $\bar{X}_2$ is also normal. Alternatively, if n is small, you should also check that $\bar{X}_1$ and $\bar{X}_2$ are both normal.

Proof.

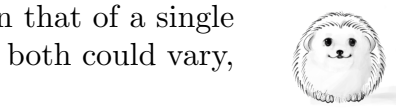
As we just shown by CLT sample mean has the following distribution: $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ If we consider two independent populations, each of them would have: $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1}$ and $\bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$ Based on the rule presented in Chapter 5 $\bar{X}_1 - \bar{X}_2$ is also a *normal* random variable, since it is a sum two independent *normal* random variables (difference is a particular case of sum).

Find the expectation and variance of this random variable:

- $E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$
- $Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(-\bar{X}_2) = Var(\bar{X}_1) + (-1)^2 Var(\bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2)$, notice the plus sign!

Standard deviation as usual equals to the square root of variance $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Note that the standard deviation of the *difference* is higher than that of a single variable. It is evident because now two variables are combined and both could vary, thus combined variability increases.



"TESTS BET" Masha is quite good at writing tests on Staistics but her results are highly volatile. Her classmate Peter has *on average* slightly lower performance although more consistent. He offered Masha a bet: "my average mark for the semester will be higher than yours!" Masha's average mark is 4.5 out of 5 with a standard deviation of 0.7 while Peter's average mark is 4.3 with a standard deviation of 0.4. There will be 50 tests during the semester. What is the probability that Peter will win the bet?

Solution:

Masha will lose the bet if her average mark will be less than the average mark of Peter.

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}) \text{ by CLT}$$

Check that sample is large! The mean of the differences is: $\mu_1 - \mu_2 = 4.5 - 4.3 = 0.2$. The standard deviation is:

$$Var(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{0.7^2}{50} + \frac{0.4^2}{50}} = 0.114$$

Probability that Masha will lose is:

$$P(\bar{X}_1 < \bar{X}_2) = P\left(\bar{X}_1 - \bar{X}_2 < 0\right) = P(z < \frac{0 - 0.2}{0.114}) = P(z < -1.75) = 0.04$$

From Table of the Normal Distribution, the area to the left of -1.75 is 0.0400. Thus there is only a 4% chance that Peter will win.

Difference in sample proportions $\hat{p}_1 - \hat{p}_2$
The difference of sample proportions has the following distribution:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}\right)$$

This holds when: both samples are SRS, $n_1 p_1 > 5$, $n_1(1-p_1) > 5$, $n_2 p_2 > 5$ $n_2(1-p_2) > 5$ and samples are independent from each other.

Single sample proportion has the distribution $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$. Exactly the same logic applies to the difference of two sample proportions. Try to proof it by yourself.

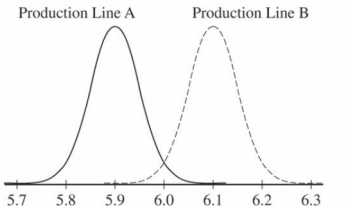
You must be able to reproduce even being half-awake

- Sample Statistic is a random variable
- Sampling distribution is the probability distribution of sample statistic
- CLT: sum of independent identically distributed random variables approaches normal distribution as the number of summands approaches infinity, in particular
- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$
- Difference in sample statistics is considered as one variable

Sample AP problems with solution

Problem 1. AP 2015 №6 d,e –проверить, не стоит ли отнести в Котопес

Corn tortillas are made at a large facility (завод) that produces 100,000 tortillas per day on each of its two production lines. The distribution of the diameters of the tortillas produced on production line A is approximately normal with mean 5.9 inches, and the distribution of the diameters of the tortillas produced on production line B is approximately normal with mean 6.1 inches. The figure below shows the distributions of diameters for the two production lines.



The tortillas produced at the factory are advertised as having a diameter of 6 inches. For the purpose of quality control, a sample of 200 tortillas is selected and the diameters are measured. From the sample of 200 tortillas, the manager of the facility wants to estimate the mean diameter, in inches, of the 200,000 tortillas produced on a given day. Two sampling methods have been proposed.

> *Method 1:* Take a random sample of 200 tortillas from the 200,000 tortillas produced on a given day. Measure the diameter of each selected tortilla.
>
> *Method 2:* Randomly select one of the two production lines on a given day. Take a random sample of 200 tortillas from the 100,000 tortillas produced by the selected production line. Measure the diameter of each selected tortilla.

Each day, the distribution of the 200,000 tortillas made that day has mean diameter 6 inches with standard deviation 0.11 inch.

(d) For samples of size 200 taken from one day's production, describe the sampling distribution of the sample mean diameter for samples that are obtained using Method 1.

(e) Suppose that one of the two sampling methods will be selected and used every day for one year (365 days). The sample mean of the 200 diameters will be recorded each day. Which of the two methods will result in less variability in the distribution of the 365 sample means? Explain.

Solution.

(d) The sampling distribution of the sample mean diameter for samples obtained using Method 1would be approximately normal with mean 6 inches and standard deviation $\frac{0.11}{\sqrt{200}} \approx \mathbf{0.0078}$ inches.

(e) Method 1 would result in less variability in the sample means (plural!) over 365 days. With Method 2 roughly half of the sample means will be around 5.9 inches and the other half will be around 6.1 inches while with Method 1

the sample means will all be very close to 6.0 inches, as indicated by very small standard deviation in part (d) (0.0078 inch)

Problem 2. AP 2014 №3 b Schools in a certain state receive funding based on the number of students who attend the school. To determine the number of students who attend a school, one school day is selected at random and the number of students in attendance that day is counted and used for funding purposes. The daily number of absences at High School A in the state is approximately normally distributed with mean of 120 students and standard deviation of 10.5 students.

(a) If more than 140 students are absent on the day the attendance count is taken for funding purposes, the school will lose some of its state funding in the subsequent year. Approximately what is the probability that High School A will lose some state funding?

(b) The principals' association in the state suggests that instead of choosing one day at random, the state should choose 3 days at random. With the suggested plan, High School A would lose some of its state funding in the subsequent year if the mean number of students absent for the 3 days is greater than 140. Would High School A be more likely, less likely, or equally likely to lose funding using the suggested plan compared to the plan described in part (a)? Justify your choice.

Solution.

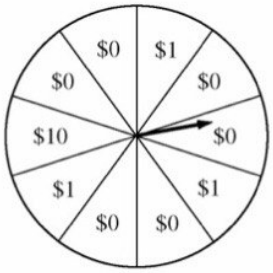
(a) Topic Normal distribution. Answer is **0.0287**.

(b) High School A would be *less* likely to lose state funding. With a random sample of 3 days, the distribution of the sample mean number of students absent would have less variability than that of a single day. With less variability, the distribution of the sample mean would concentrate more narrowly around the mean of 120 students, resulting in a smaller probability that the mean number of students absent would exceed 140.

In particular, the standard deviation of the sample mean number of absences is: $\frac{\sigma}{\sqrt{n}} = \frac{10.5}{\sqrt{3}} = 6.062$.

So, the probability that High School A loses funding using the suggested plan would be as determined as $P(\bar{X} > 140) = P(z > \frac{140-120}{6.062}) = P(z > 3.3) = 0.0005 \approx 0$, which is much less than a probability of 0.0287 obtained for the plan described in part (a).

Problem 3. AP 2012 №2 A charity fundraiser (благотворительный фонд) has a Spin the Pointer ("Покрути стрелку") game that uses a spinner (вертушку) like the one illustrated in the figure below.



A donation of $2 is required to play the game. For each $2 donation, a player spins the pointer once and receives the amount of money indicated in the sector where the pointer lands on the wheel. The spinner has an equal probability of landing in each of the 10 sectors.

(d) Based on last year's event, the charity anticipates that the Spin the Pointer game will be played 1,000 times. The charity would like to *know the probability* of obtaining a net contribution of at least $500 in 1,000 plays of the game. The mean and standard deviation of the net contribution to the charity in 1,000 plays of the game are $700 and $92.79, respectively. Use the normal distribution to approximate the probability that the charity would obtain a net contribution of at least $500 in 1,000 plays of the game.

Solution.

(d) Let random variable X be the *net contribution* the charity obtains in *one game*. It is a discrete random variable that takes three possible values 2, 1 and -8.

The charity is interested in obtaining a net contribution in 1,000 games. Then the net contribution in 1,000 games is *the sum* of 1,000 contributions in each game.

Denote the sum as Y:

$$Y = X_1 + X_2 + X_3 + \ldots + X_{1.000}$$

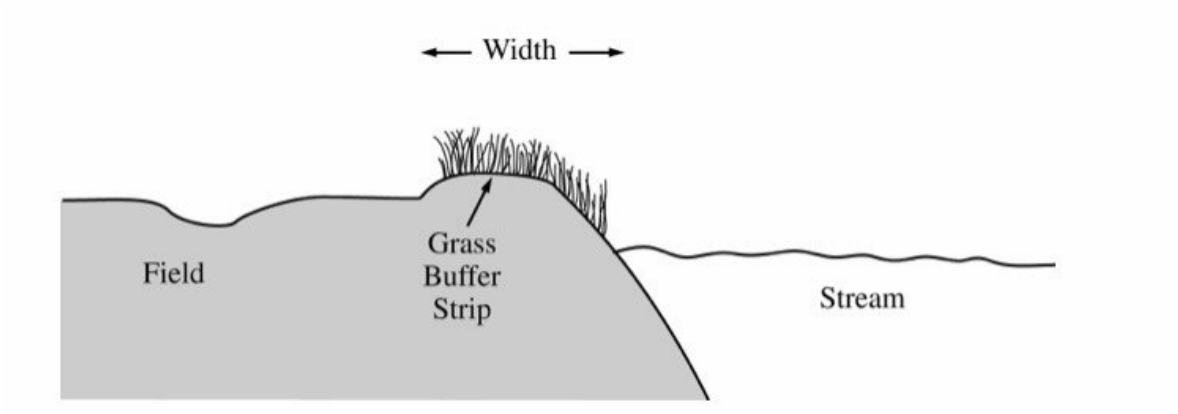
The normal approximation is appropriate here because the very large sample size ($n = 1,000$) ensures that *the central limit theorem holds*. Thus it can be said that Y is approximately normally distributed:

$$Y \sim N\left(700, 92.79^2\right)$$

Find $P(Y \geqslant 500) = P(Z \geqslant \frac{500-700}{92.79}) = P(Z \geqslant -2.155) = \mathbf{0.9844}$

Answer: The probability of obtaining a net contribution of at least $500 in 1,000 plays of the game is 0.9844 which is quite high. The charity can be very confident about gaining it.

Problem 4. AP 2011B №6 c Grass buffer strips are grassy areas that are planted between bodies of water and agricultural fields. These strips are designed to filter out sediment, organic material, nutrients, and chemicals carried in runoff water. The figure below shows a cross-sectional view of a grass buffer strip that has been planted along the side of a stream.



A study in Nebraska investigated the use of buffer strips of several widths between 5 feet and 15 feet. The following model was estimated $\hat{y} = 33.8 + 3.6x$.

A scientist in California wants to know if there is a similar relationship in her area. To investigate this, she will place a grass buffer strip between a field and a nearby stream at each of eight different locations and measure the amount of nitrogen that the grass buffer strip removes, in parts per hundred, from runoff water at each location. Each of the eight locations can accommodate a buffer strip between 6 feet and 13 feet in width. The scientist wants to investigate which combination of widths will provide the best estimate of the slope of the regression line.

Suppose the scientist decides to use buffer strips of width 6 feet at each of four locations and buffer strips of width 13 feet at each of the other four locations. Assume the model, $\hat{y} = 33.8 + 3.6x$, estimated from the Nebraska study is the true regression line in California and the observations at the different locations are normally distributed with standard deviation of 5 parts per hundred.

(c) Describe the sampling distribution of the sample mean of the observations on the amount of nitrogen removed by the four buffer strips with widths of 6 feet.

Solution.

Because the distribution of nitrogen removed for any particular buffer strip width is normally distributed with a standard deviation of 5 parts per hundred, the sampling distribution of the mean of four observations when the buffer strips are 6 feet wide will be normal with mean $33.8 + 3.6 \cdot 6 = 55.4$ parts per hundred and a standard deviation of $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{4}} = \mathbf{2.5}$ parts per hundred.

Problem 5. AP 2010 №2 A local radio station plays 40 rock-and-roll songs during each 4-hour show. The program director at the station needs to know the total amount of airtime for the 40 songs so that time can also be programmed during the show for news and advertisements. The distribution of the lengths of rock-and-roll songs, in minutes, is roughly symmetric with a mean length of 3.9 minutes and a standard deviation of 1.1 minutes.

1. Describe the sampling distribution of the sample mean song lengths for random samples of 40 rock-and-roll songs.

2. If the program manager schedules 80 minutes of news and advertisements for the 4-hour (240-minute) show, only 160 minutes are available for music. Approximately what is the probability that the total amount of time needed to play 40 randomly selected rock-and-roll songs exceeds the available airtime?

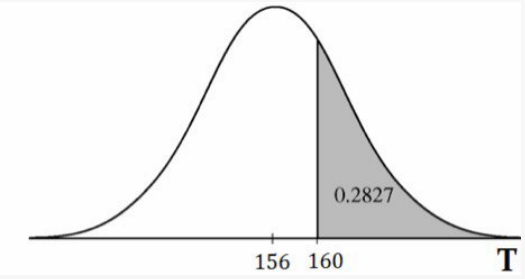
Solution.

1. The sampling distribution of the sample mean song length has mean $\mu_{\bar{X}} = \mu = 3.9$ minutes and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1.1}{\sqrt{40}} = 0.174$ minutes. The central limit theorem (CLT) applies in this case because the sample size ($n = 40$) is fairly large, especially with the population of song lengths having a roughly symmetric distribution. Thus, the sampling distribution of the sample mean song length is approximately normal.

2. Let T be the total airtime of 40 random songs, $T = X_1 + \ldots + X_{40}$. The sampling distribution of the total airtime of 40 songs is approximately normal, with mean $\mu_T = 40(3.9) = 156$ minutes and standard deviation $\sigma_T = \sqrt{Var(T)} = \sqrt{Var(X_1 + \ldots + X_{40})} = \sqrt{\sigma^2 + \ldots + \sigma^2} = \sqrt{40\sigma^2} = \sqrt{40}(1.1) = 6.96$ minutes.

   Then the probability that the total amount of time needed to play 40 randomly rock-and-roll songs exceeds the available airtime of 160 minutes is equal to:

$$P(T > 160) = P\left(z > \frac{160 - 156}{6.96}\right) = P(z > 0.57) = 0.2827$$



Problem 6. AP 2007 №3 Big Town Fisheries recently stocked a new lake in a central city park with 2,000 fish of various sizes. The distribution of the length of these fish is approximately normal.

(a) Big Town Fisheries claims that the mean length of the fish is 8 inches. If the claim is true, which of the following would be more likely?

- a random sample of 15 fish having a mean length that is greater than 10 inches

  or

- a random sample of 50 fish having a mean length that is greater than 10 inches

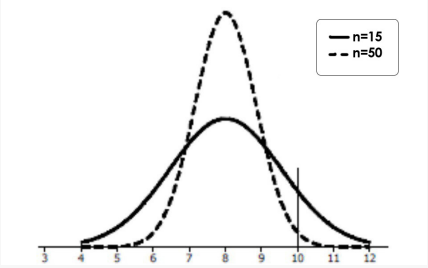
Justify your answer.

(b) Suppose the standard deviation of the sampling distribution of the sample mean for random samples of size 50 is 0.3 inch. If the mean length of the fish is 8 inches, use the normal distribution to compute the probability that a random sample of 50 fish will have a mean length less than 7.5 inches.

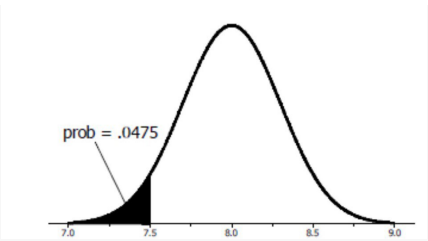
(c) Suppose the distribution of fish lengths in this lake was nonnormal but had the same mean and standard deviation. Would it still be appropriate to use the normal distribution to compute the probability in part (b)? Justify your answer.

Solution.

(a) The random sample of $n = 15$ is more likely to have a sample mean length greater than 10 inches. The sampling distribution of the sample mean $\bar{X}$ is normal with mean $\mu_{\bar{X}} = \mu = 8$ and standard deviation $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Both distributions would be centered around 8 inches, but the sampling distribution of the sample mean when $n = 15$ would have more variability than when $n = 50$ since $\frac{\sigma}{\sqrt{15}} > \frac{\sigma}{\sqrt{50}}$. The tail area $P(\bar{X} > 10$ would be larger for the distribution with more *heavy* tails that is less concentrated around mean of 8 inches.



(b) $P(\bar{X} < 7.5) = P\left(z < \frac{7.5-8}{0.3}\right) = P(z < -1.67) = \mathbf{0.0475}$



(c) Yes, it would still be appropriate. The Central Limit Theorem states that the sampling distribution of the sample mean $\bar{X}$ will become approximately normal as the sample size n increases. Here the sample size n=50 is reasonably large, so the calculation would provide a good approximation to the probability of interest even if the underlying population is nonnormal.

Problem 7. AP 2006 №3 c The depth from the surface of Earth to a refracting layer beneath the surface can be estimated using methods developed by seismologists. One method is based on the time required for vibrations to travel from a distant explosion to a receiving point. The depth measurement M is the sum of the true depth D and the random component measurement error E. The measurement error is assumed to be normally distributed with mean 0 feet and standard deviation 1.5 feet.

(c) What is the probability that the mean of the three independent depth measurements taken at the point where the true depth is 2 feet will be negative?

Solution.

(c) M – depth measurement, $M = D + E$.

$$E(M) = E(D + E) = E(D) + E(E) = E(D) + 0 = E(D)$$

$$Var(M) = Var(D + E) = Var(D) + Var(E) = 0 + Var(E) = Var(E) = 1.5^2$$

Let $\bar{M}$ be the mean of three independent depth measurements taken at the point where the true depth D is 2 feet. Since each measurement comes from a normal distribution, the distribution of $\bar{M}$ is normal. Thus we can find its mean and standard deviation:
$E(\bar{M}) = \mu = E(M) = E(D) = 2\text{feet}$ $Var(\bar{M}) = \frac{\sigma^2}{n} = \frac{Var(M)}{n} = \frac{Var(E)}{n} = \frac{1.5^2}{3}$
and standard deviation is $\frac{1.5}{\sqrt{3}} = 0.8660$

Then the probability that the mean of the three independent depth measurements taken at the point where the true depth is 2 feet will be negative is:

$$P(\bar{M} < 0) = P(z < \frac{0 - 2}{\frac{1.5}{\sqrt{3}}} = P(z < -2.31) = 0.0104$$

Practice AP Problems

Problem 1. AP 2009 №2 c –КОТОПЕС

Problem 2. AP 2008 №4 c An experiment was conducted to study the effect of temperature on the reliability of an electronic device used

in an undersea communications system. The experiment was done in a laboratory where tanks of seawater were maintained at either 10°C, 30°C, 50°C, or 70°C. After the electronic devices were submerged in the tanks for 5,000 hours, each device was inspected to determine if it was still working. The following table provides information on the number of devices tested at each temperature and the number of working devices at the end of the 5,000-hour test.

(c) An estimate of the proportion of devices that would work after 5,000 hours of submersion in 40°C seawater can be obtained by averaging the estimates at 30°C and 50°C. Compute this estimate and the associated standard error.

Problem 3. AP 2007B №2 c –КОТОПЕС

Problem 4. AP 2004B №3 c-d

Answers to practice problems:

Problem 2. 0.77, 0.0492

Problem 3. 0.0069, yes