# MAKING SENSE OF DATA

## Intermediate series

# SAMPLING DISTRIBUTIONS



**Sampling distributions**
Confidence intervals
Hypothesis tests
The t-distribution

In statistics when you take a sample of data, it's important to recognise the results will vary from sample to sample. Statistical results based on samples should include a measure of how much they expect those results to vary from sample to sample. This topic shows how to do that in terms of the sample means (for numerical data) and sample proportions (for categorical data).

# Sampling distributions

Suppose everyone in a large City rolled a single die and recorded the outcome, $X$. With all those outcomes, we'd have an entire population of values. The graph of these outcomes in the population would represent the distribution of $X$. Now suppose everyone rolled their die 10 times (a sample size of 10) and recorded the average, $\bar{x}$. With all those averages, we'd get an entirely new population — the population of sample means. The graph of this new population would represent the sampling distribution of $\bar{X}$.

When you're talking about a particular sample mean, use the notation $\bar{x}$ and for the random variable representing any sample mean in general, use the notation $\bar{X}$.

A distribution is a listing or graph of all possible values of a random variable or a population (such as $X$) and how often they occur. For example, if you roll a fair die and record the outcome and repeat an infinite number of times, the distribution of $X =$ the outcome, with numbers 1, …, 6 appearing with equal frequency.

Appling this idea to sample means, take a sample of values from your random variable $X$ (your population), find the mean of the sample, and repeat over and over again. We now have a new random variable called $\bar{X}$, with a wide range of possible values and has its own distribution.

A listing or graph of all possible values of the sample mean and how often they occur is called the *sampling distribution of the sample mean*.

For example if you roll a die 10 times, find the average, and then repeat infinite times, the average will take on values fairly close to 3.5 (halfway between 1 and 6) with values near 3.5 occurring more often than values near 1 or 6. Figure 1 shows the actual sampling distribution of $\bar{X}$, the average of 10 rolls of a die.

The term *sampling distribution* is used because data represent averages based on samples, not individual values from a population.

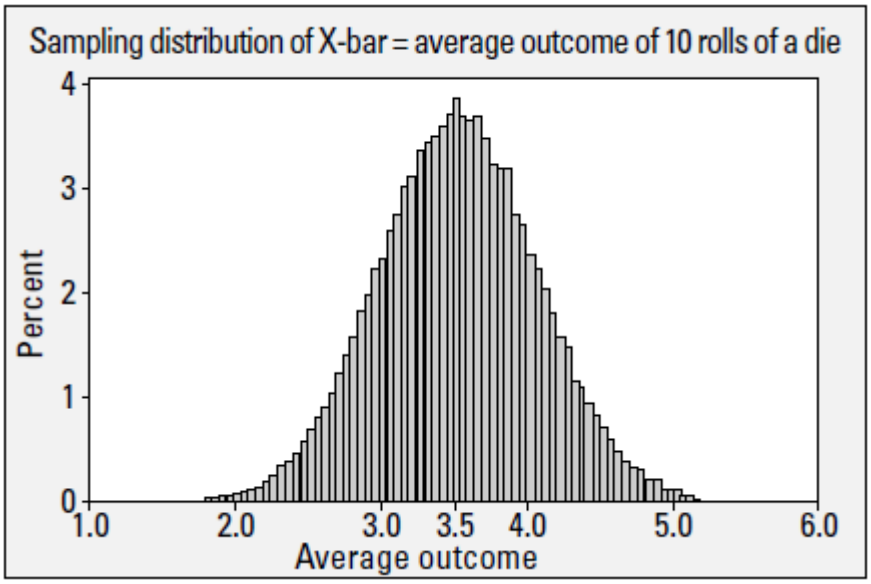


**Figure 1: Distribution of average rolls of 10 dice**

## The mean of a sampling distribution

In the die rolling example, the mean of $X$ (the outcome of a single die) is $\mu_x = 3.5$ . The mean of $\bar{X}$, denoted by $\mu_{\bar{x}}$ , equals 3.5 as well. The average of a single roll is the same as the average of all possible sample means from 10 rolls.

In general the mean of this population of all possible sample means $\mu_{\bar{x}}$ is the same as the mean of the entire population$\mu_x$, written as

$$\mu_{\bar{x}} = \mu_x \, .$$

## Standard error of a sampling distribution

The values in any population deviate from their mean (people have different heights, and so on). Variability in a population of individuals ($X$) is measured in *standard deviations*. Sample means vary because you're not sampling the whole population, only a subset. Variability in the sample mean ($\bar{X}$) is measured in terms of *standard errors*.

"Errors" don't mean there's been a mistake — it means there is a gap between the population and sample results.

The standard error of the sample means $\sigma_{\bar{x}}$ has the formula given by

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

where $\sigma_x$ is the population standard deviation and $n$ is the sample size.

## Sample size and standard error

Because $n$ is the denominator of this formula, the standard error decreases as $n$ increases. This makes sense intuitively as having more data gives less variation (and more precision) in your results.

## Population standard deviation and standard error

Suppose you have two ponds of fish (Pond #1 and Pond #2), and you want to find the average length of all the fish in each pond. Suppose you know that the fish lengths in Pond #1 have a mean of 20 inches and a standard deviation of 2 inches (see Figure 2a). Suppose the fish in Pond #2 also average 20 inches, but have a standard deviation of 5 inches (see Figure 2b). Comparing Figures 2a and 2b you see they have the same shape and mean, but the fish in Pond #2 are more variable than in Pond #1.

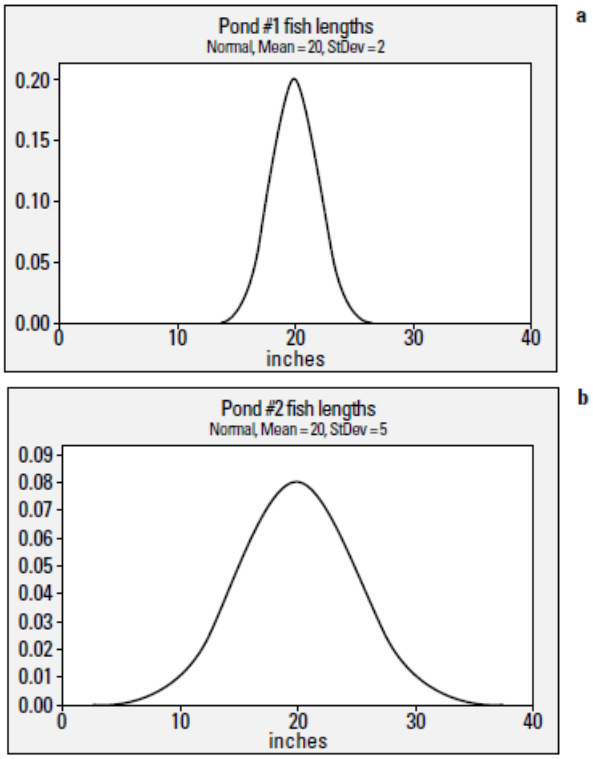


**Figure 2: Distributions of a) fish lengths in Pond #1; b) in Pond #2**

Now suppose you take a sample of 100 fish from Pond #1, find the mean length of the fish, and repeat this process over and over. Then do the same with Pond #2. Knowing that the fish in Pond #2 have more variability than Pond #1 in the first place, the means of the samples from Pond #2 will have more variability compared to Pond #1 as well. It's harder to estimate the population average when the population varies a lot to begin with — it's much easier to estimate the population average when the population values are similar.

**Distribution shape**

Now that we know the mean and standard error of $\bar{X}$, the next step is to determine the sampling distribution of $\bar{X}$ (that is, the shape of the distribution of all possible $\bar{X}$'s from all possible samples). There are two cases: 1) the original distribution for $X$ (the population) is normal; and 2) the original distribution for $X$ (the population) is not normal, or is unknown.

Case 1: Distribution of $X$ is normal
If $X$ has a normal distribution, then $\bar{X}$ does too. This is a mathematical statistics result and requires no additional tools to prove.

Case 2: Distribution of $X$ is unknown or not normal
If the $X$ distribution is any distribution that is not normal, or if its distribution is unknown, you can't automatically say the sample means ($\bar{X}$) have a normal distribution. But you can approximate $\bar{X}$'s distribution with a normal distribution — if the sample size is large enough. This result is due to the Central Limit Theorem.

*Central Limit Theorem (CLT)*

The CLT says that the sampling distribution (shape) of $\bar{X}$ is approximately normal, if the sample size is large enough. (The CLT isn't concerned what the distribution of $X$ is.)

Formally, for any population with mean $\mu$ and standard deviation $\sigma_x$, the CLT states that:

- ✓ If the distribution of $\bar{X}$ is non-normal or unknown, the sampling distribution of all possible sample means, $\bar{X}$ is approximately normal for a sufficiently large sample size.
- ✓ The larger the sample size ($n$), the closer the distribution of the sample means will be to a normal distribution.
- ✓ Statisticians tend to agree $n$ should be at least 30.

# Finding probabilities for $\bar{X}$

After you've established through Case 1 or Case 2 (from the previous section) that $\bar{X}$ has a normal or approximately normal distribution, you can find probabilities for $\bar{X}$ by converting the $\bar{x}$–value to a z-value and finding probabilities using the Z-table (see "The normal distribution" guide).

The general conversion formula is

$$Z = \frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Now substituting previous results for the mean and standard error of $\bar{X}$ the conversion formula becomes

$$Z = \frac{\bar{X} - \mu_x}{\sigma_x / \sqrt{n}}$$

*Example*

Suppose $X$ is the time it takes an administration worker to type and send 5 letters. Suppose $X$ (the times for all the workers) has a normal distribution and the reported mean is 10 minutes and the standard deviation 2 minutes. You take a random sample of 50 workers and measure their times. What is the chance that their average time is less than 9.5 minutes?

This question translates to finding $P(\bar{X} < 9.5)$. As $X$ has a normal distribution to start with, we know $\bar{X}$ also has a normal distribution.

Converting to z-value we get

$$\frac{9.5 - 10}{2 / \sqrt{50}} = -1.77$$

So we want $P(Z < -1.77)$, which equals 0.0384 from the Z-table. Therefore the chance that these 50 randomly selected workers average less than 9.5 minutes to complete this task is 3.84%.

# Sampling distribution of the sample proportion

The Central Limit Theorem (CLT) doesn't apply only to sample means. You can also use it with other statistics, including sample proportions. The *population proportion*, $p$, is the proportion of individuals in the population that have a certain characteristic of interest based on a binomial random variable. The *sample proportion*, denoted $\hat{p}$, is the proportion of individuals in the sample that have that same characteristic of interest. The sample proportion is the number of individuals in the sample who have that characteristic of interest divided by the total sample size ($n$).

If you take a sample of 100 cats and find 60 black cats, the sample proportion for black cats is $60/100 = 0.60$. This section examines the sampling distribution of all possible sample proportions, $\hat{p}$, from samples of size $n$ from a population.

The **sampling distribution of $\hat{p}$** has these properties:

- ✓ Its mean is the population proportion, denoted by $p$
- ✓ Its standard error is $\sqrt{p(1-p)/n}$ (note that because $n$ is in the denominator, standard error decreases as $n$ increases)
- ✓ Its shape is *approximately* normal, provided that the sample size is large enough. This is due to the CLT. That means you can use the normal distribution to find probabilities for $\hat{p}$
- ✓ The larger the sample size ($n$), the closer the distribution of sample proportions is to a normal distribution

How large is large enough for the CLT to work for categorical data? Most statisticians agree that both $np$ and $(1 - p)$ should be greater than or equal to 10. You want the average number of successes ($np$) and the average number of failures $(1 - p)$ to be at least 10.

**What proportion of students need maths help?**

Suppose you want to know what proportion of incoming college students would like help in maths. A survey might be issued to a sample of students entering a college with a question on whether they would like some help with maths skills. Assume past research had been conducted and that 38% of all students request help with their maths skills. That means $p = 0.38$ in this case.

The original data has a binomial distribution where success = "would like help". The "yes" responses ($p$) and "no" responses ($1 - p$) for the population are shown in Figure 3 as a bar graph.

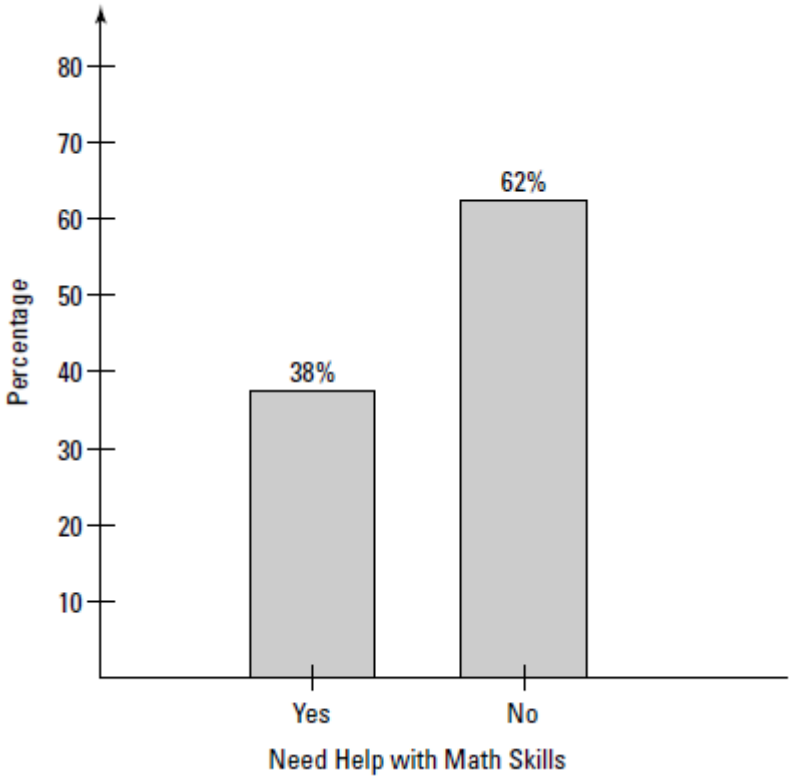


**Figure 3: Population percentages for responses to the math-help question**

Now take all possible samples of size 1,000 from this population and find the proportion in each who said they needed maths help. The distribution of these sample proportions is in Figure 4. It has an approximate normal distribution with mean $p = 0.38$ and standard error equal to

$$\sqrt{p(1-p)/n} = \sqrt{0.38(1-0.38)/1000} = 0.015$$

This approximation is valid because the two conditions for the CLT are met: 1) $np = 1{,}000(0.38) = 380$ (which is at least 10); and 2) $(1 - p) = 1{,}000(0.62) = 620$ (also at least 10).

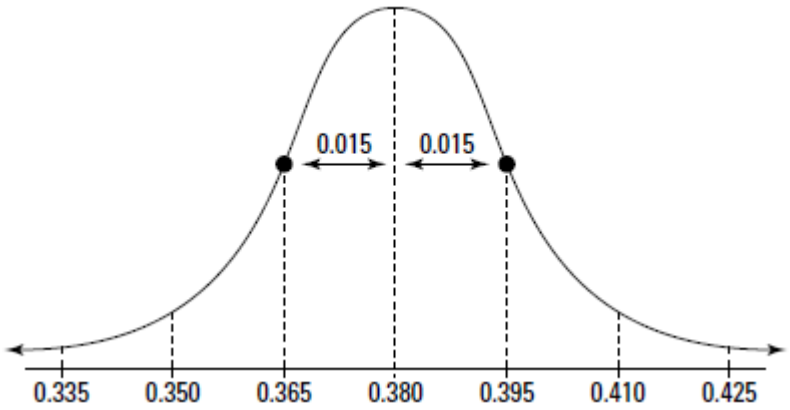


**Figure 4: Proportion of students responding "yes" to maths-help question for samples of size 1,000**

# Finding probabilities for $\hat{p}$

From our example, it's reported that 0.38 or 38% of all the students would like maths help. Suppose you took a random sample of 1,000 students. What is the chance that more than 40 percent of them say they need help?

What the question wants is the probability that the sample proportion, $\hat{p}$ is greater than 0.40; that is, $P(\hat{p} > 0.40)$. This question is answered using the normal approximation for $\hat{p}$ described in the previous section, given the stated conditions are met.

First check the conditions: 1) is $np$ at least 10? Yes because 1,000 * 0.38 = 380; 2) is $(1 - p)$ at least 10? Again yes because 1,000 * (1 – 0.38) = 620. You can use the normal approximation to answer this question.

We make the conversion of the $\hat{p}$-value to a z-value using

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

to get

$$Z = 0.40 - 0.38 \Big/ \sqrt{\frac{0.38(1-0.38)}{1000}} = 1.30$$

P($Z > 1.30$) = 1 – 0.9032 = 0.0968. So if 38 percent of students wanted help, the chance of taking a sample of 1,000 students and getting more than 40 percent needing help is approximately 0.0968 (by the CLT).

Comparing sample results to a claim about the population is called *hypothesis testing*. Because the chance of getting more than 40% of the students in our sample who requested help is 0.0968, we would not reject the claim that 38% of the population of all students request help. To reject this claim most statisticians would want a probability less than 0.05.