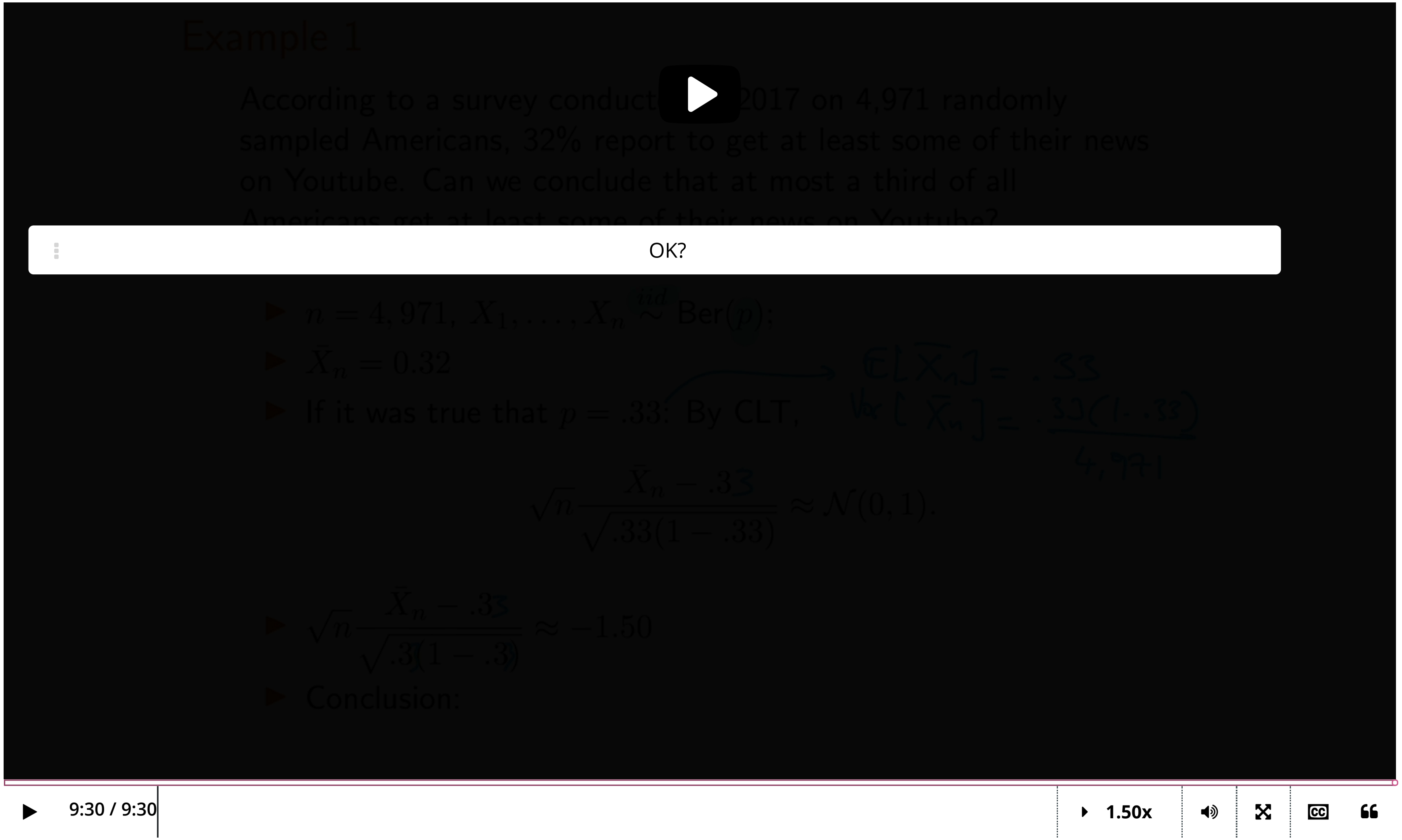
8. First Example

Does at most a third of Americans get at least some news from youtube?



Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

Intuition for Hypothesis Testing

1/1 point (graded)

The purpose of this question is not to formally outline the procedure of hypothesis testing, but rather to illustrate some of the intuition involved in answering a hypothesis testing question.

Your friend claims to you that a random variable X has the distribution $\mathcal{N}(0,1)$, and your goal is to decide whether or not this claim is true. You observe a single realization this random variable, which comes out to be $X = 3.514$.

Which of the following is the most plausible assessment of the experiment?

- It is **not** very unlikely for a standard Gaussian random variable to be at least 3.514 (*i.e.*, the event has probability larger than 5%), so you are not able to refute your friend's claim that $X \sim \mathcal{N}(0,1)$.
- It is **not** very unlikely for a standard Gaussian random variable to be at least 3.514 (*i.e.*, the event has probability larger than 5%), so you can affirm with 100% certainty your friend's claim that $X \sim \mathcal{N}(0,1)$.
- It is very unlikely for a standard Gaussian random variable to be at least 3.514 (*i.e.*, the event has probability less than 0.1%), so if indeed $X \sim \mathcal{N}(0,1)$, then you just observed a very rare event. Intuitively, it seems unlikely that your friend's claim is true.
- It is very unlikely for a standard Gaussian random variable to be at least 3.514 (*i.e.*, the event has probability less than 0.1%), so you can conclude with 100% certainty that X is **not** distributed like a Gaussian.

Solution:

The third choice is correct. We can compute using computational tools or a table that if $X \sim \mathcal{N}(0,1)$, then

$$P(X > 3.514) = \int_{3.514}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \approx .00022$$

which is smaller than 0.1%. Indeed this is a very rare event, so based on this heuristic argument, it seems unlikely that your friend's claim is true. We examine the incorrect choices in order:

- The first two choices are both incorrect. As above, $P(X \geq 3.514)$ is much smaller than 5%, so X being larger than the given observation is **not** a likely event.
  Remark: Note how the language between these two choices differs: the first one says "you are not able to refute your friend's claim," and the second says "you can affirm with 100% certainty your friend's claim". The logic of the two statements are very different. For statistical analysis, we almost always stick with the first one.

- The fourth choice is incorrect. While the observation $X \geq 3.514$ would be a rare event given that $X \sim \mathcal{N}(0,1)$, there is still some positive probability (roughly 0.02%) of it happening. Rare events can still occur, so we cannot rule out with 100% certainty that the distribution of X is $\mathcal{N}(0,1)$.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Review: Central Limit Theorem

1/1 point (graded)
Recall the central limit theorem states that if

- $X_1, \ldots, X_n$ are i.i.d.;
- $\mathbb{E}[X_1] = \mu < \infty$, and $\mathrm{Var}(X_1) = \sigma^2 < \infty$,

then a shift and a rescaling of the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to a standard Gaussian $\mathcal{N}(0,1)$ in distribution as $n \to \infty$:

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1).$$

Suppose $\mu = 0$ and $\sigma^2 = 1$. Given this assumption, which of the following limits is **strictly** between 0 and 1?

- ○ $\lim_{n\to\infty} P\left(\overline{X}_n \in (-1,1)\right)$
- ● $\lim_{n\to\infty} P\left(\overline{X}_n \in \left(-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)\right)$
- ○ $\lim_{n\to\infty} P\left(\overline{X}_n \in \left(-\frac{1}{n}, \frac{1}{n}\right)\right)$

✔

Solution:

Let $Z \sim \mathcal{N}(0,1)$ and let a_n, b_n denote sequences depending on n. By the central limit theorem (CLT),

$$\begin{aligned}\lim_{n\to\infty} P\left(\overline{X}_n \in (a_n, b_n)\right) &= \lim_{n\to\infty} P\left(\sqrt{n}\,\overline{X}_n \in (\sqrt{n}a_n, \sqrt{n}b_n)\right)\\ &= P\left(Z \in \left(\lim_{n\to\infty}\sqrt{n}a_n, \lim_{n\to\infty}\sqrt{n}b_n\right)\right)\end{aligned}$$

Now let's examine the choices in order.

- $\lim_{n\to\infty} P\left(\overline{X}_n \in (-1,1)\right) = 1$, so this choice is incorrect. Setting $a_n = -1$ and $b_n = 1$, we see that

  $$\lim_{n\to\infty}\sqrt{n}a_n = -\infty, \quad \lim_{n\to\infty}\sqrt{n}b_n = \infty.$$

  Hence, by the above calculation,

$$\lim_{n\to\infty} P\left(\overline{X}_n \in (a_n, b_n)\right) = P\left(Z \in (-\infty, \infty)\right) = 1.$$

- $\lim_{n\to\infty} P\left(\overline{X}_n \in \left(-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right)\right)$ lies strictly between 0 and 1, as we will show below. Setting $a_n = -\frac{1}{\sqrt{n}}$ and $b_n = \frac{1}{\sqrt{n}}$, we see that

$$\sqrt{n}a_n = -1, \quad \sqrt{n}b_n = 1.$$

Hence, by the above calculation,

$$\lim_{n\to\infty} P\left(\overline{X}_n \in (a_n, b_n)\right) = P\left(Z \in (-1, 1)\right)$$

Since Gaussian variables have a positive probability of being inside $(-1, 1)$ and also a positive probability of being outside $(-1, 1)$, we can also conclude without doing any computation that $0 < P\left(Z \in (-1, 1)\right) < 1$.

Remark: Alternatively we can compute, using computational tools or a table that

$$P\left(Z \in (-1, 1)\right) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \approx 0.6827.$$

- $\lim_{n\to\infty} P\left(\overline{X}_n \in \left(-\frac{1}{n}, \frac{1}{n}\right)\right) = 0$, so this choice is incorrect. Setting $a_n = -\frac{1}{n}$ and $b_n = \frac{1}{n}$, we see that

$$\lim_{n\to\infty} \sqrt{n}a_n = \lim_{n\to\infty} -\frac{1}{\sqrt{n}} = 0, \quad \lim_{n\to\infty} \sqrt{n}b_n = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0$$

Hence, by the above calculuation,

$$\lim_{n\to\infty} P(\overline{X}_n \in (a_n, b_n)) = P(Z \in (0,0)) = 0.$$

Remark: This exercise emphasizes the heuristic interpretation of the CLT which states that the sample mean $\overline{X}_n$ lives inside an interval of radius $Constant \times \frac{1}{\sqrt{n}}$ around its expectation. This heuristic will be useful for designing hypothesis tests.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Discussion

Hide Discussion

Topic: Unit 2 Foundation of Inference:Lecture 6: Introduction to Hypothesis Testing, and Type 1 and Type 2 Errors / 8. First Example

Add a Post

Show all posts ▼ by recent activity ▼

Interesting stuff
You know, as challenging as this class is, it's damn interesting. Thank you for making it this way Prof.
5