Lecture 4 practice

August 11, 2022

Identify the parameter

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(a) In a survey, one hundred college students are asked how many hours per week they spend on the Internet.

Mean. Each student reports a numerical value: a number of hours

Identify the parameter

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(b) In a survey, one hundred college students are asked: “What percentage of the time you spend on the Internet is part of your course work?”

Mean. Each student reports a number, which is a percentage, and we can average over these percentages

Identify the parameter

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(c) In a survey, one hundred college students are asked whether or not they cited information from Wikipedia in their papers.

(c) Proportion. Each student reports Yes or No, so this is a categorical variable and we use a proportion

Identify the parameter

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(d) In a survey, one hundred college students are asked what percentage of their total weekly spending is on alcoholic beverages.

(d) Mean. Each student reports a number, which is a percentage like in part (b).

Identify the parameter

For each of the following situations, state whether the parameter of interest is a mean or a proportion. It may be helpful to examine whether individual responses are numerical or categorical.

(e) In a sample of one hundred recent college graduates, it is found that 85 percent expect to get a job within one year of their graduation date.

(e) Proportion. Each student reports whether or not s/he expects to get a job, so this is a categorical variable and we use a proportion

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(a) What population is under consideration in the data set?

The sample is from all computer chips manufactured at the factory during the week of production. We might be tempted to generalize the population to represent all weeks, but we should exercise caution here since the rate of defects may change over time.

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(b) What parameter is being estimated?

The fraction of computer chips manufactured at the factory during the week of production that had defects.

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(c) What is the point estimate for the parameter?

Estimate the parameter using the data: $\hat{p} = 27 / 212 = 0.127$

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(d) What is the name of the statistic we use to measure the uncertainty of the point estimate?

Standard error (or SE)

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(e) Compute the value from part (d) for this context.

Compute the SE using p̂ = 0.127 in place of p:

$$SE \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.127(1-0.127)}{212}} = 0.023.$$

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(f) The historical rate of defects is 10%. Should the engineer be surprised by the observed rate of defects during the current week?

The standard error is the standard deviation of $\hat{p}$. A value of 0.10 would be about one standard error away from the observed value, which would not represent a very uncommon deviation. (**Usually beyond about 2 standard errors is a good rule of thumb.**) The engineer should not be surprised.

Quality control

As part of a quality control process for computer chips, an engineer at a factory randomly samples 212 chips during a week of production to test the current rate of chips with severe defects. She finds that 27 of the chips are defective.

(g) Suppose the true population value was found to be 10%. If we use this proportion to recompute the value in part (e) using p = 0.1 instead of p̂, does the resulting value change much?

Recomputed standard error using $p = 0.1\text{:}\ SE = \sqrt{\frac{0.1(1-0.1)}{212}} = 0.021$

This value isn't very different, which is typical when the standard error is computed using relatively similar proportions (and even sometimes when those proportions are quite different!)

Repeated water samples

A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(a) What is this distribution called?

Sampling distribution

Repeated water samples

A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

If the population proportion is in the 5-30% range, the success failure condition would be satisfied and the sampling distribution would be symmetric

Repeated water samples

A nonprofit wants to understand the fraction of households that have elevated levels of lead in their drinking water. They expect at least 5% of homes will have elevated levels of lead, but not more than about 30%. They randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(c) If the proportions are distributed around 8%, what is the variability of the distribution?

We use the formula for the standard error

$$SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.08(1-0.08)}{800}} = 0.0096$$

Repeated water samples

…they randomly sample 800 homes and work with the owners to retrieve water samples, and they compute the fraction of these homes with elevated lead levels. They repeat this 1,000 times and build a distribution of sample proportions.

(d) Suppose the researchers' budget is reduced, and they are only able to collect 250 observations per sample, but they can still collect 1,000 samples. They build a new distribution of sample proportions. How will the variability of this new distribution compare to the variability of the distribution when each sample contained 800 observations?

The distribution will tend to be more variable when we have fewer observations per sample

Chronic illness

In 2013, the Pew Research Foundation reported that “45% of U.S. adults report that they live with one or more chronic conditions”. However, this value was based on a sample, so it may not be a perfect estimate for the population parameter of interest on its own. The study reported a standard error of about 1.2%, and a normal model may reasonably be used in this setting.

(a) Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

Recall that the general formula is point estimate ± $z^\star$ × SE.

First, identify the three different values.

The point estimate is 45%, $z^\star = 1.96$ for a 95% confidence level, and SE = 1.2%

Chronic illness

(a) Create a 95% confidence interval for the proportion of U.S. adults who live with one or more chronic conditions. Also interpret the confidence interval in the context of the study.

Recall that the general formula is point estimate ± $z^{\star} \times$ SE.

First, identify the three different values.

The point estimate is 45%, $z^{\star} = 1.96$ for a 95% confidence level, and SE = 1.2%.

Then, plug the values into the formula: $45\% \pm 1.96 \times 1.2\% \rightarrow (42.6\%, 47.4\%)$

We are 95% confident that the proportion of US adults who live with one or more chronic conditions is between 42.6% and 47.4%.

Chronic illness

In 2013, the Pew Research Foundation reported that “45% of U.S. adults report that they live with one or more chronic conditions”, and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(a) We can say with certainty that the confidence interval from Exercise 5.7 contains the true percentage of U.S. adults who suffer from a chronic illness.

False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval “misses” about 5% of the time.

Chronic illness

In 2013, the Pew Research Foundation reported that “45% of U.S. adults report that they live with one or more chronic conditions”, and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(b) If we repeated this study 1,000 times and constructed a 95% confidence interval for each study, then approximately 950 of those confidence intervals would contain the true fraction of U.S. adults who suffer from chronic illnesses.

True. Notice that the description focuses on the true population value.

Chronic illness

In 2013, the Pew Research Foundation reported that “45% of U.S. adults report that they live with one or more chronic conditions”, and the standard error for this estimate is 1.2%. Identify each of the following statements as true or false. Provide an explanation to justify each of your answers.

(c) Since the standard error is 1.2%, only 1.2% of people in the study communicated uncertainty about their answer.

False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncertainty corresponding to individuals’ responses.

Website registration

A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(a) Check any conditions required for constructing a confidence interval.

The visitors are from a simple random sample, so independence is satisfied.

The success failure condition is also satisfied, with both 64 and 752 − 64 = 688 above 10.

Therefore, we can use a normal distribution to model $\hat{p}$ and construct a confidence interval.

Website registration

A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(b) Compute the standard error.

The sample proportion is p̂ = 64 / 752 = 0.085. The standard error is

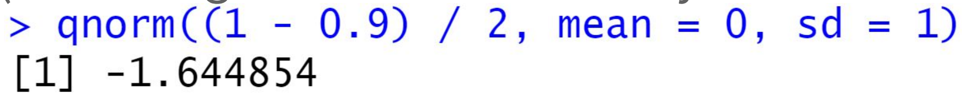
$$SE = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= \sqrt{\frac{0.085(1-0.085)}{752}} = 0.010$$

Website registration

A website is trying to increase registration for first-time visitors, exposing 1% of these visitors to a new site design. Of 752 randomly sampled visitors over a month who saw the new design, 64 registered.

(c) Construct and interpret a 90% confidence interval for the fraction of first-time visitors of the site who would register under the new design (assuming stable behaviors by new visitors over time).



```
> qnorm((1 - 0.9) / 2, mean = 0, sd = 1)
[1] -1.644854
```

For a 90% confidence interval, use $z^\star = 1.6449$.

The confidence interval is $0.085 \pm 1.6449 \times 0.010$ → (0.0683, 0.1017).

We are 90% confident that 6.83% to 10.17% of first-time site visitors will register using the new design.

Side effects of Avandia

Rosiglitazone is the active ingredient in the controversial type 2 diabetes medicine Avandia and has been linked to an increased risk of serious cardiovascular problems such as stroke, heart failure, and death. A common alternative treatment is pioglitazone, the active ingredient in a diabetes medicine called Actos. In a nationwide retrospective observational study of 227,571 Medicare beneficiaries aged 65 years or older, it was found that 2,593 of the 67,593 patients using rosiglitazone and 5,386 of the 159,978 using pioglitazone had serious cardiovascular problems.

Side effects of Avandia

(a) What are the claims being tested?

H_0: The treatment and cardiovascular problems are independent. They have no relationship, and any difference in incidence rates between the rosiglitazone and pioglitazone groups is due to chance.

H_A: The treatment and cardiovascular problems are not independent. Any difference in the incidence rates between the rosiglitazone and pioglitazone groups is not due to chance and rosiglitazone is associated with an increased risk of serious cardiovascular problems.

Side effects of Avandia

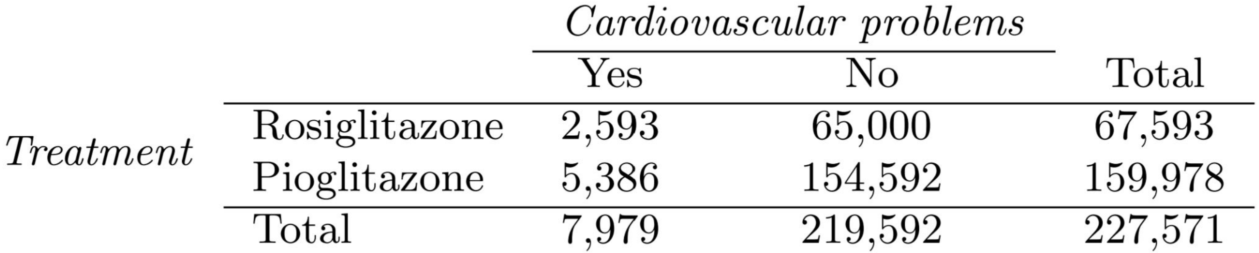
(b) If we were to find that the proportion of people who took Avandia and experience cardiovascular problems was 4 times as high as that for people who took Actos, would this be evidence for the null or alternative hypothesis?

This would be evidence in support of the alternative hypothesis, which might lead us to reject the null hypothesis.

(c) If we were to find that the proportion of people who took Avandia and experience cardiovascular problems was approximately the same as that for people who took Actos, would this be evidence for the null or alternative hypothesis?

Neither, we are always looking for evidence to support the alternative hypothesis, here we would say we fail to find evidence to support the alternative hypothesis. We would say we fail to reject the null hypothesis.

Side effects of Avandia

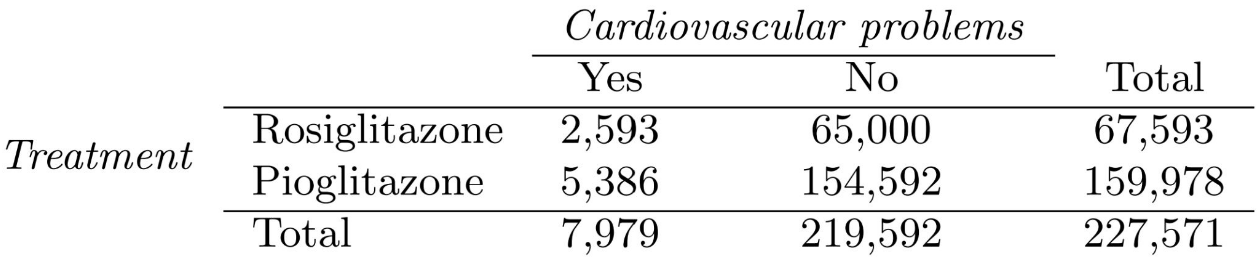


		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(d) Determine if each of the following statements is true or false. If false, explain why.

Be careful: The reasoning may be wrong even if the statement's conclusion is correct. In such cases, the statement should be considered false.

Side effects of Avandia

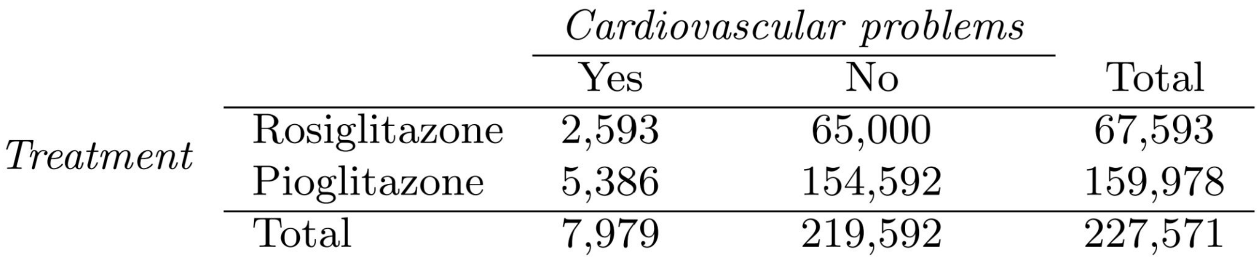


		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

i. Since more patients on pioglitazone had cardiovascular problems (5,386 vs. 2,593), we can conclude that the rate of cardiovascular problems for those on a pioglitazone treatment is higher.

False. Instead of comparing counts, we should compare percentages of people in each group who suffered cardiovascular problems.

Side effects of Avandia

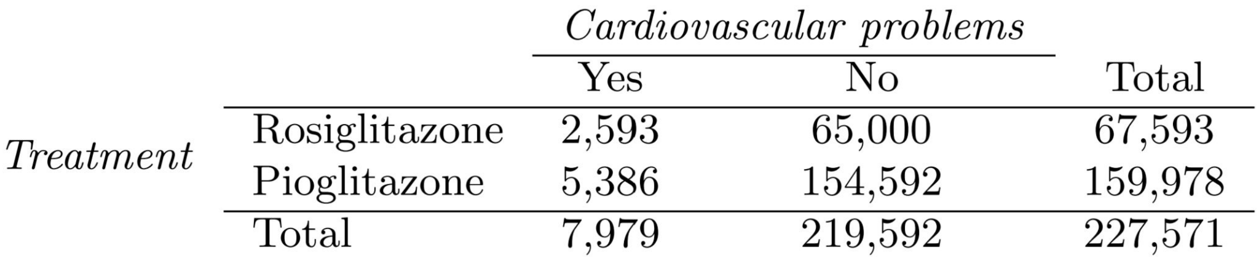


		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

ii. The data suggest that diabetic patients who are taking rosiglitazone are more likely to have cardiovascular problems since the rate of incidence was (2,593 / 67,593 = 0.038) 3.8% for patients on this treatment, while it was only (5,386 / 159,978 = 0.034) 3.4% for patients on pioglitazone.

True

Side effects of Avandia



		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

iii. The fact that the rate of incidence is higher for the rosiglitazone group proves that rosiglitazone causes serious cardiovascular problems.

False. Association does not imply causation. We cannot infer a causal relationship based on an observational study. The difference from part (ii) is subtle.

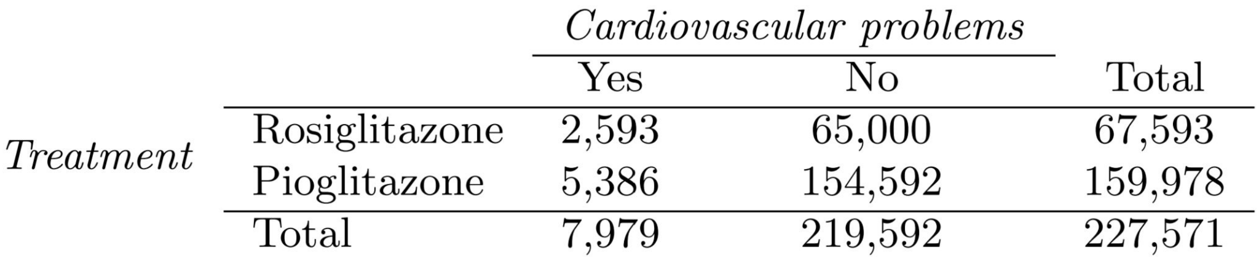
Side effects of Avandia

		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

iv. Based on the information provided so far, we cannot tell if the difference between the rates of incidences is due to a relationship between the two variables or due to chance.

True

Side effects of Avandia



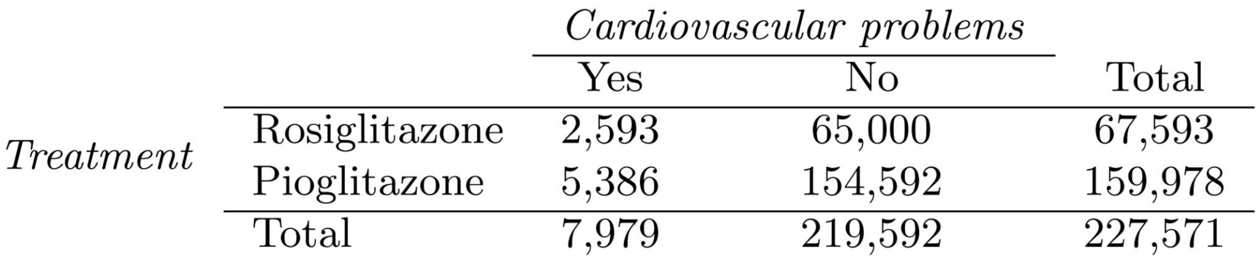
		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(e) What proportion of all patients had cardiovascular problems?

Proportion of all patients who had cardiovascular problems:

7,979 / 227,571 ≈ 0.035

Side effects of Avandia



		Cardiovascular problems		
		Yes	No	Total
Treatment	Rosiglitazone	2,593	65,000	67,593
	Pioglitazone	5,386	154,592	159,978
	Total	7,979	219,592	227,571

(f) If the type of treatment and having cardiovascular problems were independent, about how many patients in the rosiglitazone group would we expect to have had cardiovascular problems?

The expected number of heart attacks in the rosiglitazone group, if having cardiovascular problems and treatment were independent, can be calculated as the number of patients in that group multiplied by the overall cardiovascular problem rate in the study: 67, 593 ∗ 7,979 / 227,571 ≈ 2370.

Credits

Examples adapted from OpenIntro Statistics (4th edition) by David Diez, Mine Cetinkaya-Rundel, and Christopher D Barr https://www.openintro.org/book/os/ protected under the Creative Commons License